РОССИЙСКАЯ АКАДЕМИЯ НАУК

ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Журнал основан в январе 1965 г. **ISSN: 0555-2923** **Выходит 4 раза в год**

Том 58, 2022 **Вып. 2**

Апрель–Май–Июнь

М о с к в а

С О Д Е Р Ж А Н И Е

CONTENTS

ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 58 **2022** **Вып. 2**

УДК 621.391 : 536.75 : 519.722

© 2022 г. **В.А. Зорич**

ЭНТРОПИЯ В ТЕРМОДИНАМИКЕ И В ТЕОРИИ ИНФОРМАЦИИ

Обсуждается связь понятий энтропии в термодинамике и в теории информации.

Ключевые слова: энтропия, второе начало термодинамики, энтропия Больцмана, энтропия Шеннона, демон Максвелла.

DOI: 10.31857/S0555292322020016, **EDN:** DYXSYV

§ 1. Введение

Существует легенда, согласно которой Шеннон обратился к фон Нейману за советом, каким термином назвать то, что теперь называется информационной энтропией, или энтропией Шеннона. Человек широкого кругозора и глубины, фон Нейман прозорливо предложил термин *энтропия*.

Этот термин появился в классической равновесной термодинамике. Он был введен в 1865 году Клаузиусом в связи с так называемым вторым началом термодинамики, началом Карно – Клаузиуса. Функция энтропии стала одной из ключевых характеристик равновесного состояния термодинамической системы.

Огромный шаг к пониманию физического смысла энтропии был сделан Больцманом, одним из творцов статистической термодинамики.

Основные функции феноменологической термодинамики (температура, давление, ...) на самом деле являются функциями огромного числа переменных (молекул, атомов). Значения таких функций, именно по этой самой причине, оказываются почти постоянными и устойчивыми с точки зрения наблюдателя, измеряющего интегральные характеристики равновесного состояния системы, например, газа, а не детали микросостояния всей системы составляющих газ молекул.

Одному наблюдаемому макросостоянию системы может отвечать огромное множество ее микросостояний, а равновесному макросостоянию системы вообще отвечает подавляющее число всех ее возможных микросостояний.

Принцип Больцмана, или *идея Больцмана*, сопоставить классической энтропии S логарифм вероятности соответствующего состояния многочастичной термодинамической системы, в контексте известных свойств классической энтропии, диктуется уже следующим.

Вероятности независимых событий перемножаются, а логарифм даст аддитивность.

Тенденция переходить от менее вероятного состояния к более вероятному соответствует тенденции роста классической энтропии состояния при эволюции изолированной термодинамической системы к равновесию.

Максимум такой энтропии соответствует равновесному состоянию системы. Около него концентрируется почти все множество микросостояний многочастичной си-

стемы (типа молекулярного газа), отвечающих наблюдаемому равновесному макросостоянию системы!

Вот *формула Больцмана* (выгравированная на его могиле):

$$S = k \log W.$$

Здесь k – размерная постоянная (отношение R/N_A газовой постоянной и числа Авогадро), названная Планком *постоянной Больцмана*, а W – целочисленная величина, пропорциональная вероятности рассматриваемого термодинамического состояния, которую, следуя Планку, называют *статистическим весом состояния*.

Конкретнее, W – это то максимально возможное количество микросостояний, которое отвечает рассматриваемому равновесному состоянию многочастичной термодинамической системы с энтропией S.

Поясним сказанное примером. Часть выкладок этого примера мы используем ниже для демонстрации прямой связи информационной энтропии Шеннона и термодинамической энтропии Больцмана.

§ 2. Один пример

Следуя Шрёдингеру [1], рассмотрим ансамбль из N одинаковых, но перенумерованных систем, каждая из которых может находиться в одном из перенумерованных состояний $1, 2, \ldots, \ell$. Пусть $\varepsilon_1 \leqslant \varepsilon_2 \leqslant \ldots \leqslant \varepsilon_\ell$ – значения энергии индивидуальной системы в этих состояниях, а $a_1, a_2, \ldots, a_\ell$ – количество систем ансамбля, находящихся в состояниях $1, 2, \ldots, \ell$ соответственно.

Такой набор $a_1, a_2, \ldots, a_\ell$ может реализоваться многими способами. А точнее, число способов равно $\binom{N}{a_1}\binom{N-a_1}{a_2}\ldots\binom{N-a_1-\ldots-a_{\ell-1}}{a_\ell}$, т.е.

$$G = \frac{N!}{a_1!\,a_2!\ldots a_\ell!}.$$

Совокупность чисел $a_1, a_2, \ldots, a_\ell$ должна удовлетворять условиям

$$\sum_i a_i = N, \quad \sum_i \varepsilon_i a_i = E,$$

где E – полная совокупная энергия систем ансамбля.

Будем искать максимальное значение $W = \max G$ (или $\ln W = \max \ln G$) при указанных ограничениях, что даст нам наиболее вероятный набор $a_1, a_2, \ldots, a_\ell$ чисел заполнения.

(Напомним, что в интересных для термодинамики случаях, когда число N и количество возможных уровней энергии ε_i очень велики, имеет место *явление концентрации*. Можно показать, что общее число всех возможных при наших условиях состояний ансамбля почти совпадает с максимальным значением G, которое мы намерены искать. Значит, мы действительно найдем и наиболее вероятный набор $a_1, a_2, \ldots, a_\ell$ чисел заполнения.)

При больших значениях n по формуле Стирлинга $n! \simeq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$. Поэтому можно считать что

$$\ln(n!) \approx n(\ln n - 1) \approx n \ln n.$$

На самом деле нам потом будет достаточно этого замечания и приведенного выше выражения для статистического веса G, но доведем этот красивый и полезный пример до конца.

A. Действуя методом Лагранжа отыскания условного экстремума, записав, что

$$\sum_i \ln a_i da_i + \lambda \sum_i da_i + \nu \sum_i \varepsilon_i da_i = 0,$$

находим, что при любом i выполняется равенство

$$\ln a_i + \lambda + \nu\varepsilon_i = 0,$$

и значит,

$$a_i = e^{-\lambda - \nu\varepsilon_i},$$

причем λ и ν подчинены условиям

$$\sum_i e^{-\lambda-\nu\varepsilon_i} = N, \quad \sum_i \varepsilon_i e^{-\lambda-\nu\varepsilon_i} = E.$$

B. Обозначая через $E/N = U$ среднюю энергию, приходящуюся на одну систему ансамбля, запишем полученный результат в следующем виде:

$$\frac{E}{N} = U = \frac{\sum\limits_i \varepsilon_i e^{-\nu\varepsilon_i}}{\sum\limits_i e^{-\nu\varepsilon_i}} = -\frac{\partial}{\partial\nu} \ln \sum_i e^{-\nu\varepsilon_i},$$

$$a_i = N\frac{e^{-\nu\varepsilon_i}}{\sum\limits_i e^{-\nu\varepsilon_i}} = -\frac{N}{\nu}\frac{\partial}{\partial\varepsilon_i} \ln \sum_i e^{-\nu\varepsilon_i}.$$

Дополнительные рассмотрения, выясняющие физический смысл величины ν в термодинамической ситуации, приводят к тому, что

$$\nu = \frac{1}{kT},$$

где k – постоянная Больцмана, а T – абсолютная температура.

C. Возникает, как мы уже понимаем, важная величина, называемая *статистической суммой*:

$$Z = \sum_i e^{-\frac{\varepsilon_i}{kT}}.$$

Теперь можно написать, как именно числа заполнения a_i распределены по энергетическим уровням при данной абсолютной температуре:

$$a_i = N\frac{e^{-\frac{\varepsilon_i}{kT}}}{Z}.$$

Величина

$$Z^{-1}e^{-\frac{\varepsilon_i}{kT}}$$

– это вероятность малой системе оказаться в состоянии с энергией ε_i.

Мы получили *каноническое распределение Гиббса*

$$\exp\left(\frac{\psi - \varepsilon_i}{kT}\right).$$

В рассмотренной ситуации оно показывает, как распределены по энергиям ε_i малые системы (типа молекул) в термостате (большой системе, типа газа, находящейся

в равновесном термодинамическом состоянии при температуре T). Здесь введено обозначение $e^{\frac{\psi}{kT}} = Z^{-1}$.

Если найден статистический вес $W = \max G$ рассматриваемого равновесного состояния термодинамической системы, то, следуя формуле Больцмана $S = k \ln W$, можно найти энтропию S системы в этом состоянии.

§ 3. Информационная энтропия

3.1. Информация и ее количественное описание. Если случайная величина может с равной вероятностью принимать одно из двух возможных значений, например, 0 или 1, то указание конкретного состояния такой величины есть информация, которую принято называть *битом* информации. Мы напомнили определение бита информации (и неопределенности).

Вектором длины n из нулей и единиц можно задать 2^n различных объектов.

Если имеется M равновозможных значений случайной величины, или M равновероятных событий, то указание одного такого события требует $\log_2 M$ бит информации.

Если вероятность события p, то оно встречается один раз среди $M = \dfrac{1}{p}$ событий. Значит, его неопределенность, измеренная в битах, есть

$$\log_2 M = \log_2 \frac{1}{p} = -\log_2 p.$$

3.2. Энтропия случайной величины. Функцию со случайными значениями в математике принято называть *случайной величиной*. Пусть X – случайная величина. Средняя на одно значение случайной величины X (на одно событие) неопределенность в битах – это математическое ожидание

$$\mathbf{E}\, X = -\sum_{i=1}^{M} p_i \log_2 p_i$$

случайной величины X. Это математическое ожидание, вслед за Шенноном [2], называют *энтропией случайной величины* X и обозначают символом

$$H(X) = -\sum_{i=1}^{M} p_i \log_2 p_i$$

(греч. εντροπία, от др.-греч. ἐν «в» + τροπή «превращение; обращение»). Как уже было сказано, термин *энтропия* исходно появился в термодинамике.

Если единичное событие системы X имело вероятность появления p, то его осуществление несет $-\log_2 p$ бит информации. Но длительное наблюдение за появлениями такого события в единицу времени приносит всего $-p \log_2 p$ бит информации. Это же относится и к среднему числу бит информации $H(X) = -\sum\limits_{i=1}^{M} p_i \log_2 p_i$, которое за единицу времени приносит появление одного из событий системы X (значений случайной величины X). В этом смысл информационной энтропии.

Укажем сразу, в чем состоит статистический характер энтропии $H(X)$. Мы этим воспользуемся ниже.

Пусть X – произвольная дискретная случайная величина, которая может принимать M различных значений x_i с вероятностями p_i соответственно. Тогда для любых положительных чисел ε, δ найдется такое число $n_{\varepsilon\delta}$, что при $n \geqslant n_{\varepsilon\delta}$ выполняется

неравенство

$$\Pr\left\{\left|-\frac{1}{n}\sum_{i=1}^{n}\log_2 p_{x_i} - H(X)\right| < \delta\right\} > 1-\varepsilon, \tag{1}$$

где, как обычно, $\Pr\{\cdot\}$ – вероятность указанного в скобках события, но теперь x_i, $i = 1, \ldots, n$, – это n независимых значений случайной величины X, а p_{x_i} – вероятности этих значений.

Как связана энтропия с кодированием?

Рассмотрим сообщения-слова-векторы $\bar{x} = (x_1, \ldots, x_n)$, образованные n последовательными независимыми значениями случайной величины X. Вероятность $p_{\bar{x}}$ появления слова $\bar{x}$ равна $p_{\bar{x}} = p_{x_1} \ldots p_{x_n}$. В силу соотношения (1) при $n \geqslant n_{\varepsilon\delta}$ с вероятностью, большей чем $1-\varepsilon$, будем иметь

$$2^{-n(H(X)+\delta)} \leqslant p_{\bar{x}} \leqslant 2^{-n(H(X)-\delta)}. \tag{2}$$

Слово $\bar{x}$ называют *δ-типичным*, если для него выполнены эти оценки. Ясно, что существует не более $2^{n(H(X)+\delta)}$ таких δ-типичных слов, а если $n \geqslant n_{\varepsilon\delta}$, то их еще и не меньше чем $(1-\varepsilon)2^{n(H(X)-\delta)}$, и при этом все множество не δ-типичных слов имеет вероятность, не большую чем ε.

В принципе теперь уже можно использовать двоичные последовательности длины $n(H(X)+\delta)$, чтобы закодировать все δ-типичные слова. Даже если все остальные слова закодировать одним символом, вероятность ошибки при передаче слов $\bar{x}$ длины n, вызванная таким кодом, будет меньше ε.

С другой стороны (и это эффект неустойчивости экономных кодов), любой код, использующий в той же ситуации двоичные последовательности относительно чуть меньшей длины $n(H(X)-\delta)$ (например, $2\delta n$ из $n(H(X)+\delta)$ посланных символов потерялось в шуме), будет иметь асимптотически неисчезающую вероятность ошибки, стремящуюся к единице при $n \to +\infty$.

Итак, связь энтропии и кодирования информации состоит, например, в том, что эффективное кодирование асимптотически при $n \to +\infty$ требует $N \sim 2^{nH(X)}$ слов, и энтропия $H(X)$ может интерпретироваться как мера количества информации в битах на передаваемый символ, т.е. на одно значение случайной величины X.

Отсюда, в частности, следует, что энтропия источника информации не должна превышать пропускную способность канала связи, если мы хотим адекватно и без задержек передавать поступающую информацию по этому каналу связи.

Если передается M сообщений и на передачу каждого затрачивается время T секунд, то это равносильно тому, что скорость передачи в битах в секунду равна $\frac{1}{T}\log_2 M$.

Теперь посмотрим на соотношение (1) с точки зрения макро- и микросостояний. Пусть $n \gg 1$. Величина

$$-\frac{1}{n}\sum_{i=1}^{n}\log_2 p_{x_i}$$

для подавляющего числа векторов $\bar{x} = (x_1, \ldots, x_n)$ почти одна и та же, $H(X)$. Конкретному состоянию (значению, макросостоянию) величины $-\frac{1}{n}\sum\limits_{i=1}^{n}\log_2 p_{x_i}$ отвечает много микросостояний $\bar{x} = (x_1, \ldots, x_n)$.

И не просто много, а когда $n \gg 1$, почти каждое случайно взятое микросостояние $\bar{x} = (x_1, \ldots, x_n)$ реализует почти одно и то же состояние (значение) $H(X)$ наблюдаемой макровеличины.

На языке термодинамики и в терминах энтропии Больцмана это означает, что система, допускающая огромное множество микросостояний $\{\bar{x} = (x_1, \ldots, x_n)\}$, находится в равновесном макросостоянии.

§ 4. Энтропия в термодинамике и информатике

Теперь можно конкретизировать сказанное и сопоставить энтропию равновесного состояния термодинамической системы и энтропию случайной величины.

Мы увидим, что энтропия Шеннона с точки зрения математики есть просто отнесенная к одной частице (удельная) энтропия Больцмана.

Что в случае случайной величины X должно отвечать равновесному состоянию, к которому со временем эволюционирует любая изолированная термодинамическая система? По-видимому, просто надо достаточно долго наблюдать значения, которые принимает случайная величина X. Длинный вектор таких значений стабилизируется в следующем смысле. В нем каждое значение x_i случайной величины X будет присутствовать в количестве a_i, пропорциональном вероятности p_i значения x_i случайной величины X и длине N вектора значений (времени наблюдения): $a_i \simeq p_i N$.

Сами векторы $\bar{x} = (x_1, \ldots, x_N)$ большой длины N, скорее всего, будут различными, но подавляющая часть всех таких векторов будет иметь указанную статистику $a_i \simeq p_i N$ присутствия каждого отдельного значения x_i случайной величины X, если таких значений конечное число.

Количество таких векторов находится по уже известной нам формуле:

$$G = \frac{N!}{a_1! \, a_2! \ldots a_i! \ldots}.$$

Берем $\log G$. Это, с точностью до размерной постоянной (постоянной Больцмана), энтропия $\log W$ равновесного состояния по Больцману. Воспользовавшись формулой Стирлинга и уже проделанными выше выкладками, с учетом соотношений $a_i \simeq p_i N$ найдем

$$\begin{aligned}
&\log G \simeq N(\log N - 1) - \sum_i a_i(\log a_i - 1) = \\
&= N \log N - \sum_i a_i \log a_i = N \log N - \sum_i p_i N \log(p_i N) = \\
&= N \log N - \sum_i p_i N \log p_i - \log N \sum_i p_i N = \\
&= N \log N - \sum_i p_i N \log p_i - N \log N \sum_i p_i = \\
&= N \log N - \sum_i p_i N \log p_i - N \log N = -N \sum_i p_i \log p_i.
\end{aligned}$$

Приведя здесь безразмерную энтропию Больцмана к единичному элементу (в данном случае к одному значению случайной величины, поделив на N), получаем энтропию Шеннона.

Итак, энтропия Шеннона с точки зрения математики есть просто отнесенная к одной частице (удельная) энтропия Больцмана.

Заметим, что изменение основания b логарифма в определении Шеннона энтропии $H(X) = -\sum_i p_i \log_b p_i$ случайной величины X приводит лишь к появлению общего множителя – коэффициента, что равносильно переходу к соответствующему масштабу единицы измерения.

§ 5. Небольшой комментарий

Выражения

$$-\sum_i p_i \log p_i \qquad \text{и} \qquad -\int p \log p$$

в связи с энтропией в термодинамике появились, конечно, задолго до того, как они появились в теории информации. Сама теория информации как наука значительно моложе термодинамики.

На рубеже XIX–XX веков Гиббс [3] обогатил статистическую термодинамику замечательной общей математической моделью гамильтоновой динамической системы, наделенной инвариантной вероятностной мерой, отвечающей равновесному состоянию многочастичной термодинамической системы.

Равновесному состоянию изолированной термодинамической системы отвечает экстремум энтропии. Если вероятностная мера в фазовом пространстве распределена с плотностью p, то поиск плотности p, отвечающей равновесию системы, приводит, например, к задаче поиска экстремума, взятого по всему пространству интеграла $\int p \log p$ при двух естественных условиях: интеграл от p по всему пространству равен единице, и полная энергия E всей гамильтоновой системы постоянна. Дискретный вариант этой вариационной задачи мы рассмотрели выше и пришли к дискретному варианту канонического распределения Гиббса. Истоки такого распределения восходят к Максвеллу и Больцману.

§ 6. Демон Максвелла и энтропия

Мы здесь всего лишь привели математическую выкладку, демонстрирующую связь энтропии Больцмана статистической термодинамики и энтропии Шеннона информатики. Мы почти не касались сути понятия *энтропия*. Хотя бы вскользь, все же надо сказать, что физики довольно тщательно анализировали, обсуждали и продолжают обсуждать различные аспекты понятия энтропии. Не ушел от их внимания и информационный аспект термодинамической энтропии.

Еще в 1867 году Максвелл предложил свой знаменитый мысленный эксперимент, парадокс Максвелла, получивший название “демон Максвелла” (публикация 1871 года). Этот мысленный эксперимент, описанный почти во всех учебниках физики, заставил физиков глубже разобраться в двух фундаментальных принципах термодинамики.

Сначала причину парадокса искали в нарушении принципа сохранения энергии (первого начала термодинамики). Но постепенно стало ясно, что дело связано со вторым началом термодинамики и проистекающим из него принципом неубывания энтропии. Когда в 1929 году, анализируя парадокс Максвелла, Лео Силард (тот самый, кто стимулировал Эйнштейна подписать письмо Рузвельту, инициировавшее разработку атомной бомбы в США) мысленно построил свой двигатель, работавший на газе из одной молекулы, стало понятно, что дело в энтропии и, более того, в спрятанной в ней информации. Считается, что к 1984 году парадокс Максвелла был разрешен. Решение пришло как раз через информационный аспект термодинамической энтропии (см. [4, 5]).

Поясним таким сравнением. Если Архимед говорил: “дайте мне точку опоры, и я подниму Землю”, то, например, Беннет и Фейнман могли бы сказать: “дайте мне бесконечную память (для хранения информации), и я построю вам и демона Максвелла, и вечный двигатель второго рода, нарушающий второе начало термодинамики”. Ленточный двигатель Беннета, работающий как машина Тьюринга и, благодаря информации, без потерь перерабатывающий тепло в работу, представлен в [5, с. 146].

Статистическая физика объяснила бóльшую часть феноменологической термодинамики исходя из более фундаментальных законов, описывающих поведение многочастичных систем (например, молекулярного газа). В частности, как уже было сказано, Больцман дал статистическое описание понятия термодинамической энтропии S.

Позднее Бриллюэн (см., например, [6]) уже прямо связал термодинамическую энтропию с информационной энтропией следующим соотношением: $\Delta S \geqslant \Delta I$. Здесь ΔI – информация[1], полученная в результате измерения, проведенного посредством какого-то прибора, а ΔS – рост энтропии термодинамического состояния прибора в результате этого измерения. Если $T\Delta S$ через тепло характеризует соответствующие затраты энергии, то соотношение $T\Delta S \geqslant T\Delta I$ показывает, что чем тоньше (точнее) измерение (т.е. чем больше полученная в результате измерения информация ΔI), тем бóльших энергетических затрат оно требует. Это уже некоторая физическая предпосылка к фундаментальному принципу неопределенности, подробно рассматриваемому в квантовой механике.

И последнее.

Классическая термодинамика трудами Пуанкаре, Каратеодори, Борна приобрела вид достаточно последовательной в математическом отношении теории. Трудами Максвелла, Больцмана, Гиббса, Планка, Эйнштейна были созданы основы статистической термодинамики. Аксиоматизация термодинамики продолжается и сейчас (ср. [7] и, например, работу [8], посвященную формализации второго начала термодинамики и энтропии).

Возвращаясь к информационной энтропии, добавим, что Шеннон уже в своей фундаментальной для теории информации работе [2] предложил набор из трех естественных свойств – аксиом информационной энтропии, которые с неизбежностью приводят к определяющей ее формуле

$$H(X) = -\sum_{i=1}^{M} p_i \log p_i.$$

§ 7. Эпилог

Открывая сборник статей о турбулентности [9], академик О.М. Белоцерковский вспоминает, что, когда он учился на физическом факультете Московского университета, лекции по электричеству им читал профессор С.Г. Калашников, который на первой же лекции рассказал следующее. Однажды на экзамене он (Калашников) спросил студента: “Что такое электричество?” Студент заерзал, засуетился и говорит: “Вот еще вчера знал, а забыл”. На это Калашников сокрушенно заметил: “Был один человек, который знал, и тот забыл!”

Ситуация с энтропией примерно такая же.

СПИСОК ЛИТЕРАТУРЫ

1. *Шрёдингер Э.* Лекции по физике. М., Ижевск: Изд-во РХД, 2001.
2. *Shannon C.E.* A Mathematical Theory of Communication // Bell Syst. Tech. J. 1948. V. 27. № 3. P. 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x (Русск. перевод в *Шеннон К.* Работы по теории информации и кибернетике. М.: Изд-во иностр. лит., 2002.).
3. *Гиббс Дж.В.* Термодинамика. Статистическая механика. М.: Наука, 1982.

[1] В единицах термодинамической энтропии, т.е. I получается из информационной энтропии H умножением на постоянную Больцмана: $I = kH$.

4. *Беннет Ч.Г.* Демоны, двигатели и второе начало термодинамики // В мире науки. 1988. № 1. С. 52–60.
5. *Feynman R.P.* Feynman Lectures on Computation. Reading, MA: Addison-Wesley, 1996.
6. *Бриллюэн Л.* Теория информации и ее приложение к фундаментальным проблемам физики // Развитие современной физики. Сб. статей. М.: Наука, 1964. С. 324–329.
7. Развитие современной физики. Сб. статей / под ред. Б.Г. Кузнецова. М.: Наука, 1964.
8. *Lieb E.H., Yngvason J.* The Mathematics of the Second Law of Thermodynamics // Visions in Mathematics: GAFA 2000 Special Volume, Part I. Basel: Birkhäuser, 2010. P. 334–358. https://doi.org/10.1007/978-3-0346-0422-2_12
9. Этюды о турбулентности. Сб. статей. М.: Наука, 1994.

Зорич Владимир Антонович
Московский государственный университет
им. М. В. Ломоносова,
механико-математический факультет
vzorich@gmail.com

Поступила в редакцию
13.05.2022
После доработки
13.05.2022
Принята к публикации
18.05.2022

ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 58 **2022** **Вып. 2**

УДК 621.391 : 519.724.2

© 2022 г. **М. Ковачевич**

О МАКСИМАЛЬНОМ ЧИСЛЕ РАЗЛИЧИМЫХ СТРОК ПОД ДЕЙСТВИЕМ КОРОТКИХ ТАНДЕМНЫХ ДУПЛИКАЦИЙ[1]

Множество всех q-ичных строк, не содержащих повторяющихся подстрок длины $\leqslant 3$ (т.е. не содержащих подстрок вида aa, $abab$ и $abcabc$), образует код, исправляющий произвольное количество мутаций типа тандемных дупликаций длины $\leqslant 3$. Иными словами, любые две такие строки различимы в том смысле, что из них не могут образоваться одинаковые строки под действием некоторого количества тандемных дупликаций длины $\leqslant 3$. Показано, что этот код асимптотически оптимален по скорости, т.е. представляет собой максимальное множество различимых строк с точностью до субэкспоненциального множителя. Этот результат дает решение задачи о пропускной способности с нулевой ошибкой для последнего оставшегося случая каналов с тандемными дупликациями, удовлетворяющих свойству "единственности корневых строк".

Ключевые слова: тандемная дупликация, повторение тандема, ошибка дупликации, ошибка повторения, вклейка, хранение информации на основе ДНК, исправление ошибок, пропускная способность с нулевой ошибкой, код с ограничением, строка без повторений.

DOI: 10.31857/S0555292322020028, **EDN:** DZAPUX

§ 1. Введение

Тандемные дупликации – это ошибки типа "вклейки", естественным образом появляющиеся как мутации ДНК-строк и являющиеся поэтому потенциальным источником дефектов, возникающих в системах хранения информации на основе ДНК в живых организмах [1]. В то время как задача исправления тандемных дупликаций фиксированной и известной длины ℓ хорошо изучена, причем как при конечном [2, 3], так и при бесконечном [1, 4] числе ошибок, в отношении (по-видимому, гораздо более практически значимой) задачи исправления дупликаций переменной длины известно гораздо меньше. В частности, оптимальные коды, исправляющие все конфигурации дупликаций длины $\leqslant \ell$, были найдены только в специальных случаях $\ell = 1$ и $\ell = 2$ [1, теорема 32].

Нашим основным результатом является доказательство того факта, что аналогичная конструкция кодов, исправляющих неограниченное число тандемных дупликаций длины $\leqslant 3$ [1, теорема 27], также является асимптотически оптимальной по скорости. Этот результат дает решение задачи о пропускной способности с нулевой ошибкой для каналов с тандемными дупликациями во всех случаях, когда корневые строки (относительно дупликаций) любых строк единственны [1, теорема 40]. Однако при бóльших значениях ℓ корневые строки относительно дупликаций длины $\leqslant \ell$ перестают быть единственными [1, теорема 40], и следовательно, для решения за-

[1] Работа выполнена при поддержке исследовательско-инновационной программы Horizon 2020 Европейского Союза (номер гранта 856967), а также Секретариата по высшему образованию и научным исследованиям автономной провинции Воеводина, Сербия (номер проекта 142-451-2686/2021).

дачи о пропускной способности с нулевой ошибкой и смежных с ней задач в таких моделях понадобятся другие конструкции и верхние границы[2].

Помимо теоретико-информационных и кодовых вопросов упомянутого типа в литературе рассматривались и некоторые другие задачи, связанные с моделями с тандемными дупликациями переменной длины (см., например, [6–8]).

1.1. Описание модели. Для q-ичного алфавита будем использовать обозначение $\mathcal{A}_q := \{0, 1, \ldots, q-1\}$, а множество всех строк (или слов) над алфавитом $\mathcal{A}_q$ обозначим через $\mathcal{A}_q^* := \bigcup_{n=0}^{\infty} \mathcal{A}_q^n$. Длина строки $\boldsymbol{x} = x_1 \ldots x_n \in \mathcal{A}_q^n$ обозначается через $|\boldsymbol{x}| = n$. Строка, полученная конкатенацией двух строк $\boldsymbol{u}$ и $\boldsymbol{v}$, записывается в виде $\boldsymbol{uv}$. Строка $\boldsymbol{v}$ называется подстрокой (или фрагментом) строки $\boldsymbol{x}$, если существуют (возможно, пустые) строки $\boldsymbol{u}$ и $\boldsymbol{w}$, такие что $\boldsymbol{x} = \boldsymbol{uvw}$.

Пусть ℓ – фиксированное натуральное число. В *канале с тандемными* $(\leqslant \ell)$*-дупликациями* передаваемая по каналу строка $\boldsymbol{x}$ подвергается последовательному воздействию некоторого числа тандемных дупликаций длины $\leqslant \ell$ каждая, где тандемная дупликация длины k – это вставка точной копии подстроки длины k рядом с исходной подстрокой (неважно, вставляется ли она слева или справа от исходной, поскольку обе операции приводят к одной и той же новой строке). Предполагается, что число возникающих дупликаций заранее не известно ни передатчику, ни приемнику и может принимать любое значение из множества натуральных чисел $\{0, 1, 2, \ldots\}$. Более точно, канал описывается следующим образом:

- Вход: $\boldsymbol{x} \equiv \boldsymbol{x}^{(0)}$;
- Из множества $\{0, 1, 2, \ldots\}$ выбирается число t (количество дупликаций);
- Для $i = 1, \ldots, t$ повторяются следующие действия:
  - Из множества $\{1, \ldots, |\boldsymbol{x}^{(i-1)}|\}$ произвольным образом выбирается позиция дупликации j в строке $\boldsymbol{x}^{(i-1)}$;
  - Из множества $\{1, \ldots, \min\{j, \ell\}\}$ произвольным образом выбирается длина дупликации k;
  - Строка $\boldsymbol{x}^{(i)}$ получается вставкой копии подстроки $x_{j-k+1}^{(i-1)} \ldots x_j^{(i-1)}$ рядом с исходной подстрокой в $\boldsymbol{x}^{(i-1)}$, т.е.

$$\boldsymbol{x}^{(i)} = x_1^{(i-1)} \ldots \overline{x_{j-k+1}^{(i-1)} \ldots x_j^{(i-1)}}\, \underline{x_{j-k+1}^{(i-1)} \ldots x_j^{(i-1)}}\, x_{j+1}^{(i-1)} \ldots x_{|\boldsymbol{x}^{(i-1)}|}^{(i-1)},$$

    где исходная дуплицируемая подстрока выделена чертой сверху, а вставленная копия – чертой снизу;
- Выход: $\boldsymbol{y} \equiv \boldsymbol{x}^{(t)}$.

Всюду далее будем считать, что $\ell = 3$.

Пример 1. Примером того, как канал действует на передаваемую строку $\boldsymbol{x} \in \mathcal{A}_3^8$, является следующий список строк, каждая строка в котором получена из предыдущей путем тандемной дупликации длины $\leqslant 3$:

$$\boldsymbol{x} = 0\,1\,1\,2\,0\,2\,1\,0, \tag{1.1a}$$

$$\boldsymbol{x}^{(1)} = \overline{0}\,\underline{0}\,1\,1\,2\,0\,2\,1\,0, \tag{1.1b}$$

$$\boldsymbol{x}^{(2)} = 0\,0\,1\,1\,\overline{2\,0\,2}\,\underline{2\,0\,2}\,1\,0, \tag{1.1c}$$

$$\boldsymbol{x}^{(3)} = 0\,0\,1\,1\,2\,0\,2\,2\,0\,\overline{2\,1}\,\underline{2\,1}\,0, \tag{1.1d}$$

$$\boldsymbol{x}^{(4)} = 0\,0\,\overline{1\,1}\,\underline{1\,1}\,2\,0\,2\,2\,0\,2\,1\,2\,1\,0. \tag{1.1e}$$

Здесь $t = 4$, и выходом канала является строка $\boldsymbol{y} = \boldsymbol{x}^{(4)}$.

[2] Первые конструкции кодов для таких моделей (при $\ell \in \{4, 5, \ldots\}$) были приведены в [5].

Будем говорить, что строка $\boldsymbol{y}$ является t-потомком строки $\boldsymbol{x}$, или что $\boldsymbol{x}$ является t-предком $\boldsymbol{y}$, если $\boldsymbol{y}$ можно получить из $\boldsymbol{x}$ путем последовательного применения t тандемных дупликаций длины $\leqslant 3$. Множество всех t-потомков строки $\boldsymbol{x}$ обозначим через $D^t(\boldsymbol{x})$. Отметим, что строка может принадлежать различным множествам $D^t(\boldsymbol{x})$ и $D^s(\boldsymbol{x})$, $s \neq t$, поскольку в модели разрешены дупликации различных длин, т.е. множество $D^t(\boldsymbol{x}) \cap D^s(\boldsymbol{x})$ не обязано быть пустым (например, 0 1 1 1 1 является как 1-потомком строки 0 1 1, полученным при одной дупликации длины 2, так и 2-потомком этой же строки 0 1 1, полученным двумя дупликациями длины 1 каждая). Множество всех потомков строки $\boldsymbol{x}$ обозначим через $D^*(\boldsymbol{x}) := \bigcup\limits_{t\geqslant 0} D^t(\boldsymbol{x})$, где $D^0(\boldsymbol{x}) := \{\boldsymbol{x}\}$. В этих обозначениях $D^*(\boldsymbol{x})$ для заданной входной строки $\boldsymbol{x}$ – множество всех возможных выходов канала с тандемными ($\leqslant 3$)-дупликациями.

1.2. Различимые строки и безошибочная передача. Две строки $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}_q^*$ называются неразличимыми (confusable) в заданном канале связи, если при их передаче по этому каналу на его выходе может появиться одна и та же строка; в противном случае они называются различимыми (non-confusable). В нашей терминологии $\boldsymbol{x}$ и $\boldsymbol{y}$ неразличимы, если у них имеется общий потомок, т.е. $D^*(\boldsymbol{x}) \cap D^*(\boldsymbol{y}) \neq \varnothing$. Множество строк $\mathcal{C} \subseteq \mathcal{A}_q^*$ называется кодом *с нулевой ошибкой* [9] для заданного канала, если любые два несовпадающих кодовых слова $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$ различимы. Заметим, что код с нулевой ошибкой исправляет *все* конфигурации ошибок, реализуемые в канале, т.е. приемник может однозначно декодировать любую заданную строку на выходе канала, поскольку в $\mathcal{C}$ имеется лишь одно кодовое слово, способное породить эту строку. Код с нулевой ошибкой $\mathcal{C} \subseteq \mathcal{A}_q^n$ называется оптимальным, если не существует никакого другого кода с нулевой ошибкой $\mathcal{C}' \subseteq \mathcal{A}_q^n$, такого что $|\mathcal{C}'| > |\mathcal{C}|$.

Скорость кода $\mathcal{C} \subseteq \mathcal{A}_q^n$, выражаемая в битах на символ, определяется как $\dfrac{1}{n}\log_2|\mathcal{C}|$. Пропускная способность с нулевой ошибкой для канала с алфавитом $\mathcal{A}_q$ на входе определяется как $\limsup\limits_{n\to\infty}$ по всем скоростям оптимальных кодов с нулевой ошибкой в $\mathcal{A}_q^n$. Эта величина представляет собой максимальное число бит на символ, которые можно безошибочно передать по заданному каналу.

§ 2. Корневые и неприводимые строки относительно дупликаций

Последовательно применяя операцию дедупликации, т.е. удаления дуплицированных подстрок длины $\leqslant 3$, каждую строку $\boldsymbol{x}$ можно привести к ее *корневой* строке $R(\boldsymbol{x})$, не содержащей повторяющихся подряд подстрок длины $\leqslant 3$. Более того, как показано в [1, теорема 24], корневые строки единственны в том смысле, что независимо от порядка, в котором выполняются дедупликации, процесс гарантированно приведет к одной и той же строке. (Подчеркнем, что это "свойство единственности корневых строк" выполнено лишь для моделей с тандемными дупликациями длины (i) $=\ell$, (ii) $\leqslant 2$ или (iii) $\leqslant 3$. Оно не выполняется, например, в моделях с тандемными дупликациями длины $\leqslant \ell$ при $\ell \in \{4, 5, \ldots\}$; см. [1, теорема 40].)

В этом контексте строка, не содержащая повторяющихся подстрок длины $\leqslant 3$, называется *неприводимой*. Другими словами, строка неприводима[3], если она не содержит подстрок вида $a\,a$, $a\,b\,a\,b$ и $a\,b\,c\,a\,b\,c$, где $a, b, c \in \mathcal{A}_q$. Через Irr_q обозначим множество всех неприводимых строк над $\mathcal{A}_q$, через $\mathrm{Irr}_q(n)$ – множество всех неприводимых строк длины n, а через $I_q(n)$ – мощность последнего, т.е. $I_q(n) := |\mathrm{Irr}_q(n)|$. Из вышеупомянутого "свойства единственности корневых строк" следует, что любые две различные неприводимые строки $\boldsymbol{x}, \boldsymbol{y} \in \mathrm{Irr}_q$ различимы в канале с тандемными

[3] Неприводимые строки являются частным случаем строк с запрещенными конфигурациями, или строк с ограничениями [10], где множество запрещенных конфигураций имеет вид $\{a\,a, a\,b\,a\,b, a\,b\,c\,a\,b\,c\colon\ a, b, c \in \mathcal{A}_q\}$.

($\leqslant 3$)-дупликациями, т.е. $D^*(\boldsymbol{x}) \cap D^*(\boldsymbol{y}) = \varnothing$, и поэтому множество $\mathrm{Irr}_q(n)$ является кодом с нулевой ошибкой для этого канала [1, теорема 27].

Всюду далее будем предполагать, что $q \geqslant 3$, поскольку в случае двоичного алфавита рассматриваемая задача становится тривиальной. Например, над двоичным алфавитом существует лишь конечное число неприводимых строк, а именно $\mathrm{Irr}_2 = \{0, 1, 01, 10, 010, 101\}$, так что пропускная способность с нулевой ошибкой для канала с тандемными ($\leqslant 3$)-дупликациями над двоичным алфавитом равна нулю.

Для нас представляет интерес асимптотическое поведение величины $I_q(n)$ при $n \to \infty$, и в частности, экспонента ее скорости роста

$$\iota_q := \lim_{n\to\infty} \frac{1}{n} \log_2 I_q(n). \tag{2.1}$$

Величину ι_q можно охарактеризовать с помощью стандартных методов теории систем с ограничениями [10], например, как логарифм наибольшего собственного значения матрицы смежности направленного графа, представляющего собой диаграмму состояний системы, порождающей неприводимые строки. Здесь мы применим более простую характеризацию из [5, предложение 2], где было показано, что для $I_q(n)$ выполнено рекуррентное соотношение

$$I_q(n) = (q-2)I_q(n-1) + (q-3)I_q(n-2) + (q-2)I_q(n-3)$$

и поэтому

$$\iota_q = \log_2 r, \tag{2.2a}$$

где r – единственный положительный вещественный корень многочлена

$$x^3 - (q-2)x^2 - (q-3)x - (q-2),$$

т.е. r задается неявным условием

$$r^3 - (q-2)r^2 - (q-3)r - (q-2) = 0, \quad r > 0. \tag{2.2b}$$

В следующей лемме дается еще одна характеризация величины ι_q для троичного алфавита ($q = 3$), а также вытекающая из нее граница снизу на ι_q для бóльших алфавитов, которая будет использована в доказательстве нашего основного результата (теорема 1).

Лемма 1. *Для любых $q \geqslant 3$ и $\beta \in [0, 1]$ справедливо неравенство*

$$\iota_q \geqslant \frac{H(\beta)}{1+2\beta}, \tag{2.3}$$

где $H(\beta) := -\beta \log_2 \beta - (1-\beta)\log_2(1-\beta)$ – функция двоичной энтропии. Равенство в (2.3) достигается тогда и только тогда, когда $q = 3$ и $\beta = \bar{\beta}$, где $\bar{\beta}$ – единственное положительное решение уравнения $(1-x)^3 = x$.

Доказательство. Докажем соотношение

$$\iota_3 = \max_{0\leqslant\beta\leqslant 1} \frac{H(\beta)}{1+2\beta}, \tag{2.4}$$

из которого будет немедленно следовать утверждение леммы (так как ι_q монотонно возрастает по q). Приравнивая производную функции $\dfrac{H(\beta)}{1+2\beta}$ к нулю, убеждаемся, что эта функция достигает максимума в единственной положительной вещественной точке, удовлетворяющей уравнению $(1-x)^3 = x$, назовем ее $\bar{\beta}$. Тогда правую часть

равенства (2.4) можно представить в виде

$$\frac{H(\bar\beta)}{1+2\bar\beta} = \log_2\Big(\bar\beta^{\frac{-\bar\beta}{1+2\bar\beta}}(1-\bar\beta)^{\frac{-1+\bar\beta}{1+2\bar\beta}}\Big) = -\log_2(1-\bar\beta). \tag{2.5}$$

С другой стороны, мы знаем, что $\iota_3 = \log_2 r$, где r – единственный положительный вещественный корень уравнения $x^3 - x^2 - 1 = 0$ (см. (2.2)). Поэтому доказательство равенства в (2.4) равносильно доказательству того, что $-\log_2(1-\bar\beta) = \log_2 r$, т.е. что $(1-\bar\beta)^{-1}$ является решением уравнения $x^3 - x^2 - 1 = 0$. В этом можно убедиться непосредственно, подставляя $(1-\bar\beta)^{-1}$ вместо x и используя тот факт, что $(1-\bar\beta)^3 = \bar\beta$. ▲

§ 3. Различимость строк в канале с тандемными ($\leqslant 3$)-дупликациями

В этом параграфе изложим несколько фактов об эволюции строк под действием тандемных дупликаций длины $\leqslant 3$, основным из которых будет вывод верхней границы на максимальное число попарно различимых строк в заданном конусе потомков $D^*(\boldsymbol{x})$ (предложение 1). По поводу дальнейшего изучения комбинаторных и алгоритмических аспектов (не)различимости в ($\leqslant 2$)- и ($\leqslant 3$)-каналах с тандемными дупликациями отсылаем читателя к работе [11].

В следующей лемме утверждается, что множество попарно различимых строк, каждая из которых является 1-потомком данной строки $\boldsymbol{x}$, не может состоять из более чем двух элементов. В доказательстве также иллюстрируются условия, при которых две различимые строки могут быть получены применением различных мутаций к одной и той же строке $\boldsymbol{x}$ (см. уравнения (3.1) ниже).

Лемма 2. *Рассмотрим произвольную строку $\boldsymbol{x}$ и множество $D^1(\boldsymbol{x})$ ее 1-потомков, и пусть $\mathcal{C} \subseteq D^1(\boldsymbol{x})$ – код с нулевой ошибкой для канала с тандемными ($\leqslant 3$)-дупликациями. Тогда $|\mathcal{C}| \leqslant 2$.*

Доказательство. Рассмотрим $\boldsymbol{x}', \boldsymbol{x}'' \in D^1(\boldsymbol{x})$ и предположим, что мутации, переводящие $\boldsymbol{x}$ в $\boldsymbol{x}'$ и $\boldsymbol{x}''$, применяются к различным и непересекающимся подстрокам строки $\boldsymbol{x}$. Тогда $\boldsymbol{x}'$ и $\boldsymbol{x}''$ неразличимы, поскольку у них имеется общий потомок; действительно, достаточно применить к $\boldsymbol{x}'$ дупликацию, породившую $\boldsymbol{x}''$ из $\boldsymbol{x}$, и наоборот. Теперь предположим, что дупликации, переводящие $\boldsymbol{x}$ в $\boldsymbol{x}'$ и $\boldsymbol{x}''$, применяются к пересекающимся подстрокам $\boldsymbol{x}$. Оказывается, что во всех возможных случаях, *кроме одного*, то же самое рассуждение, что и для непересекающихся подстрок, показывает, что $\boldsymbol{x}'$ и $\boldsymbol{x}''$ неразличимы (мы продемонстрируем это для случаев, когда пересечение находится с правой стороны более длинной подстроки, так как остальные случаи симметричны им):

(i) Для случая пересекающихся подстрок длины 1 и 2 рассмотрим строку $\boldsymbol{x} = \boldsymbol{u}\,a\,b\,\boldsymbol{v}$ и заметим, что ее потомки $\boldsymbol{x}' = \boldsymbol{u}\,\overline{a\,b}\,\underline{a\,b}\,\boldsymbol{v}$ и $\boldsymbol{x}'' = \boldsymbol{u}\,a\,\overline{b}\,\underline{b}\,\boldsymbol{v}$ неразличимы, поскольку у них имеется общий потомок $\boldsymbol{u}\,a\,b\,a\,b\,b\,\boldsymbol{v}$;

(ii) Для случая пересекающихся подстрок длины 2 и 2 рассмотрим строку $\boldsymbol{x} = \boldsymbol{u}\,a\,b\,c\,\boldsymbol{v}$ и заметим, что ее потомки $\boldsymbol{x}' = \boldsymbol{u}\,\overline{a\,b}\,\underline{a\,b}\,c\,\boldsymbol{v}$ и $\boldsymbol{x}'' = \boldsymbol{u}\,a\,\overline{b\,c}\,\underline{b\,c}\,\boldsymbol{v}$ неразличимы, поскольку у них имеется общий потомок $\boldsymbol{u}\,a\,b\,a\,b\,c\,b\,c\,\boldsymbol{v}$;

(iii) Для случая пересекающихся подстрок длины 2 и 3, имеющих пересечение длины 1, рассмотрим строку $\boldsymbol{x} = \boldsymbol{u}\,a\,b\,c\,d\,\boldsymbol{v}$ и заметим, что ее потомки $\boldsymbol{x}' = \boldsymbol{u}\,\overline{a\,b\,c}\,\underline{a\,b\,c}\,d\,\boldsymbol{v}$ и $\boldsymbol{x}'' = \boldsymbol{u}\,a\,b\,\overline{c\,d}\,\underline{c\,d}\,\boldsymbol{v}$ неразличимы, поскольку у них имеется общий потомок $\boldsymbol{u}\,a\,b\,c\,a\,b\,c\,d\,c\,d\,\boldsymbol{v}$;

(iv) Для случая пересекающихся подстрок длины 2 и 3, имеющих пересечение длины 2, рассмотрим строку $\boldsymbol{x} = \boldsymbol{u}\,a\,b\,c\,\boldsymbol{v}$ и заметим, что ее потомки $\boldsymbol{x}' = \boldsymbol{u}\,\overline{a\,b\,c}\,\underline{a\,b\,c}\,\boldsymbol{v}$ и $\boldsymbol{x}'' = \boldsymbol{u}\,a\,\overline{b\,c}\,\underline{b\,c}\,\boldsymbol{v}$ неразличимы, поскольку у них имеется общий потомок $\boldsymbol{u}\,a\,b\,c\,a\,b\,c\,b\,c\,\boldsymbol{v}$;

(v) Для случая пересекающихся подстрок длины 3 и 3, имеющих пересечение длины 1, рассмотрим строку $\boldsymbol{x} = \boldsymbol{u}\overline{a\,b\,c}d\,e\,\boldsymbol{v}$ и заметим, что ее потомки $\boldsymbol{x}' = \boldsymbol{u}\,a\,b\,c\underline{a\,b\,c}d\,e\,\boldsymbol{v}$ и $\boldsymbol{x}'' = \boldsymbol{u}\,a\,b\,c\,d\,e\underline{c\,d\,e}\,\boldsymbol{v}$ неразличимы, поскольку у них имеется общий потомок $\boldsymbol{u}\,a\,b\,c\,a\,b\,c\,d\,e\,c\,d\,e\,\boldsymbol{v}$;

(vi) Для случая пересекающихся подстрок длины 3 и 3, имеющих пересечение длины 2, рассмотрим строку $\boldsymbol{x} = \boldsymbol{u}\,a\,b\,c\,d\,\boldsymbol{v}$ и заметим, что ее потомки $\boldsymbol{x}' = \boldsymbol{u}\,\overline{a\,b\,c}\underline{a\,b\,c}\,d\,\boldsymbol{v}$ и $\boldsymbol{x}'' = \boldsymbol{u}\,a\,\overline{b\,c\,d}\underline{b\,c\,d}\,\boldsymbol{v}$ неразличимы, поскольку у них имеется общий потомок $\boldsymbol{u}\,a\,b\,c\,a\,b\,c\,d\,b\,c\,d\,\boldsymbol{v}$;

(vii) Для случая пересекающихся подстрок длины 1 и 3 рассмотрим строку $\boldsymbol{x} = \boldsymbol{u}\,a\,b\,c\,\boldsymbol{v}$ и заметим, что ее потомки $\boldsymbol{x}' = \boldsymbol{u}\,\overline{a\,b\,c}\underline{a\,b\,c}\,\boldsymbol{v}$ и $\boldsymbol{x}'' = \boldsymbol{u}\,a\,b\,\overline{c}\,\underline{c}\,\boldsymbol{v}$ неразличимы, поскольку у них имеется общий потомок $\boldsymbol{u}\,a\,b\,c\,a\,b\,c\,c\,\boldsymbol{v}$.

Единственный случай, не вошедший в этот список, – это случай пересекающихся подстрок длины 1 и 3, когда пересечение возникает в середине более длинной подстроки. А именно, для $\boldsymbol{x} = \boldsymbol{u}\,a\,b\,c\,\boldsymbol{v}$, где символы $a, b, c \in \mathcal{A}_q$ различны, положим

$$\boldsymbol{x}' = \boldsymbol{u}\,\overline{a\,b\,c}\underline{a\,b\,c}\,\boldsymbol{v}, \tag{3.1a}$$

$$\boldsymbol{x}'' = \boldsymbol{u}\,a\,\overline{b}\,\underline{b}\,c\,\boldsymbol{v}. \tag{3.1b}$$

В этом случае мы не можем применить такое же рассуждение, как и выше, чтобы показать, что строки $\boldsymbol{x}'$ и $\boldsymbol{x}''$ неразличимы, и действительно, в общем случае это не обязательно верно. Например, если обе строки $\boldsymbol{u}$ и $\boldsymbol{v}$ – пустые, то $\boldsymbol{x}'$ и $\boldsymbol{x}''$ в (3.1) различимы, поскольку символ a не может оказаться после символа c в потомках строки $\boldsymbol{x}''$, в то время как *во всех* потомках слова $\boldsymbol{x}'$ символ a находится после c (аналогичный пример был приведен в [1]). Так происходит, поскольку фрагмент $a\,b\,c$, содержащийся в исходной строке $\boldsymbol{x}$, больше не возникнет в строке $\boldsymbol{x}''$, так как она "разбита" вставкой еще одной копии b. Наконец, в строке $\boldsymbol{x}''$ (соответственно, $\boldsymbol{x}'$) можно повторить дупликацию, породившую $\boldsymbol{x}'$ (соответственно, $\boldsymbol{x}''$) из $\boldsymbol{x}$, и таким образом показать, что $\boldsymbol{x}'$ и $\boldsymbol{x}''$ неразличимы, во всех случаях, кроме (3.1). Таким образом, код с нулевой ошибкой в $D^1(\boldsymbol{x})$ может содержать не более двух кодовых слов. ▲

Замечание 1. Не каждый случай вида (3.1) приводит к различимым потомкам. В качестве контрпримера предположим, что $\boldsymbol{u}$ – пустая строка, а $\boldsymbol{v} = a$, так что $\boldsymbol{x} = a\,b\,c\,a$, $\boldsymbol{x}' = \overline{a\,b\,c}\underline{a\,b\,c}\,a$ и $\boldsymbol{x}'' = a\,\overline{b}\,\underline{b}\,c\,a$. Тогда у $\boldsymbol{x}'$ и $\boldsymbol{x}''$ есть общий потомок $a\,b\,b\,c\,a\,b\,c\,a$, и поэтому они неразличимы. Однако для наших целей достаточно того факта, что (3.1) является *единственным* случаем, когда два потомка *могут быть* различимыми. В частности, это факт позволит нам вывести точную *верхнюю* границу на мощность оптимальных кодов с нулевой ошибкой.

Приведенное выше наблюдение справедливо в общем случае, а не только для 1-потомков строки $\boldsymbol{x}$. А именно, если $\boldsymbol{x}', \boldsymbol{x}'' \in D^*(\boldsymbol{x})$ получены применением к $\boldsymbol{x}$ двух разных конфигураций дупликаций, в каждой из этих строк можно повторить/воспроизвести дупликации, примененные ко второй, и таким образом показать, что у них имеется общий потомок. Единственный случай, когда такой процесс повторения дупликаций *может* в какой-то момент стать невозможным, это ситуация, когда к $\boldsymbol{x}'$ применяется дупликация длины 3, а к соответствующему фрагменту в $\boldsymbol{x}''$ – дупликация длины 1 (см. (3.1)), так что повторить в $\boldsymbol{x}''$ соответствующую мутацию, примененную к $\boldsymbol{x}'$, невозможно. Так происходит из-за того, что всякий раз, когда к строке применяется дупликация длины 2 или 3, *все* фрагменты длины $\leqslant 3$ исходной строки сохраняются в получившейся строке (с несколькими дополнительными подстроками, появляющимися в том месте, куда вставилась копия). Единственный случай, в котором фрагмент длины 3 исходной строки исчезает (не появляется в получившейся строке), – это после дупликации длины 1, как показано в (3.1b). На основе этого наблюдения мы выведем верхнюю границу на мощность оп-

тимальных кодов с нулевой ошибкой в множестве всех t-потомков данной строки $\boldsymbol{x}$ при любом t (предложение 1 ниже).

