Выбор метрики для оценки энтропии по методу ближайшей точки*

Е. А. ТИМОФЕЕВ
Ярославский государственный университет им. П. Г. Демидова
e-mail: TimofeevEA@gmail.com

УДК 519.72+519.234

Ключевые слова: энтропия, стохастический процесс, стационарность, непараметрическая оценка, метрика, смещение.

Аннотация

Рассматривается задача повышения эффективности непараметрических оценок энтропии стационарного эргодического процесса. Изучаемый подход основан на методе ближайшей точки. Предлагается новый широкий класс метрик на пространстве правосторонних последовательностей, состоящих из символов конечного алфавита. Эти метрики имеют параметр — невозрастающую функцию. Показано, что при некоторых условиях предлагаемые оценки имеют малую дисперсию и что при специальном выборе параметров метрики можно уменьшить смещение.

Abstract

E. A. Timofeev, Selection of a metric for the nearest neighbor entropy estimators, Fundamentalnaya i prikladnaya matematika, vol. 18 (2013), no. 2, pp. 209—227.

We consider the problem of improving the efficiency of the nonparametric entropy estimation for a stationary ergodic process. Our approach is based on the nearest-neighbor distances. We propose a broad class of metrics on the space of right-sided infinite sequences drawn from a finite alphabet. The new metric has a parameter which is a non-increasing function. We prove that, under certain conditions, our estimators have a small variance and show that a special selection of the metric parameters reduces the estimator's bias.

1. Введение

Работа посвящена повышению эффективности (точности) оценивания энтропии информационного источника с конечным числом состояний. Энтропия (энтропия на символ) является важнейшей характеристикой информационного источника. Поэтому задача оценивания энтропии представляет значительный интерес.

*Работа выполнена при поддержке гранта Правительства РФ по постановлению № 220, договор 11.G34.31.0053.

Фундаментальная и прикладная математика, 2013, том 18, № 2, с. 209—227.

Поскольку статистические характеристики информационного источника обычно не известны, наиболее широко используются так называемые *непараметрические* оценки энтропии. Однако аналитически определить точность этих оценок очень трудно, и известно очень мало результатов в этом направлении. В большинстве опубликованных работ по непараметрическому оцениванию энтропии доказывается только асимптотическая сходимость и приводятся результаты компьютерного тестирования.

Напомним, что для заданной выборки размера n наиболее важной характеристикой оценки h_n является её эффективность (точность), или квадратичная ошибка $\mathsf{E}(h_n - h)^2$, где h — энтропия источника, а n — число измерений. Напомним соотношение

$$\mathsf{E}(h_n - h)^2 = \mathsf{Var}\, h_n + (\mathsf{E}h_n - h)^2,$$

где величина $\mathsf{E}(h_n - h)$ называется *смещением*.

Цель настоящей работы — построение новой оценки, основывающейся на методе ближайшей точки и его модификациях, у которой точность будет равна $O(n^{-c})$ для широкого класса мер, где $c > 0$ — некоторая константа.

Основная идея предлагаемой конструкции состоит в следующем.

Оценки энтропии, основанные на методе ближайшей точки, используют некоторую метрику. Будет введён довольно широкий класс так называемых *слабых* [3] метрик, для которых неравенство треугольника выполняется с некоторой константой $C > 1$.

Эта метрика применяется в построении оценки энтропии по методу ближайшей точки. Построенная оценка будет линейно зависеть от параметров метрики. А эти параметры подбираются так, чтобы минимизировать среднее отклонение.

Отметим, что задачи выбора параметров с линейными ограничениями довольно часто возникают в статистике [13].

Подчеркнём, что наиболее известные непараметрические оценки энтропии основаны либо на методе сжатия Лемпеля—Зива, либо на методе ближайшей точки. Для этих оценок доказана сходимость почти всюду (наиболее общие результаты приводятся, например, в [6, 9]). К сожалению, из-за медленной скорости сходимости (не более чем $O(1/\ln n)$) их точность не является достаточной для большинства практических приложений, поскольку логарифм размера выборки ($\log_2 n$) довольно небольшой: $\log_2 n \leqslant 30 - 40$. Этим объясняется поиск новых оценок с большей скоростью сходимости.

Следует отметить, что оценки энтропии, основанные на методе Лемпеля—Зива, очень трудны для теоретического исследования их сходимости и получения аналитических выражений для точности. Так, например, первоначальной мотивацией работы [1] была задача получения асимптотического значения скорости сходимости оценок энтропии, введённых Зивом [16], но авторы только смогли указать, что нахождение смещения и дисперсии очень трудно даже для простейшей меры Бернулли. На сегодняшний день нет опубликованных работ, в которых были бы найдены смещение или дисперсия оценки энтропии, построенной по методу Лемпеля—Зива.

Далее будут рассматриваться только оценки, построенные по методу ближайшей точки.

Заметим, что в то время как многие оценки применяются к одной длинной последовательности, оценки, основанные на методе ближайшей точки, применяются к большому числу независимых коротких последовательностей. Эти два подхода эквивалентны, если информационный источник имеет память размера порядка $\ln n$. Действительно, пусть дана выборка $x_1, x_2, \ldots, x_N$, $x_i \in \mathcal{A}$, где $\mathcal{A}$ — алфавит информационного источника. Выберем достаточно большие числа $m, l \approx C \ln n$, положим $n = [(N-m)/(m+l)] + 1$ и определим точки $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n$, положив

$$\boldsymbol{\xi}_i = \big(x_{(m+l)(i-1)+1}, x_{(m+l)(i-1)+2}, \ldots, x_{(m+l)(i-1)+m}\big), \quad i = 1, 2, \ldots, n.$$

Теперь кратко опишем основные известные результаты. Сразу отметим, что удобнее оценивать не энтропию h, а обратную величину $1/h$.

Будем считать, что информационный источник, или стационарный процесс, является инвариантной относительно сдвига эргодической мерой μ на пространстве $\Omega = A^{\mathbb{N}}$, состоящем из правосторонних бесконечных последовательностей символов из конечного алфавита A, где $\mathbb{N} = \{1, 2, \ldots\}$. Таким образом, бесконечную случайную последовательность будем рассматривать как случайную точку в Ω распределённую по мере μ.

Пусть

$$\rho_0(\boldsymbol{x}, \boldsymbol{y}) = \max\{e^{-k} \colon x_k \neq y_k\} - \tag{1}$$

метрика на пространстве Ω, где $\boldsymbol{x} = (x_1, x_2, \ldots)$, $\boldsymbol{y} = (y_1, y_2, \ldots)$.

Пусть

$$\boldsymbol{\xi}_0 = (\xi_{01}, \ldots, \xi_{0m}, \ldots), \ldots, \boldsymbol{\xi}_n = (\xi_{n1}, \ldots, \xi_{nm}, \ldots) -$$

$n+1$ случайных точек в Ω. Тогда первая оценка (величины, обратной к энтропии), построенная по методу ближайшей точки [5], определяется следующим образом:

$$\tilde{h}_n = -\frac{1}{(n+1)\ln n} \sum_{j=0}^{n} \ln\left(\min_{i\colon i \neq j} \rho_0(\boldsymbol{\xi}_i, \boldsymbol{x}_j)\right), \tag{2}$$

здесь и далее в работе все логарифмы будут рассматриваться с натуральным основанием.

