Math-Net.Ru

Общероссийский математический портал

В. Г. Михайлов, Об асимптотическом поведении вероятности наличия в последовательности эквивалентных цепочек с нетривиальной структурой, *Дискрет. матем.*, 2008, том 20, выпуск 4, 113–119

DOI: <https://doi.org/10.4213/dm1031>

Использование Общероссийского математического портала Math-Net.Ru подразумевает, что вы прочитали и согласны с пользовательским соглашением
<http://www.mathnet.ru/rus/agreement>

Параметры загрузки:

IP: 54.84.234.179

26 апреля 2023 г., 11:23:09

Об асимптотическом поведении вероятности наличия в последовательности эквивалентных цепочек с нетривиальной структурой

© 2008 г. В. Г. Михайлов

В последовательности достаточной длины из дискретных случайных величин, как правило, есть s -цепочки с нетривиальной структурой, то есть цепочки, имеющие в своем составе хотя бы одно повторение знаков. Рассмотрен случай, когда последовательность состоит из $n + s - 1$ независимых случайных величин, принимающих значения $1, \dots, N$ с равными вероятностями. Показано, что при $n \rightarrow \infty$ и $ns^3N^{-2} \rightarrow 0$ вероятность наличия в этой последовательности s -цепочек с одинаковой нетривиальной структурой есть $1 - (1 + n/N)^s e^{-sn/N} (1 + o(1))$.

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проект 05-01-00035, программой президента Российской Федерации поддержки ведущих научных школ, проект НШ 4129.2006.1, и программой РАН «Современные проблемы теоретической математики».

1. Введение

Эквивалентными цепочками называют такие цепочки $(x_1, \dots, x_s)$ и $(y_1, \dots, y_s)$, для которых найдется перестановка $\sigma = (\sigma(1), \dots, \sigma(N))$ знаков алфавита $\{1, \dots, N\}$, при которой $(y_1, \dots, y_s) = (\sigma(x_1), \dots, \sigma(x_s))$. Эквивалентные цепочки всегда имеют одинаковую структуру, а именно, если $x_i = x_j$, то и $y_i = y_j$, и наоборот; подробности см. в [1]. Согласно этому определению все цепочки одинаковой длины, не имеющие повторений знаков, эквивалентны между собой. Поэтому содержательные теоремы об асимптотических свойствах распределений числа пар (числа троек и т. д.) эквивалентных цепочек в случайной последовательности удается получать только в случае $s \rightarrow \infty$ (см. [1]).

В настоящей работе предлагается рассматривать эквивалентные цепочки лишь с нетривиальной структурой, то есть цепочки, имеющие в своем составе хотя бы одно повторение знаков. Тогда традиционные задачи об асимптотическом поведении числа эквивалентных цепочек (например, вопрос о числе пар эквивалентных цепочек в последовательности) становятся содержательными для цепочек любой длины, большей единицы. Однако при большом алфавите в короткой последовательности могут отсутствовать эквивалентные цепочки заданной длины с нетривиальной структурой. Поэтому естественным является вопрос о соотношении параметров, при котором вероятность наличия таких цепочек в случайной последовательности достаточно велика. Решению этой задачи и посвящена настоящая заметка.

Рассмотрим последовательность $X_1, X_2, \dots$ независимых случайных величин, принимающих значения $1, \dots, N$ с вероятностями $p_1, \dots, p_N$ соответственно.

Пусть событие $E = E(n, s)$ состоит в том, что в отрезке $X(n) = (X_1, \dots, X_n)$ имеются эквивалентные s -цепочки с нетривиальной структурой.

Введем при $r = 2, 3$ обозначение

$$P_r = p_1^r + \dots + p_N^r.$$

Теорема 1. Пусть $n, N \rightarrow \infty$, а параметр s и вероятности $p_1, \dots, p_N$ меняются так, что

$$ns^2(sP_2^2 + P_3) \rightarrow 0. \quad (1)$$

Тогда

$$\mathbf{P}\{E(n, s)\} = 1 - e^{-(s-1)nP_2}(1 + nP_2)^{s-1} + o(1). \quad (2)$$

Приведем некоторые следствия из этого утверждения.