Пример 2. Чтобы пояснить вышесказанное, представим пример, приведенный в (1.1), в несколько ином виде (фрагмент **1 2 0** выделен, и показаны дупликации слева (соответственно, справа) от этого фрагмента, при которых копии вставляются слева (соответственно, справа) от оригинала):

$$\boldsymbol{x} = 0\,1\,\mathbf{1\,2\,0}\,2\,1\,0, \tag{3.2a}$$

$$\boldsymbol{x}^{(1)} = \underline{0}\,\overline{0}\,1\,\mathbf{1\,2\,0}\,2\,1\,0, \tag{3.2b}$$

$$\boldsymbol{x}^{(2)} = 0\,0\,1\,\mathbf{1}\,\overline{\mathbf{2\,0}\,2}\,\underline{2\,0\,2}\,1\,0, \tag{3.2c}$$

$$\boldsymbol{x}^{(3)} = 0\,0\,1\,\mathbf{1\,2\,0}\,2\,2\,0\,\overline{2\,1}\,\underline{2\,1}\,0, \tag{3.2d}$$

$$\boldsymbol{x}^{(4)} = 0\,0\,\underline{1\,1}\,\overline{1\,\mathbf{1}}\,\mathbf{2\,0}\,2\,2\,0\,2\,1\,2\,1\,0. \tag{3.2e}$$

Пусть $\boldsymbol{z} = 0\,1\,\mathbf{1\,2}\,2\,\mathbf{0}\,2\,1\,0$. Заметим, что в $\boldsymbol{z}$ можно воспроизвести все дупликации подстрок $\boldsymbol{x}$, которые либо не пересекаются с сегментом **1 2 0**, либо пересекаются с ним лишь частично[4], как в примерах, приведенных в (3.2):

$$\boldsymbol{x} = 0\,1\,\mathbf{1\,2\,0}\,2\,1\,0, \tag{3.3a}$$

$$\boldsymbol{z} = 0\,1\,\mathbf{1}\,\overline{\mathbf{2}}\,\underline{2}\,\mathbf{0}\,2\,1\,0, \tag{3.3b}$$

$$\boldsymbol{z}^{(1)} = \underline{0}\,\overline{0}\,1\,\mathbf{1\,2}\,2\,\mathbf{0}\,2\,1\,0, \tag{3.3c}$$

$$\boldsymbol{z}^{(2)} = 0\,0\,1\,\mathbf{1\,2}\,\overline{2\,\mathbf{0}\,2}\,\underline{2\,0\,2}\,1\,0, \tag{3.3d}$$

$$\boldsymbol{z}^{(3)} = 0\,0\,1\,\mathbf{1\,2}\,2\,\mathbf{0}\,2\,2\,0\,\overline{2\,1}\,\underline{2\,1}\,0, \tag{3.3e}$$

$$\boldsymbol{z}^{(4)} = 0\,0\,\underline{1\,1}\,\overline{1\,\mathbf{1}}\,\mathbf{2}\,2\,\mathbf{0}\,2\,2\,0\,2\,1\,2\,1\,0. \tag{3.3f}$$

Поэтому любая пара строк из (3.2) и (3.3) неразличима; например, общим потомком $\boldsymbol{z}$ и $\boldsymbol{x}^{(3)}$ является $\boldsymbol{z}^{(3)}$. Единственная мутация, которую нельзя повторить в $\boldsymbol{z}$, – это дупликация всего фрагмента **1 2 0**, поскольку соответствующий фрагмент в $\boldsymbol{z}$ больше не существует (он был "разбит" вставленным символом 2). Например, если бы нужно было превратить $\boldsymbol{x}^{(2)}$ из (3.2c) в строку

$$\boldsymbol{y} = 0\,0\,1\,\overline{\mathbf{1\,2\,0}}\,\underline{1\,2\,0}\,2\,2\,0\,2\,1\,0 \tag{3.4}$$

вместо $\boldsymbol{x}^{(3)}$, было бы уже невозможно применить тот же процесс, что и в (3.3).

Прежде чем сформулировать предложение 1, являющееся основным результатом этого параграфа, докажем одну полезную лемму.

Лемма 3. *Зафиксируем натуральные числа b, t, n, такие что $b \leqslant t \leqslant n$. Пусть $\mathcal{U} \subseteq \{l_1, l_3, *\}^n$ – множество строк, удовлетворяющих следующим двум условиям:*

(1) *Каждая строка в $\boldsymbol{u} \in \mathcal{U}$ имеет ровно b символов l_3, $t-b$ символов l_1, и $n-t$ символов $*$;*

(2) *Для любых двух различных строк $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{U}$ найдется позиция $i \in \{1, \ldots, n\}$, в которой $\{u_i, v_i\} = \{l_1, l_3\}$ (т.е. такая, что либо $u_i = l_1$ и $v_i = l_3$, либо $u_i = l_3$ и $v_i = l_1$).*

Тогда $|\mathcal{U}| \leqslant 2^{tH(b/t)}$.

Доказательство. Рассмотрим n бросков монеты, для которой вероятность выпадения орла равна b/t, и введем следующие события, индексированные строками из множества $\mathcal{U}$. Для $\boldsymbol{u} \in \mathcal{U}$ обозначим через $A_{\boldsymbol{u}}$ событие, состоящее в том, что на i-м

[4] Для $\boldsymbol{x}$ из (3.2a) подстроки длины $\leqslant 3$, частично пересекающиеся с подстрокой **1 2 0**, следующие: **1**, **2**, **0**, 1 **1**, **1 2**, **2 0**, **0** 2, 0 1 **1**, 1 **1 2**, **2 0** 2, **0** 2 1.

броске монеты выпал орел, если $u_i = l_3$, выпала решка, если $u_i = l_1$, и неважно, что выпало, если $u_i = *$, для всех $i = 1, \ldots, n$. Из условия (1) следует, что вероятность такого события равна

$$\Pr\{A_{\boldsymbol{u}}\} = \left(\frac{b}{t}\right)^b \left(1 - \frac{b}{t}\right)^{t-b} = 2^{-tH(b/t)}$$

для любой строки $\boldsymbol{u} \in \mathcal{U}$. При этом в силу условия (2) события $A_{\boldsymbol{u}}$ и $A_{\boldsymbol{v}}$ несовместны для любых $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{U}$, $\boldsymbol{u} \neq \boldsymbol{v}$. Отсюда $|\mathcal{U}| \Pr\{A_{\boldsymbol{u}}\} \leqslant 1$, что и требовалось доказать. ▲

Заметим, что приведенная верхняя граница на $|\mathcal{U}|$ не зависит от n (т.е. от числа символов $*$).

Предложение 1. *Рассмотрим строку $\boldsymbol{x} \in \mathcal{A}_q^n$, и пусть $\mathcal{C} \subseteq D^t(\boldsymbol{x})$ – код с нулевой ошибкой для канала с тандемными ($\leqslant 3$)-дупликациями, удовлетворяющий следующему условию: из t дупликаций, порождающих каждое кодовое слово $\boldsymbol{y} \in \mathcal{C}$ из слова $\boldsymbol{x}$, ровно b имеют длину 3. Тогда $|\mathcal{C}| \leqslant 2^{tH(b/t)}$.*

Доказательство. Как показано выше, если две строки $\boldsymbol{x}', \boldsymbol{x}'' \in D^t(\boldsymbol{x})$ различимы, то это с необходимостью означает, что к фрагменту $a\,b\,c$ в одном из предков строки $\boldsymbol{x}'$ была применена дупликация длины 3, а к среднему символу соответствующего фрагмента в некотором предке $\boldsymbol{x}''$ была применена дупликация длины 1, или наоборот. Другими словами, для любой пары кодовых слов кода с нулевой ошибкой в $D^t(\boldsymbol{x})$ найдется фрагмент, в котором они отличаются на мутацию длины 1/длины 3. Поэтому интересующий нас вопрос состоит в том, насколько велико может быть множество строк, любые две из которых отличаются на мутацию длины 1/длины 3 в некоторой позиции. (Неважно, что это за позиции и что происходит в промежутках между ними, единственное требование состоит в том, что любая пара кодовых слов в каком-то месте отличается на мутацию длины 1/длины 3, так как это единственная возможность для того, чтобы две строки могли стать различимыми.) Поэтому $|\mathcal{C}|$ можно ограничить сверху максимальной мощностью множества $\mathcal{U}$ строк над "алфавитом" {длина 1, длина 3, $*$}, удовлетворяющего следующим условиям: 1) любая строка в $\mathcal{U}$ содержит ровно b символов "длина 3" и $t - b$ символов "длина 1"; 2) любые две различные строки в $\mathcal{U}$ отличаются в некоторой позиции на символ "длина 1"/"длина 3". (Формальный символ $*$ служит для заполнения пустых позиций, которые могут возникать из-за того, что различные пары кодовых слов могут отличаться на мутацию длины 1/длины 3 в различных фрагментах; см. также пример 3 ниже.) Теперь требуемая граница получается применением леммы 3. ▲

Пример 3. Приведем пример множества строк $\mathcal{U}$ из вышеизложенного доказательства. Рассмотрим следующую строку:

$$\boldsymbol{x} = \overbrace{0\,\underbrace{1\,2}\,3}\,4\,\overbrace{5\,6\,\underbrace{7}\,8\,9}. \tag{3.5}$$

(Все символы в $\boldsymbol{x}$ обозначены по-разному для облегчения понимания, это никак не влияет на общность рассуждений.) Пусть применением тандемных дупликаций к четырем фрагментам в $\boldsymbol{x}$, отмеченных скобками сверху или снизу, получены следующие потомки в $D^3(\boldsymbol{x})$:

$$\boldsymbol{x}' = 0\,1\,2\,\underline{0\,1\,2}\,\underline{2}\,3\,4\,5\,6\,\underline{6}\,7\,8\,9, \tag{3.6a}$$

$$\boldsymbol{x}'' = 0\,1\,\underline{1}\,2\,3\,4\,5\,6\,\underline{6}\,7\,8\,9\,\underline{7\,8\,9}, \tag{3.6b}$$

$$\boldsymbol{x}''' = 0\,1\,\underline{1}\,2\,3\,4\,5\,6\,7\,\underline{5\,6\,7}\,8\,\underline{8}\,9, \tag{3.6c}$$

где подчеркнуты вставленные копии. Мутации, примененные к этим четырем фрагментам, можно описать с помощью следующих строк:

$$\boldsymbol{u}' = l_3\,l_1\,l_1\,*, \tag{3.7a}$$

$$\boldsymbol{u}'' = l_1 * l_1\, l_3, \tag{3.7b}$$
$$\boldsymbol{u}''' = l_1 * l_3\, l_1, \tag{3.7c}$$

где символ l_3 показывает, что к соответствующему фрагменту применена дупликация длины 3, символ l_1 показывает, что к среднему символу этого фрагмента применена дупликация длины 1, а $*$ означает, что к этому фрагменту ни одна из этих мутаций не применялась. Заметим, что для любой пары строк в (3.7) найдется координата, в которой одна из них равна l_1, а другая l_3.

§ 4. Пропускная способность с нулевой ошибкой для канала с тандемными ($\leqslant 3$)-дупликациями

Пусть $\mathcal{C}_q^\star(n) \subseteq \mathcal{A}_q^n$ – оптимальный код с нулевой ошибкой для канала с тандемными ($\leqslant 3$)-дупликациями. Для заданной неприводимой строки $\boldsymbol{x} \in \mathrm{Irr}_q$ положим

$$\mathcal{C}_q^\star(n;\boldsymbol{x}) := \mathcal{C}_q^\star(n) \cap D^*(\boldsymbol{x}).$$

Тогда $\mathcal{C}_q^\star(n;\boldsymbol{x})$ является оптимальным кодом с нулевой ошибкой на множестве всех потомков $\boldsymbol{x}$ длины n. Это так, поскольку в силу свойства единственности корневой строки для тандемных дупликаций длины $\leqslant 3$ любые два различных конуса потомков не пересекаются [1, следствие 26], и поэтому любые две строки, имеющие различные корневые строки, различимы. Иными словами, можно, не ограничивая общности, строить код по отдельности в конусе потомков для каждой возможной корневой/неприводимой строки. Этот факт явно сформулирован в следующей лемме; ее доказательство опущено, поскольку оно непосредственно вытекает из [1, следствие 26].

Лемма 4. *Оптимальный код длины n с нулевой ошибкой можно представить в виде несвязного объединения оптимальных кодов в каждом из конусов потомков*:

$$\mathcal{C}_q^\star(n) = \dot{\bigcup_{\boldsymbol{x}\in\mathrm{Irr}_q}} \mathcal{C}_q^\star(n;\boldsymbol{x}). \tag{4.1}$$

Следующее утверждение дает характеризацию экспоненты скорости роста мощности оптимальных кодов $\mathcal{C}_q^\star(n)$ или, эквивалентным образом, пропускной способности с нулевой ошибкой канала с тандемными ($\leqslant 3$)-дупликациями. В нем утверждается, что эта величина равна

$$\iota_q = \lim_{n\to\infty} \frac{1}{n} \log_2 I_q(n)$$

(см. (2.1) и (2.2)). Другими словами, пропускная способность с нулевой ошибкой достигается на кодах $\mathrm{Irr}_q(n)$, состоящих из неприводимых строк длины n.

Теорема 1. *Пропускная способность с нулевой ошибкой канала с тандемными ($\leqslant 3$)-дупликациями с алфавитом $\mathcal{A}_q$, $q \geqslant 3$, равна ι_q.*

Доказательство. Требуется показать, что

$$\lim_{n\to\infty} \frac{1}{n} \log_2 \bigl|\mathcal{C}_q^\star(n)\bigr| = \iota_q. \tag{4.2}$$

Так как $\mathrm{Irr}_q(n) \subseteq \mathcal{C}_q^\star(n)$, то мы знаем, что

$$\lim_{n\to\infty} \frac{1}{n} \log_2 |\mathcal{C}_q^\star(n)| \geqslant \lim_{n\to\infty} \frac{1}{n} \log_2 I_q(n) = \iota_q$$

(см. (2.1)), поэтому достаточно доказать противоположное неравенство

$$\lim_{n\to\infty}\frac{1}{n}\log_2|\mathcal{C}_q^\star(n)|\leqslant\iota_q.$$

Для этого мы упростим анализ, построив достаточно большой *подкод* $\mathcal{C}_q(n;m,t,b)\subseteq\subseteq\mathcal{C}_q^\star(n)$, имеющий ту же экспоненту скорости роста, что и оптимальный код $\mathcal{C}_q^\star(n)$, т.е.

$$\lim_{n\to\infty}\frac{1}{n}\log_2|\mathcal{C}_q^\star(n)|=\lim_{n\to\infty}\frac{1}{n}\log_2|\mathcal{C}_q(n;m,t,b)|,$$

при подходящем выборе параметров m,t,b.

Зафиксируем произвольную неприводимую строку $\boldsymbol{x}$ длины m, т.е. $\boldsymbol{x}\in\mathrm{Irr}_q(m)$, и пусть $\mathcal{C}_q(n;\boldsymbol{x},t,b)\subseteq\mathcal{C}_q^\star(n;\boldsymbol{x})$ – код, содержащий только те кодовые слова из $\mathcal{C}_q^\star(n;\boldsymbol{x})$, которые удовлетворяют следующим двум условиям: 1) каждое кодовое слово лежит в $D^t(\boldsymbol{x})$, т.е. является t-потомком строки $\boldsymbol{x}$; 2) из t дупликаций, переводящих $\boldsymbol{x}$ в заданный потомок/заданное кодовое слово, ровно b имеют длину 3. Определим такой подкод следующим образом:

$$\mathcal{C}_q(n;m,t,b):=\bigcup_{\boldsymbol{x}\in\mathrm{Irr}_q(m)}\mathcal{C}_q(n;\boldsymbol{x},t,b).\tag{4.3}$$

Из этой конструкции и леммы 4 следует, что

$$\mathcal{C}_q^\star(n)=\bigcup_{m,t,b}\mathcal{C}_q(n;m,t,b).\tag{4.4}$$

Теперь должно быть ясно, что величина $|\mathcal{C}_q(n;m,t,b)|$, максимизированная по всем возможным значениям m,t,b, имеет ту же самую экспоненту скорости роста, что и $|\mathcal{C}_q^\star(n)|$ (значения m,t,b выбираются для каждого n, т.е. оптимальные значения параметров m,t,b являются, вообще говоря, функциями от длины блока n). Это следует из равенства (4.4) и принципа Дирихле – мощность кода $\mathcal{C}_q^\star(n)$ растет экспоненциально быстро по длине блока n, а выбрать значения каждого из параметров m, t и b можно линейным количеством способов, поэтому по крайней мере для одного такого выбора код $\mathcal{C}_q(n;m,t,b)$ будет содержать экспоненциально много кодовых слов (с тем же показателем экспоненты). Таким образом, коды $\mathcal{C}_q(n;m,t,b)$ асимптотически оптимальны по скорости, т.е. на них достигается пропускная способность с нулевой ошибкой канала с тандемными ($\leqslant 3$)-дупликациями, когда параметры m,t,b выбираются правильным образом (так, чтобы максимизировать $|\mathcal{C}_q(n;m,t,b)|$).

Теперь вычислим скорость построенных кодов. Согласно (4.3) и предложению 1 (в котором утверждается, что $|\mathcal{C}_q(n;\boldsymbol{x},t,b)|\leqslant 2^{tH(b/t)}$), мощность кода $\mathcal{C}_q(n;m,t,b)$ можно оценить сверху как

$$\bigl|\mathcal{C}_q(n;m,t,b)\bigr|\leqslant I_q(m)\cdot 2^{tH(b/t)},\tag{4.5}$$

а длину этого кода можно оценить снизу как

$$n\geqslant m+3b+(t-b)=m+t+2b\tag{4.6}$$

(начальная неприводимая строка имеет длину m, и ровно b дупликаций, порождающих ее потомков, имеют длину 3). Следовательно,

$$\frac{1}{n}\log_2\bigl|\mathcal{C}_q(n;m,t,b)\bigr|\leqslant\frac{\log_2 I_q(m)+tH(b/t)}{m+t+2b}.\tag{4.7}$$

Чтобы найти асимптотику этой величины при $n \to \infty$, нужно рассмотреть два случая, соответствующие разным выборам параметров m, t, b:

- $m = o(t)$. Пусть $\lim\limits_{t\to\infty} \dfrac{b}{t} = \beta \in [0,1]$. Тогда

$$\limsup_{n\to\infty} \frac{1}{n} \log_2 \bigl|\mathcal{C}_q(n; m, t, b)\bigr| \leqslant \frac{H(\beta)}{1+2\beta} \leqslant \iota_q, \tag{4.8}$$

  где первое неравенство следует из (4.7), а второе совпадает с (2.3);
- $t = \mathcal{O}(m)$. Пусть $\liminf\limits_{m\to\infty} \dfrac{t}{m} = \tau \geqslant 0$ и $\lim\limits_{t\to\infty} \dfrac{b}{t} = \beta \in [0,1]$. Тогда

$$\limsup_{n\to\infty} \frac{1}{n} \log_2 \bigl|\mathcal{C}_q(n; m, t, b)\bigr| \leqslant \frac{\iota_q + \tau H(\beta)}{1+\tau(1+2\beta)} \leqslant \iota_q. \tag{4.9}$$

Здесь снова первое неравенство вытекает из (4.7), а второе равносильно (2.3). Итак, все выборы значений параметров m, t, b дают асимптотическую скорость кодов $\mathcal{C}_q(n; m, t, b)$, не превосходящую ι_q. Поскольку все эти коды оптимальны по скорости, как указано во втором абзаце этого доказательства, равенство (4.2) тем самым доказано. ▲

§ 5. Заключение

Эволюция строк под действием тандемных дупликаций – интересная и нетривиальная задача, важная в нескольких областях исследований. В настоящей статье исследована различимость строк при тандемных дупликациях переменной длины – задача, инспирированная исправлением ошибок в каналах связи, в которых передаваемые сообщения подвергаются мутациям такого типа. А именно, для случая дупликаций длины $\leqslant 3$ получена верхняя граница на максимальную мощность множества попарно различимых строк, которая вместе с конструкцией из [1] определяет максимальную скорость, достижимую кодами, исправляющими *произвольное* число таких дефектов.

В случаях, когда корневые строки относительно дупликаций не единственны, например, для модели ($\leqslant \ell$)-тандемных дупликаций с параметром ℓ, большим чем 3, максимальные достижимые скорости остаются неизвестными. Благодаря свойству "неединственности корневых строк" анализ эволюции и различимости строк в таких моделях более сложен, и поэтому для решения задачи о пропускной способности с нулевой ошибкой и смежных с ней задач потребуются дальнейшие исследования и, возможно, другие методы.

СПИСОК ЛИТЕРАТУРЫ

1. *Jain S., Farnoud F., Schwartz M., Bruck J.* Duplication-Correcting Codes for Data Storage in the DNA of Living Organisms // IEEE Trans. Inform. Theory. 2017. V. 63. № 8. P. 4996–5010. https://doi.org/10.1109/TIT.2017.2688361
2. *Kovačević M., Tan V.Y.F.* Asymptotically Optimal Codes Correcting Fixed-Length Duplication Errors in DNA Storage Systems // IEEE Commun. Lett. 2018. V. 22. № 11. P. 2194–2197. https://doi.org/10.1109/LCOMM.2018.2868666
3. *Lenz A., Jünger N., Wachter-Zeh A.* Bounds and Constructions for Multi-Symbol Duplication Error Correcting Codes, https://arXiv.org/abs/1807.02874v3 [cs.IT], 2018.
4. *Kovačević M.* Zero-Error Capacity of Duplication Channels // IEEE Trans. Commun. 2019. V. 67. № 10. P. 6735–6742. https://doi.org/10.1109/TCOMM.2019.2931342
5. *Chee Y.M., Chrisnata J., Kiah H.M., Nguyen T.T.* Efficient Encoding/Decoding of GC-Balanced Codes Correcting Tandem Duplications // IEEE Trans. Inform. Theory. 2020. V. 66. № 8. P. 4892–4903. https://doi.org/10.1109/TIT.2020.2981069

6. *Farnoud F., Schwartz M., Bruck J.* The Capacity of String-Duplication Systems // IEEE Trans. Inform. Theory. 2016. V. 62. № 2. P. 811–824. https://doi.org/10.1109/TIT.2015.2505735
7. *Jain S., Farnoud F., Bruck J.* Capacity and Expressiveness of Genomic Tandem Duplication // IEEE Trans. Inform. Theory. 2017. V. 63. № 10. P. 6129–6138. https://doi.org/10.1109/TIT.2017.2728079
8. *Leupold P., Martín-Vide C., Mitrana V.* Uniformly Bounded Duplication Languages // Discrete Appl. Math. 2005. V. 146. № 3. P. 301–310. https://doi.org/10.1016/j.dam.2004.10.003
9. *Shannon C.E.* The Zero Error Capacity of a Noisy Channel // IRE Trans. Inform. Theory. 1956. V. 2. № 3. P. 8–19. https://doi.org/10.1109/TIT.1956.1056798
10. *Marcus B.H., Roth R.M., Siegel P.H.* An Introduction to Coding for Constrained Systems (unpublished manuscript), 5th ed., 2001. Available online at http://www.math.ubc.ca/~marcus/Handbook/.
11. *Chee Y.M., Chrisnata J., Kiah H.M., Nguyen T.T.* Deciding the Confusability of Words under Tandem Repeats in Linear Time // ACM Trans. Algorithms. 2019. V. 15. № 3. Art. 42 (22 pp.). https://doi.org/10.1145/3338514

Ковачевич Младен
Факультет технических наук,
Университет г. Нови-Сад, Сербия
kmladen@uns.ac.rs

Поступила в редакцию
04.10.2021
После доработки
20.04.2022
Принята к публикации
21.04.2022

ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 58 **2022** **Вып. 2**

УДК 621.391 : 519.724.6 : 519.725.3

© 2022 г. **А.А. Курмукова, Ф.И. Иванов, В.В. Зяблов**

ТЕОРЕТИЧЕСКИЕ И ЭКСПЕРИМЕНТАЛЬНЫЕ ОЦЕНКИ СВЕРХУ И СНИЗУ ДЛЯ ЭФФЕКТИВНОСТИ СВЕРТОЧНЫХ КОДОВ В ДВОИЧНОМ СИММЕТРИЧНОМ КАНАЛЕ[1]

Предлагается новый подход для построения аналитических оценок вероятности появления пакета ошибок заданной длины, вероятности ошибочного декодирования и вероятности ошибки на бит для сверточных кодов в двоичном симметричном канале с декодированием Витерби. Выражения, полученные для верхней и нижней оценок вероятности ошибки на бит, а также для вероятности ошибочного декодирования, основаны на активных расстояниях и спектре активных расстояний сверточного кода. Предложенные оценки справедливы для сверточных кодов скорости $1/2$, но их можно обобщить и для сверточных кодов скорости $1/n$. Вычисление описанных оценок имеет линейную сложность от длины наиболее коротких пакетов ошибок при известных метрических характеристиках сверточного кода, при этом сложность вычисления не зависит от входной вероятности ошибки на бит. Результаты моделирования показывают, что рассматриваемые оценки достаточно точны, особенно для малых вероятностей искажений в канале.

Ключевые слова: сверточные коды, активное расстояние, вероятность ошибки на бит, решетка кода.

DOI: 10.31857/S055529232202003X, **EDN:** DZBRSC

§ 1. Введение

Сверточные коды, предложенные Элайесом в [1], являются предметом изучения уже на протяжении более 60 лет. Несмотря на то, что они уступают с точки зрения вероятности ошибки на блок при заданном уровне помех в канале передачи информации полярным кодам [2] или кодам с малой плотностью проверок на четность (МПП-кодам) [3], сверточные коды могут существенно снизить вероятность ошибки на бит при фиксированном отношении сигнал-шум [4]. Полярные коды и большинство классов МПП-кодов этим свойством не обладают [4]. Кроме сверточных кодов таким свойством обладают всего несколько классов кодов: например, коды Хэмминга [5], некоторые МПП-коды с порождающей матрицей низкой плотности [6], кодовые конструкции с повторениями [7].

Уменьшение величины ошибки на бит – это очень важное свойство для внутренних кодов в любых каскадах. Это одна из главных причин, почему сверточные коды широко используются в различных каскадных кодовых конструкциях, например, в плетеных кодах [8] и в других последовательных и параллельных каскадных конструкциях [9, 10]. Также рекурсивный систематический кодер сверточного кода часто используется в турбо-кодах [11, 12]. В каскаде с линейным кодом такой кодер

[1] Работа выполнена в рамках Программы фундаментальных исследований НИУ ВШЭ в 2022 г.

также может использоваться и показывать высокую эффективность, как, например, в параллельном каскаде с систематическим полярным кодом и итеративным декодированием [13].

Таким образом, аналитические методы расчета вероятности ошибки для сверточного кода являются важным направлением исследований. Существует ряд работ, в которых изучались вопросы теоретического анализа эффективности сверточных кодов. В работе [14] представлен классический способ подсчета вероятности ошибки с помощью производящей функции спектра кода и функции спектра ошибки на бит. Для двоичного симметричного канала используется уточненная граница Миберга [15]. В работе [16] используется простая цепь Маркова для оценки вероятности ошибки на бит сверточного кода с ограничением длины пакетов. Но экспериментальные данные, представленные в работе, показывают, что предложенная оценка является достаточно грубой и неточной для малых значений вероятности искажений в канале. Некоторые модификации этого подхода описаны в [17], где авторы используют марковскую цепь для описания конструкции сверточного кода. В этом случае процесс декодирования Витерби [18] может рассматриваться с помощью специальных переходов между состояниями марковской цепи. Оценка вероятности ошибки на бит для сверточного кода с выкалываниями была предложена в работе [19]. Наконец, в [20] показано, как оценить вероятность ошибки на бит для любого сверточного кода как для двоичного симметричного канала, так и для канала с гауссовским шумом. Но описанный в этой работе подход сложен в реализации, а также практически не применим для сверточных кодов с большой памятью из-за вычислительной сложности.

В данной статье предложен подход к оцениванию вероятности ошибки на бит и на блок для сверточного кода в двоичном симметричном канале при декодировании кода по максимуму правдоподобия. Предложенный метод основан на спектре активных расстояний сверточного кода и аналитических оценках вероятности появления пакета ошибок заданной длины [21, 22]. Вычисление вероятности ошибки на блок для терминированного сверточного кода было описано в работе [23]. В этой работе мы предлагаем вычислительно простые аналитические оценки сверху и снизу для вероятности ошибки на бит.

§ 2. Сверточные коды

2.1. Сверточный код с рекурсивным кодером. Для простоты последующих выкладок в данной статье рассмотрим двоичный сверточный код скорости $1/2$ с рекурсивным систематическим кодером. Однако сразу отметим, что полученные результаты могут быть обобщены для кода скорости $1/n$, $n > 1$, $n \in \mathbb{N}$. Также все полученные результаты верны для двоичного симметричного канала и декодера Витерби.

Порождающая матрица сверточного кода скорости $1/2$ с систематическим кодером с обратной связью может быть задана с помощью отношения многочленов:

$$\boldsymbol{G}(D) = \left(1 \quad \frac{\boldsymbol{g}^{(2)}}{\boldsymbol{g}^{(1)}}\right), \tag{1}$$

где порождающие полиномы –

$$\boldsymbol{g}^{(\ell)}(D) = g_0^{(\ell)} + g_1^{(\ell)} D + g_2^{(\ell)} D^2 + \ldots + g_m^{(\ell)} D^m, \quad g_i^{(\ell)} \in \{0, 1\},$$

для $\ell = 1, 2$. Память кода здесь обозначена через m, скорость кода $1/2$. Обратная связь в кодере задана знаменателем в отношении полиномов. Мы рассматриваем взаимно простые порождающие полиномы одинаковой степени. Состояние кодера в каждый момент времени может быть задано либо с помощью m битов, либо десятичным числом s: $0 \leqslant s < 2^m$. Всего возможно 2^m различных состояний.

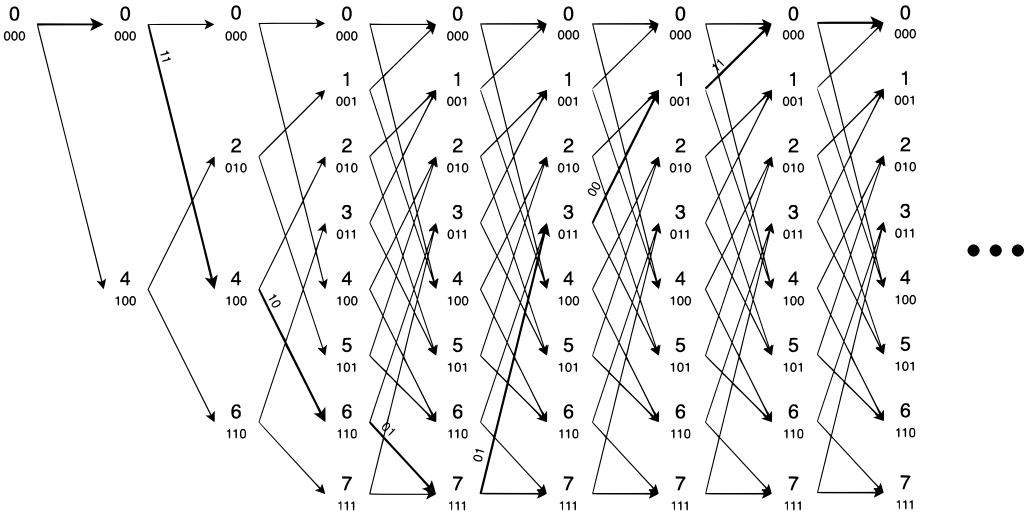


Рис. 1. Решетка для сверточного кода $(13, 17)$ с рекурсивным кодером

Кодовое слово сверточного кода может быть записано как полином с помощью оператора задержки D:

$$\boldsymbol{v}(D) = \boldsymbol{v}_0 + \boldsymbol{v}_1 D + \boldsymbol{v}_2 D^2 + \ldots, \quad \boldsymbol{v}_i \in \{0,1\}^2, \quad i \in \mathbb{N} \cup \{0\},$$

где $(\boldsymbol{v}_0, \boldsymbol{v}_1, \boldsymbol{v}_2, \ldots)$ – последовательность выходных кортежей длины 2. Кодовое слово сверточного кода полубесконечно и может быть однозначно задано последовательностью входных битов и начальным состоянием кодера.

Для представления кодовых слов сверточного кода и понимания процесса декодирования удобно использовать решетчатую диаграмму состояний кодера. Решетка изображается в виде состояний кодера, отложенных в последовательные моменты времени. Переход из состояния s_t в момент времени t в состояние s_{t+1} в момент времени $t+1$ соответствует некоторому входному биту и выходному кортежу. Начальное состояние кодера в момент времени $t = 0$ считают нулевым. Каждому кодовому слову соответствует уникальная последовательность выходных кортежей, последовательность состояний решетки и последовательность входных битов. Таким образом, каждое кодовое слово может быть записано как путь в решетке (последовательность состояний решетки). Решетчатая диаграмма состояний важна также для понимания метрических свойств сверточного кода.

На рис. 1 приведен пример решетки для сверточного кода памяти $m = 3$ с порождающими полиномами $\boldsymbol{g}^{(1)} = 1 + D + D^3$, $\boldsymbol{g}^{(2)} = 1 + D + D^2 + D^3$, которые можно также представить в восьмеричном виде как $(13, 17)$. На рис. 1 выделенный полужирным путь соответствует началу одного из кодовых слов. Более подробное изложение концепции решетчатых диаграмм приведено в работе [18].

2.2. Активные расстояния и спектр активных расстояний. Одним из важнейших метрических характеристик сверточного кода является свободное расстояние.

Определение 1. Свободное расстояние сверточного кода – это минимальный вес ненулевого кодового слова.

Кроме свободного расстояния, у сверточных кодов есть ряд других не менее важных метрических характеристик, которые определяют вероятность ошибочного декодирования.

Для блочных кодов при оценивании вероятности ошибки на блок обычно рассматривают весовой спектр кода [24]. В данной статье мы предлагаем использовать активные расстояния, предложенные в [25], и спектр активных расстояний для анализа сверточных кодов. Если рассматривать обычный спектр весов сверточного кода, то веса не будут увеличиваться при увеличении длины рассматриваемого пути, если путь начиная с некоторого момента пойдет только по нулевым состояниям решетки. Поэтому имеет смысл исключать из рассмотрения пути, как только они сливаются с нулевым путем. В этом заключается основное различие двух понятий: спектр весов и спектр весов активных расстояний.

Для начала введем дополнительное построение, чтобы определить активные расстояния.

Определение 2. Подмножество кодовых слов сверточного кода $\mathcal{C}$ с одним конечным ответвлением от нулевого пути обозначим через $\mathcal{C}_f$. Это подмножество $\mathcal{C}_f$ состоит из кодовых слов $\boldsymbol{v}(D)$, которым соответствует конечный ненулевой путь некоторой длины j, этот ненулевой путь начинается сразу после начального нулевого состояния решетки: $s_0 = 0$, $s_1 \neq 0$, и не имеет двух последовательных нулевых состояний до слияния с нулевым путем решетки $s_j = s_{j+1} = s_{j+2} = \ldots = 0$ в момент времени j.

Кодовые слова из подмножества $\mathcal{C}_f$ имеют конечный вес, и этот вес равен весу ненулевого пути в решетке. Для определения активного расстояния рассматривается только начальная ненулевая часть пути. Отметим, что подмножество $\mathcal{C}_f$ не включает в себя нулевое кодовое слово. Минимально возможная длина ненулевого пути составляет $m+1$, где m – память кода. Здесь и далее длиной пути считается количество выходных кортежей (или число переходов между состояниями решетки). Тогда кодовое слово с ненулевым путем длины j из подмножества $\mathcal{C}_f$ обозначим через $\boldsymbol{v}^{(j)}$.

Определение 3. Активное расстояние длины j сверточного кода $\mathcal{C}$ – это минимальный вес Хэмминга кодового слова $\boldsymbol{v}^{(j)}$ из подмножества $\mathcal{C}_f$, имеющего конечный ненулевой путь длины j:

$$a_j = \min_{\boldsymbol{v}^{(j)} \in \mathcal{C}_f} w(\boldsymbol{v}^{(j)}),$$

где $w(\boldsymbol{v}^{(j)})$ – вес Хэмминга кодового слова $\boldsymbol{v}^{(j)}$.

Свободное расстояние сверточного кода можно также определить через активные расстояния, так как при рассмотрении подмножества кодовых слов $\mathcal{C}_f$ мы не исключаем кодовые слова минимального веса. Тогда свободное расстояние можно ввести как минимальное активное расстояние $\min\limits_{j} a_j$, $j \geqslant m+1$. Также отметим, что свободное расстояние не всегда соответствует минимальной длине j.

Так как вес кодового слова $\boldsymbol{v}^{(j)} \in \mathcal{C}_f$ определяется весом ненулевого пути в решетке, то можно записать вес кодового слова с помощью первых j выходных кортежей:

$$w(\boldsymbol{v}^{(j)}) = \sum_{i=0}^{j-1} w(\boldsymbol{v}_i^{(j)}).$$

Число кодовых слов $\boldsymbol{v}^{(j)} \in \mathcal{C}_f$ с весом $w^{(j)}$ будем обозначать через $N_{w^{(j)}}$. Приведем определение спектра активных расстояний, впервые предложенное в работе [21].

Определение 4. Спектр $\mathcal{D}_{a_j}$ активных расстояний a_j длины j сверточного кода – это множество наборов

$$\mathcal{D}_{a_j} = \left\{ j, a_j, w^{(j)}, N_{w^{(j)}} \mid w^{(j)} = w(\boldsymbol{v}^{(j)}),\ \boldsymbol{v}^{(j)} \in \mathcal{C}_f \right\}.$$

Таблица 1

Спектр активных расстояний для сверточного кода $(13, 15)$ с рекурсивным кодером

$w\backslash j$	4	5	6	7	8	9	10	11	12	13	14	15
6	**1**		**1**									
7												
8		**1**	1	**2**	**2**	**2**	**1**	**1**				
9												
10				2	4	6	8	8	**8**	**6**	**4**	**2**
11												
12					2	5	13	19	29	34	36	34
13												
14						2	6	20	38	68	100	132
15												
16							1	8	25	64	132	230
17												
18									8	30	93	220
19												
20										6	32	121
21												
22											4	32
23												
24												2
25												

Таблица 2

Спектр активных расстояний для сверточного кода $(13, 17)$ с рекурсивным кодером

$w\backslash j$	4	5	6	7	8	9	10	11	12	13	14	15
6		**1**										
7	**1**		**1**	**1**								
8			1	1	**1**	**2**						
9					4	1	**3**	**3**				
10				1	1	3	6	4	**5**	**5**		
11				1		4	5	5	15	7	**10**	**8**
12					1	2	4	13	10	18	27	16
13						3	3	13	14	30	32	37
14					1		5	7	22	34	40	81
15							3	5	15	40	60	96
16								2	17	21	79	105
17								2	7	20	62	119
18								2		21	36	100
19									2	7	26	90
20										5	12	71
21									1		11	27
22											6	11
23												6
24												3
25												3

Минимальный вес в спектре активных расстояний $\mathcal{D}_{a_j}$ – это активное расстояние a_j длины j сверточного кода в соответствии с определением активных расстояний. Подсчет спектра кода является вычислительно более трудной задачей, чем расчет спектра активных расстояний, так как при подсчете спектра активных расстояний исключаются кодовые слова, идущие по нулевому пути. Объединение спектров активных расстояний длины j для всех возможных длин формирует весь спектр активных расстояний.

Определение 5. Спектр $\mathcal{D}_a$ активных расстояний сверточного кода – это объединение множеств наборов $\mathcal{D}_{a_j}$ для всех возможных длин j:

$$\mathcal{D}_a = \bigcup_{j=m+1}^{\infty} \mathcal{D}_{a_j},$$

где m – память кода.

Спектр активных расстояний удобно представлять в виде таблицы. Приведем примеры спектров активных расстояний для двух кодов.

Пример 1. В табл. 1 и 2 представлены спектры активных расстояний для сверточных кодов с рекурсивным кодером и памятью $m = 3$: $(13, 15)$ и $(13, 17)$. Число в таблице, находящееся на пересечении i-й строки и j-го столбца, обозначает количество кодовых слов в подмножестве $\mathcal{C}_f$ с весом, указанным в i-й строке, и длиной ненулевого пути, указанной в j-м столбце. Полужирным шрифтом выделены те коэффициенты спектра, для которых веса соответствуют активным расстояниям.

§ 3. Теоретические оценки вероятности ошибки на блок и на бит для сверточных кодов в двоичном симметричном канале

Приведем оценки сверху и снизу для вероятности появления пакета ошибок заданной длины и вероятности ошибочного декодирования терминированного сверточного кода, а также выведем оценку для вероятности ошибки на бит. Предпола-

гается, что передача данных осуществляется через двоичный симметричный канал, а в качестве декодера используется алгоритм Витерби.

3.1. Вероятность появления пакета ошибок. Вероятность появления пакета ошибок и ее оценка были подробно описаны в работе [21]. Приведем здесь вывод и итоговую формулу.

Рассматривается двоичный симметричный канал с декодированием по максимуму правдоподобия с вероятностью ошибки на бит p. Декодер Витерби всегда возвращает слово, принадлежащее исходному коду $\mathcal{C}$. Если при декодировании появился пакет ошибок, то декодированное слово $\boldsymbol{v}' \in \mathcal{C}$ не будет совпадать с переданным $\boldsymbol{v} \in \mathcal{C}$. На месте возникновения пакета ошибок будет ответвление от корректного пути на решетке. Пакеты ошибок считаются независимыми, если между ними есть хотя бы два корректных состояния решетки подряд. Так как сверточный код является линейным, то $\boldsymbol{v} + \boldsymbol{v}' \in \mathcal{C}$, и на месте пакета ошибок в сумме возникнет ненулевой путь на решетке, вес которого определяется спектром активных расстояний $\mathcal{D}_a$ сверточного кода. В декодированном слове возникнет пакет ошибок, если произошло хотя бы $\dfrac{a_j}{2}$ ошибок на позициях единиц в сумме $\boldsymbol{v} + \boldsymbol{v}'$ в месте, соответствующему пакету ошибок, где a_j – минимальный вес пакета ошибок длины j (активное расстояние).

Вероятность появления фиксированного числа ошибок среди нулей и единиц в двоичной последовательности может быть записана так:

$$P(e_{\text{all}}, e_1, j, w, p) = \binom{w}{e_1}\binom{2j-w}{e_{\text{all}}-e_1} p^{e_{\text{all}}}(1-p)^{2j-e_{\text{all}}},$$

где e_{all} – общее число ошибок, e_1 – число ошибок среди единиц, j – длина последовательности в терминах количества кортежей, w – вес последовательности, p – вероятность ошибки на бит в канале.

Как было отмечено ранее, пакет ошибок возникнет при декодировании, если на местах единиц в пакете количество ошибок больше половины веса, а количество ошибок на местах нулей может быть любым. Если число ошибок среди единиц составляет ровно половину веса пакета, то пакет ошибок возникнет с вероятностью $1/2$. Таким образом, вероятность пакета ошибок длины j и веса w для двоичного симметричного канала с вероятностью ошибки на бит p может быть записана следующим образом:

$$P_{\text{burst}}(j, w, p) = \sum_{i>\frac{w}{2}}^{2j} \sum_{i_1>\frac{w}{2}}^{\min(w,i)} P(e_{\text{all}} = i, e_1 = i_1, j, w, p) + \\ + \begin{cases} 0 & \text{для нечетного веса } w, \\ \dfrac{1}{2} \displaystyle\sum_{i=\frac{w}{2}}^{2j-\frac{w}{2}} P\Big(e_{\text{all}} = i, e_1 = \frac{w}{2}, j, w, p\Big) & \text{для четного веса } w. \end{cases} \tag{2}$$

Оценку снизу для вероятности появления пакета ошибок будем считать как вероятность появления наиболее вероятного пакета ошибок, т.е. пакета ошибок с минимальным весом. Тогда оценку сверху можно посчитать как сумму вероятностей появления всех возможных пакетов ошибок. Тогда оценки снизу и сверху для вероятности появления пакета ошибок длины j могут быть заданы с помощью активных расстояний и спектра активных расстояний соответственно:

$$P_{\text{low}}(j, p) = (1-p)^{4m} P_{\text{burst}}(j, w = a_j, p) \tag{3}$$

и

$$P_{\text{up}}(j,p) = \sum_{w_i=w_{\min}}^{w_{\max}} N_{w_i} P_{\text{burst}}(j, w = w_i, p), \tag{4}$$

где $\{w_i, N_{w_i}\} \in \mathcal{D}_{a_j}$, а m – память кода. Для нижней оценки мы также используем множитель $(1-p)^{4m}$ – вероятность того, что $2m$ битов до и после пакета ошибок будут корректными, чтобы исключить ситуации, при которых рассматриваемый участок является частью другого пакета ошибок. Более подробный вывод можно найти в работе [23].

3.2. Вероятность ошибочного декодирования. В общем случае кодовое слово сверточного кода имеет неограниченную длину, но на практике мы имеем дело с конечным объемом данных, поэтому используются усеченные сверточные коды. Мы будем использовать сверточные коды "с нулевым хвостом" (нулевым терминированием), но иногда используются также и циклически усеченные сверточные коды. В канале передаются блоки фиксированной длины, состоящие из кортежей усеченного сверточного кода, которые соответствуют конечному пути по решетке для сверточного кода, причем начальное и конечное состояния решетки нулевые. Так как переход кодера из любого состояния в нулевое гарантированно может произойти за m тактов (при последовательном записывании в память кодера нули), где m – память кода, то терминирование кода приводит к небольшому уменьшению скорости. Таким образом, мы передаем на m информационных битов меньше, что обычно мало по сравнению с длиной блока $L \gg m$.

Будем оценивать вероятность ошибочного декодирования блока длины L с помощью средней доли блоков, пораженных хотя бы одним пакетом ошибок. Нижнюю оценку вероятности ошибочного декодирования мы получим как среднюю долю блоков, в которых появился наиболее вероятный пакет ошибок. Наиболее вероятный пакет ошибок обладает минимальным возможным весом $w_{\min} = \min\limits_j a_j$. Вероятность ошибочного декодирования блока обозначается через FER (Frame Error Rate). Для вероятности появления наиболее вероятного пакета ошибок мы будем использовать нижнюю оценку (3) для $P_{\text{low}}(j,p)$. Чтобы получить более точную нижнюю оценку, мы умножаем вероятность появления в отдельно взятом кортеже наиболее вероятного пакета ошибок на количество таких различных пакетов $N_{w_{\min}}$, и умножив на длину блока L "в терминах количества кортежей", получим нижнюю оценку вероятности ошибочного декодирования:

$$\text{FER}_{\text{low}}(p) = L N_{w_{\min}} P_{\text{low}}(j = j_{w_{\min}}, p), \tag{5}$$

где $j_{w_{\min}}$ – длина пакета ошибок с минимальным весом, $\{w_{\min}, N_{w_{\min}}\} \in \mathcal{D}_a$.

Верхнюю оценку мы выводим как аддитивную границу, суммируя средние доли блоков с различными ошибками по всем возможным длинам пакетов ошибок. Для вероятности появления пакета будем использовать верхнюю оценку (4) для $P_{\text{up}}(j,p)$. Так как в нее уже входит количество пакетов, то нет необходимости в дополнительном множителе. Таким образом, получаем верхнюю оценку вероятности ошибочного декодирования:

$$\text{FER}_{\text{up}}(p) = \min\left\{1, \sum_{\ell=m+1}^{L} L P_{\text{up}}(j = \ell, p)\right\}, \tag{6}$$

где m – память кода.

Верхняя оценка может быть несколько уточнена, но так как вероятность блока, содержащего два и более пакетов ошибок, меньше на несколько порядков (так как

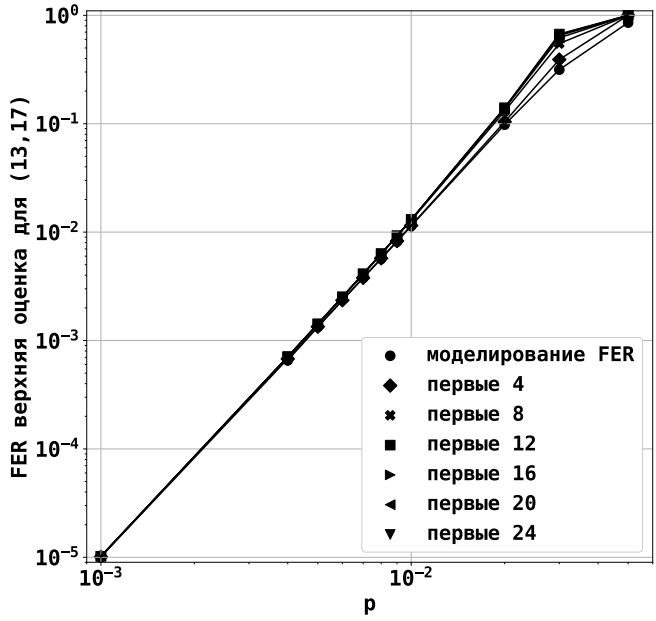


Рис. 2. Зависимость между длиной пакета j и верхней оценкой вероятности ошибочного декодирования для сверточного кода $(13, 17)$ с рекурсивным систематическим кодером и памятью $m = 3$

перемножаются вероятности пакетов ошибок по отдельности), то для упрощения выражения мы учитываем такие блоки несколько раз в нашей аддитивной оценке. Сумма в формуле (6) берется по всем возможным длинам пакетов ошибок в блоке, но для практического использования достаточно взять сумму только по самым коротким пакетам ошибок, что видно из рис. 2. Даже при рассмотрении небольшого числа самых коротких пакетов ошибок оценка $\mathrm{FER}_{\mathrm{up}}(p)$ лежит выше значений вероятности ошибочного декодирования, полученных моделированием. Для малых p рассмотрение большего числа различных длин пакетов ошибок практически не влияет на точность оценки.

3.3. Вероятность ошибки на бит. Выходная вероятность ошибки на бит для сверточного кода – это доля битов с ошибками в декодированном слове. Некорректные биты появляются только в местах пакетов ошибок.

Мы будем оценивать вероятность ошибки на бит в декодированном слове как среднюю долю битов с ошибками. Вероятность ошибочного декодирования бита обозначается через BER (Bit Error Rate). Мы выводим нижнюю оценку вероятности ошибки на бит как долю ошибочных битов в предположении появления наиболее вероятных пакетов ошибок. Тогда нижняя оценка – это вес наиболее вероятного пакета ошибок, умноженный на вероятность его появления в каком-либо бите. Для уточнения нижней границы мы используем в качестве вероятности появления пакета ошибок с минимальным весом $w_{\min} = \min\limits_{j} a_j$ – нижнюю оценку вероятности появления одного такого пакета в кортеже (3), умноженную на количество таких пакетов (они равновероятные) и нормированную по числу битов в кортеже. Таким образом, предложенная нижняя оценка вероятности ошибки на бит имеет вид

$$\mathrm{BER}_{\mathrm{low}}(p) = \frac{w_{\min} N_{w_{\min}}}{2} P_{\mathrm{low}}(j = j_{w_{\min}}, p), \tag{7}$$

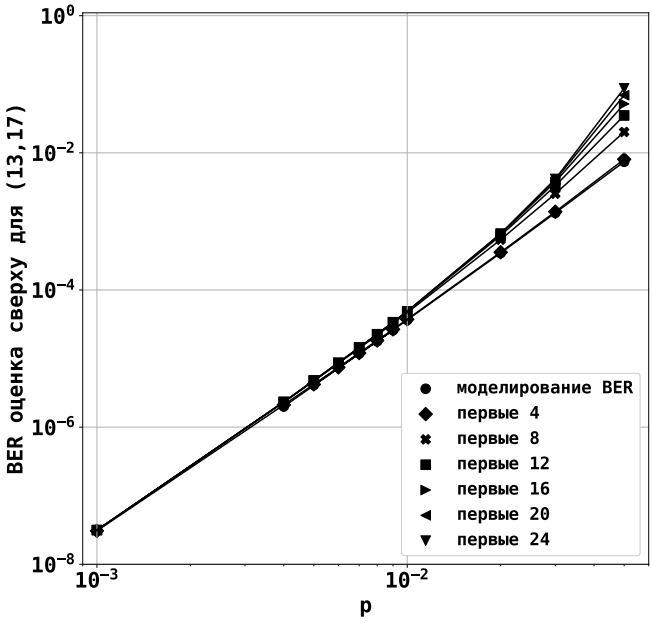


Рис. 3. Зависимость между длиной пакета j и верхней оценкой вероятности ошибки на бит для сверточного кода $(13, 17)$ с рекурсивным систематическим кодером и памятью $m = 3$

здесь $N_{w_{\min}}$ – число пакетов ошибок с минимальным весом, $j_{w_{\min}}$ – длина пакета ошибок с минимальным весом, $\{w_{\min}, N_{w_{\min}}\} \in \mathcal{D}_a$, а знаменатель $n = 2$ – число битов в кортеже.

Верхнюю оценку вероятности ошибки на бит мы строим как аддитивную оценку с помощью средних долей ошибочных битов по всем длинам пакетов ошибок с наибольшим возможным весом. Верхняя оценка вероятности появления пакета ошибок в кортеже (4) уже учитывает количество таких пакетов, поэтому нет необходимости в дополнительном множителе. Как и для случая $\mathrm{BER}_{\mathrm{low}}$, мы нормируем оценку по числу битов в выходном кортеже $n = 2$, чтобы получить вероятность появления пакета в конкретном бите. Тогда верхняя оценка вероятности ошибки на бит:

$$\mathrm{BER}_{\mathrm{up}}(p) = \min\left\{1, \sum_j \frac{\max\limits_{w_j \in \mathcal{D}_{a_j}} w_j}{2} P_{\mathrm{up}}(j, p)\right\}. \tag{8}$$

В формуле (8) берется минимум из 1 и оценки, так как оценка аддитивна и может превысить 1. В формуле (8) сумма берется по всем возможным j, в то время как для практических расчетов достаточно взять сумму лишь по самым коротким длинам. Действительно, на рис. 3 видно, что верхняя оценка сходится к одному значению с увеличением предела суммирования j в сумме (8). Более того, даже для небольшого количества рассматриваемых длин пакетов ошибок оценка лежит выше реальных значений вероятности ошибки на бит. Для малых вероятностей ошибки на бит p ($p \leqslant 0{,}01$) оказывается достаточным взять всего несколько наиболее коротких пакетов ошибок, что еще более снижает сложность вычисления оценки.

