ISSN 0555-2923

РОССИЙСКАЯ АКАДЕМИЯ НАУК

Проблемы передачи информации

НАУКА
1727

том 60 вып. 3

2024

РОССИЙСКАЯ АКАДЕМИЯ НАУК

ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

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

Том 60, 2024 **Вып. 3**

Июль–Август–Сентябрь

М о с к в а

С О Д Е Р Ж А Н И Е

CONTENTS

Coding Theory

Methods of Signal Processing

Large Systems

Communication Network Theory

ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 60 **2024** **Вып. 3**

УДК 621.391 : 519.725

© 2024 г. **А.М. Романов**

О РАНГЕ НЕЛИНЕЙНЫХ КВАЗИСОВЕРШЕННЫХ КОДОВ НАД КОНЕЧНЫМИ ПОЛЯМИ[1]

Рассматриваются нелинейные квазисовершенные коды с радиусом упаковки 1 над конечным полем из q элементов, где q – степень простого числа. Эти коды мы называем нелинейными 1-квазисовершенными q-ичными кодами. Изучаются ранг и размерность ядра нелинейных 1-квазисовершенных q-ичных кодов. Если ранг кода равен его длине, то код называется кодом полного ранга. Пусть m – положительное целое число. Доказывается, что при $n = q^m$ и при достаточно больших m и q существуют нелинейные 1-квазисовершенные q-ичные коды полного ранга длины n. Также для некоторых нелинейных 1-квазисовершенных q-ичных кодов вычисляются размерности ядра.

Ключевые слова: квазисовершенные коды, нелинейные коды, обобщенные коды Рида – Маллера, переключающая конструкция, ранг кода, ядро кода, геометрия Галуа.

DOI: 10.31857/S055529232403001X, **EDN:** OKKCTV

§ 1. Введение

Пусть $\mathbb{F}_q^n$ – векторное пространство размерности n над конечным полем $\mathbb{F}_q$ порядка q, где q – степень простого числа p. Произвольное непустое подмножество $\mathcal{C} \subseteq \mathbb{F}_q^n$ называется *q-ичным кодом с исправлением ошибок* (кратко – q-ичным кодом). *Длина* кода $\mathcal{C} \subseteq \mathbb{F}_q^n$ равна размерности пространства $\mathbb{F}_q^n$. Мы предполагаем, что нулевой вектор $\mathbf{0}$ всегда принадлежит коду, если не указано иное. Векторы, принадлежащие пространству $\mathbb{F}_q^n$, мы будем рассматривать как слова длины n над алфавитом $\mathbb{F}_q$. Слова, принадлежащие коду $\mathcal{C} \subseteq \mathbb{F}_q^n$, будем называть *кодовыми словами*. Код называется *линейным*, если он образует линейное подпространство над $\mathbb{F}_q$. В противном случае код называется *нелинейным*.

Для любой пары слов $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ и $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ из $\mathbb{F}_q^n$ определим *расстояние Хэмминга* $d(\boldsymbol{x}, \boldsymbol{y})$. Положим

$$d(\boldsymbol{x}, \boldsymbol{y}) := |\{i \mid x_i \neq y_i\}|.$$

Сферой Хэмминга радиуса r с центром в слове $\boldsymbol{x}$ называется множество

$$B_r(\boldsymbol{x}) := \{\boldsymbol{y} \in \mathbb{F}_q^n \mid d(\boldsymbol{x}, \boldsymbol{y}) \leqslant r\}.$$

Радиус упаковки $e(\mathcal{C})$ кода $\mathcal{C}$ длины n – это максимальное число $e \in \{0, 1, \ldots, n\}$ такое, что

$$B_e(\boldsymbol{u}) \cap B_e(\boldsymbol{v}) = \varnothing \quad \text{для всех} \quad \boldsymbol{u}, \boldsymbol{v} \in \mathcal{C}, \quad \boldsymbol{u} \neq \boldsymbol{v}.$$

[1] Работа выполнена в рамках государственного задания ИМ СО РАН, тема FWNF-2022-0017.

Радиус покрытия $\rho(\mathcal{C})$ кода $\mathcal{C}$ длины n – это минимальное число $\rho \in \{0, 1, \ldots, n\}$ такое, что

$$\bigcup_{\boldsymbol{c}\in\mathcal{C}} B_\rho(\boldsymbol{c}) = \mathbb{F}_q^n.$$

Код $\mathcal{C}$ называется *совершенным*, если $\rho(\mathcal{C}) = e(\mathcal{C})$, и код $\mathcal{C}$ называется *квазисовершенным*, если $\rho(\mathcal{C}) = e(\mathcal{C}) + 1$. Если радиус упаковки совершенного (квазисовершенного) кода равен e, то код называется e-совершенным (e-квазисовершенным).

В [1] установлена тесная связь 1-совершенных кодов и 1-квазисовершенных кодов с параметрами обобщенных кодов Рида – Маллера порядка

$$r = (q-1)m - 2.$$

В [1] показано, что с помощью конкатенации из 1-квазисовершенных кодов с параметрами обобщенных кодов Рида – Маллера порядка $r = (q-1)m - 2$ можно построить дважды экспоненциальное число попарно неэквивалентных 1-совершенных кодов над $\mathbb{F}_q$. В [2] показано, что с помощью переключающей конструкции из обобщенных кодов Рида – Маллера порядка $r = (q-1)m-2$ можно построить дважды экспоненциальное число попарно неэквивалентных 1-квазисовершенных кодов над $\mathbb{F}_q$. В [3] установлена тесная связь 1-совершенных кодов над смешанным алфавитом и кодов с параметрами обобщенных кодов Рида – Маллера порядка $r = (q-1)m - 2$. В [3] показано, что из кодов с параметрами обобщенных кодов Рида – Маллера порядка $r = (q-1)m - 2$ можно построить дважды экспоненциальное число попарно неэквивалентных 1-совершенных кодов над смешанным алфавитом.

Обзоры работ по квазисовершенным кодам можно найти в [4–6].

Рангом кода $\mathcal{C} \subseteq \mathbb{F}_q^n$ называется размерность подпространства, натянутого на $\mathcal{C}$. Ранг кода $\mathcal{C}$ мы будем обозначать через $\operatorname{rank}(\mathcal{C})$. Если код $\mathcal{C}$ является линейным кодом размерности k, то $\operatorname{rank}(\mathcal{C}) = k$. Если код $\mathcal{C}$ является нелинейным кодом длины n и содержит q^k слов, то $k+1 \leqslant \operatorname{rank}(\mathcal{C}) \leqslant n$. Код $\mathcal{C}$ длины n называется кодом *полного ранга*, если $\operatorname{rank}(\mathcal{C}) = n$.

Ядром кода $\mathcal{C} \subseteq \mathbb{F}_q^n$ называется множество

$$\ker(\mathcal{C}) := \left\{\boldsymbol{x} \in \mathbb{F}_q^n \mid \lambda \cdot \boldsymbol{x} + \mathcal{C} = \mathcal{C} \text{ для любого } \lambda \in \mathbb{F}_q\right\}.$$

Легко заметить, что если нулевое слово принадлежит $\mathcal{C}$, то $\ker(\mathcal{C})$ является линейным подкодом кода $\mathcal{C}$, и $\mathcal{C}$ является объединением смежных классов, образованных подпространством $\ker(\mathcal{C})$. Размерность ядра кода $\mathcal{C}$ будем обозначать через $\dim(\ker(\mathcal{C}))$. Если код $\mathcal{C}$ является линейным кодом размерности k, то $\dim(\ker(\mathcal{C})) = k$. Если код $\mathcal{C}$ является нелинейным кодом, который содержит q^k слов и содержит нулевое слово, то

$$0 \leqslant \dim(\ker(\mathcal{C})) \leqslant k-2 \quad \text{при} \quad q = 2,$$
$$0 \leqslant \dim(\ker(\mathcal{C})) \leqslant k-1 \quad \text{при} \quad q \geqslant 3$$

(см. [7]).

Пусть $\mathcal{C}$ и $\mathcal{D}$ – коды длины n над полем $\mathbb{F}_q$. Тогда коды $\mathcal{C}$ и $\mathcal{D}$ называются *эквивалентными*, если существуют n перестановок $\pi_1, \ldots, \pi_n$ из q элементов и перестановка σ, переставляющая n координатных позиций, такие что $(u_1, \ldots, u_n) \in \mathcal{C}$ тогда и только тогда, когда $\sigma(\pi_1(u_1), \ldots, \pi_n(u_n)) \in \mathcal{D}$.

При приведенном выше определении эквивалентности кодов и при $q \geqslant 5$ существуют нелинейные q-ичные коды, содержащие нулевое слово и эквивалентные линейному q-ичному коду [8, 9]. Ниже мы приведем другое определение эквивалентности q-ичных кодов, при котором ранг и размерность ядра кода будут являться его

алгебраическими инвариантами, т.е. эквивалентные коды, содержащие нулевое слово, будут иметь одинаковый ранг и будут содержать ядра одинаковой размерности. Заметим, что помимо алгебраических инвариантов изучаются также комбинаторные инварианты 1-совершенных кодов [9, 10].

Мономиальная матрица – это матрица, имеющая ровно один ненулевой элемент в каждой строке и каждом столбце.

Пусть $\mathcal{C}$ и $\mathcal{D}$ – коды длины n над полем $\mathbb{F}_q$. Тогда $\mathcal{C}$ и $\mathcal{D}$ называются *эквивалентными*, если существуют вектор $\boldsymbol{v} \in \mathbb{F}_q^n$ и мономиальная матрица M размера $n \times n$ над $\mathbb{F}_q$ такие, что

$$\mathcal{D} = \{\boldsymbol{v} + \boldsymbol{c}M \mid \boldsymbol{c} \in \mathcal{C}\}.$$

Вопрос о ранге и размерности ядра является естественным для нелинейных кодов и связан с классификацией кодов. Ранг и размерность ядра изучались, например, для совершенных кодов, кодов Адамара, $\mathbb{Z}_p\mathbb{Z}_{p^2}$-линейных кодов [7,11–14]. В [15] изучались ранг и размерность ядра некоторых подклассов 1-квазисовершенных двоичных кодов.

В данной статье доказано, что при любом $t \in \{1, 2, \ldots, m+1\}$ существуют нелинейные 1-квазисовершенные q-ичные коды длины n и ранга $n-m-1+t$, где $n = q^m$. При этом $m \geqslant 5$ при $q = 3, 4$, $m \geqslant 4$ при $5 \leqslant q \leqslant 19$ и $m \geqslant 3$ при $q \geqslant 23$. В том числе доказано, что при $n = q^m$ существуют нелинейные 1-квазисовершенные q-ичные коды полного ранга длины n. Обобщенные коды Рида – Маллера $RM_q(r, m)$ порядка

$$r = (q-1)m - 2$$

достигают границы Джонсона и являются оптимальными кодами [1]. Ранг обобщенного кода Рида – Маллера $RM_q(r, m)$ порядка $r = (q-1)m-2$ равен $n-m-1$ (см. [1]). Таким образом, в данной статье доказано существование нелинейных 1-квазисовершенных q-ичных кодов всех возможных рангов при длине $n = q^m$, мощности q^{n-m-1} и минимальном расстоянии 3.

В данной статье также при $q \geqslant 3$, $m \geqslant 2$, $n = q^m$ предложена конструкция нелинейных 1-квазисовершенных q-ичных кодов длины n и ранга $n - m$.

Кроме того, в данной статье для некоторых нелинейных 1-квазисовершенных q-ичных кодов вычисляется размерность ядра.

При $q \geqslant 3$, $m \geqslant 2$, $n = q^m$ предложена конструкция нелинейных 1-квазисовершенных q-ичных кодов длины n с ядром размерности $n - [m]_q - 1$, где $[m]_q$ является q-аналогом натурального числа m. По определению

$$[m]_q := 1 + q + \cdots + q^{m-1}.$$

В данной статье рассматривается переключающая (свитчинговая) конструкция, и с помощью этой конструкции обобщенные коды Рида – Маллера $RM_q(r, m)$ порядка $r = (q-1)m - 2$ преобразуются в нелинейные 1-квазисовершенные q-ичные коды различных рангов, а также с помощью этой конструкции обобщенные коды Рида – Маллера $RM_q(r, m)$ порядка $r = (q-1)m - 2$ преобразуются в нелинейные 1-квазисовершенные q-ичные коды с ядром разных размерностей.

В работах [7,11,12,16] переключающие конструкции применялись для построения нелинейных 1-совершенных кодов. В [2] переключающая конструкция впервые получила распространение на 1-квазисовершенные коды и при ее помощи из обобщенных кодов Рида – Маллера порядка $r = (q-1)m - 2$ было построено дважды экспоненциальное число попарно неэквивалентных нелинейных 1-квазисовершенных кодов над $\mathbb{F}_q$.

Заметим, что в двоичном случае обобщенный код Рида – Маллера $RM_q(r, m)$ порядка $r = (q-1)m - 2$ является расширенным кодом Хэмминга. Таким образом,

существование нелинейных 1-квазисовершенных двоичных кодов различных рангов и существование нелинейных 1-квазисовершенных двоичных кодов с ядром разных размерностей следует из работ [7, 11].

В данной статье раскрываются новые аспекты тесной связи 1-совершенных кодов над $\mathbb{F}_q$ и 1-квазисовершенных кодов с параметрами обобщенных кодов Рида – Маллера порядка $r = (q-1)m - 2$. Данная статья является сокращенным вариантом работы [17], в которой можно найти подробные доказательства приводимых далее утверждений.

В §2 приведены некоторые обозначения и основные определения, а также дан краткий обзор известных результатов. В §3 дано описание переключающей конструкции. В §4 показывается существование и приведена конструкция нелинейных 1-квазисовершенных q-ичных кодов различных рангов, а также приведена конструкция нелинейных 1-квазисовершенных q-ичных кодов с ядрами различных размерностей.

§ 2. Предварительные сведения

В этом параграфе мы приводим некоторые обозначения и основные определения, а также даем краткий обзор известных результатов, которые понадобятся нам в дальнейшем.

Через $\dim(L)$ обозначается размерность подпространства $L \subseteq \mathbb{F}_q^n$.

Минимальное расстояние между любыми двумя различными кодовыми словами кода $\mathcal{C}$ называется его *минимальным расстоянием*. Мы будем использовать обозначение $(n, M, d)_q$ для q-ичного кода длины n и мощности M с минимальным расстоянием d. Для линейного q-ичного кода длины n, размерности k с минимальным расстоянием d мы будем использовать обозначение $[n, k, d]_q$. В случае двоичных кодов индекс q может опускается.

Вес слова $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{F}_q^n$ равен числу ненулевых компонент в $\boldsymbol{x}$ и обозначается через $\mathrm{wt}(\boldsymbol{x})$.

Пусть $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{F}_q^n$, и пусть

$$p(\boldsymbol{x}) := \sum_{i=1}^{n} x_i.$$

Тогда слово $\boldsymbol{x}$ называется *четным*, если $p(\boldsymbol{x}) = 0$. Код $\mathcal{C} \subseteq \mathbb{F}_q^n$ называется *четным*, если он содержит только четные кодовые слова.

Для произвольного кода $\mathcal{C} \subseteq \mathbb{F}_q^n$ с параметрами $(n, M, d)_q$ определим *расширенный* код $\widehat{\mathcal{C}}$. Положим

$$\widehat{\mathcal{C}} := \left\{ \boldsymbol{x} = (x_1, x_2, \ldots, x_n, x_{n+1}) \in \mathbb{F}_q^{n+1} \mid (x_1, x_2, \ldots, x_n) \in \mathcal{C} \text{ и } p(\boldsymbol{x}) = 0 \right\}.$$

Расширенный код $\widehat{\mathcal{C}}$ имеет параметры $(n+1, M, \widehat{d})_q$, где $\widehat{d} = d$ или $d+1$.

Через $AG(m, q)$ обозначается аффинное пространство размерности m над $\mathbb{F}_q$. Векторы векторного пространства $\mathbb{F}_q^m$ являются *точками* аффинного пространства $AG(m, q)$. Смежные классы k-мерных линейных подпространств векторного пространства $\mathbb{F}_q^m$ являются k-мерными *аффинными подпространствами* аффинного пространства $AG(m, q)$. *Прямые* – это 1-мерные аффинные подпространства. Например, аффинная плоскость $AG(2, 3)$ содержит 9 точек и 12 прямых. Более подробные сведения по конечной аффинной геометрии можно найти, например, в [18].

Классические коды Рида – Маллера являются двоичными кодами [19, гл. 13]. Обобщение кодов Рида – Маллера на q-ичный случай было предложено в [20].

Пусть $\mathbb{F}_q[X_1, X_2, \ldots, X_m]$ – алгебра многочленов от m переменных над $\mathbb{F}_q$. Через $\deg(f)$ обозначим полную степень многочлена $f \in \mathbb{F}_q[X_1, X_2, \ldots, X_m]$. Пусть $n = q^m$, и пусть r – целое число, такое что $0 \leqslant r \leqslant (q-1)m$. Пусть точки $P_1, P_2, \ldots, P_n$ аффинного пространства $AG(m, q)$ некоторым образом упорядочены. *Обобщенным кодом Рида – Маллера* порядка r над $\mathbb{F}_q$ называется подпространство

$$RM_q(r, m) := \left\{ \big(f(P_1), f(P_2), \ldots, f(P_n)\big) \mid f \in \mathbb{F}_q[X_1, X_2, \ldots, X_m],\ \deg(f) \leqslant r \right\}.$$

Обобщенные коды Рида – Маллера $RM_q(r, m)$ имеют следующие параметры (см. [20; 21, теорема 5.5]):

- длина кода $RM_q(r, m)$ равна q^m;
- размерность кода $RM_q(r, m)$ равна

$$\sum_{i=0}^{r} \sum_{k=0}^{m} (-1)^k \binom{m}{k} \binom{i - kq + m - 1}{i - kq}; \tag{1}$$

- минимальное расстояние кода $RM_q(r, m)$ равно

$$(q - b)q^{m-a-1}, \tag{2}$$

где $r = (q-1)a + b$ и $0 \leqslant b < q - 1$.

При $q = 2$ и $0 \leqslant r \leqslant m$ обобщенный код Рида – Маллера порядка r является классическим двоичным кодом Рида – Маллера порядка r (см. [21, пример 5.4]).

При $r \leqslant (q-1)m-1$ обобщенный код Рида – Маллера $RM_q(r, m)$ является четным кодом.

Обобщенный код Рида – Маллера $RM_q(r, m)$ порядка $r = (q-1)m - 2$ при $q \geqslant 3$ имеет параметры $[n = q^m, n-m-1, 3]_q$ (см. [1]). В двоичном случае обобщенный код Рида – Маллера $RM_q(r, m)$ порядка $r = (q-1)m - 2$ является расширенным кодом Хэмминга и имеет параметры $[n = 2^m, n - m - 1, 4]$. Порядок кода, двойственного обобщенному коду Рида – Маллера $RM_q(r, m)$ порядка $r = (q-1)m - 2$, равен 1 [22]. Обобщенные коды Рида – Маллера $RM_q(r, m)$ порядка $r = (q-1)m - 2$ являются линейными 1-квазисовершенными кодами [1].

§ 3. Переключающая конструкция

В этом параграфе мы опишем переключающую конструкцию, применение которой к обобщенным кодам Рида – Маллера $RM_q((q-1)m - 2, m)$ позволит нам доказать существование и в некоторых случаях построить нелинейные 1-квазисовершенные q-ичные коды различных рангов, а также позволит доказать существование и в некоторых случаях построить нелинейные 1-квазисовершенные q-ичные коды с ядрами различных размерностей. Описываемая в этом параграфе переключающая конструкция основана на некоторых идеях из работ [2, 7, 11, 12, 16, 23].

Проверочная матрица H кода $RM_q((q-1)m-2, m)$ представляет собой матрицу размером $(m+1) \times q^m$, содержащую единичный вектор длины $n = q^m$, т.е. вектор, все компоненты которого равны единице, и содержащую все транспонированные векторы из $\mathbb{F}_q^m$ высоты m (см. [22]). Пусть

$$H = \big[\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_n\big],$$

где $\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_n$ – столбцы матрицы H. Пусть точки $P_1, P_2, \ldots, P_n$ аффинного пространства $AG(m, q)$ соответствуют столбцам $\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_n$ проверочной матрицы H, а также координатам $i_1, i_2, \ldots, i_n$ векторного пространства $\mathbb{F}_q^n$, и пусть эти соответствия являются взаимно-однозначными.

Пусть $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{F}_q^n$. *Носителем* вектора $\boldsymbol{x}$ называется множество

$$\operatorname{supp}(\boldsymbol{x}) := \{i \mid x_i \neq 0\}.$$

Далее в этом параграфе будем считать, что $q \geqslant 3$. Как было отмечено в § 2, минимальное расстояние кода $RM_q((q-1)m-2, m)$ равно 3. Кодовое слово веса 3 кода $RM_q((q-1)m-2, m)$ будем называть *тройкой*. Пусть $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)$ является тройкой, и пусть $\operatorname{supp}(\boldsymbol{c}) = \{i, j, k\}$. Тогда

- тройка $\boldsymbol{c}$ лежит на прямой ℓ, если $\{P_i, P_j, P_k\} \subseteq \ell$;
- соответствующие столбцы $\boldsymbol{h}_i, \boldsymbol{h}_j, \boldsymbol{h}_k$ линейно зависимы, т.е.

$$c_i\boldsymbol{h}_i + c_j\boldsymbol{h}_j + c_k\boldsymbol{h}_k = \boldsymbol{0}.$$

Поскольку каждая тройка кода $RM_q((q-1)m-2, m)$ принадлежит нулевому пространству проверочной матрицы H кода $RM_q((q-1)m-2, m)$, то непосредственно из определения прямой аффинного пространства $AG(m, q)$ и структуры проверочной матрицы H обобщенного кода Рида – Маллера $RM_q((q-1)m-2, m)$ следует, что любая тройка кода $RM_q((q-1)m-2, m)$ лежит на некоторой прямой аффинного пространства $AG(m, q)$. Очевидно, что число троек в коде $RM_q((q-1)m-2, m)$ кратно числу ненулевых элементов поля $\mathbb{F}_q$.

При $q = 3$ любая прямая аффинного пространства $AG(m, q)$ содержит 3 точки. Следовательно, при $q = 3$ носители троек кода $RM_q((q-1)m-2, m)$ соответствуют прямым аффинного пространства $AG(m, q)$. При $q = 3$ и $m = 2$ обобщенный код Рида – Маллера $RM_q((q-1)m-2, m)$ содержит 24 тройки, носители которых соответствуют 12 прямым аффинной плоскости $AG(2, 3)$.

Пусть $i \in \{1, \ldots, n\}$. Тогда через $\mathcal{R}_i$ обозначим подпространство, натянутое на множество всех троек кода $RM_q((q-1)m-2, m)$, которые имеют 1 в i-й координате. По определению минимальное расстояние линейного кода $\mathcal{R}_i$ равно 3.

Предложение 1. *Пусть $q \geqslant 3$, $m \geqslant 1$ и $n = q^m$. Тогда при любом $i \in \{1, 2, \ldots, n\}$ размерность линейного кода $\mathcal{R}_i$ равна $n - [m]_q - 1$, где $[m]_q$ является q-аналогом натурального числа m.*

Предложение 2. *Пусть $q \geqslant 3$, $m \geqslant 1$. Пусть $\boldsymbol{c}$ – тройка и $\boldsymbol{c} \in \mathcal{R}_i$. Тогда $\boldsymbol{c}$ лежит на прямой, проходящей через точку P_i.*

Код $RM_q((q-1)m-2, m)$ является 1-квазисовершенным кодом с радиусом покрытия $\rho = 2$ (см. [1]). Для заданного кода $RM_q((q-1)m-2, m)$ и $t = 0, 1, 2$ определим множества $RM_q^{(t)}((q-1)m-2, m)$. Положим

$$RM_q^{(t)}((q-1)m-2, m) := \left\{\boldsymbol{x} \in \mathbb{F}_q^n \mid d\big(\boldsymbol{x}, RM_q((q-1)m-2, m)\big) = t\right\}.$$

Множества $RM_q^{(t)}((q-1)m-2, m)$ называются *подкомпонентами* 1-квазисовершенного кода $RM_q((q-1)m-2, m)$.

Предложение 3. *Пусть $q \geqslant 3$, $m \geqslant 2$. Пусть*

$$\boldsymbol{x} \in RM_q^{(2)}((q-1)m-2, m), \quad \operatorname{wt}(\boldsymbol{x}) = 2, \quad \operatorname{supp}(\boldsymbol{x}) = \{j, k\}.$$

Пусть $i \neq j$, $i \neq k$ и точки P_i, P_j, P_k лежат на одной прямой. Тогда существует тройка $\boldsymbol{c} \in \mathcal{R}_i$, такая что $\operatorname{supp}(\boldsymbol{c}) = \{i, j, k\}$ и $d(\boldsymbol{c}, \boldsymbol{x}) = 2$.

Обозначим через $\boldsymbol{e}_i$ вектор длины n, все компоненты которого равны нулю, кроме i-й компоненты, которая равна 1.

Пусть $\boldsymbol{x} \in RM_q((q-1)m-2, m)$, $\lambda \in \mathbb{F}_q \setminus \{0\}$. Пусть $\mathcal{R}_i + \boldsymbol{x} + \lambda \cdot \boldsymbol{e}_i$ является сдвигом смежного класса $\mathcal{R}_i + \boldsymbol{x}$. Тогда замена смежного класса $\mathcal{R}_i + \boldsymbol{x}$ на сдвиг $\mathcal{R}_i + \boldsymbol{x} + \lambda \cdot \boldsymbol{e}_i$ называется *переключением*.

Предложение 4. *Пусть* $q \geqslant 3$, $m \geqslant 2$, $n = q^m$, $\boldsymbol{x} \in RM_q((q-1)m-2, m)$. *Тогда при любом* $i \in \{1, 2, \ldots, n\}$ *и при любом* $\lambda \in \mathbb{F}_q \setminus \{0\}$ *множество*

$$\mathcal{C}' = \big(RM_q((q-1)m-2, m) \setminus (\mathcal{R}_i + \boldsymbol{x})\big) \cup \big(\mathcal{R}_i + \boldsymbol{x} + \lambda \cdot \boldsymbol{e}_i\big)$$

является нелинейным 1*-квазисовершенным кодом с параметрами* $(n, q^{n-m-1}, 3)_q$.

Пусть L_1 и L_2 являются подпространствами пространства $\mathbb{F}_q^n$. Тогда через $L_1 + L_2$ обозначим сумму подпространств L_1 и L_2. По определению

$$L_1 + L_2 := \{\boldsymbol{x} + \boldsymbol{y} \mid \boldsymbol{x} \in L_1,\ \boldsymbol{y} \in L_2\}.$$

При $m = 1$ обобщенные коды Рида – Маллера представляют собой расширенные коды Рида – Соломона [19, с. 289]. При $m = 1$ аффинное пространство $AG(1, q)$ является прямой, и $\mathcal{R}_i$ совпадает с кодом $RM_q(q-3, 1)$. Следовательно, при $m = 1$

$$\dim(RM_q((q-1)m-2, m)) = \dim(\mathcal{R}_i) = q-2.$$

Предложение 5. *Пусть* $q \geqslant 3$, $m = 2$ *и* $n = q^m$. *Тогда при любых* $i, j \in \in \{1, 2, \ldots, n\}$, $i \neq j$, *справедливо неравенство*

$$\dim(\mathcal{R}_i \cap \mathcal{R}_j) \geqslant q(q-2) - 1.$$

Предложение 6. *Пусть* $q \geqslant 3$, $m \geqslant 2$ *и* $n = q^m$. *Тогда при любых* $i, j \in \in \{1, 2, \ldots, n\}$, $i \neq j$, *справедливо неравенство*

$$\dim(\mathcal{R}_i \cap \mathcal{R}_j) \geqslant n - [m]_q - q^{m-1}.$$

Доказательство. Пусть ℓ_{ij} – прямая, проходящая через точки P_i и P_j. При $m = 2$ на прямой ℓ_{ij} лежит $q-2$ линейно независимых элементов базиса подпространства $\mathcal{R}_i \cap \mathcal{R}_j$, а остальные элементы базиса подпространства $\mathcal{R}_i \cap \mathcal{R}_j$ принадлежат плоскости, проходящей через прямую ℓ_{ij}. В аффинном пространстве $AG(m, q)$ число плоскостей, проходящих через данную прямую, равно $\dfrac{q^{m-1}-1}{q-1}$. В силу предложения 5 при $m = 2$ и при любых $i, j \in \{1, 2, \ldots, n\}$, $i \neq j$, базис подпространства $\mathcal{R}_i \cap \mathcal{R}_j$ содержит не менее чем $q(q-2)-1$ линейно независимых векторов. Следовательно, при $m \geqslant 2$ и при любых $i, j \in \{1, 2, \ldots, n\}$, $i \neq j$, базис подпространства $\mathcal{R}_i \cap \mathcal{R}_j$ содержит не менее чем

$$\frac{q^{m-1}-1}{q-1}\big(q(q-2) - 1 - (q-2)\big) + (q-2)$$

элементов. Поскольку

$$\frac{q^{m-1}-1}{q-1}\big(q(q-2) - 1 - (q-2)\big) + (q-2) = n - [m]_q - q^{m-1},$$

то при $m \geqslant 2$ и при любых $i, j \in \{1, 2, \ldots, n\}$, $i \neq j$, имеем

$$\dim(\mathcal{R}_i \cap \mathcal{R}_j) \geqslant n - [m]_q - q^{m-1}. \quad \blacktriangle$$

§ 4. Ранг и ядро

В этом параграфе мы покажем существование нелинейных 1-квазисовершенных q-ичных кодов различных рангов и в некоторых случаях построим их, а также покажем существование и в некоторых случаях построим нелинейные 1-квазисовершенные q-ичные коды с ядрами различных размерностей. Доказательства приведены в работе [17].

Теорема 1. *Пусть* $q \geqslant 3$, $m \geqslant 2$, $n = q^m$. *Пусть* $i \in \{1, 2, \ldots, n\}$ *и* $\lambda \in \mathbb{F}_q \setminus \{0\}$. *Пусть* $\boldsymbol{x} \in RM_q((q-1)m-2, m) \setminus \mathcal{R}_i$ *и*

$$\mathcal{C}' = \big(RM_q((q-1)m-2, m) \setminus (\mathcal{R}_i + \boldsymbol{x})\big) \cup \big(\mathcal{R}_i + \boldsymbol{x} + \lambda \cdot \boldsymbol{e}_i\big).$$

Тогда множество $\mathcal{C}'$ *является нелинейным* 1*-квазисовершенным кодом с параметрами* $(n, q^{n-m-1}, 3)_q$. *При этом*

$$\operatorname{rank}(\mathcal{C}') = n - m,$$
$$\dim(\ker(\mathcal{C}')) = n - [m]_q - 1.$$

Пусть $t \in \{2, 3, \ldots, m+1\}$. Координаты $i_1, i_2, \ldots, i_t$ векторного пространства $\mathbb{F}_q^n$ будем называть *независимыми*, если соответствующие этим координатам столбцы $\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_t$ проверочной матрицы H кода $RM_q((q-1)m-2, m)$ являются линейно независимыми.

Заметим, что любая пара координат $i, j \in \{1, 2, \ldots, n\}$, $i \neq j$, является независимой, поскольку минимальное расстояние кода $RM_q((q-1)m-2, m)$ равно 3.

Предложение 7. *Пусть*
- $m \geqslant 5$ *при* $q = 3, 4$,
- $m \geqslant 4$ *при* $5 \leqslant q \leqslant 19$,
- $m \geqslant 3$ *при* $q \geqslant 23$.

Пусть $n = q^m$, $t \in \{2, 3, \ldots, m+1\}$, *и пусть* $i_1, i_2, \ldots, i_t$ *являются независимыми координатами векторного пространства* $\mathbb{F}_q^n$. *Тогда существуют кодовые слова* $\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}, \ldots, \boldsymbol{x}_{i_t}$, *такие что*

$$\boldsymbol{x}_{i_s} \in RM_q((q-1)m-2, m) \setminus \mathcal{R}_{i_s}$$

при $i_s \in \{i_1, i_2, \ldots, i_t\}$ *и*

$$d(\mathcal{R}_i + \boldsymbol{x}_i, \mathcal{R}_j + \boldsymbol{x}_j) \geqslant 5$$

при любых $i, j \in \{i_1, i_2, \ldots, i_t\}$, $i \neq j$.

Теорема 2. *Пусть*
- $m \geqslant 5$ *при* $q = 3, 4$,
- $m \geqslant 4$ *при* $5 \leqslant q \leqslant 19$,
- $m \geqslant 3$ *при* $q \geqslant 23$.

Пусть $n = q^m$, $t \in \{2, 3, \ldots, m+1\}$, *и пусть* $i_1, i_2, \ldots, i_t$ *являются независимыми координатами векторного пространства* $\mathbb{F}_q^n$. *Пусть кодовые слова*

$$\boldsymbol{x}_{i_s} \in RM_q((q-1)m-2, m) \setminus \mathcal{R}_{i_s}$$

при $i_s \in \{i_1, i_2, \ldots, i_t\}$ *таковы, что*

$$d(\mathcal{R}_i + \boldsymbol{x}_i, \mathcal{R}_j + \boldsymbol{x}_j) \geqslant 5$$

при любых $i, j \in \{i_1, i_2, \ldots, i_t\}$, $i \neq j$.

