Записки научных
семинаров ПОМИ
Том 510, 2022 г.

,

М. А. Лифшиц, И. М. Лялинов

ДВЕ ПРЕДЕЛЬНЫЕ ТЕОРЕМЫ О ПЕРЕСЕЧЕНИЯХ СЛУЧАЙНЫХ МНОЖЕСТВ ЦИПФА

§1. Введение и основные результаты

1.1. Случайные множества Ципфа. Эмпирический закон Ципфа (G. K. Zipf [7,8], см. также [5]) заключается в том, что самое употребительное слово языка встречается вдвое чаще следующего по частоте, втрое чаще следующего, и т.д. Ципф также применял свой закон к таким объектам экономики, как население городов, размеры индивидуальных состояний и др.[1]

Среди относительно недавних работ упомянем приложения в лингвистике [9,12] и ряд работ, посвящённых исследованию вероятностных и комбинаторных механизмов, приводящих к закону Ципфа [13–17].

Пусть имеется счётное множество объектов $\mathcal{W} = \{w_1, w_2, \dots\}$. *Случайным множеством Ципфа* с параметрами $a \in (0,1], \alpha \in (0,1]$ назовём случайное подмножество $A \subset \mathcal{W}$, для которого события $\{w_m \in A\}$ независимы и $\mathbb{P}(w_m \in A) = \frac{a}{m^\alpha}$.

В дальнейшем для краткости обозначений мы отождествляем $\mathcal{W}$ с множеством натуральных чисел $\mathbb{N}$, а w_m с m.

Отметим, что условие $a \leqslant 1$ совершенно несущественно и выбрано только для удобства работы с произвольными натуральными индексами. Все результаты статьи остаются верными для случайных множеств, удовлетворяющих асимптотическому условию

$$\lim_{m\to\infty} m^\alpha \mathbb{P}(w_m \in A) = a$$

с произвольным $a > 0$.

Ключевые слова: предельные теоремы, максимальные пересечения, множества Ципфа, случайные множества, наиболее редкий элемент пересечения.

Работа поддержана грантом РНФ No. 21-11-00047.

[1] В лингвистическом аспекте Ципф существенно опирался на данные, накопленные стенографами, в числе которых Ж.-Б. Эсту (J.-B. Estoup [2]) уже в начале XX века отмечал гиперболический характер наблюдаемых частот, см. [4, 6], а применительно к размеру городов закон сформулировал ещё физик Ф. Ауэрбах (F. Auerbach, 1913, [1]).

1.2. Максимальные элементы пересечений множеств Ципфа. Пусть A – случайное множество Ципфа с параметрами $a \in (0,1], \alpha \in (0,1]$. Очевидно, что такое множество почти наверное бесконечно (но его пересечение с независимым случайным множеством Ципфа с параметрами $x \in (0,1]$, $\beta \in (1-\alpha, 1]$ почти наверное конечно). Поэтому, прежде всего, нас будут интересовать характеристики пересечений случайных множеств Ципфа с указанным выше ограничением на параметры.

Будем писать $c_n \sim d_n$, если $\lim\limits_{n\to\infty} \frac{c_n}{d_n} = 1$ и $c_n \preceq d_n$, если выполнено $\limsup\limits_{n\to\infty} \frac{c_n}{d_n} \leqslant 1$.

Задача о максимальных характеристиках пересечений множеств возникает в математическом моделировании функционирования баз данных. Представим себе, что база данных состоит из документов $X_1, \ldots, X_n$, каждый из которых можно охарактеризовать набором слов (или ключевых слов), и имеется запрос A, который также характеризуется набором слов. Релевантность документа X запросу A можно определять различными способами. В работах [3,11] релевантность понимается как количество общих слов в запросе и документе. В них получены законы больших чисел для максимальной меры релевантности при объёме базы, стремящемся к бесконечности.

В данной работе релевантность определяется другим естественным способом – как номер самого редкого (имеющего максимальный номер) общего слова в A и X. Задачей выполнения запроса является нахождение наиболее релевантного документа, т.е.

$$\operatorname{argmax}_{1\leqslant j\leqslant n} V_j, \tag{1}$$

где через V_j обозначен максимальный элемент множества $A \cap X_j$.

Разумеется, представляет интерес и величина самого редкого общего элемента

$$V^n := \max_{1\leqslant j\leqslant n} V_j,$$

например, для того, чтобы оценивать качество приближённых решений задачи (1). Отметим, что задача о максимальном пересечении имеет многочисленные приложения: в интернет-поиске, в рекомендательных системах, онлайн-рекламе и т.д.

