

МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ
Кафедра математического моделирования и анализа данных

ШИМКО Андрей Чеславович

ОБНАРУЖЕНИЕ ВКРАПЛЕНИЙ В ДВОИЧНУЮ ЦЕПЬ
МАРКОВА НА ОСНОВЕ ЭНТРОПИЙНЫХ ХАРАКТЕРИСТИК

Дипломная работа

Научный руководитель Егор Валентинович
Вечерко
кандидат физико-математических наук

Допущена к защите
«___»_____ 2017 г.
Зав. кафедрой математического
моделирования и анализа данных
_____ Ю.С. Харин
доктор физико-математических наук,
профессор,
член-корреспондент НАН Беларуси

Минск, 2017

ОГЛАВЛЕНИЕ

ПЕРЕЧЕНЬ УСЛОВНЫХ ОБОЗНАЧЕНИЙ	3
ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ	4
ВВЕДЕНИЕ	7
1 Теоретический подход	8
1.1 Основные понятия	8
1.2 Теоретико-информационный подход к оценке стойкости систем .	10
2 Математическая модель	12
2.1 Математическая модель вкраплений на основе схемы независи- мых испытаний	12
2.2 Математическая модель вкраплений в двоичную стационарную марковскую последовательность 1-го порядка и ее свойства . . .	14
2.3 Линейный дискриминантный анализ	23
2.3.1 Линейный дискриминантный анализ для случая двух классов	23
2.3.2 Результаты линейного дискриминантного анализа	24
2.3.3 Вывод	26
2.4 Исследование реальных данных на основании изображений в формате JPEG	27
3 Компьютерные эксперименты	30
ЗАКЛЮЧЕНИЕ	37
СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ	37
ПРИЛОЖЕНИЕ А	39

ПЕРЕЧЕНЬ УСЛОВНЫХ ОБОЗНАЧЕНИЙ

ЦМ	–	цепь Маркова
ЛДА	–	линейный дискриминантный анализ
ДКП	–	дискретное косинусное преобразование

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Дипломная работа: 49 страниц, 16 иллюстраций, 2 таблицы, 10 источников, 1 приложение.

ВКРАПЛЕНИЕ, ЦЕПЬ МАРКОВА, СТЕГАНОГРАФИЯ, ЭНТРОПИЯ, ЛИНЕЙНЫЙ ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Объект работы - двоичные цепи Маркова 1-го порядка.

Цель работы - исследование методов анализа наличия вкраплений в двоичную цепь Маркова 1-го порядка.

В результате получены асимптотические оценки энтропии l -граммы при $l=2,3,4$. На основании полученных оценок установлен критерий обнаружения наличия вкраплений в двоичную цепь Маркова 1-го порядка.

Область применения: Результаты работы могут быть использованы для обнаружения наличия вкраплений в двоичную цепь Маркова, а так же для оценивания стойкости стеганографических систем.

АГУЛЬНАЯ ХАРАКТЕРЫСТЫКА ПРАЦЫ

Дыпломная праца: 49 лістоў, 16 малюнкаў, 2 табліцы, 10 крыніц, 1 прыкладанне.

Ключавыя словы: УКРАПАВАННЕ, ЛАНЦУГ МАРКАВА, СТЭГНАГРАФІЯ, ЭНТРАПІЯ, ЛІНЕЙНЫ ДЫСКРЫМІНАНТНЫ АНАЛІЗ.

Аб'ект даследавання - двайковыя ланцугі Маркава 1-га парадку.

Мэта працы - даследванне метадаў аналізу наяўнасці ўкрапаванняў ў двайковы ланцуг Маркава 1-га парадку.

У выніку атрыманы асімптатычныя ацэнкі энтрапіі l -граммы пры $l=2,3,4$. Па палучаных ацэнках усталяваны крытэрыі выяўлення наяўнасці ўкрапаванняў ў двайковы ланцуг Маркава 1-га парадку.

Вобласць прымянення: Вынікі працы могуць выкарыстоўвацца для выяўлення наяўнасці ўкрапаванняў ў двайковы ланцуг Маркава, а таксама для ацэньвання стойкасці стэганаграфічных сістэм.

SUMMARY

Graduation assignment: 49 sheets, 16 images, 2 tables, 10 sources, 1 appendices.

EMBEDDING, MARKOV CHAIN, STEGANOGRAPHY, ENTROPY,
LINEAR DISCRIMINANT ANALYSIS

Object of study is binary Markov chain with first order.

The goal of the thesis is researching of analysis methods identification of embeddings to binary Markov chain with first order.

As a result, the thesis shows asymptotic expansion of entropy l -grams where $l=2,3,4$. On the basis of the estimates obtained, the criterion for detecting the presence of inclusions in the Markov binary chain of the first order.

Application area: The results can be used for identification of embeddings to binary Markov chain and evaluation security of steganography systems.

ВВЕДЕНИЕ

Стеганография представляет собой специфическую область человеческой деятельности, связанной с разработкой и анализом методов сокрытия факта передачи информации [7]. Подобно криптографии, стеганография известна со времен античности. Но на этом аналогии, по крайней мере в контексте теоретических исследований, заканчиваются. За последние четверть века возникла и успешно развивается новая математическая дисциплина криптология, или, что то же самое, математическая криптография, изучающая математические модели криптографических схем. Попытки создания математической стеганографии (которую, быть может, следует именовать также стеганологией) предпринимаются, 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.

Актуальной является проблема посторения и анализа математических моделей вкраплений. Стеганографический термин вкрапление означает замену некоторых символов передаваемой по каналу связи открытой информации альтернативными символами, несущими маскируемую конфиденциальную информацию. Основной проблемой стеганографии является такая организация процесса вкрапления, при которой обеспечивается гарантированный уровень надежности сокрытия от противника самого факта вкрапления. С математической точки зрения здесь речь идет о различении по наблюдаемому сообщению гипотез H_0 и H_1 об отсутствии и наличии вкраплений соответственно. При этом для передающей стороны наибольший интерес представляет определение условий, при которых эти гипотезы статистически неразличимы. Решение этих задач существенно зависит от соответствующих математических моделей вкрапления [6]. В настоящей работе исследуются вероятностные свойства модели вкраплений в двоичную цепь Маркова 1-го порядка.

ГЛАВА 1

Теоретический подход

1.1 Основные понятия

Определение 1.1.1. 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, [3]как:

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

Определение 1.1.2. Контейнером $b \in B$ (носителем) называют несекретные данные, которые используют для сокрытия сообщения [3].

Определение 1.1.3. Сообщение $m \in M$ - секретная информация, наличие которой в контейнере необходимо скрыть [3].

Определение 1.1.4. Ключ $k \in K$ - секретная информация, известная только законному пользователю, которая определяет конкретный вид алгоритма сокрытия [3].

Определение 1.1.5. Пустой контейнер - контейнер, не содержащий сообщения [3].

Определение 1.1.6. Заполненный контейнер - контейнер, с внедренным в него сообщением [3].

Определение 1.1.7. Стеганографический алгоритм - два преобразования: прямое стеганографическое преобразование $F : M \times B \times K \rightarrow B$ и обратное стеганографическое преобразование $F^{-1} : B \times K \rightarrow M$, сопоставляющее соответственно тройке (сообщение, пустой контейнер, ключ) контейнер-результат и паре (заполненный контейнер, ключ) - исходное сообщение, [3]причем:

$$F(m, b, k) = b_{m,k}, F^{-1}(b_{m,k}, k) = m, m \in M, b_{m,k}, b \in B, k \in K \quad (1.1)$$

Определение 1.1.8. Под стеганографической системой будем понимать $S = (F, F^{-1}, M, B, K)$, представляющую собой совокупность сообщений, секретных ключей, контейнеров и связывающих их преобразований [3].

Определение 1.1.9. Внедрение (сокрытие) - применение прямого стеганографического преобразования к конкретным контейнеру, ключу и сообщению [3].

Определение 1.1.10. Извлечение сообщения - применение обратного стеганографического преобразования [3].

Определение 1.1.11. l-грамм - подпоследовательность из l подряд идущих элементов последовательности.

Определение 1.1.12. Под дискретным источником сообщений будем понимать устройство, порождающее последовательности, составленные из букв конечного алфавита A ($|A|=n<\infty$). При этом буквы последовательностей порождаются в дискретный момент времени: $t = 0, 1, 2, \dots; t = \dots, -2, -1, 0, 1, 2, \dots$; [2]

Определение 1.1.13. Если вероятность того, что источник порождает некоторую последовательность $a_{i_1} \dots a_{i_l}$, 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; $a_{j_1} \dots a_{j_l}$, то источник называется стационарным. [2] Стационарность означает неизменность во времени всех конечномерных распределений соответствующего случайного процесса.

Определение 1.1.14. Энтропией источника назовем величину:

$$H_{\infty} = \lim_{l \rightarrow \infty} \frac{H(C_l)}{l}; \quad (1.2)$$

если данный предел существует [2].

1.2 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 [7].

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 поражена согласно P_{Q_0} , если же верна гипотеза H_1 , тогда Q была поражена 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:

$$H(P_{Q_0}||P_{Q_1}) = \sum_q P_{Q_0}(q) \log \frac{P_{Q_0}(q)}{P_{Q_1}(q)}; \quad (1.3)$$

Относительная энтропия между двумя распределениями всегда неотрицательна и равна нулю только тогда, когда распределения равномерны. Несмотря на то, что относительная 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:

$$d(\alpha, \beta) = \alpha \log \frac{\alpha}{1 - \beta} + (1 - \alpha) \log \frac{1 - \alpha}{\beta}; \quad (1.4)$$

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

ГЛАВА 2

Математическая модель

2.1 Математическая модель вкраплений на основе схемы независимых испытаний

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

$$\mathcal{L}x_t = Bi(1, p), x_i \in V = 0, 1, i = \overline{1, T}; \quad (2.1)$$

Вкрапляемое сообщение имеет вид:

$$\mathcal{L}m_t = Bi(1, \theta), m_i \in V = 0, 1, i = \overline{1, \tau}; \quad (2.2)$$

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

$$\mathcal{L}\gamma_t = Bi(1, \delta), \gamma_i \in V = 0, 1, i = \overline{1, T}; \quad (2.3)$$

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

$$y_t = (1 - \gamma_t)x_t + \gamma_tm_{\tau_t}; \quad (2.4)$$

Лемма 2.1.1. Для модели (2.1)-(2.4)

$$P\{y_t = 1\} = (1 - \delta)p + \delta\theta; \quad (2.5)$$

$$P\{y_t = 0\} = (1 - \delta)(1 - p) + \delta(1 - \theta); \quad (2.6)$$

Доказательство. Воспользуемся формулой полной вероятности:

$$\begin{aligned} P\{y_t = 1\} &= P\{(1 - \gamma_t)x_t + \gamma_tm_{\tau_t} = 1\} = \sum_{j \in V} P\{y_t = 1, \gamma_t = j\} = \\ &= \sum_{j \in V} P\{\gamma_t = j\}P\{y_t = 1 | \gamma_t = j\} = (1 - \delta)P\{x=1, \gamma_t = 0\} + \delta P\{m_{\tau_t} = 1, \gamma_t = 1\} = (1 - \delta)p + \delta\theta; \end{aligned}$$

Тогда:

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

$$h = \frac{H(y_1, \dots, y_t)}{T} = \frac{TH(y_1)}{T} = H(y_1); \quad (2.7)$$

Воспользуемся леммой 2.1.1:

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

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

Рассмотрим модель (2.1)-(2.4).

Пусть контейнер (2.1) представляет собой цепь Маркова 1-го порядка с вектором распределения вероятностей $\pi = (\frac{1}{2}, \frac{1}{2})$ и матрицей вероятностей одношаговых переходов

$$P(\varepsilon) = \frac{1}{2} \begin{pmatrix} 1 + \varepsilon & 1 - \varepsilon \\ 1 - \varepsilon & 1 + \varepsilon \end{pmatrix}, |\varepsilon| < 1, \varepsilon \neq 0. \quad (2.8)$$

Лемма 2.2.1. Для модели (2.1)-(2.4) с условием (2.8):

$$P\{y_{t-1} = 1, y_t = 1\} = \frac{1}{4}(1 + \varepsilon)(1 - \delta)^2 + \theta\delta(1 - \delta) + \theta^2\delta^2; \quad (2.9)$$

$$P\{y_{t-1} = 1, y_t = 0\} = \frac{1}{4}(1 - \varepsilon)(1 - \delta)^2 + \frac{1}{2}\delta(1 - \delta) + \theta(1 - \theta)\delta^2; \quad (2.10)$$

$$P\{y_{t-1} = 0, y_t = 1\} = \frac{1}{4}(1 - \varepsilon)(1 - \delta)^2 + \frac{1}{2}\delta(1 - \delta) + \theta(1 - \theta)\delta^2; \quad (2.11)$$

$$P\{y_{t-1} = 0, y_t = 0\} = \frac{1}{4}(1 + \varepsilon)(1 - \delta)^2 + \delta(1 - \theta)(1 - \delta) + \delta^2(1 - \theta)^2. \quad (2.12)$$

Доказательство. Рассмотрим биграмм: $\{y_{t-1}, y_t\}$

$$(a_1, a_2) \in \{0, 1\}^2, P\{y_{t-1} = a_1, y_t = a_2\} = \sum_{(b_1, b_2) \in \{0, 1\}^2} P\{y_{t-1} = b_1, y_t = b_2, \gamma_{t-1} = a_1, \gamma_t = a_2\} = \sum_{(b_1, b_2) \in \{0, 1\}^2} P\{y_{t-1} = b_1, y_t = b_2 | \gamma_{t-1} = a_1, \gamma_t = a_2\} P\{\gamma_{t-1} = a_1, \gamma_t = a_2\}.$$

Для (2.9):

$$\sum_{(b_1, b_2) \in \{0, 1\}^2} P\{y_{t-1} = b_1, y_t = b_2 | \gamma_{t-1} = a_1, \gamma_t = a_2\} P\{\gamma_{t-1} = a_1, \gamma_t = a_2\} = \frac{1}{2} \cdot \frac{1}{2} (1 + \varepsilon)(1 - \delta)^2 + \theta\delta(1 - \theta) + \theta^2\delta^2.$$

Для случаев (2.10)-(2.12) доказывается аналогично. \square

Для формул (2.9)-(2.12) справедливо условие нормировки:

$$\sum_{(a_1, a_2) \in \{0, 1\}^2} P\{y_{t-1} = a_1, y_t = a_2\} = 1.$$

Далее полагаем, что $\theta = \frac{1}{2}$.

