О промежуточных функциях роста сложности поиска для случайных баз данных

Н. С. Кучеренко

В работе рассматриваются функции роста сложности оптимальных алгоритмов решения задачи поиска идентичных объектов в среднем по классу задач. Ранее автором были изучены случаи, когда функция роста является ограниченной функцией или имеет порядок логарифма от мощности базы данных. В данной статье описано семейство S возможных асимптотик и семейство S^* возможных порядков функций роста, которые являются неограниченно возрастающими, но по порядку меньшими, чем логарифм от мощности базы данных.

Введение

Теория хранения и поиска информации является важным разделом теории интеллектуальных систем. Одним из ключевых объектов этой теории является информационный граф (ИГ) — управляющая система, которая позволяет рассматривать имеющиеся модели данных и задачи, связанные с ними, с общих позиций [1].

В работе рассматривается задача поиска идентичных объектов (ЗПИО) на интервале $(0, 1)$. Практически такая задача возникает, когда на объектах информационного массива введен линейный порядок. Предполагается, что объекты рассортированы в соответствии с этим порядком, и по пришедшему запросу необходимо найти соответствующий ему объект массива. Алгоритмы поиска, использующие только операции сравнения, исследуются с точки зрения сложности, которая характеризует среднее время работы алгоритма.

Идея представлять такие алгоритмы поиска с помощью деревьев возникла в пятидесятые годы двадцатого века. В 1959 году Э. Н. Гильберт и Э. Ф. Мур показали, что можно построить оптимальный, в смысле введенной сложности, алгоритм поиска за $O(n^3)$ шагов (n — мощность информационного массива или базы данных), и привели оценки сложности такого алгоритма [2]. В 1971 году Д. Э. Кнут показал, что построение этого алгоритма поиска можно улучшить до порядка $O(n^2)$ шагов [3]. Дальнейшие упрощения методов построения этого алгоритма были произведены в 1977 А. М. Гарсия и М. Л. Вочем. Их метод позволяет построить оптимальный алгоритм поиска за $O(n \cdot \log_2 n)$ шагов [4].

В данной работе алгоритмы поиска, которые используют только операции сравнения, представляются с помощью информационных графов. В зависимости от конкретной задачи поиска идентичных объектов сложность оптимального ИГ может быть как логарифмом от мощности библиотеки, так и константой. Поэтому возникает вопрос о поведении сложности оптимального информационного графа на классах задач. В работе [5] автором рассмотрен класс задач, для которых база данных — случайный вектор, компоненты которого равномерно распределены на интервале $(0, 1)$, запросы также распределены равномерно на интервале $(0, 1)$. Показано, что математическое ожидание сложности оптимального ИГ незначительно отличается от сложности стандартного бинарного поиска, которая равна двоичному логарифму от мощности базы данных.

Классы задач, у которых распределение элементов данных и запросов может быть отлично от равномерного распределения, были рассмотрены в работе [6]. Распределение элементов запросов и данных задавалось с помощью функции плотности f и g соответственно. Было показано, что если функции плотности интегрируемы по Риману, ограничены и отделены от нуля на конечно-интервальном носителе, то математическое ожидание сложности оптимального ИГ асимптотически ведет себя как $c \cdot \log_2 n$, где константа c, $0 < c \leqslant 1$, зависит только от функции f и g. Также в работе [6] был изучен случай, когда средняя сложность поиска является ограниченной функцией. Было показано что для любого отрезка $[b, b + 2], b \in \mathbb{R}, b > 1$, существуют такие функции плотности f и g, что при достаточно большом объеме

данных математическое ожидание сложности поиска не выходит за пределы этого отрезка.

Описанные выше результаты относятся к двум случаям — когда функция роста сложности поиска по классу задач является ограниченной функцией, и когда функция роста имеет порядок логарифма от мощности базы данных. В статье [6] было показано, что существует класс задач, порядок функции роста сложности поиска которого не больше, чем $\log_2 \log_2 n$, и не меньше чем $\log_2 \log_2 \log_2 n$, где n — размер базы данных.

В данной работе продолжается исследование функций роста средней сложности поиска, которые, с одной стороны, являются неограниченно возрастающими функциями, с другой стороны, имеют порядок меньше, чем логарифм от мощности базы данных. Изучаются возможная асимптотика и порядок таких функций роста.

Автором рассмотрено семейство S функций вида $r(\log_2 \log_2(n))$, где возрастающая, положительная и дифференцируемая функция $r(x)$, определенная на интервале $(x_0, +\infty)$, $x_0 \geqslant 0$, сохраняет асимптотику и имеет в качестве производной монотонную, положительную и непрерывную функцию $r'(x)$, удовлетворяющую условию:

$$\exists \alpha > 0, \alpha \in \mathbb{R} : \overline{\lim_{x \to +\infty}} \frac{r'(x)}{x^\alpha} < 1.$$

Все функции из S являются неограниченно возрастающими и имеют порядок меньше, чем $\log_2 n$ при $n \to \infty$. Для каждой функции $s(n)$ из семейства S построен класс задач, асимптотика функции роста которого такая же, как у функции $s(n)$. Также в качестве примера приводится подсемейство

$$\{c \cdot (\underbrace{\log_2 \ldots \log_2}_{i+1} n)^\alpha | i \in \mathbb{N}, \alpha \in \mathbb{R}_+, c \in \mathbb{R}_+\},$$

семейства S, где $\mathbb{R}_+$ — множество положительных вещественных чисел.

Также рассмотрен класс S^* функций вида $r(\log_2(n))$, где возрастающая положительная дифференцируемая функция $r(x)$, определенная на интервале $(x_0, +\infty)$, $x_0 \geqslant 0$, сохраняет порядок и имеет в

качестве производной монотонную, положительную и непрерывную функцию $r'(x)$, удовлетворяющую условию:

$$\exists \alpha \in \mathbb{R}, 0 < \alpha < 1 : \overline{\lim_{x \to +\infty}} \frac{r'(x)}{x^{\alpha-1}} \leqslant 1.$$

Все функции из S^* являются неограниченно возрастающими и имеют порядок меньше, чем $\log_2 n$ при $n \to \infty$. В отличие от семейства S в S^* есть функции, по порядку больше, чем любая функция из S, например $(\log_2 n)^\alpha$, где $0 < \alpha < 1$. Для каждой функции $s^*(n)$ из семейства S^* построен класс задач, порядок функции роста которого такой же как у функции $s^*(n)$. Также в качестве примера приводится множество

$$\{(\log_2 n)^\alpha | 0 < \alpha < 1, \alpha \in \mathbb{R}\},$$

которое является подсемейством для S^*.

Автор выражает благодарность своему научному руководителю профессору Э. Э. Гасанову за постановку задачи и внимание к работе и академику В. Б. Кудрявцеву за ценные советы и замечания.

1. Основные понятия и результаты

В данной работе алгоритмы поиска, использующие только операции сравнения, исследуются с позиции информационно-графовой модели данных [1]. В информационно-графовой модели структура данных задается ориентированным графом, ребра и вершины которого нагружены элементами данных и функциями. Алгоритм поиска — это «волновой» процесс на графе, который управляется нагрузочными функциями. Нагрузочные функции разделены на два класса: предикаты и переключатели. Поскольку для операций сравнения выбрано представление в виде переключателей, в работе используется информационный граф, множество нагрузочных функций которого состоит только из класса переключателей. Информационный граф такого вида называется *информационным графом с переключательной нагрузкой (ИГПН).*

Перейдем к определению задачи поиска и информационного графа с переключательной нагрузкой. Задача поиска идентичных объ-

ектов — это четверка $((0,1),\ V,\ \rho_=,\ f(x))$, где $(0,1)$ — множество запросов, $V = (y_1, y_2, \ldots, y_n)$ — конечный набор точек интервала $(0,1)$, элементы которого упорядочены по возрастанию. Множество V называется библиотекой, а элементы библиотеки — записями. Отношение $\rho_=$ — отношение равенства на множестве $(0,1) \times (0,1)$. Предполагается, что запрос $x \in (0,1)$ — случайная величина на интервале $(0,1)$ с функцией плотности $f(x)$. Отношение $\rho_=$ задает функцию ответов $J(x)$, определенную на множестве запросов $(0,1)$, следующим образом

$$J(x) = \begin{cases} \{y_i\}, & \text{если } \exists i \in [1..n] : y_i \,\rho_= x, \\ \varnothing, & \text{иначе.} \end{cases}$$

Для реализации функции ответов используются операции сравнения вещественных чисел из интервала $(0,1)$, которые заданы в виде функции $p(x,y)$

$$\forall y, x \in (0,1) \quad p(y,x) = \begin{cases} 1, & \text{если } x < y, \\ 2, & \text{если } x = y, \\ 3, & \text{если } x > y. \end{cases}$$

Функция $p_a(x) = p(a,x), a \in (0,1)$ называется переключателем. Определим базовое множество переключателей $G = \{p_a(x) | a \in (0,1)\}$. Над базовым множеством G строится информационный граф с переключательной нагрузкой, который хранит в себе схему вызова переключателей из G.

