# Лекция 11: Комбинаторная сложность

А. М. Шур

Институт математики и компьютерных наук (матмех) УрФУ

28 марта 2015

# О размерах множеств

- Если даны два множества, в одном из которых пять элементов, а в другом десять — мы легко определим, какое из множеств больше.

# О размерах множеств

- Если даны два множества, в одном из которых пять элементов, а в другом десять — мы легко определим, какое из множеств больше.
- Если даны два множества, одно из которых конечно, а другое — бесконечно, задача становится еще проще.

# О размерах множеств

- Если даны два множества, в одном из которых пять элементов, а в другом десять — мы легко определим, какое из множеств больше.
- Если даны два множества, одно из которых конечно, а другое — бесконечно, задача становится еще проще.
- А если даны два бесконечных, пусть даже счетных, множества?

Теория множеств говорит, что они равны, но так ли это с практической точки зрения?

# О размерах множеств

- Если даны два множества, в одном из которых пять элементов, а в другом десять — мы легко определим, какое из множеств больше.
- Если даны два множества, одно из которых конечно, а другое — бесконечно, задача становится еще проще.
- А если даны два бесконечных, пусть даже счетных, множества?

Теория множеств говорит, что они равны, но так ли это с практической точки зрения?

★ Скорее нет: ввиду ограниченности ресурсов, нам доступны не все элементы бесконечного множества, а только те, «описания» которых удовлетворяют ресурсным ограничениям

- например, мы можем производить действия не с любыми целыми числами, а только с теми, которые помещаются в выбранный тип данных

Таких элементов конечное число, а значит, сравнение бесконечных множеств в «практическом» смысле сводится к сравнению конечных: то множество, в котором больше доступных элементов, и является бóльшим. Все сказанное относится и к множествам слов.

## Определение

(Формальным) языком над алфавитом $\Sigma$ называется произвольное множество слов над $\Sigma$. Комбинаторной сложностью языка $L \subseteq \Sigma^*$ называется функция $C_L(n) : \mathbb{N}_0 \to \mathbb{N}_0$, возвращающая число слов заданной длины в языке.

**Определение**

(Формальным) языком над алфавитом $\Sigma$ называется произвольное множество слов над $\Sigma$. Комбинаторной сложностью языка $L \subseteq \Sigma^*$ называется функция $C_L(n) : \mathbb{N}_0 \to \mathbb{N}_0$, возвращающая число слов заданной длины в языке.

Сравнение «размеров» двух языков — это сравнение их комбинаторных сложностей, причем обычно в асимптотическом смысле:

- язык $L_1$ «богаче» языка $L_2$, если $C_{L_1}(n) > C_{L_2}(n)$ для всех больших $n$

Понятие комбинаторной сложности применимо к любому языку, но его осмысленно рассматривать можно только для тех языков, для которых разрешима проблема вхождения: по заданному слову определить, принадлежит ли оно языку (такие языки называются рекурсивными).

**Определение**

(Формальным) языком над алфавитом $\Sigma$ называется произвольное множество слов над $\Sigma$. Комбинаторной сложностью языка $L \subseteq \Sigma^*$ называется функция $C_L(n) : \mathbb{N}_0 \to \mathbb{N}_0$, возвращающая число слов заданной длины в языке.

Сравнение «размеров» двух языков — это сравнение их комбинаторных сложностей, причем обычно в асимптотическом смысле:

- язык $L_1$ «богаче» языка $L_2$, если $C_{L_1}(n) > C_{L_2}(n)$ для всех больших $n$

Понятие комбинаторной сложности применимо к любому языку, но его осмысленно рассматривать можно только для тех языков, для которых разрешима проблема вхождения: по заданному слову определить, принадлежит ли оно языку (такие языки называются рекурсивными).

Мы начнем изучение комбинаторной сложности с того же частного случая, с которого начали Морс и Хэдлунд в 1938 году.

## Определение

Комбинаторной сложностью (или сложностью по подсловам) бесконечного слова $\mathbf{w}$ называется функция $C_{\mathbf{w}}(n) : \mathbb{N}_0 \to \mathbb{N}_0$, возвращающая число подслов заданной длины в $\mathbf{w}$.