3.4. Сложность вычисления теоретических оценок. Для расчета предложенных теоретических оценок необходимо вычислить активные расстояния и их спектр для

сверточного кода. Отметим, что для сверточного кода необходимо вычислить метрические характеристики всего один раз.

Расчет активных расстояний a_j для длины j основан на алгоритме Витерби и происходит следующим способом:

1. В начальный момент времени $t = 0$ инициализируем один путь в нулевом состоянии решетки длины 0.
2. В момент времени $t > 0$ продолжаем все сохраненные пути в решетке из узлов в момент времени $t - 1$ в узлы в момент времени t. Каждый путь может быть продолжен 2^k, $k = 1$, способами, не считая путей в нулевом состоянии на решетке. Из нулевого состояния решетки продолжаем путь только в ненулевое состояние решетки. Считаем вес полученных путей, всего $2^{m+k} - 1$, $k = 1$, путей (при $t > m$, где m – память кода).
3. Для каждого состояния в решетке в момент времени t сохраняем только путь с наименьшим весом, оставляя 2^m путей (при $t \geqslant m$).
4. Если $t = j$, то возвращаем вес пути, который сохранен для нулевого состояния в момент времени $t = j$. Если $t < j$, переходим к шагу 2.

Удобно представлять процесс вычисления активных расстояний как движение по решетке вправо. Так как необходимо найти ненулевой путь в решетке длины j с минимальным весом, то достаточно для каждого узла решетки в каждый момент времени хранить только "лучший" путь, т.е. путь минимального веса, входящий слева в этот узел. Сложность такого алгоритма зависит от длины ненулевого пути j и от количества состояний 2^m, где m – память кода. Из алгоритма подсчета следует, что сложность вычисления активных расстояний a_j составляет $\mathcal{O}(j(2^{m+1} - 1))$. Отметим, что вычисляя активные расстояния для длины j, мы при этом можем вычислить еще активные расстояния для всех длин, меньших j.

Расчет спектра активных расстояний имеет бóльшую алгоритмическую сложность, которая экспоненциально растет с ростом длины ненулевого пути j, для которой необходимо посчитать спектр. Спектр также вычисляется с помощью решетки, только теперь мы не отбрасываем пути с бóльшим весом, а храним все возможные пути в решетке. При продолжении путей путь из нулевого состояния также не продолжается в нулевое состояние. Вычислительную сложность алгоритма можно записать как $\mathcal{O}(2^j(2^{m+1} - 1))$, где m – память кода. Из формулы видно, что вычисление спектра активных расстояний для больших расстояний трудоемко и быстро растет с увеличением длины j.

Теперь оценим сложность вычисления предложенных оценок вероятности ошибки для сверточных кодов, если метрические характеристики кода (активные расстояния или спектр активных расстояний) уже известны. Сложность вычисления нижней $P_{\text{low}}(j, p)$ (3) и верхней $P_{\text{up}}(j, p)$ (4) оценок вероятности появления пакета ошибок длины j оценим как максимальное количество вычислений $P(e_{\text{all}}, e_1, j, w, p)$:

$$\mathcal{O}\Big(\Big(2j - \frac{a_j}{2}\Big)\frac{a_j}{2}\Big) \quad \text{и} \quad \mathcal{O}\Big(\Big(2j - \frac{w_{\min}}{2}\Big)\frac{w_{\max}}{2}(w_{\max} - w_{\min})\Big), \quad \{w_{\min}, w_{\max}\} \in \mathcal{D}_{a_j}.$$

Сложность вычисления нижних оценок (5) и (7) для $\text{FER}_{\text{low}}(p)$ и $\text{BER}_{\text{low}}(p)$ не отличаются от сложности расчета $P_{\text{low}}(j = j_{w_{\min}} p)$, где $j_{w_{\min}}$ – длина пакета ошибок с минимальным весом. Сложность вычисления верхних оценок (6) и (8) для $\text{FER}_{\text{up}}(p)$ и $\text{BER}_{\text{up}}(p)$ совпадает со сложностью вычисления верхней оценки появления пакета ошибок $P_{\text{up}}(j, p)$, умноженной на количество рассмотренных наименьших различных возможных весов пакетов ошибок. На практике достаточно рассмотреть порядка 5–8 весов, как видно из рис. 3. Таким образом, хотя сложность расчета теоретических оценок зависит от вероятности ошибки на бит p, но для практически значимых значений p, для получения достаточно точных значений величин вероятности ошибки

на бит, достаточно ограничиться небольшим числом слагаемых. Кроме того, следует отметить, что сложность расчета теоретических оценок линейна по j.

3.5. Экспериментальные результаты. Представим экспериментальные результаты и их сравнение с теоретическими оценками. В экспериментах рассматривался двоичный симметричный канал с вероятностью ошибки на бит p. По каналу передавались кодовые слова усеченного сверточного кода с рекурсивным систематическим кодером скорости 1/2 длины 1000 кортежей (2000 бит). Рассматривались различные сверточные коды с различной памятью.

Приведем результаты для вероятности появления пакета ошибок. Графики на рис. 4 приведены для трех различных вероятностей ошибки на бит p: 0,03, 0,008 и 0,004. Для вероятности $p = 0{,}004$ приведено существенно меньше точек, полученных с помощью моделирования, так как для получения экспериментальных оценок время, требуемое на моделирование, резко возрастает.

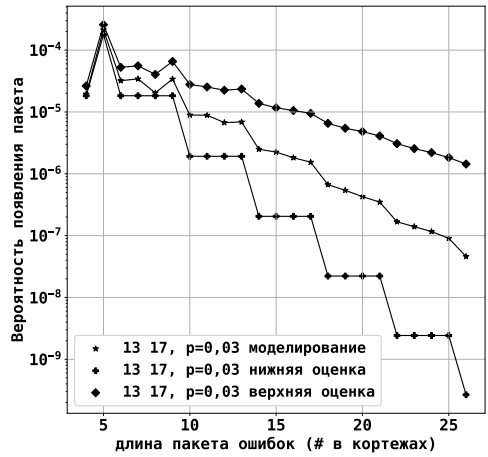
Из рис. 4 видно, что ближе всего теоретические оценки лежат к результатам моделирования при малых длинах пакетов, для которых существует всего одно или несколько слов, в которые они могут перейти. С увеличением длины пакета резко увеличивается число возможных слов, в которые может перейти слово при декодировании, что наглядно иллюстрируется таблицами 1 и 2 спектра активных расстояний. Отметим также, что пакеты ошибок небольшой длины имеют наибольшую вероятность и, соответственно, имеют наибольший вклад в вероятность ошибочного декодирования блока и в вероятность ошибки на бит. По графикам на рис. 4 также можем отметить, что с уменьшением вероятности ошибки на бит p оценки сходятся, что объясняется экспоненциальным уменьшением вероятностей длинных пакетов ошибок. Таким образом, для небольших p предложенные в статье оценки становятся достаточно точными даже для больших длин пакетов ошибок.

Принято считать, что распределение пакетов ошибок можно приблизить с помощью геометрического распределения [26, 27]. Это справедливо только для наименее вероятных ошибок, т.е. геометрическое распределение хорошо приближает вероятности появления длинных пакетов ошибок и плохо приближает наиболее вероятные пакеты с наименьшей длиной.

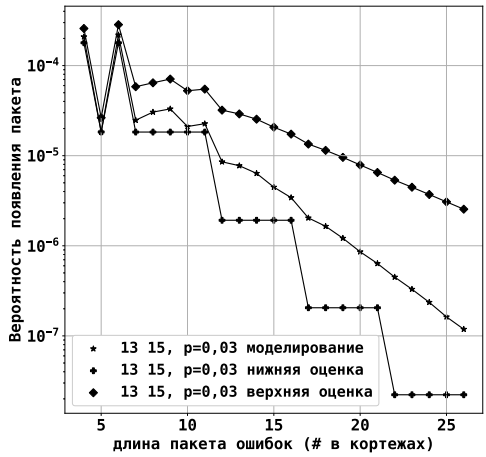
На рис. 5, 6 представлены вероятности ошибочного декодирования (FER) и выходные вероятности ошибки на бит (BER) для двух сверточных кодов $(13, 17)$ и $(13, 15)$ с памятью $m = 3$. Мы также привели на рис. 7 BER для кода $(117, 155)$ с памятью $m = 6$, чтобы показать, что наши оценки применимы и для кодов с большей памятью. Для сравнения также приведены классические аддитивные верхние оценки FER и BER из работ [14, 15], вычисленные с помощью производящей функции спектра кода и функции спектра ошибки на бит соответственно.

Оценки вероятности ошибочного декодирования были также представлены в работе [23]. Эти результаты можно сравнить с результатами из работы [20] для кода $(13, 17)$, и очевидно, что результаты практически совпадают. Но метод из работы [20] вычислительно трудоемкий, особенно с увеличением памяти кодера. Более того, в работе говорится о том, что предложенный метод не применим к кодам с памятью больше 4.

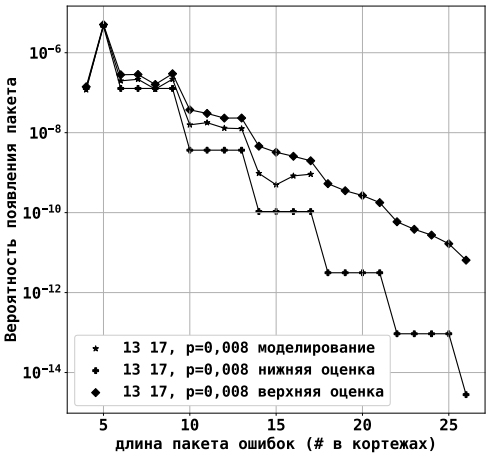
Предложенный в данной статье подход, основанный на активных расстояниях сверточного кода и их спектре, позволяет вычислить оценки для кодов большей памяти, не требуя особых вычислительных ресурсов. Для вычисления точных оценок не требуется рассматривать пакеты ошибок большой длины, достаточно рассмотреть несколько пакетов с наименьшими весами. Программы для вычисления активных расстояний и их спектра могут быть найдены на портале GitHub [28]. Здесь мы также приводим пример выходной вероятности ошибки на бит для сверточного кода $(117, 155)$ с памятью $m = 6$ на рис. 7, где для сравнения также приведены верхние оценки BER из работ [14, 15].



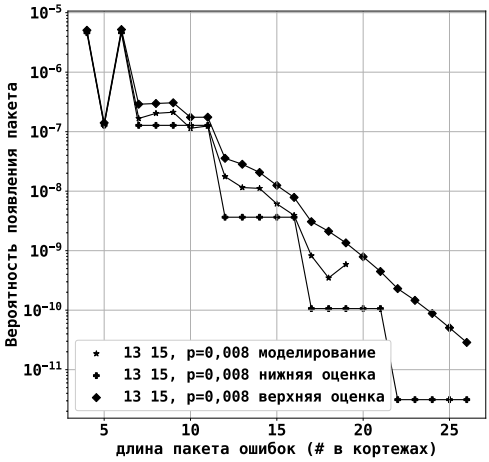
a)



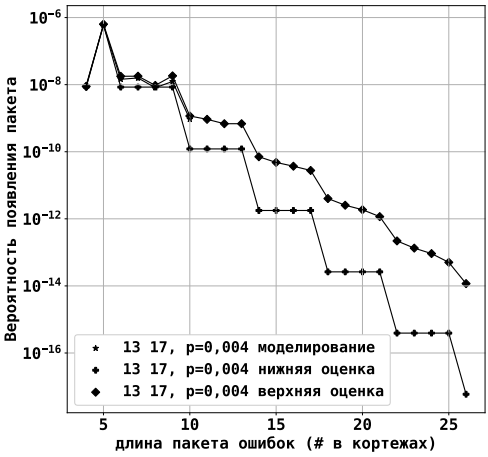
b)



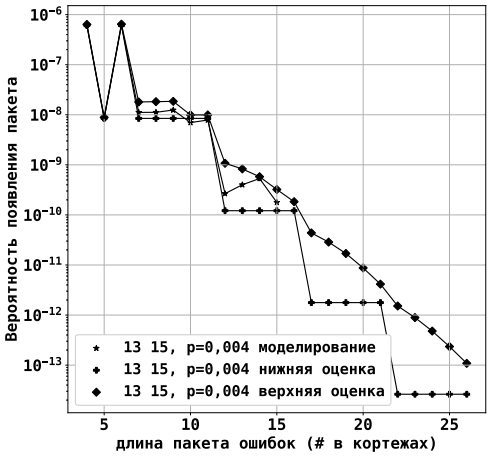
c)



d)



e)



f)

Рис. 4. Вероятностное распределение появления пакетов ошибок для сверточного кода с рекурсивным систематическим кодером и памятью $m = 3$: a) $(13, 17)$, $p = 0{,}03$, b) $(13, 15)$, $p = 0{,}03$, c) $(13, 17)$, $p = 0{,}008$, d) $(13, 15)$, $p = 0{,}008$, e) $(13, 17)$, $p = 0{,}004$, f) $(13, 15)$, $p = 0{,}004$

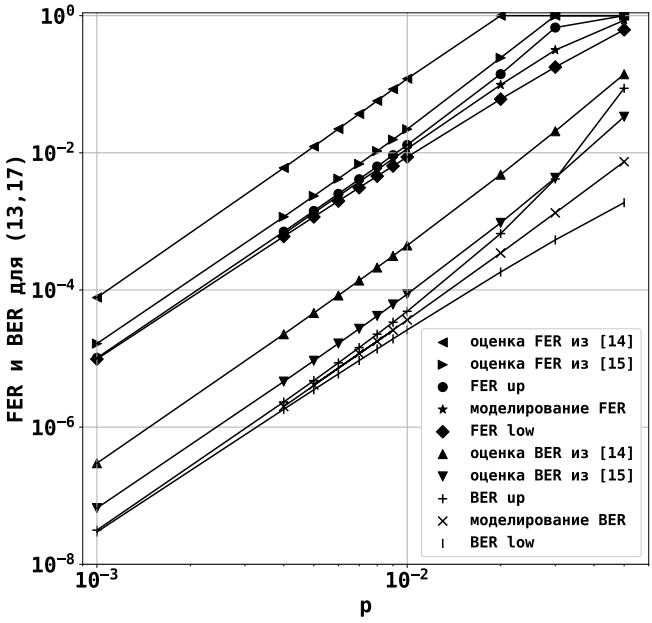


Рис. 5. Вероятности ошибочного декодирования и ошибки на бит для сверточного кода $(13, 17)$ с рекурсивным систематическим кодером и памятью $m = 3$

Из рис. 5–7 видно, что верхняя и нижняя оценки приближаются друг к другу при небольших вероятностях ошибки p. Таким образом, для этих вероятностей теоретические оценки достаточно точны, и нет необходимости использовать метод Монте-Карло для получения экспериментальных значений, что на практике может быть достаточно трудоемко для малых значений p. Мы также сравнили наши верхние оценки для вероятности ошибки на бит и для вероятности ошибочного декодирования с классическими верхними оценками из [14, 15], полученными с помощью производящих функций.

3.6. Сравнение с другими границами. Оценки вероятности ошибочного декодирования были также представлены в работе [23]. Эти результаты можно сравнить с результатами из работы [20] для кода $(13, 17)$, и очевидно, что результаты практически совпадают. Но метод из работы [20] вычислительно трудоемкий, особенно с увеличением памяти кодера. Более того, в работе говорится о том, что предложенный метод не применим к кодам с памятью больше 4, тогда как мы показали, что расчет наших оценок возможен и для кодов с большей памятью.

Предложенный в нашей статье подход, основанный на активных расстояниях сверточного кода и их спектре, позволяет вычислить оценки для кодов большей памяти, не требуя особых вычислительных ресурсов. Для вычисления точных оценок не требуется рассматривать пакеты ошибок большой длины, достаточно рассмотреть несколько пакетов с наименьшими весами. Программы для вычисления активных расстояний и их спектра могут быть найдены на портале GitHub [28].

На рис. 5–7 приведено сравнение предложенных нами оценок с классическими границами из работ [14,15]. Из результатов видно, что предложенная в нашей статье аддитивная верхняя оценка лежит ближе либо сравнима с классическими. Оценка из работы [15] лежит ближе к экспериментальным данным и к нашей границе, чем оценка из работы [14], что можно объяснить более точной оценкой сверху с помощью границы Миберга для вероятности ошибки в двоичном симметричном канале.

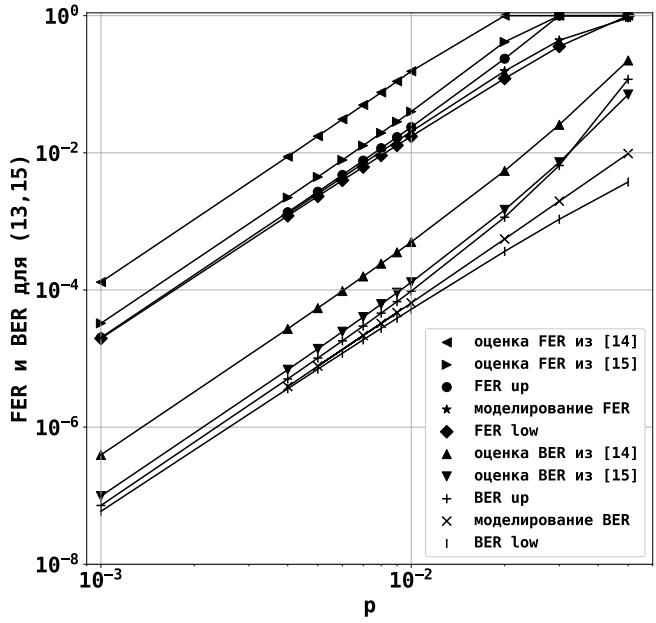


Рис. 6. Вероятности ошибочного декодирования и ошибки на бит для сверточного кода $(13, 15)$ с рекурсивным систематическим кодером и памятью $m = 3$

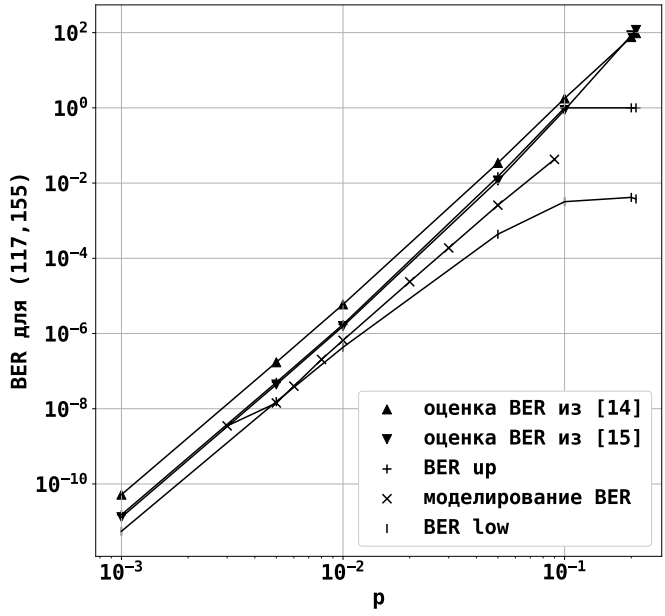


Рис. 7. Вероятность ошибки на бит для сверточного кода $(117, 155)$ с рекурсивным систематическим кодером и памятью $m = 6$

Используемые нами оценки для вероятности появления пакета ошибок и активные расстояния позволяют вычислить оценки для FER и BER более точно. Отдельно отметим, что мы также предлагаем оценку снизу, что не было ранее предложено

в работах, и оценка снизу оказывается очень близка к экспериментальным результатам.

§ 4. Заключение

В статье получены наиболее важные теоретические оценки эффективности сверточного кода в двоичном симметричном канале. А именно, получены верхние и нижние оценки вероятности появления пакета ошибок, ошибочного декодирования и вероятности ошибки на бит. Предложенные оценки основаны на активных расстояниях и спектре активных расстояний и справедливы для сверточных кодов скорости $1/2$ с декодированием Витерби. Полученные результаты могут быть обобщены для сверточных кодов скорости $1/n$. Проведен анализ сложности вычисления метрических характеристик и теоретических оценок. Показано, что для заданной длины сложность вычисления активных расстояний экспоненциально зависит от памяти кода и линейно от длины, в то время как сложность вычисления спектра активных расстояний экспоненциально зависит и от памяти кода, и от длины. Сложность вычисления теоретических оценок при известных метрических характеристиках линейно зависит от длины наиболее коротких пакетов ошибок и не зависит от входной вероятности ошибки на бит в канале. Также рассмотрены различные коды с разной памятью, и для них представлено сравнение теоретических и экспериментальных результатов. Предложенные оценки наиболее точны при небольших входных вероятностях ошибки на бит.

В дальнейшем с помощью активных расстояний и его спектра могут быть выведены оценки эффективности сверточных кодов также и для канала с гауссовским шумом.

СПИСОК ЛИТЕРАТУРЫ

1. *Elias P.* Coding for Noisy Channels // IRE Conv. Rec. 1955. V. 4. P. 37–46. Reprinted in: Key Papers in the Development of Information Theory. New York: IEEE Press, 1974. P. 102–111.
2. *Yang C., Zhan M., Deng Y., Wang M., Luo X.H., Zeng J.* Error-Correcting Performance Comparison for Polar Codes, LDPC Codes and Convolutional Codes in High-Performance Wireless // Proc. 6th Int. Conf. on Information, Cybernetics, and Computational Social Systems (ICCSS'2019). Chongqing, China. Sept. 27–30, 2019. P. 258–262. https://doi.org/10.1109/ICCSS48103.2019.9115442
3. *Deng Y., Zhan M., Wang M., Yang C., Luo X., Zeng J., Guo J.* Comparing Decoding Performance of LDPC Codes and Convolutional Codes for Short Packet Transmission // Proc. IEEE 17th Int. Conf. on Industrial Informatics (INDIN'2019). Helsinki, Finland. July 22–25, 2019. V. 1. P. 1751–1755. https://doi.org/10.1109/INDIN41052.2019.8972062
4. *Tahir B., Schwarz S., Rupp M.* BER Comparison between Convolutional, Turbo, LDPC, and Polar Codes // Proc. 24th Int. Conf. on Telecommunications (ICT'2017). Limassol, Cyprus. May 3–5, 2017. P. 1–7. https://doi.org/10.1109/ICT.2017.7998249
5. *Rurik W., Mazumdar A.* Hamming Codes as Error-Reducing Codes // Proc. 2016 IEEE Information Theory Workshop (ITW'2006). Cambridge, UK. Sept. 11–14, 2016. P. 404–408. https://doi.org/10.1109/ITW.2016.7606865
6. *Liu K., García-Frías J.* Error Floor Analysis in LDGM Codes // Proc. 2010 IEEE Int. Symp. on Information Theory (ISIT'2010). Austin, TX, USA. June 13–18, 2010. P. 734–738. https://doi.org/10.1109/ISIT.2010.5513607
7. *Gao W., Polyanskiy Y.* On the Bit Error Rate of Repeated Error-Correcting Codes // Proc. 48th Annu. Conf. on Information Sciences and Systems (CISS'2014). Princeton, NJ, USA. Mar. 19–21, 2014. P. 1–6. https://doi.org/10.1109/CISS.2014.6814087
8. *Höst S., Johannesson R., Zyablov V.V.* Woven Convolutional Codes. I. Encoder Properties // IEEE Trans. Inform. Theory. 2002. V. 48. № 1. P. 149–161. https://doi.org/10.1109/18.971745

9. *Freudenberger J., Bossert M., Shavgulidze S., Zyablov V.* Woven Turbo Codes // Proc. 7th Int. Workshop on Algebraic and Combinatorial Coding Theory (ACCT'2000). Bansko, Bulgaria. June 18–24, 2000. P. 145–150. Available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.23.345&rep=rep1&type=pdf
10. *Benedetto S., Montorsi G.* Design of Parallel Concatenated Convolutional Codes // IEEE Trans. Commun. 1996. V. 44. № 5. P. 591–600. https://doi.org/10.1109/26.494303
11. *Berrou C., Glavieux A., Thitimajshima P.* Near Shannon Limit Error-Correcting Coding and Decoding: Turbo-Codes. 1 // Proc. IEEE Int. Conf. on Communications (ICC'93). Geneva, Switzerland. May 23–26, 1993. V. 2. P. 1064–1070. https://doi.org/10.1109/ICC.1993.397441
12. *Douillard C., Berrou C.* Turbo Codes with Rate-$m/(m+1)$ Constituent Convolutional Codes // IEEE Trans. Commun. 2005. V. 53. № 10. P. 1630–1638. https://doi.org/10.1109/TCOMM.2005.857165
13. *Zhang Q., Liu A., Zhang Y., Liang X.* Practical Design and Decoding of Parallel Concatenated Structure for Systematic Polar Codes // IEEE Trans. Commun. 2016. V. 64. № 2. P. 456–466. https://doi.org/10.1109/TCOMM.2015.2502246
14. *Viterbi A.J.* Convolutional Codes and Their Performance in Communication Systems // IEEE Trans. Commun. Technol. 1971. V. 19. № 5. P. 751–772. https://doi.org/10.1109/TCOM.1971.1090700
15. *van De Meeberg L.* A Tightened Upper Bound on the Error Probability of Binary Convolutional Codes with Viterbi Decoding // IEEE Trans. Inform. Theory. 1974. V. 20. № 3. P. 389–391. https://doi.org/10.1109/TIT.1974.1055216
16. *Herro M., Hu L., Nowack J.* Bit Error Probability Calculations for Convolutional Codes with Short Constraint Lengths on Very Noisy Channels // IEEE Trans. Commun. 1988. V. 36. № 7. P. 885–888. https://doi.org/10.1109/26.2819
17. *Chiaraluce F., Gambi E., Mazzone M., Pierleoni P.* A Technique to Evaluate an Exact Formula for the Bit Error Rate of Convolutional Codes in Case of Finite Length Words // Proc. IEEE Region 10 Annu. Conf. on Speech and Image Technologies for Computing and Telecommunications (IEEE TENCON'97). Queensland Univ. of Technology, Brisbane, Australia. Dec. 2–4, 1997. V. 1. P. 113–116. https://doi.org/10.1109/TENCON.1997.647271
18. *Forney G.* The Viterbi Algorithm // Proc. IEEE. 1973. V. 61. № 3. P. 268–278. https://doi.org/10.1109/PROC.1973.9030
19. *Yoshikawa H.* Theoretical Analysis of Bit Error Probability for Punctured Convolutional Codes // Proc. 2012 IEEE Int. Sympos. on Information Theory and Its Applications (ISITA'2012). Honolulu, HI, USA. Oct. 28–31, 2012. P. 658–661.
20. *Bocharova I.E., Hug F., Johannesson R., Kudryashov B.D.* A Closed-Form Expression for the Exact Bit Error Probability for Viterbi Decoding of Convolutional Codes // IEEE Trans. Inform. Theory. 2012. V. 58. № 7. P. 4635–4644. https://doi.org/10.1109/TIT.2012.2193375
21. *Smeshko A., Ivanov F., Zyablov V.* Theoretical Estimates of Burst Error Probability for Convolutional Codes // Proc. 2020 Int. Symp. on Information Theory and Its Applications (ISITA'2020). Kapolei, HI, USA. Oct. 24–27, 2020. P. 136–140.
22. *Smeshko A., Ivanov F., Zyablov V.* The Influence of Active Distances on the Distribution of Bursts // Proc. XVI Int. Symp. "Problems of Redundancy in Information and Control Systems" (REDUNDANCY'2019). Moscow, Russia. Oct. 21–25, 2019. P. 110–114. https://doi.org/10.1109/REDUNDANCY48165.2019.9003349
23. *Smeshko A., Ivanov F., Zyablov V.* Upper and Lower Estimates of Frame Error Rate for Convolutional Codes // Proc. 2020 Int. Symp. on Information Theory and Its Applications (ISITA'2020). Kapolei, HI, USA. Oct. 24–27, 2020. P. 160–164.
24. *Кудряшов Б.Д.* Основы теории кодирования. СПб.: БХВ-Петербург, 2016.
25. *Höst S., Johannesson R., Zigangirov K., Zyablov V.* Active Distances for Convolutional Codes // IEEE Trans. Inform. Theory. 1999. V. 45. № 2. P. 658–669. https://doi.org/10.1109/18.749009
26. *Miller R.L., Deutsch L.J., Butman S.A.* On the Error Statistics of Viterbi Decoding and the Performance of Concatenated Codes // NASA STI/Recon Tech. Rep. N 81-33364. Sept. 1, 1981. Available at https://archive.org/details/nasa_techdoc_19810024821

27. *Justesen J., Andersen J.* Critical Lengths of Error Events in Convolutional Codes // IEEE Trans. Inform. Theory. 1998. V. 44. № 4. P. 1608–1611. https://doi.org/10.1109/18.681339
28. https://github.com/smeshk/Active-distances-their-spectrum-and-estimates-for-convolutional-code [GitHub online repository]. 2021.

Курмукова Анастасия Андреевна
Институт проблем передачи информации
им. А.А. Харкевича РАН
Национальный исследовательский университет
"Высшая школа экономики"
Сколковский институт науки и технологий
Anastasiia.Kurmukova@skoltech.ru
Иванов Федор Ильич
Институт проблем передачи информации
им. А.А. Харкевича РАН
Национальный исследовательский университет
"Высшая школа экономики"
fivanov@hse.ru
Зяблов Виктор Васильевич
Институт проблем передачи информации
им. А.А. Харкевича РАН
zyablov@iitp.ru

Поступила в редакцию
12.12.2021
После доработки
08.03.2022
Принята к публикации
11.03.2022

ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 58 **2022** **Вып. 2**

УДК 621.391 : 517.938 : 519.722

© 2022 г. **Г.Д. Дворкин**

ГЕОМЕТРИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ ЭНТРОПИИ ДЛЯ СИСТЕМ ДИКА

Рассматривается связь метрической энтропии с локальной скоростью деформации границ (ЛСДГ) в символическом случае. Показывается равенство ЛСДГ как предела в среднем и энтропии для широкого класса мер на системах Дика.

Ключевые слова: метрическая энтропия, локальная скорость деформации границ, символическая система, несинхронизованная система, инвариантная эргодическая мера, правильная скобочная последовательность, система Дика.

DOI: 10.31857/S0555292322020041, **EDN:** DZCGLF

§ 1. Введение

Будем рассматривать энтропию Колмогорова – Синая как предел некоторой величины (скорости деформации границ), имеющей простой геометрический смысл. Такое представление энтропии будем называть геометрической интерпретацией.

Этот подход был предложен физиком Г.М. Заславским [1], который в 1984 г. выдвинул гипотезу (точнее, сформулировал утверждение), что объем границы множества в фазовом пространстве растет экспоненциально по времени с показателем, равным метрической энтропии. Первый математический результат в этом направлении был получен Б.М. Гуревичем [2] для символических марковских систем и совместно с С.А. Комечем [3,4] обобщен на существенно более широкий класс символических систем, а также на некоторые гладкие (системы Аносова). В настоящей статье показывается возможность геометрической интерпретации энтропии для важного класса несинхронизованных систем, называемого системами (сдвигами) Дика.

Статья является продолжением работы автора [5], в которой рассматривалась геометрическая интерпретация энтропии для систем Дика, но лишь в контексте одной конкретной меры. Здесь же показывается возможность такого подхода для широкого класса мер на сдвигах Дика. Вместе с тем, результаты из [5] и настоящей статьи независимы (см. замечание 2).

§ 2. Определения и обозначения

Пусть A – конечное множество, и пусть $A^{\mathbb{Z}}$ – пространство бесконечных двусторонних последовательностей символов из A.

Множество A будем называть алфавитом, а любую конечную последовательность символов (букв) из A – (конечным) словом этого алфавита. Количество букв в слове w называют его длиной и обозначают $|w|$. Слово нулевой длины называют пустым.

Элементы $A^{\mathbb{Z}}$ будем называть бесконечными словами, при этом в данной терминологии бесконечные слова словами не являются.

Определение 1. Назовем *полным сдвигом* пару $(A^{\mathbb{Z}}, T)$, где A – конечное множество, а T – сдвиг на шаг вправо в пространстве последовательностей $A^{\mathbb{Z}}$.

Для $x \in A^{\mathbb{Z}}$ и целых a и b, таких что $a \leqslant b$, обозначим слово $(x(a), x(a+1), \ldots, x(b))$ через $x_{[a;b]}$. Если слово w равно $x_{[a;b]}$ для некоторых a и b, то будем говорить, что w является *подсловом* бесконечного слова x (аналогично определяется подслово конечного слова). Вместо $x_{[a;a]}$ будем писать $x_{[a]}$.

Введем метрику d на $A^{\mathbb{Z}}$: примем $d(x,x) = 0$, а для несовпадающих точек x и y положим $d(x,y) = (1/2)^m$, где $m = m(x,y)$ – целое неотрицательное число, равное нулю, если $x_{[0]} \neq y_{[0]}$, и удовлетворяющее условиям $x_{[-(m-1);m-1]} = y_{[-(m-1);m-1]}$, $x_{[-m;m]} \neq y_{[-m;m]}$ в противном случае. Эта метрика порождает топологию, которая далее считается фиксированной.

Определение 2. Будем называть *символической системой* пару (X,T), где X – замкнутое T-инвариантное подмножество $A^{\mathbb{Z}}$ с индуцированной топологией, а также первый элемент этой пары, считая сдвиг и топологию заданными по умолчанию.

Пусть далее в этом параграфе X – символическая система.

Множество всех подслов всех $x \in X$ называется *языком* символической системы X и обозначается $W(X)$; кроме того, обозначим через $W_n(X)$ множество всех слов из языка длины n. *Префикс* слова – любое его подслово, начинающееся с начала этого слова. *Суффикс* слова – любое его подслово, заканчивающееся в конце этого слова.

Определение 3. Множество

$$\{w\}_c^X := \big\{x \in X : \; x_{[c;c+|w|-1]} = w\big\}$$

называется *цилиндром* в X с *основанием* $w \in W(X)$. Всюду далее

$$\{w\}_c := \{w\}_c^X, \quad \big\{x_{[a;b]}\big\} := \big\{x_{[a;b]}\big\}_a.$$

Рассмотрим T-инвариантную борелевскую вероятностную меру μ на X. Под мерой $\mu(w)$ слова $w \in W(X)$ понимаем меру цилиндра $\{w\}_0$ (из T-инвариантности меры очевидно, что $\mu(\{w\}_a) = \mu(\{w\}_0)$ для любого целого a).

Введем следующие обозначения:

$k(\varepsilon) := \max\{\ell \in \mathbb{Z} : \; (1/2)^{(\ell+1)} \leqslant \varepsilon\}$ (часто будем писать просто k);

$O_\varepsilon(x)$ – открытая ε-окрестность точки x в рассматриваемом пространстве;

$O_\varepsilon(C)$ – открытая ε-окрестность множества C в рассматриваемом пространстве.

Определение 4. *Метрическая энтропия* (энтропия Колмогорова – Синая) символической динамической системы (X,T) с инвариантной мерой μ может быть определена как средняя условная энтропия настоящего при условии прошлого (общее определение для произвольной динамической системы см. в [6, 7]):

$$h_\mu(X,T) = \lim_{n\to\infty} H_\mu(x_0 \,|\, x_{-1}, x_{-2}, \ldots, x_{-n}).$$

Определение 5. Для $x \in X$ определим *локальную скорость деформации границ* (ЛСДГ):

$$P_{X,T,\mu}(x) := \lim_{\varepsilon\to 0} \frac{1}{n(\varepsilon)} \log\left(\frac{\mu\big(O_\varepsilon\big(T^{n(\varepsilon)}(O_\varepsilon(x))\big)\big)}{\mu(O_\varepsilon(x))}\right),$$

где $n(\varepsilon)$ – положительная целочисленная функция со свойствами

$$n(\varepsilon) \to \infty \quad \text{и} \quad n(\varepsilon) = o(|\log(\varepsilon)|) = o(k(\varepsilon)) \quad \text{при } \varepsilon \to 0$$

(часто будем писать просто n). Такие функции n будем называть *медленными*. Допредельное выражение будем обозначать через $P^\varepsilon_{X,T,\mu}(x)$ и называть *ε-ЛСДГ*.

Определение 6. Символическая система X *синхронизована*, если она транзитивна и существует $w \in W(X)$, такое что для любых $u, v \in W(X)$, если $uw, wv \in \in W(X)$, то и $uwv \in W(X)$. Такое слово w называется *волшебным* или *синхронизующим*.

Определение 7. Слово $w \in W(X)$ назовем *n-синхронизующим*, если для любых $u, v \in W_n(X)$ условие $uw \in W(X)$ и $wv \in W(X)$ влечет $uwv \in W(X)$.

Определение 8. *Полная правильная скобочная последовательность* (расстановка) – конечная последовательность скобок, приводящаяся к пустому слову путем сокращения стоящих подряд открывающей и закрывающей скобок одного вида. Например, в случае двух видов скобок сокращать можно только "( )" и "[ ]". Здесь и далее тип скобки – открывающая или закрывающая, вид – круглая, квадратная и т.д.

Определение 9. *Правильная скобочная последовательность* (расстановка) – конечная последовательность скобок, являющаяся подсловом некоторой полной правильной скобочной расстановки.

Определение 10. *m-язык Дика* – язык, состоящий из правильных скобочных последовательностей скобок m видов.

Определение 11. *Сдвиг Дика* – символическая система, языком которой является язык Дика.

Эта система транзитивна, но несинхронизована (см. [8]).

Далее рассматриваем систему Дика с $m \geqslant 2$ видами скобок.

Будем использовать для i-й открывающей скобки обозначение α_i, а для i-й закрывающей – β_i. Обозначим также множество всех открывающих скобок через α, а закрывающих – через β. Кроме того, при наличии меры $\mu \in M_0$ на системе Дика будем считать, что

$$\mu(\alpha) := \sum_{i=1}^{m} \mu(\alpha_i), \quad \mu(\beta) := \sum_{i=1}^{m} \mu(\beta_i).$$

Понятно, что $\mu(\alpha) + \mu(\beta) = 1$.

Пусть w – слово языка Дика. Скобка, сокращающаяся вместе с данной при описанном выше процессе сокращения, называется *парной* к ней в этом слове. Скобка, для которой нет парной в w, называется *непарной* в w. Положим

- $n_1 = n_1(w)$ – количество открывающих скобок в слове w, для которых в этом слове есть парная закрывающая;
- $n_2 = n_2(w)$ – количество непарных скобок в w.

Очевидно, $|w| = 2n_1 + n_2$.

Приведенным видом слова w будем называть слово, получающееся из w сокращением всех парных скобок. Более подробно, алгоритм получения приведенного вида слова $w = a_1 a_2 \dots a_n$, где каждое $a_i \in \alpha \cup \beta$, следующий: на первом шаге из w удаляются все пары последовательных скобок a_i, a_{i+1} такие, что $a_i = \alpha_j$ и $a_{i+1} = \beta_j$ для некоторого j. После удаления получается новое слово w_1, для которого повторяется то же действие, и так далее. Процесс завершается, когда после некоторого шага с номером s пар скобок для удаления не остается. Полученное на этом шаге слово w_s и называется приведенным видом слова w.

Из алгоритма ясно, что приведенный вид всегда является конкатенацией слова, состоящего только из закрывающих скобок, и слова, состоящего только из открывающих. Например:)]]]) + [[[[(([или))]]]] + ((((((, где символ + формально обозначает конкатенацию. Также полезно заметить, что приведенный вид слова w является пустым словом тогда и только тогда, когда само слово w – полная правильная скобочная последовательность.

Через $H(w)$ обозначим разность количества открывающих и закрывающих скобок в слове w, т.е.

$$H(w) := \sum_{i=1}^{|w|} \sum_{j=1}^{m} (\delta_{\alpha_j, w_i} - \delta_{\beta_j, w_i}),$$

где $\delta_{x,y}$ – символ Кронекера, т.е. функция, равная 1, если $x = y$, и 0 в противном случае.

Через $H^j(w)$ обозначим такую же разность, но только по j-му виду скобок:

$$H^j(w) := \sum_{i=1}^{|w|} (\delta_{\alpha_j, w_i} - \delta_{\beta_j, w_i}).$$

§ 3. Обзор основных результатов

Пусть $M(X)$ – некоторый подкласс класса борелевских вероятностных инвариантных эргодических мер на символической динамической системе (X, T) (весь этот класс будем обозначать через $M_0(X)$). Во всех случаях, когда из контекста будет ясно, что такое X, будем писать просто M и M_0.

Связь между энтропией и ЛСДГ в разных системах может быть различной и может формально выражаться одним из следующих утверждений.

Точечная гипотеза (ТГ). *$P_{X,T,\mu}(x) = h(\mu, X, T)$ для всех $\mu \in M(X)$, любой медленной функции $n(\varepsilon)$, любого $x \in X$.*

Обобщенная точечная гипотеза (ОТГ). *$P_{X,T,\mu}(x) = h(\mu, X, T)$ для всех $\mu \in M(X)$, любой медленной функции $n(\varepsilon)$, μ-почти любого $x \in X$.*

Основная гипотеза (ОГ). *$P_{X,T,\mu}(x) = h(\mu, X, T)$ для всех $\mu \in M(X)$, любой медленной функции $n(\varepsilon)$, где выражение в левой части понимается как предел в $L_1(X, \mu)$.*

ТГ очевидным образом влечет ОТГ и ОГ, при этом ОТГ и ОГ, вообще говоря, друг из друга не вытекают и не влекут ТГ.

Точечная и обобщенная точечная гипотезы верны для многих важных систем (ТГ, например, выполняется для автоморфизмов тора с мерой Лебега, а утверждения, близкие к ОТГ, верны для существенно более общих диффеоморфизмов Аносова с мерой Синая – Рюэлля – Боуэна; подробнее см. в [4]), но, как показано в [5], неверны даже для класса бернуллиевских мер на полном сдвиге (контрпримером может служить любая несимметричная бернуллиевская мера при достаточно медленном росте функции $n(\varepsilon)$).

Основная гипотеза, в отличие от двух других, подтверждается для многих символических систем. Наиболее сильный результат в этом направлении получен в [5]:

Теорема 1. *Пусть символическая динамическая система (X, T) и борелевская T-инвариантная эргодическая вероятностная мера μ на X удовлетворяют следующему условию: в X имеется синхронизованный подсдвиг L (T-инвариантное замкнутое подмножество) полной меры, обладающий хотя бы одним волшебным словом положительной меры. Тогда для X и μ верна основная гипотеза.*

В [5] также показано, что условия теоремы 1 не являются необходимыми для выполнения основной гипотезы. Это продемонстрировано на примере конкретной меры на сдвиге Дика. При этом возникает общий вопрос о возможности геометрической интерпретации энтропии для произвольной меры $\mu \in M_0$ на системе Дика. Этот вопрос и изучается ниже.

§ 4. Основная теорема для систем Дика

Теорема 2. *Пусть (X, T, μ) – сдвиг Дика с борелевской вероятностной T-инвариантной эргодической мерой μ, для которой $\mu(\alpha) \neq \mu(\beta)$. Тогда для этой системы верна основная гипотеза.*

Доказательство. Здесь, как и в [5], будем пользоваться достаточным условием выполнения основной гипотезы для борелевских вероятностных инвариантных эргодических мер на символических системах (доказано Комечем в [3]), а именно: для справедливости основной гипотезы достаточно, чтобы для $k = k(\varepsilon)$ и для любой медленной функции $n = n(\varepsilon)$ было выполнено

$$\mu(E_\varepsilon) \xrightarrow[\varepsilon \to 0]{} 1, \tag{1}$$

где

$$E_\varepsilon = \Big\{ x \in X : \, O_\varepsilon\big(T^n(O_\varepsilon(x))\big) = \big\{T^n x_{[-k+n;k]}\big\} \Big\}.$$

Зафиксируем $x \in X$ и $\varepsilon > 0$. Заметим несколько важных фактов.

Во-первых, если $x_{[-k;k-n]}$ – n-синхронизующее слово, то $x \in E_\varepsilon$. Действительно,

$$O_\varepsilon(T^n(O_\varepsilon(x))) = \bigcup_u \{u T^n x_{[-k+n;k]}\},$$

где объединение берется по всем словам $u \in W_n(X)$, таким что $u x_{[-k;k-n]} x_{[k-n+1;k]} \in \in W(X)$, но так как слово $x_{[-k;k-n]}$ – n-синхронизующее, то это объединение всех допустимых цилиндров вида $\{u T^n x_{[-k+n;k]}\}$, $u \in W_n(X)$, которое, очевидно, равно $\{T^n x_{[-k+n;k]}\}$.

Во-вторых, любое слово w, приведенный вид которого содержит не менее n скобок хотя бы одного типа (открывающие, закрывающие), является n-синхронизующим. Действительно, если в приведенном виде не менее n открывающих скобок, то все закрывающие скобки в любом слове v из определения n-синхронизованности будут иметь пару в wv. Теперь предположим, что для некоторых u и v из определения $s := uwv$ запрещено. Это означает, что пару в s составили скобки разных видов в том смысле, что после сокращения всех парных скобок по соседству оказались открывающая и закрывающая скобка (именно в таком порядке), вид которых не совпадает. Из разрешенности слов uw и wv следует, что внутри этих слов такая пара образоваться не может, а значит, остается единственная возможность: открывающая скобка лежит в u, а закрывающая – в v, но и такое невозможно, поскольку все закрывающие скобки из v имеют пару в wv – противоречие, которое и показывает, что слово w является n-синхронизующим. Случай, когда в приведенном виде содержится не менее n закрывающих скобок, симметричен.

В-третьих, если $|H(x_{[-k;k-n]})| \geqslant n$, то приведенный вид $x_{[-k;k-n]}$ содержит не меньше n скобок хотя бы одного типа – простое следствие того, что функция H не меняется при парном сокращении скобок.

Из этих трех замечаний заключаем, что $|H(x_{[-k;k-n]})| \geqslant n$ влечет $x \in E_\varepsilon$.

Теперь воспользуемся эргодической теоремой для индикатора множества β. Получим, что для почти каждого $x \in X$

$$\frac{1}{2k-n+1} \sum_{i=-k}^{k-n} I(w_i \in \beta) \xrightarrow[\varepsilon \to 0]{} \mu(\beta), \tag{2}$$

где $w = x_{[-k;k-n]}$.

Учтем очевидные равенства:

$$\sum_{i=-k}^{k-n} I(w_i \in \alpha) + \sum_{i=-k}^{k-n} I(w_i \in \beta) = 2k - n + 1$$

и

$$\sum_{i=-k}^{k-n} I(w_i \in \alpha) - \sum_{i=-k}^{k-n} I(w_i \in \beta) = H(w),$$

откуда

$$\sum_{i=-k}^{k-n} I(w_i \in \beta) = \frac{1}{2}(2k - n + 1 - H(w)),$$

а значит,

$$\frac{1}{2k-n+1} \sum_{i=-k}^{k-n} I(w_i \in \beta) = \frac{1}{2}\left(1 - \frac{H(w)}{2k-n+1}\right).$$

Подставляя это в (2), получим

$$\lim_{\varepsilon \to 0} \frac{H(x_{[-k;k-n]})}{2k-n+1} = 1 - 2\mu(\beta) := c.$$

Из условий $\mu(\alpha) + \mu(\beta) = 1$ и $\mu(\alpha) \neq \mu(\beta)$ следует, что $1 - 2\mu(\beta) = \mu(\alpha) - \mu(\beta) \neq 0$, т.е. $c \neq 0$. Также понятно, что $c \in [-1; 1]$.

Воспользуемся очевидным фактом: если функция стремится к положительному пределу $\varkappa$ в точке z, то для любого $\tau < 1$ существует окрестность точки z, в которой эта функция не меньше $\tau\varkappa$, в частности, это верно для $\tau = \frac{1}{2}$. Отсюда получаем, что для почти любого $x \in X$ существует положительное число $\gamma(x)$, такое что для всех $\varepsilon \in (0; \gamma(x)]$ выполнено

$$\left|H\left(x_{[-k;k-n]}\right)\right| \geqslant \frac{1}{2}|c|(2k - n + 1) \geqslant n. \tag{3}$$

Здесь учтено, что n мало относительно k, а значит, и относительно $2k - n + 1$. Согласно замеченному выше из неравенства (3) вытекает, что $x \in E_\varepsilon$ для всех $\varepsilon \in (0; \gamma(x)]$. Обозначим множество полной меры, на котором корректно определено $\gamma(x)$, через S. Пусть $\Gamma_\varepsilon := \{x \in S : \gamma(x) \geqslant \varepsilon\}$. Понятно, что совокупность Γ_ε монотонно возрастает при убывании ε, а $\bigcup\limits_{\varepsilon>0} \Gamma_\varepsilon = S$, значит,

$$\mu(\Gamma_\varepsilon) \xrightarrow[\varepsilon \to 0]{} 1.$$

Кроме того, $\Gamma_\varepsilon \subseteq E_\varepsilon$, откуда

$$\mu(E_\varepsilon) \xrightarrow[\varepsilon \to 0]{} 1,$$

что и завершает доказательство. ▲

Замечание 1. Повторив рассуждения для функции $H^j(w)$ вместо $H(w)$, можно ослабить дополнительное условие теоремы до $\mu(\alpha_j) \neq \mu(\beta_j)$ для некоторого j.

Замечание 2. Нетрудно понять, что меры, отвечающие условиям теоремы, существуют и в некотором смысле составляют большинство среди борелевских вероятностных T-инвариантных эргодических мер. В частности, обе меры максимальной энтропии отвечают условиям теоремы.

С другой стороны, для важной меры, рассматриваемой автором в [5], не выполняются даже ослабленные условия из замечания 1 (там все скобки имеют одинаковую меру).

СПИСОК ЛИТЕРАТУРЫ

1. *Заславский Г.М.* Стохастичность динамических систем. М.: Наука, 1984.
2. *Gurevich B.M.* Geometric Interpretation of Entropy for Random Processes // Sinai's Moscow Seminar on Dynamical Systems. Providence, RI: Amer. Math. Soc., 1996. P. 81–87.
3. *Комеч С.А.* Скорость искажения границы в синхронизованных системах: геометрический смысл энтропии // Пробл. передачи информ. 2012. Т. 48. № 1. С. 15–25. http://mi.mathnet.ru/ppi2065
4. *Гуревич Б.М., Комеч С.А.* Скорость деформации границ в системах Аносова и близких к ним // Тр. МИАН. 2017. Т. 297. С. 211–223. https://doi.org/10.1134/S037196851702011X
5. *Дворкин Г.Д.* Геометрическая интерпретация энтропии: новые результаты // Пробл. передачи информ. 2021. Т. 57. № 3. С. 90–101. https://doi.org/10.31857/S0555292321030062
6. *Синай Я.Г.* О понятии энтропии динамической системы // ДАН СССР. 1959. Т. 124. С. 768–771.
7. *Корнфельд И.П., Синай Я.Г., Фомин С.В.* Эргодическая теория. М.: Наука, 1980.
8. *Meyerovitch T.* Tail Invariant Measures of the Dyck-Shift and Non-sofic Systems. Master Thesis. Tel-Aviv Univ., Israel, 2004.

Дворкин Григорий Дмитриевич
Московский государственный университет
им. М.В. Ломоносова, механико-математический факультет,
кафедра математической статистики и случайных процессов
grisha230531415@gmail.com

Поступила в редакцию
13.09.2021
После доработки
25.11.2021
Принята к публикации
26.11.2021

ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 58 **2022** **Вып. 2**

УДК 621.391 : 519.214 : 519.218.4

© 2022 г. **А.В. Логачёв, А.А. Могульский, Е.И. Прокопенко**

ПРИНЦИП БОЛЬШИХ УКЛОНЕНИЙ ДЛЯ МНОГОМЕРНЫХ ОБОБЩЕННЫХ ПРОЦЕССОВ ВОССТАНОВЛЕНИЯ С ПРИЛОЖЕНИЕМ К СВЯЗЫВАНИЮ ПОЛИМЕРОВ[1]

Получен принцип больших уклонений для обрывающихся многомерных обобщенных процессов восстановления. Кроме того, получена асимптотика больших уклонений для случая, когда происходит гиббсовская замена исходной вероятностной меры. Рассмотренный тип случайных процессов широко используется в моделях связывания полимеров.

Ключевые слова: обобщенный процесс восстановления, принцип больших уклонений, функционал уклонений, модель связывания полимеров, гиббсовская замена меры.

DOI: 10.31857/S0555292322020053, **EDN:** DZDASV

§ 1. Введение

Статья посвящена изучению предельного поведения вероятностной меры, построенной по многомерному обобщенному процессу восстановления (ОПВ), который, вообще говоря, может обрываться (т.е. допускается возможность того, что время между моментами восстановления может быть равным ∞ с положительной вероятностью). Такие случайные процессы находят свое применение, в частности, в моделях связывания полимеров (polymer pinning models). Прежде чем привести обзор известных результатов, нам удобнее дать строгое математическое определение изучаемого объекта.

