ФГБОУ ВО
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

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва — 2017

Работа выполнена на кафедре математической логики и теории алгоритмов механико-математического факультета ФГБОУ ВО Московского государственного университета имени М.В. Ломоносова

Научный руководитель: Верещагин Николай Константинович,
доктор физико-математических наук, профессор.

Официальные оппоненты: Вьюгин Владимир Вячеславович,
доктор физико-математических наук, профессор,
заведующий лабараторией №1 им. М.С. Пинскера
"Теория передачи информации и управления"
Института проблем передачи инфорамции
им. А.А. Харкевича Российской академии наук

Мусатов Даниил Владимирович,
кандидат физико-матиматических наук,
доцент кафедры дискретной математики
московского физико-технического института
(государственного университета)

Рябко Борис Яковлевич,
доктор технических наук, профессор
заведующий лабораторией информационных
систем и защиты информации
Института вычислительных технологий СО РАН

Защита диссертации состоится 8 декабря 2017 г. в 16 часов 45 минут на заседании диссертационного совета МГУ.01.17 Московского государственного университета имени М.В. Ломоносова, по адресу: 119234, Москва, ГСП-1, Ленинские горы д.1, ФГБОУ ВО Московский Государственный Университет имени М.В. Ломоносова, механико-математический факультет, аудитория 14-08.

E-mail: msu.01.17@mail.ru

С диссертацией можно ознакомиться в читальном зале отдела диссертаций Фундаментальной библиотеки ФГБОУ ВО МГУ имени М.В. Ломоносова по адресу: Ломоносовский проспект, д. 27 и на сайте ИАС «Истина»:
`https://istina.msu.ru/dissertations/81091858/`

Автореферат разослан 8 ноября 2017 года.

Ученый секретарь диссертационного
совета МГУ.01.17 ФГБОУ ВО МГУ,
доктор физико-математических наук,
профессор, член-корр. РАН Шафаревич А.И

Общая характеристика работы

Актуальность темы

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 определения этим понятиям в 1974-ом году, 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 $\mathrm{C}(x)$. Условная сложность x относительно y, т.е. длина кратчайшей программы, которая печатает x на входе y обозначается через $\mathrm{C}(x|y)$.

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

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

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

Через некоторое время А.Н. Колмогоров задал вопрос: для любого ли слова найдётся хорошее объяснение? 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 $\mathrm{C}(A) \leq \alpha$ и $d(x|A) \leq \beta$.

А.Х. Шень доказал, 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 и нижние оценки доли нестохастических объектов заданной длины. Там же получены верхняя оценка меры таких объектов для вычислимого распределения ограниченной сложности. В.В. Вьюгин получил аналогичные верхние и нижние оценки для универсальной полувычислимой меры множества нестохастических объектов данной длины. В. В. Вьюгин выявил связь нестохастичности для конечных и бесконечных последовательностей.

В работе В.В. Вьюгина были изучены формы кривых зависимости дефекта случайности β относительно мер сложности не превосходящей α для случая когда $\alpha = o(\beta)$. Этот результат был обобщен на произвольный случай в работе Верещагина и Витаньи .

Вьюгиным и Левиным изучалась нестохастичность для бесконечных последовательностей.

Отметим также работу Гача, Тромпа и Витаньи, 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.

Цели и задачи работы

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

Основные результаты

Работа содержит четыре основных результата.

Во-первых были изучены свойства максимально нестохастических слов (существование которых было доказано Верещагиным и Витаньи), они также называются *антистохастическими*. Доказано, что эти объекты (которые являются самыми сложными для объяснения с точки зрения алгоритмической статистики), обладают свойством *голографичности*. 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 и построения контрпримеров в алгоритмической теории информации.

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 алгоритимической статистикой.

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.

В-четвёртых, были получены результаты о *нормальных* словах. Концепция нормальных объектов была введена Н.К. Верещагиным, 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 моделями 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, доказано, что они обладают свойством *наследственности*.

Основные методы исследования

В работе использованы комбинаторные и игровые методы доказательств. Использованы такие результаты теории колмогоровской сложности, как коммутативность информации, случайность (по Мартину-Лёфу), числа Чейтина.