Следствие 1. Пусть $n, N \rightarrow \infty$, параметр $s \geq 2$ фиксирован, а вероятности $p_1, \dots, p_N$ меняются так, что $nP_2 \rightarrow 0$. Тогда

$$\mathbf{P}\{E(n, s)\} \rightarrow 0.$$

Следствие 2. Пусть $n, N \rightarrow \infty$, параметр $s \geq 2$ фиксирован, а вероятности $p_1, \dots, p_N$ меняются так, что $nP_2 \rightarrow \rho$, $0 < \rho < \infty$. Тогда

$$\mathbf{P}\{E(n, s)\} \rightarrow 1 - (1 + \rho)^{s-1} e^{(1-s)\rho}.$$

Следствие 3. Пусть $n, N \rightarrow \infty$, параметр $s \geq 2$ фиксирован, а вероятности $p_1, \dots, p_N$ меняются так, что $nP_2 \rightarrow \infty$. Тогда

$$\mathbf{P}\{E(n, s)\} \rightarrow 1.$$

Следствие 4. Пусть $n, N, s \rightarrow \infty$, а вероятности $p_1, \dots, p_N$ меняются так, что выполнено условие (1) и $\sqrt{sn}P_2 \rightarrow 0$. Тогда

$$\mathbf{P}\{E(n, s)\} \rightarrow 0.$$

Следствие 5. Пусть $n, N, s \rightarrow \infty$, а вероятности $p_1, \dots, p_N$ меняются так, что выполнено условие (1) и $\sqrt{sn}P_2 \rightarrow \rho$, $0 < \rho < \infty$. Тогда

$$\mathbf{P}\{E(n, s)\} \rightarrow 1 - \exp\{-\rho^2/2\}.$$

Следствие 6. Пусть $n, N, s \rightarrow \infty$, а вероятности $p_1, \dots, p_N$ меняются так, что выполнено условие (1) и $\sqrt{sn}P_2 \rightarrow \infty$. Тогда

$$\mathbf{P}\{E(n, s)\} \rightarrow 1.$$

Замечание 1. В последнем утверждении условие (1) можно заменить более слабым условием

$$ns^2(sP_2^2 + P_3) = O(1). \quad (3)$$

2. Доказательство теоремы 1

Будем говорить, что s -цепочка является цепочкой первого типа, если в ее записи используются $s - 1$ различных знаков алфавита. Если в записи s -цепочки использовано $s - 2$ или меньше знаков, то назовем ее цепочкой второго типа. Доказательство теоремы использует то свойство, что в пределе цепочек второго типа в последовательности в данном случае нет.

Наряду с $E = E(n, s)$ рассмотрим события

$$\begin{aligned} D_1 &= D_1(n, s) = \{\text{в отрезке } X(n) \text{ присутствуют } s\text{-цепочки второго типа}\}, \\ D_2 &= D_2(n, s) = \{\text{в отрезке } X_{n+s+2}, \dots, X_n \text{ имеется повторение знаков}\}. \end{aligned}$$

Определим случайные величины

$$\eta_k = \sum_{i=1}^{n-s+1} I\{X_i = X_{i+k}\}, \quad k = 1, \dots, s-1, \quad (4)$$

и событие

$$E'_1(n, s) = \{\eta_1 \geq 2\} \cup \dots \cup \{\eta_{s-1} \geq 2\}.$$

При отсутствии в отрезке $X(n)$ s -цепочек второго типа и одновременном отсутствии повторений знаков в отрезке $X_{n+s+2}, \dots, X_n$ события $E(n, s)$ и $E'(n, s)$ эквивалентны. Поэтому

$$\begin{aligned} |\mathbf{P}\{E(n, s)\} - \mathbf{P}\{E'(n, s)\}| &\leq \mathbf{P}\{D_1(n, s) \cup D_2(n, s)\} \\ &= \mathbf{P}\{D_1(n, s) \setminus D_2(n, s)\} + \mathbf{P}\{D_2(n, s)\}. \end{aligned} \quad (5)$$