## Определение

Комбинаторной сложностью (или сложностью по подсловам) бесконечного слова $\mathbf{w}$ называется функция $C_{\mathbf{w}}(n) : \mathbb{N}_0 \to \mathbb{N}_0$, возвращающая число подслов заданной длины в $\mathbf{w}$.

Марстон Морс (1892-1977) — американец, дифференциальный геометр и тополог. Придумал «Теорию Морса» в топологии.

Вместе со своим учеником Хэдлундом впервые исследовал комбинаторную сложность бесконечных слов. Сформулировал на языке сложности критерий периодичности $\omega$-слова (1938) и критерий для слов Штурма (1940).

# Периодичность и сложность

## Определение

$\omega$-слово $\mathbf{w}$ называется

- периодическим, если у него есть период $p \in \mathbb{N}$, т.е. $\mathbf{w}[i] = \mathbf{w}[i+p]$ для всех $i \in \mathbb{N}$;
- финально периодическим, если у него есть периодический суффикс.

Таким образом,
- периодическое $\omega$-слово имеет вид $v^\omega$, где $|v| = p$;
- финально периодическое $\omega$-слово имеет вид $uv^\omega$, где $|v| = p$, а $u$ – любое.

# Периодичность и сложность

## Определение

$\omega$-слово $\mathbf{w}$ называется

- периодическим, если у него есть период $p \in \mathbb{N}$, т.е. $\mathbf{w}[i] = \mathbf{w}[i+p]$ для всех $i \in \mathbb{N}$;
- финально периодическим, если у него есть периодический суффикс.

Таким образом,
- периодическое $\omega$-слово имеет вид $v^\omega$, где $|v| = p$;
- финально периодическое $\omega$-слово имеет вид $uv^\omega$, где $|v| = p$, а $u$ – любое.

## Теорема (Морс, Хэдлунд, 1938)

$\omega$-слово $\mathbf{w}$ является финально периодическим тогда и только тогда, когда его комбинаторная сложность ограничена.

# Периодичность и сложность

## Определение

$\omega$-слово $\mathbf{w}$ называется

- периодическим, если у него есть период $p \in \mathbb{N}$, т.е. $\mathbf{w}[i] = \mathbf{w}[i+p]$ для всех $i \in \mathbb{N}$;
- финально периодическим, если у него есть периодический суффикс.

Таким образом,
- периодическое $\omega$-слово имеет вид $v^\omega$, где $|v| = p$;
- финально периодическое $\omega$-слово имеет вид $uv^\omega$, где $|v| = p$, а $u$ – любое.

## Теорема (Морс, Хэдлунд, 1938)

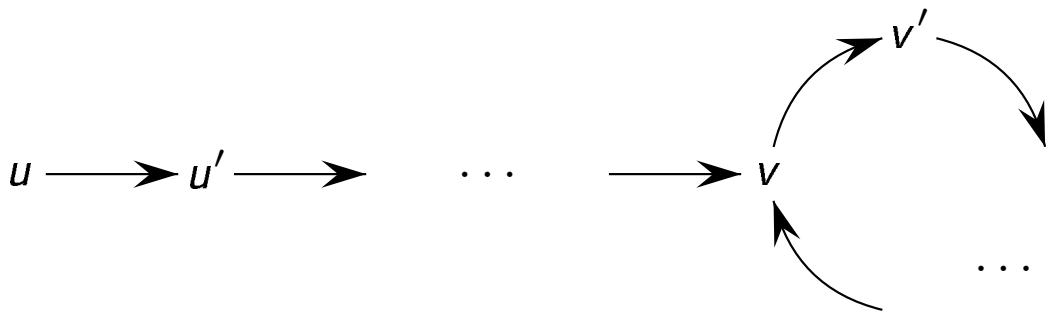
$\omega$-слово $\mathbf{w}$ является финально периодическим тогда и только тогда, когда его комбинаторная сложность ограничена.

$\Rightarrow$: В слове $v^\omega$ не более $p$ различных подслов длины $n$ для любого $n$; это подслова, начинающиеся с позиций $1, 2, \ldots, p$ в $v$. В слове $uv^\omega$ к ним добавляются не более $|u|$ подслов, начинающихся внутри $u$. $\square$

# Доказательство теоремы Морса-Хэдлунда

$\Leftarrow$: Пусть комбинаторная сложность $\omega$-слова $\mathbf{w}$ ограничена.