Позже П. Шилдс [12] показал, что оценка Грассбергера не является состоятельной для общих эргодических процессов, но справедлива для неприводимых цепей Маркова.

Две модификации оценки Грассбергера были предложены в [14]. В обозначениях настоящей работы их можно записать как $r_n^{(k)}(\rho)/\ln n$ (см. (27)) и $\eta_n^{(k)}(\rho)$ (см. (26)), где ρ — некоторая метрика.

Для оценки $r_n^{(k)}(\rho)/\ln n$ L^1-сходимость и оценка дисперсии $O(n^{-c})$ доказаны в [14] при довольно общих ограничениях на метрику ρ и меру μ. В [6] для метрики (1) ограничения на меру существенно упрощены (см. (3)) и доказана

сходимость почти всюду. Также в [6] доказана оценка дисперсии $O(n^{-c})$ для любого $c < 1$.

Для оценки $\eta_n^{(k)}(\rho)$ L^1-сходимость показана в [14] при более сильных ограничениях на меру и метрику.

Для метрики (1) вычислительный эксперимент, приведённый в [6], показал, что оценка (26) намного эффективнее, чем оценка $r_n^{(k)}(\rho_0)/\ln n$. Однако в следующей работе [7] для симметричных мер Бернулли было показано, что смещение оценки (26) является периодической функцией с периодом пропорциональным $\ln n$. В вычислительных экспериментах это смещение очень трудно заметить, поскольку его амплитуда не более чем 10^{-6} для мер с небольшой энтропией ($h < 3$).

В [15] для смещения найдено аналитическое выражение для марковских мер и метрики (1). Это смещение равно нулю, если логарифмы вероятностей перехода рационально несоизмеримы. В противном случае смещение является периодической функцией с периодом пропорциональным $\ln n$.

Этот результат показывает новую трудность в аналитических вычислениях смещения, а именно что смещение оценки может быть разрывной функцией от параметров меры.

Статья организована следующим образом:

— в разделе 2 приводится постановка задачи;
— в разделе 3 вводятся новые слабые метрики;
— в разделе 4 описаны оценки энтропии, основанные на введённых метриках;
— в разделе 5 изучаются статистики, используемые в методе ближайшей точки, и показывается, что дисперсия этих статистик оценивается как $O(n^{-1}\ln^4 n)$ для довольно широкого класса мер и рассматриваемых слабых метрик;
— в разделе 6 для симметричной меры Бернулли показано, что существуют такие параметры метрик, при которых оценка является несмещённой;
— в разделе 7 приводится описание алгоритмов для быстрого нахождения оценок и обоснование их трудоёмкости.

2. Постановка задачи

Пусть $\Omega = A^{\mathbb{N}}$ — пространство правосторонних бесконечных последовательностей символов из конечного алфавита A и μ — инвариантная относительно сдвига эргодическая вероятностная мера на Ω.

Пусть $\boldsymbol{\xi}_0, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n$ — независимые случайные точки в Ω, одинаково распределённые по мере μ.

Требуется оценить энтропию меры μ.

Дополнительно добавим следующее ограничение на меру:

$$\text{найдутся } a, b > 0, \text{ такие что } \mu\big(C_n(\boldsymbol{x})\big) \leqslant be^{-an} \quad \text{для любого } n > 0 \text{ для почти всех } \boldsymbol{x} \in \Omega, \quad (3)$$

где через

$$C_s(\boldsymbol{x}) = \{\boldsymbol{y} \in \Omega\colon y_1 = x_1, \ldots, y_s = x_s\}$$

будут обозначаться цилиндры в пространстве Ω.

Пусть случайная точка $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots)$ в Ω распределена по мере μ. Напомним, что энтропия h меры μ определяется следующим образом:

$$h = -\lim_{n\to\infty} \frac{1}{n} \mathsf{E} \ln \mu\big(C_n(\xi)\big). \tag{4}$$

3. Метрики на пространстве последовательностей

В этом разделе будет описан новый широкий класс так называемых *слабых* метрик [3], для которых неравенство треугольника выполняется с некоторой константой $C > 1$.

Пусть $\mathcal{A} = \{0, 1, \ldots, A-1\}$, и будем считать, что A чётно.

Для точки $\boldsymbol{x} = (x_1, x_2, \ldots)$ из $\Omega = \mathcal{A}^{\mathbb{N}}$, через $a\boldsymbol{x}$ будем обозначать точку $(a, x_1, x_2, \ldots)$, $a \in \mathcal{A}$.

Класс метрик ρ определяется следующим образом:

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = e^{-\alpha(\boldsymbol{x}, \boldsymbol{y})}, \tag{5}$$

где функция $\alpha(\boldsymbol{x}, \boldsymbol{y})$ определяется по следующим правилам:

$$\alpha(a\boldsymbol{x}, b\boldsymbol{y}) = \begin{cases} \alpha(\boldsymbol{x}, \boldsymbol{y}) + 1, & a = b, \\ \lambda_{a,b}\big(\alpha(\boldsymbol{x}, \boldsymbol{y})\big), & a \neq b. \end{cases} \tag{6}$$

Вспомогательные функции $\lambda_{a,b}(t)$ неубывающие и удовлетворяют условию

$$0 \leqslant \lambda_{a,b}(t) = \lambda_{b,a}(t) \leqslant 1.$$

Более того, семейство функций $\lambda_{a,b}(t)$ выбирается так, чтобы множество функций

$$S_a = \{\lambda_{a,0}(t), \ldots, \lambda_{a,a-1}(t), \lambda_{a,a+1}(t), \ldots, \lambda_{a,A-1}(t)\}$$

не зависело от a (т. е. $S_0 = S_1 = \ldots = S_{A-1}$) и $|S_a| = A - 1$. Другими словами, функции $\lambda_{a,b}(t)$ можно интерпретировать как раскраску рёбер полного графа с A вершинами $A-1$ цветами. Отметим, что такая раскраска существует только для чётных A.

Выберем хорошо известную раскраску

$$\begin{aligned} &\sigma_{0,1} = \sigma_{2,A-1} = \sigma_{3,A-2} = \ldots = \sigma_{A/2,A/2+1} = 1, \\ &\quad \ldots \\ &\sigma_{0,k} = \sigma_{k-1,k+1} = \ldots = \sigma_{1,2k-1} = \sigma_{2k,A-1} = \ldots = \sigma_{A/2+k-1,A/2+k} = k, \\ &\quad \ldots \\ &\sigma_{0,A-1} = \sigma_{1,A-2} = \sigma_{2,A-3} = \ldots = \sigma_{A/2-1,A/2} = A-1 \end{aligned} \tag{7}$$

и положим

$$\lambda_{i,j}(t) = \lambda_{\sigma_{i,j}}(t), \quad 0 \leqslant i \neq j \leqslant A-1, \tag{8}$$

где

$$\lambda_k(t) = \phi\left(\frac{k}{A} + \frac{1}{A}\phi^{-1}(t)\right), \quad k = 1, 2, \ldots, A-1, \tag{9}$$

а $\phi(t)$ — невозрастающая функция на полуинтервале $(0, 1]$, такая что

$$\phi(1) = 0, \quad \phi\left(\frac{t}{A}\right) = \phi(t) + 1. \tag{10}$$

Положим

$$\lambda_0(t) = t + 1, \tag{11}$$

тогда, очевидно, (9) выполняется и при $k = 0$.