Лемма 1. Пусть $n \geq s \geq 3$. Тогда

$$\mathbf{P}\{D_1(n, s) \setminus D_2(n, s)\} \leq \frac{1}{2}(n-s+1)(s-1)_2((s-3)P_2^2 + P_3), \quad (6)$$

$$\mathbf{P}\{D_2(n, s)\} \leq \frac{1}{2}(s-1)_2 P_2^2. \quad (7)$$

Пусть $\pi_1, \dots, \pi_{s-1}$ — независимые случайные величины, распределенные по закону Пуассона с параметром

$$\lambda = \mathbf{E}\eta_1 = \dots = \mathbf{E}\eta_{s-1} = (n-s+1)P_2.$$

Расстояние по вариации между распределениями случайных величин ξ и ζ будем обозначать $d(\xi, \zeta)$. Напомним, что для распределений на множестве неотрицательных целых чисел

$$\rho(\xi, \zeta) = \frac{1}{2} \sum_{k=0}^{\infty} |\mathbf{P}\{\xi = k\} - \mathbf{P}\{\zeta = k\}|.$$

Лемма 2. Для расстояния по вариации между распределениями случайных векторов $\eta = (\eta_1, \dots, \eta_{s-1})$ и $\pi = (\pi_1, \dots, \pi_{s-1})$ справедливо неравенство

$$d(\eta, \pi) < 2(s-1)^2(n+s-1)(3P_2^2 + P_3). \quad (8)$$

Согласно определениям

$$\left| \mathbf{P}\{E'(n, s)\} - 1 + \prod_{k=1}^{s-1} \mathbf{P}\{\Pi_k \leq 1\} \right| \leq d(\eta, \Pi).$$

Поэтому из соотношений (5) и лемм 1 и 2 вытекают оценки

$$\begin{aligned} & \left| \mathbf{P}\{E(n, s)\} - 1 + \prod_{k=1}^{s-1} ((1 + \lambda)e^{-\lambda}) \right| \\ & \leq \mathbf{P}\{D_1(n, s) \setminus D_2(n, s)\} + \mathbf{P}\{D_2(n, s)\} + d(\eta, \pi) \\ & < \frac{1}{2}(n - s + 1)(s - 1)_2((s - 3)P_2^2 + P_3) \\ & \quad + \frac{1}{2}(s - 1)_2P_2^2 + 2(s - 1)^2(n + s - 1)(3P_2^2 + P_3) \\ & < \frac{1}{2}(n - s + 1)(s - 1)^2((s + 9)P_2^2 + 5P_3) + \frac{1}{2}(s - 1)_2P_2^2. \end{aligned} \quad (9)$$

Согласно (9) и (1)

$$\mathbf{P}\{E(n, s)\} = 1 - e^{-(s-1)(n-s+1)P_2} (1 + (n - s + 1)P_2)^{s-1} + o(1). \quad (10)$$

Из условия (1) и неравенств $n \geq s$, $P_3 \geq P_2^2$ следует, что в условиях теоремы

$$(s - 1)(n - s + 1)P_2 = (s - 1)nP_2 + o(1).$$

Поэтому из (10) вытекает оценка (2). Теорема 1 доказана.

Несколько слов о доказательстве следствий из теоремы 1.

Следствия 4, 5 и 6 вытекают непосредственно из теоремы 1. При выводе следствий 1 и 2 также можно воспользоваться теоремой 1, поскольку ее условие (1) выполняется при фиксированном $s \geq 2$ и $nP_2 = O(1)$.

