УДК 004.421.6: 519.23

О СТАТИСТИЧЕСКИХ СВОЙСТВАХ ОЦЕНОК ЭНТРОПИИ ШЕННОНА, ВЫЧИСЛЕННЫХ ПО СОСЕДНИМ ЗНАЧЕНИЯМ ДЛИН ФРАГМЕНТОВ

В. Ю. ПАЛУХА, Ю. С. ХАРИН,
Н. С. ТЕГАКА

НИИ прикладных проблем математики и информатики,
Белорусский государственный университет,
г. Минск, 220030, Республика Беларусь

Введение

Генераторы случайных и псевдослучайных последовательностей являются неотъемлемой частью систем защиты информации. Необходимым условием для данных генераторов является соответствие порождаемых ими последовательностей модели равномерно распределенной случайной последовательности (РРСП). Одним из методов оценки качества генераторов является энтропийный анализ их выходных последовательностей. Наблюдаемая двоичная последовательность делится на непересекающиеся фрагменты длины s (s-граммы), по ним вычисляется статистическая оценка энтропии, для которой известно распределение вероятностей для РРСП, и по значению оценки принимается решение о (не)соответствии модели РРСП. Для уменьшения вероятности принятия ошибочного решения следует вычислить оценку при различных значениях s. Встает вопрос о том, как связаны между собой оценки энтропии, вычисленные при различных значениях s, например, для s и $s + 1$. В докладе показывается, что с ростом s значения оценок энтропии Шеннона, вычисленных по s- и $(s + 1)$-граммам, становятся статистически независимы.

Математическая модель. Пусть имеется двоичная последовательность $\{x_t\}$ длины T. Введем в рассмотрение гипотезу H_0 = {{x_t} является РРСП} = {{x_t} – н. о. р. с. в., $p_k = P\{x_t = k\} = p_k^0 = \frac{1}{2}$, $t = 1, \ldots, T$, $k = 0, 1\}$. Пусть $T = Ms(s + 1)$, тогда $\{x_t\}$ можно разбить на $M(s + 1)$ непересекающихся s-грамм и Ms непересекающихся $(s + 1)$-грамм.

Частотные оценки вероятностей s-грамм $\langle X_j \rangle$ и $(s + 1)$-грамм $\langle X'_l \rangle$ определяются соотношениями:

$$\hat{p}_i(s) = \frac{1}{M(s+1)} \sum_{j=1}^{M(s+1)} \mathbf{I}\left\{\langle X_j \rangle = i\right\} = \frac{\nu_i}{M(s+1)}, \quad i = 0, \ldots, 2^s - 1;$$

$$\hat{p}_k(s+1) = \frac{1}{Ms} \sum_{j=1}^{Ms} \mathbf{I}\left\{\langle X'_l \rangle = k\right\} = \frac{\nu_k}{Ms}, \quad k = 0, \ldots, 2^{s+1} - 1, \tag{1}$$

$$\mathbf{I}\{A\} = \begin{cases} 1, \text{ если событие } A \text{ верно,} \\ 0, \text{ в противном случае.} \end{cases}$$

Статистические оценки s- и $(s + 1)$-мерной энтропии Шеннона на основе частотных оценок вероятностей (1) имеют вид

$$\hat{H}(s) = -\sum_{i=0}^{2^s-1} \hat{p}_i(s) \ln \hat{p}_i(s) \in [0, s \ln 2],$$

$$\hat{H}(s+1) = -\sum_{k=0}^{2^{s+1}-1} \hat{p}_k(s+1) \ln \hat{p}_k(s+1) \in [0, (s+1) \ln 2]. \tag{2}$$