Тогда при любых $\lambda_{i_1}, \lambda_{i_2}, \ldots, \lambda_{i_t} \in \mathbb{F}_q \setminus \{0\}$ *множество*

$$\mathcal{C}_t = \left(RM_q((q-1)m-2, m) \setminus \bigcup_{s=1}^{t} (\mathcal{R}_{i_s} + \boldsymbol{x}_{i_s})\right) \cup \left(\bigcup_{s=1}^{t} (\mathcal{R}_{i_s} + \boldsymbol{x}_{i_s} + \lambda_{i_s} \cdot \boldsymbol{e}_{i_s})\right)$$

является нелинейным 1*-квазисовершенным кодом с параметрами* $(n, q^{n-m-1}, 3)_q$.

При этом

$$\operatorname{rank}(\mathcal{C}_t) = n - m - 1 + t.$$

Следствие. *Пусть*

- $m \geqslant 5$ *при* $q = 3, 4$,
- $m \geqslant 4$ *при* $5 \leqslant q \leqslant 19$,
- $m \geqslant 3$ *при* $q \geqslant 23$.

Пусть $n = q^m$. *Тогда существуют нелинейные* 1*-квазисовершенные* q*-ичные коды полного ранга длины* n.

Автор искренне признателен рецензентам за полезные замечания и предложения, которые позволили существенно улучшить первоначальный вариант статьи.

СПИСОК ЛИТЕРАТУРЫ

1. *Романов А.М.* О совершенных кодах и кодах Рида – Маллера над конечными полями // Пробл. передачи информ. 2021. Т. 57. № 3. С. 3–16. https://doi.org/10.31857/S0555292321030013
2. *Romanov A.M.* On the Number of q-ary Quasi-perfect Codes with Covering Radius 2 // Des. Codes Cryptogr. 2022. V. 90. № 8. P. 1713–1719. https://doi.org/10.1007/s10623-022-01070-y
3. *Romanov A.M.* Perfect Mixed Codes from Generalized Reed–Muller Codes // Des. Codes Cryptogr. 2024. V. 92. № 6. P. 1747–1759. https://doi.org/10.1007/s10623-024-01364-3
4. *Baicheva T., Bouyukliev I., Dodunekov S., Fack V.* Binary and Ternary Linear Quasi-Perfect Codes With Small Dimensions // IEEE Trans. Inform. Theory. 2008. V. 54. № 9. P. 4335–4339. https://doi.org/10.1109/TIT.2008.928277
5. *Etzion T., Mounits B.* Quasi-perfect Codes with Small Distance // IEEE Trans. Inf. Theory. 2005. V. 51. № 11. P. 3938–3946. https://doi.org/10.1109/TIT.2005.856944
6. *Etzion T.* Perfect Codes and Related Structures. Hackensack, NJ: World Sci., 2022.
7. *Phelps K.T., LeVan M.* Kernels of Nonlinear Hamming Codes // Des. Codes Cryptogr. 1995. V. 6. № 3. P. 247–257. https://doi.org/10.1007/BF01388478
8. *Heden O.* On Perfect p-ary Codes of Length $p + 1$ // Des. Codes Cryptogr. 2008. V. 46. № 1. P. 45–56. https://doi.org/10.1007/s10623-007-9133-y
9. *Romanov A.M.* Hamiltonicity of Minimum Distance Graphs of 1-Perfect Codes // Electron. J. Combin. 2012. V. 19. № 1. Article P65 (6 pp.). https://doi.org/10.37236/2158
10. *Romanov A.M.* Disjoint Hamiltonian Cycles in Minimum Distance Graphs of 1-Perfect Codes // Australas. J. Combin. 2017. V. 69. № 2. Part 2. P. 215–221. Available at https://ajc.maths.uq.edu.au/pdf/69/ajc_v69_p215.pdf
11. *Etzion T., Vardy A.* Perfect Binary Codes: Constructions, Properties and Enumeration // IEEE Trans. Inform. Theory. 1994. V.40. № 3. P. 754–763. https://doi.org/10.1109/18.335887
12. *Phelps K.T., Villanueva M.* Ranks of q-ary 1-Perfect Codes // Des. Codes Cryptogr. 2002. V. 27. № 1–2. P. 139–144. https://doi.org/10.1023/A:1016510804974
13. *Phelps K.T., Rifà J., Villanueva M.* Rank and Kernel of Binary Hadamard Codes // IEEE Trans. Inform. Theory. 2005. V. 51. № 11. P. 3931–3937. https://doi.org/10.1109/TIT.2005.856940
14. *Li X., Shi M., Wang S., Lu H., Zheng Y.* Rank and Pairs of Rank and Dimension of Kernel of $\mathbb{Z}_p\mathbb{Z}_{p^2}$-Linear Codes // IEEE Trans. Inform. Theory. 2024. V. 70. № 5. P. 3202–3212. https://doi.org/10.1109/TIT.2023.3317064
15. *Borges J., Phelps K.T., Rifà J.* The Rank and Kernel of Extended 1-Perfect $\mathbb{Z}_4$-Linear and Additive Non-$\mathbb{Z}_4$-Linear Codes // IEEE Trans. Inform. Theory. 2003. V. 49. № 8. P. 2028–2034. https://doi.org/10.1109/TIT.2003.814490
16. *Васильев Ю.Л.* О негрупповых плотно упакованных кодах // Проблемы кибернетики. Т. 8. М.: Физматлит, 1962. С. 337–339.
17. *Romanov A.M.* On Nonlinear 1-Quasi-perfect Codes and Their Structural Properties // Probl. Inf. Transm. 2024. V. 60. № 3. P. 141–154. https://doi.org/10.1134/S0032946024030013

18. *Batten L.M.* Combinatorics of Finite Geometries, Cambridge: Cambridge Univ. Press, 1997.
19. *Мак-Вильямс Ф.Дж., Слоэн Н.Дж.А.* Теория кодов, исправляющих ошибки. М.: Связь, 1979.
20. *Kasami T., Lin S., Peterson W.* New Generalizations of the Reed–Muller Codes. I: Primitive Codes // IEEE Trans. Inform. Theory. 1968. V. 14. № 2. P. 189–199. https://doi.org/10.1109/TIT.1968.1054127
21. *Assmus E.F., Key J.D.* Polynomial Codes and Finite Geometries // Handbook of Coding Theory. Amsterdam: Elsevier, 1998. Vol. II. P. 1269–1344.
22. *Delsarte P., Goethals J.-M., MacWilliams F.J.* On Generalized Reed–Muller Codes and Their Relatives // Inform. Control. 1970. V. 16. № 5. P. 403–442. https://doi.org/10.1016/S0019-9958(70)90214-7
23. *Августинович С.В., Соловъева Ф.И.* О несистематических совершенных двоичных кодах // Пробл. передачи информ. 1996. Т. 32. № 3. С. 47–50. https://www.mathnet.ru/rus/ppi344

Романов Александр Михайлович
Институт математики им. С.Л. Соболева СО РАН,
Новосибирск
rom@math.nsc.ru

Поступила в редакцию
03.06.2024
После доработки
07.11.2024
Принята к публикации
16.11.2024

ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 60 **2024** **Вып. 3**

УДК 621.391 : 519.725

© 2024 г. **Ц. Байчева**[1], **С. Топалова**[2]

НОВЫЕ РЕЗУЛЬТАТЫ ОБ ОПТИМАЛЬНЫХ ДВОИЧНЫХ ЦИКЛИЧЕСКИ ПЕРЕСТАНОВОЧНЫХ РАВНОВЕСНЫХ $(v, 4, 1)$-КОДАХ

Построены новые двоичные циклически перестановочные равновесные коды (ЦПР-коды) с параметрами $(v, 4, 1)$ для длин $v \leqslant 136$, а также исправлено несколько табличных значений из работы авторов [1].

Ключевые слова: двоичный циклически перестановочный равновесный код, оптический ортогональный код, циклическая схема, классификация, автоморфизмы циклической группы.

DOI: 10.31857/S0555292324030021, **EDN:** PDKCCX

§ 1. Введение

Начнем с обозначений для рассматриваемых комбинаторных структур. Более подробную информацию об отношениях между ними можно найти в [1–3]. Обозначим через $\mathbb{Z}_v$ аддитивную группу целых чисел по модулю v.

Пусть $V = \{P_i\}_{i=1}^{v}$ – конечное множество *точек*, а $\mathcal{B} = \{B_j\}_{j=1}^{b}$ – конечный набор k-элементных подмножеств множества V, называемых *блоками*. Тогда $D = (V, \mathcal{B})$ называется *схемой* (или *дизайном*) с параметрами 2-$(v, k, 1)$ (системой Штейнера $S(2, k, v)$), если любое подмножество размера 2 множества V содержится ровно в одном блоке из $\mathcal{B}$. Две 2-$(v, k, 1)$-схемы D и D' *изоморфны*, если существует перестановка точек, переводящая каждый блок из D в блок из D'. Схема 2-$(v, k, 1)$ называется *циклической*, если у нее имеется автоморфизм α_v, переставляющий ее точки в одном цикле, и *строго циклической*, если орбита каждого блока под действием этого автоморфизма имеет длину v (нет коротких орбит). Две циклические 2-$(v, k, 1)$-схемы D и D' *мультипликаторно эквивалентны*, если существует автоморфизм $\mathbb{Z}_v$, отображающий каждый блок из D в блок из D'.

Пусть $M = \{m_1, m_2, \ldots, m_k\}$ – некоторое подмножество размера k аддитивной группы G. Тогда через $M + t$ обозначим t-сдвиг множества M, т.е.

$$M + t = \{m_1 + t, m_2 + t, \ldots, m_k + t\},$$

где $t \in G$.

Разностным $(v, k, 1)$-семейством называется множество $\mathcal{D} = \{M_1, \ldots, M_s\}$, где $M_i = \{m_{i_1}, m_{i_2}, \ldots, m_{i_k}\}$ – k-элементные подмножества аддитивной группы G порядка v, такие что любой ненулевой элемент группы G единственным образом представляется в виде некоторой разности $m_{i_j} - m_{i_\ell}$ для $1 \leqslant i \leqslant s$ и $1 \leqslant j \neq \ell \leqslant k$. Два

[1] Работа выполнена при частичной финансовой поддержке Министерства образования и науки Республики Болгария (номер гранта D01-325/01.12.2023).

[2] Работа выполнена при частичной финансовой поддержке Национального научного фонда Болгарии (номер контракта КР-06-Н62/2/13.12.2022).

разностных семейства над G эквивалентны, если существует автоморфизм α группы G, переводящий каждое k-элементное подмножество первого семейства в некоторый сдвиг подмножества из второго семейства. Если G – циклическая группа, то и разностное семейство циклическое. В дальнейшем мы будем рассматривать циклические разностные семейства для $G = \mathbb{Z}_v$.

Двоичный циклически перестановочный равновесный код (ЦПР-код) с параметрами $(v, k, 1)$ (или оптический ортогональный $(v, k, 1)$-код) можно определить как набор $\mathcal{C} = \{C_1, \ldots, C_s\}$, состоящий из k-подмножеств группы $\mathbb{Z}_v$ (*кодовых слов*), таких что любые два сдвига одного кодового слова имеют не более одного общего элемента и любые два сдвига двух различных кодовых слов также имеют не более одного общего элемента:

$$|C_i \cap (C_i + t)| \leqslant 1, \quad 1 \leqslant i \leqslant s, \quad 1 \leqslant t \leqslant v-1, \tag{1}$$

$$|C_i \cap (C_j + t)| \leqslant 1, \quad 1 \leqslant i < j \leqslant s, \quad 0 \leqslant t \leqslant v-1. \tag{2}$$

Первое из этих свойств называется свойством автокорреляции, а второе – свойством кросс-корреляции.

Два $(v, k, 1)$-ЦПР-кода C и C' *изоморфны*, если существует перестановка φ группы $\mathbb{Z}_v$, переводящая набор сдвигов каждого кодового слова из C в набор сдвигов кодового слова из C'. Эти два кода *мультипликаторно эквивалентны*, если φ является автоморфизмом группы $\mathbb{Z}_v$.

Рассмотрим кодовое слово $C = \{c_1, c_2, \ldots, c_k\}$. Обозначим через ΔC мультимножество, состоящее из значений разностей $c_i - c_j$, $i \neq j$, $i, j = 1, 2, \ldots, k$. Свойство автокорреляции означает, что все разности кодовых слов $(v, k, 1)$-ЦПР-кода различны. Свойство кросс-корреляции означает, что $\Delta C_1 \cap \Delta C_2 = \varnothing$ для любых двух кодовых слов C_1 и C_2. Если все возможные $v-1$ ненулевых разностей покрываются разностями кодовых слов, то $(v, k, 1)$-ЦПР-код называется *совершенным* и соответствует циклическому $(v, k, 1)$-разностному семейству и циклической 2-$(v, k, 1)$-схеме. Блоки этой схемы соответствуют кодовым словам и их сдвигам. Эта схема является строго циклической. Каждое кодовое слово со своими сдвигами соответствуют одной блоковой орбите длины v под действием α_v.

Для $(v, k, 1)$-ЦПР-кода его мощность s – это число его кодовых слов. Она не может превышать

$$\left\lfloor \frac{v-1}{k(k-1)} \right\rfloor.$$

ЦПР-коды, достигающие этой границы, называются оптимальными. Если мощность кода в точности равна $(v-1)/k(k-1)$, то этот $(v, k, 1)$-ЦПР-код совершенный. Тем самым, оптимальные $(v, 4, 1)$-ЦПР-коды совершенны тогда и только тогда, когда $v = 12n + 1$.

ЦПР-коды имеют различные применения [4–6] и широко изучены (см., например, [2, 7–10]). Существование оптимальных $(v, 4, 1)$-ЦПР-кодов рассматривалось в [11–17]. Известно, что оптимальные $(v, 4, 1)$-ЦПР-коды существуют для всех рассматриваемых нами длин. Нам не известны какие-либо результаты классификации для $(v, 4, 1)$-ЦПР-кодов или циклических схем при $v > 76$. Системы $S(2, 4, v)$ представляют собой особенно интересный класс систем Штейнера, им было посвящено множество исследований [18], и в частности, изучались циклические $S(2, 4, v)$-системы [19].

Недавно на портале `arXiv` была опубликована небольшая заметка [20]. В ней было получено несколько новых результатов о разностных семействах. Ее автор также проверил некоторые уже известные результаты из [1] и [21]. Для всех проверенных случаев он получил те же самые значения, за исключением разностных семейств

с параметрами $(73, 4, 1)$. Для них он получил 1428546 не эквивалентных разностных $(73, 4, 1)$-семейств, что больше, чем число 1426986 таких семейств, полученное в [1]. Мы повторили наши вычисления, и оказалось, что значение, приведенное в [20], является верным. Неверный результат в работе [1] связан с ошибкой в программе, которая использовалась для классификации в этой работе, он не имеет отношения к алгоритму и не затрагивает наших результатов классификации для ЦПР-кодов (оптических ортогональных кодов) с другими параметрами, опубликованных в других работах. Повторение всех вычислений из [1] на компьютере с частотой 3,2 ГГц не заняло много времени, и поэтому мы решили также рассмотреть и бóльшие длины.

В настоящей статье мы вначале корректируем результаты, полученные в [1] для $(v, 4, 1)$-ЦПР-кодов с длинами 41, 46, 53, 58, 71 и 73. Затем мы приводим классификацию с точностью до изоморфизма оптимальных $(85, 4, 1)$-ЦПР-кодов и циклических $2 - (85, 4, 1)$- и $2 - (88, 4, 1)$-схем (соответствующих неэквивалентным циклическим разностным $(85, 4, 1)$- и $(88, 4, 1)$-семействам). Наконец, получено несколько тысяч оптимальных $(v, 4, 1)$-ЦПР-кодов с длинами $76 < v \leqslant 136$. В § 2 представлены исправления к [1], в § 3 – новые результаты и алгоритмы, использованные для $76 < v \leqslant 136$, а в § 4 дано краткое заключение и описание открытых вопросов. Файлы со всеми построенными нами $(v, 4, 1)$-ЦПР-кодами можно загрузить с сайта `http://www.moi.math.bas.bg/~tsonka/`, а также `https://zenodo.org/records/13771464`. Расширенное изложение результатов настоящей статьи содержится в [22].

§ 2. Исправления к работе [1]

Значения числа мультипликаторно неэквивалентных оптимальных $(v, 4, 1)$-ЦПР-кодов с длинами 41, 46, 53, 58, 71, 73 и 74 отличаются от значений, указанных в [1]. Правильные значения приведены в табл. 1.

Таблица 1

Уточненные количества оптимальных $(v, 4, 1)$-ЦПР-кодов из [1]

Длина	Мощность	Число кодов
41	3	340
46	3	10016
53	4	28680
58	4	1013340

Длина	Мощность	Число кодов
71	5	425832736
73	6	1428546
74	6	2148852

§ 3. $(v, 4, 1)$-ЦПР-коды для $76 < v \leqslant 136$

3.1. Об основном алгоритме и справедливости компьютерных результатов. Вопрос о надежности наших компьютерных программ закономерно возникает после того как были обнаружены ошибки в наших предыдущих вычислениях. Никогда нельзя быть абсолютно уверенным в том, что какая-либо сложная программа работает правильно во всех случаях, потому что обычно какие-то очень мелкие детали могут приводить к ошибкам. Чтобы проверить надежность нашей программы, мы всегда тестировали ее на всех ранее известных результатах. Для [1] это были результаты классификации для циклических $(v, 4, 1)$-схем с $v \leqslant 64$ [19], которые совпали с нашими. Позже мы усовершенствовали эту программу для построения оптимальных $(v, k, 1)$-ЦПР-кодов для любого k (не только для $k = 4$) и протестировали ее на всех известных результатах классификации, а именно на циклических $(v, 3, 1)$-схемах с $v \leqslant 57$ [23] и циклических $(v, 5, 1)$-схемах с $v \leqslant 65$ [19]. Когда в [20] появилась информация об ошибке в нашем результате для $(73, 4, 1)$, мы повторили

с помощью этой более общей программы все вычисления из [1]; полученные при этом результаты приведены в предыдущем параграфе. Также в [20] отмечено, что за исключением случая $(73, 4, 1)$ алгоритм, описанный в [20], во всех других случаях, представленных в [21], дал то же самое число неэквивалентных циклических разностных семейств.

Для классификации оптимальных ЦПР-кодов и циклических схем с точностью до мультипликаторной эквивалентности используется обратный поиск с проверкой на минимальность некоторых частичных решений. Перед началом поиска строятся все возможные кодовые слова (блоки орбит по действию α_v, если строятся схемы), которые упорядочиваются как в лексикографическом порядке, так и по действию автоморфизмов циклической группы порядка v. Это позволяет задать лексикографический порядок и на решениях, а также с легкостью проверить с помощью теста на минимальность, можно ли с помощью какого-либо автоморфизма циклической группы порядка v перевести решение в лексикографически меньшее. Если можно, то решение отбрасывается. Таким образом, строятся только мультипликаторно неэквивалентные коды (циклические схемы). Подробности можно найти в [1], а в [22] приведен небольшой пример, иллюстрирующий работу алгоритма.

3.2. Классификация циклических 2-$(85, 4, 1)$- и 2-$(88, 4, 1)$-схем с точностью до изоморфизма. Вначале была проведена классификация совершенных $(85, 4, 1)$-ЦПР-кодов (циклических 2-$(85, 4, 1)$-схем) и циклических 2-$(88, 4, 1)$-схем с точностью до мультипликаторной эквивалентности с помощью параллельной версии основного алгоритма. Каждая из этих двух задач выполнялась в 96 процессах на высокопроизводительном компьютере "Авитохол" Болгарской академии наук (подробнее см. благодарности в конце статьи), и эта задача была выполнена за четыре дня. Таким образом мы выяснили, что с точностью до мультипликаторной эквивалентности имеется 228406824 совершенных $(85, 4, 1)$-ЦПР-кода и 149494720 2-$(88, 4, 1)$-схем. Это также количества неэквивалентных разностных $(85, 4, 1)$- и $(88, 4, 1)$-семейств.

Две циклические комбинаторные структуры на $v = p{\cdot}q$ точках (где p и q – различные простые числа) изоморфны тогда и только тогда, когда они мультипликаторно эквивалентны [24]. Поэтому все эти 228406824 циклические 2-$(85, 4, 1)$-схемы не изоморфны. Две циклические схемы на $v = 88$ точках, однако, могут быть мультипликаторно не эквивалентны, но изоморфны, если их полная группа автоморфизмов имеет порядок выше, чем 88. С помощью "Авитохола" мы вычислили порядки групп автоморфизмов всех 149494720 схем и установили, что все они имеют порядок 88. Это означает, что эти схемы не изоморфны. Компьютерное время, необходимое для нахождения групп автоморфизмов схем, оказалось в 5 раз выше времени их построения.

3.3. Оптимальные ЦПР-коды для $76 < v \leqslant 136$. Обычно, когда программа пытается получить слишком много результатов или требует слишком много времени для выполнения поставленной задачи, ее просто прерывают и пытаются использовать результаты, полученные к этому моменту. Однако полученные таким образом коды, как правило, оказывались очень похожими друг на друга – они отличались лишь несколькими последними кодовыми словами. Поэтому здесь мы использовали другой подход, немного модифицировав основной алгоритм. Если найдено частичное решение из m кодовых слов, то алгоритм классификации пытается дополнить его до частичного решения из $m + 1$ кодовых слов всеми возможными способами, а модифицированный алгоритм пытается дополнить частичное решение из m кодовых слов $(m > 1)$ до частичного решения из $m + 1$ кодовых слов не всеми, а лишь не более чем e возможными способами.

Тест на минимальность отвергает код C, если он переводится некоторым автоморфизмом в лексикографически меньший код C'. При этом мы должны быть уверены, что код C' уже построен. Поэтому когда для $(m + 1)$-го кодового слова существует

более e возможностей, используются только e лексикографически наименьших. Тем самым, для $m > 1$ рассматривается не более e ветвей узлов дерева поиска. Таким образом строятся коды со всеми возможностями для первого и второго кодового слова, и каждый из полученных кодов отличается от большинства других кодов по многим кодовым словам. Построенные коды доступны онлайн.

§ 4. Заключение и открытые вопросы

Пересмотр нашей работы 2011 года показал, что более мощные параллельные компьютеры, доступные в настоящее время, позволяют получать результаты классификации для бóльших длин, но чрезвычайно большое количество кодов делает эти результаты сложными для использования. Поэтому мы пытались строить только "репрезентативную" часть кодов, а именно коды, существенно отличающиеся друг от друга. Мы надеемся, что онлайн-доступность построенных оптимальных $(v, 4, 1)$-ЦПР-кодов сделает их полезными как для приложений, так и для будущих исследований. Другой подход может заключаться в построении только тех кодов, которые обладают определенными свойствами, например, группами автоморфизмов, или свойствами, важными для того или иного конкретного приложения. Задачи полной классификации циклических разностных $(v, 4, 1)$-семейств с $v > 88$ и классификация оптимальных $(v, 4, 1)$-ЦПР-кодов с $v > 76$ и $v \neq 85$ остаются открытыми.

Авторы благодарны Ивану Гетману за оперативное информирование о его результатах, представленных в [20].

Исследования, которые привели к данным результатам, проводились с использованием инфраструктуры, приобретенной в рамках Национальной дорожной карты для исследовательской инфраструктуры, финансовую координацию которой осуществляет Министерство образования и науки Республики Болгария (грант № D01-325/01.12.2023).

СПИСОК ЛИТЕРАТУРЫ

1. *Байчева Ц., Топалова С.* Классификация оптимальных двоичных циклически перестановочных равновесных $(v, 4, 1)$-кодов и циклических 2-$(v, 4, 1)$-дизайнов для $v \leqslant 76$ // Пробл. передачи информ. 2011. V. 47. № 3. P. 10–18. https://www.mathnet.ru/rus/ppi2051
2. *Moreno O., Zhang Z., Kumar P.V., Zinoviev V.A.* New Constructions of Optimal Cyclically Permutable Constant Weight Codes // IEEE Trans. Inform. Theory. 1995. V. 41. № 2. P. 448–455. https://doi.org/10.1109/18.370146
3. *Abel R.J.R., Buratti M.* Difference Families // Handbook of Combinatorial Designs. Boca Raton: Chapman & Hall/CRC, 2007. Sec. VI.16. P. 392–410.
4. *Chung F.R.K., Salehi J.A., Wei V.K.* Optical Orthogonal Codes: Design, Analysis, and Applications // IEEE Trans. Inform. Theory. 1989. V. 35. № 3. P. 595–604. https://doi.org/10.1109/18.30982
5. *Bird I.C.M., Keedwell A.D.* Design and Applications of Optical Orthogonal Codes—A Survey // Bull. Inst. Combin. Appl. 1994. V. 11. P. 21–44.
6. *Colbourn C.J., Dinitz J.H., Stinson D.R.* Applications of Combinatorial Designs to Communications, Cryptography, and Networking // Surveys in Combinatorics, 1999. Cambridge: Cambridge Univ. Press, 1999. P. 37–100.
7. *Nguyen Q.A., Gyöfri L., Massey J.L.* Constructions of Binary Constant-Weight Cyclic Codes and Cyclically Permutable Codes // IEEE Trans. Inform. Theory. 1992. V. 38. № 3. P. 940–949. https://doi.org/10.1109/18.135636
8. *Bitan S., Etzion T.* Constructions for Optimal Constant Weight Cyclically Permutable Codes and Difference Families // IEEE Trans. Inform. Theory. 1995. V. 41. № 1. P. 77–87. https://doi.org/10.1109/18.370117

9. *Fuji-Hara R., Miao Y.* Optical Orthogonal Codes: Their Bounds and New Optimal Constructions // IEEE Trans. Inform. Theory. 2000. V. 46. № 7. P. 2396–2406. https://doi.org/10.1109/18.887852
10. *Baicheva T., Topalova S.* Classification of Optimal $(v, k, 1)$ Binary Cyclically Permutable Constant Weight Codes with $k = 5$, 6 and 7 and Small Lengths // Des. Codes Cryptogr. 2019. V. 87. № 2–3. P. 365–374. https://doi.org/10.1007/s10623-018-0534-x
11. *Brickell E.F., Wei V.K.* Optical Orthogonal Codes and Cyclic Block Designs // Congr. Numer. 1987. V. 58. P. 175–182.
12. *Chen K., Zhu L.* Existence of $(q, k, 1)$ Difference Families with q a Prime Power and $k = 4, 5$ // J. Combin. Des. 1999. V. 7. № 1. P. 21–30. https://doi.org/10.1002/(SICI)1520-6610(1999)7:1<21::AID-JCD4>3.0.CO;2-Y
13. *Buratti M.* Cyclic Designs with Block Size 4 and Related Optimal Optical Orthogonal Codes // Des. Codes Cryptogr. 2002. V. 26. № 1–3. P. 111–125. https://doi.org/10.1023/A:1016505309092
14. *Chang Y., Fuji-Hara R., Miao Y.* Combinatorial Constructions of Optimal Optical Orthogonal Codes with Weight 4 // IEEE Trans. Inform. Theory. 2003. V. 49. № 5. P. 1283–1292. https://doi.org/10.1109/TIT.2003.810628
15. *Abel R.J.R., Buratti M.* Some Progress on $(v, 4, 1)$ Difference Families and Optical Orthogonal Codes // J. Combin. Theory Ser. A. 2004. V. 106. № 1. P. 59–75. https://doi.org/10.1016/j.jcta.2004.01.003
16. *Buratti M., Pasotti A.* Further Progress on Difference Families with Block Size 4 or 5 // Des. Codes Cryptogr. 2010. V. 56. № 1. P. 1–20. https://doi.org/10.1007/s10623-009-9335-6
17. *Wang X., Chang Y.* Further Results on $(v, 4, 1)$-Perfect Difference Families // Discrete Math. 2010. V. 310. № 13-14. P. 1995–2006. https://doi.org/10.1016/j.disc.2010.03.017
18. *Reid C., Rosa A.* Steiner Systems $S(2, 4, v)$—A Survey // Electron. J. Combin., Dynamic Survey DS18. 2010. https://doi.org/10.37236/39
19. *Colbourn M.J., Mathon R.A.* On Cyclic Steiner 2-Designs // Ann. Discrete Math. 1980. V. 7. P. 215–253. https://doi.org/10.1016/S0167-5060(08)70182-1
20. *Hetman I.* Steiner Systems $S(2, 6, 121/126)$ Based on Difference Families, https://arxiv.org/abs/2401.08274 [math.CO], 2024.
21. *Baicheva T., Topalova S.* Classification Results for $(v, k, 1)$ Cyclic Difference Families with Small Parameters // Mathematics of Distances and Applications. Sofia: ITHEA, 2012. P. 24–30. Available at http://www.foibg.com/ibs_isc/ibs-25/ibs-25-p02.pdf.
22. *Baicheva T., Topalova S.* An Update on Optimal $(v, 4, 1)$ Binary Cyclically Permutable Constant Weight Codes and Cyclic 2-$(v, 4, 1)$ Designs with Small v // Probl. Inf. Transm. 2024. V. 60. № 3. P. 189–198. https://doi.org/10.1134/S0032946024030037
23. *Colbourn C.J., Rosa A.* Triple Systems, Oxford: Clarendon; New York: Oxford Univ. Press, 1999.
24. *Pálfy P.* Isomorphism Problem for Relational Structures with a Cyclic Automorphism // European J. Combin. 1987. V. 8. № 1. P. 35–43. https://doi.org/10.1016/S0195-6698(87)80018-5

Байчева Цонка
Топалова Светлана
Институт математики и информатики
Болгарской академии наук, София, Болгария
tsonka@math.bas.bg
svetlana@math.bas.bg

Поступила в редакцию
11.07.2024
После доработки
25.09.2024
Принята к публикации
07.10.2024

ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 60 2024 Вып. 3

УДК 621.391 : 519.72

 И.А. Мельников, А.Ю. Угловский, А.А. Крещук, А.А. Куреев, Е.М. Хоров

ИСПОЛЬЗОВАНИЕ МЕТОДА ИНФОРМАЦИОННОГО СЖАТИЯ ДЛЯ СНИЖЕНИЯ СЛОЖНОСТИ ДЕКОДЕРА МПП-КОДОВ[1]

Одним из способов снижения сложности алгоритмов распространения доверия для декодирования кодов с малой плотностью проверок на четность является хранение предварительно вычисленной суммы сообщений в узлах переменных. В свою очередь, объем обрабатываемой информации может быть значительно снижен с помощью метода информационного сжатия (МИС), снижающего разрядность всех обновляемых сообщений. Предлагается алгоритм построения бинарной функции на основе МИС, соответствующей вычитанию. Использование разработанной функции позволяет уменьшить количество хранимых и используемых таблиц поиска для узлов переменных.

Ключевые слова: Алгоритм распространения доверия, итеративное декодирование, коды с малой плотностью проверок на четность (МПП), метод информационного сжатия, последовательное декодирование.

DOI: 10.31857/S0555292324030033, **EDN:** QUHUCD

§ 1. Введение

Коды с малой плотностью проверок на четность (МПП) являются одними из самых широко используемых помехоустойчивых кодов в системах связи [1]. МПП-декодеры могут использовать такие подходы как алгоритм распространения доверия (РД, англ.: belief propagation) [2] и метод суммы наименьших значений (СН, англ.: min-sum) [3], которые итеративно обновляют логарифм отношения правдоподобия (ЛОП, англ.: log-likelihood ratios).

К сожалению, большим недостатком РД и СН является их высокая вычислительная сложность и большая разрядность ЛОП. Одним из методов снижения вычислительной сложности является хранение суммы нескольких сообщений ЛОП и последующее вычитание каждого сообщения ЛОП. В свою очередь, для снижения объема распространяемой информации можно использовать метод информационного сжатия (МИС, англ.: information bottleneck) [4, 5]. Этот метод использует специальные дискретные функции, определяемые таблицами поиска (англ.: lookup tables), которые заменяют обычные арифметические операции [6], например, сложение. Эти таблицы строятся с помощью метода дискретной эволюции плотности (ДЭП, англ.: discrete density evolution) [7]. МИС приводит к неравномерному квантованию ЛОП, и следовательно, полученные функции не являются коммутативными и ассоциативными. В результате вычитание, используемое для снижения вычислительной сложности, становится едва возможным.

[1] Исследование осуществлено в рамках Госзадания № FFNU-2022-0035 ИППИ РАН.

Настоящая статья посвящена решению данной проблемы и предлагает метод построения дискретной двоичной функции для построения операции вычитания. Численные результаты демонстрируют сокращение вдвое количества операций в МИС-декодере без значительной потери производительности.

Статья устроена следующим образом. В § 2 представлено описание МИС. В § 3 предлагается алгоритм построения дискретной двоичной функции вычитания. Численные результаты и оценка производительности приводятся в § 4. Наконец, § 5 завершает статью.

§ 2. Метод информационного сжатия

Для снижения размерности используемых ЛОП может применяться метод информационного сжатия. Этот метод позволяет найти компромисс между сжатием наблюдаемой случайной величины Y и информацией об истинной случайной величине X, при заданном совместном распределении вероятностей $p(x, y)$ величин X и Y. Такой подход вводит случайную величину T, такую что релевантная информация $I(T; X)$ велика, но информация сжатия $I(T; Y)$ мала. Пусть x, y и t являются элементами из пространств событий случайных величин X, Y и T соответственно, а $p(t \mid y)$ определяет отображение Y в T, т. е. $p(t \mid y) = 1$, если t является сжатым представлением y, и $p(t \mid y) = 0$ в противном случае.

