



ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 36 за 2025 г.



ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

А.А. Кислицын, Ю.Н. Орлов,
Д.Т. Тамм

О распределении расстояний
между случайными точками в
гиперкубе

Статья доступна по лицензии
[Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/)



Рекомендуемая форма библиографической ссылки: Кислицын А.А., Орлов Ю.Н., Тамм Д.Т. О распределении расстояний между случайными точками в гиперкубе // Препринты ИПМ им. М.В.Келдыша. 2025. № 36. 32 с. EDN: [CWCNXB](https://library.keldysh.ru/preprint.asp?id=2025-36)
<https://library.keldysh.ru/preprint.asp?id=2025-36>

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В. Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

А.А. Кислицын, Ю.Н. Орлов, Д.Т. Тамм

**О распределении расстояний между
случайными точками в гиперкубе**

Москва – 2025

Кислицын А.А., Орлов Ю.Н., Тамм Д.Т.

О распределении расстояний между случайными точками в гиперкубе

В работе исследуется функция распределения расстояний между случайными точками в гиперкубе. Получена точная формула для функции распределения расстояний в области $r \leq 1$. Проведены расчеты по уточнению асимптотики функции распределения расстояний в области $r \leq 1$ при больших n . Показано, что эта вероятность спадает быстрее, чем качественная ее оценка через объем n -мерного шара. Приведены результаты численного моделирования распределения расстояний в гиперкубе большой размерности. Получено предельное распределение расстояний, нормированных на $\sqrt{n}$.

Ключевые слова: гиперкуб, функция распределения расстояний, асимптотический анализ

Kislitsyn A.A., Orlov Yu.N., Tamm D.T.

On the distance distribution between random points in a hypercube

In this work the distance distribution function between random points in a hypercube is investigated. The exact formula for the distance distribution function in the region $r < 1$ is obtained. Calculations have been performed to refine the asymptotics of the distance distribution function in the region $r < 1$ for large distances. It is shown that this probability decreases faster than its qualitative assessment by the volume of an n -dimensional sphere. The results of numerical simulation of the distribution of distances in a large-dimensional hypercube are presented. The asymptotic of the distance distribution function is obtained.

Keywords: hypercube, distance distribution function, asymptotic analysis

Содержание

Введение	3
1. Распределение расстояний в кубе для малых размерностей	5
2. Плотность распределения квадрата расстояния в гиперкубе для $r < 1$	11
3. Асимптотика функции распределения расстояний для $r < 1$	17
4. Численное моделирование плотности распределения расстояний в гиперкубе при больших размерностях	22
5. О предельном распределении расстояний между точками	24
Заключение	28
Литература	30

Введение

Исследование статистических свойств расстояний между случайными точками некоторого множества представляет собой классическую задачу теории вероятностей применительно к так называемым геометрическим вероятностям. В монографии М. Кендалла и П. Морана [1] рассмотрены многие классические задачи о взаимном расположении случайных точек на плоскости и в пространстве, построены распределения расстояний между случайными точками в круге и в шаре, а также рассмотрены задачи о средних длинах секущих выпуклых тел с использованием теоремы Крофтона.

Примерно в это же время (50 лет назад) началось достаточно активное изучение расстояний между точками в многомерных объектах и, в частности, в единичном гиперкубе. Также разрабатывались и методы генерации точек со случайным распределением координат [2]. В исследованиях [3-6] были введены специальные функции, так называемые *box integrals*, с помощью которых выражаются средние расстояния между точками, средние квадраты расстояний и иные функционалы от распределения точек при их случайном равномерном координатном распределении в кубе. Поскольку мера единичного куба равна единице, то интегралы по n -мерному объему от каких-либо функций, зависящих от этих координат, доставляют средние значения функций, так что задача сводится, по сути, к многомерному интегрированию по кубу. Например, для среднего квадрата расстояний это есть интеграл от суммы квадратов разностей координат, а для среднего расстояния – интеграл от квадратного корня из указанной суммы. В частности, в работе [4] было получено выражение для среднего расстояния в трехмерном кубе (константа Роббинса). Работы по нахождению точных выражений для средних расстояний в гиперкубах в разных частных случаях продолжаются и в наше время [7, 8].

Однако в этих работах отсутствуют теоретические исследования собственно функций распределения расстояний между точками в многомерных телах, а также анализ асимптотических распределений функционалов от расстояний. В основном изучаются некоторые частные случаи, такие как нахождение среднего расстояния от вершины гиперкуба или от его центра до произвольной точки внутри него и т.п. Но если раньше задача о вероятности нахождения точки на расстоянии, не большем заданного от некоторой фиксированной точки множества, не представляла особой практической важности, то в настоящее время это принципиальный вопрос оценки точности поиска приближенного ближайшего соседа. А сам поиск ближайшего соседа привел к созданию целого класса алгоритмов, имеющих большое прикладное значение [9-14].

Вопрос о поиске приближенного первого ближайшего соседа весьма актуален, особенно если число элементов и/или размерность пространства велики и полный перебор невозможен. Для оценки точности такого поиска надо построить функцию распределения расстояний между ближайшими соседями в зависимости от того, как заданы координаты точек исследуемого множества.

Обозначим через $W_{n,N}(r)$ функцию распределения расстояний между ближайшими соседями первого порядка среди N точек в n -мерном единичном кубе. По определению, это есть вероятность того, что для заданного набора из N точек расстояние между ближайшими меньше, чем данное число r . При поиске приближенного ближайшего соседа на заданном уровне значимости требуется определить вероятность $1 - W_{n,N}(r)$ того, что на расстоянии от данной точки, меньшем r , нет других точек множества. Уровень значимости равен тогда значению функции $W_{n,N}(r)$.

Предположим, что известна функция распределения $F_n(r)$ расстояний между точками некоторого множества в n -мерном пространстве. Тогда вероятность того, что искомое расстояние больше r , равна $1 - F_n(r)$. Если рассматривается выборка из N независимых случайных расстояний, то вероятность того, что все они больше r , равна $(1 - F_n(r))^N$. Поэтому функция распределения минимального расстояния в выборке, т.е. функция распределения расстояний между ближайшими соседями, равна

$$W_{n,N}(r) = 1 - (1 - F_n(r))^N. \quad (1)$$

Следовательно, если ближайший сосед ищется на уровне значимости $W_{n,N}(r) = \alpha$, то радиус соответствующего круга r , внутри которого следует найти точку множества, которая будет рассматриваться как ближайшая к данной, определяется из уравнения

$$F_n(r) = 1 - (1 - W_{n,N}(r))^{1/N} = \frac{\alpha}{N} + o(1/N). \quad (2)$$

При этом следует учесть, что уровень значимости α не может быть задан сколь угодно малым, поскольку хотя бы один сосед должен иметь возможность быть в указанной области. Поскольку для N точек количество расстояний, по которым строится эмпирическое распределение, равно $N(N-1)/2$, то условие корректности состоит в том, что $F_n(r) \cdot N(N-1)/2 \geq 1$, т.е. $\alpha \geq 2/(N-1)$.

Таким образом, основным пунктом является нахождение распределения расстояний между точками со случайными координатами в n -мерном единичном кубе $[0;1]^n$. Хотя построение этого распределения – аналитически точно решаемая задача, явные формулы чрезвычайно громоздки и, как правило, не выписываются для пространств больших размерностей.

В настоящей работе мы показываем, в чем состоят основные трудности при построении распределения расстояний между точками в единичном гиперкубе и строим точные распределения там, где аналитические вычисления могут быть проведены явно для произвольной размерности. Точное решение для функции распределения расстояний между точками может быть получено в области $r \leq 1$. Отметим, что эта вероятность оказалась асимптотически меньше, чем объем n -мерного шара единичного радиуса, что является важным теоретическим результатом.

Также в работе делаются некоторые численные оценки асимптотического поведения распределения расстояний в гиперкубе при больших размерностях и обсуждаются вопросы, связанные с приближенным поиском ближайшего соседа среди случайных точек.

1. Распределение расстояний в кубе для малых размерностей

Дадим сначала необходимые базовые определения.

Пусть ξ – случайная величина (СВ), а P – обозначение вероятности. Функция распределения (ФР) $F_\xi(x)$ – это вероятность того, что СВ ξ принимает значения, строго меньшие заданного действительного числа x :

$$F_\xi(x) = P(\xi < x). \quad (3)$$

СВ ξ называется непрерывной, если существует такая неотрицательная функция $f_\xi(x)$, называемая плотностью функции распределения (ПФР) для СВ ξ , что

$$F_\xi(x) = \int_{-\infty}^x f_\xi(t) dt. \quad (4)$$

Тогда вероятность того, что значения СВ попали в промежуток $[x_1; x_2]$, равна

$$P(x_1 \leq \xi < x_2) = P(\xi < x_2) - P(\xi < x_1) = \int_{x_1}^{x_2} f_\xi(t) dt. \quad (5)$$

Для того чтобы две СВ ξ и η с плотностями $f_\xi(x)$ и $g_\eta(y)$ были статистически независимы, необходимо и достаточно, чтобы плотность их совместного распределения $\varphi_{\xi,\eta}(x,y)$ представлялась в виде произведения: $\varphi_{\xi,\eta}(x,y) = f_\xi(x)g_\eta(y)$ (см., напр., [15], стр. 175). Далее для краткости указание на вид СВ (нижние индексы) не пишем.

Рассмотрим распределение двух СВ – с плотностями $f(x)$ и $g(y)$. Построим плотность $h(z)$ распределения разности $z = x - y$ этих величин, которые считаем независимыми. Плотность их совместного распределения равна $\varphi(x,y) = f(x)g(y)$. Вероятность $H(z)$ того, что $x - y < z$, по определению ФР (3) равна интегралу от плотности совместного распределения по области значений $x - y < z$, т.е.

$$H(z) = \iint_{x-y < z} \varphi(x,y) dx dy.$$

Поскольку СВ независимы, то

$$H(z) = \iint_{x-y < z} f(x)g(y) dx dy.$$

Чтобы найти соответствующую плотность распределения $h(z)$, надо представить этот двойной интеграл в виде

$$H(z) = \int_{-\infty}^z h(t) dt.$$

Из условия $x - y < z$ следует, что x меняется от $-\infty$ до $y + z$, тогда как сама величина y меняется от $-\infty$ до $+\infty$. Если окажется, что по условию задачи СВ изменяется в конечных пределах, то соответствующие плотности вне такой области равны нулю, так что интегрирование все равно можно распространить на всю числовую ось. Итак, тогда имеем:

$$H(z) = \iint_{x-y < z} f(x)g(y) dx dy = \int_{-\infty}^{+\infty} g(y) \left(\int_{-\infty}^{z+y} f(x) dx \right) dy.$$

Положим в последнем интеграле $x - y = t$, так что $x = t + y$. Тогда переменная t меняется от $-\infty$ до z . В соответствии с теоремой Фубини (функции g и f непрерывны и абсолютно интегрируемы) можно свести двойной интеграл к повторному:

$$H(z) = \int_{-\infty}^{+\infty} g(y) \left(\int_{-\infty}^z f(t + y) dt \right) dy = \int_{-\infty}^z h(t) dt,$$

где

$$h(t) = \int_{-\infty}^{+\infty} f(t + y)g(y) dy. \quad (6)$$

Таким образом, плотность распределения разности двух независимых случайных величин дается сверткой двух данных плотностей. Столь подробное рассмотрение этой стандартной задачи связано прежде всего с тем, что мы хотим четко обосновать уравнение, используемое в последующих разделах для нахождения распределения расстояний между двумя случайными точками в гиперкубе.

Рассмотрим сначала распределение координат x точек с плотностью $f(x)$ на отрезке $[0; 1]$. Рассмотрим на том же множестве вторую независимую от первой случайную величину с плотностью распределения $g(y)$.