Для того, чтобы вычислить ковариацию оценок энтропии (2) $\operatorname{cov}_0\left\{\hat{H}(s), \hat{H}(s+1)\right\}$ (здесь и далее индекс «0» означает вычисление при истинной гипотезе H_0), вычислим вначале 2^{2s+1} ковариаций частотных оценок вероятностей $\operatorname{cov}_0\left\{\hat{p}_i(s), \hat{p}_k(s+1)\right\}$ для всех возможных пар s- и $(s+1)$-грамм (i, k), $i = 0, \ldots, 2^s - 1$, $k = 0, \ldots, 2^{s+1} - 1$. Обозначим:

$$a_{ik} = E_0\left\{\hat{p}_i(s)\hat{p}_k(s+1)\right\}, b_{ik} = E_0\left\{\hat{p}_i(s)\right\} E_0\left\{\hat{p}_k(s+1)\right\},$$

тогда искомая ковариация равна

$$d_{ik} = \operatorname{cov}_0\left\{\hat{p}_i(s), \hat{p}_k(s+1)\right\} = a_{ik} - b_{ik}. \tag{3}$$

Поскольку $E_0\left\{\hat{p}_i(s)\right\} = 2^{-s}, i = 0, \ldots, 2^s - 1$, $E_0\left\{\hat{p}_k(s+1)\right\} = 2^{-s-1}, k = 0, \ldots, 2^{s+1} - 1$, то $b_{ik} = 2^{-2s-1}, i = 0, \ldots, 2^s - 1, k = 0, \ldots, 2^{s+1} - 1$. Для a_{ik} справедливо выражение

$$a_{ik} = \frac{1}{M(s+1)} \frac{1}{Ms} \sum_{j=1}^{M(s+1)} \sum_{l=1}^{Ms} E_0\left\{I\left\{\left\langle X_j \right\rangle = i, \left\langle X'_l \right\rangle = k\right\}\right\} =$$

$$= \frac{1}{M(s+1)} \frac{1}{Ms} \sum_{j=1}^{M(s+1)} \sum_{l=1}^{Ms} P_0\left\{\left\langle X_j \right\rangle = i, \left\langle X'_l \right\rangle = k\right\}.$$

Значение величины

$$c_{ikjl} = P_0\left\{\left\langle X_j \right\rangle = i, \left\langle X'_l \right\rangle = k\right\},$$

входящей в это выражение для a_{ik}, зависит от взаимного расположения s-граммы $\left\langle X_j \right\rangle$ и $(s+1)$-граммы $\left\langle X'_l \right\rangle$. Если $\left\langle X_j \right\rangle$ и $\left\langle X'_l \right\rangle$ не пересекаются (т. е. не содержат общих элементов ряда $\{x_1, \ldots, x_T\}$), то в силу их независимости при H_0 имеем:

$$c_{ikjl} = P_0\left\{\left\langle X_j \right\rangle = i\right\} P_0\left\{\left\langle X'_l \right\rangle = k\right\} = 2^{-s} \cdot 2^{-s-1} = 2^{-2s-1}, i = 0, \ldots, 2^s - 1, k = 0, \ldots, 2^{s+1} - 1. \tag{4}$$

Зафиксируем $(s+1)$-грамму $\left\langle X'_l \right\rangle$. Она может пересекаться только с двумя s-граммами. Пусть их номера j'_l и $j'_l + 1$. Тогда a_{ik} представимо в виде

$$a_{ik} = \frac{1}{M(s+1)} \frac{1}{Ms} \sum_{l=1}^{Ms} \sum_{j \neq j'_l, j'_l+1} c_{ijkl} + \frac{1}{M(s+1)} \frac{1}{Ms} \sum_{l=1}^{Ms} \left(c_{ij'_l kl} + c_{i(j'_l+1)kl}\right). \tag{5}$$

Из (4) для первого слагаемого в (5) следует

$$\frac{1}{M(s+1)} \frac{1}{Ms} \sum_{l=1}^{Ms} \sum_{j \neq j'_l, j'_l+1} c_{ijkl} = \frac{1}{M(s+1)} \frac{1}{Ms} \sum_{l=1}^{Ms} \sum_{j \neq j'_l, j'_l+1} \frac{1}{2^{2s+1}} = \frac{M(s+1)-2}{2^{2s+1} M(s+1)}. \tag{6}$$

