УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО УНИВЕРСИТЕТА
*Том 154, кн. 4* Физико-математические науки *2012*

УДК 51-7

# СТЕПЕНЬ УЛЬТРАМЕТРИЧНОСТИ МЕТРИЧЕСКОГО ПРОСТРАНСТВА

*М.Д. Миссаров*

**Аннотация**

Введено понятие степени ультраметричности метрического пространства. Вычислена степень ультраметричности евклидова пространства. Приведены результаты статистического эксперимента по вычислению этой степени в пространстве строк заданной длины для расстояния Хэмминга и редакционного расстояния. Показано, что степень ультраметричности растет при увеличении длины строки.

**Ключевые слова:** коэффициент ультраметричности, степень ультраметричности, евклидова метрика, расстояние Хэмминга, редакционное расстояние.

## Введение

Напомним, что метрическое пространство $X$ с метрикой $\rho$ называется ультраметричным пространством, если неравенство треугольника заменяется на более сильное неравенство ультраметричности: для любых трех точек $x$, $y$, $z \in X$

$$\rho(x,y) \le \max(\rho(x,z), \rho(y,z)).$$

Ультраметрики и ультраметричные пространства стали появляться в различных задачах современной физики, вычислительной биологии, анализа данных и других областях [1–5]. Основными и наиболее изученными примерами ультраметричных пространств являются $p$-адические пространства. Иерархическая структура $p$-адического пространства, вытекающая из ультраметричности $p$-адической метрики, позволила получить точный анализ ренормализационной группы [6]. Выяснилось также, что жадные алгоритмы дают точные решения сложных задач комбинаторной оптимизации в $p$-адическом пространстве [7]. Это позволило описать асимптотическое поведение решений задач комбинаторной оптимизации [8].

Интересен вопрос о том, в какой степени ультраметрично то или иное метрическое пространство. Из условия ультраметричности следует, что в любом треугольнике две наибольшие «стороны треугольника» равны друг другу. Рассмотрим произвольную тройку точек $x$, $y$, $z$ в метрическом пространстве $X$ (условно назовем эту тройку «треугольником»), а также набор попарных расстояний между этими точками $(\rho(x,y), \rho(x,z), \rho(y,z))$ и обозначим через $\rho_1(x,y,z)$ наименьшее значение в этом наборе, через $\rho_2(x,y,z)$ – второе по величине значение, через $\rho_3(x,y,z)$ – наибольшее значение в наборе.

Определим коэффициент ультраметричности тройки $(x,y,z)$ как величину

$$u(x,y,z;\rho) = 2\,\frac{\rho_2(x,y,z)}{\rho_3(x,y,z)} - 1. \tag{1}$$

Если совпадают две точки из трех, то коэффициент ультраметричности равен 1 согласно определению. Если $x = y = z$ , то положим $u(x,y,z) = 1$. Заметим,

что коэффициент ультраметричности любой тройки лежит в диапазоне от 0 до 1. В случае вырожденного треугольника $\rho(x,y) = \rho(x,z) = \rho(y,z)/2$ коэффициент $u(x,y,z) = 0$. В случае ультраметричного пространства коэффициент $u(x,y,z) = 1$ для любой тройки $x,\ y,\ z$.

Назовем степенью ультраметричности $u(X;\rho)$ пространства $(X,\rho)$ среднее значение коэффициента ультраметричности в этом пространстве:

$$u(X;\rho) = \langle u(x,y,z;\rho)\rangle.$$

Здесь усреднение проводится по всем возможным тройкам вершин в пространстве $(X,\rho)$. В случае конечного метрического пространства среднее определяется как среднее арифметическое значение коэффициента ультраметричности по всем возможным тройкам. В случае, когда метрическое пространство снабжено вероятностной мерой, среднее надо понимать как среднее по этой мере. В случае, когда метрическое пространство снабжено естественной мерой Хаара (как, например, евклидово пространство с мерой Лебега), среднее надо понимать как интеграл по этой мере с учетом факторизации множества всех троек по классам троек с одинаковым значением коэффициента ультраметричности. Именно такой случай рассмотрен в разд. 2.