Пусть случайный вектор $(\tau, \boldsymbol{\zeta}, v)$ принимает значения в пространстве

$$\overline{\mathbb{R}}_+ \times \mathbb{R}^d \times \mathbb{R},$$

где $\overline{\mathbb{R}}_+ := \{t \in \mathbb{R} : \ t > 0\} \cup \{\infty\}$, так что координата $\tau > 0$ может принимать значение ∞ с вероятностью $\mathbf{P}(\tau = \infty) =: p \in [0, 1)$. Далее, пусть

$$\{(\tau_i, \boldsymbol{\zeta}_i, v_i)\}_{i \geqslant 1}$$

– последовательность независимых копий вектора $(\tau, \boldsymbol{\zeta}, v)$. Обозначим

$$T_0 := 0, \quad T_n := \sum_{i=1}^{n} \tau_i, \qquad \boldsymbol{Z}_0 := \mathbf{0}, \quad \boldsymbol{Z}_n := \sum_{i=1}^{n} \boldsymbol{\zeta}_i, \qquad V_0 := 0, \quad V_n := \sum_{i=1}^{n} v_i,$$

где $n \in \mathbb{Z}_+$.

[1] Работа выполнена при поддержке Математического Центра в Академгородке, соглашение с Министерством науки и высшего образования Российской Федерации № 075-15-2022-282.

По суммам $(T_n, \boldsymbol{Z}_n)$ для $t > 0$ построим однородный ОПВ с d-мерным фазовым пространством $\mathbb{R}^d$:

$$\boldsymbol{Z}(t) := \boldsymbol{Z}_{\nu(t)}, \quad \nu(t) := \sup\{n \geqslant 0 : T_n \leqslant t\},$$

возможно, обрывающийся (в случае, когда $p = \mathbf{P}(\tau = \infty) > 0$).

Рассмотрим семейство

$$\mathbf{P}_t(\boldsymbol{Z}(t) \in B) := \frac{\mathbf{E}\big(e^{V_{\nu(t)}}; \boldsymbol{Z}(t) \in B\big)}{\mathbf{E}\, e^{V_{\nu(t)}}}, \quad B \in \mathcal{B}_d,$$

вероятностных распределений в пространстве $\mathbb{R}^d$, где $\mathcal{B}_d$ – борелевская σ-алгебра в $\mathbb{R}^d$.

Нас будет интересовать асимптотическое поведение последовательности

$$\frac{1}{t} \ln \mathbf{P}_t\left(\frac{\boldsymbol{Z}(t)}{t} \in B\right) \tag{1.1}$$

при $t \to \infty$.

Сделаем теперь краткий обзор известных результатов, связанных с асимптотикой последовательности (1.1). Работы, в которых изучается предельное поведение последовательности (1.1), можно разделить на две группы. Первая носит чисто теоретический характер, в ней изучается асимптотика, связанная с поведением непосредственно обобщенного процесса восстановления, т.е. полагается, что $V_{\nu(t)} \equiv 0$, и изучается асимптотика последовательности

$$\frac{1}{t} \ln \mathbf{P}\left(\frac{\boldsymbol{Z}(t)}{t} \in B\right). \tag{1.2}$$

В этом направлении для случая $\mathbf{P}(\tau = \infty) = 0$ хорошо изучена как грубая асимптотика последовательности (1.2) (принцип больших уклонений (ПБУ)) [1–3], так и точная асимптотика (локальные и инегро-локальные теоремы в области нормальных, умеренно больших и больших уклонений) [4] (для одмерного случая), [5, 6] (для многомерного случая), [7] (для многомерных арифметических полумарковских ОПВ). В работах [8, 9] изучена, соответственно, грубая и точная асимптотика больших уклонений для конечномерных приращений многомерных ОПВ. Работы [10–12] посвящены принципам больших и умеренно больших уклонений для траекторий ОПВ. Отметим также работу [13], в ней получен ПБУ для мер, построенных по ОПВ. В работе [14] для обрывающихся многомерных ОПВ в некоторой области фазового пространства установлена интегро-локальная (локальная) предельная теорема.

Вторая группа работ имеет более прикладной характер, в ней изучается асимптотическое поведение последовательности (1.1) для случая, когда, вообще говоря, $V_{\nu(t)} \neq 0$ [15, 16]. В этих работах τ – целое число, которое является случайным количеством мономеров, присоединяемых к имеющемуся полимеру в процессе синтеза, $\boldsymbol{\zeta}$ – числовая характеристика, присоединяемого блока мономеров (например, количества мономеров того или иного вида в присоединяемом блоке). В частности, если τ – количество мономеров, присоединенных до выхода полимера на границу разных сред, то вероятность того, с какой стороны от границы был присоединен блок мономеров, зависит от количества энергии в этих средах (см., например, [17, гл. 1; 18, гл. 2]).

Классическим примером является случай, когда $v_i \equiv \beta$ и, соответственно, $V_{\nu(t)} = \beta\nu(t)$, где β – константа, обратно пропорциональная температуре сред (таким образом, предполагается, что температура постоянна). В этом примере средняя скорость присоединения новых блоков мономеров зависит только от температуры. Бо-

лее сложной является модель, в которой есть две различные среды, имеющие общую границу и содержащие разные мономеры. В этой модели на каждом шаге присоединяется блок, полностью состоящий из мономеров, находящихся с одной стороны от границы, т.е. τ_i – количество присоединяемых мономеров между $(i-1)$-м и i-м выходами на границу сред. Здесь $v_i = \beta - \lambda\zeta_i$ и, соответственно, $V_{\nu(t)} = \beta\nu(t) - \lambda Z_{\nu(t)}$, где константа $\lambda > 0$, ζ_i – количество присоединяемых мономеров (между $(i-1)$-м и i-м выходами на границу сред), которые находятся с одной фиксированной стороны от границы. При этом параметр λ отвечает за сложность преодоления границы сред, в частности, если $\lambda = \infty$, то мономеры, находящиеся с одной фиксированной стороны от границы, не могут быть присоединены.

Таким образом, из физических соображений возникает потребность рассматривать не исходную вероятностную меру (1.2), а некоторое ее экспоненциальное преобразование (1.1). Отметим также, что в этих моделях, в частности, событие $\{\tau = \infty\}$ может означать невозможность присоединения нового блока мономеров.

Следует отметить, что в работах из первой группы тоже приходится сталкиваться с экспоненциальным преобразованием исходной вероятностной меры, но совершенно из других соображений, связанных с техникой доказательства ПБУ. Эта техника помогает также успешно решить многие задачи, поставленные во второй группе работ.

Наиболее близкой по содержанию к нашей статье является работа [16]. В ней рассматривается случайный процесс $\boldsymbol{Z}(t)$, у которого фазовое пространство является банаховым, в том числе и бесконечномерным, и в частности, успешно изучается асимптотическое поведение последовательности (1.1) для достаточно широкого класса множеств $B \in \mathcal{B}_d$ в случае, когда выполнены дополнительные условия:

(a) *Случайная величина τ принимает целые положительные значения или ∞;*

(b) *Случайная величина v имеет вид $v = h(\tau)$ для некоторой фиксированной неслучайной функции $h = h(t)$.*

В настоящей статье проведено изучение асимптотического поведения последовательности (1.1) в общем случае, т.е. без привлечения условий (a) и (b) (см. следствие 1). Кроме того, для обрывающихся многомерных ОПВ в условиях, близких к необходимым, установлен принцип больших уклонений в фазовом пространстве (см. следствие 2). Отметим, что более общий вид функции $v = h(\tau)$ дает возможность рассматривать более сложные физические модели синтеза полимеров. В частности, модели, в которых экспоненциальное преобразование меры зависит от типа присоединенных мономеров. Отказ от целочисленности τ позволяет, в частности, рассматривать модели, в которых мы можем фиксировать только моменты времени выхода полимера на границу сред и не знаем, какое точно количество мономеров было присоединено за время между этими моментами. Вопрос о выполнении ПБУ для ОПВ в произвольном банаховом пространстве, когда условия (a) и (b) не выполнены, остается открытым.

Для формулировок и доказательства основных результатов нам потребуются некоторые дополнительные моментные условия. Везде далее будем предполагать, что выполнено следующее условие:

[$\mathbf{C}^*$] Для некоторых $\lambda > 0$ и $\varepsilon > 0$

$$\mathbf{E}\big(e^{-\lambda\tau+\varepsilon|\boldsymbol{\zeta}|+v};\, \tau < \infty\big) < \infty.$$

Заметим, что в силу того, что $-\lambda\infty = -\infty$, в условии [$\mathbf{C}^*$] можно оставить неравенство

$$\mathbf{E}\big(e^{-\lambda\tau+\varepsilon|\boldsymbol{\zeta}|+v}\big) < \infty.$$

Поскольку

$$\ln \mathbf{P}_t\left(\frac{\boldsymbol{Z}(t)}{t} \in B\right) = \ln \mathbf{E}\left(e^{V_{\nu(t)}}; \frac{\boldsymbol{Z}_{\nu(t)}}{t} \in B\right) - \ln \mathbf{E}\left(e^{V_{\nu(t)}}; \frac{\boldsymbol{Z}_{\nu(t)}}{t} \in \mathbb{R}^d\right), \tag{1.3}$$

то для того чтобы изучить предельное поведение последовательности (1.1), достаточно получить предельные теоремы для последовательности

$$\frac{1}{t} \ln \mathbf{E}\left(e^{V_{\nu(t)}}; \frac{\boldsymbol{Z}_{\nu(t)}}{t} \in B\right), \tag{1.4}$$

что и осуществлено в теореме 1. Следствием теоремы 1 в частном случае, когда $v = 0$ п.н., является ПБУ для ОПВ $\boldsymbol{Z}(t)$ (возможно, обрывающегося), который сформулирован в следствии 2.

§ 2. Основные обозначения и основной результат

Обозначим

$$A(\lambda, \boldsymbol{\mu}) := \ln \mathbf{E}\big(e^{\lambda\tau + \langle \boldsymbol{\mu}, \boldsymbol{\zeta}\rangle + v};\ \tau < \infty\big),$$

здесь и далее $\langle \cdot\,, \cdot \rangle$ – скалярное произведение. Функция $A(\lambda, \boldsymbol{\mu})$ является преобразованием Лапласа над "неполной" мерой, отличающейся от меры

$$\mathbf{E}(e^v;\ \tau \in \cdot\,,\ \boldsymbol{\zeta} \in \cdot\,) = \mathbf{E}(e^v;\ \tau \in \cdot\,,\ \boldsymbol{\zeta} \in \cdot\,,\ \tau < \infty) + \mathbf{E}(e^v;\ \boldsymbol{\zeta} \in \cdot\,,\ \tau = \infty)$$

отсутствием второго слагаемого. Однако эта "неполнота" полностью согласуется с тем очевидным обстоятельством, что изучаемая характеристика (1.1) не зависит от распределения

$$\mathbf{P}(\boldsymbol{\zeta} \in \cdot\,,\ v \in \cdot \mid \tau = \infty).$$

Рассмотрим два множества

$$\mathcal{A}^{\leqslant 0} := \{(\lambda, \boldsymbol{\mu}) :\ A(\lambda, \boldsymbol{\mu}) \leqslant 0\}, \quad \mathcal{A}^{\leqslant 0}_{\gamma} := \{(\lambda, \boldsymbol{\mu}) : \lambda < \gamma,\ A(\lambda, \boldsymbol{\mu}) \leqslant 0\},$$

где $\gamma \in \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ фиксировано. В качестве γ будут выбираться:

- либо константы $\lambda_{\pm}$, $0 \leqslant \lambda_+ \leqslant \lambda_- \leqslant \infty$, где

$$\lambda_+ := \sup\{\lambda \geqslant 0 :\ \mathbf{E}\, e^{\lambda\tau} < \infty\} = -\limsup_{t\to\infty} \frac{1}{t} \ln \mathbf{P}(\tau > t), \tag{2.1}$$

$$\lambda_- := -\liminf_{t\to\infty} \frac{1}{t} \ln \mathbf{P}(\tau > t); \tag{2.2}$$

- либо константы $\lambda^*_{\pm}$, $0 \leqslant \lambda^*_+ \leqslant \lambda^*_- \leqslant \infty$, определенные при дополнительном условии

  $$\mathbf{E}(e^v) < \infty$$

  соотношениями

$$\lambda^*_+ := \sup\{\lambda \geqslant 0 :\ \mathbf{E}\, e^{v+\lambda\tau} < \infty\} = -\limsup_{t\to\infty} \frac{1}{t} \ln \mathbf{E}(e^v;\ \tau > t),$$

$$\lambda^*_- := -\liminf_{t\to\infty} \frac{1}{t} \ln \mathbf{E}(e^v;\ \tau > t).$$

Заметим, что если выполнено условие обрываемости $p = \mathbf{P}(\tau = \infty) > 0$, то выполняется $\lambda_+ = \lambda_- = 0$.

Построим новые функции

$$A(\boldsymbol{\mu}) := -\sup\{\lambda :\ (\lambda, \boldsymbol{\mu}) \in \mathcal{A}^{\leqslant 0}\}, \quad A_\gamma(\boldsymbol{\mu}) := \max\{-\gamma, A(\boldsymbol{\mu})\}, \tag{2.3}$$

где по определению считаем

$$\sup\{\lambda :\ \lambda \in \varnothing\} = -\infty.$$

Для функции $H = H(\boldsymbol{\mu})\colon \mathbb{R}^d \to (-\infty, \infty]$ определим (см., например, [19]) преобразование Лежандра $H^{\mathfrak{L}\mathfrak{e}} = H^{\mathfrak{L}\mathfrak{e}}(\boldsymbol{\alpha})$, положив

$$H^{\mathfrak{L}\mathfrak{e}}(\boldsymbol{\alpha}) := \sup_{\boldsymbol{\mu}}\{\langle \boldsymbol{\mu}, \boldsymbol{\alpha}\rangle - H(\boldsymbol{\mu})\}, \quad \boldsymbol{\alpha} \in \mathbb{R}^d.$$

Будем называть функцию $H = H(\boldsymbol{\alpha})$, отображающую $\mathbb{R}^d$ в $[0, \infty]$, *компактной*, если для любого $c \geqslant 0$ множество $\{\boldsymbol{\alpha} :\ H(\boldsymbol{\alpha}) \leqslant c\}$ является компактом в $\mathbb{R}^d$. Легко показать, что любая компактная функция $H(\boldsymbol{\alpha})$ полунепрерывна снизу.

Определим две функции: для $\boldsymbol{\alpha} \in \mathbb{R}^d$

$$D(\boldsymbol{\alpha}) := A^{\mathfrak{L}\mathfrak{e}}(\boldsymbol{\alpha}), \quad D_\gamma(\boldsymbol{\alpha}) := A_\gamma^{\mathfrak{L}\mathfrak{e}}(\boldsymbol{\alpha}).$$

В следующей лемме содержатся необходимые нам свойства этих функций.

Л е м м а 1. (i) *Функции $A(\boldsymbol{\mu})$ и $A_\gamma(\boldsymbol{\mu})$ выпуклы и полунепрерывны снизу*;

(ii) *Функции $D(\boldsymbol{\alpha})$ и $D_\gamma(\boldsymbol{\alpha})$ выпуклы, полунепрерывны снизу и компактны*;

(iii) *Справедливы формулы*

$$A(\boldsymbol{\mu}) = D^{\mathfrak{L}\mathfrak{e}}(\boldsymbol{\mu}), \quad A_\gamma(\boldsymbol{\mu}) = D_\gamma^{\mathfrak{L}\mathfrak{e}}(\boldsymbol{\mu}), \tag{2.4}$$

так что пары $A(\boldsymbol{\mu}), D(\boldsymbol{\alpha})$ и $A_\gamma(\boldsymbol{\mu}), D_\gamma(\boldsymbol{\alpha})$ являются парами взаимно сопряженных (относительно преобразования Лежандра) функций;

(iv) *Функции $A_\gamma(\boldsymbol{\mu})$ и $A(\boldsymbol{\mu})$ совпадают (и следовательно, $D_\gamma(\boldsymbol{\alpha}) = D(\boldsymbol{\alpha})$) тогда и только тогда, когда выполнено условие*

$$\gamma \geqslant D(\mathbf{0}); \tag{2.5}$$

(v) *Для всех $\boldsymbol{\alpha} \in \mathbb{R}^d$ справедливо*

$$D_\gamma(\boldsymbol{\alpha}) = \inf_{\theta \in [0,1]}\{D(\theta, \boldsymbol{\alpha}) + \gamma(1-\theta)\}, \tag{2.6}$$

где

$$D(\theta, \boldsymbol{\alpha}) := \sup_{(\lambda, \boldsymbol{\mu}) \in \mathcal{A}^{\leqslant 0}}\{\lambda\theta + \langle \boldsymbol{\mu}, \boldsymbol{\alpha}\rangle\}.$$

Замечание 1. Дополнительные свойства функций $D(\boldsymbol{\alpha})$ и $D_\gamma(\boldsymbol{\alpha})$ будут приведены в лемме 5 (см. § 4). Поскольку доказательства лемм 1 и 5 в значительной степени повторяют доказательства аналогичных утверждений работы [20], то эти доказательства мы опускаем. Однако для удобства читателя мы приводим полные доказательства лемм 1 и 5 в расширенной версии [21] настоящей статья.

Определим теперь две функции:

$$D_+(\boldsymbol{\alpha}) := D_{\lambda_+}(\boldsymbol{\alpha}) - A_{\lambda_-}(0), \quad D_-(\boldsymbol{\alpha}) := D_{\lambda_-}(\boldsymbol{\alpha}) - A_{\lambda_+}(0), \quad \boldsymbol{\alpha} \in \mathbb{R}^d,$$

где константы λ_+ и λ_- определены формулами (2.1) и (2.2) соответственно. Заметим, что

$$A_{\lambda_\pm}(0) = \sup\{\lambda < \lambda_\pm :\ \mathbf{E}\, e^{\lambda\tau + v} \leqslant 1\}.$$

Заметим, что если выполнено условие обрываемости $p = \mathbf{P}(\tau = \infty) > 0$, то

$$D_+(\boldsymbol{\alpha}) = D_-(\boldsymbol{\alpha}) = D_0(\boldsymbol{\alpha}) - A_0(0), \quad \boldsymbol{\alpha} \in \mathbb{R}^d.$$

Для множества $B \in \mathcal{B}_d$ через $[B]$ и (B) будем обозначать его замыкание и внутренность соответственно. Положим для $B \in \mathcal{B}_d$

$$D_\gamma(B) = \inf_{\boldsymbol{\alpha} \in B} D_\gamma(\boldsymbol{\alpha}).$$

Основными результатами данной статьи являются следующие утверждения.

Т е о р е м а 1. *Для любого борелевского множества $B \subset \mathbb{R}^d$*

$$\limsup_{t\to\infty} \frac{1}{t} \ln \mathbf{E}\left(e^{V_{\nu(t)}}; \frac{\boldsymbol{Z}(t)}{t} \in B \right) \leqslant -D_{\lambda_+}([B]), \tag{2.7}$$

$$\liminf_{t\to\infty} \frac{1}{t} \ln \mathbf{E}\left(e^{V_{\nu(t)}}; \frac{\boldsymbol{Z}(t)}{t} \in B \right) \geqslant -D_{\lambda_-}((B)). \tag{2.8}$$

Используя равенство (1.3), находим оценки для последовательности (1.1).

С л е д с т в и е 1. *Для любого борелевского множества $B \subset \mathbb{R}^d$*

$$\limsup_{t\to\infty} \frac{1}{t} \ln \mathbf{P}_t\left(\frac{\boldsymbol{Z}(t)}{t} \in B \right) \leqslant -D_+([B]),$$

$$\liminf_{t\to\infty} \frac{1}{t} \ln \mathbf{P}_t\left(\frac{\boldsymbol{Z}(t)}{t} \in B \right) \geqslant -D_-((B)).$$

Оценки для последовательности (1.2) очевидным образом извлекаются из теоремы 1, накладывая условие $\mathbf{P}(v = 0) = 1$.

С л е д с т в и е 2. *Пусть $\mathbf{P}(v = 0) = 1$. Тогда для любого борелевского множества $B \subset \mathbb{R}^d$*

$$\limsup_{t\to\infty} \frac{1}{t} \ln \mathbf{P}\left(\frac{\boldsymbol{Z}(t)}{t} \in B \right) \leqslant -D_{\lambda_+}([B]), \tag{2.9}$$

$$\liminf_{t\to\infty} \frac{1}{t} \ln \mathbf{P}\left(\frac{\boldsymbol{Z}(t)}{t} \in B \right) \geqslant -D_{\lambda_-}((B)). \tag{2.10}$$

В частности, если выполнено условие обрываемости $p = \mathbf{P}(\tau = \infty) > 0$, то $\lambda_- = \lambda_+ = 0$, и тогда выполнение неравенств (2.9), (2.10) означает, что семейство $\dfrac{\boldsymbol{Z}(t)}{t}$ удовлетворяет принципу больших уклонений в $\mathbb{R}^d$ с функцией уклонений $D_0(\boldsymbol{\alpha})$.

§ 3. Доказательство теоремы 1

В основе доказательства теоремы 1 лежат следующие леммы, которые доказываются в § 4.

Л е м м а 2. *Для любого $\boldsymbol{\alpha} \in \mathbb{R}^d$ выполняется*

$$\lim_{\varepsilon\downarrow 0} \limsup_{t\to\infty} \frac{1}{t} \ln \mathbf{E}\left(e^{V_{\nu(t)}}; \frac{\boldsymbol{Z}(t)}{t} \in (\boldsymbol{\alpha})_\varepsilon \right) \leqslant -D_{\lambda_+}(\boldsymbol{\alpha}). \tag{3.1}$$

Лемма 3. *Для любых* $\boldsymbol{\alpha} \in \mathbb{R}^d$, $\varepsilon > 0$ *выполняется*

$$\liminf_{t\to\infty} \frac{1}{t} \ln \mathbf{E}\left(e^{V_{\nu(t)}}; \frac{\boldsymbol{Z}(t)}{t} \in (\boldsymbol{\alpha})_\varepsilon \right) \geqslant -D(\boldsymbol{\alpha}), \tag{3.2}$$

$$\liminf_{t\to\infty} \frac{1}{t} \ln \mathbf{E}\left(e^{V_{\nu(t)}}; \frac{\boldsymbol{Z}(t)}{t} \in (\boldsymbol{\alpha})_\varepsilon \right) \geqslant -D_{\lambda_-}(\boldsymbol{\alpha}). \tag{3.3}$$

Поскольку $A_{\lambda_-}(\boldsymbol{\mu}) \geqslant A(\boldsymbol{\mu})$, то $-D_{\lambda_-}(\boldsymbol{\alpha}) \geqslant -D(\boldsymbol{\alpha})$, и следовательно, неравенство (3.2) следует из неравенства (3.3). Однако неравенство (3.2), являющееся, вообще говоря, более грубым, чем неравенство (3.3), тем не менее представляет определенный интерес, поскольку дает содержательную оценку снизу в тех случаях, когда отсутствует информация о константе λ_-.

Лемма 4. *Для любого* $N \in (0, \infty)$ *найдется* $M \in (0, \infty)$, *такое что*

$$\limsup_{t\to\infty} \frac{1}{t} \ln \mathbf{E}\left(e^{V_{\nu(t)}}; \frac{|\boldsymbol{Z}(t)|}{t} \geqslant M \right) \leqslant -N.$$

Проведем теперь на основе лемм 2–4 доказательство теоремы 1, которое повторяет основные шаги доказательства теоремы 4.1.1 из [22, с. 259].

Доказательство теоремы 1. (i) Оценка сверху (2.7). Зафиксируем константы $\delta > 0$, $N < \infty$ и обозначим $D(B, \delta, N) := \min\{N, D_{\lambda_+}([B])\} + \delta$,

$$L_+(B) := \limsup_{t\to\infty} \frac{1}{t} \ln \mathbf{E}\left(e^{V_{\nu(t)}}; \frac{\boldsymbol{Z}(t)}{t} \in B \right).$$

В силу леммы 4 найдется компакт $K \subset \mathbb{R}^d$, такой что для $\overline{K} := \mathbb{R}^d \setminus K$ выполняется

$$L_+(\overline{K}) \leqslant -2D(B, \delta, N). \tag{3.4}$$

Далее, в силу леммы 2 для любого $\boldsymbol{\alpha}$ из компакта $K \cap [B]$ найдется $\varepsilon(\boldsymbol{\alpha}) > 0$, такое что выполняется

$$L_+((\boldsymbol{\alpha})_{\varepsilon(\boldsymbol{\alpha})}) \leqslant -D(B, \delta, N). \tag{3.5}$$

Получили открытое покрытие компакта $K \cap [B]$, из которого выделяем конечное подпокрытие

$$\left\{(\boldsymbol{\alpha}_i)_{\varepsilon(\boldsymbol{\alpha}_i)}\right\}_{i=1}^{I}: \quad K \cap [B] \subset \bigcup_{i=1}^{I} (\boldsymbol{\alpha}_i)_{\varepsilon(\boldsymbol{\alpha}_i)}, \quad I < \infty. \tag{3.6}$$

Поэтому

$$L_+([B]) \leqslant L_+\big((K \cap [B]) \cup \overline{K}\big),$$

и в силу (3.4)–(3.6) имеем

$$L_+([B]) \leqslant -D(B, \delta, N).$$

Левая часть последнего неравенства не зависит от $\delta > 0$ и $N < \infty$, поэтому неравенство сохранится, если в его правой части устремить $\delta \to 0$ и $N \to \infty$. Получили оценку сверху (2.7).

(ii) Оценка снизу (2.8). Фиксируем $\boldsymbol{\alpha} \in (B)$ и $\varepsilon > 0$ таким образом, что $(\boldsymbol{\alpha})_\varepsilon \subset (B)$, и обозначим

$$L_-(B) := \liminf_{t\to\infty} \frac{1}{t} \ln \mathbf{E}\left(e^{V_{\nu(t)}}; \frac{\boldsymbol{Z}(t)}{t} \in B \right).$$

Имеем

$$L_-(B) \geqslant L_-((B)) \geqslant L_-((\boldsymbol{\alpha})_\varepsilon),$$

поэтому, применяя лемму 3, получаем

$$L_-(B) \geqslant -D_{\lambda_-}(\boldsymbol{\alpha}).$$

Левая часть последнего неравенства не зависит от $\boldsymbol{\alpha} \in (B)$, поэтому неравенство сохранится, если его правую часть максимизировать по $\boldsymbol{\alpha} \in (B)$. Получили оценку снизу (2.8). ▲

§ 4. Доказательство лемм 2–4

Нам понадобятся следующие обозначения. Для $(\theta, \boldsymbol{\alpha}) \in \mathbb{R}^{d+1}$ обозначим

$$D_\Lambda(\theta, \boldsymbol{\alpha}) := \inf_{r>0} r\Lambda\Big(\frac{\theta}{r}, \frac{\boldsymbol{\alpha}}{r}\Big), \tag{4.1}$$

где

$$\Lambda(\theta, \boldsymbol{\alpha}) := \sup_{\lambda, \boldsymbol{\mu}} \{\lambda\theta + \langle \boldsymbol{\mu}, \boldsymbol{\alpha}\rangle - A(\lambda, \boldsymbol{\mu})\}, \quad (\theta, \boldsymbol{\alpha}) \in \mathbb{R}^{d+1},$$

– *функция уклонений*, которая определяется как преобразование Лежандра функции $A(\lambda, \boldsymbol{\mu})$:

$$\Lambda(\theta, \boldsymbol{\alpha}) = A^{\mathfrak{L}\mathfrak{e}}(\theta, \boldsymbol{\alpha}).$$

Легко убедиться, что функция $D_\Lambda(\theta, \boldsymbol{\alpha})$ выпукла и линейчата (т.е. линейна вдоль любого луча, выходящего из начала координат). Однако свойство полунепрерывности снизу для этой функции может отсутствовать.

В следующем утверждении приводятся дополнительные свойства функций $D(\boldsymbol{\alpha})$ и $D_\gamma(\boldsymbol{\alpha})$.

Лемма 5. (i) *Для всех* $\boldsymbol{\alpha} \in \mathbb{R}^d$ *справедливо*

$$D(\boldsymbol{\alpha}) = \sup_{(\lambda, \boldsymbol{\mu}) \in \mathcal{A}^{\leqslant 0}} \{\lambda + \langle \boldsymbol{\mu}, \boldsymbol{\alpha}\rangle\}, \tag{4.2}$$

$$D_\gamma(\boldsymbol{\alpha}) = \sup_{(\lambda, \boldsymbol{\mu}) \in \mathcal{A}_\gamma^{\leqslant 0}} \{\lambda + \langle \boldsymbol{\mu}, \boldsymbol{\alpha}\rangle\}; \tag{4.3}$$

(ii) *Для функции* $D_\Lambda(\theta, \boldsymbol{\alpha})$ (*см.* (4.1)) *имеет место равенство*

$$(D_\Lambda^{\mathfrak{L}\mathfrak{e}})^{\mathfrak{L}\mathfrak{e}}(\theta, \boldsymbol{\alpha}) = D(\theta, \boldsymbol{\alpha}), \tag{4.4}$$

и для всех $\theta > 0$, $\boldsymbol{\alpha} \in \mathbb{R}^d$ *выполнено*

$$D(\theta, \boldsymbol{\alpha}) = \lim_{\varepsilon \downarrow 0} \inf_{\boldsymbol{\alpha}' \in (\boldsymbol{\alpha})_\varepsilon} D_\Lambda(\theta, \boldsymbol{\alpha}'); \tag{4.5}$$

(iii) *Если* $\gamma < \infty$, *то для функции*

$$\widehat{D}_\gamma(\boldsymbol{\alpha}) := \inf_{\theta \in (0,1)} \{D(\theta, \boldsymbol{\alpha}) + \gamma(1-\theta)\}$$

имеет место равенство

$$\lim_{\varepsilon \downarrow 0} \inf_{\boldsymbol{\alpha}' \in (\boldsymbol{\alpha}_\varepsilon)} \widehat{D}_\gamma(\boldsymbol{\alpha}') = D_\gamma(\boldsymbol{\alpha}). \tag{4.6}$$

Доказательство леммы 2. Из определения процесса $\boldsymbol{Z}(t)$ вытекает, что на событии $\{T_n \leqslant t < T_n + \tau_{n+1}\}$ выполнено $\boldsymbol{Z}(t) = \boldsymbol{Z}_n$. Следовательно,

$$E(t) := \mathbf{E}\left(e^{V_{\nu(t)}}; \frac{\boldsymbol{Z}(t)}{t} \in (\boldsymbol{\alpha})_\varepsilon\right) = \sum_{n \geqslant 0} E_n(t), \tag{4.7}$$

где

$$E_n(t) := \mathbf{E}\left(e^{V_n}; \frac{\boldsymbol{Z}_n}{t} \in (\boldsymbol{\alpha})_\varepsilon,\ T_n \leqslant t < T_n + \tau_{n+1}\right).$$

Оценим сначала часть суммы в (4.7) по $n > t^2$:

$$\sum_{n > t^2} E_n(t) \leqslant \sum_{n > t^2} \mathbf{E}\big(e^{V_n}; T_n \leqslant t\big) =: \sum_{n > t^2} P_n(t).$$

Выберем число $\lambda^* > 0$ таким образом, что $A(-\lambda^*, \mathbf{0}) \leqslant 0$ (в силу условия $[\mathbf{C}^*]$ такая константа λ^* всегда найдется), и рассмотрим новые случайные независимые величины τ_j^* с распределением

$$\mathbf{P}(\tau^* \in \cdot\,) := e^{-A(-\lambda^*, \mathbf{0})}\, \mathbf{E}\big(e^{v - \lambda^* \tau};\ \tau \in \cdot\,\big).$$

Легко показать, что $\mathbf{P}(\tau^* \in (0, \infty]) = 1$, и следовательно, функция уклонений

$$\Lambda^*(\theta) := \sup_\lambda \big\{\lambda\theta - \ln \mathbf{E}\, e^{\lambda \tau^*}\big\}$$

неограниченно возрастает при монотонном приближении справа аргумента θ к началу координат, т.е. $\lim\limits_{\theta \downarrow 0} \Lambda^*(\theta) = \infty$. Далее, для $n \geqslant 1$ обозначим $T_n^* := \tau_1^* + \ldots + \tau_n^*$, так что справедливо неравенство

$$P_n(t) = e^{\pm n A(-\lambda^*, \mathbf{0})}\, \mathbf{E}\big(e^{V_n \pm \lambda^* T_n};\ T_n \leqslant t\big) = e^{n A(-\lambda^*, \mathbf{0})}\, \mathbf{E}\big(e^{\lambda^* T_n^*};\ T_n^* \leqslant t\big) \leqslant$$
$$\leqslant e^{n A(-\lambda^*, \mathbf{0}) + \lambda^* t}\, \mathbf{E}(T_n^* \leqslant t) \leqslant e^{\lambda^* t}\, \mathbf{P}(T_n^* \leqslant t),$$

где последнее неравенство справедливо в силу того, что $A(-\lambda^*, \mathbf{0}) \leqslant 0$. Поэтому в силу экспоненциального неравенства Чебышева при $\dfrac{t}{n} \leqslant \mathbf{E}\,\tau^*$ имеем

$$P_n(t) \leqslant e^{\lambda^* t - n\Lambda^*\left(\frac{t}{n}\right)}.$$

Таким образом, для $n \geqslant t^2$, $\dfrac{1}{t} \leqslant \mathbf{E}\,\tau^*$ имеем оценку

$$P_n(t) \leqslant e^{\lambda^* t} q^n(t),$$

где $q(t) := e^{-\Lambda^*\left(\frac{1}{t}\right)} < 1$ для всех достаточно больших t. Следовательно,

$$\sum_{n \geqslant t^2} E_n(t) \leqslant e^{\lambda^* t} \frac{q^{t^2}}{1 - q}.$$

Поэтому

$$\limsup_{t \to \infty} \frac{1}{t} \ln \sum_{n \geqslant t^2} E_n(t) = -\infty. \tag{4.8}$$

Приведем теперь более точную оценку сверху для $E_n(t)$, справедливую для всех $n \geqslant 1$. Для любого вектора $(\lambda, \boldsymbol{\mu}) \in (\mathcal{A}_{\lambda_+}^{\leqslant 0})$ имеем

$$E_n(t) = \mathbf{E}\left(e^{V_n \pm \lambda T_n \pm \langle \boldsymbol{\mu}, \boldsymbol{Z}_n \rangle};\ \frac{\boldsymbol{Z}_n}{t} \in (\boldsymbol{\alpha})_\varepsilon,\ T_n \leqslant t < T_n + \tau_{n+1} \right).$$

На событии

$$\left\{ \frac{\boldsymbol{Z}_n}{t} \in (\boldsymbol{\alpha})_\varepsilon,\ T_n \leqslant t < T_n + \tau_{n+1} \right\}$$

выполняется неравенство

$$e^{-\lambda T_n - \langle \boldsymbol{\mu}, \boldsymbol{Z}_n \rangle} \leqslant e^{-t(\lambda + \langle \boldsymbol{\mu}, \boldsymbol{\alpha} \rangle) + \sqrt{d}|\boldsymbol{\mu}|\varepsilon t} \max\{1, e^{\lambda \tau_{n+1}}\}.$$

Так как вектор $(\lambda, \boldsymbol{\mu}) \in \left(\mathcal{A}_{\lambda_+}^{\leqslant 0}\right)$, то $\lambda < \lambda_+$, и следовательно (учитывая, что $\lambda_+ = 0$ при $p = \mathbf{P}(\tau = \infty) > 0$), то получаем

$$\mathbf{E} \max\{1, e^{\lambda \tau_{n+1}}\} < \infty.$$

Таким образом, для $n \geqslant 1$ в любом случае имеем оценку

$$\begin{aligned} E_n(t) &\leqslant e^{-t(\lambda + \langle \boldsymbol{\mu}, \boldsymbol{\alpha} \rangle) + \sqrt{d}|\boldsymbol{\mu}|\varepsilon t}\, \mathbf{E}\Big(\max\{1, e^{\lambda \tau_{n+1}}\} e^{V_n + \lambda T_n + \langle \boldsymbol{\mu}, \boldsymbol{Z}_n \rangle} \Big) \leqslant \\ &\leqslant e^{-t(\lambda + \langle \boldsymbol{\mu}, \boldsymbol{\alpha} \rangle) + \sqrt{d}|\boldsymbol{\mu}|\varepsilon t}\, \mathbf{E} \max\{1, e^{\lambda \tau}\} e^{nA(\lambda, \boldsymbol{\mu})} \leqslant \\ &\leqslant e^{-t(\lambda + \langle \boldsymbol{\mu}, \boldsymbol{\alpha} \rangle) + \sqrt{d}|\boldsymbol{\mu}|\varepsilon t}\, \mathbf{E} \max\{1, e^{\lambda \tau}\}, \end{aligned}$$

из которой вытекает неравенство

$$\sum_{n=1}^{[t^2]} E_n(t) \leqslant \mathbf{E} \max\{1, e^{\lambda \tau}\} t^2 e^{-t(\lambda + \langle \boldsymbol{\mu}, \boldsymbol{\alpha} \rangle) + \sqrt{d}|\boldsymbol{\mu}|\varepsilon t}. \tag{4.9}$$

Наконец, оценим $E_0(t) = \mathbf{P}(\mathbf{0} \in (\boldsymbol{\alpha})_\varepsilon,\ \tau > t)$. Имеем

$$\lim_{\varepsilon \downarrow 0} \limsup_{t \to \infty} \frac{1}{t} \ln E_0(t) = \begin{cases} -\infty, & \text{если } \boldsymbol{\alpha} \neq \mathbf{0}, \\ -\lambda_+, & \text{если } \boldsymbol{\alpha} = \mathbf{0}. \end{cases} \tag{4.10}$$

Таким образом, из (4.9) и (4.10) получается неравенство

$$\limsup_{t \to \infty} \frac{1}{t} \ln \sum_{n=0}^{[t^2]} E_n(t) \leqslant -(\lambda + \langle \boldsymbol{\mu}, \boldsymbol{\alpha} \rangle) + \sqrt{d}|\boldsymbol{\mu}|\varepsilon. \tag{4.11}$$

Из соотношений (4.7), (4.8), (4.11) вытекает, что для любого $(\lambda, \boldsymbol{\mu}) \in \left(\mathcal{A}_{\lambda_+}^{\leqslant 0}\right)$

$$\lim_{\varepsilon \downarrow 0} \limsup_{t \to \infty} \frac{1}{t} \ln E(t) \leqslant -(\lambda + \langle \boldsymbol{\mu}, \boldsymbol{\alpha} \rangle). \tag{4.12}$$

Так как к любой точке $(\lambda, \boldsymbol{\mu})$ границы $\partial \mathcal{A}_{\lambda_+}^{\leqslant 0}$ можно приблизиться точками $(\lambda_n, \boldsymbol{\mu}_n)$ из внутренности $\left(\mathcal{A}_{\lambda_+}^{\leqslant 0}\right)$, то неравенство (4.12) справедливо для всех $(\lambda, \boldsymbol{\mu})$ из замкнутого выпуклого множества $\mathcal{A}_{\lambda_+}^{\leqslant 0}$. Минимизируя правую часть неравенства (4.12) по

$(\lambda, \boldsymbol{\mu}) \in \mathcal{A}_{\lambda_+}^{\leqslant 0}$ и используя утверждение (i) леммы 5, получаем

$$\lim_{\varepsilon \downarrow 0} \limsup_{t \to \infty} \frac{1}{t} \ln E(t) \leqslant -D_{\lambda_+}(\boldsymbol{\alpha}),$$

что завершает доказательство леммы 2. ▲

Доказательство леммы 3. Докажем сперва неравенство (3.2). Выберем $\lambda^* > 0$, такое что выполняется $A(-\lambda^*, \mathbf{0}) \leqslant 0$, что эквивалентно $\mathbf{E}(e^{-\lambda^* \tau + v}) \leqslant 1$ (как уже отмечалось при доказательстве леммы 2, в силу условия $[\mathbf{C}^*]$ такая константа λ^* всегда найдется). Для этого λ^* построим случайный вектор $(\tau^*, \boldsymbol{\zeta}^*, v^*) \in \mathbb{R}_+ \times \mathbb{R}^d \times \mathbb{R}$ с распределением

$$\mathbf{P}(\tau^* \in \cdot,\, \boldsymbol{\zeta}^* \in \cdot,\, v^* \in \cdot\,) := \mathbf{P}^*(\tau \in \cdot,\, \boldsymbol{\zeta} \in \cdot,\, v \in \cdot\,) :=$$
$$:= \mathbf{E}(e^{-\lambda^* \tau + v};\, \tau \in \cdot,\, \boldsymbol{\zeta} \in \cdot,\, v \in \cdot\,).$$

Заметим, что в случае, когда $A(-\lambda^*, \mathbf{0}) < 0$, распределение этого вектора будет несобственным, т.е.

$$\mathbf{P}(\tau^* \in (0, \infty),\, \boldsymbol{\zeta}^* \in \mathbb{R}^d,\, v^* \in \mathbb{R}) = e^{A(-\lambda^*, \mathbf{0})} < 1.$$

Это распределение можно произвольным образом доопределить на множестве $\{\infty\} \times$ $\times\, \mathbb{R}^d \times \mathbb{R}$. Преобразование Лапласа над распределением вектора $(\tau^*, \boldsymbol{\zeta}^*)$ обозначим через

$$e^{A^*(\lambda, \boldsymbol{\mu})} := \mathbf{E}\, e^{\lambda \tau^* + \langle \boldsymbol{\mu}, \boldsymbol{\zeta}^* \rangle} = \mathbf{E}\big(e^{\lambda \tau + \langle \boldsymbol{\mu}, \boldsymbol{\zeta} \rangle - \lambda^* \tau + v}\big) = e^{A(-\lambda^* + \lambda, \boldsymbol{\mu})},$$

т.е. положим

$$A^*(\lambda, \boldsymbol{\mu}) := A(-\lambda^* + \lambda, \boldsymbol{\mu}).$$

Определим далее последовательность $\{(\tau_i^*, \boldsymbol{\zeta}_i^*, v_i^*);\ i = 1, \ldots\}$ независимых копий случайного вектора $(\tau^*, \boldsymbol{\zeta}^*, v^*)$ и для $n = 0, 1, \ldots$ обозначим через $(T_n^*, \boldsymbol{Z}_n^*, V_n^*)$ частичные суммы этих векторов. Процесс восстановления определим естественным образом $\nu^*(t) := \sup\{n \geqslant 0 :\ T_n^* \leqslant t\}$. Тогда можно определить новую пару ОПВ:

$$(T^*(t), \boldsymbol{Z}^*(t)) := \big(T_{\nu^*(t)}^*, \boldsymbol{Z}_{\nu^*(t)}^*\big).$$

Поясним, как определяется распределение (вообще говоря, несобственное) этой новой пары ОПВ:

$$\mathbf{P}(\boldsymbol{Z}^*(t) \in \cdot,\, T^*(t) \in \cdot\,) = \sum_{n \geqslant 0} \mathbf{P}(\boldsymbol{Z}_n^* \in \cdot,\, T_n^* \in \cdot;\, T_n^* \leqslant t < T_n^* + \tau_{n+1}^*) =$$
$$= \sum_{n \geqslant 0} \mathbf{E}(e^{-\lambda^* T_{n+1} + V_{n+1}};\, \boldsymbol{Z}_n \in \cdot;\, T_n \in \cdot,\, T_n \leqslant t < T_n + \tau_{n+1}).$$

Для $i \geqslant 1$ обозначим $\overline{v}_i := -\lambda^* \tau_i + v_i$, так что выполняется

$$\mathbf{P}^*(\tau_i \in \cdot,\, \boldsymbol{\zeta}_i \in \cdot,\, v_i \in \cdot\,) := \mathbf{E}(e^{\overline{v}_i};\, \tau_i \in \cdot,\, \boldsymbol{\zeta}_i \in \cdot,\, v_i \in \cdot\,).$$

Нетрудно видеть, что найдутся такие константы

$$q > 0, \quad 0 < c < \infty, \quad 0 < R < \infty,$$

что для событий $B_i := \{c < \tau_i,\, |\boldsymbol{\zeta}_i| \leqslant R,\, |v_i| \leqslant R\}$, $i \geqslant 1$, выполняются неравенства

$$\mathbf{P}^*(B_i) \geqslant q.$$

Для $T > 0$ обозначим $k(T) := \dfrac{T}{c}$,

$B(T) := \{c < \tau_i,\ |\boldsymbol{\zeta}_i| \leqslant R,\ |v_i| \leqslant R,\ \text{для всех } i \leqslant \nu(T)+1\}$.

Лемма 6. *Для любого* $T > 0$

$\mathbf{P}^*(B(T)) \geqslant q^{k(T)+1}$.

Доказательство. Достаточно заметить, что если $\tau_i > c$, $i \geqslant 1$, то справедливо включение

$$\bigcap_{i=1}^{[k(T)]+1} B_i \subseteq B(T). \quad \blacktriangle$$

Продолжим доказательство (3.2). Очевидно, что

$$\mathbf{E}\left(e^{V_{\nu(t)}};\ \frac{\boldsymbol{Z}_{\nu(t)}}{t} \in (\boldsymbol{\alpha})_{2\varepsilon}\right) = \mathbf{E}\left(e^{\lambda^* T_{\nu(t)} + \overline{V}_{\nu(t)}};\ \frac{\boldsymbol{Z}_{\nu(t)}}{t} \in (\boldsymbol{\alpha})_{2\varepsilon}\right) \geqslant$$
$$\geqslant \int_{T=0}^{2\delta t} \mathbf{E}\left(e^{\lambda^* T_n + \overline{V}_n};\ \frac{\boldsymbol{Z}_n}{t} \in (\boldsymbol{\alpha})_{\varepsilon},\ \frac{T_n}{t} \in (1-\delta)_{\delta},\ t - T_n \in dT\right) \times I(T,\varepsilon),$$

где

$$I(T,\varepsilon) := \mathbf{E}\left(e^{\lambda^* T_{\nu(T)} - \overline{v}_{\nu(T)+1} + \overline{V}_{\nu(T)+1}};\ \frac{\boldsymbol{Z}_{\nu(T)}}{t} \in (\mathbf{0})_{\varepsilon} \cap B(T)\right).$$

Поскольку на событии $\left\{\dfrac{T_n}{n} \in \dfrac{t}{n}(1-\delta)_{\delta}\right\}$ при $T \leqslant 2\delta t$ выполняется

$\lambda^* T_n \geqslant \lambda^* t(1-2\delta)$

и на событии $B(T)$ при $T \leqslant 2\delta t$ и $2R\delta \leqslant c\varepsilon$ выполняются соотношения

$$\lambda^* T_{\nu(T)} - \overline{v}_{\nu(T)+1} \geqslant -R, \quad \left\{\frac{\boldsymbol{Z}_{\nu(T)}}{t} \in (\mathbf{0})_{\varepsilon}\right\} \cap B(T) = B(T),$$

то имеем

$$e^{-\lambda^* t(1-2\delta)}\, \mathbf{E}\left(e^{V_{\nu(t)}};\ \frac{\boldsymbol{Z}_{\nu(t)}}{t} \in (\boldsymbol{\alpha})_{2\varepsilon}\right) \geqslant$$
$$\geqslant \int_{T=0}^{2\delta t} \mathbf{E}\left(e^{\overline{V}_n};\ \frac{\boldsymbol{Z}_n}{n} \in \frac{t}{n}(\boldsymbol{\alpha})_{\varepsilon},\ \frac{T_n}{n} \in \frac{t}{n}(1-\delta)_{\delta},\ t - T_n \in dT\right) \times e^{-R} J(T,\varepsilon),$$

где

$$J(T,\varepsilon) := \mathbf{E}\big(e^{\overline{V}_{\nu(T)+1}};\ B(T)\big) = \mathbf{P}^*(B(T)) \geqslant q^{k(T)+1}.$$

В последнем неравенстве мы воспользовались леммой 6. Получаем

$$\mathbf{E}\biggl(e^{V_{\nu(t)}};\,\frac{\boldsymbol{Z}_{\nu(t)}}{t}\in(\boldsymbol{\alpha})_{2\varepsilon}\biggr)\geqslant$$

$$\geqslant e^{\lambda^* t(1-2\delta)-R}q^{\frac{2\delta t}{c}+1}\int\limits_0^{2t\delta}\mathbf{P}^*\biggl(\frac{\boldsymbol{Z}_n}{n}\in\frac{t}{n}(\boldsymbol{\alpha})_\varepsilon,\;\frac{T_n}{n}\in\frac{t}{n}(1-\delta)_\delta,\;t-T_n\in dT\biggr)=$$

$$=e^{\lambda^* t(1-2\delta)-R}q^{\frac{2\delta t}{c}+1}\,\mathbf{P}^*\biggl(\frac{\boldsymbol{Z}_n}{n}\in\frac{t}{n}(\boldsymbol{\alpha})_\varepsilon,\;\frac{T_n}{n}\in\frac{t}{n}(1-\delta)_\delta\biggr).\tag{4.13}$$

Чтобы продолжить доказательство формулы (3.2), нам понадобится следующее утверждение.

