V. G. Mikhailov, Poisson approximation for the distribution of the frequency of a given pattern in the outcome sequence of the MCV-generator,
Mat. Vopr. Kriptogr., 2014, Volume 5, Issue 4, 63–71

МАТЕМАТИЧЕСКИЕ ВОПРОСЫ КРИПТОГРАФИИ
2014 Т. 5 № 4 С. 63–71

УДК 519.212.2

Пуассоновская аппроксимация для распределения числа появлений заданной цепочки знаков в выходной последовательности генератора Пола

В. Г. Михайлов

Математический институт им. В.А. Стеклова РАН, Москва

Получено 14.I.2014

Доказана предельная теорема Пуассона (с оценками точности аппроксимации) для распределения числа появлений в выходной последовательности генератора Пола заданной не допускающей самоналожения цепочки знаков.

Ключевые слова: генератор Пола, пуассоновская аппроксимация, метода Чена – Стейна.

Poisson approximation for the distribution of the frequency of a given pattern in the outcome sequence of the MCV-generator

V. G. Mikhailov

Steklov Mathematical Institute of RAS, Moscow

Abstract. Poisson limit theorem for the distribution of the number of occurrences of a given non-overlapping pattern in the output sequence of the MCV-generator is proved along with the estimate of the convergence rate.

Keywords: the MCV-generator, Poisson approximation, the Chen – Stein method.

Citation: *Mathematical Aspects of Cryptography*, 2014, vol. 5, no. 4, pp. 63–71 (Russian)

1. Введение

Генератор Пола (см. [1]) состоит из r циклических регистров сдвига взаимно простых длин $m_1, \ldots, m_r$ над кольцом вычетов по некоторому модулю M. Обозначим через $(x_0^{(k)}, \ldots, x_{m_k-1}^{(k)})$, $k = 1, \ldots, r$, заполнения регистров генератора. Условимся использовать в индексах обозначение $t(m)$ для наименьшего неотрицательного вычета числа t по модулю m. Знаки выходной последовательности генератора образуются по формуле

$$z_t = x_{t(m_1)}^{(1)} + \ldots + x_{t(m_r)}^{(r)} \mod M. \tag{1}$$

Если заполнения регистров являются независимыми случайными величинами $X_i^{(k)}$, распределенными равномерно на множестве $\{0, \ldots, M-1\}$, то выходная последовательность Z_t является случайной и имеет период длины $L = m_1 \cdot \ldots \cdot m_r$. Очевидно, что распределение отрезка выходной последовательности длины $T \leq \max\{m_1, \ldots, m_r\}$ совпадает с распределением такого же отрезка независимых случайных величин с равномерным распределением. Считается (см. [1]), что при длине выходной последовательности генератора Пола не больше $2\sqrt{L}$ эта последовательность в значительной степени обладает свойствами равновероятной случайной последовательности. При существенно больших длинах отрезок выходной последовательности генератора со случайным заполнением регистров по своим вероятностным свойствам заметно отличается от отрезка равновероятной последовательности. Так, в [2] было показано, что при $M = 2$ распределение числа единиц на полном цикле длины L выходной последовательности генератора не является биномиальным и при неограниченном увеличении длин регистров и/или числа регистров сходится к распределениям, отличным от нормального. Там же были получены оценки скорости сходимости в соответствующих предельных теоремах.

В настоящей заметке показывается, что наряду с этим любой отрезок выходной последовательности генератора со случайными равновероятными заполнениями обладает следующим важным свойством равновероятной случайной последовательности. Пусть задана некоторая цепочка (слово) $(a_1, \ldots, a_s)$ элементов множества $\{0, \ldots, M-1\}$, удовлетворяющая условиям

$$a_1 \neq a_s, (a_1, a_2) \neq (a_{s-1}, a_s), \ldots, (a_1, \ldots, a_{s-1}) \neq (a_2, \ldots, a_s). \tag{2}$$

Другими словами, цепочка $(a_1, \ldots, a_s)$ не допускает частичного самоналожения.

В работе показано, что для распределения числа появлений в отрезке выходной последовательности конкретной цепочки, не допускающей

самоналожения, справедлива оценка точности пуассоновской аппроксимации, аналогичная оценке для распределения числа появлений такой же цепочки в последовательности независимых случайных величин, распределенных равномерно на алфавите. Эта оценка содержательна для любых достаточно длинных отрезков выходной последовательности, в том числе и для ее полного периода. Из этой оценки в работе выводятся пуассоновская и «околопуассоновская» нормальная предельные теоремы, аналогичные классическим теоремам для числа появлений цепочки, не допускающей самоналожения (см., например, [3]).