Следствие 3 выводится из следствия 2 с помощью соотношения

$$\mathbf{P}\{E(n_1, s)\} \leq \mathbf{P}\{E(n_2, s)\}, \quad n_1 \leq n_2. \quad (11)$$

Неравенство (11) используется также для того, чтобы заменить в следствии 6 условие (1) условием (3) (см. замечание 1).

3. Доказательства вспомогательных утверждений

Доказательство леммы 1. Обозначим через A_t событие, состоящее в том, что в цепочке $Y_t = (X_t, \dots, X_{t+s-1})$ есть по крайней мере две пары одинаковых знаков (причем в одну из них входит X_t), а через B_t событие, состоящее в том, что в цепочке Y_t есть по крайней мере одна тройка одинаковых знаков (причем в нее входит X_t). Тогда

$$D_1(n, s) \setminus D_2(n, s) \subseteq \bigcup_{t=1}^{n-s+1} (A_t \cup B_t). \quad (12)$$

Используя соотношение

$$A_t \subseteq \bigcup_{t < i < j < k \leq t+s-1} (E_{t,1}(i, j, k) \cup E_{t,2}(i, j, k) \cup E_{t,3}(i, j, k)),$$

где

$$\begin{aligned} E_{t,1}(i, j, k) &= \{X_t = X_i, X_j = X_k\}, \\ E_{t,2}(i, j, k) &= \{X_t = X_j, X_i = X_k\}, \\ E_{t,3}(i, j, k) &= \{X_t = X_k, X_i = X_j\}, \end{aligned}$$

и соотношение

$$B_t \subseteq \bigcup_{t < i < j \leq t+s-1} \{X_t = X_i = X_j\},$$

нетрудно показать, что

$$\begin{aligned} \mathbf{P}\{A_i\} &\leq \frac{1}{2}(s-1)_3 P_2^2, \\ \mathbf{P}\{B_i\} &\leq \frac{1}{2}(s-1)_2 P_3. \end{aligned} \quad (13)$$

Из (12) и (13) следует неравенство (6).

Аналогично доказывается неравенство (7). Как нетрудно проверить,

$$D_2(n, s) \subseteq \bigcup_{n-s+2 \leq i < j \leq n} \{X_i = X_j\}.$$

Поэтому

$$\mathbf{P}\{D_2(n, s)\} \leq (s-1)_2 \mathbf{P}\{X_i = X_j\} = (s-1)_2 P_2.$$

Лемма 1 доказана.

Доказательство леммы 2. Рассмотрим наборы случайных величин

$$\begin{aligned} Z &= (I\{X_t = X_{t+k}\}, t = 1, \dots, n-s+1, k = 1, \dots, s-1), \\ \Pi &= (\pi(t, k), t = 1, \dots, n-s+1, k = 1, \dots, s-1), \end{aligned}$$

где $\pi(t, k)$ — независимые случайные величины, распределенные по закону Пуассона с параметром

$$\mathbf{P}\{X_t = X_{t+k}\} = P_2.$$

Случайный вектор $\eta = (\eta_1, \dots, \eta_{s-1})$ является функцией вектора Z , которая задается формулой (4) (будем использовать обозначение $\eta = F(Z)$). Распределение случайного вектора $\pi = (\pi_1, \dots, \pi_{s-1})$ совпадает с распределением вектора $F(\Pi)$. Поэтому из свойств расстояния по вариации получаем неравенство

$$d(\eta, \Pi) \leq d(Z, \Pi). \quad (14)$$

Покажем, что

$$d(Z, \Pi) < 2(s-1)^2(n-s+1)(3P_2^2 + P_3). \quad (15)$$

Воспользуемся тем, что случайные события $\{X_t = X_{t+k}\}$ определяются независимыми случайными величинами. Согласно теореме 10А книги [2] в этом случае