Следующая предельная теорема характеризует распределение самого редкого общего элемента пересечения в модели, где и документы, и запрос рассматриваются как случайные множества Ципфа.

Теорема 1. *Пусть* $a, x \in (0,1]$, $\alpha \in (0,1]$, $\beta \in (1-\alpha, 1]$.

Тогда при $\alpha < 1$ *для любого* $r \geqslant 0$ *и* $n \to \infty$ *выполнено*

$$\mathbb{P}\left(V^n < r\, n^{\frac{1}{\alpha+\beta-1}}\right) \to F_{\alpha,\beta}(r) := \exp\left(-\frac{ax}{(\alpha+\beta-1)\, r^{\alpha+\beta-1}}\right),$$

а при $\alpha = 1$ *для любого* $r \geqslant 0$ *и* $n \to \infty$ *выполнено*

$$\mathbb{P}\left(V^n < r\, n^{\frac{1}{\beta}}\right) \to F_{1,\beta}(r) := \exp\left(-\frac{aI(x/r^\beta)}{\beta}\right),$$

где

$$I(v) := \int\limits_0^v \frac{1-e^{-w}}{w}\, dw, \qquad v \geqslant 0.$$

Отметим, что при $\alpha < 1$ предельное распределение $F_{\alpha,\beta}$ есть не что иное, как распределение Фреше, хорошо известное в теории предельных теорем для максимумов независимых одинаково распределённых величин [10]. Это не слишком удивительно, так как величины V_j *при фиксированном* A независимы и одинаково распределены. Однако распределение $F_{1,\beta}$ совсем не относится к классическим. Авторы сталкиваются с ним впервые.

Представляют интерес и меры релевантности, промежуточные между номером самого редкого общего элемента и количеством общих элементов. Пусть $(q_m)_{m\geqslant 1}$ – последовательность неотрицательных чисел (весов). Определим связанную с ними интегральную меру релевантности документа X запросу A как

$$Q(A,X) := \sum_{m \in A \cap X} q_m.$$

Если $q_m = 1$ при всех $m \geqslant 1$, то $Q(A,X)$ – количество общих элементов. Если же $q_m \nearrow \infty$ достаточно быстро, то асимптотическое поведение максимальной интегральной меры релевантности определяется асимптотикой самого редкого общего элемента. Сейчас мы установим этот факт для экспоненциально растущих весов.

Пусть $b > 0, q_m := e^{bm}, Q_j := Q(A, X_j), Q^n := \max_{1\leqslant j\leqslant n} Q_j$.

Теорема 2. *Пусть* $a, x \in (0,1]$, $\alpha \in (0,1]$, $\beta \in (1-\alpha, 1]$.

Тогда для любого $r \geqslant 0$ *при* $n \to \infty$ *выполнено*

$$\mathbb{P}\left(\log Q^n \leqslant b\, r\, n^\gamma\right) \to F_{\alpha,\beta}(r),$$

где функции распределения $F_{\alpha,\beta}(\,\cdot\,)$ *определены в теореме* 1 *и*

$$\gamma := (\alpha+\beta-1)^{-1}.$$

В заключение отметим, что случай степенных весов ещё предстоит исследовать.

Доказательство теоремы 1. Пользуясь независимостью A и X_j, а также независимостью появления отдельных элементов в множествах Ципфа, преобразуем интересующую нас вероятность в произведение вероятностей: для любого N верно

$$\begin{aligned}
\mathbb{P}\,(V^n < N) &= \mathbb{P}\Big(m \notin \bigcup_{j=1}^{n}(A\cap X_j), \forall m \geqslant N\Big)\\
&= \prod_{m\geqslant N}\mathbb{P}\Big(m \notin \bigcup_{j=1}^{n}(A\cap X_j)\Big)\\
&= \prod_{m\geqslant N}\Big(1-\mathbb{P}\Big(m \in \bigcup_{j=1}^{n}(A\cap X_j)\Big)\Big)\\
&= \prod_{m\geqslant N}\Big(1-\mathbb{P}(m\in A)\,\mathbb{P}\Big(m \in \bigcup_{j=1}^{n} X_j\Big)\Big)\\
&= \prod_{m\geqslant N}\Big(1-\mathbb{P}(m\in A)\Big(1-\mathbb{P}\Big(m \notin \bigcup_{j=1}^{n} X_j\Big)\Big)\Big)\\
&= \prod_{m\geqslant N}\Big(1-\frac{a}{m^\alpha}\Big(1-\Big((1-\frac{x}{m^\beta}\Big)^n\Big)\Big). \qquad (2)
\end{aligned}$$

Будем оценивать это произведение.

