



МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ М.В. ЛОМОНОСОВА

Механико-математический факультет
Кафедра математической логики и теории алгоритмов

На правах рукописи

Милованов Алексей Сергеевич

О нестохастических по Колмогорову словах

Специальность 01.01.06 – математическая логика, алгебра и теория чисел

ДИССЕРТАЦИЯ
на соискание ученой степени
кандидата физико-математических наук

Научный руководитель
доктор физико-математических наук
профессор Н. К. Верещагин

Москва — 2017

Оглавление

1	Введение	4
1.1	Актуальность темы и цели работы	4
1.2	Основные результаты	6
1.3	Апробация работы	6
1.4	План дальнейшего изложения	7
1.5	Используемые обозначения	7
	Благодарности	8
2	Основные понятия и результаты алгоритмической статистики	9
2.1	Алгоритмическая теория информации	9
2.2	Стохастичные слова	10
2.3	Профиль слова	12
2.4	Связь между дефектом случайности и дефектом оптимальности . .	14
2.5	Стандартные модели	15
2.6	Ограниченные классы гипотез	16
2.7	Сильные статистики	17
3	Антистохастические слова	22
3.1	Существование антистохастических слов	22
3.2	Антистохастические слова и числа Ω_i	24
3.3	Голографичность антистохастичных слов	25
3.4	Антистохастические слова и декодирование списком	27
3.5	Антистохастические слова и тотальная условная сложность	29
4	Предсказания и алгоритмическая статистика	31
4.1	Дискретная априорная вероятность	31
4.2	Предсказательные окрестности	32
4.2.1	Вероятностная предсказательная окрестность	32
4.2.2	Алгоритмическая предсказательная окрестность	33
4.3	Доказательство Теоремы 4.3	34
4.4	Ограниченные классы гипотез	37
5	Алгоритмическая статистика для нескольких слов	41
5.1	Профиль нескольких слов	42
5.2	Многомерный дефект случайности	44
5.3	Предсказания для неравномерных распределений	46
5.3.1	Неравномерные распределения для одного слова	47
5.3.2	Общий случай	49

6	Сильные статистики и нормальные объекты	51
6.1	Существование нормальных слов	51
6.2	Нормальные слова и стандартные модели	52
6.3	Наследственность нормальных слов	55
	Заключение	61

Глава 1

Введение

1.1 Актуальность темы и цели работы

Нестохастические слова являются основным объектом изучения алгоритмической статистики. Задачу алгоритмической статистики в первом приближении можно описать следующим образом. Пусть имеется некоторое устройство (“чёрный ящик”). Его включили, и оно выдало некоторую конечную последовательность битов. Что можно сказать об устройстве этого чёрного ящика, зная эту последовательность?

Сразу же обратим внимание на особенности этой задачи. Во-первых, у нас нет никакой априорной информации о “чёрном ящике”. Во-вторых, нет возможности ещё раз повторить эксперимент, т.е. ещё раз включить устройство. (Отметим, что случаи однократного и невоспроизводимого эксперимента довольно часто встречаются на практике.)

Может показаться, что при таких ограничениях никаких адекватных предположений о работе “чёрного ящика” выдвинуть невозможно. Тем не менее для некоторых последовательностей можно выдвинуть разумные гипотезы. Например, если устройство выдало длинную последовательность из одних нулей, то кажется разумным предположить, что кроме нулей оно ничего выдать не может. А если “чёрный ящик” выдал длинную последовательность из нулей и единиц безо всякой закономерности, то, видимо, устройство просто выдаёт биты случайным образом.

Что общего в этих двух примерах, т.е. в каком случае мы считаем объяснение приемлемым? Заметим, что каждое из объяснений можно представить в виде некоторого конечного множества: в первом случае это множество, состоящее из одного слова (т.е. последовательности битов) $0 \dots 0$, во-втором — множество всех слов той же длины, как у данного нам слова. Наши гипотезы в обоих случаях состоят в том, что устройство выбирает слово случайным образом из соответствующих множеств согласно равномерному распределению.

В общем случае, для данного нам слова x , в качестве “объяснения” можно рассматривать некоторое конечное множество A , содержащее x . При этом, такое объяснение будет считаться “хорошим”, если:

- Множество A является “простым”;
- слово x является “типичным представителем” множества A .

Как уточнить эти требования, т.е. как формализовать понятия “простое множество” и “типичный представитель” в нём? А.Н. Колмогоров дал строгие опреде-

ления этим понятиям в 1974-ом году [7, 13], используя алгоритмическую теорию информации. А именно, множество A считается простым, если у него малая *колмогоровская сложность*, т.е. если существует короткая программа, которая печатает список всех элементов A . Это определение зависит от выбора языка программирования, однако, согласно теореме Колмогорова-Соломонова, существует такой оптимальный язык программирования, что сложности всех слов (и более того, сложности всех конечных объектов, например, списка слов) минимальны с точностью до аддитивной константы. В этой работе, мы будем обозначать сложность конечного объекта x через $C(x)$. Условная сложность x относительно y , т.е. длина кратчайшей программы, которая печатает x на входе y , обозначается через $C(x|y)$.

Для того, чтобы определить типичность слова в содержащем его множестве, было введено понятие *дефекта случайности*, которое определяется следующим образом

$$d(x|A) := \log |A| - C(x|A).$$

Типичными элементами множества называют те его элементы, у которых дефект случайности мал.

Через некоторое время А.Н. Колмогоров задал вопрос: для любого ли слова найдется хорошее объяснение [6]? Объекты, для которых такое объяснение существует, Колмогоров называл *стохастическими*. Более точно, слово x называется α, β -стохастическим, если существует такое множество $A \ni x$, что $C(A) \leq \alpha$ и $d(x|A) \leq \beta$.

А.Х. Шень доказал [11], что существуют *нестохастические* слова с линейными (от длины слов) параметрами нестохастичности. Они представляют наибольший интерес для алгоритмической статистики. В этой же работе были даны верхние и нижние оценки доли нестохастических объектов заданной длины. Там же получены верхняя оценка меры таких объектов для вычислимого распределения ограниченной сложности. В работе [4] В.В. Вьюгин получил аналогичные верхние и нижние оценки для универсальной полувычислимой меры множества нестохастических объектов данной длины. Там же есть связь нестохастичности для конечных и бесконечных последовательностей.

В работе В.В. Вьюгина [3] изучены формы кривых зависимости дефекта случайности β относительно мер сложности не превосходящей α для случая когда $\alpha = o(\beta)$. То же самое для случайности относительно конечного множества сделано в статье [31]. Этот результат был обобщен на произвольный случай в статье Верещагина и Витаньи [29].

Нестохастичность для бесконечных последовательностей изучалась в статьях [2, 17, 18].

Отметим также работу Гача, Тромпа и Витаньи [14], в которой (среди прочего) и был введён термин “алгоритмическая статистика”.

В целом, сейчас эта наука является теоретической, однако некоторые практические результаты опираются на её идеи [22].

Целями данной работы являются изучение свойств нестохастических объектов, дальнейшее развитие алгоритмической статистики, а также установление её связей с другими науками.

1.2 Основные результаты

Работа содержит четыре основных результата.

Во-первых были изучены свойства максимально нестохастических слов (существование которых было доказано Верецагиным и Витаньи в [29]), они также называются *антистохастическими*. Доказано, что эти объекты (которые являются самыми сложными для объяснения с точки зрения алгоритмической статистики), обладают свойством *голографичности* (Теорема 3.4). Неформально, голографичность слова x колмогоровской сложности k означает, что x можно восстановить по любым содержащимся в нём k битам информации. Показано, как свойство голографичности антистохастических слов можно применить для теории кодирования (Раздел 3.4) и построения контрпримеров в алгоритмической теории информации (Теорема 3.9).

Во-вторых, была установлена связь алгоритмической статистики с задачей предсказания. А именно, ставится следующая задача. Пусть некоторое устройство выдало какое-то слово. Зададимся вопросом: какие слова мы ожидаем увидеть, если запустим это устройство ещё раз? Оказывается, что если должным образом формализовать эту задачу, то ответ на неё будет напрямую связан с алгоритмической статистикой — Теорема 4.3.

В-третьих, была разработана теория алгоритмической статистики для нескольких слов. Т.е. рассматривается задача нахождения приемлемого объяснения для устройства, которое выдало данные несколько слов. Оказывается, основные понятия и результаты алгоритмической статистики могут быть перенесены на этот случай (Теоремы 5.7 и 5.10).

В-четвёртых, были получены результаты о *нормальных* словах. Концепция нормальных объектов была введена Н.К. Верецагиным в [25], в связи с изучением *сильных моделей*, т. е. таких множеств A , содержащих данное слово x , что существует короткая *тотальная* (т.е. останавливающаяся на всех входах) программа, переводящая x в A . Неформально, нормальными называются те слова, для которых можно ограничиться сильными модели в качестве объяснений (т.е. другие множества будут объяснять данное слово не лучше). Автором была доказана теорема о существовании нормальных слов с наперёд заданными характеристиками (Теорема 6.1), доказано, что они обладают свойством *наследственности* (Теорема 6.5). Были получены ответы на некоторые из вопросов, поставленных в [25].

1.3 Апробация работы

Основные результаты, полученные в работе были изложены на следующих международных конференциях:

- “International Computer Science Symposium in Russia” в Листвянке в 2015-ом году;
- “Symposium on Theoretical Aspects of Computer Science” в Орлеане (Франция) в 2016-ом году;
- “International Computer Science Symposium in Russia” в Санкт-Петербурге в 2016-ом году

и опубликованы в сборниках трудов этих конференций [33–35] (Серии Lecture Notes in Computer Science и Leibniz International Proceedings in Informatics входят в систему Scopus).

Кроме того, результаты были изложены на следующих научных конференциях и семинарах:

- Международная научная конференция студентов, аспирантов и молодых учёных “Ломоносов-2016”;
- конференция “Проблемы теоретической информатики”, Москва, 2015;
- Колмогоровский семинар по сложности вычислений и сложности определений, механико-математический факультет, МГУ им. Ломоносова;
- Межкафедральный семинар МФТИ по дискретной математике;
- семинар “Логические проблемы информатики”, механико-математический факультет, МГУ им. Ломоносова;
- семинар LIRMM, Монпелье (Франция).

1.4 План дальнейшего изложения

Дальнейший текст организован следующим образом. В Главе 2 даётся обзор алгоритмической статистики и алгоритмической теории информации, формулируются теоремы, которые будут использоваться в дальнейшем (в частности в Разделе 2.7 более подробно рассказывается о теории сильных моделей). Последующие главы организованы, согласно полученным результатам.

Работа заканчивается заключением, в котором излагается план дальнейших исследований, и списком литературы, содержащим 28 наименований, в том числе 4 работы автора по теме диссертации.

1.5 Используемые обозначения

- Все логарифмы по умолчанию берутся по основанию 2.
- Двоичные слова будем обозначать маленькими латинскими буквами, например, x , y , z . Множество двоичных слов длины n будем обозначать через $\{0, 1\}^n$, а множество всех двоичных слов — через $\{0, 1\}^*$. Пустое слово будем обозначать через Λ . Длину слова x будем обозначать через $|x|$.
- Конечные множества будем обозначать большими латинскими буквами, например, A , B , D . Количество элементов в множестве A будем обозначать через $|A|$. Семейства множеств будут обозначаться каллиграфическими латинскими буквами, например, $\mathcal{A}$, $\mathcal{B}$.
- Через C_n^k мы обозначаем количество k -элементных подмножеств из n -элементного множеств (т.е. число сочетаний из n по k).
- Колмогоровская сложность x при известном y обозначается через $C(x|y)$.

- Тотальная сложность x при известном y обозначается через $CT(x|y)$.
- Безусловная колмогоровская сложность x (т.е. $C(x|\Lambda)$) обозначается через $C(x)$.

Благодарности

Автор выражает глубокую благодарность своему научному руководителю Николаю Константиновичу Верещагину за постоянную поддержку и внимание к работе. Автор также благодарит всех участников Колмогоровского семинара за внимание и интерес к докладам автора.

Работа выполнена при частичной финансовой поддержке гранта РФФИ 16-01-00362 и гранта “Молодая математика России”.

Наконец, автор глубоко признателен своей семье за постоянную поддержку.

Глава 2

Основные понятия и результаты алгоритмической статистики

В следующем разделе мы даём краткий обзор основных понятий колмогоровской сложности (более подробно ознакомиться с этой темой можно в [1, 19, 23]), которые нам потребуются в дальнейшем. В последующих разделах этой главы мы опишем основные результаты алгоритмической статистики, оставляя почти всё без доказательств (более подробный обзор с доказательствами имеется в [27] и в 14-ой главе [1]).

В Главе 3 будут использоваться результаты Разделов 2.1–2.5, в Главах 4 и 5 будут использоваться Разделы 2.1–2.6, а в Главе 6 — Разделы 2.1–2.7.

2.1 Алгоритмическая теория информации

Колмогоровская сложность — это способ формализации понятия “количество информации” для конечного объекта.

Пусть $D : \{0, 1\}^* \times \{0, 1\}^* \rightarrow \{0, 1\}^*$ — некоторая частично вычислимая функция. *Условной колмогоровской сложностью* слова x относительно слова y при способе описания D называется наименьшая длина такого кратчайшего слова p , что $D(p, y) = x$. Это число обозначают $C_D(x|y)$.

В этом контексте функцию D называют *способом описания*. Если $D(p, y) = x$, то p называют *программой*, переводящей y в x .

Способ описания D называется *универсальным* если для любого другого способа описания D' существует такое слово s , что $D'(p, y) = D(sp, y)$ для всех p, y . По теореме Колмогорова-Соломонова универсальные способы описания существуют [8].

Выберем произвольный универсальный способ описания D , будем называть величину $C_D(x|y)$ *колмогоровской сложностью* x при известном y и будем обозначать её как $C(x|y)$. Теперь определим безусловную колмогоровскую сложность $C(x)$ слова x как $C(x|\Lambda)$.

Замечание 1. Эта версия колмогоровской сложности называется *простой* сложностью; существуют также и иные варианты, такие как префиксная сложность, монотонная сложность и другие. Все эти виды совпадают с обычной сложностью с точностью до логарифмического от длины слова слагаемого. Все наши результаты имеют такую же или худшую точность, поэтому переход к другим видам

сложностей оставит наши результаты в силе.

Колмогоровскую сложность можно естественным образом определить для других конечных объектов (пар слов, конечного множества слов и так далее). Для этого нужно зафиксировать некоторую вычислимую биекцию (“кодирование”) между этими объектами и бинарными словами и определить сложность объекта как сложность соответствующей ему бинарного слова. Это определение инвариантно относительно выбора кодирования: выбор другой вычислимой биекции может изменить значение сложности не более чем на аддитивную константу.

Зафиксируем какую-нибудь вычислимую биекцию между словами и конечными подмножествами $\{0, 1\}^*$. Будем обозначать слово, которая соответствует конечному подмножеству $A \subset \{0, 1\}^*$ через $[A]$. Сложность этого конечного множества $C(A)$ определяется как $C([A])$. Аналогично, $C(x|A)$ и $C(A|x)$ определяются как $C(x|[A])$ и $C([A]|x)$.

Мы будем использовать следующие утверждения о Колмогоровской сложности, доказательства которых можно найти в [1] и в [19].

- Для любого слова x выполнено $C(x) \leq |x| + O(1)$.
- Число слов сложности меньше n заключено между $2^{n-O(1)}$ и 2^n .
- Для любой пары слов x и y :

$$C(x, y) \leq C(x) + C(y) + O(\log(C(x, y)));$$

- Для любой пары слов x и y :

$$C(x, y) = C(x) + C(y|x) + O(\log(C(x, y))).$$

Последнее равенство называют *коммутативностью информации*. Иногда этим же словосочетанием называют следующее утверждение (которое не сложно выводиться из предыдущего).

- Для любой пары слов x и y :

$$C(x) - C(x|y) = C(y) - C(y|x) + O(\log(C(x, y))).$$

2.2 Стохастичные слова

В этом разделе мы более подробно рассматриваем концепцию α, β -стохастичности, которую мы кратко обсудили во Введении.

Пусть слово x принадлежит конечному множеству A . Величина

$$d(x|A) := \log |A| - C(x|A)$$

называется *дефектом случайности* x во множестве A . Эта величина неотрицательна с точностью до аддитивной константы — в самом деле, зная A , мы можем восстановить x , по лексикографическому номеру x в A , используя не более $\log |A|$ бит.

С другой стороны, доля элементов A , для которых дефект случайности в A больше k не превосходит 2^{-k+1} , потому что число программ, у которых длина не больше $\log |A| - k$ не превосходит $|A|2^{-k+1}$. “Типичными элементами” множества A считаются те $x \in A$, для которых величина $d(x|A)$ мала.

Может возникнуть вопрос: почему в качестве объяснений рассматриваются множества (другими словами, равномерные распределения), а не произвольные вероятностные распределения? Дело в том, что для любого слова x и для любого распределения P найдётся такое множество A , которое “объясняет” x не хуже чем P . Перед тем как строго формулировать это утверждение опишем, о каких распределениях будет идти речь. Мы будем рассматривать распределение P как конечный список пар слов $(y, P(y))$, где y есть некоторое слово (таким образом, мы рассматриваем только распределения с конечным носителем). Также, мы ограничиваемся рассмотрением только тех распределений, для которых $P(y)$ рационально при любом y . Эти ограничения вводятся для того, чтобы рассматриваемые распределения были конечными объектами, у которых можно определить сложность. Дефект случайности слова x относительно распределения P определяется как $-\log P(x) - C(x|P)$ и обозначается через $d(x|P)$.

Предложение 2.1 ([29]). *Для любого слова x длины n и для любого распределения P существует такое множество $A \ni x$, что*

$$C(A) \leq C(P) + O(\log n) \text{ и} \\ d(x|A) \leq d(x|P) + O(\log n).$$

Доказательство. Рассмотрим такое целое k , что $2^{-k-1} > P(x) \geq 2^{-k}$.

Если $k > 2n$, то возьмём в качестве A множество всех слов длины n . Сложность этого множества есть $O(\log n)$, а дефект случайности $d(x|\{0,1\}^n) = n - C(x|\{0,1\}^n) \leq n$. С другой стороны, $d(x|P) > n - O(1)$, т.к. $C(x|P) \leq C(x) \leq n + O(1)$.

Если же $k \leq 2n$, то определим A как $\{y \mid P(y) \geq 2^{-k}\} \ni x$. Ясно, что $C(A|P) = O(\log n)$, из чего и следуют требуемые неравенства. $\square$

Сейчас мы строго сформулируем результаты о существовании слов, для которых не существует простых множеств, в которых эти слова были бы “типичными представителями”. Напомним, что слово x называется α, β -стохастическим, если существует такое множество $A \ni x$, что $C(A) \leq \alpha$ и $d(x|A) \leq \beta$.

Теорема 2.2 ([11]). *При $2\alpha + \beta < n - O(\log n)$ существуют слово длины n , не являющиеся α, β -стохастическим.*

(То есть существует такая константа c , что для всех n и всех таких α и β , что $2\alpha + \beta < n - c \log n$ существует слово длины n , не являющееся α, β -стохастическим.)

Этот результат был улучшен следующим образом.

Теорема 2.3 ([29]). *При $\alpha + \beta < n - O(\log n)$ существуют слово длины n , не являющиеся α, β -стохастическим.*

Следующая теорема показывает, что нестохастических слов довольно мало.

Теорема 2.4 ([27]). *Доля слов длины n не являющихся α, α -стохастическими не превосходит $2^{-\alpha + O(\log n)}$.*

2.3 Профиль слова

Можно оценивать качество объяснения множества A для содержащегося в нём слова x в несколько другой системе координат.

А именно, будем говорить, что множество A является *оптимальной* моделью для слова $x \in A$, если сложность и размер A малы настолько, насколько это возможно. Заметим, что для любого $x \in A$ выполнено следующее неравенство

$$C(A) + \log |A| \geq C(x) - O(\log C(x))$$

так как x можно восстановить по множеству A и лексикографическому номеру x в A (дополнительное логарифмическое слагаемое нужно для кодирования пары [1]).

Таким образом оба параметра $C(A)$ и $\log |A|$ не могут быть малыми одновременно, если сложность x велика.

Величина $C(A) + \log |A| - C(x)$ называется *дефектом оптимальности* x в A и обозначается как $\delta(x, A)$. Чем дефект оптимальности меньше, тем гипотеза предпочтительней. Но как сравнить две гипотезы, у которых одинаковый дефект оптимальности, но разные сложности? Ответ на этот вопрос даёт следующее

Предложение 2.5 ([1]). *Пусть конечное множество A содержит слово x . Тогда для любого числа $i < \log |A|$ существует такое конечное множество $A' \ni x$, что $|A'| \leq |A|/2^i$ и $C(A') \leq C(A) - i + O(\log i)$.*

Доказательство. Разобьём все элементы A на 2^i частей мощности $|A|/2^i$ (если A не является степенью двойки, то частей будет чуть больше). Обозначим ту часть, в которой находится x через A' . Чтобы её задать нужно знать A , число i и ещё i битов — номер части A' в A . $\square$

Таким образом, из простых моделей можно получать более сложные с таким же (примерно) дефектом оптимальности. Обратное же утверждение не верно — мы это увидим в Теореме 2.6. Таким образом, при равенстве дефектов оптимальности та модель ценнее, у которой меньше сложность (и значит, больше мощность).

Удобно рассмотреть множество параметров всех моделей, содержащих данное слово.

Определение 1. *Профилем* слова x называют множество P_x , состоящее из пар таких целых неотрицательных чисел (m, l) , для которых существует конечное множество $A \ni x$, для которого $C(A) \leq m$ и $\log |A| \leq l$.

Рисунок 2.1 показывает как может выглядеть профиль слова x длины n и сложности k .

Профиль любого слова x длины n и сложности k обладает следующими тремя свойствами.

- Во-первых, множество P_x замкнуто вверх: если P_x содержит пару (m, l) , то P_x также содержит все пары (m', l') , где $m' \geq m$ и $l' \geq l$.
- Во-вторых, P_x содержит множество

$$P_{\min} = \{(m, l) \mid m + l \geq n \text{ или } m \geq k\}$$

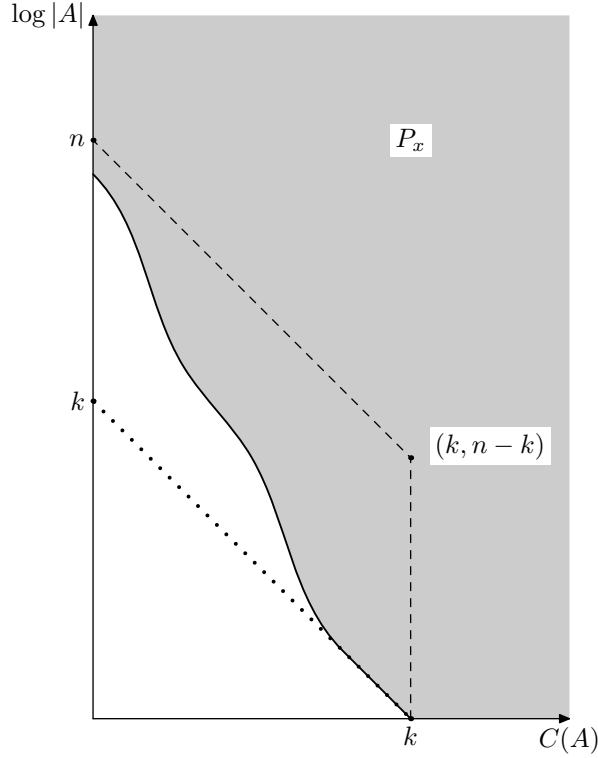


Рис. 2.1: Возможный профиль P_x слова x длины n и сложности k .