Подчеркнём, что функции $\lambda_k(t)$ такие, что

$$\lambda_0(t_0) > \lambda_1(t_1) > \ldots > \lambda_{A-1}(t_{A-1}) \ \text{ для любых } \ t_0, t_1, \ldots, t_{A-1} \geqslant 0. \tag{12}$$

Итак, можно переписать (6) как

$$\alpha(a\boldsymbol{x}, b\boldsymbol{y}) = \lambda_{\sigma_{a,b}}\big(\alpha(\boldsymbol{x}, \boldsymbol{y})\big),$$

где

$$\sigma_{a,a} = 0.$$

Применяя (9), получаем

$$\alpha(a\boldsymbol{x}, b\boldsymbol{y}) = \phi\left(\frac{\sigma_{a,b}}{A} + \frac{1}{A}\phi^{-1}\big(\alpha(\boldsymbol{x}, \boldsymbol{y})\big)\right). \tag{13}$$

Отметим, что функция $\phi(t)$ является произвольной на интервале $(1/A, 1]$ (невозрастающей и удовлетворяющей условию $0 \leqslant \phi(t) \leqslant 1$).

Если $\phi(t) = 0$, $1/A < t \leqslant 1$, то получаем метрику ρ_0 (см. (1)).

Если $\phi(t) = 1$, $1/A \leqslant t \leqslant 1$, получаем метрику $(1/A)\rho_0$.

Отметим также, что метрика ρ с произвольной функцией $\phi(t)$ билипшицево эквивалентна метрике ρ_0, т. е.

$$e^{-1}\rho_0(\boldsymbol{x}, \boldsymbol{y}) \leqslant \rho(\boldsymbol{x}, \boldsymbol{y}) \leqslant \rho_0(\boldsymbol{x}, \boldsymbol{y}). \tag{14}$$

Следовательно, ρ является слабой метрикой [3], т. е. неравенство треугольника выполняется только с некоторой константой.

Каждая точка $\boldsymbol{x}$ имеет бесконечное число координат, однако в практических вычислениях нужно использовать только конечное число из них. Сделаем это введением понятия сечения метрики, которое будет использовать только первые m координат каждой точки.

Определим $\rho^{(m)}$ — *сечение метрики* ρ, положив

$$\rho^{(m)}(\boldsymbol{x}, \boldsymbol{y}) = e^{-\alpha^{(m)}(\boldsymbol{x}, \boldsymbol{y})}, \tag{15}$$

где

$$\begin{aligned} &\alpha^{(0)}(\boldsymbol{x}, \boldsymbol{y}) = 0, \\ &\alpha^{(m)}(a\boldsymbol{x}, b\boldsymbol{y}) = \lambda_{\sigma_{a,b}}\big(\alpha^{(m-1)}(\boldsymbol{x}, \boldsymbol{y})\big). \end{aligned} \tag{16}$$

Утверждение 1. *Множество значений* $\alpha^{(m)}(\boldsymbol{x}, \boldsymbol{y})$ *есть*

$$\Phi_m = \left\{ \phi\left(\frac{N}{A^m}\right),\ N = 1, 2, \ldots, A^m \right\}. \tag{17}$$

Доказательство. Проведём доказательство индукцией по m.

При $m = 0$ утверждение очевидно.

Применяя определение (16), получаем

$$\Phi_m = \bigcup_{k=0}^{A-1} \lambda_k(\Phi_{m-1}). \tag{18}$$

По предположению индукции и (9), (11) имеем

$$\bigcup_{k=0}^{A-1} \lambda_k(\Phi_{m-1}) = \left\{ \phi\left(\frac{k}{A} + \frac{N}{A^m}\right),\ N = 1, 2, \ldots, A^{m-1},\ k = 0, 1, 2, \ldots, A-1 \right\}.$$

Подставляя в (18), получаем (17). □

Для прикладных вычислений также нужно ограничить число параметров, которые используются для нахождения функции $\phi(t)$. Для этого введём сечение функции $\phi(t)$.

Определим $\phi_l(t)$, *сечение функции* $\phi(t)$, следующим образом:

$$\begin{aligned} &\phi_l(t) = \phi\left(\frac{N}{A^l}\right), \qquad \frac{N-1}{A^l} < t \leqslant \frac{N}{A^l}, \quad N = A^{l-1}+1, A^{l-1}+2, \ldots, A^l, \\ &\phi_l\left(\frac{t}{A}\right) = \phi_l(t) + 1, \quad 0 < t \leqslant 1. \end{aligned} \tag{19}$$

Следует подчеркнуть, что функции $\phi_l(t)$ имеют множество значений

$$\left\{ \phi\left(\frac{N}{A^l}\right) + k,\ N = A^{l-1}+1, \ldots, A^l,\ k = 0, 1, \ldots \right\}. \tag{20}$$

Следовательно, обратная функция $\phi_l^{-1}(x)$ определена на этом множестве и имеет множество значений

$$\left\{ \frac{N}{A^{l+k}},\ N = A^{l-1}+1, \ldots, A^l,\ k = 0, 1, \ldots \right\}.$$

Применяя утверждение 1, получаем следствие.

Следствие 1. *Пусть* $\phi(t) = \phi_l(t)$ *и* $m \geqslant l$. *Тогда множество значений* $\alpha^{(m)}(\boldsymbol{x}, \boldsymbol{y})$ *совпадает с*

$$\begin{aligned} &\left\{ \phi\left(\frac{N}{A^l}\right) + k,\ N = A^{l-1}+1, \ldots, A^l,\ k = 0, 1, \ldots, m-l \right\} \cup \\ &\qquad \cup \bigcup_{k=1}^{l-1} \left\{ \phi\left(\frac{N}{A^k}\right) + m - k,\ N = A^{k-1}+1, \ldots, A^k \right\} \cup \{m\}. \end{aligned} \tag{21}$$

Утверждение 2. *Пусть* $\phi(t) = \phi_l(t)$, $m \geqslant l$ *и* $\alpha(\boldsymbol{x}, \boldsymbol{y}) < m - l + 1$. *Тогда*

$$\alpha^{(m)}(\boldsymbol{x}, \boldsymbol{y}) = \alpha(\boldsymbol{x}, \boldsymbol{y}).$$

Доказательство. Итерируя (13), получаем

$$\alpha(\boldsymbol{x}, \boldsymbol{y}) = \phi_l\left(\frac{N}{A^j} + \frac{1}{A^j}\phi_l^{-1}\big(\alpha(\mathfrak{S}^j\boldsymbol{x}, \mathfrak{S}^j\boldsymbol{y})\big)\right), \tag{22}$$

где через $\mathfrak{S}$ обозначается сдвиг на пространстве Ω, а N — некоторое целое число, $0 \leqslant N < A^j$, $j \geqslant 1$.