Пусть сначала $\alpha<1$. Фиксируем $r>0$ и положим $N := r\,n^{\frac{1}{\alpha+\beta-1}}$.

При $\alpha<1$ в зоне $\{m : m\geqslant N\}$ верно $m^\beta \gg n$. Поэтому

$$1-\Big(1-\frac{x}{m^\beta}\Big)^n \sim \frac{xn}{m^\beta}$$

и

$$\begin{aligned}
\log\mathbb{P}\,(V^n<N) &\sim -\sum_{m\geqslant N}\frac{anx}{m^{\alpha+\beta}} \sim -\frac{anx}{(\alpha+\beta-1)N^{\alpha+\beta-1}}\\
&= -\frac{ax}{(\alpha+\beta-1)r^{\alpha+\beta-1}}.
\end{aligned}$$

Последнее соотношение эквивалентно утверждению теоремы при $\alpha < 1$.

В случае $\alpha = 1$ для фиксированного $r > 0$ положим $N := rn^{\frac{1}{\beta}}$.

Нам нужно показать, что величина $p_n := -\log \mathbb{P}(V^n < N)$ имеет соответствующий предел. В силу соотношения (2), имеем

$$p_n \sim \sum_{m \geqslant N} \frac{a}{m} \left(1 - \left(1 - \frac{x}{m^\beta}\right)^n\right) := \widetilde{p}_n. \tag{3}$$

Рассмотрим близкую к $\widetilde{p}_n$ сумму ряда

$$s_n := \sum_{m \geqslant N} \frac{a}{m} \left(1 - e^{\frac{-nx}{m^\beta}}\right)$$

и покажем, что

$$\lim_{n\to\infty} s_n = \frac{a\, I(x/r^\beta)}{\beta}. \tag{4}$$

Действительно, заменяя сумму интегралом и делая замену переменной $u := \frac{nx}{t^\beta}$, получим

$$\sum_{m \geqslant N} \frac{a}{m} \left(1 - e^{\frac{-nx}{m^\beta}}\right) \sim \int_N^\infty \frac{a dt}{t} \left(1 - e^{\frac{-nx}{t^\beta}}\right) = \frac{a}{\beta} \int_0^{\frac{x}{r^\beta}} \frac{1 - e^{-u}}{u}\, du = \frac{a\, I(x/r^\beta)}{\beta}.$$

Остаётся показать, что

$$\lim_{n\to\infty} (\widetilde{p}_n - s_n) = 0. \tag{5}$$

Для этого напомним несколько элементарных оценок. При достаточно малых y имеем $\log(1-y) \geqslant -y - y^2$, откуда при всех натуральных n выполнено

$$(1-y)^n \geqslant \exp(-ny - ny^2)$$

и если $y > 0$, то

$$0 \leqslant \exp(-ny) - (1-y)^n \leqslant \exp(-ny)\left(1 - \exp(-ny^2)\right) \leqslant ny^2.$$

Применяя эту оценку к $y = \frac{x}{m^\beta}$ для $m \geqslant N$, получим

$$\begin{aligned} 0 \quad &\leqslant \quad \widetilde{p}_n - s_n = \sum_{m \geqslant N} \frac{a}{m} \left(e^{\frac{-nx}{m^\beta}} - \left(1 - \frac{x}{m^\beta}\right)^n\right) \\ &\leqslant \quad \sum_{m \geqslant N} \frac{a}{m}\, n\, \frac{x^2}{m^{2\beta}} \sim \frac{ax^2}{2\beta}\, n\, N^{-2\beta} = \frac{ax^2}{2\beta r^{2\beta}}\, n^{-1} \to 0. \end{aligned}$$

Теперь из соотношений (3), (4) и (5) следует

$$\lim_{n\to\infty} p_n = \frac{a\, I(x/r^\beta)}{\beta},$$

что и требовалось доказать. □

Доказательство теоремы 2. С одной стороны, очевидно, что $Q_j \geqslant e^{bV_j}$, $Q^n \geqslant e^{bV^n}$. Поэтому, с учётом теоремы 1,

$$\mathbb{P}(\log Q^n \leqslant brn^\gamma) \leqslant \mathbb{P}(bV^n \leqslant brn^\gamma) = \mathbb{P}(V^n \leqslant rn^\gamma) \to F_{\alpha,\beta}(r).$$

С другой стороны,

$$Q_j = \sum_{m\in A\cap X} e^{bm} \leqslant \sum_{m=1}^{V_j} e^{bm} \leqslant Be^{bV_j},$$

где $B := \frac{e^b}{e^b-1}$.