- Найдется $n_0$ такое, что $C_{\mathbf{w}}(n_0+1) \leqslant C_{\mathbf{w}}(n_0)$
- С другой стороны, если $u$ – подслово в $\mathbf{w}$, выберем любое его вхождение; если за ним в $\mathbf{w}$ следует буква $a$, то $ua$ – подслово в $\mathbf{w}$
- Такое отображение $u \to ua$ инъективно, откуда $C_{\mathbf{w}}(n+1) \geqslant C_{\mathbf{w}}(n)$ для любого $n$; значит, $C_{\mathbf{w}}(n_0+1) = C_{\mathbf{w}}(n_0)$
- Тогда для любого подслова $u$ длины $n_0$ следующая за ним буква $a$ определяется однозначно, т.е. на одну позицию правее вхождения $u$ всегда начинается вхождение одного и того же подслова $u'$ (а именно, если $u = bx$, где $b$ – буква, то $u' = xa$)
- Так как подслов длины $n_0$ в $\mathbf{w}$ конечное число, через определенное число шагов мы встретим подслово $v$, которое уже встречалось раньше:




Очевидно, $v^\omega$ – суффикс $\mathbf{w}$. □

## Теорема Морса–Хэдлунда (уточненная формулировка)

Для произвольного $\omega$-слова выполняется одна из двух альтернатив:

1. оно является финально периодическим и его комбинаторная сложность равна константе, начиная с некоторого $n$
2. оно не является финально периодическим и его комбинаторная сложность возрастает в каждой точке

## Теорема Морса–Хэдлунда (уточненная формулировка)

Для произвольного $\omega$-слова выполняется одна из двух альтернатив:

1. оно является финально периодическим и его комбинаторная сложность равна константе, начиная с некоторого $n$
2. оно не является финально периодическим и его комбинаторная сложность возрастает в каждой точке

- Константа в первой альтернативе появляется из следующего наблюдения: если любое подслово в $\mathbf{w}$ длины $n$ продолжается вправо единственным образом, то и любое подслово большей длины продолжается вправо единственным образом (поскольку у него есть суффикс длины $n$)
- Поскольку унарные слова являются финально периодическими, вторая альтернатива означает, в частности, $C_{\mathbf{w}}(n) \geqslant n+1$ (а все более медленно растущие функции не могут являться комбинаторной сложностью бесконечных слов)

# Уточнения и обобщения

## Теорема Морса–Хэдлунда (уточненная формулировка)

Для произвольного $\omega$-слова выполняется одна из двух альтернатив:

1. оно является финально периодическим и его комбинаторная сложность равна константе, начиная с некоторого $n$
2. оно не является финально периодическим и его комбинаторная сложность возрастает в каждой точке

- Константа в первой альтернативе появляется из следующего наблюдения: если любое подслово в $\mathbf{w}$ длины $n$ продолжается вправо единственным образом, то и любое подслово большей длины продолжается вправо единственным образом (поскольку у него есть суффикс длины $n$)
- Поскольку унарные слова являются финально периодическими, вторая альтернатива означает, в частности, $C_{\mathbf{w}}(n) \geqslant n+1$ (а все более медленно растущие функции не могут являться комбинаторной сложностью бесконечных слов)

Вопрос: существуют ли $\omega$-слова с комбинаторной сложностью $n+1$ («самые простые» непериодические $\omega$-слова)?

# Слова Фибоначчи

Бинарный морфизм $\phi : 0 \to 01, 1 \to 0$ порождает слова Фибоначчи:

$$\phi^0(0) = f_0 = 0,$$
$$\phi^1(0) = f_1 = 01,$$
$$\phi^2(0) = f_2 = 010,$$
$$\phi^3(0) = f_3 = 01001,$$
$$\phi^4(0) = f_4 = 01001010,$$
$$\dots$$
$$\phi^\infty(0) = \mathbf{f} = 01001010010010100 1010 \cdots$$

★ $\mathbf{f}$ – второе по «знаменитости» бинарное $\omega$-слово после слова Туэ-Морса; оно не 3-свободно, в нем есть подслова $f_n^3$ для всех $n \geqslant 2$

