



О НЕКОТОРЫХ ПОДХОДАХ К СТАТИСТИЧЕСКОМУ ТЕСТИРОВАНИЮ КРИПТОГРАФИЧЕСКИХ ГЕНЕРАТОРОВ СЛУЧАЙНЫХ И ПСЕВДОСЛУЧАЙНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Харин Ю.С., Палуха В.Ю.

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



1. Введение
2. Проблемы статистического тестирования и существующие подходы к их разрешению
3. Подход к статистическому тестированию с использованием сложной нулевой гипотезы
 - 3.1. Некорректность модели простой нулевой гипотезы
 - 3.2. Простая нулевая гипотеза H_0
 - 3.3. Сложная нулевая гипотеза H_0^ε
 - 3.4. Критерий отношения правдоподобия для проверки сложных гипотез $H_0^\varepsilon, H_1^\varepsilon$
 - 3.5. Тест для проверки $H_0^\varepsilon, H_1^\varepsilon$ с использованием асимптотических разложений
 - 3.6. Вероятностные характеристики теста
 - 3.7. Численные результаты
 - 3.8. Подход к статистическому оцениванию уровней безопасности (стойкости) криптографических генераторов
4. Подход к тестированию на основе малопараметрических цепей Маркова порядка s (МПЦМ(s))
 - 4.1. Малопараметрические модели цепей Маркова порядка s для выходных последовательностей криптографических генераторов
 - 4.2. Способы построения малопараметрических моделей
 - 4.3. Цепь Маркова s -го порядка с r частичными связями ЦМ(s, r)
 - 4.4. МПЦМ(s) на основе стандартных дискретных распределений вероятностей
 - 4.5. МПЦМ(s) с использованием искусственных нейронных сетей
 - 4.6. МПЦМ(s) на основе экспоненциальных семейств и достаточных статистик
5. Подход к статистическому тестированию на основе энтропийных профилей
 - 5.1. Функционалы энтропии и их статистические оценки
 - 5.2. Асимптотические свойства статистических оценок
 - 5.3. Тесты на основе энтропийных профилей
 - 5.4. Программный Комплекс статистического тестирования «ЭАДП»
6. Компьютерные эксперименты по оценке близости выходных последовательностей физических и программных генераторов к равномерно распределенной последовательности
7. Заключение
8. Список литературы



1. Введение

Наряду с разработкой государственных стандартов криптографической защиты информации актуальным направлением криптологии является разработка и анализ качества криптографических генераторов случайных и псевдослучайных последовательностей [1, 2]. Генерация равномерно распределенной случайной последовательности (РРСП) представляет собой важную задачу не только для генерации «гаммы» в поточных криптосистемах, но и в других системах криптографической защиты информации для выработки ключей и векторов инициализации, одноразовых чисел (Nonce), в криптографических протоколах и схемах ЭЦП.

В криптографии генерация случайности сводится к задаче генерации равномерно распределенной случайной последовательности (РРСП) $x_t \in \{0, 1\}$, $t = 1, 2, \dots$, для которой при любом $s = 1, 2, \dots$ s -мерное распределение вероятностей является равномерным: $\mathbf{P}\{x_1 = j_1, x_2 = j_2, \dots, x_s = j_s\} = 2^{-s}$. Для анализа качества выходных последовательностей длины T криптографических генераторов используются наборы (так называемые «батареи») статистических тестов проверки нулевой гипотезы $H_0 = \{x_t \text{ есть РРСП}\}$ против альтернативы $H_1 = \overline{H_0}$ [1 – 3].



2. Проблемы статистического тестирования и существующие подходы к их разрешению

Проведенный в [2] обзор существующих тестов выявляет следующие недостатки: 1) многие из тестов ориентированы на проверку лишь частных случаев альтернативы $\overline{H}_0$; 2) многие тесты не фиксируют семейство альтернатив и не имеют оценок мощности w_T ; 3) при включении нескольких тестов в «батарею» не удастся учесть вероятностную зависимость тестовых статистик; 4) известны примеры генераторов псевдослучайных последовательностей, «уверенно» преодолевающих батареи тестов [4]. Особенно значимой критике подвергнута «батарея» тестов NIST STS, что привело к решению Национального института США по стандартизации о начале «ревизии» «батареи тестов» [5]. Для преодоления этих недостатков криптографическое сообщество предлагает определенные подходы к статистическому тестированию. В [6] предлагается подход, основанный на теоретическом исследовании вероятностной зависимости тестовых статистик, входящих в батарею тестов. В [6] предлагается подход, основанный на двухэтапной проверке гипотез H_0 , $\overline{H}_0$, включая на 2-ом этапе проверку соответствия стандартному равномерному закону распределения P -значений, вычисляемых на 1-ом этапе.

Для преодоления указанных недостатков в докладе предлагается еще 3 подхода к статистическому тестированию криптографических генераторов: подход с использованием сложной нулевой гипотезы; подход на основе сложных марковских моделей; подход на основе энтропийных профилей.



3. Подход к статистическому тестированию с использованием сложной нулевой гипотезы

3.1. Некорректность модели простой нулевой гипотезы

В дополнение к указанным выше недостаткам существующих статистических тестов выделим **еще один их общий существенный недостаток**: некорректность модели простой нулевой гипотезы H_0 и необходимость использования сложной нулевой гипотезы при статистическом тестировании [8].

Этот недостаток проявляется в следующем парадоксе: при увеличении длины $T \rightarrow +\infty$ наблюдаемых последовательностей, порождаемых реально существующими генераторами случайности, вероятность принятия альтернативы $\overline{H_0}$ стремится к единице: $w_T \rightarrow 1$. Иначе говоря, «увеличением длины исследуемой последовательности T можно забраковать любой реальный генератор случайности». Объяснение этого «парадокса» состоит в следующем. Реальные генераторы случайности – не идеальны, они отклоняются от гипотезы H_0 по своему вероятностному распределению на некоторую величину $\varepsilon > 0$. Все тесты, включаемые в батареи тестов, как известно, обладают оптимальным свойством проверки простой гипотезы H_0 против сложной альтернативы $\overline{H_0}$ – свойством состоятельности: $w_T \rightarrow 1$ при $T \rightarrow +\infty$. Таким образом, этот «парадокс» порожден некорректностью математической модели нулевой гипотезы H_0 . Для того, чтобы избежать этого «парадокса» нулевая гипотеза должна быть сложной: она должна задавать величину допуска $\varepsilon > 0$ отклонений от простой гипотезы H_0 .



3.2. Простая нулевая гипотеза H_0

Пусть наблюдается двоичная случайная последовательность длины $T = n \cdot s$, разбитая на n последовательных s -грамм ($s \in N$):

$$X = X_1^T = (x_1, \dots, x_T) \equiv (X_1^s, X_{s+1}^{2s}, \dots, X_{(n-1)s+1}^T) \in V^T, V = \{0, 1\}; \quad (1)$$

эти s -граммы независимы в совокупности и одинаково распределены с некоторым распределением $p = (p_{J_1^s}) \in \mathcal{P}$:

$$\mathbf{P}\{X_{(i-1)s+1}^{is} = J_1^s\} ::= \mathbf{P}\{x_{(i-1)s+1} = j_1, \dots, x_{is} = j_s\} = p_{J_1^s}, J_1^s = (j_k) \in V^s, i = 1, \dots, n, \quad (2)$$

$$\mathcal{P} = \left\{ p = (p_{J_1^s}) : p_{J_1^s} \geq 0, J_1^s \in V^s, \sum_{J_1^s \in V^s} p_{J_1^s} = 1 \right\} - \quad (3)$$

семейство (симплекс) всевозможных s -мерных вероятностных распределений на V^s . Для упрощения обозначений (1) – (3) перейдем от мультииндекса J_1^s к одномерному индексу $k \in \{0, 1, \dots, 2^s - 1\}$:

$$k = \langle J_1^s \rangle ::= \sum_{i=1}^s j_i \cdot 2^{i-1} \in \{0, 1, \dots, K-1\}, K = 2^s;$$

$$J_1^s = (j_1, \dots, j_s) = \langle k \rangle, j_{s-i+1} = \left[\left(k - \sum_{l=s-i+2}^s j_l \cdot 2^{l-1} \right) / 2^{s-i} \right], i = 1, 2, \dots, s.$$

В существующих «батареях тестов» [3, 4], подвергнутых многочисленным модификациям и критике [4 – 8], используются тесты статистической проверки простой нулевой гипотезы

$$H_0 = \{p = p^{(0)}\}, p^{(0)} = (p_k^{(0)}), p_k^{(0)} \equiv \frac{1}{K}, \quad (4)$$

против сложной альтернативы $\overline{H_0}$, приводящие к указанному выше парадоксу.



3.3. Сложная нулевая гипотеза H_0^ε

Во избежание парадокса «неадекватности модели (4) простой нулевой гипотезы» введем в рассмотрение сложную нулевую гипотезу об s -мерной равномерности:

$$H_0^\varepsilon = \left\{ p = (p_k) \in \mathcal{P}_0^\varepsilon \right\}, \quad \mathcal{P}_0^\varepsilon = \left\{ p \in \mathcal{P} : \|p - p^{(0)}\| = \sqrt{\sum_{k=0}^{K-1} \left(p_k - \frac{1}{K} \right)^2} \leq \varepsilon \right\} - \quad (5)$$

пересечение симплекса $\mathcal{P}$ с гипершаром в R^K заданного радиуса ε с центром в точке p_* s -мерного равномерного распределения; $\varepsilon \in (0, \sqrt{1-1/K})$ – достаточно малый параметр нулевой гипотезы (5), определяющий допустимые отклонения от простой гипотезы H_0 (рис. 1). Гипотеза H_0 является предельной по отношению к $H_0^\varepsilon : H_0^\varepsilon \xrightarrow{\varepsilon \rightarrow 0} H_0$. Альтернатива также является сложной $H_1^\varepsilon = \overline{H_0^\varepsilon} = \{ p \in \mathcal{P} : \|p - p^{(0)}\| > \varepsilon \}$.