(это множество всех точек, находящихся сверху и справа от пунктирной линии на Рис. 2.1) и включено во множество

$$P_{\max} = \{(m, l) \mid m + l \geq k\}$$

(это множество расположено сверху и справа от линии из точек на Рис. 2.1). Другими словами, граница множества P_x (А. Н. Колмогоров называл её *структурной функцией* x), находится между пунктирной линией и линией из точек.

Оба включения понимаются с логарифмической точностью: множество $P_{\min}$ включено в $O(\log n)$ -окрестность множества P_x , множество P_x включено в $O(\log n)$ -окрестность множества $P_{\max}$.

В самом деле, P_x содержит $P_{\min}$ в силу того, что дефект оптимальности не отрицателен (с логарифмической точностью). Для того, чтобы показать, что P_x содержится в $P_{\max}$ заметим, что x содержится в $\{x\}$, а также, для каждого $i < k$ слово x принадлежит множеству всех слов длины n , чьи первые i битов такие же как у x .

- В-третьих, согласно Предложению 2.5 профиль P_x обладает следующим свойством:

$$\begin{aligned} &\text{если пара } (m, l) \in P_x, \text{ то} \\ &\text{пара } (m + i + O(\log n), l - i) \text{ также принадлежит } P_x \text{ для всех } i \leq l. \end{aligned}$$

Приведём примеры слов, чей профиль близок к множеству $P_{\max}$. Такими будут слова, состоящие из конкатенации случайного слова длины k (т.е. такого, у

которого колмогоровская сложность приблизительно равняется k) и слова из $n - k$ нулей.

Сложнее показать, что другие множества между $P_{\min}$ и $P_{\max}$ являются профилями некоторых слов. В [29] было показано, что для любого множества, которое удовлетворяет трём упомянутым свойствам профиля, существует слово длины n и сложности k , чей профиль совпадает с этим множеством с логарифмической точностью.

Теорема 2.6 ([29]). *Пусть замкнутое вверх множество пар натуральных чисел P содержит $P_{\min}$ и содержится в $P_{\min}$. Предположим также, что для всех $(m, l) \in P$ и всех $i \leq l$ выполнено $(m + i, l - i) \in P$.*

Тогда существует слово x длины n и сложности $k + O(\log n)$ чей профиль $C(P) + O(\log n)$ -близок к P .

В этой теореме мы говорим, что два подмножества $\mathbb{N}^2$ являются ε -близкими, если каждое из них содержится в ε -окрестности другого. В этом результате говорить о сложности множества P , которое не является конечным объектом. Тем не менее, если множество P удовлетворяет предположению теоремы, то его можно задать с помощью функции $h(l) = \min\{m \mid (m, l) \in P\}$. У этой функции лишь конечное число ненулевых значений, так как $h(k) = h(k + 1) = \dots = 0$. Таким образом h является конечным объектом, а значит мы можем определить $C(P)$ как сложность конечного объекта h .

2.4 Связь между дефектом случайности и дефектом оптимальности

Между подходами к измерению качества гипотез, которые обсуждались в предыдущих двух разделах, существует прямая связь. Во-первых, дефект случайности всегда не на много больше дефекта оптимальности, как показывает следующее

Предложение 2.7. *Для любого слова x и для любого содержащего его множества A выполнено следующее неравенство*

$$d(x|A) \leq \delta(x, A) + O(\log |x|).$$

Доказательство. Это предложение напрямую следует из очевидного неравенства $C(A) + C(x|A) \geq C(x) - O(\log |x|)$. $\square$

Обратное неравенство верно не всегда: дефект оптимальности может быть гораздо больше дефекта случайности, как показывает следующий пример.

Пример 1. Пусть x — случайное слово длины n , т.е. такое слово, у которого сложность приблизительно равняется n .

Пусть y — другое слово длины n , которое является случайным относительно x , т.е. $C(x|y) \approx n$.

Рассмотрим множество $A := \{0, 1\}^n \setminus \{y\}$, содержащее x . Для этого множества $\delta(x, A) = C(A) + \log |A| - C(x) \approx n + n - n = n$.

С другой стороны $d(x|A) = \log |A| - C(x|A) \approx n - n = 0$, так как

$$C(x|A) \approx C(x|y) \approx n.$$

Тем не менее, для любого множества A и любого содержащегося в нём слова x существует множество $B \ni x$, у которого сложность не сильно больше чем $C(A)$, а дефект оптимальности $\delta(x, B)$ не сильно больше дефекта случайности $d(x|A)$ (в примере выше в качестве B можно взять, например, $\{0, 1\}^n$). А именно, верна следующая

Теорема 2.8 ([29]). *Для любого конечного множества A и любого содержащегося в нём слова x , существует такое множество $B \ni x$, что*

$$\delta(x, B) \leq d(x|A) + O(\log |x|) \text{ и}$$

$$C(B) \leq C(A) + O(\log |x|).$$

Таким образом профиль слова определяет (с логарифмической точностью) его стохастичность (в частности, из Теоремы 2.6 следует Теорема 2.3). Это позволяет говорить для слов x и y одинаковой длины и сложности, что если $P_x \subset P_y$, то слово y стохастичнее x . У стохастических слов профиль близок $P_{\max}$. Те же слова, чей профиль близок $P_{\min}$, называются *антистохастическими*. Мы изучаем их свойства в Главе 3.

В Главе 5 мы докажем некоторое обобщение Теоремы 2.8.

2.5 Стандартные модели

В этом разделе мы покажем, как можно построить множества, содержащие данное слово x , чьи параметры лежат на границе профиля P_x . Таким образом, это будут наилучшие объяснения для x в смысле определённых выше параметров.

Рассмотрим некоторое число m . Обозначим через Ω_m количество слов, чья сложность не превосходит m . Рассмотрим некоторый простой алгоритм (длины $O(\log m)$), который перечисляет все такие слова. Теперь разложим число Ω_m в двоичной системе исчисления:

$$\Omega_m = 2^{t_1} + 2^{t_2} + \dots,$$

где $t_1 > t_2 > \dots$

Для данного перечисляющего алгоритма в соответствии с этим разложением в двоичную запись все слова сложности не больше m распадаются на порции размера 2^{t_1} , 2^{t_2} и так далее. Эти порции можно рассматривать как описания для соответствующих слов. Таким образом для каждого слова x и для каждого $m \geq C(x)$ мы получили некоторое множество, содержащее x . Такие множества мы будем называть *стандартными моделями* или *стандартными описаниями* для x . Заметим, что, зная слово x , перечисляющий алгоритм и размер 2^t порции, содержащий x , мы можем найти стандартную модель $B \ni x$, соответствующую x (запуская перечисляющий алгоритм до нахождения x , и подождя в случае надобности до тех пор, пока не перечислится очередная порция размера 2^t), т.е. $C(B|x) \approx 0$.

Теорема 2.9 ([14]). *Для любого слова x и любого конечного множества $A \ni x$ существует такая стандартная модель $B \ni x$, что*

$$\delta(x, B) \leq \delta(x, A) + O(\log |x|) \text{ и}$$

$$C(B|A) = O(\log |x|).$$

(В частности, $C(B)$ не превосходит $C(A) + O(\log |x|)$.)

Таким образом, по каждой модели $A \ni x$ можно получить стандартную модель $B \ni x$, которая будет не хуже A , а как обсуждалось в Разделе 2.3 из двух моделей с одинаковым дефектом оптимальности предпочтительнее та, у которой сложность меньше.

Два конечных объекта x и y называются ε -информационно эквивалентными, если величины $C(x|y)$ и $C(y|x)$ меньше ε .

Оказывается, что любая стандартная модель эквивалента Ω_i (так обозначается количество слов, чья колмогоровская сложность не превосходит i) для некоторого i .

Предложение 2.10 ([27]). Пусть B — стандартная модель сложности i , полученная перечислением списка слов сложности m . Тогда множество B является $O(\log m)$ -информационно эквивалентным числу Ω_i .

Мы будем пользоваться следующими свойствами чисел Ω_i .

Предложение 2.11 ([27]).

- (1) Для любого i выполнено $C(\Omega_i) = i + O(\log i)$.
- (2) Пусть $l \leq m$. Обозначим через $(\Omega_m)_{1:l}$ первые l бит Ω_m . Тогда числа $(\Omega_m)_{1:l}$ и Ω_l являются $O(\log l)$ -информационно эквивалентными. В частности, в этом случае $C(\Omega_l|\Omega_m) = O(\log m)$.

Тот факт, что стандартные модели (которые можно назвать наилучшими) имеют такой простой вид, не есть повод для оптимизма, а скорее наоборот. Эти описания не выглядят адекватными с интуитивной точки зрения. Поэтому, наличие таких множеств свидетельствует скорее о том, что “правильная” теория алгоритмической статистики нуждается в доработке. В следующих двух разделах мы покажем, как это можно сделать.

2.6 Ограниченные классы гипотез

В этом разделе мы будем предполагать, что нам известна некоторая информация о “чёрном ящике” который выдал данное нам слово x .

А именно, мы предполагаем, что существует некоторое семейство конечных множеств $\mathcal{A}$, и данное нам слово x было получено случайным выбором в одном из множеств из $\mathcal{A}$ (среди которых мы и будем искать подходящее объяснение).

В [30] было показано, как можно обобщить основные результаты предыдущих разделов для семейства $\mathcal{A}$, если оно обладает некоторыми естественными свойствами. Главным свойством, которое нужно потребовать, является перечислимость $\mathcal{A}$ (это означает, что существует алгоритм, который печатает по очереди множества (как списки слов) из $\mathcal{A}$). Также могут потребоваться следующие два свойства.

- Для всех n семейство $\mathcal{A}$ содержит множество $\{0, 1\}^n$;
- для некоторого полинома p , для каждого множества $A \in \mathcal{A}$ и для любых натуральных чисел n и $c < |A|$ выполнено следующее: существует покрытие $\{0, 1\}^n \cap A$, состоящее из не более $p(n)|A|/c$ множеств из $\mathcal{A}$, причем мощность каждого из покрывающих множеств не превосходит c .

Смысл последнего требования заключается в том, что для семейств, обладающих этим свойством, выполнен аналог Предложения 2.5.

Семейства, обладающие всеми тремя перечисленными свойствами называются *приемлемыми*.

Пример 2. Рассмотрим семейство, состоящее из всех ‘цилиндров’, где каждый ‘цилиндр’ есть множество вида $\{uv \mid v \in \{0, 1\}^m\}$, где u — произвольное бинарное слово, а m — произвольное целое неотрицательное число. Очевидно, что такое семейство является приемлемым.

В [1] показано, что множество всех шаров Хемминга также является приемлемым.

Для перечислимых семейств верен следующий аналог Теоремы 2.8.

Теорема 2.12 ([30]). *Для любого перечислимого семейства конечных множеств $\mathcal{A}$ верно следующее. Для любого множества $A \in \mathcal{A}$ и для любого слова x длины n , содержащегося в A , существует такое множество B , что выполнены следующие два неравенства*

$$\delta(x, B) \leq d(x|A) + O(\log n + \log C(A) + \log \log |A|);$$

$$C(B) \leq C(A) + O(\log n + \log C(A) + \log \log |A|).$$

Для приемлемых семейств множеств выполнен также аналог Теоремы 2.6. Перед тем как его формулировать, обозначим через P_x^A множество таких пар (m, l) , что существует $A \in \mathcal{A}$, содержащее x , для которого $C(A) \leq m$ и $\log |A| \leq l$.

Из приемлемости $\mathcal{A}$ следует, что множество P_x^A обладает теми же тремя свойствами (которые обсуждались в Разделе 2.3), что и “обычный” профиль P_x .

Теорема 2.13 ([30]). *Пусть замкнутое вверх множество пар натуральных чисел P содержит $P_{\min}$ и содержится в $P_{\max}$. Предположим также, что для всех $(m, l) \in P$ и всех $i \leq l$ выполнено $(m + i, l - i) \in P$.*

Тогда существует слово x длины n и сложности $k + O(\log n)$, для которой оба множества P_x и P_x^A являются $C(P) + \sqrt{n \log n}$ -близкими к P .

Замечание 2. В [30] говорится лишь про близость P к P_x^A . Однако в [27] было замечено, что доказательство из [30] также показывает близость P к P_x .

2.7 Сильные статистики

Как мы видели в Разделе 2.5, существует некоторый набор стандартных моделей, которые хороши с точки зрения сложности и дефекта оптимальности (а значит, благодаря Теореме 2.8, и с точки зрения дефекта случайности), но которые не кажутся адекватными объяснениями с интуитивной точки зрения. Чтобы можно было забраковать такие модели по какому-то формальному признаку, в [25] было предложено добавить ещё один параметр качества гипотезы. Перед тем как его описывать напомним понятие *тотальной условной сложности*, которое неформально определяется, как длина кратчайшей программы, которая переводит одно слово в другое, и при этом останавливается на любом входе.

Определение 2. Пусть $D : \{0, 1\}^* \times \{0, 1\}^* \rightarrow \{0, 1\}^*$ — некоторая частично вычислимая функция (способ описания). Определим

$$\text{CT}_D(x|y) := \min\{|p| \mid D(p, y) = x \text{ и величина } D(p, y') \text{ определена для любого } y'\}.$$

(При этом, сама функция D не обязана быть определена всюду.)

Существует такой универсальный способ описания D , что величина CT_D минимальна с точностью до аддитивной константы [1, 19]. Зафиксировав такой универсальный способ описания, определим *тотальную условную сложность* слова x относительно слова y как $\text{CT}_D(x|y)$. Опуская индекс D , эту величину обозначают через $\text{CT}(x|y)$.

Величина $\text{CT}(x|y)$ может быть гораздо больше чем $C(x|y)$ — см., например, [24].

Определение тотальной условной сложности переносится на произвольные конечные объекты — это делается таким же образом, как и для обычной условной сложности (см. Раздел 2.1)

Если для слов x и y величины $\text{CT}(x|y)$ и $\text{CT}(y|x)$ малы, то их профили P_x и P_y близки друг к другу.

Предложение 2.14 ([27]). Пусть для слов x и y выполнено $\text{CT}(x|y) < \varepsilon$ и $\text{CT}(y|x) < \varepsilon$. Тогда профили этих слов $O(\varepsilon)$ -близки друг к другу.

Доказательство. Пусть пара (a, b) принадлежит профилю P_x , т.е. существует такое множество $A \ni x$, что $C(A) \leq a$ и $\log |A| \leq b$. Рассмотрим тотальную программу длины не больше ε , которая переводит x в y . Рассмотрим множество $\{p(z) \mid z \in A\} \ni y$. Его мощность не превосходит мощность A , а его сложность не больше $a + O(\varepsilon)$. $\square$

Пусть слово x принадлежит множеству A . Н.К. Верещагин предложил рассмотреть новый параметр, измеряющий качество множества A , как объяснения для x , а именно $\text{CT}(A|x)$. Для “хороших” объяснений эта величина должна быть мала. Множества $A \ni x$, для которых $\text{CT}(A|x) \approx 0$ называются *сильными гипотезами* или *сильными моделями* для слова x .

Определение 3. Конечное множество A , содержащее слово x , называется ε -сильной моделью (или статистикой) для x , если $\text{CT}(A|x) < \varepsilon$.

Пример 3. Пусть x — произвольное слово длины n , а y — какой-нибудь префикс x . Тогда множество всех слов длины n , начинающихся на y , будет $O(\log n)$ -сильной статистикой для x .

Для сильных моделей существует аналог Предложения 2.5, который доказывается таким же образом.

Предложение 2.15 ([25]). Пусть конечное множество A является ε -сильной моделью для слова x . Тогда для любого числа $i \leq \log |A|$ существует такая $\varepsilon + O(\log i)$ -сильная модель A' для x , что $|A'| \leq |A|/2^i$ и $C(A') \leq C(A) + i + O(\log i)$.

Введем понятие *сильного профиля* слова x , как множество параметров всех сильных статистик для x .

Определение 4. Множество пар целых чисел

$$P_x(\varepsilon) := \{(m, l) \mid \exists A \ni x : \text{CT}(x|A) < \varepsilon, C(A) \leq m \text{ и } \log |A| \leq l\}$$

называется ε -*сильным профилем* слова x .

Очевидно, для всех слов длины n и для любого ε выполнены включения:

$$P_x(\varepsilon) \subseteq P_x = P_x(n + O(1)).$$

Те слова x , для которых профиль P_x близок к сильному профилю $P_x(\varepsilon)$ называются *нормальными*.

Определение 5. Слово x называется ε, δ -нормальным, если множество P_x содержится в δ -окрестности множества $P_x(\varepsilon)$, т.е. из того, что пара (a, b) принадлежит P_x следует, что пара $(a + \delta, b + \delta)$ принадлежит $P_x(\varepsilon)$.

Слова, не являющиеся ε, δ -нормальными называются ε, δ -*странными*. В [25] Н. К. Верещагин показал, что существуют странные слова с линейными параметрами.

Теорема 2.16 ([25]). Пусть натуральные числа k, n, ε удовлетворяют следующим неравенствам

$$O(1) \leq \varepsilon \leq k \leq n.$$

Тогда существует такое слово x длины n и сложности $k + O(\log n)$, что множества P_x и $P_x(\varepsilon)$ являются $O(\log n)$ -близкими к множествам, изображенным на Рис. 2.2. (Множество P_x расположено справа от пунктирной линии. Множество $P_x(\varepsilon)$ расположено справа от сплошной линии. Разница между этими множествами изображена на рисунке в виде параллелограмма.)

Сильные статистики с малым дефектом оптимальности обладают некоторыми интересными свойствами. Такие модели называются сильными достаточными статистиками. Сформулируем более точно, что означает последнее.

Определение 6. Конечное множество $A \ni x$ называется ε -*достаточной статистикой* для x , если $\delta(x, A) \leq \varepsilon$.

Достаточные статистики для слова x обладают следующим неформальным свойством: в них сохранена вся “существенная информация” об x , но удален некоторый “шум”. Оказывается, если множество A является сильной и достаточной статистикой для x , то между профилями P_x и $P_{[A]}$ существует прямая связь (см. Рисунок 2.7).

Теорема 2.17 ([25]). Пусть x — это слово длины n , A — ε -сильная и ε -достаточная статистика для x . Тогда для всех $b \geq \log |A|$ выполнено

$$(a, b) \in P_x \Leftrightarrow (a + O(\varepsilon + \log n), b - \log |A| + O(\varepsilon + \log n)) \in P_{[A]}$$

и для всех $b \leq \log |A|$ выполнено $(a, b) \in P_x \Leftrightarrow a + b \geq C(x) - O(\log n)$.

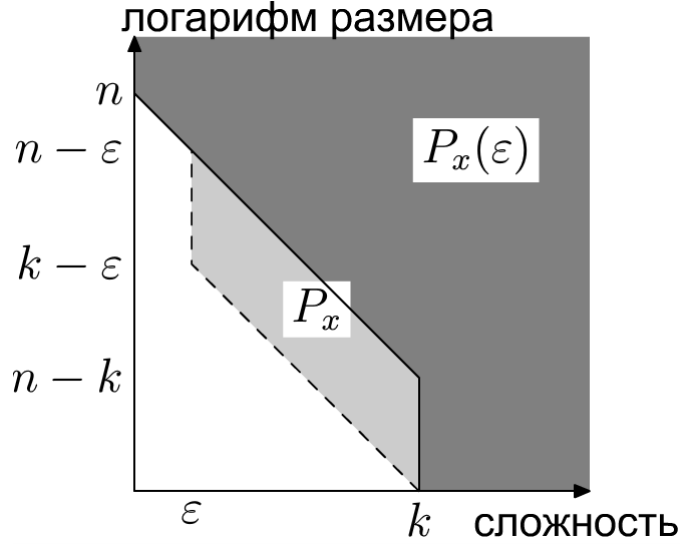


Рис. 2.2: Множества P_x и $P_x(\varepsilon)$ для странных слов из Теоремы 2.16 (с логарифмической точностью). Множество P_x расположено справа от пунктирной линии. Множество $P_x(\varepsilon)$ расположено справа от сплошной линии.

(Профиль множества A определяется как профиль слова $[A]$, где $[A]$ есть образ A при вычислимой биекции между конечными множествами слов и словами.)

Кроме достаточности важным свойством моделей является свойство *минимальности*.

Определение 7. Модель $A \ni x$ называется $(\delta, \varkappa)$ -минимальной, если не существует такого множества $B \ni x$, что $C(B) \leq C(A) - \delta$ и $\delta(x, B) \leq \delta(x, A) + \varkappa$. Чем меньше δ и больше $\varkappa$, тем сильнее свойство $(\delta, \varkappa)$ -минимальности.

Теорема 2.18 ([25]). Для некоторого $\varkappa = O(\log n)$ верно следующее. Пусть A и B являются ε -достаточными статистиками для слова x длины n . Предположим также, что A является $(\delta, \varepsilon + \varkappa)$ -минимальной статистикой для x . Тогда $C(A|B) = O(\delta + \log n)$. Более того, если к тому же A является ε -сильной моделью для x , то $CT(A|B) = O(\varepsilon + \delta + \log n)$.

В Главе 6 мы изучаем свойства нормальных слов. Мы доказываем следующие утверждения.

- Для любого множества пар P , удовлетворяющему требованиям к профилю из Теоремы 2.6 существует нормальное слово, чей профиль близок к P (Теорема 6.1).
- Существует нормальное слово, для которого есть сильная минимальная достаточная статистика, однако никакая стандартная статистика не является таковой — Теорема 6.3. (Отметим, что гипотеза о наличии таких слов и моделей являлась мотивацией для создания теории сильных статистик.)

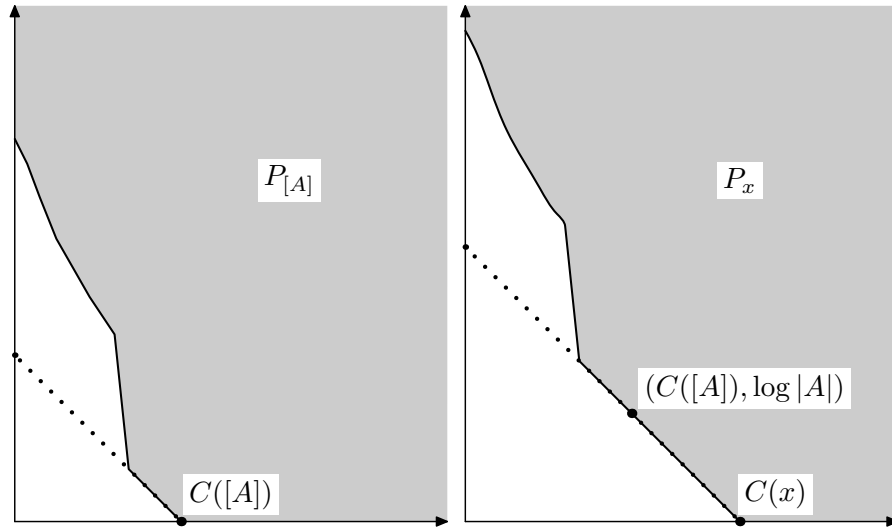


Рис. 2.3: Профили $P_{[A]}$ и P_x для множества A и слова x из Теоремы 2.17

- Нормальные статистики обладают свойством *наследственности*: любая сильная минимальная достаточная статистика для нормального слова x сама является нормальной (Теорема 6.5).

Глава 3

Антистохастические слова

В этой главе мы изучаем слова, чей профиль близок к множеству $P_{\min}$ на рисунке 2.1.