# Слова Фибоначчи

Бинарный морфизм $\phi : 0 \to 01, 1 \to 0$ порождает слова Фибоначчи:

$$\phi^0(0) = f_0 = 0,$$
$$\phi^1(0) = f_1 = 01,$$
$$\phi^2(0) = f_2 = 010,$$
$$\phi^3(0) = f_3 = 01001,$$
$$\phi^4(0) = f_4 = 01001010,$$
$$\dots$$
$$\phi^\infty(0) = \mathbf{f} = 01001010010010100 1010 \cdots$$

★ $\mathbf{f}$ – второе по «знаменитости» бинарное $\omega$-слово после слова Туэ-Морса; оно не 3-свободно, в нем есть подслова $f_n^3$ для всех $n \geqslant 2$

Простые свойства слов Фибоначчи:

- $f_n = f_{n-1} f_{n-2}$
- длина $f_n$ равна числу Фибоначчи $\Phi_n$ ($\Phi_0 = 1$, $\Phi_1 = 2$)
- две последние буквы в $f_n$ различны (по индукции из первого свойства)
- слова $f_n$ разбиваются на блоки 01 и 0; значит, в них нет подслов 11 и 000 (прообраз слова, содержащего 000, содержит 11)

## Лемма о периоде

Для всех $n \geqslant 2$, $\mathrm{per}(f_n) = \Phi_{n-1}$

## Лемма о периоде

Для всех $n \geqslant 2$, $\mathrm{per}(f_n) = \Phi_{n-1}$

Число $\Phi_{n-1}$ очевидно является периодом слова $f_n = f_{n-1}f_{n-2} = f_{n-2}f_{n-3}f_{n-2}$, поскольку префикс и суффикс $f_{n-2}$ входят в $f_n$ на расстоянии $\Phi_{n-1}$. Докажем по индукции, что у $f_n$ нет меньших периодов. База индукции ($n = 2, 3$) очевидна. Для шага индукции достаточно доказать, что у слова $f_n$ нет периодов $p$ таких, что $\Phi_{n-2} \leqslant p < \Phi_{n-1}$ (уже префикс $f_{n-1}$ слова $f_n$ не имеет периодов, меньших $\Phi_{n-2}$, по предположению индукции).

## Лемма о периоде

Для всех $n \geqslant 2$, $\mathrm{per}(f_n) = \Phi_{n-1}$

Число $\Phi_{n-1}$ очевидно является периодом слова $f_n = f_{n-1}f_{n-2} = f_{n-2}f_{n-3}f_{n-2}$, поскольку префикс и суффикс $f_{n-2}$ входят в $f_n$ на расстоянии $\Phi_{n-1}$. Докажем по индукции, что у $f_n$ нет меньших периодов. База индукции ($n = 2, 3$) очевидна. Для шага индукции достаточно доказать, что у слова $f_n$ нет периодов $p$ таких, что $\Phi_{n-2} \leqslant p < \Phi_{n-1}$ (уже префикс $f_{n-1}$ слова $f_n$ не имеет периодов, меньших $\Phi_{n-2}$, по предположению индукции).

Замечание: слова $f_nf_{n-1}$ и $f_{n-1}f_n$ отличаются только в двух последних буквах (легко доказать по индукции, поскольку $f_nf_{n-1} = f_{n-1}f_{n-2}f_{n-1}$ и $f_{n-1}f_n = f_{n-1}f_{n-1}f_{n-2}$)

