Министерство образования республики беларусь
Белорусский Государственный Университет
Факультет Прикладной Математики и Информатики
Кафедра математического моделирования и анализа данных

Обнаружение вкраплений в марковскую последовательность на основе энтропийных характеристик

Курсовая работа

Шимко Андрея Чеславовича
студента 4 курса,
специальность «Компьютерная безопасность»

Научный руководитель:
ассистент кафедры ММАД
Е. В. Вечерко

Минск, 2016

Содержание

1 Введение

Стеганография представляет собой специфическую область человеческой деятельности, связанной с разработкой и анализом методов сокрытия факта передачи информации. Подобно криптографии, стеганография известна со времен античности. Но на этом аналогии, по крайне мере в контексте теоретических исследований, заканчиваются. За последние четверть века возникла и успешно раз- вивается новая математическая дисциплина криптология, или, что то же самое, математическая криптография, изучающая математические модели криптографических схем. Попытки создания ма- тематической стеганографии (которую, быть может, следует именовать также стеганологией) пред- принимаются, но исследования здесь находятся лишь в зачаточном состоянии.

Такое положение дел обусловлено прежде всего сложностью возникающих в стеганографии задач. Всякая попытка построения математических моделей стеганографических систем сопряжена с необходимостью рассмотрения большого количества случаев и подслучаев, не допускающих простой и единообразной трактовки. Другими словами, внешняя среда, в которой должны функционировать стеганографические системы, имеет гораздо большее, по сравнению с внешней средой криптографических схем, количество степеней свободы.

Основными критериями для оценки и сравнения различных методов построения стеганографических систем являются их стойкость и емкость. В отличие от достаточно исследованных криптографических систем, оценки стойкости стегосистем более сложны и само понятие стойкости имеет большое число различных формулировок, что объясняется разнообразием задач стеганографи-ческой защиты данных. В настоящей работе исследуются методы построения стегосистем, предназначенных для скрытия факта передачи конфиденциальных сообщений. Говоря о стойкости криптографических систем, важно упомянуть о принципе Керкхоффса, который заключается в том, что система защиты информации должна обеспечивать свои функции даже при полной информированности противника о ее структуре и алгоритмах, и вся секретность системы должна заключаться в ключе. Этот принцип также можно соотнести с определением стойкости стегосистем. В данном случае, ключом может являться, например, секретная последовательность, определяющая порядок прохода, элементов контейнера при внедрении бит информации, что имеет место в алгоритмах рассеянного заполнения контейнеров. Второй критерий, емкость метода, определяет максимальное количество встраиваемой информации, и может выражаться в единицах бит на пиксель.

2 Основные понятия

Определение 2.1. *Стеганография - наука о способах передачи (хранения) скрытой информации, при которых скрытый канал организуется на базе и внутри открытого канала с использованием особенностей восприятия информации, причем для этой цели могут использоваться такие приемы, [2]как:*

1. *Полное сокрытие факта существования скрытого канала связи*
2. *Создание трудностй для обнаружения, извлечения и модификации передаваемых скрытых сообщений внутри открытых сообщений-контейнеров*
3. *Маскировки скрытой информации в протоколе*

Определение 2.2. *Контейнером $b \in B$ (носителем) называют несекретные данные, которые используют для сокрытия сообщения [2].*

Определение 2.3. *Сообщение $m \in M$ - секретная информация, наличие которой в контейнере необходимо скрыть [2].*

Определение 2.4. *Ключ $k \in K$ - секретная информация, известная только законному пользователю, которая определяет конкретный вид алгоритма сокрытия [2].*

Определение 2.5. *Пустой контейнер - контейнер, не содержащий сообщения [2].*

Определение 2.6. *Заполненный контейнер - контейнер, с внедренным в него сообщением [2].*

Определение 2.7. *Стеганографический алгоритм - два преобразования: прямое стеганографическое преобразование $F : M \times B \times K \to B$ и обратное стеганографическое преобразование $F^{-1} : B \times K \to M$, сопоставляющее соответственно тройке (сообщение, пустой контейнер, ключ) контейнер-результат и паре (заполненный контейнер, ключ) - исходное сообщение, [2]причем:*

$$F(m, b, k) = b_{m,k}, F^{-1}(b_{m,k}, k) = m, m \in M, b_{m,k}, b \in B, k \in K \quad (1)$$

Определение 2.8. *Под стеганографической системой будем понимать $S = (F, F^{-1}, M, B, K)$, представляющую собой совокупность сообщений, секретных ключей, контейнеров и связывающих их преобразований [2].*

Определение 2.9. *Внедрение (сокрытие) - применение прямого стегонаграфического преобразования к конкретным контейнеру, ключу и сообщению [2].*

Определение 2.10. *Извлечение сообщения - применение обатного стегонаграфического преобразованияю [2].*

Определение 2.11. *l-грамм - подпоследовательность из l подряд иущих элементов последовательности.*

Определение 2.12. *Под дискретным источником сообщений будем понимать устройство, порождающее последовательности, составленные из букв конечного алфавита A ($|A|=n<\infty$). При этом буквы последовательностей пораждаются в дискретный момент времени:$t = 0, 1, 2, ...; t = ..., -2, -1, 0, 1, 2, ...$; [1]*

Определение 2.13. *Если вероятность того, что источник порождает некоторую последовательность $a_{i_1}...a_{i_l}$, составленную из букв алфавита A, в момент времени 1, 2, ... l, равна вероятности того, что порождается точно такая же последовательность в момент времени j+1, ..., j+l для любых j,l;$a_{j_1}...a_{j_l}$, то источник называется стационарным.[1] Стационарность означает неизменность во времени всех конечномерных распределений соответствующего случайного процесса.*

Определение 2.14. *Энтропией источника назовем величину:*

$$H_\infty = \lim_{l\to\infty} \frac{H(C_l)}{l}; \tag{2}$$

если данный предел существует[1].

3 Теоретико-информационный подход к оценке стойкости систем

В настоящем разделе рассматривается теоретико-информационный подход к определению стойкости стеганосистем для стеганографических каналов без повторений в присутствии пассивного противника, обладающего неограниченными вычислительными возможностями.

В основе всех известных определений стойкости стеганосистем лежит требование неотличимости распределения вероятностей на множестве стего от распределения вероятностей на множестве пустых контейнеров. Рассматривается статистическая неотличимость, или, иначе говоря, неотличимость относительно произвольных алгоритмов.

Парадигма неотличимости распределений вероятностей заимствована из математической криптографии. Заметим, однако, что ее адекватность для стеганографии не очевидна. По крайней мере, в случае стеганографического канала без повторений не ясно, насколько оправданными будут усилия отправителя по имитации распределения вероятностей на множестве пустых контейнеров. Не следует ли вместо этого стремиться передать скрытое сообщение в одном из наиболее вероятных контейнеров?

Вполне очевидна идея создания стеганографического канала путем маскировки скрытого сообщения под шум, вносимый алгоритмом шифрования в исходный контейнер.

Проверка гипотезы о стойкости системы состоит в том, чтобы определить, какая из двух гипотез - H_0 или H_1 - является верным толкованием наблюдаемой величины Q. Есть два возможных распределения вероятностей, которые принято обозначать P_{Q_0}, P_{Q_1}, над пространством возможных наблюдений. Если верна гипотеза H_0, тогда Q была порождена согласно P_{Q_0}, если же верна гипотеза H_1, тогда Q была порождена согласно P_{Q_1}. Правило принятия решения - это двоичное отображение, заданное на пространстве возможных наблюдений, которое составляет одну из двух возможных гипотез для каждого возможного элемента q. Основной мерой проверки гипотезы является относительная энтропия или различие между двумя распределениями вероятности P_{Q_0} и P_{Q_1}, определяемое следующим выражением:

$$H(P_{Q_0}||P_{Q_1}) = \sum_q P_{Q_0}(q) log \frac{P_{Q_0}(q)}{P_{Q_1}(q)}; \tag{3}$$

Относительная энтропия между двумя распределениями всегда неотрицательна и равна нулю только тогда, когда распределения равномерны. Несмотря на то, что остиельная энтропия не является метрикой с точки зрения математики (так как не симметрична и не удовлетворяет аксиоме треугольника), полезно считать ее таковой. Двоичная относительная энтропия $d(\alpha, \beta)$ определяется как:

$$d(\alpha, \beta) = \alpha \log \frac{\alpha}{1-\beta} + (1-\alpha) \log \frac{1-\alpha}{\beta}; \tag{4}$$

где α - вероятность ошибки первого рода, β - вероятность ошибки второго рода.

4 Математическая модель вкраплений на основе схемы независимых испытаний

Контейнер представляет собой последовательность случайных величин распределенных по закону Бернулли с параметром p:

$$\mathcal{L}x_t = Bi(1,p), x_i \in V = 0,1, i = \overline{1,T}; \tag{5}$$

Вкрапляемое сообщение имеет вид:

$$\mathcal{L}m_t = Bi(1,\theta), m_i \in V = 0,1, i = \overline{1,\tau}; \tag{6}$$

Ключ γ_t определяет момент времени вкрапления i-того бита сообщения в исходный контейнер:

$$\mathcal{L}\gamma_t = Bi(1,\delta), \gamma_i \in V = 0,1, i = \overline{1,T}; \tag{7}$$

Вкрапление битов m_t производится по правилу, заданному следующим функциональным преобразованием:

$$y_t = (1-\gamma_t)x_t + \gamma_t m_{\tau_t} = \sum_{j=1}^{t} \gamma_j; \tag{8}$$

Лемма 4.1. *Для модели 5-8*

$$P\{y_t = 1\} = (1-\delta)p + \delta\theta; \tag{9}$$

$$P\{y_t = 0\} = (1-\delta)(1-p) + \delta(1-\theta); \tag{10}$$

Доказательство. Воспользуемся формулой полной вероятности:
$P\{y_t = 1\} = P\{(1-\gamma_t)x_t + \gamma_t m_{\tau_t} = 1\} = \sum_{j\in V} P\{y_t = 1, \gamma_t = j\} = \sum_{j\in V} P\{\gamma_t = j\}P\{y_t = 1|\gamma_t = j\} = (1-\delta)P\{x_=1, \gamma_t = 0\} + \delta P\{m_{\tau_t} = 1, \gamma_t = 1\} = (1-\delta)p + \delta\theta;$
Тогда:
$P\{y_t = 0\} = 1 - P\{y_t = 1\} = (1-\delta)(1-p) + \delta(1-\theta);$ □

$$h = \frac{H(y_1,...,y_t)}{T} = \frac{TH(y_1)}{T} = H(y_1); \tag{11}$$

Воспользуемся леммой 4.1:

$$h = -P\{y_t = 1\}\log_2 P(y_t = 1) - P\{y_t = 0\}\log_2 P(y_t = 0) = -((1-\delta)p+\delta\theta)\log_2((1-\delta)p+\delta\theta) - ((1-\delta) \tag{12}$$

5 Математическая модель вкраплений в двоичную стационарную марковскую последовательность 1-го порядка и ее свойства

Рассмотрим модель 5-8.