3.1 Существование антистохастических слов

Определение 8. Слово x длины n и сложности k называется ε -антистохастическим, если для всех пар $(m, l) \in P_x$ выполнено $m > k - \varepsilon$ или $m + l > n - \varepsilon$ (т.е., если P_x достаточно близок к множеству $P_{\min}$, см. Рисунок 3.1).

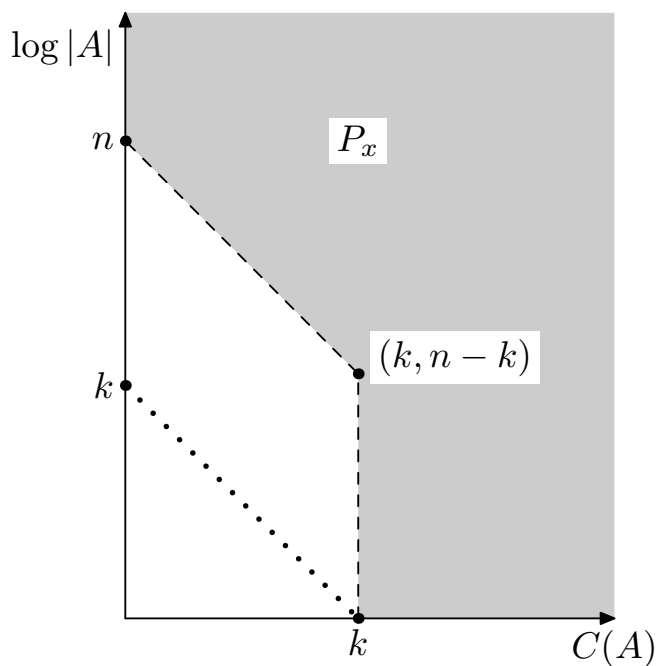


Рис. 3.1: Профиль ε -антистохастического слова x при маленьком ε близок к множеству $P_{\min}$.

Согласно Теореме 2.6 антистохастические слова существуют. Более точно, у Теоремы 2.6 имеется следующее:

Следствие 3.1. Для всех n и для всех $k \leq n$ существует $O(\log n)$ -антистохастическое слово x длины n и сложности $k + O(\log n)$.

Это следствие можно доказать гораздо проще, чем общее утверждение Теоремы 2.6, поэтому мы его докажем непосредственно.

Доказательство. Во-первых, сформулируем достаточное условие антистохастичности.

Лемма 3.2. *Если профиль слова x длины n и сложности k не содержит пару $(k - \varepsilon, n - k)$, тогда x является $\varepsilon + O(\log n)$ -антистохастичным.*

Заметим, что условие этой леммы является частным случаем определения ε -антистохастичности. Таким образом, Лемма 3.2 может рассматриваться как эквивалентное (с логарифмической точностью) определение ε -антистохастичности.

Доказательство Леммы 3.2. Пусть для слова x выполнены условия леммы. Предположим, что пара (m, l) содержится в профиле x . Мы покажем, что $m > k - \varepsilon$ или $m + l > n - \varepsilon - O(\log n)$. Предположим, что $m \leq k - \varepsilon$ и следовательно $l > n - k$. Докажем, что тогда $m + l > n - \varepsilon$ (с логарифмической точностью). Согласно третьему свойству профиля (из Раздела 2.3) пара

$$(m + (l - (n - k)) + O(\log n), n - k)$$

также содержится в профиле. Таким образом, мы получаем, что

$$m + l - (n - k) + O(\log n) > k - \varepsilon$$

и, следовательно,

$$m + l > n - \varepsilon - O(\log n).$$

□

Вернемся к доказательству Следствия 3.1. Возьмём $k' = k + D$, где D — некоторая константа, точное значение которой мы определим позже. Рассмотрим семейство $\mathcal{A}$ состоящее из всех конечных множеств A , чья сложность не превосходит k' , а логарифм размера не превосходит $n - k'$. Количество таких множеств меньше чем $2^{k'}$ (все такие множества имеют описание меньшей длины чем k'), а значит, общее количество слов во всех этих множествах меньше чем $2^{k'} 2^{n - k'} = 2^n$. Следовательно, существует слово длины n , которое не принадлежит никакому множеству из $\mathcal{A}$. Возьмём в качестве x первое в лексикографическом порядке такое слово.

Покажем, что сложность x равняется $k + O(\log n)$. Ясно, что она не меньше чем $k' - O(1)$, так как по построению синглетон $\{x\}$ имеет сложность не меньше k' . Определим D таким образом, чтобы сложность x была не меньше $k = k' - D$. С другой стороны, сложность x не больше $\log |\mathcal{A}| + O(\log n) \leq k' + O(\log n)$. В самом деле, семейство $\mathcal{A}$ может быть описано с помощью k', n и $|\mathcal{A}|$, так как можно перечислять $\mathcal{A}$ до того момента, как будут получены $|\mathcal{A}|$ множеств. Сложность $|\mathcal{A}|$ ограничена $\log |\mathcal{A}| + O(1)$, а сложности k' и n ограничены $O(\log n)$.

По построению x удовлетворяет условию Леммы 3.2 для k равного сложности x и $\varepsilon = O(\log n)$. Следовательно, x является $O(\log n)$ -антистохастичным словом. □

3.2 Антистохастические слова и числа Ω_i

Для того, чтобы доказать голографичность антистохастических слов, нам потребуется некоторая техника, которую мы изложим в этом разделе. Напомним, что число Ω_i обозначает количество слов, чья сложность не превосходит i .

Каждое антистохастическое слово x сложности $k < |x| - O(\log |x|)$ содержит ту же самую информацию, что и Ω_k :

Лемма 3.3. *Для некоторой константы c выполнено следующее.*

Любое ε -антистохастическое слово x длины n и сложности $k < n - \varepsilon - c \log n$ является $\varepsilon + c \log n$ -информационно эквивалентным числу Ω_k .

Связь между нестохастическими словами с числами Ω была найдена ещё в [14].

Доказательство Леммы 3.3. Нам нужно показать, что $C(\Omega_k|x)$ и $C(x|\Omega_k)$ малы.

Покажем это для первой величины.

Зафиксируем алгоритм, который на входе k перечисляет все слова сложности не больше k . Обозначим через N количество слов, которые появятся после x в этом перечислении (если x окажется последним словом, то $N = 0$).

Зная x , k и N , можно найти Ω_k подождая пока появятся N слов после x . Если $N = 0$, то утверждение $C(\Omega_k|x) = O(\log k)$ очевидно, поэтому далее мы будем предполагать, что $N > 0$. Обозначим $l = \lfloor \log N \rfloor$. Мы утверждаем, что $l \leq \varepsilon + O(\log n)$ потому что x является ε -антистохастическим. Разделим множество всех перечисляемых слов в порции размера 2^l . Последняя порция может быть не полной, но тогда x в неё не входит, потому что после x идёт ещё $N \geq 2^l$ элементов. Каждая полная порция может быть описана своим порядковым номером и k . Общее число полных порций меньше чем $O(2^k/2^l)$. Таким образом, профиль P_x содержит пару $(k - l + O(\log k), l)$. Из антистохастичности x получается, что $k - l + O(\log k) \geq k - \varepsilon$ или $k - l + O(\log k) + l \geq n - \varepsilon$. Первое неравенство означает, что $l \leq \varepsilon + O(\log k)$. Второе неравенство невозможно при достаточно большой константе c .

Как мы видим, для того, чтобы получить Ω_k из x требуется $\varepsilon + O(\log n)$ бит, так как N может быть описано с помощью $\log N = l$ бит, и k может быть описано с помощью $O(\log k)$ бит.

Теперь покажем, что $C(\Omega_k|x) < \varepsilon + O(\log n)$. Для этого нужно воспользоваться коммутативностью информации.

$$C(x) + C(\Omega_k|x) = C(x|\Omega_k) + C(\Omega_k) + O(\log k).$$

Слова x и Ω_k имеют одну и ту же сложность с логарифмической точностью, поэтому $C(\Omega_k|x) = C(x|\Omega_k) + O(\log n)$. $\square$

Замечание 3. Из этой леммы следует, что существует не более $2^{\varepsilon + O(\log n)}$ слов сложности k и длины n , являющихся ε -антистохастическими. В самом деле, мы получили, что каждое такое слово удовлетворяет неравенству

$$C(x|\Omega_k) \leq \varepsilon + O(\log n).$$

3.3 Голографичность антистохастичных слов

Перед тем как формулировать основной результат этой главы (Теорему 3.4), рассмотрим её частный случай в качестве примера. Докажем, что каждое $O(\log n)$ -антистохастическое слово x длины n и сложности k можно восстановить по его первым k битам с использованием $O(\log n)$ дополнительных битов. В самом деле, рассмотрим множество A , состоящее из слов длины n , у которых первые k битов такие же как у x . Сложность A не превосходит $k + O(\log n)$. С другой стороны, профиль x содержит пару $(C(A), n - k)$ (потому что $\log |A| = n - k$). Так как x является $O(\log n)$ -антистохастичным, получаем, что $C(A) \geq k - O(\log n)$ (или $C(A) + (n - k) \geq nO(\log n)$, но неравенство сводится к предыдущему). Таким образом, $C(A) = k + O(\log n)$. Учитывая, что $C(A|x) = O(\log n)$, и используя коммутативность информации получаем, что $C(x|A)$ тоже равняется $O(\log n)$.

Тот же аргумент работает, если вместо первых k битов взять любое простое k -элементное подмножество индексов $I \subset \{1, \dots, n\}$: если $C(I) = O(\log n)$, то $C(x|x_I) = O(\log n)$. Здесь x_I обозначает слово, которое получается из x путём замены каждого символа, номер которого не входит в I , на пустой символ (некий фиксированный символ, отличный от 0 и 1). Таким образом, x_I содержит информацию об I и о битах x в I -позициях.

Оказывается, тот же самый результат верен для *каждого* k -элементного подмножества индексов, даже если это подмножество сложно устроено: $C(x|x_I) = O(\log n)$. Следующая теорема утверждает даже более общий результат.

Теорема 3.4. Пусть x является ε -антистохастическим словом длины n и сложности k . Предположим, что множество A содержит x и $|A| \leq 2^{n-k}$. Тогда $C(x|A) \leq 2\varepsilon + O(\log C(A) + \log n)$.

Неформально, это теорема утверждает, что любых k битов информации об x , которые ограничивают x неким множеством размера 2^{n-k} , достаточно для восстановления x . Количество дополнительных битов, нужных для восстановления x , зависит от сложности A , которая может быть очень большой, но эта зависимость логарифмическая.

Например, если I — это k -элементное множество индексов, и A есть множество всех слов длины n , которые совпадают с x на I , тогда сложность A равняется $O(n)$ и, следовательно, $C(x|A) \leq 2\varepsilon + O(\log n)$.

Доказательство. Мы можем считать, что $k < n - \varepsilon - c \log n$, где c — константа из Леммы 3.3. В самом деле, в противном случае A настолько мало ($n - k \leq \varepsilon + c \log n$), что x может быть восстановлен по своему индексу в A , длина которого составляет $\varepsilon + c \log n$ битов.

Таким образом (при $k < n - \varepsilon - c \log n$), для x выполнена Лемма 3.3, т.е. $C(\Omega_k|x)$ и $C(x|\Omega_k)$ не превосходят $\varepsilon + O(\log n)$.

Во всех неравенствах ниже мы опускаем аддитивные слагаемые порядка $O(\log C(A) + \log n)$. Но мы, однако, оставляем слагаемые с ε (мы не требуем, чтобы ε было мало, хоть это и наиболее интересный случай).

Изложим план доказательства. Будем различать два случая: A является либо “достаточно нестохастичным”, либо “достаточно стохастичным” множеством. В первом случае A содержит информацию об Ω_k , по которой можно восстановить x .

Во втором случае A содержится в некотором простом семействе множеств $\mathcal{A}$. Мы рассматриваем множество всех y , которые покрываются большим количеством

множеств из $\mathcal{A}$, как модель для x и, используя антистохастичность x , оцениваем значения параметров этой модели.

Теперь опишем обе части доказательства более подробно. Запустим алгоритм, перечисляющий все конечные множества сложности не выше $C(A)$. Рассмотрим число $\Omega_{C(A)}$ — количество множеств сложности не выше $C(A)$. Обозначим через N номер A в этом перечислении (таким образом, $N \leq \Omega_{C(A)}$). Пусть m есть количество общих первых битов в двоичной записи у N и у $\Omega_{C(A)}$, а l — число оставшихся битов. Таким образом, $N = a2^l + b$ и $\Omega_{C(A)} = a2^l + c$ для некоторых натуральных $a < 2^m$ и $b \leq c < 2^l$. Для $l > 0$ мы можем оценить b и c более точно: $b < 2^{l-1} \leq c < 2^l$. Заметим, что $l + m$ равно длине $\Omega_{C(A)}$ в двоичной записи, то есть, $C(A) + O(1)$. Теперь рассмотрим два упомянутых случая.

Случай 1: $m \geq k$. В этом случае мы используем неравенство $C(x|\Omega_k) \leq \varepsilon$. (Напомним, что мы опускаем слагаемые порядка $O(\log C(A) + \log n)$ здесь и далее.) Число Ω_k может быть получено из Ω_m согласно Предложению 2.11 так как $m \geq k$, а Ω_m может быть получено из m первых битов $\Omega_{C(A)}$. Наконец, m первых битов $\Omega_{C(A)}$ восстанавливаются по A , как m первых битов номера N у A в перечислении всех слов сложности не больше $C(A)$. Таким образом, сложность x при известном A ограничена сверху $\varepsilon + O(\log C(A) + \log n)$.

Случай 2: $m < k$. Для рассмотрения этого случая нам понадобится дополнительная конструкция.

Лемма 3.5. *Пара $(m, l + n - k - C(A|x) + \varepsilon)$ принадлежит P_x .*

Как обычно, мы опускаем члены порядка $O(\log C(A) + \log n)$, которые должны были быть в формулировке этой леммы.

Доказательство Леммы 3.5. Мы построим множество $B \ni x$ сложности m , чей логарифм размера равен $l + n - k - C(A|x) + \varepsilon$, в два шага.

Первый шаг. Наша цель: построить семейство множеств $\mathcal{A} \ni A$, для которого $C(A) \leq m$, $C(A|x) \leq \varepsilon$ и $|\mathcal{A}| \leq 2^l$. Для этого разделим все слова сложности не больше $C(A)$ на порции размера 2^{l-1} (или размера 1, если $l = 0$) в порядке их перечисления. Последняя порция может быть неполной, однако, в этом случае A содержится в предыдущей (полной) порции, как это следует из определения m как длины общего префикса $\Omega_{C(A)}$ и N .

Определим $\mathcal{A}$ как семейство тех конечных множеств мощности не более 2^{nk} , которые принадлежат порции, содержащей A . По построению $|\mathcal{A}| \leq 2^l$. Так как $\mathcal{A}$ может быть получено из a (общих первых битов N и $\Omega_{C(A)}$), получаем, что $C(\mathcal{A}) \leq m$. Чтобы доказать неравенство $C(\mathcal{A}|x) \leq \varepsilon$, достаточно показать, что $C(a|x) \leq \varepsilon$. Это следует из того, что $C(\Omega_k|x) \leq \varepsilon$, а из Ω_k может быть получено Ω_m , а значит и число a как первые m битов $\Omega_{C(A)}$ (Предложение 2.11).

Второй шаг. Мы утверждаем, что x содержится, по крайней мере, в $2^{C(A|x)-\varepsilon}$ множествах из $\mathcal{A}$. В самом деле, предположим, что x попадает в K из них. Если мы знаем x , то нам требуются дополнительные $C(\mathcal{A}|x) \leq \varepsilon$ битов, чтобы описать $\mathcal{A}$ плюс $\log K$ чтобы описать A по его номеру в списке элементов $\mathcal{A}$ содержащих x . Таким образом, $C(A|x) \leq \log K + \varepsilon$.

Пусть B есть множество всех слов, которые содержатся, по крайней мере, в $2^{C(A|x)-\varepsilon}$ множествах из $\mathcal{A}$. Как было показано, x принадлежит B . Так как B может быть получено из $\mathcal{A}$, то $C(B) \leq m$. Для завершения доказательства Леммы 3.5, осталось оценить мощность B . Общее количество слов во всех множествах из $\mathcal{A}$

не больше $2^l \cdot 2^{n-k}$, и каждый элемент из B покрыт не менее $2^{C(A|x)-\varepsilon}$ раз, таким образом B содержит не более $2^{l+n-k-C(A|x)+\varepsilon}$ слов. $\square$

Так как x является ε -антистохастичным, то из Леммы 3.5 следует, что либо $m \geq k - \varepsilon$, либо $m + l + n - k - C(A|x) + \varepsilon \geq n - \varepsilon$. Если $m \geq k - \varepsilon$, то мы можем просто повторить аргументы из Случая 1 и показать, что $C(x|A) \leq 2\varepsilon$.

В случае $m + l + n - k - C(A|x) + \varepsilon \geq n - \varepsilon$ напомним, что $m + l = C(A)$, а значит $C(A)C(A|x) \geq k2\varepsilon$. В силу коммутативности информации, это неравенство можно переписать, как $C(x)C(x|A) \geq k2\varepsilon$, то есть, $C(x|A) \leq 2\varepsilon$ (вспомним, что k — это сложность x), что и требовалось доказать. $\square$

Замечание 4. Заметим, что каждое слово, удовлетворяющее заключению Теоремы 3.4 является δ -антистохастичным для $\delta \approx 2\varepsilon$. В самом деле, если у x длина n , сложность k , и при этом x не является δ -антистохастичным для некоторого δ , тогда x принадлежит некоторому множеству A , у которого 2^{n-k} элементов и чья сложность меньше чем $k - \delta + O(\log n)$ (Лемма 3.2). Тогда величина $C(x|A)$ будет большой, так как

$$k = C(x) \leq C(x|A) + C(A) + O(\log n) \leq C(x|A) + k - \delta + O(\log n)$$

и, следовательно, $C(x|A) \geq \delta - O(\log n)$, тогда как заключение Теоремы 3.4 утверждает, что $C(x|A) \leq 2\varepsilon + O(\log C(A) + \log n)$.

3.4 Антистохастические слова и декодирование списком

Теорема 3.4 влечёт существование хороших кодов, исправляющих ошибки. Мы не можем использовать антистохастические слова непосредственно как кодовые слова, потому что их очень мало. Вместо этого мы рассмотрим более слабое свойство, которым обладают все антистохастические слова, а потом покажем, что слов с таким свойством должно быть много, а значит, они могут быть использованы как кодовые слова.

Определение 9. Слово x длины n называется (ε, k) -голографичным, если для каждого k -элементного множества индексов $I \subset \{1, \dots, n\}$ выполняется неравенство $C(x|_{x_I}) < \varepsilon$.

Теорема 3.6. Для всех n и для всех $k \leq n$ существует по крайней мере 2^k слов длины n , являющихся $(O(\log n), k)$ -голографичными.

Доказательство. Согласно Следствию 3.1 и Теореме 3.4 для всех n и $k \leq n$ существует $(O(\log n), k)$ -голографичное слово x длины n и сложности k (с точностью $O(\log n)$). Из этого следует, что голографичных слов много. В самом деле, множество $(O(\log n), k)$ -голографичных слов длины n может быть идентифицировано по n и k . Более точно, по данным n и k можно перечислять все $(O(\log n), k)$ -голографичные слова, а следовательно, x может быть идентифицирован по k, n и номеру в этом перечислении. Сложность x не меньше $k - O(\log n)$, таким образом длина его номера должна быть не меньше $k - O(\log n)$, таким образом, существует $2^{k-O(\log n)}$ голографичных слов.

Мы, однако, должны доказать несколько более сильное утверждение: что существует 2^k голографичных слов, а не $2^{k-O(\log n)}$. Для этого возьмём некоторое k' , равное $k+O(\log n)$, и получим 2^k слов являющихся $(O(\log n), k')$ -голографичными. Разница между k и k' уходит в $O(\log n)$, так как первые $k' - k$ удалённых битов могут рассматриваться как добавка логарифмического размера. $\square$

С помощью Теоремы 3.6 можно определить код, который обладает декодированием стираний небольшим списком. В самом деле, рассмотрим 2^k слов длины n , являющихся $(O(\log n), k)$ -голографичными, как кодовые слова. Этот код позволяет исправлять $n - k$ стираний списком размера $2^{O(\log n)} = \text{poly}(n)$. Действительно, предположим, что противник стёр $n - k$ битов кодового слова x , т.е. от x осталось только x_I для некоторого множества I из k индексов. Тогда x может быть получено из оставшегося x_I с помощью программы длины $O(\log n)$. Применяя все такие программы к x_I , мы получим список размера $\text{poly}(n)$, содержащий x .

Хотя существование кода с такими параметрами может быть доказано с помощью вероятностного метода [15, Теорема 10.9 на стр. 258], связь антистохастических слов с теорией кодирования выглядит неожиданной. На самом деле, можно доказать более общее утверждение.

Теорема 3.7. *Пусть k и n — натуральные числа, при этом $k \leq n$. Пусть $\mathcal{A}$ — разрешимое семейство, состоящее из множеств слов длины n и каждое множество из семейства имеет мощность 2^{n-k} . Пусть мощность $\mathcal{A}$ ограничена $2^{\text{poly}(n)}$. Тогда существует такое множество S размера не меньше $2^{k-O(\log n)}$, что каждое $A \in \mathcal{A}$ содержит не более $2^{O(\log n)}$ слов из S .*

В предыдущем случае (для результата о декодировании списком) семейство $\mathcal{A}$ состояло из множеств, которые получались из слов фиксации k координат в некотором смысле.

Доказательство. Мы используем ту же идею, что и в доказательстве Теоремы 3.6. Мы можем предположить, не умаляя общности, что объединение множеств в $\mathcal{A}$ содержит все слова длины n — мы можем добавить 2^k множеств к $\mathcal{A}$. Это можно сделать таким образом, чтобы семейство $\mathcal{A}$ осталось разрешимым, и его размер оставался равным $2^{\text{poly}(n)}$.

Положим x равным какому-нибудь $O(\log n)$ -антистохастическому слову длины n и сложности $k+O(\log n)$. В соответствии с нашим предположением x принадлежит некоторому множеству в $\mathcal{A}$. Семейство $\mathcal{A}$ разрешимо и не является очень большим (по условию), следовательно, для каждого $A \in \mathcal{A}$ имеем $C(A) \leq \text{poly}(n) = 2^{O(\log n)}$. По Теореме 3.4 для каждого $A \in \mathcal{A}$, содержащего x , выполнено неравенство $C(x|A) \leq D \log n$ для некоторой константы D .

Определим S как множество всех таких слов y , что $C(y|A) \leq D \log n$ для каждого $A \in \mathcal{A}$, содержащего y . В S есть не более $2^{D \log n+1}$ слов, которые принадлежат некоторому множеству $A \in \mathcal{A}$, таким образом, нам сейчас осталось только доказать, что $|S| \geq 2^{k-O(\log n)}$.

Так как $\mathcal{A}$ разрешимо, мы можем перечислять S короткой программой длины $O(\log n)$. Антистохастическое слово x принадлежит S ; с другой стороны, x может быть идентифицировано по своему порядковому номеру в перечислении S , таким образом логарифм этого порядкового номера (а значит, и логарифм размера S) не меньше чем $k - O(\log n)$. $\square$

С помощью вероятностного метода можно усилить эту теорему.

Теорема 3.8. Пусть k, n — натуральные числа, при этом $k \leq n$. Пусть $\mathcal{A}$ — семейство, состоящее из множеств размера 2^{n-k} слов длины n . Пусть мощность $\mathcal{A}$ ограничена $2^{\text{poly}(n)}$. Тогда существует такое множество S размера не меньше 2^k , что каждое $A \in \mathcal{A}$ содержит не более $O(n)$ слов из S .