Цель метода информационного сжатия состоит в том, чтобы найти такое распределение $p(t \mid y)$, которое минимизирует следующий функционал L:

$$L[p(t \mid y)] = I(T; Y) - \beta I(T; X) \to \min, \tag{1}$$

где β определяет компромисс между $I(T; X)$ и $I(T; Y)$. Если $\beta = 0$, то сжатие будет максимальным, а $\beta \to \infty$ указывает на максимальное сохранение релевантной информации [8]. Глобально оптимальное решение (1) можно найти с помощью алгоритма динамического программирования для метода информационного сжатия [9].

В контексте задачи декодирования случайная величина X представляет собой сообщение на входе канала, наблюдаемая случайная величина Y – сообщение на выходе канала, а случайная величина T – сжатое представление Y. При декодировании ЛОП дискретизируются, а стандартные операции суммирования заменяются операциями, описываемыми таблицами поиска, которые возвращают сжатую сумму $t_\ell = y_\ell \oplus t_{\ell-1}$ между ЛОП y_ℓ и частичной сжатой суммы $t_{\ell-1}$ предыдущих $\ell - 1$ ЛОП, где $\ell > 1$, $t_1 = y_1$. Другими словами, таблица поиска ℓ принимает на вход y_ℓ и $t_{\ell-1}$ и возвращает t_ℓ.

Таблицы поиска строятся последовательно для каждого ℓ, начиная с $\ell = 2$. На каждом шаге таблица поиска строится путем решения задачи метода информационного сжатия для распределения

$$p(y_\ell, t_{\ell-1} \mid x_1 = \ldots = x_\ell),$$

где x_ℓ равно 0 или 1. Стоит отметить, что $x_1 = \ldots = x_\ell$, потому что наблюдаемые величины соответствуют одному и тому же биту входных данных канала.

Используя таблицы поиска, можно представить работу алгоритма распространения доверия следующим образом. Алгоритм распространения доверия присваивает значения ЛОП $P_{\boldsymbol{v}}$ на выходе канала каждому узлу переменных $\boldsymbol{v} \in V$ и два значения ЛОП $Q_{\boldsymbol{vc}}$ и $R_{\boldsymbol{cv}}$ ребрам $(\boldsymbol{v}, \boldsymbol{c})$, где $\boldsymbol{c} \in C$ и $\boldsymbol{v} \in V$. При этом $Q_{\boldsymbol{vc}}$ является сообщением, отправляемым от $\boldsymbol{v}$ к $\boldsymbol{c}$, а $R_{\boldsymbol{cv}}$ – сообщением, отправляемым в противоположном направлении. Затем алгоритм рекурсивно обновляет $Q_{\boldsymbol{vc}}$ и $R_{\boldsymbol{cv}}$, учитывая только

соседние узлы:

$$
\begin{aligned}
Q_{\boldsymbol{vc}} &\leftarrow P_{\boldsymbol{v}} \oplus \left(\bigoplus_{\boldsymbol{c}' \in N(\boldsymbol{v}) \setminus \boldsymbol{c}} R_{\boldsymbol{c}'\boldsymbol{v}} \right), \\
R_{\boldsymbol{cv}} &\leftarrow -s_{\boldsymbol{cv}} \left(\bigoplus_{\boldsymbol{v}' \in N(\boldsymbol{c}) \setminus \boldsymbol{v}} Q_{\boldsymbol{v}'\boldsymbol{c}} \right),
\end{aligned}
\tag{2}
$$

где $N(k)$ – множество узлов соседей узла k, а

$$
s_{\boldsymbol{cv}} = \prod_{\boldsymbol{v}' \in N(\boldsymbol{c}) \setminus \boldsymbol{v}} \operatorname{sign}(Q_{\boldsymbol{v}'\boldsymbol{c}}).
$$

§ 3. Вычитание с помощью метода информационного сжатия

Определим функцию $\ominus$, которая вычитает одно значение ЛОП из общей суммы ЛОП, и реализуем ее, используя таблицы поиска. Для этого перепишем первое уравнение (2) следующим образом:

$$
\begin{aligned}
&A_{\boldsymbol{v}} \leftarrow P_{\boldsymbol{v}} \oplus \left(\bigoplus_{\boldsymbol{c}' \in N(\boldsymbol{v})} R_{\boldsymbol{c}'\boldsymbol{v}} \right), \\
&Q_{\boldsymbol{vc}} \leftarrow A_{\boldsymbol{v}} \ominus R_{\boldsymbol{cv}}.
\end{aligned}
\tag{3}
$$

Таблица поиска для операции вычитания $\ominus$ наблюдаемой величины Y из частичной суммы строится с помощью метода информационного сжатия для распределения

$$
p(t, y_j \,|\, x_i = b,\, i \neq j),
$$

где t – сумма ЛОП по множеству узлов N, x_i – i-й бит на входе канала, j – индекс вычитаемого входа, b равно 0 или 1. Пусть $s_j = t \ominus y_j$. Для простоты далее примем $j = 1$. Аналогично $I(T; X)$ в (1) рассмотрим следующую взаимную информацию:

$$
I(s_1; x_i = b) = \sum_{(s_1, b)} P(s_1, b \,|\, x_i = b) \log \frac{P(s_1, b \,|\, x_i = b)}{P(s_1) P(b \,|\, x_i = b)},
\tag{4}
$$

где $i \neq 1$. Тогда в случае использования вычитания в декодере и с параллельным расписанием (ПрР) имеем

$$
P(s_1, b \,|\, x_i = b) = \sum_{\left(\bigoplus_{k=1}^{n} y_k \right) \ominus y_1 = s_1} P(y_1) P(b \,|\, x_i = b) \prod_{r=2}^{n} P(y_r, b \,|\, x_r = b).
\tag{5}
$$

Остаточная информация $I(s_1; x_i = b)$ показывает объем информации о "невычтенных" битах, оставшийся после вычисления общей суммы и вычитания первого входа преобразования. Она является релевантной информацией, которую необходимо максимизировать.

$I(s_1; x_1 = b)$ – взаимная информация между значениями, которые должны быть независимыми в методе дискретной эволюции плотности. Она соответствует информации о первом вычитаемом бите после его вычитания. Для $i = 1$ имеем

$$
I(s_1; x_1 = b) = \sum_{(s_1, b)} P(s_1, b \,|\, x_1 = b) \log \frac{P(s_1, b \,|\, x_1 = b)}{P(s_1) P(b \,|\, x_1 = b)}.
\tag{6}
$$

В случае ПрР получим

$$P(s_1, b \,|\, x_1 = b) = \sum_{\left(\bigoplus\limits_{k=1}^{n} y_k\right) \ominus y_1 = s_1} P(y_1, b \,|\, x_1 = b) \prod_{r=2}^{n} P(y_r).$$

Для оценки предложенного метода также рассматривается взаимная информация $I(a_1; x_i = b)$, $i \neq 1$, которая соответствует информации $I(s_1; x_i = b)$ в базовом методе, где a_1 – результат суммирования всех чисел, кроме первого:

$$I(a_1; x_i = b) = \sum_{(a_1, b)} P(a_1, b \,|\, x_i = b) \log \frac{P(a_1, b \,|\, x_i = b)}{P(a_1) P(b \,|\, x_i = b)}. \tag{7}$$

В случае ПрР $P(a_1, b \,|\, x_i = b)$ может быть выражено следующим образом:

$$P(a_1, b \,|\, x_i = b) = \sum_{\bigoplus\limits_{k=2}^{n} y_k = a_1} P(b \,|\, x_i = b) \prod_{r=2}^{n} P(y_r, b \,|\, x_r = b).$$

В лучшем случае значения, получаемые с помощью (3), должны быть равны значениям, получаемым с помощью (2). Однако из-за сжатия эти значения могут отличаться. Близость (4) и (7) показывает, насколько метод вычитания близок к сумме без одного, а близость (6) к нулю показывает, насколько успешно информация о вычтенном бите удалена. Все вышесказанное приводит к задаче минимизации следующего функционала:

$$L[p(t \,|\, y)] = I(s_1; t) + \gamma I(s_1; x_1 = b) - \beta I(s_1; x_i = b,\ i \neq 1) \to \min, \tag{8}$$

где $I(s_1, t)$ – взаимная информация между общей суммой t и суммой без одного s_1.

§ 4. Численные результаты

Рассмотрим регулярную протоматрицу проверок на четность $\{3, 6\}$ размера 1536 на 3072. Каждый переменный узел связан с тремя проверочными узлами, и мощность входных распределений равна $|y_i| = 2^{10} - 1$. Для метода информационного сжатия выбрано $\beta = \infty$, чтобы сохранить максимальное количество релевантной информации.

Как показывает рис. 1, построенная функция вычитания дает взаимную информацию (4), очень близкую к (7), что означает сохранение значительного количества информации об оставшихся битах. Незначительное расхождение наблюдается только при высоком отношении сигнал/шум (ОСШ).

Остаточная информация при низких значениях ОСШ достаточно близка к нулю, но с увеличением ОСШ она быстро растет. На самом деле после каждой итерации декодера уровень шума фактически уменьшается, поэтому предлагаемый алгоритм должен хорошо работать в первых итерациях, а на последних итерациях появляется слишком много информации о вычитаемом бите и точность декодера снижается.

Результаты, продемонстрированные на рис. 2,*а*, показывают, что эффективность коррекции ошибок декодера ПрР с разработанным методом вычитания практически не отличается от базового декодера без использования разработанной оптимизации сложности.

Рассмотрим предлагаемые декодеры для регулярных протоматриц $\{3, 6\}$, $\{4, 8\}$ и $\{5, 10\}$ с учетом 4- и 5-битных сообщений. Разница в разрядности рассматрива-

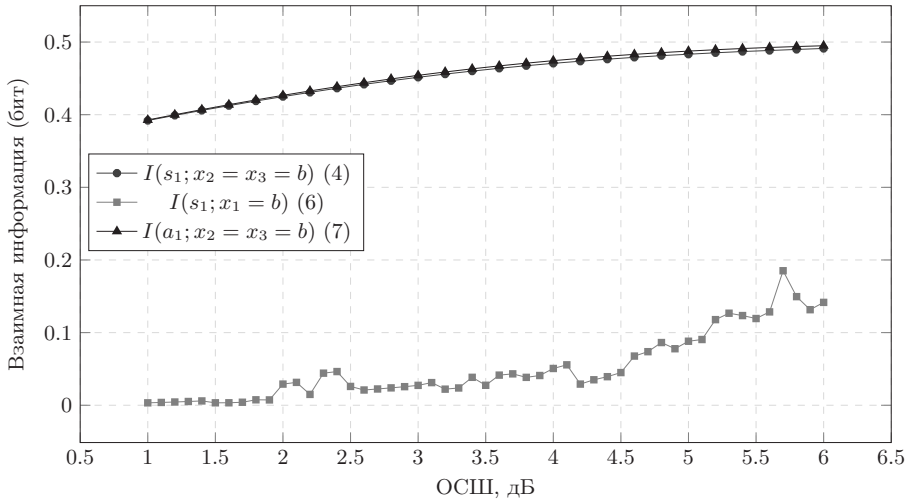


Рис. 1. Анализ взаимной информации

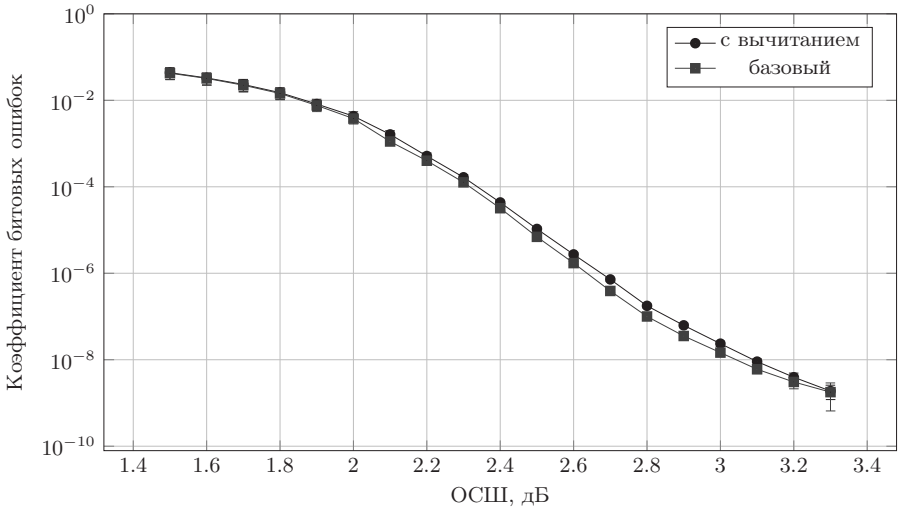
емых декодеров объясняется затратами на память, которые растут с увеличением разрядности сообщений и весов матрицы.

Результаты моделирования для регулярной протоматрицы $\{3, 6\}$ и 5-битных сообщений показаны на рис. 2,*б*. Предлагаемые и стандартные методы имеют одинаковую производительность исправления ошибок на пороге декодирования, их небольшое различие объясняется чувствительностью метода МИС к точности. Однако, что более важно, предлагаемый метод сопоставим со стандартным и имеет более простую архитектуру. Различие в уровнях ошибок объясняется тем, что "вычитание" с увеличением ОСШ начинает оставлять слишком много остаточной информации о вычитаемом бите, и точность декодера начинает снижаться (см. рис. 1).

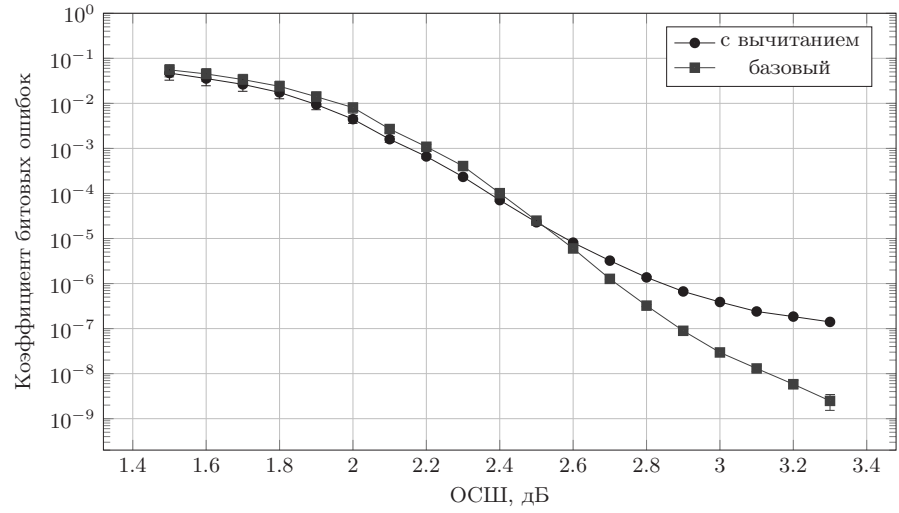
Результаты доказывают эффективность разработанного алгоритма. Однако его использование приводит к экспоненциальной по памяти сложности оптимизации параметров декодера, так как требует хранения $(N + 1)$-мерных распределений во время работы, где N зависит от весов матрицы проверок на четность. Нахождение алгоритма с меньшей сложностью является задачей дальнейших исследований. Расширенная версия представленных результатов опубликована в работе [10].

§ 5. Заключение

Данное исследование посвящено разработке методы снижения вычислительной и пространственной сложности МПП-декодеров. В статье предложен и оценен метод реализации оптимизации вычитания для уменьшения сложности операций в узлах и количества хранимых таблиц поиска. Предложенный метод вычитания реализован в декодерах с использованием различных типов расписаний: параллельного и последовательного расписания. Это позволило значительно снизить сложность и количество обращений к памяти декодера при потере примерно 0,1 дБ в эффективности коррекции ошибок при использовании последовательного расписания и без потерь при использовании параллельного расписания. Выигрыш в количестве используемых таблиц поиска увеличивается с весом столбца матрицы проверок на четность и для веса 5 составляет 50% для обоих типов расписаний. Однако оптимизация таблицы поиска для вычитания в случае последовательного расписания требует памяти, которая растет экспоненциально с весом столбца матрицы проверки на четность.



а) ПРР, 10 итераций с 5-битными сообщениями с регулярной {3,6} протоматрицей.



б) ПСР, 5 итераций с 5-битными сообщениями с регулярной {3,6} протоматрицей

Рис. 2. Сравнение базового декодера и декодера с предложенной оптимизацией

СПИСОК ЛИТЕРАТУРЫ

1. *Угловский А.Ю., Мельников И.А., Алексеев И.А., Кureев А.А.* Оценка низкого уровня ошибок с помощью выборки по значимости с равномерным распределением // Пробл. передачи информ. 2023. Т. 59. № 4. С. 3–12. https://doi.org/10.31857/S0555292323040010
2. *Gallager R.* Low-Density Parity-Check Codes // IRE Trans. Inform. Theory. 1962. V. 8. № 1. P. 21–28. https://doi.org/10.1109/TIT.1962.1057683
3. *Fossorier M.P.C., Mihaljevic M., Imai H.* Reduced Complexity Iterative Decoding of Low-Density Parity Check Codes Based on Belief Propagation // IEEE Trans. Commun. 1999. V. 47. № 5. P. 673–680. https://doi.org/10.1109/26.768759

4. *Lewandowsky J., Bauch G.* Information-Optimum LDPC Decoders Based on the Information Bottleneck Method // IEEE Access. 2018. V. 6. P. 4054–4071. https://doi.org/10.1109/ACCESS.2018.2797694
5. *Lewandowsky J., Bauch G., Stark M.* Information Bottleneck Signal Processing and Learning to Maximize Relevant Information for Communication Receivers // Entropy. 2022. V. 24. № 7. Paper No. 972 (30 pp.). https://doi.org/10.3390/e24070972
6. *Stark M., Lewandowsky J., Bauch G.* Information-Bottleneck Decoding of High-Rate Irregular LDPC Codes for Optical Communication Using Message Alignment // Appl. Sci. 2018. V. 8. № 10. Paper No. 1884 (17 pp.). https://doi.org/10.3390/app8101884
7. *Kurkoski B.M., Yamaguchi K., Kobayashi K.* Noise Thresholds for Discrete LDPC Decoding Mappings // Proc. 2008 IEEE Global Telecommunications Conf. (GLOBECOM'08). New Orleans, LA, USA. Nov. 30 – Dec. 4, 2008. P. 1–5. https://doi.org/10.1109/GLOCOM.2008.ECP.214
8. *Фернандес М., Кабатянский Г.А., Круглик С.А., Мяо И.* Коды для точного нахождения носителя разреженного вектора по ошибочным линейным измерениям и их декодирование // Пробл. передачи информ. 2023. Т. 59. № 1. С. 17–24. https://doi.org/10.31857/S0555292323010023
9. *He X., Cai K., Song W., Mei Z.* Dynamic Programming for Sequential Deterministic Quantization of Discrete Memoryless Channels // IEEE Trans. Commun. 2021. V. 69. № 6. P. 3638–3651. https://doi.org/10.1109/TCOMM.2021.3062838
10. *Melnikov I.A., Uglovskii A.Yu., Kreshchuk A.A., Kureev A.A., Khorov E.M.* Reducing the Complexity of the Layer Scheduled LDPC Decoder Based on the Information Bottleneck Method // Probl. Inf. Transm. 2024. V. 60. № 3. P. 199–208. https://doi.org/10.1134/S0032946024030049

Мельников Игнат Алексеевич
Угловский Артем Юрьевич
Крещук Алексей Андреевич
Институт проблем передачи информации
им. А.А. Харкевича РАН, Москва
melnikov@iitp.ru
uglovski@iitp.ru
krsh@iitp.ru
Кureев Алексей Андреевич
Московский физико-технический институт
(национальный исследовательский университет), Москва
Институт проблем передачи информации
им. А.А. Харкевича РАН, Москва
kureev@wireless.iitp.ru
Хоров Евгений Михайлович
Институт проблем передачи информации
им. А.А. Харкевича РАН, Москва
khorov@wireless.iitp.ru

Поступила в редакцию
20.09.2024
После доработки
20.09.2024
Принята к публикации
30.11.2024

ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 60 **2024** **Вып. 3**

УДК 621.391 : 519.216.3

© 2024 г. **Н.Г. Докучаев**

СПЕКТРАЛЬНОЕ ПРЕДСТАВЛЕНИЕ НЕЗАТУХАЮЩИХ СИГНАЛОВ В ДИСКРЕТНОМ ВРЕМЕНИ И ЗАДАЧА ПРОГНОЗИРОВАНИЯ

Изучается спектральное представление для дискретных по времени сигналов из ℓ_∞, т.е. для ограниченных дискретных по времени сигналов, включая сигналы, не затухающие на $\pm\infty$. Это представление применено к проблеме предсказуемости и восстановления сигналов. С этой целью понятия передаточных функций, спектральных разрывов, ограниченности полосы и фильтров распространены на сигналы данного общего типа.

Ключевые слова: незатухающие сигналы, спектральное представление, передаточные функции, прогнозирование, восстановление данных.

DOI: 10.31857/S0555292324030045, **EDN:** WPDDTT

§ 1. Введение

Наиболее важные инструменты, используемые для обработки сигналов, основаны на представлении сигналов в частотной области. Это включает в себя, в частности, понятия передаточных функций, спектральных разрывов, фильтров, используемых для условий предсказуемости и восстанавливаемости данных.

Для сигналов в непрерывном времени $x(t)|_{t\in\mathbb{R}}$ спектральное представление осуществляется через преобразование Фурье для двусторонних процессов,затухающих при $t \to \pm\infty$, и через преобразование Лапласа для односторонних процессов, равных нулю на одной половине временной оси, но не обязательно затухающих на другой половине временной оси.

Аналогичная ситуация и для дискретных временных процессов $\{x(t)\}_{t=-\infty}^{+\infty}$ и их спектрального представления через Z-преобразование. Для двусторонних дискретных процессов, стремящихся к нулю достаточно быстро на $\pm\infty$, таких как процессы из ℓ_2, это Z-преобразование определено на единичной окружности $z \in \mathbb{C} : \ |z| = 1$. В случае ограниченных односторонних процессов $\{x(t)\}_{t=-\infty}^{+\infty}$, т.е. таких, что либо $x(t) = 0$ для $t < 0$, либо $x(t) = 0$ для $t > 0$, это Z-преобразование определено либо в области $z \in \mathbb{C} : \ |z| < 1$, либо в области $z \in \mathbb{C} : \ |z| > 1$.

Можно заметить, что любой ненулевой сигнал из ℓ_∞ может быть преобразован в сигнал из ℓ_1 без потери информации, например, заменой $x(t)$ на $e^{-|t|}x(t)$. Однако, по крайней мере для сигналов из ℓ_2, такие преобразования представляют собой свертки на окружности $z \in \mathbb{C} : \ |z| = 1$ в частотной области со сглаживающими ядрами. К сожалению, такие преобразования устраняют вырождения спектра.

Стандартные методы прогнозирования и восстановления данных в обработке сигналов в основном основываются на предположениях о вырождении спектра. Следовательно, исключение незатухающих сигналов из спектральных методов может быть неудобным для практических реализаций; это потребует введения некоторых

искусственных предположений о реальных сигналах только для того, чтобы удовлетворить этим ограничениям. Этим мотивировано наше изучение спектрального представления и спектральных методов для общих типов незатухающих ограниченных сигналов.

Формально спектральное представление для ограниченных незатухающих дискретных во времени сигналов определяется как преобразования Фурье для псевдомер на $[-\pi, \pi]$, представленных как элементы пространства ℓ_∞ (см. [1, гл. III]). Однако это определение не приводит к частотным понятиям и методам, таким как фильтрация, предсказание и восстановление данных, применимым к незатухающим сигналам. В данной статье предлагается более конструктивное определение спектрального представления для незатухающих сигналов. На основе этого определения в статье расширяются понятия передаточных функций, каузальных передаточных функций, спектральных разрывов и фильтров для незатухающих сигналов (§ 3).

В частности, показано, что класс допустимых передаточных функций для незатухающих сигналов меньше, чем для сигналов из ℓ_2, и что стандартные идеальные фильтры низких частот не могут быть напрямую применены к незатухающим сигналам. Далее, в статье представлены частотные условия предсказуемости для незатухающих сигналов с вырождением спектра в одной точке. Для этого используется адаптированная для незатухающих сигналов "предсказывающая" передаточная функция из [2], введенная для прогнозирования сигналов из ℓ_1. В данной статье установлена применимость этих функций для незатухающих сигналов. Это может быть интересно, поскольку доказывает, что влияние прошлой истории незатухающих сигналов из ℓ_∞ уменьшается со временем, аналогично случаю сигналов из ℓ_1.

Следует отметить, что гипотеза о вырождении спектра широко используется в теории обработки сигналов; однако было бы нереалистично ожидать, что реальный сигнал будет иметь вырождение спектра в чистой форме, учитывая шумовые искажения и пропуски выборок. Некоторая устойчивость предикторов из [2], используемых в данной статье, к шумовым искажениям и усечению выборок была установлена в [2]. Численные эксперименты для этих предикторов при работе с искажениями данных описаны в [3].

Наконец, в статье представлены результаты восстановления сигналов с разрывами в спектре определенного размера с неизвестным расположением (§ 4).

Расширенная версия настоящей статьи представлена в [4].

Некоторые обозначения. Обозначим через $\mathbb{Z}$, $\mathbb{R}$ и $\mathbb{C}$ множества всех целых, вещественных и комплексных чисел соответственно.

Пусть

$$\mathbb{T} := \{z \in \mathbb{C} : \ |z| = 1\},$$
$$\mathbb{D} := \{z \in \mathbb{C} : \ |z| > 1\},$$
$$\bar{\mathbb{D}} := \{z \in \mathbb{C} : \ |z| \geqslant 1\}.$$

Обозначим через ℓ_∞ множество всех процессов (сигналов) $x\colon \mathbb{Z} \to \mathbb{C}$, таких что

$$\|x\|_{\ell_\infty} := \sup_{t \in \mathbb{Z}} |x(t)| < +\infty.$$

Для $r \in [1, \infty)$ обозначим через ℓ_r множество всех процессов (сигналов) $x\colon \mathbb{Z} \to \mathbb{C}$, таких что

$$\|x\|_{\ell_r} := \left(\sum_{t \in \mathbb{Z}} |x(t)|^r \right)^{1/r} < +\infty.$$

Обозначим через $C([-\pi,\pi])$ стандартное линейное пространство непрерывных функций $f\colon [-\pi,\pi]\to\mathbb{C}$ с равномерной нормой $\|f\|_C := \sup\limits_{\omega}|f(\omega)|$.

Обозначим через $\overset{p}{W}{}_2^1\ (-\pi,\pi)$ пространство Соболева функций $f\colon [-\pi,\pi]\to\mathbb{C}$, принадлежащих пространству $L_2(-\pi,\pi)$ вместе с обобщенными производными до первого порядка, таких что $f(-\pi)=f(\pi)$.

Для линейного нормированного пространства $\mathcal{X}$ обозначим через $\mathcal{X}^*$ его сопряженное пространство.

Обозначим через $\mathbb{I}$ индикаторную функцию.

§ 2. Спектральное представление процессов из ℓ_∞

Пусть $\mathcal{A}$ – пространство функций $f\in C([-\pi,\pi])$ с конечной нормой

$$\|f\|_{\mathcal{A}} := \sum_{k\in\mathbb{Z}}|\widehat{f}_k|,$$

где

$$\widehat{f}_k = \frac{1}{2\pi}\int\limits_{-\pi}^{\pi} e^{-i\omega s}f(s)\,ds$$

– коэффициенты Фурье функции f. Другими словами, $\mathcal{A}$ – пространство абсолютно сходящихся рядов Фурье на $[-\pi,\pi]$. Благодаря выбору нормы это сепарабельное банахово пространство, изоморфное ℓ_1.

Очевидно, что вложения

$$\mathcal{A}\subset C([-\pi,\pi])\subset L_1([-\pi,\pi])$$

являются непрерывными. Отсюда следует, что вложения

$$L_1([-\pi,\pi])^*\subset C([-\pi,\pi])^*\subset\mathcal{A}^*$$

являются непрерывными.

Можно заметить, что существуют функции из $C([-\pi,\pi])$, которые не принадлежат пространству $\mathcal{A}$ (см., например, [5, с. 113]).

Мы предполагаем, что $X\in L_1([-\pi,\pi])$ представляет элемент сопряженного пространства $C([-\pi,\pi])^*$, такой что

$$\langle X,f\rangle = \frac{1}{2\pi}\int\limits_{-\pi}^{\pi} X(\omega)f(\omega)\,d\omega$$

для $f\in C([-\pi,\pi])$. Мы будем использовать это же обозначение $\langle\cdot\,,\cdot\rangle$ для расширения этой билинейной формы на $\mathcal{A}^*\times\mathcal{A}$.

Предложение 1. *Справедливы следующие утверждения*:

1. *Если* $f\in\mathcal{A}$ *и* $g\in\mathcal{A}$, *то* $h=fg\in\mathcal{A}$ *и* $\|h\|_{\mathcal{A}}\leqslant\|f\|_{\mathcal{A}}\|g\|_{\mathcal{A}}$;
2. *Вложение* $\overset{p}{W}{}_2^1\ (-\pi,\pi)\subset\mathcal{A}$ *является непрерывным.*

Можно заметить, что двойственность между $\mathcal{A}$ и двойственным ему пространством $\mathcal{A}^*$ формально приводит к спектральному представлению

$$\langle X,f\rangle = \sum_{k\in\mathbb{Z}} x(k)\widehat{f}_k$$

для $X \in \mathcal{A}^*$ и $f \in \mathcal{A}$, где $\{\widehat{f_k}\} \in \ell_1$ – ряд коэффициентов Фурье для f, аналогично [1, гл. III], где элементы $x \in \ell_\infty$ рассматривались как преобразования Фурье для псевдомер на $[-\pi, \pi]$. Для сигналов в непрерывном времени определение преобразования Фурье для $L_\infty(\mathbb{R})$ через подобную двойственность дано в [5, гл. VI]. Однако для целей, связанных с задачами восстанавливаемости и прогнозирования цифровых сигналов, нам понадобится прямое определение спектральных представлений для $x \in \ell_\infty$, основанное на следующей лемме.

Лемма 1. *Для любого $x \in \ell_\infty$ существует слабо* предельный элемент $X \in \mathcal{A}^*$ последовательности функций*

$$X_m(\omega) := \sum_{t=-m}^{m} e^{-i\omega t} x(t),$$

определенных на $[-\pi, \pi]$ для $m = 1, 2, \ldots$ Этот элемент X таков, что

$$\|X\|_{\mathcal{A}^*} = \|x\|_{\ell_\infty}$$

и

$$\langle X, f \rangle = \sum_{t \in \mathbb{Z}} x(k) \widehat{f_k},$$

где $\widehat{f_k} \in \ell_1$ является рядом коэффициентов Фурье для f.

Можно отметить, что в приведенной выше лемме

$$X_m \in L_1([-\pi, \pi]) \subset C([-\pi, \pi])^* \subset \mathcal{A}^*.$$

Зададим отображение $\mathcal{F}\colon \ell_\infty \to \mathcal{A}^*$ так, что $X = \mathcal{F}x$ для $x \in \ell_\infty$ является пределом в $\mathcal{A}^*$, введенным в лемме 1. Согласно лемме 1 это отображение линейно и непрерывно.

Зададим отображение $\mathcal{G}\colon \mathcal{A}^* \to \ell_\infty$, такое что

$$x(t) = \langle X, e^{i \cdot t} \rangle \quad \text{для} \quad x = \mathcal{G}X, \quad X \in \mathcal{A}^*, \quad t \in \mathbb{Z}.$$

Очевидно, что оператор $\mathcal{G}\colon \mathcal{A}^* \to \ell_\infty$ является линейным и непрерывным. Кроме того, $x_m(t) = x(t)\mathbb{I}_{|t| \leqslant m}$ для $x = \mathcal{G}X$ и $x_m := \mathcal{G}X_m$.

Теорема 1. *Отображения $\mathcal{F}\colon \ell_\infty \to \mathcal{A}^*$ и $\mathcal{G}\colon \mathcal{A}^* \to \ell_\infty$ являются непрерывными изометрическими биекциями, такими что $\mathcal{F} = \mathcal{G}^{-1}$ и $\mathcal{G} = \mathcal{F}^{-1}$.*

На этом основании мы будем использовать обозначение $\mathcal{F}^{-1}$ вместо $\mathcal{G}$. Мы рассматриваем $\mathcal{X} = \mathcal{F}x \in \mathcal{A}^*$ как спектральное представление $x \in \ell_\infty$.

Замечание 1. Из леммы 1 следует, что для $x \in \ell_\infty$, $y \in \ell_1$ и $X = \mathcal{F}x$, $Y = \mathcal{F}y$ справедливо, что $Y \in \mathcal{A}$ и

$$\sum_{t \in \mathbb{Z}} x(t) y(t) = \langle X, Y \rangle.$$

Далее мы говорим, что $X \in \mathcal{A}^*$ является действительным (мнимым), если $\langle X, f \rangle$ является действительным (мнимым) для любого действительного $f \in \mathcal{A}$.

Очевидно, что любое $X \in \mathcal{A}^*$ допускает единственное представление $X = \bar{X} + i\widetilde{X}$, где $\bar{X}, \widetilde{X} \in \mathcal{A}^*$ являются действительными. Мы будем использовать обозначения $\operatorname{Re} X$ для $\bar{X}$ и $\operatorname{Im} X$ для $\widetilde{X}$.