Тогда:

$$P\{y_{t-1} = 1, y_t = 1\} = P\{y_{t-1} = 0, y_t = 0\} = \delta^2 \frac{\varepsilon}{4} - \delta \frac{\varepsilon}{2} + \frac{1 + \varepsilon}{4}; \quad (2.13)$$

$$P\{y_{t-1} = 1, y_t = 0\} = P\{y_{t-1} = 0, y_t = 1\} = -\delta^2 \frac{\varepsilon}{4} + \delta \frac{\varepsilon}{2} + \frac{1 - \varepsilon}{4}. \quad (2.14)$$

Определение 2.2.1. [2] Дискретный стационарный источник называется марковским источником порядка m , если для любого $l(l > m)$ и любой последовательности $c_l = (a_{i_1}, \dots, a_{i_l})$ выполняется: $P\{a_{i_l} | a_{i_{l-1}}, \dots, a_{i_1}\} = P\{a_{i_l} | a_{i_{l-1}}, \dots, a_{i_{l-m+1}}\}$.

Определение 2.2.2. [2] Величина: $H^{(k)} = \sum_{\{c_k\}} P\{a_{i_1}, \dots, a_{i_k}\} \log P\{a_{i_k} | a_{i_{k-1}}, \dots, a_{i_1}\}$ называется шаговой энтропией марковского источника порядка k .

Введем понятие энтропии на знак для l -граммы:

$$H_l(\delta) = -\frac{1}{l} \sum_{(a_1, \dots, a_l) \in \{0,1\}^l} P\{y_{t-l} = a_1, \dots, y_{t-1} = a_l\} \log P\{y_{t-l} = a_1, \dots, y_{t-1} = a_l\}. \quad (2.15)$$

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

$$H_l(0) = -\frac{1}{l} (H\{x_1\} + (l-1)H\{x_2 | x_1\}); \quad (2.16)$$

$$\lim_{l \rightarrow \infty} H_l(0) = \lim_{l \rightarrow \infty} -\frac{1}{l} (H\{x_1\} + (l-1)H\{x_2 | x_1\}) = H\{x_2 | x_1\}. \quad (2.17)$$

Рассмотрим случайную величину $\xi \in B = \{b_1, \dots, b_m\}$, заданную на вероятностном пространстве $(\Omega, \mathcal{F}, \mathcal{P})$, $P\{\xi = b_i\} = p_i$.

Определение 2.2.3. [2] Величина

$$I\{b_i\} = -\log p_i \quad (2.18)$$

называется собственной информацией, содержащейся в исходе $b_i \in B$.

Величина $I\{b_i\}$ изменяется от нуля в случае реализации достоверного исхода до бесконечности, когда $p(b_i) = p_i \rightarrow 0$. Величину $I\{b_i\}$ можно интерпретировать как априорную неопределённость события $\{\xi = b_i\}$.

Случайная величина $I\{\xi\}$ имеет математическое ожидание

$$EI\{\xi\} = - \sum_{b_i \in B} p_i \log p_i. \quad (2.19)$$

Определение 2.2.4. Величина $EI\{\xi\}$ называется средней собственной информацией.

Средняя собственная информация равна энтропии: $EI\{\xi\} = H\{\xi\}$.

Для представления функции логарифма воспользуемся формулой Маклорена первого порядка:

$$f(\delta) = f(\delta_0) + (\delta - \delta_0)f'(\delta_0) + o((\delta - \delta_0)^2). \quad (2.20)$$

Для краткости, в обозначении функции \log будем использовать основание b .

Согласно (2.20) в точке $\delta = 0$ справедливо асимптотическое разложение 1-го порядка:

$$\begin{aligned} \log_b(a_0(\varepsilon) + \delta a_1(\varepsilon) + \delta^2 a_2(\varepsilon) + o(\delta^2)) &= \\ = \log a_0(\varepsilon) + \delta \left(\log(a_0(\varepsilon) + \delta a_1(\varepsilon) + \delta^2 a_2(\varepsilon) + o(\delta^2)) \right)' \big|_{\delta=0} + o(\delta) &= \\ = \log a_0(\varepsilon) + \delta \frac{1}{\ln b} \frac{a_1(\varepsilon)}{a_0(\varepsilon)} + o(\delta). \end{aligned}$$

Учитывая (2.20), найдем асимптотические выражения при $\delta \rightarrow 0$ для собственной информации $I\{y_{t-1} = i_1, y_t = i_2\}, i_1, i_2 \in \{0, 1\}$:

$$\begin{aligned} I\{y_{t-1} = 0, y_t = 0\} &= I\{y_{t-1} = 1, y_t = 1\} = -\log\left(\frac{1}{4}(1+\varepsilon)(1-\delta)^2 + \frac{1}{2}\delta(1-\delta) + \frac{1}{4}\delta^2\right) = \\ &= -\log \frac{1+\varepsilon}{4} + \delta \frac{1}{\ln b} \frac{2\varepsilon}{1+\varepsilon} + o(\delta); \\ I\{y_{t-1} = 0, y_t = 1\} &= I\{y_{t-1} = 1, y_t = 0\} = -\log\left(\frac{1}{4}(1-\varepsilon)(1-\delta)^2 + \frac{1}{2}\delta(1-\delta) + \frac{1}{4}\delta^2\right) = \\ &= -\log \frac{1-\varepsilon}{4} - \delta \frac{1}{\ln b} \frac{2\varepsilon}{1-\varepsilon} + o(\delta). \end{aligned}$$

Лемма 2.2.2. Если имеет место монобитная модель вкраплений (2.1)-(2.4), то для энтропии при $l = 2$ справедливо асимптотическое разложение 1-го порядка

$$H_2(\delta) = H_2(0) + 2\delta\varepsilon \log \frac{1+\varepsilon}{1-\varepsilon} + O(\delta^2). \quad (2.21)$$

Доказательство. $H_2(\delta) = -(P\{y_{t-1} = 0, y_t = 0\} \log(P\{y_{t-1} = 0, y_t = 0\}) + P\{y_{t-1} = 0, y_t = 1\} \log(P\{y_{t-1} = 0, y_t = 1\}) + P\{y_{t-1} = 1, y_t = 0\} \log(P\{y_{t-1} = 1, y_t = 0\}) + P\{y_{t-1} = 1, y_t = 1\} \log(P\{y_{t-1} = 1, y_t = 1\})) = -2(P\{y_{t-1} = 0, y_t = 0\} \log(P\{y_{t-1} = 0, y_t = 0\}) + P\{y_{t-1} = 0, y_t = 0\} \log(P\{y_{t-1} = 0, y_t = 1\}) + P\{y_{t-1} = 0, y_t = 1\} \log(P\{y_{t-1} = 0, y_t = 0\}) + P\{y_{t-1} = 0, y_t = 1\} \log(P\{y_{t-1} = 0, y_t = 1\}) + P\{y_{t-1} = 1, y_t = 0\} \log(P\{y_{t-1} = 1, y_t = 0\}) + P\{y_{t-1} = 1, y_t = 0\} \log(P\{y_{t-1} = 1, y_t = 1\}) + P\{y_{t-1} = 1, y_t = 1\} \log(P\{y_{t-1} = 1, y_t = 0\}) + P\{y_{t-1} = 1, y_t = 1\} \log(P\{y_{t-1} = 1, y_t = 1\})) = -2((\delta^2 \frac{\varepsilon}{4} - \delta \frac{\varepsilon}{2} + \frac{1+\varepsilon}{4})(-\log(\frac{1+\varepsilon}{4}) + \delta \frac{1}{\ln b} \cdot \frac{2\varepsilon}{1+\varepsilon} + o(\delta^2)) + (-\delta^2 \frac{\varepsilon}{4} + \delta \frac{\varepsilon}{2} + \frac{1-\varepsilon}{4})(-\log(\frac{1-\varepsilon}{4}) - \delta \frac{1}{\ln b} \cdot \frac{2\varepsilon}{1-\varepsilon} + o(\delta^2))) = \frac{1}{2}(-(1+\varepsilon) \log(\frac{1+\varepsilon}{4}) - (1-\varepsilon) \log(\frac{1-\varepsilon}{4}) + 2\delta\varepsilon \log(\frac{1+\varepsilon}{1-\varepsilon}))) + O(\delta^2) = H_2(0) + 2\delta\varepsilon \log(\frac{1+\varepsilon}{1-\varepsilon}) + O(\delta^2). \quad \square$

Определение 2.2.5. [2] Величина $\lim_{k \rightarrow \infty} H^{(k)} = \lim_{k \rightarrow \infty} H_k = H_\infty \geq 0$ называется энтропией марковского источника, где H_k - энтропия на знак.

Рассмотрим условные вероятности появления трехграммы $(0, 0, 0)$ при условии всевозможных стегоключей.

$$\begin{aligned} P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0\} &= (1-\delta)^3 P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0 | \gamma_{i-1} = 0, \gamma_i = 0, \gamma_{i+1} = 0\} + \\ &+ \delta(1-\delta)^2 (P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0 | \gamma_{i-1} = 1, \gamma_i = 0, \gamma_{i+1} = 0\} + P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0 | \gamma_{i-1} = 0, \gamma_i = 1, \gamma_{i+1} = 0\} + P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0 | \gamma_{i-1} = 1, \gamma_i = 1, \gamma_{i+1} = 0\} + P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0 | \gamma_{i-1} = 0, \gamma_i = 0, \gamma_{i+1} = 1\} + P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0 | \gamma_{i-1} = 1, \gamma_i = 0, \gamma_{i+1} = 1\} + P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0 | \gamma_{i-1} = 0, \gamma_i = 1, \gamma_{i+1} = 1\} + P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0 | \gamma_{i-1} = 1, \gamma_i = 1, \gamma_{i+1} = 1\}) \end{aligned}$$

$$\begin{aligned}
& 0|\gamma_{i-1} = 0, \gamma_i = 0, \gamma_{i+1} = 1\}) + \\
& + \delta^2(1 - \delta)(P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0|\gamma_{i-1} = 1, \gamma_i = 1, \gamma_{i+1} = 0\} + P\{y_{i-1} = \\
& 0, y_i = 0, y_{i+1} = 0|\gamma_{i-1} = 0, \gamma_i = 1, \gamma_{i+1} = 1\} + P\{y_{i-1} = 0, y_i = 0, y_{i+1} = \\
& 0|\gamma_{i-1} = 1, \gamma_i = 0, \gamma_{i+1} = 1\}) + \\
& + \delta^3(P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0|\gamma_{i-1} = 1, \gamma_i = 1, \gamma_{i+1} = 1\}).
\end{aligned}$$

$$\begin{aligned}
P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0|\gamma_1 = 0, \gamma_2 = 0, \gamma_3 = 0\} &= P\{x_{i-1} = 0, x_i = \\
0, x_{i+1} = 0\} &= P\{x_{i-1} = 0\}P\{x_i = 0, x_{i+1} = 0|x_{i-1} = 0\} = P\{x_{i-1} = 0\}P\{x_i = \\
0\}P\{x_{i+1} = 0|x_i = 0\} &= \frac{1}{2} \cdot \frac{1}{2}(1 + \varepsilon) \cdot \frac{1}{2}(1 + \varepsilon) = \frac{1}{8}(1 + \varepsilon)^2;
\end{aligned}$$

$$\begin{aligned}
P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0|\gamma_1 = 1, \gamma_2 = 0, \gamma_3 = 0\} &= P\{\xi = 0, x_i = \\
0, x_{i+1} = 0\} &= P\{\xi = 0\}P\{x_i = 0, x_{i+1} = 0\} = P\{\xi = 0\}P\{x_i = 0\}P\{x_{i+1} = \\
0|x_i = 0\} &= \frac{1}{2} \cdot \frac{1}{2}(1 + \varepsilon) \cdot \frac{1}{2} = \frac{1}{8}(1 + \varepsilon);
\end{aligned}$$

$$\begin{aligned}
P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0|\gamma_1 = 0, \gamma_2 = 1, \gamma_3 = 0\} &= P\{x_{i-1} = 0, \xi = \\
0, x_{i+1} = 0\} &= P\{\xi = 0\}P\{x_{i-1} = 0, x_{i+1} = 0\} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}(1 + \varepsilon^2) = \frac{1}{8}(1 + \varepsilon^2);
\end{aligned}$$

$$\begin{aligned}
P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0|\gamma_1 = 0, \gamma_2 = 0, \gamma_3 = 1\} &= P\{x_{i-1} = 0, x_i = \\
0, \xi = 0\} &= P\{\xi = 0\}P\{x_{i-1} = 0, x_i = 0\} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}(1 + \varepsilon) = \frac{1}{8}(1 + \varepsilon);
\end{aligned}$$