Доказательство. Покажем, что случайно выбранное подмножество $\{0, 1\}^n$ размера приблизительно 2^k обладает необходимым свойством с положительной вероятностью. Более точно, мы определяем S как такое случайное множество, что каждое слово длины n содержится в S с вероятностью $\frac{1}{2^{n-k-1}}$ (независимо от других слов).

Математическое ожидание количества элементов в S равняется $2^n \cdot \frac{1}{2^{n-k-1}} = 2^{k+1}$. По неравенству Маркова $|S| < \frac{1}{2} \cdot 2^{k+1} = 2^k$ с вероятностью меньше чем $\frac{1}{2}$. Поэтому, теперь достаточно показать, что событие “существует множество в $\mathcal{A}$, содержащее более чем $O(\log n)$ слов в S ” случается с вероятностью меньше чем $\frac{1}{2}$.

Для этого достаточно показать, что для каждого $A \in \mathcal{A}$ вероятность события “ A содержит больше чем $O(n)$ элементов S ” меньше чем $\frac{1}{2} \cdot \frac{1}{|\mathcal{A}|}$.

Другими словами, у нас имеется 2^{n-k} независимых испытаний с вероятностями успеха $\frac{1}{2^{n-k-1}}$, т.е. ожидаемое количество успехов равняется 2. Нам нужно оценить хвост этого распределения, начиная с cn для некоторого c и показать, что он меньше чем $\frac{1}{2} \cdot \frac{1}{|\mathcal{A}|}$ для достаточно большого c . Этот хвост ограничен

$$|A| \cdot C_A^i 2^{-(n-k+1)i} \leq \frac{|A|^{i+1}}{i!} 2^{-(n-k+1)i} = \frac{2^{n+i-k}}{i!}$$

Эта величина меньше чем $\frac{1}{2} \frac{1}{|\mathcal{A}|} = \frac{1}{\text{poly}(n)}$ для $i = O(n)$.

□

3.5 Антистохастические слова и тотальная условная сложность

Напомним, что тотальная условная сложность $\text{CT}(a|b)$ определяется как длина кратчайшей тотальной программы p , которая переводит b в a : $\text{CT}(a|b) = \min\{|p| \mid D(p, b) = a \text{ и } D(p, y) \text{ определено для всех } y\}$.

В [24] было показано, что тотальная условная сложность может быть гораздо больше чем стандартная условная сложность. А именно, существуют такие слова x и y длины n , что $\text{CT}(x|y) \geq n$ и $\text{C}(x|y) = O(1)$. Антистохастические слова помогают обобщить этот результат (к сожалению, с меньшей точностью):

Теорема 3.9. Для всех k и n существуют такие слова $x_1, \dots, x_k$ длины n , что:

- (1) $\text{C}(x_i|x_j) = O(\log k + \log n)$ для всех i и j .
- (2) $\text{CT}(x_i|x_1 \dots x_{i-1}x_{i+1} \dots x_k) \geq n - O(\log k + \log n)$ для каждого i .

Доказательство. Пусть x есть $O(\log(kn))$ -антистохастическое слово длины kn и сложности n . Рассмотрим x как конкатенацию k слов длины n :

$$x = x_1 \dots x_k \in \{0, 1\}^n.$$

Покажем, что слова $x_1, \dots, x_k$ удовлетворяют условию теоремы.

Первое утверждение есть простое следствие антистохастичности x . Теорема 3.4 влечёт $C(x|x_j) = O(\log(kn))$ для каждого j . Разумеется, $C(x_i|x) = O(\log(kn))$ для каждого i , т.е. мы получаем, $C(x_i|x_j) = O(\log(kn))$ для каждого i и j .

Чтобы доказать второе утверждение, рассмотрим такую тотальную программу p , что $p(x_1 \dots x_{i-1}x_{i+1} \dots x_k) = x_i$. Наша цель — показать, что p является длинной. Заменим p на такую тотальную программу $\tilde{p}$, что $\tilde{p}(x_1 \dots x_{i-1}x_{i+1} \dots x_k) = x$ и $|\tilde{p}| \leq |p| + O(\log(kn))$. Рассмотрим множество

$$A := \{\tilde{p}(y) \mid y \in \{0, 1\}^{k(n-1)}\}.$$

Заметим, что A содержит антистохастическое слово x длины kn и сложности n и $\log |A| \leq k \cdot (n-1)$. По определению антистохастичности получаем, что $C(A) \geq n - O(\log(kn))$. Из конструкции A следует, что

$$C(A) \leq |\tilde{p}| + O(\log(kn)) \leq |p| + O(\log(kn)).$$

Таким образом, мы получили, что $|p| \geq n - O(\log(kn))$, т.е.,

$$CT(x_i|x_1 \dots x_{i-1}x_{i+1} \dots x_k) \geq n - O(\log(kn)).$$

□

Замечание 5. Этот пример, так же как пример из [26], показывает, что для тотальной условной сложности коммутативность информации не имеет места. В самом деле, заметим, что для любого a выполнено $CT(a) = CT(a|\Lambda) = C(a) + O(1)$. Тогда $CT(x_1) - CT(x_1|x) > CT(x) - CT(x|x_1) + n - O(\log kn)$ для слов x, x_1 из Теоремы 3.9.

Большим вопросом является выполнимость коммутативность информации для сложности с ограничением на время. Частичные ответы на этот вопрос есть в [16, 20, 21].

Пусть тотальная программа p переводит b в a . Время её работы ограничено некоторой тотально вычислимой функцией f_p . Таким образом, тотальная условная сложность может рассматриваться как вариант сложности с ограничением на время. Однако, верхняя граница на f_p может зависеть от p невычислимым образом. Таким образом $CT(b|a)$ есть довольно далекое приближение к колмогоровской сложности с ограничением на время.

Глава 4

Предсказания и алгоритмическая статистика

Алгоритмическая статистика занимается изучением хороших объяснений для некоторого слова x . Но иногда нас интересуют не сами объяснения некоторого процесса, а то, как этот процесс поведёт себя в будущем. Более конкретно: пусть некий чёрный ящик выдал слово x . Рассмотрим какое-нибудь слово y . Какая вероятность того, что этот же чёрный ящик может выдать y ? Чтобы этот вопрос имел смысл, нужно задать некоторое распределение на чёрных ящиках, точнее, на распределениях, которые эти устройства задают. При этом, мы хотим сохранить общий принцип — простые объяснения должны иметь большую вероятность, чем сложные. Чтобы задать такое распределение формально, нам потребуется ещё один инструмент из алгоритмической теории сложности — *дискретная априорная вероятность*.

4.1 Дискретная априорная вероятность

Пусть U есть некоторая вероятностная машина без входа, которая выдаёт некоторое бинарное слово и останавливается. Зафиксируем какое-нибудь слово x , и обозначим через $\mathbf{m}_U(x)$ вероятность того, что машина U выдаст x . Эта вероятность, конечно, сильно зависит от U . Оказывается, существует такая универсальная машина U , что величина $\mathbf{m}_U(x)$ не меньше чем $\mathbf{m}_V(x)$ для любой машины V с точностью до мультипликативной константы.

Теорема 4.1 ([5, 9]). *Существует такая машина U , что для любой машины V существует такая положительная константа D , что*

$$\mathbf{m}_V(x) \leq \mathbf{m}_U(x) \cdot D$$

для любого слова x .

Такие машины U называются *универсальными*. Зафиксируем какую-нибудь универсальную машину U и определим значение $\mathbf{m}(x)$ как $\mathbf{m}_U(x)$. Величину $\mathbf{m}(x)$ называют *дискретной априорной вероятностью x* .

Функцию $\mathbf{m}$ можно определить и по-другому. Вещественнозначная функция f на бинарных словах называется *перечислимой снизу полумерой*, если множество пар (x, r) , где x — бинарное слово, r — рациональное число, и $r < f(x)$, является

перечислимым, и $\sum f(x) \leq 1$. Оказывается, $\mathbf{m}$ является наибольшей (с точностью до мультипликативной константы) перечислимой полумерой [1, 19].

Так же как колмогоровскую сложность, дискретную априорную вероятность можно определить для любых конечных объектов (пар слов, конечных множеств слов, ...), зафиксировав соответствующую вычислимую биекцию.

Величина $\mathbf{m}(x)$ напрямую связана с колмогоровской сложностью x .

Теорема 4.2 ([5, 9]). *Для любого слова x выполнено равенство*

$$-\log \mathbf{m}(x) = C(x) + O(\log C(x)).$$

Замечание 6. На самом деле, эта теорема верна даже с константной точностью, если вместо простой колмогоровской сложности использовать префиксную [1, 19].

4.2 Предсказательные окрестности

4.2.1 Вероятностная предсказательная окрестность

Вернёмся к задаче предсказания. Мы будем рассматривать чёрные ящики следующего типа — каждому чёрному ящику сопоставляется некоторое множество, из которого равномерно и случайно выбирается элемент при запуске этого чёрного ящика.

Теперь рассмотрим следующий двухэтапный процесс. На первом этапе случайным образом выбирается конечное множество слов: множество A выбирается с вероятностью $\mathbf{m}(A)$. На втором этапе из этого множества выбирается случайный элемент x согласно равномерному распределению. Таким образом, каждое слово x получается с вероятностью

$$\sum_{A \ni x} \mathbf{m}(A)/|A|.$$

Нетрудно видеть, что эта вероятность равняется $\mathbf{m}(x)$ с точностью до мультипликативной константы. В самом деле, формула выше определяет перечислимую снизу полумеру от x , т.е. эта величина не превосходит $\mathbf{m}(x)$ с точностью до мультипликативной константы. С другой стороны, вся сумма не превосходит $\mathbf{m}(\{x\})/1$, что уже равняется $\mathbf{m}(x)$ (с точностью до мультипликативной константы).

Теперь рассмотрим для данных x и y следующую условную вероятность.

$$p_x(y) = \Pr[y \in A \mid \text{результатом двухэтапного процесса является } x].$$

Другими словами, по определению,

$$p_x(y) = \frac{\sum_{A \ni x, y} \mathbf{m}(A)/|A|}{\sum_{A \ni x} \mathbf{m}(A)/|A|}. \quad (4.1)$$

Как мы уже видели, знаменатель равняется $\mathbf{m}(x)$ с точностью до мультипликативной константы, т.е.

$$p_x(y) = \frac{\sum_{A \ni x, y} \mathbf{m}(A)/|A|}{\mathbf{m}(x)} \quad (4.2)$$

с точностью до мультипликативной константы. По данному слову x и некоторому пороговому значению d можно составить множество всех таких слов y , что $p_x(y) \geq 2^{-d}$. При небольшом d это множество можно рассматривать, как те слова, которые с большой вероятностью могут появиться при повторении эксперимента неизвестной природы, который однажды выдал x .

Главный результат этой главы состоит в том, что это множество слов можно описать в терминах алгоритмической статистики с логарифмической точностью. По техническим причинам нам придётся немного изменить процесс, определяющий величину $p_x(y)$. А именно, в качестве множеств A мы будем рассматривать подмножества $\{0, 1\}^n$ для некоторого n . При этом, мы ограничимся рассмотрением таких множеств, чья сложность не превосходит $\text{poly}(n)$; для некоторого достаточно большого полинома (на самом деле, $4n + O(1)$ достаточно). Т.е. мы теперь предполагаем, что суммы в (4.1) и (4.2), и в подобных формулах далее всегда ограничены множествами $A \subseteq \{0, 1\}^n$ и имеющих сложность не больше $4n + O(1)$, и рассматриваем эту модифицированную версию (4.1) как финальное определение $p_x(y)$.

Определение 10. Пусть x — бинарное слово, а d — натуральное число. Множество всех таких слов y , что $p_x(y) \geq 2^{-d}$ называется *вероятностной предсказательной d -окрестностью x* .

4.2.2 Алгоритмическая предсказательная окрестность

Пусть слово x содержится в конечном множестве A . Напомним, что величина $C(A) + \log |A| - C(x)$ называется дефектом оптимальности и обозначается $\delta(x, A)$. Чем эта величина меньше, тем множество A предпочтительней, как объяснение для x . Напомним, что дефект оптимальности неотрицателен с логарифмической точностью. Объединение элементов всех множеств A , для которых $\delta(x, A)$ мало, называется *алгоритмической предсказательной окрестностью x* .

Определение 11. Пусть x — бинарное слово длины n , и d есть некоторое натуральное число. Объединение всех таких конечных множеств $A \subseteq \{0, 1\}^n$, что $x \in A$ и $\delta(x, A) \leq d$ называется алгоритмической предсказательной d -окрестностью x .

Очевидно, что при увеличении d алгоритмическая предсказательная d -окрестность также увеличивается. Она становится тривиальной (совпадающей с $\{0, 1\}^n$) когда $d = n$ (при таком d множество $\{0, 1\}^n$ будет одним из множеств в объединении).

Пример 4. Если $x = 0 \dots 0$ (слово, состоящее из n нулей), тогда x' принадлежит d -окрестности x , если и только если $C(x') \lesssim d$.

Пример 5. Если x случайное слово длины n (т.е. $C(x) \approx n$) тогда d -окрестность x содержит все слова длины n , если d больше некоторого значения порядка $O(\log n)$.

Если x есть результат работы некоторого устройства, то разумно предположить, что при следующих запусках этого же устройства будут появляться элементы из алгоритмической предсказательной окрестности x . Если распределение “чёрных ящиков” устроено так, как мы это описали в предыдущем пункте, то наше предположение будет истинным. Точнее, верно следующее утверждение.

Теорема 4.3. (а) Для каждого слова x длины n и для каждого d алгоритмическая предсказательная d -окрестность x содержится в вероятностной предсказательной $d + O(\log n)$ -окрестности x .

(б) Для каждой слова x длины n и для каждого d вероятностная предсказательная d -окрестность x содержится в алгоритмической предсказательной $d + O(\log n)$ -окрестности x .

Следующий раздел посвящён доказательству этой теоремы. Далее мы показываем, как её можно обобщить.

4.3 Доказательство Теоремы 4.3

Доказательство Теоремы 4.3 (а). Этот пункт доказывается несложно. Пусть некоторое слово y принадлежит алгоритмической предсказательной d -окрестности x , т.е. существует такое множество A , содержащее x и y , что $C(A) + \log |A| \leq C(x) + d$. Не умаляя общности, мы можем предположить, что $d \leq 2n$, иначе каждое слово длины n принадлежит вероятностной предсказательной d -окрестности x (возьмём $A = \{0, 1\}^n$). Тогда неравенство для $C(A) + \log |A|$ влечёт, что сложность A не превышает $4n$, т.е. множество A включено в сумму. Это неравенство влечёт также, что

$$\frac{\mathbf{m}(A)/|A|}{\mathbf{m}(x)} \geq 2^{-d-O(\log n)}$$

(напомним, что $-\log \mathbf{m}(u)$ равняется $C(u) + O(\log C(u))$). Эта дробь является одним из слагаемых в сумме, которая определяет $p_x(y)$, т.е. y принадлежит вероятностной предсказательной $d + O(\log n)$ -окрестности x . $\square$

Перед тем, как доказывать вторую часть (б), нам нужно доказать одну техническую лемму. Она является обобщением результата из [30, Лемма 6], где было показано, что если слово принадлежит большому количеству множеств некоторой ограниченной сложности, то одно из этих множеств имеет ещё меньшую сложность.

Лемма 4.4. Пусть множества L и R состоят из конечных объектов (в частности, для каждого элемента $v \in L$ определена его колмогоровская сложность $C(v)$). Предположим, что R содержит не более 2^n элементов. Пусть G есть конечный двудольный граф, в котором L и R есть множества левых и правых вершин соответственно. Пусть у правой вершины x есть по крайней мере 2^k соседей, чья колмогоровская сложность не превосходит i .

Тогда у x есть сосед сложности не больше $i - k + O(C(G) + \log(k + i + n))$.

Здесь $C(G)$ обозначает длину кратчайшей программы, которая, получив на вход любую вершину $v \in L$, выводит список её соседей.

Доказательство. Будем перечислять все левые вершины сложности не больше i . Во время перечисления некоторые из вершин мы будем отмечать. При этом процесс выбора отмеченных вершин должен удовлетворять следующим требованиям.

- в каждый момент времени, если у правой вершины есть по крайней мере 2^k соседа из уже перечисленных вершин, то среди них есть и отмеченная;

- общее количество отмеченных вершин не превосходит $2^{i-k}p(i, k, n)$ для некоторого полинома p (который мы зададим позднее).

Предположим, что существует стратегия выбора отмечаемых вершин, удовлетворяющая этим двум требованиям, сложности не больше $C(G) + O(\log(i + k + n))$. Из этого следует, что у правой вершины x есть сосед сложности не больше

$$i - k + O(C(G) + \log(k + i + n)),$$

а именно, его отмеченный сосед (потому что отмеченный сосед может быть задан своим номером в списке всех отмеченных вершин).

Таким образом, для доказательства леммы нам достаточно доказать существование такой стратегии. Для этого рассмотрим следующую игру. В игре участвуют два игрока, которые ходят по очереди. Максимальное количество шагов равняется 2^i . Первый игрок, совершая ход, выбирает какую-нибудь левую вершину, а второй на следующем ходу отвечает, берёт ли он эту вершину или нет. Второй игрок проигрывает, если количество отмеченных вершин превышает $2^{i-k+1}(n + 1) \ln 2$, или если после некоторого его хода существует правая вершина, у которой есть по крайней мере 2^k левых соседа из отмеченных первым игроком, но среди них нет отмеченных (выбор значения $2^{i-k+1}(n + 1) \ln 2$ будет ясен из некоторой оценки далее). В противном случае второй игрок выигрывает.

Предположим сперва, что множество L левых вершин конечно (напомним, что множество правых вершин конечно из условия). Тогда наша игра есть конечная игра с полной информацией, а значит, один из игроков имеет в ней выигрышную стратегию. Мы утверждаем, что второй игрок может выиграть. Если это не так, то у первого игрока есть выигрышная стратегия. Мы получим противоречие показав, что у второго игрока есть вероятностная стратегия, которая приводит его к выигрышу с положительной вероятностью при любой стратегии первого игрока. Зафиксируем какую-нибудь стратегию первого игрока и рассмотрим следующую вероятностную стратегию второго: каждая вершина, которую отмечает первый игрок, объявляется отмеченной с вероятностью $p := 2^{-k}(n + 1) \ln 2$. Математическое ожидание количества отмеченных вершин не больше $p2^i = 2^{i-k}(n + 1) \ln 2$. Согласно неравенству Маркова, количество отмеченных вершин превышает своё математическое ожидание в два раза с вероятностью меньшей чем $\frac{1}{2}$. Таким образом, достаточно показать, что второй плохой случай (после нескольких ходов существует правая вершина y , у которой 2^k выбранных первым игроком соседей, но ни одна из них не является отмеченной вторым игроком) случается с вероятностью меньшей чем $\frac{1}{2}$.

Для этого, в свою очередь, достаточно показать, что для каждой правой вершины y вероятность такого плохого события меньше чем $\frac{1}{2}$, делённая на количество $|R|$ правых вершин.

Оценим эту вероятность. Если y имеет 2^k (или более) соседей, то второй игрок мог (по крайней мере) 2^k раз отметить какого-нибудь соседа y , и вероятность того, что всеми этими 2^k возможностями второй игрок не воспользуется, не превосходит $(1 - p)^{2^k}$. Выбор значения p гарантирует, что эта вероятность меньше чем $2^{-n-1} \leq (1/2)/|R|$. В самом деле, используя неравенство $1 - x \leq e^{-x}$, нетрудно видеть, что

$$(1 - p)^{2^k} \leq e^{\ln 2 \cdot (-n-1)} = 2^{-n-1}.$$

Мы доказали, что выигрышная стратегия существует, но не оценили её сложность.

Выигрышную стратегию можно найти из списка всех возможных стратегий перебором. Множество всех стратегий конечно, а игра задаётся по G , i и k . Следовательно, сложность первой найденной выигрышной стратегии не превосходит $C(G) + O(\log(i + k))$.

Таким образом, Лемма 4.4 доказана в том случае, когда L конечное множество. Чтобы доказать лемму в общем случае заметим, что выигрышное условие зависит только от соседей каждой правой вершины, а их в каждой игре используется не больше чем 2^i . Худшим случаем для второго игрока является следующий “модельный” граф, состоящий из 2^{2^n+i} левых вершин и 2^n правых вершин. Зафиксируем 2^n правых вершин. Из левой вершины можно провести к ним рёбра 2^{2^n} способами. В модельном графе будут по 2^i правых вершин каждого типа. Выигрышная стратегия для такого графа может быть найдена по n , i и k , и следовательно её сложность зависит от $n + i + k$ логарифмическим образом. Эта стратегия может быть транслирована для игры с изначальным графом следующим образом. Каждой вершине, выбранной первым игроком, будем сопоставлять вершину модельного графа такого же типа таким образом, чтобы разным вершинам исходного графа соответствовали разные вершины модельного графа (ограничение на число ходов в игре и запас вершин в модельном графе гарантируют возможность соблюдения этого условия). Это транслирование увеличит сложность не более чем на $C(G)$, потому что каждой вершине, выбранной первым игроком, нужно сопоставить некоторую вершину из модельного графа. $\square$

Имея ввиду будущие приложения в следующих разделах, в следствии ниже мы будем рассматривать произвольное разрешимое семейство $\mathcal{A}$ конечных множеств, хотя для доказательства Теоремы 4.3 нам нужен будет только тот случай, когда $\mathcal{A}$ содержит все конечные множества.

Следствие 4.5. Пусть $\mathcal{A}$ — разрешимое семейство конечных множеств. Пусть $x_1, \dots, x_l$ — слова длины n . Обозначим через $\mathcal{A}_m$ все множества из $\{0, 1\}^n$ принадлежащие $\mathcal{A}$ чья сложность не превосходит m . Тогда сумма

$$S := \sum_{A \in \mathcal{A}_m, x_1, \dots, x_l \in A} \frac{\mathbf{m}(A)}{|A|}$$

равняется своему максимальному слагаемому с точностью до умножения на $2^{O(\log(n+m+l))}$.

Доказательство следствия. Обозначим через M максимальное слагаемое в S . Разложим S в слагаемые следующего вида для каждого i и j :

$$\sum_{\substack{A \in \mathcal{A}_m \\ C(A)=i \\ \log |A|=j \\ x_1, \dots, x_l \in A}} \frac{\mathbf{m}(A)}{|A|}. \quad (4.3)$$

Всего есть $(m + 1)(n + 1)$ таких сумм, нам нужно показать, что каждая из них не больше $M \cdot 2^{O(\log n + m + l)}$. Иными словами, нам нужно показать, что для всех i, j существует такое множество $H \in \mathcal{A}_m$, содержащее $x_1, \dots, x_l$, что $\frac{\mathbf{m}(H)}{|H|}$ больше соответствующей суммы (4.3) с точностью до умножения на $2^{O(\log(n+m+l))}$.

Для этого зафиксируем i и j . Так как $\mathbf{m}(u) = 2^{-C(u)-O(\log C(u))}$, то сумма (4.3) равняется

$$\sum_{\substack{A \in \mathcal{A}_m^n \\ C(A)=i \\ \log |A|=j \\ x_1, \dots, x_l \in A}} 2^{-C(A)-\log |A|+O(\log(n+m))} = \sum_{\substack{A \in \mathcal{A}_m^n \\ C(A)=i \\ \log |A|=j \\ x_1, \dots, x_l \in A}} 2^{-i-j+O(\log(n+m))} \quad (4.4)$$

Все слагаемые в сумме (4.4) совпадают, а значит вся сумма (4.4) равняется $2^{-i-j+O(\log(n+m))}$, умноженное на количество таких множеств $A \in \mathcal{A}_m^n$, что $C(A) = i$, $\log |A| = j$ и $x_1, \dots, x_l \in A$. Обозначим через k целую часть бинарного логарифма этого числа.