$$d(Z, \Pi) \leq S_1 + S_2, \quad (16)$$

где S_1 есть сумма произведений вероятностей всех тех событий $\{X_t = X_{t+k}\}$ и $\{X_u = X_{u+l}\}$, которые зависят между собой (в эту сумму входят и квадраты вероятностей событий $\{X_t = X_{t+k}\}$), а S_2 представляет собой сумму совместных вероятностей всех различных событий $\{X_t = X_{t+k}\}$ и $\{X_u = X_{u+l}\}$, которые зависят между собой. Заметим, что в общем случае цитируемая оценка имеет более сложный вид. В нашем случае сумма S_1 допускает оценку

$$S_1 < \sum_{k=1}^{s-1} \sum_{t=1}^{n-s+1} \sum_{l=1}^{s-1} \sum_{u=t, t+k, t-l, t-l+k} \mathbf{P}\{X_t = X_{t+k}\} \mathbf{P}\{X_u = X_{u+l}\}. \quad (17)$$

Считаем, что в сумму входят лишь те слагаемые, для которых значения индексов $t+k$, $t-l$, $t-l+k$ принадлежат множеству $\{1, \dots, n\}$. Из (17) следует, что

$$S_1 < 4(s-1)^2(n-s+1)P_2^2. \quad (18)$$

Аналогичным образом оцениваем сумму S_2 :

$$S_2 \leq S_{2,1} + S_{2,2} + S_{2,3} + S_{2,4}, \quad (19)$$

где

$$\begin{aligned} S_{2,1} &= \sum_{k=1}^{s-1} \sum_{t=1}^{n-s+1} \sum_{l \neq k} \mathbf{P}\{X_t = X_{t+k}, X_t = X_{t+l}\}, \\ S_{2,2} &= \sum_{k=1}^{s-1} \sum_{t=1}^{n-s+1} \sum_{l \neq k} \mathbf{P}\{X_t = X_{t+k}, X_{t+k-l} = X_{t+k}\}, \\ S_{2,3} &= \sum_{k=1}^{s-1} \sum_{t=1}^{n-s+1} \sum_{l=1}^{s-1} \mathbf{P}\{X_t = X_{t+k} = X_{t+k+l}\}, \\ S_{2,4} &= \sum_{k=1}^{s-1} \sum_{t=1}^{n-s+1} \sum_{l=1}^{s-1} \sum_{u=t+k, t-l} \mathbf{P}\{X_{t-l} = X_t = X_{t+k}\}. \end{aligned}$$

В этих формулах также считаем, что в суммы входят лишь те слагаемые, для которых значения индексов принадлежат множеству $\{1, \dots, n\}$.

Используя равенства

$$\begin{aligned} \mathbf{P}\{X_t = X_{t+k}\} &= P_2, \\ \mathbf{P}\{X_t = X_{t+k} = X_{t+k+l}\} &= P_3, \end{aligned}$$

получаем оценки

$$\begin{aligned} S_{2,r} &< P_2^2 \sum_{k=1}^{s-1} \sum_{t=1}^{n-s+1} (s-2) \leq (s-1)_2(n-s+1)P_2^2, \quad r = 1, 2, \\ S_{2,2} &< P_3 \sum_{k=1}^{s-1} \sum_{t=1}^{n-s+1} (s-1) \leq (s-1)^2(n-s+1)P_3, \quad r = 3, 4. \end{aligned}$$

Подставив эти оценки в (19), получаем неравенство

$$S_2 < 2(s-1)^2(n-s+1)(P_2^2 + P_3). \quad (20)$$

Из (16), (18) и (20) следует (15). В свою очередь, из (15) и (14) следует оценка (8). Лемма 2 доказана.

Автор благодарен рецензенту за полезные замечания.

Список литературы

1. Михайлов В. Г., Шойтов А. М., Структурная эквивалентность цепочек в случайных дискретных последовательностях. *Дискретная математика* (2003) **15**, №4, 7–34.
2. Barbour A. D., Holst L., Janson S., *Poisson approximation*. Oxford Univ. Press, Oxford, 1992.

Статья поступила 28.11.2006.

Переработанный вариант поступил 15.09.2008.