Научная новизна

Все результаты, полученные в диссертационной работе, являются новыми.

Исследованы свойства максимально нестохастических слов, разработан новых подход в теории предсказаний, получены ответы на некоторые вопросы из теории сильных моделей алгоритмической статистики.

Теоретическая и практическая ценность

Работа имеет теоретический характер. Полученные результаты могут представлять интерес для специалистов в области алгоритмической теории информации и математической статистики. Результаты диссертации могут быть использованы в исследованиях, проводимых в Московском государственном университете имени М.В. Ломоносова, Московском физико-техническом институте и в других научно-исследовательских центрах.

Апробация работы

Основные результаты, представленные в диссертации, докладывались автором и обсуждались на следующих конференциях и научных семинарах:

- “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
- 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).

Публикации

Основные результаты работы опубликованы в сборниках трудов конференции конференций “International Computer Science Symposium in Russia” и “Symposium on Theoretical Aspects of Computer Science”[2-4] (Серии Lecture Notes in Computer Science и Leibniz International Proceedings in Informatics входят в систему Scopus). Также список работ включает опубликованный тезис Международной научной конференции студентов, аспирантов и молодых учёных “Ломоносов-2016”[1].

Список работ приведён в конце автореферата.

Структура диссертации

Диссертация состоит из введения, обзора основных понятий, четырёх глав с изложением результатов работы, заключения и списка литературы. Полный объём диссертации составляет 65 страниц. Список литературы содержит 35 наименований. В диссертации приведено 7 рисунков.

Основное содержание работы

Глава 1 является введением диссертации. Она содержит описание актуальности темы, цели работы, список основных результатов, сведения об апробации работы и список используемых обозначений.

Глава 2 представляет собой обзор основных результатов алгоритмической статистики (известных до работ автора). Мы приведём здесь лишь некоторые ключевые понятия и результаты.

Профиль слова

Рассмотрим множество слов A содержащее некоторое слово x. Мы уже обсуждали, что приемлемость A, как объяснения для x, можно измерить с помощью параметров $\mathrm{C}(A)$ и $d(x|A) = \log|A| - \mathrm{C}(x|A)$. Однако, адекватность объяснения A можно измерить и в несколько иной системе координат: $\mathrm{C}(A)$ и $\log|A|$. (На самом деле, между этими двумя подходами существует прямая связь.) Величина $\mathrm{C}(A) + \log|A| - \mathrm{C}(x)$ называется *дефектом оптимальности* x в A и обозначается как $\delta(x, A)$.

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 $\mathrm{C}(A) \le m$ и $\log|A| \le l$.

Рисунок 1 показывает как может выглядеть профиль слова x длины n и сложности k.

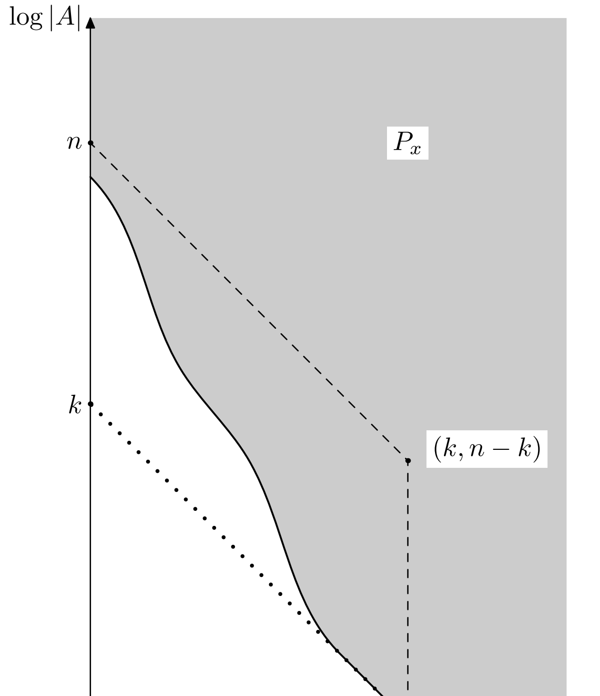

Рис. 1: Возможный профиль P_x слова x длины n и сложности k.