Будем говорить, что действительное $X \in \mathcal{A}^*$ является нечетным (четным), если $\langle X, f \rangle = 0$ для любого действительного четного (нечетного) $f \in \mathcal{A}$.

Легко показать, что если $x \in \ell_\infty$ является действительным, то $\operatorname{Re} X$ четно, а $\operatorname{Im} X$ нечетно для $X = \mathcal{F}x$; если $X \in \mathcal{A}^*$ является действительным и четным, то $\operatorname{Im} X = 0$.

Вырождение спектра. Введенное выше спектральное представление позволяет описывать сигналы из ℓ_∞, имеющие спектральные разрывы, такие как полосовые процессы, а также более слабые типы вырождения спектра.

Определение 1. (i) Для измеримого по Борелю множества $D \subset [-\pi, \pi]$ с непустой внутренностью пусть $x \in \ell_\infty$ таково, что $\langle \mathcal{F}x, f\rangle = 0$ для любого $f \in \mathcal{A}$, такого что $f|_{[-\pi,\pi]\setminus D} \equiv 0$. В этом случае будем говорить, что D является спектральным разрывом для $x \in \ell_\infty$ и $X = \mathcal{F}X$.

(ii) Пусть Γ – некоторое множество, и пусть функция $G\colon [-\pi, \pi] \times \Gamma \to \mathbb{C}$ такова, что $G(\cdot\,, g) \in \mathcal{A}$ для каждого $g \in \Gamma$ и $\sup\limits_{g\in\Gamma} \|G(\cdot\,, g)\|_{\mathcal{A}} = +\infty$. Пусть $x \in \ell_\infty$ таково, что $\sup\limits_{g\in\Gamma} \|G(\cdot\,, g)X\|_{\mathcal{A}^*} < +\infty$. Тогда будем говорить, что сигнал x обладает вырождением спектра, компенсирующим G.

Пример 1. Пусть $r > 0$, $\Gamma = (0, 1)$ и $G(\omega, \nu) = (|\omega|^r + \nu)^{-1}$. Пусть $x \in \ell_1 \cap \ell_\infty$ таково, что

$$\operatorname*{ess\,sup}_{\omega\in[-\pi,\pi]} |X(\omega)|/G(\omega, 0) < +\infty$$

для $X = \mathcal{F}x$. Тогда x обладает вырождением спектра, компенсирующим G.

В приведенном выше примере соответствующее вырождение спектра является вырождением спектра в одной точке. Можно отметить, что $G(\cdot\,, \nu) \in \mathcal{A}$ для $\nu > 0$, поскольку $G(\cdot\,, \nu) \in \overset{p}{W}{}_2^1\,(-\pi, \pi)$.

Замечание 2. Из леммы 1 и замечания 1 следует, что $x \in \ell_\infty$ имеет спектральный разрыв $D \subset [-\pi, \pi]$ тогда и только тогда, когда

$$\sum_{t\in\mathbb{Z}} x(t)y(t) = 0$$

для любого $y \in \ell_1$, такого что $Y(\omega) = 0$ для $Y = \mathcal{F}y$ и $\omega \notin D$. Аналогично, сигнал $x \in \ell_\infty$ имеет вырождение спектра, компенсирующее G, как в определении 1(ii), тогда и только тогда, когда

$$\sup_{\nu>0,\eta_g} |\sum_{t\in\mathbb{Z}} x(t)y(t)| < +\infty,$$

где $\eta_g = \mathcal{F}^{-1}(G(\cdot\,, g)Y_g)$ и $Y_g = \mathcal{F}y_g$.

§ 3. Применения для обработки сигналов

Введенное выше спектральное представление позволяет распространить некоторые стандартные инструменты обработки сигналов на случай сигналов из ℓ_∞. В частности, оно позволяет характеризовать сигналы из ℓ_∞, обладающие спектральными разрывами, такие как процессы с ограниченной полосой. Также оно поддерживает реализацию важных инструментов, таких как передаточные функции и фильтр нижних частот, и помогает получать алгоритмы прогнозирования и восстановления данных для этих сигналов.

3.1. Передаточные функции. Существующая теория не рассматривает передаточные функции, применяемые к общему типу двусторонних незатухающих сигналов процессов $x(t)$ из ℓ_∞. Предложенное выше спектральное представление $x \in \ell_\infty$

через элементы из $\mathcal{A}^*$ позволяет реализовать передаточные функции для всех $x \in \ell_\infty$.

Определение 2. Пусть функция $H\colon \mathbb{T} \to \mathbb{C}$ такова, что функция $H(e^{i\cdot})$, определенная на $[-\pi, \pi]$, принадлежит $\mathcal{A}$. Тогда будем говорить, что $H(e^{i\cdot})$ является передаточной функцией на ℓ_∞. Для $x \in \ell_\infty$ и $X = \mathcal{F}x$ определим $\widehat{X} = HX \in \mathcal{A}^*$ и $\widehat{x} = \mathcal{F}^{-1}\widehat{X}$ таким образом, что

$$\langle \widehat{X}, f\rangle := \langle X, H(e^{i\cdot})f\rangle.$$

Следует отметить, что согласно предложению 1(i) для любого $f \in \mathcal{A}$ также верно, что $H(e^{i\cdot})f \in \mathcal{A}$. Следовательно, $\widehat{X} = HX$ корректно определено как элемент пространства $\mathcal{A}^*$.

Лемма 2. *Пусть $H\colon \mathbb{T} \to \mathbb{C}$ такова, что $H(e^{i\cdot}) \in \mathcal{A}$, и пусть*

$$H(e^{i\omega}) = \sum_{k\in\mathbb{Z}} \widehat{h}_k e^{i\omega k},$$

где $\{\widehat{h}_k\}_{k\in\mathbb{Z}} \in \ell_1$. Пусть $X \in \mathcal{A}^*$, $\widehat{X} = HX$ и $\widehat{x} = \mathcal{F}^{-1}\widehat{X}$. Тогда

$$\widehat{x}(t) = \sum_{q\in\mathbb{Z}} \widehat{h}_{t-q}x(q).$$

Этот ряд абсолютно сходится равномерно на любом ограниченном множестве $x \in \ell_\infty$.

3.2. Фильтры. К сожалению, идеальные низкочастотные и высокочастотные фильтры с прямоугольным профилем передаточной функции не принадлежат пространству $\mathcal{A}$. Поэтому они не покрываются определением 2 передаточных функций, применимых к сигналам из ℓ_∞. Тем не менее, некоторые аппроксимации этих идеальных фильтров могут быть достигнуты с использованием трапецеидальных передаточных функций из $\mathcal{A}$. Например, пусть

$$H_{p,q}(e^{i\omega}) := \mathbb{I}_{\omega\in[-p,p]} + \frac{q-\omega}{q-p}\mathbb{I}_{\omega\in(p,q]} + \frac{q+\omega}{q-p}\mathbb{I}_{\omega\in[-q,-p)},$$

где $0 < p < q < \pi$. Так как $\overset{p}{W}{}_2^1(-\pi, \pi) \subset \mathcal{A}$, функции $H_{p,q}(e^{i\cdot})$ принадлежат $\mathcal{A}$, следовательно, они являются допустимыми передаточными функциями на ℓ_∞. Очевидно, что для любого $x \in \ell_\infty$ и $X = \mathcal{F}x$ функция $H_{p,q}X$ имеет спектральный разрыв на $(-\pi, -q) \cup (q, \pi]$; в этом смысле фильтрованный процесс $\widehat{x} = \mathcal{F}^{-1}(H_{p,q}X)$ является процессом с ограниченной полосой.

При $q \to p+$ указанные функции аппроксимируют идеальный низкочастотный фильтр с полосой пропускания $[-p, p]$, т.е. с прямоугольной передаточной функцией $H_{p,p}(e^{i\omega}) = \mathbb{I}_{\omega\in[-p,p]}$. Как было указано выше, $H_{p,p}(e^{i\cdot}) \notin \mathcal{A}$, т.е. это не передаточная функция, применимая к сигналам из ℓ_∞.

Замечание 3. Для любых $p > 0$ и $q > p$ и для $\Gamma = \{\nu : \nu > 0\}$ каждый сигнал x со спектральным разрывом $[-q, q]$ имеет вырождение спектра, компенсирующее $G(\omega, \nu) = \nu H_{p,q}(\omega)$.

3.3. Каузальные передаточные функции.

Определение 3. Будем говорить, что передаточная функция H, описанная в определении 2, является каузальной на ℓ_∞, если для любого $\tau \in \mathbb{Z}$ и любого $x \in \ell_\infty$, такого что $x(t) = 0$ для всех $t < \tau$, выполняется $\widehat{x}(\tau) = 0$, где $\widehat{x} = \mathcal{F}^{-1}\widehat{X}$ и $\widehat{X} = HX$.

Отметим, что в приведенном выше определении $\widehat{X} \in \mathcal{A}^*$.

Теорема 2. *Предположим, что функция $H\colon \bar{\mathbb{D}} \to \mathbb{C}$ является непрерывной на $\bar{\mathbb{D}}$ и аналитической на $\mathbb{D}$, и что $H(e^{i\cdot}) \in \mathcal{A}$. Тогда $H(e^{i\cdot})$ является передаточной функцией на ℓ_∞, которая является каузальной в смысле определения* 3.

§ 4. Применения для задач восстановления данных и прогнозирования

Рассмотрим задачу восстановления ненаблюдаемых значений $x(t_k)|_{k\in\mathcal{M}}$ по наблюдаемым значениям $x(t_k)|_{k\in\mathbb{Z}\setminus\mathcal{M}}$ для сигналов из некоторых подмножеств ℓ_∞.

4.1. Восстановление конечного набора пропущенных значений. Пусть $D \subset [-\pi,\pi]$ – борелевское множество с непустой внутренностью. Обозначим через $\mathcal{V}(D)$ (или $\mathcal{V}_R(D)$, или $\mathcal{V}_I(D)$) множество всех сигналов $x \in \ell_\infty$, таких что $x = x_v + x_s$, где

- D является спектральным разрывом для $X_v = \mathcal{F}x_v$ (соответственно, для $\operatorname{Re} X_v$, или для $\operatorname{Im} X_v$);
- $X_s = \mathcal{F}x_s$ таково, что $X_s \in C([-\pi,\pi])^*$ – мера Радона на $[-\pi,\pi]$, сингулярная относительно меры Лебега.

В частности, соответствующие сигналы включают сигналы

$$x_s = \sum_{k\in\mathbb{Z}} \alpha_k e^{i\omega_k t}$$

для всех последовательностей $\{\alpha_k\}_{k\in\mathbb{Z}}^{\infty} \in \ell_1$. В этом случае

$$X_s = \mathcal{F}x_s = \sum_{k\in\mathbb{Z}} \alpha_k \delta(\cdot - \omega_k),$$

где $\delta(\cdot - \omega_k)$ – дельта-функции, т.е. $\langle \delta(\cdot - \omega_k), f\rangle = f(\omega_k)$ для $f \in C([-\pi,\pi])$.

Очевидно, что $\mathcal{V}(D) \subset \mathcal{V}_R(D) \cap \mathcal{V}_I(D)$.

Теорема 3. *Для любого борелевского множества $D \subset [-\pi,\pi]$ с непустой внутренностью и любого конечного множества $\mathcal{M} \subset \mathbb{Z}$ значения $x(t)|_{t\in\mathcal{M}}$ для любого $x \in \mathcal{V}_R(D) \cup \mathcal{V}_I(D)$ однозначно определяются значениями $x(t)|_{t\in\mathbb{Z}\setminus\mathcal{M}}$.*

Теорема 3 формально подразумевает метод восстановления для $x|_{\mathcal{M}}$, так как тригонометрический многочлен

$$X_{\mathcal{M}}(\omega) = \sum_{t\in\mathcal{M}} e^{-it\omega} x(t)$$

наблюдаем на D в следующем смысле: для любого $f \in \mathcal{A}$, поддержанного на D, выполнено

$$\langle X_{\mathcal{M}}, f\rangle = -\langle X_{\mathbb{Z}\setminus\mathcal{M}}, f\rangle,$$

где $X_{\mathbb{Z}\setminus\mathcal{M}} := \mathcal{F}\big(\mathbb{I}_{\cdot\notin\mathcal{M}} x(\cdot)\big)$. Однако это потребовало бы вычисления $X_{\mathbb{Z}\setminus\mathcal{M}}$, что кажется численно трудоемким.

Следующая теорема предлагает альтернативный подход, основанный на применении явно заданных каузальных передаточных функций и применимый к сигналам из более узкого класса $\mathcal{V}(D)$.

4.2. Задача прогнозирования. Пусть заданы $c > 0$ и $q > 1$. Для $\widehat{\omega} \in (-\pi,\pi]$, $\nu \in (0,1)$ определим

$$G(\omega,\widehat{\omega},\nu) := \exp\frac{c}{|e^{i\omega} - e^{i\widehat{\omega}}|^q + \nu}.$$

Функции $G(\cdot\,,\widehat{\omega},\nu)$ принадлежат $\overset{p}{W}{}_2^1(-\pi,\pi)$ для любых $\widehat{\omega}$ и ν. Таким образом, они принадлежат $\mathcal{A}$. Кроме того, $\|G(\cdot\,,\widehat{\omega},\nu)\|_{\mathcal{A}} \to +\infty$ для любого $\widehat{\omega}$ при $\nu \to 0$.

Пусть $\mathcal{X}_{\widehat{\omega}}$ – множество всех процессов $x \in \ell_\infty$ с точечным вырождением спектра в $\widehat{\omega}$, компенсирующим G, т.е. таких, что

$$\|x\|_{\mathcal{X}_{\widehat{\omega}}} := \sup_{\nu\in(0,1)} \|XG(\cdot\,,\widehat{\omega},\nu)\|_{\mathcal{A}^*} < +\infty, \quad X = \mathcal{F}x.$$

Рассмотрим $\mathcal{X}_{\widehat{\omega}}$ как линейное нормированное пространство с соответствующей нормой.

В частности, это множество включает все процессы с любым спектральным разрывом $D \subset [-\pi,\pi]$ с непустой внутренностью, такой что $e^{i\widehat{\omega}}$ принадлежит внутренней части дуги $\left\{e^{i\omega}\right\}_{\omega\in D}$.

Пусть задано $r \in (0,1)$. Для всех $\gamma > 0$ определим

$$H_\gamma(z) := z\left(1 - \exp\left[-\frac{\gamma}{z+1-\gamma^{-r}}\right]\right). \tag{1}$$

Имеем $H_\gamma(e^{i\cdot}) \in \overset{p}{W}{}_2^1(-\pi,\pi) \subset \mathcal{A}$.

Теорема 4. *Функции $\{H_\gamma(e^{i\cdot})\}_{\gamma>0} \subset \mathcal{A}$ являются каузальными передаточными функциями, определенными на ℓ_∞. При этом для любого $\widehat{\omega} \in (-\pi,\pi]$ существует $\bar{\gamma} > 0$, такое что*

$$\sup_{t\in\mathbb{Z}} |x(t+1) - \widehat{x}_\gamma(t)| \leqslant \varepsilon \quad \forall\gamma \geqslant \bar{\gamma}$$

для любого $x \in \mathcal{X}_{\widehat{\omega}}$, такого что $\|x\|_{\mathcal{X}_{\widehat{\omega}}} \leqslant 1$. Здесь

$$\widehat{x}_\gamma(t) = e^{i(\widehat{\omega}-\pi)t}\sum_{s\in\mathbb{Z}}^{t} h_\gamma(t-s)e^{i(\pi-\widehat{\omega})s}x(s), \quad h_\gamma = \mathcal{F}^{-1}H_\gamma.$$

В [4] показано, что функции H_γ аппроксимируют $e^{i\omega}$ на $\mathbb{T}$ для $\omega \in (-\pi,\pi)$ при $\gamma \to +\infty$, т.е. они представляют собой одношаговый предиктор.

Эти передаточные функции были введены в [2] для прогнозирования сигналов из ℓ_1. Некоторые численные эксперименты для этих предикторов с различными значениями r были описаны в [3].

4.3. Восстановимость данных в случае неизвестного спектрального зазора. Для $\Omega > 0$ пусть $\mathcal{U}(\Omega)$ (или $\mathcal{U}_R(\Omega)$, или $\mathcal{U}_I(\Omega)$) обозначает множество всех сигналов $x \in \ell_\infty$, таких что для каждого x существует борелевское измеримое множество $D = D(x(\cdot)) \subset [-\pi,\pi]$, такое что $\operatorname{mes} D \geqslant \Omega$ и $x \in \mathcal{V}(D)$ (или $\mathcal{V}_R(D)$, или $\mathcal{V}_I(D)$, соответственно). Очевидно, что $\mathcal{U}(\Omega) \subset \mathcal{U}_R(\Omega) \cap \mathcal{U}_I(\Omega)$.

Пусть $\lfloor r \rfloor$ обозначает целую часть числа $r > 0$, а $|\mathcal{M}|$ обозначает количество элементов множества $\mathcal{M}$.

Теорема 5. *Для любого конечного множества $\mathcal{M} \subset \mathbb{Z}$, для любого $x \in \mathcal{U}_R(\Omega) \cup \mathcal{U}_I(\Omega)$, для заданного множества наблюдений $x(t)|_{t\in\mathbb{Z}\setminus\mathcal{M}}$ количество возможных различных упорядоченных множеств $\{x(t)\}_{t\in\mathcal{M}}$ не может превышать $N := \lfloor 2\pi/\Omega \rfloor$.*

Следует отметить, что в теореме 5 оценка возможного количества значений в $\mathbb{C}^{|M|}$ для ненаблюдаемого вектора $\{x(t)\}_{t\in\mathcal{M}}$ не зависит от $\mathcal{M}$ или $|\mathcal{M}|$.

В частности, если $\Omega > \pi/2$, то вектор $\{x(t)\}_{t\in\mathcal{M}}$ однозначно определяется наблюдениями $x(t)|_{t\in\mathbb{Z}\setminus\mathcal{M}}$. Если $\Omega > \pi/4$, то вектор $\{x(t)\}_{t\in\mathcal{M}}$ может принимать не

более двух возможных значений в $\mathbb{C}^M$ для любого заданного множества наблюдений $x(t)|_{t\in\mathbb{Z}\setminus\mathcal{M}}$.

Доказательства результатов можно найти в [4].

СПИСОК ЛИТЕРАТУРЫ

1. *Kahane J.-P.* Séries de Fourier absolument convergentes. Berlin: Springer, 1970. https://doi.org/10.1007/978-3-662-59158-1
2. *Dokuchaev N.* Predictors for Discrete Time Processes with Energy Decay on Higher Frequencies // IEEE Trans. Signal Process. 2012. V. 60. № 11. P. 6027–6030. https://doi.org/10.1109/TSP.2012.2212436
3. *Dokuchaev N.* Near-ideal Causal Smoothing Filters for the Real Sequences // Signal Process. 2016. V. 118. № 1. P. 285–293. https://doi.org/10.1016/j.sigpro.2015.07.002
4. *Dokuchaev N.G.* Spectral Representation of Two-Sided Signals from ℓ_∞ and Applications to Signal Processing // Probl. Inf. Transm. 2024. V. 60. № 2. P. 113-126. https://doi.org/10.1134/S0032946024020030
5. *Katznelson Y.* An Introduction to Harmonic Analysis. Cambridge, UK: Cambridge Univ. Press, 2004. https://doi.org/10.1017/CBO9781139165372

Докучаев Николай Геннадьевич
Институт ZJU-UIUC (Чжэцзянский университет – Иллинойский университет в Урбане-Шампейне), Чжэцзянский университет, Хайнин, провинция Чжэцзян, Китай
Dokuchaev@intl.zju.edu.cn

Поступила в редакцию 23.12.2023
После доработки 14.09.2024
Принята к публикации 11.10.2024

ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 60 2024 Вып. 3

УДК 621.391.6 : 004.725.5

© 2024 г. **А.И. Пойда, И.А. Буртаков, А.А. Куреев, Е.М. Хоров**

ФОРМИРОВАНИЕ ШИРОКИХ ОТРАЖЕННЫХ ЛУЧЕЙ РЕКОНФИГУРИРУЕМЫМИ ИНТЕЛЛЕКТУАЛЬНЫМИ ПОВЕРХНОСТЯМИ[1]

Реконфигурируемая интеллектуальная поверхность (англ.: reconfigurable intelligent surface, RIS) является перспективной технологией для увеличения зоны покрытия и пропускной способности существующих и будущих беспроводных систем за счет настройки коэффициентов отражения. Среди предложенных в научном сообществе алгоритмов настройки RIS выделяется иерархический поиск луча (ИПЛ) за счет высокой скорости и возможности использовать простые канальные метрики, такие как мощность сигнала. Для алгоритмов ИПЛ необходимо формировать отраженные лучи разной ширины и с разными направлениями. Существующие методы синтеза широких лучей обладают рядом недостатков, таких как высокая вычислительная сложность и использование элементов RIS со сложной архитектурой. В данной статье рассматриваются эти проблемы и предлагается метод синтеза отраженных двумерных лучей с заданными шириной и направлением, который основан на фазовом сшивании узких лучей. Разработанный метод синтеза имеет низкую вычислительную сложность и позволяет RIS формировать лучи с десятками и сотнями тысяч элементов.

Ключевые слова: реконфигурируемые интеллектуальные поверхности, 5G, LTE, синтез лучей.

DOI: 10.31857/S0555292324030057, **EDN:** XATDIP

§ 1. Введение

Реконфигурируемая интеллектуальная поверхность (англ.: reconfigurable intelligent surface, RIS) является ключевой технологией для повышения области покрытия и пропускной способности систем связи шестого поколения (6G) [1]. RIS представляет собой двумерную поверхность, состоящую из множества пассивных элементарных ячеек (ЭЯ), которые могут воздействовать на распространение беспроводных сигналов путем изменения коэффициентов отражения. Например, RIS может перенаправлять и фокусировать отраженный сигнал от базовой станции (БС) на пользовательском устройстве (ПУ) для увеличения отношения сигнал/шум (англ.: signal-to-noise ratio, SNR) что приводит к увеличению пропускной способности, области покрытия и энергоэффективности беспроводных систем связи. Кроме того, это позволяет повысить защищенность беспроводных каналов связи [2, 3].

Для достижения описанных преимуществ необходимо осуществлять настройку коэффициентов отражения RIS. Применение традиционных подходов настройки на основе информации о состоянии канала [4] требует слишком большого числа пилотных сигналов, даже с учетом методов их снижения, таких как, например, compressed sensing [5, 6]. Поэтому важным требованием к алгоритмам настройки RIS является

[1] Исследование осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ.

использование простых канальных метрик, таких как SNR или мощность принимаемого опорного сигнала (англ.: reference signals received power, RSRP) [7]. Распространенным классом алгоритмов, основанных на использовании вышеуказанных метрик, являются алгоритмы поиска луча [8]. Они основаны на сканировании окружающей среды с помощью лучей различной формы и направления и выборе такого, который, например, максимизирует SNR.

Среди алгоритмов поиска луча выделяется алгоритм иерархического поиска луча (ИПЛ). Алгоритм ИПЛ начинает сканирование с широких лучей, а затем последовательно уменьшает их ширину, чтобы найти наилучшее направление. Поскольку технология RIS имеет ограничения, такие как дискретный фазовый сдвиг на ЭЯ и невозможность варьировать амплитуду отраженного сигнала, методы синтеза луча с заданным направлением и шириной, которые используются в технологии massive MIMO, неприменимы. В некоторых исследованиях предлагаются новые методы синтеза луча [9,10], такие как настройка фазы с использованием метода Герхберга – Сакстона [9] и формирование луча путем деактивации подмножества ЭЯ [10]. Однако они обладают высокой вычислительной сложностью или требуют слишком сложных архитектур RIS с поглощающими ЭЯ.

В данной статье предложен новый метод синтеза широкого луча, который учитывает архитектурные ограничения RIS. В основе предложенного метода лежит принцип фазовой сшивки нескольких узких лучей.

Дальнейшее изложение построено следующим образом. В § 2 представлен обзор литературы. Далее, § 3 посвящен методу синтеза широкого луча. В § 4 приведены полученные численные результаты. Наконец, § 5 завершает статью.

§ 2. Обзор литературы

Для настройки RIS с помощью алгоритмов ИПЛ необходимо формировать лучи с заданными направлением и шириной. Распространенным подходом является рассмотрение синтеза луча как решения оптимизационной задачи.

Например, авторы статьи [11] предлагают использовать генетический алгоритм для построения многонаправленных лучей или одного широкого луча от RIS. В работе [12] предлагается оптимальный алгоритм для дискретных фазовых сдвигов на ЭЯ и локально оптимальное решение для непрерывных.

Некоторые работы используют итеративный метод Герхберга – Сакстона [13] для поиска фазы луча, чтобы формировать луч с заданной шириной. В исследовании [14] предлагается двумерная иерархическая схема синтеза луча для работы RIS в ближнем поле. Авторы предлагают алгоритм на основе метода Герхберга – Сакстона для получения первичной конфигурации каждой ЭЯ. Затем на основе первичной конфигурации предлагается использовать алгоритм попеременной оптимизации для получения конечной конфигурации всей RIS.

Существуют более сложные подходы к синтезу луча, которые не формируют луч заданной ширины, но позволяют RIS охватывать заданную область произвольной формы. Например, в работе [15] предлагается подход к синтезу луча для покрытия слепых зон в миллиметровом диапазоне. В работе [16] предлагается схема на основе мажоризации-максимизации для максимизации минимальной пропускной способности множества ПУ в определенной области.

Основным недостатком методов из работ [9,14–16] является крайне высокая вычислительная сложность, не позволяющая применить предложенные алгоритмы для двумерных RIS с большим числом элементов.

Авторы работы [17] предлагают использовать детерминированный вид фазового сдвига на ЭЯ в явной форме, а именно, фазовый сдвиг каждой ЭЯ пропорционален квадрату номера ЭЯ по каждой оси. Однако в [17] рассматривается только

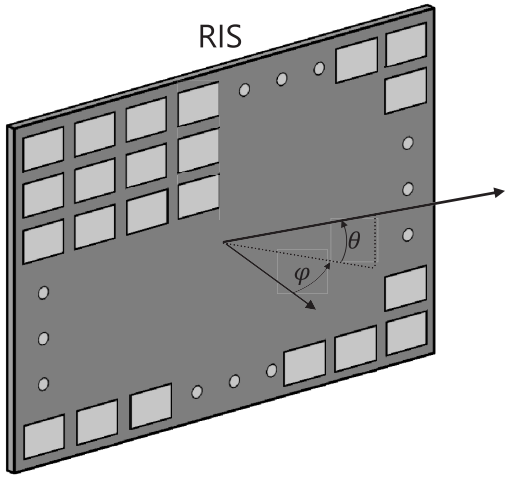


Рис. 1. Геометрия системы с RIS

RIS с непрерывными фазовыми сдвигами, что не позволяет напрямую использовать предложенный метод в RIS с дискретными фазовыми сдвигами. Кроме того, авторы не исследуют поведение боковых лепестков синтезированных лучей.

Таким образом, необходимо разработать метод синтеза широкого двумерного луча, учитывающий дискретные фазовые сдвиги и обладающий низкой вычислительной сложностью.

§ 3. Синтез лучей

В данном параграфе сначала рассматривается синтез узкого луча, основанный на классических подходах антенных решеток. Затем рассматривается метод синтеза широкого луча, который состоит из двух частей: синтез одномерного (1D) и двумерного (2D) широких лучей.

Для описания системы с RIS используется сферическая система координат с азимутальным углом φ и углом подъема θ, представленная на рис. 1.

3.1. Синтез узкого луча. Рассмотрим RIS с ЭЯ, обладающими непрерывными фазовыми сдвигами. Узким лучом назовем такой отраженный луч от RIS, который имеет максимальный амплитудный коэффициент усиления в желаемом направлении (φ, θ) при фиксированном числе ЭЯ.

Чтобы описать узкий луч, рассмотрим коэффициент усиления AF одномерной решетки [18]. В соответствии с геометрией, представленной на рис. 1,

$$\mathrm{AF}_x(\varphi,\theta) = \left| \sum_{n_x=0}^{N_x-1} e^{jn_x(kd_x \cos\theta \sin\varphi + \delta_x)} \right|, \tag{1}$$

$$\mathrm{AF}_y(\varphi,\theta) = \left| \sum_{n_y=0}^{N_y-1} e^{jn_y(kd_y \sin\theta + \delta_y)} \right|, \tag{2}$$

где k — волновое число, $d_{x/y}$ — расстояние между элементами, а $\delta_{x/y}$ — набег фазы между элементами в x/y-осях соответственно.

Если N_y одномерных решеток по оси x размещены с шагом $d_y = d_x = d$ вдоль оси y с набегом фазы δ_y, то образуется двумерная решетка с AF, равным

$$\mathrm{AF}(\varphi,\theta) = \left| \sum_{n_y=0}^{N_y-1} \sum_{n_x=0}^{N_x-1} e^{jn_x(kd\cos\theta\sin\varphi+\delta_x)} e^{jn_y(kd\sin\theta+\delta_y)} \right|. \tag{3}$$

Чтобы сформировать узкий луч, необходимо максимизировать значение $\mathrm{AF}(\varphi,\theta)$ для заданных φ, θ. Для достижения максимального значения (3) необходимо, чтобы значения δ_x и δ_y были равны

$$\delta_x = -kd\cos\theta\sin\varphi \pm 2\pi r_x, \quad r_x = 0,1,2,\ldots, \tag{4}$$

$$\delta_y = -kd\sin\theta \pm 2\pi r_y, \quad r_y = 0,1,2,\ldots \tag{5}$$

Преобразуем φ,θ в $\widehat{\varphi},\theta$ так, чтобы δ_x формально не зависел от θ. Для этого введем

$$\widehat{\varphi} = g_\theta(\varphi) = \arcsin(\sin\varphi\cos\theta). \tag{6}$$

Подставив $\sin\widehat{\varphi}$ в (4), получим

$$\delta_x = -kd\sin\widehat{\varphi} \pm 2\pi r_x. \tag{7}$$

Теперь набег фазы вдоль оси x/y формально зависит только от соответствующего угла φ/θ. Поскольку угол $\widehat{\varphi}$ введен, чтобы заменить φ, преобразование φ в $\widehat{\varphi}$ следует рассматривать как параметрическое преобразование с θ в качестве параметра.

Обозначим векторы конфигурации одномерных RIS вдоль осей x и y через $\boldsymbol{\alpha}^{\varphi} \in \mathbb{C}^{N_x}$ и $\boldsymbol{\alpha}^{\theta} \in \mathbb{C}^{N_y}$ соответственно. В свою очередь, вектор конфигурации двумерной RIS в плоскости xy обозначим через $\boldsymbol{\alpha}^{\varphi,\theta} \in \mathbb{C}^{N_xN_y}$. В результате сделаем следующее

Утверждение. *Чтобы сформировать узкий луч с конфигурацией $\boldsymbol{\alpha}^{\varphi,\theta}$ в направлении (φ,θ), достаточно сформировать две независимые конфигурации RIS $\boldsymbol{\alpha}^{\widehat{\varphi}}$ и $\boldsymbol{\alpha}^{\theta}$ в соответствующих направлениях $(\widehat{\varphi},0)$ и $(0,\theta)$ согласно* (6). *Тогда конечная конфигурация вычисляется как*

$$\boldsymbol{\alpha}^{\varphi,\theta} = \mathrm{vec}\left((\boldsymbol{\alpha}^{\theta})^T \boldsymbol{\alpha}^{\widehat{\varphi}}\right),$$

где

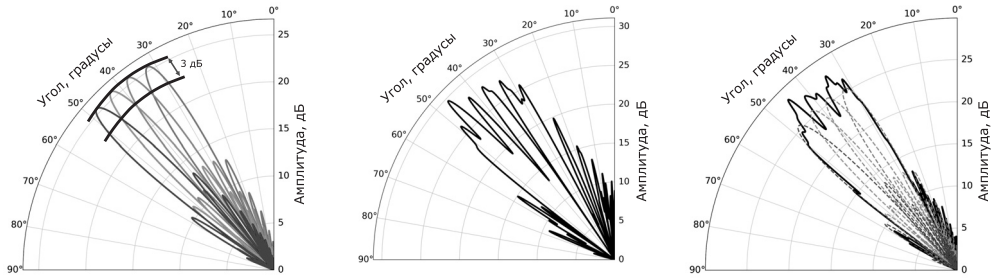
$$\boldsymbol{\alpha}^{\theta} = (0,\delta_y,\ldots,(N_y-1)\delta_y), \quad \boldsymbol{\alpha}^{\widehat{\varphi}} = (0,\delta_x,\ldots,(N_x-1)\delta_x).$$

Дополнительно введем вектор конфигурации RIS $\boldsymbol{\alpha}^{\varphi,\theta,\psi}$, в котором фаза всех элементов повернута на одинаковый угол ψ. Отметим, что добавление ψ никак не влияет на AF луча, но изменяет его фазу, что позволяет контролировать суперпозицию нескольких лучей.

3.2. Синтез широких одномерных лучей. Сначала рассмотрим синтез широкого одномерного луча, например, в оси φ. Широкий одномерный луч полностью определяется двумя параметрами: ширина W^{φ} и угол центра φ^c.