Информационный граф с переключательной нагрузкой — это связный ориентированный конечный граф без кратных ребер и петель, вершины и ребра которого размечены следующим образом: выделены два подмножества вершин P и L, так что $P \cap L = \varnothing$. Вершины множества P называются *переключательными,* а вершины множества L — *листьями.* Ребра, исходящие из переключательных вершин, называются также *переключательными.* Среди переключательных вершин выделена одна вершина v_0, которая называется *корневой.* Переключательным вершинам сопоставлены переключатели из G. Переключатель $p_a(x)$, сопоставленный переключательной вершине v, называется ее нагрузкой и обозначается $[v]_P = p_a(x)$. Ли-

стьям сопоставляются записи библиотеки. Это сопоставление обозначается $[v]_L = y_i, v \in L$; y_i называется нагрузкой листа v. Переключательным ребрам сопоставляются элементы множества $\{\Lambda, 1, 2, 3\}$, при этом из переключательной вершины не выходит двух ребер с одинаковыми номерами. Если ребру (v_{i_1}, v_{i_2}) сопоставлен номер $i \in \{1, 2, 3\}$, то номер i называется нагрузкой ребра (v_{i_1}, v_{i_2}) и обозначается $[(v_{i_1}, v_{i_1})] = i$.

Определим *функционирование* информационного графа с переключательной нагрузкой. Переключательное ребро (v_{i_1}, v_{i_2}), которому приписан номер i, *проводит запрос* $x \in X$, если переключатель, приписанный началу этого ребра, принимает значение i на запросе x, то есть, если $[(v_{i_1}, v_{i_2})] = [v_{i_1}]_P(x)$. Если $[(v_{i_1}, v_{i_2})] \neq [v_{i_1}]_P(x)$ то переключательное ребро (v_{i_1}, v_{i_2}), *не проводит запрос* $x \in X$. Полагаем, что переключательное ребро, которому был сопоставлен символ Λ, не проводит запросы. Также не проводят запросы все остальные ребра, отличные от переключательных. Ориентированная цепочка ребер проводит запрос $x \in X$, если каждое ребро цепочки проводит запрос x. Запрос $x \in X$ проходит в вершину v ИГПН, если существует ориентированная цепочка, ведущая из корневой вершины в вершину v, которая проводит запрос x. Запись y, приписанная листу v', называется *результатом функционирования информационного графа с переключательной нагрузкой на запросе* x, если запрос x проходит в лист v'. *Результат функционирования информационного графа с переключательной нагрузкой* U — это функция $R_U(x)$, которая каждому запросу x сопоставляет множество всех записей, которые являются результатом функционирования ИГПН на запросе x.

Информационный граф с переключательной нагрузкой U *решает* ЗПИО $((0,1), V, \rho_=, f(x))$, если результат функционирования ИГПН U совпадает с функцией ответов $J(x)$, то есть $R_U(x) = J(x), \forall x \in (0,1)$.

Если в информационном графе с переключательной нагрузкой U удалить «не функциональные» элементы, а именно, не переключательные ребра и ребра, которым сопоставлен символ Λ, а затем оставить только одну компоненту связности, которой принадлежит корневая вершина, то результат функционирования $R_U(x)$ не изменится. Поэтому далее рассматриваются ИГПН с точностью до этих операций удаления.

Сложностью ИГПН U на запросе x называется число

$$T(U,x)=\sum_{v\in P}\varphi_v(x),$$

где P — множество переключательных вершин, а $\varphi_v(x)$ — предикат, принимающий значение единица, если x проходит в вершину v, и нуль в противном случае. Предикат $\varphi_v(x)$ называется *функцией фильтра вершины* v. Величина $T(U,x)$ как функция от x является измеримой случайной величиной [1]. *Сложностью ИГПН* $T(U)$ называется математическое ожидание величины $T(U,x)$, которое можно записать в виде

$$T(U)=\int\limits_0^1 T(U,x)f(x)dx.$$

Множество ИГПН, которые решают ЗПИО $I=((0,1),\ V,\ \rho_=,\ f)$, обозначим через S_I. Сложностью ЗПИО I называется величина

$$T(I)=\inf_{U\in S_I}T(U).$$

Информационный граф, на котором достигается инфимум, называется *оптимальным* информационным графом с переключательной нагрузкой. В статье [5] показано, что для любой ЗПИО существует оптимальный ИГПН.

В работе рассмотрен класс задач

$$\Upsilon_n(f,g)=\{I(V)=((0,1),V,\rho_=,f):V=(y_{(1)},y_{(2)},\ldots,y_{(n)})\},$$

где $y_{(1)},y_{(2)},\ldots,y_{(n)}$ — вариационный ряд, составленный из независимых случайных величин $y_1,y_2,\ldots,y_n$, с функцией плотности $g(x)$.

Обозначим через $T_n^{(f,g)}(V)$ сложность ЗПИО $I(V)$ из класса $\Upsilon_n(f,g)$. Величина $T_n^{(f,g)}(V)$ является случайной как функция от случайной библиотеки V. Через $\mathbf{M}_V(T_n^{(f,g)}(V))$ обозначим математическое ожидание этой случайной величины. Заметим, что любую задачу поиска идентичных объектов можно решить с помощью ИГПН, который реализует бинарный поиск и имеет сложность не больше

верхней целой части числа $log_2(n+1)$. Поэтому при любых функциях плотности f, g и при любом $n > 0$ верно следующее неравенство

$$\mathbf{M}_V(T_n^{(f,g)}(V)) \leqslant] \log_2(n+1)[.$$

В работах [5], [6] были изучены случаи, когда величина $\mathbf{M}_V(T_n^{(f,g)}(V))$ как функция от n является ограниченной функцией или имеет порядок $\log_2 n$. В данной работе изучаются возможная асимптотика и порядок функций роста величины $\mathbf{M}_V(T_n^{(f,g)}(V))$ в случае, когда эта величина при $n \to \infty$ является неограниченно возрастающей функцией по порядку меньшей, чем $\log_2 n$.

Для формулировки результатов понадобятся следующие определения. Положительная, возрастающая функция $r(x)$ называется *сохраняющей асимптотику,* если выполнено условие

$$\forall c \in \mathbb{R} \quad r(x+c) \sim r(x) \quad (x \to \infty).$$

Положительная, возрастающая функция $r(x)$ называется *сохраняющей порядок,* если

$$\forall c \in \mathbb{R}, c \neq 0, \quad r(c \cdot x) \asymp r(x) \quad (x \to \infty).$$

Автором рассмотрено семейство S функций вида $r(\log_2 \log_2(n))$, где возрастающая, положительная и дифференцируемая функция $r(x)$, определенная на интервале $(x_0, +\infty), x_0 \geqslant 0$, сохраняет асимптотику и имеет в качестве производной монотонную, положительную и непрерывную функцию $r'(x)$, удовлетворяющую условию:

$$\exists \alpha > 0, \alpha \in \mathbb{R} : \overline{\lim_{x \to +\infty}} \frac{r'(x)}{x^\alpha} < 1.$$

Все функции из S являются неограниченно возрастающими и имеют порядок меньше, чем $\log_2 n$ при $n \to \infty$. Для функций семейства S верна следующая теорема

Теорема 1. *Для любой функции $s(n) = r(\log_2 \log_2(n))$ из семейства S существуют функции плотности f и g, такие что*

$$\mathbf{M}_V T_n^{(f,g)}(V)) \sim r(\log_2 \log_2(n)) \quad (n \to \infty).$$

В качестве примера показано, что множество

$$\{c \cdot (\underbrace{\log_2 \dots \log_2}_{i+1} n)^\alpha | i \in \mathbb{N}, \alpha \in \mathbb{R}_+, c \in \mathbb{R}_+\},$$

где $\mathbb{R}_+$ — множество положительных вещественных чисел, является подсемейством для S.

Следствие 1. *Для любого натурального i и для любых положительных, вещественных α и c существуют функции плотности f и g, такие что*

$$\mathbf{M}_V(T_n^{(f,g)}(V)) \sim (c \cdot \underbrace{\log_2 \dots \log_2}_{i+1} n)^\alpha \quad (n \to \infty).$$

Также рассмотрен класс S^* функций вида $r(\log_2(n))$, где возрастающая положительная дифференцируемая функция $r(x)$, определенная на интервале $(x_0, +\infty)$, $x_0 \geqslant 0$, сохраняет порядок и имеет в качестве производной монотонную, положительную и непрерывную функцию $r'(x)$, удовлетворяющую условию:

$$\exists \alpha \in \mathbb{R}, 0 < \alpha < 1 : \overline{\lim_{x \to +\infty}} \frac{r'(x)}{x^{\alpha-1}} \leqslant 1.$$

Все функции из S^* являются неограниченно возрастающими и имеют порядок меньше, чем $\log_2 n$ при $n \to \infty$. В отличие от класса S в классе S^* есть функции, по порядку больше, чем любая функция из S, например $(\log_2 n)^\alpha$, $0 < \alpha < 1$. Для функций семейства S^* верна следующая теорема

Теорема 2. *Для любой функции $s^*(n) = r(\log_2(n))$ из семейства S^* существуют функции плотности f и g, такие что*

$$\mathbf{M}_V T_n^{(f,g)}(V)) \asymp r(\log_2(n)) \quad (n \to \infty).$$

При доказательстве теоремы искомые функций плотности f и g построены. В качестве примера показано, что множество

$$\{(\log_2 n)^\alpha | 0 < \alpha < 1, \alpha \in \mathbb{R}\},$$

является подсемейством для S^*.