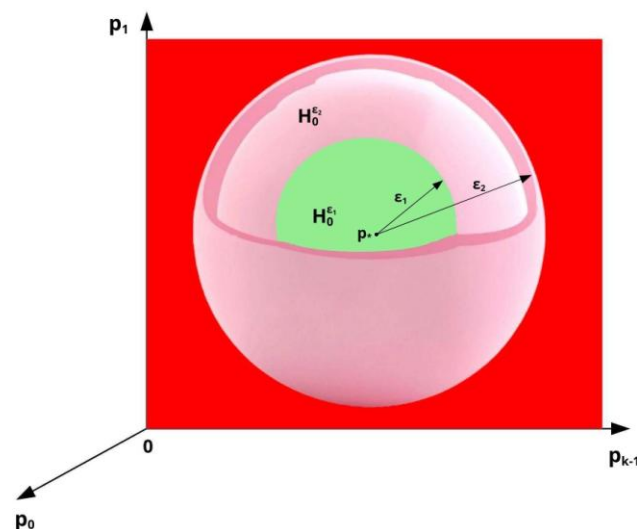


Рис. 1. Иллюстрация простой H_0 и двух сложных гипотез $H_0^{\varepsilon_1}, H_0^{\varepsilon_2}$



3.4. Критерий отношения правдоподобия для проверки сложных гипотез $H_0^\varepsilon, H_1^\varepsilon$

Построим критерий отношения правдоподобия по наблюдаемой последовательности $X = X_1^T$ для проверки сложной нулевой гипотезы H_0^ε против сложной альтернативы H_1^ε .

Примем следующие обозначения: $\mathbf{1}\{C\}$ – индикатор события C , $\mathbf{1}_N$ – вектор-столбец, все N компонент которого равны единице;

$$\hat{p} = (\hat{p}_k), \hat{p}_k = \frac{\nu_k}{n}, \nu_k = \sum_{i=1}^n \mathbf{1}\{\langle X_{(i-1)s+1}^{is} \rangle = k\} \geq 0, k = 0, 1, \dots, K-1, - \quad (6)$$

несмещенная, строго состоятельная, асимптотически нормальная и эффективная оценка максимального правдоподобия для p ; ν_k – частота встречаемости s -граммы $\langle k \rangle$ в последовательности n фрагментов $X_1^s, X_{s+1}^{2s}, \dots, X_{(n-1)s+1}^T$

$$l(X; p) = \sum_{k=0}^{K-1} \nu_k \ln p_k = n \sum_{k=0}^{K-1} \hat{p}_k \ln p_k - \quad (7)$$

логарифмическая функция правдоподобия (ЛФП) для наблюдаемой последовательности $X = X_1^T$;

$$S_\varepsilon = p \in \mathcal{P} : \|p - p^{(0)}\| = \varepsilon \subset \mathcal{P}_0^\varepsilon - \quad (8)$$

пересечение гиперсферы радиуса ε с центром в точке $p^{(0)} = \frac{1}{K} \mathbf{1}_K$ с симплексом $\mathcal{P}$;

$$l_{S_\varepsilon}^*(X) = \max_{p \in S_\varepsilon} l(X; p) - \quad (9)$$

максимальное значение ЛФП на гиперсфере S_ε , определяемой (8).



3.5. Тест для проверки $H_0^\varepsilon, H_1^\varepsilon$ с использованием асимптотических разложений

Логарифмическая статистика отношения правдоподобия при проверке сложных гипотез для рассматриваемой модели наблюдений (1) – (5) имеет вид [9]:

$$\lambda(X) ::= \sup_{p \in \mathcal{P}_0^\varepsilon} l(X; p) - \sup_{p \in \mathcal{P}} l(X; p) \leq 0,$$

$$\lambda(X) = \begin{cases} l_{S_\varepsilon}^*(X) - n \sum_{k=0}^{K-1} \hat{p}_k \ln \hat{p}_k, & \text{если } \|p - p^{(0)}\| > \varepsilon, \\ 0, & \text{если } \|\hat{p} - p^{(0)}\| \leq \varepsilon. \end{cases} \quad (10)$$

Семейство всевозможных решающих правил (статистических тестов), основанное на логарифмической статистике отношения правдоподобия (10), определяется порогом (критическим значением) $\Delta_n \leq 0$:

$$d = d(X) = \begin{cases} 0, & \lambda(X) \geq \Delta_n, \\ 1, & \lambda(X) < \Delta_n. \end{cases} \quad (11)$$

Основная трудность в реализации и теоретическом анализе решающего правила (10), (11) состоит в вычислении статистики (9), сводящемся к отысканию условного максимума $p = (p_k)$ по логарифмической функции правдоподобия (7) на гиперсфере (8) при дополнительных ограничениях: $\sum_{k=0}^{K-1} p_k = 1, p_k \geq 0$. В [8] доказано, что при $\varepsilon \rightarrow 0$ удастся получить асимптотические разложения статистики $l_{S_\varepsilon}^*(X)$ с погрешностью $O(\varepsilon^2)$ и $O(\varepsilon^3)$, в частности, согласно Следствию 2 [8]:

$$\lambda(X) = (nH(\hat{p}) - \ln K + \varepsilon K \|p - p^{(0)}\|) + O(\varepsilon^2), \quad (12)$$

где $H(\hat{p}) = -\sum_{k=0}^{K-1} \hat{p}_k \ln \hat{p}_k \geq 0$ – подстановочная оценка энтропии Шеннона s-граммы, а $\hat{p} = (\hat{p}_k)$ определяется (6).



3.6. Вероятностные характеристики теста

Теорема 1. Если имеет место математическая модель двоичной последовательности (1) – (5), то при использовании аппроксимации (12) и $n \rightarrow +\infty$ заданный асимптотический размер $0 < \alpha_0 < 1/2$ достигается для теста

$$d = d(X) = \begin{cases} 0, & \|p - p^{(0)}\| \leq \gamma_n, \\ 1, & \|p - p^{(0)}\| > \gamma_n, \end{cases} \quad (13)$$

$$\gamma_n = \varepsilon + \frac{\Phi^{-1}(1 - \alpha_0)}{\sqrt{n}} \sqrt{1/K + \varepsilon(e_K^+ - \varepsilon)}, \quad e_K^+ = \frac{1 - 2/K}{\sqrt{1 - 1/K}} \leq 1,$$

где $\Phi^{-1}(\cdot)$ – квантильная функция стандартного нормального закона $N_1(0,1)$.

Теорема 2. В условиях Теоремы 1 при $n \rightarrow +\infty$ для частной альтернативы $h^{\varepsilon_+} ::= \{p \in S_{\varepsilon_+}\} \subset H_1^{\varepsilon}$ и фиксированных $\varepsilon, \varepsilon_+$, таких что $\varepsilon_+ > \varepsilon$, справедливо следующее асимптотическое приближение мощности $w(\varepsilon_+)$ решающего правила (13):

$$w(\varepsilon_+) = \Phi \left(\sqrt{\frac{n}{K}} \frac{K(\varepsilon_+ - \varepsilon)}{\sqrt{1 + K\varepsilon_+(e_K - \varepsilon_+)}} - \sqrt{\frac{1 + K\varepsilon(e_K^+ - \varepsilon)}{1 + K\varepsilon_+(e_K - \varepsilon_+)}} \Phi^{-1}(1 - \alpha_0) \right) \quad (14)$$

где e_K^+ определено в (13), а $e_K = \sum_{k=0}^{K-1} (p_k - p_k^{(0)})^3 / \varepsilon^3, |e_K| \leq e_K^+$.

Следствие 1. Выполняются свойство состоятельности решающего правила (13):

$$w(\varepsilon_+) \xrightarrow{n \rightarrow +\infty} 1, \quad \forall \varepsilon_+ > \varepsilon,$$

и свойство монотонности его мощности:

$$w(\varepsilon_+) \geq w(\varepsilon) = \alpha_0, \quad \forall \varepsilon_+ \geq \varepsilon.$$



3.7. Численные результаты

Для иллюстрации работоспособности построенного теста при $s = 8$ (тестирование байтовой равномерности), $K = 2^s = 256$, $\varepsilon = 0.9/K = 0.00352$ имитировалось по 100 реализаций двоичной последовательности (1) – (5) для $T \in \{2^{19}, 2^{20}\}$, $\alpha_0 = 0.05$. Графики зависимости оценки вероятности принятия альтернативы $\hat{P}_1$ от $\|p - p^{(0)}\|$ представлены на рис. 2, 3 соответственно.

Рис. 2, 3 иллюстрируют состоятельность построенного решающего правила (13). Видно также, что при увеличении длительности T последовательности X график зависимости $\hat{P}_1$ от $\|p - p^{(0)}\|$ приближается к ступенчатой функции со ступенькой в точке $\|p - p^{(0)}\| = \varepsilon$: при $\|p - p^{(0)}\| < \varepsilon$ величина $\hat{P}_1 < \alpha_0$, а при $\|p - p^{(0)}\| > \varepsilon$, величина $\hat{P}_1$ приближается к единичному значению.

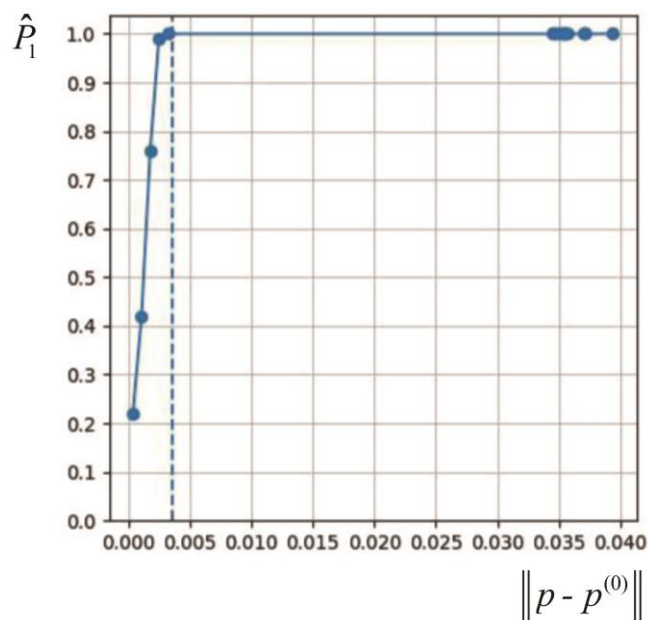


Рис. 2. Случай $T = 2^{19}$

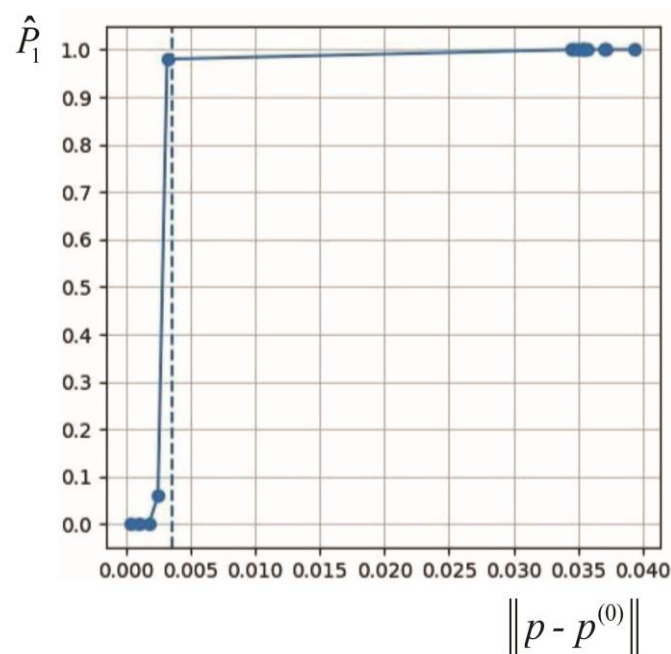


Рис. 3. Случай $T = 2^{20}$



3.8. Подход к статистическому оцениванию уровней безопасности (стойкости) криптографических генераторов