## 1. Степень ультраметричности евклидова пространства

Заметим, что в случае евклидова пространства коэффициент ультраметричности тройки инвариантен относительно сдвигов, растяжений и поворотов:

$$\begin{aligned} &u(x,y,z;\rho_E) = u(x+a,y+a,z+a;\rho_E), \quad a\in R^d,\\ &u(x,y,z;\rho_E) = u(\lambda x,\lambda y,\lambda z;\rho_E), \quad \lambda\in R, \quad \lambda\neq 0,\\ &u(Ox,Oy,Oz;\rho_E) = u(x,y,z;\rho_E), \quad O\in O(d). \end{aligned} \tag{2}$$

Другими словами, все подобные треугольники в евклидовом пространстве имеют один и тот же коэффициент ультраметричности. Если $d>1$ и треугольник невырожденный, то его можно вписать в окружность. Пусть $x,\ y,\ z$ – заданные точки. Выберем в каждом классе подобных треугольников треугольник, вписанный в заданную окружность единичного радиуса так, что точка $x$ занимает фиксированное положение в заданной точке окружности $A$. Усреднение будем проводить по точкам $y$ и $z$, независимо и равномерно распределенным по этой окружности. Тогда степень ультраметричности определяется как среднее по $y$ и $z$: $u(R^d;\rho_E) = Eu(A,y,z;\rho_E)$. Обозначим дисперсию коэффициента ультраметричности как $\sigma^2(R^d;\rho_E) = Du(A,y,z;\rho_E)$. Здесь $\rho_E$ обозначает евклидову метрику.

**Теорема 1.** *Имеют место соотношения* $u(R^1;\rho_E) = 1/2$, $\sigma^2(R^1;\rho_E) = 1/12$. *В случае* $d>1$

$$u(R^d;\rho_E) = \frac{24\ln 2}{\pi^2} - 1, \quad \sigma^2(R^d;\rho_E) = 4\left(\frac{3}{\pi^2}(1+\ln 4) - \left(\frac{12}{\pi^2}\ln 2\right)^2\right).$$

**Доказательство.** Пусть $d=1$. В этом случае все треугольники являются вырожденными и без ограничения общности мы можем считать $x=0$, $z=1$, $0\le y\le 1$. Усреднение проводится по $y$, где $y$ равномерно распределено на отрезке $[0,1]$. Тогда

$$E\,\frac{\rho_{E,2}}{\rho_{E,3}} = \int\limits_0^1 \max(y,1-y)\,dy = \frac{3}{4},$$

$$E\left(\frac{\rho_{E,2}}{\rho_{E,3}}\right)^2 = \int\limits_0^1 (\max(y, 1-y))^2\, dy = \frac{7}{12}.$$

Отсюда следует первая часть теоремы.

Пусть $d > 1$, $x$ занимает фиксированное положение на окружности, дуга между $x$ и $y$ задается углом $\varphi_1$, а между $x$ и $z$ – углом $\varphi_2$, где $\varphi_1$ и $\varphi_2$ независимо и равномерно распределены от 0 до $2\pi$. Так как бóльшая сторона треугольника лежит против бóльшего угла, достаточно рассматривать соотношения между этими дугами. Три дуги по величине могут располагаться шестью возможными способами, и из соображений симметрии достаточно рассматривать один из вариантов:

$$\varphi_1 \le \varphi_2 - \varphi_1 \le 2\pi - \varphi_2. \tag{3}$$

Фигура в квадрате $0 \le \varphi_1 \le 2\pi$, $0 \le \varphi_2 \le 2\pi$, задаваемая неравенствами (3), имеет вид треугольника с площадью $\pi^2/3$. Из (3) следует, что $0 \le \varphi_1 \le 2/3\pi$, $2\varphi_1 \le \varphi_2 \le \pi + \varphi_1/2$. Тогда

$$\frac{\rho_{E,2}}{\rho_{E,3}} = \frac{\sin(\varphi_2 - \varphi_1)/2}{\sin(2\pi - \varphi_2)/2} = \frac{\sin(\varphi_2 - \varphi_1)/2}{\sin \varphi_2/2}.$$