Согласно (6), плотность $h(z)$ распределения разности $z = x - y$ этих величин является сверткой двух плотностей:

$$h(z) = \int_0^1 f(z + y)g(y) dy, \quad (7)$$

с областью определения $z \in [-1; 1]$. При $z \geq 0$ функция $f(z + y)$ равна нулю, если $z + y > 1$, так что y меняется тогда от 0 до $1 - z$. При $z < 0$ функция $f(z + y)$ равна нулю, если $z + y < 0$, так что y меняется в этом случае от $-z$ до 1. В результате свертка (7) представляется в виде следующей суммы двух интегралов для положительных и отрицательных значений разности $x - y$:

$$\begin{aligned}
h(z) &= \int_0^{1-z} f(z+y)g(y)dy + \int_{-z}^1 f(z+y)g(y)dy = \\
&= \int_0^{1-z} f(z+y)g(y)dy + \int_0^{1+z} f(y)g(y-z)dy \equiv U^+(z) + U^-(z).
\end{aligned} \tag{8}$$

Поэтому одномерное распределение $\rho_1(r)$ модуля разности $r = |x - y|$ СВ, заданных на отрезке $[0;1]$, дается формулой

$$\rho_1(r) = U^+(z) + U^-(z), \quad r = |z|. \tag{9}$$

Тогда из (8) и (9) получаем

$$\rho_1(r) = \int_0^{1-r} f(y+r)g(y)dy + \int_0^{1-r} f(y)g(y+r)dy, \quad r \in [0; 1]. \tag{10}$$

Отсюда следует, что если по каждой координате плотности распределения различны, то получить какие-либо обозримые результаты для многомерного пространства не представляется возможным. Мы рассмотрим простейший случай плотности распределения модуля разности двух независимых одинаково распределенных случайных величин (когда $f = g$). Тогда из (10) следует, что плотность распределения расстояний дается формулой

$$\rho_1(r) = 2 \int_0^{1-r} f(x)f(x+r)dx, \quad r \in [0; 1]. \tag{11}$$

В частности, если $f(x) = 1$, т.е. распределение координат равномерно, получаем плотность распределения расстояний в виде

$$\rho_1(r) = 2(1-r). \tag{12}$$

Соответствующая функция распределения расстояний на отрезке $[0; 1]$ есть

$$F_1(r) = \int_0^r \rho_1(t)dt = 2r - r^2. \tag{13}$$

Среднее значение расстояния между двумя точками на единичном отрезке равно

$$\langle r \rangle = \int_0^1 r \rho_1(r)dr = \int_0^1 (2r - 2r^2)dr = \frac{1}{3}.$$

Средний квадрат расстояния равен

$$\langle r^2 \rangle = \int_0^1 r^2 \rho_1(r)dr = \int_0^1 (2r^2 - 2r^3)dr = \frac{1}{6}.$$

Следовательно, дисперсия распределения расстояний на отрезке с равномерным распределением координат равна $\sigma^2 = \frac{1}{6} - \left(\frac{1}{3}\right)^2 = \frac{1}{18}$.

Рассмотрим теперь расстояние между точками в квадрате $[0; 1]^2$. Пусть по обеим координатам плотности распределения равномерны. Согласно (12),

величины модулей разностей $u = |x - x'|$ и $v = |y - y'|$ имеют плотности распределения $\rho_1(u)$ и $\rho_1(v)$.

Тогда случайные величины $z = u^2$ и $s = v^2$ имеют плотности распределения $\chi_1(z)$ и $\chi_1(s)$ соответственно, где

$$\chi_1(z) = \rho_1(\sqrt{z}) / (2\sqrt{z}) = \begin{cases} \frac{1}{\sqrt{z}} - 1, & z \in [0; 1] \\ 0, & z \notin [0; 1] \end{cases}. \quad (14)$$

Формула (14) получается при замене переменных $t = \sqrt{z}$ в интеграле, представляющем функцию распределения расстояний на отрезке. Поскольку для того, чтобы значения СВ ξ^2 были меньше r^2 , необходимо и достаточно, чтобы значения СВ ξ были меньше r , то $P(\xi < r) = P(\xi^2 < r^2)$. Тогда имеем

$$P(\xi < r) = \int_0^r \rho(t) dt = \int_0^{r^2} \frac{\rho(\sqrt{z})}{2\sqrt{z}} dz.$$

Но

$$P(\xi^2 < r^2) = \int_0^{r^2} \chi(z) dz,$$

откуда и следует (14).

Тогда случайная величина $w = z + s \equiv u^2 + v^2$ определена на отрезке $[0; 2]$ и формально имеет плотность распределения

$$\chi_2(w) = \int_0^w \chi_1(w-s) \chi_1(s) ds. \quad (15)$$

Однако надо учесть, что плотность χ_1 равна нулю вне отрезка $[0; 1]$. Поэтому в зависимости от величины w получаем для плотности $\chi_2(w)$ распределения квадрата расстояний в двумерном кубе $[0; 1]^2$ выражения

$$\chi_2(w) = \begin{cases} \int_0^w \frac{(1-\sqrt{s})(1-\sqrt{w-s})}{\sqrt{(w-s)s}} ds, & 0 \leq w < 1; \\ \int_{w-1}^1 \frac{(1-\sqrt{s})(1-\sqrt{w-s})}{\sqrt{(w-s)s}} ds, & 1 \leq w \leq 2. \end{cases} \quad (16)$$

Первый интеграл в (16) преобразуется к виду

$$\begin{aligned}\chi_2(w) &= \int_0^w \frac{ds}{\sqrt{s(w-s)}} - \int_0^w \frac{ds}{\sqrt{w-s}} - \int_0^w \frac{ds}{\sqrt{s}} + \int_0^w ds = \\ &= \int_0^w \frac{ds}{\sqrt{s(w-s)}} + 2\sqrt{w-s} \Big|_0^w - 2\sqrt{s} \Big|_0^w + w = \int_0^w \frac{ds}{\sqrt{s(w-s)}} + w - 4\sqrt{w}.\end{aligned}$$

Оставшийся в последнем выражении интеграл вычисляется с помощью выделения полного квадрата под корнем:

$$\int_0^w \frac{ds}{\sqrt{s(w-s)}} = \int_0^w \frac{ds}{\sqrt{\frac{w^2}{4} - \left(s - \frac{w}{2}\right)^2}} = \frac{2}{w} \int_{-w/2}^{w/2} \frac{dt}{\sqrt{1 - \left(\frac{2t}{w}\right)^2}} = 2 \int_0^1 \frac{d\xi}{\sqrt{1 - \xi^2}} = \pi.$$

В результате получаем для плотности квадрата расстояния в первой области выражение

$$\chi_2(w) = \pi - 4\sqrt{w} + w, \quad 0 \leq w < 1. \quad (17)$$

Плотность распределения расстояний в этой области есть

$$\rho_2(r) = 2r\chi_2(r^2) = 2\pi r - 8r^2 + 2r^3, \quad 0 \leq r < 1. \quad (18)$$

Аналогично вычисляются интегралы в (16) при $1 < w < 2$:

$$\chi_2(w) = 2 \arcsin\left(\frac{2}{w} - 1\right) + 4\sqrt{w-1} - w - 2,$$

так что

$$\rho_2(r) = 4r \arcsin\left(\frac{2}{r^2} - 1\right) + 8r\sqrt{r^2 - 1} - 2r^3 - 4r, \quad 1 < r \leq \sqrt{2}. \quad (19)$$

Тогда функция распределения расстояний между точками в квадрате $[0; 1]^2$ с равномерным распределением координат дается формулой

$$F_2(r) = \begin{cases} \pi r^2 - \frac{8}{3}r^3 + \frac{1}{2}r^4, & 0 \leq r \leq 1; \\ \frac{1}{3} + 2r^2 \arcsin\left(\frac{2}{r^2} - 1\right) + 4\sqrt{r^2 - 1} + \frac{8}{3}(r^2 - 1)^{3/2} - 2r^2 - \frac{r^4}{2}, & 1 < r \leq \sqrt{2}. \end{cases} \quad (20)$$

Как видно, эта функция уже не слишком проста для аналитического исследования. То есть выражения сами по себе не особенно сложные, но их громоздкость затрудняет понимание. В частности, среднее расстояние между точками в единичном квадрате равно $\langle r \rangle = (2 + \sqrt{2})/15 + \frac{1}{3} \ln(1 + \sqrt{2}) \approx 0,5214$.

Эта числовая величина уже не имеет наглядной геометрической интерпретации.

Для трехмерного случая надо рассмотреть распределение суммы трех случайных величин (квадратов разностей координат), каждая из которых имеет плотность $\chi_1(x)$, т.е. плотность распределения квадрата расстояния в кубе $[0; 1]^3$ может быть построена в соответствии с уравнением:

$$\chi_3(w) = \int_0^w \chi_1(u) du \int_0^{w-u} \chi_1(w-u-v) \chi_1(v) dv = \int_0^w \chi_1(u) \chi_2(w-u) du. \quad (21)$$

При этом аргументы функций χ_1 в (21) неотрицательны и не больше единицы, тогда как аргумент w суммы трех квадратов меняется от 0 до 3. Тогда интеграл в формуле (21) разбивается на следующие варианты, заметно различающиеся между собой.

1. Если $0 \leq w \leq 1$, то интеграл имеет последний вид (21). Явная формула плотности распределения квадрата расстояния тогда следующая:

$$\chi_3(r^2) = 2\pi r - 3\pi r^2 + 4r^3 - r^4 / 2, \quad 0 \leq r < 1. \quad (22)$$

2. Если $1 \leq w < 2$, то u в первом интеграле (21) меняется от нуля до единицы, а v во втором интеграле (21) меняется либо от нуля до $w-u$, если $w-u \leq 1$, либо от нуля до единицы, если $w-u > 1$. Следовательно, в этом случае получаем выражение

$$\chi_3(w) = \int_0^{w-1} \psi(u) du \int_0^1 \psi(w-u-v) \psi(v) dv + \int_{w-1}^1 \psi(u) du \int_0^{w-u} \psi(w-u-v) \psi(v) dv.$$

Явно получающаяся при этом формула для плотности распределения имеет вид

$$\chi_3(r^2) = 3\pi - \frac{1}{2} - 4\pi r + 3r^2 + r^4 + 12r^2 \arctg \sqrt{r^2 - 1} - 4r(1 + 2r^2) \sqrt{r^2 - 1}, \quad 1 < r \leq \sqrt{2}.$$

3. Если $2 < w \leq 3$, то всегда $w-u > 1$, так что v меняется от нуля до единицы. Поскольку же при этом $w-1 > 1$, то и u меняется от нуля до единицы. Тогда плотность вероятности квадрата расстояния имеет вид

$$\chi_3(w) = \int_0^1 \psi(u) du \int_0^1 \psi(w-u-v) \psi(v) dv.$$

Явная формула имеет достаточно сложный вид. Тем не менее ее полезно привести для демонстрации громоздкости получающихся выражений и для понимания того, что для размерностей выше трех эмпирические распределения оказываются содержательнее, чем аналитические.

$$\begin{aligned} \chi_3(r^2) = & \frac{\pi}{2} - 1 + \sqrt{r^2 - 2} - \sqrt{r^2 - 1} + \frac{4}{3} \left(r^3 - 2(r^2 - 1)^{3/2} + (r^2 - 2)^{3/2} \right) + \\ & + (r^2 + 3) \arcsin \left(\frac{1}{\sqrt{r^2 - 1}} \right) - 4r \cdot \arctg \left(r \cdot \tg \left(\frac{1}{\sqrt{r^2 - 1}} \right) \right) + \\ & + 2r^2 \left(\arctg(r^2 - 1) - \arctg(r^2 - 2) \right) - r^2 \arcsin \frac{1}{r}, \quad \sqrt{2} < r \leq \sqrt{3}. \end{aligned}$$

Среднее расстояние между случайными точками в трехмерном кубе равно

$$\langle r \rangle = \frac{4 + 17\sqrt{2} - 6\sqrt{3} - 42 \ln 2 - 7\pi + 21 \ln(1 + \sqrt{2}) + 84 \ln(1 + \sqrt{3})}{105} \approx 0,6617,$$

это и есть константа Роббинса.

Далее в п. 2 и 3 рассматриваем распределения расстояний в области $r < 1$.

2. Плотность распределения квадрата расстояния в гиперкубе для $r < 1$

В области $0 \leq r < 1$ и в пределе при $r \rightarrow 1 - 0$ аналитические выражения для функций распределения расстояний выглядят относительно просто. В качестве примера получим их для размерностей $n = 4$ и $n = 5$.