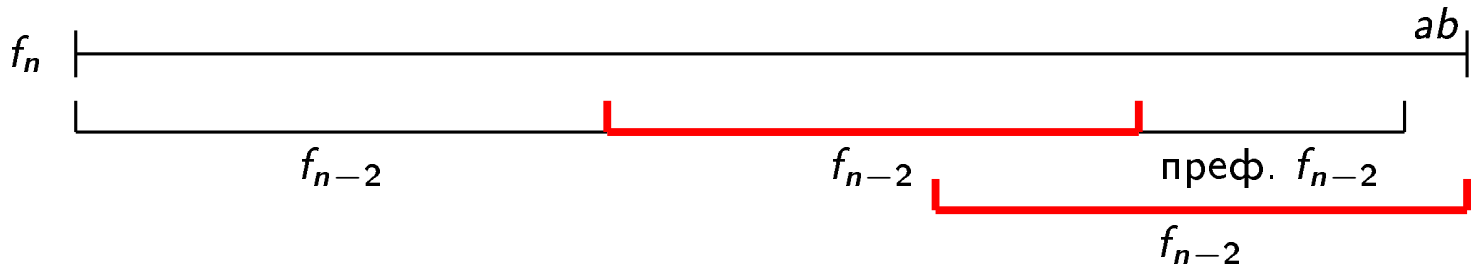
## Лемма о периоде

Для всех $n \geqslant 2$, $\mathrm{per}(f_n) = \Phi_{n-1}$

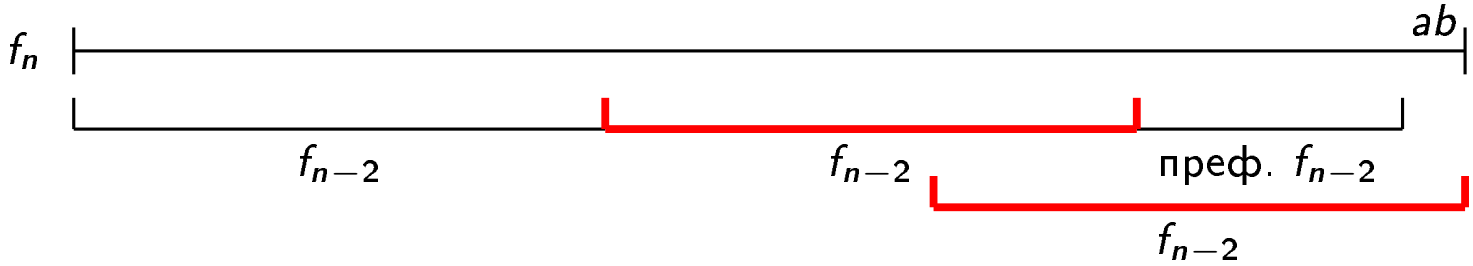
Число $\Phi_{n-1}$ очевидно является периодом слова $f_n = f_{n-1}f_{n-2} = f_{n-2}f_{n-3}f_{n-2}$, поскольку префикс и суффикс $f_{n-2}$ входят в $f_n$ на расстоянии $\Phi_{n-1}$. Докажем по индукции, что у $f_n$ нет меньших периодов. База индукции ($n = 2, 3$) очевидна. Для шага индукции достаточно доказать, что у слова $f_n$ нет периодов $p$ таких, что $\Phi_{n-2} \leqslant p < \Phi_{n-1}$ (уже префикс $f_{n-1}$ слова $f_n$ не имеет периодов, меньших $\Phi_{n-2}$, по предположению индукции).

Замечание: слова $f_nf_{n-1}$ и $f_{n-1}f_n$ отличаются только в двух последних буквах (легко доказать по индукции, поскольку $f_nf_{n-1} = f_{n-1}f_{n-2}f_{n-1}$ и $f_{n-1}f_n = f_{n-1}f_{n-1}f_{n-2}$)

Слово $f_{n-2}f_{n-2}f_{n-3}$, отличающееся от $f_n$ двумя последними буквами, имеет период $\Phi_{n-2}$; это означает, во-первых, что у $f_n$ нет такого периода, а во-вторых, что $f_n$ имеет такую структуру:

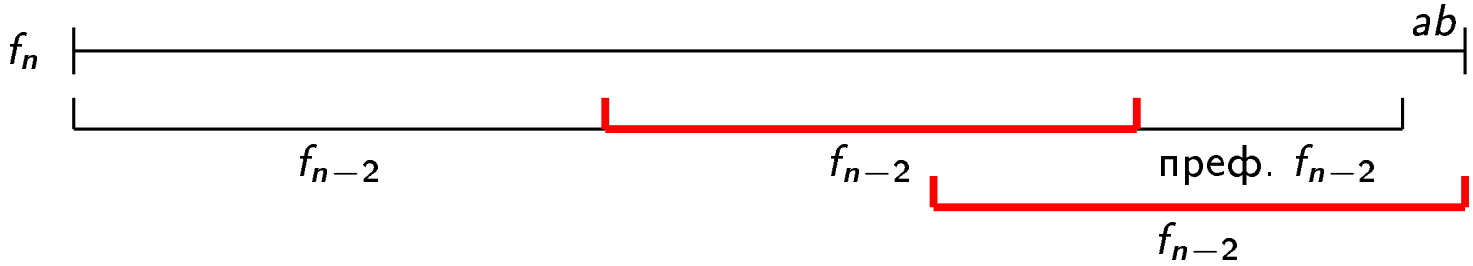


# Лемма о периоде (окончание) и лемма о продолжении




Если у $f_n$ есть период $p$, $\Phi_{n-2} < p < \Phi_{n-1}$, то слово $f_{n-2}$ встречается в $f_n$ где-то между красными вхождениями; но оно не может заканчиваться в предпоследней позиции, т.к. $a \neq b$, и не может заканчиваться раньше: в противном случае $f_{n-2}$ совпадает с одним из своих циклических сдвигов и непримитивно, что противоречит тому, что по предположению индукции его минимальный период равен $\Phi_{n-3}$ $(> \Phi_{n-2}/2)$. □

# Лемма о периоде (окончание) и лемма о продолжении




Если у $f_n$ есть период $p$, $\Phi_{n-2} < p < \Phi_{n-1}$, то слово $f_{n-2}$ встречается в $f_n$ где-то между красными вхождениями; но оно не может заканчиваться в предпоследней позиции, т.к. $a \neq b$, и не может заканчиваться раньше: в противном случае $f_{n-2}$ совпадает с одним из своих циклических сдвигов и непримитивно, что противоречит тому, что по предположению индукции его минимальный период равен $\Phi_{n-3}$ $(> \Phi_{n-2}/2)$. □

## Лемма о продолжении

Не существует слова $x$ такого, что $0x0$ и $1x1$ – подслова в $\mathbf{f}$.

Пусть $F$ – множество подслов в $\mathbf{f}$, $x$ – минимальное по длине слово такое, что $0x0, 1x1 \in F$. Тогда $x \neq \lambda$, $x \neq 0$, $x = 0y0$ для некоторого $y$. Поскольку $1x1$ – не префикс $\mathbf{f}$, $01x1 \in F$. Тогда $\phi^{-1}(01x1) = \phi^{-1}(010y01) = 0\phi^{-1}(0y)0$, откуда $0\phi^{-1}(0y)0 \in F$. Теперь рассмотрим слово $0x0 = 00y00$. Тогда $0x01 \in F$; взяв $\phi^{-1}(0x01) = \phi^{-1}(00y001) = 1\phi^{-1}(0y)10$, получаем $1\phi^{-1}(0y)1 \in F$. В результате имеем $0z0, 1z1 \in F$ для $z = \phi^{-1}(0y)$; но $|z| < |x|$, противоречие. □

# Комбинаторная сложность слова Фибоначчи

**Теорема**

$C_{\mathbf{f}}(n) = n + 1.$

# Комбинаторная сложность слова Фибоначчи

**Теорема**

$C_{\mathbf{f}}(n) = n + 1$.

Из леммы о периоде следует, что слово Фибоначчи не является финально периодическим, т.е. $C_{\mathbf{f}}(n) \geqslant n + 1$. Равенство будем доказывать по индукции. База индукции очевидна: $C_{\mathbf{f}}(1) = 2$.

**Теорема**

$C_{\mathbf{f}}(n) = n + 1$.

Из леммы о периоде следует, что слово Фибоначчи не является финально периодическим, т.е. $C_{\mathbf{f}}(n) \geqslant n + 1$. Равенство будем доказывать по индукции. База индукции очевидна: $C_{\mathbf{f}}(1) = 2$.