Следствие 2. *Для любого вещественного* α, $0 < \alpha < 1$, *существуют функции плотности* f *и* g, *такие что*

$$\mathbf{M}_V(T_n^{(f,g)}(V)) \asymp (\log_2 n)^\alpha \quad (n \to \infty).$$

2. Доказательство теоремы 1

Доказательство теоремы заключается в непосредственном построении искомых функций плотности f и g.

Рассмотрим произвольную функцию $r(\log_2 \log_2 n)$ из семейства S. По определению S функция $r(x)$ определена на интервале $(x_0, +\infty)$, $x_0 \geqslant 0$, и имеет в качестве производной монотонную, положительную и непрерывную функцию $r'(x)$, удовлетворяющую условию:

$$\exists \alpha > 0, \alpha \in \mathbb{R} : \overline{\lim_{x \to +\infty}} \frac{r'(x)}{x^\alpha} < 1.$$

Значит существует такое $x_1 \in \mathbb{R}$, $x_1 > x_0$, что выполнено

$$\forall x > x_1 \quad r'(x) < x^\alpha.$$

Обозначим $z_1 = 2^{2^{x_1}}$. Функция $r(\log_2 \log_2 z)$, определена на интервале $(z_1, +\infty)$. Рассмотрим интеграл

$$\int_{z_1}^{\infty} \frac{r'(\log_2 \log_2 z)}{z \log_2^2 z}\, dz.$$

Для подинтегрального выражения при $z > z_1$ верна следующая оценка

$$\frac{r'(\log_2 \log_2 z)}{z \log_2^2 z} < \frac{(\log_2 \log_2 z)^\alpha}{z \log_2^2 z}.$$

Также для любого $0 < \beta < 1$ существует такое $z_2 \in \mathbb{R}$, $z_2 > z_1$, что при $\forall z > z_2$ выполнено

$$\frac{(\log_2 \log_2 z)^\alpha}{z \log_2^2 z} < \frac{\log_2^\beta z}{z \log_2^2 z} = \frac{1}{z} (\log_2 z)^{\beta - 2}.$$

Получается, что для интегралов верно неравенство

$$\int\limits_{z_2}^{\infty} \frac{r'(\log_2 \log_2 z)}{z \log_2^2 z}\, dz < \int\limits_{z_2}^{\infty} (\log_2 z)^{\beta-2} z^{-1}\, dz.$$

Поскольку интеграл в правой части неравенства сходится, то сходится и интеграл в левой части.

Возьмем значение $z_3 \in \mathbb{N}$, такое, что $z_3 > \max(z_2, 2)$ и

$$(\ln 2)^{-2} \cdot \int\limits_{z_3}^{\infty} \frac{r'(\log_2 \log_2 z)}{z \log_2^2 z}\, dz < 1.$$

Определим величины a_i следующим образом. Если функция $r'(x)$ убывающая, то

$$a_i = \frac{(\ln 2)^{-2}}{2} \cdot \frac{r'(\log_2 \log_2 (z_3 + i))}{(z_3 + i) \log_2^2 (z_3 + i)},\ i = 1, 2 \ldots$$

Если функция $r'(x)$ возрастающая, то

$$a_i = \frac{(\ln 2)^{-2}}{2} \cdot \frac{r'(\log_2 \log_2 (z_3 + i - 1))}{(z_3 + i - 1) \log_2^2 (z_3 + i - 1)},\ i = 1, 2 \ldots$$

Заметим, что $\sum_{i=1}^{\infty} a_i < 1/2$, поскольку

$$\sum_{i=1}^{\infty} a_i < \frac{(\ln 2)^{-2}}{2} \cdot \int\limits_{z_3}^{\infty} \frac{r'(\log_2 \log_2 z)}{z \log_2^2 z}\, dz.$$

Положим $a_0 = 1/2 - \sum_{i=1}^{\infty} a_i$. Используем введенные величины для построения разбиения интервала $(0, 1)$.

$$\Delta_0 = (0, a_0),\ \Delta_1 = (a_0, a_0 + a_0),\ \ldots,$$
$$\Delta_{2k-1} = (2\sum_{i=0}^{k-2} a_i + a_{k-1}, 2\sum_{i=0}^{k-1} a_i),\ \Delta_{2k} = (2\sum_{i=0}^{k-1} a_i, 2\sum_{i=0}^{k-1} a_i + a_k),\ \ldots$$

Функция плотности f определяется так, что вероятность нечетных интервалов равняется нулю, а вероятность четного интервала Δ_{2k}, $k = 0, 1, \ldots$, включая концевые точки, равняется величине $(2 \cdot a_k)$. Например, функцию плотности f определим так, что на нечетных интервалах она равна нулю, а на четных представляет из себя равнобедренный треугольник высотой 4. Площадь равнобедренного треугольника на интервале Δ_{2k}, $k = 0, 1, \ldots$, обозначим через p_k, заметим, что $p_k = 2a_k$. Если в точках 0 и 1 функцию f доопределить нулем, то она будет интегрируема по Риману на отрезке $[0, 1]$, и интеграл равен единице.

Функция плотности g определяется так, что вероятность четных интервалов равняется нулю, а вероятность нечетного интервала Δ_{2k+1}, $k = 0, 1, \ldots$, включая концевые точки, равняется величине $\frac{1}{(k+1)(k+2)}$. Например, определим функцию плотности g, равной нулю на четных интервалах, а на нечетных интервалах Δ_{2k+1}, $k = 0, 1, \ldots$, она имеет вид равнобедренного треугольника высотой h_k равной

$$h_k = \frac{4a \log_2^2(k+2)}{k+1}.$$

Площадь этого треугольника обозначим через q_k.

$$q_k = \frac{1}{(k+1)(k+2)}.$$

Ряд $\sum_{i=0}^{\infty} q_i$ сходится, и сумма его равна единице. Получившаяся функция плотности g также интегрируема по Риману на отрезке $[0, 1]$, если в концах отрезка ее доопределить нулем.

Построение функций плотности f и g закончено. Покажем что они удовлетворяют условиям теоремы. Для этого нам понадобится лемма о свойствах распределения записей библиотеки, когда вероятность попасть в интервал Δ_{2k+1}, $k = 0, 1, \ldots$, равна $\frac{1}{(k+1)(k+2)}$, а вероятность попадания в остальные интервалы равна нулю.

Лемма 1. *Пусть событие A_n состоит в том, что ни в одном из интервалов Δ_{2k+1}, где $k > n^2 - 2$, нет ни одной из n записей библиотеки. А событие B_n, заключается в том, что в каждом из интервалов $\Delta_{2i+1}, i = 0, 1, \ldots, [n^{1/3}]$, есть хотя бы одна из n записей библиотеки. Тогда*

$$\mathbf{P}\,(A_n) = 1 - o\left(\frac{1}{n}\right) \quad (n \to \infty),$$
$$\mathbf{P}\,(B_n) = 1 - o(1) \quad (n \to \infty).$$

Доказательство. Рассмотрим событие A_n. Вероятность того, что запись библиотеки не попадет в интервалы Δ_{2k+1}, где $k > n^2 - 2$, равна

$$\sum_{k=0}^{n^2-2} \frac{1}{(k+1)(k+2)} = 1 - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} \cdots - \frac{1}{n^2} = 1 - \frac{1}{n^2}$$

Поскольку записи библиотеки независимы, мощность множества A_n равна

$$\mathbf{P}\,(A_n) = \left(1 - \frac{1}{n^2}\right)^n = \left(\left(1 - \frac{1}{n^2}\right)^{n^2}\right)^{\frac{1}{n}} \to 1 \quad (n \to \infty)$$

Оценим скорость сходимости к единице

$$1 - \mathbf{P}\,(A_n) = 1 - \left(\left(1 - \frac{1}{n^2}\right)^{n^2}\right)^{\frac{1}{n}} \sim 1 - e^{-1/n} = o\left(\frac{1}{n}\right).$$

Первое равенство леммы доказано.

Перейдем к событию B_n, которое состоит в том, что в каждом из интервалов Δ_{2i+1}, $i = 0, 1, \ldots, [n^{1/3}]$, есть хотя бы одна запись библиотеки. Рассмотрим событие C_n, которое является отрицанием события B_n

$$C_n = \{\text{пуст хотя бы один из отрезков } \Delta_{2i+1}, i = 0, 1, \ldots, [n^{1/3}]\}.$$

Вероятности этих события B_n можно представить так: $\mathbf{P}\,(B_n) = 1 - \mathbf{P}\,(C_n)$. Оценим вероятность события C_n

$$\mathbf{P}\,(C_n) < \sum_{s=1}^{[n^{1/3}]} \mathbf{P}\,(\{\text{пуст отрезок с номером } s\}).$$

Вероятность того, что пуст отрезок с номером s равна

$$\mathbf{P}\left(\{\text{пуст отрезок с номером } s\}\right) = (1-q_s)^n, \quad q_s = \frac{1}{(s+1)(s+2)}.$$

Величина q_s принимает минимальное значение при $s = [n^{1/3}]$, поэтому ее можно оценить следующим образом

$$\forall s = 1, \ldots, [n^{1/3}] \quad q_s > \frac{1}{([n^{1/3}]+2)^2}.$$

Используя эту оценку, получаем, что при любом s от 1 до $[n^{1/3}]$ верно неравенство

$$\mathbf{P}\left(\{\text{пуст отрезок с номером } s\}\right) < \left(1 - \frac{1}{([n^{1/3}]+2)^2}\right)^n.$$

Вероятность множества B_n оценивается так

$$\mathbf{P}\left(B_n\right) > 1 - [n^{1/3}] \cdot \left(1 - \frac{1}{([n^{1/3}]+2)^2}\right)^n.$$

Подсчитаем предел вычитаемого выражения в правой части

$$\lim_{n\to\infty} \left(([n^{1/3}]) \cdot \left(\left(1 - \frac{1}{[n^{2/3}]+2}\right)^{[n^{2/3}]} \right)^{[n^{1/3}]} \right) =$$

$$= \lim_{n\to\infty} \left(n^{1/3} \cdot e^{-n^{1/3}} \right) = 0.$$

Получаем, что $\mathbf{P}\left(B_n\right) > 1 - o(1)$ при $n \to \infty$.