Используя (14) и (22), имеем для распределения квадрата расстояний в кубе $[0; 1]^4$:

$$\begin{aligned}\chi_4(w) &= \int_0^w \chi_1(t) \chi_3(w-t) dt = \int_0^w \chi_3(t) \chi_1(w-t) dt = \\ &= \int_0^w \left(2\pi\sqrt{t} - 3\pi t + 4t^{3/2} - \frac{t^2}{2} \right) \left(-1 + \frac{1}{\sqrt{w-t}} \right) dt = \\ &= \pi^2 w - \frac{16}{3} \pi w^{3/2} + 3\pi w^2 - \frac{32}{15} w^{5/2} + \frac{1}{6} w^3.\end{aligned}\quad (23)$$

После этого находим

$$\begin{aligned}\rho_4(r) &= 2r\chi_4(r^2) = 2\pi^2 r^3 - \frac{32}{3} \pi r^4 + 6\pi r^5 - \frac{64}{15} r^6 + \frac{1}{3} r^7, \\ F_4(r) &= \int_0^r \rho_4(t) dt = \frac{\pi^2}{2} r^4 - \frac{32}{15} \pi r^5 + \pi r^6 - \frac{64}{105} r^7 + \frac{1}{24} r^8.\end{aligned}$$

Аналогично имеем для куба $[0; 1]^5$ в области $0 \leq r < 1$:

$$\begin{aligned}\chi_5(w) &= \int_0^w \chi_4(t) \left(-1 + \frac{1}{\sqrt{w-t}} \right) dt = \\ &= \frac{4\pi^2}{3} w^{3/2} - \frac{5}{2} \pi^2 w^2 + \frac{16\pi}{3} w^{5/2} - \frac{5\pi}{3} w^3 + \frac{16}{21} w^{7/2} - \frac{1}{24} w^4. \\ \rho_5(r) &= 2r\chi_5(r^2) = \\ &= \frac{8\pi^2}{3} r^4 - 5\pi^2 r^5 + \frac{32\pi}{3} r^6 - \frac{10\pi}{3} r^7 + \frac{32}{21} r^8 - \frac{1}{12} r^9. \\ F_5(r) &= \int_0^r \rho_5(t) dt = \frac{8\pi^2}{15} r^5 - \frac{5}{6} \pi^2 r^6 + \frac{32\pi}{21} r^7 - \frac{5\pi}{12} r^8 + \frac{32}{189} r^9 - \frac{1}{120} r^{10}.\end{aligned}\quad (24)$$

Получим теперь общее выражение для функции распределения квадрата расстояния между случайными точками в гиперкубе при $0 \leq r < 1$, а также следующее из него выражение для функции распределения расстояний.

Из предыдущего рассмотрения следует, что в этом промежутке имеет место рекуррентное соотношение между величинами $\chi_n(w)$ в виде

$$\chi_{n+1}(w) = \int_0^w \chi_n(t) \left(-1 + \frac{1}{\sqrt{w-t}} \right) dt. \quad (25)$$

Выражение для $\chi_n(w)$ в указанной области представляет собой многочлен по степеням $\sqrt{w}$ вида

$$\chi_n(w) = \sum_{k=n-2}^{2n-2} a_{n,k} w^{k/2}. \quad (26)$$

Коэффициенты $a_{n,k}$ в (26) подлежат определению. Их нахождение – одна из основных задач данного исследования.

Соответствующая функция распределения расстояний имеет вид

$$F_n(r) = 2 \int_0^r t \chi_n(t^2) dt = \sum_{k=n-2}^{2n-2} \frac{2}{k+2} a_{n,k} r^{k+2} = \sum_{k=n}^{2n} \frac{2}{k} a_{n,k-2} r^k. \quad (27)$$

Подставим выражение (26) в (25). Тогда, с одной стороны, согласно (26) имеем представление

$$\chi_{n+1}(w) = \sum_{k=n-1}^{2n} a_{n+1,k} w^{k/2},$$

а с другой стороны, после подстановки (26) в (25) получаем

$$\begin{aligned} \chi_{n+1}(w) &= \int_0^w \left(-1 + \frac{1}{\sqrt{w-t}} \right) \sum_{k=n-2}^{2n-2} a_{n,k} t^{k/2} dt = \\ &= - \sum_{k=n-2}^{2n-2} a_{n,k} \frac{2w^{(k+2)/2}}{k+2} + \int_0^w \sum_{k=n-2}^{2n-2} a_{n,k} \left(t^{k/2} / \sqrt{w-t} \right) dt. \end{aligned} \quad (28)$$

В последнем интеграле сделаем замену $t/w = q$. Тогда он приводится к виду

$$\int_0^1 \sum_{k=n-2}^{2n-2} a_{n,k} w^{(k+1)/2} \left(q^{k/2} / \sqrt{1-q} \right) dq = 2 \sum_{k=n-2}^{2n-2} a_{n,k} w^{(k+1)/2} \int_0^1 (1-\xi^2)^{k/2} d\xi.$$

Введем обозначения для получающихся интегралов в зависимости от четности k :

$$J_{k/2} = \int_0^1 \frac{x^{k/2}}{\sqrt{1-x}} dx = 2 \int_0^1 (1-t^2)^{k/2} dt = \begin{cases} 2 \frac{(2n)!!}{(2n+1)!!}, & k = 2n; \\ \pi \frac{(2n-1)!!}{(2n)!!}, & k = 2n-1. \end{cases} \quad (29)$$

Тогда из (28) после взятия интегралов в правой части получаем

$$\begin{aligned} \chi_{n+1}(w) &= \sum_{k=n-1}^{2n} a_{n+1,k} w^{k/2} = - \sum_{k=n-2}^{2n-2} \frac{2}{k+2} a_{n,k} w^{(k+2)/2} + \sum_{k=n-2}^{2n-2} a_{n,k} J_{k/2} w^{(k+1)/2} = \\ &= - \sum_{k=n}^{2n} \frac{2}{k} a_{n,k-2} w^{k/2} + \sum_{k=n-1}^{2n-1} a_{n,k-1} J_{(k-1)/2} w^{k/2}. \end{aligned}$$

Это равенство удобно записать в следующем виде, выделив первые и последние слагаемые в суммах:

$$\begin{aligned} a_{n+1,n-1} w^{(n-1)/2} + \sum_{k=n}^{2n-1} a_{n+1,k} w^{k/2} + a_{n+1,2n} w^n &= a_{n,n-2} J_{(n-2)/2} w^{(n-1)/2} - \\ - \frac{1}{n} a_{n,2n-2} w^n + \sum_{k=n}^{2n-1} a_{n,k-1} J_{(k-1)/2} w^{k/2} &- \sum_{k=n}^{2n-1} \frac{2}{k} a_{n,k-2} w^{k/2}. \end{aligned} \quad (30)$$

Приравнивая в (30) коэффициенты при одинаковых степенях $\sqrt{w}$, получаем систему для определения коэффициентов $a_{n,k}$:

$$\begin{aligned} w^{(n-1)/2} : \quad a_{n+1,n-1} &= a_{n,n-2} J_{\frac{n-1}{2}}; \\ w^n : \quad a_{n+1,2n} &= -\frac{1}{n} a_{n,2n-2}; \\ w^{k/2}, n \leq k \leq 2n-1 : \quad a_{n+1,k} &= -\frac{2}{k} a_{n,k-2} + a_{n,k-1} J_{\frac{k-1}{2}}. \end{aligned} \quad (31)$$

Для решения этих рекуррентных соотношений используем начальные значения коэффициентов $a_{n,k}$ с учетом выражений (22-24) и (27). Из них получаем:

$$\begin{aligned} a_{1,-1} &= 1; \quad a_{1,0} = -1; \\ a_{2,0} &= \pi; \quad a_{2,1} = -4; \quad a_{2,2} = 1; \\ a_{3,1} &= 2\pi; \quad a_{3,2} = -3\pi; \quad a_{3,3} = 4; \quad a_{3,4} = -1/2. \end{aligned} \quad (32)$$

2.1. Первое уравнение рекурсии.

Первое рекуррентное соотношение в (31) приводится к виду

$$a_{n,n-2} = a_{2,0} J_0 J_{\frac{1}{2}} \cdots J_{\frac{n-1}{2}-1} = 2\pi J_{\frac{1}{2}} \cdots J_{\frac{n-1}{2}-1}. \quad (33)$$

Поскольку $J_{k-1/2} J_k = 2\pi / (2k+1)$, то для произведения таких величин получаем

$$(J_{1/2} J_1)(J_{3/2} J_2) \cdots (J_{n-1/2} J_n) = \frac{(2\pi)^n}{(2n+1)!!}. \quad (34)$$

Отсюда следует, что если n в (34) есть нечетное число, т.е. $n = 2m+1$, то

$$a_{2m+1,2m-1} = 2\pi J_{\frac{1}{2}} J_1 \cdots J_{m-\frac{3}{2}} J_{m-1} = 2\pi \frac{(2\pi)^{m-1}}{(2m-1)!!}.$$

Если же в (34) $n = 2m$, то с учетом (28) получаем

$$a_{2m,2m-2} = 2\pi J_{\frac{1}{2}} J_1 \cdots J_{m-2} J_{m-\frac{3}{2}} = 2\pi \frac{(2\pi)^{m-2}}{(2m-3)!!} J_{m-\frac{3}{2}} = \pi \frac{(2\pi)^{m-1}}{(2m-2)!!}.$$

Поскольку для гамма-функции имеет место формула

$$\Gamma\left(n + \frac{1}{2}\right) = \sqrt{\pi} \frac{(2n-1)!!}{2^n}, \text{ и } (2m-2)!! = 2^{m-1} (m-1)! = 2^{m-1} \Gamma(m), \text{ то}$$

$$a_{2m+1,2m-1} = \frac{\pi^{m+1/2}}{\Gamma(m+1/2)}, \quad a_{2m,2m-2} = \frac{\pi^m}{\Gamma(m)}.$$

Видно, что обе последние формулы объединяются в одну:

$$a_{n,n-2} = \frac{\pi^{n/2}}{\Gamma(n/2)}. \quad (35)$$

Следовательно, первый коэффициент в представлении (27) функции распределения расстояний равен

$$\frac{2}{n} a_{n,n-2} = \frac{\pi^{n/2}}{\frac{n}{2} \Gamma\left(\frac{n}{2}\right)} = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2} + 1\right)} = V_n, \quad (36)$$

где V_n есть объем n -мерного шара единичного радиуса.

Через объем гипершара соответствующей размерности можно также представить и произведения величин (34). Например,

$$(J_{1/2} J_1)(J_{3/2} J_2) \dots (J_{n-1/2} J_n) = \frac{(2\pi)^n}{(2n+1)!!} = \frac{(2\pi)^n \sqrt{\pi}}{2^{n+1} \Gamma\left(\frac{2n+1}{2} + 1\right)} = \frac{1}{2} V_{2n+1}.$$

Поскольку же $J_0 = 2$, то $J_0(J_{1/2} J_1)(J_{3/2} J_2) \dots (J_{n-1/2} J_n) = V_{2n+1}$.

Аналогично для нечетного произведения получаем

$$(J_0 J_{1/2})(J_1 J_{3/2}) \dots (J_n J_{n+1/2}) = \frac{\pi^n}{n!} = V_{2n}.$$

Следовательно,

$$V_n = J_0 J_{1/2} \dots J_{(n-1)/2}. \quad (37)$$

Тогда при $k \leq n$ имеет место представление

$$V_n = V_k J_{k/2} \dots J_{(n-1)/2}. \quad (38)$$

Формулу (38) удобно использовать далее при решении рекурсий (31).

2.2. Второе уравнение рекурсии.

Второе рекуррентное соотношение в (31) также разрешается непосредственно:

$$a_{n,2n-2} = \frac{(-1)^n}{(n-1)!}. \quad (39)$$

Тогда последний коэффициент в функции распределения равен $(-1)^n / n!$, что мы видели ранее в соответствии с явными формулами (23, 24).

2.3. Предпоследний коэффициент рекурсии.