Отсюда

$$E\frac{\rho_{E,2}}{\rho_{E,3}} = \frac{3}{\pi^2} \int\limits_0^{2/3\pi} d\varphi_1 \int\limits_{2\varphi_1}^{\pi+\varphi_1/2} d\varphi_2 \cdot \frac{\sin(\varphi_2 - \varphi_1)/2}{\sin \varphi_2/2} = \frac{12 \ln 2}{\pi^2}.$$

Можно также вычислить

$$E\left(\frac{\rho_{E,2}}{\rho_{E,3}}\right)^2 = \frac{3}{\pi^2} \int\limits_0^{2/3\pi} d\varphi_1 \int\limits_{2\varphi_1}^{\pi+\varphi_1/2} d\varphi_2 \left(\frac{\sin(\varphi_2 - \varphi_1)/2}{\sin \varphi_2/2}\right)^2 = (1 + \ln 4)\frac{3}{\pi^2}\,.$$

Отсюда среднее значение коэффициента ультраметричности равно

$$u(R^d; \rho_E) = \frac{24 \ln 2}{\pi^2} - 1 \approx 0.6854.$$

Дисперсия коэффициента ультраметричности есть

$$\sigma^2(R^d; \rho_E) = 4\left(\frac{3}{\pi^2}(1 + \ln 4) - \left(\frac{12}{\pi^2} \ln 2\right)^2\right) \approx 0.06136.$$

Стандартное отклонение $\sigma(R^d; \rho_E) \approx 0.2477$. Теорема доказана. □

На рис. 1 изображена гистограмма выборки из $10^6$ независимых наблюдений коэффициента ультраметричности в евклидовом случае при $d > 1$. По результатам этого эксперимента выборочное среднее равно 0.6857, а стандартное отклонение – 0.2458.

## 2. Степень ультраметричности пространства строк

Пусть дан некоторый алфавит $G$. Рассмотрим множество $X$ строк символов из алфавита $G$ длины $n$:

$$a \in X, \quad a = a_1 a_2 \dots a_n, \quad a_i \in G, \quad i = 1, \dots, n.$$

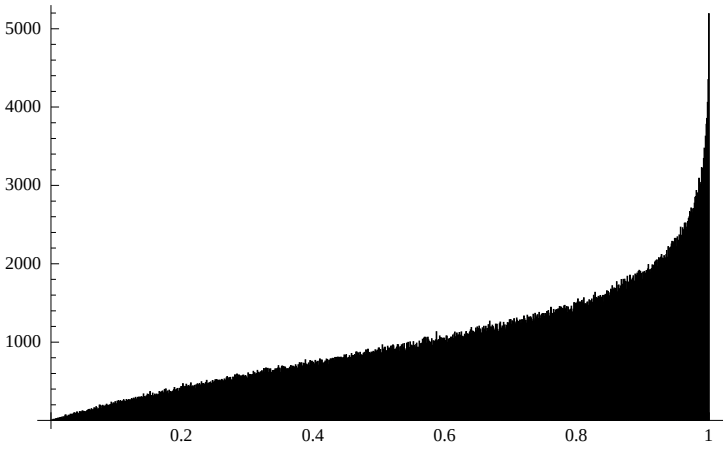


Рис. 1. Гистограмма коэффициента ультраметричности для евклидова расстояния

Табл. 1

| $n$ | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0.776 | 0.826 | 0.852 | 0.869 | 0.880 | 0.889 | 0.897 | 0.903 | 0.907 | 0.912 |
| | 0.186 | 0.137 | 0.115 | 0.102 | 0.092 | 0.085 | 0.079 | 0.075 | 0.075 | 0.068 |

Расстояние по Хэммингу [2] между строками $A = a_1 \dots a_n$ и $B = b_1 \dots b_n$ определяется как количество несовпадающих позиций в этих строках:

$$\rho_H(A, B) = |I|,$$

где $I = \{i \in (1, \dots, n) : a_i \neq b_i\}$. Тогда множество $X$ с метрикой $\rho_H$ является метрическим пространством. В дальнейшем мы будем предполагать, что алфавит состоит из 4 букв (алфавит генетического кода определяется 4 нуклеотидами).