Рассмотрим двудольный граф, чьи левые вершины есть конечные множества из $\mathcal{A}^n$ мощности не больше 2^j , а правые вершины есть наборы из l слов длины n , и левая вершина A соединена с правой вершиной $\langle x_1, \dots, x_l \rangle$ если все $x_1, \dots, x_l$ принадлежат A . Сложность этого графа равняется $O(\log(n+l+j))$, логарифм числа правых вершин равен nl . По Лемме 4.4 существует множество $H \in \mathcal{A}_m^n$ размера 2^j , чья сложность не превосходит $i - k + O(\log(i+j+k+n+l)) = i - k + O(\log(l+m+n))$, и при этом $x_1, \dots, x_l \in H$. Дробь $\frac{\mathbf{m}(H)}{|H|}$ равняется $2^{-(i-k)-j}$ с точностью до умножения на $2^{O(\log(n+m+l))}$.

Напомним, что сумма (4.4) равняется $2^k 2^{-i-j}$ с точностью до умножения на такой же множитель, что показывает, что множество H удовлетворяет нашему требованию. $\square$

Замечание 7. Рассмотрим частный случай Следствия 4.5, когда $\mathcal{A}$ есть семейство всех конечных множеств и $l = 1$. Как было показано, сумма $\sum_{A \ni x} \mathbf{m}(A)/|A|$ равняется $\mathbf{m}(x)$ с точностью до константного множителя.

По этой причине возникает предположение, что точность в Следствии 4.5 может быть улучшена.

Доказательство Теоремы 4.3 (b). Пусть слово y принадлежит вероятностной предсказательной d -окрестности x . В соответствии с (4.2), это влечёт следующее неравенство

$$\sum_{A \ni x, y} \frac{\mathbf{m}(A)}{|A|} \geq \mathbf{m}(x) 2^{-d-O(\log n)} = 2^{-d-C(x)-O(\log n)}.$$

Теперь воспользуемся Следствием 4.5 для $l = 2$, $x_1 = x$, $x_2 = y$, $m = 4n$ и семейства всех множеств в качестве $\mathcal{A}$. Согласно следствию, существует такое множество $A \ni x, y$, что $\mathbf{m}(A)/|A| = 2^{-d-C(x)-O(\log n)}$, т.е.:

$$C(A) + \log |A| - C(x) \leq d + O(\log n),$$

т.е. y принадлежит алгоритмической $d + O(\log n)$ -окрестности x . $\square$

4.4 Ограниченные классы гипотез

В этом разделе мы будем рассматривать в качестве объяснений не произвольные конечные множества, а множества из некоторого фиксированного семейства (так же, как это делалось в Разделе 2.6). Мы покажем, что Теорема 4.3 имеет аналог

для разрешимого семейства множеств $\mathcal{A}$. Будем считать, что $\mathcal{A}$ состоит из множеств строк одинаковой длины, т.е. для любого $A \in \mathcal{A}$ длины всех строк в A имеют одинаковую длину.

Сначала мы будем рассматривать такие семейства $\mathcal{A}$, что для любого слова x семейство $\mathcal{A}$ содержит синглетон $\{x\}$.

Определим вероятностную предсказательную окрестность для слова x длины n относительно семейства $\mathcal{A}$. Для этого мы опять рассмотрим двухэтапный процесс: во-первых, некоторое множество A слов длины n из $\mathcal{A}^n$ выбирается с вероятностью $\mathbf{m}(A)$. Во-вторых, элемент из A выбирается согласно равномерному распределению. Опять мы вынуждены добавить условие, что сложность выбираемых множеств не превосходит $4n + O(1)$. Число $p_{x,\mathcal{A}}(y)$ определяется как вероятность события $y \in A$ относительно события “ x есть результат двухэтапного процесса”:

$$p_{x,\mathcal{A}}(y) = \frac{\sum_{A \ni x,y} \mathbf{m}(A)/|A|}{\sum_{A \ni x} \mathbf{m}(A)/|A|}. \quad (4.5)$$

Здесь сумма берётся по всем множествам из $\mathcal{A}$, чья сложность не превосходит $4n + O(1)$.

Так же как в Разделе 4.2, знаменатель равняется $\mathbf{m}(x)$ с точностью до умножения на константу (потому что $\{x\} \in \mathcal{A}$), таким образом:

$$p_{x,\mathcal{A}}(y) = \frac{\sum_{A \ni x,y} \mathbf{m}(A)/|A|}{\mathbf{m}(x)} \quad (4.6)$$

с точностью до умножения на константу.

Определим $\mathcal{A}$ -вероятностную предсказательную d -окрестность естественным образом: слово y принадлежит этой окрестности, если $p_{x,\mathcal{A}}(y) \geq 2^{-d}$. Теперь определим $\mathcal{A}$ -алгоритмическую предсказательную d -окрестность для x как множество таких слов y , для которых существует такое множество $A \ni x, y$, содержащееся в $\mathcal{A}$, что $\delta(x, A) \leq d$.

Теперь всё готово для формулировки аналога Теоремы 4.3:

Теорема 4.6. Пусть $\mathcal{A}$ есть разрешимое семейство множеств двоичных слов, содержащее все синглетоны. Тогда:

(а) Для каждого слова x длины n и для каждого d $\mathcal{A}$ -алгоритмическая предсказательная d -окрестность x содержится в $\mathcal{A}$ -вероятностной предсказательной $d + O(\log n)$ -окрестности x .

(б) Для каждого слова x длины n и для каждого d $\mathcal{A}$ -вероятностная предсказательная d -окрестность x содержится в $\mathcal{A}$ -алгоритмической предсказательной $d + O(\log n)$ -окрестности x .

Доказательство Теоремы 4.6 (а). Доказательство похоже на доказательство Теоремы 4.3 (а). Пусть слово y принадлежит $\mathcal{A}$ -алгоритмической предсказательной d -окрестности x , т.е. существует множество $A \in \mathcal{A}$, содержащее x и такое y , что $C(A) + \log |A| \leq C(x) + d$. Если $d > 3n$, тогда утверждение тривиально. В самом деле, существует такое множество $A' \in \mathcal{A}$, которое содержит x и y , что $\delta(x, A') \leq 3n$. Для того, чтобы это доказать, мы не можем взять множество $A' = \{0, 1\}^n$ как раньше, потому что это множество может не принадлежать $\mathcal{A}$.

Однако, мы можем в качестве A' взять первое множество в $\mathcal{A}$, содержащее x и y . Сложность этого множества не больше чем $|x| + |y| \leq 2n$, а логарифм размера не превосходит n . Таким образом, $\delta(x, A') \leq 3n$. Дальнейшее доказательство этой теоремы полностью повторяют рассуждение Теоремы 4.3 (а). $\square$

Доказательство Теоремы 4.6 (b). Доказательство почти дословно повторяет доказательство Теоремы 4.3 (b). $\square$

Теперь мы сформулируем и докажем Теорему 4.6 в общем случае (для семейств $\mathcal{A}$, которые могут не содержать всех синглетонов). В случае если x принадлежит некоторому множеству из $\mathcal{A}$, определение $\mathcal{A}$ -вероятностной предсказательной окрестности остаётся тем же. В противном случае, если x не принадлежит никакому множеству из $\mathcal{A}$, слово x не может появиться в результате двухэтапного процесса, поэтому в этом случае мы определяем $\mathcal{A}$ -вероятностную предсказательную d -окрестность x как пустое множество для каждого d . Отметим, что теперь мы не можем переписать (4.5) как (4.6), потому что $\{x\}$ может не принадлежать $\mathcal{A}$.

Теперь определим $\mathcal{A}$ -алгоритмическую предсказательную окрестность. Тут есть некоторая тонкость, которую нужно иметь ввиду: может случиться, что в $\mathcal{A}$ не содержится такого множества $A \ni x$, что $\delta(x, A) \approx 0$. По этой причине мы включаем в алгоритмическую предсказательную окрестность x объединение всех таких множеств A из $\mathcal{A}$, что $\delta(x, A)$ мало настолько, насколько это возможно:

Определение 12. Пусть x есть бинарное слово длины n , d — натуральное число и $\mathcal{A}$ — семейство конечных множеств. Объединение всех таких конечных множеств в $\mathcal{A}$, что $x \in A$ и для каждого $B \in \mathcal{A}$, содержащего x , выполняется неравенство: $\delta(x, A) \leq \delta(x, B) + d$ называется $\mathcal{A}$ -алгоритмической предсказательной d -окрестностью x . (Иными словами, d -окрестность включает все множества A , для которых величина $\delta(x, A)$ превосходит минимально возможную величину не более чем на d .)

Теорема 4.7. Пусть $\mathcal{A}$ есть разрешимое семейство множеств двоичных слов. Тогда:

- (а) для каждого слова x длины n и для каждого d $\mathcal{A}$ -алгоритмическая предсказательная d -окрестность x содержится в $\mathcal{A}$ -вероятностной предсказательной $d + O(\log n)$ -окрестности x ;
- (б) для каждого слова x длины n и для каждого d $\mathcal{A}$ -вероятностная предсказательная d -окрестность x содержится в $\mathcal{A}$ -алгоритмической предсказательной $d + O(\log n)$ -окрестности x .

Заметим, что если $x \notin \mathcal{A}$, то обе алгоритмические и предсказательные окрестности пустые, и в этом случае утверждение становится тривиальным. По этой причине в доказательстве мы этот случай рассматривать не будем.

Доказательство Теоремы 4.7 (а). Пусть слово y принадлежит алгоритмической предсказательной d -окрестности x . Т.е. для некоторого множества B , содержащего x и y величина $\frac{\mathbf{m}(B)}{|B|}$ больше чем $2^{-d-O(\log n)} \cdot \max\{\frac{\mathbf{m}(A)}{|A|} \mid A : x, y \in A \in \mathcal{A}\}$. (Напомним, что $\frac{\mathbf{m}(A)}{|A|} = 2^{-C(A)-\log |A|}$ с точностью до умножения на $2^{O(\log n)}$.) Согласно Следствию 4.5 сумма $\sum_{A \ni x} \frac{\mathbf{m}(A)}{|A|}$ равняется своему наибольшему слагаемому (опять, с точностью до умножения на $2^{O(\log n)}$). Таким образом мы получили

$$\sum_{A \ni x, y} \frac{\mathbf{m}(A)}{|A|} \geq \frac{\mathbf{m}(B)}{|B|} \geq 2^{-d-O(\log n)} \sum_{A \ni x} \frac{\mathbf{m}(A)}{|A|}.$$

Это означает требуемое включение. $\square$

Доказательство Теоремы 4.7 (b). Теперь мы начинаем со слова y для которого

$$\sum_{A \ni x, y} \frac{\mathbf{m}(A)}{|A|} \geq 2^{-d} \sum_{A \ni x} \frac{\mathbf{m}(A)}{|A|}. \quad (4.7)$$

Обозначим

$$A_x := \arg \max \{ \mathbf{m}(A)/|A| \mid x \in A \in \mathcal{A} \}$$

и

$$A_{xy} := \arg \max \{ \mathbf{m}(A)/|A| \mid x, y \in A \in \mathcal{A} \}.$$

Согласно Следствию 4.5 суммы в обеих частях неравенства равняются своим наибольшим членам. Таким образом,

$$2^{-C(A_{x,y}) - \log |A_{x,y}|} \geq 2^{-d - O(\log n)} 2^{-C(A_x) - \log |A_x|},$$

что означает следующие неравенство $\delta(x, A_{x,y}) \leq \delta(x, A_x) + d + O(\log n)$. Следовательно, y принадлежит $\mathcal{A}$ -алгоритмической предсказательной $d + O(\log n)$ -окрестности x . $\square$

Глава 5

Алгоритмическая статистика для нескольких слов

До этого момента мы рассматривали только равномерные вероятностные распределения в качестве статистических гипотез. Причиной этому было Предложение 2.1: для любого распределения P и для любого слова x существует множество $A \ni x$, которое объясняет x не хуже чем P . Однако, для данных, состоящих более чем из одного слова, это утверждение (как мы вскоре убедимся) уже не будет верным.

Качество вероятностного распределения P , как объяснения для данного набора слов $\vec{x} = x_1, \dots, x_l$, может быть измерено с помощью следующих двух параметров:

- сложность $C(P)$ распределения P ;
- $-\log(P(x_1) \dots P(x_l))$ (чем меньше этот параметр, тем с большей вероятностью можно получить набор $\vec{x}$ при случайном независимом выборе l слов согласно распределению P).

Так же как раньше, мы рассматриваем только распределения с конечным носителем, для которых вероятность любого события есть рациональное число. Сложность такого распределения P определяется как сложность списка пар $\langle y, P(y) \rangle$ (отсортированных в лексикографическом порядке).

Если P является равномерным распределением на конечном множестве A , то первый параметр равняется $C(A)$, а второй равняется $-l \log |A|$. Если $l = 1$, тогда для каждой пары $\langle x, P \rangle$ существует такое конечное множество $A \ni x$, что числа $C(A), \log |A|$ не больше $C(P), -\log P(x)$ с точностью $O(\log |x|)$.

Предложение 5.1 ([29]). *Для каждого слова x длины n и для каждого распределения P существует такое множество $A \ni x$, что $C(A) \leq C(P) + O(\log n)$ и $\log |A| \leq -\log P(x) + 1$.*

(Это утверждение доказывается таким же образом, что и Предложение 2.1.)

При $l = 2$ ситуация меняется.

Пример 6. Пусть x_1 — случайное слово длины $2n$, а $x_2 = 00 \dots 0y$ слово длины $2n$, где y — случайное слово длины n , независимое от x_1 (т.е., $C(x_1, x_2) = 3n + O(1)$). Для этой пары слов есть следующее правдоподобное объяснение: слова x_1, x_2 получились независимо и случайно согласно распределению P , где половина распределения рассредоточена равномерно на всех словах длины $2n$, а вторая половина — равномерна распределена на словах длины $2n$, начинающихся на n

нулей. Сложность распределения P мала ($O(\log n)$), при этом второй параметр $-\log(P(x_1)P(x_2))$ равняется примерно $3n$. С другой стороны, не существует простого множества A , содержащего x_1 и x_2 , для которого $2 \log |A|$ близко к $3n$. В самом деле, для каждого множества A , содержащего x_1 имеем $C(A) + \log |A| \geq 2n - O(\log n)$, а значит, $2 \log |A| \geq 4n - 2C(A) - O(\log n) \gg 3n$ (последнее неравенство будет верным при маленьком $C(A)$).

Поэтому мы не будем ограничивать себя равномерными распределениями. В следующем разделе мы определяем и изучаем понятие профиля для нескольких слов. Далее мы докажем аналог утверждения о связи дефектов случайности и оптимальности (Теорема 2.8) — Теорему 5.4. Также мы покажем, как можно обобщить результаты предыдущей главы для неравномерных распределений (Теорема 5.10).

5.1 Профиль нескольких слов

Рассмотрим набор слов $x_1, \dots, x_l \in \{0, 1\}^n$, который мы будем обозначать через $\vec{x}$. Дефект оптимальности для $\vec{x}$ определяется следующей формулой

$$\delta(\vec{x}, P) = C(P) - \log(P(x_1) \dots P(x_l)) - C(\vec{x}).$$

Эта величина неотрицательна с точностью $O(\log(n + l))$, потому что по данным P и l можно описать $\vec{x}$, используя $-\log(P(x_1) \dots P(x_l)) + O(1)$ битов с помощью кодирования Шеннона-Фано.

Определение 13. Профиль $P_{\vec{x}}$ набора $\vec{x}$ определяется как множество всех таких пар $\langle a, b \rangle$ натуральных чисел, что существует такое вероятностное распределение P сложности не выше a , что $\delta(\vec{x}, P) \leq b$.

(Такое определение профиля будет нам удобней чем то, что мы рассматривали для одномерного случая (сложность, логарифм размера) по техническим причинам.) Набор слов $\vec{x}$ называется стохастическим, если существует такое простое распределение P , что $\delta(\vec{x}, P) \approx 0$; в противном случае он называется нестохастическим. (Для одномерного случая мы определяли стохастичность через дефект случайности, однако, благодаря Теореме 2.8, эти подходы совпадают.)

В многомерном случае, также как в одномерном (Теоремы 2.2, 2.3, 2.6), существуют нестохастические объекты. При этом в одномерном случае мы не можем предъявить нестохастические слова *явно*. В двумерном случае ситуация отличается: пусть x_1 — случайное слово длины n , и $x_2 = x_1$. Тогда для пары x_1, x_2 не существует простого распределения P с маленьким значением $\delta(\langle x_1, x_2 \rangle, P)$. В самом деле, для каждого вероятностного распределения P выполняется неравенство $C(P) - \log P(x_i) \geq C(x_i) = n$ для $i = 1, 2$ (с точностью $O(\log n)$). Складывая эти неравенства, получаем

$$2C(P) - \log(P(x_1)P(x_2)) \geq 2n.$$

Следовательно $\delta(\langle x_1, x_2 \rangle, P) \geq 2n - C(P) - C(x_1, x_2) = n - C(P)$. Эта величина большая, если $C(P) \ll n$.

Вообще, если у слов x_1 и x_2 много *общей информации* (т. е. $C(x_1, x_2) \ll C(x_1) + C(x_2)$), то пара $\langle x_1, x_2 \rangle$ является нестохастичной. Формально общая информация

слов x_1 и x_2 определяется как $C(x_2) - C(x_2|x_1)$. Согласно коммутативности информации эта же величина равняется $C(x_1) - C(x_1|x_2)$ с точностью $O(\log C(x_1, x_2))$.

Есть ещё неявный пример нестохастической пары слов: рассмотрим любую пару, у которой первый элемент нестохастичный. Для первого элемента этой пары хорошего объяснения нет, а значит, хорошего объяснения нет и для всей пары.

Зададимся вопросом: верно ли, что профиль пары слов x_1, x_2 определяется по $C(x_1), C(x_2), C(x_1, x_2), P_{x_1}, P_{x_2}$ и $P_{[x_1, x_2]}$? Здесь $[x_1, x_2]$ обозначает конкатенацию слов x_1 и x_2 . Отметим, что $P_{[x_1, x_2]}$ определяет одномерный профиль слова $[x_1, x_2]$, который не нужно путать с P_{x_1, x_2} — двумерным профилем пары слов x_1, x_2 .

Следующая теорема даёт отрицательный ответ на этот вопрос.

Теорема 5.2. *Для каждого n существуют такие слова x_1, x_2, y_1 и y_2 длины $2n$, что:*

- (1) Множества P_{x_1} и P_{y_1}, P_{x_2} и $P_{y_2}, P_{[x_1, x_2]}$ и $P_{[y_1, y_2]}$ являются $O(\log n)$ -близкими.
- (2) $C(x_1) = C(y_1) + O(\log n)$, $C(x_2) = C(y_2) + O(\log n)$, $C(x_1, x_2) = C(y_1, y_2) + O(\log n)$.
- (3) Тем не менее, расстояние между P_{x_1, x_2} и P_{y_1, y_2} больше чем $0.5n - O(\log n)$.

(Напомним, что два множества R и Q называются ε -близкими, если R содержится в ε -окрестности Q , и наоборот.)

Доказательство. Пример нужных нам слов взят из [12], где были даны несколько примеров пар слов с невыделяемой общей информацией. При этом важно, что в [12] приведены примеры *стохастических* пар слов с таким свойством. Одним из этих примеров мы и воспользуемся. Рассмотрим конечное поле $\mathbb{F}$ мощности 2^n и плоскость (двумерное векторное пространство) над $\mathbb{F}$. Возьмём в качестве y_1 случайную прямую на этой плоскости, а в качестве y_2 — случайную точку на этой прямой. Тогда

$$C(y_1) = 2n, C(y_2) = 2n, C(y_1, y_2) = 3n$$

(всё с логарифмической точностью). У слов y_1, y_2 общая информация равняется n . С другой стороны [12, Теорема 8] утверждает, что эта информация невыделяема:

Теорема 5.3 ([12]). *Не существует такого слова z , что $C(z) = n + O(\log n)$, $C(y_1|z) = n + O(\log n)$, $C(y_2|z) = n + O(\log n)$ (такое слово z можно было бы рассматривать как представление общей информации слов y_1 и y_2).*

Более того, для любого z выполнено

$$C(z) + C(y_1|z)/2 + \max\{C(y_1|z)/2, C(y_2|z)\} \geq 3n - O(\log n), \quad (5.1)$$

$$C(z) + C(y_2|z)/2 + \max\{C(y_2|z)/2, C(y_1|z)\} \geq 3n - O(\log n). \quad (5.2)$$

Покажем вначале, что неравенства (5.1) и (5.2) влекут

$$C(z) + C(y_1|z) + C(y_2|z) \geq \min\{4n - C(z)/3, 5n - C(z)\} - O(\log n). \quad (5.3)$$

В самом деле, если $C(y_1|z)$ и $C(y_2|z)$ отличаются друг от друга не более чем в два раза, то максимум в обоих неравенствах (5.1) и (5.2) равняется второму члену; суммируя (5.1) и (5.2) получаем

$$2C(z) + 3C(y_1|z)/2 + 3C(y_2|z)/2 \geq 6n - O(\log n),$$

что может быть переписано как

$$C(z) + C(y_1|z) + C(y_2|z) \geq 4n - C(z)/3 - O(\log n).$$

В противном случае, когда, скажем, $C(y_1|z) > 2C(y_2|z)$, максимум в неравенстве (5.1) равняется первому члену. Тогда суммируя это неравенство с $C(z) + C(y_2|z) \geq C(y_2) = 2n$ мы получаем

$$2C(z) + C(y_1|z) + C(y_2|z) \geq 5n - O(\log n),$$

что эквивалентно

$$C(z) + C(y_1|z) + C(y_2|z) \geq 5n - C(z) - O(\log n).$$

В обоих случаях мы получаем (5.3).

Из этого следует, что профиль P_{y_1, y_2} пар слов из Теоремы 5.3 обладает следующим свойством

$$\langle a, b \rangle \in P_{y_1, y_2} \Rightarrow b \geq \min\{n - a/3, 2n - a\} - O(\log n). \quad (5.4)$$

В самом деле, для каждого вероятностного распределения P выполняется неравенство $C(y|P) \leq -\log P(y) + O(1)$, а значит,

$$\delta(\langle y_1, y_2 \rangle, P) \geq C(P) + C(y_1|P) + C(y_2|P) - 3n - O(1). \quad (5.5)$$

Комбинируя неравенство (5.3) для $z = P$ и неравенство (5.5) мы получаем (5.4).

Таким образом, профиль пары y_1, y_2 не содержит пару $(1.5n, 0.5n - O(\log n))$. С другой стороны, все слова $y_1, y_2, [y_1, y_2]$ являются стохастическими, т.е. множества $P_{y_1}, P_{y_2}, P_{[y_1, y_2]}$ содержат почти все пары (a, b) (более точно, все такие пары, для которых $a, b \geq O(\log n)$).