Третье уравнение системы (31) последовательно рекуррентно разрешимо, так как оставшаяся система представляет собой треугольную матрицу. Например, с учетом (39) и (29) предпоследний коэффициент $a_{n+1,2n-1}$ в этой системе удовлетворяет уравнению

$$a_{n+1,2n-1} = -\frac{2}{2n-1} a_{n,2n-3} + a_{n,2n-2} J_{n-1} = -\frac{2}{2n-1} a_{n,2n-3} + \frac{(-1)^n 2^n}{(2n-1)!!}.$$

Обозначим для сокращения промежуточных вычислений свободный член рекурсии как известную величину через $b_n = \frac{(-1)^n 2^n}{(2n-1)!!}$. Тогда получаем

$$a_{n+1,2n-1} = b_n - \frac{2}{2n-1} b_{n-1} + \frac{2}{2n-1} \frac{2}{2n-3} b_{n-2} + \dots$$

Заметим, что

$$-\frac{2}{2n-1}b_{n-1}=b_n, \quad \frac{2}{2n-1}\frac{2}{2n-3}b_{n-2}=b_n, \dots$$

В результате получаем

$$a_{n+1,2n-1} = \frac{(-1)^n 2^n (n+1)}{(2n-1)!!}$$

или

$$a_{n,2n-3} = \frac{(-1)^{n-1} 2^{n-1} n}{(2n-3)!!}. \quad (40)$$

2.4. Нахождение коэффициента $a_{n+1,n}$.

Вернемся теперь к началу рекурсии в третьем равенстве (31) и найдем коэффициент $a_{n+1,n}$. Для него из (31) имеем

$$a_{n+1,n} = -\frac{2}{n}a_{n,n-2} + a_{n,n-1}J_{(n-1)/2}.$$

Поскольку из (36) следует, что $\frac{2}{n}a_{n,n-2} = V_n$, то имеем

$$a_{n+1,n} = -V_n + a_{n,n-1}J_{(n-1)/2} = -V_n + J_{\frac{n-1}{2}} \left(-V_{n-1} + J_{\frac{n-1}{2}} \left(-V_{n-2} + J_{\frac{n-3}{2}} (\dots) \right) \right).$$

Но, согласно (31), $V_n = V_{n-1}J_{(n-1)/2} = V_{n-2}J_{(n-2)/2}J_{(n-1)/2} = \dots$, так что $a_{n+1,n} = -(n+1)V_n$, т.е.

$$a_{n,n-1} = -nV_{n-1}. \quad (41)$$

2.5. Нахождение коэффициента $a_{n+1,n+1}$.

Аналогично поступаем и далее. Используя (41), имеем из (31) для $a_{n+1,n+1}$:

$$a_{n+1,n+1} = -\frac{2}{n+1}a_{n,n-1} + a_{n,n}J_{n/2} = \frac{2n}{n+1}V_{n-1} + a_{n,n}J_{n/2}.$$

Обозначим для сокращения промежуточных вычислений свободный член рекурсии как известную величину через $c_n = \frac{2n}{n+1}V_{n-1}$. Тогда получаем

$$a_{n+1,n+1} = c_n + c_{n-1}J_{n/2} + c_{n-2}J_{n/2}J_{(n-1)/2} + \dots + c_2J_{n/2}\dots J_{3/2} + a_{2,2}J_{n/2}\dots J_1, \quad (42)$$

причем $a_{2,2} = 1$, согласно (32).

В последнем выражении произведения интегралов (29) вида $J_{k/2}$ преобразуем с помощью формул (37, 38), и с учетом того, что $J_0J_{1/2} = 2 \cdot (\pi/2) = \pi$. Тогда

$$J_{n/2} \cdot \dots \cdot J_{(n-k+1)/2} = \frac{J_{n/2} \cdot \dots \cdot J_1}{J_{(n-k)/2} \cdot \dots \cdot J_1} = \frac{V_{n+1}/\pi}{V_{n-k+1}/\pi} = \frac{V_{n+1}}{V_{n-k+1}}.$$

При этом

$$\begin{aligned}
c_n &= \frac{2n}{n+1} V_{n-1} = \frac{2n}{n+1} V_{n+1} \frac{V_{n-1}}{V_{n+1}} = \frac{2n}{n+1} V_{n+1} \frac{1}{\pi} \frac{\Gamma((n+1)/2 + 1)}{\Gamma((n-1)/2 + 1)} = \\
&= \frac{2n}{n+1} V_{n+1} \frac{(n+1)/2}{\pi} = \frac{n V_{n+1}}{\pi}.
\end{aligned}$$

Тогда (42) преобразуется к виду:

$$a_{n+1,n+1} = \sum_{k=0}^{n-1} c_{n-k} \frac{V_{n+1}}{V_{n-k+1}} = \frac{V_{n+1}}{\pi} \sum_{k=0}^{n-1} (n-k) = \frac{V_{n+1}}{\pi} \sum_{k=1}^n k = \frac{n(n+1)}{2\pi} V_{n+1},$$

т.е.

$$a_{n,n} = \frac{n(n-1)}{2\pi} V_n = \frac{1}{\pi} C_n^2 V_n. \quad (43)$$

2.6. Нахождение коэффициента $a_{n+1,n+2}$.

Используя (43), из третьего уравнения системы (31) получаем:

$$a_{n+1,n+2} = -\frac{2}{n+2} a_{n,n} + a_{n,n+1} J_{(n+1)/2} = -\frac{2}{n+2} \frac{C_n^2 V_n}{\pi} + a_{n,n+1} J_{(n+1)/2}.$$

Обозначим для сокращения промежуточных вычислений свободный член рекурсии как известную величину через $d_n = -\frac{2}{n+2} \frac{C_n^2 V_n}{\pi}$. Тогда получаем:

$$a_{n+1,n+2} = d_n + d_{n-1} J_{(n+1)/2} + d_{n-2} J_{(n+1)/2} J_{n/2} + \dots$$

Поскольку же

$$\begin{aligned}
d_{n-k} J_{(n+1)/2} J_{n/2} \dots J_{(n-k+2)/2} &= d_{n-k} \frac{V_{n+2}}{V_{n-k+2}} = -\frac{2}{n-k+2} C_{n-k}^2 \frac{V_{n-k} V_{n+2}}{V_{n-k+2}} = \\
&= -\frac{V_{n+2}}{\pi} \frac{2}{n-k+2} \frac{(n-k)(n-k+1)}{2} \frac{1}{\pi} \frac{\Gamma((n-k+2)/2 + 1)}{\Gamma((n-k)/2 + 1)} = -\frac{V_{n+2}}{2\pi^2} (n-k)(n-k+1),
\end{aligned}$$

то

$$\begin{aligned}
a_{n+1,n+2} &= -\frac{V_{n+2}}{2\pi^2} \sum_{k=0}^{n-1} (n-k)(n-k+1) = -\frac{V_{n+2}}{2\pi^2} \sum_{k=1}^n k(k+1) = \\
&= -\frac{V_{n+2}}{2\pi^2} \left(\frac{n(n+1)}{2} + \frac{n(n+1)(2n+1)}{6} \right) = -\frac{V_{n+2}}{6\pi^2} n(n+1)(n+2) = -C_{n+2}^3 \frac{V_{n+2}}{\pi^2}.
\end{aligned}$$

Таким образом,

$$a_{n,n+1} = -\frac{1}{\pi^2} C_{n+1}^3 V_{n+1}. \quad (44)$$

Аналогично для остальных коэффициентов в (31) получаем

$$a_{n,n+p} = \frac{(-1)^p}{\pi^{p+1}} C_n^{p+2} V_{n+p}, \quad p = 0, 1, \dots, n-4. \quad (45)$$

Итак, все коэффициенты представления (26, 27) функции распределения расстояний в n -мерном кубе найдены. Эти формулы обобщают частные случаи, приведенные выше для $n = 2 \div 5$ (формулы (12), (17), (23), (24)).

Заметим теперь, что все вообще коэффициенты разложения (27) определяются формулами (45). Действительно, если положить в (45) $k = n$, то получаем

$$\frac{2}{n+k} \frac{(-1)^k}{\pi^{k-1}} C_n^k V_{n+k-2} \Big|_{k=n} = \frac{(-1)^n}{n\pi^{n-1}} \frac{\pi^{(2n-2)/2}}{\Gamma(n)} = \frac{(-1)^n}{n!},$$

что соответствует (39).

Аналогично, если положить в (45) $k = n - 1$, то имеем

$$\frac{2}{n+k} \frac{(-1)^k}{\pi^{k-1}} C_n^k V_{n+k-2} \Big|_{k=n-1} = \frac{2}{2n-1} \frac{(-1)^{n-1}}{\pi^{n-2}} \frac{\pi^{(2n-3)/2} n}{\Gamma(n-1+1/2)}.$$

Поскольку же

$$\Gamma\left(n + \frac{1}{2}\right) = \sqrt{\pi} \frac{(2n-1)!!}{2^n},$$

то получаем выражение, соответствующее (40).

Точно так же второе слагаемое в (45) соответствует выражению (41):

$$\frac{2}{n+k} \frac{(-1)^k}{\pi^{k-1}} C_n^k V_{n+k-2} \Big|_{k=1} = \frac{-2n}{n+1} V_{n-1},$$

а первое слагаемое при $k = 0$ равно $\frac{2}{n} a_{n,n-2} = V_n$ в соответствии с (36).

Итак, в итоге мы получаем точную формулу для представления функции распределения расстояний между случайными точками в гиперкубе в области $0 \leq r < 1$. Соответствующая сумма (27) имеет вид

$$F_n(r) = \sum_{k=n}^{2n} \frac{2}{k} a_{n,k-2} r^k = \sum_{k=0}^n \frac{2}{n+k} \frac{(-1)^k}{\pi^{k-1}} C_n^k V_{n+k-2} r^{n+k}. \quad (46)$$

При $r \rightarrow 1$ из (46) получаем выражение для вероятности того, что расстояние между двумя точками в гиперкубе не больше единицы, т.е. две точки лежат в шаре единичного радиуса. Поскольку сумма (46) знакопеременная, то выделение главной части коэффициента при больших n для оценки главной части суммы не работает, и нужны более тонкие методы.

3. Асимптотика функции распределения расстояний для $r < 1$

Оценим асимптотику функции распределения расстояний между точками с равномерно распределенными координатами в n -мерном кубе в области $0 < r < 1$. Численные расчеты показывают, что значение функции распределения расстояний в точке $r = 1$ очень быстро спадает с увеличением размерности. На рис. 1 показана зависимость $F_n(r)$ с шагом по длине 10^{-2} , а на рис. 2 приведена зависимость значения $F_n(1)$ от размерности гиперкуба.

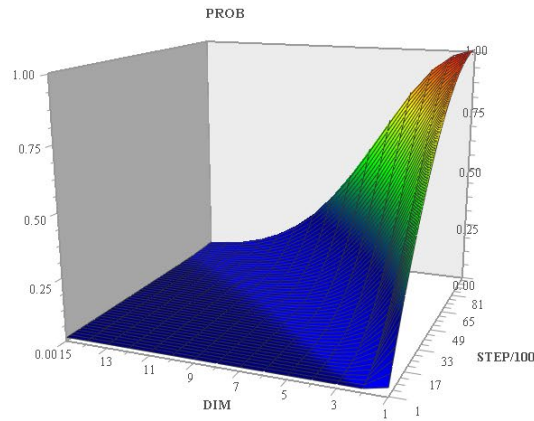


Рис. 1. Функция распределения расстояний в гиперкубе

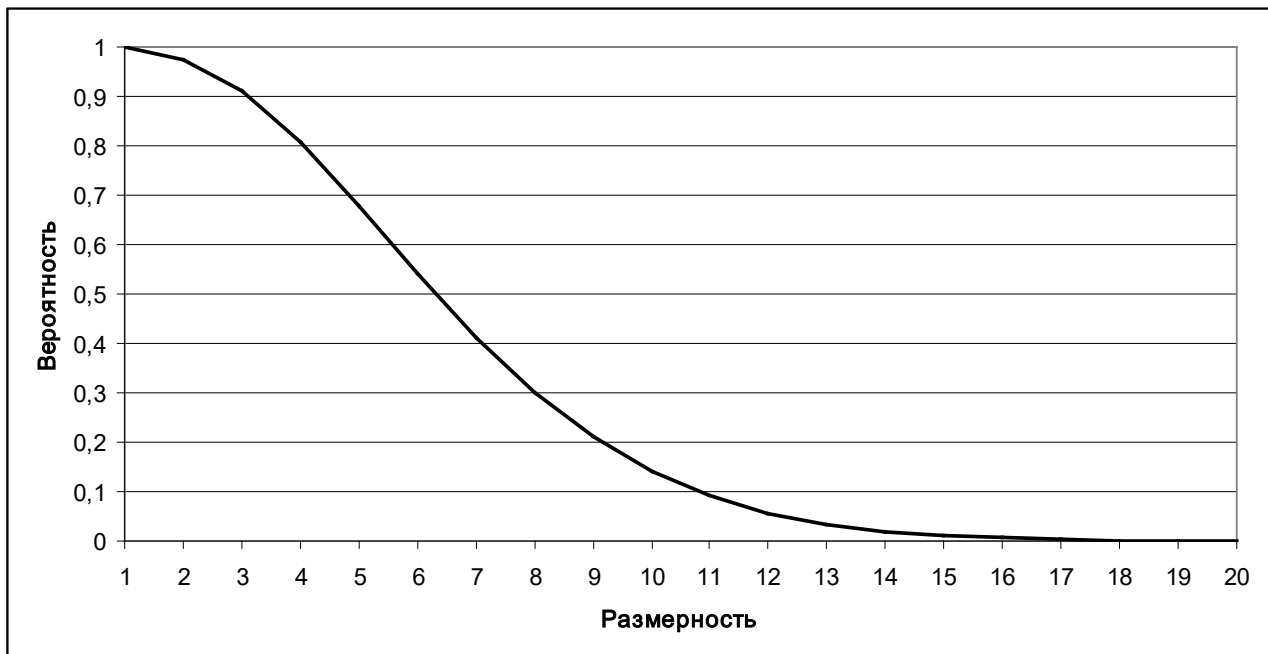


Рис. 2. Зависимость функции распределения расстояний от размерности гиперкуба при $r = 1$