Естественным является предположение о том, что аналогичными асимптотическими свойствами обладают распределения и многих других классов редких событий в выходной последовательности генератора Пола.

2. Формулировки результатов

Рассматривается генератор с r регистрами взаимно простых длин $m_1, \ldots, m_r$. Пусть $X^{(i)} = (X_0^{(i)}, \ldots, X_{m_i-1}^{(i)})$ — случайное заполнение i-го регистра, $i = 1, \ldots, r$. Все случайные величины $X_j^{(i)}$ в этих наборах независимы в совокупности и имеют равномерное распределение на множестве $\{0, \ldots, M-1\}$. Выходная последовательность генератора задается формулой

$$Z_t = X_{t(m_1)}^{(1)} + \ldots + X_{t(m_r)}^{(r)} \mod M. \tag{3}$$

Рассмотрим некоторую цепочку $a_1, \ldots, a_s$ из элементов множества $\{0, \ldots, M-1\}$, удовлетворяющую условию (2). Появлением в последовательности $\{Z_t\}$ цепочки $a_1, \ldots, a_s$ с началом в точке t естественно считать событие

$$\{X_t = a_1, \ldots, X_{t+s-1} = a_s\}.$$

Нас интересует распределение числа $\xi(s, T)$ тех появлений цепочки $a_1, \ldots, a_s$, начала которых лежат в заданном отрезке выходной последовательности длины $T \leq m_1 \ldots m_r$. В обозначении случайной величины $\xi(s, T)$ не указаны ни расположение рассматриваемого отрезка, ни выбор знаков $a_1, \ldots, a_s$, поскольку распределение величины $\xi(s, T)$ от этих параметров не зависит.

Условимся использовать обозначение $\mathcal{L}(W)$ для распределения случайной величины W и обозначение $\mathrm{Po}(\lambda)$ для распределения Пуассона с параметром λ. Напомним, что расстояние по вариации между распределениями случайных величин U и V на множестве неотрицательных

целых чисел выражается формулой

$$\rho(\mathcal{L}(U),\mathcal{L}(V)) = \frac{1}{2}\sum_{k=0}^{\infty} |\mathbf{P}\{U=k\} - \mathbf{P}\{V=k\}|\,.$$

Положим

$$\Lambda = TM^{-s}. \tag{4}$$

Если $s < \min\{m_1,\dots,m_r\}$, то $\Lambda = \mathbf{E}\xi(s,T)$.

Будем использовать обозначение $]b[$ для наименьшего целого числа, превосходящего число b или равного ему.

Теорема 1. *Пусть $M \geq 2$, $2s-1 \leq \min\{m_1,\dots,m_r\}$, $T \leq m_1\dots m_r$. Тогда*

$$\rho(\mathcal{L}(\xi(s,T)),\mathrm{Po}(\Lambda)) \leq \frac{(M+1)(2s-1)}{M^s}\left(\left]\frac{T}{m_1}\right[+ \dots + \left]\frac{T}{m_r}\right[\right). \tag{5}$$

Нетрудно показать, что оценка (5) грубее аналогичной оценки для распределения числа $\eta(s,T)$ появлений цепочки $a_1,\dots,a_s$ в последовательности из T независимых случайных величин, распределенных равномерно на множестве $\{0,\dots,M-1\}$. Для распределения $\eta(s,T)$ аналогично (5) выводится неравенство

$$\rho(\mathcal{L}(\eta(s,T)),\mathrm{Po}(\Lambda)) \leq \frac{(M+1)(2s-1)}{M^s}. \tag{6}$$

Следствие 1. *Пусть $M \geq 2$, $T \leq m_1\dots m_r$, а $m_1,\dots,m_r,s,T \to \infty$ так, что $\Lambda = TM^{-s} \to \lambda \in (0,\infty)$ и*

$$\max_{i,j}\left(\frac{\ln m_i}{m_j}\right) \to 0. \tag{7}$$

Тогда распределение случайной величины $\xi(s,T)$ сходится к распределению Пуассона с параметром λ.