Пусть контейнер 5 пердставляет собой цепь Маркова 1-го порядка, с вектором распределения вероятностей $\pi = (\frac{1}{2}, \frac{1}{2})$, и матрицей вероятностей одношаговых переходов

$$P(\epsilon) = \begin{pmatrix} \epsilon, (1-\epsilon) \\ (1-\epsilon), \epsilon \end{pmatrix} \tag{13}$$

Лемма 5.1. *Для модели 5-8 с условием 13:*

$$P\{y_{t-1} = 1, y_t = 1\} = \frac{1}{2}\epsilon(1-\delta)^2 + \theta\delta(1-\delta) + \theta^2\delta^2; \tag{14}$$

$$P\{y_{t-1} = 1, y_t = 0\} = \frac{1}{2}(1-\epsilon)(1-\delta)^2 + \frac{1}{2}\delta(1-\delta) + \theta(1-\theta)\delta^2; \tag{15}$$

$$P\{y_{t-1} = 0, y_t = 1\} = \frac{1}{2}(1-\epsilon)(1-\delta)^2 + \frac{1}{2}\delta(1-\delta) + \theta(1-\theta)\delta^2; \tag{16}$$

$$P\{y_{t-1} = 0, y_t = 0\} = \frac{1}{2}\epsilon(1-\delta)^2 + \delta(1-\theta)(1-\delta) + \delta^2(1-\theta)^2; \tag{17}$$

Доказательство. Рассмотрим биграмм: $\{y_{t-1}, y_t\}$
$(a_1, a_2) \in \{0,1\}, P\{y_{t-1} = a_1, y_t = a_2\} = \sum_{(b_1,b_2)\in\{0,1\}} P\{y_{t-1} = b_1, y_t = b_2, \gamma_{t-1} = a_1, \gamma_t = a_2\} = \sum_{(b_1,b_2)\in\{0,1\}} P\{y_{t-1} = b_1, y_t = b_2 | \gamma_{t-1} = a_1, \gamma_t = a_2\} P\{\gamma_{t-1} = a_1, \gamma_t = a_2\}$;
Для 14:
$\sum_{(b_1,b_2)\in\{0,1\}} P\{y_{t-1} = b_1, y_t = b_2 | \gamma_{t-1} = a_1, \gamma_t = a_2\} P\{\gamma_{t-1} = a_1, \gamma_t = a_2\} = \frac{1}{2}\epsilon(1-\delta)^2 + \theta\delta(1-\theta) + \theta^2\delta^2$;
Для случаев 15-17 доказывается аналогично. □

Для формул 17-14 справедливо условие нормировки:
$\sum_{(a_1,a_2)\in\{0,1\}} P\{y_{t-1} = a_1, y_t = a_2\} = 1$;

Определение 5.1. *Дискретный стационарный источник называется марковским источником порядка m, если для любого l(l>m) и любой последовательности* $c_l = (a_{i_1}, ..., a_{i_l})$ *выполняется:* $P\{a_{i_l} | a_{i_{l-1}}, ..., a_{i_1}\} = P\{a_{i_l} | a_{i_{l-1}}, ..., a_{i_{l-m+1}}\}$

Определение 5.2. *Величина:* $H^{(k)} = \sum_{C_k} P\{a_{i_1}, ..., a_{i_k}\} \log P\{a_{i_k} | a_{i_{k-1}} ..., a_{i_1}\}$ *называется шаговой энтропией марковского источника порядка k.*

Введем понятие энтропии на знак для биграмма:

$$H_2^{(\delta)} = -\frac{1}{2} \sum_{(a_1,a_2)\in\{0,1\}} P\{y_{t-1} = a_1, y_t = a_2\} \log P\{y_{t-1} = a_1, y_t = a_2\}; \tag{18}$$

Аналогично введем понятие энтропии на знак для l-грамма:

$$H_l^{(\delta)} = -\frac{1}{l} \sum_{(a_1,...,a_l)\in\{0,1\}} P\{y_{t-l} = a_1, ..., y_{t-1} = a_l\} \log P\{y_{t-l} = a_1, ..., y_{t-1} = a_l\}; \tag{19}$$

При $\delta = 0$ y совпадает с x, тогда:

$$H_l^{(0)} = -\frac{1}{l}(H\{x_1\} + (l-1)H\{x_2|x_1\}); \tag{20}$$

$$\lim_{l\to\infty} H_l^{(0)} = \lim_{l\to\infty} -\frac{1}{l}(H\{x_1\} + (l-1)H\{x_2|x_1\}) = H\{x_2|x_1\}; \tag{21}$$

Лемма 5.2. $H_2^{(\delta)} = -2\Bigg(\frac{\varepsilon}{2}\log(\varepsilon) + \frac{(1-\varepsilon)}{2}\log(1-\varepsilon) - \frac{1}{2} + \delta\Bigg(\frac{(1-2\varepsilon)}{2}\log(\varepsilon)$
$- \frac{(1-2\varepsilon)}{2}\log(1-\varepsilon)\Bigg) + \delta^2\Bigg(2\frac{2\epsilon+1}{\epsilon} - \log\frac{\epsilon}{2} - \frac{\epsilon^2-\frac{1}{2}\epsilon}{(1-\epsilon)^2} + \log\frac{1}{2}(1-\epsilon) - 2\frac{2\epsilon+1}{(1-\epsilon)}\Bigg)\Bigg) + O(\delta^3)$

Доказательство. Данное утверждение получено разложением 18 в ряд Тейлора. □

Определение 5.3. *Величина* $\lim_{k\to\infty} H^{(k)} = \lim_{k\to\infty} H_k = H_\infty \geqslant 0$ *- называется энтропий марковского источника, где* H_k *- энтропия на знак.*

Теорема 5.1. *(Первая теорема Шеннона для марковский источников порядка m=1) Для любых* $\epsilon > 0$ *и* $\eta > 0$ *существует* l_0 *такое, что при* $l > l_0$ *все реализации длины* l *марковского источника могут быть разбиты на два класса:* $C_l = C_l' + C_l''$*. Для любой последовательности* $c_l' \in C_l'$ *имеет место:*

$$|\frac{1}{l}\log\frac{1}{p(c_l')} - H_\infty| < \eta \tag{22}$$

, где $H_\infty = -\sum_{(i,j)} p_i p_{ij}\log p_{ij} = -\sum_{(i,j)} p(ij)\log p_{ij}$ *- энтропия источника сообщений.*
Суммарная вероятность последовательностей из класса C_l'' *меньше* ϵ

Рассмотрим триграмм:

Рассмотрим вероятности появления всех возможных шаблонов:
$P\{y_{i-1}=0, y_i=0, y_{i+1}=0\} = (1-\delta)^3 P\{y_{i-1}=0, y_i=0, y_{i+1}=0|\gamma_{i-1}=0, \gamma_i=0, \gamma_{i+1}=0\} +$
$+ \delta(1-\delta)^2(P\{y_{i-1}=0, y_i=0, y_{i+1}=0|\gamma_{i-1}=1, \gamma_i=0, \gamma_{i+1}=0\} + P\{y_{i-1}=0, y_i=0, y_{i+1}=0|\gamma_{i-1}=0, \gamma_i=1, \gamma_{i+1}=0\} + P\{y_{i-1}=0, y_i=0, y_{i+1}=0|\gamma_{i-1}=0, \gamma_i=0, \gamma_{i+1}=1\}) +$
$+ \delta^2(1-\delta)(P\{y_{i-1}=0, y_i=0, y_{i+1}=0|\gamma_{i-1}=1, \gamma_i=1, \gamma_{i+1}=0\} + P\{y_{i-1}=0, y_i=0, y_{i+1}=0|\gamma_{i-1}=0, \gamma_i=1, \gamma_{i+1}=1\} + P\{y_{i-1}=0, y_i=0, y_{i+1}=0|\gamma_{i-1}=1, \gamma_i=0, \gamma_{i+1}=1\}) +$
$+ \delta^3(P\{y_{i-1}=0, y_i=0, y_{i+1}=0|\gamma_{i-1}=1, \gamma_i=1, \gamma_{i+1}=1\})$

$P\{y_{i-1}=0, y_i=0, y_{i+1}=0\} = \frac{1}{8}\big(2\delta^3 + \delta^2(4\varepsilon^2-3) + \delta(2-8\varepsilon^2) + 4\varepsilon^2\big)$

Доказательство. $P\{y_{i-1}=0, y_i=0, y_{i+1}=0|\gamma_1=0, \gamma_2=0, \gamma_3=0\} = P\{x_{i-1}=0, x_i=0, x_{i+1}=0\} = P\{x_{i-1}=0\}P\{x_i=0, x_{i+1}=0|x_{i-1}=0\} = P\{x_{i-1}=0\}P\{x_i=0\}P\{x_{i+1}=0|x_i=0\} = \frac{1}{2}\varepsilon\varepsilon = \frac{\varepsilon^2}{2}$

$P\{y_{i-1}=0, y_i=0, y_{i+1}=0|\gamma_1=1, \gamma_2=0, \gamma_3=0\} = P\{\xi=0, x_i=0, x_{i+1}=0\} = P\{\xi=0\}P\{x_i=0, x_{i+1}=0\} = P\{\xi=0\}P\{x_i=0\}P\{x_{i+1}=0|x_i=0\} = \frac{1}{2}\frac{1}{2}\varepsilon = \frac{\varepsilon}{4}$

$P\{y_{i-1}=0, y_i=0, y_{i+1}=0|\gamma_1=0, \gamma_2=1, \gamma_3=0\} = P\{x_{i-1}=0, \xi=0, x_{i+1}=0\} = P\{\xi=0\}P\{x_{i-1}=0, x_{i+1}=0\} = \frac{1}{2}\frac{1}{2}(\varepsilon^2 + (1-\varepsilon)^2) = \frac{\varepsilon^2+(1-\varepsilon)^2}{4}$

$P\{y_{i-1}=0, y_i=0, y_{i+1}=0|\gamma_1=0, \gamma_2=0, \gamma_3=1\} = P\{x_{i-1}=0, x_i=0, \xi=0\} = P\{\xi=0\}P\{x_{i-1}=0, x_i=0\} = \frac{1}{2}\frac{1}{2}\varepsilon = \frac{\varepsilon}{4}$

