Модел. и анализ информ. систем. Т.**13**, №**1** (2006) 41–45

УДК 519.987

Смещение статистической оценки энтропии для простейшей меры Бернулли

Тимофеев Е.А.
Ярославский государственный университет
150 000, Ярославль, Советская, 14
E-mail : tim@uniyar.ac.ru

получена 15 января 2006

Аннотация

В работе найдены асимптотические и приближенные формулы для смещения оценки энтропии, предложенной в работе [2], в случае простейшей бернуллиевской меры с $p = q = 1/2$.

В работе [2] предложена статистическая оценка энтропии h случайного процесса, принимающего значения в конечном множестве $\mathcal{A}$, если заданы n его независимых реализаций. В настоящей работе найдено смещение этой оценки для бернуллиевской меры с $\mathcal{A} = \{0, 1\}$ и $p = q = 1/2$.

Полученные значения смещения показывают, что результаты оценивания смещения и дисперсии в [2] являются асимптотически точными для бернуллиевской меры.

Интересно, что смещение содержит небольшое (порядка 10^{-6}) слагаемое, которое является функцией с изменяющимся периодом, равным $\frac{2\pi \ln(n-1)}{h}$.

1. Приведем результаты работы [2] и введем обозначения. Пусть $\xi = (X_1, X_2, \dots)$ – стационарный процесс, принимающий значения в конечном алфавите $\mathcal{A}$. Через μ обозначим индуцированную борелевскую меру на пространстве односторонних последовательностей $\Omega = \mathcal{A}^{\mathbb{N}}$, где $\mathbb{N} = \{1, 2, \dots\}$. Энтропия h (средняя энтропия на символ) определяется как

$$h = -\lim_{n\to\infty} \frac{1}{n} \mathsf{E} \ln p(X_1, \dots, X_n), \tag{1}$$

где

$$p(x_1, \dots, x_n) = \mathsf{P}\{X_1 = x_1, \dots, X_n = x_n\}.$$

Отметим, что выбор натурального логарифма (ln) позволяет избежать дополнительных множителей при построении оценок.

Введем на Ω метрику

$$\rho(x, y) = \frac{1}{\max\{k : x_k \neq y_k\}}. \tag{2}$$

Пусть $\xi_1, \xi_2, \dots, \xi_n$ – независимые одинаково распределенные случайные величины, принимающие значения в Ω и имеющие общее распределение – меру μ.

Статистическая оценка $\eta_n(\rho)$ энтропии строится следующими простыми вычислениями:

- находим вспомогательную случайную величину

$$r_n^{(k)} = \frac{1}{n} \sum_{j=1}^{n} \left(\min_{i:i\neq j}{}^{(k)} \rho(\xi_i, \xi_j) \right)^{-1}, \tag{3}$$

  где $\min^{(k)}\{x_1, \dots, x_N\} = x_k$, если $x_1 \leqslant x_2 \leqslant \dots \leqslant x_N$;

- полагаем оценкой энтропии случайную величину

$$\eta_n^{(k)}(\rho) = \frac{\ln n}{r_n^{(k)}}. \tag{4}$$

Через

$$C_n(x) = \{y \in \mathcal{A}^{\mathbb{N}} : y_i = x_i,\ i = 1, \ldots, n\},\quad C_0(x) = \Omega$$

будем обозначать цилиндры в пространстве последовательностей $\Omega = \mathcal{A}^{\mathbb{N}}$, а через $B(x, r) = \{y \in \Omega : \rho_0(x, y) < r\}$ будем обозначать открытый шар радиуса r с центром в точке x. Подчеркнем, что для метрики (2) цилиндры совпадают с шарами

$$C_n(x) = B(x, r),\quad \frac{1}{n+1} < r \leqslant \frac{1}{n}. \tag{5}$$

Для рассматриваемой бернуллиевской меры

$$\mu(C_n(x)) = 2^{-n}, \tag{6}$$

а энтропия

$$h = \ln 2. \tag{7}$$

Введем вспомогательную функцию

$$\phi(s) = \sum_{k=0}^{\infty} \left(1 - \left[1 - e^{-kh}\right]^s\right). \tag{8}$$