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 $m' \ge m$ и $l' \ge l$.

- Во-вторых, P_x содержит множество

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

(это множество всех точек, находящихся сверху и справа от пунктирной линии на Рис. 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. 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-третьих, профиль P_x обладает следующим свойством:

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

Оказывается, 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 1 (Верещагин, Витаньи). *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 $\mathrm{C}(P) + O(\log n)$-близок к P.

Стандартные модели

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:

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

где $t_1 > t_2 > \ldots$

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 $m \geq \mathrm{C}(x)$ мы получили некоторое множество, содержащее x. Такие множества называются *стандартными моделями* или *стандартными описаниями* для x.

Оказывается, по любому множеству $A \ni x$ можно построить стандартную модель $B \ni x$, которая будет не хуже A в качестве объяснения для x (в смысле параметров обсуждемых ранее). Тем не менее, стандартные описания не выглядят адекватными с интуитивной точки зрения. Чтобы решить эту проблему нужно изменить определение хорошего описания для слова.

Сильные модели

Неформально, тотальная условная сложность определяется как длина кратчайшей программы, которая переводит одно слово в другое и при этом останавливается на любом входе. Дадим точное определение.

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

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

(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. Зафиксировав такой универсальный способ описания, определим *тотальную условную сложность* слова x относительно слова y как $\mathrm{CT}_D(x|y)$. Опуская индекс D, эту величину обозначают через $\mathrm{CT}(x|y)$.

Величина $\mathrm{CT}(x|y)$ может быть гораздо больше чем $\mathrm{C}(x|y)$.

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

Н.К. Верещагин предложил рассмотреть новый параметр, измеряющий качество множества $A \ni x$, как объяснения для x, а именно $\mathrm{CT}(A|x)$. Для "хороших" объяснений эта величина должна быть мала. Множества $A \ni x$, для которых $\mathrm{CT}(A|x) \approx 0$ 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 $\mathrm{CT}(A|x) < \varepsilon$.

Сильный профиль слова x определяется как множество параметров всех сильных статистик для x.

Определение 4. Множество пар целых чисел

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

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*. Н.К. Верещагин показал, что существуют странные слова с линейными параметрами.

Теорема 2 (Верещагин). *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. (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.)

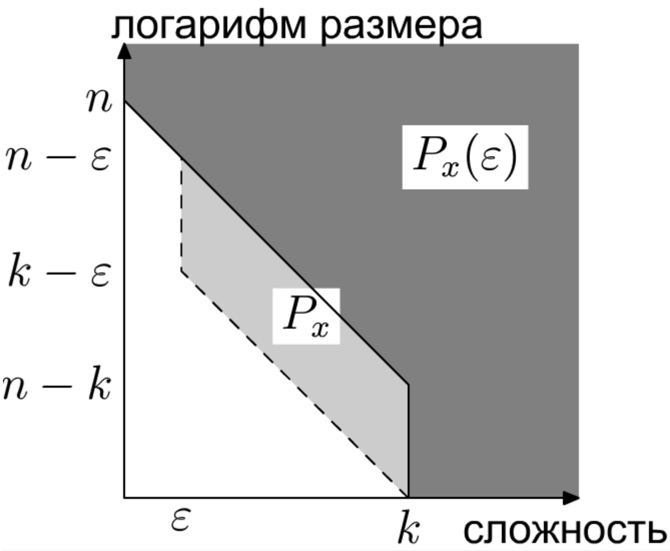

Рис. 2: Множества P_x и $P_x(\varepsilon)$ для странных слов из Теоремы 2 (с логарифмической точностью). Множество P_x расположено справа от пунктирной линии. Множество $P_x(\varepsilon)$ расположено справа от сплошной линии.

Глава 3 посвящена свойствам максимально нестохастических или *антистохастических* слов. Определим это понятие формально.

Определение 6. Слово x длины n и сложности k называется ε-*антистохастичным*, 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).