Доказательство леммы закончено.

Вернемся к доказательству теоремы. Для ЗПИО $I(V) = ((0,1), V, \rho_=, f)$, $V = (y_{(1)}, y_{(2)}, \ldots, y_{(n)})$ из класса $\Upsilon_n(f,g)$ определим энтропию $H_n^{(f,g)}(V)$ следующим образом. Введем вероятности $P_i, i = 0, \ldots, n$ того, что запрос x попадет в интервал $(y_{(i)}, y_{(i+1)})$,

$$P_i = \mathbf{P}\left(x \in (y_{(i)}, y_{(i+1)})\right),$$

полагаем $y_{(0)} = 0, y_{(n+1)} = 1$. Энтропия $H_n^{(f,g)}(V)$ есть функция от этих вероятностей

$$H_n^{(f,g)}(V) = H(P_0, P_1, \ldots, P_n) = \sum_{i=0}^{n} -P_i \cdot \log_2 P_i,$$

если P_i равно нулю или единице, полагаем выражение $(P_i \cdot \log_2 P_i)$ равным нулю.

Покажем, что $\mathbf{M}_V T_n^{(f,g)}(V))$ асимптотически не больше, чем $r(\log_2 \log_2 n)$. Математическое ожидание сложности рассматриваемого класса задач можно оценить с помощью энтропии [7]

$$\mathbf{M}_V T_n^{(f,g)}(V)) < \mathbf{M}_V(H_n^{(f,g)}(V)) + 2.$$

Рассмотрим подкласс $\Upsilon(A_n)$ класса задач $\Upsilon_n(f,g)$, для библиотек которых выполнено следующее условие: ни в одном из интервалов с номером $2k+1$, где $k > n^2 - 2$, нет записей библиотеки. Поскольку записи библиотеки и запросы как случайные величины независимы, класс $\Upsilon(A_n)$ определяется событием A_n из леммы выше. Используя условные математические ожидания, математическое ожидание энтропии можно записать следующим образом

$$\begin{aligned}\mathbf{M}_V(H_n^{(f,g)}(V)) = \mathbf{M}_V(H_n^{(f,g)}(V)|A_n) \cdot \mathbf{P}(A_n) + \\ + \mathbf{M}_V(H_n^{(f,g)}(V)|\overline{A_n}) \cdot \mathbf{P}(\overline{A_n}).\end{aligned}$$

Оценим второе слагаемое в правой части равенства. Положительная величина $\mathbf{M}_V(H_n^{(f,g)}(V)|\overline{A_n})$ при любом $n \in \mathbb{N}$ ограничена числом $]\log_2(n+1)[$. В силу леммы вероятность $\mathbf{P}(\overline{A_n}) = o(1/n)$ при $n \to \infty$. Получаем, что второе слагаемое

$$\mathbf{M}_V(H_n^{(f,g)}(V)|\overline{A_n}) \cdot \mathbf{P}(\overline{A_n}) = o(1) \quad (n \to \infty).$$

Учитывая, что $\lim_{n\to\infty} \mathbf{P}(A_n) = 1$, получаем

$$\mathbf{M}_V(H_n^{(f,g)}(V)) \sim \mathbf{M}_V(H_n^{(f,g)}(V)|A_n) \quad (n \to \infty).$$

Все точки библиотеки задачи из класса A_n, с вероятностью единица находятся вне интервалов Δ_i с четными номерами. А на интервалах с нечетными номерами функция f равна нулю. Поэтому

вероятности P_i для задачи поиска из класса A_n составляются следующим образом. Вероятность P_n для всех задач из этого класса равна $1 - \sum_{i=0}^{n^2-1} p_i$. А вероятности P_i, $i = 0, 1, \ldots, (n-1)$, имеют вид

$$P_i = \sum_{k=j_i}^{j_{i+1}} p_k,\ i = 0, 1, \ldots, n-1,\ j_0 = 0,\ j_n = n^2 - 1,$$

где индексы $0 \leqslant j_1 \leqslant j_2 \leqslant \ldots \leqslant j_n \leqslant (n^2 - 1)$ зависят от библиотеки V.

Для функции энтропии верно следующее свойство [7]. При фиксированных значениях $t_1, t_2, \ldots, t_{n-k}$ выполнено

$$H(t_1, t_2, \ldots, t_{n-k}, t_{n-k+1}, \ldots, t_n) \geqslant H(t_1, t_2, \ldots, t_{n-k}, q),$$

где $q = t_{n-k+1} + \cdots + t_n$. Поэтому энтропия любой ЗПИО из класса A_n, не превосходит значения

$$H(p_0, p_1, \ldots, p_{n^2-1}, 1 - \sum_{i=0}^{n^2-1} p_i).$$

Для условного математического ожидания энтропии получаем следующую оценку

$$\mathbf{M}_V(H_n^{(f,g)}(V)|A_n) \leqslant H\left(p_0, p_1, \ldots, p_{n^2-1}, 1 - \sum_{i=0}^{n^2-1} p_i\right).$$

Последнее слагаемое в функции энтропии оценим единицей. Получим следующее неравенство

$$H(p_0, p_1, \ldots, p_{n^2-1}, 1 - \sum_{i=0}^{n^2-1} p_i) < -\sum_{i=0}^{n^2-1} p_i \log_2 p_i + 1.$$

Оценим выражение $H_{n^2-1} = \sum_{i=1}^{n^2-1} p_i \log_2 p_i$ при $n \to \infty$. Если функция $r'(x)$ возрастающая, то по построению

$$p_i = (\ln 2)^{-2} \cdot \frac{r'(\log_2 \log_2(z_1 + i - 1))}{(z_1 + i - 1)\log_2^2(z_1 + i - 1)},\ i = 1, 2, \ldots, n^2 - 1.$$

Подставляя эти значения и раскрывая скобки, получим

$$H_{n^2-1} = (\ln 2)^{-2} \sum_{i=z_3}^{z_3+n^2-2} \frac{r'(\log_2 \log_2 i)}{i \log_2 i} + \\ + (\ln 2)^{-2} \sum_{i=z_3}^{z_3+n^2-2} \frac{r'(\log_2 \log_2 i) \cdot \log_2 \log_2 i}{i \log_2^2 i} - \\ - (\ln 2)^{-2} \sum_{i=z_3}^{z_3+n^2-2} \frac{r'(\log_2 \log_2 i) \cdot (\log_2 [r'(\log_2 \log_2 i)])}{i \log_2^2 i}.$$

Если функция $r'(x)$ убывающая, то подставляя

$$p_i = (\ln 2)^{-2} \frac{r'(\log_2 \log_2 (z_3 + i))}{(z_3 + i) \log_2^2 (z_3 + i)}, \ i = 1, 2, \ldots, n^2 - 1,$$

получаем

$$H_{n^2-1} = (\ln 2)^{-2} \sum_{i=z_3+1}^{z_3+n^2-1} \frac{r'(\log_2 \log_2 i)}{i \log_2 i} + \\ + (\ln 2)^{-2} \sum_{i=z_3+1}^{z_3+n^2-1} \frac{r'(\log_2 \log_2 i) \cdot \log_2 \log_2 i}{i \log_2^2 i} - \\ - (\ln 2)^{-2} \sum_{i=z_3+1}^{z_3+n^2-1} \frac{r'(\log_2 \log_2 i) \cdot (\log_2 [r'(\log_2 \log_2 i)])}{i \log_2^2 i}.$$

Как видно, выражение H_{n^2-1} для возрастающей и убывающей функции различается лишь пределами суммирования. Поскольку нам нужна верхняя оценка, отбросим последний ряд и получим оценку

$$H_{n^2-1} \leqslant (\ln 2)^{-2} \sum_{i=z_3+c}^{z_3+n^2-2+c} \frac{r'(\log_2 \log_2 i)}{i \log_2 i} \\ + (\ln 2)^{-2} \sum_{i=z_3+c}^{z_3+n^2-2+c} \frac{r'(\log_2 \log_2 i) \cdot \log_2 \log_2 i}{i \log_2^2 i},$$

где $c = 0$, если $r'(x)$ возрастает, и $c = 1$, если $r'(x)$ убывает.