$P\{y_{i-1}=0, y_i=0, y_{i+1}=0|\gamma_1=0, \gamma_2=1, \gamma_3=1\} = P\{y_{i-1}=0, y_i=0, y_{i+1}=0|\gamma_1=1, \gamma_2=0, \gamma_3=1\} = P\{y_{i-1}=0, y_i=0, y_{i+1}=0|\gamma_1=1, \gamma_2=1, \gamma_3=0\} = P\{y_{i-1}=0, y_i=0, y_{i+1}=0|\gamma_1=1, \gamma_2=1, \gamma_3=1\} = P\{\xi=0, \xi=0, \xi=0\} = P\{\xi=1\}P\{x_{i-1}=1\}P\{x_i=0\} = \frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{8}$

□

$P\{y_{i-1}=1,y_i=1,y_{i+1}=1\}=\frac{1}{8}\big(2\delta^3+\delta^2(4\varepsilon^2-3)+\delta(2-8\varepsilon^2)+4\varepsilon^2\big)$

Доказательство. $P\{y_{i-1}=1,y_i=1,y_{i+1}=1|\gamma_1=0,\gamma_2=0,\gamma_3=0\}=P\{x_{i-1}=1,x_i=1,x_{i+1}=1\}=P\{x_{i-1}=1\}P\{x_i=1,x_{i+1}=1|x_{i-1}=1\}=P\{x_{i-1}=1\}P\{x_i=1\}P\{x_{i+1}=1|x_i=1\}=\frac{1}{2}\varepsilon\varepsilon=\frac{\varepsilon^2}{2}$

$P\{y_{i-1}=1,y_i=1,y_{i+1}=1|\gamma_1=1,\gamma_2=0,\gamma_3=0\}=P\{\xi=1,x_i=1,x_{i+1}=1\}=P\{\xi=1\}P\{x_i=1,x_{i+1}=1\}=P\{\xi=1\}P\{x_i=1\}P\{x_{i+1}=1|x_i=1\}=\frac{1}{2}\frac{1}{2}\varepsilon=\frac{\varepsilon}{4}$

$P\{y_{i-1}=1,y_i=1,y_{i+1}=1|\gamma_1=0,\gamma_2=1,\gamma_3=0\}=P\{x_{i-1}=1,\xi=1,x_{i+1}=1\}=P\{\xi=1\}P\{x_{i-1}=1,x_{i+1}=1\}=\frac{1}{2}\frac{1}{2}(\varepsilon^2+(1-\varepsilon)^2)=\frac{\varepsilon^2+(1-\varepsilon)^2}{4}$

$P\{y_{i-1}=1,y_i=1,y_{i+1}=1|\gamma_1=0,\gamma_2=0,\gamma_3=1\}=P\{x_{i-1}=1,x_i=1,\xi=1\}=P\{\xi=1\}P\{x_{i-1}=1,x_i=1\}=\frac{1}{2}\frac{1}{2}\varepsilon=\frac{\varepsilon}{4}$

$P\{y_{i-1}=1,y_i=1,y_{i+1}=1|\gamma_1=0,\gamma_2=1,\gamma_3=1\}=P\{y_{i-1}=1,y_i=1,y_{i+1}=1|\gamma_1=1,\gamma_2=0,\gamma_3=1\}=P\{y_{i-1}=1,y_i=1,y_{i+1}=1|\gamma_1=1,\gamma_2=1,\gamma_3=0\}=P\{y_{i-1}=1,y_i=1,y_{i+1}=1|\gamma_1=1,\gamma_2=1,\gamma_3=1\}=P\{\xi=1,\xi=1,\xi=1\}=P\{\xi=1\}P\{x_{i-1}=1\}P\{x_i=0\}=\frac{1}{2}\frac{1}{2}\frac{1}{2}=\frac{1}{8}$

□

$P\{y_{i-1}=1,y_i=1,y_{i+1}=1\}=P\{y_{i-1}=0,y_i=0,y_{i+1}=0\}\Rightarrow$
$P\{y_{i-1}=1,y_i=1,y_{i+1}=0\}=P\{y_{i-1}=0,y_i=0,y_{i+1}=1\}$
$P\{y_{i-1}=1,y_i=0,y_{i+1}=1\}=P\{y_{i-1}=0,y_i=1,y_{i+1}=0\}$
$P\{y_{i-1}=0,y_i=1,y_{i+1}=1\}=P\{y_{i-1}=1,y_i=0,y_{i+1}=0\}$

$P\{y_{i-1}=1,y_i=1,y_{i+1}=0\}=\frac{1}{8}\big(2\delta^3+\delta^2(4\varepsilon-4\varepsilon^2)+2\delta(2\varepsilon-1)^2-4(\varepsilon-1)^2\big)$

$P\{y_{i-1}=0,y_i=1,y_{i+1}=1\}=\frac{1}{8}\big(2\delta^3+\delta^2(4\varepsilon-4\varepsilon^2)+2\delta(2\varepsilon-1)^2-4(\varepsilon-1)^2\big)$

$P\{y_{i-1}=1,y_i=0,y_{i+1}=1\}=\frac{1}{8}\big(2\delta^3+\delta^2(4\varepsilon^2-8\varepsilon+1)+\delta(-8\varepsilon^2+16\varepsilon-6)+4(\varepsilon-1)^2\big)$

Если рассматривать ассимптотическое представление первого порядка вероятности появления шаблонов, достаточно рассмотреть шаблоны '000', '100', '010', '001'. Тогда
$P_3\{y\}=(1-3\delta)P\{y|\gamma_{i-1}=0,\gamma_i=0,\gamma_{i+1}=0\}+\delta(P\{y|\gamma_{i-1}=1,\gamma_i=0,\gamma_{i+1}=0\}+P\{y|\gamma_{i-1}=0,\gamma_i=1,\gamma_{i+1}=0\}+P\{y|\gamma_{i-1}=0,\gamma_i=0,\gamma_{i+1}=1\})$

Получив вероятности можно определить ассимптотическое выражение для энтропии для триграмма:

Лемма 5.3. $H_3(\delta)=-2\Bigg(\varepsilon\log(\varepsilon)+(1-\varepsilon)\log(1-\varepsilon)-\frac{1}{2}+\delta\big((1-2\varepsilon)\log(\varepsilon)-(1-2\varepsilon)\log(1-\varepsilon)+\frac{(1-2\varepsilon)^2}{2\ln(2)}\big)\Bigg)+O(\delta^2)$

Доказательство. Распишем по формуле (19), подставляя полученные выше вероятности.
□

Аналогично рассмотрим 4-грамм:
Рассмотрим вероятности появления всех возможных шаблонов:

$P\{y_{i-1}y_iy_{i+1}y_{i+2}\} = (1-\delta)^4 P\{y_{i-1}y_iy_{i+1}y_{i+2}|'0000'\} + \delta(1-\delta)^3\big(P\{y_{i-1}y_iy_{i+1}y_{i+2}|'0001'\} + P\{y_{i-1}y_iy_{i+1}y_{i+2}|'0010'\} + P\{y_{i-1}y_iy_{i+1}y_{i+2}|'0100'\} + P\{y_{i-1}y_iy_{i+1}y_{i+2}|'1000'\}\big) + \delta^2(1-\delta)^2\big(P\{y_{i-1}y_iy_{i+1}y_{i+2}|'1100'\} + P\{y_{i-1}y_iy_{i+1}y_{i+2}|'1010'\} + P\{y_{i-1}y_iy_{i+1}y_{i+2}|'1001'\} + P\{y_{i-1}y_iy_{i+1}y_{i+2}|'0110'\} + P\{y_{i-1}y_iy_{i+1}y_{i+2}|'0101'\} + P\{y_{i-1}y_iy_{i+1}y_{i+2}|'0011'\}\big) + + \delta^3(1-\delta)\big(P\{y_{i-1}y_iy_{i+1}y_{i+2}|'1110'\} + P\{y_{i-1}y_iy_{i+1}y_{i+2}|'1101'\} + P\{y_{i-1}y_iy_{i+1}y_{i+2}|'1011'\} + P\{y_{i-1}y_iy_{i+1}y_{i+2}|'0111'\}\big) + \delta^4 P\{y_{i-1}y_iy_{i+1}y_{i+2}|'1111'\}$

$P\{y_{i-1}=0, y_i=0, y_{i+1}=0\, y_{i+2}=0\} = \frac{1}{16}\big(\delta^4(4\varepsilon^2-6\varepsilon+4) + \delta^3(-16\varepsilon^2+20\varepsilon-7) + \delta^2(8\varepsilon^3+20\varepsilon^2-22\varepsilon+4) + \delta(-16\varepsilon^3-8\varepsilon^2+8\varepsilon) + 8\varepsilon^3\big)$

Доказательство. $P\{y_{i-1}=0, y_i=0, y_{i+1}=0\, y_{i+2}=0|\gamma_{i-1}=0, \gamma_i=0, \gamma_{i+1}=0, \gamma_{i+2}=0\} = \frac{\varepsilon^3}{2}$

$P\{y_{i-1}=0, y_i=0, y_{i+1}=0\, y_{i+2}=0|\gamma_{i-1}=1, \gamma_i=0, \gamma_{i+1}=0, \gamma_{i+2}=0\} = P\{y_{i-1}=0, y_i=0, y_{i+1}=0\, y_{i+2}=0|\gamma_{i-1}=0, \gamma_i=0, \gamma_{i+1}=0, \gamma_{i+2}=1\} = \frac{\varepsilon^2}{4}$

$P\{y_{i-1}=0, y_i=0, y_{i+1}=0\, y_{i+2}=0|\gamma_{i-1}=0, \gamma_i=1, \gamma_{i+1}=0, \gamma_{i+2}=0\} = P\{y_{i-1}=0, y_i=0, y_{i+1}=0\, y_{i+2}=0|\gamma_{i-1}=0, \gamma_i=0, \gamma_{i+1}=1, \gamma_{i+2}=0\} = \frac{\varepsilon(\varepsilon^2+(1-\varepsilon)^2)}{4}$

$P\{y_{i-1}=0, y_i=0, y_{i+1}=0\, y_{i+2}=0|\gamma_{i-1}=1, \gamma_i=1, \gamma_{i+1}=0, \gamma_{i+2}=0\} = P\{y_{i-1}=0, y_i=0, y_{i+1}=0\, y_{i+2}=0|\gamma_{i-1}=0, \gamma_i=0, \gamma_{i+1}=1, \gamma_{i+2}=1\} = P\{y_{i-1}=0, y_i=0, y_{i+1}=0\, y_{i+2}=0|\gamma_{i-1}=1, \gamma_i=0, \gamma_{i+1}=0, \gamma_{i+2}=1\} = \frac{\varepsilon}{8}$

$P\{y_{i-1}=0, y_i=0, y_{i+1}=0\, y_{i+2}=0|\gamma_{i-1}=0, \gamma_i=1, \gamma_{i+1}=0, \gamma_{i+2}=1\} = P\{y_{i-1}=0, y_i=0, y_{i+1}=0\, y_{i+2}=0|\gamma_{i-1}=1, \gamma_i=0, \gamma_{i+1}=1, \gamma_{i+2}=0\} = \frac{\varepsilon^2+(1-\varepsilon)^2}{8}$