Следствие 2. *Пусть $M \geq 2$, $T \leq m_1\dots m_r$, а $m_1,\dots,m_r,s,T \to \infty$ так, что $\Lambda = TM^{-s} \to \infty$ и*

$$\frac{T}{M^s}\max_{i,j}\left(\frac{\ln m_i}{m_j}\right) \to 0. \tag{8}$$

Тогда распределение случайной величины $(\xi(s,T)-\Lambda)\Lambda^{-1/2}$ сходится к стандартному нормальному распределению.

3. Доказательства

Доказательство теоремы 1. Введем случайные события

$$E_t(a) = \{Z_t = a\} = \{X^{(1)}_{t(m_1)} + \ldots + X^{(r)}_{t(m_r)} \equiv a \mod M\}.$$

Набор значений $i_1 = t(m_1), \ldots, i_r = t(m_r)$ и условие $0 \le t < m_1 \ldots m_r$ однозначно определяют величину t. Поэтому для индикатора появления цепочки $a_1, \ldots, a_s$ в момент t можно использовать обозначение

$$W(i_1, \ldots, i_r) = I\{E_t(a_1) \cap \ldots \cap E_{t+s-1}(a_s)\},$$

где $I\{E\}$ — индикатор случайного события E.

Введем множество

$$\Gamma(T) = \{(t(m_1), \ldots, t(m_r)) : \ t = 0, \ldots, T-1\}.$$

Используя тот факт, что распределение случайной величины $\xi(s,T)$ не зависит от расположения отрезка наблюдений, заключаем, что задача свелась к изучению распределения случайной величины

$$\xi'(s,T) = \sum_{(i_1,\ldots,i_r)\in\Gamma(T)} W(i_1, \ldots, i_r). \tag{9}$$

Это распределение совпадает с распределением случайной величины $\xi(s,T)$.

Мы воспользуемся оценкой локального варианта метода Чена – Стейна (см. [4], [5] и ссылки в этих работах). Пусть

$$\Gamma^{ind}_{(i_1,\ldots,i_r)}(T) =$$

$$= \{(i'_1, \ldots, i'_r) \in \Gamma(T) : \ \min\{|i'_k - i_k|, |m_k - (i'_k - i_k)|\} \ge s, \ k = 1, \ldots, r\}.$$

Положим

$$\Gamma_{(i_1,\ldots,i_r)}(T) = (\Gamma(T)\backslash\{(i_1, \ldots, i_r)\})\backslash\Gamma^{ind}_{(i_1,\ldots,i_r)}(T). \tag{10}$$

Нетрудно проверить, что случайная величина $W(i_1, \ldots, i_r)$ не зависит от совокупности случайных величин

$$\{W(i'_1, \ldots, i'_r) : \ (i'_1, \ldots, i'_r) \in \Gamma^{ind}_{(i_1,\ldots,i_r)}(T)\}.$$

Поэтому оценка по методу Чена – Стейна (см., например, теорему 6.1 в [5]) принимает вид

$$\rho(\mathcal{L}(\xi'(s,T)), \mathrm{Po}(\Lambda)) \le$$

$$\leq \frac{1-e^{-\Lambda}}{\Lambda}\left(\sum_{(i_1,\ldots,i_r)\in\Gamma(T)}\mathbf{E}W(i_1,\ldots,i_r)\big(\mathbf{E}W(i_1,\ldots,i_r)+\mathbf{E}U(i_1,\ldots,i_r)\big)+\right.$$

$$\left.+\sum_{(i_1,\ldots,i_r)\in\Gamma(T)}\mathbf{E}W(i_1,\ldots,i_r)U(i_1,\ldots,i_r)\right), \tag{11}$$

где

$$U(i_1,\ldots,i_r)=\sum_{(i'_1,\ldots,i'_r)\in\Gamma_{(i_1,\ldots,i_r)}(T)}W(i'_1,\ldots,i'_r). \tag{12}$$

Лемма 1. *Пусть* $2s-1\leq\min\{m_1,\ldots,m_r\}$. *Тогда*

$$|\{(i_1,\ldots,i_r)\}\cup\Gamma_{(i_1,\ldots,i_r)}(T)|=|\Gamma(T)\backslash\Gamma^{ind}_{(i_1,\ldots,i_r)}(T)|\leq$$

$$\leq(2s-1)\left(\left]\frac{T}{m_1}\right[+\ldots+\left]\frac{T}{m_r}\right[\right). \tag{13}$$