В получившейся оценке для H_{n^2-1} ряд вида

$$\sum_{i=z_3+c}^{z_3+n^2-2+c} \frac{r'(\log_2 \log_2 i) \cdot \log_2 \log_2 i}{i \log_2^2 i}$$

ограничен константой, поскольку мажорируется сходящимся рядом

$$\sum_{i=z_3+c}^{z_3+n^2-2+c} \frac{r'(\log_2 \log_2 i) \cdot \log_2 \log_2 i}{i \log_2^2 i} < \sum_{i=z_3}^{\infty} \frac{(\log_2 \log_2 i)^{\alpha} \cdot \log_2 \log_2 i}{i \log_2^2 i}.$$

В выражении H_{n^2-1} для суммы вида

$$(\ln 2)^{-2} \sum_{i=z_3+c}^{z_3+n^2-2+c} \frac{r'(\log_2 \log_2 i)}{i \log_2 i}, c \in \{0, 1\},$$

верны следующие оценки. Если функция $r'(x)$ возрастающая, то

$$(\ln 2)^{-2} \sum_{i=z_3}^{z_3+n^2-2} \frac{r'(\log_2 \log_2 i)}{i \log_2 i} \leqslant (\ln 2)^{-2} \int_{z_3}^{z_3+n^2-2} \frac{r'(\log_2 \log_2 x)}{x \log_2 x}\, dx.$$

Если функция $r'(x)$ убывающая, то

$$(\ln 2)^{-2} \sum_{i=z_3+1}^{z_3+n^2-1} \frac{r'(\log_2 \log_2 i)}{i \log_2 i} \leqslant (\ln 2)^{-2} \int_{z_3+1}^{z_3+n^2-1} \frac{r'(\log_2 \log_2 x)}{x \log_2 x}\, dx + \frac{r'(\log_2 \log_2(z_3+1))}{(z_3+1) \log_2(z_3+1)}.$$

Обозначим $a = z_3 + c$, $b = z_3 + n^2 - 2 + c$. Тогда интеграл в правой части неравенства равен

$$(\ln 2)^{-2} \int_a^b \frac{r'(\log_2 \log_2 x)}{x \log_2 x}\, dx = \int_a^b r'(\log_2 \log_2 x)\, d(\log_2 \log_2 x) = r(\log_2 \log_2(b)) - r(\log_2 \log_2(a)).$$

Заметим, что выражение $r(\log_2 \log_2(a))$ — константа. Собирая вместе полученные оценки, получаем что $\exists C > 0$, такое, что $\forall n$ выполнено

$$H_{n^2-1} < r(\log_2 \log_2(z_3 + n^2 - 2 + c)) + C.$$

Покажем, что

$$r(\log_2 \log_2(z_3 + n^2 - 2 + c)) \sim r(\log_2 \log_2 n) \quad (n \to \infty),$$

где $c = 0, 1$. Представим внутренний логарифм в выражении $r(\log_2 \log_2(z_3 + n^2 - 2 + c))$ в виде суммы

$$\log_2(z_3 + n^2 - 2 + c) = 2 \log_2 n + \log_2 \left(1 + \frac{z_3 - 2 + c}{n^2}\right).$$

Обозначим второе слагаемое в правой части через $sl(n)$. Заметим, что

$$sl(n) = \log_2 \left(1 + \frac{z_3 - 2 + c}{n^2}\right) = o(1) \quad (n \to \infty).$$

Подставим выражение для внутреннего логарифма во внешний логарифм, и преобразуем его

$$\log_2 (2 \log_2 n + sl(n)) = \log_2 \log_2 n + \log_2 \left(2 + \frac{sl(n)}{\log_2 n}\right).$$

Начиная с некоторого n_0 второе слагаемое в правой части меньше 2. Поэтому при $n > n_0$ верно неравенство

$$r(\log_2 \log_2(z_3 + n^2 - 2 + c)) < r(\log_2 \log_2 n + 2).$$

По условию теоремы

$$r(\log_2 \log_2 n + 2) \sim r(\log_2 \log_2 n) \quad (n \to \infty).$$

Мы показали, что верхняя оценка для математического ожидания сложности ЗПИО асимптотически не больше, чем $r(\log_2 \log_2 n)$ при $(n \to \infty)$.

Покажем, что нижняя оценка асимптотически не меньше чем $r(\log_2 \log_2 n)$. Нижнюю оценку математического ожидания сложности ЗПИО по рассматриваемому классу задач также можно оценить с помощью энтропии [5]

$$\mathbf{M}_V \left(T_n^{(f,g)}(V) \right) \geqslant \mathbf{M}_V \left(H^{(f,g)}(V) \right).$$

В классе ЗПИО $\Upsilon_n(f,g)$ рассмотрим подкласс задач $\Upsilon(B_n)$, у которых в каждом из отрезков $\Delta_{2i+1}, i = 0, 1, \ldots, [n^{1/3}]$ есть хотя бы одна запись библиотеки. Поскольку записи библиотеки и запросы как случайные величины независимы, класс $\Upsilon(B_n)$ определяется событием B_n из леммы 1. Используя условные математические ожидания, математическое ожидание энтропии можно записать следующим образом

$$\begin{aligned}\mathbf{M}_V(H_n^{(f,g)}(V)) = \\ = \mathbf{M}_V(H_n^{(f,g)}(V)|B_n) \cdot \mathbf{P}\,(B_n) + \mathbf{M}_V(H_n^{(f,g)}(V)|\overline{B_n}) \cdot \mathbf{P}\,(\overline{B_n}).\end{aligned}$$

Поскольку по определению функция энтропии неотрицательна, получаем неравенство

$$\mathbf{M}_V(H_n^{(f,g)}(V)) \geqslant \mathbf{M}_V(H_n^{(f,g)}(V)|B_n) \cdot \mathbf{P}\,(B_n).$$

Энтропия для задачи из класса $\Upsilon(B_n)$ всегда содержит слагаемые $(-p_i \log_2 p_i)$, где $i = 1, 2, \ldots, [n^{2/3}]$, поэтому

$$\mathbf{M}_V(H_n^{(f,g)}(V)|B_n) \geqslant -\sum_{i=1}^{[n^{2/3}]} p_i \log_2 p_i.$$

Обозначим выражение $(-\sum_{i=1}^{[n^{2/3}]} p_i \log_2 p_i)$ через $H_{[n^{2/3}]}$.

Подставим в выражение $H_{[n^{2/3}]}$ значение

$$p_i = (\ln 2)^{-2} \frac{r'(\log_2 \log_2 (z_3 + i - c))}{(z_3 + i - c) \log_2^2 (z_3 + i - c)},\ i = 1, 2, \ldots, [n^{2/3}],$$

где $c = 0$, если функция $r'(x)$ убывает, и $c = 1$, если функция $r'(x)$ возрастает. Получаем

$$H_{[n^{2/3}]} = (\ln 2)^{-2} \sum_{i=z_3+1-c}^{z_3+[n^{2/3}]-c} \frac{r'(\log_2 \log_2 i)}{i \log_2 i}$$
$$+ (\ln 2)^{-2} \sum_{i=z_3+1-c}^{z_3+[n^{2/3}]-c} \frac{r'(\log_2 \log_2 i) \cdot \log_2 \log_2 i}{i \log_2^2 i}$$
$$- (\ln 2)^{-2} \sum_{i=z_3+1-c}^{z_3+[n^{2/3}]-c} \frac{r'(\log_2 \log_2 i) \cdot (\log_2 [r'(\log_2 \log_2 i)])}{i \log_2^2 i}.$$

Поскольку нам нужна только нижняя оценка, можно отбросить второе слагаемое

$$H_{[n^{2/3}]} \geqslant (\ln 2)^{-2} \sum_{i=z_3+1-c}^{z_3+[n^{2/3}]-c} \frac{r'(\log_2 \log_2 i)}{i \log_2 i}$$
$$- (\ln 2)^{-2} \sum_{i=z_3+1-c}^{z_3+[n^{2/3}]-c} \frac{r'(\log_2 \log_2 i) \cdot (\log_2 [r'(\log_2 \log_2 i)])}{i \log_2^2 i}.$$

В получившейся оценке для $H_{[n^{2/3}]}$ ряд вида

$$(\ln 2)^{-2} \sum_{i=m}^{m+l} \frac{r'(\log_2 \log_2 i) \cdot (\log_2 [r'(\log_2 \log_2 i)])}{i \log_2^2 i}$$

с положительными членами ограничен константой, поскольку мажорируется сходящимся рядом

$$\sum_{i=m}^{m+l} \frac{r'(\log_2 \log_2 i) \cdot (\log_2 [r'(\log_2 \log_2 i)])}{i \log_2^2 i} <$$
$$< \sum_{i=m}^{+\infty} \frac{(\log_2 \log_2 i)^{\alpha} \cdot (\alpha \log_2 \log_2 \log_2 i)}{i \log_2^2 i}.$$

Для ряда вида

$$(\ln 2)^{-2} \sum_{i=z_3+1-c}^{z_3+[n^{2/3}]-c} \frac{r'(\log_2 \log_2 i)}{i \log_2 i}, c \in \{0, 1\},$$

верны следующие оценки. Если функция $r'(x)$ возрастающая, то

$$(\ln 2)^{-2} \sum_{i=z_3}^{z_3+[n^{2/3}]-1} \frac{r'(\log_2 \log_2 i)}{i \log_2 i} \geqslant (\ln 2)^{-2} \int_{z_3}^{z_3+[n^{2/3}]-2} \frac{r'(\log_2 \log_2 x)}{x \log_2 x}\, dx.$$

Если функция $r'(x)$ убывающая, то

$$(\ln 2)^{-2} \sum_{i=z_3+1}^{z_3+[n^{2/3}]} \frac{r'(\log_2 \log_2 i)}{i \log_2 i} \geqslant (\ln 2)^{-2} \int_{z_3+1}^{z_3+[n^{2/3}]} \frac{r'(\log_2 \log_2 x)}{x \log_2 x}\, dx.$$