Расчеты по формуле (27) дают значение $F_{20}(1) \approx 5 \cdot 10^{-4}$, а $F_{30}(1) \approx 2 \cdot 10^{-7}$. Это, в частности, означает, что из множества равномерно распределенных точек в 30-мерном кубе лишь одна из 5 млн точек имеет соседа на расстоянии, меньшем чем 1 (длина ребра куба).

Аппроксимация значений $F_n(1)$ при $n > 10$ с детерминацией выше 0,99 дает снижение вероятности вида $F_n(1) \propto \exp(-0,5n \ln n)$, что соответствует тому, что объем n -мерного шара убывает примерно как $1/\sqrt{n!}$.

Оценим более точно скорость убывания вероятности того, что расстояние между точками в шаре находится в области $r < 1$. Для этого надо выделить главный по n член в асимптотике величины $F_n(1)$ в сумме (27) при $n \gg 1$.

Используем формулу Стирлинга для гамма-функции, согласно которой при $n \gg 1$ имеет место представление

$$\Gamma(z) \propto \left(\frac{z}{e}\right)^z \sqrt{\frac{2\pi}{z}}. \quad (47)$$

Тогда из (37) заключаем, что объем n -мерного шара асимптотически равен

$$V_n = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2} + 1\right)} \propto \sqrt{\frac{e}{\pi(n+2)}} \left(\frac{2\pi e}{n+2}\right)^{n/2} \approx \sqrt{\frac{e}{\pi n}} \left(\frac{2\pi e}{n}\right)^{n/2}. \quad (48)$$

Таким образом, при больших n объем шара убывает как $n^{-(n+1)/2}$ (рис. 3), максимальное значение достигается при $n=5$. При $n=30$ объем шара равен примерно $2 \cdot 10^{-5}$, а при $n=50$ – примерно $2 \cdot 10^{-13}$.

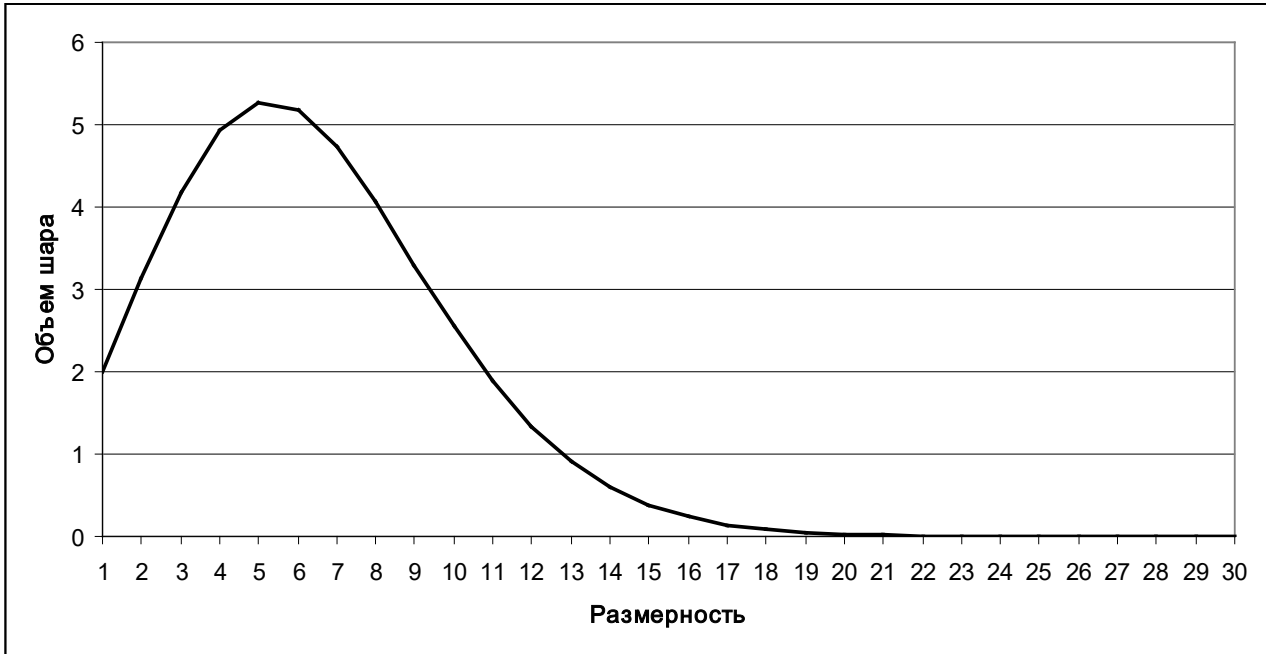


Рис. 3. Зависимость объема шара единичного радиуса от размерности

Поведение общего члена в сумме (46) нетривиально. Типичная зависимость от индекса суммирования абсолютной величины коэффициента

$$A_k(n) = \frac{2}{n+k} \frac{1}{\pi^{k-1}} C_n^k V_{n+k-2} \quad (49)$$

приведена на рис. 4. При малых k коэффициент (49) как функция n растет с увеличением индекса суммирования, так что $\frac{A_{k+1}(n)}{A_k(n)} \propto \sqrt{n}$. Это следует из

того, что биномиальный коэффициент при малых k возрастает примерно в n раз при увеличении k на единицу, а объем уменьшается в $\sqrt{n}$ раз. С другой стороны, при $k \approx n$ биномиальный коэффициент уменьшается в n раз с ростом k , так что получаем $\frac{A_{k+1}(n)}{A_k(n)} \propto \frac{1}{n^{3/2}}$.

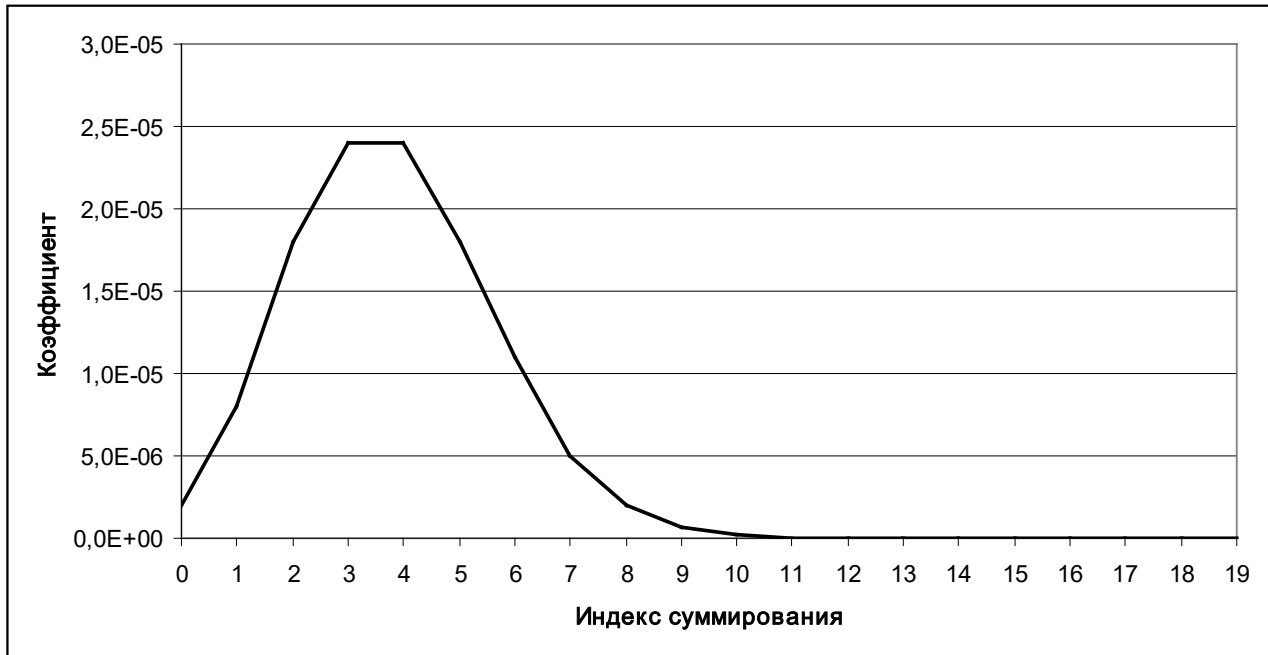


Рис. 4. Значения коэффициентов A_k ($n = 30$)

Поскольку слагаемые в сумме (46) знакопеременные, то асимптотика по n вероятности $F_n(1)$ определяется не главным по n слагаемым, а всей суммой. Частичная сумма в начале суммирования знакопеременная, но по мере уменьшения слагаемых выходит на положительную асимптотику. На рис. 5 показан пример суммы (46) для $n = 30$ при частичном суммировании.

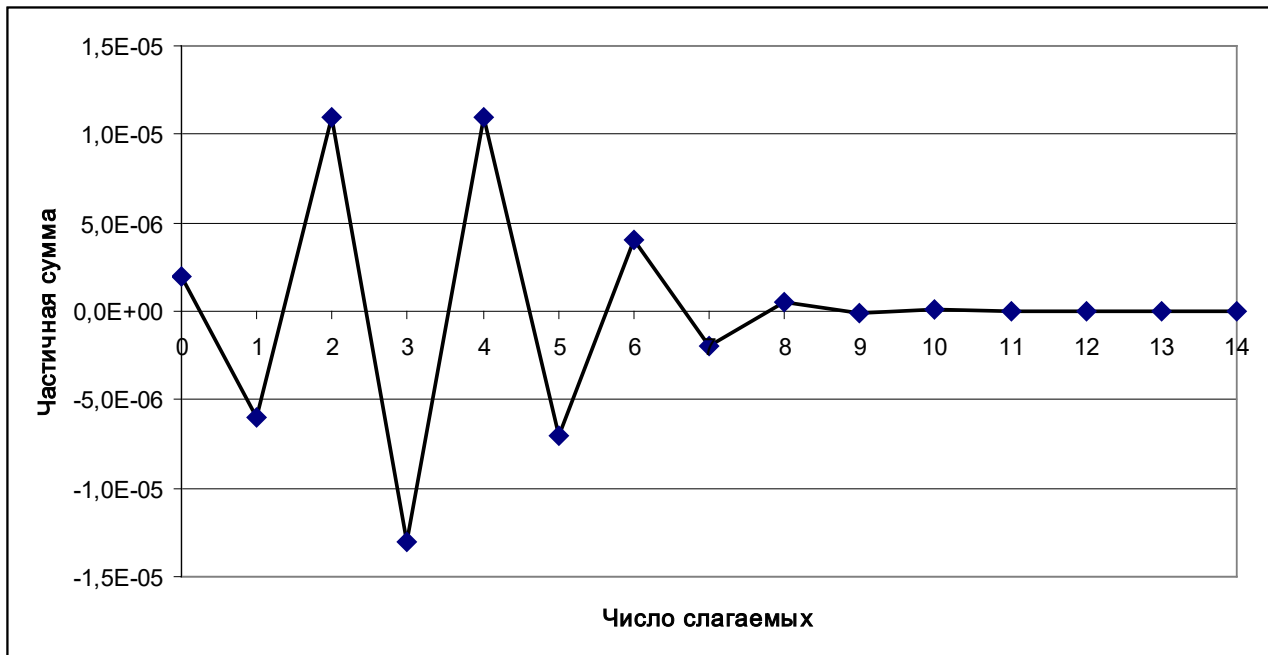


Рис. 5. Значения частичной суммы функции распределения для $n = 30$

Для получения оценки скорости убывания вероятности $F_n(1)$ с увеличением размерности поступим следующим образом. Формулу (46) можно представить в виде

$$F_n(r) = \sum_{k=0}^n \frac{2}{n+k} \frac{(-1)^k}{\pi^{k-1}} C_n^k V_{n+k-2} r^{n+k} = 2\pi \int_0^r \sum_{k=0}^n \frac{(-1)^k}{\pi^k} C_n^k V_{n+k-2} x^{n+k-1} dx.$$