Лемма 7. *Для любых $\varepsilon>0$, $\delta>0$, $r>0$ имеет место следующая оценка снизу в принципе больших уклонений для сумм $\Bigl(\frac{T_n}{n},\frac{\boldsymbol{Z}_n}{n}\Bigr)$, $n:=[rt]$, для несобственного, вообще говоря, распределения $\mathbf{P}^*(\cdot)$:*

$$\liminf_{t\to\infty}\frac{1}{t}\ln\mathbf{P}^*\biggl(\biggl(\frac{T_n}{n},\frac{\boldsymbol{Z}_n}{n}\biggr)\in\frac{t}{n}(1-\delta)_\delta\times(\boldsymbol{\alpha})_\varepsilon\biggr)\geqslant-\Lambda_r^*\bigl((1-\delta)_\delta\times(\boldsymbol{\alpha})_\varepsilon\bigr),\tag{4.14}$$

где $\Lambda_r^*(\theta,\boldsymbol{\alpha}):=r\Lambda^*\Bigl(\frac{(\theta,\boldsymbol{\alpha})}{r}\Bigr)$ и где для множества $B\subset\mathbb{R}_+\times\mathbb{R}^d$

$$\Lambda_r^*(B):=\inf_{(\theta,\boldsymbol{\alpha})\in B}\Lambda_r^*(\theta,\boldsymbol{\alpha}).$$

Доказательство. Наряду с несобственным распределением $\mathbf{P}^*(\tau\in\cdot,\,\boldsymbol{\zeta}\in\cdot\,)$ рассмотрим собственное распределение

$$\widehat{\mathbf{P}}(\tau\in\cdot,\,\boldsymbol{\zeta}\in\cdot\,):=e^{-C}\,\mathbf{P}^*(\tau\in\cdot,\,\boldsymbol{\zeta}\in\cdot\,),\tag{4.15}$$

где $C:=\ln\mathbf{E}\,e^{\overline{v}}$. Функцию уклонений, отвечающую $\widehat{\mathbf{P}}$-распределению вектора $(\tau,\boldsymbol{\zeta})$, обозначим

$$\widehat{\Lambda}(\theta,\boldsymbol{\alpha}):=\sup_{(\lambda,\boldsymbol{\mu})}\bigl\{\lambda\theta+\langle\boldsymbol{\mu},\boldsymbol{\alpha}\rangle-\ln\widehat{\mathbf{E}}e^{\lambda\tau+\langle\boldsymbol{\mu},\boldsymbol{\zeta}\rangle}\bigr\}.$$

Очевидно, что справедливо равенство

$$\widehat{\Lambda}(\theta,\boldsymbol{\alpha})=\Lambda^*(\theta,\boldsymbol{\alpha})+C.\tag{4.16}$$

Воспользуемся теперь известной (см., например, [22, теорема 1.2.1]) оценкой снизу в принципе больших уклонений для сумм $\Bigl(\frac{T_n}{n},\frac{\boldsymbol{Z}_n}{n}\Bigr)$, $n:=[rt]$, для собственного распределения $\widehat{\mathbf{P}}(\cdot)$:

$$\liminf_{t\to\infty}\frac{1}{t}\ln\widehat{\mathbf{P}}\biggl(\biggl(\frac{T_n}{n},\frac{\boldsymbol{Z}_n}{n}\biggr)\in\frac{t}{n}(1-\delta)_\delta\times(\boldsymbol{\alpha})_\varepsilon\biggr)\geqslant-\widehat{\Lambda}_r\bigl((1-\delta)_\delta\times(\boldsymbol{\alpha})_\varepsilon\bigr),\tag{4.17}$$

где $\widehat{\Lambda}_r(\theta,\boldsymbol{\alpha}):=r\widehat{\Lambda}\Bigl(\frac{(\theta,\boldsymbol{\alpha})}{r}\Bigr)$ и где для множества $B\subset\mathbb{R}_+\times\mathbb{R}^d$

$$\widehat{\Lambda}_r(B):=\inf_{(\theta,\boldsymbol{\alpha})\in B}\widehat{\Lambda}_r(\theta,\boldsymbol{\alpha}).$$

Остается заметить, что в силу (4.15) и (4.16) левая (правая) часть (4.14) отличается от левой (правой) части (4.17) на слагаемое $-rC$. ▲

Продолжим доказательство неравенства (3.2). Используя (4.13) и лемму 7, получаем

$$L_-(\boldsymbol{\alpha},2\varepsilon) := \varliminf_{t\to\infty}\frac{1}{t}\ln\mathbf{E}\Big(e^{V_{\nu(t)}}:\ \frac{Z_{\nu(t)}}{t}\in(\boldsymbol{\alpha})_{2\varepsilon}\Big)\geqslant$$
$$\geqslant -\Lambda_r^*((1-\delta)_\delta\times(\boldsymbol{\alpha})_\varepsilon)+\lambda^*-W\delta, \tag{4.18}$$

где $W := \Big(4\lambda^*+\frac{4}{c}|\ln q|\Big)$. Максимизируя правую часть (4.18) по $r>0$, используя обозначения

$$D^*_{\Lambda^*}(\theta,\boldsymbol{\beta}) := \inf_{r>0}\Lambda_r^*(\theta,\boldsymbol{\beta}),\quad \theta>0,\quad \boldsymbol{\beta}\in\mathbb{R}^d,$$
$$D^*_{\Lambda^*}(B) := \inf_{(\theta,\boldsymbol{\beta})\in B} D^*_{\Lambda^*}(\theta,\boldsymbol{\beta}),\quad B\subset(0,\infty)\times\mathbb{R}^d,$$

и равенство

$$\inf_{r>0}\Lambda_r^*\big((1-\delta)_\delta\times(\boldsymbol{\alpha})_\varepsilon\big) = D^*_{\Lambda^*}\big((1-\delta)_\delta\times(\boldsymbol{\alpha})_\varepsilon\big),$$

получаем

$$L_-(\boldsymbol{\alpha},2\varepsilon)\geqslant -D^*_{\Lambda^*}\big((1-\delta)_\delta\times(\boldsymbol{\alpha})_\varepsilon\big)+\lambda^*-W\delta.$$

Поскольку для любого $u>0$ выполняется

$$D^*_{\Lambda^*}(u\theta,u\boldsymbol{\beta}) = uD^*_{\Lambda^*}(\theta,\boldsymbol{\beta}),$$

то из последнего неравенства выводим

$$L_-(\boldsymbol{\alpha},2\varepsilon)\geqslant -(1-\delta)D^*_{\Lambda^*}\big((1)_{\delta'}\times(\boldsymbol{\beta})_{\varepsilon'}\big)+\lambda^*-W\delta\geqslant$$
$$\geqslant -(1-\delta)D^*_{\Lambda^*}\big(\{1\}\times(\boldsymbol{\beta})_{\varepsilon'}\big)+\lambda^*-W\delta, \tag{4.19}$$

где $\delta' := \dfrac{\delta}{1-\delta}$, $\varepsilon' := \dfrac{\varepsilon}{1-\delta}$, $\boldsymbol{\beta} := \dfrac{\boldsymbol{\alpha}}{1-\delta}$. Заметим, что для любого $\varepsilon>0$ найдется $\delta_0=\delta_0(\varepsilon)>0$, такое что для всех $\delta\in(0,\delta_0)$ выполняется

$$(\boldsymbol{\alpha})_{\varepsilon/2}\subset(\boldsymbol{\beta})_{\varepsilon'},$$

и следовательно,

$$-D^*_{\Lambda^*}\big(\{1\}\times(\boldsymbol{\beta})_{\varepsilon'}\big)\geqslant -D^*_{\Lambda^*}\big(\{1\}\times(\boldsymbol{\alpha})_{\varepsilon/2}\big).$$

Используя последнее неравенство для оценки снизу правой части (4.19), получаем

$$L_-(\boldsymbol{\alpha},2\varepsilon)\geqslant -(1-\delta)D^*_{\Lambda^*}\big(\{1\}\times(\boldsymbol{\alpha})_{\varepsilon/2}\big)+\lambda^*-W\delta;$$

устремляя $\delta\downarrow 0$, имеем для любого $N\geqslant 2$

$$L_-(\boldsymbol{\alpha},2\varepsilon)\geqslant -D^*_{\Lambda^*}\big(\{1\}\times(\boldsymbol{\alpha})_{\varepsilon/2}\big)+\lambda^*\geqslant -D^*_{\Lambda^*}\big(\{1\}\times(\boldsymbol{\alpha})_{\varepsilon/N}\big)+\lambda^*;$$

устремляя $N\to\infty$ и используя (4.5), получаем неравенство

$$L_-(\boldsymbol{\alpha},2\varepsilon)\geqslant -\big(D^*_{\Lambda^*}(1,\boldsymbol{\alpha})+\lambda^*\big). \tag{4.20}$$

Осталось установить взаимосвязь между функциями $D^*(u,\boldsymbol{\alpha})$ и $D(\boldsymbol{\alpha})$. Для любого $0<u\leqslant 1$ воспользуемся представлениями

$$D^*(u,\boldsymbol{\alpha}) = \sup_{A^*(\lambda,\boldsymbol{\mu})\leqslant 0}\{\lambda u+\langle\boldsymbol{\mu},\boldsymbol{\alpha}\rangle\},\quad A^*(\lambda,\boldsymbol{\mu}) = A(\lambda-\lambda^*,\boldsymbol{\mu}),$$

в силу которых получаем равенство

$$D^*(u,\boldsymbol{\alpha}) - \lambda^* u = \sup_{A(\lambda-\lambda^*,\boldsymbol{\mu})\leqslant 0} \{\lambda u + \langle \boldsymbol{\mu}, \boldsymbol{\alpha}\rangle\} - \lambda^* u = D(u,\boldsymbol{\alpha}). \tag{4.21}$$

Применяя (4.21) при $u = 1$ к неравенству (4.20) и используя (4.18), получаем доказательство неравенства (3.2). При этом получено доказательство неравенства (3.3) в случае, когда выполнено

$$D(\boldsymbol{\alpha}) = D_{\lambda_-}(\boldsymbol{\alpha}). \tag{4.22}$$

Докажем теперь неравенство (3.3) в случае, когда условие (4.22) не выполнено, т.е. когда

$$D(\boldsymbol{\alpha}) > D_{\lambda_-}(\boldsymbol{\alpha}). \tag{4.23}$$

Заметим, что в случае (4.23) выполняется $\lambda_- < D(\mathbf{0})$ (см. лемму 1, п. (iv)), и следовательно,

$$\lambda_- < \infty. \tag{4.24}$$

Поэтому достаточно доказать неравенство (3.3) в случае (4.24). Проведем это доказательство. В этом случае последний "большой скачок" $\tau_{\nu(t)+1}$ вносит некоторый вклад в асимптотику исследуемой вероятности. Очевидно, что

$$\mathbf{E}\left(e^{V_{\nu(t)}};\ \frac{\boldsymbol{Z}_{\nu(t)}}{t} \in (\boldsymbol{\alpha})_{2\varepsilon}\right) = \mathbf{E}\left(e^{\lambda^* T_{\nu(t)} + \overline{V}_{\nu(t)}};\ \frac{\boldsymbol{Z}_{\nu(t)}}{t} \in (\boldsymbol{\alpha})_{2\varepsilon}\right) \geqslant$$
$$\geqslant \mathbf{E}\left(e^{\lambda^* T_n + \overline{V}_n};\ \frac{\boldsymbol{Z}_n}{t} \in (\boldsymbol{\alpha})_{\varepsilon},\ \frac{T_n}{t} \in (u-\delta)_{\delta}\right) \times \mathbf{P}(\tau > t(1-u+2\delta)),$$

где число $u \in (0,1)$ фиксировано. Поскольку на событии $\left\{\frac{T_n}{n} \in \frac{t}{n}(u-\delta)_{\delta}\right\}$ выполняется

$$\lambda^* T_n \geqslant \lambda^* t(u-2\delta),$$

то имеем

$$\mathbf{E}\left(e^{V_{\nu(t)}};\ \frac{\boldsymbol{Z}_{\nu(t)}}{t} \in (\boldsymbol{\alpha})_{2\varepsilon}\right) \geqslant$$
$$\geqslant e^{\lambda^* t(u-2\delta)}\, \mathbf{E}\left(e^{\overline{V}_n};\ \frac{\boldsymbol{Z}_n}{n} \in \frac{t}{n}(\boldsymbol{\alpha})_{\varepsilon}, \frac{T_n}{n} \in \frac{t}{n}(u-\delta)_{\delta}\right) \times \mathbf{P}(\tau > t(1-u+2\delta)). \tag{4.25}$$

Далее, повторяя с очевидными изменениями вывод из (4.13) неравенства (4.20), выводим из (4.25) для всех $u \in (0,1)$, $\boldsymbol{\alpha} \in \mathbb{R}^d$ неравенство

$$L_-(\boldsymbol{\alpha}, 2\varepsilon) \geqslant -(D^*(u,\boldsymbol{\alpha}) - \lambda^* u + \lambda_-(1-u)). \tag{4.26}$$

Применяя (4.21) к правой части неравенства (4.26), получаем

$$L_-(\boldsymbol{\alpha}, 2\varepsilon) \geqslant -(D(u,\boldsymbol{\alpha}) + \lambda_-(1-u)).$$

Максимизируя правую часть последнего неравенства по $u \in (0,1)$, получаем для всех $\boldsymbol{\alpha} \in \mathbb{R}^d$

$$L_-(\boldsymbol{\alpha}, 2\varepsilon) \geqslant -\widehat{D}_{\lambda_-}(\boldsymbol{\alpha}), \tag{4.27}$$

где функция $\widehat{D}_{\lambda_-}(\boldsymbol{\alpha})$ определена в п. (iii) леммы 5. Выберем теперь произвольные $\boldsymbol{\alpha}' \in (\boldsymbol{\alpha})_{\varepsilon}$ и $\varepsilon' > 0$, такие что выполняется $(\boldsymbol{\alpha}')_{\varepsilon'} \subset (\boldsymbol{\alpha})_{\varepsilon}$. Применяя (4.27) для $\boldsymbol{\alpha}'$

и ε', получаем

$$L_-(\boldsymbol{\alpha}, 2\varepsilon) \geqslant L_-(\boldsymbol{\alpha}', 2\varepsilon') \geqslant -\widehat{D}_{\lambda_-}(\boldsymbol{\alpha}'). \tag{4.28}$$

Максимизируя далее правую часть (4.28) по $\boldsymbol{\alpha}' \in (\boldsymbol{\alpha})_\varepsilon$, для любого $\varepsilon' \in (0, \varepsilon]$ получаем

$$L_-(\boldsymbol{\alpha}, 2\varepsilon) \geqslant -\inf_{\boldsymbol{\alpha}' \in (\boldsymbol{\alpha})_{\varepsilon'}} \widehat{D}_{\lambda_-}(\boldsymbol{\alpha}').$$

Осталось воспользоваться равенством (4.6) и получить утверждение (3.3) при дополнительном условии $\lambda_- < \infty$, что завершает доказательство леммы 3. ▲

Доказательство леммы 4. Легко видеть, что для любых $\gamma > 0$, $\widetilde{\lambda} > 0$ п.н. справедливы неравенства

$$\mathbf{I}\biggl(\frac{|\boldsymbol{Z}(t)|}{t} \geqslant M\biggr) \leqslant \mathbf{I}\bigl(\gamma|\boldsymbol{Z}(t)| - \widetilde{\lambda} T_{\nu(t)} \geqslant M\gamma t - \widetilde{\lambda} t\bigr) \leqslant \frac{e^{\gamma|\boldsymbol{Z}_{\nu(t)}| - \widetilde{\lambda} T_{\nu(t)}}}{e^{M\gamma t - \widetilde{\lambda} t}}, \tag{4.29}$$

$$\mathbf{I}(\nu(t) = k) \leqslant \mathbf{I}(T_k \leqslant t) = \mathbf{I}(e^{-T_k} \geqslant e^{-t}) \leqslant \frac{e^{-T_k}}{e^{-t}}. \tag{4.30}$$

Из условия [$\mathbf{C}^*$] следует, что найдутся $\gamma > 0$ и $\widetilde{\lambda} > 0$, такие что

$$u := \mathbf{E}\, e^{v + \gamma|\boldsymbol{\zeta}| - (\widetilde{\lambda}+1)\tau} < 1. \tag{4.31}$$

Выбирая $\gamma > 0$ и $\widetilde{\lambda} > 0$ так, чтобы было выполнено неравенство (4.31), используя неравенства (4.29), (4.30) и лемму Беппо Леви, получаем

$$\begin{aligned}
&\mathbf{E}\biggl(e^{V_{\nu(t)}}; \frac{|\boldsymbol{Z}(t)|}{t} \geqslant M\biggr) \leqslant \mathbf{E}\biggl(\frac{e^{V_{\nu(t)} + \gamma|\boldsymbol{Z}_{\nu(t)}| - \widetilde{\lambda} T_{\nu(t)}}}{e^{M\gamma t - \widetilde{\lambda} t}}\biggr) \leqslant \\
&\leqslant e^{-M\gamma t + \widetilde{\lambda} t} + \sum_{k=1}^{\infty} \mathbf{E}\biggl(\frac{e^{V_k + \gamma|\boldsymbol{Z}_k| - \widetilde{\lambda} T_k}}{e^{M\gamma t - \widetilde{\lambda} t}};\ \nu(t) = k\biggr) \leqslant \\
&\leqslant e^{-M\gamma t + \widetilde{\lambda} t} + \sum_{k=1}^{\infty} \mathbf{E}\biggl(\frac{e^{V_k + \gamma|\boldsymbol{Z}_k| - (\widetilde{\lambda}+1) T_k}}{e^{M\gamma t - (\widetilde{\lambda}+1) t}}\biggr) \leqslant \\
&\leqslant e^{-M\gamma t + \widetilde{\lambda} t} + e^{-t(M\gamma - \widetilde{\lambda} - 1)} \sum_{k=1}^{\infty} \bigl(\mathbf{E}\, e^{v + \gamma|\boldsymbol{\zeta}| - (\widetilde{\lambda}+1)\tau}\bigr)^k \leqslant \\
&\leqslant \Bigl(1 + \frac{u}{1-u}\Bigr) e^{-t(M\gamma - \widetilde{\lambda} - 1)}.
\end{aligned} \tag{4.32}$$

Используя неравенство (4.32), выбирая $M = \dfrac{N + \widetilde{\lambda} + 1}{\gamma}$, получаем

$$\limsup_{t \to \infty} \frac{1}{t} \ln \mathbf{E}\biggl(e^{V_{\nu(t)}}; \frac{|\boldsymbol{Z}(t)|}{t} \geqslant M\biggr) \leqslant -M\gamma + \widetilde{\lambda} + 1 = -N. \quad ▲$$

СПИСОК ЛИТЕРАТУРЫ

1. *Могульский А.А., Прокопенко Е.И.* Принцип больших уклонений в фазовом пространстве для многомерного первого обобщенного процесса восстановления // Сиб. электрон. матем. изв. 2019. Т. 16. С. 1464–1477. https://doi.org/10.33048/semi.2019.16.101
2. *Могульский А.А., Прокопенко Е.И.* Принцип больших уклонений в фазовом пространстве для многомерного второго обобщенного процесса восстановления // Сиб. электрон. матем. изв. 2019. Т. 16. С. 1478–1492. https://doi.org/10.33048/semi.2019.16.102

3. *Tsirelson B.* From Uniform Renewal Theorem to Uniform Large and Moderate Deviations for Renewal-Reward Processes // Electron. Commun. Probab. 2013. V. 18. № 52. P. 1–13. https://doi.org/10.1214/ECP.v18-2719
4. *Боровков А.А., Могульский А.А.* Интегро-локальные предельные теоремы для обобщенных процессов восстановления при выполнении условия Крамера. I, II // Сиб. матем. журн. 2018. Т. 59. № 3. С. 491–513. https://doi.org/10.17377/smzh.2018.59.302; № 4. С. 736–758. https://doi.org/10.17377/smzh.2018.59.402
5. *Могульский А.А., Прокопенко Е.И.* Интегро-локальные теоремы для многомерных обобщенных процессов восстановления при моментном условии Крамера. I, II, III // Сиб. электрон. матем. изв. 2018. Т. 15. С. 475–502. https://doi.org/10.17377/semi.2018.15.041; С. 503–527. https://doi.org/10.17377/semi.2018.15.042; С. 528–553. https://doi.org/10.17377/semi.2018.15.043
6. *Могульский А.А., Прокопенко Е.И.* Локальные теоремы для арифметических многомерных обобщенных процесов восстановления при выполнении условия Крамера // Матем. тр. 2019. Т. 22. № 2. С. 106–133. https://doi.org/10.33048/mattrudy.2019.22.207
7. *Logachov A., Mogulskii A., Prokopenko E., Yambartsev A.* Local Theorems for (Multidimensional) Additive Functionals of Semi-Markov Chains // Stochastic Process. Appl. 2021. V. 137. P. 149–166. https://doi.org/10.1016/j.spa.2021.03.011
8. *Могульский А.А., Прокопенко Е.И.* Принцип больших уклонений для конечномерных распределений многомерных обобщенных процессов восстановления // Матем. тр. 2020. V. 23. № 2. С. 148–176. https://doi.org/10.33048/mattrudy.2020.23.206
9. *Логачёв А.В., Могульский А.А.* Локальные теоремы для конечномерных приращений арифметических многомерных обобщенных процессов восстановления при выполнении условия Крамера // Сиб. электрон. матем. изв. 2020. Т. 17. С. 1766–1786. https://doi.org/10.33048/semi.2020.17.120
10. *Боровков А.А., Могульский А.А.* Принципы больших уклонений для траектории обобщенных процессов восстановления. I, II // Теория вероятн. и ее примен. 2015. Т. 60. № 2. С. 227–247. https://doi.org/10.4213/tvp4617; № 3. С. 417–438. https://doi.org/10.4213/tvp4631
11. *Logachov A.V., Mogulskii A.A.* Anscombe-type Theorem and Moderate Deviations for Trajectories of a Compound Renewal Process // J. Math. Sci. (N.Y.). 2018. V. 229. P. 36–50. https://doi.org/10.1007/s10958-018-3661-z
12. *Могульский А.А.* Расширенный принцип больших уклонений для траекторий обобщенного процесса восстановления // Матем. тр. 2021. Т. 24. № 1. С. 142–174. https://doi.org/10.33048/mattrudy.2021.24.106
13. *Lefevere R., Mariani M., Zambotti L.* Large Deviations for Renewal Processes // Stochastic Process. Appl. 2011. V. 121. № 10. P. 2243–2271. https://doi.org/10.1016/j.spa.2011.06.005
14. *Бакай Г.А.* Большие уклонения обрывающихся многомерных обобщенных процессов восстановления // Теория вероятн. и ее примен. 2021. Т. 66. № 2. С. 261–283. https://doi.org/10.1016/j.spa.2011.06.005
15. *Zamparo M.* Large Deviations in Renewal Models of Statistical Mechanics // J. Phys. A: Math. Theor. 2019. V. 52. № 49. P. 495004 (31 pp.). https://doi.org/10.1088/1751-8121/ab523f
16. *Zamparo M.* Large Deviations in Discrete-Time Renewal Theory // Stochastic Process. Appl. 2021. V. 139. P. 80–109. https://doi.org/10.1016/j.spa.2021.04.014
17. *Giacomin G.* Random Polymer Models. London: Imperial College Press, 2007.
18. *den Hollander F.* Random Polymers. New York: Springer, 2009.
19. *Zălinescu C.* Convex Analysis in General Vector Spaces. River Edge, N.J.; London: World Sci., 2002.
20. *Могульский А.А., Прокопенко Е.И.* Функция уклонений и базовая функция для многомерного обобщенного процесса восстановления // Сиб. электрон. матем. изв. 2019. Т. 16. С. 1449–1463. https://doi.org/10.33048/semi.2019.16.100

21. *Logachov A., Mogulskii A., Prokopenko E.* Large Deviations Principle for Terminating Multidimensional Compound Renewal Processes with Application to Polymer Pinning Models, arxiv.org/abs/2112.09640 [math.PR], 2021.
22. *Боровков А.А.* Асимптотический анализ случайных блужданий. Быстроубывающие распределения приращений. М.: Физматлит, 2013.

Логачёв Артём Васильевич
Институт математики им. С.Л. Соболева СО РАН, Новосибирск
Новосибирский государственный университет
Новосибирский государственный технический университет
omboldovskaya@mail.ru
Могульский Анатолий Альфредович
Прокопенко Евгений Игоревич
Институт математики им. С.Л. Соболева СО РАН, Новосибирск
Новосибирский государственный университет
mogul@math.nsc.ru
evgenii.prokopenko@gmail.com

Поступила в редакцию
23.12.2021
После доработки
28.03.2022
Принята к публикации
30.03.2022

ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 58 **2022** **Вып. 2**

УДК 621.391.1 : 519.713 : 517.977.5

© 2022 г. **А.В. Колногоров**

ПУАССОНОВСКИЙ ДВУРУКИЙ БАНДИТ: НОВЫЙ ПОДХОД[1]

Рассматривается новый подход к задаче о двуруком бандите с непрерывным временем, в которой доходы описываются пуассоновскими процессами. Для этого, во-первых, горизонт управления разбивается на равные последовательные полуинтервалы, на которых стратегия остается постоянной, а доходы поступают пакетами, соответствующими этим полуинтервалам. Для нахождения оптимальной кусочно-постоянной байесовской стратегии и соответствующего ей байесовского риска получено рекуррентное разностное уравнение. Установлено существование предельной величины байесовского риска, если количество полуинтервалов неограниченно растет, и получено дифференциальное уравнение в частных производных для его нахождения. Во-вторых, в отличие от рассмотренных ранее постановок этой задачи мы исследуем зависимость стратегии от текущей предыстории управляемого процесса, а не от эволюции апостериорного распределения. Это позволяет снять требование конечности множества допустимых параметров, которое накладывалось в прежних постановках. Численные эксперименты показывают, что для практического нахождения байесовских и минимаксных стратегий и рисков достаточно разбить поступающие доходы на 30 пакетов. В случае минимаксной постановки показано, что оптимальная обработка поступающих доходов по одному не является более эффективной, чем оптимальная пакетная обработка, если горизонт управления неограниченно растет.

Ключевые слова: пуассоновский двурукий бандит, байесовский и минимаксный подходы, асимптотическая минимаксная теорема, пакетная обработка.

DOI: 10.31857/S0555292322020065, **EDN:** DZKOCQ

§ 1. Введение

Рассматривается задача о двуруком бандите [1,2], известная также как задача об адаптивном управлении [3, 4] и целесообразном поведении в случайной среде [5, 6], имеющая приложения в медицине, интернет-технологиях, обработке данных. Двурукий бандит – это устройство с двумя рукоятками, называемыми также действиями. Каждый выбор одного из действий сопровождается случайным доходом, распределение которого зависит только от выбранного действия, фиксировано, но неизвестно игроку. Количество игр против двурукого бандита определено заранее и известно. Требуется, наблюдая статистику игры, определить лучшее действие и обеспечить его преимущественное применение с целью максимизации математического ожидания полного дохода, т.е. это задача оптимального управления.

Особенностью рассматриваемой постановки являются непрерывное время и описание доходов пуассоновскими процессами. Такая постановка естественно дополняет

[1] Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (номер проекта 20-01-00062).

классическую формулировку задачи о бернуллиевском двуруком бандите, в которой доходы могут принимать значения 0 и 1. Наиболее известными исследованиями по пуассоновскому двурукому бандиту являются работы [2, 7], в которых установлен ряд важных результатов, в частности, существование оптимальной стратегии и ее пороговый характер, а также описана процедура синтеза оптимального управления. Отметим, что постановка задачи в [2] является даже более общей, чем задача о двуруком бандите. Однако существенным недостатком работ [2,7] является то, что в них рассматриваются только конечные множества допустимых значений параметра управляемого процесса. Это вызвано тем, что в этих работах стратегии управления зависят от эволюции апостериорного распределения, описываемой системой обыкновенных дифференциальных уравнений, размерность которой как раз равна количеству параметров. В качестве других постановок задачи о двуруком бандите с непрерывным временем отметим [1, 8], где рассмотрено управление винеровскими процессами. В [8] исходная задача о бернуллиевском одноруком бандите ставится в байесовской постановке и дискретном времени, а непрерывное время возникает в результате предельного перехода, когда горизонт управления неограниченно растет. В [1] также рассмотрена задача об одноруком бандите в байесовской постановке, однако здесь время сразу предполагается непрерывным. Наконец, в [9] рассмотрена постановка с непрерывным временем для многоруких бандитов, у которых априорные распределения для различных действий являются независимыми, а доходы дисконтируются на бесконечном горизонте управления.

Формально, пуассоновский двурукий бандит – это непрерывный справа скачкообразный управляемый случайный процесс $\{X(t),\ 0 \leqslant t \leqslant T\}$, значения которого интерпретируются как текущие доходы, увеличивающиеся на единицу в моменты скачков. Управление осуществляется с использованием двух действий. Будем использовать обозначение $y((t, t+\varepsilon]) = \ell$, если на полуинтервале $t' \in (t, t+\varepsilon]$, $\varepsilon > 0$, постоянно выбиралось действие $y(t') = \ell$ $(\ell = 1, 2)$. При использовании такого постоянного управления приращения процесса $X(t)$ зависят от выбираемых действий следующим образом:

$$\Pr\Big(X(t+\varepsilon) - X(t) = i \,|\, y((t, t+\varepsilon]) = \ell\Big) = p(i, \varepsilon; \lambda_\ell) = \frac{(\lambda_\ell \varepsilon)^i}{i!} e^{-\lambda_\ell \varepsilon}, \tag{1.1}$$

$i = 0, 1, 2, \ldots$, $\ell = 1, 2$. Величину $X(t+\varepsilon) - X(t)$ будем интерпретировать как пакет доходов, полученных на полуинтервале $(t, t+\varepsilon]$. Как известно, математическое ожидание и дисперсия $X(t+\varepsilon) - X(t)$ в этом случае равны

$$\begin{aligned} &\mathbf{E}\big(X(t+\varepsilon) - X(t) \,|\, y((t, t+\varepsilon]) = \ell\big) = \\ &= \mathbf{D}\big(X(t+\varepsilon) - X(t) \,|\, y((t, t+\varepsilon]) = \ell\big) = \lambda_\ell \varepsilon, \end{aligned} \tag{1.2}$$

$\ell = 1, 2$. Кроме того, при малых ε справедливы приближенные формулы

$$p(0, \varepsilon; \lambda_\ell) = 1 - \lambda_\ell \varepsilon + o(\varepsilon), \quad p(1, \varepsilon; \lambda_\ell) = \lambda_\ell \varepsilon + o(\varepsilon), \quad p(i, \varepsilon; \lambda_\ell) = o(\varepsilon), \tag{1.3}$$

$i = 2, 3, \ldots$, $\ell = 1, 2$. Таким образом, векторный параметр $\theta = (\lambda_1, \lambda_2)$, где λ_1, λ_2 характеризуют интенсивности поступления единичных доходов при выборе первого и второго действий, полностью описывает пуассоновский двурукий бандит. Множество Θ допустимых значений параметра предполагается известным, измеримым относительно меры Лебега и ограниченным, т.е. $\lambda_\ell \leqslant C < \infty$, $\ell = 1, 2$.

Для управления используются кусочно-постоянные стратегии. Для этого горизонт управления разбивается на равные временны́е полуинтервалы длины ε, на которых выбранные действия не меняются, т.е. рассматривается дискретное приближение задачи с непрерывным временем. Стратегия управления σ в момент времени $t = n\varepsilon$, соответствующий началу очередного временно́го полуинтервала, определяет выбор (вообще говоря, рандомизированный) действия $y((t, t+\varepsilon])$ в зависимости от

известной текущей предыстории. Такая предыстория имеет достаточно общий вид, состоящий из последовательности примененных действий и полученных в ответ доходов

$$y((0,\varepsilon]),\quad X(\varepsilon)-X(0),\quad y((\varepsilon,2\varepsilon]),\quad X(2\varepsilon)-X(\varepsilon),\quad \ldots,$$
$$y(((n-1)\varepsilon,n\varepsilon]),\quad X(n\varepsilon)-X((n-1)\varepsilon).$$

Однако можно показать, что в качестве предыстории можно ограничиться достаточной статистикой вида (X_1,t_1,X_2,t_2), где t_1,t_2 – текущие полные времена применения обоих действий ($t_1+t_2=t$), а X_1,X_2 – соответствующие полные доходы. В частности, для рассматриваемых в данной статье байесовских стратегий это следует из представленных в § 3 уравнений, описывающих оптимальное управление. Более подробно кусочно-постоянные стратегии обсуждаются в § 2.

Обозначим текущие значения доходов X_1 и X_2 в момент времени t через $X_1(t)$ и $X_2(t)$. Если бы значения интенсивностей λ_1,λ_2 были известны, то следовало бы всегда применять действие, соответствующее большей из них; при этом полный ожидаемый доход на всем горизонте управления T был бы равен $T\max(\lambda_1,\lambda_2)$. Но поскольку для управления используется стратегия σ, то полный ожидаемый доход меньше максимального на величину

$$L_T(\sigma,\theta)=T\max(\lambda_1,\lambda_2)-\mathbf{E}_{\sigma,\theta}\left(X_1(T)+X_2(T)\right), \tag{1.4}$$

которая называется функцией потерь и вызвана неполнотой информации об управляемом процессе. Здесь $\mathbf{E}_{\sigma,\theta}$ обозначает математическое ожидание, вычисленное по мере, порожденной стратегией σ и параметром θ. Отметим, что из ограниченности множества Θ следует ограниченность функции потерь (1.4).

Зададим априорную плотность распределения $\mu(\theta)=\mu(\lambda_1,\lambda_2)$ на множестве параметров Θ. Тогда математическое ожидание потерь, вычисленных относительно априорной плотности распределения $\mu(\theta)$, равно

$$L_T(\sigma,\mu)=\int\limits_{\Theta} L_T(\sigma,\theta)\mu(\theta)\,d\theta. \tag{1.5}$$

Байесовский риск, вычисленный относительно плотности $\mu(\theta)$, равен

$$R_T^B(\mu)=\inf_{\{\sigma\}} L_T(\sigma,\mu), \tag{1.6}$$

и соответствующая оптимальная стратегия σ^B называется байесовской. Минимаксный риск на множестве Θ равен

$$R_T^M(\Theta)=\inf_{\{\sigma\}}\sup_{\Theta} L_T(\sigma,\theta), \tag{1.7}$$

и соответствующая оптимальная стратегия σ^M называется минимаксной.

Прямого метода для нахождения минимаксных стратегии и риска не существует. Однако их можно найти с использованием основной теоремы теории игр, согласно которой имеет место равенство

$$R_T^M(\Theta)=R_T^B(\mu_0)=\sup_{\{\mu\}} R_T^B(\mu), \tag{1.8}$$

т.е. минимаксный риск равен байесовскому риску, вычисленному относительно наихудшего априорного распределения, на котором байесовский риск достигает максимума, а минимаксная стратегия совпадает с соответствующей байесовской. Отметим, что в случае конечного множества Θ численное нахождение минимаксного

риска в соответствии с равенством (1.8) не представляет труда, поскольку байесовский риск является вогнутой функцией априорного распределения.

Известны общие асимптотические оценки для функции потерь, байесовского и минимаксного рисков при $T \to \infty$, которые справедливы для рассматриваемого дискретного приближения задачи. Асимптотические оценки для функции потерь (1.4) в случае фиксированного, но неизвестного параметра θ даны, например, в [10] и имеют порядок $\ln(T)$. Там же при некоторых ограничениях на априорную плотность распределения $\mu(\theta)$ даны асимптотические оценки для байесовского риска (1.6), которые имеют порядок $\ln^2(T)$. Отметим, что в [2, 7] в ряде случаев байесовский риск оказался асимптотически ограничен (см. замечание 2 в § 3), что противоречит приведенной оценке, однако во всех этих случаях не были выполнены ограничения на априорное распределение, указанные в [10]. Наконец, асимптотическая оценка для минимаксного риска (1.7) вытекает из результатов работы [11] и имеет порядок $T^{1/2}$.

Статья имеет следующую структуру. Параграф 2 посвящен более подробной характеристике рассматриваемых кусочно-постоянных стратегий, в том числе показано, что для этих стратегий выполнены условия основной теоремы теории игр. Стандартное рекуррентное уравнение типа уравнения Беллмана для нахождения функции потерь, а также байесовских стратегии и риска дано в § 3. Отметим, что рассматриваемый подход отличается от представленного в [2,7], поскольку пересчет байесовского риска основан на текущей известной предыстории процесса (X_1, t_1, X_2, t_2), в то время как в [2, 7] байесовский риск рассматривается как функция апостериорного распределения. В § 3 также представлена другая, более удобная для анализа версия рекуррентного уравнения для вычисления байесовских стратегии и риска. В § 4 установлен пороговый характер байесовской стратегии управления и рассмотрен предельный случай, когда число полуинтервалов, на которых определена кусочно-постоянная стратегия, неограниченно растет. Такое управление можно интерпретировать как обработку поступающих доходов по одному. В этом случае доказано существование предела байесовского риска и получено дифференциальное уравнение в частных производных для его нахождения. В § 5 для случая, когда горизонт управления T неограниченно растет, с помощью выбора подходящего априорного распределения получена асимптотическая оценка снизу для минимаксного риска. Согласно этой оценке оптимальная обработка поступающих доходов по одному не является более эффективной, чем оптимальная пакетная обработка, если горизонт управления и количество пакетов неограниченно растут. В § 6 приведены результаты численных экспериментов, которые показывают, что нахождение близкого к оптимальному управления не требует больших вычислительных ресурсов. Например, число полуинтервалов, на которых задана кусочно-постоянная стратегия и формируются поступающие пакеты доходов, достаточно выбрать равным 30, а наихудшее априорное распределение при рассмотрении умеренных горизонтов управления можно сконцентрировать на шести парах параметров. В § 7 содержится заключение.

§ 2. Кусочно-постоянные стратегии управления

В этом параграфе вводится класс непрерывных слева стратегий, кусочно-постоянных на множестве последовательных полуинтервалов. Разобьем горизонт управления T на N последовательных полуинтервалов одинаковой длины ε, так что $T = N\varepsilon$. На каждом из этих полуинтервалов выбранное действие не меняется. Именно, для любых целых n_1, n_2, таких что $n_1 \geqslant 0$, $n_2 \geqslant 0$ и $n_1 + n_2 < N$, положим

$$t_1 = n_1\varepsilon, \quad t_2 = n_2\varepsilon, \quad t = t_1 + t_2.$$

Тогда

$$\Pr\big(y((t,t+\varepsilon])=\ell\,|\,X_1,t_1,X_2,t_2\big)=\sigma_\ell(X_1,t_1,X_2,t_2),$$

где $\sigma_\ell(X_1,t_1,X_2,t_2)$ определяется текущей статистикой (X_1,t_1,X_2,t_2) и постоянна на временно́м полуинтервале $t'\in(n\varepsilon,(n+1)\varepsilon]$. В частности, если стратегия предписывает выбор одного из действий с вероятностью 1, то это действие и будет применяться на всем временно́м полуинтервале.

Если же на временно́м полуинтервале $(t,t+\varepsilon]$ требуется выбирать действия рандомизированно с вероятностями $\varkappa_1,\varkappa_2$, то решение о том, какое действие будет применяться, принимается в начале этого полуинтервала: либо с вероятностью $\varkappa_1$ на всем полуинтервале будет применяться первое действие, обеспечивающее поток событий (единичных доходов) с интенсивностью λ_1, либо с вероятностью $\varkappa_2$ будет применяться второе действие, обеспечивающее поток событий с интенсивностью λ_2, причем этот выбор не зависит от предыстории процесса. В [2] предложен другой способ решения этой проблемы: считать, что в этом случае применяются оба действия одновременно с распределением ресурса между ними, обеспечивая интенсивности потоков событий $\varkappa_1\lambda_1$ и $\varkappa_2\lambda_2$ соответственно, и следовательно, суммарную интенсивность $\varkappa_1\lambda_1+\varkappa_2\lambda_2$. Ясно, что эти способы не эквивалентны, хотя дают одинаковое математическое ожидание полного числа событий на полуинтервале длины ε, равное $(\varkappa_1\lambda_1+\varkappa_2\lambda_2)\varepsilon$. Отметим также, что определение стратегии в [2] не предполагает разбиения горизонта управления на полуинтервалы.

Покажем, что предлагаемый способ в рамках рассматриваемого в данной статье дискретного приближения задачи эквивалентен предложенному в [2], если количество полуинтервалов, на которые разбивается горизонт управления, неограниченно растет. Справедлива следующая

Лемма 1. *Пусть исходный полуинтервал длины ε, на котором оба действия применяются с вероятностями $\varkappa_1$ и $\varkappa_2$, разбит на K последовательных полуинтервалов длины ε/K, на каждом из которых независимо осуществляется данное смешанное управление, и пусть $K\to\infty$. Тогда распределение, характеризующее поток событий на указанном полуинтервале, слабо сходится к распределению пуассоновского процесса с интенсивностью $\varkappa_1\lambda_1+\varkappa_2\lambda_2$.*

Доказательство. В этом случае полное число событий, соответствующих применению ℓ-го действия, равно $\xi_{\ell,1}+\ldots+\xi_{\ell,K}$, где $\xi_{\ell,j}$ соответствует доходу за применение ℓ-го действия на j-м полуинтервале и с учетом (1.3) характеризуется распределением

$$\begin{aligned}
&\Pr(\xi_{\ell,j}=1)=\varkappa_\ell\lambda_\ell\varepsilon/K+o(\varepsilon/K),\\
&\Pr(\xi_{\ell,j}=0)=1-\varkappa_\ell\lambda_\ell\varepsilon/K+o(\varepsilon/K),\\
&\Pr(\xi_{\ell,j}=i)=o(\varepsilon/K),\quad i=2,3,\ldots
\end{aligned}$$

Поэтому соответствующая характеристическая функция равна

$$\varphi_{\xi_{\ell,j}}(t)=\mathbf{E}\,e^{it\xi_{\ell,j}}=1+(e^{it}-1)\varkappa_\ell\lambda_\ell\varepsilon/K+o(\varepsilon/K).$$

Так как все $\{\xi_{\ell,j}\}$ независимы, то характеристическая функция полного дохода $\xi_{\ell,1}+\ldots+\xi_{\ell,K}$ за применение ℓ-го действия на полуинтервале длины ε равна

$$\varphi_{\xi_{\ell,1}+\ldots+\xi_{\ell,K}}(t)=\big(1+(e^{it}-1)\varkappa_\ell\lambda_\ell\varepsilon/K+o(\varepsilon/K)\big)^K,$$

поэтому

$$\varphi_{\xi_{\ell,1}+\ldots+\xi_{\ell,K}}(t)\to e^{\varkappa_\ell\lambda_\ell\varepsilon(e^{it}-1)}\quad\text{при } K\to\infty,$$

т.е. сходится к характеристической функции пуассоновского процесса с интенсивностью $\varkappa_\ell\lambda_\ell$ на полуинтервале длины ε. Поскольку события в обоих потоках независимы, то результирующий поток также имеет распределение Пуассона с суммарной интенсивностью $\varkappa_1\lambda_1 + \varkappa_2\lambda_2$. ▲

Еще один аргумент в пользу кусочно-постоянных стратегий следует из результатов [9], где так же, как и в [2], допускается одновременное применение нескольких действий. В [9] установлено, что для многоруких бандитов со случайными доходами диффузионного типа одновременное применение нескольких действий возможно только с нулевой вероятностью. По-видимому, пуассоновский двурукий бандит обладает сходным свойством. В § 6 приведена типичная траектория отклонений текущих доходов X_1 от их пороговых стратегий, которая ведет себя так же, как соответствующая траектория для диффузионных многоруких бандитов.

Покажем теперь, что для рассматриваемой задачи при использовании кусочно-постоянных стратегий справедлива основная теорема теории игр.

Лемма 2. *Пусть* Θ *– компактное множество, а в качестве стратегий используются определенные выше кусочно-постоянные стратегии. Тогда минимаксный риск* (1.7) *и байесовские риски* (1.6) *связаны равенством*

$$R_T^M(\Theta) = \sup_{\{\mu\}} R_T^B(\mu). \tag{2.1}$$

Доказательство. Рассмотрим следующее сужение множества стратегий. Для достаточно большого M зафиксируем вероятности выбора действий для предысторий (X_1, t_1, X_2, t_2), на которых выполнено условие

$$\max(X_1, X_2) > M$$

(например, в этом случае они всегда выбирают только первое действие). Отметим, что за счет выбора M вероятность появления таких предысторий может быть сделана сколь угодно малой. Так как интенсивности λ_1, λ_2 ограничены, то из малости вероятности наступления события $\max(X_1, X_2) > M$ следует близость функций потерь, вычисленных по всем стратегиям и по рассматриваемому сужению. Поэтому функцию потерь, вычисляемую на исходном множестве стратегий, можно сколь угодно точно приблизить с помощью стратегий из данного более узкого класса.

Покажем, что для этого класса стратегий выполнено равенство (1.8). В соответствии с первой фундаментальной теоремой теории игр (см., например, [12]) для этого достаточно показать, что множество таких стратегий $\{\sigma\}$ является компактным, а функция потерь (1.4) непрерывна по совокупности переменных θ, σ. Эти свойства следуют из замечания 1 (см. § 3). Поэтому равенство (1.8) выполнено для указанного сужения множества стратегий. Устремляя M к бесконечности, получаем, что для кусочно-постоянных стратегий без введенного ограничения выполнено (2.1). ▲

Отметим, что хотя доказательство леммы 2 не гарантирует существования наихудшей априорной плотности распределения в соответствии с равенством (1.8), это можно установить, если ввести подходящее расстояние на исходном классе стратегий, превращающее его в компактное множество. Такой подход использован, например, в [13], где справедливость основной теоремы теории игр установлена для двурукого бандита с нормально распределенными доходами. Отметим также следующее свойство функции потерь, связанное с использованием смешанных стратегий: для любых двух стратегий σ_1, σ_2 и любого $0 < \varkappa < 1$ существует такая стратегия σ, что равенство

$$L_{\varepsilon,T}(\sigma, \theta) = \varkappa L_{\varepsilon,T}(\sigma_1, \theta) + (1 - \varkappa) L_{\varepsilon,T}(\sigma_2, \theta)$$

выполнено при всех θ. Это равенство может быть проверено непосредственно, если выписать полное выражение для $L_{\varepsilon,T}(\sigma,\theta)$ с использованием (3.3)–(3.6). Оно означает, что для выбора смешанной стратегии не требуется выполнять рандомизацию в начале управления. Данное свойство имеет место для двуруких бандитов с любыми распределениями одношаговых доходов. По-видимому, впервые оно отмечено в [14] для бернуллиевского двурукого бандита.

§ 3. Рекуррентные уравнения для нахождения байесовских потерь и риска

В этом параграфе сперва дается стандартное рекуррентное уравнение для вычисления потерь, соответствующих применению некоторой кусочно-постоянной стратегии. Это уравнение позволяет определить оптимальную стратегию, минимизирующую эти потери, т.е. вычислить байесовский риск, для нахождения которого также приведено стандартное рекуррентное уравнение. Затем эти уравнения преобразованы в формы, более удобные для дальнейшего анализа.

Пусть к моменту времени $t = t_1 + t_2$ первое и второе действия применялись на промежутках времени общей длины t_1 и t_2 соответственно, при этом полные доходы за применение первого и второго действий оказались равны X_1 и X_2. Тогда апостериорная плотность распределения в момент времени t равна

$$\mu(\lambda_1,\lambda_2\,|\,X_1,t_1,X_2,t_2)=\frac{p(X_1,t_1;\lambda_1)p(X_2,t_2;\lambda_2)\mu(\lambda_1,\lambda_2)}{\mu(X_1,t_1,X_2,t_2)}, \tag{3.1}$$

где $p(X_\ell,t_\ell;\lambda_\ell)$, $\ell=1,2$, определены в (1.1),

$$\mu(X_1,t_1,X_2,t_2)=\iint\limits_{\Theta} p(X_1,t_1;\lambda_1)p(X_2,t_2;\lambda_2)\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2. \tag{3.2}$$

Так как $p(0,0;\lambda)=1$, то формула (3.1) сохраняется при $t_1=0$ и/или $t_2=0$.

Пусть $\sigma_\ell(X_1,t_1,X_2,t_2)$ – определенная в § 2 кусочно-постоянная стратегия, причем дополнительно предположим, что на начальном промежутке времени длины $2t_0$ действия применяются по очереди – каждое на отрезке времени длины t_0. Данное условие в предположении, что t_0 достаточно мало, удобно использовать при рассмотрении предельного и асимптотического описаний байесовского риска, представленных в §§ 4, 5. Обозначим через

$$L_{T-t}(\sigma,(\lambda_1,\lambda_2))=(T-t)\max(\lambda_1,\lambda_2)-\mathbf{E}_{\sigma,\theta}\left(X_1(T)-X_1(t)+X_2(T)-X_2(t)\right)$$

функцию потерь на горизонте управления $(T-t,T]$, а через

$$L^B_\varepsilon(\sigma;X_1,t_1,X_2,t_2)=\iint\limits_{\Theta} L_{T-t}(\sigma,(\lambda_1,\lambda_2))\mu(\lambda_1,\lambda_2\,|\,X_1,t_1,X_2,t_2)\,d\lambda_1\,d\lambda_2$$

– ее математическое ожидание, вычисленное относительно апостериорной плотности распределения $\mu(\lambda_1,\lambda_2\,|\,X_1,t_1,X_2,t_2)$. Запишем стандартное рекуррентное уравнение для нахождения функции потерь (1.5) относительно апостериорного распределения (3.1). Обозначим $x^+=\max(x,0)$. Тогда

$$L^B_\varepsilon(\sigma;X_1,t_1,X_2,t_2)=\sum_{\ell=1}^{2}\sigma_\ell(X_1,t_1,X_2,t_2)\times L^{B,\ell}_\varepsilon(\sigma;X_1,t_1,X_2,t_2), \tag{3.3}$$

где

$$L^{B,1}_\varepsilon(\sigma;X_1,t_1,X_2,t_2)=L^{B,2}_\varepsilon(\sigma;X_1,t_1,X_2,t_2)=0, \tag{3.4}$$

если $t_1 + t_2 = T$, и далее

$$
\begin{aligned}
L_\varepsilon^{B,1}(\sigma; X_1, t_1, X_2, t_2) &= \iint\limits_\Theta \mu(\lambda_1, \lambda_2 \,|\, X_1, t_1, X_2, t_2) \times \\
&\times \left((\lambda_2 - \lambda_1)^+ \varepsilon + \sum_{j=0}^{\infty} L_\varepsilon^B(\sigma; X_1 + j, t_1 + \varepsilon, X_2, t_2) p(j, \varepsilon; \lambda_1) \right) d\lambda_1 \, d\lambda_2, \\
L_\varepsilon^{B,2}(\sigma; X_1, t_1, X_2, t_2) &= \iint\limits_\Theta \mu(\lambda_1, \lambda_2 \,|\, X_1, t_1, X_2, t_2) \times \\
&\times \left((\lambda_1 - \lambda_2)^+ \varepsilon + \sum_{j=0}^{\infty} L_\varepsilon^B(\sigma; X_1, t_1, X_2 + j, t_2 + \varepsilon) p(j, \varepsilon; \lambda_2) \right) d\lambda_1 \, d\lambda_2
\end{aligned}
\tag{3.5}
$$

при $2t_0 \leqslant t < T$. Здесь $\{L_\varepsilon^{B,\ell}(\sigma; X_1, t_1, X_2, t_2)\}$ описывают ожидаемые потери на оставшемся горизонте управления, если сначала на горизонте длины ε применялось ℓ-е действие, а затем управление осуществлялось в соответствии со стратегией σ. В частности, $(\lambda_2 - \lambda_1)^+ \varepsilon$ и $(\lambda_1 - \lambda_2)^+ \varepsilon$ в силу (1.2) описывают ожидаемые потери дохода на полуинтервале длины ε за применение первого и второго действий соответственно. Нижний индекс ε указывает, что для управления используются кусочно-постоянные стратегии на промежутках времени длины ε. Потери (1.5) вычисляются по формуле

$$
\begin{aligned}
L_{\varepsilon,T}(\sigma, \mu) &= t_0 \iint\limits_\Theta |\lambda_1 - \lambda_2| \mu(\lambda_1, \lambda_2) \, d\lambda_1 \, d\lambda_2 + \\
&+ \sum_{X_1=0}^{\infty} \sum_{X_2=0}^{\infty} L_\varepsilon^B(\sigma; X_1, t_0, X_2, t_0) \mu(X_1, t_0, X_2, t_0),
\end{aligned}
\tag{3.6}
$$

где первое слагаемое описывает потери на начальном горизонте длины $2t_0$, когда действия применяются по очереди, а второе слагаемое – потери на заключительном горизонте длины $T - 2t_0$. Ясно, что если $t_0 = 0$, то

$$L_{\varepsilon,T}(\sigma, \mu) = L_\varepsilon^B(\sigma; 0, 0, 0, 0).$$

Отметим, что для нахождения функции потерь (1.4) надо взять вырожденную априорную плотность распределения, сосредоточенную на параметре (λ_1, λ_2), при этом все апостериорные плотности также останутся вырожденными.

Замечание 1. Рассмотрим стратегии, у которых фиксированы вероятности выбора действий для предысторий, характеризуемых условием $\max(X_1, X_2) > M$, а остальные вероятности $\{\sigma_\ell(X_1, t_1, X_2, t_2)\}$ могут произвольно меняться от 0 до 1 при условии, что при всех предысториях выполнено равенство

$$\sigma_1(X_1, t_1, X_2, t_2) + \sigma_2(X_1, t_1, X_2, t_2) = 1.$$

Определим расстояние между двумя такими стратегиями $\sigma^{(1)}$ и $\sigma^{(2)}$ как

$$\max \left| \sigma_1^{(1)}(X_1, t_1, X_2, t_2) - \sigma_1^{(2)}(X_1, t_1, X_2, t_2) \right|,$$

где максимум берется по всем предысториям (X_1, t_1, X_2, t_2), удовлетворяющим условию $\max(X_1, X_2) \leqslant M$. Расстояние между параметрами $\theta^{(1)} = \left(\lambda_1^{(1)}, \lambda_2^{(1)}\right)$ и $\theta^{(2)} = \left(\lambda_1^{(2)}, \lambda_2^{(2)}\right)$ определим как

$$\max \left(\left| \lambda_1^{(1)} - \lambda_1^{(2)} \right|, \left| \lambda_2^{(1)} - \lambda_2^{(2)} \right| \right).$$

Так как при $\max(\lambda_1, \lambda_2) \leqslant C$ все функции потерь ограничены величиной TC, то все бесконечные суммы в (3.5), (3.6) сходятся равномерно. Поэтому из (3.3)–(3.6) следует, что функция потерь (1.4), вычисляемая при вырожденной априорной плотности распределения, непрерывна относительно введенных расстояний. При этом стратегии полностью определяются вероятностями $\{\sigma_1(X_1, t_1, X_2, t_2)\}$, заданными для предысторий, удовлетворяющим условию $\max(X_1, X_2) \leqslant M$, а их множество эквивалентно единичному кубу соответствующей размерности, который является компактным.

Уравнения (3.3)–(3.5) позволяют найти стратегию, минимизирующую полные потери. Для этого надо, начиная с момента $t_1 + t_2 = T - \varepsilon$ и заканчивая моментом $t_1 + t_2 = 2t_0$, при каждой предыстории (X_1, t_1, X_2, t_2) выбирать то действие, которому соответствует меньшее из значений $L^{B,\ell}(\sigma; X_1, t_1, X_2, t_2)$. Полученные полные потери характеризуют байесовский риск и могут быть вычислены с помощью стандартного рекуррентного уравнения

$$R_\varepsilon^B(X_1, t_1, X_2, t_2) = \min\bigl(R_\varepsilon^{B,1}(X_1, t_1, X_2, t_2), R_\varepsilon^{B,2}(X_1, t_1, X_2, t_2)\bigr), \tag{3.7}$$

где

$$R_\varepsilon^{B,1}(X_1, t_1, X_2, t_2) = R_\varepsilon^{B,2}(X_1, t_1, X_2, t_2) = 0, \tag{3.8}$$

если $t_1 + t_2 = T$, и далее

$$\begin{aligned}
R_\varepsilon^{B,1}(X_1, t_1, X_2, t_2) &= \iint\limits_\Theta \mu(\lambda_1, \lambda_2 \,|\, X_1, t_1, X_2, t_2) \times \\
&\times \left((\lambda_2 - \lambda_1)^+ \varepsilon + \sum_{j=0}^{\infty} R_\varepsilon^B(X_1 + j, t_1 + \varepsilon, X_2, t_2) p(j, \varepsilon; \lambda_1) \right) d\lambda_1\, d\lambda_2, \\
R_\varepsilon^{B,2}(X_1, t_1, X_2, t_2) &= \iint\limits_\Theta \mu(\lambda_1, \lambda_2 \,|\, X_1, t_1, X_2, t_2) \times \\
&\times \left((\lambda_1 - \lambda_2)^+ \varepsilon + \sum_{j=0}^{\infty} R_\varepsilon^B(X_1, t_1, X_2 + j, t_2 + \varepsilon) p(j, \varepsilon; \lambda_2) \right) d\lambda_1\, d\lambda_2
\end{aligned} \tag{3.9}$$

при $2t_0 \leqslant t < T$. Здесь $\{R_\varepsilon^{B,\ell}(X_1, t_1, X_2, t_2)\}$ описывают ожидаемые потери на горизонте управления $(t, T]$, вычисленные относительно апостериорной плотности распределения $\mu(\lambda_1, \lambda_2 \,|\, X_1, t_1, X_2, t_2)$, если сначала на горизонте длины ε применялось ℓ-е действие, а затем управление осуществлялось оптимально, а $\{R_\varepsilon^B(X_1, t_1, X_2, t_2)\}$ описывают соответствующие байесовские риски. Байесовский риск (1.6) вычисляется по формуле

$$\begin{aligned}
R_{\varepsilon,T}^B(\mu) &= t_0 \iint\limits_\Theta |\lambda_1 - \lambda_2| \mu(\lambda_1, \lambda_2)\, d\lambda_1\, d\lambda_2 + \\
&+ \sum_{X_1=0}^{\infty} \sum_{X_2=0}^{\infty} R_\varepsilon^B(X_1, t_0, X_2, t_0) \mu(X_1, t_0, X_2, t_0).
\end{aligned} \tag{3.10}$$

Если $t_0 = 0$, то

$$R_{\varepsilon,T}(\mu) = R_\varepsilon^B(0, 0, 0, 0).$$

Наряду с байесовским риском уравнения (3.7)–(3.9) позволяют найти байесовскую стратегию. Байесовская стратегия предписывает выбирать ℓ-е действие, т.е.

$$\sigma_\ell(X_1,t_1,X_2,t_2)=1 \quad \text{при} \quad t'\in(t,t+\varepsilon],$$

если меньшую величину имеет $R_\varepsilon^{B,\ell}(X_1,t_1,X_2,t_2)$ $(\ell=1,2)$. В случае равенства $R_\varepsilon^{B,1}(X_1,t_1,X_2,t_2)=R_\varepsilon^{B,2}(X_1,t_1,X_2,t_2)$ выбор действия может быть произвольным.