Как видно из слайда 5 и известно из практики, существующие криптографические генераторы случайных и псевдослучайных последовательностей различаются по уровню ε отклонения от «истинной случайности». Это позволяет ввести в рассмотрение различные уровни безопасности криптографических генераторов: чем больше ε , тем ниже уровень безопасности. Построенный статистический тест (13) с учетом установленных в Теоремах 1, 2 его вероятностных свойств можно применить для статистического оценивания уровней безопасности (стойкости) исследуемого криптографического генератора.

Проиллюстрируем этот подход в ситуации, когда заданы $M = 3$ уровня безопасности (рис. 4). Здесь $0 < \varepsilon_1 < \varepsilon_2 < 1$ – два заданных критических значения допустимых отклонений $\|p - p^{(0)}\|$ распределения вероятностей s -грамм $p = (p_k)$ от равномерного распределения $p^{(0)} = (p_k^{(0)})$. Определим вложенные сложные гипотезы $H_0^{\varepsilon_1} \subset H_0^{\varepsilon_2}$. С помощью статистического теста (13) по X_1^T последовательно проверяем пары гипотез $(H_0^{\varepsilon_1}, \overline{H_0^{\varepsilon_1}})$ и $(H_0^{\varepsilon_2}, \overline{H_0^{\varepsilon_2}})$. Анализируя полученные решения, находим оценки для уровня безопасности: генератор имеет уровень безопасности 2, если $H_0^{\varepsilon_1}$ верна; генератор имеет уровень безопасности 1, если $H_0^{\varepsilon_2}$ верна, но $H_0^{\varepsilon_1}$ не верна; если $H_0^{\varepsilon_2}$ не верна, то генератор бракуется (уровень безопасности 0).

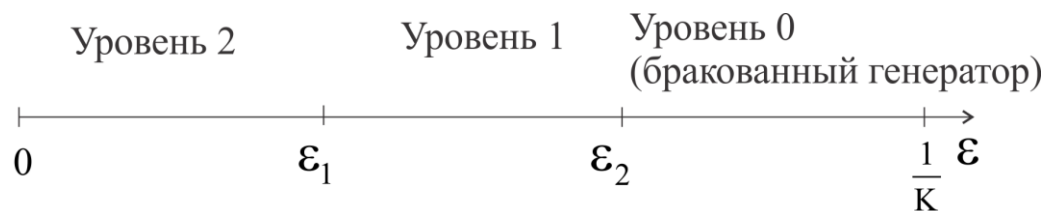


Рис. 4. Уровни безопасности генератора



4. Подход к тестированию на основе малопараметрических цепей Маркова порядка s (МПЦМ(s))

4.1. Малопараметрические модели цепей Маркова порядка s для выходных последовательностей криптографических генераторов

Выходные последовательности x_t криптографических генераторов при альтернативах H_1 , встречающихся на практике, могут иметь стохастические зависимости большой глубины между элементами последовательности. Особенно часта такая ситуация для **генераторов псевдослучайных последовательностей**. Универсальная модель – предложенная Дж. Дубом однородная цепь Маркова (MC(s)) достаточно большого порядка $s \in N$ на вероятностном пространстве $(\Omega, F, \mathbf{P})$, определяемая порождающим уравнением для условного распределения вероятностей:

$$\mathbf{P}\{x_t = j_t \mid F_{t-1}\} = \mathbf{P}\{x_t = j_t \mid X_{t-s}^{t-1} = J_{t-s}^{t-1}\} = p_{J_{t-s}^{t-1}, j_t}, \quad t \in Z,$$

где $F_{t-1} = \sigma(\{x_\tau : \tau \leq t-1\})$ – σ -алгебра, $X_{t-s}^{t-1} = (x_{t-s}, \dots, x_{t-1})' \in V^s$, $J_{t-s}^{t-1} = (j_{t-s}, \dots, j_{t-1})' \in V^s$, $V = \{0, 1\}$. Эта модель определяется матричным параметром – $(s+1)$ -мерной матрицей переходных вероятностей $P = (p_{J_1^{s+1}})$, $J_1^{s+1} \in V^{s+1}$. Для ЦМ(s) число независимых параметров $D_{\text{ЦМ}(s)} = 2^s$ увеличивается экспоненциально с ростом порядка s . Для преодоления экспоненциальной сложности этой универсальной модели мы предлагаем для аппроксимации цепей Маркова высокого порядка использовать малопараметрические модели, для которых матрица P имеет следующий параметрический вид [10]:

$$p_{J_1^{s+1}} =: p_\alpha(J_1^{s+1}), \quad \alpha = (\alpha_1, \dots, \alpha_d)' \in R^d,$$

где α – векторный параметр некоторой малой размерности $d \ll D_{\text{ЦМ}(s)}$. При этом семейство параметрических моделей должно содержать РРСП при некотором значении параметра $\alpha^{(0)}$, а при $\alpha \neq \alpha^{(0)}$ адекватно описывать встречающиеся на практике варианты альтернативы H_1 . Тест для проверки гипотез H_0 , H_1 основывается на статистике $\|\hat{\alpha} - \alpha^{(0)}\|$, где $\hat{\alpha}$ – состоятельная статистическая оценка параметра малопараметрической модели по наблюдаемой выходной последовательности криптографического генератора.



4.2. Способы построения малопараметрических моделей

Предложены четыре основных способа построения малопараметрических цепей Маркова высокого порядка s (МПЦМ(s)) [11 – 15]:

- 1) сжатие множества возможных значений элементов матрицы P ;
- 2) использование параметрических семейств стандартных дискретных распределений вероятностей;
- 3) использование искусственных нейронных сетей для аппроксимации зависимости от предыстории;
- 4) подход на основе достаточных статистик и информационной геометрии.

На основе этих подходов удалось построить следующие новые малопараметрические вероятностные модели, применимые для двоичных временных рядов $x_t \in V$:

- цепь Маркова порядка s с r частичными связями MC(s, r) [11];
- цепь Маркова условного порядка MCCO(s, L);
- двоичная условно нелинейная авторегрессионная модель BCNAR(s) [12];
- биномиальная условно нелинейная авторегрессионная модель BiCNAR(s) [13];
- семибиномиальная условно нелинейная авторегрессионная модель SBiCNAR(s);
- нейросетевая модель [14];
- семейство малопараметрических моделей на основе экспоненциальных семейств вероятностных распределений и достаточных статистик [15].

Разработан [12] метод FBE (Frequencies Based Estimation) статистического оценивания параметров α малопараметрических моделей на основе условных многомерных частот, позволяющий строить состоятельные статистические оценки параметров $\hat{\alpha}$ в явном виде. Доказана состоятельность и асимптотическая нормальность построенных статистических оценок $\hat{\alpha}$ при возрастании длины T наблюдаемого дискретного временного ряда. Количественная оценка стойкости криптографического генератора определяется по величине уклонения матрицы $P_{\hat{\alpha}}$ от матрицы P^0 для РПСР.

Представим некоторые МПЦМ(s) и методы статистического оценивания параметров α по исследуемой двоичной последовательности $X_1^n = (x_1, \dots, x_n)' \in V^n$.



4.3. Цепь Маркова s -го порядка с r частичными связями ЦМ(s, r)

Матрица переходов P для ЦМ(s, r) имеет следующий параметрический вид [11]:

$$P_{j_1, \dots, j_s, j_{s+1}} = q_{j_{m_1^0}, \dots, j_{m_r^0}, j_{s+1}}, \quad J_1^{s+1} \in V^{s+1}, \quad (15)$$

где $r \in \{1, 2, \dots, s\}$ – число связей; $M_r^0 = (m_1^0, m_2^0, \dots, m_r^0)$ – шаблон связей, $1 \leq m_1^0 < m_2^0 < \dots < m_r^0 \leq s$; $Q = (q_{J_1^{r+1}})$, $J_1^{r+1} \in V^{r+1}$ – некоторая $(r+1)$ -мерная стохастическая матрица. Если $r = s$, то ЦМ(s, s) $\equiv$ ЦМ(s) – цепь Маркова с полными связями.

Введем обозначения: $F(X_t^{t+s}; M_r) = (x_{t+m_1-1}, \dots, x_{t+m_r-1}, x_{t+s})'$ – селекторная функция порядка $r+1$; символ “ $\bullet$ ”, стоящий вместо любого индекса, означает суммирование по всем его значениям;

$$\nu_{J_1^{r+1}}(M_r) = \sum_{t=1}^{n-s} \mathbf{1}\{F(X_t^{t+s}; M_r) = J_1^{r+1}\}, \quad J_1^{r+1} \in V^{r+1}; \quad (16)$$

$$\hat{\mu}_{J_1^{r+1}}(M_r) = \nu_{J_1^{r+1}}(M_r) / (n-s);$$

$$\hat{I}_{r+1}(M_r) = \sum_{J_1^{r+1} \in V^{r+1}} \hat{\mu}_{J_1^{r+1}}(M_r) \ln \frac{\hat{\mu}_{J_1^{r+1}}(M_r)}{\hat{\mu}_{J_1^r \bullet}(M_r) \hat{\mu}_{\bullet, j_{r+1}}} \geq 0 \quad -$$

подстановочная оценка количества информации по Шеннону.