Неформально целью синтеза широкого луча является обеспечение высоких значений AF внутри области широкого луча

$$\Phi_{\mathrm{beam}} = \left[\varphi^c - \frac{W^{\varphi}}{2}, \varphi^c + \frac{W^{\varphi}}{2}\right]$$



а) *б*) *в*)

Рис. 2. Диаграмма широкого одномерного луча с параметрами $W^{\varphi} = 20°$, $\varphi^c = 40°$ для RIS с $N = 100$, $b = 4$: *а*) диаграмма узких лучей, сшитых по амплитуде; *б*) диаграмма широкого луча без сшивки по фазе; *в*) диаграмма широкого луча со сшивкой по фазе

и низких значений вне этой области

$$\Phi_{\text{sl}} = \Big[-\frac{\pi}{2}, \varphi^c - \frac{W^{\varphi}}{2}\Big) \cup \Big(\varphi^c + \frac{W^{\varphi}}{2}, \frac{\pi}{2}\Big].$$

Для формирования широкого луча RIS разделяется на M сегментов, где каждый сегмент формирует свой собственный узкий луч так, что направления центров лучей соседних сегментов сдвинуты (сшиты) по порогу половины мощности (англ.: half power beam width, HPBW) узких лучей. Разбиение на сегменты происходит так, чтобы все M узких лучей имели одинаковые HPBW и пиковые значения AF. Описанная схема представлена на рис. 2,*а*).

Для определения значения M используется тот факт [19], что HPBW для одномерного RIS с N ЭЯ равен

$$W_0 = \frac{0{,}886\lambda}{Nd\cos\varphi^c}. \tag{8}$$

Таким образом, после разделения RIS на M равных частей суммарная ширина луча W^{φ} равна

$$W^{\varphi} \approx \frac{0{,}886\lambda}{\frac{N}{M}d\cos\varphi^c} M = \frac{0{,}886\lambda}{Nd\cos\varphi^c} M^2. \tag{9}$$

Поэтому для требуемого значения W^{φ} необходимое число сегментов равно

$$M = \left\lceil \sqrt{W^{\varphi} \frac{Nd\cos\varphi^c}{0{,}886\lambda}} \right\rceil. \tag{10}$$

Операция округления применяется для того, чтобы сформировать луч *не* уже, чем требуемое W^{φ}.

Теперь необходимо сшить полученные узкие лучи для формирования широкого луча. Сшивка состоит из двух этапов.

На первом этапе узкие лучи сшиваются по порогу HPBW. Для этого центры M узких лучей располагаются равномерно внутри области широкого луча. Обозначим центры узких лучей через φ_i^c, $i = 1, \ldots, M$:

$$\varphi_i^c = \varphi^c - \frac{M/2 - i}{M} W^{\varphi}, \quad i = 1, \ldots, M. \tag{11}$$

Теперь M сегментов с $N_i = \dfrac{N}{M}$ ЭЯ в каждом формируют M узких подлучей в различных направлениях (11). Следовательно, согласно (8) их HPBW немного отличаются. Для компенсации этого эффекта производится балансировка ЭЯ между сегментами, чтобы выровнять амплитуды и HPBW их узких лучей:

$$N_i \cos\varphi_i^c = \text{const}, \quad i = 1, \ldots, M.$$

Это приводит к тому, что число ЭЯ в каждом сегменте равно

$$N_i = \left[\frac{N}{M} \frac{\cos\varphi_i^c}{\sum\limits_{i=1}^{M} \cos\varphi_i^c} \right], \quad i = 1, \ldots, M. \tag{12}$$

В результате получены M узких лучей с одинаковыми амплитудами и HPBW, центры φ_i^c которых обеспечивают амплитудную сшивку по порогу -3 дБ. На рис. 2,*б*) представлен пример такой сшивки.

Можно заметить, что взаимная интерференция между узкими лучами может приводить к сильным пульсациями внутри широкого луча и высоким боковым лепесткам. Для борьбы с интерференцией на втором этапе производится фазовая сшивка узких лучей. Фазовая сшивка обеспечивается упомянутой ранее добавленной фазой ψ для каждого узкого луча.

Обозначим вектор фаз узких лучей через $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_M)$. Тогда соответствующие векторы конфигурации каждого сегмента обозначаются как $\boldsymbol{\alpha}^{\varphi_i^c, \psi_i}$, а полный вектор конфигурации – как

$$\boldsymbol{\alpha}^{W^\varphi, \varphi^c, \boldsymbol{\psi}} = \left[\boldsymbol{\alpha}^{\varphi_1^c, \psi_1}, \ldots, \boldsymbol{\alpha}^{\varphi_M^c, \psi_M} \right].$$

Рассмотрим коэффициент усиления одномерного RIS $\mathrm{AF}(\varphi)$ как функцию от $\boldsymbol{\alpha}$: $\mathrm{AF}[\boldsymbol{\alpha}](\varphi)$. Тогда фазовая сшивка может быть выполнена с минимизацией следующей функции потерь $f_{\text{loss}}(\boldsymbol{\alpha}^{W^\varphi, \varphi^c, \boldsymbol{\psi}})$:

$$\begin{aligned} \min_{\boldsymbol{\psi}} f_{\text{loss}}\left(\boldsymbol{\alpha}^{W^\varphi, \varphi^c, \boldsymbol{\psi}}\right) &= \gamma \,\mathrm{std}\left(\left|\mathrm{AF}\left[\boldsymbol{\alpha}^{W^\varphi, \varphi^c, \boldsymbol{\psi}}\right](\Phi_{\text{beam}})\right|\right) - \\ &- (1-\gamma)\left(\min\left(\left|\mathrm{AF}\left[\boldsymbol{\alpha}^{W^\varphi, \varphi^c, \boldsymbol{\psi}}\right](\Phi_{\text{beam}})\right|\right) - \max\left(\left|\mathrm{AF}\left[\boldsymbol{\alpha}^{W^\varphi, \varphi^c, \boldsymbol{\psi}}\right](\Phi_{\text{sl}})\right|\right)\right), \end{aligned} \tag{13}$$

где $\mathrm{std}(\cdot)$ — стандартное отклонение $\mathrm{AF}[\boldsymbol{\alpha}]$ в области Φ, $\max(\cdot)$ — максимальное значение $\mathrm{AF}[\boldsymbol{\alpha}]$ в области Φ. Отметим, что величина $\mathrm{std}(\mathrm{AF}[\boldsymbol{\alpha}^{W^\varphi, \varphi^c, \boldsymbol{\psi}}](\Phi_{\text{beam}}))$ отражает уровень пульсаций внутри широкого луча, а оставшаяся часть отражает уровень боковых лепестков относительно луча. В результате минимизации f_{loss}, уровень пульсаций широкого одномерного луча значительно снижается, как показано на рис. 2,*в*). Кроме того, амплитуда боковых лепестков также снижается.

Задача (13) является невыпуклой [20], поэтому предлагается решать ее одним из двух подходов. Первым является жадный алгоритм (англ.: greedy algorithm, GA), при котором поиск осуществляется по конечному множеству Ψ для каждого ψ_i. Алгоритм GA для каждого ψ_i перебирает все значения ψ из Ψ и выбирает то ψ, которое минимизирует (13). Отметим, что этот алгоритм может более одного раза проходить по всем ψ_i для обеспечения еще лучшего результата.

Вторым из этих подходов является случайное семплирование (англ.: random sampling, RS), при котором случайным образом семплируются векторы $\boldsymbol{\psi}$ из Ψ^M, и в качестве приближенного решения выбирается тот, который минимизирует (13). Полностью синтез широкого одномерного луча с применением GA представлен в алгоритме 1.

Алгоритм 1 Синтез широкого одномерного луча

Вход: ширина луча W^{φ}, центр луча φ^c.
Выход: вектор конфигурации $\boldsymbol{\alpha}^{\star}$.

Инициализация: нулевой вектор $\boldsymbol{\alpha}^{\star}$, $Q \leftarrow f_{\text{loss}}(\boldsymbol{\alpha}^{\star})$, нулевой вектор $\boldsymbol{\psi}$.
Вычислить число узких лучей M через (10).
Вычислить центры узких лучей φ_i^c, $i = 1, \ldots, M$, используя (11).
Вычислить число ЭЯ N_i для каждого сегмента, используя (12), $i = 1, \ldots, M$. /* Найти приближенное решение $\boldsymbol{\psi}^{\star}$ для (13), используя GA, и построить соответствующую конфигурацию RIS $\boldsymbol{\alpha}^{\star}$. */
for $i \in 2 \ldots M$ **do**
    $\psi^{\star} \leftarrow 0$.
    for $j \in 1 \ldots |\Psi|$ **do**
        $\psi_i \leftarrow \Psi_j$.
        $\boldsymbol{\alpha}^{W^{\varphi}, \varphi^c, \boldsymbol{\psi}} \leftarrow [\boldsymbol{\alpha}^{\varphi_1^c, \psi_1}, \ldots, \boldsymbol{\alpha}^{\varphi_M^c, \psi_M}]$.
        if $f_{\text{loss}}(\boldsymbol{\alpha}^{W^{\varphi}, \varphi^c, \boldsymbol{\psi}}) < Q$ **then**
            $Q \leftarrow f_{\text{loss}}(\boldsymbol{\alpha}^{W^{\varphi}, \varphi^c, \boldsymbol{\psi}})$.
            $\psi^{\star} \leftarrow \Psi_j$.
    $\psi_i \leftarrow \psi^{\star}$.
return: $\boldsymbol{\alpha}^{\star}$.

Сравнение GA и RS рассматривается в § 4.

Синтез широкого двумерного луча. Широкий двумерный луч полностью определяется следующими параметрами: W^{φ} — ширина луча в оси φ, W^{θ} — ширина луча в оси θ, φ^c и θ^c — направление центра луча. Чтобы оставить вычислительную сложность синтеза такой же, как в случае широкого одномерного луча, предлагается формировать двумерный луч как композицию двух одномерных, одного в оси φ, а другого в оси θ.

Для этого воспользуемся утверждением, представленным в п. 3.1, где показано, что конфигурация узкого луча $\boldsymbol{\alpha}^{\varphi,\theta}$ для двумерного RIS может рассматриваться как произведение двух конфигураций одномерного RIS $\boldsymbol{\alpha}^{\varphi,\theta} = \text{vec}\left((\boldsymbol{\alpha}^{\theta})^T \boldsymbol{\alpha}^{\widehat{\varphi}}\right)$, где $\boldsymbol{\alpha}^{\widehat{\varphi}}$ – вектор конфигурации одномерного RIS для угла

$$\widehat{\varphi} = g_{\theta}(\varphi) = \arcsin(\sin\varphi\cos\theta).$$

Этот же подход можно применить и к широким лучам, так как широкие лучи формируются из множества узких лучей.

Однако в случае широких лучей преобразование φ^c с помощью (6) оказывается недостаточным, так как $g_{\theta}(\cdot)$ является нелинейной операцией относительно φ, и поэтому исходный геометрический центр луча не преобразуется в геометрический центр преобразованного широкого луча. Вместо преобразования φ^c и W^{φ} предлагается оперировать с левой φ^l и правой φ^r границами широкого луча, которые вычисляются как

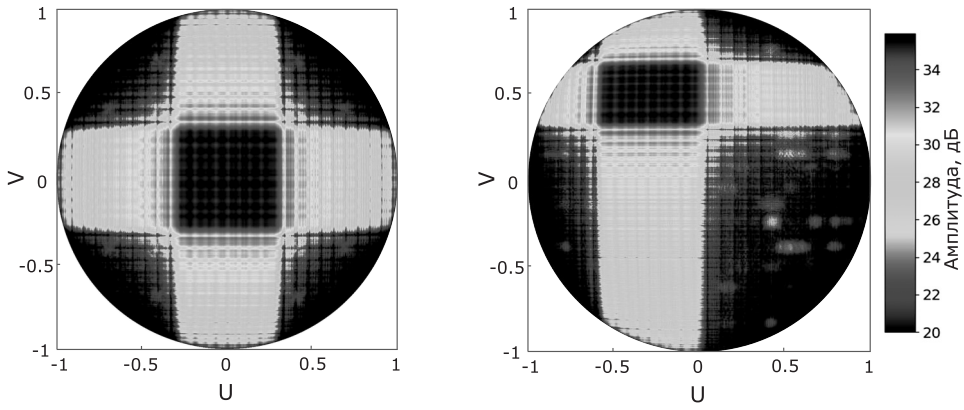
$$\varphi^l = \varphi^c - \frac{W^{\varphi}}{2}, \tag{14}$$

$$\varphi^r = \varphi^c + \frac{W^{\varphi}}{2}. \tag{15}$$

После преобразования φ^l и φ^r новые значения $\widehat{\varphi^c}$ и $\widehat{W^{\varphi}}$ вычисляются как

$$\widehat{\varphi^c} = \frac{g_{\theta}(\varphi^l) + g_{\theta}(\varphi^r)}{2}, \tag{16a}$$

$$\widehat{W^{\varphi}} = \frac{g_{\theta}(\varphi^l) - g_{\theta}(\varphi^r)}{2}. \tag{16b}$$



а) *б)*

Рис. 3. Диаграммы широких двумерных лучей с различными параметрами: *а)* $W^{\varphi} = W^{\theta} = 30°$, $\varphi^c = \theta^c = 0°$; *б)* $W^{\varphi} = 20$, $W^{\theta} = 40°$, $\varphi^c = 20$, $\theta^c = 30°$

В результате вектор конфигурации для широкого двумерного луча с требуемыми $(W^{\varphi}, W^{\theta}, \varphi^c, \theta^c)$ вычисляется как

$$\boldsymbol{\alpha}^{W^{\varphi},\varphi^c,W^{\theta},\theta^c} = \mathrm{vec}\Big(\big(\boldsymbol{\alpha}^{W^{\theta},\theta^c}\big)^T \boldsymbol{\alpha}^{\widehat{W^{\varphi}},\widehat{\varphi^c}}\Big). \tag{17}$$

Примеры диаграмм широкого двумерного луча для RIS с $N = 200 \times 200$, $b = 4$ представлены на рис. 3. Можно заметить, что диаграмма имеет клеточную структуру, что обусловлено наличием, хоть и малых, но пульсаций в одномерных лучах.

Полностью синтез широкого двумерного луча представлен в алгоритме 2.

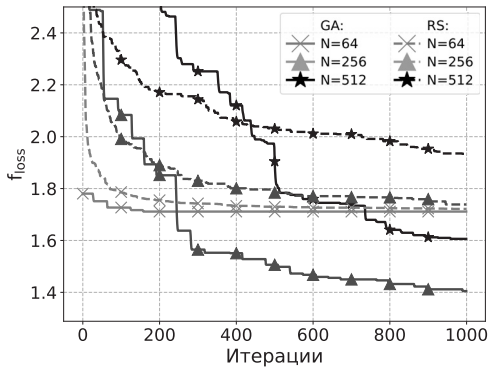
Алгоритм 2 Синтез широкого двумерного луча

Вход: ширины лучей и их центры в осях φ и θ: $W^{\varphi}, \varphi^c, W^{\theta}, \theta^c$.
Выход: вектор конфигурации $\boldsymbol{\alpha}^{\star}$.
1: Вычислить границы лучей φ^l, φ^r, используя (3.2).
2: Вычислить $\widehat{\varphi^c}$ и $\widehat{W^{\varphi}}$, используя (16).
3: Вычислить $\boldsymbol{\alpha}^{\widehat{W^{\varphi}},\widehat{\varphi^c}}, \boldsymbol{\alpha}^{W^{\theta},\theta^c}$, используя алгоритм 1.
4: $\boldsymbol{\alpha}^{\star} \leftarrow \mathrm{vec}\big((\boldsymbol{\alpha}^{W^{\theta},\theta^c})^T \boldsymbol{\alpha}^{\widehat{W^{\varphi}},\widehat{\varphi^c}}\big)$.
5: **return:** $\boldsymbol{\alpha}^{\star}$.

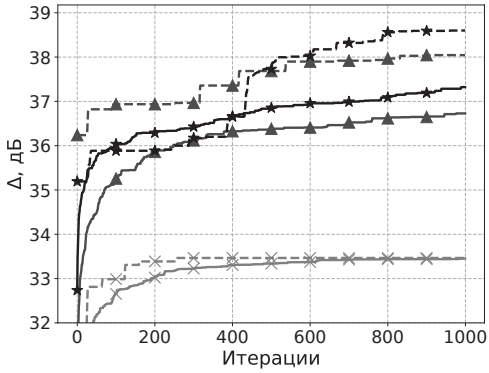
§ 4. Численные результаты

В данном параграфе проводится сравнение синтеза широких одномерных лучей с помощью GA и RS на основе различных метрик, изображенных на рис. 4, причем сплошные кривые соответствуют GA, а штрихпунктирные – RS.

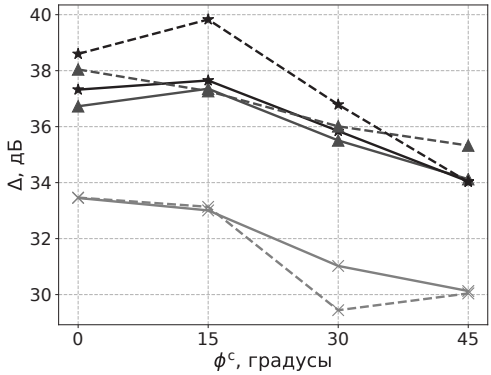
На рис. 4,*а*) представлена зависимость f_{loss} с $\gamma = 0,9$ от числа итераций (количества вычислений диаграмм луча) для широкого одномерного луча с $W^{\varphi} = 20°$, $\varphi^c = 0°$. Для малых RIS с количеством элементов равным $N = 64$ GA превосходит RS при любом числе итераций. Однако при $N = 256$ или $N = 512$ GA проигрывает RS при малом числе итераций, но потом GA начинает выигрывать. Это связано с тем, что большее число ЭЯ N приводит к большему числу подлучей M, и GA не успевает оптимизировать все подлучи.



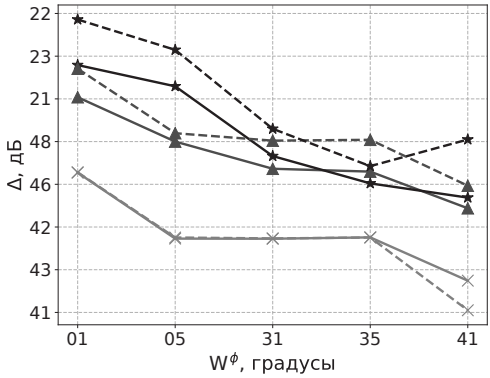
а) Зависимость f_{loss} от числа итераций



б) Зависимость Δ от числа итераций



в) Зависимость Δ от φ^c после 1000 итераций



г) Зависимость Δ от W^{φ} после 1000 итераций

Рис. 4. Синтеза луча с помощью RS и GA

В дополнение к f_{loss} также рассматривается альтернативная метрика качества луча $\Delta = \text{mean}(\text{AF}(\Phi_{\text{beam}})) - \text{mean}(\text{AF}(\Phi_{\text{sl}}))$, которая отражает разницу между средней мощностью луча и боковых лепестков, что показано на рис. 4,*б*). Поведение Δ похоже на поведение f_{loss}: при большом размере RIS и малом количестве итераций RS превосходит GA, в остальных случаях GA демонстрирует лучший результат.

Наконец, рассмотрим результат синтеза для различных φ^c и W^{φ} после 1000 итераций. Как показано на рис. 4,*в*), в среднем для различных φ^c и $N > 64$ метод GA обеспечивает лучшее Δ. Однако для малого RIS с $N = 64$ RS обходит GA. В то же время для больших φ^c, например, $\varphi^c = 45°$, эффективность RS и GA становится близкой. Эти же свойства остаются верными и для различных W^{φ}, изображенных на рис. 4,*г*).

В результате можно сделать следующие эмпирические рекомендации по выбору GA и RS. Если RIS состоит из большого числа ЭЯ и время синтеза сильно ограничено, то следует использовать RS. В иных случаях следует выбирать GA.

§ 5. Заключение

В данной статье рассмотрена задача синтеза широких двумерного и одномерного лучей с помощью RIS. Предложенный метод обладает низкой вычислительной сложностью и позволяет формировать широкий луч с низкими уровнями пульса-

ций и боковых лепестков. Кроме того, эти методы могут быть применимы на RIS с простейшей архитектурой, подразумевающей, что ЭЯ обладают дискретной фазой и постоянной амплитудой.

СПИСОК ЛИТЕРАТУРЫ

1. Framework and Overall Objectives of the Future Development of IMT for 2030 and Beyond. International Telecommunication Union (ITU) Recommendation (ITU-R), 2023.
2. *Hua M., Wu Q., Chen W., Dobre O.A., Swindlehurst A.L.* Secure Intelligent Reflecting Surface-Aided Integrated Sensing and Communication // IEEE Trans. Wirel. Commun. 2023. V. 23. № 1. P. 575–591. https://doi.org/10.1109/TWC.2023.3280179
3. *Гао В., Башконуш Х.М., Каттани К.* Эффективность передачи данных при атаках с точки зрения варианта изолированной жесткости // Пробл. передачи информ. 2023. Т. 59. № 2. С. 83–101. https://doi.org/10.31857/S0555292323020067
4. *Kim I.-S., Bennis M., Oh J., Chung J., Choi J.* Bayesian Channel Estimation for Intelligent Reflecting Surface-Aided mmWave Massive MIMO Systems with Semi-Passive Elements // IEEE Trans. Wirel. Commun. 2023. V. 22. № 12. P. 9732–9745. https://doi.org/10.1109/TWC.2023.3273284
5. *Wang P., Fang J., Duan H., Li H.* Compressed Channel Estimation for Intelligent Reflecting Surface-Assisted Millimeter Wave Systems // IEEE Signal Process. Lett. 2020. V. 27. P. 905–909. https://doi.org/10.1109/LSP.2020.2998357
6. *Фернандес М., Кабатянский Г.А., Круглик С.А., Мяо И.* Коды для точного нахождения носителя разреженного вектора по ошибочным линейным измерениям и их декодирование // Пробл. передачи информ. 2023. Т. 59. № 1. С. 17–24. https://doi.org/10.31857/S0555292323010023
7. *Yan G., Zhu L., Zhang R.* Channel Autocorrelation Estimation for IRS-Aided Wireless Communications Based on Power Measurements // Proc. 2023 IEEE Globecom Workshops (GC Wkshps). Kuala Lumpur, Malaysia. Dec. 4–8, 2023. P. 1457–1462. https://doi.org/10.1109/GCWkshps58843.2023.10465210
8. *Zhang H., Shlezinger N., Guidi F., Dardari D., Imani M.F., Eldar Y.C.* Beam Focusing for Near-Field Multiuser MIMO Communications // IEEE Trans. Wirel. Commun. 2022. V. 21. № 9. P. 7476–7490. https://doi.org/10.1109/TWC.2022.3158894
9. *Chen Y., Dai L.* Coded Beam Training for RIS Assisted Wireless Communications, https://www.arxiv.org/abs/2406.15802 [cs.IT], 2024.
10. *Wu C., You C., Liu Y., Chen L., Shi S.* Two-Stage Hierarchical Beam Training for Near-Field Communications // IEEE Trans. Veh. Technol. 2024. V. 73. № 2. P. 2032–2044. https://doi.org/10.1109/TVT.2023.3311868
11. *Bagheri A., Safaei M., Araghi A., Mahdi Shahabi S.M., Wang F., Khalily M., Tafazolli R.* Mathematical Model and Real-World Demonstration of Multi-Beam and Wide-Beam Reconfigurable Intelligent Surface // IEEE Access. 2023. V. 11. P. 19613–19621. https://doi.org/10.1109/ACCESS.2023.3248501
12. *Ghanem W.R., Jamali V., Schellmann M., Cao H., Eichinger J., Schober R.* Optimization-Based Phase-Shift Codebook Design for Large IRSs // IEEE Commun. Lett. 2022. V. 27. № 2. P. 635–639. https://doi.org/10.1109/LCOMM.2022.3225585
13. *Gerchberg R.W., Saxton W.O.* A Practical Algorithm for the Determination of Phase from Image and Diffraction Plane Picture // Optik. 1972. V. 35. № 2. P. 237–246.
14. *Lu Y., Zhang Z., Dai L.* Hierarchical Beam Training for Extremely Large-Scale MIMO: From Far-Field to Near-Field // IEEE Trans. Commun. 2024. V. 72. № 4. P. 2247–2259. https://doi.org/10.1109/TCOMM.2023.3344600
15. *Torkzaban N., Khojastepour M.A., Farajzadeh-Tehrani M., Baras J.S.* RIS-Aided mmWave Beam-Forming for Two-Way Communications of Multiple Pairs // ITU J. Future Evol. Technol. 2023. V. 4. № 1. P. 87–101. https://doi.org/10.52953/VBEX2484
16. *Shen D., Dai L., Su X., Suo S.* Multi-Beam Design for Near-Field Extremely Large-Scale RIS-Aided Wireless Communications // IEEE Trans. Green Commun. Netw. 2023. V. 7. № 3. P. 1542–1553. https://doi.org/10.1109/TGCN.2023.3259579

17. *Jamali V., Najafi M., Schober R., Vincent Poor H.* Power Efficiency, Overhead, and Complexity Tradeoff of IRS Codebook Design—Quadratic Phase-Shift Profile // IEEE Commun. Lett. 2021. V. 25. № 6. P. 2048–2052. https://doi.org/10.1109/LCOMM.2021.3058063
18. *Balanis C.A.* Antenna Theory: Analysis and Design. Hoboken, NJ: Wiley, 2016.
19. *Björnson E., Demir Ö.T., Sanguinetti L.* A Primer on Near-Field Beamforming for Arrays and Reconfigurable Intelligent Surfaces // Conf. Record: 55th Asilomar Conf. on Signals, Systems & Computers. Pacific Grove, CA, USA. Oct. 31 – Nov. 3, 2021. P. 105–112. https://doi.org/10.1109/IEEECONF53345.2021.9723331
20. *Zhang S., Zhang R.* Capacity Characterization for Intelligent Reflecting Surface Aided MIMO Communication // IEEE J. Sel. Areas Commun. 2020. V. 38. № 8. P. 1823–1838. https://doi.org/10.1109/JSAC.2020.3000814

Пойда Арсений Игоревич
Институт проблем передачи информации
им. А.А. Харкевича РАН, Москва
poyda@wnlab.ru
Буртаков Илья Андреевич
Институт проблем передачи информации
им. А.А. Харкевича РАН, Москва
Московский физико-технический институт
(национальный исследовательский университет), Москва
burtakov@wnlab.ru
Куреев Алексей Андреевич
Хоров Евгений Михайлович
Национальный исследовательский университет
"Высшая школа экономики", Москва
Институт проблем передачи информации
им. А.А. Харкевича РАН, Москва
kureev@wnlab.ru
khorov@wnlab.ru

Поступила в редакцию
26.08.2024
После доработки
26.11.2024
Принята к публикации
29.11.2024

ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 60 **2024** **Вып. 3**

УДК 621.391 : 519.7

© 2024 г. **М.Н. Вялый, А.А. Рубцов**

ЗАДАЧИ РЕГУЛЯРНОЙ РЕАЛИЗУЕМОСТИ ДЛЯ ОПИСАНИЙ КОНЕЧНЫХ ОТНОШЕНИЙ[1]

Рассмотрены описания конечных отношений на множестве неотрицательных целых чисел в формате, предложенном П. Вольф и Х. Фернау, и связанные с ними задачи регулярной реализуемости. Доказана универсальность этих задач относительно дизъюнктных сводимостей за полиномиальное время для унарных отношений; относительно дизъюнктных сводимостей на полиномиальной памяти для инвариантных бинарных отношений; а также относительно сводимостей по Тьюрингу с NP-оракулом для инвариантных унарных отношений, заданных в унарном алфавите.

Ключевые слова: регулярная реализуемость, сводимость, NP-оракул.

DOI: 10.31857/S0555292324030069, **EDN:** ZXXKCJ

§ 1. Введение

Класс регулярных языков является одним из важнейших объектов изучения в теории формальных языков. Хорошо известно, что основные алгоритмические задачи теории формальных языков – проверка пустоты, конечности, принадлежности слова регулярному языку – разрешимы, и более того, разрешимы за полиномиальное время, если регулярный язык задан описанием конечного автомата, распознающего язык.

Однако регулярные условия, накладываемые на слова некоторого языка, могут оказаться гораздо сложнее. Соответствующая алгоритмическая задача называется *задачей регулярной реализуемости*. Это целый класс задач, параметризованный языками. Зафиксируем некоторый язык F, называемый *фильтром*. Входом задачи регулярной реализуемости является регулярный язык $L(\mathcal{A})$, распознаваемый автоматом $\mathcal{A}$. Необходимо проверить выполнение условия $L(\mathcal{A}) \cap F \neq \varnothing$. Задачу регулярной реализуемости обозначаем $\mathrm{DRR}(F)$ или $\mathrm{NRR}(F)$ в зависимости от вида автомата $\mathcal{A}$: детерминированный или недетерминированный.

На задачи регулярной реализуемости можно смотреть как на задачи достижимости в графе при ограничении на пути: считаем, что ребра графа помечены символами и интересуемся существованием пути, вдоль которого прочитывается слово из некоторого языка. Наиболее изучена такая задача для ограничений в виде КС языка [1–4]. В этом случае задача регулярной реализуемости разрешима за полиномиальное время.

Общая формулировка задачи регулярной реализуемости дана в [5]. Там же приведены примеры фильтров, для которых задачи регулярной реализуемости полны для некоторых типичных классов сложности или для всего класса перечислимых языков.

[1] Работа выполнена при поддержке Российского научного фонда (грант № 20-11-20203).

Более общим вопросом, чем существование алгоритмически трудных задач, является вопрос об "универсальности" задач некоторого класса среди всех возможных задач разрешения. Более точно, *свойство универсальности* класса языков $\mathscr{C}$ относительно сводимостей $(\leqslant_1, \leqslant_2)$ состоит в том, что для любого непустого языка X существует такой язык $L_X \in \mathscr{C}$, что

$$X \leqslant_1 L_X \leqslant_2 X.$$

Получаем целое семейство свойств универсальности, в котором чем слабее сводимости, тем сильнее свойство. С точки зрения теории алгоритмов имеет смысл рассматривать сводимости, которые не сильнее сводимостей по Тьюрингу. Для применений к теории вычислительной сложности требуются более слабые сводимости.

В [6] доказано свойство универсальности класса DRR относительно пары сводимостей $(\leqslant_{\mathrm{m}}^{\log}, \leqslant_{\mathrm{dtt}}^{\mathrm{FNL}})$ (определения см. в п. 2.2 и в работе [6]). В той же работе доказано, что для каждого класса полиномиальной иерархии существует задача DRR, полная в этом классе относительно полиномиальных сводимостей. Заметим также, что результаты [6] легко переносятся на NRR.

Вопрос об универсальности имеет смысл и для задач регулярной реализуемости с ограниченными классами фильтров. Такого рода результаты можно интерпретировать как универсальность подходящих моделей вычисления. В [6] доказана универсальность DRR относительно той же пары сводимостей, что и в общем случае, для префиксно-замкнутых фильтров. Этот результат говорит о представительности модели автоматов с обобщенным недетерминизмом (см. [7]). Этот результат был уточнен в [8], где была доказана универсальность задачи регулярной реализуемости для фильтров, образованных корректными протоколами работы конечных автоматов с дополнительными структурами данных относительно пары сводимостей $(\leqslant_{\mathrm{m}}^{\log}, \leqslant_{\mathrm{T}}^{\mathrm{P}})$, где $\leqslant_{\mathrm{T}}^{\mathrm{P}}$ – полиномиальные сводимости по Тьюрингу. Отсюда следует, что любой достаточно сильный класс сложности можно описать как класс языков, распознаваемых машинами Тьюринга на логарифмической памяти, к которым добавили структуру данных, соответствующую языку протоколов.

В этой статье мы уточняем границу между классами фильтров со свойствами универсальности и без них. На этот раз мотивация выбора класса фильтров другая и не связана с моделями вычислений.

Задачи регулярной реализуемости (под другим названием – regular intersection emptiness problems) изучались в работах [9, 10]. Там изучен такой круг вопросов: зафиксируем некоторое свойство графов и способ описания графов в виде слов конечного алфавита. Интересует существование графа, обладающего данным свойством, одно из описаний которого удовлетворяет некоторому регулярному ограничению. Другими словами, на вход задачи о графах подается не один граф, а множество графов, описания которых удовлетворяют некоторому регулярному условию – входу задачи NRR(F). Интересует существование хотя бы одного графа в этом множестве, обладающего искомым свойством (которое задает фильтр F). При такой постановке крайне существен способ задания графа. В [9] предложен очень интересный способ кодировки, при котором вершины графа – натуральные числа, а ребра – пары чисел, причем числа записываются в унарной системе. Самой важной особенностью этого способа представления является свобода в выборе порядка записи ребер и возможность повторения элементов списка. Для задач регулярной реализуемости это означает отсутствие "лишних" трудностей, которые могут возникнуть при фиксированном порядке записи ребер (например, см. [8], где порядок записи блоков в протоколе существен).

Для такого способа представления графов в [9] доказана разрешимость задач регулярной реализуемости для многих типичных для теории графов свойств. Ряд родственных результатов получен в работах [10, 11].

Неформально говоря, результаты из [9–11] указывают на структурную простоту выбранного в этих работах способа описания графов. Поэтому неочевиден ответ на вопрос об универсальности задач регулярной реализуемости для таких описаний: раз уж свойства графов легко поддаются алгоритмическому анализу при выбранном способе описания, то строить трудные примеры должно быть сложнее.