В [2] при доказательстве теоремы 1 показано, что

$$\mathsf{E} r_n^{(k)} = \sum_{m=0}^{k-1} (-1)^m \binom{n-1}{m} \Delta^{(m)} \phi(n-1), \tag{9}$$

а в теореме 2 из [2] показано, что $\forall c < 1$

$$\mathsf{D} r_n^{(k)} = \mathcal{O}(n^{-c}), \tag{10}$$

где через

$$\Delta\phi(s) = \phi(s) - \phi(s-1)$$

обозначается конечная разность первого порядка функции $\phi(s)$.

2. Асимптотические и приближенные формулы. Для функции $\phi(s)$ справедливо асимптотическое разложение

$$\phi(s) = \frac{H_s}{h} + \frac{1}{2} - \frac{1}{h} \sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} \Gamma\left(\frac{2\pi k i}{h}\right) s^{-2\pi k i/h} + \mathcal{O}(s^{-1}), \tag{11}$$

где через H_s обозначаются гармонические числа

$$H_s = \sum_{i=1}^{s} \frac{1}{i}. \tag{12}$$

Первое приближение получается из (11) отбрасыванием суммы

$$\phi(s) \approx \frac{H_s}{h} + \frac{1}{2}.$$

Точность этого приближения равна $1.8 \cdot 10^{-6}$.

Второе приближение получается из (11) отбрасыванием слагаемых с $|k| > 1$

$$\phi(s) \approx \frac{H_s}{h} + \frac{1}{2} + \frac{10^{-6}}{\pi} \left[3.7861079986 \cos\left(\frac{2\pi \ln s}{h}\right) + 3.1766226452 \sin\left(\frac{2\pi \ln s}{h}\right)\right]. \tag{13}$$

Точность этого приближения примерно 10^{-8}.

При $m > 0$ для m-ой разности функции $\phi(s)$ справедливо асимптотическое разложение

$$\frac{(-1)^{m-1} s! \Delta^{(m)} \phi(s)}{(s-m)!} = \frac{1}{h} \sum_{k=-\infty}^{\infty} \Gamma\left(m + \frac{2\pi k i}{h}\right) s^{-2\pi k i/h} + \mathcal{O}(s^{-1}). \tag{14}$$

Приближенное значение для математического ожидания случайной величины $r_n^{(k)}$ имеет следующий вид

$$\mathsf{E}r_{n+1}^{(k)} \approx \frac{H_n - H_{k-1}}{h} + \frac{1}{2} + \frac{1}{h}\sum_{m=0}^{k-1} \frac{(-1)^{m-1}}{m!}\left(\Gamma\left(m + \frac{2\pi i}{h}\right) n^{-2\pi i/h} + \Gamma\left(m - \frac{2\pi i}{h}\right) n^{2\pi i/h}\right). \tag{15}$$

3. Обоснование формул (11) – (15).

3.1. Вычисление $\phi(s)$. Для нахождения суммы (8) применим формулу суммирования Пуассона [1, (13.8.4)]

$$\phi(s) = \sum_{k=0}^{\infty} f(k) = \int_0^{\infty} f(t)\,dt + \frac{1}{2}f(0) - \frac{1}{12}f'(0) - \frac{1}{2\pi^2}\sum_{k=1}^{\infty}\frac{1}{k^2}\int_0^{\infty} f''(t)\cos 2\pi kt\,dt$$

для функции $f(x) = 1 - (1 - e^{-xh})^s$.

Отметим, что формула Пуассона получается при подстановке в формулу Эйлера-Маклорена [1, (13.6.3)] разложения в ряд Фурье многочлена Бернулли $B_2(\{t\})$.