Теорема 3. *Оценки максимального правдоподобия (ОМП) $\hat{M}_r$, $\hat{Q} = (\hat{q}_{J_1^{r+1}})$, $J_1^{r+1} \in V^{r+1}$ для параметров M_r^0 , Q^0 определяются следующими выражениями:*

$$\hat{M}_r = \arg \max_{M_r \in M} \hat{I}_{r+1}(M_r), \quad (17)$$

$$\hat{q}_{J_1^{r+1}} = \begin{cases} \hat{\mu}_{J_1^{r+1}}(\hat{M}_r) / \hat{\mu}_{J_1^r}(\hat{M}_r), & \text{если } \hat{\mu}_{J_1^r}(\hat{M}_r) > 0, \\ 1/2, & \text{если } \hat{\mu}_{J_1^r}(\hat{M}_r) = 0. \end{cases}$$

Теорема 4. *Если $ЦМ(s, r)$, определяемая (15), стационарна и шаблон связей $M_r^0 \in M$ удовлетворяет условию идентифицируемости, то при увеличении длительности наблюдения $n \rightarrow +\infty$ ОМП $\hat{M}_r$, $\hat{Q}$, определяемые (17), состоятельны:*

$$\hat{M}_r \xrightarrow{P} M_r^0, \quad \hat{Q} \xrightarrow{L_2} Q^0, \quad (18)$$

$$\Delta_n^2 = \mathbf{E} \left\{ \|\hat{Q} - Q^0\|^2 \right\} = \frac{1}{n-s} \sum_{J_1^{r+1} \in V^{r+1}} \frac{(1 - q_{J_1^{r+1}}^0) q_{J_1^{r+1}}^0}{\mu_{J_1^r}(M_r^0)} + o\left(\frac{1}{n}\right).$$



4.4. МПЦМ(s) на основе стандартных дискретных распределений вероятностей

Пусть $\{q_j(\theta): j \in A\}$ – некоторое стандартное дискретное распределение вероятностей на дискретном множестве A с параметрами $\theta = (\theta_j) \in \Theta \subseteq R^L$; $\theta = \theta(j_1, \dots, j_s; \alpha): A^s \times R^m \rightarrow R^L$ – некоторая функция с параметром $\alpha \in R^m$, которая определяет зависимость от предыстории. Порождающее уравнение для переходных вероятностей имеет вид:

$$p_{J_1^{s+1}} = q_{j_{s+1}}(\theta(J_1^s; \alpha)), J_1^{s+1} \in A^{s+1}. \quad (19)$$

Проиллюстрируем наши результаты на модели **биномиального условно нелинейного авторегрессионного временного ряда BiCNAR(s) порядка $s \in N$** , который порождается условным биномиальным распределением вероятностей с параметром p_t на множестве $A = \{0, 1, \dots, N-1\}$:

$$\mathbf{P}\{x_t = j_t \mid X_{-\infty}^{t-1} = J_{-\infty}^{t-1}\} \equiv \mathbf{P}\{x_t = j_t \mid X_{t-s}^{t-1} = J_{t-s}^{t-1}\} = C_{N-1}^{j_t} p_t^{j_t} (1-p_t)^{N-1-j_t}, j_k \in A, k \in Z; \quad (20)$$

$$p_t = p(J_{t-s}^{t-1}) = F_0(J_{t-s}^{t-1}), t \in Z,$$

где $X_{t-s}^{t-1} = (x_{t-s}, \dots, x_{t-1})' \in A^s$ – вектор-столбец s -предыстории процесса к моменту времени t ; s – глубина предыстории $1 \leq s < +\infty$; $p_t \in [0, 1]$ – параметр биномиального распределения вероятностей; $F_0(\cdot): A^s \rightarrow [0, 1]$ – некоторая неизвестная «функция обратной связи», задающая зависимость от s -предыстории. Рассмотрим случай аппроксимаций $F_0(\cdot)$ с использованием базисных функций:

$$F_0(J_{t-s}^{t-1}) ::= F\left(\sum_{i=1}^m b_i \psi_i(J_{t-s}^{t-1})\right), J_{t-s}^{t-1} \in A^s, \quad (21)$$

где $F(\cdot): R^1 \rightarrow [0, 1]$ – некоторая абсолютно непрерывная функция распределения; $\{\psi_i(\cdot)\}$ – заданный набор m линейно независимых на A^s базисных функций, $m \leq N^s$; $B = (b_i)$ – неизвестные параметры (коэффициенты) модели.



Пусть наблюдается реализация $X_1^T = (x_1, \dots, x_T)' \in A^T$ длины T временного ряда, порожденного моделью (19) – (21) с неизвестными параметрами $B \in R^m$. Для оценивания параметров B применим разработанный нами FBE (Frequencies Based Estimation)-метод – метод, основанный на многомерных частотах [12, 13]. Примем обозначения: $\mathbf{1}(U)$ – индикаторная функция события U ,

$$\nu_s^T(J_1^s) = \sum_{t=s}^T \mathbf{1}\{X_{t-s+1}^T = J_1^s\} - \quad (22)$$

частота встречаемости s -граммы $J_1^s = (j_1, \dots, j_s) \in A^s$ в реализации X_1^T ,

$$\hat{\pi}_{J_1^s} = \frac{\nu_s^T(J_1^s)}{(T-s+1)}, \quad \hat{\pi}_{J_1^{s+1}} = \frac{\nu_{s+1}^T(J_1^{s+1})}{(T-s)} - \quad (23)$$

частотные оценки s и $(s+1)$ -мерных стационарных распределений вероятностей,

$$J^{(T)} = \{J_1^s \in V^s: \nu_s^T(J_1^s) > 0\} \subseteq A^s - \quad (24)$$

подмножество последовательных s -грамм в реализации X_1^T , имеющих ненулевые частоты; $J^{(T)} \xrightarrow{\text{п.н.}} A^s$ при $T \rightarrow +\infty$.



Построим подстановочную частотную оценку для условной вероятности (20):

$$\hat{p}(J) = \begin{cases} \frac{\sum_{t=s+1}^T \mathbf{1}\{X_{t-s}^{t-1} = J\}}{(N-1)\hat{\pi}_{J_1^s}}, & J \in J^{(T)}, \\ 1/N, & J \in \bar{J}^{(T)}. \end{cases} \quad (25)$$

Введем в рассмотрение статистики

$$\hat{u}(J_1^s) = F^{-1}(\hat{p}(J_1^s)), \quad J_1^s \in J^{(T)}, \quad (26)$$

и согласно FBE-методу [12] минимизируем квадратичную функцию уклонений левых частей от правых частей уравнений (26) при подстановке значений $u(J_1^s) = F^{-1}(\hat{p}(J_1^s)), J_1^s \in J^{(T)}$:

$$w = w(B) ::= \sum_{J_1^s \in J^{(T)}} (B' \Psi(J_1^s) - \hat{u}(J_1^s))^2 \rightarrow \min_B.$$

Обозначим:

$$D = (d_{kl}) = \sum_{J_1^s \in J^{(T)}} \Psi(J_1^s) \Psi'(J_1^s) \in R^{m \times m}, \quad E = (e_k) = \sum_{J_1^s \in J^{(T)}} \hat{u}(J_1^s) \Psi(J_1^s) \in R^m. \quad (27)$$

Теорема 5. Если ДВР x_t , определяемый (19) – (21), эргодичен и $|D| \neq 0$, то при увеличении длительности наблюдения $T \rightarrow +\infty$ существует FBE-оценка $\hat{B} = (\hat{b}_i) = D^{-1}E$, которая состоятельна:

$$\hat{B} \xrightarrow{P} B^0,$$

и асимптотически нормальна:

$$\sqrt{T}(\hat{B} - B^0) \xrightarrow{D} \mathcal{N}_m(O_m, \Sigma), \quad \Sigma = (\Psi^T \Psi)^{-1} \Psi^T \Lambda \Psi (\Psi^T \Psi)^{-1},$$

где Λ – известная $N^s \times N^s$ -матрица, $O_m \in R^m$ – нулевой вектор.



4.5. МПЦМ(s) с использованием искусственных нейронных сетей (ИНС)

Проиллюстрируем применение ИНС для построения МПЦМ(s) на примере BiCNAR(s)-модели (20). Существенный недостаток предыдущей модели (20) – отсутствие рекомендаций по выбору базиса $\{\psi_i\}$. Вместо аппроксимации (21) применим нейросетевую аппроксимацию [14]:

$$F_0(J_1^s) = F\left(\sum_{i=1}^m b_i F_i\left(\sum_{k=1}^s a_{ik} j_k\right)\right), J_1^s \in A^s, \quad (28)$$

где $F(\cdot), F_1(\cdot), \dots, F_m(\cdot): R^1 \rightarrow [0, 1]$ – некоторые заданные абсолютно непрерывные функции распределения; параметры модели: $2 \leq m < +\infty$, $B = (b_1, \dots, b_m)' \in R^m$, $A_i = (a_{i1}, \dots, a_{is})' \in R^s$, $i \in \{1, \dots, m\}$ – неизвестные вектор-столбцы коэффициентов. Нелинейная зависимость (28) реализуется двухслойной ИНС (рис. 5) с s входами, одним выходом, m нейронами на 1-ом слое и одним нейроном на 2-ом слое; число параметров $d = m(s+1)$ линейно зависит от глубины памяти s .

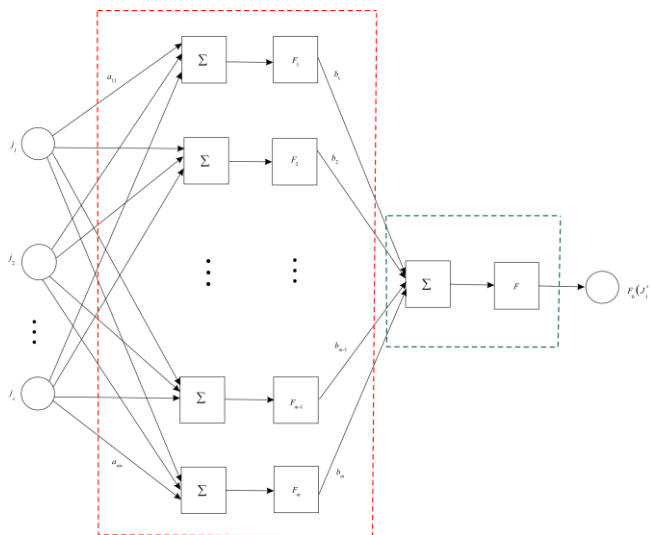


Рис. 5. Двухслойная ИНС, реализующая преобразование $F_0(\cdot)$

Нейросетевая модель (28) обладает известным свойством универсальности: при увеличении параметра сложности $m \rightarrow +\infty$ она позволяет достичь заданной точности аппроксимации любой «функции обратной связи» $F_0(J_1^s)$.