Перед рассмотрением пересечений s- и $(s+1)$-грамм введем обозначение:

$$l' = \begin{cases} l \bmod s, \text{ если } l \bmod s > 0; \\ s, \text{ если } l \bmod s = 0 \end{cases} \in \{1, 2, \ldots, s\}.$$

Пусть длина пересечения $(s+1)$-граммы $\langle X'_l \rangle$ и s-граммы $\langle X_{j'_l+1} \rangle$ равна m_l, а длина пересечения $\langle X'_l \rangle$ и s-граммы $\langle X_{j'_l} \rangle$ равна соответственно $s+1-m_l$. В [1] показано, что справедливо соотношение $m_l = l'$. При пересечении $(s+1)$-граммы $\langle X'_l \rangle$ и s-граммы $\langle X_{j'_l} \rangle$ полученный вектор имеет длину $s+1+s-(s+1-m_l) = s+l'$, соответствующие пересечению индексы i и k обозначим соответственно мультииндексами $I_{l'}^s$ и $K_1^{s+1-l'}$. При пересечении $(s+1)$-граммы $\langle X'_l \rangle$ и s-граммы $\langle X_{j'_l+1} \rangle$ полученный вектор имеет длину $s+1+s-m_l = 2s+1-l'$, а соответствующие пересечению индексы i и k обозначим соответственно мультииндексами $I_1^{l'}$ и $K_{s+2-l'}^{s+1}$.

Теорема 1. [1] *При истинной гипотезе H_0 для ковариаций (3) справедливо следующее выражение:*

$$d_{ik} = \frac{1}{M(s+1)} \cdot \frac{1}{2^s} \left(\frac{1}{s} \sum_{l=1}^{s} \left(\frac{1}{2^l} \mathbf{I}\left\{ I_l^s = K_1^{s+1-l} \right\} + \frac{1}{2^{s+1-l}} \mathbf{I}\left\{ I_1^l = K_{s+2-l}^{s+1} \right\} \right) - \frac{1}{2^s} \right). \tag{7}$$

Заметим, что индикаторы в (7) представимы в виде

$$\mathbf{I}\left\{ I_l^s = K_1^{s+1-l} \right\} = \prod_{j=1}^{s+1-l} \left(i_{l+j-1} \oplus k_j \oplus 1 \right), \quad \mathbf{I}\left\{ I_1^l = K_{s+2-l}^{s+1} \right\} = \prod_{j=1}^{l} \left(i_j \oplus k_{j+s+1-l} \oplus 1 \right).$$

Следствие 1. *Справедлива следующая точная двухсторонняя оценка для ковариаций (3):*

$$-\frac{1}{2^{2s} M(s+1)} \le \operatorname{cov}_0 \left\{ \hat{p}_i(s), \hat{p}_k(s+1) \right\} \le \frac{1}{2^{s-1} s(s+1) M} \left(1 - \frac{s+2}{2^{s+1}} \right). \tag{8}$$

Нижняя граница достигается в случае, если i состоит из одних нулей, а k из одних единиц, или наоборот, тогда все индикаторы в (7) будут равны 0. Верхняя граница достигается в случае, если i и k состоят из одних нулей (или единиц), тогда все индикаторы в (7) будут равны 1.

Следствие 2. *Коэффициент корреляции частотных оценок*

$$Corr_0 \left\{ \hat{p}_i(s), \hat{p}_k(s+1) \right\} = \frac{\operatorname{cov}_0 \left\{ \hat{p}_i(s), \hat{p}_k(s+1) \right\}}{\sqrt{D_0 \left\{ \hat{p}_i(s) \right\} D_0 \left\{ \hat{p}_k(s+1) \right\}}}. \tag{9}$$

удовлетворяет двухстороннему неравенству

$$-2\sqrt{\frac{s}{(s+1)\left(2^s-1\right)\left(2^{s+1}-1\right)}} \le Corr_0 \left\{ \hat{p}_i(s), \hat{p}_k(s+1) \right\} \le \frac{2^{s+2}-2s-4}{\sqrt{s(s+1)\left(2^s-1\right)\left(2^{s+1}-1\right)}}. \tag{10}$$