Замечание 2. В случае конечного множества параметров

$$\{\theta_i=(\lambda_{1,i},\lambda_{2,i}),\ i=1,\dots,K\}$$

из результатов [2] вытекает следующее интересное свойство. Если в каждом из множеств $\{\lambda_{1,1},\dots,\lambda_{1,K}\}$ и $\{\lambda_{2,1},\dots,\lambda_{2,K}\}$ нет совпадающих элементов, то байесовский риск $R_{\varepsilon,T}^B(\mu)$ асимптотически конечен при $T\to\infty$.

Замечание 3. Для некоторого $k>0$ рассмотрим следующую замену переменных:

$$\begin{aligned}
&\lambda'_\ell=k\lambda_\ell, \quad \mu'(\lambda'_1,\lambda'_2)=k^{-2}\mu(\lambda_1,\lambda_2), \quad \varepsilon'=k^{-1}\varepsilon,\\
&t'_0=k^{-1}t_0, \quad T'=k^{-1}T, \quad t'_\ell=k^{-1}t_\ell, \quad X'_\ell=X_\ell,\\
&\sigma'_\ell(X'_1,t'_1,X'_2,t'_2)=\sigma_\ell(X_1,t_1,X_2,t_2), \quad \ell=1,2.
\end{aligned} \tag{3.11}$$

Тогда при всех X_1,t_1,X_2,t_2 справедливы равенства

$$\begin{aligned}
&R'^{B,\ell}_{\varepsilon'}(X'_1,t'_1,X'_2,t'_2)=R_\varepsilon^{B,\ell}(X_1,t_1,X_2,t_2),\\
&L'^{B,\ell}_{\varepsilon'}(\sigma';X'_1,t'_1,X'_2,t'_2)=L_\varepsilon^{B,\ell}(\sigma;X_1,t_1,X_2,t_2), \quad \ell=1,2,
\end{aligned} \tag{3.12}$$

где через $R'_{\varepsilon'}(\cdot)$ и $L'_{\varepsilon'}(\cdot)$ обозначены байесовские риски и потери, вычисленные относительно априорной плотности распределения μ'. Из (3.12) следует также выполнение равенств

$$R'^{B}_{\varepsilon'}(X'_1,t'_1,X'_2,t'_2)=R_\varepsilon^{B}(X_1,t_1,X_2,t_2)$$

и

$$L'^{B}_{\varepsilon'}(\sigma';X'_1,t'_1,X'_2,t'_2)=L_\varepsilon^{B}(\sigma;X_1,t_1,X_2,t_2)$$

при всех X_1,t_1,X_2,t_2. В том числе справедливы равенства

$$R^B_{\varepsilon',T'}(\mu')=R^B_{\varepsilon,T}(\mu), \quad L^B_{\varepsilon',T'}(\sigma',\mu')=L^B_{\varepsilon,T}(\sigma,\mu). \tag{3.13}$$

Равенства (3.12), (3.13) устанавливаются выполнением замены переменных (3.11) в формулах (3.3)–(3.6) и (3.7)–(3.10). Равенства (3.12), (3.13) означают, что задачу о пуассоновском двуруком бандите всегда можно рассматривать на единичном горизонте управления $T=1$.

Получим более удобную для вычислений и анализа форму рекуррентного уравнения (3.7)–(3.9). Положим

$$\begin{aligned}
&\widetilde{p}(X_\ell,t_\ell;\lambda_\ell)=\lambda_\ell^{X_\ell}e^{-\lambda_\ell t_\ell}=\frac{p(X_\ell,t_\ell;\lambda_\ell)X_\ell!}{t_\ell^{X_\ell}},\\
&\widetilde{\mu}(X_1,t_1,X_2,t_2)=\iint\limits_\Theta \widetilde{p}(X_1,t_1;\lambda_1)\widetilde{p}(X_2,t_2;\lambda_2)\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2,\\
&R_\varepsilon(X_1,t_1,X_2,t_2)=R_\varepsilon^B(X_1,t_1,X_2,t_2)\widetilde{\mu}(X_1,t_1,X_2,t_2).
\end{aligned} \tag{3.14}$$

Отметим, что

$$\mu(\lambda_1,\lambda_2\,|\,X_1,t_1,X_2,t_2)=\frac{\widetilde{p}(X_1,t_1;\lambda_1)\widetilde{p}(X_2,t_2;\lambda_2)\mu(\lambda_1,\lambda_2)}{\widetilde{\mu}(X_1,t_1,X_2,t_2)}. \tag{3.15}$$

Справедлива следующая

Теорема 1. *Рассмотрим рекуррентное разностное уравнение*

$$R_\varepsilon(X_1,t_1,X_2,t_2)=\min\bigl(R_\varepsilon^{(1)}(X_1,t_1,X_2,t_2),R_\varepsilon^{(2)}(X_1,t_1,X_2,t_2)\bigr), \tag{3.16}$$

где

$$R_\varepsilon^{(1)}(X_1,t_1,X_2,t_2)=R_\varepsilon^{(2)}(X_1,t_1,X_2,t_2)=0, \tag{3.17}$$

если $t_1+t_2=T$, *и далее*

$$\begin{aligned}
R_\varepsilon^{(1)}(X_1,t_1,X_2,t_2)&=\varepsilon g^{(1)}(X_1,t_1,X_2,t_2)+\mathbf{T}_\varepsilon^{(1)}R_\varepsilon(X_1,t_1+\varepsilon,X_2,t_2),\\
R_\varepsilon^{(2)}(X_1,t_1,X_2,t_2)&=\varepsilon g^{(2)}(X_1,t_1,X_2,t_2)+\mathbf{T}_\varepsilon^{(2)}R_\varepsilon(X_1,t_1,X_2,t_2+\varepsilon)
\end{aligned} \tag{3.18}$$

при $2t_0\leqslant t<T$. *Здесь функции* $\{g^{(\ell)}(X_1,t_1,X_2,t_2)\}$ *и операторы* $\{\mathbf{T}_\varepsilon^{(\ell)}\}$ *таковы*:

$$\begin{aligned}
g^{(1)}(X_1,t_1,X_2,t_2)&=\iint\limits_\Theta(\lambda_2-\lambda_1)^+\lambda_1^{X_1}e^{-\lambda_1t_1}\lambda_2^{X_2}e^{-\lambda_2t_2}\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2,\\
g^{(2)}(X_1,t_1,X_2,t_2)&=\iint\limits_\Theta(\lambda_1-\lambda_2)^+\lambda_1^{X_1}e^{-\lambda_1t_1}\lambda_2^{X_2}e^{-\lambda_2t_2}\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2,\\
\mathbf{T}_\varepsilon^{(1)}F(X_1,t_1,X_2,t_2)&=\sum_{j=0}^{\infty}F(X_1+j,t_1,X_2,t_2)\times\frac{\varepsilon^j}{j!},\\
\mathbf{T}_\varepsilon^{(2)}F(X_1,t_1,X_2,t_2)&=\sum_{j=0}^{\infty}F(X_1,t_1,X_2+j,t_2)\times\frac{\varepsilon^j}{j!}.
\end{aligned} \tag{3.19}$$

Байесовская стратегия предписывает выбирать ℓ*-е действие* (*иными словами*, $\sigma_\ell(X_1,t_1,X_2,t_2)=1$), *если* $R_\varepsilon^{(\ell)}(X_1,t_1,X_2,t_2)$ *имеет меньшую величину* ($\ell=1,2$). *В случае равенства* $R_\varepsilon^{(1)}(X_1,t_1,X_2,t_2)=R_\varepsilon^{(2)}(X_1,t_1,X_2,t_2)$ *выбор действия может быть произвольным. Байесовский риск* (1.6) *вычисляется по формуле*

$$\begin{aligned}
R_{\varepsilon,T}(\mu)&=t_0\iint\limits_\Theta|\lambda_1-\lambda_2|\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2+\\
&+\sum_{X_1=0}^{\infty}\sum_{X_2=0}^{\infty}R_\varepsilon(X_1,t_0,X_2,t_0)\frac{t_0^{X_1}t_0^{X_2}}{X_1!\,X_2!},
\end{aligned} \tag{3.20}$$

в частности, $R_{\varepsilon,T}(\mu)=R_\varepsilon(0,0,0,0)$ *при* $t_0=0$.

Доказательство. Левую и правую части первого уравнения в (3.9) домножим на $\widetilde{\mu}(X_1,t_1,X_2,t_2)$. С учетом (3.15) получим первое уравнение из (3.18), где $g^{(1)}(X_1,t_1,X_2,t_2)$ имеет вид (3.19). Далее,

$$\mathbf{T}_\varepsilon^{(1)}R_\varepsilon(X_1,t_1+\varepsilon,X_2,t_2)=\sum_{j=0}^{\infty}R_\varepsilon(X_1+j,t_1+\varepsilon,X_2,t_2)\times h_\varepsilon(j),$$

где $h_\varepsilon(j)$ с учетом (3.9), (3.14) при $t_1 > 0$ равна

$$\frac{\iint\limits_\Theta \widetilde{p}(X_1,t_1;\lambda_1)\widetilde{p}(X_2,t_2;\lambda_2)p(j,\varepsilon;\lambda_1)\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2}{\iint\limits_\Theta \widetilde{p}(X_1+j,t_1+\varepsilon;\lambda_1)\widetilde{p}(X_2,t_2;\lambda_2)\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2} =$$
$$= \frac{\widetilde{p}(X_1,t_1;\lambda_1)p(j,\varepsilon;\lambda_1)}{\widetilde{p}(X_1+j,t_1+\varepsilon;\lambda_1)} = \frac{\varepsilon^j}{j!},$$

что соответствует (3.19). При $t_1 = 0$ также $X_1 = 0$, поэтому $h_\varepsilon(j)$ равна

$$\frac{\iint\limits_\Theta \widetilde{p}(X_2,t_2;\lambda_2)p(j,\varepsilon;\lambda_1)\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2}{\iint\limits_\Theta \widetilde{p}(j,\varepsilon;\lambda_1)\widetilde{p}(X_2,t_2;\lambda_2)\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2} = \frac{p(j,\varepsilon;\lambda_1)}{\widetilde{p}(j,\varepsilon;\lambda_1)} = \frac{\varepsilon^j}{j!},$$

что также соответствует (3.19). Проверка второго равенства (3.18) выполняется аналогично. Равенство (3.20) следует из (3.10) с учетом (3.14). ▲

Запишем также новую форму рекуррентного уравнения для нахождения потерь. Справедлива следующая

Теорема 2. *Для заданной стратегии $\sigma(X_1,t_1,X_2,t_2)$ рассмотрим рекуррентное уравнение*

$$L_\varepsilon(\sigma;X_1,t_1,X_2,t_2) = \sum_{\ell=1}^{2} \sigma_\ell(X_1,t_1,X_2,t_2) \times L_\varepsilon^{(\ell)}(\sigma;X_1,t_1,X_2,t_2), \tag{3.21}$$

где

$$L_\varepsilon^{(1)}(\sigma;X_1,t_1,X_2,t_2) = L_\varepsilon^{(2)}(\sigma;X_1,t_1,X_2,t_2) = 0, \tag{3.22}$$

если $t_1 + t_2 = T$, и далее

$$\begin{aligned} L_\varepsilon^{(1)}(\sigma;X_1,t_1,X_2,t_2) &= \varepsilon g^{(1)}(X_1,t_1,X_2,t_2) + \mathbf{T}_\varepsilon^{(1)} L_\varepsilon(\sigma;X_1,t_1+\varepsilon,X_2,t_2),\\ L_\varepsilon^{(2)}(\sigma;X_1,t_1,X_2,t_2) &= \varepsilon g^{(2)}(X_1,t_1,X_2,t_2) + \mathbf{T}_\varepsilon^{(2)} L_\varepsilon(\sigma;X_1,t_1,X_2,t_2+\varepsilon) \end{aligned} \tag{3.23}$$

при $2t_0 \leqslant t < T$, где $\{g^{(\ell)}(X_1,t_1,X_2,t_2)\}$ и $\{\mathbf{T}_\varepsilon^{(\ell)}\}$ определены в (3.19). *Тогда полные потери* (1.5) *вычисляются по формуле*

$$\begin{aligned} L_{\varepsilon,T}(\sigma,\mu) &= t_0 \iint\limits_\Theta |\lambda_1-\lambda_2|\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2 +\\ &+ \sum_{X_1=0}^{\infty}\sum_{X_2=0}^{\infty} L_\varepsilon(\sigma;X_1,t_0,X_2,t_0)\frac{t_0^{X_1}t_0^{X_2}}{X_1!\,X_2!}, \end{aligned} \tag{3.24}$$

в частности, $L_{\varepsilon,T}(\sigma,\mu) = L_\varepsilon(0,0,0,0)$ при $t_0 = 0$.

Доказательство проводится аналогично доказательству теоремы 1, если подставить $L(\sigma;X_1,t_1,X_2,t_2) = L^B(\sigma;X_1,t_1,X_2,t_2)\widetilde{\mu}(X_1,t_1,X_2,t_2)$ в формулы (3.3)–(3.6).

Замечание 4. Нетрудно проверить, что при использовании замены переменных (3.11) для любых X_1,t_1,X_2,t_2 будут выполнены равенства

$$\begin{aligned} R_{\varepsilon'}^{\prime(\ell)}(X_1',t_1',X_2',t_2') &= k^{X_1+X_2} R_\varepsilon^{(\ell)}(X_1,t_1,X_2,t_2),\\ L_{\varepsilon'}^{\prime(\ell)}(\sigma';X_1',t_1',X_2',t_2') &= k^{X_1+X_2} L_\varepsilon^{(\ell)}(\sigma;X_1,t_1,X_2,t_2), \end{aligned}$$

$\ell = 1, 2$, где через $R'_{\varepsilon'}(\cdot)$ и $L'_{\varepsilon'}(\cdot)$ обозначены риски и потери, вычисленные относительно априорной плотности μ'. При этом равенства (3.13) сохранятся.

§ 4. Предельное описание

В этом параграфе сначала устанавливается пороговый характер стратегии управления. Далее устанавливается существование непрерывного по t_1, t_2 предела риска $R_\varepsilon(X_1, t_1, X_2, t_2)$ при $\varepsilon \to +0$. Наконец, для предельного риска получено дифференциальное уравнение в частных производных.

Установим пороговый характер байесовской стратегии управления. Он выражается в том, что байесовскую стратегию всегда можно выбрать так, что на множествах значений статистики $(X_1, t_1 - t, X_2, t_2 + t)$, где X_1, t_1, X_2, t_2 фиксированы, а t возрастает, смена оптимального действия со второго на первое произойдет не более чем при одном t. Аналогичным свойством обладают статистики $(X_1 + x, t_1, X_2 - x, t_2)$, где X_1, t_1, X_2, t_2 фиксированы, а x возрастает. Обозначим

$$\Delta R_\varepsilon(X_1, t_1, X_2, t_2) = R_\varepsilon^{(1)}(X_1, t_1, X_2, t_2) - R_\varepsilon^{(2)}(X_1, t_1, X_2, t_2).$$

Ясно, что критериями выбора первого и второго действий являются неравенства $\Delta R_\varepsilon(X_1, t_1, X_2, t_2) < 0$ и $\Delta R_\varepsilon(X_1, t_1, X_2, t_2) > 0$ соответственно, а в случае, когда $\Delta R_\varepsilon(X_1, t_1, X_2, t_2) = 0$, действия можно выбирать произвольно. Справедлива следующая

Теорема 3. *При любой априорной плотности распределения $\mu(\lambda_1, \lambda_2)$ функции*

$$\Delta R_\varepsilon(X_1, t_1 - t, X_2, t_2 + t) \quad \textit{и} \quad \Delta R_\varepsilon(X_1 + x, t_1, X_2 - x, t_2) \tag{4.1}$$

являются монотонно невозрастающими функциями t и x соответственно.

Доказательство. Обозначим через $R_\varepsilon^{(12)}(X_1, t_1, X_2, t_2)$ и $R_\varepsilon^{(21)}(X_1, t_1, X_2, t_2)$ потери, если сначала по очереди применяются первое и второе (соответственно, второе и первое) действия, а затем управление осуществляется оптимально. Положим

$$\Delta R_\varepsilon^{(1)}(X_1, t_1, X_2, t_2) = R_\varepsilon^{(1)}(X_1, t_1, X_2, t_2) - R_\varepsilon^{(12)}(X_1, t_1, X_2, t_2),$$
$$\Delta R_\varepsilon^{(2)}(X_1, t_1, X_2, t_2) = R_\varepsilon^{(2)}(X_1, t_1, X_2, t_2) - R_\varepsilon^{(21)}(X_1, t_1, X_2, t_2).$$

Ясно, что

$$\Delta R_\varepsilon(X_1, t_1, X_2, t_2) = \Delta R_\varepsilon^{(1)}(X_1, t_1, X_2, t_2) - \Delta R_\varepsilon^{(2)}(X_1, t_1, X_2, t_2).$$

Введем также обозначения $x^+ = \max(x, 0)$ и $x^- = \max(-x, 0)$. Из первого уравнения в (3.18) следует, что

$$R_\varepsilon^{(12)}(X_1, t_1, X_2, t_2) = \varepsilon g^{(1)}(X_1, t_1, X_2, t_2) + \mathbf{T}_\varepsilon^{(1)} R_\varepsilon^{(2)}(X_1, t_1 + \varepsilon, X_2, t_2),$$

поэтому, вычитая это уравнение из первого уравнения (3.18), получаем

$$\Delta R_\varepsilon^{(1)}(X_1, t_1, X_2, t_2) = -\mathbf{T}_\varepsilon^{(1)} \Delta R_\varepsilon^{-}(X_1, t_1 + \varepsilon, X_2, t_2). \tag{4.2}$$

Аналогично,

$$-\Delta R_\varepsilon^{(2)}(X_1, t_1, X_2, t_2) = \mathbf{T}_\varepsilon^{(2)} \Delta R_\varepsilon^{+}(X_1, t_1, X_2, t_2 + \varepsilon). \tag{4.3}$$

Проверим монотонность функции $\Delta R_\varepsilon(X_1, t_1 - t, X_2, t_2 + t)$ по t по индукции. При $t_1 + t_2 = T - \varepsilon$ имеем

$$\Delta R_\varepsilon(X_1, t_1 - t, X_2, t_2 + t) = \varepsilon \iint\limits_\Theta \Big((\lambda_2 - \lambda_1)^+ \lambda_1^{X_1} e^{-\lambda_1 t_1} \lambda_2^{X_2} e^{-\lambda_2 t_2} e^{-(\lambda_2 - \lambda_1)t} -$$

$$- (\lambda_1 - \lambda_2)^+ \lambda_1^{X_1} e^{-\lambda_1 t_1} \lambda_2^{X_2} e^{-\lambda_2 t_2} e^{(\lambda_1 - \lambda_2)t} \Big) \mu(\lambda_1, \lambda_2)\, d\lambda_1\, d\lambda_2.$$

Так как

$$(\lambda_2 - \lambda_1)^+ \lambda_1^{X_1} e^{-\lambda_1 t_1} \lambda_2^{X_2} e^{-\lambda_2 t_2} e^{-(\lambda_2 - \lambda_1)t} - (\lambda_1 - \lambda_2)^+ \lambda_1^{X_1} e^{-\lambda_1 t_1} \lambda_2^{X_2} e^{-\lambda_2 t_2} e^{(\lambda_1 - \lambda_2)t}$$

является монотонно невозрастающей функцией t, то $\Delta R_\varepsilon(X_1, t_1 - t, X_2, t_2 + t)$ – также монотонно невозрастающая функция t при $t_1 + t_2 = T - \varepsilon$. Далее, если $\Delta R_\varepsilon(X_1, t_1 - t, X_2, t_2 + t)$ – монотонно невозрастающая функция t, то такими же являются функции $\Delta R_\varepsilon^+(X_1, t_1 - t, X_2, t_2 + t)$ и $-\Delta R_\varepsilon^-(X_1, t_1 - t, X_2, t_2 + t)$. Поэтому из (4.2), (4.3) и равенства

$$\Delta R_\varepsilon(X_1, t_1 - t, X_2, t_2 + t) = \Delta R_\varepsilon^{(1)}(X_1, t_1 - t, X_2, t_2 + t) -$$

$$- \Delta R_\varepsilon^{(2)}(X_1, t_1 - t, X_2, t_2 + t)$$

следует, что если $\Delta R_\varepsilon(X_1, t_1 - t, X_2, t_2 + t)$ является монотонно невозрастающей функцией t при некоторых $t_1 + t_2 = \tau + \varepsilon$, то это свойство сохранится и при $t_1 + t_2 = \tau$.

Проверка того, что $\Delta R_\varepsilon(X_1 + x, t_1, X_2 - x, t_2)$ является монотонно невозрастающей функцией x, выполняется аналогично по индукции, причем при $t_1 + t_2 = T - \varepsilon$ следует рассмотреть выражение

$$\Delta R_\varepsilon(X_1 + x, t_1, X_2 - x, t_2) = \varepsilon \iint\limits_\Theta \Big((\lambda_2 - \lambda_1)^+ \lambda_1^{X_1} e^{-\lambda_1 t_1} \lambda_2^{X_2} e^{-\lambda_2 t_2} (\lambda_1/\lambda_2)^x -$$

$$- (\lambda_1 - \lambda_2)^+ \lambda_1^{X_1} e^{-\lambda_1 t_1} \lambda_2^{X_2} e^{-\lambda_2 t_2} (\lambda_1/\lambda_2)^x \Big) \mu(\lambda_1, \lambda_2)\, d\lambda_1\, d\lambda_2. \quad \blacktriangle$$

Таким образом, для обеспечения порогового характера стратегии достаточно обеспечить его при $\Delta R_\varepsilon(X_1, t_1, X_2, t_2) = 0$. Например, можно в этом случае всегда выбирать первое действие. При этом возможно, что одно из действий не будет выбрано ни при какой предыстории (X_1, t_1, X_2, t_2); например, так будет, если $\mu(\lambda_1, \lambda_2) > 0$ только при $\lambda_1 > \lambda_2$. Отметим также, что теорема 3 является аналогом теоремы 5 из [14], в которой установлен пороговый характер стратегии для бернуллиевского двурукого бандита.

Перейдем к оценкам рисков $R_\varepsilon(X_1, t_1, X_2, t_2)$.

Лемма 3. *При всех допустимых X_1, t_1, X_2, t_2 справедливы оценки*

$$R_\varepsilon(X_1, t_1, X_2, t_2) \leqslant (T - t) g(X_1, t_1, X_2, t_2), \tag{4.4}$$

где $t = t_1 + t_2$, $g(X_1, t_1, X_2, t_2) = \min\big(g^{(1)}(X_1, t_1, X_2, t_2), g^{(2)}(X_1, t_1, X_2, t_2)\big)$, а функции $g^{(1)}(X_1, t_1, X_2, t_2)$ и $g^{(2)}(X_1, t_1, X_2, t_2)$ определены в (3.19).

Доказательство. Выражение в (4.4) справа характеризует потери, обеспечиваемые стратегией, которая для текущей статистики (X_1, t_1, X_2, t_2) на всем оставшемся горизонте управления длины $T - t$ выбирает действие, которому соответствует меньшее из значений $g^{(1)}(X_1, t_1, X_2, t_2)$, $g^{(2)}(X_1, t_1, X_2, t_2)$. Ясно, что для оптимальной стратегии потери будут не больше указанных. $\blacktriangle$

Следующая лемма является вспомогательной.

Лемма 4. *При $\ell = 1, 2$ справедливы равенства*

$$\mathbf{T}^{(\ell)}_{\varepsilon_1}\mathbf{T}^{(\ell)}_{\varepsilon_2}F(X_1,t_1,X_2,t_2) = \mathbf{T}^{(\ell)}_{\varepsilon_1+\varepsilon_2}F(X_1,t_1,X_2,t_2), \tag{4.5}$$

где $F(X_1,t_1,X_2,t_2)$ – произвольная функция,

$$\begin{aligned}
&\mathbf{T}^{(1)}_{\varepsilon}g^{(\ell)}(X_1,t_1,X_2,t_2) = g^{(\ell)}(X_1,t_1-\varepsilon,X_2,t_2) \quad \textit{при } t_1 > 0,\\
&\mathbf{T}^{(2)}_{\varepsilon}g^{(\ell)}(X_1,t_1,X_2,t_2) = g^{(\ell)}(X_1,t_1,X_2,t_2-\varepsilon) \quad \textit{при } t_2 > 0,
\end{aligned} \tag{4.6}$$

а также оценки

$$\begin{aligned}
&\mathbf{T}^{(\ell)}_{\varepsilon}R_\varepsilon(X_1,t_1,X_2,t_2) - R_\varepsilon(X_1,t_1,X_2,t_2) \leqslant (e^{C\varepsilon}-1)(T-t)g(X_1,t_1,X_2,t_2),\\
&\mathbf{T}^{(1)}_{\varepsilon}R_\varepsilon(X_1,t_1,X_2,t_2) - R_\varepsilon(X_1,t_1,X_2,t_2) - \varepsilon R_\varepsilon(X_1+1,t_1,X_2,t_2) \leqslant\\
&\leqslant (e^{C\varepsilon}-1-\varepsilon C)(T-t)g(X_1,t_1,X_2,t_2),\\
&\mathbf{T}^{(2)}_{\varepsilon}R_\varepsilon(X_1,t_1,X_2,t_2) - R_\varepsilon(X_1,t_1,X_2,t_2) - \varepsilon R_\varepsilon(X_1,t_1,X_2+1,t_2) \leqslant\\
&\leqslant (e^{C\varepsilon}-1-\varepsilon C)(T-t)g(X_1,t_1,X_2,t_2),
\end{aligned} \tag{4.7}$$

где $C = \max\limits_{\Theta}\max\limits_{\ell=1,2}\lambda_\ell$.

Доказательство. Проверим равенство (4.5). Достаточно рассмотреть случай $\ell = 1$. Тогда

$$\begin{aligned}
&\mathbf{T}^{(1)}_{\varepsilon_1}\mathbf{T}^{(1)}_{\varepsilon_2}F(X_1,t_1,\cdot) = \sum_{i=0}^{\infty}\left(\sum_{j=0}^{\infty}F(X_1+j+i,t_1,\cdot)\times\frac{\varepsilon_2^j}{j!}\right)\times\frac{\varepsilon_1^i}{i!} =\\
&= \sum_{k=0}^{\infty}F(X_1+k,t_1,\cdot)\times\left(\sum_{j=0}^{k}\frac{\varepsilon_2^j}{j!}\times\frac{\varepsilon_1^{k-j}}{(k-j)!}\right) = \sum_{k=0}^{\infty}F(X_1+k,t_1,\cdot)\times\frac{(\varepsilon_1+\varepsilon_2)^k}{k!},
\end{aligned}$$

что соответствует (4.5). Проверим равенство (4.6). Достаточно рассмотреть $\mathbf{T}^{(1)}_{\varepsilon}$ при $\ell = 1$. В этом случае

$$\begin{aligned}
&\mathbf{T}^{(1)}_{\varepsilon}g^{(1)}(X_1,t_1,X_2,t_2) =\\
&= \sum_{j=0}^{\infty}\frac{\varepsilon^j}{j!}\iint\limits_{\Theta}(\lambda_2-\lambda_1)^+\lambda_1^{X_1+j}e^{-\lambda_1 t_1}\lambda_2^{X_2}e^{-\lambda_2 t_2}\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2 =\\
&= \iint\limits_{\Theta}(\lambda_2-\lambda_1)^+\left(\sum_{j=0}^{\infty}\frac{(\lambda_1\varepsilon)^j}{j!}\right)\lambda_1^{X_1}e^{-\lambda_1 t_1}\lambda_2^{X_2}e^{-\lambda_2 t_2}\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2 =\\
&= g^{(1)}(X_1,t_1-\varepsilon,X_2,t_2),
\end{aligned}$$

что соответствует (4.6).

Проверим первую оценку в (4.7). Достаточно рассмотреть случай $\ell = 1$. С учетом (3.19), (4.4) и неравенства $g(X_1,t_1,X_2,t_2) \leqslant g^{(1)}(X_1,t_1,X_2,t_2)$ получаем

$$\begin{aligned}
&\mathbf{T}^{(1)}_{\varepsilon}R_\varepsilon(X_1,t_1,X_2,t_2) - R_\varepsilon(X_1,t_1,X_2,t_2) \leqslant\\
&\leqslant (T-t)\sum_{j=1}^{\infty}\frac{\varepsilon^j}{j!}\times g^{(1)}(X_1+j,t_1,X_2,t_2) =
\end{aligned}$$

$$= (T-t)\iint\limits_{\Theta}(\lambda_2-\lambda_1)^+\left(\sum_{j=1}^{\infty}\frac{(\lambda_1\varepsilon)^j}{j!}\right)\lambda_1^{X_1}e^{-\lambda_1 t_1}\lambda_2^{X_2}e^{-\lambda_2 t_2}\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2 \leqslant$$
$$\leqslant (T-t)\left(e^{C\varepsilon}-1\right)g^{(1)}(X_1,t_1,X_2,t_2).$$

Поскольку также выполнено неравенство

$$\mathbf{T}_{\varepsilon}^{(1)}R_{\varepsilon}(X_1,t_1,X_2,t_2)-R_{\varepsilon}(X_1,t_1,X_2,t_2)\leqslant (T-t)\left(e^{C\varepsilon}-1\right)g^{(2)}(X_1,t_1,X_2,t_2),$$

то отсюда следует первая оценка в (4.7). Вторая и третья оценки в (4.7) проверяются аналогично. ▲

Установим условия Липшица для $R_{\varepsilon}(X_1,t_1,X_2,t_2)$ по t_1,t_2. Справедлива следующая

Лемма 5. *Пусть $\delta=K\varepsilon$, где K – целое число. Тогда имеют место оценки*

$$\begin{aligned}
&|R_{\varepsilon}(X_1,t_1,X_2,t_2)-R_{\varepsilon}(X_1,t_1+\delta,X_2,t_2)|\leqslant\\
&\leqslant \delta g^{(1)}(X_1,t_1,X_2,t_2)+(e^{C\delta}-1)(T-t-\delta)g(X_1,t_1+\delta,X_2,t_2),\\
&|R_{\varepsilon}(X_1,t_1,X_2,t_2)-R_{\varepsilon}(X_1,t_1,X_2,t_2+\delta)|\leqslant\\
&\leqslant \delta g^{(2)}(X_1,t_1,X_2,t_2)+(e^{C\delta}-1)(T-t-\delta)g(X_1,t_1,X_2,t_2+\delta).
\end{aligned}\tag{4.8}$$

Доказательство. Достаточно установить первую оценку. Обозначим через $R_{\varepsilon}^{(1,K)}(X_1,t_1,X_2,t_2)$ потери, если сначала K раз применялось первое действие, а затем управление осуществлялось оптимально. Справедливо уравнение

$$\begin{aligned}
&R_{\varepsilon}^{(1,i)}(X_1,t_1+(K-i)\varepsilon,X_2,t_2)=\varepsilon g^{(1)}(X_1,t_1+(K-i)\varepsilon,X_2,t_2)+\\
&+\mathbf{T}_{\varepsilon}^{(1)}R_{\varepsilon}^{(1,i-1)}(X_1,t_1+(K-i+1)\varepsilon,X_2,t_2),
\end{aligned}\tag{4.9}$$

где $i=1,2,\ldots,K$, причем $R_{\varepsilon}^{(1,0)}(X_1,t_1+K\varepsilon,X_2,t_2)=R_{\varepsilon}(X_1,t_1+\delta,X_2,t_2)$. Из (4.9) следует, что

$$\begin{aligned}
&R_{\varepsilon}^{(1,1)}(X_1,t_1+(K-1)\varepsilon,X_2,t_2)=\varepsilon g^{(1)}(X_1,t_1+(K-1)\varepsilon,X_2,t_2)+\\
&+\mathbf{T}_{\varepsilon}^{(1)}R_{\varepsilon}(X_1,t_1+\delta,X_2,t_2).
\end{aligned}$$

Далее, с учетом (4.5), (4.6) получаем, что

$$\begin{aligned}
&R_{\varepsilon}^{(1,i)}(X_1,t_1+(K-i)\varepsilon,X_2,t_2)=i\varepsilon g^{(1)}(X_1,t_1+(K-i)\varepsilon,X_2,t_2)+\\
&+\mathbf{T}_{i\varepsilon}^{(1)}R_{\varepsilon}(X_1,t_1+\delta,X_2,t_2)\quad\text{при } i=2,\ldots,K.
\end{aligned}$$

При $i=K$ получаем

$$R_{\varepsilon}^{(1,K)}(X_1,t_1,X_2,t_2)=\delta g^{(1)}(X_1,t_1,X_2,t_2)+\mathbf{T}_{\delta}^{(1)}R_{\varepsilon}(X_1,t_1+\delta,X_2,t_2).\tag{4.10}$$

С другой стороны, справедлива оценка

$$R_{\varepsilon}(X_1,t_1,X_2,t_2)\geqslant\mathbf{T}_{\delta}^{(1)}R_{\varepsilon}(X_1,t_1+\delta,X_2,t_2),\tag{4.11}$$

которая следует из того, что байесовский риск на меньшем горизонте управления и при наличии дополнительной информации, обусловленной K-кратным применением первого действия, не превосходит исходного байесовского риска. Из (4.10), (4.11) и неравенства

$$R_{\varepsilon}(X_1,t_1,X_2,t_2)\leqslant R_{\varepsilon}^{(1,K)}(X_1,t_1,X_2,t_2)$$

следует, что

$$\left|R_\varepsilon(X_1,t_1,X_2,t_2)-\mathbf{T}_\delta^{(1)}R_\varepsilon(X_1,t_1+\delta,X_2,t_2)\right|\leqslant\delta g^{(1)}(X_1,t_1,X_2,t_2).$$

С учетом первой оценки (4.7) отсюда следует (4.8). ▲

Далее считаем, что $\varepsilon\to 0$. Отметим, что при малых ε управление соответствует обработке доходов по одному. Справедлива следующая

Теорема 4. *При $t_1\geqslant t_0$, $t_2\geqslant t_0$ существует предел*

$$R(X_1,t_1,X_2,t_2)=\lim_{\varepsilon\to+0}R_\varepsilon(X_1,t_1,X_2,t_2). \tag{4.12}$$

Этот предел ограничен в соответствии с оценкой (4.4) *и удовлетворяет условиям Липшица по t_1,t_2 в соответствии с оценками* (4.8). *Байесовский риск* (1.6) *равен*

$$R_T(\mu)=\lim_{t_0\to+0}R(0,t_0,0,t_0). \tag{4.13}$$

Доказательство. Для некоторого ε рассмотрим последовательность $\varepsilon_i=2^{-i}\varepsilon$, $i=1,2,\ldots$ Так как уменьшение величины ε_i означает, что действия можно менять чаще, то $R_{\varepsilon_i}(X_1,t_1,X_2,t_2)$ при фиксированных X_1,t_1,X_2,t_2 является неубывающей функцией ε_i. Поскольку $R_{\varepsilon_i}(X_1,t_1,X_2,t_2)\geqslant 0$, то предел (4.12) существует при всех $\{t_\ell\}$ вида $t_\ell=t_0+k\varepsilon_i$, $\ell=1,2$, $i=1,2,\ldots$ Выполнение оценок (4.4) и (4.8) для $R(X_1,t_1,X_2,t_2)$ устанавливается предельным переходом по $\varepsilon_i\to 0$. Поэтому полученный предел можно по непрерывности доопределить на все $t_1\geqslant t_0$, $t_2\geqslant t_0$. Формула (4.13) следует из (3.20) и (4.4), так как первое слагаемое в правой части (3.20) и все

$$\frac{R(X_1,t_0,X_2,t_0)t_0^{X_1+X_2}}{X_1!\,X_2!}$$

при $X_1+X_2>0$ стремятся к нулю при $t_0\to 0$. ▲

Покажем, что $R(X_1,t_1,X_2,t_2)$ удовлетворяет дифференциальному уравнению в частных производных

$$\begin{aligned}&\min\biggl(\frac{\partial R}{\partial t_1}+R(X_1+1,t_1,X_2,t_2)+g^{(1)}(X_1,t_1,X_2,t_2),\\&\frac{\partial R}{\partial t_2}+R(X_1,t_1,X_2+1,t_2)+g^{(2)}(X_1,t_1,X_2,t_2)\biggr)=0\end{aligned} \tag{4.14}$$

с начальным условием

$$R(X_1,t_1,X_2,t_2)=0\quad\text{при } t_1+t_2=T, \tag{4.15}$$

при этом байесовский риск (1.6) вычисляется по формуле (4.13). Дифференциальное уравнение (4.14) одновременно описывает не только эволюцию байесовского риска $R(X_1,t_1,X_2,t_2)$, но и байесовскую стратегию, которая предписывает выбирать ℓ-е действие, если ℓ-й член в левой части (4.14) имеет меньшее значение; в случае их равенства выбор действия может быть произвольным. Существования такой предельной стратегии, в свою очередь, достаточно для строгого вывода уравнения (4.14). Однако пока этого сделать не удалось, хотя можно, как это сделано в [15], доказать, что этот предел существует в некоторых областях.

Зафиксируем некоторое $\varepsilon>0$, и пусть $\varepsilon_i=2^{-i}\varepsilon$, $i=1,2,\ldots$ Из (4.10) и второй оценки в (4.7) следует уравнение

$$\begin{aligned}&R_{\varepsilon_i}^{(1,2^i)}(X_1,t_1,X_2,t_2)=\varepsilon g^{(1)}(X_1,t_1,X_2,t_2)+\\&+R_{\varepsilon_i}(X_1,t_1+\varepsilon,X_2,t_2)+\varepsilon R_{\varepsilon_i}(X_1+1,t_1+\varepsilon,X_2,t_2)+\alpha(\varepsilon),\end{aligned} \tag{4.16}$$

где $|\alpha(\varepsilon)| \leqslant (e^{C\varepsilon} - 1 - \varepsilon C)(T-t)g(X_1,t_1,X_2,t_2) = O(\varepsilon^2)$. Аналогично,

$$R_{\varepsilon_i}^{(2,2^i)}(X_1,t_1,X_2,t_2) = \varepsilon g^{(2)}(X_1,t_1,X_2,t_2) + \\ + R_{\varepsilon_i}(X_1,t_1,X_2,t_2+\varepsilon) + \varepsilon R_{\varepsilon_i}(X_1,t_1,X_2+1,t_2+\varepsilon) + \alpha(\varepsilon), \tag{4.17}$$

Если при всех t_1', t_2', таких что $t_1' \geqslant t_1$, $t_2' \geqslant t_2$, $t_1' + t_2' < t_1 + t_2 + \varepsilon$, оптимальным является одно и то же действие, то уравнение (3.16), которым следует дополнить уравнения (4.16), (4.17), можно записать в виде

$$\min\Big(R_{\varepsilon_i}^{(1,2^i)}(X_1,t_1,X_2,t_2) - R_{\varepsilon_i}(X_1,t_1,X_2,t_2), \\ R_{\varepsilon_i}^{(2,2^i)}(X_1,t_1,X_2,t_2) - R_{\varepsilon_i}(X_1,t_1,X_2,t_2)\Big) = 0. \tag{4.18}$$

Выполняя предельные переходы сначала по $i \to \infty$, а затем по $\varepsilon \to 0$, из (4.16)–(4.18) получаем уравнение (4.14).

Отметим, что для численного решения уравнения (4.14) с начальным условием (4.15) следует использовать уравнения (3.16)–(3.18), в которых для вычисления операторов $\{\mathbf{T}_\varepsilon^{(\ell)}\}$ надо ограничиться слагаемыми порядка не выше ε.

§ 5. Асимптотическая оценка минимаксного риска снизу

Рассмотрим теперь асимптотическое описание байесовского риска при $T \to \infty$, которое в значительной степени аналогично приведенному в [16] описанию для бернуллиевского двурукого бандита. Будет показано, что при подходящем выборе априорного распределения он описывается тем же дифференциальным уравнением в частных производных второго порядка, что и байесовский риск для гауссовского двурукого бандита. Поскольку минимаксный риск не меньше любого байесовского, а гауссовский двурукий бандит описывает пакетную обработку, эти результаты означают, что минимаксный риск для пуассоновского двурукого бандита при обработке доходов по одному не может быть сделан меньше минимаксного риска, соответствующего оптимальной пакетной обработке, если $T \to \infty$.

Отметим, что в [16] асимптотическое описание было получено для риска

$$\widehat{R}(X_1,t_1,X_2,t_2) = R^B(X_1,t_1,X_2,t_2)\mu(X_1,t_1,X_2,t_2). \tag{5.1}$$

В этом случае при подходящей нормировке переменных X_1,t_1,X_2,t_2 и самого риска можно получить дифференциальное уравнение в частных производных второго порядка. Однако получить соответствующее дифференциальное уравнение для $R_\varepsilon(X_1,t_1,X_2,t_2)$, описываемого разностными уравнениями (3.16)–(3.19), а в предельном случае – дифференциальным уравнением (4.14), не удается по той причине, что этот риск отличается от риска $\widehat{R}_\varepsilon(X_1,t_1,X_2,t_2)$ множителем $X_1!\,X_2!\,t_1^{-X_1}t_2^{-X_2}$, который не является медленно меняющимся при любых изменениях X_1 и X_2.

Получим дифференциальное уравнение для $\widehat{R}(X_1,t_1,X_2,t_2)$. Снова предполагаем, что на начальном этапе управления оба действия применяются по t_0 раз, а затем при $t_1 > t_0$, $t_2 > t_0$ управление осуществляется оптимально в соответствии с уравнением (4.14). Если $t_0 \ll T$, то такая стратегия практически не приводит к увеличению байесовского риска. Справедлива следующая

Теорема 5. *Риск* (5.1) *удовлетворяет дифференциальному уравнению*

$$\min_{\ell=1,2}\left(\frac{\partial\widehat{R}}{\partial t_\ell} + D^{(\ell)}\widehat{R}(X_1,t_1,X_2,t_2) + \widehat{g}^{(\ell)}(X_1,t_1,X_2,t_2)\right) = 0 \tag{5.2}$$

с начальным условием

$$\widehat{R}(X_1,t_1,X_2,t_2)=0 \quad \textit{при } t_1+t_2=T, \tag{5.3}$$

где

$$\begin{aligned}
&D^{(1)}\widehat{R}(X_1,t_1,X_2,t_2)=\big(\widehat{R}(X_1+1,t_1,X_2,t_2)(X_1+1)-\widehat{R}(X_1,t_1,X_2,t_2)X_1\big)t_1^{-1},\\
&D^{(2)}\widehat{R}(X_1,t_1,X_2,t_2)=\big(\widehat{R}(X_1,t_1,X_2+1,t_2)(X_2+1)-\widehat{R}(X_1,t_1,X_2,t_2)X_2\big)t_2^{-1},\\
&\widehat{g}^{(1)}(X_1,t_1,X_2,t_2)=\iint\limits_{\Theta}(\lambda_2-\lambda_1)^+p(X_1,t_1;\lambda_1)p(X_2,t_2;\lambda_2)\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2,\\
&\widehat{g}^{(2)}(X_1,t_1,X_2,t_2)=\iint\limits_{\Theta}(\lambda_1-\lambda_2)^+p(X_1,t_1;\lambda_1)p(X_2,t_2;\lambda_2)\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2.
\end{aligned} \tag{5.4}$$

Байесовский риск (1.6) *вычисляется по формуле*

$$R_T(\mu)=t_0\iint\limits_{\Theta}|\lambda_1-\lambda_2|\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2+\sum_{X_1=0}^{\infty}\sum_{X_2=0}^{\infty}\widehat{R}(X_1,t_0,X_2,t_0). \tag{5.5}$$

Доказательство. Из (5.1) и (3.14) следует, что

$$R(X_1,t_1,X_2,t_2)=\widehat{R}(X_1,t_1,X_2,t_2)X_1!\,X_2!\,t_1^{-X_1}t_2^{-X_2}.$$

Подставляя это выражение в первый член в левой части (4.14), получаем при $X_1>0$

$$\begin{aligned}
&\left(\widehat{R}(X_1,t_1,X_2,t_2)\frac{X_1!\,X_2!}{t_1^{X_1}t_2^{X_2}}\right)'_{t_1}+\widehat{R}(X_1+1,t_1,X_2,t_2)\frac{(X_1+1)!\,X_2!}{t_1^{X_1+1}t_2^{X_2}}+\\
&+g(X_1,t_1,X_2,t_2)=\Big(\widehat{R}'_{t_1}(X_1,t_1,X_2,t_2)+D^{(1)}\widehat{R}(X_1,t_1,X_2,t_2)+\\
&+\widehat{g}(X_1,t_1,X_2,t_2)\Big)\frac{X_1!\,X_2!}{t_1^{X_1}t_2^{X_2}},
\end{aligned}$$

что с точностью до множителя $X_1!\,X_2!\,t_1^{-X_1}t_2^{-X_2}$ равно члену в левой части (5.2) при $\ell=1$. При $X_2>0$ так же связаны второй член в левой части (4.14) и член в левой части (5.2) при $\ell=2$. Проверка показывает, что при $X_1=0$ и/или $X_2=0$ эти выражения сохраняются. Поэтому из (4.14), (4.15) следуют (5.2), (5.3). Формула (5.5) следует из (5.1) и (3.10). ▲

Для некоторого $\lambda>0$ положим

$$\begin{aligned}
&\lambda_1=\lambda+(m+w)T^{-1/2},\quad \lambda_2=\lambda+(m-w)T^{-1/2},\\
&X_\ell=\lambda t_\ell+x_\ell T^{1/2},\quad \tau_\ell=t_\ell/T,\quad \ell=1,2.
\end{aligned}$$

В качестве множества параметров выберем

$$\Theta=\big\{\theta=(\lambda+(m+w)T^{-1/2},\lambda+(m-w)T^{-1/2}):\ |w|\leqslant c,\ |m|\leqslant a_T\big\},$$

где $c>0$ – достаточно большая фиксированная константа, $a_T=T^\alpha$, $0<\alpha<1/2$ и T достаточно велико. В дальнейшем удобно изменить параметризацию и от $\theta=(\lambda_1,\lambda_2)$ перейти к $\theta'=(m,w)$. Априорную плотность распределения выберем в виде $T\varkappa_a(m)\varrho(w)$, где $\varkappa_a(m)=(2a_T)^{-1}$ – плотность равномерного распределения на отрезке $|m|\leqslant a_T$, а $\varrho(w)$ – произвольная плотность на отрезке $|w|\leqslant c$.

Далее положим

$$D_1 = D_2 = \lambda, \quad t_\ell^* = t_\ell D_\ell, \quad t'_\ell = t_\ell / D_\ell, \quad \tau_\ell^* = \tau_\ell D_\ell, \quad \tau'_\ell = \tau_\ell / D_\ell, \quad \tau' = \tau'_1 + \tau'_2,$$
$$\varepsilon_0 = t_0/T, \quad \varepsilon = T^{-1}, \quad \delta = T^{-1/2}, \quad \widehat{R}(X_1, t_1, X_2, t_2) = T^{-1/2}\widehat{r}(x_1, \tau_1, x_2, \tau_2).$$

Отметим, что в этом параграфе ε и δ не характеризуют длины полуинтервалов. Будем писать

$$x_T \sim y_T, \quad \text{если } \lim_{T\to\infty} \frac{x_T}{y_T} = 1,$$
$$x_T \lesssim y_T, \quad \text{если } \lim_{T\to\infty} \frac{x_T}{y_T} \leqslant 1.$$

Если t_0 достаточно велико, то при $t_1 \geqslant t_0$, $t_2 \geqslant t_0$ справедливы оценки

$$\begin{aligned} &p(X_1, t_1; \lambda_1) \sim T^{-1/2} f_{\tau_1^*}(x_1 \,|\, (m+w)\tau_1), \\ &p(X_2, t_2; \lambda_2) \sim T^{-1/2} f_{\tau_2^*}(x_2 \,|\, (m-w)\tau_2). \end{aligned} \tag{5.6}$$

Действительно, в силу центральной предельной теоремы для плотностей имеем

$$p(X_\ell, t_\ell; \lambda_\ell) \sim f_{t_\ell^*}(X_\ell \,|\, \lambda_\ell t_\ell) = f_{t_\ell^*}(X_\ell - \lambda t_\ell \,|\, \lambda_\ell t_\ell - \lambda t_\ell).$$

Отсюда с учетом сделанной замены переменных следует (5.6).