$P\{y_{i-1}=0, y_i=0, y_{i+1}=0\, y_{i+2}=0\} = P\{y_{i-1}=0, y_i=0, y_{i+1}=0\, y_{i+2}=0|\gamma_{i-1}=0, \gamma_i=1, \gamma_{i+1}=1, \gamma_{i+2}=0\} = \frac{\varepsilon^3+3\varepsilon(1-\varepsilon)^2}{8}$

$P\{y_{i-1}=0, y_i=0, y_{i+1}=0\, y_{i+2}=0|\gamma_{i-1}=1, \gamma_i=1, \gamma_{i+1}=1, \gamma_{i+2}=0\} = P\{y_{i-1}=0, y_i=0, y_{i+1}=0\, y_{i+2}=0|\gamma_{i-1}=0, \gamma_i=1, \gamma_{i+1}=1, \gamma_{i+2}=1\} = P\{y_{i-1}=0, y_i=0, y_{i+1}=0\, y_{i+2}=0|\gamma_{i-1}=1, \gamma_i=0, \gamma_{i+1}=1, \gamma_{i+2}=1\} = P\{y_{i-1}=0, y_i=0, y_{i+1}=0\, y_{i+2}=0|\gamma_{i-1}=1, \gamma_i=1, \gamma_{i+1}=0, \gamma_{i+2}=1\} = P\{y_{i-1}=0, y_i=0, y_{i+1}=0\, y_{i+2}=0|\gamma_{i-1}=1, \gamma_i=1, \gamma_{i+1}=1, \gamma_{i+2}=1\} = \frac{1}{16}$

Преобразовав получим:

$P\{y_{i-1}=0, y_i=0, y_{i+1}=0\, y_{i+2}=0\} = \frac{1}{16}\big(\delta^4(4\varepsilon^2-6\varepsilon+4) + \delta^3(-16\varepsilon^2+20\varepsilon-7) + \delta^2(8\varepsilon^3+20\varepsilon^2-22\varepsilon+4) + \delta(-16\varepsilon^3-8\varepsilon^2+8\varepsilon) + 8\varepsilon^3\big)$

□

$P\{y_{i-1}=1, y_i=1, y_{i+1}=1\, y_{i+2}=1\} = \frac{1}{16}\big(\delta^4(4\varepsilon^2-6\varepsilon+4) + \delta^3(-16\varepsilon^2+20\varepsilon-7) + \delta^2(8\varepsilon^3+20\varepsilon^2-22\varepsilon+4) + \delta(-16\varepsilon^3-8\varepsilon^2+8\varepsilon) + 8\varepsilon^3\big)$

Доказательство. $P\{y_{i-1}=1, y_i=1, y_{i+1}=1\, y_{i+2}=1|\gamma_{i-1}=0, \gamma_i=0, \gamma_{i+1}=0, \gamma_{i+2}=0\} = \frac{\varepsilon^3}{2}$

$P\{y_{i-1}=1,y_i=1,y_{i+1}=1\,y_{i+2}=1|\gamma_{i-1}=1,\gamma_i=0,\gamma_{i+1}=0,\gamma_{i+2}=0\}=P\{y_{i-1}=1,y_i=1,y_{i+1}=1\,y_{i+2}=1|\gamma_{i-1}=0,\gamma_i=0,\gamma_{i+1}=0,\gamma_{i+2}=1\}=\frac{\varepsilon^2}{4}$

$P\{y_{i-1}=1,y_i=1,y_{i+1}=1\,y_{i+2}=1|\gamma_{i-1}=0,\gamma_i=1,\gamma_{i+1}=0,\gamma_{i+2}=0\}=P\{y_{i-1}=1,y_i=1,y_{i+1}=1\,y_{i+2}=1|\gamma_{i-1}=0,\gamma_i=0,\gamma_{i+1}=1,\gamma_{i+2}=0\}=\frac{\varepsilon(\varepsilon^2+(1-\varepsilon)^2)}{4}$

$P\{y_{i-1}=1,y_i=1,y_{i+1}=1\,y_{i+2}=1|\gamma_{i-1}=1,\gamma_i=1,\gamma_{i+1}=0,\gamma_{i+2}=0\}=P\{y_{i-1}=1,y_i=1,y_{i+1}=1\,y_{i+2}=1|\gamma_{i-1}=0,\gamma_i=0,\gamma_{i+1}=1,\gamma_{i+2}=1\}=P\{y_{i-1}=1,y_i=1,y_{i+1}=1\,y_{i+2}=1|\gamma_{i-1}=1,\gamma_i=0,\gamma_{i+1}=0,\gamma_{i+2}=1\}=\frac{\varepsilon}{8}$

$P\{y_{i-1}=1,y_i=1,y_{i+1}=1\,y_{i+2}=1|\gamma_{i-1}=0,\gamma_i=1,\gamma_{i+1}=0,\gamma_{i+2}=1\}=P\{y_{i-1}=1,y_i=1,y_{i+1}=1\,y_{i+2}=1|\gamma_{i-1}=1,\gamma_i=0,\gamma_{i+1}=1,\gamma_{i+2}=0\}=\frac{\varepsilon^2+(1-\varepsilon)^2}{8}$

$P\{y_{i-1}=1,y_i=1,y_{i+1}=1\,y_{i+2}=1\}=P\{y_{i-1}=1,y_i=1,y_{i+1}=1\,y_{i+2}=1|\gamma_{i-1}=0,\gamma_i=1,\gamma_{i+1}=1,\gamma_{i+2}=0\}=\frac{\varepsilon^3+3\varepsilon(1-\varepsilon)^2}{8}$

$P\{y_{i-1}=1,y_i=1,y_{i+1}=1\,y_{i+2}=1|\gamma_{i-1}=1,\gamma_i=1,\gamma_{i+1}=1,\gamma_{i+2}=0\}=P\{y_{i-1}=1,y_i=1,y_{i+1}=1\,y_{i+2}=1|\gamma_{i-1}=0,\gamma_i=1,\gamma_{i+1}=1,\gamma_{i+2}=1\}=P\{y_{i-1}=1,y_i=1,y_{i+1}=1\,y_{i+2}=1|\gamma_{i-1}=1,\gamma_i=0,\gamma_{i+1}=1,\gamma_{i+2}=1\}=P\{y_{i-1}=1,y_i=1,y_{i+1}=1\,y_{i+2}=1|\gamma_{i-1}=1,\gamma_i=1,\gamma_{i+1}=0,\gamma_{i+2}=1\}=P\{y_{i-1}=1,y_i=1,y_{i+1}=1\,y_{i+2}=1|\gamma_{i-1}=1,\gamma_i=1,\gamma_{i+1}=1,\gamma_{i+2}=1\}=\frac{1}{16}$

Преобразовав получим:

$P\{y_{i-1}=1,y_i=1,y_{i+1}=1\,y_{i+2}=1\}=\frac{1}{16}\big(\delta^4(4\varepsilon^2-6\varepsilon+4)+\delta^3(-16\varepsilon^2+20\varepsilon-7)+\delta^2(8\varepsilon^3+20\varepsilon^2-22\varepsilon+4)+\delta(-16\varepsilon^3-8\varepsilon^2+8\varepsilon)+8\varepsilon^3\big)$

□

$P\{y_{i-1}=1,y_i=1,y_{i+1}=1\,y_{i+2}=1\}=P\{y_{i-1}=0,y_i=0,y_{i+1}=0,y_{i+2}=0\}\Rightarrow$

Вероятности симметричных шаблонов будут совпадать.

$P\{y_{i-1}=1,y_i=1,y_{i+1}=1,y_{i+2}=0\}=P\{y_{i-1}=0,y_i=0,y_{i+1}=0,y_{i+2}=1\}$
$P\{y_{i-1}=1,y_i=1,y_{i+1}=0,y_{i+2}=1\}=P\{y_{i-1}=0,y_i=0,y_{i+1}=1,y_{i+2}=0\}$
$P\{y_{i-1}=1,y_i=0,y_{i+1}=1,y_{i+2}=1\}=P\{y_{i-1}=0,y_i=1,y_{i+1}=1,y_{i+2}=0\}$
$P\{y_{i-1}=0,y_i=1,y_{i+1}=1,y_{i+2}=1\}=P\{y_{i-1}=1,y_i=0,y_{i+1}=0,y_{i+2}=0\}$
$P\{y_{i-1}=1,y_i=1,y_{i+1}=0,y_{i+2}=0\}=P\{y_{i-1}=0,y_i=0,y_{i+1}=1,y_{i+2}=1\}$
$P\{y_{i-1}=1,y_i=0,y_{i+1}=0,y_{i+2}=1\}=P\{y_{i-1}=0,y_i=1,y_{i+1}=1,y_{i+2}=0\}$
$P\{y_{i-1}=1,y_i=0,y_{i+1}=1,y_{i+2}=0\}=P\{y_{i-1}=0,y_i=1,y_{i+1}=0,y_{i+2}=1\}$

$P\{y_{i-1}=1,y_i=1,y_{i+1}=1,y_{i+2}=0\}=\frac{1}{16}\Big(\delta^4(-4\varepsilon^2+4\varepsilon+2)+\delta^3(16\varepsilon^2-16\varepsilon+1)+\delta^2(-8\varepsilon^3-12\varepsilon^2+20\varepsilon-6)+\delta(16\varepsilon^3-8\varepsilon^2-8\varepsilon+4)-8\varepsilon^3+8\varepsilon^2\Big)$

$P\{y_{i-1}=1,y_i=1,y_{i+1}=0,y_{i+2}=1\}=\frac{1}{16}\Big(\delta^4(-2\varepsilon^2+2\varepsilon+2)+\delta^3(12\varepsilon^2-12\varepsilon+1)+\delta^2(8\varepsilon^3-34\varepsilon^2+26\varepsilon-6)+\delta(-16\varepsilon^3+40\varepsilon^2-24\varepsilon+4)-8\varepsilon^3-16\varepsilon^2+8\varepsilon\Big)$

$P\{y_{i-1}=1,y_i=0,y_{i+1}=1,y_{i+2}=1\}=\frac{1}{16}\Big(\delta^4(4\varepsilon^3-12\varepsilon^2+8\varepsilon+2)+\delta^3(-12\varepsilon^3+40\varepsilon^2-$

$28\varepsilon + 1) + \delta^2(20\varepsilon^3 - 60\varepsilon^2 + 40\varepsilon - 6) + \delta(-20\varepsilon^3 + 48\varepsilon^2 - 28\varepsilon + 4) + 8\varepsilon^3 - 16\varepsilon^2 + 8\varepsilon\Big)$