Следовательно, сумма, которая требует оценивания, имеет вид

$$S_n(x) = \sum_{k=0}^n \frac{(-1)^k}{\pi^k} C_n^k V_{n+k-2} x^k.$$

С учетом (48) эта сумма при $n \gg 1$ представляется в виде

$$\begin{aligned} S_n(x) &\propto \sqrt{\frac{e}{\pi n}} \sum_{k=0}^n \frac{(-1)^k}{\pi^k} C_n^k \left(\frac{2\pi e}{n+k} \right)^{(n+k)/2} x^k = \\ &= \sqrt{\frac{e}{\pi n}} \left(\frac{2\pi e}{n} \right)^{n/2} \sum_{k=0}^n (-1)^k C_n^k \frac{(x \sqrt{2e/(\pi n)})^k}{(1+k/n)^{(n+k)/2}}. \end{aligned} \quad (50)$$

Если эта сумма вычислена, то функция распределения выражается через нее по формуле

$$F_n(r) = 2\pi \int_0^r x^{n-1} S_n(x) dx. \quad (51)$$

Поскольку, как было показано выше, члены (49) при больших k много меньше, чем члены при малых k , то в сумме (50) последние слагаемые можно оценить приближенно. Тогда знаменатель в сумме (50) оцениваем асимптотически, что дает для суммы оценку сверху:

$$(1+k/n)^{n/2+o(k/n)} \propto e^{k/2}.$$

В результате оценка суммы (50) принимает вид

$$S_n(x) \propto \sqrt{\frac{e}{\pi n}} \left(\frac{2\pi e}{n} \right)^{n/2} \sum_{k=0}^n (-1)^k C_n^k (x \sqrt{2/\pi n})^k = \sqrt{\frac{e}{\pi n}} \left(\frac{2\pi e}{n} \right)^{n/2} \left(1 - x \sqrt{\frac{2}{\pi n}} \right)^n. \quad (52)$$

Подставляя результат (52) в интеграл (51), получаем

$$\begin{aligned} F_n(r) &\propto \sqrt{2} \left(\frac{2\pi e}{n} \right)^{(n+1)/2} \int_0^r x^{n-1} \left(1 - x \sqrt{\frac{2}{\pi n}} \right)^n dx = \\ &= \sqrt{2} \left(\frac{2\pi e}{n} \right)^{(n+1)/2} \left(\sqrt{\frac{\pi n}{2}} \right)^{n/2} \int_0^{r\sqrt{2/\pi n}} t^{n-1} (1-t)^n dt. \end{aligned}$$

Поскольку $n \gg 1$, то главное значение интеграла в последнем выражении оценивается значением подынтегрального выражения на верхнем пределе, умноженным на малый промежуток интегрирования:

$$\begin{aligned}
F_n(r) &\propto \sqrt{2} \left(\frac{2\pi e}{n} \right)^{(n+1)/2} r^n \left(1 - r \sqrt{\frac{2}{\pi n}} \right)^n = \\
&= \sqrt{2} \left(\frac{2\pi e}{n} \right)^{(n+1)/2} r^n \exp \left(n \ln \left(1 - r \sqrt{\frac{2}{\pi n}} \right) \right) \propto 2\pi e^{-r^2/\pi} V_n r^n e^{-r\sqrt{2n/\pi} + o(1/\sqrt{n})}.
\end{aligned} \tag{53}$$

Следовательно, вероятность того, что расстояние между двумя случайными точками в n -мерном единичном кубе меньше единицы, убывает в соответствии с оценкой

$$F_n(1) \propto \sqrt{2} \left(\frac{2\pi e}{n} \right)^{(n+1)/2} e^{-1/\pi - \sqrt{2n/\pi}} = 2\pi e^{-1/\pi} V_n e^{-\sqrt{2n/\pi}}. \tag{54}$$

При $n \gg 1$ эта оценка существенно меньше, чем V_n , за счет отрицательного члена $-\sqrt{2n/\pi}$ в экспоненте, так что проведенный уточняющий анализ был необходим.

4. Численное моделирование плотности распределения расстояний в гиперкубе при больших размерностях

Рассмотрим распределение расстояний между точками с равномерно распределенными случайными координатами в кубе $[0;1]^n$ при $n \gg 1$ с помощью вычислительного эксперимента. На рис. 6 показаны плотности распределения расстояний для различных размерностей по результатам генерации 10 тыс. случайных векторов.

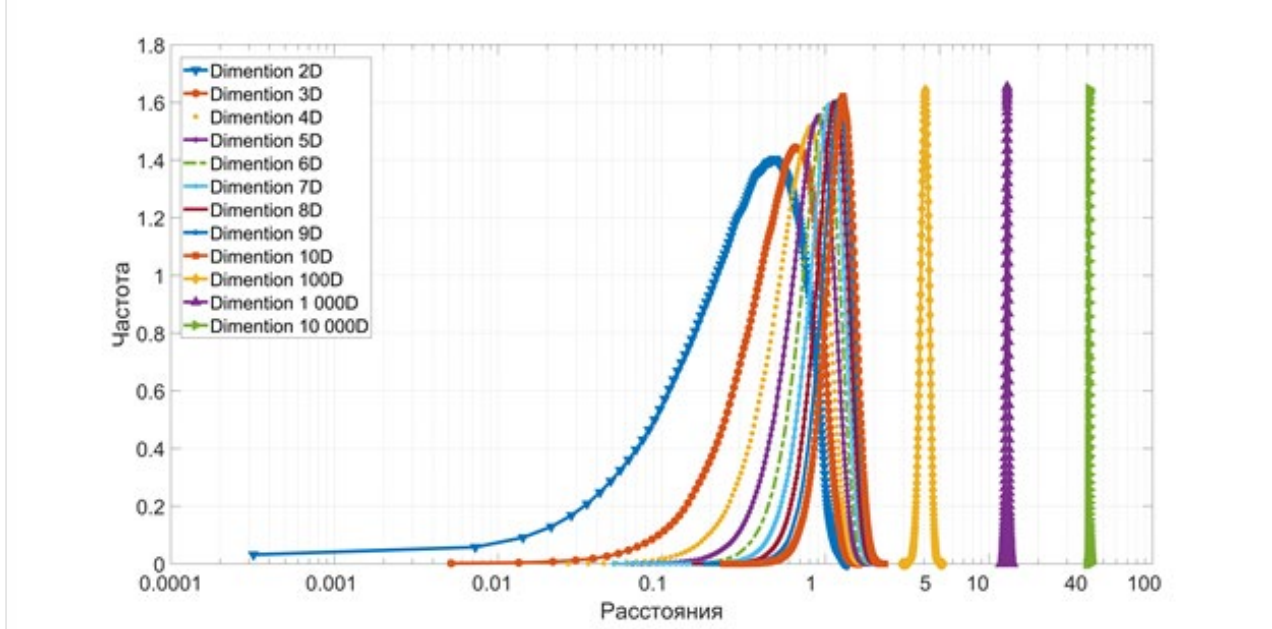


Рис. 6. Распределение расстояний в единичном n -мерном кубе

В частности, для размерностей $n = 100$ и $n = 10000$ распределения показаны ниже на рис. 7-8.

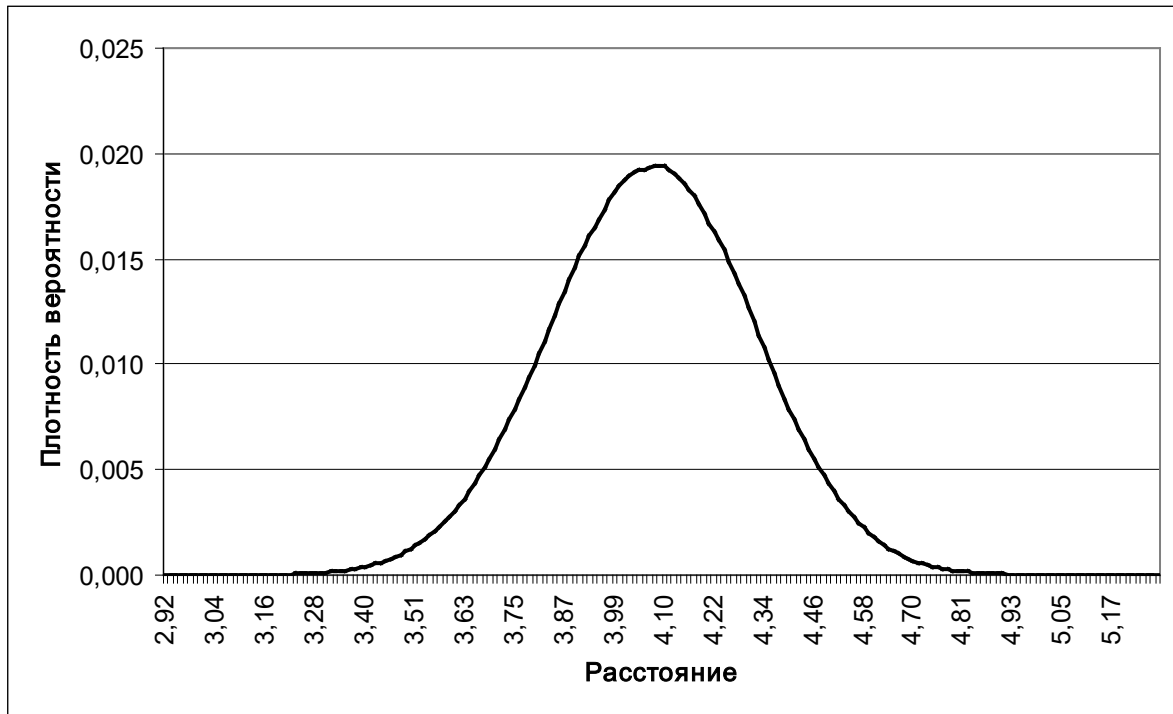


Рис. 7. Распределение расстояний в 100-мерном единичном кубе

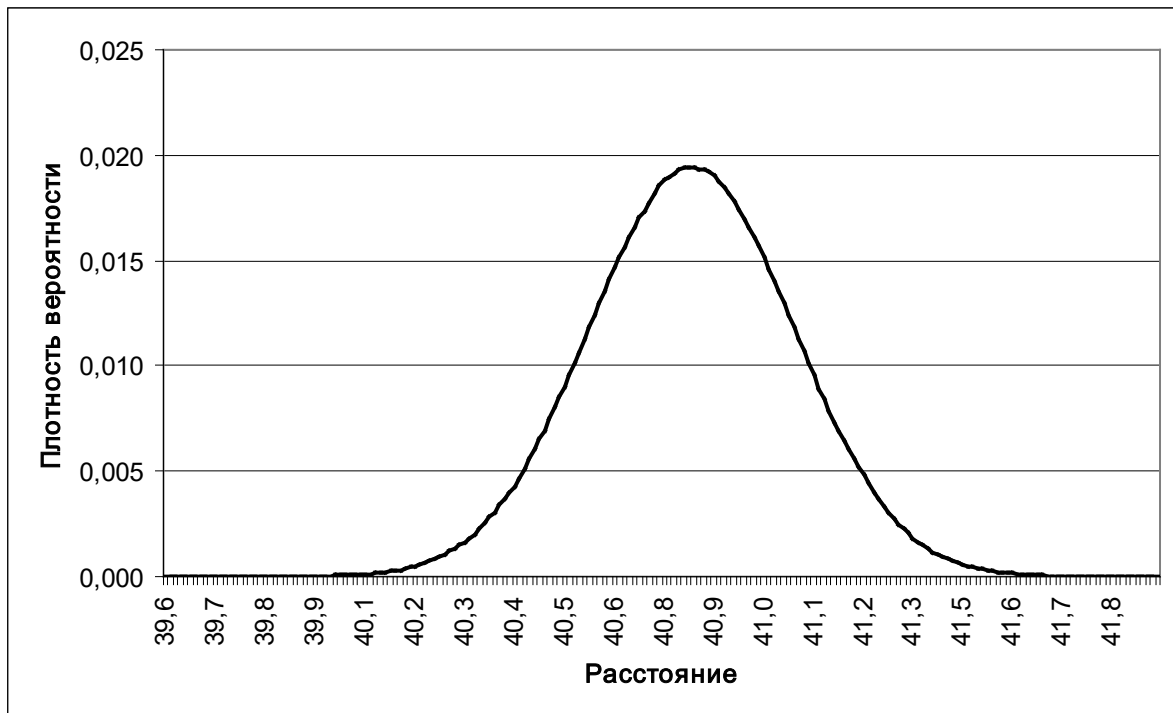


Рис. 8. Распределение расстояний в 10 000 -мерном единичном кубе

Видно, что распределения подобны с коэффициентом растяжения носителя в $\sqrt{n}$ раз. Аппроксимация этих распределений при больших размерностях показывает, что плотности нормальны на уровне детерминации более чем 0,99. Среднее значение приблизительно находится на уровне $0,408\sqrt{n}$. Плотность