Доказательство. Для дисперсий частотных оценок справедливо

$$D_0\{\hat{p}_i(s)\} = D_0\left\{\frac{1}{M(s+1)}\sum_{j=1}^{M(s+1)}\mathbf{I}\left\{\left\langle X_j\right\rangle = i\right\}\right\} = \frac{1}{2^s M(s+1)}\left(1-\frac{1}{2^s}\right), i = 0,\ldots,2^s-1;$$
$$D_0\{\hat{p}_k(s+1)\} = \frac{1}{2^{s+1}Ms}\left(1-\frac{1}{2^{s+1}}\right), \quad k = 0,\ldots,2^{s+1}-1. \tag{11}$$

Тогда

$$\sqrt{D_0\{\hat{p}_i(s)\}D_0\{\hat{p}_k(s+1)\}} = \frac{\sqrt{\left(2^s-1\right)\left(2^{s+1}-1\right)}}{2^{2s+1}M\sqrt{s(s+1)}}.$$

Подставим в (9), получим

$$Corr_0\{\hat{p}_i(s),\,\hat{p}_k(s+1)\} = \frac{2^{2s+1}M\sqrt{s(s+1)}\,\mathrm{cov}_0\{\hat{p}_i(s),\,\hat{p}_k(s+1)\}}{\sqrt{\left(2^s-1\right)\left(2^{s+1}-1\right)}},$$

и основе двухсторонней оценки (8) получим двухстороннюю оценку коэффициента корреляции (10).

Вернемся к оценкам энтропии (2).

Теорема 2. *Пусть* $0 < \hat{p}_i \le 2^{1-s}, i = 0,\ldots,2^s-1$. *Тогда при истинной гипотезе* H_0 *для ковариации и коэффициента корреляции оценок энтропии* $\hat{H}(s)$ *и* $\hat{H}(s+1)$ *справедливы следующие асимптотические выражения при* $M \to \infty$*:*

$$\mathrm{cov}_0\left\{\hat{H}(s),\hat{H}(s+1)\right\} = \frac{2^{s+1}-s-2}{M^2s^2(s+1)^2}, \tag{12}$$

$$corr_0\left\{\hat{H}(s),\hat{H}(s+1)\right\} = \frac{2\left(2^{s+1}-s-2\right)}{s(s+1)\sqrt{(2^s-1)(2^{s+1}-1)}}. \tag{13}$$

Схема доказательства. Из разложения логарифма в ряд Тейлора следует, что главный член асимптотического разложения $\mathrm{cov}_0\left\{\hat{H}(s),\hat{H}(s+1)\right\}$ равен $2^{2s}\sum_{i=0}^{2^s-1}\sum_{k=0}^{2^{s+1}-1}d_{ik}^2$, а остаточным членом при $M \to \infty$ можно пренебречь. В свою очередь, для суммы квадратов ковариаций (3) по всем возможным парам *s*- и (*s* + 1)-грамм справедливо выражение $\sum_{i=0}^{2^s-1}\sum_{k=0}^{2^{s+1}-1}d_{ik}^2 = \frac{2^{s+1}-s-2}{2^{2s}M^2s^2(s+1)^2}$, откуда следует выражение (12) для ковариации оценок энтропии.

Также из разложения логарифма в ряд Тейлора с учетом (11) следует, что главные члены асимптотических разложений $D_0\left\{\hat{H}(s)\right\}$ и $D_0\left\{\hat{H}(s+1)\right\}$ равны соответсвенно $\frac{2^s-1}{2M^2(s+1)^2}$

и $\frac{2^{s+1}-1}{2M^2s^2}$, следовательно, $\sqrt{D_0\left\{\hat{H}(s)\right\}D_0\left\{\hat{H}(s+1)\right\}} = \frac{\sqrt{(2^s-1)(2^{s+1}-1)}}{2M^2s(s+1)}$, откуда следует выражение (13) для коэффициента корреляции оценок энтропии.