Выберем $b_T > 0$ из условий $b_T \to +\infty$, $b_T/a_T \to 0$ при $T \to \infty$. Сделаем замену переменных $y = (\bar{x}_1\tau'_1 + \bar{x}_2\tau'_2)/\tau'$, $z = x_1\tau_2 - x_2\tau_1$, где $\bar{x}_1 = x_1/\tau_1$, $\bar{x}_2 = x_2/\tau_2$. Так же, как в [16], с учетом (5.6) устанавливаются оценки

$$\begin{aligned} &\widehat{g}^{(\ell)}(x_1, t_1, X_2, t_2) \sim T^{-3/2}(2a_T)^{-1}\widehat{g}^{(\ell)}(z, \tau_1, \tau_2) \quad \text{при } |y| \leqslant a_T - b_T, \\ &\widehat{g}^{(\ell)}(x_1, t_1, X_2, t_2) \lesssim T^{-3/2}(2a_T)^{-1}\widehat{g}^{(\ell)}(z, \tau_1, \tau_2) \quad \text{при } a_T - b_T < |y| \leqslant a_T + b_T, \\ &\widehat{g}^{(\ell)}(x_1, t_1, X_2, t_2) = T^{-3/2}(2a_T)^{-1} o(e^{-\gamma(y-a_T)^2}) \quad \text{при } |y| > a_T + b_T, \ \gamma > 0, \end{aligned} \tag{5.7}$$

где

$$\begin{aligned} \widehat{g}^{(1)}(z, \tau_1, \tau_2) &= \int_{-c}^{0} 2|w| f_{\tau_1^*\tau_2^*\tau'}(z - 2w\tau_1\tau_2)\varrho(w)\,dw, \\ \widehat{g}^{(2)}(z, \tau_1, \tau_2) &= \int_{0}^{c} 2|w| f_{\tau_1^*\tau_2^*\tau'}(z - 2w\tau_1\tau_2)\varrho(w)\,dw. \end{aligned} \tag{5.8}$$

Кроме того, так же, как в [16], устанавливаются оценки

$$\begin{aligned} &\widehat{r}(x_1, \tau_1, x_2, \tau_2) = (2a_T)^{-1} O(1) \quad \text{при } |y| \leqslant a_T + b_T, \\ &\widehat{r}(x_1, \tau_1, x_2, \tau_2) = (2a_T)^{-1} o(e^{-\gamma(y-a_T)^2}) \quad \text{при } |y| > a_T + b_T, \ \gamma > 0. \end{aligned} \tag{5.9}$$

Покажем, что для

$$\mathrm{r}(z, \tau_1, \tau_2) = (2a_T)\widehat{r}(x_1, \tau_1, x_2, \tau_2)$$

при $|y| \leqslant a_T - b_T$ и $T \to \infty$ справедливо дифференциальное уравнение в частных производных второго порядка

$$\min_{\ell=1,2}\Big(\mathrm{r}'_{\tau_\ell} + \tau_\ell^{-1}\mathrm{r} + z\tau_\ell^{-1}\mathrm{r}'_z + 0{,}5D_\ell(\tau_{3-\ell})^2\mathrm{r}''_{zz} + \widehat{g}^{(\ell)}(z, \tau_1, \tau_2)\Big) = 0 \tag{5.10}$$

с начальным условием

$$\mathrm{r}(z,\tau_1,\tau_2)=0 \quad \text{при } \tau_1+\tau_2=1. \tag{5.11}$$

При этом байесовский риск равен

$$R_N^B(\lambda)\sim T^{1/2}\biggl(\varepsilon_0\int\limits_{-c}^{c}2|w|\varrho(w)\,dw+\int\limits_{-\infty}^{\infty}\mathrm{r}(z,\varepsilon_0,\varepsilon_0)\,dz\biggr). \tag{5.12}$$

Запишем уравнение (5.2) с использованием переменных x_1,τ_1,x_2,τ_2. Чтобы выразить в новых переменных выражение $D^{(\ell)}\widehat{R}(X_1,t_1,X_2,t_2)$, заметим, что паре (X_ℓ,t_ℓ) соответствует (x_ℓ,τ_ℓ) по определению, а паре $(X_\ell+1,t_\ell)$ соответствует $(x_\ell+\delta,\tau_\ell)$. Действительно,

$$X_\ell+1=\lambda t_\ell+x_\ell T^{1/2}+1=\lambda t_\ell+x'_\ell T^{1/2},$$

откуда $x'_\ell=x_\ell+\delta$. Обозначим $\widetilde{x}_\ell=x_\ell/\tau_\ell$, $\ell=1,2$. Пусть $\ell=1$. Так как

$$(X_1+1)/t_1=(\lambda\tau_1T+x_1T^{1/2}+1)/(\tau_1T)=\lambda+\widetilde{x}_1\delta+\varepsilon/\tau_1,\quad X_1/t_1=\lambda+\widetilde{x}_1\delta,$$

то

$$\begin{aligned}&\widehat{R}(X_1+1,t_1,\cdot\,)(X_1+1)/t_1=\delta\times\widehat{r}(x_1+\delta,\tau_1,\cdot\,)(\lambda+\widetilde{x}_1\delta+\varepsilon/\tau_1),\\&\widehat{R}(X_1,t_1,\cdot\,)(X_1/t_1)=\delta\times\widehat{r}(x_1,\tau_1,\cdot\,)(\lambda+\widetilde{x}_1\delta).\end{aligned} \tag{5.13}$$

В этих обозначениях опущена зависимость $\widehat{R}(X_1,t_1,X_2,t_2)$ от X_2,t_2 и зависимость $\widehat{r}(x_1,\tau_1,x_2,\tau_2)$ от x_2,τ_2, поскольку при $\ell=1$ эти переменные не меняются. Если $r(x_1,\tau_1,\cdot\,)$ – достаточно гладкая функция x_1, то разлагая ее в ряд Тейлора до членов порядка ε и учитывая (5.13), получаем

$$\begin{aligned}&D^{(1)}\widehat{R}(X_1,t_1,X_2,t_2)=\widehat{R}(X_1+1,t_1,\cdot\,)(X_1+1)/t_1-\widehat{R}(X_1,t_1,\cdot\,)(X_1/t_1)=\\&=\delta(\widehat{r}+\widehat{r}'_{x_1}\delta+0{,}5\varepsilon\widehat{r}''_{x_1x_1}+O(\varepsilon^{3/2}))(\lambda+\widetilde{x}_1\delta+\varepsilon/\tau_1)-\delta\times\widehat{r}\times(\lambda+\widetilde{x}_1\delta)=\\&=\delta\Bigl(\widehat{r}\varepsilon/\tau_1+\widehat{r}'_{x_1}(\lambda\delta+\widetilde{x}_1\varepsilon)+0{,}5\varepsilon\lambda\widehat{r}''_{x_1x_1}+O(\varepsilon^{3/2})\Bigr),\end{aligned} \tag{5.14}$$

где $\widehat{r}=\widehat{r}(x_1,\tau_1,x_2,\tau_2)$. Так как $x_1=T^{-1/2}(X_1-\lambda t_1)$, $\tau_1=t_1/T$, то

$$\begin{aligned}&\widehat{R}'_{t_1}(X_1,t_1,X_2,t_2)=T^{-1/2}\widehat{r}'_{t_1}(T^{-1/2}(X_1-\lambda t_1),t_1/T,x_2,\tau_2)=\\&=T^{-1/2}\Bigl(\widehat{r}'_{x_1}(x_1,\tau_1,x_2,\tau_2)(-\lambda T^{-1/2})+\widehat{r}'_{\tau_1}(x_1,\tau_1,x_2,\tau_2)T^{-1}\Bigr)=-\lambda\varepsilon\widehat{r}'_{x_1}+\delta\varepsilon\widehat{r}'_{\tau_1}.\end{aligned}$$

Поэтому с учетом (5.14), (5.7) левая часть (5.2) при $\ell=1$ принимает вид

$$\begin{aligned}&\widehat{R}'_{t_1}(X_1,t_1,X_2,t_2)+D^{(1)}\widehat{R}(X_1,t_1,X_2,t_2)+g^{(1)}(X_1,t_1,X_2,t_2)=\\&=-\lambda\varepsilon r'_{x_1}+\delta\varepsilon r'_{\tau_1}+\delta\Bigl(r\varepsilon/\tau_1+r'_{x_1}(\lambda\delta+\widetilde{x}_1\varepsilon)+0{,}5\lambda r''_{x_1x_1}\varepsilon+O(\varepsilon^{3/2})\Bigr)+\\&+\varepsilon\delta(2a_T)^{-1}\widehat{g}^{(1)}(z,\tau_1,\tau_2)=\\&=\varepsilon\delta\Bigl(r'_{\tau_1}+r/\tau_1+r'_{x_1}\widetilde{x}_1+0{,}5\lambda r''_{x_1x_1}+(2a_T)^{-1}\widehat{g}^{(1)}(z,\tau_1,\tau_2)+O(\varepsilon^{1/2})\Bigr).\end{aligned} \tag{5.15}$$

Положим

$$\widehat{r}(x_1,\tau_1,x_2,\tau_2)=(2a_T)^{-1}\mathrm{r}(z,\tau_1,\tau_2),\quad \text{где}\quad z=x_1\tau_2-x_2\tau_1.$$

Таблица 1

Байесовский риск в зависимости от ε и $\Delta\lambda$

t_0	$\varepsilon \diagdown \Delta\lambda$	0,08	0,16	0,20	0,24	0,28	0,32	0,36	0,40
0	1	0,3478	0,5310	0,5731	0,5904	0,5891	0,5748	0,5519	0,5246
0	0,5	0,3469	0,5276	0,5678	0,5830	0,5794	0,5627	0,5377	0,5082
0	0,25	0,3467	0,5268	0,5666	0,5813	0,5772	0,5599	0,5343	0,5042
0	0,125	0,3467	0,5266	0,5663	0,5809	0,5766	0,5592	0,5334	0,5033
1	0,125	0,3472	0,5297	0,5716	0,5890	0,5881	0,5748	0,5540	0,5295

Тогда

$$\widehat{r}'_{\tau_1} = (2a_T)^{-1}(-x_2\mathrm{r}'_z + \mathrm{r}'_{\tau_1}), \quad \widehat{r}'_{x_1} = (2a_T)^{-1}\mathrm{r}'_z\tau_2, \quad \widehat{r}''_{x_1x_1} = (2a_T)^{-1}\mathrm{r}''_{zz}\tau_2^2.$$

Так как $a_T = T^\alpha$, $0 < \alpha < 1/2$, то (5.15) принимает вид

$$\varepsilon\delta(2a_T)^{-1}\Big(\mathrm{r}'_{\tau_1} + \mathrm{r}/\tau_1 + (z/\tau_1)\mathrm{r}'_z + 0{,}5\lambda\tau_2^2\mathrm{r}''_{zz} + \widehat{g}^{(1)}(z,\tau_1,\tau_2) + O(\varepsilon^{1/2-\alpha})\Big). \tag{5.16}$$

Аналогично,

$$\widehat{R}'_{t_2}(X_1,t_1,X_2,t_2) + D^{(2)}\widehat{R}(X_1,t_1,X_2,t_2) + g^{(2)}(X_1,t_1,X_2,t_2) = \varepsilon\delta(2a_T)^{-1} \times$$
$$\times \Big(\mathrm{r}'_{\tau_2} + \mathrm{r}/\tau_2 + (z/\tau_2)\mathrm{r}'_z + 0{,}5\lambda\tau_1^2\mathrm{r}''_{zz} + \widehat{g}^{(2)}(z,\tau_1,\tau_2) + O(\varepsilon^{1/2-\alpha})\Big). \tag{5.17}$$

Подставляя (5.16), (5.17) в (5.2), учитывая, что $\lambda = D_1 = D_2$, и переходя к пределу при $T \to \infty$, что соответствует $\varepsilon \to 0$, получаем уравнение (5.10). Начальные условия (5.11) следуют из (5.3). Получение формулы (5.12) из (5.5) выполняется с учетом (5.9) так же, как формулы (6.18) в [16]. При этом в [16] показано, что (5.10)–(5.12) описывают байесовский риск, вычисленный для гауссовского двурукого бандита относительно наихудшего априорного распределения (см. [16, теорема 8]).

§ 6. Численные эксперименты

О качестве управления в зависимости от значения ε можно судить по соответствующей величине байесовского риска. Результаты представлены в табл. 1. Априорное распределение выбрано равномерным на множестве из 18 параметров

$$\theta_{1,i} = (\lambda_i + \Delta\lambda, \lambda_i - \Delta\lambda), \quad \theta_{2,i} = (\lambda_i - \Delta\lambda, \lambda_i + \Delta\lambda),$$

где $\lambda_i = 0{,}8 + 0{,}05\,(i-1)$, $i = 1,\ldots,9$. Горизонт управления выбран равным $T = 30$, а возможные значения $\{\Delta\lambda\}$ и $\{\varepsilon\}$ представлены в верхней строке и втором слева столбце таблицы. В первых четырех строках при задании стратегии принято $t_0 = 0$, в последней строке $t_0 = 1$ (для сравнения результатов). Отметим, что значение $\varepsilon = 1$ соответствует 30 полуинтервалам переключения стратегии, что уже обеспечивает высокое качество управления.

О поведении стратегии на конкретной траектории управления можно судить по рис. 1. В этом случае байесовская стратегия была вычислена для указанного выше априорного распределения при $\Delta\lambda = 0{,}1$. Кроме того, выбраны $T = 120$, $t_0 = \varepsilon = 2$. Затем моделировалось управление с использованием найденной стратегии при $\lambda_1 = 1 + \Delta\lambda$, $\lambda_2 = 1 - \Delta\lambda$. Для текущей статистики (X_1, t_1, X_2, t_2), наблюдаемой в момент времени $t = t_1 + t_2$, величина x определялась из условия $x = X_1 - \widetilde{X}_1$ (или, эквивалентно, $x = \widetilde{X}_2 - X_2$), где $(\widetilde{X}_1, t_1, \widetilde{X}_2, t_2)$ – статистика, при которой происходит переключение оптимального действия с первого на второе (см. теорему 3). При этом при $x > 0$ выбирается первое действие, а при $x < 0$ – второе. На рис. 1 сплошной линией представлена типичная траектория отклонений X_1 от пороговых

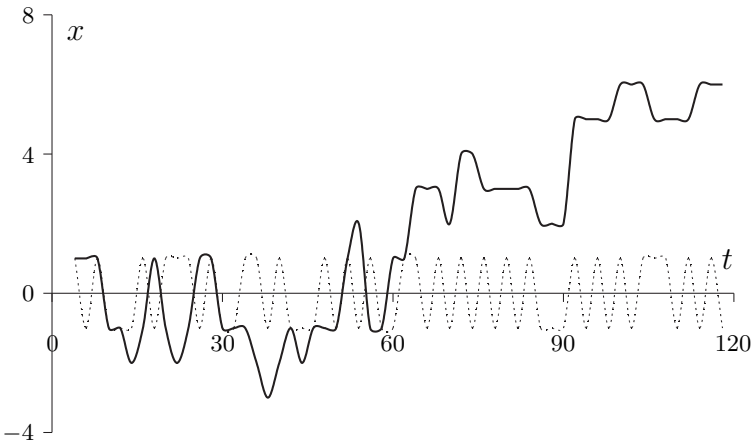


Рис. 1. Динамика отклонений X_1 от пороговых значений

значений $\widetilde{X}_1$. Такое поведение отклонений с учетом аргументации, представленной в [9, раздел 6], позволяет предположить, что в рассматриваемой постановке одновременное применение действий (с распределением ресурса между ними) практически не требуется. Такая необходимость в распределении ресурса возникает, когда отклонения минимальны при выборе сколь угодно малых значений ε, как это было бы в случае траектории, представленной пунктирной линией.

В качестве примера приближенного определения минимаксных стратегии и риска рассмотрим множество параметров

$$\Theta = \big\{(\lambda_1, \lambda_2) :\ 0{,}8 \leqslant (\lambda_1 + \lambda_2)/2 \leqslant 1{,}2;\ |\lambda_1 - \lambda_2|/2 \leqslant 0{,}4\big\}.$$

Приблизительно наихудшее априорное распределение было выбрано симметричным на множестве из шести пар параметров

$$\theta_{1,i} = (\lambda_i + \Delta\lambda_i, \lambda_i - \Delta\lambda_i), \quad \theta_{2,i} = (\lambda_i - \Delta\lambda_i, \lambda_i + \Delta\lambda_i), \quad i = 1, \ldots, 6,$$

т.е. $\Pr(\theta_{1,i}) = \Pr(\theta_{2,i}) = \mu_i$. Значения $\lambda_1 = 0{,}8$, $\lambda_2 = 0{,}85$, $\lambda_3 = 0{,}95$, $\lambda_4 = 1{,}05$, $\lambda_5 = 1{,}15$, $\lambda_6 = 1{,}2$ были фиксированы, а $\{\Delta\lambda_i\}$, $\{\mu_i\}$ требовалось найти в соответствии с условием (1.8).

В рассматриваемом случае байесовский риск становится функцией конечного числа переменных $\{\Delta\lambda_i\}$, $\{\mu_i\}$, поэтому для поиска максимума этой функции был использован градиентный метод. Начальные значения переменных были выбраны из условий $\mu_i = 1/12$, $\Delta\lambda_i = 1{,}6\,(\lambda_i/T)^{1/2}$, $i = 1, \ldots, 6$ (такие $\{\Delta\lambda_i\}$ соответствуют наихудшему распределению при больших T). При $T = 150$, $t_0 = 0$, $\varepsilon = 5$ параметры приблизительно наихудшего априорного распределения оказались следующими: $\Delta\lambda_1 \approx 0{,}138$, $\Delta\lambda_2 \approx 0{,}134$, $\Delta\lambda_3 \approx 0{,}127$, $\Delta\lambda_4 \approx 0{,}122$, $\Delta\lambda_5 \approx 0{,}123$, $\Delta\lambda_6 \approx 0{,}129$, $\mu_1 \approx 0{,}043$, $\mu_2 \approx 0$, $\mu_3 \approx 0{,}140$, $\mu_4 \approx 0{,}036$, $\mu_5 \approx 0$, $\mu_6 \approx 0{,}281$. Соответствующее значение нормированного байесовского риска равно $T^{-1/2} R_T^B(\mu) \approx 0{,}627$. Затем для найденной байесовской стратегии σ^B были вычислены нормированные потери

$$l_i(\Delta\lambda) = T^{-1/2} L_T^B(\sigma^B, (\lambda_i + \Delta\lambda, \lambda_i - \Delta\lambda)),$$

которые представлены на рис. 2. Номера линий $i = 1, 2, 3, 4, 5, 6$ соответствуют индексам $l_i(\Delta\lambda)$, при этом линии *4–6* почти совпадают. Максимальные потери, как и байесовский риск, приблизительно равны 0,627, поэтому найденные стратегию и риск можно считать приблизительно минимаксными. Отметим, что нормированный байесовский риск для начального распределения приблизительно равен 0,613,

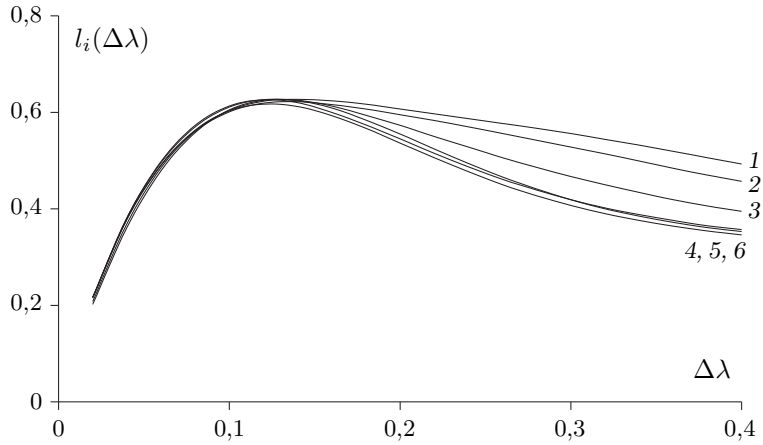


Рис. 2. Потери, обеспечиваемые приблизительно минимаксной стратегией

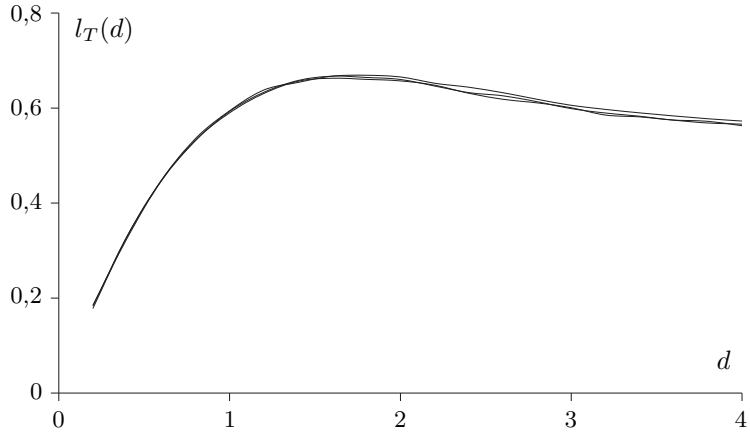


Рис. 3. Потери, обеспечиваемые пакетной обработкой

а максимальные нормированные потери $\{l_i(\Delta\lambda)\}$ приблизительно равны 0,659, т.е. эти параметры уже обеспечивают неплохое приближение минимаксного управления.

Минимаксное управление на больших горизонтах управления T выполнялось с использованием стратегий, полученных для гауссовского двурукого бандита, описывающего пакетную обработку. В [16] показано, что в этом случае максимальные потери достигаются на множестве "близких" распределений, для которых математические ожидания одношаговых доходов различаются на величину порядка $T^{-1/2}$, а нормированный минимаксный риск не меньше величины 0,637, к которой стремится, если количество пакетов неограниченно растет. При моделировании пакетной обработки нормированные потери определялись как

$$l_T(d) = (\lambda T)^{-1/2} L_T(\sigma, \theta_d),$$

где

$$\theta_d = (\lambda + d(\lambda/T)^{1/2}, \lambda - d(\lambda/T)^{1/2}),$$

а σ – минимаксная стратегия для гауссовского двурукого бандита. Интенсивность λ всегда выбиралась равной единице, а T выбиралось равным 600, 3000 и 15000.

Количество пакетов (полуинтервалов переключения стратегии) было равно 30, т.е. t_0 и ε выбирались равными 20, 100 и 500 соответственно. Так как стратегия пакетной обработки требует знания дисперсий одношаговых доходов, то на первых двух этапах, когда действия применяются поровну, делались оценки интенсивности как

$$\widehat{\lambda} = \frac{X_1 + X_2}{2t_0},$$

где X_1, X_2 – начальные доходы за применение первого и второго действий. На оставшихся 28 этапах $\widehat{\lambda}$ использовалась в качестве оценки обеих дисперсий. На рис. 3 представлены полученные результаты моделирования нормированных потерь методом Монте-Карло. Видно, что все кривые близки, хотя кривая, соответствующая $T = 600$, проходит чуть выше. Если оценку интенсивности не делать, а сразу принять $\lambda = 1$, то все три кривые практически совпадают с нижними двумя кривыми на рис. 3. Отметим, что максимум нормированных потерь при этом приблизительно равен 0,667, т.е. превышает минимально возможный менее чем на 5%.

§ 7. Заключение

Рассмотрено использование кусочно-постоянных стратегий в задаче о пуассоновском двуруком бандите в байесовской и минимаксной постановках. Такое управление соответствует пакетной обработке поступающих доходов. Результаты численных экспериментов показывают, что разбиение горизонта управления на 30 равных полуинтервалов, на которых стратегия остается постоянной, обеспечивает высокое качество управления. Для умеренных горизонтов управления минимаксные стратегию и риск можно искать как байесовские, вычисленные относительно наихудшего априорного распределения, которое с высокой степенью точности можно считать сосредоточенным на конечном множестве параметров. Для больших горизонтов можно использовать пакетные стратегии управления, ранее полученные для гауссовского двурукого бандита. При этом оптимальное управление, найденное для пуассоновского двурукого бандита, не позволяет уменьшить минимаксный риск, обеспечиваемый оптимальными пакетными стратегиями, если горизонт управления и количество пакетов неограниченно растут.

Автор выражает глубокую признательность рецензенту за внимание к статье и полезные замечания.

СПИСОК ЛИТЕРАТУРЫ

1. *Berry D.A., Fristedt B.* Bandit Problems: Sequential Allocation of Experiments. London, New York: Chapman & Hall, 1985.
2. *Пресман Э.Л., Сонин И.М.* Последовательное управление по неполным данным. Байесовский подход. М.: Наука, 1982.
3. *Срагович В.Г.* Адаптивное управление. М.: Наука, 1981.
4. *Назин А.В., Позняк А.С.* Адаптивный выбор вариантов: рекуррентные алгоритмы. М.: Наука, 1986.
5. *Цетлин М.Л.* Исследования по теории автоматов и моделированию биологических систем. М.: Наука, 1969.
6. *Варшавский В.И.* Коллективное поведение автоматов. М.: Наука, 1973.
7. *Пресман Э.Л.* Пуассоновский вариант задачи о «двуруком бандите» с дисконтированием // Теория вероятн. и ее примен. 1990. Т. 35. № 2. С. 318–328. `http://mi.mathnet.ru/tvp999`
8. *Chernoff H., Ray S.N.* A Bayes Sequential Sampling Inspection Plan // Ann. Math. Statist. 1965. V. 36. № 5. P. 1387–1407. `https://doi.org/10.1214/aoms/1177699898`

9. *Mandelbaum A.* Continuous Multi-Armed Bandits and Multiparameter Processes // Ann. Probab. 1987. V. 15. № 4. P. 1527–1556. https://doi.org/10.1214/aop/1176991992
10. *Lai T.L.* Adaptive Treatment Allocation and the Multi-Armed Bandit Problem // Ann. Statist. 1987. V. 15. № 3. P. 1091–1114. https://doi.org/10.1214/aos/1176350495
11. *Vogel W.* An Asymptotic Minimax Theorem for the Two Armed Bandit Problem // Ann. Math. Statist. 1960. V. 31. P. 444–451. https://doi.org/10.1214/aoms/1177705907
12. *Боровков А.А.* Математическая статистика. Дополнительные главы. М.: Наука, 1984.
13. *Колногоров А.В.* Нахождение минимаксных стратегии и риска в случайной среде (задаче о двуруком бандите) // АиТ. 2011. № 5. С. 127–138. http://mi.mathnet.ru/at1708
14. *Fabius J., van Zwet W.R.* Some Remarks on the Two-Armed Bandit // Ann. Math. Statist. 1970. V. 41. № 6. P. 1906–1916. https://doi.org/10.1214/aoms/1177696692
15. *Колногоров А.В.* К предельному описанию робастного параллельного управления в случайной среде // АиТ. 2015. № 7. С. 111–126. http://mi.mathnet.ru/at14258
16. *Колногоров А.В.* Гауссовский двурукий бандит: предельное описание // Пробл. передачи информ. 2020. Т. 56. № 3. С. 86–111. https://doi.org/10.31857/S0555292320030055

Колногоров Александр Валерианович
Новгородский государственный университет
им. Ярослава Мудрого, кафедра
прикладной математики и информатики
kolnogorov53@mail.ru

Поступила в редакцию
31.05.2021
После доработки
09.04.2022
Принята к публикации
18.04.2022

ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 58 **2022** **Вып. 2**

УДК 621.391 : 004.056.5 : 519.725

© 2022 г. **В.В. Зяблов, Ф.И. Иванов, Е.А. Крук, В.Р. Сидоренко**[1]

О НОВЫХ ЗАДАЧАХ В АСИММЕТРИЧНОЙ КРИПТОГРАФИИ, ОСНОВАННОЙ НА ПОМЕХОУСТОЙЧИВОМ КОДИРОВАНИИ[2]

Рассматривается задача построения криптосистем с открытым ключом на основе помехоустойчивых кодов. Данный класс криптосистем на сегодняшний день является устойчивым к атакам с использованием квантового компьютера и потому может быть отнесен к методам постквантовой криптографии. Основным недостатком кодовой криптографии является очень большая длина открытого ключа. Большинство усилий по преодолению этого недостатка сводилось к замене кода Гоппы, который использовался в исходной криптосистеме, на код из другого множества, позволяющего описать открытый ключ более компактно, при этом сохранив стойкость криптосистемы к различным атакам. Здесь предложен другой подход к сокращению длины ключа – мы ставим задачу простого описания множества исправимых кодом ошибок, вес которых превосходит половину его минимального расстояния или которые не могут быть исправлены без знания некоторого скрытого преобразования. Если структура кода позволяет дать такое описание множества ошибок, то сложность большинства атак на зашифрованный текст (например, атака по информационным совокупностям) существенно возрастает.

Ключевые слова: криптографическая система Мак-Элиса, декодирование по информационным совокупностям, обобщенные коды Рида – Соломона, постквантовая криптография.

DOI: 10.31857/S0555292322020077, **EDN:** DZRXPW

§ 1. Введение

Методы теории помехоустойчивого кодирования давно используются в криптографии. С их помощью были построены одни из первых криптосистем с открытым ключом [1] и системы цифровой подписи [2]. Однако в отличие от алгебраических криптосистем, основанных на задачах факторизации [3] и вычисления дискретного логарифма [4], кодовые криптосистемы фактически не применяются на практике. Хотя кодовые криптосистемы выигрывают у алгебраических по времени шифрования/дешифрования [5], их использование в значительной степени ограничивается рядом объективных и субъективных факторов.

Во-первых, алгебраические системы возникли несколько раньше кодовых и сразу прошли через многочисленные испытания их безопасности. Последнее обстоятельство, в условиях отсутствия доказательной стойкости криптосистем с открытым ключом, является определенной гарантией безопасности.

[1] Работа В.Р. Сидоренко выполнена при поддержке европейского исследовательского совета ERC в рамках инновационной программы “Горизонт 2020” (номер гранта 801434).

[2] В статье использованы результаты проекта “Разработка методов достоверной и целостной передачи информации в многопользовательских системах с использованием помехоустойчивых кодов и цифровых водяных знаков”, выполненного в рамках Программы фундаментальных исследований НИУ ВШЭ в 2021 г.

Во-вторых, для первых кодовых криптосистем (система Мак-Элиса) было характерно наличие открытого ключа, существенно более длинного, чем для алгебраических систем (например, системы RSA).

Дальнейшие исследования кодовых криптосистем позволили существенно уменьшить размер открытого ключа [6, 7], а разработка новых методов решения задачи факторизации [8] заставила увеличить длину открытого ключа алгебраических криптосистем настолько, что эти величины открытых ключей стали соизмеримы [9]. Однако алгебраические криптосистемы остаются и сегодня основным инструментом организации криптографической безопасности.

Такое положение начало меняться в последние годы в связи с появлением понятия постквантовой криптографии [10] и развитием ряда новых областей использования криптографических методов.

Активные исследования в области создания так называемого квантового компьютера, моделирующего вычислительные процессы на квантовом уровне, привело к построению алгоритмов, ориентированных на этот компьютер. Одним из главных достижений теории квантовых алгоритмов явилась разработка Шором полиномиального алгоритма решения задачи факторизации на квантовом компьютере [11]. Получение этого алгоритма означает, что после создания достаточно мощного квантового компьютера системы защиты, основанные на алгебраических криптосистемах (а они составляют абсолютное большинство), окажутся скомпрометированными. В связи с этим появилось понятие постквантовой криптографии, т.е. криптографии, стойкость которой не подвергнется сомнению в связи с появлением квантового компьютера. Задача декодирования линейных кодов, лежащая в основе кодовых криптосистем, является *NP*-трудной [12] и, по-видимому, не будет решена за полиномиальное время даже с помощью квантовых компьютеров.

Современная практика сенсорных сетей, облачных вычислений и ряда других направлений инфокоммуникационных технологий выдвигает задачу создания так называемой "легкой криптографии" – криптографических алгоритмов, обеспечивающих достаточный уровень безопасности при использовании устройств с ограниченными вычислительными ресурсами [13]. Для целей легкой криптографии криптосистемы, основанные на теории кодирования, оказываются более перспективными [14], чем алгебраические. Кодовые криптосистемы требуют меньшего числа операций и используют операции линейной алгебры, реализация которых предпочтительна по сравнению с арифметическими операциями.

Все это определило новый всплеск интереса к кодовой криптографии и, возможно, новый прикладной этап в ее развитии.

Как уже было отмечено выше, существует ряд попыток преодолеть главный недостаток кодовых криптосистем – большую длину открытого ключа. Основная идея этих улучшений состоит в замене двоичного кода Гоппы, который используется в исходной криптосистеме Мак-Элиса, на какой-то другой с определенной структурой, позволяющей уменьшить размер открытого ключа. Например, в работе [6] коды Гоппы заменены подкодами квазициклических обобщенных кодов Рида – Соломона. Это позволяет получить криптосистему с размером ключа от 6000 до 11000 бит и уровнем безопасности от 2^{80} до 2^{107}. В [15] предлагается использовать квазициклические коды с умеренной плотностью проверок (QC-MDPC). Это приводит к значительному уменьшению размера ключа до 0,6 КБайт, что делает криптосистему на основе таких кодов практически осуществимой. Очень похожая криптосистема, основанная на квазициклических кодах с малой плотностью проверок (QC-LDPC), была предложена в [16].

Главный недостаток замены кодов Гоппы кодами QC-LDPC или QC-MDPC заключается в том, что практически реализуемые алгоритмы их итеративного декодирования не гарантируют исправления ошибок заданного веса t даже при сравнитель-

но небольших значениях данной величины. Более того, практически используемые коды QC-LDPC или QC-MDPC обычно имеют сравнительно небольшое минимальное расстояние (порядка десятков для кодов скорости $R = 1/2$ и длины в несколько тысяч).

В данной статье мы предлагаем иной подход к выбору кода, который будет использован в качестве компоненты кодовой криптосистемы: вместо задачи компактного описания открытого ключа за счет структурности кода мы поставим задачу выбора тройки $(C_0, \mathcal{E}, \varphi)$, где C_0 – секретный (n, k, d)-код, $\mathcal{E}$ – множество ошибок, вносимых на этапе шифрования, а φ – преобразование полиномиальной сложности, отображающее множество вносимых ошибок в множество ошибок, исправимых кодом. Таким образом, ставится задача описания множества ошибок (не обязательно малого веса), исправимых кодом C_0. Структура данного кода должна быть спрятана от криптоаналитика. Сокращение длины открытого ключа в этом случае будет достигаться за счет того, что классические атаки на зашифрованный текст (например, атака по информационным совокупностям [17]) столкнутся со значительно более сложной, нежели задача исправления ошибок веса до $\dfrac{d-1}{2}$, задачей исправления ошибок веса, большего чем $\dfrac{d-1}{2}$. Это, в свою очередь, позволит перейти к значительно более коротким кодам с сохранением при этом требуемой сложности атаки.

§ 2. Криптосистема Мак-Элиса

Первой кодовой криптосистемой была система Мак-Элиса, предложенная в 1978 году в [1]. В качестве открытого ключа в ней использовалась двоичная матрица $\boldsymbol{G}$ размера $k \times n$, представляющая собой произведение матриц

$$\boldsymbol{G} = \boldsymbol{S}\boldsymbol{G}_0\boldsymbol{P}, \tag{1}$$

где $\boldsymbol{S}$ – невырожденная двоичная $(k \times k)$-матрица, $\boldsymbol{G}_0$ – порождающая матрица двоичного (n, k, d)-кода C_0, для которого известен "простой" (как правило, полиномиальный) алгоритм ξ декодирования ошибок кратности до половины кодового расстояния $t = \dfrac{d-1}{2}$, а $\boldsymbol{P}$ – перестановочная $(n \times n)$-матрица. Следует отметить, что матрица $\boldsymbol{S}\boldsymbol{G}_0$ порождает то же множество кодовых слов, что и $\boldsymbol{G}_0$, т.е. код C_0.

Шифрограмма для сообщения $\boldsymbol{x}$ в системе Мак-Элиса вычисляется следующим образом:

1. Генерируется случайный вектор $\boldsymbol{e}$ длины n из множества $\mathcal{E}_t$ векторов веса t;
2. Вычисляется шифрограмма

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{G} + \boldsymbol{e}. \tag{2}$$

Предполагается, что легальный получатель сообщения знает матрицы $\boldsymbol{S}$, $\boldsymbol{G}$, $\boldsymbol{P}$ в произведении (1), которые являются закрытым ключом криптосистемы. В этом случае легальный пользователь находит зашифрованное сообщение по следующему алгоритму:

1. Умножает $\boldsymbol{y}$ на $\boldsymbol{P}^{-1}$:

$$\boldsymbol{y}\boldsymbol{P}^{-1} = \boldsymbol{x}\boldsymbol{S}\boldsymbol{G}_0 + \boldsymbol{e}\boldsymbol{P}^{-1};$$

2. Декодирует полученный вектор кодом C_0 с порождающей матрицей $\boldsymbol{S}\boldsymbol{G}_0$. Поскольку вектор $\boldsymbol{e}\boldsymbol{P}^{-1}$ имеет вес t, то в результате декодирования будет получен вектор $\boldsymbol{x}\boldsymbol{S}$;
3. Получает $\boldsymbol{x}$ как $\boldsymbol{x} = \boldsymbol{x}\boldsymbol{S}\boldsymbol{S}^{-1}$.

Идея системы Мак-Элиса состоит в том, что после применения обратного преобразования $\boldsymbol{P}^{-1}$ к шифрограмме $\boldsymbol{y}$ вектор ошибки $\boldsymbol{e}\boldsymbol{P}^{-1}$ не меняет своего веса,

т.е. принадлежит тому же множеству векторов, которому принадлежал исходный вектор ошибки $\boldsymbol{e}$, используемый при шифровании. Поэтому для дешифрования легальному пользователю не надо было знать конкретный вектор ошибки – любой вектор ошибки веса до t декодировался в коде C_0 одним и тем же алгоритмом.

Открытым ключом системы Мак-Элиса является матрица $\boldsymbol{G}$. Действительно, стойкость описанной системы основана на сложности декодирования линейного кода C произвольной структуры. После преобразования матрицы, описываемого формулой (1), алгебраическая структура матрицы кода с простым декодированием ξ маскируется. Умножение справа на $\boldsymbol{P}$ переводит исходный код в эквивалентный код C, и простое декодирование ξ к нему применить нельзя.

§ 3. Атаки на кодовую криптосистему

В этом параграфе мы рассмотрим классификацию атак на кодовую криптосистему (безотносительно выбора конкретного кода, лежащего в основе криптосистемы).

Выделяют два основных типа атак на кодовую криптосистему: атака декодирования и структурная. Их основное отличие заключается в том, что атака декодирования направлена на извлечение зашифрованного сообщения $\boldsymbol{x}$ из шифротекста $\boldsymbol{y} = \boldsymbol{x}\boldsymbol{G} + \boldsymbol{e}$ путем декодирования $\boldsymbol{y}$ некоторым специально выбранным алгоритмом. Структурная атака направлена на восстановление секретного ключа $(\boldsymbol{S}, \boldsymbol{G}_0, \boldsymbol{P})$ из открытого ключа $\boldsymbol{G} = \boldsymbol{S}\boldsymbol{G}_0\boldsymbol{P}$ на основе некоторой доступной информации о структуре кода C_0 с порождающей матрицей $\boldsymbol{G}_0$. Отметим, что не всегда требуется найти первоначальное разложение $\boldsymbol{G} = \boldsymbol{S}\boldsymbol{G}_0\boldsymbol{P}$. Зачастую достаточным оказывается найти некоторое разложение $(\boldsymbol{S}', \boldsymbol{G}_0', \boldsymbol{P}')$, такое что $\boldsymbol{G} = \boldsymbol{S}'\boldsymbol{G}_0'\boldsymbol{P}'$, как это сделано, например, в атаке Сидельникова – Шестакова на криптосистему на основе обобщенного кода Рида – Соломона [18]. В данной статье мы не будем подробно останавливаться на анализе структурных атак применительно к предложенной нами криптосистеме, более детально фокусируясь на атаках по декодированию. Более того, нельзя гарантировать, что для заданной криптосистемы не найдется структурной атаки. Известно, что большинство успешных (полиномиальных по сложности атак) на различные варианты кодовых криптосистем относятся именно к классу структурных атак.

Атаки на основе декодирования подразделяются на следующие:

1. Атака на основе перебора информационных векторов – осуществляется путем перебора всех возможных векторов $\boldsymbol{x}$ до тех пор, пока не будет получено значение $\mathrm{wt}(\boldsymbol{x}\boldsymbol{G} - \boldsymbol{y}) = t$. Число попыток, необходимое для реализации данной атаки для (n, k)-кода C, оценивается сверху как $2^{\min(k, n-k)}$;
2. Атака на основе перебора векторов ошибки – осуществляется путем перебора всех возможных векторов $\boldsymbol{e}$ до тех пор, пока не будет получено $\mathrm{wt}\big((\boldsymbol{y} - \boldsymbol{e})\boldsymbol{H}^{\mathrm{T}}\big) = 0$, где $\boldsymbol{H}$ – проверочная матрица, соответствующая публичной порождающей матрице $\boldsymbol{G}$. Сложность данной атаки зависит от мощности множества ошибок, вносимых на этапе шифрования сообщения. Если в основе криптосистемы лежит двоичный (n, k)-код, исправляющий t ошибок, которые случайным образом вносятся на этапе шифрования, то среднее число попыток при данной атаке оценивается сверху величиной $\binom{n}{t}$;
3. Атака на основе поиска свободных от ошибок информационных совокупностей. Подробное описание данной атаки приведено в следующем параграфе.

Следует отметить, что цель любой атаки – это поиск простого декодирования в классе эквивалентных кодов, если известно, что по крайней мере для одного из кодов существует простое декодирование ξ. Разница двух подходов – структурного и на основе декодирования – состоит в том, что в случае атак на основе декодирования применяются общие методы декодирования линейных кодов с произвольной

структурой, а успешное применение структурной атаки позволяет использовать для декодирования полиномиальный декодер ξ.

3.1. Декодирование по информационным совокупностям. Задача декодирования по минимуму расстояния кода с произвольной структурой, как уже было отмечено во введении, является *NP*-трудной [12]. Декодирование ошибок кратности до t (до половины кодового расстояния), конечно, является относительно более простой, но тоже, по-видимому, экспоненциальной по сложности задачей, хотя доказательства ее *NP*-трудности на данный момент не представлено. Во всяком случае, полиномиальных алгоритмов решения этой задачи на данный момент не известно.

Далее мы дадим описание алгоритма декодирования по информационным совокупностям (ISD), на основе которого реализованы наиболее "перспективные" атаки на криптосистему Мак-Элиса.

Цель алгоритмов ISD – восстановить сообщение $\boldsymbol{x}$ из заданного вектора $\boldsymbol{y} = \boldsymbol{x}\boldsymbol{G}+\boldsymbol{e}$, где $\boldsymbol{G}$ – порождающая матрица (n,k)-кода C с минимальным расстоянием $d = 2t+1$ и $\mathrm{wt}(\boldsymbol{e}) \leqslant t$.

Пусть $\mathcal{I}$ является k-подмножеством набора координат $[n] := \{1,2,\ldots,n\}$, такое что $\mathcal{I}$ является информационной совокупностью кода C, а $\boldsymbol{G}_{\mathcal{I}}$ – подматрица $\boldsymbol{G}$, состоящая из столбцов с индексами из $\mathcal{I}$. Аналогично, пусть $\boldsymbol{e}_{\mathcal{I}}$ – вектор, состоящий из координат вектора $\boldsymbol{e}$ с индексами из $\mathcal{I}$.

Алгоритм декодирования ISD работает следующим образом:

1. Выбирается случайная информационная совокупность $\mathcal{I} \subset \{1,2,\ldots,n\}$;
2. Если $\mathrm{wt}(\boldsymbol{y}-\boldsymbol{y}_{\mathcal{I}}\boldsymbol{G}_{\mathcal{I}}^{-1}\boldsymbol{G}) \leqslant t$, то $\boldsymbol{y}_{\mathcal{I}}$ не содержит ошибок, а значит, $\mathrm{wt}(\boldsymbol{e}_{\mathcal{I}}) = 0$. Тогда $\boldsymbol{u} = \boldsymbol{y}_{\mathcal{I}}\boldsymbol{G}_{\mathcal{I}}^{-1}$. Иначе возврат к шагу 1.

Легко заметить, что вероятность P_k того, что заданная информационная совокупность не содержит ошибок, оценивается снизу как

$$P_k \leqslant \frac{\binom{n-t}{k}}{\binom{n}{k}} = \frac{\binom{n-k}{t}}{\binom{n}{t}}. \tag{3}$$

Это значит, что среднее число попыток поиска свободной от ошибок информационной совокупности не превышает $\dfrac{\binom{n}{t}}{\binom{n-k}{t}}$, что значительно меньше, нежели перебор по всем векторам ошибок. Поэтому для достижения требуемой стойкости криптосистемы Мак-Элиса приходится выбирать коды большой длины n. В частности, сам Мак-Элис предлагал использовать $(1024, 524, 101)$-код Гоппы, исправляющий $t = 50$ ошибок, для которого длина открытого ключа равна 536576 бит.

ISD-атака упоминалась еще в [1] и получила дальнейшее развитие в многочисленных публикациях (см., например, работу [19] и библиографию в ней). Существуют разные интерпретации и модификации исходного алгоритма ISD. Было предложено несколько различных улучшений, например, основанных на обобщенном парадоксе дней рождения. В работе [20] показано, что асимптотическая сложность декодирования по информационным совокупностям не превосходит $\widetilde{\mathcal{O}}(2^{0{,}0494n})$, что на данный момент является наилучшей известной оценкой.

3.2. Сложность атак на основе декодирования для криптосистемы Мак-Элиса. Приведем пример вычисления сложности атак. (*Здесь и далее под сложностью атаки мы будем понимать среднее число элементарных операций, необходимых для поиска сообщения $\boldsymbol{x}$ по зашифрованному сообщению $\boldsymbol{y}$*). Для классической крипто-

системы Мак-Элиса на основе $(1024, 524, 101)$-кода Гоппы, исправляющего $t = 50$ ошибок:

- Сложность атаки на основе перебора информационных векторов $2^{n-k}(n-k)k = 2^{518}$;
- Сложность атаки на основе перебора векторов ошибки $k(n-k)\binom{n}{t} = 524 \cdot 500 \cdot \binom{1024}{50} \approx 2^{302}$;
- Сложность атаки на основе поиска свободной от ошибки информационной совокупности оценивается сверху как

$$\frac{\binom{n}{t}}{\binom{n-k}{t}} k(n-k) = \frac{\binom{1024}{50}}{\binom{500}{50}} 524 \cdot 500 \approx 2^{72}.$$

Как уже было отмечено выше, сложность последней атаки оказалась наименьшей. Таким образом, стойкость (*под стойкостью криптосистемы мы будем понимать наименьшую сложность среди известных атак*) классической криптосистемы Мак-Элиса можно оценить минимумом сложности среди рассмотренных атак, который равен 2^{72}.

Далее будут обсуждаться способы увеличения сложности этой атаки путем модификации множества вносимых ошибок.

§ 4. Задача построения множества исправимых кодом ошибок

Все описанные в ранее указанных работах алгоритмы декодирования так или иначе опираются на то, что они исправляют "легкие" ошибки, вес которых значительно меньше длины кодового слова. Однако любой линейный код способен исправлять значительное количество ошибок большого веса. Пусть линейный код C_0 длины n задан над полем $\mathbb{F}$. Пространство $\mathbb{F}^n$ разобьем на смежные классы по C_0. Ясно, что код C_0 может исправлять только один вектор ошибки из каждого смежного класса, но это может быть любой вектор из этого класса. Это означает, что в формуле (2) можно использовать в качестве случайного маскирующего вектора (вектора ошибки $\boldsymbol{e}$) любые (не обязательно "легкие") векторы с условием, что они принадлежат разным смежным классам кода. При этом существенно возрастает сложность алгоритма декодирования по информационным совокупностям (она растет экспоненциально вместе с t), и для обеспечения требуемой стойкости системы можно использовать код существенно меньших размеров.

Однако здесь возникает другая задача, которая легко решалась (и даже не рассматривалась как задача) в исходной системе Мак-Элиса. Как генерировать множество вносимых ошибок, которые легальный пользователь будет декодировать с помощью кода C_0 с порождающей матрицей $\boldsymbol{G}_0$? Фактически мы должны задать множество ошибок $\mathcal{E}$, из которого будет случайным образом выбираться вектор ошибок, используемый в шифровании (2).

Вначале сформулируем свойства, которыми должно обладать это множество. Обозначим через $\mathcal{E}_0$ множество векторов ошибок $\boldsymbol{e}$, исправляемых кодом C_0 с помощью полиномиального алгоритма декодирования ξ, а через $\varphi(C_0)$ – некоторое обратимое преобразование кода (его порождающей матрицы, т.е. базиса). В результате этого преобразования получается код $C = \varphi(C_0)$. Пусть, кроме того, ω – требуемая стойкость криптосистемы.

Тогда мы можем сформулировать требования, предъявляемые к множеству $\mathcal{E}$:

(а) Существует алгоритм V генерации случайного вектора $\boldsymbol{e} \in \mathcal{E}$, имеющий полиномиальную сложность;

(b) Существует обратимое линейное преобразование $\varphi(C_0)$ полиномиальной сложности, отображающее код C_0 в код C;

(c) $\varphi^{-1}(\mathcal{E}) \subset \mathcal{E}_0$, причем $|\varphi^{-1}(\mathcal{E})| \geqslant \omega$;

(d) Для кода C_0 существует алгоритм ξ исправления ошибок из множества $\mathcal{E}$, имеющий полиномиальную сложность;

(e) Сложность декодирования ошибок из множества $\mathcal{E}$ в коде C не меньше чем ω, в частности, $|\mathcal{E}| \geqslant \omega$ – т.е. мощность множества должна препятствовать перебору по всем его элементам с целью вскрытия криптосистемы.

С учетом введенных обозначений предлагаемая обобщенная схема шифрования $\mathbb{S}$ может быть описана следующим образом:

- Публичный ключ в предлагаемой системе: порождающая матрица $\boldsymbol{G} = \varphi(\boldsymbol{G}_0)$ и алгоритм V;
- Секретный ключ: обратное преобразование φ^{-1}, матрица $\boldsymbol{G}_0$ и декодер ξ;
- Алгоритм шифрования:
  1. С помощью алгоритма V выбирается случайный вектор $\boldsymbol{e} \in \mathcal{E}$;
  2. По сообщению $\boldsymbol{x}$ вычисляется шифрограмма

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{G} + \boldsymbol{e}; \tag{4}$$

- Алгоритм дешифрования:
  1. Вычисляется $\boldsymbol{y}' = \varphi^{-1}(\boldsymbol{y}) = \boldsymbol{x}\varphi^{-1}(\boldsymbol{G}) + \varphi^{-1}(\boldsymbol{e}) = \boldsymbol{x}\boldsymbol{G}_0 + \boldsymbol{e}'$, где $\boldsymbol{e}' \in \mathcal{E}_0$;
  2. С помощью алгоритма ξ находится $\boldsymbol{x}$.

Хотя визуально задача декодирования (4) полностью совпадает с задачей (1), она должна быть сложнее за счет того, что исправление ошибок из $\mathcal{E}$ в коде C с произвольной структурой является более сложной задачей, чем исправление ошибок малой кратности. Этот факт следует непосредственно из того, что при заданных n, k вероятность P_k в (3) является монотонно убывающей функцией аргумента t, т.е. из $t < t' \leqslant \dfrac{n}{2}$ следует, что $P_k(t) \gg P_k(t')$.

Оценка эффективности кодовых криптосистем. Стойкость криптосистемы, заданной алгоритмом шифрования (4) с учетом условий (a)–(e), накладываемых на множество $\mathcal{E}$, определяется при прямой атаке (т.е. атаке, основанной на декодировании ошибок из $\mathcal{E}$ в коде C) величиной ω. Если ошибки из $\mathcal{E}$ не являются лидерами смежных классов (самыми "легкими" в смежном классе), то их декодирование в коде C (без знания обратного преобразования φ^{-1}) возможно только перебором – либо по словам кода C, либо по множеству ошибок из $\mathcal{E}$. Таким образом,

$$\omega = \min\{2^k, 2^{n-k}, |\mathcal{E}|\}.$$

По построению $|\mathcal{E}| \leqslant 2^{n-k}$. С учетом того, что при декодировании полным перебором по словам кода C сложность декодирования не зависит от $\mathcal{E}$, естественно называть оптимальной криптосистему (4), для которой $|\mathcal{E}| = 2^{n-k}$, и оценивать то, насколько "полно" используются корректирующие свойства кода, величиной

$$\tau = \frac{\log_2 |\mathcal{E}|}{n-k}. \tag{5}$$

Сформулированный нами критерий "полноты" кодовой криптосистемы не является достаточным и даже наиболее важным с точки зрения практического использования криптосистемы. Он не учитывает размеры открытого ключа криптосистемы, которые обычно обсуждались как главный недостаток кодовых криптосистем, – размеры открытого ключа системы, т.е. используемого кода. Поэтому наряду с параметром при оценке криптосистемы мы будем оценивать и размеры применяемого

кода. Далее, кроме атаки на основе полного перебора существует целый ряд неперебо́рных атак, сложность которых приходится учитывать при оценке криптосистем. Кроме того, малое значение параметра τ свидетельствует о том, что "ресурсы" кода, лежащего в основе криптосистемы, используются не в полной мере, а значит, потенциально возможно ее улучшение за счет расширения множества вносимых ошибок, которые впоследствии будут исправляться декодером. И напротив, близкие к единице значения τ позволяют утверждать, что лежащий в основе криптосистемы код используется достаточно эффективно, а значит, дальнейшее расширение множества $\mathcal{E}_0$ за счет внесения в него "тяжелых" векторов ошибок позволит лишь незначительно уменьшить длину ключа.

Например, для классической криптосистемы Мак-Элиса на основе $(1024, 524)$-кода Гоппы τ_{ME} будет оцениваться как

$$\tau_{\mathrm{ME}} \geqslant \frac{\log_2 \binom{1024}{50}}{500} \approx 0{,}5681,$$

а длина открытого ключа при этом $1024 \cdot 524 = 536576$ бит.

§ 5. Криптосистема, основанная на двоичном образе обобщенного кода Рида – Соломона

5.1. Обобщенные коды Рида – Соломона и их двоичные образы. Здесь и далее будем предполагать, что рассматриваемые коды заданы над полем $\mathbb{F}_q$, $q = 2^m$, $m > 0$.

Задача построения множества $\mathcal{E}$, удовлетворяющего условиям (а)–(е) из § 4, для произвольного кода C является достаточно сложной задачей. Тем не менее, двоичный образ обобщенного кода Рида – Соломона (РС-кода), заданного над полем $\mathbb{F}_q$, имеет полиномиальный алгоритм V построения таких множеств достаточно большой мощности.

Вначале напомним определение обобщенного РС-кода $GRS_{n,k}(\boldsymbol{\alpha}, \boldsymbol{v})$.

Определение 1. Пусть задано конечное поле $\mathbb{F}_q$. Выберем ненулевые элементы $v_1, \ldots, v_n \in \mathbb{F}_q$ и различные элементы $\alpha_1, \ldots, \alpha_n \in \mathbb{F}_q$. Пусть $\boldsymbol{v} = (v_1, \ldots, v_n)$ и $\alpha = (\alpha_1, \ldots, \alpha_n)$. Для любого $0 \leqslant k \leqslant n$ определим обобщенный код Рида – Соломона как

$$GRS_{n,k}(\alpha, \boldsymbol{v}) = \big\{(v_1 f(\alpha_1), v_2 f(\alpha_2), \ldots, v_n f(\alpha_n)) \mid f(x) \in F_k[x]\big\},$$

где под $F_k[x]$ подразумевается множество полиномов $f(x)$ над полем $\mathbb{F}_q$, степени которых не превосходят $k - 1$.

Известно, что наряду с обычным РС-кодом $GRS_{n,k}(\boldsymbol{\alpha}, \boldsymbol{v})$ также является кодом с максимально достижимым расстоянием (МДР-кодом), т.е. имеет минимальное расстояние $d = n - k + 1$. Основная причина, по которой в данной статье рассматривается именно обобщенный РС-код, заключается в том, что для заданных n и k мощность множества различных кодов $GRS_{n,k}(\boldsymbol{\alpha}, \boldsymbol{v})$ существенно превосходит количество различных кодов Рида – Соломона, что препятствует структурной атаке на криптосистему, в основе которой лежат $GRS_{n,k}(\boldsymbol{\alpha}, \boldsymbol{v})$. Порождающую матрицу кода $GRS_{n,k}(\boldsymbol{\alpha}, \boldsymbol{v})$ будем обозначать через $\boldsymbol{G}'$. Данная матрица задана над полем $\mathbb{F}_q$ и имеет размер $k \times n$.

Зафиксируем некоторый базис $\mathbb{F}_q/\mathbb{F}_2$. Рассмотрим двоичное представление кода $GRS_{n,k}(\boldsymbol{\alpha}, \boldsymbol{v})$, т.е. код, слова которого получаются из слов кода $GRS_{n,k}(\boldsymbol{\alpha}, \boldsymbol{v})$ в результате замены символов поля $\mathbb{F}_q$ их двоичным представлением. В результате мы получим двоичный (nm, km)-код с порождающей матрицей $\boldsymbol{G}'_b$ размера $km \times nm$. Обозначим данный код через C_b.

Код C_b в двоичной метрике Хэмминга имеет минимальное расстояние, не меньшее чем у $GRS_{n,k}(\boldsymbol{\alpha}, \boldsymbol{v})$, и при этом способен исправлять любые пакеты ошибок при условии, что данные пакеты ошибок покрывают не более чем t символов принятого слова, если рассматривать их как элементы поля $\mathbb{F}_q$. Если ограничивать только длину пакетов ошибок величинами $1 \leqslant \ell_i \leqslant m$, но при этом снять ограничения на позиции начала и конца пакетов, то число гарантированно исправимых пакетов ошибок будет $\left\lfloor \dfrac{n-k}{4} \right\rfloor = \left\lfloor \dfrac{t}{2} \right\rfloor$. Это следует из того, что никакие $\left\lfloor \dfrac{t}{2} \right\rfloor$ пакетов ошибок длин $1 \leqslant \ell_i \leqslant m$ не исказят более чем t символов поля $\mathbb{F}_q$, а значит, принятый вектор будет исправлен кодом $GRS_{n,k}(\boldsymbol{\alpha}, \boldsymbol{v})$ при обратном преобразовании $\mathbb{F}_2^{mn} \mapsto \mathbb{F}_q^n$.