Доказательство. Проекции множества $\{(i_1,\ldots,i_r)\}\cup\Gamma_{(i_1,\ldots,i_r)}(T)$ на «окружности» $\{0,\ldots,m_1-1\},\ldots,\{0,\ldots,m_r-1\}$ представляют собой отрезки из $2s-1$ подряд идущих точек. Последовательность $t(m_1)$ с ростом параметра t от 0 до $T-1$ проходит все точки множества $\{0,\ldots,m_1-1\}$ подряд (по кругу) не более $]T/m_1[$ раз. В этом процессе значение $t(m_1)$ попадет на отрезок из $2s-1$ точек не более $(2s-1)]T/m_1[$ раз. Это число является оценкой сверху для числа тех элементов множества $\{(i_1,\ldots,i_r)\}\cup\Gamma_{(i_1,\ldots,i_r)}(T)$, которые попали в это множество за счет нарушения условия $\min\{|i'_1-i_1|,|m_1-(i'_1-i_1)|\}\geq s$. То же самое выполнено для остальных проекций. Из этих оценок следует неравенство (13). Лемма доказана. □

Продолжим доказательство теоремы. Из определений вытекает, что при всех $(i_1,\ldots,i_r)\in\Gamma(T)$

$$\mathbf{E}W(i_1,\ldots,i_r)=\frac{1}{M^s}. \tag{14}$$

Из (13) и (14) следует, что

$$\mathbf{E}W(i_1,\ldots,i_r)+\mathbf{E}U(i_1,\ldots,i_r)=|\{(i_1,\ldots,i_r)\}\cup\Gamma_{(i_1,\ldots,i_r)}(T)|\frac{1}{M^s}\leq$$

$$\leq\frac{2s-1}{M^s}\left(\left]\frac{T}{m_1}\right[+\ldots+\left]\frac{T}{m_r}\right[\right).$$

Поэтому первая сумма в правой части (11) допускает оценку

$$\sum_{(i_1,\dots,i_r)\in\Gamma(T)} \mathbf{E}W(i_1,\dots,i_r)(\mathbf{E}W(i_1,\dots,i_r)+\mathbf{E}U(i_1,\dots,i_r)) \le$$

$$\le \frac{\Lambda(2s-1)}{M^s}\left(\left]\frac{T}{m_1}\right[+ \dots + \left]\frac{T}{m_r}\right[\right). \tag{15}$$

Оценим вторую сумму в правой части (11).

Лемма 2. *Пусть* $2s-1 \le \min\{m_1,\dots,m_r\}$,

$$(i'_1,\dots,i'_r)\in\Gamma_{(i_1,\dots,i_r)}(T), \quad (i_1,\dots,i_r)\in\Gamma(T).$$

Тогда $\mathbf{E}W(i_1,\dots,i_r)W(i'_1,\dots,i'_r) \le M^{1-2s}$.

К лемме 2 мы вернемся в конце статьи, а пока продолжим доказательство теоремы. Согласно (12), (13) и оценке леммы 2

$$\sum_{(i_1,\dots,i_r)\in\Gamma(T)} \mathbf{E}W(i_1,\dots,i_r)U(i_1,\dots,i_r) < \sum_{(i_1,\dots,i_r)\in\Gamma(T)} |\Gamma_{(i_1,\dots,i_r)}(T)|\frac{1}{M^{2s-1}} \le$$

$$\le \frac{\Lambda(2s-1)}{M^{s-1}}\left(\left]\frac{T}{m_1}\right[+ \dots + \left]\frac{T}{m_r}\right[\right). \tag{16}$$

Из (11), (15) и (16) следует, что

$$\rho(\mathcal{L}(\xi'(s,T)),\mathrm{Po}(\Lambda)) \le$$

$$\le \frac{(M+1)(2s-1)}{M^s}\left(\left]\frac{T}{m_1}\right[+ \dots + \left]\frac{T}{m_r}\right[\right). \tag{17}$$

Из (17) получаем неравенство (5). Теорема доказана. □

Доказательство следствий 1 и 2. Из (4), неравенства $T \le m_1 \dots m_r$ и условия $\Lambda \to \lambda \in (0,\infty]$ следует, что $s = O(\ln m_1 + \dots + \ln m_r)$, а правая часть в (5) допускает при переходе к пределу оценку

$$O\left(\frac{sT}{M^s}\left(\frac{1}{m_1}+\dots+\frac{1}{m_r}\right)\right) = O\left(\Lambda \max_{i,j}\left(\frac{\ln m_i}{m_j}\right)\right). \tag{18}$$

Поэтому по условиям следствий она стремится к нулю, и из условия $\Lambda \to \lambda \in (0,\infty)$ и теоремы 1 вытекает сходимость распределения величины $\xi(s,T)$ к распределению Пуассона $\mathrm{Po}(\lambda)$. Следствие 1 доказано.