Построим теперь пару слов x_1, x_2 , которая будет обладать теми же свойствами, однако пара $(n + O(1), O(1))$ будет внутри P_{x_1, x_2} . Возьмём x_1, x_2 случайными словами длины $2n$, у которых совпадают первые n символов: $x_1 = x^*x_1^*$, $x_2 = x^*x_2^*$ и $C(x^*x_1^*x_2^*) = 3n + O(1)$. Тогда $C(x_1) = 2n + O(1)$, $C(x_2) = 2n + O(1)$, $C(x_1x_2) = 3n + O(1)$ (аналогично y_i). И также все слова $x_1, x_2, [x_1, x_2]$ являются стохастическими. Чтобы показать, что пара $(n + O(\log n), O(\log n))$ принадлежит P_{x_1, x_2} , рассмотрим равномерное распределение P на всех словах длины $2n$, чья первая половина равняется x^* . У этого распределения такая же сложность как у x^* , таким образом, $C(P) = n + O(1)$ и следовательно, $C(P) - \log P(x_1) - \log P(x_2) = 3n + O(1) = C(x_1x_2)$. А значит, пара $(n + O(1), O(1))$ принадлежит P_{x_1, x_2} . $\square$

5.2 Многомерный дефект случайности

Понятие дефекта случайности можно обобщить для нескольких слов.

Для набора слов $\vec{x} = x_1, \dots, x_l$ и распределения P *многомерный дефект случайности* $d(\vec{x}|P)$ определяется следующим образом

$$d(\vec{x}|P) = -\log(P(x_1) \dots P(x_l)) - C(x_1, \dots, x_l|P).$$

Если $l = 1$, а P — равномерное распределение на конечном множестве, то это определение эквивалентно одномерному случаю (который мы рассматривали в Главе 2).

Множество всех таких пар (a, b) , что существует такое распределение P сложности не больше a , что $d(\vec{x} | P) \leq b$, называется *профилем стохастичности* $\vec{x}$ и обозначается $Q_{\vec{x}}$. Во избежание путаницы, профиль $P_{\vec{x}}$ из предыдущего раздела мы будем называть *профилем оптимальности*.

Теорема 2.8 и Предложение 2.7 вместе означают, что для одномерного случая профили стохастичности и оптимальности совпадают с логарифмической точностью.

Оказывается, то же самое верно для многомерного случая.

Теорема 5.4. *Для каждого набора $\vec{x} = x_1, \dots, x_l$ слов длины n расстояние между множествами $P_{\vec{x}}$ и $Q_{\vec{x}}$ не превосходит $O(\log(n + l))$.*

Доказательство Теоремы 5.4. Доказательство похоже на доказательство Предложения 2.7 и Теоремы 2.8 в [30]. Во-первых, заметим, что для каждого распределения P выполнено

$$d(\vec{x} | P) \leq \delta(\vec{x}, P) + O(\log(n + l)). \text{ В самом деле:}$$

$$\begin{aligned} d(\vec{x} | P) &= -\log(P(x_1) \dots P(x_n)) - C(\vec{x} | P) \\ &\leq -\log(P(x_1) \dots P(x_n)) + C(P) - C(\vec{x}) = \delta(\vec{x}, P). \end{aligned}$$

Таким образом, множество $Q_{\vec{x}}$ содержится во множестве $P_{\vec{x}}$ (с точностью $O(\log(n + l))$).

Осталось показать обратное включение. Из неравенства выше ясно, что разница между $\delta(\vec{x}, P)$ и $d(\vec{x} | P)$ равняется

$$(C(P) - C(\vec{x})) + C(\vec{x} | P) = C(P | \vec{x}),$$

где равенство следует из коммутативности информации.

Оказывается, что если величина $C(P | \vec{x})$ большая, то существует объяснение $\tilde{P}$ для $\vec{x}$ с гораздо лучшими параметрами:

Лемма 5.5. *Для каждого распределения P и для каждого набора $\vec{x} = x_1 \dots x_l$ слов длины n существует такое распределение $\tilde{P}$, что:*

- (1) $-\log(\tilde{P}(x_1) \dots \tilde{P}(x_l)) \leq -\log(P(x_1) \dots P(x_l)) + O(\log(n + l))$ и
- (2) $C(\tilde{P}) \leq C(P) - C(P | \vec{x}) + O(\log(n + l))$.

Для доказательства этой леммы нам понадобится другая:

Лемма 5.6. *Пусть $x_1, \dots, x_l \in \{0, 1\}^n$. Предположим, что существует 2^k таких распределений P , что:*

- (1) $-\log(P(x_1) \dots P(x_l)) \leq b$.
- (2) $C(P) \leq a$.

Тогда существует такое распределение $\tilde{P}$ сложности не больше $a - k + O(\log(n + l + a + b))$, что $-\log(\tilde{P}(x_1) \dots \tilde{P}(x_l)) \leq b$.

Доказательство Леммы 5.6. В Лемме 4.4 возьмём в качестве L множество вероятностных распределений, а в качестве R множество наборов из l слов длины n . Будем проводить ребро между $\langle x_1, \dots, x_l \rangle$ и распределением Q , если $\log(Q(x_1) \dots Q(x_l)) \geq -b$. Тогда существование нужного нам распределения следует из заключения Леммы 4.4. $\square$

Доказательство Леммы 5.5. Пусть нам дан некоторый набор слов $\vec{x}$. Будем перечислять все такие распределения Q , что $C(Q) \leq a = C(P)$ и $-\log(Q(x_1) \dots Q(x_l)) \leq b = -\log(P(x_1) \dots P(x_l))$. Мы можем восстановить P из $\vec{x}$ и порядкового номера P в этом перечислении. Таким образом, логарифм этого числа должен быть больше чем $C(P|\vec{x})$ (с логарифмической точностью). По Лемме 5.6 для $k = C(P|\vec{x})$ в перечислении существует вероятностное распределение $\tilde{P}$ сложности не больше $a - k$ (с логарифмической точностью). $\square$

Теперь мы готовы завершить доказательство теоремы. Рассмотрим некоторое распределение P . Нам нужно показать, что существует такое распределение $\tilde{P}$, что: $C(\tilde{P}) \leq C(P) + O(\log(n + l))$ и $\delta(\vec{x}, \tilde{P}) \leq d(\vec{x}|P) + O(\log(n + l))$. Для этого рассмотрим распределение $\tilde{P}$ из 5.5. По построению сложность $\tilde{P}$ не больше сложности P (с логарифмической точностью). Дефект оптимальности $\delta(\vec{x}, \tilde{P})$ может быть ограничен сверху следующим образом

$$\begin{aligned} \delta(\vec{x}, \tilde{P}) &= C(\tilde{P}) - \log(\tilde{P}(x_1) \dots \tilde{P}(x_l)) - C(x_1, \dots, x_l) \\ &\leq C(P) - C(P|\vec{x}) - \log(P(x_1) \dots \tilde{P}(x_l)) - C(x_1, \dots, x_l) \\ &= \delta(P, \vec{x}) - C(P|\vec{x}) = d(\vec{x}|P). \quad \square \end{aligned}$$

Напомним, что в одномерном случае связь между дефектом случайности и дефектом оптимальности верна также для ограниченного класса гипотез — Теорема 2.12.

Возникает естественный вопрос: может ли Теорема 5.4 быть обобщена для любого разрешимого класса распределений? Утвердительным ответом является Теорема 5.7 ниже. Сначала дадим необходимые определения.

Определение 14. Пусть $\mathcal{A}$ — разрешимое семейство распределений. Профиль оптимальности $P_{\vec{x}}^{\mathcal{A}}$ набора $\vec{x}$ определяется как множество всех таких пар $\langle a, b \rangle$ натуральных чисел, что существует такое вероятностное распределение $P \in \mathcal{A}$ сложности не больше a , что $\delta(\vec{x}, P) \leq b$.

Определение 15. Пусть $\mathcal{A}$ — разрешимое семейство распределений, и $\vec{x}$ есть набор слов. Множество всех таких пар (a, b) , что существует такое распределение $P \in \mathcal{A}$ сложности не больше a , что $d(\vec{x}|P) \leq b$ называется $\mathcal{A}$ -профилем стохастичности $\vec{x}$ и обозначается как $Q_{\vec{x}}^{\mathcal{A}}$.

Теорема 5.7. Для каждого разрешимого семейства распределений $\mathcal{A}$ и для каждого набора $\vec{x} = x_1, \dots, x_l$ слов длины n расстояние между множествами $P_{\vec{x}}^{\mathcal{A}}$ и $Q_{\vec{x}}^{\mathcal{A}}$ не превосходит $O(\log(n + l))$.

Доказательство почти дословно повторяет доказательство Теоремы 5.4. $\square$

5.3 Предсказания для неравномерных распределений

В Главе 4 мы использовали множества, т.е. равномерные распределения для определения алгоритмических и вероятностных окрестностей. Согласно Предложению 5.1 другие распределения не могут иметь качественно лучшие параметры объяснения для одного объекта. Это частично оправдывает наше пренебрежение

к неравномерным распределениям для случая одного слова. Но Пример 6 показывает, что уже для двух слов использование неравномерных распределений существенно и пренебрегать ими нельзя.

В этом разделе мы показываем, как обобщить результаты Раздела 4.2 на случай произвольных распределений. Удивительно, но несмотря на Предложение 5.1, такой подход оказывается полезным даже для случая одного положительного примера.

5.3.1 Неравномерные распределения для одного слова

Начнём с определения вероятностной окрестности для одного слова x . Для этого рассмотрим схожий двухэтапный процесс: в начале выбирается распределение P (на словах длины n) с вероятностью $\mathbf{m}(P)$, далее выбирается слово x с вероятностью $P(x)$. Раньше мы определяли $p_x(y)$ как

$$\Pr[y \in A \mid x \text{ является выходом двухэтапного процесса}].$$

Чем сейчас заменить выражение “ $y \in A$ ”? Его можно заменить на “ $P(y) > 0$ ”, но это ни к чему интересному не приведёт: $p_x(y)$ будет большим при любом y . В самом деле, при таком определении это значение будет равняться

$$\frac{\sum_{P: P(y) > 0} \mathbf{m}(P) \cdot P(x)}{\sum_P \mathbf{m}(P) \cdot P(x)}.$$

Знаменатель этой дроби равняется (с точностью до мультипликативной константы) $\mathbf{m}(x)$ (из тех же соображений, что и раньше). Чтобы оценить числитель рассмотрим распределение P_x^n , для которого $P_x^n(x) = \frac{1}{2}$, а оставшаяся $\frac{1}{2}$ равномерно распределена по оставшимся словам длины n . Так как $\mathbf{m}(P_x^n)$ равняется $\mathbf{m}(x)$ (с точностью до мультипликативной константы), то вообще вся дробь не меньше некоторой положительной константы (зависящей только от выбора универсального $\mathbf{m}$).

Поэтому определение следует изменить. Первый этап остаётся без изменений: мы выбираем распределение P с вероятностью $\mathbf{m}(P)$. Далее мы повторяем второй этап процесса (т.е. получаем слова согласно распределению P) t раз. Обозначим через $p_x^t(y)$ вероятность того, что после t повторений хотя бы один раз выйдет y . При $t = 1$ получаем

$$p_x^1(y) = \frac{\sum_P \mathbf{m}(P) \cdot P(x) \cdot P(y)}{\sum_P \mathbf{m}(P) \cdot P(x)}.$$

Знаменатель этой дроби равняется $\mathbf{m}(x)$ с точностью до мультипликативной константы, а числитель — $\mathbf{m}(x, y)$. В самом деле, для нижней оценки нужно рассмотреть распределение, для которого вероятности x и y равны по $\frac{1}{2}$, а верхняя оценка числителя получается так же как верхняя оценка для знаменателя. В свою очередь

$$\frac{\mathbf{m}(x, y)}{\mathbf{m}(x)} = 2^{-C(x, y) + C(y) + O(\log n)} = 2^{-C(x|y) + O(\log n)}$$

(согласно коммутативности информации).

Для произвольного t получаем

$$p_x^t(y) = \frac{\sum_P \mathbf{m}(P) \cdot P(x) \cdot (1 - (1 - P(y))^t)}{\sum_P \mathbf{m}(P) \cdot P(x)}.$$

Как и в Разделе 4 мы суммируем только по тем вероятностным распределениям P , чьи сложности не превосходят $4n + O(1)$. Заметим, что $p_x^t(y)$ не убывает при увеличении t .

Определим вероятностную d, t -окрестность x , как объединение всех слов y , для которых $p_x^t(y) \geq 2^{-d}$.

Для формулировки результата, сходного с результатами Главы 4 нам придется поправить определение алгоритмической окрестности. Будем говорить, что слово y содержится в алгоритмической d, p -окрестности x , если существует распределение P для которого $\delta(x, P) \leq d$ и $P(y) \geq p$. Отметим, что алгоритмическая d, p -окрестность увеличивается при увеличении d и при уменьшении p , а вероятностная d, t -окрестность увеличивается при увеличении обоих аргументов.

Теорема 5.8. (а) Пусть слово y содержится в алгоритмической d, p -окрестности слова x длины n . Тогда y также содержится в вероятностной λ, t -окрестности x при любых λ и t , удовлетворяющих неравенству $\lambda \geq d - \min(0, \log pt) + O(\log n)$.

(б) Пусть слово y содержится в вероятностной d, t -окрестности слова x длины n . Тогда y также содержится в алгоритмической λ, p -окрестности x , при некотором p и любом $\lambda \geq d + \min(0, \log pt) + O(\log n)$.

Сравним эту теорему с Теоремой 4.3. В Теореме 4.3 утверждается, что если некоторое устройство выдало x , то вероятность (в соответствующей модели) того, что это же устройство может выдать y велика, если y содержится в некотором множестве $A \ni x$ для которого $\delta(x, A) \approx 0$. В Теореме 5.8 утверждается, что если устройство выдало x , то вероятность (в соответствующей модели) того, что это же устройство может выдать y за t шагов мала, если существует такое распределение P , что $\delta(x, P) \approx 0$ и $P(y) \geq 2^{-t}$.

Доказательство Теоремы 5.8 (а). Пусть слово y содержится в d, p -окрестности слова x , т.е. существует такое распределение P , что $P(y) \geq p$ и $\delta(x, P) \leq d$. Как и раньше, существенным является случай, когда $d \geq 2n$ (при большем d для доказательства включения y в соответствующую вероятностную окрестность x нужно рассмотреть равномерное распределение на словах длины n).

Неравенство $\delta(x, P) = C(P) - \log P(x) - C(x) \leq d$ влечёт

$$\frac{\mathbf{m}(P)P(x)}{\mathbf{m}(x)} \geq 2^{-d-O(\log n)}, \text{ что позволяет сделать следующую оценку.}$$

$$p_x^t(y) = \frac{\sum_P \mathbf{m}(P) \cdot P(x) \cdot (1 - (1 - P(y))^t)}{\mathbf{m}(x) \cdot O(1)} \geq (1 - (1 - P(y))^t) 2^{-d-O(\log n)}.$$

$$\text{Так как } P(y) \geq p, \text{ а } (1 - p)^{\frac{1}{p}} < \frac{1}{e} \text{ то } p_x^t(y) \geq (1 - e^{-pt}) 2^{-d-O(\log n)}.$$

$$\text{При } pt \geq 1 \text{ получаем } p_x^t(y) \geq 2^{-d-O(\log n)}.$$

$$\text{Иначе } 1 - e^{-pt} > \frac{pt}{2}, \text{ а значит } p_x^t(y) \geq 2^{\log pt - d - O(\log n)}.$$

Последние две оценки для $p_x^t(y)$ дают утверждаемое включение y в вероятностную окрестность x . $\square$

Перед тем как доказывать второе утверждение теоремы сформулируем еще одно следствие Леммы 4.4.

Следствие 5.9. Пусть $\mathcal{A}$ – разрешимое множество распределений, $x_1, \dots, x_l, y$ – слова длины n , и t – натуральное число. Обозначим через $\mathcal{A}_m^n$ подмножество распределений из $\mathcal{A}$ сложности не больше m , сосредоточенных на $\mathbb{B}^n$. Тогда сумма

$$\sum_{P \in \mathcal{A}_m^n} \mathbf{m}(P) P(x_1) \dots P(x_l) (1 - (1 - P(y))^t)$$

равна своему максимальному члену с точностью до мультипликативного множителя $2^{O(\log(n+m+l))}$.

Это следствие доказывается точно так же как Следствие 4.5.

Доказательство Теоремы 5.8 (b). Пусть $p_x^t(y) \geq 2^{-d}$. Воспользовавшись Следствием 5.9 (взяв в качестве l единицу, а в качестве m взяв $4n + O(1)$) получаем, что для некоторого P верно

$$\frac{\mathbf{m}(P) P(x) (1 - (1 - P(y))^t)}{\mathbf{m}(x)} \geq 2^{-d - O(\log n)}.$$

Следовательно, y принадлежит алгоритмической λ, p -окрестности x , где $\lambda = C(P) - \log P(x) - C(x)$, $p = P(y)$. При этом

$$\lambda \leq d + \log(1 - (1 - p)^t) + O(\log n).$$

Положим $p = P(y)$ и рассмотрим распределение P в качестве одного из распределений в определении алгоритмической окрестности. Из последнего неравенства следует, что y принадлежит алгоритмической λ, p -окрестности x при любом $\lambda > d + \log(1 - (1 - p)^t) + O(\log n)$.

Для завершения доказательства осталось лишь проверить, что

$$\log(1 - (1 - p)^t) \leq -\min(0, \log pt),$$

то есть,

$$(1 - (1 - p)^t) \leq \min(1, pt).$$

То, что левая часть не превосходит 1, очевидно, а то, что она не больше pt следует из неравенства Бернулли $(1 - a)^b > 1 - ab$. $\square$

5.3.2 Общий случай

Обобщим предыдущий результат для произвольного количества слов. Ограничим также множество рассматриваемых распределений некоторым разрешимым множеством $\mathcal{A}$.

Для определения алгоритмической окрестности мы обобщаем процесс, описанный в прошлом пункте.

Во-первых, выбирается распределение $P \in \mathcal{A}$ на словах длины n (как обычно сложности не больше $4n + O(1)$) с вероятностью $\mathbf{m}(P)$. Далее согласовано с этим распределением выбираются l слов $x_1, \dots, x_l = \vec{x}$. Запустим распределение P ещё t раз. Обозначим через $p_{\vec{x}, \mathcal{A}}^t(y)$ вероятность того, что за эти t шагов хотя бы один раз выйдет y . Получаем

$$p_{\vec{x}, \mathcal{A}}^t(y) = \frac{\sum_{P \in \mathcal{A}} \mathbf{m}(P) P(x_1) \dots P(x_l) \cdot (1 - (1 - P(y))^t)}{\sum_{P \in \mathcal{A}} \mathbf{m}(P) \cdot P(x_1) \dots P(x_l)}. \quad (5.6)$$

Определим вероятностную $\mathcal{A}, t, d$ -окрестность $\vec{x}$ как множество всех слов y для которых $p_{\vec{x}, \mathcal{A}}^t(y) \geq 2^{-d}$.

Новое определение алгоритмической окрестности схоже с определением из Раздела 4.4.

Определение 16. Пусть $\vec{x} = x_1, \dots, x_l$ — вектор слов длины n , d — некоторое число, и $\mathcal{A}$ — семейство распределений. Тогда слово y содержится в $\mathcal{A}, d, p$ -алгоритмической окрестности $\vec{x}$, если существует такое распределение $P \in \mathcal{A}$, при котором

- $P(y) \geq 2^{-p}$
- $\delta(\vec{x}, P) \leq \delta(\vec{x}, Q) + d$ для любого $Q \in \mathcal{A}$.

Теперь всё готово, чтобы сформулировать теорему о связи вероятностной и алгоритмической окрестностей в наиболее общей форме.

Теорема 5.10. Пусть $\vec{x}$ — набор из l слов длины n . Тогда

(а) если слово y содержится в алгоритмической $\mathcal{A}, d, p$ -окрестности $\vec{x}$, тогда y также содержится в вероятностной $\mathcal{A}, \lambda, t$ -окрестности $\vec{x}$ при любых λ и t , удовлетворяющих неравенству $\lambda \geq d - \min(0, \log pt) + O(\log n)$,

(б) если y содержится в вероятностной $\mathcal{A}, d, t$ -окрестности $\vec{x}$, то y также содержится в некоторой алгоритмической $\mathcal{A}, \lambda, p$ -окрестности $\vec{x}$ при некотором p и любом $\lambda \geq d + \min(0, \log pt) + O(\log n)$.

(Чтобы получить из этой теоремы результаты о множествах нужно в качестве $\mathcal{A}$ взять семейство равномерных распределений, а t устремить к бесконечности).

Доказательство этой теоремы повторяет идеи доказательства Теоремы 5.8, а также Теоремы 4.7.

Доказательство Теоремы 5.10 (а). Слово y принадлежит алгоритмической $\mathcal{A}, d, p$ -окрестности $\vec{x}$, т.е. максимальное слагаемое числителя в (5.6) не меньше чем максимальное слагаемое в знаменателе, умноженное на 2^{-d} . Благодаря Следствию 5.9 получаем

$$\sum_{P \in \mathcal{A}} \mathbf{m}(P) P(x_1) \cdots P(x_l) \geq 2^{-d - O(\log(n+l))} \sum_{P \in \mathcal{A}} \mathbf{m}(P) \cdot P(x_1) \cdots P(x_l).$$

Из этого неравенства следует (также как в предыдущем доказательстве) требуемое включение y в вероятностную окрестность $\vec{x}$. $\square$

Доказательство Теоремы 5.10 (б). Теперь y принадлежит вероятностной $\mathcal{A}, d, p$ -окрестности, т.е. числитель в (5.6) не меньше чем 2^{-d} [знаменатель в (5.6)]. А значит, и

$$\sum_{P \in \mathcal{A}} \mathbf{m}(P) P(x_1) \cdots P(x_l) \cdot (1 - (1 - P(y))^t) \geq 2^{-d - O(\log(n+l))} D_{\max},$$

где $D_{\max}$ — максимальное слагаемое знаменателя в (5.6). Левая сумма также равна своему максимальному слагаемому. Оценив (как в предыдущем доказательстве) выражение $1 - (1 - P(y))^t$, получаем требуемое включение y в алгоритмическую окрестность $\vec{x}$. $\square$

Глава 6

Сильные статистики и нормальные объекты

Эта глава посвящена нормальным словам — это одно из основных понятий теории сильных моделей, о которой шла речь в Разделе 2.7. Напомним основные определения этой теории.

- Множество $A \ni x$ называется ε -сильной моделью для x , если $\text{CT}(A|x) < \varepsilon$.
- Множество пар целых чисел

$$P_x(\varepsilon) := \{(m, l) \mid \exists A \ni x : \text{CT}(A|x) < \varepsilon, C(A) \leq m \text{ и } \log |A| \leq l\}$$

называется ε -сильным профилем слова x .

- Слово x называется ε, δ -нормальным, если множество P_x содержится в δ -окрестности множества $P_x(\varepsilon)$ (т.е. из того, что пара (a, b) принадлежит P_x следует, что пара $(a + \delta, b + \delta)$ принадлежит $P_x(\varepsilon)$).

6.1 Существование нормальных слов

Покажем, что для любого множества пар натуральных чисел P , удовлетворяющему условию Теоремы 2.6, существует нормальное слово, чей профиль близок к P . Напомним, что для данных n и k множества $P_{\min}$ и $P_{\max}$ определяются следующим образом.

$$P_{\min} = \{(m, l) \mid m + l \geq n \text{ или } m \geq k\}$$
$$P_{\max} = \{(m, l) \mid m + l \geq k\}$$

Теорема 6.1. Пусть замкнутое вверх множество пар натуральных чисел P содержит $P_{\min}$ и содержится в $P_{\max}$. Предположим также, что для всех $(m, l) \in P$ и всех $i \leq l$ выполнено $(m + i, l - i) \in P$.

Тогда существует $(O(\log n), O(\sqrt{n \log n}))$ -нормальное слово x длины n и сложности $k + O(\log n)$, чей профиль $C(P) + O(\sqrt{n \log n})$ -близок к P .