Обозначим $a = z_3+1-c$, $b = z_3+[n^{2/3}]-2c$. Верна следующая цепочка равенств

$$(\ln 2)^{-2} \int_a^b \frac{r'(\log_2 \log_2 x)}{x \log_2 x}\, dx = \int_a^b r'(\log_2 \log_2 x)\, d(\log_2 \log_2 x) = \\ = r(\log_2 \log_2(b)) - r(\log_2 \log_2(a)).$$

Заметим, что выражение $r(\log_2 \log_2(a))$ — константа. Собирая вместе полученные оценки, получаем что $\exists C > 0$, такое, что $\forall n$ выполнено

$$H_{[n^{2/3}]} < r(\log_2 \log_2(z_3 + [n^{2/3}] - 2c)) - C.$$

Покажем, что

$$r(\log_2 \log_2(z_3 + [n^{2/3}] - 2c)) \sim r(\log_2 \log_2 n) \quad (n \to \infty),$$

где $c \in \{0, 1\}$. Оценим внутренний логарифм в исследуемом выражении

$$\log_2(z_3 + [n^{2/3}] - 2c) \geqslant \frac{2}{3} \cdot \log_2 n + \log_2 \left(1 + \frac{z_3 - 2c - 1}{n^{2/3}}\right).$$

Обозначим второе слагаемое в правой части через $vs(n)$. Заметим, что

$$vs(n) = \log_2 \left(1 + \frac{z_3 - 2c - 1}{n^{2/3}}\right) = o(1) \quad (n \to \infty).$$

Подставим выражение для внутреннего логарифма во внешний логарифм, и преобразуем его

$$\log_2\left(\frac{2}{3}\cdot\log_2 n + vs(n)\right) = \log_2\log_2 n + \log_2\left(\frac{2}{3} + \frac{vs(n)}{\log_2 n}\right).$$

Начиная с некоторого n_0 второе слагаемое в правой части больше $\log_2(1/2) = -1$. Поэтому при $n > n_0$ верно неравенство

$$r(\log_2\log_2(z_3 + [n^{2/3}] - 2c)) > r(\log_2\log_2 n - 1).$$

По условию теоремы

$$r(\log_2\log_2 n - 1) \sim r(\log_2\log_2 n) \quad (n \to \infty).$$

Мы показали, что нижняя оценка для математического ожидания сложности ЗПИО асимптотически не меньше, чем $r(\log_2\log_2 n)$ при $(n \to \infty)$.

Доказательство теоремы закончено.

3. Доказательство следствия 1

Покажем, что любую из функций семейства

$$\{c\cdot(\underbrace{\log_2\ldots\log_2 n}_{i})^{\alpha} |\ i \geqslant 2, i \in \mathbb{N},\ \alpha > 0, \alpha \in \mathbb{R}\},$$

можно представить в виде $r(\log_2\log_2 n)$, где $r(x)$ возрастающая, положительная и дифференцируемая функция, определенная на некотором интервале $(x_0, +\infty)$, $x_0 \geqslant 0$. Функция $r(x)$ сохраняет асимптотику и имеет в качестве производной монотонную, положительную и непрерывную функцию $r'(x)$, удовлетворяющую условию:

$$\exists\beta > 0, \alpha \in \mathbb{R} : \overline{\lim_{x\to+\infty}}\ \frac{r'(x)}{x^{\beta}} < 1.$$

В случае, когда $i = 2$, в качестве $r(x)$ возьмем функцию $r(x) = c\cdot x^{\alpha}$, $\alpha > 0$. Эта функция положительная, дифференцируемой и

монотонно возрастает на интервале $(0, +\infty)$. Также выполнено для любого вещественного b выполнено

$$c \cdot (x+b)^{\alpha} \sim c \cdot (x)^{\alpha} \quad (x \to +\infty).$$

Производная $r'(x) = c \cdot x^{\alpha-1}$ является монотонной, положительной и непрерывной функцией на $(0, +\infty)$. Положим $\beta = \alpha$ и проверим условие

$$\overline{\lim_{x \to +\infty}} \frac{c \cdot x^{\alpha-1}}{x^{\alpha}} = \overline{\lim_{x \to +\infty}} \frac{c}{x} = 0.$$

В случае, когда $i > 2$, в качестве $r(x)$ возьмем функцию

$$r(x) = c \cdot (\underbrace{\log_2 \dots \log_2}_{i-2} x)^{\alpha},$$

определенную на интервале $(x_0, +\infty)$, где

$$x_0 = \underbrace{2^{2^{2^{\cdot^{\cdot^{\cdot^{2}}}}}}}_{i-3}.$$

Эта функция положительная, дифференцируемой и монотонно возрастает на интервале $(x_0, +\infty)$. Покажем, что для любого вещественного b верно

$$c \cdot (\underbrace{\log_2 \dots \log_2}_{i-2}(x+b))^{\alpha} \sim c \cdot (\underbrace{\log_2 \dots \log_2}_{i-2} x)^{\alpha} \quad (x \to +\infty).$$

Выражение $\log_2(x+b)$ при $x \to +\infty$ представим в виде

$$\log_2(x+b) = \log_2 \left(x \cdot \left(1 + \frac{b}{x} \right) \right) = \log_2 x + o(1) \quad (x \to \infty).$$

Представив таким образом все вложенные логарифмы, получим

$$\underbrace{\log_2 \dots \log_2}_{i-2}(x+b) = \underbrace{\log_2 \dots \log_2}_{i-2} x + o(1) \quad (x \to \infty).$$

Из этой оценки следует, что функция

$$r(x) = c \cdot (\underbrace{\log_2 \dots \log_2}_{i-2} x)^{\alpha}$$

сохраняет асимптотику.

Производная функции $r(x)$ на интервале $(x_0, +\infty)$ равна

$$r'(x) = c \cdot (\underbrace{\log_2 \dots \log_2}_{i-2} x)^{\alpha-1} \cdot \left(x \cdot \log_2 x \cdot \dots \cdot \underbrace{\log_2 \dots \log_2}_{i-3} x \right)^{-1}.$$

Это непрерывная, монотонно убывающая и положительная функция. Положим $\beta = \alpha$. Предел отношения производной и функции x^{α} равен нулю. Выбранная функция $r(x)$ удовлетворяет условиям из определения семейства S.

Доказательство следствия закончено.

4. Доказательство теоремы 2

Также как и доказательство теоремы 1 доказательство этой теоремы заключается в непосредственном построении искомых функций плотности f и g.

Рассмотрим произвольную функцию $r(\log_2 n)$ из семейства S^*. По определению S^* функция $r(x)$ определена на интервале $(x_0, +\infty)$, $x_0 \geqslant 0$, и имеет в качестве производной монотонную, положительную и непрерывную функцию $r'(x)$, удовлетворяющую условию:

$$\exists \alpha \in \mathbb{R}, 0 < \alpha < 1 : \overline{\lim_{x \to +\infty}} \frac{r'(x)}{x^{\alpha-1}} \leqslant 1.$$

Значит существует такое $x_1 \in \mathbb{R}$, $x_1 > x_0$, что выполнено

$$\forall x > x_1 \quad r'(x) \leqslant x^{\alpha-1}.$$

Обозначим $z_1 = 2^{x_1}$. Рассмотрим интеграл

$$\int_{z_1}^{\infty} \frac{r'(\log_2 z)}{z \log_2 z}\, dz.$$

Внесем $(1/z)$ под знак дифференциала и сделаем замену переменной

$$\ln 2 \cdot \int\limits_{z_1}^{\infty} \frac{r'(\log_2 z)}{\log_2 z}\, d(\log_2 z) = \ln 2 \cdot \int\limits_{x_1}^{\infty} \frac{r'(x)}{x}\, dx.$$

Для последнего интеграла верны следующие оценки

$$\int\limits_{x_1}^{\infty} \frac{r'(x)}{x}\, dx \leqslant \int\limits_{x_1}^{\infty} \frac{(x)^{\alpha-1}}{x}\, dx = \int\limits_{x_1}^{\infty} x^{\alpha-2}\, dx < \infty,$$

поэтому сходится интеграл

$$\int\limits_{z_1}^{\infty} \frac{r'(\log_2 z)}{z \log_2 z}\, dz.$$

Возьмем значение $z_2 \in \mathbb{N}$, такое, что $z_2 > \max(z_1, 2)$ и

$$(\ln 2)^{-1} \cdot \int\limits_{z_2}^{\infty} \frac{r'(\log_2 x)}{x \log_2 x}\, dx < 1.$$

Заметим, что функция $r'(x)$ убывающая, и определим следующие величины

$$a_i = \frac{(\ln 2)^{-2}}{2} \cdot \frac{r'(\log_2 \log_2 (z_2 + i))}{(z_2 + i) \log_2^2 (z_2 + i)},\ i = 1, 2 \ldots$$

Выражение $\sum_{i=1}^{\infty} a_i$ меньше $1/2$, поскольку

$$\sum_{i=1}^{\infty} a_i < \frac{(\ln 2)^{-1}}{2} \cdot \int\limits_{z_2}^{\infty} \frac{r'(\log_2 x)}{x \log_2 x}\, dx.$$

Положим $a_0 = 1/2 - \sum_{i=1}^{\infty} a_i$. Используем введенные величины для построения разбиения интервала $(0, 1)$.

$$\Delta_0 = (0, a_0),\ \Delta_1 = (a_0, a_0 + a_0),\ \ldots,$$

$$\Delta_{2k-1} = (2\sum_{i=0}^{k-2} a_i + a_{k-1}, 2\sum_{i=0}^{k-1} a_i),\ \Delta_{2k} = (2\sum_{i=0}^{k-1} a_i, 2\sum_{i=0}^{k-1} a_i + a_k),\ \ldots$$

Функция плотности f определяется так, что вероятность нечетных интервалов равняется нулю, а вероятность четного интервала Δ_{2k}, $k = 0, 1, \ldots$, включая концевые точки, равняется величине $2a_k$. Например, функция плотности f определим так, что на нечетных интервалах она равна нулю, а на четных представляет из себя равнобедренный треугольник высотой 4. Площадь равнобедренного треугольника на интервале Δ_{2k}, $k = 0, 1, \ldots$, обозначим через p_k, заметим, что $p_k = 2a_k$. Если в точках 0 и 1 функцию f доопределить нулем, то она будет интегрируема по Риману на отрезке $[0, 1]$, и интеграл равен единице.