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

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

Найдем вероятности появления всевозможных трехграмм в стежоконтейнере $\{y_t\}$:

$$\begin{aligned}
P\{y_{i-1} = 1, y_i = 1, y_{i+1} = 1\} &= P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0\} = \\
&= \frac{1}{8}(\varepsilon(\varepsilon + 2)\delta^2 - 2\varepsilon(\varepsilon + 2)\delta + (1 + \varepsilon)^2), \\
P\{y_{i-1} = 1, y_i = 1, y_{i+1} = 0\} &= \\
= P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 1\} &= P\{y_{i-1} = 0, y_i = 1, y_{i+1} = 1\} = \\
= P\{y_{i-1} = 1, y_i = 0, y_{i+1} = 0\} &= \frac{1}{8}(-\varepsilon^2\delta^2 + 2\varepsilon^2\delta - \varepsilon^2 + 1), \\
P\{y_{i-1} = 1, y_i = 0, y_{i+1} = 1\} &= P\{y_{i-1} = 0, y_i = 1, y_{i+1} = 0\} = \\
&= \frac{1}{8}(\varepsilon(\varepsilon - 2)\delta^2 - 2\varepsilon(\varepsilon - 2)\delta + (1 - \varepsilon)^2).
\end{aligned}$$

Теорема 2.2.1. Если имеет место монобитная модель вкраплений (2.1)-(2.4), то для энтропии при $l = 3$ справедливо асимптотическое разложение 1-го

порядка:

$$H_3(\delta) = H_3(0) + 2\varepsilon\delta \log \frac{1+\varepsilon}{1-\varepsilon} + O(\delta^2); \quad (2.22)$$

собственная информация имеет вид:

$$\begin{aligned} I\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0\} &= -\log \frac{(1+\varepsilon)^2}{8} + \delta \frac{1}{\ln b} \cdot \frac{2\varepsilon^2 + 4\varepsilon}{(1+\varepsilon)^2} + O(\delta^2), \\ I\{y_{i-1} = 1, y_i = 0, y_{i+1} = 0\} &= I\{y_{i-1} = 0, y_i = 0, y_{i+1} = 1\} = \\ &= -\log \frac{1-\varepsilon^2}{8} - \delta \frac{1}{\ln b} \cdot \frac{2\varepsilon^2}{1-\varepsilon^2} + O(\delta^2), \\ I\{y_{i-1} = 0, y_i = 1, y_{i+1} = 0\} &= -\log \frac{(1-\varepsilon)^2}{8} + \delta \frac{1}{\ln b} \cdot \frac{2\varepsilon^2 - 4\varepsilon}{(1-\varepsilon)^2} + O(\delta^2); \\ I\{y_{i-1} = j_1, y_i = j_2, y_{i+1} = j_3\} &= I\{y_{i-1} = 1 - j_1, y_i = 1 - j_2, y_{i+1} = 1 - j_3\}, \\ & \quad j_1, j_2, j_3 \in \{0, 1\}. \end{aligned}$$

Доказательство. Подставляя в (2.20) найденные выражения для вероятностей трехграмм, получим асимптотические выражения для собственной информации.

Используя выражения для собственной информации, получим:

$$\begin{aligned} H_3(\delta) &= -2\left(\frac{1}{8}(\varepsilon(\varepsilon+2)\delta^2 - 2\varepsilon(\varepsilon+2)\delta + (1+\varepsilon)^2)\left(\log\left(\frac{(1+\varepsilon)^2}{8}\right) + \delta \frac{1}{\ln b} \cdot \frac{-2\varepsilon^2-4\varepsilon}{(1+\varepsilon)^2}\right) + \right. \\ & \quad \left. 2\frac{1}{8}(-\varepsilon^2\delta^2 + 2\varepsilon^2\delta - \varepsilon^2 + 1)\left(\log\left(\frac{1-\varepsilon^2}{8}\right) + \delta \frac{1}{\ln b} \cdot \frac{2\varepsilon^2}{1-\varepsilon^2}\right) + \frac{1}{8}(\varepsilon(\varepsilon-2)\delta^2 - 2\varepsilon(2+\varepsilon)\delta + (1-\varepsilon)^2)\right. \\ & \quad \left. \left(\log\left(\frac{(1-\varepsilon)^2}{8}\right) + \delta \frac{1}{\ln b} \cdot \frac{-2\varepsilon^2+4\varepsilon}{(1-\varepsilon)^2}\right)\right) + O(\delta^2) = -((1-\varepsilon)\log(1-\varepsilon) + (1+\varepsilon)\log(1+\varepsilon) + \\ & \quad \varepsilon) + \log\left(\frac{1}{8}\right) + 2\varepsilon\delta \log \frac{1-\varepsilon}{1+\varepsilon} + O(\delta^2) = H_3(0) - 2\varepsilon\delta \log \frac{1-\varepsilon}{1+\varepsilon} + O(\delta^2). \quad \square \end{aligned}$$

Рассмотрим условные вероятности появления четырехграммы $(0, 0, 0, 0)$ при условии всевозможных стегоключей.

$$\begin{aligned} P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0, y_{i+2} = 0\} &= (1-\delta)^4 P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0, y_{i+2} = 0 | \gamma_{i-1} = 0, \gamma_i = 0, \gamma_{i+1} = 0, \gamma_{i+2} = 0\} + \\ & \quad \delta(1-\delta)^3 \left(P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0, y_{i+2} = 0 | \gamma_{i-1} = 0, \gamma_i = 0, \gamma_{i+1} = 0, \gamma_{i+2} = 1\} + P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0, y_{i+2} = 0 | \gamma_{i-1} = 0, \gamma_i = 0, \gamma_{i+1} = 1, \gamma_{i+2} = 0\} + \right. \\ & \quad \left. P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0, y_{i+2} = 0 | \gamma_{i-1} = 0, \gamma_i = 1, \gamma_{i+1} = 0, \gamma_{i+2} = 0\} + P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0, y_{i+2} = 0 | \gamma_{i-1} = 1, \gamma_i = 0, \gamma_{i+1} = 0, \gamma_{i+2} = 0\} \right) + \\ & \quad \delta^2(1-\delta)^2 \left(P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0, y_{i+2} = 0 | \gamma_{i-1} = 0, \gamma_i = 0, \gamma_{i+1} = 1, \gamma_{i+2} = 1\} + P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0, y_{i+2} = 0 | \gamma_{i-1} = 0, \gamma_i = 1, \gamma_{i+1} = 1, \gamma_{i+2} = 0\} + \right. \\ & \quad \left. P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0, y_{i+2} = 0 | \gamma_{i-1} = 1, \gamma_i = 1, \gamma_{i+1} = 0, \gamma_{i+2} = 0\} + P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0, y_{i+2} = 0 | \gamma_{i-1} = 1, \gamma_i = 0, \gamma_{i+1} = 1, \gamma_{i+2} = 0\} + \right. \\ & \quad \left. P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0, y_{i+2} = 0 | \gamma_{i-1} = 0, \gamma_i = 1, \gamma_{i+1} = 0, \gamma_{i+2} = 1\} + P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0, y_{i+2} = 0 | \gamma_{i-1} = 1, \gamma_i = 0, \gamma_{i+1} = 0, \gamma_{i+2} = 1\} \right) + \\ & \quad \delta^3(1-\delta) \left(P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0, y_{i+2} = 0 | \gamma_{i-1} = 0, \gamma_i = 0, \gamma_{i+1} = 0, \gamma_{i+2} = 1\} + \right. \end{aligned}$$

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

$$P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0, y_{i+2} = 0 | \gamma_{i-1} = 0, \gamma_i = 0, \gamma_{i+1} = 0, \gamma_{i+2} = 0\} = \frac{(1+\varepsilon)^3}{16};$$

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

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

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

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

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

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

Найдем вероятности появления всевозможных четырехграмм в стежоконттей-

нере $\{y_t\}$:

$$\begin{aligned}
P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0, y_{i+2} = 1\} &= P\{y_{i-1} = 1, y_i = 1, y_{i+1} = 1, y_{i+2} = 0\} = \\
P\{y_{i-1} = 1, y_i = 0, y_{i+1} = 0, y_{i+2} = 0\} &= P\{y_{i-1} = 0, y_i = 1, y_{i+1} = 1, y_{i+2} = 1\} = \\
\frac{1}{16}(\delta^4(-\varepsilon^2) + \delta^3 \cdot 4\varepsilon^2 + \delta^2(-\varepsilon^3 - 6\varepsilon^2 + \varepsilon) + \delta(2\varepsilon^3 + 4\varepsilon^2 - 2\varepsilon) - \varepsilon^3 - \varepsilon^2 + \varepsilon + 1); \\
P\{y_{i-1} = 0, y_i = 1, y_{i+1} = 0, y_{i+2} = 0\} &= P\{y_{i-1} = 1, y_i = 0, y_{i+1} = 1, y_{i+2} = 1\} = \\
P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 1, y_{i+2} = 0\} &= P\{y_{i-1} = 1, y_i = 1, y_{i+1} = 0, y_{i+2} = 1\} = \\
\frac{1}{16}(\delta^4(-\varepsilon^2) + \delta^3 \cdot 4\varepsilon^2 + \delta^2(\varepsilon^3 - 6\varepsilon^2 - \varepsilon) + \delta(-2\varepsilon^3 + 4\varepsilon^2 + 2\varepsilon) + \varepsilon^3 - \varepsilon^2 - \varepsilon + 1); \\
P\{y_{i-1} = 0, y_i = 0, y_{i+1} = 1, y_{i+2} = 1\} &= P\{y_{i-1} = 1, y_i = 1, y_{i+1} = 0, y_{i+2} = 0\} = \\
\frac{1}{16}(\delta^4\varepsilon^2 + \delta^3(-4\varepsilon^2) + \delta^2(-\varepsilon^3 + 4\varepsilon^2 + \varepsilon) + \delta(2\varepsilon^3 - 2\varepsilon) - \varepsilon^3 - \varepsilon^2 + \varepsilon + 1); \\
P\{y_{i-1} = 0, y_i = 1, y_{i+1} = 1, y_{i+2} = 0\} &= P\{y_{i-1} = 1, y_i = 0, y_{i+1} = 0, y_{i+2} = 1\} = \\
\frac{1}{16}(\delta^4\varepsilon^2 + \delta^3(-4\varepsilon^2) + \delta^2(\varepsilon^3 + 4\varepsilon^2 - \varepsilon) + \delta(-2\varepsilon^3 + 2\varepsilon) + \varepsilon^3 - \varepsilon^2 - \varepsilon + 1); \\
P\{y_{i-1} = 1, y_i = 0, y_{i+1} = 1, y_{i+2} = 0\} &= P\{y_{i-1} = 0, y_i = 1, y_{i+1} = 0, y_{i+2} = 1\} = \\
= \frac{1}{16}(\delta^4\varepsilon^2 + \delta^3(-4\varepsilon^2) + \delta^2(-\varepsilon^3 + 8\varepsilon^2 - 3\varepsilon) + \delta(2\varepsilon^3 - 8\varepsilon^2 + 6\varepsilon) \\
- \varepsilon^3 + 3\varepsilon^2 - 3\varepsilon + 1).
\end{aligned}$$

Теорема 2.2.2. Если имеет место монобитная модель вкраплений (2.1)-(2.4), то для энтропии при $l = 4$ справедливо асимптотическое разложение 1-го порядка:

$$H_4(\delta) = H_4(0) + \frac{24\varepsilon\delta}{16} \log \frac{1+\varepsilon}{1-\varepsilon} + O(\delta^2); \quad (2.23)$$