Доказательство. Нам потребуется вспомнить результаты Раздела 2.6 об ограниченных классах гипотез. А именно, Теорему 2.13 о том, что для любого приемлемого семейства множеств $\mathcal{A}$ для любого множества пар P , удовлетворяющему

условию Теоремы 2.6, существует такое слово x длины n , что оба профиля P_x и P_x^A являются $C(P) + \sqrt{n \log n}$ -близкими к P .

Воспользуемся Теоремой 2.13, взяв в качестве $\mathcal{A}$ семейство цилиндров (т.е. множеств вида $\{uv \mid v \in \{0,1\}^m\}$, где u — произвольное бинарное слово, а m — произвольное целое неотрицательное число). Как мы обсуждали в Разделе 2.6, семейство цилиндров является приемлемым (а значит, к нему в самом деле можно применить Теорему 2.13). Таким образом, для любого множества пар P (удовлетворяющему условию Теоремы 2.6) существует такое слово x длины n , что множества P_x и P_x^A являются $C(P) + O(\sqrt{n \log n})$ -близкими к P . Покажем, что слово x является $(O(\log n), O(\sqrt{n \log n}))$ -нормальным (что завершит доказательство теоремы). Для этого заметим, что любой цилиндр $A \ni x$ является $O(\log n)$ -сильной статистикой для x : в самом деле, множество A можно получить из x , зафиксировав несколько первых его бит. Таким образом $P_x(O(\log n))$ включает P_x^A , а P_x^A близко (с точностью $O(\sqrt{n \log n})$) к P_x , так как оба этих множества близки к P с такой же точностью. $\square$

Это доказательство основано на довольно сложной Теореме 2.13. Между тем, для некоторых множеств P Теорему 6.1 можно доказать гораздо проще и даже с лучшей точностью. А именно, можно показать, что антистохастические слова из Главы 3 являются нормальными. (Напомним, что слово x длины n и сложности k называется антистохастичным, если её профиль P_x близок к множеству $P_{\min}$, см. Рисунок 3.1.)

Предложение 6.2. *Пусть x является ε -антистохастическим словом длины n и сложности k . Тогда x также является $(O(\log n), O(\log n) + \varepsilon)$ -нормальным.*

Доказательство. Чтобы доказать это утверждение, достаточно предъявить для каждой точки (i, j) на границе профиля P_x на Рисунке 3.1 такую $O(\log n)$ -сильную модель A для x , что $C(A) \leq i + O(\log n)$ и $\log |A| \leq j$.

Для $i \geq k$ в качестве A возьмём $\{x\}$. Для $i < k$, возьмём в качестве A цилиндр, состоящий из слов длины n , у которых первые i битов совпадают с первыми i битами x . По построению $C(A) \leq i + O(\log n)$ и $\log |A| = n - i = j$. $\square$

6.2 Нормальные слова и стандартные модели

Напомним, что сильные модели были введены, чтобы отделить хорошие модели от плохих. При этом, под плохими статистиками мы имеем ввиду стандартные модели из Раздела 2.5. Следующая теорема показывает, что существует такое слово x и её сильная модель A , что у любой сильной *стандартной* модели для x параметры гораздо хуже чем у A .

Теорема 6.3. *Для некоторого положительного c для всех достаточно больших k существует такое $O(\log k), O(\log k)$ -нормальное слово x длины $n = 4k$ чей профиль $O(\log n)$ -близок к серому множеству, изображённому на Рисунке 6.1, что*

- для x существует $O(\log n)$ -сильная модель A сложности $k + O(\log n)$ и размера 2^{2k} (т.е. модель соответствующая точке $(k, 2k)$ на границе P_x), однако

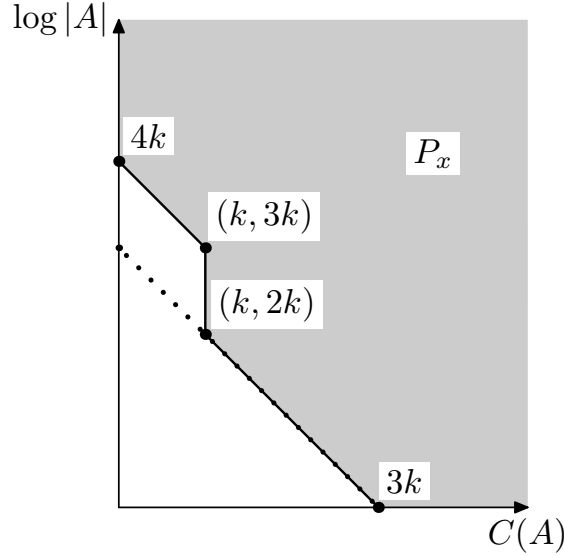


Рис. 6.1: Профиль P_x слова x из Теоремы 6.3.

- для каждого $t \geq C(x)$ и для каждого простого перечисления слов сложности не больше t стандартная модель B для x , полученная при этом перечислении, либо не является сильной для x , либо её параметры далеки от точки $(k, 2k)$. Более точно, если B является моделью для x , полученной из перечисления программой q , то, по крайней мере, одно из следующих значений $\text{CT}(B|x)$, $C(q)$, $|C(B) - k|$, $|\log |B| - 2k|$ больше чем ck .

Доказательство Теоремы 6.3. Рассмотрим $O(\log k)$ -антистохастическое слово y сложности $k + O(\log k)$ и длины $2k$, которое существует согласно Следствию 3.1. Пусть z такое слово длины $2k$, что $C(z|y) \geq 2k$. Наконец, возьмём в качестве x конкатенацию слов y и z . Очевидно, что $C(x) = 3k + O(\log k)$.

Рассмотрим множество $A = \{yz' \mid l(z') = 2k\}$. Заметим, что A является $O(\log k)$ -сильной статистикой для x . Также оно является $O(\log k)$ -достаточной статистикой для x (напомним, что множество $B \ni x$ называется ε -достаточной статистикой для x , если $\delta(x, B) < \varepsilon$). Очевидно, что профиль $P_{[A]}$ совпадает с профилем антистохастического слова y с точностью $O(\log k)$. Поэтому по Теореме 2.17 профиль P_x совпадает с серым множеством на Рисунке 6.1 с точностью $O(\log k)$. Из нормальности y несложно следует $O(\log k)$, $O(\log k)$ -нормальность x .

Теперь нам нужно показать, что каждая стандартная модель, чьи параметры (сложность, логарифм размера), близки к параметрам A не является сильной.

Пусть q есть некоторая программа, перечисляющая все слова сложности не больше t . Пусть B — стандартная модель для x , полученная из этого перечисления. Нам нужно показать, что для некоторого положительного c и для всех достаточно больших k выполняется неравенство

$$\min\{\text{CT}(B|x), C(q), |C(B) - k|, |\log |B| - 2k|\} \geq ck. \quad (6.1)$$

Зафиксируем небольшое положительное c . Предположим противное: для бесконечно многих k существуют такие t, q, B , что неравенство (6.1) не выполняется.

Опишем вкратце, как из этого прийти к противоречию. Мы предполагаем, что $C(B)$ и $\log |B|$ близки к k и $2k$ соответственно. Из формы P_x следует, что в этом

случае B является достаточной минимальной статистикой для x . Мы также предполагаем, что $C(B|x)$ мало, т.е. B является сильной статистикой для x . Значит, мы можем применить Теорему 2.18 к B и A и вывести, что $CT(B|A)$ мало. Так как $CT(A|y) \approx 0$, то $CT(B|y)$ также мало. С другой стороны мы покажем, что тотальная условная сложность любой стандартной статистики B , полученной из перечисления слов сложности не больше m , при условии любого слова y больше чем $\min\{C(B), m - |y|\}$. В нашем случае m не меньше $C(x) \approx 3k$, $|y| = 2k$ и $C(B) \approx k$, таким образом, $CT(B|y) \geq k$ с точностью $O(\log k)$, что приводит к противоречию если c достаточно маленькое.

Теперь опишем доказательство подробнее. Покажем сначала, что $m = O(k)$. Напомним, что B — стандартная модель, полученная из списка слов сложности не больше m , который перечисляется программой q . Покажем, что

$$C(B) + \log |B| = m + O(C(q) + \log m). \quad (6.2)$$

В самом деле, B можно задать с помощью q и первыми $m - \log |B|$ битами числа Ω_m (количества всех слов сложности не больше m), поэтому левая часть (6.2) не превосходит правой. С другой стороны, по q , B и $\log |B|$ битам (задающим число слов, оставшихся в перечислении после всех слов из B) можно восстановить Ω_m (чья сложность равняется $m + O(\log m)$ согласно Предложению 2.11), поэтому неравенство верно и в другую сторону.

Следовательно, $m \leq (k + ck) + (2k + ck) = O(k)$ с точностью $O(C(q) + \log m)$. Мы предположили, что $C(q) < ck = O(k)$, а значит $m = O(k)$.

Теперь мы будем применять Теорему 2.18. Напомним её формулировку.

Для некоторого значения $\varkappa = O(\log n)$ верно следующее. Пусть A и B являются ε -достаточными статистиками для слова x длины n . Предположим также, что B является $(\delta, \varepsilon + \varkappa)$ -минимальной статистикой для x , а также ε -сильной моделью для x . Тогда $CT(A|B) = O(\varepsilon + \delta + \log n)$.

Зафиксируем такое $\varkappa = O(\log n) = O(\log k)$ (напомним, что $n = 4k$).

Модели A, B являются ε -достаточными, а B к тому же ε -сильна для

$$\varepsilon = \max\{\delta(x, B), \delta(x, A), CT(B|x)\} \leq 2ck + O(\log k).$$

Тем не менее, форма P_x гарантирует, что B является δ, t -минимальной для

$$\begin{aligned} \delta &= |C(B) - k| + O(\log k) \leq ck + O(\log k), \\ t &= k - \delta(x, B) - O(\log k) \geq k - 2ck - O(\log k). \end{aligned}$$

Если $c < 1/4$, то для всех достаточно больших k верно $t \geq \varepsilon + \varkappa$, а значит B является $\delta, \varepsilon + \varkappa$ -минимальной и, следовательно, мы можем применить Теорему 2.18. Таким образом $CT(B|A) \leq O(\varepsilon + \delta + \log k) = O(ck + \log k)$ и, следовательно, $CT(B|y) = O(ck + \log k)$

Противоречие получается из последнего неравенства и следующей леммы.

Лемма 6.4. Пусть множество B является некоторой стандартной моделью, полученной перечислением слов сложности не больше m программой q , а y — некоторое слово. Тогда $CT(B|y) \geq \min\{C(B), m - |y|\} - O(C(q) + \log n)$, где $n = \max\{|y|, m\}$.

Доказательство. Обозначим через b лексикографически первый элемент в B . Можно оценить тотальную условную сложность b при условии y следующим образом:

$$CT(b|y) \leq CT(b|B) + CT(B|y) + O(\log n).$$

Первое слагаемое здесь ограничено константой, а значит, $CT(b|y) \leq CT(B|y) + O(\log n)$. Обозначим через p тотальную программу длины $CT(b|y)$, которая преобразует y в b и рассмотрим следующее множество

$$D := \{p(y') \mid |y'| = |y|\}.$$

Ясно, что $C(D) \leq |p| + O(\log n)$. Таким образом, достаточно доказать, что $C(D) \geq \min\{C(B), m - |y|\} - O(\log n)$.

Сложность каждого элемента из D не превосходит $C(D) + \log |D| + O(\log n) \leq C(D) + |y| + O(\log n)$. В случае, когда правая часть этого неравенства больше чем m получается, что $C(D) \geq m - |y| - O(\log n)$, из чего сразу же следует утверждение леммы.

Теперь рассмотрим случай, когда у каждого элемента из D сложность не превосходит m . Запустим программу q и дождёмся того момента, когда она напечатает все элементы из D . Так как $b \in D$ и $b \in B$, то существует не более $2|B|$ слов, чья сложность не превосходит m и которые не были напечатаны (все элементы в B которые идут после b в перечислении, и не более $|B|$ элементов, которые идут после перечисления всего множества B). Таким образом, мы можем найти список всех слов сложности не больше m по D , q и некоторым $\log |B| + 1$ битам. Так как сложность этого списка равняется $m - O(\log m)$, мы получаем

$$C(D) + C(q) + \log |B| \geq m - O(\log m).$$

Напомним, что для каждой стандартной модели B выполнено равенство (6.2). Складывая это равенство с предыдущим неравенством получаем

$$C(D) + O(C(q) + \log m) \geq C(B).$$

□

Лемма 6.4 таким образом влечёт, что

$$\min\{C(B), m - 2k\} - O(C(q) + \log k) \leq CT(B|y) = O(ck + \log k).$$

Так как $C(B) \geq k - ck$ и $m \geq C(x) = 3k - O(\log k)$, получаем

$$k \leq O(ck + \log k).$$

Обозначим константу, спрятанную в $O(ck + \log k)$, через C . Таким образом, при $c < 1/C$ мы получаем противоречие. □

6.3 Наследственность нормальных слов

В этом разделе мы показываем, что нормальные слова обладают следующим свойством: минимальная достаточная статистика любого нормального слова сама является нормальной. Формально, мы доказываем следующий результат.

Теорема 6.5. Для некоторого $\varkappa = O(\log n)$ выполнено следующее. Пусть A является ε -сильной и ε -достаточной статистикой для $(\varepsilon, \varepsilon)$ -нормального слова x длины n . Предположим также, что A является $(\delta, \varkappa)$ -минимальной моделью для x . Тогда A является $(O(\delta + (\varepsilon + \log n)\sqrt{n}), O(\delta + (\varepsilon + \log n)\sqrt{n}))$ -нормальным словом¹.

Перед тем как перейти к доказательству этой теоремы, сформулируем одно общее утверждение о сильных статистиках: дадим им эквивалентное определение.

Конечное семейство множеств $\mathcal{A}$ называется *разбиением*, если для любых $A_1, A_2 \in \mathcal{A}$ выполняется $A_1 \cap A_2 \neq \emptyset \Rightarrow A_1 = A_2$. Так как любое разбиение является конечным объектом, то можно определить его сложность. Оказывается, сильными статистиками являются те множества, которые принадлежат простым разбиениям.

Предложение 6.6. Пусть A — модель для x , принадлежащая разбиению сложности ε . Тогда A является $\varepsilon + O(1)$ -сильной моделью для x . Обратно, предположим, что A является ε -сильной статистикой для слова x длины n . Тогда существует разбиение $\mathcal{A}$ сложности не больше $\varepsilon + O(\log n)$ и такая модель $A' \in \mathcal{A}$ для x , что: $|A'| \leq |A|$ и обе величины $\text{CT}(A|A')$ и $\text{CT}(A'|A)$ не превосходят $\varepsilon + O(\log n)$.

Доказательство. Пусть A — модель для x , принадлежащая разбиению сложности ε . Программа, которая переводит слово x во множество $A \in \mathcal{A}$, которому принадлежит x (если x не принадлежит никакому множеству в $\mathcal{A}$, тогда программа переводит x , скажем, в пустое множество) является тотальной, и её длина не превосходит $\varepsilon + O(1)$.

Обратно, предположим A является ε -сильной статистикой для слова x длины n . Тогда существует такая тотальная программа p , что $p(x) = A$ и $|p| \leq \varepsilon$.

Рассмотрим множество X , состоящее из всех таких слов x' , что $x' \in p(x')$. Очевидно, что $x \in X$. Разобьём X в соответствии со значением $p(x')$: слова x' и x'' принадлежат одному и тому же элементу разбиения, если $p(x') = p(x'')$. Это сконструированное разбиение $\mathcal{A}$ имеет сложность не больше $\varepsilon + O(\log n)$. Оно включает множество $A' = \{x' \in X \mid p(x') = A\}$, которое содержит x , и которое может быть получено из x с помощью тотальной программы длины не более $\varepsilon + O(\log n)$, которая переводит данное ей слово x' во множество $\{x'' \in X \mid p(x'') = p(x')\}$.

Так как $A' \subset A$, то $|A'| \leq |A|$. Осталось показать, что обе величины $\text{CT}(A|A')$ и $\text{CT}(A'|A)$ меньше чем $|p| + O(\log n) = \varepsilon + O(\log n)$. В самом деле, A' может быть получено из A с помощью тотальной программы длины $|p| + O(\log n)$, которая переводит данное ей множество B в $\{x' \in X \mid p(x') = B\}$. С другой стороны, A может быть получено из A' при помощи тотальной программы длины $|p| + O(1)$, которая из данного ему множества B выбирает любой элемент $x' \in B$ и вычисляет $p(x')$. $\square$

Для доказательства Теоремы 6.5 нам потребуются две леммы. В начале мы формулируем эти леммы неформально, затем дадим эскиз доказательства теоре-

¹Строго говоря, A — это множество слов, а не слово, поэтому в последней фразе речь идет о коде A , то есть, о слове, сопоставленном A при зафиксированной вычислимой биекции между двоичными словами и конечными множествами двоичных слов.

мы. Затем мы сформулируем леммы формально, и дадим строгое доказательство теоремы.

По Теореме 2.9 и Предложению 2.10 для каждого $A \ni x$ существует такое $B \ni x$ (стандартная модель), что $C(B|\Omega_{C(B)}) \approx 0$, причём параметры B не хуже параметров A . Нам будет нужен похожий результат для нормальных слов и сильных моделей.

Лемма 6.7 (нестрогая формулировка). *Пусть A минимальная статистика для некоторого слова. Тогда $C(\Omega_{C(A)}|A) \approx 0$.*

Лемма 6.8 (нестрогая формулировка). *Для каждого нормального слова x и для каждой модели A для x существует такая сильная статистика H для x , что*

- (1) $\delta(x, H) \lesssim \delta(x, A)$,
- (2) $C(H|\Omega_{C(H)}) \approx 0$ и
- (3) $C(H) \leq C(A)$.

Теперь мы сделаем набросок доказательства Теоремы 6.5, используя эти леммы.

Эскиз доказательства Теоремы 6.5. Пусть x — нормальное слово, и A — сильная минимальная достаточная статистика для x . Нам нужно показать, что профиль A близок к своему сильному профилю.

В начале покажем, что не умаляя общности, можно считать, что A принадлежит простому разбиению. В самом деле, так как A является сильной статистикой для x , мы можем применить Предложение 6.6 к A и x . Пусть $\mathcal{A}$ — простое разбиение, и $A_1 \in \mathcal{A}$ — модель для x из Предложения 6.6. Так как тотальные условные сложности $CT(A_1|A)$ и $CT(A|A_1)$ малы, то профили A и A_1 близки друг к другу (согласно Предложению 2.14). Нетрудно видеть, что то же самое верно для сильных профилей. Таким образом, достаточно показать, что A_1 нормально.

Для этого рассмотрим любую модель $\mathcal{G}$ (семейство множеств) для A_1 . Наша цель — найти сильную модель $\mathcal{F}$ для A_1 , чьи параметры (сложность, логарифм размера) не хуже, чем у $\mathcal{G}$. Для этого мы найдем такую модель M_1 для x , что

$$\begin{aligned} CT(M_1|A_1) &\approx 0, \quad \log |(M_1 \cap A_1)| \approx \log |A_1|, \\ C(M_1) &\leq C(\mathcal{G}), \quad C(M) + \log |M_1| \leq C(\mathcal{G}) + \log |\mathcal{G}| + \log |A_1|. \end{aligned} \tag{6.3}$$

Затем мы рассмотрим такое семейство

$$\mathcal{F} = \{A' \in \mathcal{A} \mid \log |(A' \cap M_1)| = \log |(A_1 \cap M_1)|\}.$$

(Здесь и далее под $\log$ мы имеем ввиду целую часть бинарного логарифма.) Семейство $\mathcal{F}$ может быть вычислено по M_1 , $\mathcal{A}$ и $\log |(A_1 \cap M_1)|$. Так как $\mathcal{A}$ простое, мы заключаем, что $C(\mathcal{F}) \leq C(M_1) \leq C(\mathcal{G})$.

Более того, $CT(\mathcal{F}|M_1) \approx 0$, потому что отображение

$$M' \mapsto \{A' \in \mathcal{A} \mid \log |(A' \cap M')| = \log |(A_1 \cap M')|\}$$

тотально. Так как $CT(M_1|A_1) \approx 0$, получаем, что $CT(\mathcal{F}|A_1) \approx 0$, т.е., $\mathcal{F}$ является сильной моделью для A_1 .

Наконец, $\log |\mathcal{F}| \leq \log |M_1| - \log |A_1|$, потому что $\mathcal{A}$ — разбиение, и значит, оно содержит лишь небольшое число множеств, у которых $\log |(A_1 \cap M_1)| \approx \log |A_1|$

общих элементов с M_1 . Таким образом, сумма сложности и логарифма размера $\mathcal{F}$ не превосходит

$$\begin{aligned} C(M_1) + (\log |M_1| - \log |A_1|) &\leq (C(\mathcal{G}) + \log |\mathcal{G}| + \log |A_1|) - \log |A_1| \\ &= C(\mathcal{G}) + \log |\mathcal{G}|. \end{aligned}$$

Следовательно, $\mathcal{F}$ является сильной моделью для A_1 , чьи параметры (сложность, сложность + логарифм размера) не хуже требуемых. Из Предложения 2.15 следует, что сильный профиль A_1 содержит точку $(C(\mathcal{G}), \log |\mathcal{G}|)$.

Как найти модель M_1 для x , удовлетворяющую (6.3)? Мы сделаем это в три шага. На первом шаге мы конструируем такую модель L для x , что $C(L) \leq C(\mathcal{G})$ и $\log |L| \leq \log |\mathcal{G}| + \log |A_1|$. Более точно, определим L как

$$L = \bigcup \{A' \in \mathcal{G} \mid \log |A'| = \log |A_1|\}.$$

По построению $C(L) \leq C(\mathcal{G})$ и $\log |L| \leq \log |\mathcal{G}| + \log |A_1|$ (см. Рисунок 6.2).