Отметим также, что дальнейшее разложение (по следующим производным) не имеет смысла, поскольку максимальные значения всех производных функции $f(x)$ примерно постоянны при всех s, а $f^{(m)}(0) = 0$, при $1 \leqslant m < s$. Поэтому при замене второй производной на m-ую получим такое же значение остаточного члена.

Подставляя в полученную формулу значения $f(0) = 1$, $f'(0) = 0$, вычисляя первый интеграл и интегрируя по частям второй, получим

$$\phi(s) = \frac{H_s}{h} + \frac{1}{2} - \sum_{k=1}^{\infty}\frac{1}{\pi k}\int_0^{\infty} f'(t)\sin 2\pi kt\,dt.$$

Подставляя производную, получим

$$\phi(s) = \frac{H_s}{h} + \frac{1}{2} + \sum_{k=1}^{\infty}\frac{hs}{\pi k}\int_0^{\infty} e^{-ht}\left(1 - e^{-ht}\right)^{s-1}\sin 2\pi kt\,dt.$$

Применяя формулу [4, (8.60)] для вычисления оставшегося интеграла, получим

$$\phi(s) = \frac{H_s}{h} + \frac{1}{2} + \sum_{k=1}^{\infty}\frac{is}{2\pi k}\left[\frac{\Gamma\left(1 + \frac{2\pi ki}{h}\right)\Gamma(s)}{\Gamma\left(s + 1 + \frac{2\pi ki}{h}\right)} - \frac{\Gamma\left(1 - \frac{2\pi ki}{h}\right)\Gamma(s)}{\Gamma\left(s + 1 - \frac{2\pi ki}{h}\right)}\right].$$

Применяя формулу $\Gamma(x+1) = x\Gamma(x)$, перепишем полученное выражение как

$$\phi(s) = \frac{H_s}{h} + \frac{1}{2} - \frac{1}{h}\sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty}\frac{\Gamma\left(\frac{2\pi ki}{h}\right)\Gamma(s+1)}{\Gamma\left(s + 1 + \frac{2\pi ki}{h}\right)}. \tag{16}$$

Применяя формулу Стирлинга, нетрудно вывести

$$\frac{\Gamma(s+1)}{\Gamma(s+1+\alpha)} = s^{-\alpha} + \mathcal{O}(s^{-1}). \tag{17}$$

Отметим, что оценка является равномерной по α.

Подставляя эту оценку для $\alpha = \frac{2\pi ki}{h}$ в (16), получим (11).

3.2. Вывод формулы (13). Гамма-функция убывает по мнимой оси довольно быстро, поэтому для получения приближенной формулы достаточно оставить только первый член ряда ($k = 1$). Вычисление гамма-функции с применением пакета *Mathematica* дает

$$\Gamma\left(1 + \frac{2\pi i}{h}\right) = (3.1766226452 - 3.7861079986i)10^{-6}.$$

Подставляя эти числа в (11), получим (13).

3.3. Вывод формулы (14). Нетрудно проверить, что

$$\Delta^{(m)}\left(\frac{\Gamma(s+1)}{\Gamma(s+1+\alpha)}\right)=(-1)^m\frac{\Gamma(\alpha+m)\Gamma(s-m+1)}{\Gamma(\alpha)\Gamma(s+1+\alpha)}. \tag{18}$$

Подставляя это равенство для $\alpha=\frac{2\pi ki}{h}$ в (16) и применяя оценку (17), получим (14).

Отметим, что при $m>0$ гамма-функция в (14) определена и для $k=0$, поэтому слагаемое $\frac{1}{h}$ внесено под знак суммы.

3.4. Вывод формулы (15). Подставляя асимптотические оценки (14), (11) в (9) и оставляя только первые члены рядов ($k=0,\pm1$), получим формулу (15).

4. Численные вычисления. На рис. 1 для меры Бернулли при $p=q=1/2$ приведены график функции $\phi(s)$ и график функции $H_s/h+0.5$, который сливается с предыдущим.