Поскольку $\alpha(\boldsymbol{x}, \boldsymbol{y}) < m - l + 1$, то для некоторого $k \leqslant m - l$

$$\alpha(\boldsymbol{x}, \boldsymbol{y}) = \phi_l\left(\frac{N'}{A^l}\right) + k,$$

где $A^{l-1} + 1 \leqslant N' \leqslant A^l$. Подставляя $j = l + k$ в (22), имеем

$$\alpha(\boldsymbol{x}, \boldsymbol{y}) = \phi_l\left(\frac{N}{A^{l+k}} + \frac{1}{A^{l+k}}\phi_l^{-1}(t)\right), \tag{23}$$

где $t = \alpha(\mathfrak{S}^{l+k}\boldsymbol{x}, \mathfrak{S}^{l+k}\boldsymbol{y})$. Поэтому

$$\phi_l\left(\frac{N'}{A^l}\right) + k = \phi_l\left(\frac{N}{A^{l+k}} + \frac{1}{A^{l+k}}\phi_l^{-1}(t)\right).$$

Отсюда следует, что $N < A^l$ и

$$\phi_l\left(\frac{N'}{A^l}\right) = \phi_l\left(\frac{N}{A^l} + \frac{1}{A^l}\phi_l^{-1}(t)\right).$$

Поскольку $0 < \phi_l^{-1}(t) \leqslant 1$, имеем

$$N = N' - 1.$$

Итерируя (16), получаем

$$\alpha^{(m)}(\boldsymbol{x}, \boldsymbol{y}) = \phi_l\left(\frac{N}{A^{l+k}} + \frac{1}{A^{l+k}}\phi_l^{-1}(t_m)\right),$$

где $t_m = \alpha^{(m-l-k)}(\mathfrak{S}^{l+k}\boldsymbol{x}, \mathfrak{S}^{l+k}\boldsymbol{y})$ и N такое же, как в (23). Следовательно,

$$\alpha^{(m)}(\boldsymbol{x}, \boldsymbol{y}) = \phi_l\left(\frac{N}{A^l} + \frac{1}{A^l}\phi_l^{-1}(t_m)\right) + k = \phi_l\left(\frac{N+1}{A^l}\right) + k = \alpha(\boldsymbol{x}, \boldsymbol{y}). \qquad \square$$

4. Оценки энтропии, построенные по методу ближайшей точки

Пусть $n + 1$ точек $\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_n$ заданы своими первыми m координатами. Применим сечение $\rho^{(m)}$ метрики (5)—(10).

Будем считать, что сечение функции (10) задаётся функцией $\phi_l(t)$, $l < m$.

Пусть функция $\phi_l(t)$ задана параметрами

$$\beta_i = \phi\left(\frac{A^{l-1}+i}{A^l}\right), \quad i = 1, 2, \ldots, A^l - A^{l-1} - 1. \tag{24}$$

Подчеркнём, что

$$1 \geqslant \beta_1 > \beta_2 > \ldots > \beta_{A^l - A^{l-1}-1} \geqslant 0. \tag{25}$$

Кроме точек и метрики, будем использовать вспомогательный параметр k, который служит для контроля применимости. Оценки, полученные для различных значений k, являются оценками одной и той же величины.

Оценка $\eta_n^{(k)}\big(\rho^{(m)}\big)$ величины, обратной к энтропии $1/h$, определяется следующим образом [14]:

$$\eta_n^{(k)}\big(\rho^{(m)}\big) = k\Big(r_n^{(k)}\big(\rho^{(m)}\big) - r_n^{(k+1)}\big(\rho^{(m)}\big)\Big), \tag{26}$$

где

$$r_n^{(k)}\big(\rho^{(m)}\big) = \frac{1}{n+1}\sum_{j=0}^{n} \max_{i:\, i\neq j}^{(k)} \alpha^{(m)}(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) \tag{27}$$

и $\max^{(k)}\{X_1, \ldots, X_N\} = X_k$, если $X_1 \geqslant X_2 \geqslant \ldots \geqslant X_N$.

Применяя следствие 1, получаем основное свойство нашей метрики.

Следствие 2. *Оценка $\eta_n^{(k)}\big(\rho^{(m)}\big)$ и статистика $r_n^{(k)}\big(\rho^{(m)}\big)$ являются линейными функциями от параметров β_i, $i = 1, 2, \ldots, A^l - A^{l-1} - 1$.*

5. Статистические свойства оценок

В этом разделе будут изучены статистические свойства статистик (27) и оценок (26).

Обозначим открытый шар радиуса r с центром в точке $\boldsymbol{x} \in \Omega$ через

$$B(\boldsymbol{x}, r, \rho) = \{\boldsymbol{y} \in \Omega\colon \rho(\boldsymbol{x}, \boldsymbol{y}) < r\}.$$

Обозначим обратную функцию к $t = \mu\big(B(\boldsymbol{x}, e^{-u}, \rho)\big)$ через $u = \nu(t, \boldsymbol{x}, \rho)$, т. е.

$$\nu(t, \boldsymbol{x}, \rho) = \inf\big\{u\colon \mu\big(B(\boldsymbol{x}, e^{-u}, \rho)\big) < t\big\}. \tag{28}$$

Для краткости будем использовать обозначения

$$\nu(t, \boldsymbol{x}) = \nu(t, \boldsymbol{x}, \rho), \quad B(\boldsymbol{x}, r) = B(\boldsymbol{x}, r, \rho).$$

Следует подчеркнуть, что функции $t = \mu\big(B(\boldsymbol{x}, r)\big)$ и $r = \nu(t, \boldsymbol{x})$ могут иметь интервалы постоянства и точки разрыва (интервал постоянства одной функции соответствует разрыву другой).

Лемма 1. *Пусть μ — инвариантная относительно сдвига эргодическая мера на Ω и ρ — метрика* (5)*. Тогда для μ-почти всех точек $\boldsymbol{x} \in \Omega$*

$$\lim_{r\to 0} \frac{\ln \mu\big(B(\boldsymbol{x}, r, \rho)\big)}{\ln r} = h,$$

где h — энтропия μ.

Доказательство. Сначала рассмотрим случай $\rho = \rho_0$. Шары в метрике ρ_0 являются цилиндрами, т. е.

$$B(\boldsymbol{x}, r, \rho_0) = C_n(\boldsymbol{x}), \quad e^{-n-1} < r \leqslant e^{-n}.$$

Поэтому

$$\lim_{r\to 0} \frac{\ln \mu\big(B(\boldsymbol{x}, r, \rho_0)\big)}{\ln r} = -\lim_{n\to\infty} \frac{\ln \mu\big(C_n(\boldsymbol{x})\big)}{n}.$$

Применяя теорему Шеннона—Макмиллана—Бреймана [10, 2.10], получаем, что для μ-почти всех точек $\boldsymbol{x} \in \Omega$

$$\lim_{r\to 0} \frac{\ln \mu\big(B(\boldsymbol{x}, r, \rho_0)\big)}{\ln r} = -\lim_{n\to\infty} \frac{\ln \mu\big(C_n(\boldsymbol{x})\big)}{n} = h.$$

Теперь рассмотрим произвольную метрику (5). Поскольку ρ билипшицево эквивалентна (14) метрике (1), имеем

$$B(\boldsymbol{x}, e^{-1}r, \rho_0) \subset B(\boldsymbol{x}, r, \rho) \subset B(\boldsymbol{x}, er, \rho_0).$$

Поэтому

$$\mu\big(B(\boldsymbol{x}, e^{-1}r, \rho_0)\big) \leqslant \mu\big(B(\boldsymbol{x}, r, \rho)\big) \leqslant \mu\big(B(\boldsymbol{x}, er, \rho_0)\big).$$

