УДК 519.214+519.212.3

Предельные распределения экстремальных расстояний до ближайшего соседа

А. М. Зубков[*†], **О. П. Орлов**[†]

[*] *Математический институт им. В. А. Стеклова РАН, ул. Губкина 8, Москва, Россия, 119991*

[†] *Московский государственный университет имени М. В. Ломоносова, Ленинские горы 1, Москва, Россия, 119992*

Аннотация. В докладе представлены теоремы о предельных распределениях минимального и максимального расстояний до ближайшего соседа в совокупности случайных независимых точек, имеющих в определенном смысле равномерное распределение на произвольном метрическом пространстве. В качестве примеров таких пространств рассмотрены многомерный тор и двоичный куб. При определенных условиях утверждения теорем удается свести к исследованию суммы индикаторов, которая допускает пуассоновскую аппроксимацию.

Ключевые слова: предельные теоремы, экстремальные значения, ближайший сосед, метод моментов.

1. Введение

Пусть на метрическом пространстве (B, ρ) задана вероятностная мера Q, удовлетворяющая условию

$$Q\left(\{y \in B : \rho(y, x) \leqslant z\}\right) = w(z) \quad \text{для любого} \quad x \in B, \tag{1}$$

и $\xi_1, \ldots, \xi_n$ — независимые случайные элементы B с распределением Q, удовлетворяющим условию (1). Естественными примерами таких мер являются равномерные распределения на многомерных сферах и (непрерывных или дискретных) торах; существуют также примеры неравномерных распределений, удовлетворяющих условию (1) при некоторых значениях z.

Для каждого $i \in \{1, \ldots, n\}$ введём случайную величину $\zeta_i = \min_{j \neq i} \rho(\xi_i, \xi_j)$ — расстояние от точки ξ_i до ее ближайшего соседа. Пусть $\zeta_{(1)} \leqslant \ldots \leqslant \zeta_{(n)}$ — вариационный ряд, составленный из величин $\zeta_1, \ldots, \zeta_n$. Тогда величина $\phi_n = \zeta_{(1)} = \min_{1 \leqslant i < j \leqslant n} \rho(\xi_i, \xi_j)$ является минимальным попарным расстоянием между точками выборки (минимальным расстоянием до ближайшего соседа), а величина $\psi_n = \zeta_{(n)} = \max_{1 \leqslant i \leqslant n} \zeta_i$ является максимальным расстоянием до ближайшего соседа.

В п. 2 сформулированы теоремы о предельном распределении величин ϕ_n и ψ_n. В этих теоремах рассматриваются схемы серий, в которых при изменении параметра n могут изменяться пространство B, метрика ρ и мера Q. Доказательства основаны на модификации метода моментов, предложенной Б. А. Севастьяновым [1]. В пп. 3 и 4 в

качестве примеров расмотрены случаи, когда B — многомерный тор или двоичный куб.

Случайные величины, связанные с расстояниями до ближайших соседей (как правило, для многомерных торов), изучались многими авторами (см., например, [2–5]); расстояния до ближайших соседей используются при построении статистических критериев равномерности распределения (см., например, [6–9]), в алгоритмах классификации, поиска и т. п.

2. Формулировки общих теорем

Теорема 1 *Если* $C_n^2 w(r_n) \to \lambda \in (0,\infty)$ *при* $n \to \infty$, *то*

$$\mathsf{P}\left(\phi_n > r_n\right) \to e^{-\lambda}, \quad n \to \infty.$$

В формулировке следующей теоремы используются величины

$$L_k(z) = \min_{\substack{x_1,\ldots,x_k \in B \\ \rho(x_i,x_j)>z \\ 1\leqslant i<j\leqslant k}} Q\left(\bigcup_{i=1}^{k} \{y \in B : \rho(y,x_i) \leqslant z\}\right), \quad k \geqslant 1, z \geqslant 0,$$

равные минимальной вероятностной мере объединения k таких шаров радиуса z, что центр каждого шара не принадлежит другим шарам.