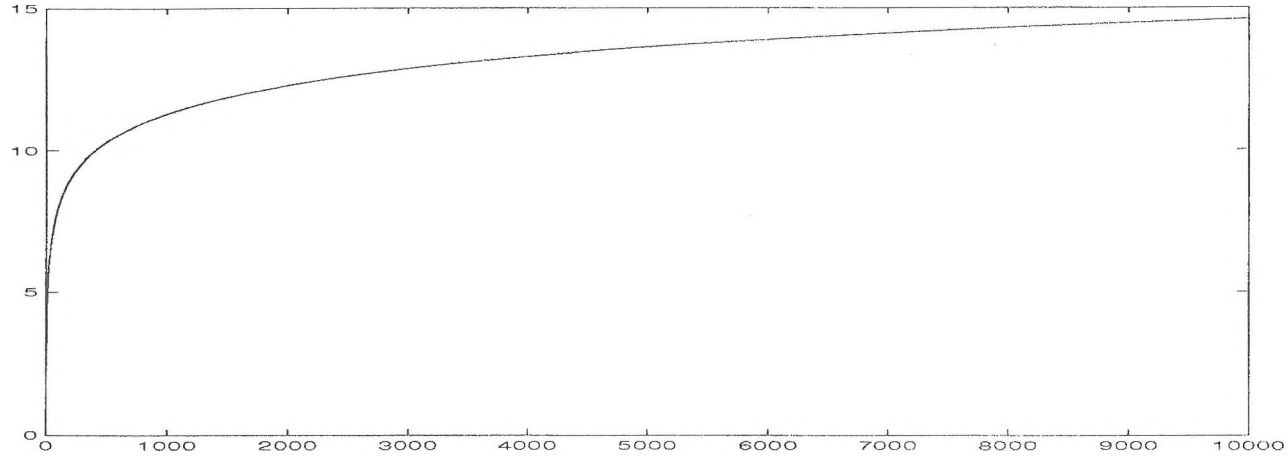

Рис. 1. Графики функций $\phi(s)$ и $H_s/h+0.5$ в зависимости от s

На рис. 2 для меры Бернулли при $p=q=1/2$ приведены графики функций $\phi(s)-H_s/h-0.5$ и ее приближения по формуле (13) (пунктир)

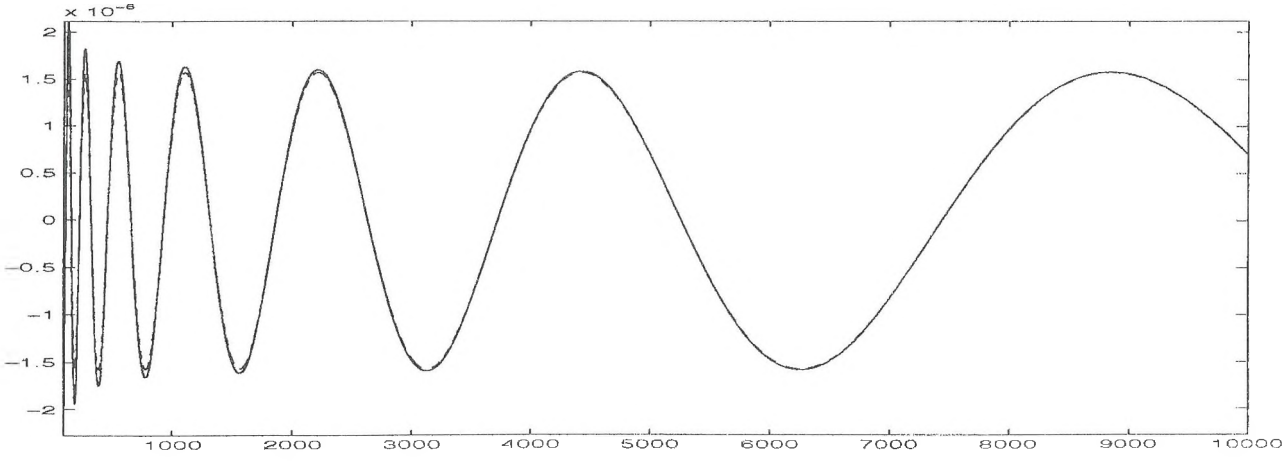

Рис. 2. Графики функций $\phi(s)-H_s/h-0.5$ и ее приближения по формуле (13) (пунктир) в зависимости от s