распределения расстояний, нормированных на $\sqrt{n}$, имеет аппроксимацию с детерминацией на уровне 0,99:

$$\rho_n(x) \propto e^{-0,111n(x-0,408)^2}, \quad 0 \leq x = r/\sqrt{n} \leq 1. \quad (55)$$

Таким образом, численный эксперимент показывает, что плотность распределения нормированных расстояний с увеличением размерности стремится к дельта-функции от некоторого среднего значения. Это накладывает ограничения на точность нахождения ближайших соседей в многомерных пространствах. Теоретически при любой конечной размерности проблемы не возникает, так как упорядоченность расстояний в конкретной выборке не зависит от их функции распределения. Однако на практике в силу конечности регистра нет возможности рассматривать распределения с неограниченным носителем, что приводит к необходимости кластеризации расстояний и следующей из этого невозможности различать соседние точки, даже если они в определенном смысле далеки. Количество же ложных ближайших соседей будет возрастать с увеличением размерности, что и является главным препятствием к распознаванию образов в пространствах большой размерности.

5. О предельном распределении расстояний между точками

Обсудим некоторые соображения, обосновывающие аппроксимацию (55). Рассмотрим точно решаемую задачу о распределении расстояний между случайными точками в R^n , если каждая координата является нормальной случайной величиной с нулевым средним и дисперсией σ_0^2 . Итак, положим $f(x)$ в (6) равной

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{x^2}{2\sigma_0^2}\right). \quad (56)$$

Свертка (6) легко вычисляется:

$$h(t) = \frac{1}{2\pi\sigma_0^2} \int_{-\infty}^{+\infty} \exp\left(-\frac{y^2 + (t+y)^2}{2\sigma_0^2}\right) dy = \frac{1}{2\sigma_0\sqrt{\pi}} \exp\left(-\frac{t^2}{4\sigma_0^2}\right).$$

Тогда плотность распределения расстояний в R^1 равна

$$\rho_1(r) = h(t \leq 0) + h(t > 0) = \frac{1}{\sigma_0\sqrt{\pi}} \exp\left(-\frac{r^2}{4\sigma_0^2}\right), \quad r \geq 0. \quad (57)$$

Соответствующая плотность распределения квадрата расстояния равна

$$\chi_1(z) = \frac{\rho_1(\sqrt{z})}{2\sqrt{z}} = \frac{1}{2\sigma_0\sqrt{\pi}} \frac{1}{\sqrt{z}} \exp\left(-\frac{z}{4\sigma_0^2}\right), \quad z \geq 0. \quad (58)$$

Далее в соответствии с (15) и с учетом условия, что плотность (58) равна нулю при отрицательных значениях аргумента, имеем для распределения квадрата расстояний в R^2 :

$$\chi_2(w) = \int_0^{+\infty} \chi_1(w-s) \chi_1(s) ds = \int_0^w \chi_1(w-s) \chi_1(s) ds = \frac{1}{4\pi\sigma_0^2} \exp\left(-\frac{w}{4\sigma_0^2}\right) \int_0^w \frac{ds}{\sqrt{s(w-s)}}.$$

Последний интеграл нам уже встречался ранее в (16), он равен π . В итоге получаем:

$$\chi_2(w) = \frac{1}{4\sigma_0^2} \exp\left(-\frac{w}{4\sigma_0^2}\right). \quad (59)$$

Действуя тем же рекуррентным путем, что и при выводе (26), (28), строим для плотности квадрата расстояний в R^n зависимость вида

$$\chi_n(w) = C_n w^{(n-2)/2} \exp\left(-\frac{w}{4\sigma_0^2}\right),$$

где C_n – коэффициенты, подлежащие определению.

Тогда из уравнения, аналогичного (25), следует

$$\chi_{n+1}(w) = C_{n+1} w^{(n-1)/2} \exp\left(-\frac{w}{4\sigma_0^2}\right) = \frac{C_n}{2\sigma_0\sqrt{\pi}} \exp\left(-\frac{w}{4\sigma_0^2}\right) \int_0^w \frac{s^{(n-2)/2} ds}{\sqrt{w-s}}.$$

Последний интеграл встречался выше в (29), он равен $\frac{J_{n-2}}{2}$. В результате

для определения коэффициентов C_n имеем рекурсию:

$$C_{n+1} = \frac{C_n}{2\sigma_0\sqrt{\pi}} \frac{J_{n-2}}{2} = \dots = \frac{C_2}{(2\sigma_0\sqrt{\pi})^{n-1}} J_0 \dots J_{n-2} = \frac{1}{4\sigma_0^2} \frac{V_{n-1}}{(2\sigma_0\sqrt{\pi})^{n-1}}.$$

Последнее равенство получено с учетом (59) и (37). Итак,

$$\chi_n(w) = \frac{1}{4\sigma_0^2} \frac{1}{(2\sigma_0)^{n-2} \Gamma(n/2)} w^{(n-2)/2} \exp\left(-\frac{w}{4\sigma_0^2}\right). \quad (60)$$

Соответствующая плотность распределения собственно расстояний равна

$$\rho_n(r) = \frac{1}{4\sigma_0^2} \frac{2}{(2\sigma_0)^{n-2} \Gamma(n/2)} r^{n-1} \exp\left(-\frac{r^2}{4\sigma_0^2}\right). \quad (61)$$

Это распределение носит название χ -распределения с n степенями свободы (см. [16], стр. 126).

Получим слабый предел распределения (61) при $n \rightarrow \infty$ как предел моментов. С учетом интегрального представления гамма-функции в виде

$$\Gamma(z) = \int_0^{+\infty} e^{-t} t^{z-1} dt$$

имеем для момента второго порядка:

$$\langle r^2 \rangle_n = \int_0^{+\infty} w \chi_n(w) dw = 2n\sigma_0^2. \quad (62)$$

Из (62) следует, что распределение расстояний, нормированных на $\sqrt{n}$, при $n \rightarrow \infty$ будет иметь конечный второй момент, равный

$$\tilde{\mu}_{2,n} = 2\sigma_0^2 \rightarrow 2\sigma_0^2, \quad n \rightarrow \infty. \quad (63)$$

Здесь и далее тильдой обозначены функции от нормированных расстояний. Рассмотрим теперь первый момент распределения (61) для таких же нормированных координат. Имеем с использованием формулы Стирлинга (47):

$$\begin{aligned} \tilde{\mu}_{1,n} &= \left\langle \frac{r}{\sqrt{n}} \right\rangle_n = \frac{1}{\sqrt{n}} \int_0^{+\infty} \sqrt{w} \chi_n(w) dw = \frac{2\sigma_0}{\sqrt{n}} \frac{\Gamma((n+1)/2)}{\Gamma(n/2)} \propto \\ &\propto \frac{2\sigma_0}{\sqrt{n}} \left(\frac{n+1}{2e} \right)^{(n+1)/2} \left(\frac{2e}{n} \right)^{n/2} \sqrt{\frac{n}{n+1}} = \sqrt{\frac{2}{e}} \sigma_0 \left(\frac{n+1}{n} \right)^{n/2} \rightarrow \sqrt{2} \sigma_0, \quad n \rightarrow \infty. \end{aligned} \quad (64)$$

Сравнивая (64) с (63), замечаем, что в пределе квадрат среднего значения расстояний равен среднему квадрату, т.е. $\lim_{n \rightarrow \infty} (\tilde{\mu}_{1,n}^2 - \tilde{\mu}_{2,n}) = 0$. Поэтому дисперсия предельного распределения (61) равна нулю. Это означает, что предельное распределение нормированных расстояний имеет плотность

$$\tilde{\rho}_\infty(t) = \delta(t - \sigma_0 \sqrt{2}), \quad t = r / \sqrt{n}. \quad (65)$$

Вернемся теперь к эмпирическим оценкам распределения нормированных расстояний в гиперкубе, когда координаты равномерно распределены на отрезке $[0;1]$. Дисперсия равномерного распределения равна, как известно, величине $\sigma_0^2 = 1/12$, так что $\sigma_0 \sqrt{2} = 1/\sqrt{6} \approx 0,408$, что как раз и отвечает модели (55). Наблюдаемый эффект, выражаемый формулой (63), является достаточно общим. Рассмотрим его более детально.

Пусть имеется набор из n независимых одинаково распределенных одномерных случайных величин (в данном случае наших координат) с плотностью $f(x)$, имеющей конечные первый момент μ_0 и дисперсию σ_0^2 . Рассмотрим выборку из N таких наборов, т.е. N векторов с координатами $(x_{1,i}, \dots, x_{n,i})$, $i=1, \dots, N$ из данного распределения. Пусть расстояние между векторами вычисляется в норме L_2 . Тогда для квадрата расстояния имеем:

$$r_{ij}^2 = \sum_{k=1}^n (x_{k,i} - x_{k,j})^2.$$

При усреднении этой величины по распределению $f(x_1) \dots f(x_n)$ воспользуемся теоремой Гливенко о том, что эмпирическая функция распределения $F_N(x)$ стационарной случайной величины равномерно по x сходится по вероятности к соответствующему распределению $F(x)$ генеральной совокупности при $N \rightarrow \infty$, т.е.

$$P \left\{ \lim_{N \rightarrow \infty} \sup_x |F_N(x) - F(x)| = 0 \right\} = 1.$$

Тогда усреднение эмпирических выражений типа функций от расстояния между двумя точками в R^n трактуется при определенных условиях как усреднение по генеральной совокупности. В соответствии с теоремой Леви–Линдберга, обобщающей центральную предельную теорему Муавра–

Лапласа, если $\{\xi_N\}$ – последовательность взаимно независимых и одинаково распределенных случайных величин, имеющих конечные математическое ожидание μ и дисперсию σ^2 , то последовательность функций распределения нормированных сумм $z_N = \frac{1}{\sigma\sqrt{N}}\left(\sum_{k=1}^N \xi_k - N\mu\right)$ сходится равномерно на R к нормальному стандартизованному распределению.