Тем не менее в данной статье получены результаты об универсальности задач регулярной реализуемости для описаний графов и унарных отношений на множестве натуральных чисел. Эти результаты используют более сильные сводимости, чем общая теорема об универсальности, что отражает специфику данного класса задач регулярной реализуемости.

При получении таких результатов одним из основных препятствий является свобода в выборе порядка записи элементов множества. Для преодоления этой трудности приходится использовать довольно сложные и нетипичные для теории формальных языков конструкции – эффективные асимптотически хорошие коды. Мы также вводим некоторую новую модель вычисления булевых функций (или, что то же самое, семейств подмножеств некоторого фиксированного множества).

Статья организована следующим образом. В § 2 приводятся точные формулировки неформально поставленных выше задач и напоминаются основные определения, относящиеся к автоматам и алгоритмическим сводимостям. В § 3 описано основное техническое средство анализа задач регулярной реализуемости для таких способов описания: переход к автоматам в алфавите $\mathbb{N}^k$, где k – арность рассматриваемых отношений[2]. Задание таких автоматов опирается на структурную простоту регулярных языков в унарном алфавите.

В § 4 описан наш подход к получению результатов универсальности для различных классов фильтров. Для применения этого подхода к интересующему нас случаю описаний конечных отношений требуется, как уже говорилось, анализ общей задачи представления семейств подмножеств заданного множества (“универсума”) ориентированными графами. Эта задача формулируется в § 5. Именно при построении разделенных семейств множеств возникает необходимость в эффективных асимптотически хороших кодах.

Поскольку основная мотивация статьи связана с задачами регулярной реализуемости для теоретико-графовых свойств, мы рассматриваем отдельно классы отношений на $\mathbb{N}$, инвариантные относительно биекций $\mathbb{N}$. Ведь обычно под свойством графов понимают свойство, которое сохраняется при изоморфизме. Доказательство универсальности для инвариантных свойств удается получить только для весьма сильных сводимостей (дизъюнктные сводимости на полиномиальной памяти).

В § 6 содержатся основные результаты статьи: универсальность описаний унарных отношений (т.е. конечных подмножеств $\mathbb{N}$), универсальность инвариантных бинарных отношений и универсальность инвариантных унарных отношений. Последний результат доказан только для языков в унарном алфавите, поскольку единственный инвариант конечного подмножества – это его размер, что не позволяет строить короткие кодировки двоичных слов числами в унарной системе.

В заключительном параграфе обсуждаются результаты этой статьи и перспективы дальнейших исследований в этом направлении.

Расширенный вариант этой статьи см. в [12].

§ 2. Определения и основные конструкции

Напомним основные определения, связанные с автоматами и трансдьюсерами, сводимостями и классами сложности, а также зафиксируем обозначения. В этом

[2] Под натуральными числами мы понимаем целые неотрицательные числа.

параграфе также фиксируется точный формат описания конечных отношений, используемый в настоящей статье.

2.1. Автоматы и трансдьюсеры. (Недетерминированный) автомат $\mathcal{A}$ в алфавите Σ задается конечным множеством состояний Q, начальным состоянием q_0, множеством принимающих состояний Q_a и отношением переходов

$$\delta_{\mathcal{A}} \subseteq Q \times (\{\varepsilon\} \cup \Sigma) \times Q,$$

где ε – пустое слово. Отношение переходов продолжается до отношения на $Q \times \Sigma^* \times Q$ естественным правилом: $(q_1, w\sigma, q_2) \in \delta_{\mathcal{A}}$ равносильно тому, что существует такое q_3, что $(q_1, w, q_3) \in \delta_{\mathcal{A}}$ и $(q_3, \sigma, q_2) \in \delta_{\mathcal{A}}$. Здесь $w \in \Sigma^*$, $\sigma \in \{\varepsilon\} \cup \Sigma$.

Обратите внимание, что мы допускаем бесконечные алфавиты для автоматов. Это нестандартное определение оказывается здесь удобным. Если алфавит конечный, то возможно задание автомата списком элементов отношения переходов. В случае бесконечного алфавита задание отношения переходов оговаривается особо (см. пример в § 3 ниже).

Слово w *принимается* автоматом $\mathcal{A}$, если $(q_0, w, q_a) \in \delta_{\mathcal{A}}$ для $q_a \in Q_a$. *Языком* называется любое подмножество слов в конечном алфавите. Язык $L(\mathcal{A})$ слов, распознаваемый автоматом $\mathcal{A}$, состоит из слов, принимаемых $\mathcal{A}$. Автоматы $\mathcal{A}'$ и $\mathcal{A}''$ называются *эквивалентными*, если $L(\mathcal{A}') = L(\mathcal{A}'')$.

Переходы (q_1, ε, q_2) называются ε-переходами. По описанию автомата $\mathcal{A}$ возможно построить описание эквивалентного автомата $\mathcal{A}'$ без ε-переходов за полиномиальное время. Поэтому в дальнейшем по умолчанию предполагается, что в автоматах нет ε-переходов (при необходимости применяется описанное выше преобразование).

Ходом автомата $\mathcal{A}$ при чтении слова $w = w_1 w_2 \dots w_n$ называется такая последовательность состояний $q_0, q_1, \dots, q_n$, что $(q_{i-1}, w_i, q_i) \in \delta_{\mathcal{A}}$ для всех $1 \leqslant i \leqslant n$. Ход автомата из начального состояния в принимающее называется *принимающим*.

Всюду далее по умолчанию предполагается, что в автоматах нет *избыточных* состояний, т.е. каждое состояние лежит хотя бы на одном принимающем ходе автомата. Удаление избыточных состояний возможно за полиномиальное от размера описания автомата время.

Трансдьюсер – это конечный автомат, дополненный лентой результата, и для каждого перехода указано, какое слово в *выходном алфавите* трансдьюсер должен дописать на ленту результата к уже записанным там символам. Трансдьюсер T задает бинарное отношение на словах входного и выходного алфавитов: uTv истинно, если при чтении слова u возможен результат v на некотором принимающем ходе. Отношение *рационального доминирования* $A \leqslant_{\mathrm{rat}} B$ на языках означает, что существует такой трансдьюсер T, что

$$A = T(B) = \{v \mid \exists u:\ uTv,\ u \in B\}.$$

2.2. Сводимости. Язык A m-сводится к языку B относительно класса функций $\mathscr{C}$ (обозначение $\leqslant_{\mathrm{m}}^{\mathscr{C}}$), если существует такая функция $f \in \mathscr{C}$, что $x \in A$ равносильно $f(x) \in B$. Чтобы отношение сводимости было предпорядком (транзитивным и рефлексивным отношением), необходимо и достаточно, чтобы класс $\mathscr{C}$ содержал тождественное отображение и был замкнут относительно композиции. Используем обозначения $\leqslant_{\mathrm{m}}^{\mathrm{P}}$ для сводимости за полиномиальное время (сводимость по Карпу) и $\leqslant_{\mathrm{m}}^{\log}$ для сводимости на логарифмической памяти.

Аналогично определяется сводимость по Тьюрингу (сводимость с оракулом): $A \leqslant_{\mathrm{T}}^{\mathscr{C}} B$, если A разрешается алгоритмом из класса $\mathscr{C}$ с оракулом B. (Алгоритм разрешает A, если он вычисляет характеристическую функцию A). В этой статье используется сводимость $\leqslant_{\mathrm{T}}^{\mathrm{FNP}}$, где функции из класса FNP вычислимы за полиномиальное от длины входа время с NP-оракулом. Алгоритм вычисления такой функ-

ции может за один такт работы выяснять у оракула принадлежность слова-запроса к любому языку из NP. Нетрудно видеть, что равносильное требование состоит в том, чтобы алгоритм проверял принадлежность некоторому NP-полному языку.

Пусть R – бинарное отношение из класса P, и пусть для некоторого многочлена $q(\cdot)$ выполняется $|y| \leqslant q(|x|)$ для всех $(x, y) \in R$. Именно такие отношения используются в определении класса NP, и y называется *сертификатом* для x, если $(x, y) \in R$. В построении сводимостей мы используем хорошо известное соображение: алгоритм с NP-оракулом может за полиномиальное время не только проверить существование сертификата, но и найти его.

Частным случаем сводимости по Тьюрингу является табличная сводимость. В этом случае алгоритм разрешения строит список запросов $q_1, \ldots, q_s$ к оракулу и описание булевой функции от s аргументов, которая вычисляет результат по ответам оракула на запросы. Нам потребуется еще более ограниченный тип табличных сводимостей: *дизъюнктная сводимость* (dtt), когда вычисляется дизъюнкция ответов на запросы (т.е. ответ положительный, если хотя бы одно слово-запрос принадлежит языку-оракулу).

Мы будем использовать два вида дизъюнктных сводимостей. Более сильная из них, $\leqslant_{\mathrm{dtt}}^{\mathrm{PSPACE}}$, – дизъюнктная сводимость на полиномиальной памяти. Заметим, что в этом случае список запросов может быть экспоненциально велик по сравнению с длиной входа. Такая сводимость имеет смысл для достаточно трудных языков. Например, соответствующий ей предпорядок различает классы арифметической иерархии.

Более слабая, $\leqslant_{\mathrm{dtt}}^{\mathrm{P}}$, использует функции, вычислимые за полиномиальное время.

2.3. Описания конечных отношений. Нас интересуют фильтры, которые задают конечные отношения (например, графы). Обобщая кодировки из [8, 9], будем рассматривать фильтры следующего вида. Пусть $R \subseteq \mathbb{N}^k$ – конечное k-арное отношение. Кодируем последовательность $x = (x_1, x_2, \ldots, x_k) \in \mathbb{N}^k$ словом

$$\langle x \rangle = \lhd a^{x_1} \# a^{x_2} \# \ldots \# a^{x_k} \rhd$$

в алфавите $\{a, \lhd, \rhd, \#\}$. Будем называть такие слова *блоками*. *Описанием* $\langle R \rangle$ конечного отношения R является любое слово вида

$$\prod_{i=1}^{t} \langle x(i) \rangle, \quad \text{где} \quad \{x(i) : \, 1 \leqslant i \leqslant t\} = R.$$

Здесь $\prod$ обозначает конкатенацию слов. Заметим, что порядок записи кодировок различных $x \in R$ произвольный и допустимы повторения. Таким образом, кодировка неоднозначна.

Сопоставим семейству конечных отношений $\mathcal{R}$ язык

$$\langle \mathcal{R} \rangle = \{\langle R \rangle : \, R \in \mathcal{R}\},$$

состоящий из всех описаний всех отношений из $\mathcal{R}$. Эти языки будут фильтрами в задачах регулярной реализуемости, которые рассматриваются в данной статье.

Как и в [9], нас будут также интересовать описания простых неориентированных графов без изолированных вершин. Следуя [9], полагаем, что ребра графа перечисляются в произвольном порядке, порядок вершин в описании ребра не важен, диагональные блоки $\lhd a^x \# a^x \rhd$ допустимы, но игнорируются. Более точно, произвольному бинарному отношению R сопоставляем граф $G_R = (V, E)$ по правилу

$$\begin{aligned} V &= \{v : \, \exists u \, (u, v) \in R\} \cup \{u : \, \exists v \, (u, v) \in R\}, \\ E &= \big\{\{u, v\} : \, (u \neq v) \wedge \big(((u, v) \in R) \vee ((v, u) \in R)\big)\big\}. \end{aligned} \tag{1}$$

По определению описанием $\langle G\rangle$ графа G является описание $\langle R\rangle$ любого отношения R, для которого $G = G_R$.

Наши кодировки несколько отличаются от кодировок в других работах на эту тему, см. [9–11]. Различие состоит в формате разделителей между описаниями чисел в унарной системе. Поэтому одни кодировки переводятся в другие с помощью трансдьюсерного преобразования, и наоборот. Нас интересуют задачи регулярной реализуемости, их сложность не изменяется при таких преобразованиях, так как в [3] доказано, что из $A \leqslant_{\mathrm{rat}} B$ следует $\mathrm{NRR}(A) \leqslant_{\mathrm{m}}^{\log} \mathrm{NRR}(B)$. Поэтому эти различия в кодировках несущественны.

Обычно интересуются свойствами графов, сохраняющимися при изоморфизме. Будем далее называть *инвариантными* семейства отношений на $\mathbb{N}$, замкнутые относительно биекций $\mathbb{N}$. Из определения (1) сразу следует, что описания семейства графов с вершинами из $\mathbb{N}$, замкнутого относительно изоморфизма, образуют инвариантное семейство отношений.

§ 3. Автоматы в бесконечном алфавите

Рассмотрим автомат $\mathcal{A}$ в алфавите $\{a, \lhd, \rhd, \#\}$. Обозначим через $\mathcal{R}^k(\mathcal{A})$ класс тех k-арных отношений на $\mathbb{N}$, описания которых принимаются автоматом $\mathcal{A}$: хотя бы одно описание для каждого отношения. Слова в унарном алфавите находятся во взаимно-однозначном соответствии с натуральными числами, слову сопоставляется его длина. В дальнейшем анализе удобно избавиться от символов-разделителей и рассматривать автоматы в бесконечном алфавите $\mathbb{N}^k$.

Для этого по автомату $\mathcal{A}$ построим автомат $\widetilde{\mathcal{A}}^k_{\mathbb{N}}$ с тем же множеством состояний, теми же начальными и принимающими состояниями и алфавитом $\mathbb{N}^k$. Отношение переходов определим так: для всех $\boldsymbol{i} = (i_1, i_2, \ldots, i_k) \in \mathbb{N}^k$

$$(q_1, \boldsymbol{i}, q_2) \in \delta_{\widetilde{\mathcal{A}}^k_{\mathbb{N}}} \iff (q_1, \lhd a^{i_1} \# a^{i_2} \# \ldots \# a^{i_k} \rhd, q_2) \in \delta_{\mathcal{A}}\,. \tag{2}$$

Из такого определения следует, что если непустое слово в алфавите $\mathbb{N}^k$ принадлежит $L(\widetilde{\mathcal{A}}^k_{\mathbb{N}})$, то множество входящих в него символов (т.е. последовательностей неотрицательных целых чисел длины k) принадлежит $\mathcal{R}^k(\mathcal{A})$. Верно и обратное: любое отношение $R \in \mathcal{R}^k(\mathcal{A})$ является множеством символов некоторого слова из $L(\widetilde{\mathcal{A}}^k_{\mathbb{N}})$. В автомате $\widetilde{\mathcal{A}}^k_{\mathbb{N}}$ возможны избыточные состояния. После их удаления получаем автомат $\mathcal{A}^k_{\mathbb{N}}$, для которого выполняются те же соотношения с множеством $\mathcal{R}^k(\mathcal{A})$. Об отношениях из $\mathcal{R}^k(\mathcal{A})$ будем говорить, что они принимаются автоматом $\mathcal{A}^k_{\mathbb{N}}$. Аналогично, граф G принимается автоматом, если для некоторого отношения R выполняется $G = G_R$ и $R \in L(\mathcal{A}^2_{\mathbb{N}})$.

Представление автомата в бесконечном алфавите в виде конечного объекта требует уточнения в описании отношения переходов. В данном случае это довольно просто сделать в силу ограниченной структуры регулярных языков в унарном алфавите. Теорема Хробака–Мартинеза (см. [13,14] и уточнения в [15,16]) утверждает, что по автомату в унарном алфавите, имеющему n состояний, можно за полиномиальное время построить описание множества длин принимаемых ими слов в виде $O(n^2)$ арифметических прогрессий $\{a + bt : t \in \mathbb{N}\}$, где $a = O(n^2)$, $b = O(n)$. Если не оговорено противное, далее мы предполагаем, что одномерные полулинейные множества представлены в виде объединения арифметических прогрессий с указанными ограничениями на начальные члены и разности.

Сопоставим паре состояний (q_1, q_2) автомата $\mathcal{A}^k_{\mathbb{N}}$ множество $L_{q_1 q_2} \subseteq \mathbb{N}^k$ тех наборов из $\mathbb{N}^k$, чтение которых переводит q_1 в q_2.

Лемма 1. *Множество* $L_{q_1q_2} \subseteq \mathbb{N}^k$, *где* $k = O(1)$, *является объединением декартовых произведений одномерных полулинейных множеств. Размер описания этого объединения* $\mathrm{poly}(\mathrm{size}(\mathcal{A}))$, *где* $\mathrm{size}(\mathcal{A})$ – *размер описания* $\mathcal{A}$.

Отношение переходов автомата $\mathcal{A}^k_{\mathbb{N}}$, т.е. множество $\{(q_1, \boldsymbol{i}, q_2) : \boldsymbol{i} \in L_{q_1q_2}\}$, задаем соответствующими $L_{q_1q_2}$ объединениями декартовых произведений одномерных полулинейных множеств. Размер такого описания полиномиально ограничен размером описания исходного автомата $\mathcal{A}$, как следует из леммы 1. Построение описания автомата $\mathcal{A}^k_{\mathbb{N}}$ по $\mathcal{A}$ возможно за полиномиальное время в силу теоремы Хробака – Мартинеса.

В дальнейшем будут важны те автоматы, для которых $\mathcal{R}^k(\mathcal{A})$ конечно. Поэтому потребуется следующий факт.

Лемма 2. $\mathcal{R}^k(\mathcal{A})$ *конечно тогда и только тогда, когда* $L_{q_1q_2}$ *конечно для любой пары* (q_1, q_2). *Аналогичное утверждение выполняется также и для множества принимаемых автоматом* $\mathcal{A}$ *графов.*

Следствие 1. *Конечность* $\mathcal{R}^k(\mathcal{A})$ *проверяется за полиномиальное время.*

§ 4. Общая схема доказательства универсальности

Универсальность описаний отношений заданного типа состоит в том, что для любого непустого языка двоичных слов X существует такой класс $\mathcal{U}_X$ конечных k-арных отношений этого типа, что

$$X \leqslant_1 \mathrm{NRR}\big(\langle \mathcal{U}_X \rangle\big) \leqslant_2 X. \tag{3}$$

Точный вид сводимостей и требования к классу отношений будут в разных случаях различными. Однако общая схема доказательства одна и та же, и мы ее сейчас опишем.

Определяем функцию $X \mapsto \mathcal{U}_X$ следующим образом. Обозначим через $P_f(\mathbb{N}^k)$ семейство всех конечных подмножеств $\mathbb{N}^k$, т.е. конечных k-арных отношений на $\mathbb{N}$. Выберем кодировку $\varphi\colon \{0,1\}^* \to P_f(\mathbb{N}^k)$ двоичных слов конечными отношениями, удовлетворяющую следующим свойствам:

1. φ инъективна;
2. φ вычислима за полиномиальное время;
3. (частичная) функция φ^{-1} также вычислима за полиномиальное время;
4. $x \in X$ равносильно $\varphi(x) \in \mathcal{U}_X$;
5 (условие конечности). Для всякого автомата $\mathcal{A}$ из бесконечности $\mathcal{R}^k(\mathcal{A})$ следует, что $\mathcal{A}$ принимает хотя бы одно описание $\langle R \rangle$ для некоторого $R \notin \varphi(\{0,1\}^*)$.

Для выполнения условия 4 потребуем

$$\mathcal{U}_X = \varphi(X) \cup \overline{\varphi(\{0,1\}^*)} \tag{4}$$

Тогда

$$\mathcal{U}_X \cap \varphi\big(\{0,1\}^*\big) = \varphi(X),$$

и сводимость $X \leqslant \mathrm{NRR}\big(\langle \mathcal{U}_X \rangle\big)$ возможно определить как $x \mapsto \mathcal{A}_{\langle \varphi(x) \rangle}$, где автомат $\mathcal{A}_{\langle \varphi(x) \rangle}$ принимает в точности одно слово из множества описаний отношения $\varphi(x)$ (такие сводимости в [6] названы *моносводимостями*). Из условия 2 следует, что в таком случае

$$X \leqslant^{\mathrm{P}}_{\mathrm{m}} \mathrm{NRR}\big(\langle \mathcal{U}_X \rangle\big).$$

Сводимость $\mathrm{NRR}\big(\langle \mathcal{U}_X \rangle\big) \leqslant X$ определяется по-разному на двух классах входов. Мы разделяем автоматы, т.е. входы $\mathrm{NRR}\big(\langle \mathcal{U}_X \rangle\big)$, на *тривиальные* и *нетривиальные*.

Определения тривиальности варьируются. Однако всегда для тривиального автомата $\mathcal{A}$ выполняется

$$\mathcal{R}^k(\mathcal{A}) \cap \overline{\varphi(\{0,1\}^*)} \neq \varnothing,$$

т.е. $\mathcal{A}$ принимает хотя бы одно слово $\langle R \rangle$ для отношения R, которое не является кодировкой двоичного слова. Важный частный случай этого условия – бесконечность $\mathcal{R}^k(\mathcal{A})$, так как тогда из условия конечности 5 следует

$$\mathcal{R}^k(\mathcal{A}) \cap \overline{\varphi(\{0,1\}^*)} \neq \varnothing.$$

В силу (4) ответ в задаче $\mathrm{NRR}\big(\langle \mathcal{U}_X \rangle\big)$ для тривиального автомата $\mathcal{A}$ положительный, т.е. $\mathcal{A} \in \mathrm{NRR}\big(\langle \mathcal{U}_X \rangle\big)$. Это позволяет определять вторую сводимость

$$\mathrm{NRR}\big(\langle \mathcal{U}_X \rangle\big) \leqslant X$$

на тривиальных автоматах простейшим способом: отображать каждый тривиальный автомат в некоторый фиксированный элемент $x_0 \in X$. Такое определение сводимости возможно для тех сводимостей, сложность которых достаточна для проверки тривиальности автомата. Формально, дизъюнктная сводимость на тривиальном автомате выдает список запросов, состоящий из единственного элемента x_0.

Для нетривиальных автоматов возможны оба ответа в задаче $\mathrm{NRR}\big(\langle \mathcal{U}_X \rangle\big)$. Пусть $\mathcal{A}$ – нетривиальный автомат, в частности, $|\mathcal{R}^k(\mathcal{A})| < \infty$. Список запросов к оракулу X состоит из слов

$$\varphi^{-1}(\mathcal{R}^k(\mathcal{A})) = \{\varphi^{-1}(R) : \ R \in \mathcal{R}^k(\mathcal{A})\},$$

здесь φ – та же функция, что и при построении первой сводимости. Поэтому

$$\mathcal{A} \in \mathrm{NRR}\big(\langle \mathcal{U}_X \rangle\big) \iff \varphi^{-1}(\mathcal{R}^k(\mathcal{A})) \cap X \neq \varnothing,$$

поскольку хотя бы один запрос возвращает положительный ответ в точности при $\mathcal{A} \in \mathrm{NRR}\big(\langle \mathcal{U}_X \rangle\big)$. Количество запросов конечно, так как φ – инъекция по условию 1 и $|\mathcal{R}^k(\mathcal{A})| < \infty$.

Так как множество $\mathcal{R}^k(\mathcal{A})$ конечно, автомат $\mathcal{A}^k_{\mathbb{N}}$ принимает только отношения размера $\mathrm{poly}(\mathrm{size}(\mathcal{A}))$ в силу лемм 1 и 2. Все эти отношения можно перечислить на полиномиальной памяти. Это уже дает сводимость на полиномиальной памяти. В некоторых случаях она может быть ослаблена до дизъюнктной полиномиальной $\leqslant^{\mathrm{P}}_{\mathrm{dtt}}$.

В случае инвариантных отношений эту схему нужно модифицировать. В этом случае нужны классы изоморфизма конечных отношений, поэтому $\mathcal{U}_X$ должно быть замкнуто относительно биекций $\mathbb{N}$. Значит, нетривиальны те автоматы $\mathcal{A}$, для которых $\mathcal{R}^k(\mathcal{A})$ конечно и содержит лишь отношения, изоморфные отношениям из $\varphi(\{0,1\}^*)$. Для второй сводимости нужен список всех классов изоморфизма отношений из $\mathcal{R}^k(\mathcal{A})$.

Заметим также, что для универсальности инвариантных унарных отношений (теорема 3 ниже) условие конечности 5 требуется ослабить.

§ 5. Пути на размеченных графах

Для реализации изложенного выше плана важно находить все отношения из $\mathcal{R}^k(\mathcal{A})$, если это множество конечно. По лемме 2 равносильное условие состоит в том, что множества $L_{q_1 q_2}$ конечные для всех пар q_1, q_2 состояний автомата $\mathcal{A}^k_{\mathbb{N}}$ (их может быть меньше, чем состояний $\mathcal{A}$, см. § 3).

Поскольку порядок перечисления элементов в описании отношения несущественен, характеризация принимаемых автоматом $\mathcal{A}^k_{\mathbb{N}}$ отношений дается следующей общей моделью задания семейств конечных множеств ориентированными графами.

Рассматриваем ориентированные графы с петлями и кратными ребрами. *Граф переходов* $G = (V, s, t, E, \ell)$ задается множеством вершин V, множеством ориентированных ребер E, начальной вершиной s, конечной вершиной t и функцией разметки

$$\ell\colon E \to \{\varnothing\} \cup U,$$

которая отображает ребра графа переходов в пустое множество или в элемент конечного множества U, именуемого далее *универсумом меток*. Размер $|G|$ графа переходов – это размер описания G списками вершин, ребер и таблицей функции разметки.

Значение функции $\ell(e)$, $e \in E$, называем *меткой* ребра e. Для пути P *множеством меток пути* $\ell(P)$ называем множество всех непустых меток его ребер. Говорим также, что P *несет* множество меток $\ell(P)$. Через $\mathcal{F}(G)$ обозначаем семейство множеств меток путей из s в t (в дальнейшем (s,t)-пути).

Связь с автоматами $\mathcal{A}^k_{\mathbb{N}}$, у которых все множества $L_{q_1q_2}$ конечные, прямолинейна. Вершинами графа G являются состояния автомата и дополнительная терминальная вершина, куда ведут ребра из принимающих состояний, помеченные пустым множеством. Каждому $\boldsymbol{i} \in L_{q_1q_2}$ соответствует ребро (q_1, q_2) в графе переходов G (разным $\boldsymbol{i}$ соответствуют разные ребра). Меткой этого ребра является $\boldsymbol{i}$. Начальной вершиной является начальное состояние $\mathcal{A}^k_{\mathbb{N}}$. Так как $\mathcal{R}^k(\mathcal{A})$ совпадает с семейством множеств символов непустых слов языка $L(\mathcal{A}^k_{\mathbb{N}})$, то

$$\mathcal{R}^k(\mathcal{A}) = \mathcal{F}(G).$$

Проверка принадлежности множества S семейству $\mathcal{F}(G)$ алгоритмически трудна.

Задача All-labels. Вход: граф переходов $G = (V, s, t, E, \ell)$. Требуется выяснить, существует ли такой (s,t)-путь P в G, что $\ell(P) = \ell(E)$.

Лемма 3. *Задача* All-labels *является* NP*-полной.*

Следствие 2. *Проверка принадлежности множества S семейству $\mathcal{F}(G)$ является* NP*-полной.*

Хотя проверка $S \in \mathcal{F}(G)$ является NP-полной, множества меток путей возможно эффективно порождать в смысле инкрементального порождения с полиномиальной задержкой, когда каждый следующий элемент множества порождается за время, полиномиальное от длины входа и количества уже найденных элементов множества (см., например, [17]). Более точно, для инкрементального порождения $\mathcal{F}(G)$ требуется решение следующей задачи.

Задача Next-label. Вход: граф переходов $G = (V, s, t, E, \ell)$ и семейство множеств меток

$$\mathcal{S} = \{S_0, \ldots, S_{t-1}\}, \quad \mathcal{S} \subseteq \mathcal{F}(G).$$

Требуется проверить равенство $\mathcal{F}(G) = \mathcal{S}$, и в случае отрицательного ответа предъявить множество $S_t \in \mathcal{F}(G) \setminus \mathcal{S}$.

Для ациклических графов переходов задача Next-label разрешима за полиномиальное от длины входа время. Поэтому эффективное порождение $\mathcal{F}(G)$ возможно для ациклического графа переходов, если $|\mathcal{F}(G)|$ полиномиально ограничено $|G|$.

Ограничение ациклическими графами несущественно.

Лемма 4. *По графу переходов G за полиномиальное время можно построить такой ациклический граф переходов G', что $\mathcal{F}(G') = \mathcal{F}(G)$.*

Лемма 5. *Задача* Next-label *разрешима за полиномиальное время.*

Общий план построения сводимостей из § 4 требует порождения всех множеств из $\mathcal{F}(G)$. Как следует из леммы 5, это возможно за полиномиальное время, если $|\mathcal{F}(G)|$ полиномиально ограничено относительно $|G|$. Для выполнения этого условия нам потребуются ε-разделенные семейства множеств. Через $S_1 \oplus S_2$ обозначаем симметрическую разность множеств. Семейство конечных подмножеств $\mathcal{F}$ назовем *ε-разделенным*, если для каждой пары различных множеств $S_1 \neq S_2$ из $\mathcal{F}$ выполняется

$$|S_1 \oplus S_2| > \varepsilon \cdot \max\big(|S_1|, |S_2|\big)$$

при

$$\varepsilon \cdot \max\big(|S_1|, |S_2|\big) \geqslant 6.$$

Последнее ограничение носит технический характер, добавлено для упрощения доказательств. Из него сразу следует, что $\varepsilon > 0$. Заметим также, что $\varepsilon < 2$ для любого ε-разделенного семейства. Эта оценка достигается: семейство непересекающихся множеств, мощности которых одинаковы и не меньше 3, является $(2-\delta)$-разделенным при любом δ, таком что $2 > \delta > 0$.

Лемма 6. *Пусть G – такой граф переходов с n вершинами и универсумом размера m, что $\mathcal{F}(G)$ является ε-разделенным для некоторого ε. Тогда*

$$|\mathcal{F}(G)| \leqslant m \cdot n^{3/\varepsilon} + m^{6/\varepsilon}.$$

Сформулируем алгоритмический вариант леммы 6.

Лемма 7. *Пусть $\mathscr{C}$ – такое семейство подмножеств, что $\mathscr{C} \in \mathrm{P}$, и пусть для любого графа переходов G из $\mathcal{F}(G) \subseteq \mathscr{C}$ следует, что $\mathcal{F}(G)$ является ε-разделенным для некоторого $\varepsilon = \varepsilon(G)$, где функция $\varepsilon(G)$ полиномиально вычислима. Тогда существует алгоритм, который по графу переходов G проверяет, что $\mathcal{F}(G) \subseteq \mathscr{C}$, и в случае положительного ответа порождает список множеств из $\mathcal{F}(G)$ за время*

$$\mathrm{poly}(|G|^{O(1/\varepsilon(G))}).$$

Для приложения к доказательству универсальности описаний унарных отношений потребуются *обильные* разделенные семейства множеств, т.е. такие семейства, в которые входит экспоненциально много множеств относительно размера универсума m. Кроме того, для применения леммы 7 необходимо, чтобы такие семейства были эффективно разрешимыми. Для этих целей понадобятся эффективные асимптотически хорошие двоичные коды.

Напомним, что линейным двоичным (n, k, d)-кодом C называется такое подпространство размерности k координатного векторного пространства $\mathbb{F}_2^n$, что среди координат каждого ненулевого вектора этого подпространства не менее d единиц. Асимптотически хорошим называется такой код, для которого

$$k = \Omega(n) \quad \text{и} \quad d = \Omega(n).$$

Эффективность в данном случае означает лишь существование полиномиального алгоритма кодирования: такого семейства линейных отображений

$$c_k \colon \mathbb{F}_2^k \to \mathbb{F}_2^n,$$

что $c_k(\mathbb{F}_2^k) = C$ и функция $f(k, x) = c_k(x)$ вычислима за время $\mathrm{poly}(k, n)$. Более подробные сведения о корректирующих кодах можно найти в [18, 19]. Там же описаны конструкции эффективных асимптотически хороших кодов.

Мы используем эффективные асимптотически хорошие коды для построения кодировки φ, описанной в § 4.

§ 6. Результаты об универсальности

Теорема 1. *Для любого языка $\varnothing \subset X \subseteq \{0,1\}^*$ существует такое семейство $\mathcal{U}_X$ конечных подмножеств $\mathbb{N}$, что*

$$X \leqslant_{\mathrm{m}}^{\mathrm{P}} \mathrm{NRR}\big(\langle \mathcal{U}_X \rangle\big) \leqslant_{\mathrm{dtt}}^{\mathrm{P}} X.$$

Теорема 2. *Для любого языка $\varnothing \subset X \subseteq \{0,1\}^*$ существует такое семейство графов $\mathcal{G}_X$, замкнутое относительно изоморфизма, что*

$$X \leqslant_{\mathrm{m}}^{\mathrm{P}} \mathrm{NRR}\big(\langle \mathcal{G}_X \rangle\big) \leqslant_{\mathrm{dtt}}^{\mathrm{PSPACE}} X.$$

Единственный инвариант конечных подмножеств $\mathbb{N}$ относительно биекций $\mathbb{N}$ – это размер (мощность) множества. Поэтому инвариантное семейство $\mathcal{I}_L \subset 2^{\mathbb{N}}$ конечных унарных отношений задается подмножеством $L \subseteq \mathbb{N}$ и состоит из таких конечных подмножеств натуральных чисел, мощности которых принадлежат L. Это ограничивает возможности кодировки двоичных слов, так как для слов длины n потребуются экспоненциально большие числа, а числа здесь представляются в унарной кодировке. Поэтому в данном случае мы ограничимся более слабым результатом и докажем универсальность инвариантных унарных отношений для языков в унарном алфавите.