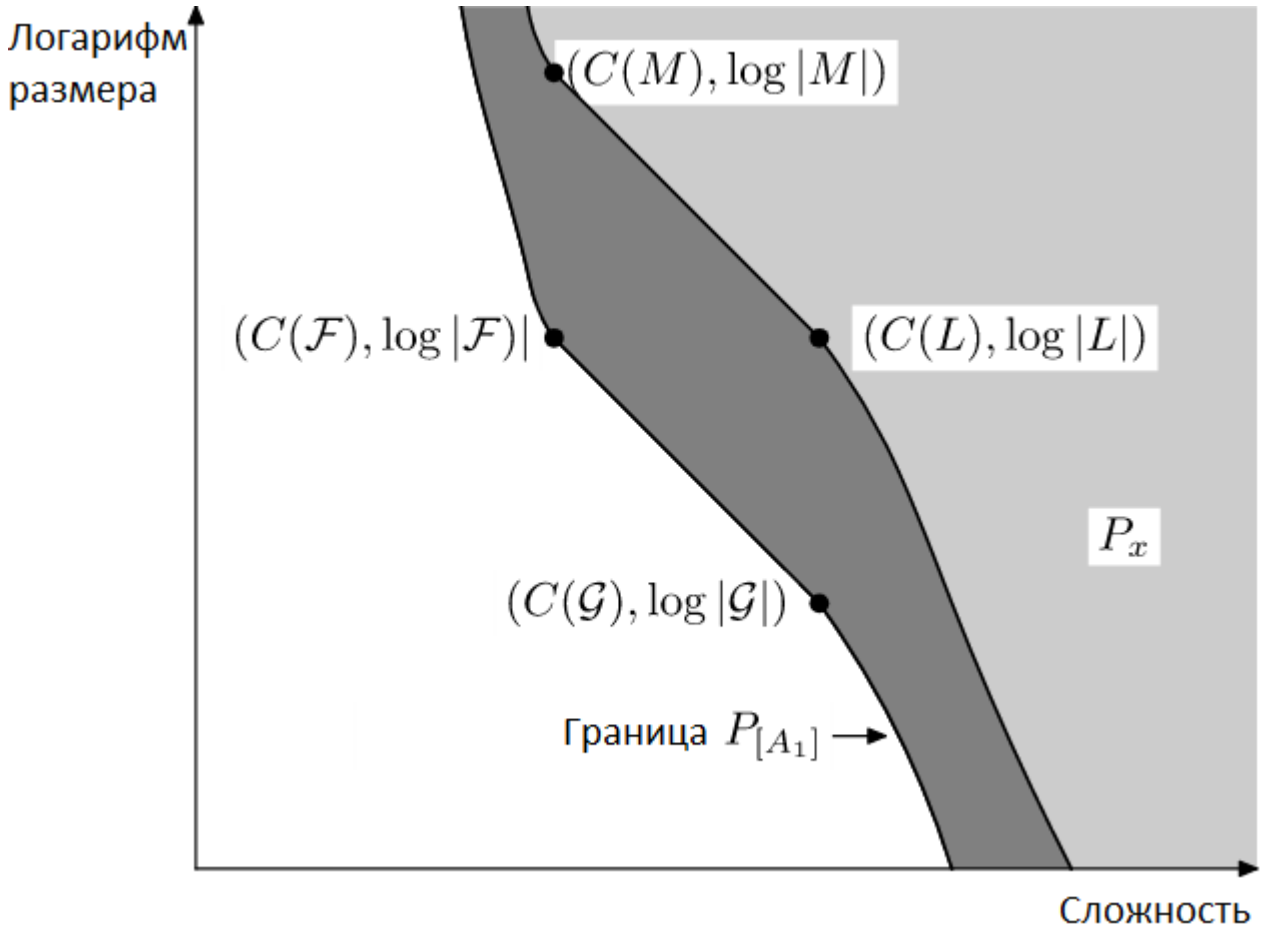


Рис. 6.2: Рисунок показывает параметры (сложность, логарифм размера) моделей $\mathcal{G}, \mathcal{F}$ (для A) и L, M (для x).

На втором шаге мы находим сильную модель M для x , чьи параметры (сложность, сложность + логарифм размера) не хуже, чем у L , для которой $C(M|\Omega_{C(M)}) \approx$

0; такая модель существует согласно Лемме 6.8. На третьем шаге мы находим такую модель M_1 с такими же параметрами как у M , которая принадлежит простому разбиению $\mathcal{M}$, для которой $CT(M_1|M) \approx 0$; такая модель существует согласно Предложению 6.6.

По построению имеем

$C(M_1) \leq C(M) \leq C(L) \leq C(\mathcal{G})$. Нам нужно показать, что $CT(M_1|A_1) \approx 0$ и $\log |(M_1 \cap A_1)| \approx \log |A_1|$.

Чтобы доказать первое равенство, мы сначала покажем, что $C(M|A_1)$ мало. В самом деле, по A_1 можно вычислить A (короткой программой), по A можно вычислить $\Omega_{C(A)}$ (Лемма 6.7), по $\Omega_{C(A)}$ можно вычислить $\Omega_{C(M)}$ (в самом деле, согласно Лемме 6.8 $C(M) \leq C(L) \leq C(\mathcal{G})$ и, не умаляя общности, мы можем предположить, что $C(\mathcal{G}) \leq C(A)$, так как $\mathcal{G}$ является моделью для A) и затем вычислить M (так как $C(M|\Omega_{C(M)}) \approx 0$ согласно Лемме 6.8). Согласно Предложению 6.6 множество M_1 является простым относительно M . Так как $C(M_1|M) \approx 0$ и $C(M|A_1) \approx 0$, получаем, что $C(M_1|A_1)$ также приблизительно равняется нулю.

Чтобы показать более сильно равенство $CT(M_1|A_1) \approx 0$, рассмотрим модель $A_1 \cap M_1$ для x . Мы утверждаем, что у этой модели мощность не может быть гораздо меньше чем у A_1 . В самом деле, так как $C(M_1|A_1) \approx 0$, имеем $C(A_1 \cap M_1) \leq C(A_1)$. Ясно, что $|(A_1 \cap M_1)| \leq |A_1|$. Таким образом, параметры $M_1 \cap A_1$ не хуже чем у A_1 . Но модель $M_1 \cap A_1$ не может иметь намного лучшие параметры чем A_1 , так как A_1 достаточная статистика для x (напомним, что параметры A_1 не хуже параметров A , и A является достаточной статистикой для x). Следовательно, $\log |(A_1 \cap M_1)| \approx \log |A_1|$.

Напомним, что M_1 принадлежит простому разбиению $\mathcal{M}$.

Модель можно получить M_1 с помощью тотальной программы из A_1 и своего порядкового номера среди таких множеств $M' \in \mathcal{M}$, что $\log |(A_1 \cap M')| \approx \log |A_1|$. Так как $\mathcal{M}$ — разбиение, то таких множеств $M' \in \mathcal{M}$ будет немного. Следовательно, $CT(M_1|A_1) \approx 0$. $\square$

Теперь мы строго сформулируем и докажем используемые леммы.

Лемма 6.7 (строгая формулировка). Для некоторого $\kappa = O(\log n)$ выполнено следующее. Пусть A является (δ, κ) -минимальной статистикой для слова x длины n . Тогда $C(\Omega_{C(A)}|A) = O(\delta + \log n)$.

Доказательство. Пусть B — стандартная модель x , являющаяся улучшением A , т.е. $\delta(x, B) \leq \delta(x, A) + O(\log n)$; такая модель существует согласно Предложению 2.9. Если $\varkappa$ выбрано надлежащим образом, то $C(B) > C(A) - \delta$. Мы можем оценить $C(\Omega_{C(A)}|A)$ следующим образом

$$C(\Omega_{C(A)}|A) \leq C(\Omega_{C(A)}|\Omega_{C(B)}) + C(\Omega_{C(B)}|B) + C(B|A).$$

Покажем, что каждое слагаемое в правой части этого неравенства равняется $O(\delta + \log n)$. Для третьего слагаемого это верно по построению. Второе слагаемое равняется $O(\log n)$, так как B — стандартная модель. Для первого слагаемого это выполнено, потому что $C(A) < C(B) + \delta$. $\square$

Лемма 6.8 (строгая формулировка). Пусть A является моделью ε , α -нормального слова x длины n и при этом, $\varepsilon \leq n$, $\alpha < \sqrt{n}/2$. Тогда существует такое множество H , что:

- 1) H является ε -сильной статистикой для x ,
- 2) $\delta(x, H) \leq \delta(x, A) + O((\alpha + \log n) \cdot \sqrt{n})$,
- 3) $C(H|\Omega_{C(H)}) = O(\sqrt{n})$,
- 4) $C(H) \leq C(A) + \alpha$.

Доказательство. Рассмотрим последовательность $B_0, A_1, B_1, A_2, B_2, \dots$ моделей для x , которые определяются следующим образом. Определим $B_0 := A$. Для всех i множество A_{i+1} есть сильная статистика для x , полученная из B_i с помощью условия о ε, α -нормальности x :

$$C(A_{i+1}) \leq C(B_i) + \alpha, \quad \log |A_{i+1}| \leq \log |B_i| + \alpha, \quad \text{CT}(A_{i+1}|x) \leq \varepsilon.$$

(См. Рисунок 6.3.)

Для каждого i определим B_i , как стандартную модель, являющуюся улучшением A_i , существующую согласно Предложению 2.9:

$$\delta(B_i, x) \leq \delta(A_i, x) + O(\log n), \quad C(B_i|A_i) = O(\log n).$$

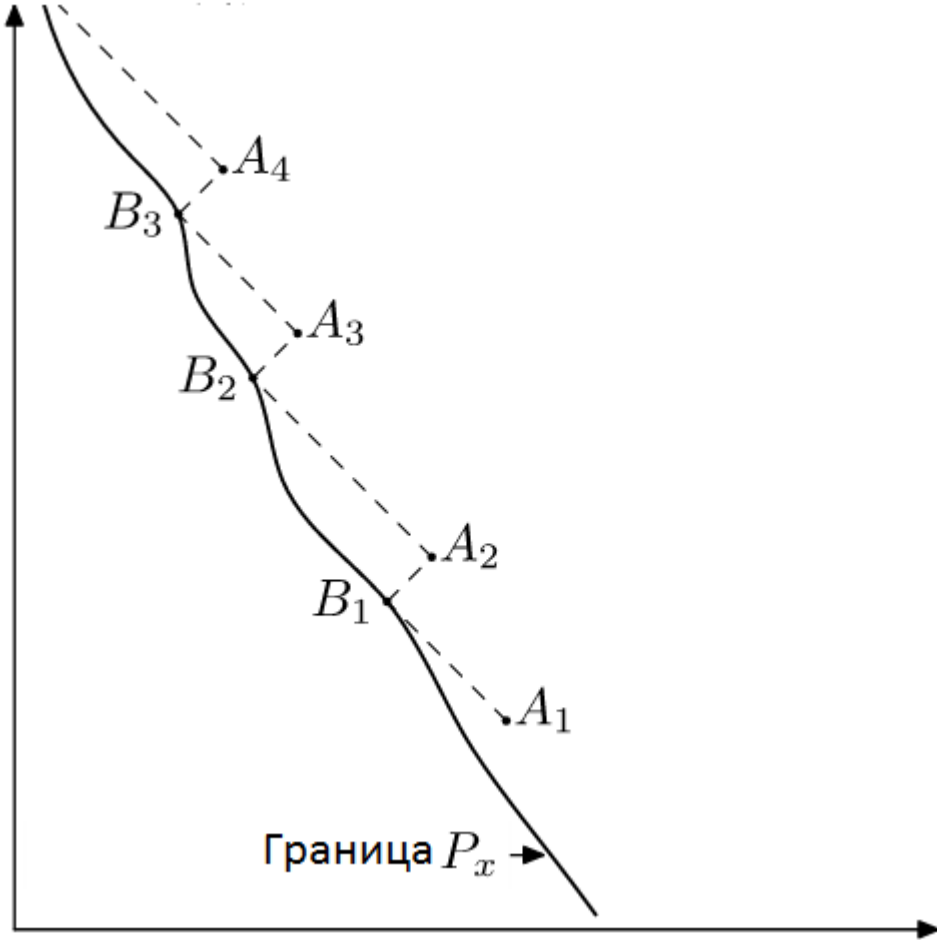


Рис. 6.3: Параметры статистик A_i и B_i

Обозначим через N минимальное число, при котором $C(A_N) - C(B_N) \leq \sqrt{n}$. Таким образом, при всех $i < N$ имеем $C(B_i) < C(A_i) - \sqrt{n}$. С другой стороны, сложность A_{i+1} превосходит сложность B_i не более чем на $\alpha < \sqrt{n}/2$. Таким

образом для всех $i < N$ имеем $C(A_{i+1}) < C(A_i) - \sqrt{n}/2$. Так как $C(A_1) = O(n)$ (напомним, что $CT(A_1|x) \leq \varepsilon \leq n$) и $C(A_N) \geq 0$, получаем $N = O(\sqrt{n})$.

Возьмём $H := A_N$. По построению H сильная модель (первое утверждение леммы) и $C(H) \leq C(A_1) \leq C(A) + \alpha$ (последнее). Из $N = O(\sqrt{n})$ следует, что второе утверждение также выполнено.

Осталось оценить $C(H|\Omega_{C(H)})$. Для этого воспользуемся следующим неравенством:

$$C(A_N|\Omega_{C(A_N)}) \leq C(A_N|B_N) + C(B_N|\Omega_{C(B_N)}) + C(\Omega_{C(B_N)}|\Omega_{C(A_N)}).$$

Нам нужно показать, что все слагаемые в правой части равняются $O(\sqrt{n})$. Это верно для первого слагаемого, потому что N выбрано таким образом, что $C(A_N) - C(B_N) \leq \sqrt{n}$. По построению $C(B_N|A_N) = O(\log n)$. Теперь мы можем оценить $C(A_N|B_N)$, используя коммутативность информации:

$$C(A_N|B_N) = C(A_N) + C(B_N|A_N) - C(B_N) = C(A_N) - C(B_N) \leq \sqrt{n}$$

(с логарифмической точностью).

Второе слагаемое есть $O(\log n)$ по построению (напомним, что B_N является стандартной моделью). Чтобы оценить третье слагаемое заметим, что по построению $C(B_N|A_N) = O(\log n)$, а значит, $C(B_N) - C(A_N) = O(\log n)$. $\square$

Доказательство Теоремы 6.5. Для завершения доказательства Теоремы 6.5 нам нужно проверить, что все приближительные равенства и неравенства, используемые в эскизе доказательства выполняются с точностью $O(\delta + (\varepsilon + \log n)\sqrt{n})$.

В построении A_1 мы используем Предложение 6.6. Следовательно, в неравенствах, связанных с A и A_1 добавочное слагаемое имеет порядок $O(\varepsilon + \log n)$. Сложность разбиения имеет такой же порядок.

Построение L : добавочное слагаемое в неравенстве $C(L) \leq C(\mathcal{G})$ имеет порядок $O(\log \log |A_1|) = O(\log n)$, а добавочное слагаемое во втором слагаемом оценивается константой.

Для конструкции M мы используем Лемму 6.8 для $\alpha = \varepsilon$ (мы можем считать, что условие $\varepsilon < \sqrt{n}/2$ для этой леммы выполнено, в противном случае утверждение теоремы очевидно).

Добавочные слагаемые в Лемме 6.8 имеют порядок $O((\varepsilon + \log n)\sqrt{n})$ и, следовательно, параметры M (сложность, сложность + логарифм размера) превышают параметры L не более чем на $O((\varepsilon + \log n)\sqrt{n})$.

Далее мы используем Лемму 6.7, чтобы оценить $C(\Omega_{C(A)}|A)$ как $O(\delta + \log n)$. Таким образом, добавочное слагаемое в равенстве $C(M|A_1) \approx 0$ равняется $O(\delta + (\varepsilon + \log n)\sqrt{n})$.

В конструкции M_1 мы используем Предложение 6.6. Следовательно, оба выражения $CT(M_1|M)$ и $CT(M|M_1)$ равняются $O(\varepsilon + \log n)$. Сложность $A_1 \cap M_1$ превышает сложность A_1 не более чем на $O(\delta + (\varepsilon + \log n)\sqrt{n})$, а значит, равенство $\log |(A_1 \cap M_1)| \approx \log |A_1|$ выполняется с точностью $O(\delta + (\varepsilon + \log n)\sqrt{n})$. Из этого следует, что равенство $CT(M_1|A_1) \approx 0$ выполняется с той же точностью.

Конструкция $\mathcal{F}$: как мы увидели, равенство $CT(M_1|A_1) \approx 0$ выполняется с точностью $O(\delta + (\varepsilon + \log n)\sqrt{n})$. Равенство $CT(\mathcal{F}|M_1) \approx 0$ выполняется с точностью $O(\varepsilon + \log n)$, так как сложность $\mathcal{A}$ равняется $O(\varepsilon + \log n)$. Добавочные члены в других неравенствах такие же, как в предыдущих этапах доказательства и, таким образом, имеют порядок $O(\delta + (\varepsilon + \log n)\sqrt{n})$. $\square$

Заключение

Дальнейшее развитие алгоритмической статистики автор видит в следующих трёх направлениях.

Во-первых, это решение открытых вопросов в теории сильных статистик и в теории ограниченных классах гипотез. Эти вопросы подробно описаны в [27]. Также можно попытаться улучшить точность в имеющихся результатах (например, в Теоремах 2.13 и 6.5). Как было отмечено в [10], в алгоритмической теории информации редко встречаются естественные утверждения, которые выполнены с точностью $o(\cdot)$, но не выполнены с логарифмической точностью — это даёт надежду на усиление приведённых теорем.

Второе направление связано со сложностью с ограничением на ресурсы. Как отмечал А.Н. Колмогоров [8], объект, который может быть порождён хоть и короткой, но долго работающей программой, не следует считать простым. То же самое относится и к “простым” объяснениям в алгоритмической статистике. Поэтому резонно ограничиться программами, которые работают разумное время и используют не очень много памяти. Этот подход в алгоритмической статистике только начинает развиваться, здесь ещё очень много открытых вопросов. Некоторые из них связаны с фундаментальными проблемами сложности вычислений (такими как проблема перебора).

Наконец, третьим (наиболее туманным) направлением является установление связей между алгоритмической статистикой и машинным обучением, задачами предсказания. Ведь поводом для создания этой теории была попытка формализации понятия “хорошее объяснение для наблюдаемых данных”, а не теория рекурсии или сложность вычислений. Поэтому хочется надеяться, что алгоритмическая статистика сможет вернуться к истокам и, обогащенная новыми результатами, помочь статистикам в решении их задач. Вопросы, на которые здесь хочется ответить не являются математическими гипотезами: почему мы предпочитаем простые объяснения сложным? Есть ли у этого физическое обоснование? Какой способ измерения сложности является самым “правильным” (простая сложность, сложность с ограничением на время или нечто совсем другое)? На эти вопросы у нас пока нет ответов.

Литература

1. Н.К. Верещагин, В.А. Успенский, А. Шень, *Колмогоровская сложность и алгоритмическая случайность*. МЦНМО, 2013.—576 с.
2. В. В. Вьюгин, Алгебра инвариантных свойств двоичных последовательностей, *Проблемы передачи информации*, **18**(2), 83–100 (1982).
3. В.В. Вьюгин, О дефекте случайности конечного объекта относительно мер с заданными границами их сложности, *Теория вероятностей и ее применение*, **32**(6), 558–563, (1987).
4. В. В. Вьюгин, О нестохастических объектах, *Проблемы передачи информации*, **21**(2), 3–9, (1985).
5. Гач П., О симметрии алгоритмической информации. *Доклады Академии наук СССР*, **218**(6), 1265–1267 (1974).
6. А.Н. Колмогоров. Доклад на семинаре механико-математического факультета Московского Государственного Университета кафедры Математической логики, 26 ноября 1981.
7. А.Н.Колмогоров, Сложность алгоритмов и объективное определение случайности. Заседание Московского математического общества, *Успехи математических наук*, **29**(4[178]), 155 (1974).<http://mi.mathnet.ru/rus/umn/v29/i4/p153>.
8. Колмогоров А.Н., Три подхода к определению понятия “количество информации”, *Проблемы передачи информации*, **1**(1), 4–11 (1965).
9. Левин Л. А., О понятии случайной последовательности. *Доклады Академии наук СССР*, **212**(3), с. 548–550 (1973).
10. Мучник Ан.А., Ромащенко А.Е., Устойчивость колмогоровских свойств при релятивизации. *Проблемы передачи информации*, **46**(1), 42–67 (2010).
11. А. Х. Шень. Понятие (α, β) -стохастичности по Колмогорову и его свойства. *Доклады Академии наук СССР*, **271**(6), 1337–1340 (1983).
12. A. Chernov, An. Muchnik, A. Romashchenko, A. Shen, and N. Vereshchagin. Upper semi-lattice of binary strings with the relation “x is simple conditional to y”. *Theoretical Computer Science* **271**, 69–95, (2002).
13. Thomas M. Cover, Peter Gács, Robert M. Gray, Kolmogorov’s Contributions to Information Theory and Algorithmic Complexity, *The Annals of Probability*, **17**(3), 840–865 (1989).

14. P. Gács, J. Tromp, P.M.B. Vitányi, Algorithmic statistics, *IEEE Transactions on Information Theory*, **47**(6), 2443–2463 (2001).
15. V. Guruswami, *List decoding of error-correcting codes: winning thesis of the 2002 ACM doctoral dissertation competition*, Springer, 2004.
16. T. Lee and A. Romashchenko, Resource bounded symmetry of information revisited, *Theoretical Computer Science*, **345**(2–3), 386–405 (2005).
17. Leonid A. Levin, Randomness conservation inequalities; information and independence in mathematical theories, *Information and Control* **61**(1), 15–37, (1984).
18. Levin L.A., V'yugin V.V. Invariant properties of informational bulks, *Lecture Notes on Computer Science*, 1977, **53**, 359–364.
19. Li M., Vitányi P., *An Introduction to Kolmogorov complexity and its applications*, 3rd ed., Springer, 2008 (1 ed., 1993; 2 ed., 1997), xxiii+790 pp. ISBN 978-0-387-49820-1.
20. L. Longpré and S. Mocas, Symmetry of information and one-way functions. *Information Processing Letters*, 46(2), 95–100 (1993).
21. L. Longpré and O. Watanabe, On symmetry of information and polynomial time invertibility. *Information and Computation*, **121**(1):1–22 (1995).
22. S. de Rooij, P.M.B. Vitanyi, Approximating Rate-Distortion Graphs of Individual Data: Experiments in Lossy Compression and Denoising, *IEEE Transaction on Computers*, **61**(3), 395–407 (2012).
23. A. Shen, Around Kolmogorov complexity: basic notions and results. *Measures of Complexity. Festschrift for Alexey Chervonenkis*. Editors: V. Vovk, H. Papadoupoulos, A. Gammerman. Springer, 2015. ISBN: 978-3-319-21851-9.
24. A. Shen, Game Arguments in Computability Theory and Algorithmic Information Theory. *Proceedings of Computability in Europe (CiE) 2012*, 655–666.
25. N. Vereshchagin, Algorithmic Minimal Sufficient Statistics: a New Approach. *Theory of Computing Systems*, **58**(3), 463–481 (2016).
26. Nikolay Vereshchagin. On Algorithmic Strong Sufficient Statistics.. In: *9th Conference on Computability in Europe, CiE 2013, Milan, Italy, July 1–5, 2013. Proceedings*, LNCS 7921, 424–433.
27. Nikolay K. Vereshchagin, Alexander Shen, Algorithmic Statistics: Forty Years Later. *Computability and Complexity*: 669–737 (2017).
28. N.K. Vereshchagin, P.M.B. Vitányi, Kolmogorov’s structure functions and model selection, *FOCS 2002*: 751–760.
29. N.K. Vereshchagin, P.M.B. Vitányi, Kolmogorov’s structure functions and model selection, *IEEE Transactions on Information Theory*, **50**(12), 3265–3290 (2004).

30. N.K. Vereshchagin, P.M.B. Vitányi. Rate Distortion and Denoising of Individual Data Using Kolmogorov Complexity. *IEEE Transactions on Information Theory*, **56**(7), 3438–3454 (2010).
31. V.V. V'yugin, Algorithmic complexity and stochastic properties of finite binary sequences, *The Computer Journal*, 1999, **42**(4), 294–317 (1999).

Работы автора по теме диссертации

32. А.С. Милованов, Некоторые свойства антистохастических слов. Материалы XXI Международной научной конференции “Ломоносов- 2016”, Издательство “Макс-пресс” (2016), http://universiade.msu.ru/archive/Lomonosov_2016/data/8440/uid57721_report.pdf.
33. A. Milovanov, Some properties of antistochastic strings. In: *Computer Science – Theory and Applications, 10th International Computer Science Symposium in Russia, Russia, July 13–17, 2015*. (CSR 2015), Lecture Notes in Computer Science, **9139**, 339–349.
34. A. Milovanov, Algorithmic statistic, prediction and machine learning, *33rd Symposium on Theoretical Aspects of Computer Science* (STACS 2016), Leibnitz International Proceedings in Informatics (LIPIcs), **47**, 2016, DOI 10.4230/LIPIcs.STACS.2016.54, <http://drops.dagstuhl.de/opus/volltexte/2016/5755/>, 54:1–54:13.
35. A. Milovanov, Algorithmic Statistics: Normal Objects and Universal Models, *Computer Science – Theory and Applications*, Proceedings of CSR 2016 conference, Lecture Notes in Computer Science, **9691**, 280–293.