$P\{y_{i-1} = 0, y_i = 1, y_{i+1} = 1, y_{i+2} = 1\} = \frac{1}{16}\Big(\delta^4(-4\varepsilon^2 + 4\varepsilon + 2) + \delta^3(16\varepsilon^2 - 16\varepsilon + 1) +$
$\delta^2(-8\varepsilon^3 - 12\varepsilon^2 + 20\varepsilon - 6) + \delta(16\varepsilon^3 - 8\varepsilon^2 - 8\varepsilon + 4) - 8\varepsilon^3 + 8\varepsilon^2\Big)$

$P\{y_{i-1} = 1, y_i = 1, y_{i+1} = 0, y_{i+2} = 0\} = \frac{1}{16}\Big(\delta^4(4\varepsilon^2 - 4\varepsilon + 4) + \delta^3(-16\varepsilon^2 + 16\varepsilon - 7) +$
$\delta^2(-8\varepsilon^3 + 28\varepsilon^2 - 20\varepsilon + 4) + \delta(16\varepsilon^3 - 24\varepsilon^2 + 8\varepsilon) - 8\varepsilon^3 + 8\varepsilon^2\Big)$

$P\{y_{i-1} = 0, y_i = 1, y_{i+1} = 1, y_{i+2} = 0\} = \frac{1}{16}\Big(\delta^4(4\varepsilon^2 - 4\varepsilon + 4) + \delta^3(-16\varepsilon^2 + 16\varepsilon - 7) +$
$\delta^2(8\varepsilon^3 + 4\varepsilon^2 - 12\varepsilon + 4) + \delta(-16\varepsilon^3 + 24\varepsilon^2 - 8\varepsilon) + 8\varepsilon^3 - 16\varepsilon^2 + 8\varepsilon\Big)$

$P\{y_{i-1} = 1, y_i = 0, y_{i+1} = 1, y_{i+2} = 0\} = \frac{1}{16}\Big(\delta^4(4\varepsilon^2 - 4\varepsilon + 4) + \delta^3(= 16\varepsilon^2 + 16\varepsilon - 7) +$
$\delta^2(-8\varepsilon^3 + 44\varepsilon^2 - 44\varepsilon + 12) + \delta(16\varepsilon^3 - 56\varepsilon^2 + 56\varepsilon - 16) - 8\varepsilon^3 + 24\varepsilon^2 - 24\varepsilon + 8\Big)$

Если рассматривать ассимптотическое представление первого порядка вероятности появления шаблонов, достаточно рассмотреть шаблоны '0000', '1000', '0100', '0010', '0001'. Тогда:

$P_4\{y\} = (1 - 4\delta)P\{y|\gamma_{i-1} = 0, \gamma_i = 0, \gamma_{i+1} = 0, \gamma_{i+2} = 0\} + \delta\Big(P\{y|\gamma_{i-1} = 1, \gamma_i = 0, \gamma_{i+1} =$
$0, \gamma_{i+2} = 0\} + P\{y|\gamma_{i-1} = 0, \gamma_i = 1, \gamma_{i+1} = 0, \gamma_{i+2} = 0\} + P\{y|\gamma_{i-1} = 0, \gamma_i = 0, \gamma_{i+1} = 1, \gamma_{i+2} =$
$0\} + P\{y|\gamma_{i-1} = 0, \gamma_i = 0, \gamma_{i+1} = 0, \gamma_{i+2} = 1\}\Big)$

Получив вероятности можно определить ассимптотическое выражение для энтропии 4-грамма:

Лемма 5.4. $H_4(\delta) = -2\Big(-\frac{1}{2} + \frac{3}{2}(1 - \varepsilon)\log(1 - \varepsilon) + \frac{3}{2}\varepsilon\log(\varepsilon) +$
$\delta\Big(\frac{1}{4\ln(2)} + 2(1 - 2\varepsilon)\log(\varepsilon) + (3\varepsilon - \frac{5}{4})\log(1 - \varepsilon) - \frac{1}{4}\Big)\Big) + O(\delta^2)$

Доказательство. Данное выражение получается при подстановке полученных выше вероятностей в формулу (19) и разложением логарифма в ряд Тейлора. □

6 Компьютерные эксперименты

Экспериментально построим $H_2^{(\delta)}$ и $H_2^{(0)}$

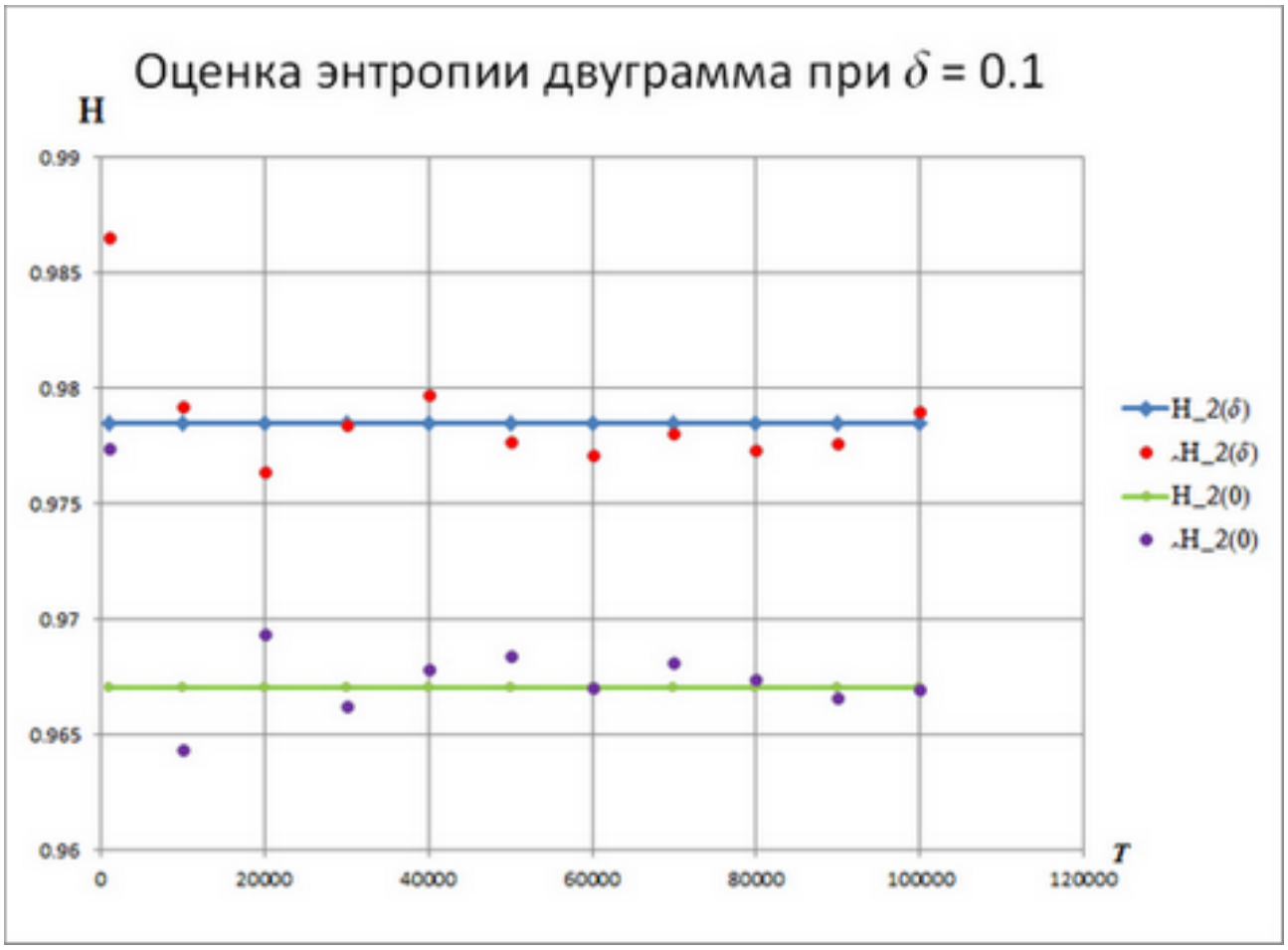

Рис. 1: График зависимости энтропии $H_2^{(\delta)}$ от длины последовательности при $\delta = 0.1$

Рис. 2: График зависимости энтропии $H_2^{(\delta)}$ от длины последовательности при $\delta = 0.3$

Рис. 3: График зависимости энтропии $H_2^{(\delta)}$ от длины последовательности при $\delta = 0.5$

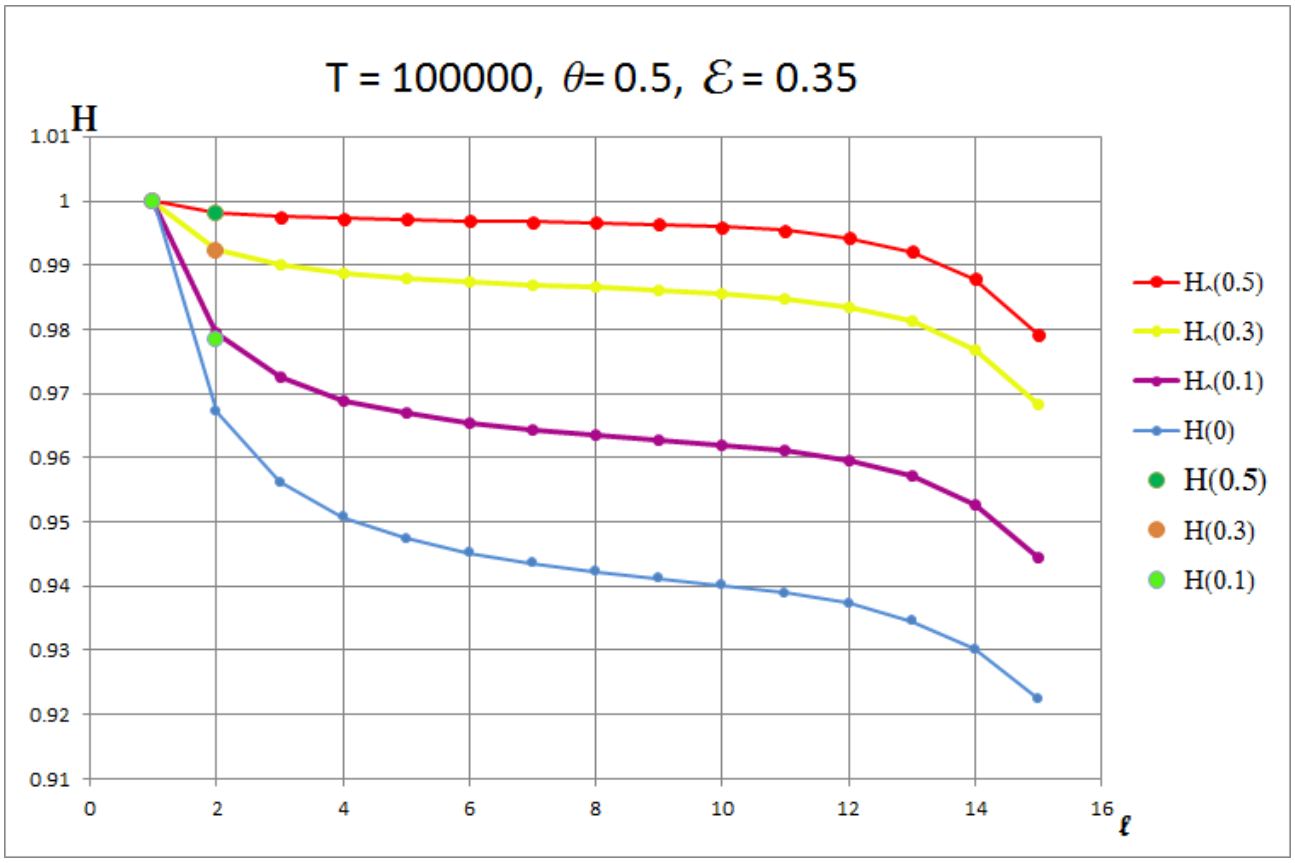

Рис. 4: Семейство графиков зависимости энтропии $H_l^{(\delta)}$ от L при различных δ