Теорема 3. *Для любого языка $\varnothing \subset X \subseteq \{a\}^*$ существует такое множество $L \subset \mathbb{N}$, что*

$$X \leqslant_{\mathrm{m}}^{\mathrm{P}} \mathrm{NRR}\big(\langle \mathcal{I}_L \rangle\big) \leqslant_{\mathrm{T}}^{\mathrm{FNP}} X.$$

§ 7. Заключение

Эта статья расширяет список классов фильтров, для которых выполняется свойство универсальности. Сравнивая известные примеры – произвольные и префиксно-замкнутые фильтры из [6], фильтры, которые являются языками корректных протоколов работы конечных автоматов с дополнительными структурами данных, из [8] – с новыми классами фильтров, можно отметить следующие особенности. Во-первых, усиление ограничений на фильтры приводит к усилению требуемых сводимостей, т.е. дает более слабые свойства универсальности. Это ожидаемое явление. Во-вторых, для доказательства универсальности новых классов фильтров потребовались более сложные технические инструменты. В частности, существенную роль в доказательствах этой статьи играют асимптотически хорошие эффективные коды. Это необычно для теории формальных языков. Кроме того, для доказательств универсальности пришлось изучить представление семейств множеств путями в ориентированных графах. Это новая неоднородная модель вычислений, свойства которой существенно отличаются от большинства стандартных неоднородных моделей вычисления. Мы надеемся, что эта модель найдет и другие приложения в теоретической информатике. В-третьих, во всех известных примерах классы фильтров, несмотря на структурные ограничения, параметризованы произвольными языками или их аналогами (например, семействами конечных отношений в данной статье).

Возможно ли наложить такие сильные структурные ограничения при сохранении параметризации произвольными языками, чтобы полученный класс фильтров не был универсальным даже в самом слабом разумном смысле общих сводимостей по Тьюрингу? Ответ на этот вопрос неизвестен, хотя кажется правдоподобным, что такой пример возможно построить, используя методы теории алгоритмов. Однако при таком построении пример окажется весьма искусственным.

Второй интересный вопрос, возникающий при сравнении полученных результатов, связан со сложностью сводимостей, требуемых для универсальности. Возможно ли доказать универсальность инвариантных свойств графов, ограничиваясь полиномиальными дизъюнктными сводимостями (или хотя бы полиномиальными сводимостями по Тьюрингу)? Ответ на такой вопрос, даже по модулю стандартных теоретико-сложностных гипотез, был бы чрезвычайно интересным. Нам представляется, что для этого необходимы новые технические средства.

СПИСОК ЛИТЕРАТУРЫ

1. *Yannakakis M.* Graph-Theoretic Methods in Database Theory // Proc. 9th ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems (PODS'90). Nashville, TN, USA. Apr. 2–4, 1990. New York: ACM, 1990. P. 230–242. https://doi.org/10.1145/298514.298576
2. *Bouajjani A., Esparza J., Maler O.* Reachability Analysis of Pushdown Automata: Application to Model-Checking // Proc. Int. Conf. on Concurrency Theory (CONCUR'97). Warsaw, Poland. July 1–4, 1997. Lect. Notes Comput. Sci. V. 1243. Berlin, Heidelberg: Springer, 1997. P. 135–150. https://doi.org/10.1007/3-540-63141-0_10
3. *Rubtsov A.A., Vyalyi M.N.* Regular Realizability Problems and Context-Free Languages // Proc. 17th Int. Workshop on Descriptional Complexity of Formal Systems (DCFS'2015). Waterloo, ON, Canada. June 25–27, 2015. Lect. Notes Comput. Sci. V. 9118. Berlin, Heidelberg: Springer, 2015. P. 256–267. https://doi.org/10.1007/978-3-319-19225-3_22
4. *Chistikov D., Majumdar R., Schepper P.* Subcubic Certificates for CFL Reachability // Proc. ACM Program. Lang. 2022. V. 6. Article 41 (29 pp.). https://doi.org/10.1145/3498702
5. *Вялый М.Н.* О задачах регулярной реализуемости // Пробл. передачи информ. 2011. Т. 47. № 4. С. 43–54. https://www.mathnet.ru/rus/ppi2059
6. *Вялый М.Н.* О выразительной силе задач регулярной реализуемости // Пробл. передачи информ. 2013. Т. 49. № 3. С. 86–104. https://www.mathnet.ru/rus/ppi2117
7. *Vyalyi M.N.* On Models of a Nondeterministic Computation // Computer Science – Theory and Applications: Proc. 4th Int. Computer Science Symp. in Russia (CSR'09). Novosibirsk, Russia. Aug. 18–23, 2009. Lect. Notes Comput. Sci. V. 5675. Berlin, Heidelberg: Springer, 2009. P. 334–345. https://doi.org/10.1007/978-3-642-03351-3_31
8. *Rubtsov A., Vyalyi M.* Automata Equipped with Auxiliary Data Structures and Regular Realizability Problems, https://arxiv.org/abs/2210.03934 [cs.FL], 2022.
9. *Wolf P., Fernau H.* Regular Intersection Emptiness of Graph Problems: Finding a Needle in a Haystack of Graphs with the Help of Automata, https://arxiv.org/abs/2003.05826 [cs.FL], 2020.
10. *Wolf P.* From Decidability to Undecidability by Considering Regular Sets of Instances // Theor. Comput. Sci. 2022. V. 899. P. 25–38. https://doi.org/10.1016/j.tcs.2021.11.006
11. *Diekert V., Fernau H., Wolf P.* Properties of Graphs Specified by a Regular Language // Acta Inform. 2022. V. 59. № 4. P. 357–385. https://doi.org/10.1007/s00236-022-00427-z
12. *Rubtsov A., Vyalyi M.* On Universality of Regular Realizability Problems // Probl. Inf. Transm. 2024. V. 60. № 3. P. 209–232. https://doi.org/10.1134/S0032946024030050
13. *Chrobak M.* Finite Automata and Unary Languages // Theor. Comput. Sci. 1986. V. 47. P. 149–158. https://doi.org/10.1016/0304-3975(86)90142-8
14. *Martinez A.* Efficient Computation of Regular Expressions from Unary NFAs // Pre-proc. 4th Int. Workshop on Descriptional Complexity of Formal Systems (DCFS 2002). London, Canada. Aug. 21–24, 2002. Dept. Comput. Sci., Univ. of Western Ontario, Canada, 2002. P. 174–187.
15. *Chrobak M.* Errata to: "Finite Automata and Unary Languages": [Theoret. Comput. Sci. 47 (1986) 149–158] // Theor. Comput. Sci. 2003. V. 302. № 1–3. P. 497–498. https://doi.org/10.1016/S0304-3975(03)00136-1

16. *To A.W.* Unary Finite Automata vs. Arithmetic Progressions // Inform. Process. Lett. 2009. V. 109. № 17. P. 1010–1014. https://doi.org/10.1016/j.ipl.2009.06.005
17. *Gurvich V.* Complexity of Generation // Computer Science – Theory and Applications: Proc. 13th Int. Computer Science Symp. in Russia (CSR 2018). Moscow, Russia. June 6–10, 2018. Lect. Notes Comput. Sci. V. 10846. Berlin, Heidelberg: Springer, 2018. P. 1–14. https://doi.org/10.1007/978-3-319-90530-3_1
18. *Мак-Вильямс Ф.Дж., Слоэн Н.Дж.А.* Теория кодов, исправляющих ошибки. М.: Связь, 1979.
19. *Ромащенко А.Е., Румянцев А.Ю., Шень А.* Заметки по теории кодирования. М.: МЦНМО, 2017.

Вялый Михаил Николаевич
Рубцов Александр Александрович
Национальный исследовательский университет
“Высшая школа экономики”, Москва
vyalyi@gmail.com
rubtsov99@gmail.com

Поступила в редакцию
05.09.2024
После доработки
17.10.2024
Принята к публикации
22.10.2024

ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 60 2024 Вып. 3

УДК 621.391 : 004.725.5

© 2024 г. **Д.В. Банков, А.И. Ляхов, Е.А. Степанова, Е.М. Хоров**

АНАЛИЗ ВЛИЯНИЯ МЕХАНИЗМА RESTRICTED TARGET WAKE TIME НА ПРОПУСКНУЮ СПОСОБНОСТЬ СЕТИ Wi-Fi[1]

С каждым годом в сетях Wi-Fi увеличивается доля трафика приложений реального времени (англ.: real time application, RTA), предъявляющих строгие требования к задержкам и надежности доставки данных. Для удовлетворения данных требований часто используется резервирование канальных ресурсов с последующим их выделением для передач RTA-кадров. Однако в случае нерегулярного RTA-трафика часть зарезервированных канальных ресурсов может оказаться неиспользованной при его отсутствии, что приводит к снижению общей пропускной способности сети. Для решения этой проблемы был разработан механизм пробуждения по расписанию с ограниченным резервированием (англ.: restricted target wake time, R-TWT), запрещающий всем станциям передачу в заданный момент времени R-TWT. Механизм R-TWT одновременно позволяет RTA-трафику в момент R-TWT быстрее получить доступ к каналу и не запрещает его другим станциям после момента R-TWT. В статье представлена аналитическая модель сети Wi-Fi, при помощи которой были получены зависимости пропускной способности пользователей от параметров передачи и периода резервирования R-TWT.

Ключевые слова: резервирование, R-TWT, трафик VR, QoS, Wi-Fi.

DOI: 10.31857/S0555292324030070, **EDN:** AETZUT

§ 1. Введение

В последние годы растет число приложений реального времени (англ.: real time application, RTA), требующих для своей работы передачу данных с высокой надежностью и низкой задержкой. Спектр применения таких приложений достаточно широк и включает в себя как сети Индустриального Интернета вещей, так и устройства дополненной или виртуальной реальности (ВР). Для таких устройств целесообразно использовать беспроводные технологии передачи данных, среди которых одной из самых распространенных является технология Wi-Fi.

Для передачи данных станции Wi-Fi используют механизм случайного доступа к каналу, что может приводить к коллизиям, т.е. к потерям передаваемых данных при одновременной передаче несколькими станциями. В случае коллизии станция выполняет повторную передачу данных через случайное время, что позволяет уменьшить вероятность повторной коллизии, однако увеличивает задержку доставки данных.

Изначально в сетях Wi-Fi использовалась одна очередь кадров данных, которые обслуживались в порядке поступления. Однако в случае высокой загрузки станции

[1] Исследование выполнено в ИППИ РАН за счет гранта Российского научного фонда № 24-19-00816, https://rscf.ru/project/24-19-00816/.

низкоприоритетным фоновым трафиком при таком подходе невозможно удовлетворить строгие требования RTA-трафика к задержкам, поэтому возникла потребность в механизме приоритезации доставки кадров. Первым шагом для обеспечения приоритетной доставки трафика в сетях Wi-Fi стала разработка улучшенного метода распределенного доступа к каналу (англ.: enhanced distributed channel access, EDCA), позволяющего распределить поступающий трафик в зависимости от приоритета по четырем очередям, обладающим различными параметрами доступа к каналу (подробнее механизм EDCA описан в п. 2.1). Такой подход позволил при освобождении канала высокоприоритетному трафику получать доступ к каналу с меньшей задержкой и большей вероятностью, чем другие виды трафика. Однако использование EDCA не позволяет гарантировать выполнение ограничений на задержку доставки кадра данных, так как в момент его поступления канал может оказаться занят чужой передачей. Отметим, что длительность передачи станций можно ограничить при помощи параметров EDCA [1] и сделать достаточно короткой, чтобы с большей вероятностью выполнялись требования по задержке RTA-трафика, однако это может заметно уменьшить пропускную способность сети из-за уменьшения объема передаваемых полезных данных при фиксированном объеме накладных расходов.

Чтобы решить описанные проблемы и обеспечить выполнение требований к качеству обслуживания RTA-трафика, рабочей группой IEEE 802.11 был предложен принципиально другой подход к приоритезации – резервирование канальных ресурсов для доставки приоритетного трафика. На этом подходе основано множество механизмов резервирования канальных ресурсов, в частности, механизмы HCCA, MCCA и RAW. В общем виде резервирование при помощи данных механизмов часто представляет собой периодическую последовательность временных промежутков фиксированной длительности, в пределах которых применяются специфические правила доступа к каналу, а доступ предоставляется одной станции или определенной группе. Все эти механизмы успешно справляются с обеспечением высокой надежности и низких задержек доставки кадров RTA, однако в случае неидеального канала резервирование должно выполняться с учетом возможных повторных попыток передачи. Отметим, что случайный шум в канале может приводить к потере данных даже при использовании новейших СКК [2, 3]. Такое резервирование может приводить к простою канала в случае, если повторные попытки передачи не выполнялись, или в случае непериодического трафика RTA, который может представлять собой, к примеру, неординарный пуассоновский поток кадров [4, 5]. В случае если в сети Wi-Fi кроме RTA-устройств присутствуют пользовательские станции, излишнее резервирование может негативно сказываться на общей пропускной способности сети, так как значительная часть зарезервированного времени канала не будет использоваться.

В связи с этим в настоящий момент разрабатывается новый механизм доступа к каналу – механизм пробуждения по расписанию с ограниченным резервированием (англ.: restricted target wake time, R-TWT) [6]. Отличительной особенностью механизма R-TWT является то, что он резервирует не временной интервал, а только момент времени, к которому все станции в сети должны завершить свои передачи. Такой подход позволяет RTA-станции при наличии данных в зарезервированный момент начать получение доступа к каналу без ожидания освобождения канала. В то же время в случае отсутствия данных канал остается свободен для передач других станций в сети. Механизм R-TWT построен таким образом, чтобы минимизировать влияние его использования на пользовательский трафик, под которым далее будем понимать низкоприоритетный трафик станций, требовательных к пропускной способности сети. Однако запрет передач в определенный момент времени может снизить пропускную способность пользовательских станций, особенно если моменты R-TWT будут назначаться достаточно часто.

Данная статья посвящена анализу механизма R-TWT и исследованию влияния его использования на пропускную способность пользовательских станций. Для этого разработана аналитическая модель передачи данных в сети Wi-Fi, использующей механизм R-TWT, позволяющая оценить пропускную способность пользователей в зависимости от периода назначения моментов R-TWT и параметров передачи пользователей, таких как ограничение длительности передачи T_{lim} и используемая сигнально-кодовая конструкция (СКК). С помощью модели показано, что зависимость пропускной способности пользователей от периода R-TWT является немонотонной.

Дальнейшее изложение построено следующим образом. В § 2 представлено описание объекта исследования, § 3 содержит обзор литературы. В § 4 описывается разработанная аналитическая модель, а в § 5 представлены результаты моделирования. Наконец, в § 6 формулируются основные результаты статьи.

Отметим, что в расширенной версии данной статьи [7] представлено описание всех основных механизмов резервирования канальных ресурсов в сетях Wi-Fi, подробно описаны все виды механизма пробуждения по расписанию, а также сформулированы открытые вопросы в исследовании механизма R-TWT.

§ 2. Объект исследования

Отметим, что механизм R-TWT является одним из видов механизма пробуждения по расписанию (англ.: target wake time, TWT), и в частности механизма широковещательного TWT (англ.: broadcast TWT, BTWT), поэтому наследует многие их особенности. К примеру, передача данных при помощи R-TWT основана на использовании механизма EDCA. Поэтому чтобы полноценно описать механизм R-TWT и все его особенности, повествование в данном параграфе будет построено следующим образом. Начнем с краткого описания в п. 2.1 стандартного механизма случайного доступа к каналу EDCA, используемого в сетях Wi-Fi. Далее в п. 2.2 представим описание механизмов TWT и BTWT. В завершение в п. 2.3 будет представлено подробное описание механизма R-TWT.

2.1. Механизм EDCA. Механизм EDCA основан на методе множественного доступа с прослушиванием несущей и избеганием коллизий (англ.: carrier sense multiple access with collision avoidance, CSMA/CA), согласно которому передачи кадров выполняются с использованием случайной отсрочки для уменьшения вероятности коллизии передач станций.

Для передачи кадра станция равновероятно выбирает значение отсрочки из промежутка от 0 до $W_r - 1$, где W_r – конкурентное окно, а r – число выполненных попыток передачи. Далее значение отсрочки уменьшается на единицу каждый раз, когда канал оказывается свободным в течение определенного промежутка времени, зависящего от предыдущего состояния канала. После успешной передачи в канале отсрочка уменьшается спустя T_{AIFS}, после задетектированной неудачной передачи – спустя T_{EIFS}, а если канал был свободен, то спустя слот отсрочки T_e. Передача кадра происходит, когда значение отсрочки становится равным 0. Для подтверждения успешного получения данных в ответ отправляется специальный кадр подтверждения (англ.: acknowlegement, Ack) через T_{SIFS} после завершения приема кадра данных. Если станция, отправившая данные, не получает в ответ кадр Ack в течение $T_{\text{AckTimeout}}$ после завершения передачи, то станция считает попытку передачи неуспешной и выполняет повторную попытку передачи, выбирая новое значение отсрочки из интервала $[0,\ W_r - 1]$, где

$$W_r = \begin{cases} W_{\min}, & r = 0, \\ \min\left\{2 \cdot W_{r-1}, W_{\max}\right\}, & 0 < r < R, \end{cases} \tag{1}$$

где R – наибольшее число попыток передачи одного кадра, $W_{\min}$ и $W_{\max}$ – минимальное и максимальное значения конкурентного окна.

После получения доступа к каналу станция может передать несколько кадров в течение $T_{\lim}$ без промежуточного выполнения случайной отсрочки при условии, что передачи кадров разделены между собой интервалом SIFS. Такие передачи называют TXOP (англ.: transmission opportunity – возможность передачи). Для уменьшения накладных расходов на отправку кадров Ack может использоваться механизм блочного подтверждения успешности доставки кадров. При использовании данного механизма после череды переданных кадров данных отправляется кадр запроса блочного подтверждения (англ.: block Ack request, BAR). В ответ на кадр BAR отправляется блочное подтверждение (англ.: block Ack, BAck), в котором передается номер первого подтверждаемого кадра и последовательность битов фиксированной длины, которые обозначают, успешно были приняты соответствующие кадры или нет. Отметим, что для упорядоченной доставки кадров используется механизм скользящего окна.

Также для снижения накладных расходов на передачу кадров может использоваться механизм агрегации. В данной статье используется механизм A-MPDU, в котором каждому кадру, пришедшему с вышестоящего уровня, добавляется контрольная сумма, информация о его длине и разделительная последовательность. Далее эти кадры объединяются в один кадр физического уровня с общим заголовком и кодируются, а разделительная последовательность используется получателем при декодировании для определения начала вложенных кадров. Таким образом, метод A-MPDU в случае короткого шума в канале позволяет успешно декодировать незатронутые шумом блоки. Для подтверждения успешности доставки кадров A-MPDU используется механизм блочного подтверждения, при этом кадр BAR часто отправляется вместе с агрегированным кадром данных. За счет такого объединения кадров сокращается количество заголовков физического уровня и передается только один кадр подтверждения BAck.

Отметим, что если агрегированный кадр попадет в коллизию, то из-за его размера коллизия окажется длительной. Это может негативно сказываться на общей пропускной способности сети, поэтому при использовании агрегации передача данных обычно предваряется обменом короткими кадрами RTS/CTS, которые в случае успеха резервируют канал для дальнейшего обмена кадрами данных, а в случае неудачи минимизируют длительность коллизии.

2.2. Механизмы Target Wake Time. Впервые механизм TWT был представлен в стандарте IEEE 802.11ah в качестве механизма энергосбережения. Механизм TWT позволяет запланировать моменты пробуждения и последующие интервалы обслуживания TWT (англ.: TWT service period, TWT SP), в течение которых возможен обмен кадрами между двумя станциями. Для передачи данных в течение TWT SP используется механизм EDCA. Процесс энергосбережения при использовании механизма TWT главным образом состоит в том, что вне TWT SP станция может отключать радиоинтерфейс и экономить энергию. Заметим, что станции в сети ничего не знают об используемых TWT других станций, не существует запрета на пересечение TWT SP различных станций или запрета на одновременное назначение TWT нескольким станциям. Поэтому использование механизма TWT априори не гарантирует защиту передач от коллизий в течение TWT SP.

Несмотря на данные недостатки, идея TWT оказалась настолько перспективной, что получила развитие в стандарте IEEE 802.11ax. Данный стандарт разрабатывался для обеспечения связи в плотных городских сценариях, поэтому его развитие пошло по пути организации детерминированного доступа к каналу. Это привело к появлению в IEEE 802.11ax механизма ортогонального множественного доступа с частотным разделением (англ.: orthogonal frequency-division multiple access,

OFDMA), позволяющего точке доступа разделить частотный диапазон на ресурсные блоки и выделять их станциям для одновременной передачи данных в разных частотах. Для более полной адаптации механизма TWT к идеям многопользовательских передач и использования механизма OFDMA в стандарте IEEE 802.11ax появился новый вид TWT – механизм широковещательного TWT (англ.: broadcast TWT, BTWT), который в отличие от описанного выше индивидуального TWT из IEEE 802.11ah, устанавливается между точкой доступа и группой станций. Параметры BTWT широковещательно рассылаются по сети в биконах.

Следует отметить, что хотя передача кадров в BTWT SP может выполняться при помощи механизма случайного доступа к каналу, в качестве основного способа обмена кадрами в BTWT SP стандарт предполагает использовать механизм OFDMA. При его использовании в начале BTWT SP точка доступа должна отправлять триггер-кадр, резервирующий канал для передачи конкретных станций и одновременно выделяющий им канальные ресурсы. Несмотря на это, BTWT SP все еще не защищен от передач других станций в сети: триггер-кадр может попасть в коллизию с передачей пользователя, а также начало BTWT SP может быть занято чужой передачей, что увеличит задержку доставки данных.

2.3. Механизм R-TWT. Теперь мы перейдем к краткому описанию механизма R-TWT, представленного в рамках стандарта IEEE 802.11be. Механизм R-TWT основан на механизме BTWT и перенял все его параметры и особенности, но при этом R-TWT является механизмом с ограниченной защитой доступа к каналу и предназначен для доставки трафика, чувствительного к задержке. Для достижения этой цели механизм R-TWT определяет момент времени (далее момент R-TWT), к которому все станции должны завершить обмен кадрами и освободить канал. R-TWT настраивается не просто для передачи данных между станциями, а для передачи конкретного типа или типов трафика, определяемых по специальным идентификаторам (англ.: traffic identifier, TID) и фиксируемых при настройке R-TWT. Это означает, что станция, для которой настроен R-TWT, должна передавать в R-TWT SP трафик с заданным TID, что дополнительно подчеркивает отличие R-TWT от предшественников, нацеленных на энергосбережение.

Современные станции, поддерживающие R-TWT, должны завершать свой TXOP до начала каждого R-TWT SP, установленного их точкой доступа, даже если они являются участниками этого R-TWT. Если станция, закончив отсчет отсрочки, понимает, что не успеет выполнить обмен кадрами до начала R-TWT SP, то станция должна отложить передачу, выбрав новое значение отсрочки из того же окна W_r, т.е. значение счетчика попыток передачи r не изменяется. Следует отметить, что точка доступа также должна завершить свой TXOP до начала R-TWT SP, за исключением случая, когда начало R-TWT пересекается передачей трафика с TID, соответствующим трафику, для которого назначался данный R-TWT.

Отметим, что в одной сети Wi-Fi могут одновременно работать устройства разных поколений, поэтому важно учитывать, что устройства, существующие на рынке в настоящий момент, не смогут поддерживать функционал механизма R-TWT, который еще только разрабатывается. Поэтому для обеспечения работы механизма R-TWT в сетях с устаревшими станциями стандарт предлагает использовать механизм Интервалов тишины (англ.: quieting interval, QI), который является обязательным для реализации в устройствах, начиная со стандарта IEEE 802.11n. Механизм QI изначально создавался для того, чтобы обеспечить возможность точке доступа определять состояние канала до станций, поэтому QI объявляются в биконах и запрещают станциям на заданном промежутке времени передачу кадров. Для обеспечения работы R-TWT QI предлагается использовать следующим образом: точка доступа планирует QI, чтобы начало QI совпадало с моментом R-TWT, при этом длительность QI выбирается минимально возможной и равной 1 TU (англ.: time unit, 1024 мкс).

Одновременно с этим современным станциям, обладающим функционалом R-TWT, разрешено игнорировать такие QI.

Анализируя описанные выше правила, можно сделать вывод, что механизм R-TWT в случае отсутствия RTA-трафика не запрещает пользователям получать доступ к каналу и выполнять передачи после наступления момента R-TWT, что с точки зрения использования канальных ресурсов выгодно отличает его от стандартных механизмов резервирования в сетях Wi-Fi.

§ 3. Обзор литературы

Как говорилось ранее, механизм R-TWT является разновидностью механизма TWT, который разрабатывался для обеспечения низкого уровня энергопотребления устройств. Однако существует исследование механизма TWT [8], предлагающее способ его использования для обслуживания RTA-трафика. Для этого авторы предлагают при помощи методов машинного обучения определять категории передаваемого трафика, используя которые далее адаптивно настраивать параметры механизма TWT. К сожалению, механизм TWT сможет защитить RTA-трафик от коллизий только в том случае, если все станции сети используют механизм TWT и передают кадры только в течение TWT SP. В то же время существует много факторов, таких как эффект дрейфа часов [9], из-за которых станции не всегда могут строго следовать установленному расписанию TWT.

Переходя к механизму R-TWT, отметим, что он еще находится в стадии разработки, поэтому существует множество работ, в которых R-TWT упоминается или описывается общими словами [10–16], однако работ по его изучению достаточно мало. К примеру, авторы [17] реализовали механизм R-TWT в платформе OpenWiFi и подтвердили, что его использование позволяет удовлетворять требования по задержке сразу для нескольких потоков трафика различного приоритета, однако авторы не исследовали пропускную способность пользовательских станций, не использующих R-TWT.

Многие авторы, рассматривающие механизм R-TWT, отмечают, что один из главных его недостатков проявляется в сценарии нескольких сетей Wi-Fi с пересекающимися областями покрытия. Авторы [18] разработали модель сети Wi-Fi в среде имитационного моделирования ns-3 для такого сценария и показали, что в нем эффективность передачи данных при помощи механизма R-TWT аналогична эффективности стандартного метода EDCA. Это связано с тем, что станции одной сети не знают о назначенных моментах R-TWT в другой сети и передают кадры независимо от них, поэтому в момент R-TWT канал может оказаться занят передачей станции соседней сети. Такая некоординированная работа сетей приводит к нарушению резервирования моментов R-TWT и потере эффективности данного механизма. Отметим, что проблема влияния передач в одной сети на передачи в запланированном TWT SP станций другой сети не является новой, и исследователи сходятся в том, что ее решение лежит в области координирования расписания TWT/R-TWT соседних точек доступа [18,19]. Причем такое координирование планируется сделать в рамках стандарта IEEE 802.11bn, разработка которого уже началась.

Кроме имитационных исследований R-TWT, существуют и аналитические оценки эффективности использования данного механизма. Авторы [20] предложили для удовлетворения требований к качеству обслуживания высокоприоритетного трафика использовать R-TWT совместно с планированием доступа в рамках механизма, определенного в дополнении к стандарту IEEE 802.1Qbv, позволяющего в сетях Ethernet передавать кадры из очереди в соответствии с заранее построенным расписанием. Для представленного подхода авторы разработали аналитическую модель, позволяющую оценить вероятность доставки кадров с заданной задержкой. Однако авторы лишь моделируют передачу данных по расписанию и не учитывают особен-

ности метода случайного доступа, используемого в сетях Wi-Fi, и его взаимодействия с механизмом R-TWT, поэтому разработанная ими модель не может использоваться для точной оценки пропускной способности пользователей сети Wi-Fi.

В работе [21] представлена аналитическая модель передачи RTA-трафика при помощи механизма R-TWT в сети Wi-Fi с пользовательскими станциями. Разработанная модель позволяет определить такие параметры R-TWT, как длительность R-TWT SP и период их следования, позволяющие удовлетворять требования к качеству обслуживания RTA-трафика при максимизации общей пропускной способности сети. Данная модель разработана в предположении, что, во-первых, в моменты R-TWT станции всегда имеют RTA-трафик для передачи, а во-вторых, передачи кадров внутри R-TWT SP защищены от коллизий с передачами станций, использующих для доставки данных случайный доступ к каналу. Эти предположения позволяют авторам оценивать влияние R-TWT на пропускную способность пользовательских станций через долю времени, занимаемого R-TWT SP. В то же время такой подход к оценке пропускной способности не учитывает долю времени, в течение которого канал может простаивать перед моментом R-TWT из-за необходимости всем станциям в сети завершить свои передачи до его наступления. Также отметим, что в сценариях с нерегулярным RTA-трафиком представленная модель может оказаться малоэффективной для оценки пропускной способности пользователей.

Обобщая все вышесказанное, можно сделать вывод, что ни одна из перечисленных аналитических моделей не позволяет оценить пропускную способность сети с несколькими пользователями, в которой используется механизм R-TWT, в то время как целью данной статьи является исследование влияния механизма R-TWT на пропускную способность пользователей. Отметим, что при использовании механизма R-TWT трафик RTA может влиять на трафик пользователей, к примеру, кадры RTA и кадры пользователей могут попадать в коллизии. В данной статье исследуется снижение пропускной способности пользователей только из-за невозможности пересечения собственной передачей момента R-TWT, поэтому не рассматриваются передачи кадров RTA в канале. Для достижения поставленной цели в статье ставится и решается следующая задача: *разработать аналитическую модель передачи данных в сети Wi-Fi с настроенным механизмом R-TWT, позволяющую оценить пропускную способность пользовательских станций в зависимости от периода R-TWT, используемой сигнально-кодовой конструкции и ограничения на длительность TXOP* (T_{lim}).

§ 4. Аналитическая модель

Рассмотрим сеть Wi-Fi, состоящую из точки доступа и N пользовательских станций, передающих кадры данных точке доступа в режиме насыщения с использованием фиксированной СКК. Станции передают кадры, используя механизм EDCA. При формировании кадров для передачи станции используют механизм агрегации A-MPDU совместно с механизмом блочного подтверждения. Размер скользящего окна, используемого в механизме блочных подтверждений, достаточен для передачи всех агрегируемых кадров. Для уменьшения длительности коллизии передача кадров данных начинается с обмена кадрами RTS/CTS. Точка доступа ожидает передачу в сети RTA-трафика, поэтому настраивает для его передачи механизм R-TWT с периодом T_p. Рассматривается временной промежуток работы сети, в течение которого не происходит передача трафика RTA. В рассматриваемой сети нет устаревших станций, т.е. все станции обладают функционалом R-TWT.

Считаем, что канал идеальный, т.е. шума в канале нет и потери кадров происходят только в случае коллизии с передачей другой станции. Все станции находятся в зоне радиовидимости друг друга и синхронно отсчитывают отсрочку.

Рассмотрим произвольную станцию, закончившую отсчет отсрочки за время T_q до ближайшего момента R-TWT. В зависимости от того, как соотносятся T_q, максимальная длительность обмена кадрами $T_s^{\max}$ и минимальная длительность обмена кадрами $T_s^{\min}$, могут возникнуть следующие ситуации:

1. $T_q \geqslant T_s^{\max}$. В этом случае момент R-TWT никак не повлияет на возможность передачи в выбранном слоте.
2. $T_s^{\min} \leqslant T_q < T_s^{\max}$. В этом случае момент R-TWT не повлияет на возможность передачи в выбранном слоте, но может ограничить длительность передачи.
3. $T_q < T_s^{\min}$. В этом случае в течение времени T_q станция не может начать передачу, но продолжает отсчитывать отсрочку. При завершении отсчета отсрочки в течение времени T_q станция выбирает новое значение случайной отсрочки и продолжает отсчет.

Разобьем интервал между моментами R-TWT на виртуальные слоты – интервалы времени между двумя последовательными изменениями значения счетчика отсрочки. Считаем, что вероятность передачи станции в виртуальном слоте τ является не зависящей от времени величиной (при условии, что передача в рассматриваемом виртуальном слоте не запрещена механизмом R-TWT) и вычисляется путем решения известной системы уравнений из [22, 23]:

$$\begin{cases} \tau = \dfrac{2(1-2p)(1-p^{R-1})}{W_{\min}(1-(2p)^{R-1})(1-p)+(1-2p)(1-p^{R-1})}, \\ p = 1-(1-\tau)^N, \end{cases} \tag{2}$$

где p – вероятность коллизии станции в виртуальном слоте при условии ее передачи.

Следуя подходу [23], найдем вероятности того, что слот окажется пустым (π_e), успешным (π_s) или коллизионным (π_c), следующим образом:

$$\pi_e = (1-\tau)^N, \tag{3}$$

$$\pi_s = N\tau(1-\tau)^{N-1}, \tag{4}$$

$$\pi_c = 1-\pi_e-\pi_s. \tag{5}$$

Длительность пустого виртуального слота равна длительности слота отсрочки T_e, а при определении длительностей успешного и коллизионного виртуальных слотов учитываются использование при передаче кадров RTS/CTS, а также использование механизма агрегации при формировании кадра данных:

$$T_s(T) = T_{\text{RTS}} + T_{\text{SIFS}} + T_{\text{CTS}} + T_{\text{SIFS}} + T_{\text{Data}}(T) + T_{\text{SIFS}} + T_{\text{BAck}} + T_{\text{AIFS}}, \tag{6}$$

$$T_c = T_{\text{RTS}} + T_{\text{EIFS}}, \tag{7}$$

где T_{RTS}, T_{CTS}, T_{BAck} – длительности передачи кадров RTS, CTS и кадра подтверждения BAck, $T_{\text{EIFS}} = T_{\text{SIFS}} + T_{\text{BAck}} + T_{\text{AIFS}}$, T_{SIFS} и T_{AIFS} – длительности соответствующих межкадровых интервалов, $T_{\text{Data}}(T)$ – длительность передачи агрегированного кадра пользовательской станции, которая зависит от времени T между началом передачи и следующим моментом R-TWT.