На рис. 3 для меры Бернулли при $p = q = 1/2$ приведен график разности функций $\phi(s)$ и ее приближения по формуле (13)

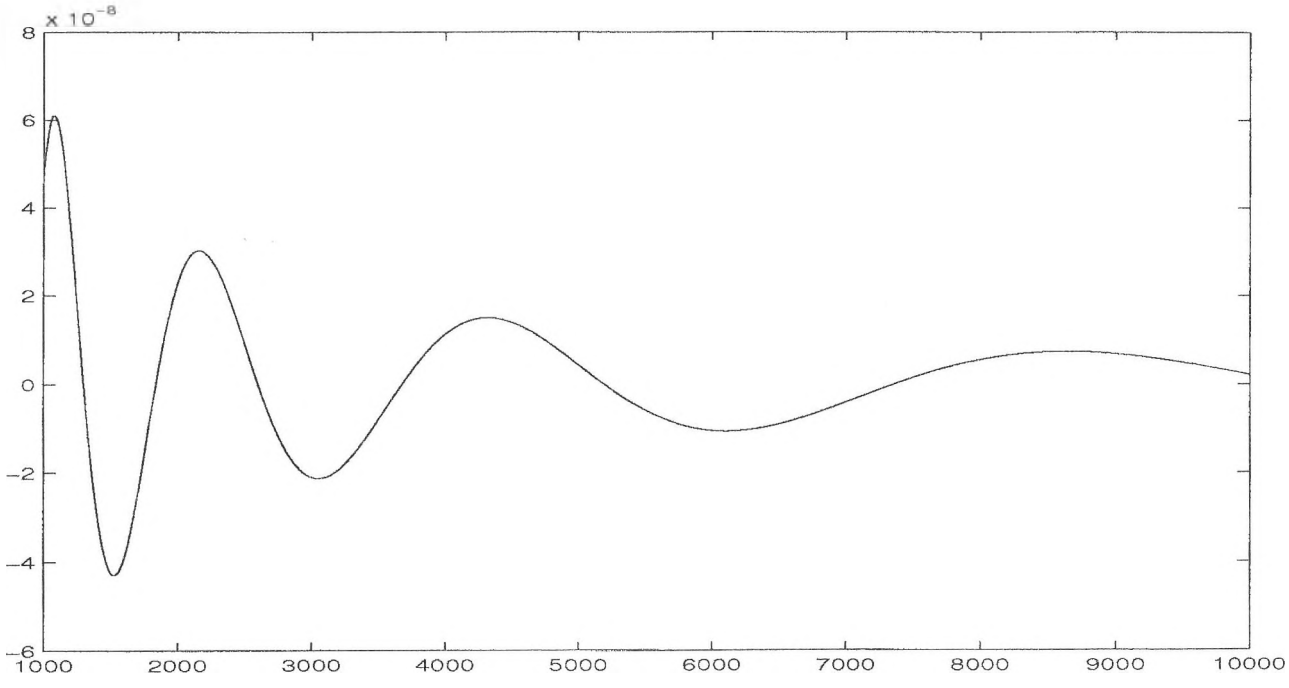

Рис. 3. График разности функций $\phi(s)$ и ее приближения по формуле (13) в зависимости от s

Список литературы

1. Харди Г. Расходящиеся ряды. М.: ИЛ, 1951.

2. Тимофеев Е.А. Состоятельная оценка энтропии мер и динамических систем //Мат. заметки. 2005. Т. 77, №6. С. 903 – 916.

3. Градштейн И.С.,Рыжик И.М. Таблицы интегралов, сумм, рядов и произведений. М.:Наука, 1971.

4. Диткин В.А., Прудников А.П. Интегральные преобразования и операционное исчисление. М. : Наука. 1974.