собственная информация имеет вид:

$$\begin{aligned}
I\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0, y_{i+2} = 0\} &= I\{y_{i-1} = 1, y_i = 1, y_{i+1} = 1, y_{i+2} = 1\} = \\
= -\left(\log \frac{(1+\varepsilon)^3}{16} + \delta \frac{-2\varepsilon^3 - 8\varepsilon^2 - 6\varepsilon}{(1+\varepsilon)^3 \ln b} \right) + O(\delta^2); \\
I\{y_{i-1} = 0, y_i = 0, y_{i+1} = 0, y_{i+2} = 1\} &= I\{y_{i-1} = 1, y_i = 1, y_{i+1} = 1, y_{i+2} = 0\} = \\
I\{y_{i-1} = 1, y_i = 0, y_{i+1} = 0, y_{i+2} = 0\} &= I\{y_{i-1} = 0, y_i = 1, y_{i+1} = 1, y_{i+2} = 1\} = \\
= -\left(\log \frac{(1-\varepsilon)(1+\varepsilon)^2}{16} + \delta \frac{2\varepsilon^3 + 4\varepsilon^2 - 2\varepsilon}{(1-\varepsilon)(1+\varepsilon)^2 \ln b} \right) + O(\delta^2); \\
I\{y_{i-1} = 0, y_i = 0, y_{i+1} = 1, y_{i+2} = 0\} &= I\{y_{i-1} = 1, y_i = 1, y_{i+1} = 0, y_{i+2} = 1\} = \\
I\{y_{i-1} = 0, y_i = 1, y_{i+1} = 0, y_{i+2} = 0\} &= I\{y_{i-1} = 1, y_i = 0, y_{i+1} = 1, y_{i+2} = 1\} = \\
= -\left(\log \frac{(1-\varepsilon)^2(1+\varepsilon)}{16} + \delta \frac{-2\varepsilon^3 + 4\varepsilon^2 + 2\varepsilon}{(1-\varepsilon)^2(1+\varepsilon) \ln b} \right) + O(\delta^2);
\end{aligned}$$

$$\begin{aligned}
I\{y_{i-1} = 0, y_i = 0, y_{i+1} = 1, y_{i+2} = 1\} &= I\{y_{i-1} = 1, y_i = 1, y_{i+1} = 0, y_{i+2} = 0\} = \\
&= -\left(\log \frac{(1-\varepsilon)(1+\varepsilon)^2}{16} + \delta \frac{2\varepsilon^3 - 2\varepsilon}{(1-\varepsilon)(1+\varepsilon)^2 \ln b}\right) + O(\delta^2); \\
I\{y_{i-1} = 0, y_i = 1, y_{i+1} = 1, y_{i+2} = 0\} &= I\{y_{i-1} = 1, y_i = 0, y_{i+1} = 0, y_{i+2} = 1\} = \\
&= -\left(\log \frac{(1-\varepsilon)^2(1+\varepsilon)}{16} + \delta \frac{-2\varepsilon^3 + 2\varepsilon}{(1-\varepsilon)^2(1+\varepsilon) \ln b}\right) + O(\delta^2); \\
I\{y_{i-1} = 1, y_i = 0, y_{i+1} = 1, y_{i+2} = 0\} &= I\{y_{i-1} = 0, y_i = 1, y_{i+1} = 0, y_{i+2} = 1\} = \\
&= -\left(\log \frac{(1-\varepsilon)^3}{16} + \delta \frac{2\varepsilon^3 - 8\varepsilon^2 + 6\varepsilon}{(1-\varepsilon)^3 \ln b}\right) + O(\delta^2).
\end{aligned}$$

Доказательство. Подставляя в (2.20) найденные выражения для вероятностей четырехграмм, получим асимптотические выражения для собственной информации.

Используя выражения для собственной информации, получим:

$$\begin{aligned}
H_4(\delta) &= -2\left(\frac{1}{16}(\delta^4\varepsilon^2 + \delta^3(-4\varepsilon^2) + \delta^2(\varepsilon^3 + 8\varepsilon^2 + 3\varepsilon) + \delta(-2\varepsilon^3 - 8\varepsilon^2 - 6\varepsilon) + \varepsilon^3 + 3\varepsilon^2 + 3\varepsilon + 1)\left(\log \frac{(1+\varepsilon)^3}{16} + \delta \frac{-2\varepsilon^3 - 8\varepsilon^2 - 6\varepsilon}{(1+\varepsilon)^3 \ln b} + O(\delta^2)\right) + 2\frac{1}{16}(\delta^4(-\varepsilon^2) + \delta^3 \cdot 4\varepsilon^2 + \delta^2(-\varepsilon^3 - 6\varepsilon^2 + \varepsilon) + \delta(2\varepsilon^3 + 4\varepsilon^2 - 2\varepsilon) - \varepsilon^3 - \varepsilon^2 + \varepsilon + 1)\left(\log \frac{(1-\varepsilon)(1+\varepsilon)^2}{16} + \delta \frac{2\varepsilon^3 + 4\varepsilon^2 - 2\varepsilon}{(1-\varepsilon)(1+\varepsilon)^2 \ln b} + O(\delta^2)\right) + 2\frac{1}{16}(\delta^4(-\varepsilon^2) + \delta^3 \cdot 4\varepsilon^2 + \delta^2(\varepsilon^3 - 6\varepsilon^2 - \varepsilon) + \delta(-2\varepsilon^3 + 4\varepsilon^2 + 2\varepsilon) + \varepsilon^3 - \varepsilon^2 - \varepsilon + 1)\left(\log \frac{(1-\varepsilon)^2(1+\varepsilon)}{16} + \delta \frac{-2\varepsilon^3 + 4\varepsilon^2 + 2\varepsilon}{(1-\varepsilon)^2(1+\varepsilon) \ln b} + O(\delta^2)\right) + \frac{1}{16}(\delta^4\varepsilon^2 + \delta^3(-4\varepsilon^2) + \delta^2(-\varepsilon^3 + 4\varepsilon^2 + \varepsilon) + \delta(2\varepsilon^3 - 2\varepsilon) - \varepsilon^3 - \varepsilon^2 + \varepsilon + 1)\left(\log \frac{(1-\varepsilon)(1+\varepsilon)^2}{16} + \delta \frac{2\varepsilon^3 - 2\varepsilon}{(1-\varepsilon)(1+\varepsilon)^2 \ln b} + O(\delta^2)\right) + \frac{1}{16}(\delta^4\varepsilon^2 + \delta^3(-4\varepsilon^2) + \delta^2(\varepsilon^3 + 4\varepsilon^2 - \varepsilon) + \delta(-2\varepsilon^3 + 2\varepsilon) + \varepsilon^3 - \varepsilon^2 - \varepsilon + 1)\left(\log \frac{(1-\varepsilon)^2(1+\varepsilon)}{16} + \delta \frac{-2\varepsilon^3 + 2\varepsilon}{(1-\varepsilon)^2(1+\varepsilon) \ln b} + O(\delta^2)\right) + \frac{1}{16}(\delta^4\varepsilon^2 + \delta^3(-4\varepsilon^2) + \delta^2(-\varepsilon^3 + 8\varepsilon^2 - 3\varepsilon) + \delta(2\varepsilon^3 - 8\varepsilon^2 + 6\varepsilon) - \varepsilon^3 + 3\varepsilon^2 - 3\varepsilon + 1)\left(\log \frac{(1-\varepsilon)^3}{16} + \delta \frac{2\varepsilon^3 - 8\varepsilon^2 + 6\varepsilon}{(1-\varepsilon)^3 \ln b} + O(\delta^2)\right)\right) = \\
&= \frac{(1+\varepsilon)^3}{16} \log \frac{(1+\varepsilon)^3}{16} + 3\frac{(1+\varepsilon)^2(1-\varepsilon)}{16} \log \frac{(1+\varepsilon)^2(1-\varepsilon)}{16} + 3\frac{(1+\varepsilon)(1-\varepsilon)^2}{16} \log \frac{(1+\varepsilon)(1-\varepsilon)^2}{16} + \\
&+ \frac{(1-\varepsilon)^3}{16} \log \frac{(1-\varepsilon)^3}{16} + \frac{24\varepsilon\delta}{16} \log \frac{1+\varepsilon}{1-\varepsilon} + O(\delta^2) = H_4(0) + \frac{24\varepsilon\delta}{16} \log \frac{1+\varepsilon}{1-\varepsilon} + O(\delta^2). \quad \square
\end{aligned}$$

Оценим остаточный член для асимптотического выражения энтропии биграммы:

$$r_n(\delta) = \frac{f^{(n+1)}(\bar{\delta})}{(n+1)!}(\delta - \delta_0), \bar{\delta} \in [\delta_0, \delta]. \quad (2.24)$$

Для асимптотического разложения 1-го порядка при $\delta_0 = 0$ остаточный член имеет вид:

$$r_n(\delta) = \frac{f''(\bar{\delta})}{2}\delta, \bar{\delta} \in [0, \delta]. \quad (2.25)$$

$$(\log(P_{00}))'' = \frac{-2\varepsilon(\delta^2\varepsilon - 2\delta\varepsilon + \varepsilon - 1)}{\ln b(\delta^2\varepsilon - 2\delta\varepsilon + \varepsilon + 1)^2};$$

$$(\log(P_{01}))'' = \frac{-2\varepsilon(\delta^2\varepsilon - 2\delta\varepsilon + \varepsilon + 1)}{\ln b(\delta^2\varepsilon - 2\delta\varepsilon + \varepsilon - 1)^2};$$

Тогда остаточный член для $\log(P_{00}) = \log(P_{11})$ равен:

$$r_{n_{00}}(\delta) = \frac{\delta}{2} \cdot \frac{-2\varepsilon(\delta^2\varepsilon - 2\delta\varepsilon + \varepsilon - 1)}{\ln b(\delta^2\varepsilon - 2\delta\varepsilon + \varepsilon + 1)^2} \Big|_{\delta=\bar{\delta}} \quad (2.26)$$

Остаточный член для $\log(P_{10}) = \log(P_{01})$ равен:

$$r_{n_{10}}(\delta) = \frac{\delta}{2} \cdot \frac{-2\varepsilon(\delta^2\varepsilon - 2\delta\varepsilon + \varepsilon + 1)}{\ln b(\delta^2\varepsilon - 2\delta\varepsilon + \varepsilon - 1)^2} \Big|_{\delta=\bar{\delta}} \quad (2.27)$$

Тогда на основании (2.26) и (2.27) остаточный член для энтропии будет равен:

$$\begin{aligned} R_n(\delta) = & -2(P_{00}r_{n_{00}}(\delta) + P_{10}r_{n_{01}}(\delta)) = -2\left((\delta^2\frac{\varepsilon}{4} - \delta\frac{\varepsilon}{2} + \frac{1+\varepsilon}{4}) \right. \\ & \left. \left(\frac{\delta}{2} \cdot \frac{-2\varepsilon(\delta^2\varepsilon - 2\delta\varepsilon + \varepsilon - 1)}{\ln b(\delta^2\varepsilon - 2\delta\varepsilon + \varepsilon + 1)^2} \Big|_{\delta=\bar{\delta}}\right) + \right. \\ & \left. \left(-\delta^2\frac{\varepsilon}{4} + \delta\frac{\varepsilon}{2} + \frac{1-\varepsilon}{4}\right) \left(\frac{\delta}{2} \cdot \frac{-2\varepsilon(\delta^2\varepsilon - 2\delta\varepsilon + \varepsilon + 1)}{\ln b(\delta^2\varepsilon - 2\delta\varepsilon + \varepsilon - 1)^2} \Big|_{\delta=\bar{\delta}}\right)\right) \end{aligned}$$

2.3 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 Correspondence Analysis). [1]

2.3.1 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 :

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

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.

2.3.2 Результаты линейного дискриминантного анализа

Линейный дискриминантный анализ применен для классификации последовательностей с вкраплениями и без вкраплений при фиксированном параметре ε .

Пусть имеется последовательность $Y = \{y_1, \dots, y_T\}$, на основании Y вычисляем $(H_3(\delta), H_4(\delta))$ при фиксированном ε , 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 вкраплениями) последовательностей можно провести дискриминантный анализ и оценить вероятность правильной классификации и мощность критерия.

Тогда:

$$\hat{\alpha} = \frac{n_0 - \nu_0}{n_0} - \text{оценка вероятности ошибки первого рода;} \quad (2.28)$$

$$\hat{\beta} = \frac{n_1 - \nu_1}{n_1} - \text{оценка вероятности ошибки второго рода;} \quad (2.29)$$

ν_0 - количество верно определенных пустых последовательностей, ν_1 - количество верно определенных последовательностей с вкраплениями.

Мощность критерия:

$$\hat{w} = \frac{\nu_1}{n_1} \quad (2.30)$$

Таблица 2.1 — Результаты дискриминантного анализа при $\varepsilon = 0.55$.

δ	$\hat{\alpha}$	$\hat{\beta}$	\hat{w}
0.03	0.42	0.31	0.69
0.07	0.21	0.14	0.86
0.08	0.1	0.13	0.87
0.09	0.14	0.08	0.92
0.1	0.13	0.05	0.95
0.3	0.05	0.02	0.98

Таблица 2.2 — Результаты дискриминантного анализа при $\varepsilon = 0.15$.

δ	$\hat{\alpha}$	$\hat{\beta}$	\hat{w}
0.01	0.47	0.4	0.6
0.03	0.3	0.2	0.8
0.05	0.18	0.16	0.84
0.07	0.1	0.08	0.92
0.1	0.06	0.05	0.95
0.3	0.02	0.02	0.98

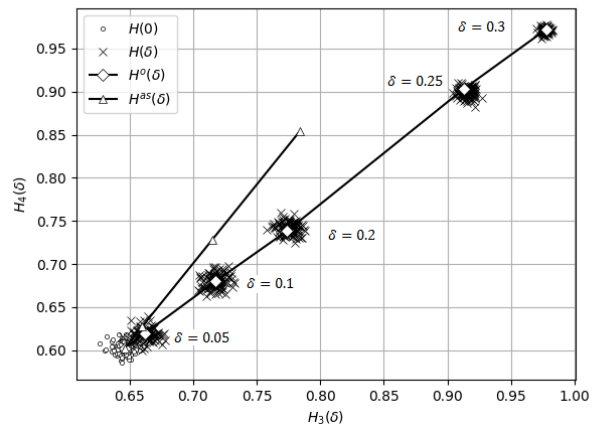


Рисунок 2.1 — График зависимости энтропии $H_4(\delta)$ от $H_3(\delta)$ при различных долях вкраплений

2.3.3 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 0.05 при доле вкраплений $\delta \geq 0.1$.