В табл. 1 приведены результаты статистического эксперимента для расстояния по Хэммингу $10^6$ независимых наблюдений для каждого столбца таблицы. Здесь $n$ обозначает длину строки, верхнее число в ячейке является выборочным средним коэффициента ультраметричности, нижнее – выборочным стандартным отклонением. Мы видим, что степень ультраметричности для метрики Хэмминга растет с длиной строки, в то время как стандартное отклонение уменьшается. На рис. 2 изображена гистограмма выборки из $10^6$ наблюдений для пространства строк длины 100. В последнем случае степень ультраметричности близка к 1 (равна 0.912).

Рассмотрим теперь случай редакционного расстояния между строками, которое широко используется в биоинформатике для сравнения геномов различных видов и построения филогенетических деревьев [2]. Мы будем предполагать, что строки $A$ и $B$ имеют длину $n$. При процедуре выравнивания в обе строки вставляются пробелы "–" так, что растянутые строки имеют одинаковую длину $l$. Получаемые строки обозначим как $A' = a'_1 \dots a'_l$ и $B' = b'_1 \dots b'_l$. При сравнении двух выравниваемых строк штрафуется несовпадение символов с помощью штрафной функции $d(a, b)$. В дальнейшем мы положим, что $d(a, b) = 0$, если $a = b$; $d(a, b) = \gamma$, $0 \leq \gamma \leq 1$, если $a$ – буква из алфавита, $b$ – пробел и наоборот; $d(a, b) = 1$, если $a$ и $b$ – несовпадающие буквы. Редакционное расстояние между строками $A$ и $B$ определяется как