В свою очередь, в случае $\Lambda \to \infty$ из сходимости к нулю выражений (18) и теоремы 1 вытекает сходимость к нулю расстояния по вариации между распределением случайной величины $\xi(s,T)$ и распределением Пуассона $\mathrm{Po}(\Lambda)$ с растущим к бесконечности параметром Λ. Последнее распределение при указанных в формулировке центрировке и нормировке стремится к стандартному нормальному распределению. Поэтому выполнено утверждение следствия 2. Итак, оба следствия доказаны. $\square$

Доказательство леммы 2. Пусть $1 \le t < t+l \le m_1 \dots m_r - 1$. Событие $B_t \cap B_{t+l}$ можно трактовать как выполнение системы из $2s$ сравнений

$$X^{(1)}_{t(m_1)} + \dots + X^{(r)}_{t(m_r)} \equiv a_t \mod M, \quad t = 1, \dots, s, \tag{19}$$

$$X^{(1)}_{(t+l)(m_1)} + \dots + X^{(r)}_{(t+l)(m_r)} \equiv a_t \mod M, \quad t = 1, \dots, s. \tag{20}$$

Система (19) – (20) обладает следующими свойствами.

1) Согласно условию $2s - 1 \le \min\{m_1, \dots, m_r\}$ множества неизвестных в уравнениях каждой из систем (19) и (20) не пересекаются.

2) Комбинация, выражающая уравнение системы (19) через уравнения системы (20), если она существует, может быть получена лишь из одного уравнения системы (20) (иначе в ней оказалось бы более r ненулевых слагаемых). То же самое верно и для уравнения системы (20). А значит, возможны лишь следующие варианты:

2а) системы (19) и (20) имеют общее уравнение,

2б) ни одно уравнение системы (19) не может быть выражено через уравнения системы (20), и обратно, ни одно уравнение системы (20) не может быть выражено через уравнения системы (19) – (20).

В случае 2а) события B_t и B_{t+l} несовместны, так как цепочка $a_1, \dots, a_s$ не может наложиться сама на себя. Следовательно (здесь и ниже в (22) $i_k = t(m_k)$, $i'_k = (t+l)(m_k)$, $k = 1, \dots, r$),

$$\mathbf{E} W(i_1, \dots, i_r) W(i'_1, \dots, i'_r) = 0. \tag{21}$$

В случае 2б) каждое следующее уравнение системы (19) – (20) (уравнения можно брать в любом порядке) содержит хотя бы одно неизвестное, не входящее в предыдущие уравнения. Это означает, что, по крайней мере, $2s - 1$ неизвестных системы (19) – (20) однозначно определяются остальными неизвестными. Следовательно,

$$\mathbf{E} W(i_1, \dots, i_r) W(i'_1, \dots, i'_r) < \mathbf{P}\{B_t \cap B_{t'}\} \le M^{1-2s}. \tag{22}$$

Из (21) и (22) вытекает, что неравенство (22) выполнено в обоих случаях. Лемма 2 доказана. $\square$

Автор признателен А. М. Зубкову за полезные замечания.

Список литературы

[1] Pohl P., "Description of MCV, a pseudo-random number generator", *Scand. Actuarial J.*, 1976, № 1, 1–14.

[2] Меженная Н. М., Михайлов В. Г., "Оценки и предельные теоремы нормального типа для числа единиц в выходной последовательности генератора Пола", *Математические вопросы криптографии*, **4**:4 (2013), 95–107.

[3] Chryssaphinou O., Papastavridis S., "A limit theorem for the number of non-overlapping occurrence of a pattern in a sequence of independent trials", *J. Appl. Prob.*, **25**:2 (1988), 428–431.

[4] Barbour A. D., Holst L., Janson S., *Poisson Approximation*, Oxford: Oxford University Press, 1992.

[5] Михайлов В. Г., "Явные оценки в предельных теоремах для сумм случайных индикаторов", *Обозр. прикл. и промышл. матем.*, **1**:4 (1994), 580–617.