Существование антистохастических слов следует из Теоремы 1. Оказывается, что антистохастическое слово сложности k можно восстановить по любым содержащимся в ней k битам информации. А именно, верно следующее утверждение.

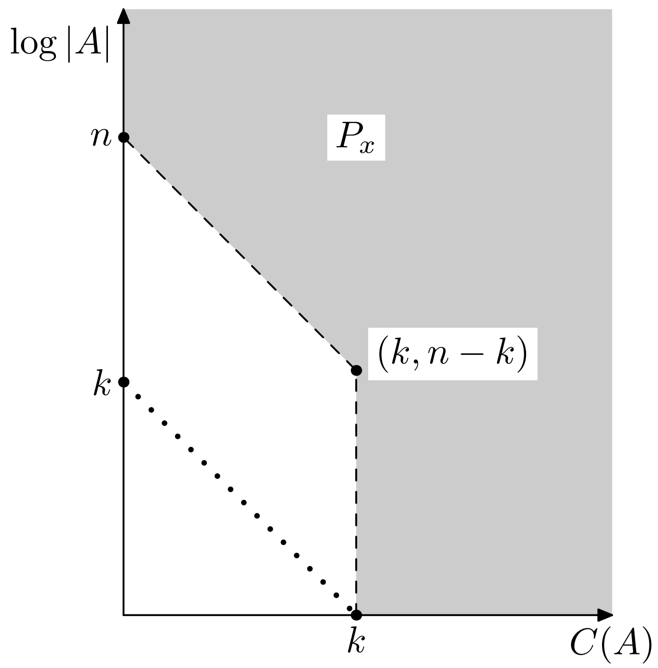

Рис. 3: Профиль ε-антистохастического слова x при маленьком ε близок к множеству $P_{\min}$.

Теорема 3. *Пусть x является ε-антистохастическим словом длины n и сложности k. Предположим, что множество A содержит x и $|A| \leq 2^{n-k}$. Тогда $\mathrm{C}(x|A) \leq 2\varepsilon + O(\log \mathrm{C}(A) + \log n)$.*

Этот результат позволяет использовать антистохастические объекты для построения кодов, исправляющих стирания списком. Также их можно использовать для получения результатов в области тотальной условной сложности.

Теорема 4. *Для всех k и n существуют такие слова $x_1, \ldots, x_k$ длины n, что:*
(1) $\mathrm{C}(x_i|x_j) = O(\log k + \log n)$ *для всех i и j.*
(2) $\mathrm{CT}(x_i|x_1 \ldots x_{i-1}x_{i+1} \ldots x_k) \geq n - O(\log k + \log n)$ *для каждого i.*

Глава 4 посвящена связи алгоритмической статистики с теорией предсказания.

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 бо́льшую вероятность, чем сложные. Чтобы задать такое распределение формально, нам потребуется ещё один инструмент из алгоритмической теории сложности — *дискретная априорная вероятность*.

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 5 (Левин)**.** *Существует такая машина U, что для любой машины V существует такая положительная константа D, что*

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

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*.

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

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

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,

$$p_x(y) = \frac{\sum_{A \ni x,y} \mathbf{m}(A)/|A|}{\sum_{A \ni x} \mathbf{m}(A)/|A|}. \qquad (1)$$

По данному слову x и некоторому пороговому значению d можно составить множество всех таких слову y, что $p_x(y) \ge 2^{-d}$. 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 четвёртой 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 $\mathrm{poly}(n)$; для некоторого достаточно большого полинома (на самом деле, $4n + O(1)$ достаточно). Т.е. мы теперь предполагаем, что суммы в (1) ограничены множествами $A \subseteq \{0,1\}^n$ и имеющих сложность не больше $4n + O(1)$, и рассматриваем эту модифицированную версию (1) как финальное определение $p_x(y)$.

Определение 7. 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.*

Алгоритмическая предсказательная окрестность

Пусть слово x содержится в конечном множестве A. Напомним, что величина $\mathrm{C}(A) + \log|A| - \mathrm{C}(x)$ 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 8. 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 1. Если $x = 0\ldots0$ (слово, состоящее из n нулей), тогда x' принадлежит d-окрестности x, если и только если $\mathrm{C}(x') \lesssim d$.