Итак, для μ-почти всех точек $\boldsymbol{x} \in \Omega$

$$\lim_{r\to 0} \frac{\log \mu\big(B(\boldsymbol{x}, r, \rho)\big)}{\log r} = h. \qquad \square$$

Лемма 2.

$$\mathsf{E} r_n^{(k)}(\rho) = n\binom{n-1}{k-1} \int\limits_{\Omega} \int\limits_{0}^{1} \nu(u, \boldsymbol{x}, \rho) u^{k-1}(1-u)^{n-k}\, du\, d\mu(\boldsymbol{x}). \tag{29}$$

Доказательство. Поскольку $\mathsf{P}\{\alpha(\boldsymbol{x}, \boldsymbol{\xi_i}) > t\} = \mu\big(B(\boldsymbol{x}, e^{-t}, \rho)\big)$, имеем

$$\mathsf{E} \max_{1\leqslant j\leqslant n}{}^{(k)} \alpha(\boldsymbol{x}, \boldsymbol{\xi_j}) = n\binom{n-1}{k-1} \int\limits_{0}^{\infty} t\, \mu\big(B(\boldsymbol{x}, e^{-t}, \rho)\big)^{k-1} \times$$

$$\times \Big(1 - \mu\big(B(\boldsymbol{x}, e^{-t}, \rho)\big)\Big)^{n-k} d_t \mu\big(B(\boldsymbol{x}, e^{-t}, \rho)\big).$$

Следовательно,

$$\mathsf{E} r_n^{(k)}(\rho) = \mathsf{E} \max_{1\leqslant j\leqslant n}{}^{(k)} \alpha(\boldsymbol{\xi}_0, \boldsymbol{\xi}_j) = \mathsf{E}[\mathsf{E} \max_{1\leqslant j\leqslant n}{}^{(k)} \alpha(\boldsymbol{\xi}_0, \boldsymbol{\xi}_j) \mid \boldsymbol{\xi}_0] =$$

$$= n\binom{n-1}{k-1} \int\limits_{\Omega} \int\limits_{0}^{\infty} t\, \mu\big(B(\boldsymbol{x}, e^{-t}, \rho)\big)^{k-1} \times$$

$$\times \Big(1 - \mu\big(B(\boldsymbol{x}, e^{-t}, \rho)\big)\Big)^{n-k} d_t \mu\big(B(\boldsymbol{x}, e^{-t}, \rho)\big)\, d\mu(\boldsymbol{x}).$$

Применяя формулу замены переменных, получаем (29). $\square$

Лемма 3. *Пусть мера μ удовлетворяет условию* (3). *Пусть ρ — метрика* (5) *с $\phi(t) = \phi_l(t)$. Тогда существуют такие константы c_1, c_2, что верно неравенство*

$$0 < \mathsf{E}r_n^{(k)}(\rho) - \mathsf{E}r_n^{(k)}\big(\rho^{(m)}\big) \leqslant c_1 n^{-1} \quad \text{для} \quad m \geqslant l + c_2 \log n. \tag{30}$$

Доказательство. Согласно утверждению 2

$$B(\boldsymbol{x}, e^{-t}, \rho) = B\big(\boldsymbol{x}, e^{-t}, \rho^{(m)}\big), \quad t \leqslant m - l,$$

поэтому

$$\nu(u, \boldsymbol{x}, \rho) = \nu\big(u, \boldsymbol{x}, \rho^{(m)}\big), \quad u \geqslant \varepsilon_m,$$

где

$$\varepsilon_m = \mu\big(B(\boldsymbol{x}, e^{-m+l}, \rho)\big).$$

Применив лемму 2, получим

$$\begin{aligned}
&\mathsf{E}r_n^{(k)}(\rho) - \mathsf{E}r_n^{(k)}\big(\rho^{(m)}\big) = \\
&\quad = n\binom{n-1}{k-1} \int\limits_{\Omega} \int\limits_{0}^{\varepsilon_m} \big[\nu(u, \boldsymbol{x}, \rho) - \nu\big(u, \boldsymbol{x}, \rho^{(m)}\big)\big] u^{k-1}(1-u)^{n-k}\, du\, d\mu(\boldsymbol{x}) \leqslant \\
&\quad \leqslant n\binom{n-1}{k-1} \int\limits_{\Omega} \int\limits_{0}^{\varepsilon_m} \nu(u, \boldsymbol{x}, \rho) u^{k-1}\, du\, d\mu(\boldsymbol{x}).
\end{aligned}$$

Используя условие (3) и неравенство (14), имеем

$$\mu\big(B(\boldsymbol{x}, e^{-t}, \rho)\big) \leqslant c_3 e^{-at}.$$

Поэтому

$$\nu(u, \boldsymbol{x}, \rho) \leqslant -\frac{1}{a} \ln \frac{u}{c_3}, \quad \varepsilon_m \leqslant c_3 e^{-a(m-l)}.$$

Подставляя эти неравенства, получаем

$$\mathsf{E}r_n^{(k)}(\rho) - \mathsf{E}r_n^{(k)}\big(\rho^{(m)}\big) \leqslant -\frac{n^k}{a} \int\limits_{0}^{\varepsilon_m} \ln \frac{u}{c_3} u^{k-1}\, du.$$

Вычислив интеграл, получим

$$\mathsf{E}r_n^{(k)}(\rho) - \mathsf{E}r_n^{(k)}\big(\rho^{(m)}\big) = O\left(n^k m e^{-ak(m-l)}\right).$$

Возьмём $c_2 > (k+1)/(ak)$. Тогда неравенство (30) следует из полученного равенства. □

Применяя лемму 1 и повторяя доказательство теоремы 1 из [14], получаем следующее утверждение.

Утверждение 3. *Пусть $\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_n$ — $n+1$ независимых точек в пространстве Ω, распределённых по мере μ и $k = O(\ln n)$. Тогда*

$$\lim_{n\to\infty} \frac{\mathsf{E}r_n^{(k)}(\rho)}{\ln n} = \frac{1}{h}.$$

Следствие 3. *Пусть выполнены условия леммы* 3 *и* $k = O(\ln n)$. *Тогда*

$$\lim_{n\to\infty} \frac{\mathsf{E} r_n^{(k)}\bigl(\rho^{(m)}\bigr)}{\ln n} = \frac{1}{h}.$$

Теорема 4. *Пусть* $r_n^{(k)}\bigl(\rho^{(m)}\bigr)$ *— статистика, определённая в* (27). *Тогда выполняется следующее неравенство*:

$$\mathsf{Var}\, r_n^{(k)}\bigl(\rho^{(m)}\bigr) \leqslant \frac{m^2(km+1)^2A^2}{4(n+1)}. \tag{31}$$

Доказательство. Применим метод Макдиармида [11].