Опишем, как вычисляется длительность передачи агрегированного кадра данных. С вышестоящего уровня приложений на передачу поступает поток кадров по L байт каждый. Каждый такой кадр дополняется заголовком уровня доступа к каналу размером 36 байт и 4 байтами контрольной суммы. Далее к полученным кадрам добавляется последовательность из 4 байт, позволяющая получателю разделять кадры при получении одного агрегированного кадра данных, а также добавляется до 3 байт нагрузки для того, чтобы размер получившегося сегмента агрегированного кадра был кратен 4 байтам. Также в конец агрегированного кадра

добавляется кадр запроса блочного подтверждения BAR, к которому также добавляется разделяющая последовательность и нагрузка для кратности размера 4 байт, поэтому итоговый размер последнего сегмента получается 28 байт. Таким образом, размер кадра данных, передаваемого на физический уровень, составляет в байтах величину

$$L_{\text{mac}}(k) = k \cdot \left(11 + \left\lceil \frac{L}{4} \right\rceil\right) \cdot 4 + 28, \tag{8}$$

где k – число сегментов данных, агрегированных в общий кадр для передачи.

Для передачи в канале к полученным ранее L_{mac} байтам добавляются еще 22 служебных бита и дополнительная нагрузка, дополняющая передачу до целого числа OFDM-символов. Также добавляется заголовок физического уровня длительностью T_{pr}, состоящий из базового заголовка длительностью $T_{\text{pr}}^{*} = 20\,\text{мкс}$ и дополнительных полей длительностью $T_{\text{pr}}^{\text{HE}} = 16\,\text{мкс}$, введенных в стандарте IEEE 802.11ax:

$$T_{\text{pr}} = T_{\text{pr}}^{*} + T_{\text{pr}}^{\text{HE}} + N_{\text{HE-LTF}} \cdot T_{\text{HE-LTF}},$$

где $N_{\text{HE-LTF}}$ – число используемых в заголовке полей HE-LTF, которое зависит от числа используемых пространственных потоков для передачи кадра, $T_{\text{HE-LTF}}$ – длительность передачи поля HE-LTF. Таким образом, длительность передачи кадра A-MPDU равна

$$T_{\text{data}}^{*}(k) = T_{\text{pr}} + \left\lceil \frac{8L_{\text{mac}}(k) + 22}{K_{\text{sym}}} \right\rceil \cdot T_{\text{sym}}, \tag{9}$$

где K_{sym} и T_{sym} – число бит данных в OFDM-символе кадра данных и длительность OFDM-символа соответственно.

Длительности передачи кадров RTS, CTS и BAck рассчитываются аналогично (9), за исключением того, что $L_{\text{mac}}(k)$ соответствует их собственным размерам, используется только базовая часть заголовка физического уровня, т.е. $T_{\text{pr}} = T_{\text{pr}}^{*}$, а СКК выбирается из определенного стандартом Wi-Fi набора, соответствующего кадрам управления.

Отметим, что согласно (9) при фиксированном размере L кадров данных и параметров K_{sym}, T_{sym} длительность передачи $T_{\text{data}}^{*}(k)$ зависит только от числа сегментов k, из которых сформирован агрегированный кадр. Считаем, что процесс агрегации кадров происходит достаточно быстро и кадр может пересобираться несколько раз в процессе отсчета отсрочки. Тогда для выполнения ограничений, задаваемых одновременно временем T_{lim} и временем T, оставшимся до момента R-TWT, станция выбирает такое максимальное $k(T)$, при котором

$$\begin{aligned} &k(T) = \arg\max_{k} T_{\text{data}}^{*}(k) \quad \text{при условии} \\ &T_{\text{RTS}} + T_{\text{CTS}} + T_{\text{data}}^{*}(k) + T_{\text{BAck}} + 3T_{\text{SIFS}} \leqslant \min(T; T_{\text{lim}}). \end{aligned} \tag{10}$$

Таким образом получаем, что длительность агрегированного кадра данных $T_{\text{data}}(T)$ в уравнении (6) вычисляется при помощи уравнений (10) и (9) и равна $T_{\text{data}}(T) = T_{\text{data}}^{*}(k(T))$.

Пропускную способность станций с учетом использования в сети механизма R-TWT будем считать как

$$S = \frac{\overline{L}}{T_p}, \tag{11}$$

где $\overline{L}$ – среднее число байт полезной информации, успешно переданной за период R-TWT T_p.

Пронумеруем виртуальные слоты t от момента R-TWT до последнего момента, в который может начаться передача в данном периоде $T_p - T_s^{\min}$, где $T_s^{\min}$ – минимально возможная длительность успешного обмена кадрами в сети без учета последующего межкадрового интервала AIFS:

$$T_s^{\min} = T_{\text{RTS}} + T_{\text{CTS}} + T^*_{\text{data}}(1) + T_{\text{BAck}} + 3 \cdot T_{\text{SIFS}}.$$

Отметим, что число таких слотов не превышает величину

$$t_{\max} = \left\lceil \frac{T_p - T_s^{\min}}{T_e} \right\rceil.$$

Если в периоде R-TWT будет выполнено несколько успешных передач кадров данных, то только длительность последней передачи может варьироваться и оказаться короче предыдущих, длительность которых ограничивается только временем T_{lim}. Поэтому далее будем вычислять вероятность $P_s(t+1, c, s)$ того, что в слоте $t+1$ произойдет последняя успешная передача в периоде R-TWT с учетом того, что s и c – число успешных и коллизионных слотов, предшествующих рассматриваемому слоту. Важно отметить, что время T_{tail} от начала слота $t+1$ до момента R-TWT зависит от состояний всех предыдущих слотов, так как длительности успешного, пустого и коллизионного слотов отличаются, поэтому

$$T_{\text{tail}}(t, c, s) = T_p - T_s(T_{\text{lim}}) \cdot s - T_c \cdot c - T_e \cdot (t - c - s).$$

При этом вероятность и длительность передачи в слоте $t+1$ не зависит от последовательности успешных, пустых и коллизионных слотов среди предшествующих t, а только от количества слотов каждого вида, так как это влияет на длительность $T_{\text{tail}}(t, c, s)$ и, как следствие, на возможность передачи в рассматриваемом слоте. Для того чтобы передача в слоте $t+1$ была возможна, длительность T_{tail} должна быть достаточна для ее выполнения, т.е. $T_{\text{tail}} \geqslant T_s^{\min}$. Если рассматриваемая передача закончится до наступления последнего виртуального слота в периоде R-TWT, в котором возможна передача, т.е. при $T_{\text{tail}} > T_s(T_{\text{tail}})$, то для того чтобы рассматриваемая передача оказалась последней в R-TWT периоде, после нее должны быть только пустые или коллизионные слоты.

Таким образом, вероятность выполнения в слоте $t+1$ последней успешной передачи в рассматриваемом периоде при $T_{\text{tail}}(t, c, s) \geqslant T_s^{\min}$ рассчитывается как

$$P_s(t+1, c, s) = \pi_s C_t^s C_{t-s}^c \pi_e^{t-s-c} \pi_s^s \pi_c^c P_{\text{tail}}(T_{\text{tail}}(t, c, s) - T_s(T_{\text{tail}}(t, c, s))), \tag{12}$$

где C_t^s учитывает число вариантов выбрать s успешных слотов из t предшествующих рассматриваемому, C_{t-s}^c – число вариантов c коллизионных слотов из оставшихся $t-s$, а $P_{\text{tail}}(T^*)$ – вероятность того, что в течение интервала времени T^* все виртуальные слоты были пустыми или коллизионными. В противном случае, т.е. при $T_{\text{tail}}(t, c, s) < T_s^{\min}$, принимаем $P_s(t+1, c, s) = 0$. Отметим, что если рассматриваемая последняя в периоде успешная передача будет совершена в последнем возможном для нее виртуальном слоте, т.е. $T_{\text{tail}} \leqslant T_s(T_{\text{tail}})$, то вероятность P_{tail} принимаем равной единице.

Теперь перейдем к подсчету вероятности $P_{\text{tail}}(T^*)$, которая соответствует тому, что за время T^* до момента R-TWT не произойдет ни одной успешной передачи. Подсчитывать данную вероятность будем аналогично вероятности успеха P_s, т.е. разобьем интервал времени $T^* - T_s^{\min}$ на виртуальные слоты $\widehat{t}$, число которых

не превосходит

$$\widehat{t}_{\max} = \left\lfloor \frac{T^* - T_s^{\min}}{T_e} \right\rfloor + 1,$$

и будем считать вероятность того, что в слоте с номером $\widehat{t}+1$ произойдет последняя коллизия на интервале $T^* - T_s^{\min}$. Отметим, что рассматриваются только те виртуальные слоты, которые начинаются на интервале $T^* - T_s^{\min}$, так как в них возможна передача и, как следствие, коллизия, а оставшиеся виртуальные слоты окажутся пустыми из-за особенностей механизма R-TWT, поэтому они не рассматриваются. Если в рассматриваемом слоте $(\widehat{t}+1)$ происходит последняя коллизия на интервале T^*, то после нее до момента $T^* - T_s^{\min}$ все рассматриваемые виртуальные слоты должны быть пустыми. Вероятность этого можно учесть путем подсчета числа таких пустых слотов $\widehat{e}$ после рассматриваемого коллизионного слота $(\widehat{t}+1)$ следующим образом:

$$\widehat{e}(\widehat{t},\widehat{c}) = \max\left(0, \left\lfloor \frac{T^* - T_s^{\min} - (\widehat{c}+1)T_c - (\widehat{t}-\widehat{c})T_e}{T_e} \right\rfloor + 1\right),$$

где $\widehat{c}$ – число коллизионных слотов из $\widehat{t}$ предшествующих рассматриваемому. Таким образом, вероятность отсутствия за время T^* успешных передач определяется следующим образом:

$$P_{\text{tail}}(T^*) = \begin{cases} \pi_c \sum\limits_{t=0}^{\widehat{t}_{\max}} \sum\limits_{\widehat{c}=0}^{\widehat{c}_{\max}} C_{\widehat{t}}^{\widehat{c}} \pi_e^{\widehat{t}-\widehat{c}+\widehat{e}(\widehat{t},\widehat{c})} \pi_c^{\widehat{c}}, & \text{если } T^* \geq T_s^{\min}, \\ 1 & \text{в противном случае,} \end{cases} \tag{13}$$

где

$$\widehat{c}_{\max} = \left\lfloor \frac{T^* - T_s^{\min}}{T_c} \right\rfloor$$

– максимальное возможное число коллизий за время $T^* - T_s^{\min}$, а $C_{\widehat{t}}^{\widehat{c}}$ – число способов выбрать $\widehat{c}$ коллизионных слотов из $\widehat{t}$ виртуальных слотов, прошедших к началу рассматриваемого.

С учетом (10), (12) и (13) можно рассчитать среднее количество переданных в периоде T_p данных $\overline{L}$:

$$\overline{L} = \sum_{t=0}^{t_{\max}} \sum_{s=0}^{s_{\max}} \sum_{c=0}^{c_{\max}} P_s(t+1,c,s)(k(T_{\lim})\cdot s + k(T_{\text{tail}}(t,c,s)))L, \tag{14}$$

где

$$s_{\max} = \min\left(t, \left\lfloor \frac{T_p}{T_s(T_{\lim})} \right\rfloor\right), \quad c_{\max} = \min\left(t-s, \left\lfloor \frac{T_p - s\cdot T_s(T_{\lim})}{T_c} \right\rfloor\right)$$

– ограничения по возможному числу успешных и коллизионных слотов среди t слотов, прошедших к рассматриваемому моменту.

Отметим следующее: механизм R-TWT требует, чтобы обмен кадрами завершился до момента R-TWT. При этом механизм случайного доступа к каналу требует, чтобы передачи разных станций в сети были разделены по крайней мере интервалом AIFS. Таким образом, если подтверждение на последнюю успешную передачу было получено менее, чем за T_{AIFS} до окончания периода R-TWT, то эффективная длительность следующего периода R-TWT уменьшается, ведь первый виртуальный

слот в периоде начнется не в момент R-TWT, а спустя T_{AIFS} после получения кадра BAck. Рассчитаем время $\overline{\mathcal{T}}$, на которое передача в одном периоде R-TWT может уменьшить эффективную длительность следующего периода R-TWT, для чего вернемся к рассмотрению последнего успешного виртуального слота в периоде и уравнению (12). Рассмотрим передачи, для которых $T_{\text{tail}}(t,c,s) < T_s(T_{\text{tail}}(t,c,s))$. Для таких передач

$$\mathcal{T} = T_s(T_{\text{tail}}(t,c,s)) - T_{\text{tail}}(t,c,s)$$

с вероятностью $P_s(t+1,c,s)$. Если обмен кадрами завершается не позднее чем за T_{AIFS} до окончания периода R-TWT, то влияние на следующий период R-TWT ограничивается тем, что границы виртуальных слотов и моменты R-TWT не синхронизированы, а передача может начаться только на границе виртуального слота. В среднем такое влияние можно оценить как $\mathcal{T} = \frac{T_e}{2}$. Таким образом, время $\overline{\mathcal{T}}$ вычисляется аналогично (14) следующим образом:

$$\begin{aligned}
&\overline{\mathcal{T}} = \sum_{t=0}^{t_{\max}} \sum_{s=0}^{s_{\max}} \sum_{c=0}^{c_{\max}} P_s(t+1,c,s)\mathcal{T}(t,c,s), \\
&\mathcal{T}(t,c,s) = \\
&= \begin{cases} T_s(T_{\text{tail}}(t,c,s)) - T_{\text{tail}}(t,c,s), & \text{если } T_{\text{tail}}(t,c,s) < T_s(T_{\text{tail}}(t,c,s)); \\ \dfrac{T_e}{2} & \text{в противном случае.} \end{cases}
\end{aligned} \tag{15}$$

Время $\overline{\mathcal{T}}$ следует учитывать во всех предыдущих расчетах вероятностей и длительностей передачи, заменив T_p на $T_p^* = T_p - \overline{\mathcal{T}}$.

Для нахождения пропускной способности системы в модели используется итерационный подход, при котором на первом шаге на вход модели подается $T_p^* = T_p$, а далее на каждом новом шаге i на вход модели подается

$$T_p^* = T_p - \overline{\mathcal{T}}_{i-1},$$

где $\overline{\mathcal{T}}_{i-1}$ – значение $\overline{\mathcal{T}}$, определенное на предыдущем шаге итерации $i-1$. Пропускная способность системы $\overline{S}_i$ на шаге i вычисляется как отношение суммы $\sum\limits_{j=0}^{i} \overline{L}_j$ всей переданной полезной информации за все предыдущие i шагов к числу периодов:

$$\overline{S}_i = \frac{\sum\limits_{j=0}^{i} \overline{L}_j}{i \cdot T_p},$$

где $\overline{L}_j$ – среднее число байт полезной информации, успешно переданной на шаге j итерации. Последним шагом итерации считаем тот, при котором значение пропускной способности системы незначительно отличается от значения, полученного на предыдущем шаге, т.е. при

$$\left| \frac{\overline{S}_i - \overline{S}_{i-1}}{\overline{S}_{i-1}} \right| < \varepsilon.$$

Важно отметить, что в (11) не выполняется замена T_p на T_p^*, так как в нем вычисляется среднее число полезной информации именно за период между последовательными моментами R-TWT.

Таблица 1

Параметры сценария

Величина	Значение
Число пользовательских станций, N	5
Ширина канала	20 МГц
Количество пространственных потоков	2
Количество полей HE-LTF, $N_{\text{HE-LTF}}$	2
Длительность передачи поля HE-LTF, $T_{\text{HE-LTF}}$	4 мкс
Битовая скорость:	
• кадров данных при использовании СКК 4	103,2 Мбит/с
• кадров данных при использовании СКК 11	286,8 Мбит/с
• служебных кадров при использовании СКК 3	18 Мбит/с
Длительность OFDM-символа T_{sym} для:	
• служебного кадра	4 мкс
• кадра данных	13,6 мкс
Число бит данных K_{sym} в OFDM-символе:	
• кадров данных при использовании СКК 4	1404 бита
• кадров данных при использовании СКК 11	3900 бит
• служебных кадров при использовании СКК 3	72 бита
Длительность межкадрового интервала:	
• SIFS, T_{SIFS}	16 мкс
• AIFS, T_{AIFS}	43 мкс
Длительность слота отсрочки, T_e	9 мкс
Минимальное конкурентное окно, $W_{\min}$	16
Максимальное конкурентное окно, $W_{\max}$	1024
Максимальное количество попыток передачи одного кадра, R	7
Полезная нагрузка в одном сегменте данных, L	1500 байт
Размер кадра на уровне доступа к каналу:	
• BAck, L_{BAck}	32 байта
• RTS, L_{RTS}	20 байт
• CTS, L_{CTS}	14 байт

§ 5. Результаты моделирования

Для оценки точности разработанной аналитической модели была разработана дискретно-событийная имитационная модель передачи данных в сети Wi-Fi, учитывающая особенности использования в сети механизма R-TWT. В отличие от аналитической модели, имитационная модель не использует предположение о независимости вероятности передачи станции в виртуальном слоте от предыстории, а подробно моделирует процесс отсчета отсрочки станции. Сравнение результатов аналитического и имитационного моделирования в рамках описанного в начале § 4 сценария было проведено при использовании приведенных в табл. 1 параметров. Для завершения итерационного расчета использовался параметр $\varepsilon = 0{,}001$.

На рис. 1 представлена зависимость пропускной способности S сети Wi-Fi от периода R-TWT T_p при $T_{\text{lim}} = 1$ мс и используемой СКК 4. На данном графике маркерами обозначены значения пропускной способности, полученные при помощи имитационной модели (“им.”), а линия соединяет значения, полученные при помощи аналитической модели (“ан.”). Сравнивая полученные при помощи данных моделей результаты, можно отметить, что они совпадают с высокой точностью. Далее перейдем к подробному анализу самой зависимости.

При увеличении периода R-TWT от 400 мкс до 1 мс пропускная способность сети увеличивается. Это связано с тем, что при таких периодах R-TWT в каждом из них передается не более одного кадра данных, а при увеличении периода R-TWT увеличивается размер агрегированного кадра и размер полезной нагрузки при неизменных накладных расходах. В то же время следует отметить некоторую

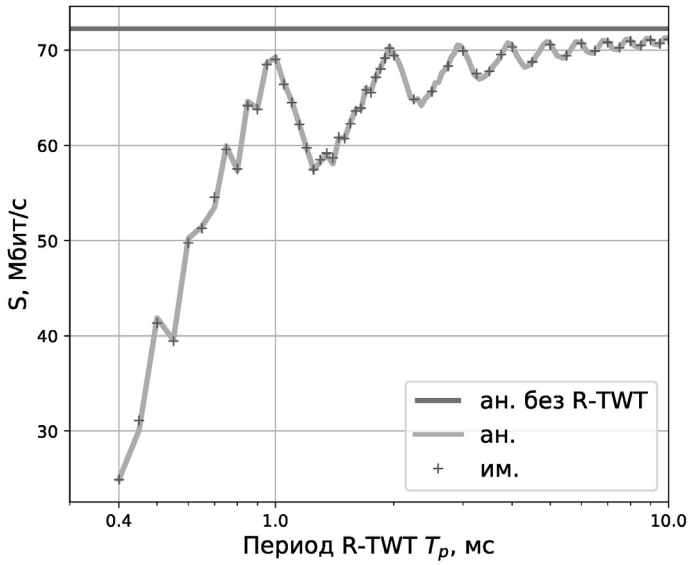


Рис. 1. Зависимость пропускной способности S сети Wi-Fi от периода R-TWT T_p при $T_{\text{lim}} = 1$ мс и СКК 4

ступенчатость зависимости на данном интервале, которая объясняется дискретностью размера и, соответственно, длительности передаваемых в сети кадров данных. Первый максимум пропускной способности достигается при минимальном периоде R-TWT, в котором помещается передача кадра данных, ограниченная только длительностью T_{lim}. При дальнейшем увеличении периода R-TWT пропускная способность сети уменьшается, так как длительность периода все еще вмещает только одну передачу. Уменьшение пропускной способности прекращается в тот момент, когда в периоде R-TWT становится возможно выполнить две передачи данных. Отметим, что от первого максимума до минимума падение пропускной способности составляет 17%. Дальнейшее увеличение числа передаваемых в периоде R-TWT кадров приводит к уменьшению влияния на пропускную способность накладных расходов на передачу последнего в периоде кадра данных, размер которого может быть меньше максимально разрешенного T_{lim}. Именно поэтому с увеличением периода R-TWT падение пропускной способности между последовательными максимумами и минимумами уменьшается, а в пределе при $T_p \to \infty$ значение пропускной способности стремится к пропускной способности сети без механизма R-TWT, которое на графике показано горизонтальной прямой (“ан. без R-TWT”).

На рис. 2 представлены аналогичные зависимости пропускной способности сети от периода R-TWT для $T_{\text{lim}} = 1$ мс и $T_{\text{lim}} = 2,5$ мс и используемых станциями при передаче сигнально-кодовых конструкций СКК 4 и СКК 11. Сравнение данных зависимостей при фиксированной СКК показывает, что увеличение T_{lim} позволяет уменьшить максимальный относительный перепад пропускной способности, что связано с относительным уменьшением накладных расходов. Как было сказано ранее, максимумы пропускной способности достигаются при минимальных периодах R-TWT, в которых все передаваемые кадры имеют максимальный размер. Поэтому при увеличении T_{lim} от 1 мс до 2,5 мс число локальных максимумов пропускной способности на рассматриваемом интервале времени уменьшилось в два раза.

Перейдем к сравнению зависимостей пропускной способности сети от периода R-TWT для СКК 4 и СКК 11 при фиксированном значении T_{lim}. Увеличение по-

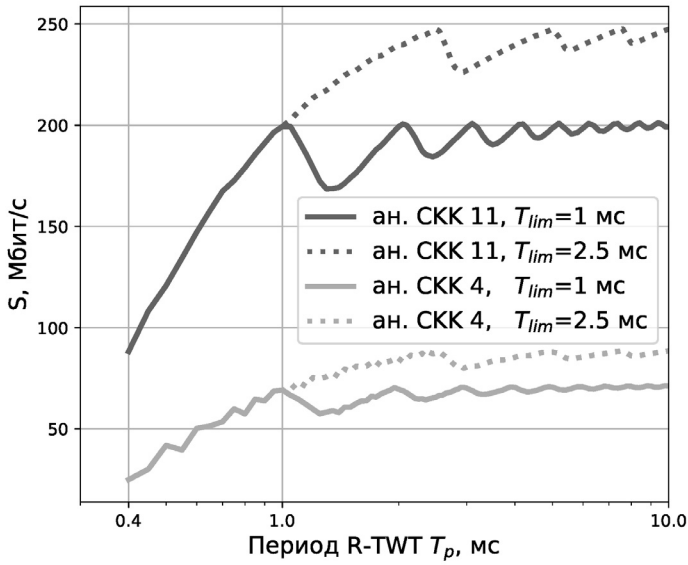


Рис. 2. Сравнение зависимостей пропускной способности S сети Wi-Fi от периода R-TWT T_p при $T_{\text{lim}} = 1\,\text{мс}$ и $T_{\text{lim}} = 2,5\,\text{мс}$, СКК 4 и СКК 11

рядка СКК приводит к уменьшению длительности передачи сегментов агрегированного кадра и, соответственно, минимальной длительности передачи $T_s^{\min}$. Это приводит к уменьшению времени простоя канала, из-за чего уменьшаются относительные перепады пропускной способности, а также зависимость становится более гладкой при $T_p < 1\,\text{мс}$. В то же время абсолютное значение перепадов пропускной способности растет при увеличении СКК, что связано с увеличением объема передаваемых данных в единицу времени и, как следствие, увеличением влияния неиспользуемого канального времени на пропускную способность сети. Это также приводит к увеличению периода осцилляции зависимости пропускной способности от периода R-TWT.

Подводя итог вышесказанному, подчеркнем, что зависимость пропускной способности сети от периода R-TWT является немонотонной, т.е. увеличение периода R-TWT не гарантирует увеличения пропускной способности сети. Это связано с запретом для станций пересечения передачей момента R-TWT, что может приводить как к ограничению длительности передачи, так и к запрету начала передачи в конце периода R-TWT. Таким образом, выбор наибольшего периода R-TWT, задающего минимально необходимое число моментов R-TWT для передачи RTA-кадра, может приводить к получению меньшей пропускной способности сети, чем при более частом назначении R-TWT.

§ 6. Заключение

В данной статье был исследован новейший механизм ограниченной защиты доступа к каналу R-TWT в сетях Wi-Fi 7, который предназначен для доставки высокоприоритетного трафика, чувствительного к задержкам. Разработана аналитическая модель передачи данных в сети Wi-Fi, использующей механизм R-TWT, а также показана ее высокая точность при помощи имитационной модели. С помощью разработанной модели оценивалась зависимость пропускной способности пользовательских станций, передающих низкоприоритетные потоки данных, от периода резерви-

рований R-TWT, используемых станциями сигнально-кодовых конструкций и ограничения на длительность обмена кадрами T_{lim} Показано, что данная зависимость немонотонная, а точки локальных минимумов и максимумов зависят не только от периода R-TWT, но и от установленного для станций ограничения T_{lim}.

СПИСОК ЛИТЕРАТУРЫ

1. *Chemrov K., Bankov D., Khorov E., Lyakhov A.* Smart Preliminary Channel Access to Support Real-Time Traffic in Wi-Fi Networks // Future Internet. 2022. V. 14. № 10. Paper No. 296 (14 pp.). https://doi.org/10.3390/fi14100296
2. *Угловский А.Ю., Мельников И.А., Алексеев И.А., Куреев А.А.* Оценка низкого уровня ошибок с помощью выборки по значимости с равномерным распределением // Пробл. передачи информ. 2023. Т. 59. № 4. С. 3–12. https://doi.org/10.31857/S0555292323040010
3. *Фернандес М., Кабатянский Г.А., Круглик С.А., Мяо И.* Коды для точного нахождения носителя разреженного вектора по ошибочным линейным измерениям и их декодирование // Пробл. передачи информ. 2023. Т. 59. № 1. С. 17–24. https://doi.org/10.31857/S0555292323010023
4. *Лихтциндер Б.Я., Привалов А.Ю., Моисеев В.И.* Неординарные пуассоновские модели трафика мультисервисных сетей // Пробл. передачи информ. 2023. Т. 59. № 1. С. 71–79. https://doi.org/10.31857/S0555292323010060
5. *Лихтциндер Б.Я., Привалов А.Ю.* Обобщение формул для моментов очереди при неординарном пуассоновском потоке для очередей пакетов в системах телекоммуникаций // Пробл. передачи информ. 2023. Т. 59. № 4. С. 32–37. https://doi.org/10.31857/S0555292323040046
6. *Hu C., Torab P., Cherian G., Ho D.* Protected TWT Enhancement for Latency Sensitive Traffic. IEEE 802.11-20/1046r1. Jul. 29, 2020. https://mentor.ieee.org/802.11/dcn/20/11-20-1046-01-00be-prioritized-edca-channel-access-slot-management.pptx.
7. *Bankov D.V., Lyakhov A.I., Stepanova E.A., Khorov E.M.* Performance Evaluation of Wi-Fi 7 Networks with Restricted Target Wake Time // Probl. Inf. Transm. 2024. V. 60. № 3. P. 233–254. https://doi.org/10.1134/S0032946024030062
8. *Qiu W., Chen G., Nguyen K.N., Sehgal A., Nayak P., Choi J.* Category-Based 802.11ax Target Wake Time Solution // IEEE Access. 2021. V. 9. P. 100154–100172. https://doi.org/10.1109/ACCESS.2021.3096940
9. *Степанова Е.А., Банков Д.В., Хоров Е.М., Ляхов А.И.* Влияние дрейфа часов на эффективность механизмов энергосбережения в сетях Wi-Fi // Информационные процессы. 2022. Т. 22. № 4. С. 261–275. http://www.jip.ru/2022/261-275-2022.pdf
10. *Cavalcanti D., Cordeiro C., Smith M., Regev A.* WiFi TSN: Enabling Deterministic Wireless Connectivity over 802.11 // IEEE Commun. Stand. Mag. 2022. V. 6. № 4. P. 22–29. https://doi.org/10.1109/MCOMSTD.0002.2200039
11. *Adame T., Carrascosa-Zamacois M., Bellalta B.* Time-Sensitive Networking in IEEE 802.11be: On the Way to Low-Latency WiFi 7 // Sensors. 2021. V. 21. № 15. Paper No. 4954 (20 pp.). https://doi.org/10.3390/s21154954
12. *Galati-Giordano L., Geraci G., Carrascosa M., Bellalta B.* What will Wi-Fi 8 Be? A Primer on IEEE 802.11bn Ultra High Reliability // IEEE Commun. Mag. 2024. V. 62. № 8. P. 126–132. https://doi.org/10.1109/MCOM.001.2300728
13. *Shaji George A., Hovan George A.S., Baskar T.* Wi-Fi 7: The Next Frontier in Wireless Connectivity // Partners Univ. Int. Innov. J. 2023. V. 1. № 4. P. 133–145. https://doi.org/10.5281/zenodo.8266217
14. *Reshef E., Cordeiro C.* Future Directions for Wi-Fi 8 and Beyond // IEEE Commun. Mag. 2022. V. 60. № 10. P. 50–55. https://doi.org/10.1109/MCOM.003.2200037
15. *Avallone S., Imputato P., Magrin D.* Controlled Channel Access for IEEE 802.11-Based Wireless TSN Networks // IEEE Internet Things Mag. 2023. V. 6. № 1. P. 90–95. https://doi.org/10.1109/IOTM.001.2200070

16. *Guérin E., Begin T., Lassous I.G.* An Overview of MAC Energy-Saving Mechanisms in Wi-Fi // Comput. Commun. 2023. V. 203. P. 129–145. https://doi.org/10.1016/j.comcom.2023.03.003
17. *Haxhibeqiri J., Jiao X., Shen X., Pan C., Jiang X., Hoebeke J.* Coordinated SR and Restricted TWT for Time Sensitive Applications in WiFi 7 Networks // IEEE Commun. Mag. 2024. V. 62. № 8. P. 118–124. https://doi.org/10.1109/MCOM.001.2300431
18. *Liu S., Wang D., Yang M., Yan Z., Li B.* A Channel Reservation Mechanism in IEEE 802.11be for Multi-cell Scenarios // Smart Grid and Internet of Things: Proc. 6th EAI Int. Conf. (SGIoT 2022). TaiChung, Taiwan. Nov. 19–20, 2022. Lect. Notes Inst. Comput. Sci. Soc.-Inform. Telecommun. Eng. V. 497. Cham: Springer, 2023. P. 241–253. https://doi.org/10.1007/978-3-031-31275-5_23
19. *Peng X., Fang Y., Li C., Guo L.* Access Point Coordination Based TWT Scheduling for the Next Generation WLAN // 2024 IEEE 13th Int. Conf. on Communications, Circuits and Systems (ICCCAS 2024). Xiamen, China. May 10–12, 2024. P. 238–243. https://doi.org/10.1109/ICCCAS62034.2024.10652675
20. *Barroso-Fernández C., Martín-Pérez J., Ayimba C., De La Oliva A.* Aligning rTWT with 802.1Qbv: A Network Calculus Approach // Proc. 24th Int. Symp. on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc'23). Washington, DC, USA. Oct. 23–26, 2023. P. 352–354. https://doi.org/10.1145/3565287.3617606
21. *Belogaev A., Shen X., Pan C., Jiang X., Blondia C., Famaey J.* Dedicated Restricted Target Wake Time for Real-Time Applications in Wi-Fi 7, https://arxiv.org/abs/2402.15900 [cs.NI], 2024.
22. *Bianchi G.* Performance Analysis of the IEEE 802.11 Distributed Coordination Function // IEEE J. Sel. Areas Commun. 2000. V. 18. № 3. P. 535–547. https://doi.org/10.1109/49.840210
23. *Vishnevsky V.M., Lyakhov A.I.* IEEE 802.11 Wireless LAN: Saturation Throughput Analysis with Seizing Effect Consideration // Cluster Comput. 2002. V. 5. P. 133–144. https://doi.org/10.1023/A:1013977425774

Банков Дмитрий Викторович
Ляхов Андрей Игоревич
Степанова Екатерина Алексеевна
Хоров Евгений Михайлович
Институт проблем передачи информации
им. А.А. Харкевича РАН, Москва
bankov@iitp.ru
lyakhov@iitp.ru
stepanova@iitp.ru
khorov@iitp.ru

Поступила в редакцию
13.11.2024
После доработки
13.11.2024
Принята к публикации
28.11.2024

Адрес редакции: 127051, Москва, Б. Каретный пер., 19, стр. 1, тел. (495) 650-47-39

Оригинал-макет подготовил *Д.Ю. Ногин*

**Москва
ФГБУ «Издательство «Наука»**