2.4 Исследование реальных данных на основании изображений в формате JPEG

Формат файла JPEG (Joint Photographic Experts Group - Объединенная экспертная группа по фотографии, произносится "джейпег") был разработан компанией C-Cube Microsystems как эффективный метод хранения изображений с большой глубиной цвета, например, получаемых при сканировании фотографий с многочисленными едва уловимыми (а иногда и неуловимыми) оттенками цвета. Основное отличие формата JPEG от других форматов состоит в том, что в JPEG используется алгоритм сжатия с потерями (а не алгоритм без потерь) информации. Алгоритм сжатия без потерь так сохраняет информацию об изображении, что распакованное изображение в точности соответствует оригиналу. При сжатии с потерями теряется часть информации об изображении, чтобы достичь большего сжатия. Распакованное изображение JPEG редко соответствует оригиналу абсолютно точно, но очень часто эти различия столь незначительны, что их едва можно (если вообще можно) обнаружить.

Ключевым компонентом работы алгоритма является дискретное косинусное преобразование. Дискретное косинусное преобразование (2.31) представляет собой разновидность преобразования Фурье и, так же как и оно, имеет обратное преобразование. Графическое изображение можно рассматривать как совокупность пространственных волн, причем оси X и Y совпадают с шириной и высотой картинки, а по оси Z откладывается значение цвета соответствующего пикселя изображения. Дискретное косинусное преобразование позволяет переходить от пространственного представления картинки к ее спектральному представлению и обратно. Воздействуя на спектральное представление картинки, состоящее из "гармоник", то есть, отбрасывая наименее значимые из них, можно балансировать между качеством воспроизведения и степенью сжатия [5].

$$Y[u, v] = \frac{1}{\sqrt{2N}} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} C(i, u) C(j, v) y(i, j), \quad (2.31)$$

где $C(i, u) = A(u) \cos\left(\frac{(2i+1)u\pi}{2N}\right)$ – гармоника сигнала,

$$A(u) = \begin{cases} \frac{1}{\sqrt{2}}, u=0; \\ 1, u \neq 0; \end{cases} \quad \text{– постоянная составляющая.}$$

Обратное дискретное косинусное преобразование:

$$y(i, j) = \frac{1}{\sqrt{2N}} \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} C(i, u) C(j, v) Y[u, v]. \quad (2.32)$$

В получившейся матрице коэффициентов низкочастотные компоненты расположены ближе к левому верхнему углу, а высокочастотные - справа и внизу. Это важно потому, что большинство графических образов на экране компьютера состоит из низкочастотной информации. Высокочастотные компоненты не так важны для передачи изображения. Таким образом, дискретное косинусное преобразование позволяет определить, какую часть информации можно безболезненно выбросить, не внося серьезных искажений в картинку.

Далее происходит квантование коэффициентов ДКП. Здесь каждое число из матрицы делится на элемент из таблицы квантования, а результат округляется до ближайшего целого:

$$Y^q(u, v) = \text{Round}\left(\frac{Y(u, v)}{q(u, v)}\right), \quad (2.33)$$

где $q(u, v)$ - элемент таблицы квантования

Стандарт JPEG даже допускает использование собственных таблиц квантования, которые, однако, необходимо будет передавать декодеру вместе со сжатыми данными, что увеличит общий размер файла. Понятно, что пользователю сложно самостоятельно подобрать 64 коэффициента, поэтому стандарт JPEG использует два подхода для матриц квантования. Первый заключается в том, что в стандарт JPEG включены две рекомендуемые таблицы квантования: одна для яркости, вторая для цветности. Второй подход заключается в синтезе (вычислении на лету) таблицы квантования, зависящей от одного параметра, который задается пользователем. Сама таблица строится по формуле:

$$Q(i, j) = 1 + (i + j)R, \quad (2.34)$$

где R - коэффициент сжатия.

Чем больше коэффициент квантования, тем больше данных теряется, поскольку реальное ДКП-значение представляется все менее и менее точно. Кроме того, для данных яркости и цветности применяются отдельные таблицы квантования, позволяющие квантовать данные цветности с большими коэффициентами, чем данные яркости. Таким образом, JPEG использует различную чувствительность глаза к яркости и цветности изображения.

Далее полученная матрица 8×8 переводится в 64-элементный вектор при помощи "зигзаг" -сканирования (Рис. 2.2). Таким образом, в начале вектора мы получаем коэффициенты матрицы, соответствующие низким частотам, а в конце - высоким.

Заключительная стадия алгоритма сжатия JPEG - кодирование. Полученный вектор обрабатывается с помощью алгоритмов Хаффмана или арифметического кодирования, в зависимости от реализации.

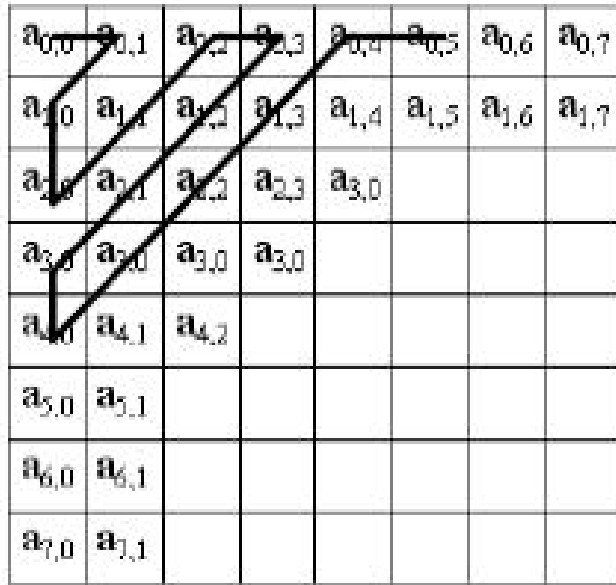


Рисунок 2.2 — "Зигзаг" -сканирование матрицы.

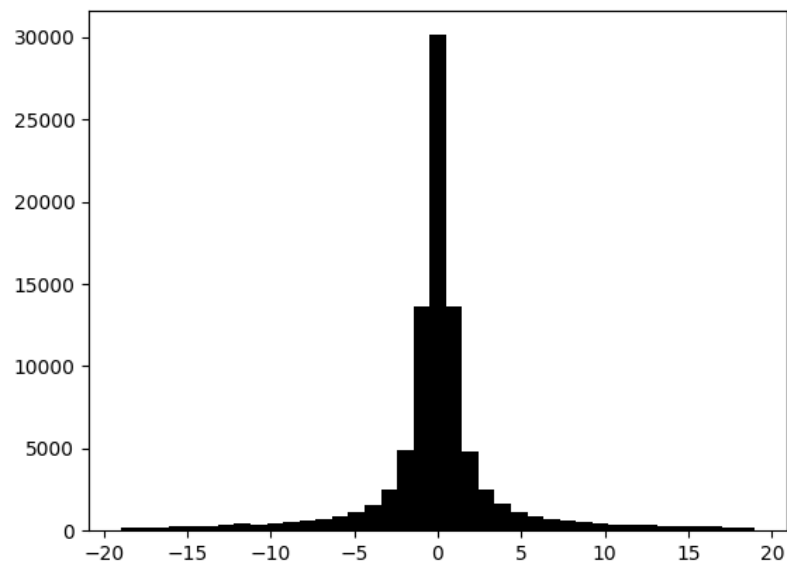


Рисунок 2.3 — Гистограмма ДКП коэффициентов jpeg изображения.

ГЛАВА 3

Компьютерные эксперименты

Исходный код компьютерных экспериментов отображен в ПРИЛОЖЕНИИ А.

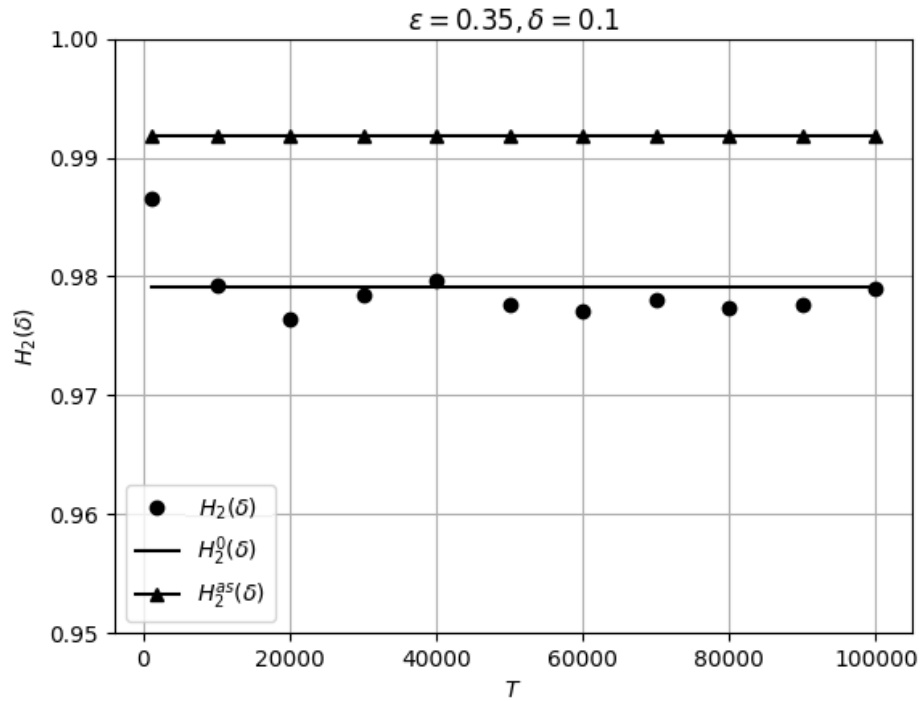


Рисунок 3.1 — График зависимости энтропии $H_2(\delta)$ от длины последовательности при $\delta = 0.1$

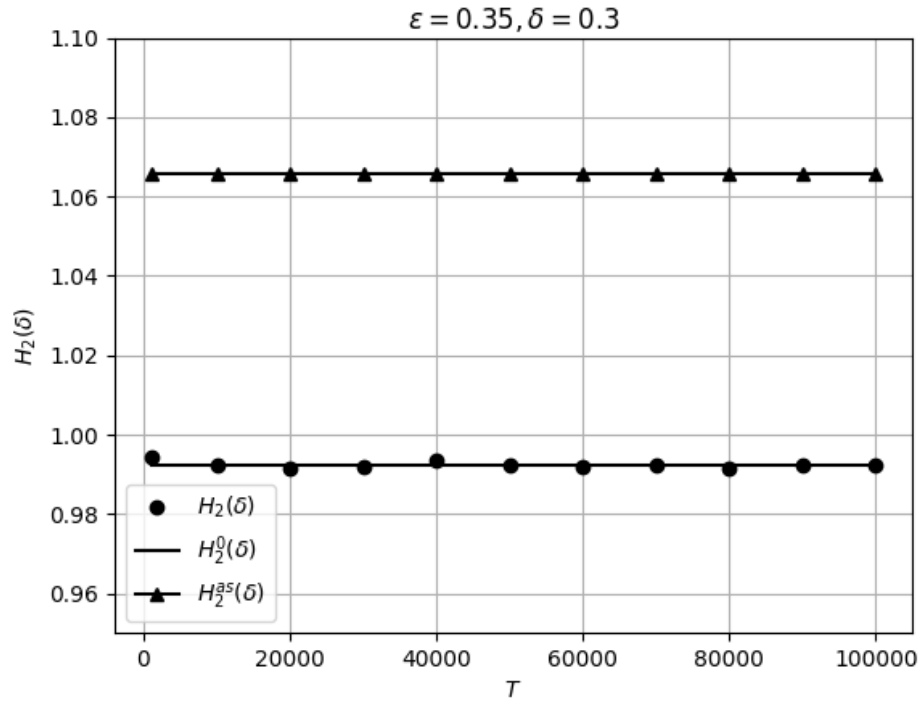


Рисунок 3.2 — График зависимости энтропии $H_2(\delta)$ от длины последовательности при $\delta = 0.3$

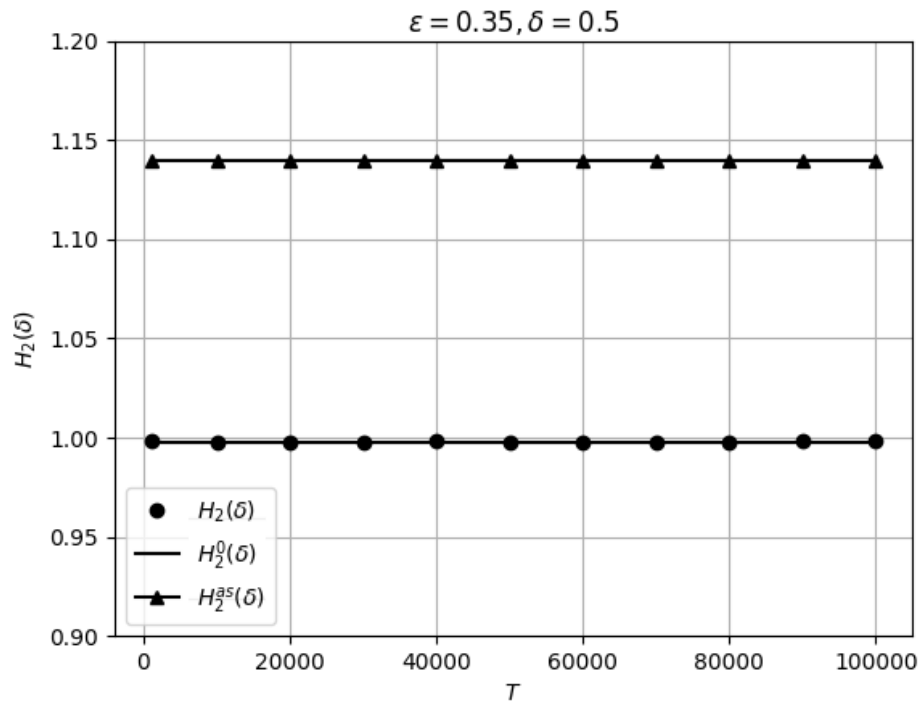


Рисунок 3.3 — График зависимости энтропии $H_2(\delta)$ от длины последовательности при $\delta = 0.5$