Следствие 3. *Из (13) следует, что с ростом s коэффициент корреляции стремится к нулю.*

Компьютерные эксперименты. Для демонстрации справедливости формул (12) – (13) и следствия 3 проведена серия компьютерных экспериментов. Рассматривалась псевдослучайная последовательность, полученная алгоритмом BelT [2] в режиме счетчика. Для $s = 2, \ldots, 10$ из наблюдаемой последовательности брались $K = 1000$ фрагментов длины $T = Ms(s + 1)$ с фиксированным значением $M = 10000$. По полученным K фрагментам для каждого значения s вычислялись оценки энтропии (2), по которым затем вычислялись выборочные ковариации и коэффициенты корреляции. Также для указанных значений параметров вычислены теоретические значения ковариации и коэффициента корреляции оценок энтропии (2). На рисунках 1 и 2 приведено сравнение выборочных и теоретических значений ковариации и коэффициента корреляции оценок энтропии соответственно. Как видно из рисунков, полученные экспериментально значения ковариации и коэффициента корреляции оценок энтропии (2) близки к теоретическим значениям, а также коэффициент корреляции стремится к 0 с ростом s.

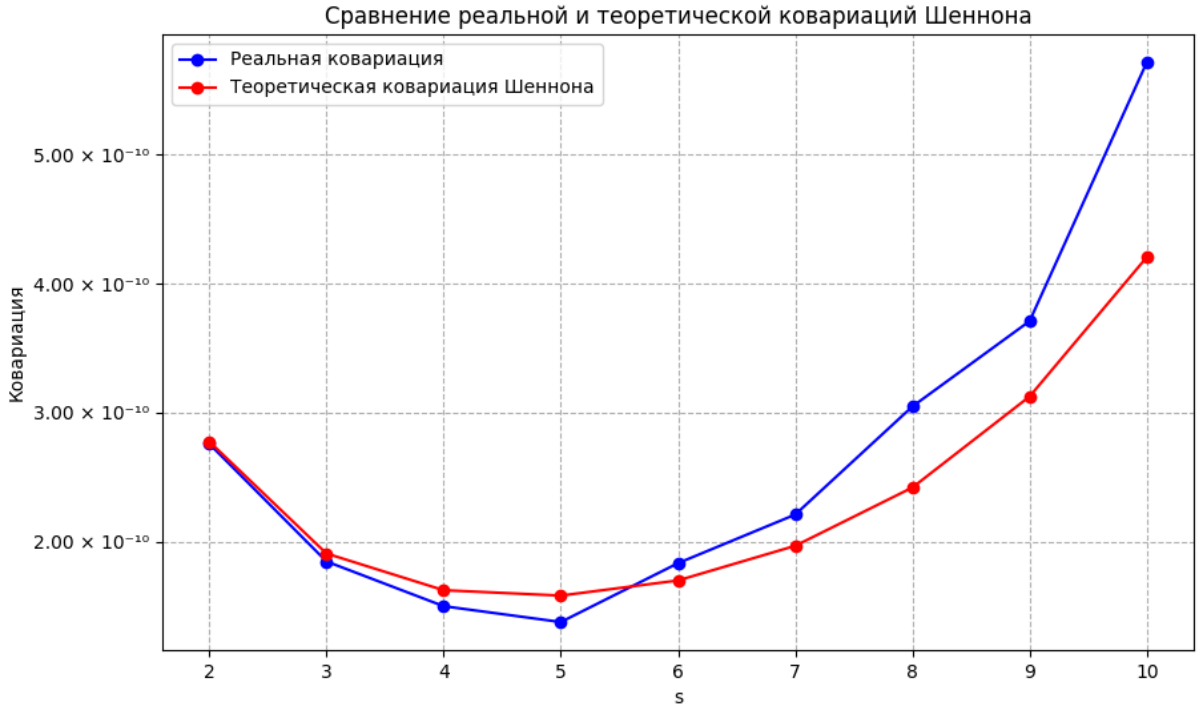

Рисунок 1 – Сравнение выборочной и теоретической ковариации оценок энтропии Шеннона, вычисленных по s- и (s + 1)-граммам, в зависимости от s