Теорема 2 *Если выполнены условия*

1) $n(1-w(r_n))^n \xrightarrow[n\to\infty]{} \lambda, \quad \lambda \in (0,\infty)$,

2) $w(2r_n) \xrightarrow[n\to\infty]{} 0$,

3) *при любых фиксированных натуральных* $1 \leqslant k < m$

$$n^m (w(2r_n) - w(r_n))^{m-k} \max_{i_1+\ldots+i_k=m} \left(1 - L_{i_1}(r_n) - \ldots - L_{i_k}(r_n)\right)^n \xrightarrow[n\to\infty]{} 0,$$

то

$$\mathsf{P}(\psi_n \leqslant r_n) \xrightarrow[n\to\infty]{} e^{-\lambda}.$$

3. Случайные точки на торе

Пусть d — фиксированное натуральное число. Рассмотрим d-мерный тор $T^d = S^1 \times \ldots \times S^1$ (прямое произведение d окружностей единичной длины). Введём метрику $\rho(\bar{x},\bar{y}) = \max_{1\leqslant i\leqslant d} \rho^1(x_i,y_i)$, где $\bar{x} = (x_1,\ldots,x_d), \bar{y} = (y_1,\ldots,y_d)$, а $\rho^1(x_1,y_1)$ — длина наименьшей дуги, соединяющей точки x_1 и y_1 на окружности. В этом случае $w(z) = \min\{(2z)^d, 1\}$.

Пусть $\xi_1,\ldots,\xi_n$ — независимые точки на этом торе, имеющие на нём равномерное распределение. Обозначим, как и прежде, через $\zeta_i =$

$\min_{j\neq i}\{\rho(\xi_i,\xi_j)\}$ расстояние от i-й точки до ближайшего соседа, а через $\phi_n = \min_{1\leqslant i\leqslant n}\zeta_i$ и $\psi_n = \max_{1\leqslant i\leqslant n}\zeta_i$ — минимальное и максимальное расстояния до ближайшего соседа.

Теорема 3 *При $d = \text{const}$, $n\to\infty$ справедливы соотношения*

$$\mathsf{P}\left(\phi_n > \frac{1}{2}\left(\frac{2y}{n^2}\right)^{1/d}\right)\to e^{-y},\ \mathsf{P}\left(\psi_n \leqslant \left(\frac{y+\ln n}{2^d n}\right)^{1/d}\right)\to e^{-e^{-y}},\ y\in\mathbf{R}.$$

Замечание. Утверждения теоремы 3 можно вывести из результатов работ [2], [9]; теорема приводится как пример применения теорем 1 и 2. В [3] найдено предельное распределение максимального расстояния до ближайшего соседа в случае евклидовой метрики и некоторых ограничений на распределение случайных точек.

4. Двоичные строки

Пусть $V_T = \{0,1\}^T$ — пространство двоичных строк длины T с метрикой Хемминга $\rho(\mathbf{x},\mathbf{y}) = \sum_{i=1}^T \mathbb{I}\{x_i\neq y_i\}$, где $\mathbf{x} = (x_1,\dots,x_T), \mathbf{y} = (y_1,\dots,y_T)\in V_T$. Пусть $\xi_1,\dots,\xi_n$ — независимые элементы V_T, имеющие равновероятное распределение на V_T. Тогда

$$w(z) = \mathsf{P}(\rho(\xi_1,\mathbf{x})\leqslant r_n) = \frac{1}{2^T}\sum\nolimits_{k=0}^{[z]} C_T^k \text{ для всех } \mathbf{x}\in V_T.$$

Как и раньше, пусть $\zeta_i = \min_{j\in\{1,\dots,n\}\setminus\{i\}}\rho(\xi_i,\xi_j)$ — расстояние от строки ξ_i до её ближайшего соседа, $i = 1,\dots,n$, а $\phi_n = \min_{1\leqslant i\leqslant n}\zeta_i$ и $\psi_n = \max_{1\leqslant i\leqslant n}\zeta_i$ — минимальное и максимальное расстояния до ближайшего соседа.

Статистики ϕ_n и ψ_n можно использовать для проверки гипотезы о равновероятности и независимости элементов выборки $\xi_1,\dots,\xi_n$.