Фиксируем $\varepsilon \in (0,1)$. Если $V_j \leqslant (1-\varepsilon)rn^\gamma$, то $Q_j \leqslant Be^{b(1-\varepsilon)rn^\gamma}$. При достаточно больших n имеем $Be^{-\varepsilon brn^\gamma} < 1$, так что $Q_j \leqslant e^{brn^\gamma}$.

Соответственно из $V^n \leqslant (1-\varepsilon)rn^\gamma$ следует $Q^n \leqslant e^{brn^\gamma}$, $\log Q^n \leqslant brn^\gamma$. По теореме 1 получаем

$$\mathbb{P}(\log Q^n \leqslant brn^\gamma) \geqslant \mathbb{P}(V^n \leqslant (1-\varepsilon)rn^\gamma) \to F_{\alpha,\beta}((1-\varepsilon)r).$$

Наконец, устремляя $\varepsilon \searrow 0$, приходим к

$$\liminf_{n\to\infty} \mathbb{P}(\log Q^n \leqslant brn^\gamma) \geqslant F_{\alpha,\beta}(r).$$ □

Список литературы

1. F. Auerbach, *Das Gesetz der Bevölkerungskonzentration*. In: Petermanns Geogr. Mitteilungen **59** (1913), 73–76.
2. J.-B. Estoup, *Gammes Sténographiques*, various editions, Paris, 1908–1916.
3. B. Hoffmann, M. Lifshits, Yu. Lifshits, D. Nowotka, *Maximal intersection queries in randomized input models*. — Theory Comput. Syst. **46**, No. 1 (2010), 104–119.
4. A. Lelu, *Jean-Baptiste Estoup and the origins of Zipf's law: a stenographer with a scientific mind* (1868–1950). — Boletn de Estadstica e Investigación Operativa **30**, No. 1 (2014), 66–77.
5. C. D. Manning, H. Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, 1999.
6. M. Petruszewycz, *L'histoire de la loi d'Estoup–Zipf: documents*. — Math. Sci. Hum. **44** (1973), 41–56.

7. G. K. Zipf, *Selected Studies of the Principle of Relative Frequency in Language*. Harvard Univ. Press, 1932.
8. G. K. Zipf, *The Psycho-Biology of Language*, Mifflin, Houghton, 1935.
9. М. П. Бакулина, *Использование закона Ципфа для сжатия текстов.* — Дискретн. анализ и исслед. опер. **14**, No. 2 (2007), 3–13.
10. Я. Галамбош, *Асимптотическая теория экстремальных порядковых статистик*, Наука, М., 1984.
11. М. А. Лифшиц, И. М. Лялинов, *Вероятностные свойства множеств Ципфа и их максимальных пересечений*, в печати, 2022.
12. В. П. Маслов, Т. В. Маслова, *О законе Ципфа и ранговых распределениях в лингвистике и семиотике.* — Матем. заметки **80**, No. 5 (2006), 718–732.
13. В. П. Маслов, *Фазовые переходы нулевого рода и квантование закона Ципфа.* — Теор. матем. физ. **150**, No. 1 (2007), 118–142.
14. В. П. Маслов, *Об одной общей теореме теории множеств, приводящей к распределению Гиббса, Бозе–Эйнштейна, Парето и закону Ципфа–Мандельброта для фондового рынка.* — Матем. заметки **78**, No. 6 (2005), 870–877.
15. Ю. И. Манин, *Закон Ципфа и вероятностные распределения Левина.* — Функц. анализ и его прил. **48**, No. 2 (2004), 51–66.
16. Ю. А. Шрейдер, *О возможности теоретического вывода статистических закономерностей текста (к обоснованию закона Ципфа).* — Пробл. передачи информ. **3**, No. 1 (1967), 57–63.
17. Б. А. Трубников, И. А. Румынский, *Простейший вывод закона Ципфа–Крылова для слов и возможность его "эволюционной" интерпретации.* — Докл. АН СССР **321**, No. 2 (1991), 270–275.

Lifshits M. A., Lyalinov I. M. Two limit theorems on the intersections of random Zipf sets.

In this work we study the asymptotic behaviour of the rarest element in the intersections of a random Zipf set with a large number of independent random sets of the same type but, eventually, with different parameters. The same problem is solved for the maximum of the integral measure of intersection associated with exponentially growing weights.

С.-Петербургский
государственный университет,
Университетская наб. 7/9,
191023, Санкт-Петербург,
Россия
E-mail: mikhail@lifshits.org
E-mail: lyalinov239@yandex.ru

Поступило 29 августа 2022 г.