Введём функцию

$$f\colon \Omega^{n+1} \to \mathbb{R},$$

положив

$$f(\boldsymbol{x}_0,\dots,\boldsymbol{x}_n) = \frac{1}{n+1}\sum_{j=0}^{n} \max_{i\colon i\neq j}{}^{(k)}\, \alpha^{(m)}(\boldsymbol{x}_i,\boldsymbol{x}_j).$$

Для того чтобы применить метод Макдиармида, нужно показать, что f удовлетворяет неравенству

$$\sup_{\boldsymbol{x}_0,\dots,\boldsymbol{x}_n,\boldsymbol{y}\in\Omega} |f(\boldsymbol{x}_0,\dots,\boldsymbol{x}_n) - f(\boldsymbol{x}_0,\dots,\boldsymbol{x}_{i-1},\boldsymbol{y},\boldsymbol{x}_{i+1},\dots,\boldsymbol{x}_n)| \leqslant c_i \tag{32}$$

для $0 \leqslant i \leqslant n$. Докажем это неравенство для

$$c_i = \frac{m(km+1)A}{n+1}, \quad 0 \leqslant i \leqslant n. \tag{33}$$

Для краткости введём следующие обозначения:

$$\begin{aligned}
&\boldsymbol{X} = (\boldsymbol{x}_0,\dots,\boldsymbol{x}_n),\\
&\tilde{\boldsymbol{X}} = (\boldsymbol{x}_0,\dots,\boldsymbol{x}_{i-1},y,\boldsymbol{x}_{i+1},\dots,\boldsymbol{x}_n),\\
&g_j(\boldsymbol{X}) = \max_{l\colon l\neq j}{}^{(k)}\, \alpha^{(m)}(\boldsymbol{x}_l,\boldsymbol{x}_j).
\end{aligned}$$

Положим

$$J = \{j \neq i\colon g_j(\boldsymbol{X}) \neq g_j(\tilde{\boldsymbol{X}})\}.$$

Поскольку

$$\begin{aligned}
&g_j(\boldsymbol{X}) = g_j(\tilde{\boldsymbol{X}}), \quad j \notin J, \quad j \neq i,\\
&|g_j(\boldsymbol{X}) - g_j(\tilde{\boldsymbol{X}})| \leqslant m \ \text{ для всех } \ j,
\end{aligned}$$

имеем

$$|f(\boldsymbol{X}) - f(\tilde{\boldsymbol{X}})| \leqslant \frac{m(|J|+1)}{n+1}.$$

По утверждению 5 $|J| \leqslant kmA$, поэтому выполняется (33).

Как показано в [2], метод Макдиармида даёт следующую оценку дисперсии $f(\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_n)$:

$$\mathsf{Var}[f(\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_n)] \leqslant \frac{1}{4}\sum_{i=0}^{n} c_i^2. \tag{34}$$

Подставляя (33) в (34), получаем (31). □

Следствие 4.

$$P\left\{\left|r_n^{(k)}\left(\rho^{(m)}\right) - \mathsf{E}r_n^{(k)}\left(\rho^{(m)}\right)\right| > \delta\right\} \leqslant 2e^{-2(n+1)\delta^2/m^2(km+1)^2A^2}. \tag{35}$$

Доказательство. Это неравенство получается применением следующего неравенства из [11]:

$$P\{|f(\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_n) - \mathsf{E}f(\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_n)| > \delta\} \leqslant 2e^{-2\delta^2/\sum\limits_{i=0}^{n} c_i^2},$$

где c_i заданы в (33). Подставляя (33) в это неравенство, получаем

$$P\{|f(\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_n) - \mathsf{E}f(\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_n)| > \delta\} \leqslant 2e^{-2(n+1)\delta^2/m^2(km+1)^2A^2}.$$

Подставляя $r_n^{(k)}\left(\rho^{(m)}\right) = f(\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_n)$, получаем (35). □

Следствие 5. *Пусть мера μ удовлетворяет условию* (3), *ρ — метрика* (5) *с $\phi(t) = \phi_l(t)$, $m \geqslant c_1 \ln n + l$, $m \leqslant c_2 \ln n$, где c_1, c_2 — некоторые константы, и $k = O(\ln n)$. Тогда $\mathsf{E}r_n^{(k)}\left(\rho^{(m)}\right)/\ln n$ почти всюду сходится к $1/h$.*

Применяя теорему 4 и неравенство $\mathsf{Var}(X+Y) \leqslant \left(\sqrt{\mathsf{Var}\,X} + \sqrt{\mathsf{Var}\,Y}\right)^2$, получаем следующее утверждение.

Утверждение 4. *Пусть ρ — метрика* (5) *с $\phi(t) = \phi_l(t)$, $m \geqslant c_1 \ln n + l$, $m \leqslant c_2 \ln n$, где c_1, c_2 — некоторые константы, k, A фиксированные. Тогда*

$$\mathsf{Var}\,\eta_n^{(k)}\left(\rho^{(m)}\right) = O\left(\frac{\ln^4 n}{n}\right).$$

6. Симметричная мера Бернулли

Покажем, что для симметричной меры Бернулли (с A равновероятными символами) можно найти такую функцию $\phi(t)$ для метрики (5)—(10), что $\eta_n^{(k)}(\rho)$ является несмещённой оценкой.

Для симметричной меры Бернулли функция $\nu(t, \omega)$ не зависит от точки ω. Поэтому $\chi(t) = \nu(t, \omega)$ задаётся следующим рекуррентным уравнением:

$$\chi(t) = \begin{cases} \chi(At) + 1, & t \leqslant \dfrac{1}{A}, \\ \phi\left(\dfrac{k}{A} + \dfrac{1}{A}\phi^{-1}(\chi(At-k)\right), & \dfrac{k}{A} < t \leqslant \dfrac{k+1}{A},\ k = 1, 2, \ldots, A-1. \end{cases} \tag{36}$$

Нетрудно убедиться, что решением этого уравнения является $\chi(t) = \phi(t)$.

Покажем, что для

$$\chi(t) = \phi(t) = -\frac{\ln t}{\ln A}$$

$\eta_n^{(k)}(\rho)$ будет несмещённой оценкой.

Подставляя $\chi(t)$ в (29), имеем

$$\mathsf{E}r_n^{(k)}(\rho) = -\frac{n!}{(k-1)!\,(n-k)!\,\ln A}\int\limits_0^1 \ln(t)\,t^{k-1}(1-t)^{n-k}\,dt.$$

Применяя [4, 4.253.1], получаем

$$\mathsf{E}r_n^{(k)}(\rho) = \frac{\psi(n+1) - \psi(k)}{\ln A}, \tag{37}$$

где

$$\psi(t) = \frac{d}{dt}\ln\Gamma(t) -$$

пси-функция. Следовательно,

$$\mathsf{E}\eta_n^{(k)}(\rho) = k\frac{\psi(k+1) - \psi(k)}{\ln A} = \frac{1}{\ln A}. \tag{38}$$

7. Алгоритмы

В этом разделе будет описано несколько алгоритмов для нахождения статистики (27).

7.1. Поиск k-ближайшей

Пусть $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n$ — n слов $\boldsymbol{\xi}_i = \xi_{i1}\ldots\xi_{im}$, $\xi_{ij} \in \mathcal{A}$, $i = 1, 2, \ldots, n$.

Для данного слова $\boldsymbol{\xi}_0$ и целого k нужно найти следующий максимум:

$$\max_{1\leqslant i\leqslant n}{}^{(k)}\alpha^{(m)}(\boldsymbol{\xi}_i, \boldsymbol{\xi}_0).$$

Все слова $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n$ будем хранить как бор T [8]. Бор — это A-арное дерево, вершинами которого служат векторы длины A. Будем обозначать через T массив размера $A \times N$, столбцами которого служат эти N векторов. Таким образом, $T(i, j)$ будет вершиной на следующем уровне от вершины i, если существует слово в словаре с заданным префиксом и следующим символом j, иначе $T(i, j) = 0$. Припишем каждой вершине i бора T дополнительный параметр $L(i)$, равный числу листьев в подборе с корнем i.