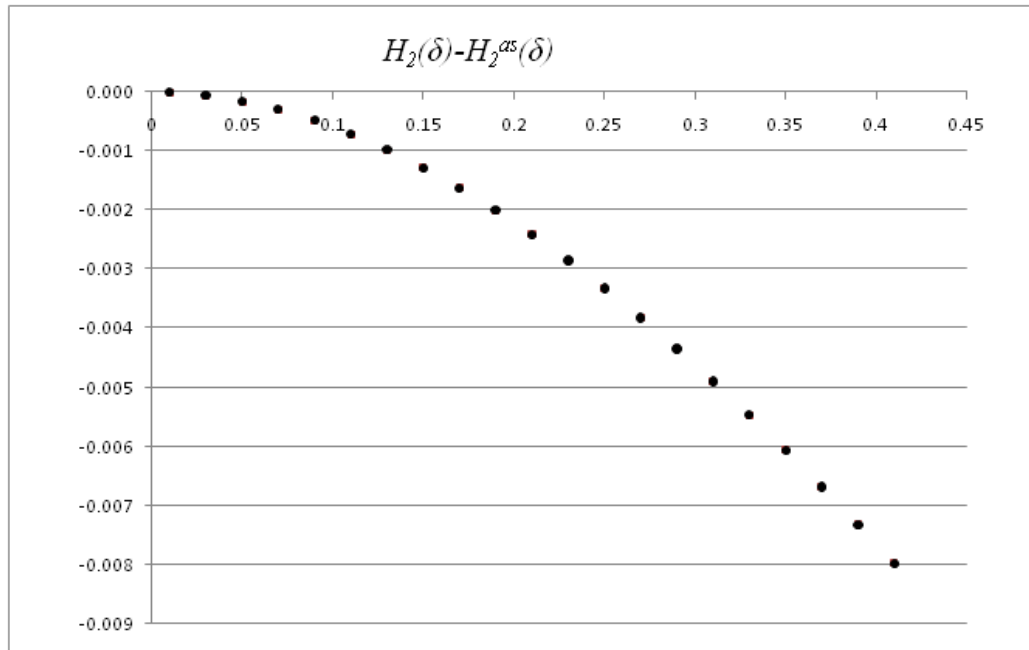


Рисунок 3.4 — График зависимости разности асимптотического и точного значений энтропии биграммы от доли вкрапления

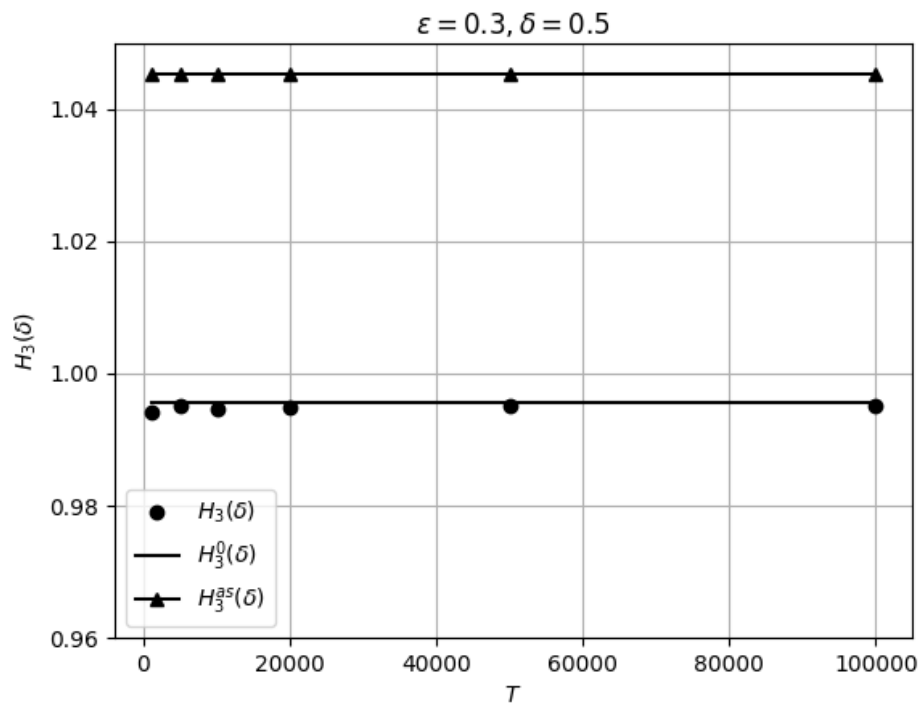


Рисунок 3.5 — График зависимости энтропии $H_3(\delta)$ от длины последовательности при $\delta = 0.5$

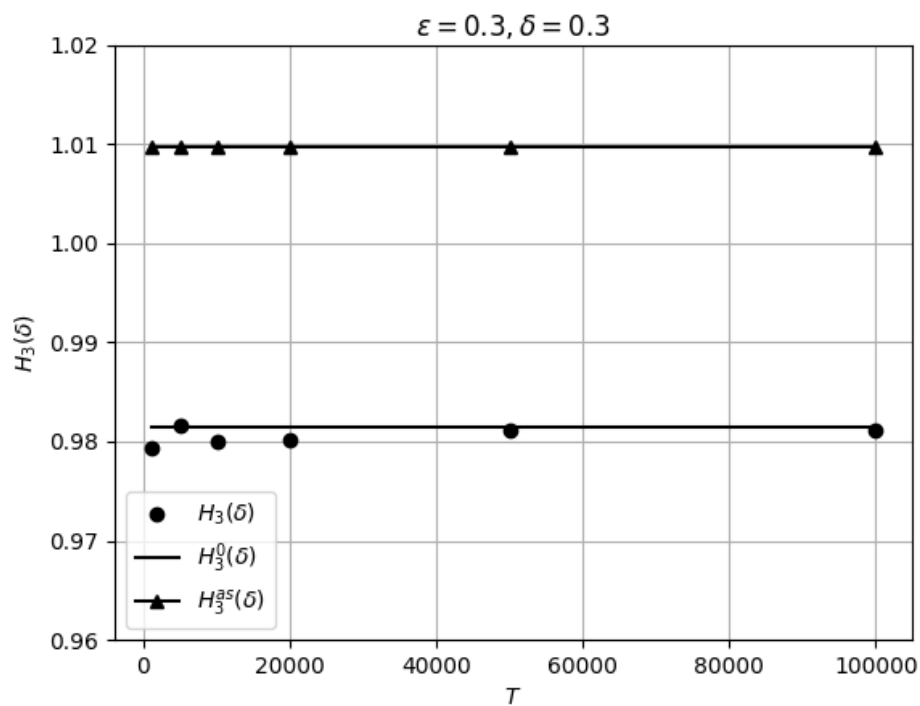


Рисунок 3.6 — График зависимости энтропии $H_3(\delta)$ от длины последовательности при $\delta = 0.3$

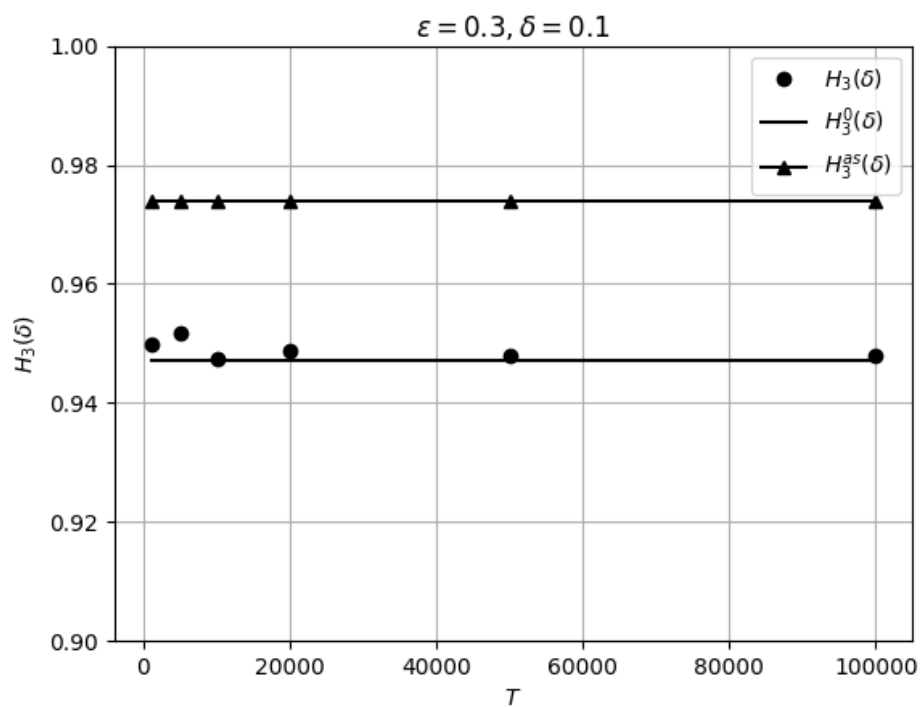


Рисунок 3.7 — График зависимости энтропии $H_3(\delta)$ от длины последовательности при $\delta = 0.1$

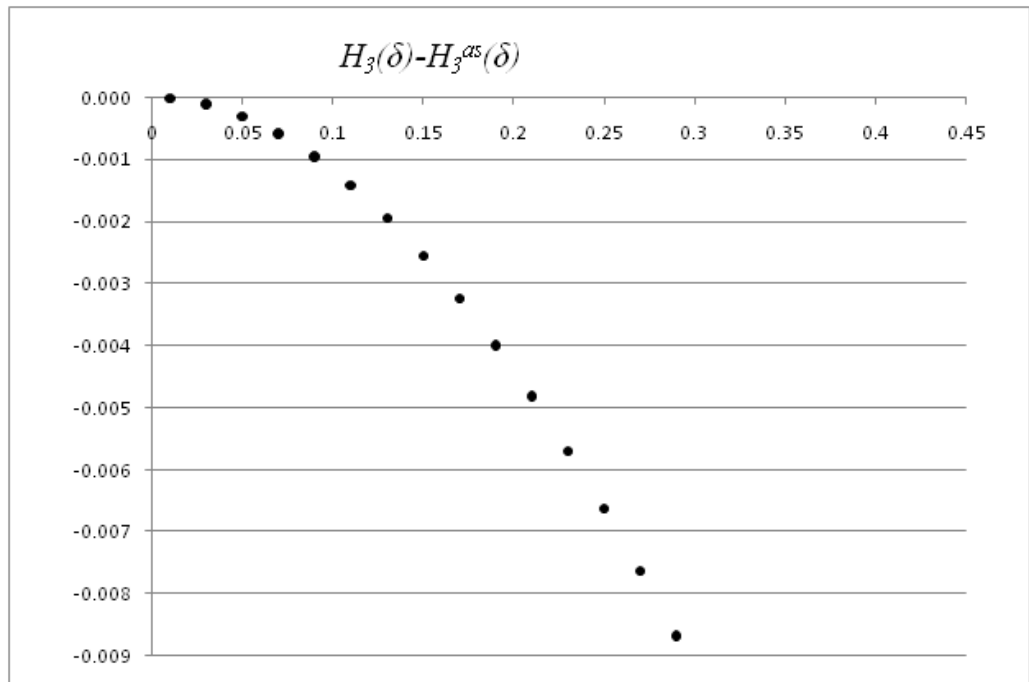


Рисунок 3.8 — График зависимости разности асимптотического и точного значений энтропии 3-граммы от доли вкрапления

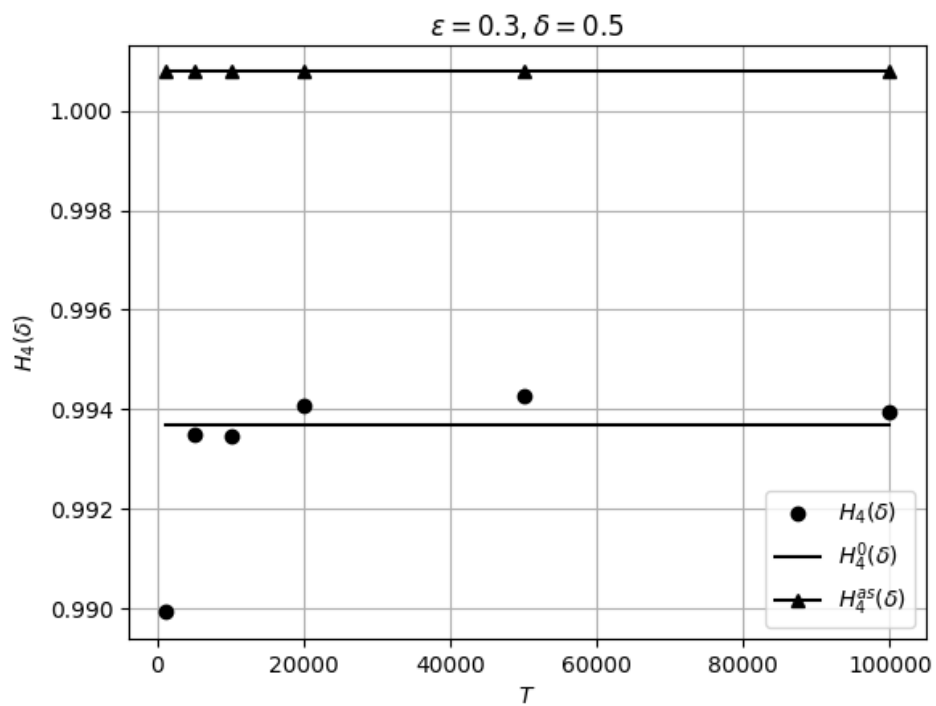


Рисунок 3.9 — График зависимости энтропии $H_4(\delta)$ от длины последовательности при $\delta = 0.5$