Функцию плотности g определяется также как и в предыдущей теореме. Вероятность четных интервалов равняется нулю, а вероятность нечетного интервала Δ_{2k+1}, $k = 0, 1, \ldots$, включая концевые точки, равняется величине $\frac{1}{(k+1)(k+2)}$. К примеру рассмотрим функцию плотности g, равной нулю на четных интервалах, а на нечетных интервалах Δ_{2k+1}, $k = 0, 1, \ldots$, она имеет вид равнобедренного треугольника высотой h_k равной

$$h_k = \frac{4a \log_2^2(k+2)}{k+1}.$$

Площадь этого треугольника обозначим через q_k. Заметим, что

$$q_k = \frac{1}{(k+1)(k+2)}.$$

Ряд $\sum_{i=0}^{\infty} q_i$ сходится, и сумма его равна единице. Получившаяся функция плотности g также интегрируема по Риману на отрезке $[0, 1]$, если в концах отрезка ее доопределить нулем.

Построение функций плотности f и g закончено. Вероятность того, что запись библиотеки попадет в интервал Δ_{2k+1}, $k = 0, 1, \ldots$, равна $\frac{1}{(k+1)(k+2)}$, а вероятность попадания записи в остальные интервалы равна нулю. Поэтому для события A_n, которое состоит в том, что ни в одном из интервалов Δ_{2k+1}, где $k > n^2 - 2$, нет ни одной из n записей библиотеки, и для события B_n, которое заключается в том, что в каждом из интервалов $\Delta_{2i+1}, i = 0, 1, \ldots, [n^{1/3}]$, есть хотя бы одна из n записей библиотеки, верна лемма 1.

$$\mathbf{P}\,(A_n) = 1 - o\left(\frac{1}{n}\right) \quad (n \to \infty),$$
$$\mathbf{P}\,(B_n) = 1 - o(1) \quad (n \to \infty).$$

Покажем, что $\mathbf{M}_V T_n^{(f,g)}(V))$ по порядку не больше, чем $r(\log_2 n)$. Математическое ожидание сложности рассматриваемого класса задач можно оценить с помощью энтропии [7]

$$\mathbf{M}_V T_n^{(f,g)}(V)) < \mathbf{M}_V(H_n^{(f,g)}(V)) + 2.$$

Используя условные математические ожидания, математическое ожидание энтропии можно записать следующим образом

$$\mathbf{M}_V(H_n^{(f,g)}(V)) = \mathbf{M}_V(H_n^{(f,g)}(V)|A_n) \cdot \mathbf{P}\,(A_n) + \\ + \mathbf{M}_V(H_n^{(f,g)}(V)|\overline{A_n}) \cdot \mathbf{P}\,(\overline{A_n}).$$

Величина $(\mathbf{M}_V(H_n^{(f,g)}(V)|\overline{A_n}) \cdot \mathbf{P}\,(\overline{A_n}))$ есть бесконечно малое при $n \to \infty$, поскольку $\mathbf{P}\,(\overline{A_n}) = o(1/n)$, а энтропия всегда ограничена верхней частью от $\log_2(n+1)$. Учитывая, что $\lim_{n\to\infty} \mathbf{P}\,(A_n) = 1$, получаем

$$\mathbf{M}_V(H_n^{(f,g)}(V)) \asymp \mathbf{M}_V(H_n^{(f,g)}(V)|A_n) \quad (n \to \infty).$$

Для функции энтропии верно следующее свойство [7]. При фиксированных значениях $t_1, t_2, \ldots, t_{n-k}$ выполнено

$$H(t_1, t_2, \ldots, t_{n-k}, t_{n-k+1}, \ldots, t_n) \geqslant H(t_1, t_2, \ldots, t_{n-k}, q),$$

где $q = t_{n-k+1} + \cdots + t_n$. Поэтому энтропия любой ЗПИО из класса A_n, не превосходит значения

$$H(p_0, p_1, \ldots, p_{n^2-1}, 1 - \sum_{i=0}^{n^2-1} p_i).$$

Для условного математического ожидания энтропии получаем следующую оценку

$$\mathbf{M}_V(H_n^{(f,g)}(V)|A_n) \leqslant H\left(p_0, p_1, \ldots, p_{n^2-1}, 1 - \sum_{i=0}^{n^2-1} p_i\right).$$

Последнее слагаемое в функции энтропии оценим единицей. Получим следующее неравенство

$$H(p_0, p_1, \ldots, p_{n^2-1}, 1 - \sum_{i=0}^{n^2-1} p_i) < - \sum_{i=0}^{n^2-1} p_i \log_2 p_i + 1.$$

Оценим выражение $H_{n^2-1} = \sum_{i=1}^{n^2-1} p_i \log_2 p_i$ при $n \to \infty$. По построению функции f

$$p_i = (\ln 2)^{-1} \frac{r'(\log_2(z_2 + i))}{(z_2 + i)\log_2(z_2 + i)},\ i = 1, 2, \ldots, n^2 - 1,$$

Подставляя эти значения и раскрывая скобки, получим

$$\begin{aligned} H_{n^2-1} = {} & (\ln 2)^{-1} \sum_{i=z_2+1}^{z_2+n^2-1} \frac{r'(\log_2 i)}{i} \\ & + (\ln 2)^{-2} \sum_{i=z_2+1}^{z_2+n^2-1} \frac{r'(\log_2 i) \cdot \log_2 \log_2 i}{i \log_2 i} \\ & - (\ln 2)^{-2} \sum_{i=z_2+1}^{z_2+n^2-1} \frac{r'(\log_2 i) \cdot \log_2 [r'(\log_2 i)]}{i \log_2 i}. \end{aligned}$$

Поскольку нам нужна верхняя оценка, отбросим последний ряд и получим оценку

$$\begin{aligned} H_{n^2-1} \leqslant {} & (\ln 2)^{-2} \sum_{i=z_2+1}^{z_2+n^2-1} \frac{r'(\log_2 \log_2 i)}{i \log_2 i} \\ & + (\ln 2)^{-2} \sum_{i=z_2+1}^{z_2+n^2-1} \frac{r'(\log_2 \log_2 i) \cdot \log_2 \log_2 i}{i \log_2^2 i}. \end{aligned}$$

В получившейся оценке для H_{n^2-1} ряд вида

$$\sum_{i=z_2+1}^{z_2+n^2-1} \frac{r'(\log_2 i) \cdot \log_2 \log_2 i}{i \log_2 i}$$

ограничен константой, поскольку мажорируется сходящимся рядом

$$\sum_{i=z_2+1}^{z_2+n^2-1} \frac{r'(\log_2 i)\cdot \log_2\log_2 i}{i\log_2 i} < \sum_{i=z_2}^{\infty} \frac{(\log_2 i)^{\alpha}\cdot \log_2\log_2 i}{i\log_2 i},$$

где $0<\alpha<1$. Функция $r'(x)$ убывающая, поэтому верны следующие оценки

$$(\ln 2)^{-1}\sum_{i=z_2+1}^{z_2+n^2-1} \frac{r'(\log_2 i)}{i} \leqslant (\ln 2)^{-1}\int\limits_{z_2+1}^{z_2+n^2-1} \frac{r'(\log_2 x)}{x}\,dx + \frac{r'(\log_2(z_2+1))}{(z_2+1)}.$$

Обозначим $a = z_2+1$, $b = z_2+n^2-1$. Интеграл в правой части неравенства равен

$$(\ln 2)^{-1}\int\limits_a^b \frac{r'(\log_2 x)}{x}\,dx = \int\limits_a^b r'(\log_2 x)\,d(\log_2 x) = r(\log_2(b)) - r(\log_2(a)).$$

Заметим, что выражение $r(\log_2(a))$ — константа. Собирая вместе полученные оценки, получаем что $\exists C>0$, такое, что $\forall n$ выполнено

$$H_{n^2-1} < r(\log_2(z_2+n^2-1)) + C.$$

Покажем, что

$$r(\log_2(z_2+n^2-1)) \asymp r(\log_2 n) \quad (n\to\infty).$$

Представим внутренний логарифм исследуемой функции в виде суммы

$$\log_2(z_2+n^2-1) = 2\log_2 n + \log_2\left(1+\frac{z_2-1}{n^2}\right).$$

Обозначим второе слагаемое в правой части через $sl(n)$. Заметим, что

$$sl(n) = \log_2\left(1+\frac{z_2-1}{n^2}\right) = o(1) \quad (n\to\infty)$$

Подставим выражение для внутреннего логарифма в функцию

$$r(\log_2(z_2+n^2-1)) = r(2\log_2 n + sl(n)).$$

Начиная с некоторого n_0 выражение $(2\log_2 n + sl(n))$ можно оценить сверху $3\log_2 n$

$$r(2\log_2 n + sl(n)) < r(3\log_2 n).$$

По условию теоремы

$$r(3\log_2 n) \asymp r(\log_2 n) \quad (n \to \infty).$$

Мы показали, что верхняя оценка для математического ожидания сложности ЗПИО по порядку не больше, чем $r(\log_2 n)$ при $(n \to \infty)$.