Теорема 4 *Если $n, T\to\infty$ и s_n меняются так, что*

$$C_n^2\frac{1}{2^T}\sum\nolimits_{k=0}^{s_n} C_T^k\to\lambda,$$

$$\frac{1}{2^T}\sum\nolimits_{k=0}^{s_n} C_T^k = (1+o(1))\frac{2\lambda}{n^2},$$

то

$$\mathsf{P}(\phi_n > s_n)\to e^{-\lambda}.$$

Теорема 5 *Если $n, T \to \infty$ и r_n меняются так, что*

$$n\left(1 - \frac{1}{2^T}\sum\nolimits_{k=0}^{r_n} C_T^k\right)^n \to \lambda \in (0,\infty), \quad r_n = o(T),$$

то

$$\mathsf{P}(\psi_n \leqslant r_n) \to e^{-\lambda}.$$

Замечание. При условиях теоремы 5 необходимо $\ln n \sim T \ln 2$. Если, например, $n = 2^{T-T^\alpha}$, $\alpha \in (0,1)$, то первое условие теоремы принимает вид $C_T^{r_n}/2^{T^\alpha} - \ln\left(2^{T-T^\alpha}/\lambda\right) \to 0$; и так как при $1 > \beta > \alpha$ верно $C_T^{[T^\beta]}/2^{T^\alpha} - \ln\left(2^{T-T^\alpha}/\lambda\right) \to \infty$, поэтому существует такая последовательность $r_n = o(T)$, что выполняются условия теоремы 5.

Литература

1. *Севастьянов Б. А.* Предельный закон Пуассона в схеме сумм зависимых случайных величин // Теория вероят. и ее примен. — 1972. — Т. 17, вып. 4. — С. 733-738.
2. *Silverman B., Brown T.* Short distances, flat triangles and Poisson limits // J. Appl. Probab. — 1978. — Vol. 15. — P. 815–825.
3. *Henze N.* The limit distribution for maxima of «weighted» r-th nearest-neighbor distances // J. Appl. Prob. — 1982. — Vol. 19. — P. 344–354.
4. *Penrose M. D., Yukich J. E.* Laws of large numbers and nearest neighbor distances / Advances in Directional and Linear Statistics. — Berlin: Physica-Verlag HD, 2011. — P. 189–199
5. *Baryshnikov Yu., Penrose M. D., Yukich J. E.* Gaussian limits for generalized spacings // Ann. Appl. Probab. — 2009. — Vol. 19, no. 1. — P. 158-185.
6. *Bickel P. J., Breiman L.* Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test // Ann. Probab. — 1983. — Vol. 11, no. 1 — P. 185–214.
7. *Schilling M. F.* Goodness of fit testing in R^m based on the weighted empirical distribution of certain nearest neighbor statistics // Ann. Statist. — 1983. — Vol. 11, no. 1. — P. 1–12.
8. *Schilling M. F.* An infinite-dimensional approximation for nearest neighbor goodness of fit tests // Ann. Statist. — 1983. — Vol.1, no. 1. — P. 13–24.
9. *L'écuyer P., Cordeau J.-F., Simard R.* Close-point spatial tests and their application to random number generators // Oper. Res. — 2000. — Vol. 48, no. 2. — P. 308–317.

UDC 519.214+519.212.3

Limit distributions of extreme distances to the nearest neighbor

A. M. Zubkov[*†], **O. P. Orlov**[†]

[*] *Steklov Mathematical Institute of RAS,*
Gubkina st. 8, Moscow, 119991, Russia
[†] *Moscow State University,*
Leninskie Gory 1, Moscow, 119992, Russia

The paper presents theorems on the limit distributions of the minimal and maximal distances to the nearest neighbor in a set of random independent points, that in a certain sense have a uniform distribution on an arbitrary metric space. As examples of such spaces, we consider a multidimensional torus and a binary cube. Under certain conditions, the theorems can be reduced to the study of the sum of indicators, which admits Poisson approximation.

Keywords: limit theorem, extreme values, nearest neighbor, method of moments.