Отметим, что в нашем боре T все листья находятся на уровне m. Будем считать, что номер каждого листа равняется индексу i соответствующего слова $\boldsymbol{\xi}_i$.

Метрика (5) задаётся функцией $\phi(t)$ (10) и перестановками (7)

$$\boldsymbol{\sigma}_i = (\sigma_{i,0}, \ldots, \sigma_{i,A-1}).$$

Будем также использовать обратные перестановки

$$\bar{\boldsymbol{\sigma}}_i = (\bar{\sigma}_{i,0}, \ldots, \bar{\sigma}_{i,A-1}), \quad i = 0, 1, \ldots, A-1.$$

Приведём псевдокод алгоритма поиска.

- Полагаем $V := \mathrm{root}$.
- **for** $l := 1$ **to** $m-1$ **do**
- **for** $i := 0$ **to** $A-1$ **do**
  if $L\big(T(\bar{\sigma}(\xi_{0,l}, i), V)\big) \geqslant k$
  then $V := T(\bar{\sigma}(\xi_{0,l}, i), V)$; **break**;
  else $k := k - L\big(T(\bar{\sigma}(\xi_{0,l}, i), V)\big)$.

Номер вершины V (на уровне m) является индексом слова $\boldsymbol{\xi}_V$, для которого

$$\alpha^{(m)}(\boldsymbol{\xi}_V, \boldsymbol{\xi}_0) = \\ = \phi\left(\frac{1}{A}\sigma_{\xi_{V,1},\xi_{0,1}} + \frac{1}{A}\left(\frac{1}{A}\sigma_{\xi_{V,2},\xi_{0,2}} + \ldots + \frac{1}{A}\left(\frac{1}{A}\sigma_{\xi_{V,m},\xi_{0,m}} + \frac{1}{A}\right)\ldots\right)\right). \quad (39)$$

Корректность алгоритма вытекает из следующего простого свойства метрики.

Утверждение 5. *Пусть $\boldsymbol{Y}$ — такое слово, что*

$$\alpha^{(m)}(\boldsymbol{Y}, \boldsymbol{\xi}_0) = \max_{1 \leqslant i \leqslant n}{}^{(k)} \alpha^{(m)}(\boldsymbol{\xi}_i, \boldsymbol{\xi}_0),$$

пусть $\boldsymbol{Y}$ лежит в подборе T_0, таком что число листьев T_0 не менее k, и пусть $r_0, \ldots, r_{A-1}$ — потомки корня T_0, занумерованные индексами соответствующих функций $\lambda_i(t)$. Тогда $\boldsymbol{Y}$ принадлежит подбору с корнем r_i, таким что

$$L(r_0) + \ldots + L(r_{i-1}) < k, \quad L(r_0) + \ldots + L(r_i) \geqslant k.$$

Доказательство следует непосредственно из свойства (12).

Временна́я трудоёмкость алгоритма равна $\mathcal{O}(mA)$. Требуемая память — $\mathcal{O}(mAn)$.

7.2. Эффективный алгоритм нахождения $r_n^{(k)}\big(\rho^{(m)}\big)$

Для того чтобы найти $r_n^{(k)}\big(\rho^{(m)}\big)$, построим бор T с дополнительным параметром L из слов $\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_n$. Затем для $i = 0, 1, \ldots, n$ найдём $(k+1)$-ближайшее слово в боре T для слова $\boldsymbol{\xi}_i$. Суммируя, получаем $r_n^{(k)}\big(\rho^{(m)}\big)$.

Временна́я сложность равна $\mathcal{O}(mnA)$.

7.3. Эффективный алгоритм нахождения $r_n^{(k)}\big(\rho^{(m)}\big)$ для всех n

В этом разделе рассмотрим следующую задачу. Пусть $\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_n$ — $n+1$ слов $\boldsymbol{\xi}_i = \xi_{i1} \ldots \xi_{im}$, $\xi_{ij} \in \mathcal{A}$, $i = 0, 1, \ldots, n$. Требуется найти значения $r_j^{(i)}\big(\rho^{(m)}\big)$ для всех i, j, где $k \leqslant j \leqslant n$, $1 \leqslant i \leqslant k$.

Введём вспомогательный параметр

$$R(i,j) = \max_{1\leqslant s\leqslant n}^{(i)} \alpha^{(m)}(\boldsymbol{\xi}_s, \boldsymbol{\xi}_j). \tag{40}$$

Для данной таблицы слов ξ строим вышеописанный бор T и находим вспомогательный параметр R.

Приведём псевдокод алгоритма.

- Полагаем $R := 0$, $T := \varnothing$.
- **for** $j := 1$ **to** n **do**
- вставляем слово $\boldsymbol{\xi_j}$ в бор T;
- находим $R(i,j)$ для $i = 1, 2, \ldots, k$, применяя алгоритм из раздела 7.1;
- пусть $p_0, \ldots, p_m$ — путь в боре от листа j к корню и $Q := 0$;
  for $t := 1$ **to** m **do**
    for каждого сына S вершины p_t **do**
    if $L(S) \leqslant k$ **then**
    for каждого листа V of S **do**
      for $i := L(S)$ **to** k **do**
        полагаем $Q(i) := Q(i) - R(i,V)$;
        пересчитываем $R(i,V)$, применяя алгоритм из раздела 7.1;
        полагаем $Q(i) := Q(i) + R(i,V)$;
- **for** $i := 1$ **to** k **do**
    полагаем $r_j^{(i)}\big(\rho^{(m)}\big) = (1/j)\big((j-1)r_{j-1}^{(i)}\big(\rho^{(m)}\big) + Q(i) + R(i,j)\big)$.

Корректность алгоритма вытекает из следующего утверждения.

Утверждение 6. *Пусть V — такая вершина в боре T, что $L(V) > k$, и T_0 — подбор с корнем V. Тогда для любого слова ξ_i, соответствующего листу i подбора T_0, имеем*

$$\max_{0\leqslant j\neq i\leqslant n}^{(k)} \alpha^{(m)}(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) = \alpha^{(m)}(\boldsymbol{\xi}_i, \boldsymbol{\xi}_s),$$

где лист s также принадлежит подбору T_0.

Доказательство непосредственно следует из утверждения 5.

Временна́я трудоёмкость равна $\mathcal{O}(mk^2An)$.

7.4. Символическое вычисление $r_n^{(k)}\big(\rho^{(m)}\big)$ для всех n

Предположим, что сечение $\phi_l(t)$ задано параметрами β_i (24), где β_i неизвестны, $i = 1, 2, \ldots, M$,

$$M = A^l - A^{l-1} - 1. \tag{41}$$

Алгоритм из раздела 7.1 применим и в этом случае, но в (39) получим

$$\alpha^{(m)}(\boldsymbol{\xi}_V, \boldsymbol{\xi}_0) = U\beta_I + V,$$

где U, V не зависят от β.