$$\rho(A, B) = \min \sum_{i=1}^{l} d(a'_i, b'_i),$$

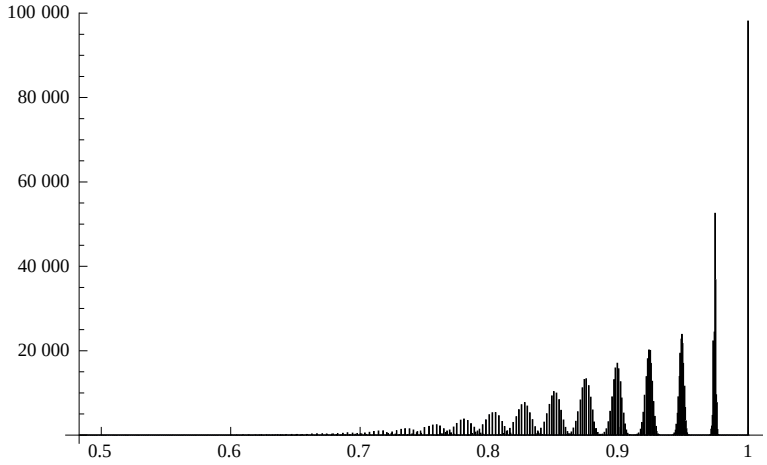


Рис. 2. Гистограмма коэффициента ультраметричности для расстояния Хэмминга

Табл. 2

| $\gamma/n$ | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.1 | 0.761 | 0.817 | 0.848 | 0.865 | 0.880 | 0.891 | 0.899 | 0.906 | 0.912 | 0.919 |
| | 0.227 | 0.159 | 0.129 | 0.112 | 0.097 | 0.089 | 0.082 | 0.077 | 0.071 | 0.066 |
| 0.2 | 0.762 | 0.815 | 0.846 | 0.867 | 0.882 | 0.891 | 0.901 | 0.906 | 0.913 | 0.918 |
| | 0.225 | 0.160 | 0.130 | 0.109 | 0.098 | 0.090 | 0.081 | 0.076 | 0.070 | 0.067 |
| 0.3 | 0.759 | 0.815 | 0.847 | 0.866 | 0.879 | 0.892 | 0.899 | 0.907 | 0.914 | 0.917 |
| | 0.225 | 0.161 | 0.129 | 0.110 | 0.098 | 0.088 | 0.081 | 0.075 | 0.071 | 0.066 |
| 0.4 | 0.763 | 0.815 | 0.849 | 0.865 | 0.880 | 0.890 | 0.899 | 0.908 | 0.911 | 0.917 |
| | 0.225 | 0.159 | 0.128 | 0.110 | 0.097 | 0.088 | 0.082 | 0.076 | 0.071 | 0.067 |
| 0.5 | 0.761 | 0.814 | 0.847 | 0.866 | 0.880 | 0.894 | 0.898 | 0.907 | 0.912 | 0.917 |
| | 0.226 | 0.161 | 0.130 | 0.111 | 0.097 | 0.080 | 0.082 | 0.075 | 0.071 | 0.066 |
| 0.6 | 0.762 | 0.817 | 0.852 | 0.873 | 0.884 | 0.896 | 0.903 | 0.911 | 0.915 | 0.921 |
| | 0.182 | 0.135 | 0.112 | 0.097 | 0.087 | 0.080 | 0.074 | 0.068 | 0.065 | 0.061 |
| 0.7 | 0.767 | 0.822 | 0.856 | 0.874 | 0.888 | 0.897 | 0.905 | 0.912 | 0.917 | 0.922 |
| | 0.172 | 0.131 | 0.108 | 0.095 | 0.085 | 0.078 | 0.073 | 0.067 | 0.063 | 0.060 |
| 0.8 | 0.773 | 0.828 | 0.860 | 0.878 | 0.891 | 0.901 | 0.907 | 0.914 | 0.920 | 0.924 |
| | 0.171 | 0.127 | 0.106 | 0.093 | 0.083 | 0.075 | 0.071 | 0.065 | 0.062 | 0.058 |
| 0.9 | 0.772 | 0.833 | 0.860 | 0.881 | 0.894 | 0.903 | 0.911 | 0.916 | 0.923 | 0.926 |
| | 0.174 | 0.127 | 0.106 | 0.090 | 0.081 | 0.075 | 0.068 | 0.063 | 0.059 | 0.056 |
| 1.0 | 0.785 | 0.839 | 0.865 | 0.882 | 0.897 | 0.905 | 0.911 | 0.919 | 0.923 | 0.928 |
| | 0.190 | 0.135 | 0.110 | 0.096 | 0.082 | 0.076 | 0.070 | 0.065 | 0.061 | 0.057 |

где минимум берется по всем возможным выравниваниям. Редакционное расстояние вычисляется методом динамического программирования. Известен следующий алгоритм [2]: пусть $A = a_1 \dots a_n$, $B = b_1 \dots b_n$, определим

$$\rho_{i,j} = \rho(a_1 \dots a_i; b_1 \dots b_j); \quad \rho_{0,0} = 0;$$

$$\rho_{0,j} = \sum_{k=1}^{j} d(-, b_k) = \gamma j, \quad \rho_{i,0} = \sum_{k=1}^{i} d(a_k, -) = \gamma i.$$

Тогда

$$\rho_{i,j} = \min\{\rho_{i-1,j} + d(a_i, -), \rho_{i-1,j-1} + d(a_i, b_j), \rho_{i,j-1} + d(-, b_j)\}.$$

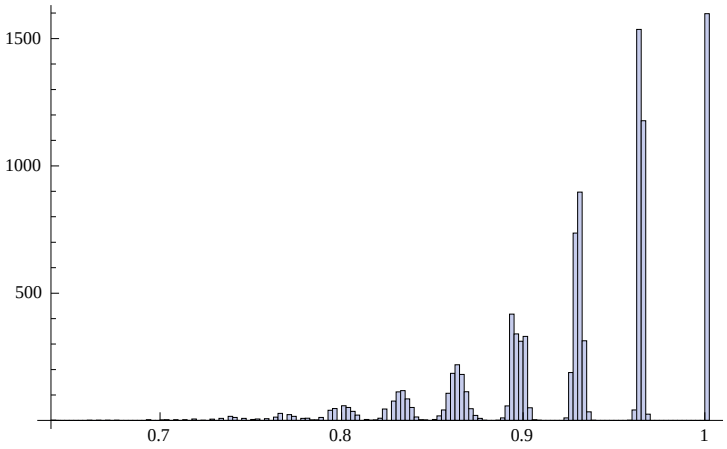


Рис. 3. Гистограмма коэффициента ультраметричности для редакционного расстояния

Если $d(a,b)$ – метрика на алфавите, то $\rho(A,B)$ – метрика на множестве строк в этом алфавите, $\rho(A,B)=\rho_{n,n}$.