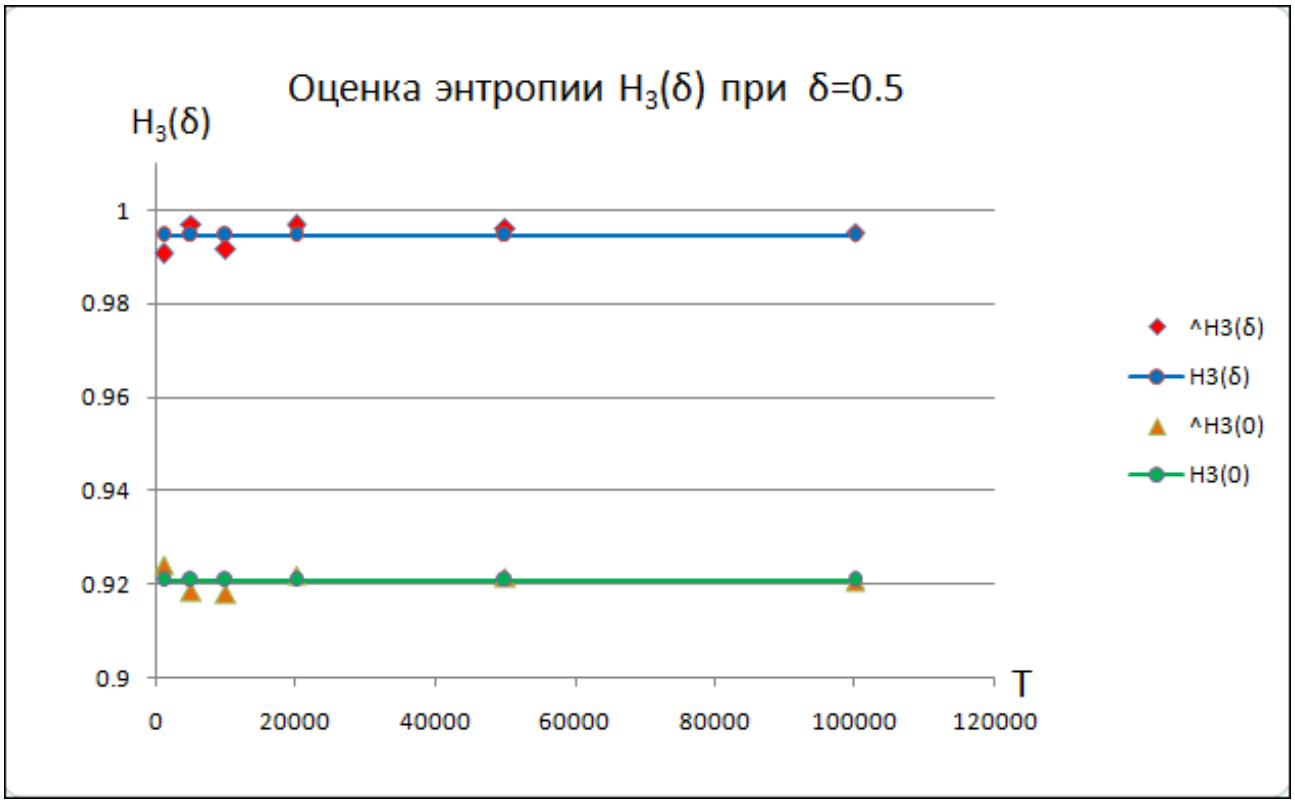

Рис. 5: График зависимости энтропии $H_3^{(\delta)}$ от длины последовательности при $\delta = 0.5$

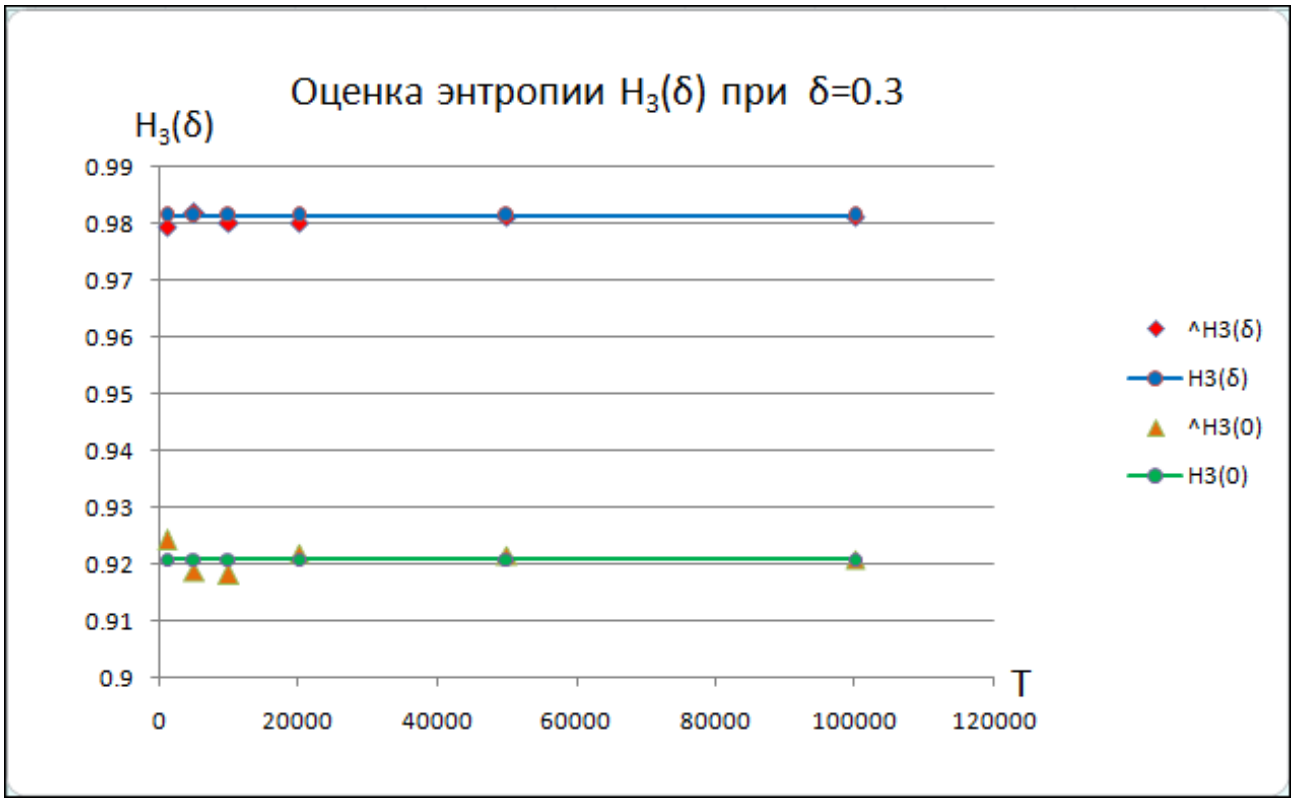

Рис. 6: График зависимости энтропии $H_3^{(\delta)}$ от длины последовательности при $\delta = 0.3$

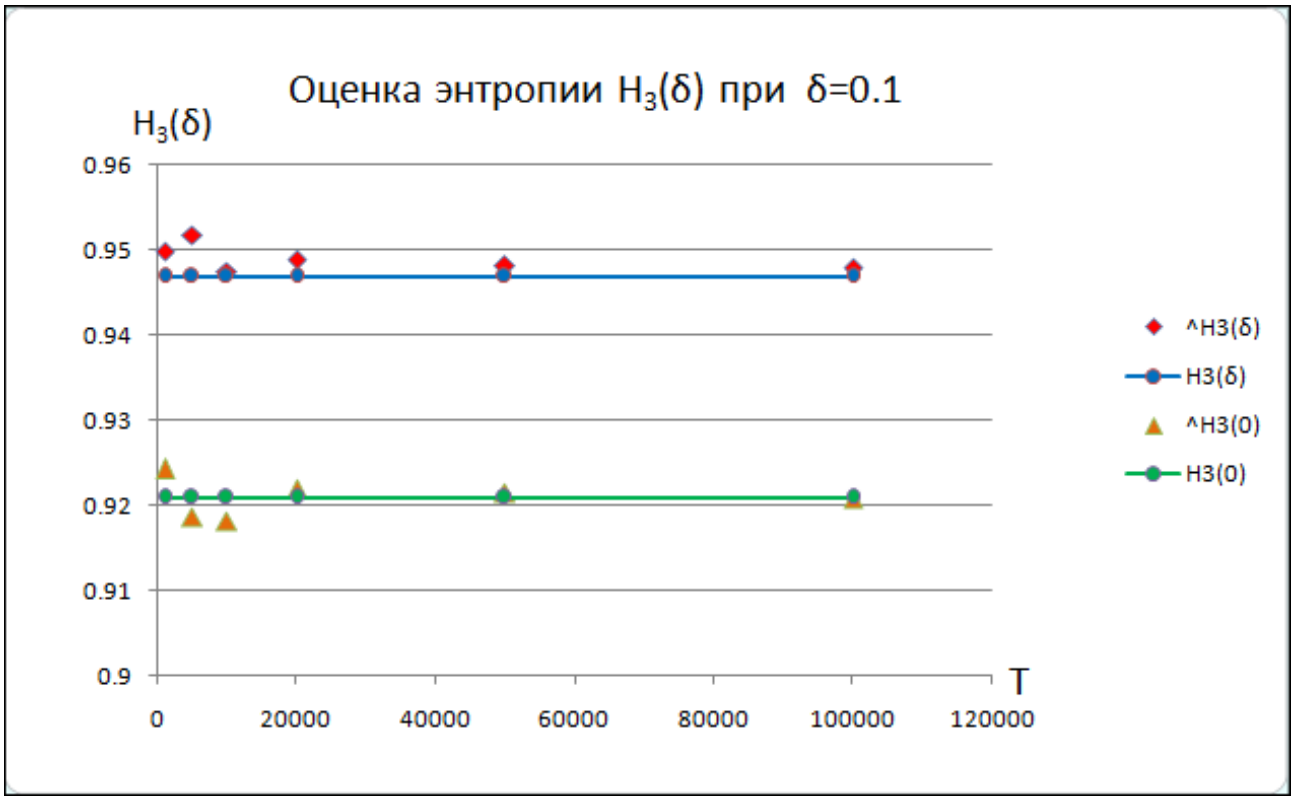

Рис. 7: График зависимости энтропии $H_3^{(\delta)}$ от длины последовательности при $\delta = 0.1$