Покажем, что нижняя оценка по порядку не меньше, чем $r(\log_2 n)$. Для математического ожидания сложности ЗПИО по классу задач верна следующая оценка

$$\mathbf{M}_V\left(T_n^{(f,g)}(V)\right) \geqslant \mathbf{M}_V\left(H^{(f,g)}(V)\right).$$

Используя условные математические ожидания, математическое ожидание энтропии можно записать следующим образом

$$\mathbf{M}_V(H_n^{(f,g)}(V)) = \\ = \mathbf{M}_V(H_n^{(f,g)}(V)|B_n) \cdot \mathbf{P}\,(B_n) + \mathbf{M}_V(H_n^{(f,g)}(V)|\overline{B_n}) \cdot \mathbf{P}\,(\overline{B_n}).$$

Поскольку по определению функция энтропии неотрицательна и вероятность события B_n стремится к единице, получаем

$$\mathbf{M}_V(H_n^{(f,g)}(V)) \asymp \mathbf{M}_V(H_n^{(f,g)}(V)|B_n) \quad (n \to \infty).$$

Энтропия для задачи из класса B_n всегда содержит слагаемые $(-p_i \log_2 p_i)$, где $i = 1, 2, \ldots, [n^{2/3}]$,

$$\mathbf{M}_V(H_n^{(f,g)}(V)|B_n) \geqslant \sum_{i=1}^{[n^{2/3}]} p_i \log_2 p_i.$$

Обозначим выражение $\sum_{i=1}^{[n^{2/3}]} p_i \log_2 p_i$ через $H_{[n^{2/3}]}$. Подставим в выражение $H_{[n^{2/3}]}$ значение

$$p_i = (\ln 2)^{-1} \frac{r'(\log_2(z_2+i))}{(z_2+i)\log_2(z_2+i)},\ i = 1, 2, \ldots, [n^{2/3}].$$

Получаем

$$H_{[n^{2/3}]} = (\ln 2)^{-1} \sum_{i=z_2+1}^{z_2+[n^{2/3}]} \frac{r'(\log_2 i)}{i} + \\ + (\ln 2)^{-1} \sum_{i=z_2+1}^{z_2+[n^{2/3}]} \frac{r'(\log_2 i) \cdot \log_2 \log_2 i}{i \log_2 i} - \\ - (\ln 2)^{-1} \sum_{i=z_2+1}^{z_2+[n^{2/3}]} \frac{r'(\log_2 i) \cdot (\log_2 [r'(\log_2 i)])}{i \log_2 i}.$$

Поскольку нам нужна только нижняя оценка, можно отбросить второе слагаемое

$$H_{[n^{2/3}]} \geqslant (\ln 2)^{-1} \sum_{i=z_2+1}^{z_2+[n^{2/3}]} \frac{r'(\log_2 i)}{i} - \\ - (\ln 2)^{-1} \sum_{i=z_2+1}^{z_2+[n^{2/3}]} \frac{r'(\log_2 i) \cdot (\log_2 [r'(\log_2 i)])}{i \log_2 i}.$$

В получившейся оценке для $H_{[n^{2/3}]}$ ряд вида

$$(\ln 2)^{-1} \sum_{i=z_2+1}^{z_2+[n^{2/3}]} \frac{r'(\log_2 i) \cdot (\log_2 [r'(\log_2 i)])}{i \log_2 i}$$

с положительными членами ограничен константой, поскольку мажорируется сходящимся рядом

$$\sum_{i=z_2+1}^{z_2+[n^{2/3}]} \frac{r'(\log_2 i) \cdot (\log_2 [r'(\log_2 i)])}{i \log_2 i} < \sum_{i=z_2+1}^{+\infty} \frac{(\log_2 i)^{\alpha} \cdot (\alpha \log_2 \log_2 i)}{i \log_2 i}.$$

Для ряда вида

$$(\ln 2)^{-1} \sum_{i=z_2+1}^{z_2+[n^{2/3}]} \frac{r'(\log_2 i)}{i}$$

верны следующие оценки

$$(\ln 2)^{-1} \sum_{i=z_2+1}^{z_2+[n^{2/3}]} \frac{r'(\log_2 i)}{i} \geqslant (\ln 2)^{-1} \int_{z_2+1}^{z_2+[n^{2/3}]} \frac{r'(\log_2 x)}{x}\, dx.$$

Обозначим $a = z_2 + 1$, $b = z_2 + [n^{2/3}]$. Верна следующая цепочка равенств

$$(\ln 2)^{-1} \int_a^b \frac{r'(\log_2 x)}{x}\, dx = \int_a^b r'(\log_2 x)\, d(\log_2 x) = r(\log_2(b)) - r(\log_2(a)).$$

Заметим, что выражение $r(\log_2(a))$ — константа. Собирая вместе полученные оценки, получаем что $\exists C > 0$, такое, что $\forall n$ выполнено

$$H_{[n^{2/3}]} < r(\log_2(z_2 + [n^{2/3}])) - C.$$

Покажем, что

$$r(\log_2(z_2 + [n^{2/3}])) \asymp r(\log_2 n) \quad (n \to \infty).$$

Оценим внутренний логарифм в выражении $r(\log_2(z_2 + [n^{2/3}]))$

$$\log_2(z_2 + [n^{2/3}]) \geqslant \frac{2}{3} \cdot \log_2 n + \log_2 \left(1 + \frac{z_2 - 1}{n^{2/3}}\right) \geqslant \frac{2}{3} \cdot \log_2 n.$$

Подставим оценку для внутреннего логарифма в возрастающую функцию r

$$r(\log_2(z_2 + [n^{2/3}])) \geqslant r\left(\frac{2}{3} \log_2 n\right).$$

По условию теоремы

$$r\left(\frac{2}{3} \log_2 n\right) \asymp r(\log_2 n).$$

Мы показали, что нижняя оценка для математического ожидания сложности ЗПИО по порядку не меньше, чем $r(\log_2 n)$ при $(n \to \infty)$.

Доказательство теоремы закончено.

5. Доказательство следствия 2

Покажем, что любую функция из семейства

$$\{(\log_2 n)^\alpha | 0 < \alpha < 1, \alpha \in \mathbb{R}\},$$

можно представить в виде $r(\log_2(n))$, где функция $r(x)$ — возрастающая, положительная, дифференцируемая функция, определенная на интервале $(x_0, +\infty)$, $x_0 > 0$. Функция $r(x)$ сохраняет порядок и имеет в качестве производной монотонную, положительную и непрерывную функцию $r'(x)$, удовлетворяющую условию:

$$\exists \beta \in \mathbb{R}, 0 < \beta < 1 : \overline{\lim_{x \to +\infty}} \frac{r'(x)}{x^{\beta-1}} \leqslant 1.$$

Положим $r(x) = x^\alpha$, $0 < \alpha < 1$. Функция $r(x)$ — возрастающая, положительная, дифференцируемая функция на интервале $(0, +\infty)$. Из равенства

$$(x \cdot c)^\alpha = x^\alpha c^\alpha,$$

верном при любом вещественном $c \neq 0$, следует, что функция $r(x)$ сохраняет порядок. Производная $r'(x) = x^{\alpha-1}$ — монотонно убывает, непрерывна и положительна на интервале $(0, 1)$. В качестве β возьмем любое вещественное число из интервала $(\alpha, 1)$. Проверим условие

$$\overline{\lim_{x \to +\infty}} \frac{x^{\alpha-1}}{x^{\beta-1}} = \overline{\lim_{x \to +\infty}} \frac{1}{x^{\beta-\alpha}} = 0.$$

Доказательство следствия закончено.

Список литературы

[1] Гасанов Э. Э., Кудрявцев В. Б. Теория хранения и поиска информации. М.: Физматлит, 2002.

[2] Gilbert E. N, Moore E. F. Variable-length binary encodings // Bell System Tech. J. 1959. **38**. P. 933–968.

[3] Knuth D. E. Optimum binary search trees // Acta Informatica. 1971. **1**. P. 14–25.

[4] Garsia A. M., Wachs M. L. A new algorithm for minimum cost binary trees // SICOMP. 1977. **4**. P. 622–642.

[5] Кучеренко Н. С. Сложность поиска идентичных объектов в случайных базах данных // Интеллектуальные системы. 2007. Т. 11, вып. 1–4. С. 495–516.

[6] Кучеренко Н. С. Средняя сложность поиска идентичных объектов для случайных неравномерных баз данных // Дискретная математика — в печати.

[7] Кнут Д. Э. Искусство программирования. М.: Издательский дом «Вильямс», 2000. Т. 3.