Подставляя в (27), имеем

$$r_n^{(k)}\bigl(\rho^{(m)}\bigr) = \frac{1}{n} D_{n,k} + \frac{1}{n} \sum_{s=1}^{M} C_{n,k,s} \beta_s. \tag{42}$$

В этом подразделе рассмотрим задачу нахождения величин $D_{j,i}$, $C_{j,i,s}$ для всех i, j, где $k \leqslant j \leqslant n$, $1 \leqslant i \leqslant k$, $1 \leqslant s \leqslant M$.

Введём вспомогательные параметры I, U, V:

$$\max_{1 \leqslant s \leqslant n}{}^{(i)} \alpha^{(m)}(\boldsymbol{\xi_s}, \boldsymbol{\xi_j}) = U(i,j)\beta_{I(i,j)} + V(i,j). \tag{43}$$

Пусть бор T задан вышеописанным образом. Тогда по алгоритму последовательно находим вспомогательные параметры I, U, V и D, C.

Приведём псевдокод алгоритма.

- Полагаем $C := 0$, $D := 0$, $U := 0$, $V := 0$.
- **for** $j := 1$ **to** n **do**
- **for** $i := 1$ **to** k **do**

  находим $I(i,j)$, $U(i,j)$, $V(i,j)$, применяя алгоритм из раздела 7.1;

  полагаем $\tilde{D}(i) := V(i,j)$;

  полагаем $\tilde{C}\bigl(i, I(i,j)\bigr) := U(i,j)$;
- пусть $p_0, \ldots, p_m$ — путь в боре T от листа j до корня;

  for $t := 1$ **to** m **do**

  for каждого сына S вершины p_t **do**

  if $L(S) \leqslant k$ **then**

  for каждого листа F вершины S **do**

  for $i := L(S)$ **to** k **do**

  полагаем $\tilde{D}(i) := \tilde{D}(i) - V(i,F)$,

  $\tilde{C}\bigl(i, I(i,F)\bigr) := \tilde{C}\bigl(i, I(i,F)\bigr) - U(i,F)$;

  пересчитываем $I(i,F)$, $U(i,F)$, $V(i,F)$,

  применяя алгоритм из раздела 7.1;

  полагаем $\tilde{D}(i) := \tilde{D}(i) + V(i,F)$,

  $\tilde{C}\bigl(i, I(i,F)\bigr) := \tilde{C}\bigl(i, I(i,F)\bigr) + U(i,F)$;
- **for** $i := 1$ **to** k **do**

  полагаем $D(j,i) := D(j,i) + \tilde{D}(i)$;

  for $s := 1$ **to** M **do**

  полагаем $C(j,i,s) := C(j,i,s) + \tilde{C}(i,s)$.

Корректность алгоритма следует из утверждения 6.

Временна́я трудоёмкость равна $\mathcal{O}(mk^2An + kMn)$.

7.5. Нахождение параметров метрики

Параметры метрики будем выбирать так, чтобы минимизировать среднее отклонение. По следствию 2 оценка является линейной функцией от параметров, поэтому получаем задачу минимизации положительной квадратичной формы на симплексе (25). Действительно, по (42) имеем

$$\eta_n^{(k)}\big(\rho^{(m)}\big) = \eta_n^{(k)}\big(\rho_0^{(m)}\big) + \sum_{i=1}^{M} \beta_i \tilde{C}_{n,k,i}, \tag{44}$$

где

$$\tilde{C}_{n,k,i} = k(C_{n,k,i} - C_{n,k+1,i}).$$

Будем минимизировать функцию

$$F(\beta) = \frac{1}{n-k+1} \sum_{j=k}^{n} \Big(\eta_j^{(k)}\big(\rho^{(m)}\big) - \bar{\eta}_j^{(k)}\big(\rho^{(m)}\big)\Big)^2, \tag{45}$$

где

$$\bar{\eta}_n^{(k)}\big(\rho^{(m)}\big) = \frac{1}{n-k+1} \sum_{j=k}^{n} \eta_j^{(k)}\big(\rho^{(m)}\big). \tag{46}$$

Подставляя (44) в (45), (46), получаем, что $F(\beta)$ является квадратичной формой от своих параметров. Следовательно, задача минимизации этой функции является задачей минимизации положительной квадратичной формы на симплексе (25).

7.6. Алгоритм нахождения оценки энтропии

Алгоритм состоит из двух этапов.

1. Выбираем часть заданных строк.
   Находим коэффициенты линейной функции (44).
   Находим минимум положительной квадратичной формы (45) на симплексе (25).
2. Находим оценку (26) энтропии по оставшимся заданным строкам с параметрами, найденными на первом этапе.

Автор пользуется возможностью выразить признательность В. Л. Дольникову за полезные обсуждения.

Литература

[1] Aldous D., Shields P. A diffusion limit for a class of randomly-growing binary trees // Probab. Th. Rel. Fields. — 1988. — Vol. 79. — P. 509—542.

[2] Devroye L. Exponential inequalities in nonparametric estimation // Nonparametric Functional Estimation and Related Topics / G. Roussas, ed. — Dordrecht: Kluwer Academic, 1991. — P. 31—44.

[3] Deza M., Deza T. Encyclopedia of Distances. — Berlin: Springer, 2009.

[4] Gradshtein I. S., Ryzhik I. M. Table of Integrals, Series, and Products. — London: Academic Press, 1994.

[5] Grassberger P. Estimating the information content of symbol sequences and efficient codes // IEEE Trans. Inform. Theory. — 1989. — Vol. 35. — P. 669—675.

[6] Kaltchenko A., Timofeeva N. Entropy estimators with almost sure convergence and an $O(n^{-1})$ variance // Adv. Math. Commun. — 2008. — Vol. 2. — P. 1—13.

[7] Kaltchenko A., Timofeeva N. Rate of convergence of the nearest neighbor entropy estimator // AEU — Int. J. Electron. Commun. — 2010. — Vol. 64. — P. 75—79.

[8] Knuth D. E. The Art of Computer Programming. Vol. 3. Sorting and Searching. — Reading: Addison-Wesley, 1975.

[9] Kontoyiannis I., Suhov Yu. M. Prefixes and the entropy rate for long-range sources // Probability Statistics and Optimization / F. P. Kelly, ed. — New York: Wiley, 1994. — P. 89—98.

[10] Martin N., England J. Mathematical Theory of Entropy. — Cambridge: Cambridge Univ. Press, 1984.

[11] McDiarmid C. On the method of bounded differences // Surveys in Combinatorics. — Cambridge: Cambridge Univ. Press, 1989. — P. 148—188.

[12] Shields P. C. Entropy and prefixes // Ann. Probab. — 1992. — Vol. 20. — P. 403—409.

[13] Silvapulle M. J., Sen P. K. Constrained Statistical Inference: Inequality, Order and Shape Restrictions. — New York: Wiley, 2005.

[14] Timofeev E. A. Statistical estimation of measure invariants // St. Petersburg Math. J. — 2006. — Vol. 17, no. 3. — P. 527—551.

[15] Timofeev E. A. Bias of a nonparametric entropy estimator for Markov measures // J. Math. Sci. — 2011. — Vol. 176, no. 2. — P. 255—269.

[16] Ziv J., Lempel A. Compression of individual sequences by variable rate coding // IEEE Trans. Inform. Theory. — 1978. — Vol. 24. — P. 530—536.