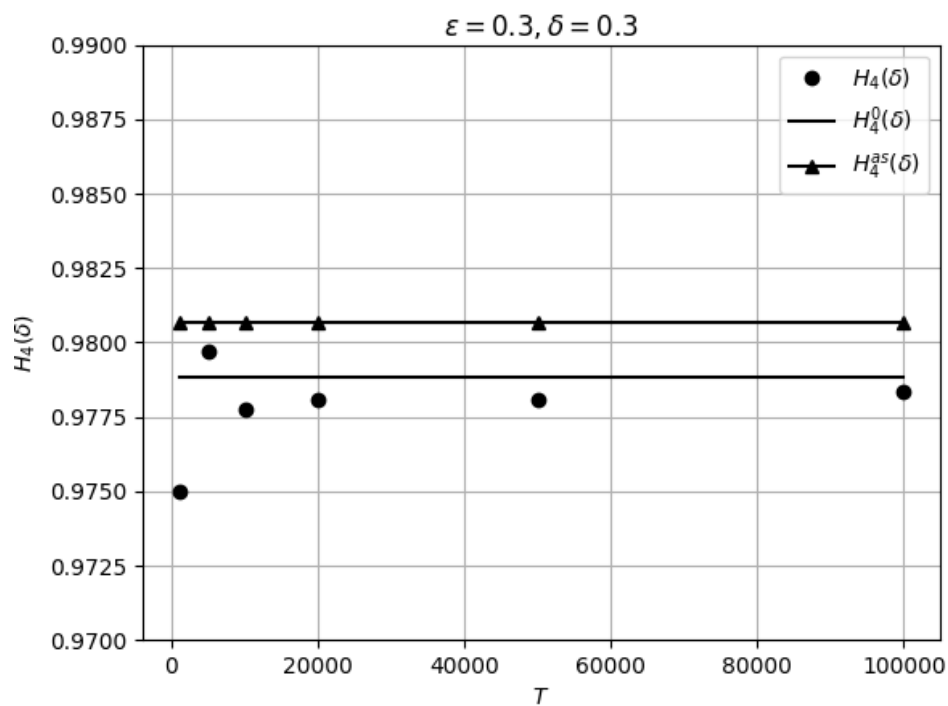


Рисунок 3.10 — График зависимости энтропии $H_4(\delta)$ от длины последовательности при $\delta = 0.3$

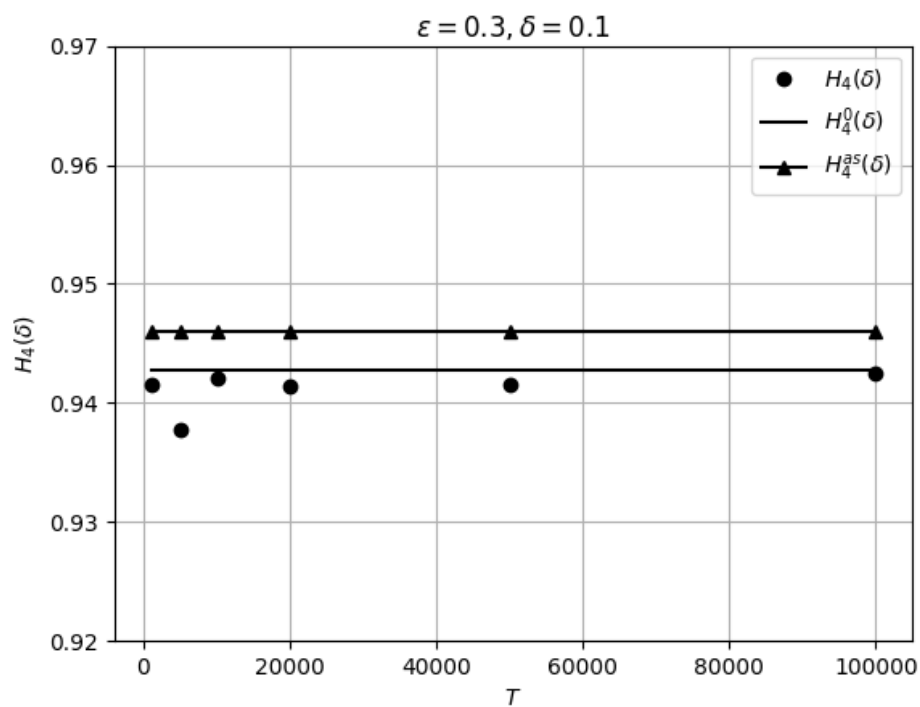


Рисунок 3.11 — График зависимости энтропии $H_4(\delta)$ от длины последовательности при $\delta = 0.1$

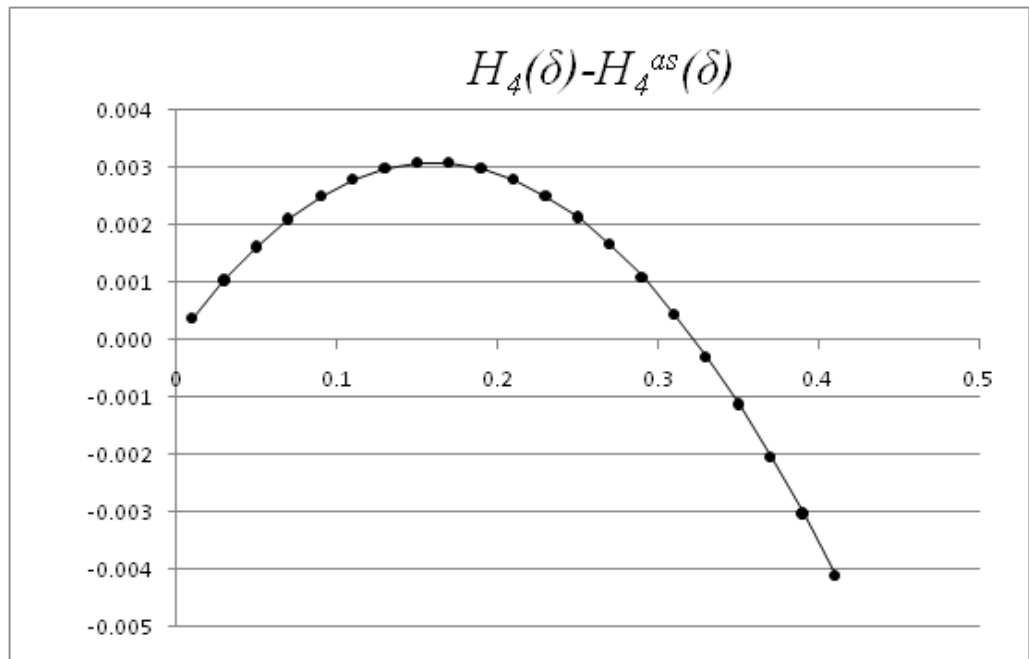


Рисунок 3.12 — График зависимости разности асимптотического и точного значений энтропии биграммы от доли вкрапления

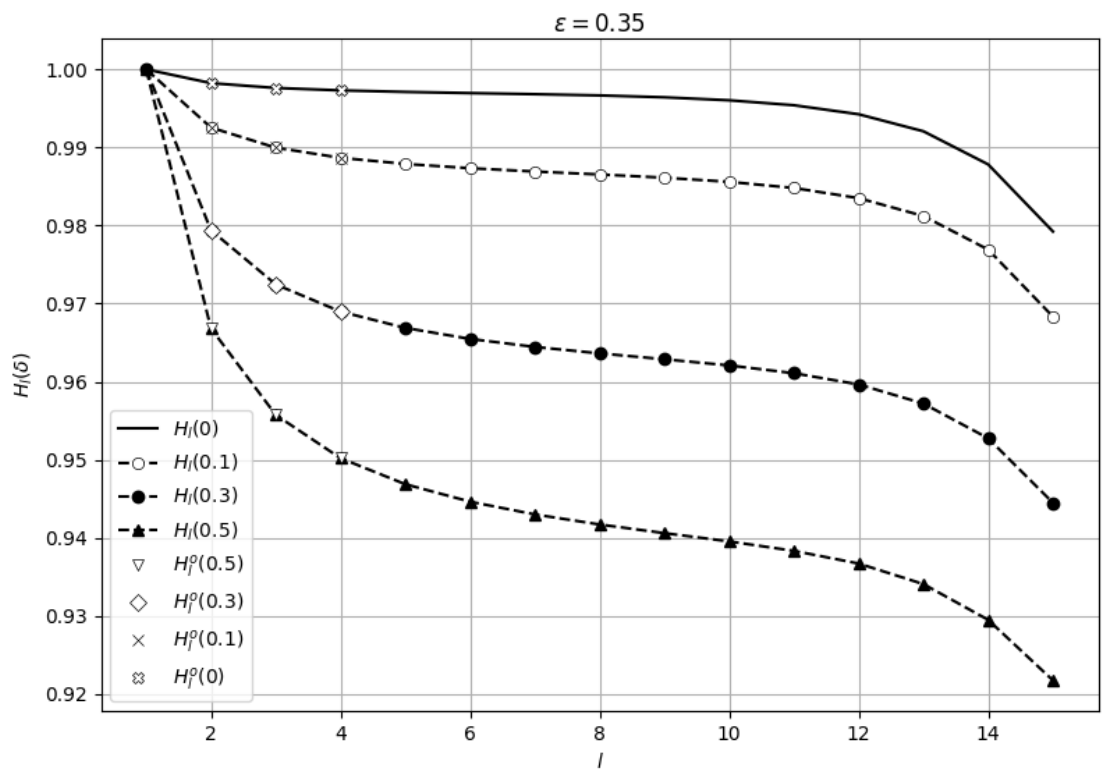


Рисунок 3.13 — Семейство графиков зависимости энтропии $H_l(\delta)$ от l при различных δ

ЗАКЛЮЧЕНИЕ

В работе получены следующие основные результаты:

1. Исследована математическая модель вкраплений в цепь Маркова 1-го порядка.
2. Получены точные значения вероятностей для всевозможных шаблонов l -граммы при $l=2,3,4$.
3. Получены асимптотические оценки первого порядка для энтропии l -граммы при $l=2,3,4$.
4. Проведен линейный дискриминантный анализ на основании асимптотических оценок для энтропии 3-граммы и 4-граммы.
5. Проведен ряд компьютерных экспериментов.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. R.O. Duda, P.E. Hart, D.H. Stork Pattern Classification (2nd ed.) // Wiley Interscience. — 2000.
2. А. А. Духин: Теория информации - М.: "Гелиос АРВ", 2007.
3. А.В. Аграновский, А. В. Балакин: Стеганография, цифровые водяные знаки о стегоанализе - М.: Вузовская книга, 2009.
4. В. Г. Грибунин, И. Н. Оков, И. В. Туринцев: Цифровая стеганография - М.: Солон-Прессб, 2002.
5. Дж. Миано: Форматы и алгоритмы сжатия изображений в действии - М.: Триумф, 2003.
6. К. И. Пономарев "Параметрическая модель вкрапления и ее статистический анализ", Дискрет. матем., 21:4 (2009), 148-157.
7. 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.
8. Ю. С. Харин [и др.]: Криптология - Минск: БГУ, 2013.
9. Ю. С. Харин, Е. В. Вечерко "Статистическое оценивание параметров модели вкраплений в двоичную цепь Маркова", Дискрет. матем., 25:2 (2013), 135-148.
10. Ю. С. Харин, Е. В. Вечерко "Распознавание вкраплений в двоичную цепь Маркова", Дискрет. матем., 27:3 (2015), 123–144.