Вероятностные свойства нейросетевой модели

Обозначим: $\theta = (B', A'_1, \dots, A'_m)' \in R^{m(s+1)}$ – составной вектор-столбец параметров модели:

$$\Theta_\theta = \left\{ \tilde{\theta} = (\tilde{B}', \tilde{A}'_1, \dots, \tilde{A}'_m)' : \tilde{B} = (b_{\pi_1}, \dots, b_{\pi_m})', \tilde{A}_1 = A_{\pi_1}, \dots, \tilde{A}_m = A_{\pi_m}, \right. \quad (29)$$

$\pi = (\pi_1, \dots, \pi_m) \in S_m$ – произвольная подстановка на $\{1, 2, \dots, m\}$.

Теорема 6. В условиях эргодичности, если $0 < F(z) < 1$, $z \in R$, то нейросетевая модель (28), с любыми параметрами $\tilde{\theta} \in \Theta_\theta$ эквивалентна этой модели с параметрами θ , $|\Theta_\theta| = |S_m| = m!$.

Статистическое оценивание параметров нейросетевой модели

Применим разработанный нами ранее FBE (Frequencies Based Estimation)-метод, основанный на многомерных частотах [12].

Обозначения: $F^{-1}(\cdot)$ – квантильная функция для функции распределения $F(\cdot)$; $J = (j_1, \dots, j_s) \in A^s$; m – вектор-столбец

$$G = (g_k) = G(A_1, \dots, A_m; J) = (F_1(A'_1 J), \dots, F_m(A'_m J))' \in R^m;$$

$X_1^T = (x_1, \dots, x_T) \in A^T$ – наблюдаемая реализация длины T ; частотная оценка для $p(J) = (N-1)^{-1} \mathbf{E} \{x_t | X_{t-s}^{t-1} = J\}$ по X_1^T :

$$\hat{p}(J) = \frac{1}{N-1} \sum_{t=s+1}^T x_t \mathbf{1} \{X_{t-s}^{t-1} = J\} / \sum_{t=s+1}^T \mathbf{1} \{X_{t-s}^{t-1} = J\}.$$

Согласно FBE-методу [12] определим квадратичную функцию потерь:

$$w = w(B, \{A_i\}) ::= \sum_{J \in J^{(T)}} (B'G - u(J))^2 \geq 0, u(J) = F^{-1}(\hat{p}(J)),$$

и построим статистические оценки $\hat{B}, \{\hat{A}_i\}$, минимизирующие эту функцию потерь:

$$w(B, \{A_i\}) \rightarrow \min_{B, \{A_i\}}. \quad (30)$$



Оценивание B при известных $\{A_i^0\}$

Примем обозначения: δ_{kl} – символ Кронеккера; $f_i(z) = F'_i(z) \geq 0$; $D = (d_{kl}) = \sum_{J \in J^{(T)}} GG' \in R^{m \times m}$,
 $D^{-1} = (\bar{d}_{kl}) \in R^{m \times m}$ – симметричные неотрицательно определенные $(m \times m)$ -матрицы;
 $E = (e_k) = \sum_{J \in J^{(T)}} u(J)G(\{A_i\}; J) \in R^m$.

Теорема 7. Если истинные значения $\{A_i^0\}$ известны, $|D| \neq 0$ и дискретный временной ряд X_1^T , определяемый (20), (28), эргодичен, то при увеличении длительности наблюдения $T \rightarrow +\infty$ оценка $\hat{B} = (\hat{b}_i) = D^{-1}E$ сходится к истинному значению B^0 по вероятности:

$$\hat{B} = D^{-1}E \xrightarrow{P} B^0.$$

Совместное оценивание B , $\{A_i^0\}$

Для построения оценок $\{\hat{A}_i\}$ приходим к задаче максимизации:

$$w_2(\{A_i\}) := E'D^{-1}E \rightarrow \max_{\{A_i\}}. \quad (31)$$

Лемма 1. В принятых обозначениях частные производные первого порядка целевой функции (31) вычисляются по формулам ($i \in \{1, \dots, m\}$, $\nu \in \{1, \dots, s\}$):

$$\frac{\partial w_2(\{A_k\})}{\partial a_{i\nu}} = 2\hat{b}_i \sum_{J \in J^{(T)}} j_\nu f_i(A'_i J) (u(J) - \hat{B}'G(\{A_k\}; J)),$$

где $\hat{B} = (\hat{b}_i) = D^{-1}E$.



Алгоритм вычисления оценок $\hat{B}, \{\hat{A}_i\}$ и их свойства

Лемма 1 дает явный вид всех частных производных первого порядка по элементам матриц $\{A_i\}$, поэтому для максимизации целевой функции (31) можно применить градиентный метод наискорейшего подъема ($\tau \in 1, 2, \dots$):

$$\hat{A}_i^{(\tau)} = \hat{A}_i^{(\tau-1)} + \lambda V_i^{(\tau)}, \quad i = 1, \dots, m, \quad (32)$$

$$V_i^{(\tau)} = (v_{iv}^{(\tau)}), \quad v_{iv}^{(\tau)} = 2\hat{b}_i^{(\tau-1)} \sum_{J \in J^{(T)}} j_v f_i(\hat{A}_i^{(\tau-1)'J}) \left(u(J) - \hat{B}^{(\tau-1)'} G\left(\{\hat{A}_k^{(\tau-1)'}\}; J\right) \right), \quad v = 1, \dots, s,$$

$\hat{A}_i^{(\tau)} = (\hat{a}_{iv}^{(\tau)})$ – оценка вектора A_i на τ -ой итерации: $\{A_i^{(0)}\}$ – задаваемые начальные значения, $\lambda > 0$ – задаваемая величина шага. Для нахождения глобального максимума в (31) целесообразно задавать случайным образом несколько начальных значений $\{A_i^{(0)}\}$.

Теорема 8. В условиях Теоремы 6, если $|D| \neq 0$, то составной вектор-столбец построенных оценок $\hat{\theta} = (\hat{B}', \hat{A}_1', \dots, \hat{A}_m')' \in R^{m(s+1)}$ при $T \rightarrow +\infty$ сходится по вероятности:

$$\hat{\theta} \xrightarrow{P} \theta^*,$$

где θ^* – некоторая точка из множества Θ_{θ^0} , определенного формулой (29). При этом

$$F_0(J_1^s; \theta^*) = F_0(J_1^s; \theta^0), \quad J_1^s \in A^s.$$



4.6. МПЦМ(s) на основе экспоненциальных семейств и достаточных статистик

$\mathbf{x}_t \in A$, $t \in \mathbb{Z}$, – наблюдаемый дискретный временной ряд с конечным пространством состояний мощности $N = |A| < +\infty$, задаваемый экспоненциальными семействами условных распределений с каноническим параметром $\boldsymbol{\eta} = (\eta_i)_{i=1}^m \in R^m$:

$$\mathbf{P}\{\mathbf{x}_t = x \mid \mathbf{x}_{t-s}^{t-1} = q\} = \mathcal{E}^{(q)}(x; \boldsymbol{\eta}) ::= \exp\left(h_0^{(q)}(x) + \langle \boldsymbol{\eta}, \mathbf{h}^{(q)}(x) \rangle - \phi^{(q)}(\boldsymbol{\eta})\right), \quad (33)$$

$$\phi^{(q)}(\boldsymbol{\eta}) = \ln \sum_{x \in A} \exp\left(h_0^{(q)}(x) + \langle \boldsymbol{\eta}, \mathbf{h}^{(q)}(x) \rangle\right), \quad (34)$$

$$\mathbf{h}^{(q)}(x) = \left(h_i^{(q)}(x)\right)_{i=1}^m \in R^m, \quad q \in A^s, \quad m \leq \mathbf{M} = N^s (N - 1), \quad (35)$$

где опорная функция $h_0^{(q)}(\cdot)$ и базисные функции $\{h_i^{(q)}(\cdot) : i = 1, \dots, m\}$ полностью определяют модель (33) – (35), которую будем называть MCSS(s) (от англ. **Markov Chain with Sufficient Statistics**). Определим: $\mathbf{LIND}_U\{f_1, \dots, f_k\}$ – условие линейной независимости функций $\{f_i : U \rightarrow R\}_{i=1}^k$.

Условие регулярности R.0: $\mathbf{LIND}_{A^{s+1}}\{h_1, \dots, h_m, \mathbf{1}_{q'}, q' \in A^s\}$,

где $\mathbf{1}_{q'}^{(q)}(x) ::= \mathbf{1}\{q = q'\}$, $q, q' \in A^s$, $x \in A$, – функция на $A^{s+1} = \{(q, x) : q \in A^s, x \in A\}$, индикатор события $\{q = q'\}$.

Условие регулярности R.1: $\mathcal{M} ::= \{q \in A^s : \mathbf{LIND}_A\{\mathbf{1}, h_1^{(q)}, \dots, h_m^{(q)}\}\} \neq \emptyset$, где $\mathbf{1}(x) \equiv 1, x \in A$.

Условие **R.1** усиленная версия условия **R.0**: **R.1** $\Rightarrow$ **R.0**.

Лемма 2. При условии **R.0** модель (33) – (35) идентифицируема.



Статистическое оценивание параметров модели MCSS(s)

Теорема 9. Пусть $X_1^T = (\mathbf{x}_1, \dots, \mathbf{x}_T) \in A^T$ – наблюдаемый временной ряд (33) – (35) длины $T \in \mathbb{N}$. Функция логправдоподобия модели MCSS(s) допускает следующее представление:

$$L(\boldsymbol{\eta}) = \langle \boldsymbol{\eta}, \mathbf{H} \rangle - (T-s) \sum_{q \in A^s} \hat{\pi}_q^{(s)} \phi^{(q)}(\boldsymbol{\eta}) + \sum_{t=s+1}^T h_0^{(\mathbf{x}_{t-s}^{t-1})}(\mathbf{x}_t), \quad (36)$$

где $\mathbf{H} = (\mathbf{H}_i)_{i=1}^T := \sum_{t=s+1}^T h^{(\mathbf{x}_{t-s}^{t-1})}(\mathbf{x}_t) \in R^m$, $\hat{\pi}_q^{(s)} = \frac{1}{T-s} \sum_{t=s+1}^T \mathbf{1}\{\mathbf{x}_{t-s}^{t-1} = q\}$, $q \in A^s$, – достаточные статистики.