В табл. 2 приведены результаты статистического эксперимента при различных значениях штрафа $\gamma$ и различной длине строк $n$. Количество экспериментов в каждом случае равнялось 10000. Верхнее число в ячейке является выборочным средним коэффициентом ультраметричности, нижнее – выборочным стандартным отклонением. Мы видим, что при любом значении величины штрафа $\gamma$ степень ультраметричности для редакционного расстояния растет с длиной строки, в то время как стандартное отклонение уменьшается. Это означает, что коэффициент ультраметричности в среднем растет и все больше концентрируется вокруг среднего значения. При фиксированной длине строки коэффициент ультраметричности в среднем почти не зависит от величины штрафа $\gamma$ (растет очень незначительно). На рис. 3 изображена гистограмма выборки из $10^4$ наблюдений для пространства строк длины 100 и величиной $\gamma=1$. В этом случае статистическая оценка степени ультраметричности равна 0.928 при стандартном отклонении 0.057. Можно предположить, что при достаточно большом значении $n$ заданное множество строк имеет (с некоторой погрешностью) иерархическую структуру относительно редакционного расстояния.

Автор выражает благодарность студенту ИВМиИТ Казанского федерального университета Илье Калинину за помощь в проведении расчетов в пакете Mathematica.

**Summary**

*M.D. Missarov.* The Degree of Ultrametricity of a Metric Space.

The notion of the degree of ultrametricity of a metric space is introduced. The degree of ultrametricity of the Euclidean space is obtained. The results of a statistical experiment on the computation of the degree of ultrametricity in the space of strings of given lengths for Hamming and edit distances are presented. It is shown that the degree of ultrametricity grows with an increase in the string length.

**Key words:** coefficient of ultrametricity, degree of ultrametricity, Euclidean metric, Hamming distance, edit distance.

**Литература**

1. *Mezard M., Parisi G., Virasoro M.A.* Spin Glass Theory and Beyond. – Singapore: World Sci., 1987. – 461 p.

2. *Гасфилд Д.* Строки, деревья и последовательности в алгоритмах. – СПб.: Невский Диалект, 2003. – 653 с.

3. *Vladimirov V.S., Volovich I.V., Zelenov E.I.* $p$-Adic Analysis and Mathematical Physics. – Singapore: World Sci., 1994. – 340 p.

4. *Dragovich B., Khrennikov A.Yu., Kozyrev S.V., Volovich I.V.* On $p$-adic mathematical physics // $p$-Adic Numbers Ultrametric Anal. Appl. – 2007. – V. 1, No 1. – P. 1–17.

5. *Lerner E.Yu., Missarov M.D.* $P$-adic Feynman and string amplitudes // Commun. Math. Phys. – 1989. – V. 121, No 1. – P. 35–48.

6. *Missarov M.D.* Renormalization group solution of fermionic Dyson model // Asymptotic Combinatorics with Application to Mathematical Physics / Eds. V.A. Malyshev, A.M. Vershik. – Springer, 2002. – P. 151–166.

7. *Миссаров М.Д., Степанов Р.Г.* О задачах комбинаторной оптимизации в ультраметричных пространствах // Теор. и мат. физика. – 2003. – Т. 136, № 1. – С. 164–176.

8. *Missarov M.D., Stepanov R.G.* Asymptotic properties of combinatorial optimization problems in $p$-adic space // $p$-Adic Numbers Ultrametric Anal. Appl. – 2011. – V. 3, No 2. – P. 114–128.

Поступила в редакцию
20.11.12

**Миссаров Мукадас Дмухтасибович** – доктор физико-математических наук, заведующий кафедрой анализа данных и исследования операций Казанского (Приволжского) федерального университета.

E-mail: *moukadas.missarov@ksu.ru*