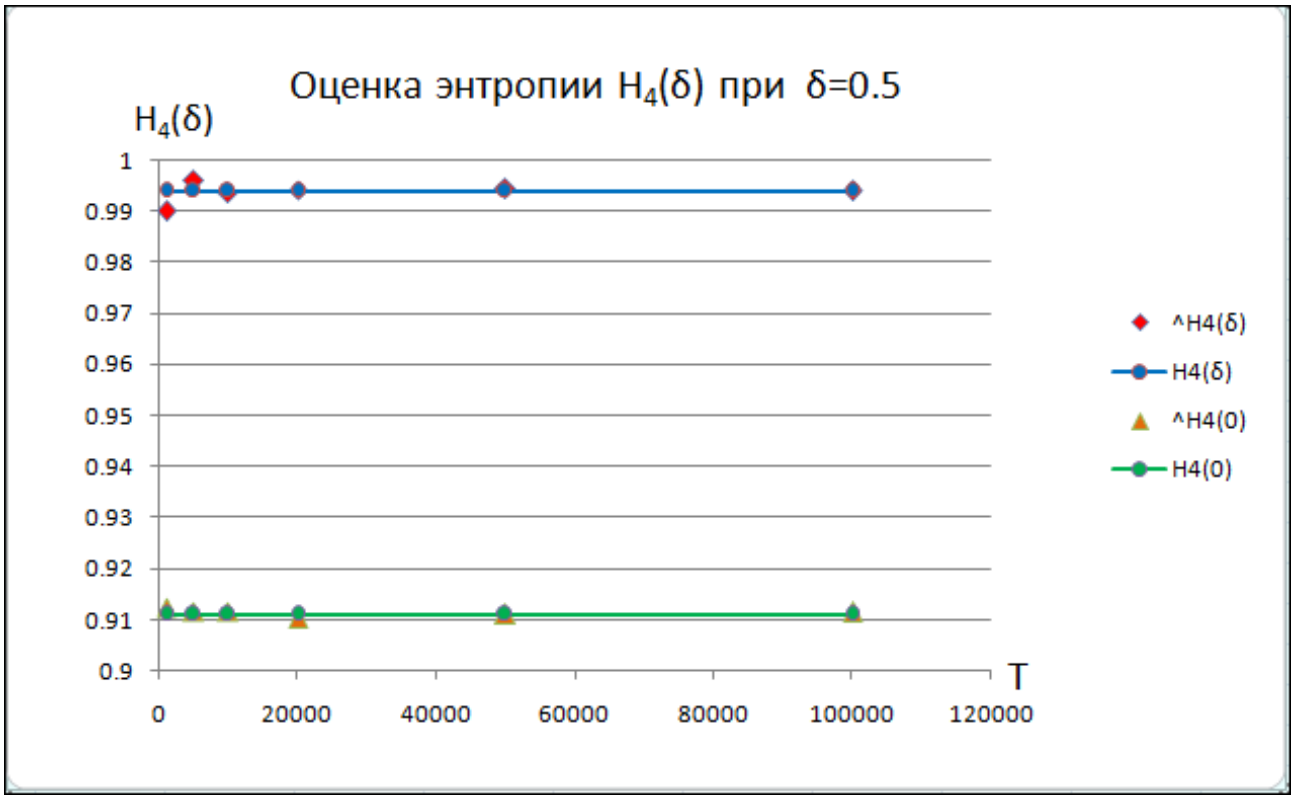

Рис. 8: График зависимости энтропии $H_4^{(\delta)}$ от длины последовательности при $\delta = 0.5$

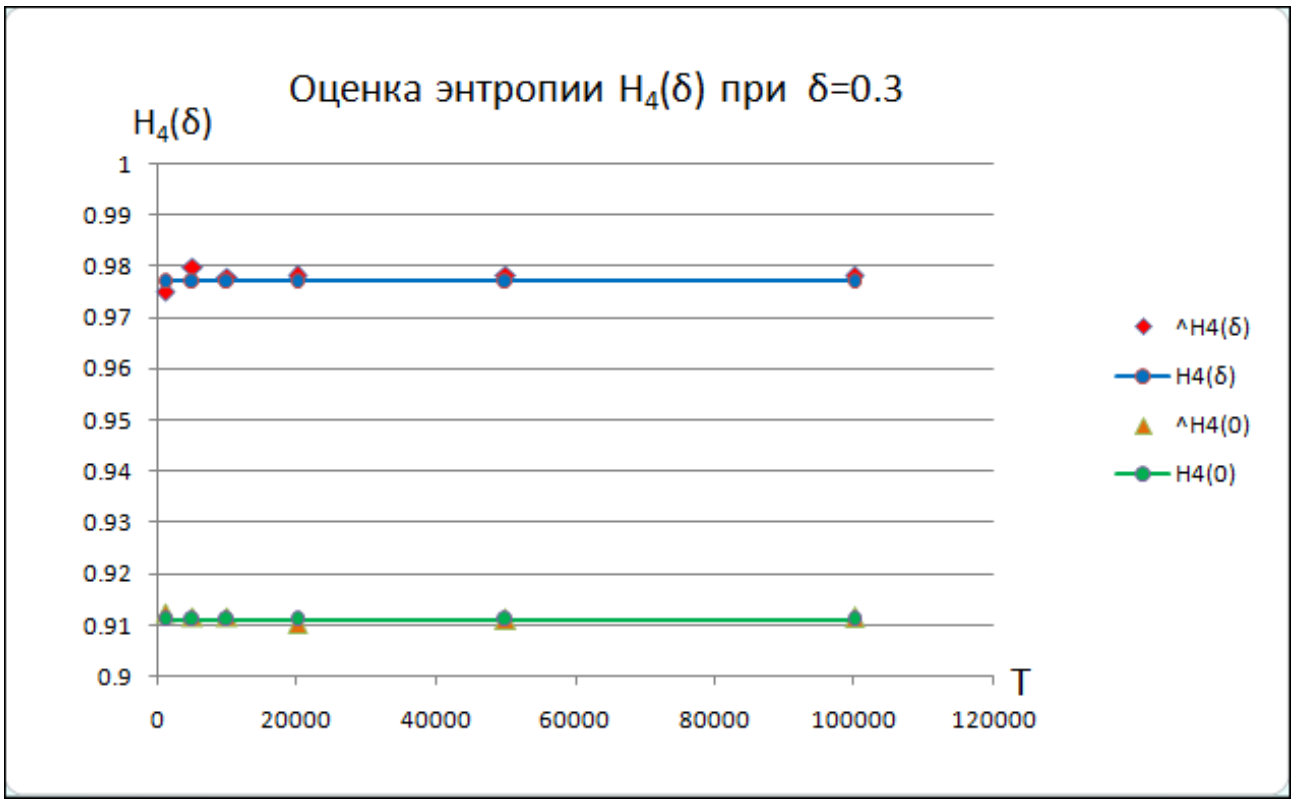

Рис. 9: График зависимости энтропии $H_4^{(\delta)}$ от длины последовательности при $\delta = 0.3$

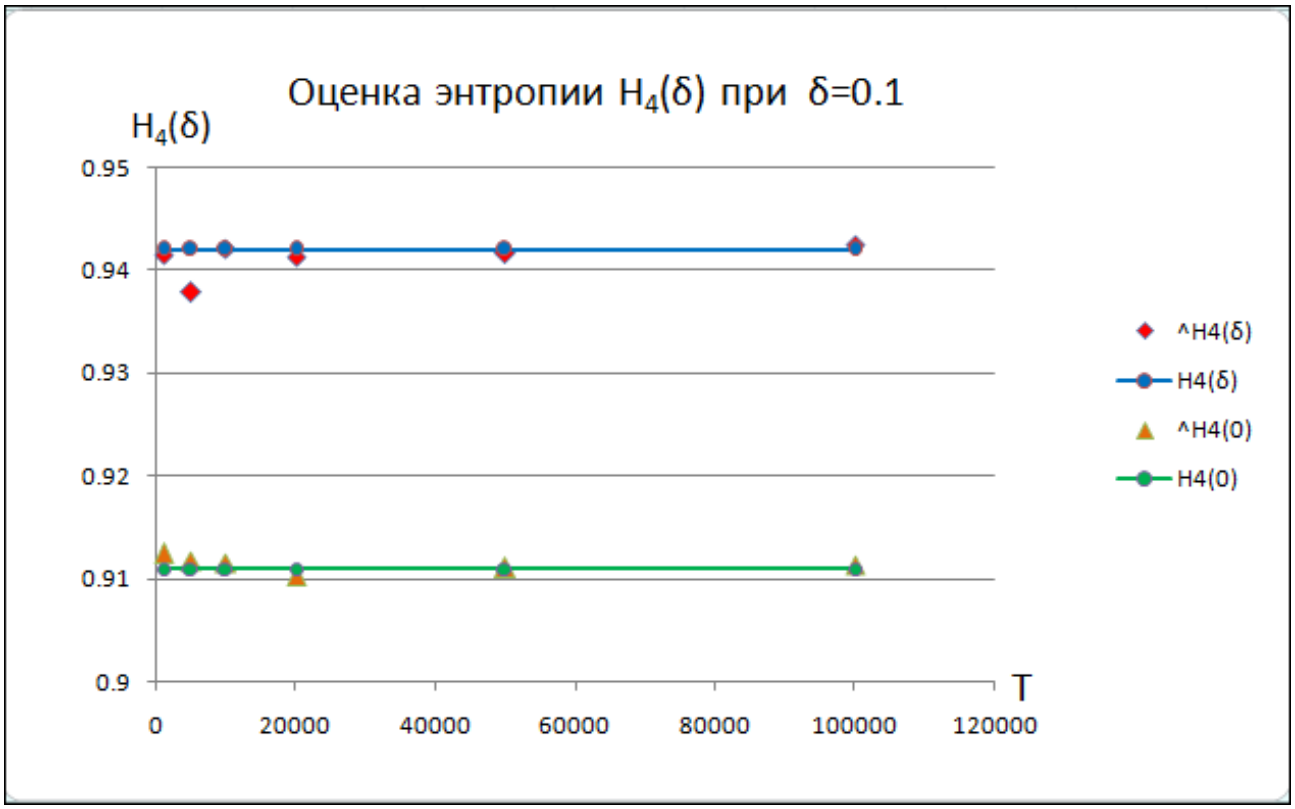

Рис. 10: График зависимости энтропии $H_4^{(\delta)}$ от длины последовательности при $\delta = 0.1$