Теорема 10. При условии **R.1** функция логправдоподобия $L(\boldsymbol{\eta})$ при $T \rightarrow +\infty$ почти наверное строго вогнута и ОМП $\hat{\boldsymbol{\eta}} = \arg \max_{\boldsymbol{\eta} \in R^m} L(\boldsymbol{\eta})$ при $T \rightarrow +\infty$ состоятельна, асимптотически нормально распределена и эффективна:

$$\sqrt{T}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \xrightarrow[T \rightarrow +\infty]{D} \mathcal{N}_m(0_m, \mathcal{J}^{-1}(\boldsymbol{\eta})),$$

где $\mathcal{J}(\boldsymbol{\eta}) \in R^{m \times m}$ – информационная матрица Фишера параметра $\boldsymbol{\eta}$ модели MCSS(s).

ОМП $\hat{\boldsymbol{\eta}}$ вычисляется алгоритмом градиентного подъема ($k = 0, 1, \dots$) с шагами величиной $\alpha_k > 0$:

$$\boldsymbol{\eta}^{(k+1)} = \boldsymbol{\eta}^{(k)} + \alpha_k \frac{d}{d\boldsymbol{\eta}} \tilde{L}(\boldsymbol{\eta}) \Big|_{\boldsymbol{\eta} = \boldsymbol{\eta}^{(k)}}, \quad (37)$$

$$\frac{d}{d\boldsymbol{\eta}} \tilde{L}(\boldsymbol{\eta}) = \sum_{t=s+1}^T \mathbf{h}^{(\mathbf{x}_{t-s}^{t-1})}(\mathbf{x}_t) - (T-s) \sum_{q \in A^s} \hat{\pi}_q^{(s)} \frac{\sum_{x \in A} \mathbf{h}^{(q)}(x) \exp(h_0^{(q)}(x) + \langle \boldsymbol{\eta}, \mathbf{h}^{(q)}(x) \rangle)}{\sum_{x \in A} \exp(h_0^{(q)}(x) + \langle \boldsymbol{\eta}, \mathbf{h}^{(q)}(x) \rangle)}.$$



5. Подход к статистическому тестированию на основе энтропийных профилей

5.1. Функционалы энтропии и их статистические оценки

Пусть, как и в разделе 2, наблюдается случайная выборка $X = \{x_t : t = 1, \dots, n\}$ объема n из распределения вероятностей $\{p_k : k \in A\}$, $p_k = \mathbf{P}\{x_t = k\}$, где $A = \{0, \dots, K-1\}$, $K = 2^s$, порождённая исследуемым генератором случайных или псевдослучайных последовательностей. Построим частотные оценки:

$$\hat{p}_k = \frac{v_k}{n}, \quad v_k = \sum_{t=1}^n \mathbf{1}\{x_t = k\} \geq 0, \quad k \in A.$$

Будем использовать специальную **асимптотику «растущей размерности»**:

$$n, K \rightarrow \infty, n/K \rightarrow \lambda, 0 < \lambda < \infty, \quad (38)$$

которая отличается от классической ($n \rightarrow \infty, K < \infty$) тем, что длительность наблюдения n и число значений (сложность модели) K растут синхронно.

Таблица 1: Исследуемые функционалы энтропии ($x^2 ::= x(x-1)$)

Тип	Формула	Статистическая оценка
Энтропия Шеннона	$H(p) = -\sum_{k \in A} p_k \ln p_k$	$\hat{H}(n, K) = \ln n - \frac{1}{n} \sum_{k \in A} v_k \ln v_k$
Энтропия Реньи	$H_2(p) = -\ln \left(\sum_{k \in A} p_k^2 \right)$	$\hat{H}_2(n, K) = 2 \ln n - \ln \left(\sum_{k \in A} v_k^2 \right)$
Энтропия Тсаллиса	$S_2(p) = 1 - \sum_{k \in A} p_k^2$	$\hat{S}_2(n, K) = 1 - \frac{1}{n^2} \sum_{k \in A} v_k^2$



5.2. Асимптотические свойства статистических оценок

Доказана асимптотическая нормальность оценок $\hat{H}$, $\hat{H}_2(n, K)$, $\hat{S}_2(n, K)$ при простой нулевой гипотезе H_0 и оценки $\hat{S}_2(n, K)$ при сложной нулевой гипотезе $H_0^{(\varepsilon)}$, определяемой (5) в разделе 2. В таблице 2 приведены формулы асимптотических математических ожиданий и дисперсий статистических оценок энтропии Шеннона, Реньи и Тсаллиса при справедливости гипотезы H_0 из [16, 17] и Тсаллиса при справедливости гипотезы $H_0^{(\varepsilon)}$ из [18].

Таблица 2: Параметры асимптотически нормального распределения оценок энтропии

Тип энтропии; гипотеза	Математическое ожидание	Дисперсия
Шеннон; H_0	$\mu_H = \ln n - e^{-\lambda} \sum_{k=1}^{+\infty} \frac{\lambda^k \ln(k+1)}{k!}$	$\sigma_H^2 = \frac{e^{-\lambda}}{n} \sum_{k=1}^{+\infty} \frac{(k+1)\lambda^k}{k!} \ln^2(k+1) -$ $-\frac{e^{-2\lambda}}{N} \left(\sum_{k=1}^{+\infty} \frac{\ln(k+1)\lambda^k}{k!} \right)^2 -$ $-\frac{e^{-2\lambda}}{n} \left(\sum_{k=1}^{+\infty} \ln(k+1) \frac{\lambda^k}{k!} (k+1-\lambda) \right)^2$
Реньи; H_0	$\mu_{H,2} = \ln K$	$\sigma_{H,2}^2 = \frac{2}{n\lambda}$
Тсаллис; H_0	$\mu_{S,2} = 1 - \frac{1}{K}$	$\sigma_{S,2}^2 = \frac{2}{Kn^2}$
Тсаллис; $H_0^{(\varepsilon)}, 0 \leq \varepsilon \leq \varepsilon_+$	$\mu_{S,2}^{(\varepsilon_+)} = \mu_{S,2} - \varepsilon_+^2$	$\sigma_{S,2}^{2(\varepsilon_+)} = \sigma_{S,2}^2 + \frac{2}{n^2} (2\lambda + 1) \varepsilon_+^2 + \frac{4\varepsilon_+^3}{n} (1 - \varepsilon_+)$



5.3. Тесты на основе энтропийных профилей

Введём обозначения для проверки гипотез H_0 , $H_0^{(\varepsilon)}$ об s -мерной равномерности ($1 \leq s \leq s_+$): $\hat{h}(s)$ – статистическая оценка энтропии Шеннона, Реньи или Тсаллиса, μ_h и σ_h^2 – соответственно её асимптотические математическое ожидание и дисперсия при истинной гипотезе H_0 . Статистический критерий асимптотического уровня значимости $\alpha \in \{0, 1\}$ проверки гипотез H_0 , $\overline{H_0} \left(H_0^\varepsilon, \overline{H_0^\varepsilon} \right)$ на основе статистики $\hat{h}(s)$ имеет вид:

$$\text{принимается} \begin{cases} H_0 \left(H_0^\varepsilon \right), \text{ если } \Delta_- < \hat{h} < \Delta_+, \\ \overline{H_0} \left(\overline{H_0^\varepsilon} \right), \text{ в противном случае,} \end{cases} \quad (39)$$

где для простой гипотезы $\Delta_\pm = \mu_h \pm \sigma_h \Phi^{-1}(1 - \alpha/2)$, для сложной гипотезы $\Delta_- = \mu_{s,2}^{(\varepsilon_+)} - \sigma_{s,2}^{(\varepsilon_+)} \Phi^{-1}(1 - \alpha/2)$, $\Delta_+ = \mu_{s,2} + \sigma_{s,2}^{(\varepsilon_+)} \Phi^{-1}(1 - \alpha/2)$; $\Phi(\cdot)$ – функция распределения стандартного нормального закона.

Для каждого значения $s = 1, \dots, s_+$ вычислим оценку энтропии $\hat{h}(s)$, нормированные статистики $\tilde{h}_0$ и $\tilde{h}_\varepsilon$ для простой и сложной гипотез соответственно и вспомогательную статистику $\tilde{h}_\varepsilon$ для вычисления двустороннего p -значения: $\tilde{h}_0 = (\hat{h} - \mu_h) / \sigma_h$, $\tilde{h}_\varepsilon = (\hat{S}_2 - \mu_{s,2}) / \sigma_{s,2}^{(\varepsilon_+)}$, $\tilde{h}_\varepsilon = (\hat{S}_2 - \mu_{s,2}^{(\varepsilon_+)}) / \sigma_{s,2}^{(\varepsilon_+)}$. Формулы P -значений для простой H_0 и сложной H_0^ε гипотез равны соответственно:

$$P_0 = 2(1 - \Phi(|\tilde{h}_0|)), \quad P_\varepsilon = \begin{cases} 2(1 - \Phi(|\tilde{h}_0|)), & \text{если } \hat{S}_2 > \mu_{s,2}, \\ 2 - \Phi(|\tilde{h}_0|) - \Phi(|\tilde{h}_0|), & \text{если } \mu_{s,2}^{(\varepsilon_+)} \leq \hat{S}_2 \leq \mu_{s,2}, \\ 2(1 - \Phi(|\tilde{h}_0|)), & \text{если } \hat{S}_2 < \mu_{s,2}^{(\varepsilon_+)}. \end{cases}$$



5.4. Программный Комплекс статистического тестирования «ЭАДП»

В НИИ ППМИ БГУ разработан Программный Комплекс «Энтропийный анализ дискретных последовательностей» (ПК «ЭАДП») [19], позволяющий вычислять оценки функционалов энтропии Шеннона, Реньи и Тсаллиса дискретной последовательности и на основе критерия (39) в асимптотике (38) принимать или отклонять нулевую гипотезу. ПК «ЭАДП» позволяет для заданного уровня значимости α вычислить последовательность статистик $\bar{h}(s) = \tilde{h}(s) / (\Phi^{-1}(1 - \alpha/2))$ для $s = 1, \dots, s_+$, и представлять ее в виде **графика, называемого энтропийным профилем выходной последовательности криптографического генератора** (см. рис. 6). Простая нулевая гипотеза H_0 принимается, если $\bar{h}_0(s) \in [-1; 1]$. Сложная нулевая гипотеза H_0^ε принимается, если $\bar{h}_\varepsilon(s) \in \left[-1 - \varepsilon_+^2 / (\sigma_{s,2}^{(\varepsilon_+)} \Phi^{-1}(1 - \alpha/2)), 1\right]$.