Прежде чем приступать к описанию криптосистемы, введем понятие синхронного и несинхронного пакетов ошибок.

Определение 2. Пакет ошибок длины $1 \leqslant \ell_i \leqslant m$ будем называть синхронным, если для позиции его начала i найдется такое $r \in \mathbb{N} \cup 0$, что одновременно $i \geqslant mr + 1$ и $i + \ell_i - 1 \leqslant m(r+1)$, т.е. все ненулевые элементы пакета попадают в один и только один подвектор $\boldsymbol{e}_{r+1}$ вектора $\boldsymbol{e} = (\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$. В противном случае пакет ошибок будем называть несинхронным.

5.2. Базовое описание протокола криптосистемы. Теперь представим описание криптосистемы с открытым ключом, основанной на двоичном образе обобщенного кода Рида – Соломона. На самом деле, в данном разделе будет представлено высокоуровневое описание предложенной криптосистемы, а представление ее отдельных компонент будет дано в последующих разделах статьи.

Публичная порождающая матрица криптосистемы имеет вид

$$\boldsymbol{G} = \boldsymbol{S}\boldsymbol{G}_b'\boldsymbol{Q}, \tag{6}$$

где $\boldsymbol{G}_b'$ – секретная двоичная порождающая матрица кода C_b, а $\boldsymbol{S}$ – произвольная невырожденная двоичная $(mk \times mk)$-матрица. Двоичная матрица $\boldsymbol{Q}$ размера $mn \times mn$ выбирается согласно теореме 1 из семейства матриц, описание которых представлено в п. 5.3.

Теперь опишем процедуры генерации ключей, шифрования и дешифрования.

- Генерация секретного и публичного ключей:
  1. Выбирается порождающая матрица $\boldsymbol{G}'$ кода $GRS_{n,k}(\boldsymbol{\alpha}, \boldsymbol{v})$ и строится ее двоичный образ $\boldsymbol{G}_b'$;
  2. Строится случайная невырожденная двоичная матрица $\boldsymbol{S}$ размера $mk \times mk$;
  3. В соответствии с теоремой 1 строится матрица $\boldsymbol{Q}$ размера $mn \times mn$ и выбирается соответствующий класс вносимых векторов ошибок $\mathcal{V}_i$;
  4. Вычисляется публичная порождающая матрица $\boldsymbol{G} = \boldsymbol{S}\boldsymbol{G}_b'\boldsymbol{Q}$;
  5. Публичным ключом криптосистемы является $(\boldsymbol{G}, \mathcal{V}_i)$;
  6. Секретным ключом криптосистемы является набор $(\boldsymbol{Q}, \boldsymbol{G}_b', \boldsymbol{S})$.
- Шифрование открытого текста $\boldsymbol{x} \in \mathbb{F}_2^{km}$ осуществляется следующим образом:
  1. Выбирается случайный вектор $\boldsymbol{e} \in \mathcal{V}_i \subset \mathbb{F}_2^{mn}$, согласованный с матрицей $\boldsymbol{Q}$, так что вектор ошибок $\boldsymbol{e}\boldsymbol{Q}^{-1}$ исправим кодом с порождающей матрицей $\boldsymbol{G}_b'$;
  2. Вычисляется зашифрованное сообщение $\boldsymbol{y} \in \mathbb{F}_2^{mn}$:

     $$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{G} + \boldsymbol{e}.$$

- Дешифрование вектора $\boldsymbol{y} \in \mathbb{F}_2^{mn}$ осуществляется следующим образом:
  1. Производится умножение $\boldsymbol{y}$ на $\boldsymbol{Q}^{-1}$:

     $$\boldsymbol{y}\boldsymbol{Q}^{-1} = \boldsymbol{x}\boldsymbol{S}\boldsymbol{G}_b' + \boldsymbol{e}\boldsymbol{Q}^{-1};$$

2. Вектор $\boldsymbol{yQ}^{-1}$ преобразуется в q-ичный вектор и декодируется декодером кода $GRS_{n,k}(\boldsymbol{\alpha}, \boldsymbol{v})$, исправляющим t ошибок, откуда находится $\boldsymbol{x}' = \boldsymbol{xS} \in \mathbb{F}_q^k$ – вектор длины k над полем $\mathbb{F}_q$;
3. Вектор $\boldsymbol{x}' \in \mathbb{F}_q^k$ отображается в двоичный вектор $\boldsymbol{x}''$;
4. Зашифрованный текст $\boldsymbol{x}$ находится как

$$\boldsymbol{x} = \boldsymbol{x}''\boldsymbol{S}^{-1}.$$

Как уже было отмечено выше, ключевое требование, которым должна удовлетворять пара $(\boldsymbol{Q}, \boldsymbol{e})$, заключается в том, что вектор $\boldsymbol{eQ}^{-1}$ должен быть исправим кодом с порождающей матрицей $\boldsymbol{G}_b'$, т.е. содержать не более чем t синхронных или $\left\lfloor \frac{t}{2} \right\rfloor$ несинхронных пакетов ошибок длины до m. Далее будет показано, каким образом согласуются структуры векторов $\boldsymbol{e}$ и матриц $\boldsymbol{Q}$ так, чтобы $\boldsymbol{eQ}^{-1} \in \mathcal{E}_0$, где $\mathcal{E}_0$ – множество исправимых кодом $GRS_{n,k}(\boldsymbol{\alpha}, \boldsymbol{v})$ ошибок.

5.3. Выбор пары $(\boldsymbol{Q}, \boldsymbol{e})$ в предложенной криптосистеме. Покажем, каким образом нужно выбирать пару $(\boldsymbol{Q}, \boldsymbol{e})$ так, чтобы вектор ошибок $\boldsymbol{eQ}^{-1}$ был исправим двоичным образом обобщенного кода Рида – Соломона (q-ичное представление вектора $\boldsymbol{eQ}^{-1}$ было исправимо кодом $GRS_{n,k}(\boldsymbol{\alpha}, \boldsymbol{v})$) – в этом случае будем говорить, что *вектор $\boldsymbol{e}$ согласован с матрицей $\boldsymbol{Q}$*.

Введем следующее обозначение: символ $\mathcal{A}(\boldsymbol{Q}, \boldsymbol{e})$ означает, что вектор $\boldsymbol{e}$ согласован с матрицей $\boldsymbol{Q}$, т.е. $\boldsymbol{eQ}^{-1}$ содержит не более чем $\left\lfloor \frac{t}{2} \right\rfloor$ несинхронных пакетов ошибок длины до m.

Для простоты введем также следующие обозначения для различных семейств матриц $\boldsymbol{Q}$ и векторов $\boldsymbol{e}$.

Семейства матриц $\boldsymbol{Q}$:

- Будем считать, что матрица $\boldsymbol{Q}$ принадлежит семейству $\mathcal{Q}_1$, если $\boldsymbol{Q} = \operatorname{diag}(\boldsymbol{M})$ – двоичная матрица размера $mn \times mn$, где под $\operatorname{diag}(\boldsymbol{M})$ подразумевается блочно-диагональная $(mn \times mn)$-матрица, на главной диагонали которой стоят невырожденные нижнетреугольные матрицы $\boldsymbol{M}_i$ размеров $m_i \times m_i$, $m+1 \leqslant m_i \leqslant 2m+2$, $\sum m_i = mn$.
- Будем считать, что матрица $\boldsymbol{Q}$ принадлежит семейству $\mathcal{Q}_2$, если $\boldsymbol{Q} = \operatorname{diag}(\boldsymbol{M})$ – двоичная матрица размера $mn \times mn$, где под $\operatorname{diag}(\boldsymbol{M})$ подразумевается блочно-диагональная $(mn \times mn)$-матрица, на главной диагонали которой стоят невырожденные матрицы $\boldsymbol{M}_i$, причем для любых двух соседних матриц $\boldsymbol{M}_{i_1}$ и $\boldsymbol{M}_{i_2}$, стоящих на главной диагонали

  $$\begin{pmatrix} \boldsymbol{M}_{i_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M}_{i_2} \end{pmatrix}$$

  и имеющих размеры $m_{i_1} \times m_{i_1}$ и $m_{i_2} \times m_{i_2}$, выполняется следующее:
  - $m_{i_1} + m_{i_2} = 2m$;
  - В матрице $\boldsymbol{Q}$ матрицы размеров $m_{i_1} \times m_{i_1}$ и $m_{i_2} \times m_{i_2}$ чередуются;
  - Пусть $\boldsymbol{M}_1$ имеет размер $m_1 \times m_1$, а $\boldsymbol{M}_2$ – размер $m_2 \times m_2$, тогда если $m_1 < m_2$, то в каждом блоке из двух подряд идущих матриц

    $$\begin{pmatrix} \boldsymbol{M}_{i_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M}_{i_2} \end{pmatrix}$$

    матрицы большего размера являются верхнетреугольными. Если $m_1 > m_2$, то матрицы большего размера являются нижнетреугольными.

Заметим, что ключевое различие между матрицами $\boldsymbol{Q}$ из семейств $\mathcal{Q}_1$ и $\mathcal{Q}_2$ заключается в том, что на матрицы из $\mathcal{Q}_1$ накладываются ограничения на структуру блоков $\boldsymbol{M}_i$ (они должны быть нижнетреугольными), в то время как выбор размеров m_i каждого из блоков остается достаточно гибким: $m+1 \leqslant m_i \leqslant 2m+2$, $\sum m_i = mn$. На элементы $\mathcal{Q}_2$ накладываются ограничения как на размеры соседних матриц $m_{i_1} + m_{i_2} = 2m$, так и на структуру больших матриц $\boldsymbol{M}_i$. Далее будет показано, что накладываемые ограничения на структуру матриц из семейства $\mathcal{Q}_2$ позволяют на этапе шифрования вносить ошибки большего веса, нежели при шифровании с использованием матриц из семейства $\mathcal{Q}_1$.

Семейства векторов $\boldsymbol{e}$:

- Будем считать, что вектор $\boldsymbol{e}$ принадлежит семейству $\mathcal{V}_1$, если $\boldsymbol{e}$ содержит до $\left\lfloor \dfrac{t}{4} \right\rfloor$ несинхронных пакетов ошибок длины до m;
- Будем считать, что вектор $\boldsymbol{e}$ принадлежит семейству $\mathcal{V}_2$, если $\boldsymbol{e}$ содержит до $\left\lfloor \dfrac{t}{3} \right\rfloor$ несинхронных пакетов ошибок длины до m;
- Будем считать, что вектор $\boldsymbol{e}$ принадлежит семейству $\mathcal{V}_3$, если $\boldsymbol{e}$ содержит до $\left\lfloor \dfrac{t}{2} \right\rfloor$ несинхронных пакетов ошибок длины до m.

Согласование $\boldsymbol{e}$ с $\boldsymbol{Q} \in \boldsymbol{\mathcal{Q}_1}$. Ключевым этапом при проектировании криптосистемы, приведенной в п. 5.2, является выбор матрицы $\boldsymbol{Q}$, являющейся составной частью публичного и секретного ключей, а также вектора ошибки $\boldsymbol{e}$, который вносится на этапе шифрования. Для того чтобы согласовать между собой введенные семейства матриц $\mathcal{Q}_1$ и $\mathcal{Q}_2$ с типами векторов ошибки $\mathcal{V}_1$, $\mathcal{V}_2$ и $\mathcal{V}_3$, докажем ряд лемм.

Лемма 1. *Пусть вектор $\boldsymbol{e}$ представляет собой несинхронный m-пакет. Пусть $\boldsymbol{e}' = \boldsymbol{e}\boldsymbol{Q}^{-1}$, где матрица $\boldsymbol{Q} \in \mathcal{Q}_1$. Тогда $\boldsymbol{e}'$ содержит не более четырех синхронных m-пакетов.*

Доказательство. Рассмотрим наихудший случай. Зададим двоичный вектор $\boldsymbol{e}$ длины mn, такой что этот вектор содержит $mn - m$ нулей, а пакет ошибок имеет координату начала, кратную $m-1$. Пусть для простоты данный пакет состоит из единиц. Тогда, если представить $\boldsymbol{e}$ в виде $\boldsymbol{e} = (\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$, $\boldsymbol{e}_i = (e_{i_1}, \ldots, e_{i_m})$, $e_{i_j} \in \mathbb{F}_2$, то $\boldsymbol{e}$ содержит два последовательных вектора $\boldsymbol{e}_i$, $\boldsymbol{e}_{i+1}$, таких что $\boldsymbol{e}_i = (0, 0, \ldots, 0, 1)$, $\boldsymbol{e}_{i+1} = (1, 1, \ldots, 1, 0)$, причем $\mathrm{wt}(\boldsymbol{e}_{i+1}) = m-1$. Все прочие $\boldsymbol{e}_j$, $j \notin \{i, i+1\}$, являются нулевыми векторами длины m. Пусть в матрице $\boldsymbol{Q}^{-1}$, соответствующей ненулевому участку вектора $\boldsymbol{e}$, находятся две матрицы $\boldsymbol{M}_{i_1}$, $\boldsymbol{M}_{i_2}$ размеров $m_{i_1} \times m_{i_1}$ и $m_{i_2} \times m_{i_2}$ соответственно. Рассмотрим векторы $\boldsymbol{e}'_i = (0, 0, \ldots 0, \boldsymbol{e}_i)$, $\boldsymbol{e}'_{i+1} = (\boldsymbol{e}_{i+1}, 0, \ldots, 0)$ длин m_{i_1} и m_{i_2} соответственно. При вычислении $\boldsymbol{e}' = \boldsymbol{e}\boldsymbol{Q}^{-1}$ участок вектора $\boldsymbol{e}'$, соответствующий произведению $(\boldsymbol{e}'_i, \boldsymbol{e}'_{i+1})$ на $\boldsymbol{Q}^{-1}$, будет иметь вид

$$(\widehat{\boldsymbol{e}}_i, \widehat{\boldsymbol{e}}_{i+1}) = (\boldsymbol{e}'_i \boldsymbol{M}_{i_1}, \boldsymbol{e}'_{i+1} \boldsymbol{M}_{i_2}).$$

Так как вектор $\widehat{\boldsymbol{e}}_i$ содержит единицу на последней позиции, то $\boldsymbol{e}'_i \boldsymbol{M}_{i_1}$ совпадает с последней строкой матрицы $\boldsymbol{M}_{i_1}$, имеющей вес до m_{i_1}. В худшем случае (с точки зрения распространения пакетов ошибок) вектор $\widehat{\boldsymbol{e}}_i$ начинается с 1. Вектор $\widehat{\boldsymbol{e}}_{i+1}$ представляет собой произведение вектора, содержащего первые $m-1$ единицы, а остальные $m_{i_2} - m + 1$ его символов равны 0. Таким образом, $\boldsymbol{e}'_{i+1} \boldsymbol{M}_{i_2}$ имеет вид

$$\widehat{\boldsymbol{e}}_{i+1} = (\widehat{e}_{i+1,1}, \widehat{e}_{i+1,2}, \ldots, \widehat{e}_{i+1,m-1}, 0, \ldots, 0),$$

где $\widehat{e}_{i+1,m-1}$ могут быть ненулевыми. В худшем случае $\widehat{e}_{i+1,m-1} = 1$.

Таким образом, вектор $(\widehat{\boldsymbol{e}}_i, \widehat{\boldsymbol{e}}_{i+1})$ длины $m_{i_1} + m_{i_2}$, где $2m+2 \leqslant m_{i_1} + m_{i_2} \leqslant 4m+4$, содержит пакет ошибок длины до $m_{i_1} + m - 1 \leqslant 3m+1$. Очевидно, что данный пакет ошибок покрывается максимум четырьмя синхронными m-пакетами, а значит, преобразование $\boldsymbol{Q}^{-1}$ не более чем в 4 раза увеличивает вес вектора ошибки

(в q-ичной метрике Хэмминга), который впоследствии должен быть декодирован кодом $GRS_{n,k}(\boldsymbol{\alpha}, \boldsymbol{v})$.

Если участку вектора $\boldsymbol{e}$, на котором расположены $(\boldsymbol{e}_i, \boldsymbol{e}_{i+1})$, в матрице $\boldsymbol{Q}^{-1}$ соответствует единственная матрица $\boldsymbol{M}_i$ размера не более чем $2m+2$, то вектор $(\boldsymbol{e}'_i, \boldsymbol{e}'_{i+1})\boldsymbol{M}_i$ покрывает не более чем четыре символа поля $\mathbb{F}_q$.

По построению матрицы $\boldsymbol{Q}$ никакому пакету ошибок длины не более m на участке $(\boldsymbol{e}_i, \boldsymbol{e}_{i+1})$ длины $2m$ вектора $\boldsymbol{e}$ не может соответствовать более двух блочных подматриц $\boldsymbol{M}_{i_1}$, $\boldsymbol{M}_{i_2}$.

Таким образом, никакой пакет ошибок длины m при преобразовании $\boldsymbol{Q}^{-1}$ не покрывает более чем четыре символа поля $\mathbb{F}_q$. ▲

Таким образом, если $\boldsymbol{Q} \in \mathcal{Q}_1$ и $\boldsymbol{e} \in \mathcal{V}_1$, то выполнено $\mathcal{A}(\boldsymbol{Q}, \boldsymbol{e})$.

Покажем теперь, какие ограничения сверху на m_i в матрицах $\boldsymbol{Q} \in \mathcal{Q}_1$ необходимо накладывать, чтобы произвольный пакет ошибок длины до m при преобразовании $\boldsymbol{Q}^{-1}$ покрывал как можно меньше символов поля $\mathbb{F}_q$, что позволило бы увеличить число вносимых ошибок. Ограничение снизу $m_i \geqslant m+1$ будет сохранено для гарантии того, что никакому пакету ошибок длины не более m на участке $(\boldsymbol{e}_i, \boldsymbol{e}_{i+1})$ длины $2m$ не может соответствовать более двух блочных матриц $\boldsymbol{M}_{i_1}$, $\boldsymbol{M}_{i_2}$ в $\boldsymbol{Q}^{-1}$. Напомним, что в худшем случае умножение пакета ошибок на матрицу $\boldsymbol{Q}^{-1}$ формирует пакет длины $m_i + m - 1 \geqslant 2m$. Ясно, что таким пакетом можно покрыть не более чем три последовательных символа поля $\mathbb{F}_q$. При длине пакета ошибок не более чем $2m+1$, число последовательно покрытых символов поля векторов длины m в $\boldsymbol{e}\boldsymbol{Q}^{-1}$ не превышает трех. Таким образом, если получить ограничение сверху на m_i из

$$m_i + m - 1 \leqslant 2m + 1,$$

т.е. $m_i \leqslant m+2$, то вместо внесения $\left\lfloor \frac{t}{4} \right\rfloor$ пакетов ошибок длины до m в вектор $\boldsymbol{e}$ можно вносить $\left\lfloor \frac{t}{3} \right\rfloor$ пакетов ошибок длины до m. Отменим, что внесение $\left\lfloor \frac{t}{2} \right\rfloor$ ошибок уже не гарантирует декодируемость вектора $\boldsymbol{e}\boldsymbol{Q}^{-1}$ при описанной ранее структуре матрицы $\boldsymbol{Q}$.

Таким образом, справедлива

Лемма 2. *Пусть вектор $\boldsymbol{e}$ представляет собой несинхронный m-пакет. Пусть $\boldsymbol{e}' = \boldsymbol{e}\boldsymbol{Q}^{-1}$, где матрица $\boldsymbol{Q} \in \mathcal{Q}_1$, и при этом для размеров m_i блоков $\boldsymbol{M}_i$ справедливо неравенство $m+1 \leqslant m_i \leqslant m+2$. Тогда $\boldsymbol{e}'$ содержит не более трех синхронных m-пакетов.*

Таким образом, если $m+1 \leqslant m_i \leqslant m+2$, $\boldsymbol{Q} \in \mathcal{Q}_1$ и $\boldsymbol{e} \in \mathcal{V}_2$, то $\mathcal{A}(\boldsymbol{Q}, \boldsymbol{e})$.

В общем случае при внесении не более чем $\left\lfloor \frac{t}{\ell} \right\rfloor$, $\ell \geqslant 2$, пакетов ошибок длины до m для декодируемости вектора $\boldsymbol{e}\boldsymbol{Q}^{-1}$ (т.е. согласованности вектора $\boldsymbol{e}$ с матрицей $\boldsymbol{Q} \in \mathcal{Q}_1$) при ограничении снизу $m_i \geqslant m+1$ следует ограничение сверху

$$m_i \leqslant (\ell - 1)m - m + 2.$$

Согласование $\boldsymbol{e}$ с $\boldsymbol{Q} \in \mathcal{Q}_2$. Ранее было показано, что размеры m_i блоков $\boldsymbol{M}_i$ матрицы $\boldsymbol{Q} \in \mathcal{Q}_1$ существенно влияют на число пакетов ошибок, которые можно вносить при шифровании. Ясно также, что при отсутствии ограничений на индексы начала пакетов ошибок можно исправлять до $\left\lfloor \frac{t}{2} \right\rfloor$ пакетов ошибок длины до m, где t – число ошибок, исправимых кодом $GRS_{n,k}(\boldsymbol{\alpha}, \boldsymbol{v})$. Однако ранее было показано, что при единственном ограничении $m_i \geqslant m+1$, где m_i – размеры квадратных матриц, входящих в состав $\boldsymbol{Q} \in \mathcal{Q}_1$, наибольшее число пакетов ошибок, вносимых на этапе шифрования, не может превышать $\left\lfloor \frac{t}{3} \right\rfloor$. Только в этом случае можно га-

рантировать возможность их исправления кодом $GRS_{n,k}(\boldsymbol{\alpha}, \boldsymbol{v})$ после применения преобразования $\boldsymbol{Q}^{-1}$.

Покажем, что если матрица $\boldsymbol{Q} \in \mathcal{Q}_2$ и при этом для размеров m_{i_1} и m_{i_2} любых двух соседних невырожденных матриц $\boldsymbol{M}_{i_1}$ и $\boldsymbol{M}_{i_2}$ выполняется $m_{i_1} + m_{i_2} = 2m$, то с учетом ограничений на большие матрицы из семейства $\mathcal{Q}_2$ матрица $\boldsymbol{Q}$ согласована с $\boldsymbol{e} \in \mathcal{V}_3$, т.е. на этапе шифрования допустимым было бы внесение максимального числа $\left\lfloor \dfrac{t}{2} \right\rfloor$ несинхронных пакетов ошибок.

Лемма 3. *Если матрица $\boldsymbol{Q} \in \mathcal{Q}_2$, а вектор $\boldsymbol{e}$ представляет собой несинхронный m-пакет, то вектор $\boldsymbol{e}\boldsymbol{Q}^{-1}$ содержит не более двух синхронных m-пакетов.*

Доказательство. Очевидно, что для того чтобы $\boldsymbol{e}\boldsymbol{Q}^{-1}$ содержал не более двух синхронных m-пакетов, необходимо и достаточно, чтобы умножение вектора $(\boldsymbol{e}_i, \boldsymbol{e}_{i+1})$ длины $2m$, содержащего на произвольных m подряд идущих позициях пакет ошибок длины до m, на соответствующий участок длины $2m$ матрицы $\boldsymbol{Q}^{-1}$ не приводило к "размножению" пакетов.

Это, очевидно, достигается в том случае, когда соответствующий участок матрицы $\boldsymbol{Q}^{-1}$ имеет вид

$$\begin{pmatrix} \boldsymbol{M}_{i_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M}_{i_2} \end{pmatrix},$$

где $\boldsymbol{M}_{i_1}$ и $\boldsymbol{M}_{i_2}$ – квадратные матрицы размеров $m_{i_1} \times m_{i_1}$ и $m_{i_2} \times m_{i_2}$, и $m_{i_1} + m_{i_2} = 2m$. Если при этом в $\boldsymbol{Q}$ матрицы размеров $m_{i_1} \times m_{i_1}$ и $m_{i_2} \times m_{i_2}$ чередуются, то никакой пакет ошибок веса m не будет стоять на пересечении более чем двух матриц в $\boldsymbol{Q}$.

Если, кроме того, выполняются дополнительные ограничения на структуру больших матриц $\boldsymbol{M}_i$ (они являются верхнетреугольными, если размер первой блочной матрицы $\boldsymbol{M}_1$ меньше размера $\boldsymbol{M}_2$, и нижнетреугольными, если размер первой блочной матрицы $\boldsymbol{M}_1$ больше размера $\boldsymbol{M}_2$), то каков бы ни был пакет ошибок, лежащий в $(\boldsymbol{e}_i, \boldsymbol{e}_{i+1})$, при умножении на $\boldsymbol{Q}^{-1}$ он не "размножится" на соседние символы, а потому вектор $\boldsymbol{e}' = \boldsymbol{e}\boldsymbol{Q}^{-1}$ будет иметь ту же структуру, что и вектор $\boldsymbol{e}$, который генерируется на этапе шифрования.

Единственное отличие вектора $\boldsymbol{e}'$ от $\boldsymbol{e}$ будет заключаться в том, что длины пакетов ошибок в $\boldsymbol{e}'$ могут достигать $2m$, однако вес Хэмминга вектора $\boldsymbol{e}'$, вычисленный над полем $\mathbb{F}_q$, не будет превышать t, что гарантирует его исправимость кодом $GRS_{n,k}(\boldsymbol{\alpha}, \boldsymbol{v})$. ▲

Согласование $\boldsymbol{e}$ с $\boldsymbol{Q}$ – основной результат. Объединяя леммы 1–3, сформулируем теорему, связывающую между собой семейства $\mathcal{Q}_1$ и $\mathcal{Q}_2$ с семействами $\mathcal{V}_1$, $\mathcal{V}_2$, $\mathcal{V}_3$ векторов $\boldsymbol{e}$ так, чтобы $\boldsymbol{e}\boldsymbol{Q}^{-1}$ содержал не более чем t синхронных пакетов ошибок длины до m, т.е. был исправим кодом $GRS_{n,k}(\boldsymbol{\alpha}, \boldsymbol{v})$.

Теорема 1. *Справедливы следующие утверждения*:

- *Если $\boldsymbol{Q} \in \mathcal{Q}_1$, $\boldsymbol{e} \in \mathcal{V}_1$ и для всех блоков $\boldsymbol{M}_i$ размеров $m_i \times m_i$ выполняется $m + 1 \leqslant m_i \leqslant 2m + 2$, $\sum m_i = mn$, то имеет место $\mathcal{A}(\boldsymbol{Q}, \boldsymbol{e})$;*
- *Если $\boldsymbol{Q} \in \mathcal{Q}_1$, $\boldsymbol{e} \in \mathcal{V}_1 \cup \mathcal{V}_2$ и для всех блоков $\boldsymbol{M}_i$ размеров $m_i \times m_i$ выполняется $m + 1 \leqslant m_i \leqslant m + 2$, $\sum m_i = mn$, то имеет место $\mathcal{A}(\boldsymbol{Q}, \boldsymbol{e})$;*
- *Если $\boldsymbol{Q} \in \mathcal{Q}_2$ и $\boldsymbol{e} \in \mathcal{V}_1 \cup \mathcal{V}_2 \cup \mathcal{V}_3$, то имеет место $\mathcal{A}(\boldsymbol{Q}, \boldsymbol{e})$.*

Таким образом, на этапе проектирования криптографической системы, основанной на двоичном образе обобщенного кода Рида – Соломона, представленной в п. 5.2, разработчик выбирает согласованную в соответствии с теоремой 1 пару $(\boldsymbol{Q}, \boldsymbol{e})$. Выбор пары позволяет регулировать гибкость параметров, определяющих матрицу $\boldsymbol{Q}$, и число вносимых на этапе шифрования ошибок.

Следует особо отметить, что в отличие от классической криптосистемы Мак-Элиса, где публичная порождающая матрица задает линейный код, эквивалентный секретному, в нашем случае это не так: умножение порождающей матрицы $\boldsymbol{G}'_b$ на $\boldsymbol{Q} \in \mathcal{Q}_1 \cup \mathcal{Q}_2$ справа задает преобразование столбцов в $\boldsymbol{G}'_b$, а потому матрица $\boldsymbol{S}\boldsymbol{G}'_b\boldsymbol{Q}$ не является порождающей матрицей двоичного образа кода $GRS_{n,k}(\boldsymbol{\alpha}, \boldsymbol{v})$. Эквивалентность кодов сохранилась бы в том случае, если бы $\boldsymbol{Q}$ являлась блочной перестановкой длины n, а длина блока была равна m, т.е. задавала перестановку символов поля $\mathbb{F}_q$. Исходя из того, что код C_b исправляет максимальное число пакетов ошибок длины до m, с большой вероятностью данное множество пакетов будет не исправимо кодом с порождающей матрицей $\boldsymbol{S}\boldsymbol{G}'_b\boldsymbol{Q}$, что является ключевым фактором, на котором основана предложенная криптосистема.

Далее будут рассмотрены атаки декодирования на предложенный класс криптосистем.

5.4. Анализ некоторых атак. При рассмотрении атак мы будем отталкиваться от того, что на этапе шифрования в вектор $\boldsymbol{e}$ вносится до $\left\lfloor \frac{t}{4} \right\rfloor$ пакетов ошибок длины до m, хотя все полученные результаты легко обобщаются для произвольного числа пакетов $\left\lfloor \frac{t}{\ell} \right\rfloor$, $\ell \geqslant 2$.

Прямые атаки. Напомним, что прямые атаки сводятся к перебору либо информационных векторов $\boldsymbol{x}$, либо векторов ошибок $\boldsymbol{e}$. Сложность прямых атак можно оценить сверху следующим образом:

- Максимальное число раундов для восстановления $\boldsymbol{x}$: $\min\{2^{mk}, 2^{m(n-k)}\}$, на каждом раунде происходит умножение вектора длины mk или $m(n-k)$ на публичную проверочную или порождающую матрицу, требующее $m^2k(n-k)$ операций;
- Максимальное число раундов для восстановления $\boldsymbol{e}$: $\binom{mn}{\lfloor t/4 \rfloor} 2^{m-1}$, на каждом раунде вектор $\boldsymbol{e}$ вычитается из принятого вектора $\boldsymbol{y}$ и считается синдром, это требует $m^2k(n-k)$ операций.

Таким образом, сложность C_{dir} прямой атаки можно оценить как

$$C_{\text{dir}} = \mathcal{O}\left(m^2k(n-k) \cdot \min\left\{ 2^{mk},\ 2^{m(n-k)},\ \binom{mn}{\left\lfloor \frac{t}{4} \right\rfloor} 2^{m-1} \right\} \right).$$

Атака на основе декодирования по информационным совокупностям. Как известно, для классической криптосистемы Мак-Элиса атака декодирования по информационным совокупностям является наиболее эффективной – именно она определяет сложность раскрытия криптосистемы и влияет на выбор параметров кодов (а значит, и длины публичного и секретного ключей), необходимых для достижения заданного уровня стойкости.

Поскольку в вектор $\boldsymbol{e}$ длины mn вносится $\left\lfloor \frac{t}{4} \right\rfloor$ пакетов ошибок длины до m, то для поиска информационной совокупности, свободной от ошибок, необходимо найти $\left\lfloor \frac{t}{4} \right\rfloor$ индексов начала каждого из пакетов ошибок и считать, что длина каждого пакета ошибок равна m. Количество раундов для нахождения данных позиций не превосходит $\binom{mn}{\lfloor t/4 \rfloor}$. Таким образом, сложность нахождения информационной совокупности, свободной от ошибок, есть

$$C_{\text{ISD}} = \binom{mn}{\left\lfloor \frac{t}{4} \right\rfloor} m^2k(n-k).$$

Если искать информационную совокупность среди дополнения к множеству из $\left\lfloor \frac{t}{4} \right\rfloor$ непересекающихся пакетов ошибок длины m, то последнюю оценку можно уточнить:

$$C_{\mathrm{ISD}} = \\ = \frac{m(n-1)(m(n-1)-m)(m(n-1)-2m)\dots\left(m(n-1)-m\left\lfloor \frac{t}{4} \right\rfloor\right)}{\left(\left\lfloor \frac{t}{4} \right\rfloor\right)!} m^2 k(n-k).$$

Синдромная атака. Суть синдромной атаки состоит в том, чтобы по публичной порождающей матрице $\boldsymbol{G}$ вычислить публичную проверочную матрицу $\boldsymbol{H}$, а затем свести задачу нахождения $\boldsymbol{x}$ из соотношения $\boldsymbol{y} = \boldsymbol{x}\boldsymbol{G} + \boldsymbol{e}$ к решению соответствующего синдромного уравнения за счет умножения обеих частей на $\boldsymbol{H}^{\mathrm{T}}$. Так как

$$\boldsymbol{G} = \boldsymbol{S}\boldsymbol{G}'_b\boldsymbol{Q},$$

то

$$\boldsymbol{H} = \boldsymbol{L}\boldsymbol{H}'_b(\boldsymbol{Q}^{-1})^{\mathrm{T}},$$

где $\boldsymbol{H}'_b$ – проверочная матрица, соответствующая порождающей матрице $\boldsymbol{G}'_b$, а $\boldsymbol{L}$ – некоторая невырожденная матрица размера $m(n-k)\times m(n-k)$ над полем $\mathbb{F}_2$. Ясно, что матрица $\boldsymbol{L}$ не влияет на свойства кода. Поэтому будем полагать, что $\boldsymbol{L} = \boldsymbol{I}$. В этом случае синдром $\boldsymbol{Z}$ зашифрованного текста $\boldsymbol{y}$ имеет вид

$$\boldsymbol{Z} = \boldsymbol{y}\boldsymbol{H}^{\mathrm{T}} = \boldsymbol{e}\boldsymbol{Q}^{-1}(\boldsymbol{H}'_b)^{\mathrm{T}}.$$

Однако ввиду произвольности в выборе $\boldsymbol{Q}$ проверочная матрица $\boldsymbol{H}$ может соответствовать коду с расстоянием значительно меньшим, чем расстояние кода C_b. Таким образом, применение синдромной атаки не позволяет гарантированно найти вектор пакетов ошибок, сгенерированный на этапе шифрования.

5.5. Длина ключей. Сложность C_{comp} криптоанализа предложенной в статье криптосистемы мы будем оценивать сверху величиной

$$C_{\mathrm{comp}} = \mathcal{O}(\min\{C_{\mathrm{dir}}, C_{\mathrm{ISD}}\}).$$

Таким образом, для получения заданной стойкости W системы необходимо выбрать такой (n,k)-код (возможно, укороченный) $GRS_{n,k}(\boldsymbol{\alpha},\boldsymbol{v})$ над полем $\mathbb{F}_q$, чтобы $W \leqslant C_{\mathrm{comp}}$. Длина публичного ключа при этом составит $L_{\mathrm{pub}} = knm^2$. Таким образом, оптимальная с точки зрения длины ключа криптосистема, имеющая стойкость W, определяется тройкой параметров (n,k,m), $n \leqslant 2^m - 1 = q-1$, $k < n$, для которых

$$\begin{cases} knm^2 \to \min, \\ C_{\mathrm{comp}} \geqslant W, \\ n \leqslant 2^m - 1, \\ 0 < k < n. \end{cases}$$

Рассмотрим несколько примеров.

Пример 1. Пусть $W = 2^{72}$. Рассмотрим укороченный обобщенный $(76,18)$-код Рида–Соломона над полем $\mathbb{F}_q$, $q = 2^7$, полученный из обобщенного кода Рида–Соломона над полем $\mathbb{F}_q$. Данный код имеет расстояние 59 и исправляет 29 любых независимых q-ичных ошибок. Рассмотрим далее двоичный образ этого кода, представив каждый элемент поля $\mathbb{F}_q$ в виде двоичного вектора длины 7. В результате получим двоичный $(532,126)$-код C, исправляющий до 29 пакетов ошибок длины

до 7. Если взять данный код в качестве основы для построения описанной выше криптосистемы, то сложность криптоанализа системы оценивается как

1. $2^{mk}m^2k(n-k) > 2^{141}$ – сложность атаки на основе перебора информационных векторов;
2. $2^{m(n-k)}m^2k(n-k) > 2^{421}$ – сложность атаки на основе перебора информационных векторов для двойственного кода;
3. $\binom{mn}{\lfloor t/4 \rfloor}2^{m-1}m^2k(n-k) > 2^{72}$ – сложность атаки на основе перебора всех векторов ошибок;
4. Сложность атаки по информационным совокупностям оценивается как

$$C_{\text{ISD}} = \frac{7 \cdot 75 \cdot (7 \cdot 75 - 7) \cdot (7 \cdot 75 - 14) \cdot \ldots \cdot (7 \cdot 75 - 49)}{7!} \cdot 49 \cdot 18 \cdot 58 > 2^{75}.$$

Таким образом, стойкость криптосистемы $W_c \approx 2^{72} \approx W$. При этом длина ключа составляет $L_{\text{pub}} = 76 \cdot 18 \cdot 7^2 = 67032$ бит, что более чем в 8 раз меньше длины ключа криптосистемы Мак-Элиса, основанной на $(1024, 524, 101)$-коде Гоппы и имеющей стойкость 2^{72}.

"Полнота" множества вносимых ошибок согласно формуле (5) оценивается снизу величиной

$$\tau_{GRS} = \frac{\log_2 |\mathcal{E}|}{m(n-k)} = \frac{\log_2 \left(\binom{mn}{\lfloor t/4 \rfloor} 2^{m-1} \right)}{m(n-k)} \approx 0{,}1405,$$

что значительно уступает оценке данной величины для криптосистемы Мак-Элиса $\tau_{\text{ME}} \approx 0{,}5681$. Это в первую очередь говорит о том, что двоичный образ обобщенного кода Рида – Соломона способен исправлять значительно более широкое множество ошибок, нежели генерируемое в рамках данной криптосистемы. А значит, в перспективе возможно дальнейшее уменьшение длины публичного ключа, что и будет сделано в последующих примерах.

Приведем еще один пример параметров криптосистемы в предположении того, что на этапе шифрования вносится $\left\lfloor \frac{t}{3} \right\rfloor$ пакета ошибок длины до m. Напомним, что при этом размеры матриц $\boldsymbol{M}_i$ выбираются из множества $\{m+1, m+2\}$. Оценки сложности криптоанализа при этом очевидно получаются из аналогичных соотношений для случая внесения $\left\lfloor \frac{t}{4} \right\rfloor$ пакетов ошибок.

Пример 2. Пусть $W = 2^{72}$. Рассмотрим обобщенный $(63, 15)$-код Рида – Соломона над полем $\mathbb{F}_q$, $q = 2^6$, полученный из обобщенного кода Рида – Соломона над полем $\mathbb{F}_q$. Данный код имеет расстояние 48 и исправляет 24 любые q-ичные независимые ошибки. Рассмотрим далее двоичный образ этого кода, представив каждый элемент поля $\mathbb{F}_q$ в виде двоичного вектора длины 6. В результате получим двоичный $(378, 90)$-код C, исправляющий до 24 пакетов ошибок длины до 6. Если взять данный код в качестве основы для построения описанной выше криптосистемы, то сложность криптоанализа системы оценивается как

1. $2^{mk}m^2k(n-k) > 2^{104}$ – сложность атаки на основе перебора информационных векторов;
2. $2^{m(n-k)}m^2k(n-k) > 2^{302}$ – сложность атаки на основе перебора информационных векторов для двойственного кода;
3. $\binom{mn}{\lfloor t/3 \rfloor}2^{m-1}m^2k(n-k) > 2^{72}$ – сложность атаки на основе перебора всех векторов ошибок;

4. Сложность атаки по информационным совокупностям оценивается как

$$C_{\text{ISD}} = \frac{6 \cdot 62 \cdot (6 \cdot 62 - 6) \cdot (6 \cdot 62 - 12) \cdot \ldots \cdot (6 \cdot 62 - 96)}{8!} \cdot 36 \cdot 15 \cdot 48 > 2^{75}.$$

Таким образом, стойкость криптосистемы $W_c \approx 2^{72} \approx W$. При этом длина ключа составляет $L_{\text{pub}} = 63 \cdot 15 \cdot 6^2 = 34020$ бит, что более чем в 15 раз меньше длины ключа криптосистемы Мак-Элиса, основанной на $(1024, 524, 101)$-коде Гоппы и имеющей стойкость 2^{72}.

"Полнота" множества вносимых ошибок для этой криптосистемы согласно формуле (5) оценивается снизу величиной

$$\tau_{GRS} = \frac{\log_2 |\mathcal{E}|}{m(n-k)} = \frac{\log_2 \left(\binom{mn}{\lfloor t/3 \rfloor} 2^{m-1} \right)}{m(n-k)} \approx 0{,}19147,$$

что по-прежнему меньше, чем у криптосистемы Мак-Элиса, но больше, чем у криптосистемы, основанной на укороченном обобщенном $(76, 18)$-коде Рида – Соломона.

В заключение этого параграфа рассмотрим еще один пример параметров криптосистемы в предположении того, что на этапе шифрования вносится $\left\lfloor \frac{t}{2} \right\rfloor$ пакетов ошибок длины до m. Напомним, что при этом накладываются некоторые дополнительные ограничения на матрицу $\boldsymbol{Q}$, которые были рассмотрены ранее. Оценки сложности криптоанализа при этом очевидно получаются из аналогичных соотношений для случая внесения $\left\lfloor \frac{t}{4} \right\rfloor$ или $\left\lfloor \frac{t}{3} \right\rfloor$ пакетов ошибок.

Пример 3. Пусть $W = 2^{72}$. Рассмотрим укороченный обобщенный $(46, 10)$-код Рида – Соломона над полем $\mathbb{F}_q$, $q = 2^6$. Данный код имеет расстояние 37 и исправляет 18 любых q-ичных независимых ошибок. Рассмотрим далее двоичный образ этого кода, представив каждый элемент поля $\mathbb{F}_q$ в виде двоичного вектора длины 6. В результате получим двоичный $(276, 60)$-код C, исправляющий до 18 пакетов ошибок длины до 6. Если взять данный код в качестве основы для построения описанной выше криптосистемы, то сложность криптоанализа системы оценивается как

1. $2^{mk} m^2 k(n-k) > 2^{73}$ – сложность атаки на основе перебора информационных векторов;
2. $2^{m(n-k)} m^2 k(n-k) > 2^{229}$ – сложность атаки на основе перебора информационных векторов для двойственного кода;
3. $\binom{mn}{\lfloor t/2 \rfloor} 2^{m-1} m^2 k(n-k) \approx 2^{73}$ – сложность атаки на основе перебора всех векторов ошибок;
4. Сложность атаки по информационным совокупностям оценивается как

$$C_{\text{ISD}} = \frac{6 \cdot 45 \cdot (6 \cdot 45 - 6) \cdot (6 \cdot 45 - 12) \cdot \ldots \cdot (6 \cdot 45 - 63)}{9!} \cdot 36 \cdot 10 \cdot 36 > 2^{74}.$$

Таким образом, стойкость криптосистемы $W_c \approx 2^{73} > W$. При этом длина ключа составляет $L_{\text{pub}} = 46 \cdot 10 \cdot 6^2 = 16560$ бит, что более чем в 32 раза меньше длины ключа криптосистемы Мак-Элиса, основанной на $(1024, 524, 101)$-коде Гоппы и имеющей стойкость 2^{72}.

"Полнота" множества вносимых ошибок для такой криптосистемы согласно формуле (5) оценивается снизу величиной

$$\tau_{GRS} = \frac{\log_2 |\mathcal{E}|}{m(n-k)} = \frac{\log_2 \left(\binom{mn}{\lfloor t/2 \rfloor} 2^{m-1} \right)}{m(n-k)} \approx 0{,}2746,$$

что по-прежнему меньше, чем у криптосистемы Мак-Элиса, но больше, чем у криптосистем, основанных на укороченных обобщенных $(76, 18)$- и $(63, 15)$-кодах Рида – Соломона.

§ 6. Заключение

В статье рассмотрена общая постановка задачи построения криптосистемы с открытым ключом на основе кодов, исправляющих ошибки.

Сформулированы свойства, которым должна соответствовать кодовая криптосистема для того, чтобы обеспечивать требуемый уровень стойкости.

Предложен критерий τ для сравнения криптосистем между собой, оценивающий взаимосвязь между мощностью множества ошибок, вводимых в криптосистему для обеспечения ее стойкости, и количеством проверочных символов у кода, лежащего в основе криптосистемы. Таким образом, величину $1-\tau$ можно трактовать как меру потенциала улучшения криптосистемы за счет дальнейшего расширения множества вносимых ошибок.

Чтобы продемонстрировать теоретическую возможность построения криптосистем, для которых возможно выполнение сформулированных в статье свойств, описана конструкция, удовлетворяющая предложенному набору условий. Эта конструкция основана на двоичных образах обобщенных кодов Рида – Соломона. Продемонстрировано, что данная конструкция имеет меньшую длину ключа для заданных параметров стойкости по сравнению с криптосистемой Мак-Элиса на основе двоичных кодов Гоппы.

В результате анализа свойств криптосистемы было показано, что наиболее перспективными являются криптосистемы, где количество вносимых ошибок значительно больше, чем половина минимального расстояния, при этом декодирование по информационным совокупностям перестает быть наиболее эффективной стратегией атаки по декодированию.

Результаты статьи показывают, что построение кодовых криптосистем на основе использования маскирующих векторов (векторов ошибки) малого веса Хэмминга не является эффективным, поскольку такие системы чувствительны к атаке на основе поиска свободной от ошибок информационной совокупности. Использование векторов ошибок, не являющихся самыми легкими в своих смежных классах, позволяют существенно снизить эффективность атаки по информационным совокупностям. При этом встает задача поиска преобразований, отображающих легкие представители смежных классов кодов в более тяжелые. Задача эта в теории кодирования, насколько нам известно, не решалась. Мы надеемся, что эта задача может оказаться плодотворной как в криптографии, так и в других приложениях теории помехоустойчивого кодирования.

СПИСОК ЛИТЕРАТУРЫ

1. *McEliece R.J.* A Public-Key Cryptosystem Based on Algebraic Coding Theory // DSN Progress Report 42-44. Jet Propulsion Lab., California Inst. of Technology, Pasadena, CA. 1978. P. 114–116. Available at `https://tmo.jpl.nasa.gov/progress_report2/42-44/44N.PDF`
2. *Kabatianskii G., Krouk E., Smeets B.* A Digital Signature Scheme Based on Random Error-Correcting Codes // Crytography and Coding (Proc. 6th IMA Int. Conf. on Cryptography and Coding. Cirencester, UK. Dec. 17–19, 1997). Lect. Notes Comput. Sci. V. 1355. Berlin: Springer, 1997. P. 161–167. `https://doi.org/10.1007/BFb0024461`
3. *Rivest R.L., Shamir A., Adleman L.* A Method for Obtaining Digital Signatures and Public-Key Cryptosystems // Commun. ACM. 1978. V. 21. № 2. P. 120–126. `https://doi.org/10.1145/359340.359342`

4. *El Gamal T.* A Public Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms // IEEE Trans. Inform. Theory. 1985. V. 31. № 4. P. 469–472. https://doi.org/10.1109/TIT.1985.1057074
5. *Véron P.* Code Based Cryptography and Steganography // Algebraic Informatics (Proc. 5th Int. Conf. on Algebraic Informatics (CAI'2013). Porquerolles, France. Sept. 3–6, 2013). Lect. Notes Comput. Sci. V. 8080. Berlin: Springer, 2013. P. 9–46. https://doi.org/10.1007/978-3-642-40663-8_5
6. *Berger T.P., Cayrel P.L., Gaborit P., Otmani A.* Reducing Key Length of the McEliece Cryptosystem // Progress in Cryptology — AFRICACRYPT 2009 (Proc. 2nd Int. Conf. on Cryptology in Africa. Gammarth, Tunisia. June 21–25, 2009). Lect. Notes Comput. Sci. V. 5580. Berlin: Springer, 2009. P. 77–97. https://doi.org/10.1007/978-3-642-02384-2_6
7. *Faugère J.-C., Otmani A., Perret L., Tillich J.-P.* Algebraic Cryptanalysis of McEliece Variants with Compact Keys // Advances in Cryptology — EUROCRYPT 2010 (Proc. 29th Annu. Int. Conf. on the Theory and Applications of Cryptographic Techniques. French Riviera. May 30 – June 3, 2010). Lect. Notes Comput. Sci. V. 6110. Berlin: Springer, 2010. P. 279–298. https://doi.org/10.1007/978-3-642-13190-5_14
8. *Kocher P.C.* Timing Attacks on Implementations of Diffie–Hellman, RSA, DSS, and Other Systems // Advances in Cryptology — CRYPTO'96 (Proc. 16th Annu. Int. Cryptology Conf. Santa Barbara, CA, USA. Aug. 18–22, 1996). Lect. Notes Comput. Sci. V. 1109. Berlin: Springer, 1996. P. 104–113. https://doi.org/10.1007/3-540-68697-5_9
9. *Barker E.* NIST Special Publication (SP) 800-57 Part 1 Revision 4. Recommendation for Key Management – Part 1: General. National Inst. of Standards and Technology, Gaithersburg, MD, USA, 2016. https://doi.org/10.6028/NIST.SP.800-57pt1r4
10. *Chen L., Jordan S., Liu Y.-K., Moody D., Peralta R., Perlner R., Smith-Tone D.* Report on Post-Quantum Cryptography. NIST Internal Report 8105. National Inst. of Standards and Technology, Gaithersburg, MD, USA, 2016. https://doi.org/10.6028/NIST.IR.8105
11. *Shor P.W.* Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer // SIAM Rev. 1999. V. 41. № 2. P. 303–332. https://doi.org/10.1137/S0036144598347011
12. *Berlekamp E., McEliece R., van Tilborg H.* On the Inherent Intractability of Certain Coding Problems // IEEE Trans. Inform. Theory. 1978. V. 24. № 3. P. 384–386. https://doi.org/10.1109/TIT.1978.1055873
13. *Eisenbarth T., Kumar S., Paar C., Poschmann A., Uhsadel L.* A Survey of Lightweight-Cryptography Implementations // IEEE Des. Test Comput. 2007. V. 24. № 6. P. 522–533. https://doi.org/10.1109/MDT.2007.178
14. *Ivanov F., Krouk E., Kreshchuk A.* On the Lightweight McEliece Cryptosystem for Low-Power Devices // Proc. 2019 XVI Int. Symp. "Problems of Redundancy in Information and Control Systems" (REDUNDANCY). Moscow, Russia. Oct. 21–25, 2019. P. 133–138. https://doi.org/10.1109/REDUNDANCY48165.2019.9003324
15. *Misoczki R., Tillich J.-P., Sendrier N., Barreto P.S.L.M.* MDPC-McEliece: New McEliece Variants from Moderate Density Parity-Check Codes // Proc. 2013 IEEE Int. Symp. on Information Theory (ISIT'2013). Istanbul, Turkey. July 7–12, 2013. P. 2069–2073. https://doi.org/10.1109/ISIT.2013.6620590
16. *Baldi M., Chiaraluce F., Garello R., Mininni F.* Quasi-cyclic Low-Density Parity-Check Codes in the McEliece Cryptosystem // Proc. 2007 IEEE Int. Conf. on Communications (ICC'2007). Glasgow, UK. June 24–28, 2007. P. 951–956. https://doi.org/10.1109/ICC.2007.161
17. *Крук Е.А.* Граница для сложности декодирования линейных блоковых кодов // Пробл. передачи информ. 1989. Т. 25. № 3. С. 103–107. http://mi.mathnet.ru/ppi665
18. *Сидельников В.М., Шестаков С.О.* О системе шифрования, построенной на основе обобщенных кодов Рида – Соломона // Дискрет. матем. 1992. Т. 4. № 3. С. 57–63. http://mi.mathnet.ru/dm747
19. *Bernstein D.J., Lange T., Peters C.* Attacking and Defending the McEliece Cryptosystem // Post-Quantum Cryptography (Proc. 2nd Int. Workshop on Post-Quantum Cryp-

tography (PQCrypto 2008). Cincinnati, OH, USA. Oct. 17–19, 2008). Lect. Notes Comput. Sci. V. 5299. Berlin: Springer, 2008. P. 31–46. https://doi.org/10.1007/978-3-540-88403-3_3

20. *Becker A., Joux A., May A., Meurer A.* Decoding Random Binary Linear Codes in $2^{n/20}$: How $1 + 1 = 0$ Improves Information Set Decoding // Advances in Cryptology — EUROCRYPT 2012 (Proc. 31st Annu. Int. Conf. on the Theory and Applications of Cryptographic Techniques. Cambridge, UK. Apr. 15–19, 2012). Lect. Notes Comput. Sci. V. 7237. Berlin: Springer, 2012. P. 520–536. https://doi.org/10.1007/978-3-642-29011-4_31

Зяблов Виктор Васильевич
Институт проблем передачи информации
им. А.А. Харкевича РАН
zyablov@iitp.ru
Иванов Федор Ильич
Институт проблем передачи информации
им. А.А. Харкевича РАН
Национальный исследовательский университет
"Высшая школа экономики"
fivanov@hse.ru
Крук Евгений Аврамович
Национальный исследовательский университет
"Высшая школа экономики"
ekrouk@hse.ru
Сидоренко Владимир Рэмович
Институт проблем передачи информации
им. А.А. Харкевича РАН
Технический университет Мюнхена, Германия
vladimir.sidorenko@tum.de

Поступила в редакцию
30.09.2020
После доработки
14.04.2022
Принята к публикации
16.04.2022

Адрес редакции: 127051, Москва, Б. Каретный пер., 19, стр. 1, тел. (495) 650-47-39

Оригинал-макет подготовил *Д.Ю. Ногин*
по контракту с ООО «Тематическая редакция»

Москва
ООО «Тематическая редакция»