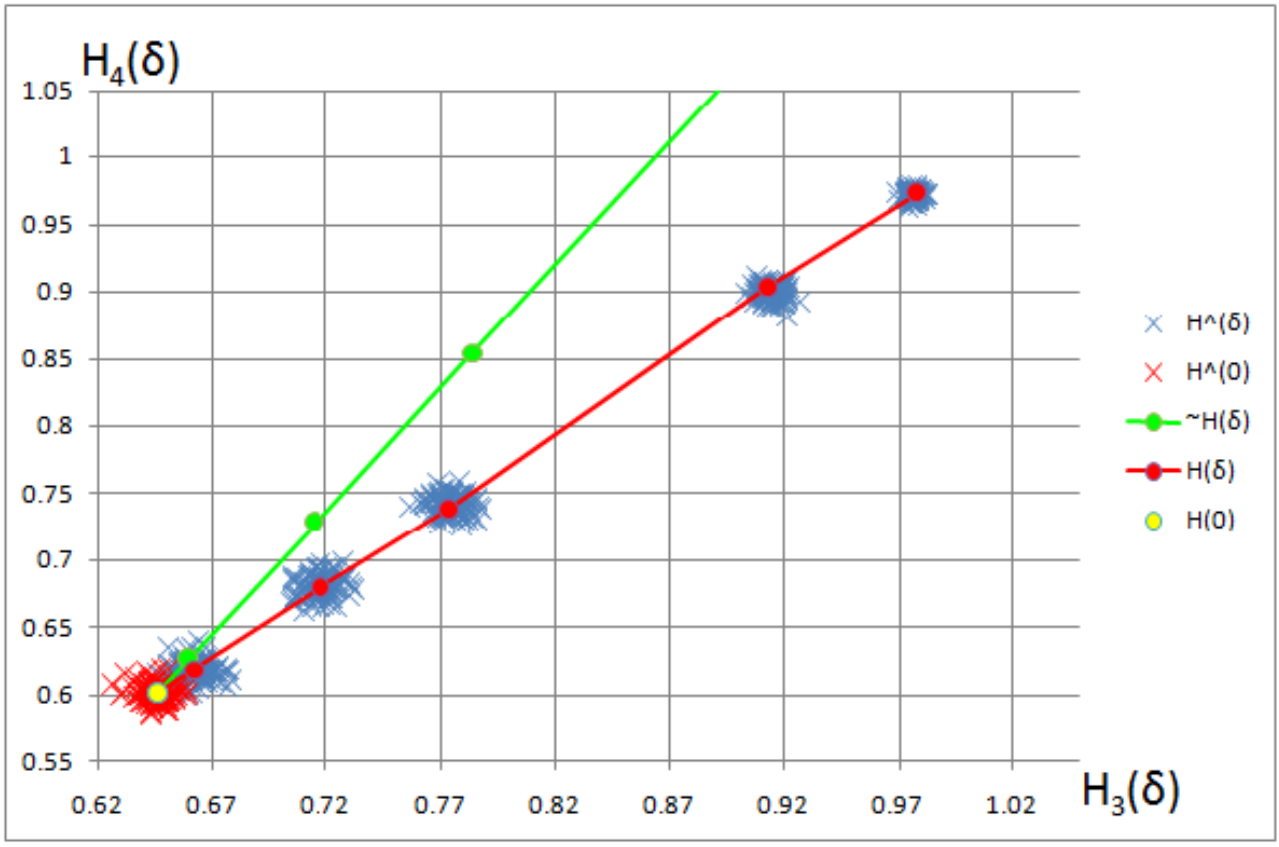

Рис. 11: График зависимости $H_4^{(\delta)}$ от $H_3^{(\delta)}$

Отметим, точки ассимптотических оценок $H_4^{(\delta)}$ и $H_3^{(\delta)}$ находятся на прямой:

$$\begin{cases} H_3(\delta) = -2\Big(\varepsilon\log(\varepsilon) + (1-\varepsilon)\log(1-\varepsilon) - \frac{1}{2} + \delta\big((1-2\varepsilon)\log(\varepsilon) - (1-2\varepsilon)\log(1-\varepsilon) + \\ \frac{(1-2\varepsilon)^2}{2\ln(2)}\big)\Big) \\ H_4(\delta) = -2\Big(-\frac{1}{2} + \frac{3}{2}(1-\varepsilon)\log(1-\varepsilon) + \frac{3}{2}\varepsilon\log(\varepsilon) + \delta\Big(\frac{1}{4\ln(2)} + 2(1-2\varepsilon)\log(\varepsilon) + \\ (3\varepsilon - \frac{5}{4})\log(1-\varepsilon) - \frac{1}{4}\Big)\Big) \end{cases}$$

7 Линейный дискриминантный анализ

Линейный дискриминантный анализ (ЛДА), а также связанный с ним линейный дискриминант Фишера — методы статистики и машинного обучения, применяемые для нахождения линейных комбинаций признаков, наилучшим образом разделяющих два или более класса объектов или событий. Полученная комбинация может быть использована в качестве линейного классификатора или для сокращения размерности пространства признаков перед последующей классификацией. ЛДА тесно связан с дисперсионным анализом и регрессионным анализом, также пытающимися выразить какую-либо зависимую переменную через линейную комбинацию других признаков или измерений. В этих двух методах зависимая переменная — численная величина, а в ЛДА она является величиной номинальной (меткой класса). Помимо того, ЛДА имеет схожие черты с методом главных компонент и факторным анализом, которые ищут линейные комбинации величин, наилучшим образом описывающие данные. Для использования ЛДА признаки должны быть непрерывными величинами, иначе следует использовать анализ соответствий (англ. Discriminant Correspondece Analysis).

7.1 Линейный дискриминантный анализ для случая двух классов

Для каждого образца объекта или события с известным классом y рассматривается набор наблюдений x (называемых ещё признаками, переменными или измерениями). Набор таких образцов называется обучающей выборкой (или набором обучения, обучением). Задачи классификации состоит в том, чтобы построить хороший прогноз класса y для всякого так же распределённого объекта (не обязательно содержащегося в обучающей выборке), имея только наблюдения x.
При ЛДА предполагается, что функции совместной плотности распределения вероятностей $p(\vec{x}|y=1)$ и $p(\vec{x}|y=0)$ - нормальны. В этих предположениях оптимальное байесовское решение - относить точки ко второму классу если отношение правдоподобия ниже некоторого порогового значения T:

$$(\vec{x}-\vec{\mu}_0)^T\Sigma_{y=0}^{-1}(\vec{x}-\vec{\mu}_0)+\ln|\Sigma_{y=0}|-(\vec{x}-\vec{\mu}_1)^T\Sigma_{y=1}^{-1}(\vec{x}-\vec{\mu}_1)-\ln|\Sigma_{y=0}|<T$$

Если не делается никаких дальнейших предположений, полученную задачу классификации называют квадратичным дискриминантным анализом (англ. quadratic discriminant analysis, QDA). В ЛДА делается дополнительное предположение о гомоскедастичности (т.е. предполагается, что ковариационные матрицы равны, $\Sigma_{y=0}=\Sigma_{y=1}=\Sigma$) и считается, что ковариационные матрицы имеют полный ранг. При этих предположениях задача упрощается и сводится к сравнению скалярного произведения с пороговым значением

$$\vec{\omega}\cdot\vec{x}<c$$

для некоторой константы c, где

$$\vec{\omega}=\Sigma^{-1}(\vec{\mu_1}-\vec{\mu_0}).$$

Это означает, что вероятность принадлежности нового наблюдения x к классу y зависит исключительно от линейной комбинации известных наблюдений.

7.2 Результаты линейного дискриминантного анализа

Линейный дискриминантный анализ применен для классификации последовательностей с вкраплениями и без вкраплений при фиксированном параметре ε.
Пусть имеется последовательность $Y=\{y_1,...,y_T\}$, на основании Y вычисляем $(H_3(\delta),H_4(\delta))$

при фиксированном $\varepsilon = 0.55$, тогда:
H_0: последовательность Y имеет вкрапления
H_1: последовательность Y не имеет вкраплений
тогда для $n = n_0 + n_1$, (где n_0 - количество заведомо пустых последовательностей, n_1 - количество последовательностей с вкрапениями) последовательностей можно провести дискриминантный анализ и оценить вероятновсть правильной классификации и мощность критерия.
Тогда вероятности ошибок первого и второго рода:

$$\alpha = \frac{n_0 - \nu_0}{n_0} - ; \tag{23}$$

$$\beta = \frac{n_1 - \nu_1}{n_1} - ; \tag{24}$$

ν_0 - количество верно отпределенных пустых последоваательностей, ν_1 - количество верно отпределенных последоваательностей с вкраплениями.
Мощность критерия:

$$w = \frac{\nu_1}{n_1} \tag{25}$$

δ	α	β	Мощность критерия
0.07	21%	14%	0.86
0.08	10%	13%	0.87
0.09	14%	8%	0.92
0.1	13%	5%	0.95

Таблица 1: Результаты дискриминантного анализа.

8 Вывод

Методом линейного дискриминантного анализа на основании энтропийных характеристик $(H_3(\delta), H_4(\delta))$ можно определить наличие вкраплений в последовательность с вероятностью ошибки второго рода 5% при доле вкраплений $\delta \geq 0.1$.

Список литературы

[1] А. А. Духин: Теория информации - М.: "Гелиос АРВ 2007.

[2] А.В. Аграновский, А. В. Балакин: Стеганография, цифровые водяные знаки о стегоанализ - М.: Вузовская книга, 2009.

[3] В. Г. Грибунин, И. Н. Оков, И. В. Туринцев: Цифровая стеганография - М.: Солон-Прессб, 2002.

[4] Н. П. Варновский, Е. А. Голубев, О. А. Логачев: Современные направления стеганографии. Математика и безопасность информационных технологий. Материалы конференции в МГУ 28-29 октября 2004 г., МЦМНО, М., 2005, с. 32-64.

[5] Ю. С. Харин [и др.]: Криптология - Минск: БГУ, 2013.

[6] Ю. С. Харин, Е. В. Вечерко "Статистическое оценивание параметров модели вкраплений в двоичную цепь Маркова Дискрет. матем., 25:2 (2013), 135-148.