Пример 2. Если x случайное слово длины n (т.е. $\mathrm{C}(x) \approx n$) 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 6. (a) *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.*

(b) *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.*

Глава 5 посвящена алгоритмической статистикой для нескольких слов. Оказывается, основные результаты алгоритмической статистики (например, связь дефектов случайности и оптимальности и Теорему 6) можно обобщить на этот случай.

Глава 6 посвящена результатам о нормальных словах. Во-первых, это результат об их существовании — аналог Теоремы 1.

Теорема 7. *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 $\mathrm{C}(P) + O(\sqrt{n \log n})$-близок к P.

Напомним, что сильные модели были введены, чтобы разделить хорошие модели от плохих. При этом, под плохими статистиками имеются ввиду стандартные модели, о которых говорилось ранее. Следующая теорема показывает, что существует такое слово x и его сильная модель A, что у любой сильной *стандартной* модели для x параметры гораздо хуже чем у A.

Теорема 8. *Для некоторого положительного с 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 4, что*

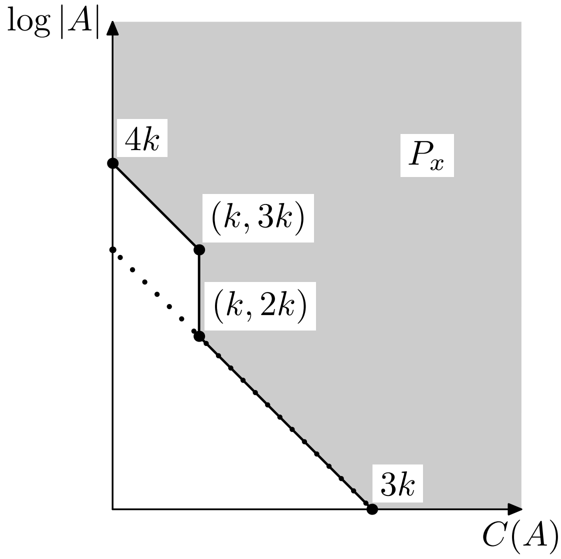

Рис. 4: Профиль P_x слова x из Теоремы 8.

- *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*

- *для каждого $m \geq \mathrm{C}(x)$ и для каждого простого перечисления слов сложности не больше m 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 $\mathrm{CT}(B|x), \mathrm{C}(q), |\mathrm{C}(B) - k|, |\log|B| - 2k|$ больше чем ck.*

Последним результатом является утверждение о наследственности нормальных слов. Чтобы его сформулировать введём следующие определения.

Определение 9. Конечное множество $A \ni x$ называется *ε-достаточной статистикой* для x, если $\delta(x, A) \leq \varepsilon$.

Определение 10. Модель $A \ni x$ называется $(\delta, \varkappa)$-*минимальной*, если не существует такого множества $B \ni x$, что $\mathrm{C}(B) \le \mathrm{C}(A) - \delta$ и $\delta(x, B) \le \delta(x, A) + \varkappa$. Чем меньше δ и больше $\varkappa$, тем сильнее свойство $(\delta, \varkappa)$-минимальности.

Теорема 9. *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 словом*[1].

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. Также можно пытаться улучшить точность в имеющихся результатах (например, в Теореме 9). 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 вышеупомянутой теоремы.

Второе направление связано со сложностью с ограничением на ресурсы. Как отмечал А.Н. Колмогоров, 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 вернутся 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.

[1] Строго 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 биекциии между двоичными словами и конечными множествами двоичных слов.

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.

Публикации автора по теме диссертации

1. А.С. Милованов, Некотороые свойства антистохастических слов. Материалы XXI Международнародной научной конференции “Ломоносов- 2016”, Издательство “Макс-пресс” (2016), http://universiade.msu.ru/archive/Lomonosov_2016/data/8440/uid57721_report.pdf.

2. 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.

3. 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.

4. 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.