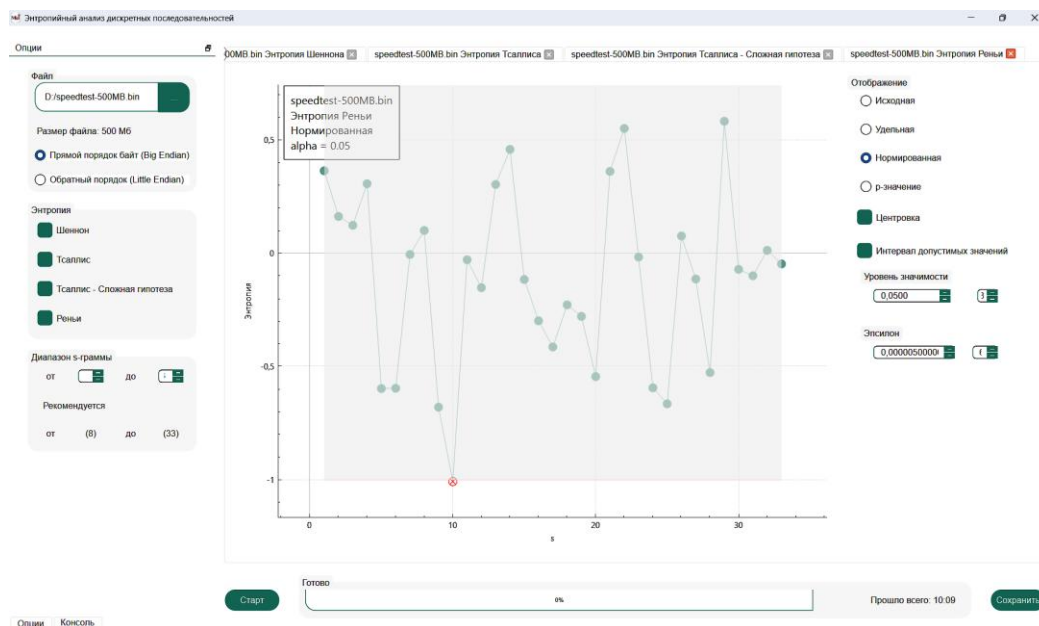


Рис. 6. Скриншот Программного Комплекса ЭАДП



6. Компьютерные эксперименты по оценке близости выходных последовательностей физических и программных генераторов к равномерно распределенной последовательности

Пример 1: SG-генератор

Статистический анализ выходной последовательности длины $T = 10^5$ SG-генератора с помощью $MC(s, r)$ ($s < 20$, $r < 9$) дал следующие результаты: $\hat{s} = 15$, $\hat{r} = 4$, $\hat{M}_{\hat{r}} = (1, 2, 3, 4)$, матрица $\hat{Q}$ визуализирована на рис. 7.



Рис. 7. Оценка матрицы Q (**красный** – оценка вероятности перехода в 0, **зеленый** – в 1)



Пример 2: Генератор со случайной обратной связью

Параметры генератора: характеристический многочлен $p(\lambda) = \lambda^{63} + \lambda^z + 1$, с одинаковой вероятностью на каждом такте случайная величина z принимает значения из множества $\{1, 5, 31\}$.

При статистическом анализе выходной последовательности длины $T = 10^5$ с помощью $MC(s, r)$ ($s \leq 80$, $r \leq 6$) получено: $\hat{s} = 63$, $\hat{r} = 4$, $\hat{M}_{\hat{r}} = (1, 2, 6, 32)$, $\hat{Q}$ – на рис. 8.



Рис. 8. Оценка матрицы Q (**красный** – оценка вероятности перехода в 0, **зеленый** – в 1)



Пример 3: Генератор Bluetooth

При статистическом анализе выходной последовательности длины $T = 10^5$ генератора Bluetooth с помощью $MC(s, r)$ ($s < 20, r < 9$) получено: $\hat{s} = 5$, $\hat{r} = 5$, $\hat{M}_r = (1, 2, 3, 4, 5)$, $\hat{Q}$ – на рис. 9.



Рис. 9. Оценка матрицы Q при $s = 32$, $r = 4$, $M_r = (1, 4, 17, 20)$



Пример 4: Генератор на основе BELT (разработка НИИ ПМИ)

При статистическом анализе последовательности длины $T = 8 \cdot 10^6$ генератора BELT в режиме счетчика с помощью $MC(s, r)$ (в предположении, что $s \leq 50$, $r \leq 9$) не удалось отличить эту последовательность от РПСР, $\hat{Q}$ – на рис. 10.

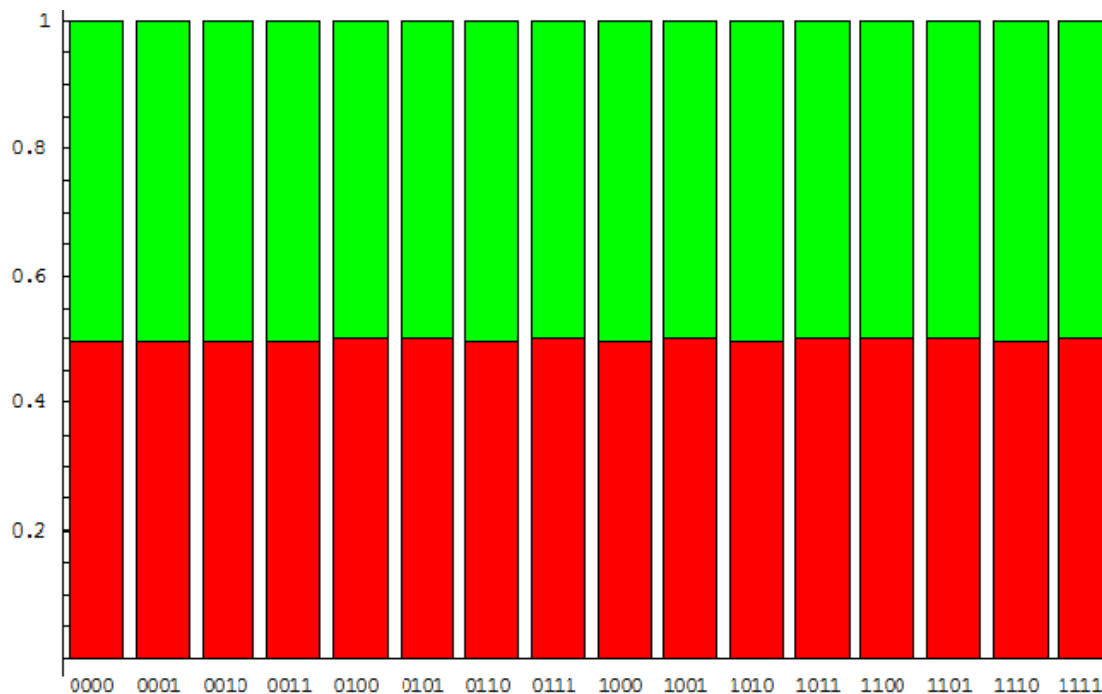


Рис. 10. Оценка матрицы Q (**красный** – оценка вероятности перехода в 0, **зеленый** – в 1)



Пример 5: Физический генератор (Берлинский университет)