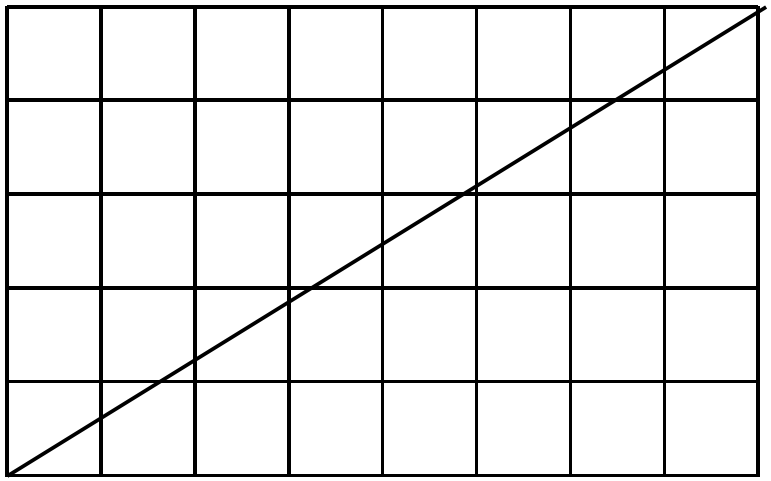
Слово $u \in F$ назовем свободным, если $u0, u1 \in F$. Так как к каждому подслову длины $n$ можно приписать справа букву, получая подслово длины $n{+}1$, легко заметить, что $C_{\mathbf{f}}(n{+}1))$ равно $C_{\mathbf{f}}(n)$ плюс количество свободных слов длины $n$ (про которое мы уже знаем, что оно больше нуля). Таким образом, предположение индукции означает, что для каждого $k = 1, \ldots, n{-}1$ имеется ровно одно свободное слово длины $k$. Рассмотрим слова длины $n$. Если слово длины $n$ свободно, то его суффикс длины $n{-}1$ также свободен. Но такой суффикс – единственный по предположению индукции, обозначим его через $x$. Если в $F$ имеется лишь одно слово длины $n$ с суффиксом $x$, то оно, очевидно, и будет единственным свободным словом длины $n$. Если же оба слова $0x$ и $1x$ принадлежат $F$, то хотя бы одно из слов $0x0, 1x1$ не принадлежит $F$ по лемме о продолжении; следовательно, лишь одно из слов $0x, 1x$ свободно. □

# Слова Штурма

Итак, слова с комбинаторной сложностью $n+1$ есть; можно доказать, что множество таких слов совпадает с весьма примечательным классом двоичных слов – словами Штурма. Слова Штурма определяются пересечениями прямой с иррациональным наклоном с линиями целочисленной решетки на плоскости:

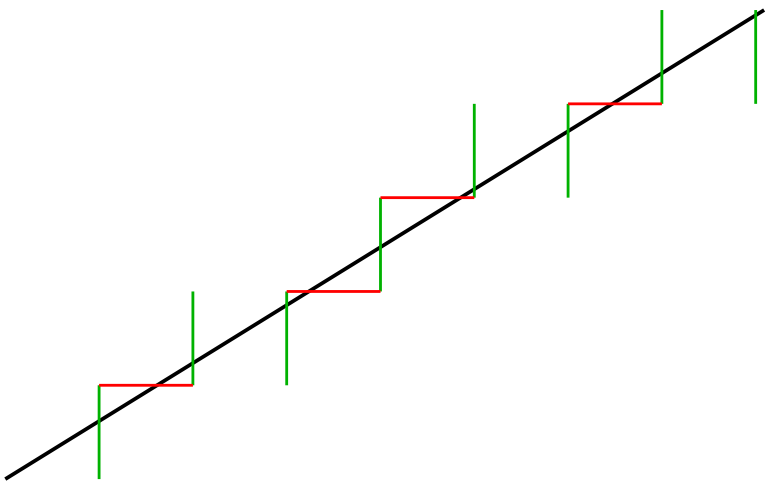
# Слова Штурма

Итак, слова с комбинаторной сложностью $n+1$ есть; можно доказать, что множество таких слов совпадает с весьма примечательным классом двоичных слов – словами Штурма. Слова Штурма определяются пересечениями прямой с иррациональным наклоном с линиями целочисленной решетки на плоскости:



# Слова Штурма

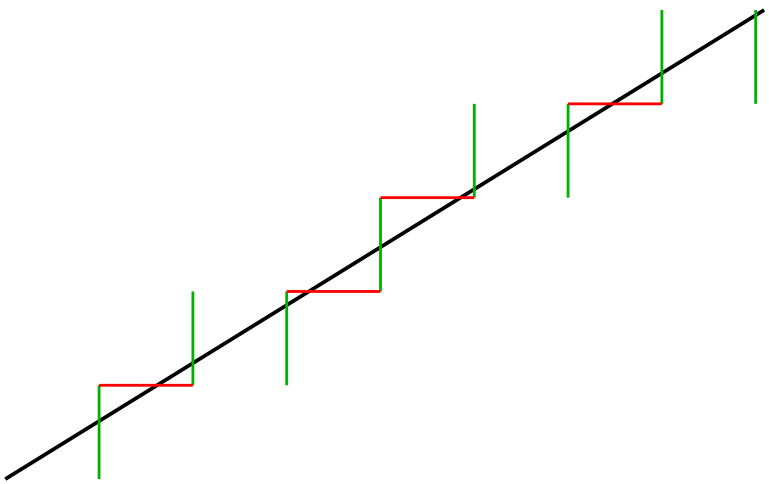
Итак, слова с комбинаторной сложностью $n+1$ есть; можно доказать, что множество таких слов совпадает с весьма примечательным классом двоичных слов – словами Штурма. Слова Штурма определяются пересечениями прямой с иррациональным наклоном с линиями целочисленной решетки на плоскости:



0 1 0 0 1 0 1 0 0 1 0 0 $\cdots$

# Слова Штурма

Итак, слова с комбинаторной сложностью $n+1$ есть; можно доказать, что множество таких слов совпадает с весьма примечательным классом двоичных слов – словами Штурма. Слова Штурма определяются пересечениями прямой с иррациональным наклоном с линиями целочисленной решетки на плоскости:



$0\,1\,0\,0\,1\,0\,1\,0\,0\,1\,0\,0\cdots$

Слова Штурма имеют много различных характеризаций и применений; например, среди слов, у которых угол наклона – квадратичная иррациональность, удалось найти следы $\left(\frac{k}{k-1}\right)^+$-свободных слов для $k = 7, \ldots, 14$, продвинув доказательство гипотезы Дежан.

# Общие замечания о комбинаторной сложности $\omega$-слов

- Легко построить двоичное слово с максимально возможной комбинаторной сложностью $2^n$: например, записать друг за другом двоичные записи всех натуральных чисел в порядке возрастания
- Упражнение: доказать, что если хотя бы одно двоичное слово не является подсловом двоичного $\omega$-слова $\mathbf{w}$, то $C_{\mathbf{w}}(n) \leqslant C \cdot \alpha^n$ для некоторого $\alpha < 2$ и некоторой константы $C$
- слова, порождаемые морфизмами, имеют очень низкую сложность; теорема Пансьё (того же самого) говорит, что сложность слова, порожденного морфизмом, принадлежит одному из классов $\Theta(1)$, $\Theta(n)$, $\Theta(n \log\log n)$, $\Theta(n \log n)$, $\Theta(n^2)$
- слово Туэ-Морса имеет линейную сложность; график сложности является фракталом, основная фигура которого – ломаная из двух звеньев с разными углами наклона

# Общие замечания о комбинаторной сложности $\omega$-слов

- Легко построить двоичное слово с максимально возможной комбинаторной сложностью $2^n$: например, записать друг за другом двоичные записи всех натуральных чисел в порядке возрастания
- Упражнение: доказать, что если хотя бы одно двоичное слово не является подсловом двоичного $\omega$-слова $\mathbf{w}$, то $C_{\mathbf{w}}(n) \leqslant C \cdot \alpha^n$ для некоторого $\alpha < 2$ и некоторой константы $C$
- слова, порождаемые морфизмами, имеют очень низкую сложность; теорема Пансьё (того же самого) говорит, что сложность слова, порожденного морфизмом, принадлежит одному из классов $\Theta(1)$, $\Theta(n)$, $\Theta(n \log\log n)$, $\Theta(n \log n)$, $\Theta(n^2)$
- слово Туэ-Морса имеет линейную сложность; график сложности является фракталом, основная фигура которого – ломаная из двух звеньев с разными углами наклона

Если не ограничиваться бесконечными словами, а перейти к более общим классам языков, ситуация со сложностью становится действительно сложной, но весьма увлекательной.