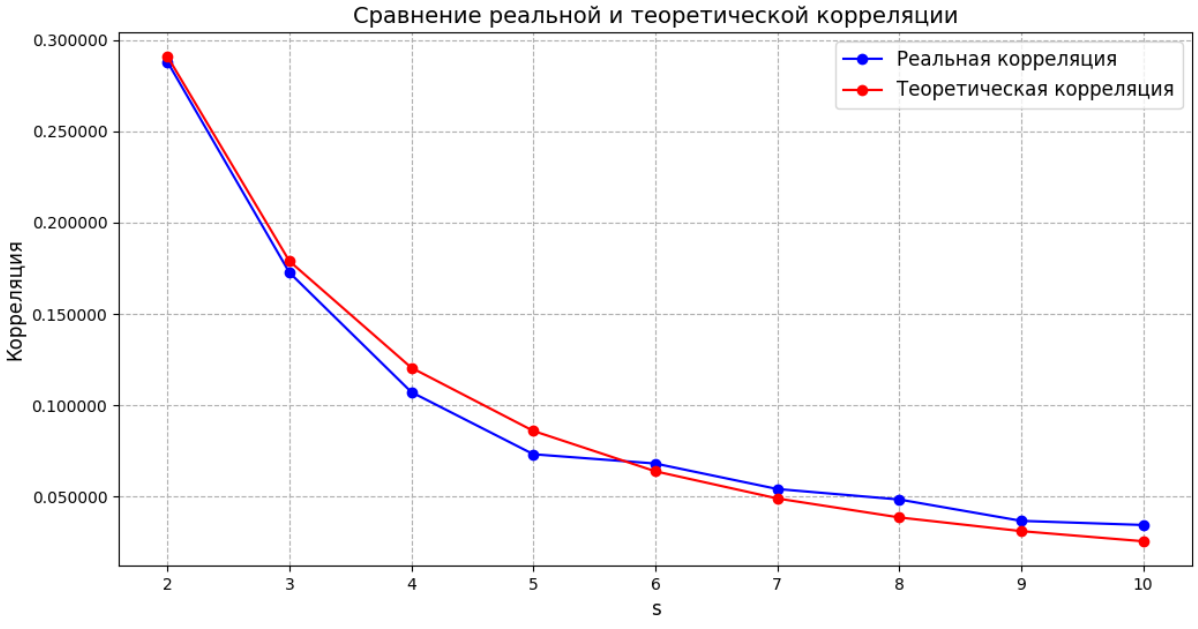

Рисунок 2 – Сравнение выборочного и теоретического коэффициента корреляции оценок энтропии Шеннона, вычисленных по s- и (s + 1)-граммам, в зависимости от s

Заключение

В докладе рассмотрены статистические свойства оценок энтропии двоичной последовательности, вычисленных по фрагментам длины s и $s + 1$, в случае справедливости гипотезы H_0 о том, что наблюдаемая последовательность является РРСП. Найдены асимптотические выражения для ковариации и коэффициента корреляции оценок энтропии, показано, что с ростом s коэффициент корреляции стремится к 0, что свидетельствует о слабой зависимости между оценками энтропии. Проведена серия компьютерных экспериментов, подтверждающая полученные результаты.

Литература

1. Palukha, U. Yu. On statistical testing of random and pseudorandom sequences based on entropy functionals / U. Yu. Palukha, Yu. S. Kharin, A. I. Siarheeu, A. A. Arlou // Computer Data Analysis and Modeling: Stochastics and Data Science. Proceedings of the XIII International Conference, Minsk, September 6–10, 2022 / BSU; eds.: Yu. Kharin [et al.]. – Minsk: BSU, 2022. – P. 148–162.

2. Информационные технологии и безопасность. Криптографические алгоритмы генерации псевдослучайных чисел = Інфармацыйныя тэхналогіі і бяспека. Крыптаграфічныя алгарытмы генерацыі псеўдавыпадковых лікаў: СТБ 34.101.47–2017. – Взамен СТБ 34.101.47–2012; введен 01.09.2017. – Минск: Госстандарт, 2017. – III, 21 с.