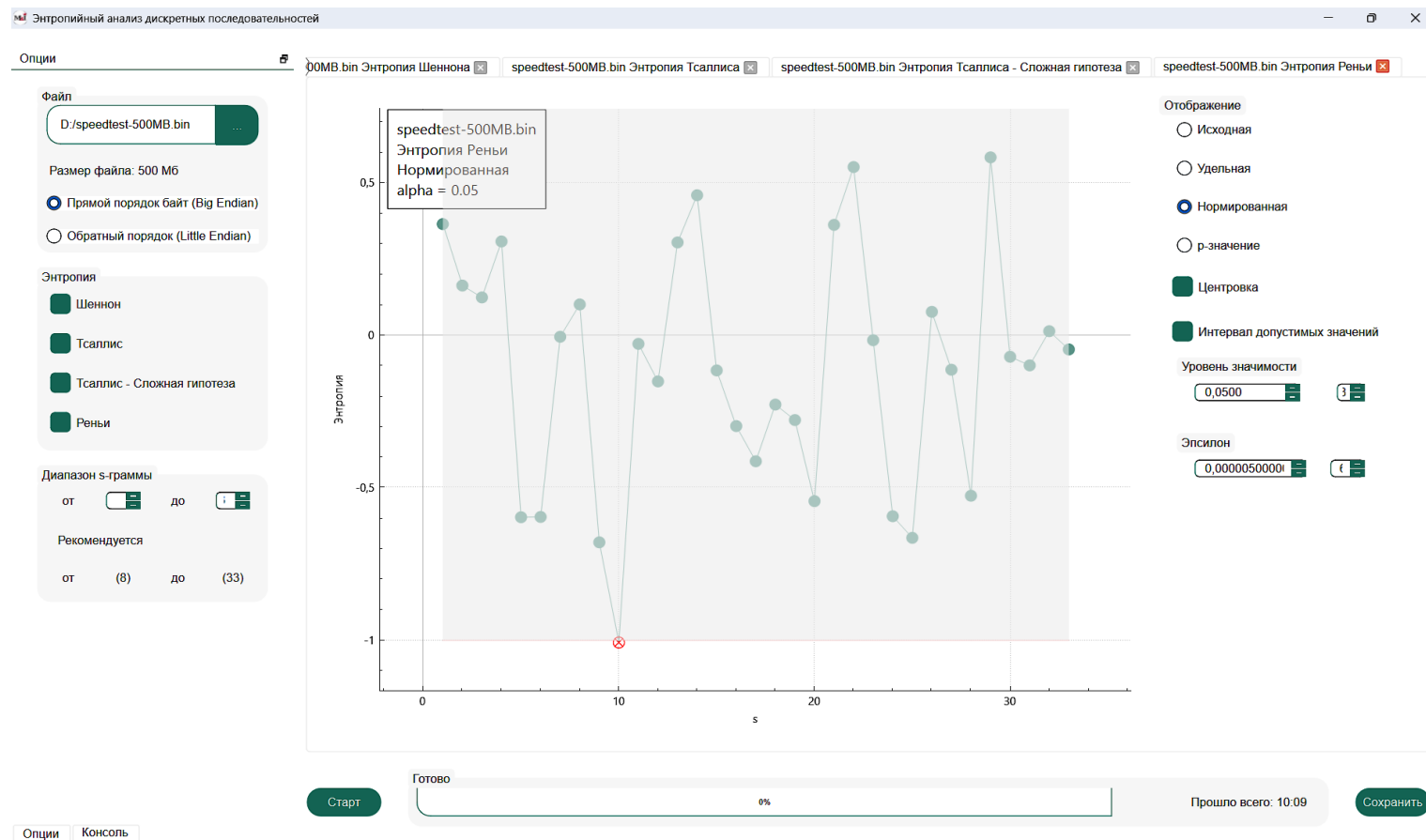


Рис. 11. Энтропийный профиль физического генератора



Пример 6: Программный генератор: линейный регистр сдвига 32-го порядка

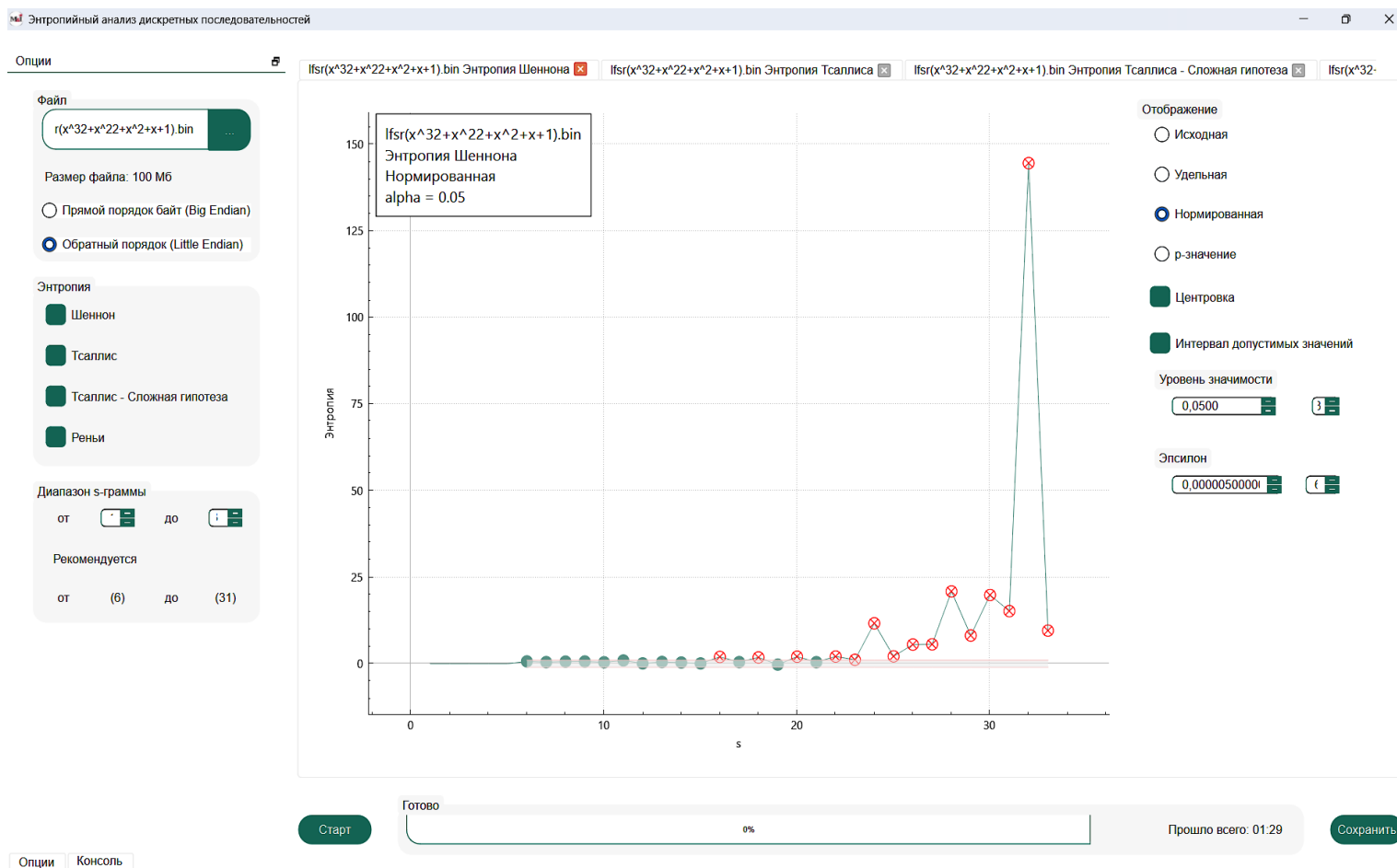


Рис. 12. Энтропийный профиль линейного регистра сдвига с многочленом $x^{32} + x^{22} + x + 1$



7. Заключение



В докладе представлены следующие основные результаты по статистическому тестированию криптографических генераторов.

1. Представлен подход к статистическому тестированию, в котором нулевая гипотеза предполагается сложной и допускает отклонения распределения вероятностей s -грамм от равномерного, не превышающие некоторого фиксированного уровня ε . Построен статистический критерий отношения правдоподобия с заданным асимптотическим уровнем значимости.
2. Разработан подход к статистическому оцениванию уровня безопасности (стойкости) криптографического генератора на основе статистического тестирования его выходных последовательностей с использованием сложных нулевых гипотез. Предложен алгоритм статистического оценивания уровня безопасности криптографического генератора по M -значной шкале, основанный на $(M - 1)$ -кратном использовании теста (13).
3. Представлен подход к статистическому тестированию на основе аппроксимации выходной последовательности криптографического генератора малопараметрическими цепями Маркова высокого порядка, преодолевающими экспоненциальную сложность, и проиллюстрирована целесообразность применения этого подхода при исследовании генераторов случайных и псевдослучайных последовательностей.
4. Разработан подход к статистическому тестированию на основе статистических оценок функционалов энтропии Шеннона, Реньи и Тсаллиса в «асимптотике растущей размерности», реализованный в виде Программного Комплекса «ЭАДП».
5. Теоретические результаты проиллюстрированы результатами компьютерных экспериментов.



8. Список литературы



- [1] А.П. Алферов, А.Ю. Зубов, А.С. Кузьмин, А.В. Черемушкин, *Основы криптографии*, Гелиос АРВ, Москва, 2005, 480 с.
- [2] Ю.С. Харин, С.В. Агиевич, Д.В. Васильев, Г.В. Матвеев, *Криптология*, БГУ, Минск, 2023, 512 с.
- [3] NIST SP 800-22: *Download Documentation and Software*, <https://csrc.nist.gov/projects/random-bit-generation/documentation-and-software>.
- [4] А.М. Zubkov, A.A. Serov, “Testing the NIST Statistical Test Suite on artificial pseudorandom sequences”, *Математические вопросы криптографии*, **10:2** (2019), 89-96.
- [5] *Decision to Revise NIST SP 800-22 Rev. 1a*, <https://csrc.nist.gov/news/2022/decision-to-revise-nist-sp-800-22-rev-1a>.
- [6] М.П. Савелов, «Предельное совместное распределение статистик критериев пакета NIST “Monobit Test”, “Frequency Test within a Block” и обобщения критерия “Approximate Entropy Test”», *Дискретная математика*, **35:2** (2023), 93-108.
- [7] A.A. Serov, “Reliability of two-level testing approach of the NIST test suite”, *Computer data analysis and modeling: stochastics and data science: Proceedings of the XIV international Conference, September 24-27, 2025, BSU, Minsk*, 240-243.
- [8] Ю.С. Харин, А.М. Зубков, «О статистической проверке сложных гипотез об s-мерном равномерном распределении вероятностей двоичных последовательностей», *Дискретная математика*, **36:1** (2024), 116-135.
- [9] Г.И. Ивченко, Ю.И. Медведев, *Математическая статистика*, Высшая школа, Москва, 1992, 304 с.
- [10] Ю.С. Харин, В.А. Волошко, «Об аппроксимации двоичных цепей Маркова высокого порядка малопараметрическими моделями», *Дискретная математика*, **34:3** (2022), 114-135.
- [11] Ю.С. Харин, А.И. Петлицкий, «Цепь Маркова s-го порядка с r-частичными связями и статистические выводы о ее параметрах», *Дискретная математика*, **19:2** (2007), 109-130.
- [12] Yu.S. Kharin, V.A. Voloshko, E.A. Medved, “Statistical estimation of parameters for binary conditionally nonlinear autoregressive time series”, *Mathematical Methods of Statistics*, **27:2** (2018), 103-118.
- [13] Yu.S. Kharin, V.A. Voloshko, “Robust estimation for Binomial conditionally nonlinear autoregressive time series based on multivariate conditional frequencies”, *Journal of Multivariate Analysis*, **185:2** (2022), 104777.
- [14] Yu.S. Kharin, “Neural net models of binomial time series in data analysis”, *Reports of national Academy of Sciences of Belarus*, **65:6** (2021), 654-660.
- [15] Yu.S. Kharin, V.A. Voloshko, “Statistical analysis of parsimonious high-order multivariate finite Markov chains based on sufficient statistics”, *Journal of Multivariate Analysis*, **208:2** (2025), 105422.
- [16] В.Ю. Палуха, Ю.С. Харин, «Статистические оценки энтропии Реньи и Тсаллиса и их использование для проверки гипотез о «чистой случайности», *Весці НАН Беларусі. Серыя фіз.-мат. навук*, **52:2** (2016), 37-47.
- [17] В.Ю. Палуха, «Статистические тесты на основе оценок энтропии для проверки гипотез о равномерном распределении случайной последовательности», *Весці НАН Беларусі. Серыя фіз.-мат. навук*, **53:1** (2017), 79-88.
- [18] В.Ю. Палуха, Ю.С. Харин, «Статистическое тестирование равномерности дискретного распределения вероятностей на основе энтропии Тсаллиса», *Дискретная математика*, **36:4** (2024), 101-116.
- [19] В.Ю. Палуха, Ю.С. Харин, «О применении программного комплекса «ЭАДП» для статистического тестирования выходных последовательностей криптографических генераторов», *Комплексная защита информации: материалы XXVIII научно-практической конференции, Гомель, 25-27 мая 2023 г.*, УГЗ МЧС, Минск, 2023, 148-153.



**Благодарю
за Ваше внимание!**