Исходный код проведенных компьютерных экспериментов

А.1 Вычисление эмпирических оценок энтропии

```
public class Initializator {
    private static Random rand = new Random();

    private final double[] pi = {0.5, 0.5};
    private final double teta = 0.5;

    private double eps;
    private double[][] p = new double[2][2];
    private double delta;
    private int T;

    private int[] x;
    private int[] y;
    private int[] gamma;
    private int[] m;

    public Initializator(double eps, double delta) {
        this.eps = eps;
        double[][] p = {{0.5*(1.+eps), 0.5*(1 - eps)}, { 0.5*(1 - eps),
            0.5*(1.+eps)}};
        this.p = p;
        this.delta = delta;
        this.T = 100000;

        x = Markov(pi, p, T);
        m = binomial(teta, T);
        gamma = binomial(delta, T);
        y = func(T);
    }

    public Initializator(double eps, double delta, int _T) {
        this.eps = eps;
        double[][] p = {{0.5*(1.+eps), 0.5*(1 - eps)}, { 0.5*(1 - eps),
            0.5*(1.+eps)}};
        this.p = p;
        this.delta = delta;
        this.T = _T;

        x = Markov(pi, p, T);
        m = binomial(teta, T);
        gamma = binomial(delta, T);
        y = func(T);
    }
}
```

```

}

public Initializator() {

    eps = 0.12;
    p = new double[][]{{0.5*(1.+eps), 0.5*(1 - eps)}, { 0.5*(1 - eps),
        0.5*(1.+eps)}};
    delta = 0.01;
    T = 100000;

    x = Markov(pi, p, T);
    m = binomial(teta, T);
    gamma = binomial(delta, T);
    y = func(T);
}

public int[] getY() {
    return y;
}

public int getT() {
    return T;
}

int[] binomial(double p, int T) {
    int[] x = new int[T];
    double tmp = rand.nextDouble();
    for (int i = 0; i < T; i++) {
        if (tmp < p)
            x[i] = 1;
        else x[i] = 0;
        tmp = rand.nextDouble();
    }
    return x;
}

int[] Markov(double[] pi, double[][] probabilities, int size) {
    int[] mark = new int[size];
    mark[0] = (int) get(pi);
    for (int i = 1; i < size; ++i) {
        mark[i] = (int) get(probabilities[mark[i - 1]]);
    }
    return mark;
}

int get(double[] pi) {
    Random rnd = new Random();
    double probability = rnd.nextDouble();
    double sum = 0;
    for (int i = 0; i < pi.length - 1; ++i) {
        sum += pi[i];
        if (probability <= sum) {

```



```

return i;
}
}
return pi.length - 1;
}

public int[] func(int T) {
int[] y = new int[T];
for (int i = 0; i < T; ++i) {
int ind = 0;
for (int j = 0; j < i; ++j)
ind += gamma[j];
y[i] = ((1 - gamma[i]) * x[i] + gamma[i] * m[ind]);
}
return y;
}
}

public class Counter{
private int l;
private Initializator initializator;
private ArrayList<Integer> decimal;

public Counter(Initializator init, int l){
this.l = l;
this.initializator = init;
this.decimal = initDecimal(l);
}

public Counter(int l) {
this.l = l;
this.initializator = new Initializator();
this.decimal = initDecimal(l);
}
public Counter(int l, double eps, double delta) {
this.l = l;
this.initializator = new Initializator(eps, delta);
this.decimal = initDecimal(l);
}

public Initializator getInitializator() {
return initializator;
}

public ArrayList<Integer> getDecimal() {
return decimal;
}

public void setDecimal(ArrayList<Integer> decimal) {
this.decimal = decimal;
}
}

```

```

public ArrayList<Integer> initDecimal ( int l){
ArrayList<String> binaryWords =
    converterToBinaryWords(initializator.getY(), l);
decimal = convertToDecimal(binaryWords);
return decimal;
}

public double showEmpiricEntropy(int l) {
ArrayList<Long> frequencies = countFrequencies(decimal);
ArrayList<Double> probabilities = countProbabilities(frequencies,
    initializator.getT(), l);
return countEmpiricEntropy(probabilities, l);
}

public ArrayList<Integer> bitShift( ArrayList<Integer> decimal){
ArrayList<Integer> result = new ArrayList<Integer>();
for ( Integer decim : decimal){
result.add(decim >> 1);
}
return result;
}

private double countEmpiricEntropy(ArrayList<Double> probabilities,
    int l) {
double entropy = 0.;
for (Double prob : probabilities) {
entropy += nLogN(prob);
}
return entropy/l;
}

private ArrayList<Double> countProbabilities(ArrayList<Long>
    frequencies, int size, int l) {
ArrayList<Double> probabilities = new ArrayList<Double>();
for (Long freq : frequencies) {
probabilities.add(freq.doubleValue() /(double) (size-l+1));
}
return probabilities;
}

private ArrayList<Long> countFrequencies(ArrayList<Integer> list) {
Map<Integer, Long> counts =
list.stream().collect(Collectors.groupingBy(e -> e,
    Collectors.counting()));
return new ArrayList<Long>(counts.values());
}

public ArrayList<String> converterToBinaryWords(int[] x, int l) {
ArrayList<String> binaryWords = new ArrayList<String>();
StringBuffer stringBuffer = new StringBuffer();
for (int i : x){

```

```

        stringBuffer.append(i);
    }
    for (int i = 0; i < x.length - 1; ++i) {
        binaryWords.add(stringBuffer.substring(i, i+1));
    }
    return binaryWords;
}

private ArrayList<Integer> convertToDecimal(ArrayList<String>
        binaryWords) {
    ArrayList<Integer> decimal = new ArrayList<Integer>();
    for (String binaryWord : binaryWords) {
        decimal.add(Integer.parseInt(binaryWord, 2));
    }
    return decimal;
}

public static double nLogN(double x) {
    if (x == 0.)
        return 0;
    return -x * Math.log(x) / Math.log(2);
}
}

```

A.2 Линейный дискриминантный анализ

```

/**
 * Данные в формате:
 * число строк класса
 * x y class
 */
public class LinearDescrimination {
    private static int n;
    private static int nData;
    private static double[][] class1;
    private static double[][] class2;
    private static double[][] data;
    private static double[][] coefB;
    private static double A;

    public static int getN() {
        return n;
    }

    public static double[][] getClass1() {
        return class1;
    }

    public static double[][] getClass2() {
        return class2;
    }
}

```

```

public static double[][] getCoefB() {
return coefB;
}

public static double getA() {
return A;
}
/**
 * data in format
 * N
 * H3;H4;c
 */

static void readStudingData() {
String[] strs = null;
String str = null;
try {
BufferedReader br = new BufferedReader(new
    FileReader("LinearDescriminationData.csv"));
n = Integer.parseInt(br.readLine());
class1 = new double[n][2];
class2 = new double[n][2];
int y = 0;
for (int i = 0; i < 2 * n; i++) {
str = br.readLine();
strs = str.split(";");
y = Integer.parseInt(strs[2]);
if (y == 0) {
class1[i%n][0] = Double.parseDouble(strs[0]);
class1[i%n][1] = Double.parseDouble(strs[1]);
} else {
class2[i%n][0] = Double.parseDouble(strs[0]);
class2[i%n][1] = Double.parseDouble(strs[1]);
}
}
br.close();

} catch (FileNotFoundException e) {
e.printStackTrace();
} catch (IOException e) {
e.printStackTrace();
}
}

//data in format
/*
N
H3;H4
* */
static void readResearchingData() {
String[] strs = null;

```

```

String str = null;
try {
    BufferedReader br = new BufferedReader(new FileReader("Data.csv"));
    nData = Integer.parseInt(br.readLine());
    data = new double[nData][2];
    for (int i = 0; i < nData; i++) {
        str = br.readLine();
        strs = str.split(";");
        data[i][0] = Double.parseDouble(strs[0]);
        data[i][1] = Double.parseDouble(strs[1]);
    }
    br.close();

} catch (FileNotFoundException e) {
    e.printStackTrace();
} catch (IOException e) {
    e.printStackTrace();
}
}

static double[][] multiplyMatrix(double[][] A, double[][] B) {
    int aRows = A.length;
    int aColumns = A[0].length;
    int bRows = B.length;
    int bColumns = B[0].length;

    if (aColumns != bRows) {
        throw new IllegalArgumentException("A:Rows: " + aColumns + " did not
            match B:Columns " + bRows + ".");
    }

    double[][] C = new double[aRows][bColumns];
    for (int i = 0; i < aRows; i++) {
        for (int j = 0; j < bColumns; j++) {
            C[i][j] = 0.00000;
        }
    }

    for (int i = 0; i < aRows; i++) { // aRow
        for (int j = 0; j < bColumns; j++) { // bColumn
            for (int k = 0; k < aColumns; k++) { // aColumn
                C[i][j] += A[i][k] * B[k][j];
            }
        }
    }

    return C;
}

public static void print(double[][] matr) {
    int n = matr.length;
    int m = matr[0].length;

```

```

for (int i = 0; i < n; ++i) {
for (int j = 0; j < m; ++j) {
System.out.print(matr[i][j] + " ");
}
System.out.println();
}
}

public static double[][] transpose(double[][] a){
int n = a.length;
int m = a[0].length;
double[][] res = new double[m][n];
for( int i =0 ; i < n; ++i){
for ( int j = 0; j < m; ++j){
res[j][i] = a[i][j];
}
}
return res;
}

public static double[][] aMinusB(double[][] a, double[][] b){
int n = a.length;
int m = a[0].length;
double[][] res = new double[n][m];
for( int i =0 ; i < n; ++i){
for ( int j = 0; j < m; ++j){
res[i][j] = a[i][j]-b[i][j];
}
}
return res;
}

public static double[][] aPlusB(double[][] a, double[][] b){
int n = a.length;
int m = a[0].length;
double[][] res = new double[n][m];
for( int i =0 ; i < n; ++i){
for ( int j = 0; j < m; ++j){
res[i][j] = a[i][j]+b[i][j];
}
}
return res;
}

public static double[][] inverse(double[][] a){
double det = a[0][0] * a[1][1] - a[0][1]*a[1][0];
double[][] res = new double[2][2];
res[0][0] = a[1][1]/det;
res[0][1]= -a[0][1]/det;
res[1][0]= -a[1][0]/det;
res[1][1]= a[0][0]/det;
return res;
}

//значение дискриминантной функции. принимает столбец B
public static double countD(double[][] b, double x[][]){
return b[0][0]*x[0][0] + b[1][0]*x[0][1];
}

```

```
}
```

```
public static double countD(double [][] b, double [] x){  
return b[0][0]*x[0] + b[1][0]*x[1];  
}
```

```
//X без крышки
```

```
public static double [][] countC(double [][] x){  
double midX =0;  
double midY =0;  
int n = x.length;
```

```
for(int i = 0; i < n; ++i){  
midX+=x[i][0];  
midY+=x[i][1];  
}  
midX/=(double)n;  
midY/=(double)n;  
double [][] res = new double[n][2];  
for(int i = 0; i < n; ++i){  
res[i][0] = x[i][0] - midX;  
res[i][1] = x[i][1] - midY;  
}  
return res;  
}
```

```
//матрица ковариации
```

```
public static double [][] countCOV(double [][] x1, double [][] x2){
```

```
double [][] x1c=countC(x1);  
double [][] x2c=countC(x2);
```

```
double [][] res =  
    aPlusB(multiplyMatrix(transponate(x1c),x1c),multiplyMatrix(transponate(x2c),x2c));  
for (int i = 0; i < res.length; ++i){  
for ( int j = 0; j< res[0].length; ++j){  
res[i][j] /= x1.length + x2.length - 2.;  
}  
}  
return res;  
}
```

```
//вектор матожидания X с крышкой
```

```
public static double [][] countE(double [][] x){  
double midX =0;  
double midY =0;  
int n = x.length;
```

```
for(int i = 0; i < n; ++i){  
midX+=x[i][0];  
midY+=x[i][1];  
}  
midX/=(double)n;
```

```

midY/=(double)n;
double [][] res = new double[1][2];
res[0][0]=midX;
res[0][1]=midY;
return res;
}

public static void countCoef(){
double [][] cov = countCOV(class1, class2);
coefB = multiplyMatrix(inverse(cov),
    transponate(aMinusB(countE(class1), countE(class2))));
}

public static void countA(){
A =(countD(coefB, countE(class1)) + countD(coefB, countE(class2))) /
    2.;
}

public static void researchData() throws FileNotFoundException {

PrintWriter pw =new PrintWriter("descriminationRES.csv");
pw.println("=CYMM(A3:A103)");
pw.println("=CYMM(A104:A302)");
for ( int i = 0; i < data.length; ++i){
if(countD(coefB, data[i]) > A){
pw.println("1");
}else pw.println("2");
}
pw.flush();
pw.close();
}

public static void run() throws FileNotFoundException {
readStudingData();
double [][] cov = countCOV(class1,class2);
coefB = multiplyMatrix(inverse(cov),
    transponate(aMinusB(countE(class1), countE(class2))));
countA();

readResearchingData();
researchData();
}

public static void main(String[] args) {
try {
run();
} catch (FileNotFoundException e) {
e.printStackTrace();
}
}
}

```



```

public class LinearRegressionDataPreparator {

    private static int[] L = {3, 4};
    private static double[] eps = {0.15, 0.55};
    private static double[] delta = {0., 0.005, 0.01, 0.4, 0.5};
    private static int[] T = {10000};
    private static int M = 100;

    public static void main(String[] args) {
        empiricGraphics();
    }

    private static void empiricGraphics() {
        try {
            PrintWriter pw = new PrintWriter("LinearRegressionData2.csv");
            ArrayList<Double> entH3 = new ArrayList<Double>();
            ArrayList<Double> entH4 = new ArrayList<Double>();
            double entr = 0.;
            for (int l : L) {
                pw.println("L=" + l);
                for (double e : eps) {
                    for (double d : delta) {
                        for (int len : T) {
                            for (int i = 0; i < M; i++) {
                                Initializator init = new Initializator(e, d, len);
                                Counter counter = new Counter(init, 1);
                                entr = counter.showEmpiricEntropy(1);
                                if (l == 3){
                                    entH3.add(entr);
                                }else { entH4.add(entr);}
                                pw.println("T;" + len + ";eps;" + e + "; delta;" + d + ";entropy;" +
                                    entr);
                            }
                        }
                    }
                }
            }

            pw.flush();
            pw.close();
        } catch (FileNotFoundException e) {
            e.printStackTrace();
        }
    }
}

```