То же относится и к выборочным моментам $m_k(N)$ порядка k по выборке объема N :

$$m_k(N) = \langle x^k \rangle_N = \frac{1}{N} \sum_{i=1}^N (x_i)^k \rightarrow \mu_k.$$

Моменты $m_k(N)$ являются несмещенными состоятельными оценками соответствующих моментов μ_k теоретического распределения, если последние существуют. Распределение величин $m_k(N)$ при $N \rightarrow \infty$ асимптотически нормально с параметрами

$$Mm_k(N) = \frac{1}{N} \sum_{i=1}^N M(x_i)^k = \frac{1}{N} \sum_{i=1}^N \mu_k = \mu_k, \quad Dm_k(N) = M(m_k(N) - \mu_k^2)^2 = \frac{\mu_{2k} - \mu_k^2}{N}.$$

В частности, если μ и σ^2 есть генеральные среднее и дисперсия, то

$$Mm_1(N) = \mu, \quad Dm_1(N) = \sigma^2 / N.$$

Тогда при усреднении эмпирического расстояния с учетом того, что по каждой координате распределения совпадают, получаем

$$\langle r_{ij}^2 \rangle = \left\langle \sum_{k=1}^n ((x_{k,i} - \mu_0) - (x_{k,j} - \mu_0))^2 \right\rangle = \left\langle \sum_{k=1}^n ((x_{k,i} - \mu_0)^2 + (x_{k,j} - \mu_0)^2) \right\rangle = 2n\sigma_0^2.$$

Следовательно, нормированные на $\sqrt{n}$ расстояния имеют второй момент, равный $2\sigma_0^2$, как и указано в (63).

Далее, согласно (25), между величинами $\chi_n(w)$ имеет место рекуррентное соотношение в виде свертки с заданным ядром, определяемым плотностью функции распределения расстояний между точками в R^1 :

$$\chi_{n+1}(w) = \int_0^w \chi_n(t) \chi_1(w-t) dt.$$

Последовательность функций $\chi_n(w)$ такова, что все эти функции неотрицательны и нормированы на единицу. Однако предельная функция χ_∞ при $n \rightarrow \infty$ не принадлежит пространству, в котором были заданы элементы исходной функциональной последовательности. В пределе для χ_∞ получается интегральное уравнение Абеля II рода:

$$\chi_\infty(w) = \int_0^w \chi_\infty(t) \chi_1(w-t) dt,$$

которое в классе ограниченных функций не имеет иных решений, кроме нулевого (см. [17], стр. 37). Но все нормированные моменты существуют и конечны при каждом конечном значении размерности n . Поэтому естественно ожидать, что χ_∞ будет представляться сингулярной обобщенной функцией. Рассмотрим нормированное расстояние между некоторой случайной парой точек в R^n :

$$\tilde{r}_{ij} = \frac{r_{ij}}{\sqrt{n}} = \sqrt{\frac{1}{n} \sum_{k=1}^n (x_{i,k} - x_{j,k})^2}.$$

При $n \rightarrow \infty$ сумма под корнем есть просто среднее значение квадрата разности двух одинаково распределенных случайных величин, т.е. $2\sigma_0^2$. Поэтому $\lim_{n \rightarrow \infty} \tilde{r}_{ij} = \langle \tilde{r} \rangle = \sigma_0 \sqrt{2}$, т.е. квадрат момента первого порядка совпадает с моментом второго порядка. Таким образом, дисперсия предельного распределения равна нулю, что соответствует постоянной случайной величине, чему отвечает распределение в виде дельта-функции.

Заключение

В данной работе представлены новые результаты, касающиеся вероятности распределения расстояний между случайными точками в R^n и, в частности, в гиперкубе.

Получена точная формула для функции распределения расстояний в гиперкубе в области $r \leq 1$. Сложность рассмотрения вероятности при $r > 1$ обусловлена техническими трудностями при вычислении соответствующего объема пересечения двух n -мерных тел (куба и шара). Эта задача приводит к очень громоздким формулам, примеры которых приведены для размерностей 2 и 3. Поэтому важным представляется нахождение предельного по размерности распределения расстояний между точками. Это предельное распределение – второй основной результат работы. Показано, что для n одинаково распределенных случайных величин (координат) с плотностью распределения $f(x)$, которая имеет конечные математическое ожидание μ и дисперсию σ^2 , распределение нормированных на $\sqrt{n}$ расстояний в норме L_2 стремится к дельта-функции $\delta(\tilde{r} - \sigma\sqrt{2})$.

На основе доказанного предельного свойства объяснены результаты численного моделирования распределения расстояний в гиперкубе большой размерности. Поскольку равномерное распределение на отрезке $[0; 1]$ имеет дисперсию $\sigma^2 = 1/12$, то наблюдаемое в вычислительном эксперименте асимптотическое распределение расстояний в гиперкубе при больших размерностях становится все более нормальным и в пределе стремится к дельта-функции $\delta\left(\tilde{r} - \frac{1}{\sqrt{6}}\right)$.

Также в работе получено уточнение асимптотики функции распределения расстояний в гиперкубе в области $r < 1$ при больших n . Показано, что эта вероятность спадает быстрее, чем качественная ее оценка через объем n -мерного шара. Найденный уточняющий фактор имеет оценку сверху в виде $\exp(-\sqrt{2n/\pi})$.

Перечисленные результаты существенны при разработке алгоритмов поиска ближайших соседей в пространствах больших размерностей, что является одной из весьма актуальных практических задач, связанных с распознаванием образов с помощью технологий машинного обучения и искусственного интеллекта. В частности, проведенный асимптотический анализ позволяет оценить соотношение между числом N случайных векторов и размерностью n пространства, в котором они заданы, чтобы находить ближайшего соседа на расстоянии не более чем r на уровне значимости α . Согласно (2), для этого надо решить уравнение

$$F_n(r) = \frac{\alpha}{N},$$

где $F_n(r)$ есть функция распределения расстояний между точками в данном множестве из R^n . В настоящей работе получены две оценки для функции распределения расстояний – при $r \leq 1$ для любых n и при $n \rightarrow \infty$ для любых r . Первую оценку по формуле (54) удобно использовать при относительно небольших размерностях, но которые все же много больше единицы – например, при $n \propto 10 \div 100$. Тогда нахождение соседней точки на расстоянии порядка длины ребра куба имеет практический смысл. При $n > 100$ для определения вероятности нахождения ближайшего соседа более эффективна асимптотика (65). Получим соответствующую оценку.

Из проведенного анализа для гиперкуба следует, что при больших n предельная плотность распределения (65) может быть аппроксимирована нормальным распределением вида (55)

$$\rho_n(\tilde{r}) \propto \exp\left(-\lambda n(\tilde{r} - \sigma\sqrt{2})^2\right),$$

где λ – это дисперсия одномерного распределения расстояний, а σ^2 – дисперсия одномерного распределения координат (не расстояний!). Нормированное на единицу выражение для $\rho_n(\tilde{r})$ (с учетом того, что $\tilde{r} \geq 0$) имеет вид

$$\rho_n(\tilde{r}) = \sqrt{\frac{\lambda n}{\pi}} \frac{2}{1 + \operatorname{erf}\left(\sqrt{2\lambda n}\sigma\right)} \exp\left(-\lambda n(\tilde{r} - \sigma\sqrt{2})^2\right).$$

Здесь введен так называемый интеграл вероятностей $\operatorname{erf}(z)$ согласно определению:

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-x^2} dx.$$

Обозначим для краткости нормировочную постоянную через Z_n :

$$\frac{1}{Z_n} = \sqrt{\frac{\lambda n}{\pi}} \frac{2}{1 + \operatorname{erf}\left(\sqrt{2\lambda n}\sigma\right)}.$$

Тогда при $\tilde{r} < \sigma\sqrt{2}$ получаем:

$$\begin{aligned} F_n(\tilde{r}) &= \frac{1}{Z_n} \int_0^{\tilde{r}} \exp\left(-\lambda n(x - \sigma\sqrt{2})^2\right) dx = \frac{1}{Z_n} \int_{-\sigma\sqrt{2}}^{\tilde{r}-\sigma\sqrt{2}} \exp(-\lambda n t^2) dt = \\ &= \frac{1}{Z_n \sqrt{\lambda n}} \left(\int_0^{\sigma\sqrt{2\lambda n}} - \int_0^{\sigma\sqrt{2\lambda n} - \tilde{r}\sqrt{\lambda n}} \right) e^{-x^2} dx = \frac{\operatorname{erf}\left(\sqrt{2\lambda n}\sigma\right) - \operatorname{erf}\left(\sqrt{2\lambda n}\sigma^2 - \tilde{r}\sqrt{\lambda n}\right)}{1 + \operatorname{erf}\left(\sqrt{2\lambda n}\sigma\right)}. \end{aligned}$$

Для приближенного решения уравнения $F_n(r) = \frac{\alpha}{N}$ воспользуемся аппроксимацией интеграла вероятностей при большом значении аргумента (см. [18], стр. 122):

$$\operatorname{erf}(x) \approx 1 - \frac{ae^{-x^2}}{1 + bx}, \quad a = 0,34802, \quad b = 0,47047.$$

Отсюда и из оценок коэффициентов в (55) находим, что расстояние, на котором имеет смысл искать ближайшего соседа в гиперкубе, оценивается величиной

$$r(n, N; \alpha) \approx 1,2\alpha \frac{\sqrt{n}}{N} e^{0,02n}.$$

Поскольку максимальное расстояние между точками в гиперкубе равно $\sqrt{n}$, то число точек N должно возрасть с увеличением размерности, во всяком случае по экспоненте, т.е. $N \propto e^{0,02n}$. В противном случае возможность различения ближайших соседей отсутствует.

Разделы 1 и 2 написаны Д.Т. Таммом, раздел 3 написан Д.Т. Таммом А.А. Кислицыным, раздел 4 написан А.А. Кислицыным, остальные разделы написаны совместно.

Благодарности

Работа выполнена при поддержке гранта РНФ, проект № 25-71-30001.

Литература

1. Кендалл М., Моран П. Геометрические вероятности (пер. с англ.). — М.: Наука, 1972. — 192 с.
2. Соболев И.М., Левитан Ю.Л. Получение точек, равномерно расположенных в многомерном кубе // Препринты ИПМ им. М.В. Келдыша. 1976. № 40.

3. Anderssen R., Brent R., Daley D., and Moran P. Concerning $\int_0^1 \dots \int_0^1 (x_1^2 + \dots + x_n^2)^{1/2} dx_1 \dots dx_n$ and a Taylor series method // SIAM J. Applied Math., 30, 1, (1976), 22-30.
4. Robbins D.P. Average distance between two points in a box // Amer. Mathematical Monthly, 85 (1978), 278.
5. Robbins D.P. Problem E2629: Find the Average Distance between Two Points in a Box // Amer. Math. Monthly, 1977, 57.
6. Bolis T.S. Solution of problem E2629: Average Distance between Two Points in a Box // Amer. Math. Monthly Vol. 85, No. 4. pp. 277-78, April 1978.
7. Bailey D.H., Borwein J.M., Crandall R.E. Box integrals // Journal of Computational and Applied Mathematics Vol. 206, Issue 1, Sept. 2007, pp. 196-208.
8. Burgstaller B., Pillichshammer F. The average distance between two points // Bull. Aust. Math. Soc. 80 (2009), 353–359. doi:10.1017/S0004972709000707
9. Kumar R.R., Viswanath P., Bindu C.S. Nearest neighbor classifiers: a review // Int. J. Comput. Intell. Res. 2017. Vol. 13, no. 2. P. 303-311.
10. 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.
11. Muja M., Lowe D.G. Scalable nearest neighbor algorithms for high dimensional data. IEEE transactions on pattern analysis and machine intelligence, 2014, 36(11):2227–2240.
12. Malkov Y.A., Yashunin D.A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs // IEEE transactions on pattern analysis and machine intelligence. 2018. V. 42. №. 4. P. 824-836.
13. Malkov Y., Ponomarenko A., Logvinov A., and Krylov V. Approximate nearest neighbor algorithm based on navigable small world graphs // Information Systems, 45:61–68, 2014.
14. Горячкин Б.С., Павловская А.А., Григорьев Ю.А. Сравнительный анализ алгоритмов поиска ближайших соседей с использованием графов // Информатика, вычислительная техника и управление, 2024. № 5. С. 55-65. DOI 10.37882/2223-2966.2024.05.09
15. Гмурман В.Е. Теория вероятностей и математическая статистика. М.: Высшая школа, 1972. – 368 с.
16. Корольок В.С., Портенко Н.И., Скороход А.В., Турбин А.Ф. Справочник по теории вероятностей и математической статистике. – М.: Наука, 1985. – 640 с.
17. Мюнтц Г. Интегральные уравнения. Т. I. – М. – Л., ОНТИ, 1934. – 330 с.
18. Абрамовиц М., Стиган И. Справочник по специальным функциям (пер. с англ.). – М.: Наука, 1979. – 832 с.