А. Гасников, Е. Черноусова, Т. Нагапетян, О. Федько, Стохастический анализ в задачах, *Матем. просв.*, 2012, выпуск 16, 181–213

Преподавание математики

Стохастический анализ в задачах

А. Гасников Е. Черноусова Т. Нагапетян О. Федько

Курсы «Теория вероятностей», «Случайные процессы», «Математическая статистика» читаются студентам ФУПМ МФТИ вот уже более 30 лет. Во многом эти курсы сформировались под влиянием и при самом непосредственном участии профессора Андрея Александровича Натана (06.02.1918 – 09.01.2009) [40].

На данный момент уже выпущены пособия по материалам прочитанных курсов, написанные А. А. Натаном в соавторстве с учениками (ныне лекторами по этим дисциплинам) доцентами С. А. Гузом и О. Г. Горбачевым.

Натан А.А., Горбачев О.Г., Гуз С.А. Теория вероятностей: Учебное пособие. — М.: МЗ Пресс — МФТИ, 2007. — 253 с.

Натан А.А., Горбачев О.Г., Гуз С.А. Основы теории случайных процессов: Учебное пособие. — М.: МЗ Пресс — М.: МФТИ, 2003. — 165 с.

Натан А.А., Горбачев О.Г., Гуз С.А. Математическая статистика: Учебное пособие. — М.: МЗ Пресс — М.: МФТИ, 2004. — 156 с.

Пособия получили положительные отзывы специалистов. Однако, как отмечал Андрей Александрович, пособия представляют собой лишь запись (конспект) курса лекций, и содержат мало разобранных примеров. Ввиду вышесказанного кафедра Математических основ управления ФУПМ, которую долгие годы возглавлял Андрей Александрович, решила подготовить настоящую статью, в основу которой положены задачи (в том числе повышенной сложности), предлагавшиеся в разные годы, в основном студентам ФУПМ, на различных семинарах, сдачах заданий и экзаменах. Главными отличительными особенностями являются: а) широкий спектр представленного материала, б) отражение ряда современных течений и в) нацеленность на приложения. При работе над задачами в 2010 г. было решено организовать учебно-методический кафедральный семинар

«Стохастический анализ в задачах» [38]. На этом семинаре обсуждаются важные вопросы, как правило, не входящие в обязательную программу цикла стохастических дисциплин.

Отметим несколько важных современных течений, нашедших отражения в виде докладов на этом семинаре и в виде задач: понятие равновесия макросистемы = эргодическая теорема + явление концентрации меры (Пуанкаре – Леви – Мильман – Громов – Талагран), стохастические транспортные (компьютерные) сети при термодинамическом предельном переходе, вероятностные методы в комбинаторике и теории чисел, вероятностные и рандомизированные алгоритмы.

Задачи специально подбирались таким образом, чтобы с одной стороны часть из них была доступна продвинутым школьникам старших классов, а с другой стороны — отражали основной инструментарий и наиболее важные современные приложения вероятностных методов. Некоторые задачи трудны, и их решение предполагает серьёзное самостоятельное исследование (такие задачи помечены звёздочкой). Причём понимание условий ряда таких задач, в свою очередь, требует значительных усилий. Наконец, имеется пара задач, полные решения которых авторам не известны.

От читателей предполагается знакомство с азами теории вероятностей, случайных процессов и математической статистики, например, в объёме замечательной книги [12]. Рекомендуем также ознакомиться с материалами [24, 36, 44], содержащими много интересных вероятностных задач.

Авторы благодарны О. Г. Горбачеву и С. А. Гузу за ряд ценных замечаний, а также М. Н. Вялому и А. Х. Шеню, способствовавшим улучшению первоначальной версии текста. Работа выполнена при поддержке Лаборатории структурных методов анализа данных в предсказательном моделировании, МФТИ, грант правительства РФ дог. 11 11.G34.31.0073.

Неравенство Чебышёва и закон больших чисел

Задача (о выборах). В некотором городе прошел второй тур выборов. Выбор был между двумя кандидатами A и B (графы «против всех» на этих выборах не было). Сколько человек надо опросить на выходе с избирательных участков, чтобы, исходя из ответов, можно было определить долю проголосовавших за кандидата A с точностью 0.01 и с вероятностью не меньшей 0.95.

Задача (о количестве экзаменационных билетов). В некотором вузе проходит экзамен. Количество экзаменационных билетов $N \gg 1$. Перед экзаменационной аудиторией выстроилась очередь из студентов, которые не знают чему равно N. Согласно этой очереди студенты вызываются на экзамен (второй студент заходит в аудиторию, после того как из нее выйдет первый и т. д.). Каждый студент с равной вероятностью может выбрать любой из N билетов (в независимости от других студентов). Проэкзаменованные студенты, выходя из аудитории, сообщают оставшейся очереди номера своих билетов. Оцените (сверху) сколько студентов должно быть проэкзаменовано, чтобы оставшаяся к этому моменту очередь смогла оценить число экзаменационных билетов с точностью 5% с вероятностью не меньшей 0.95.

Задача (изогнутая игла Бюффона). Любопытный студент швейного техникума решил повторить опыты Бюффона по бросанию иглы (студент хочет оценить число π). Для этого он подготовил горизонтально расположенный лист бумаги, разлинованный параллельными прямыми так, что расстояние между соседними прямыми равно 1. Однако в распоряжении студента оказалось только погнутая иголка. Иголка имеет форму кочерги, но студент не имеет точного представления о том, как именно погнута иголка. Ему известно лишь то, что длина погнутой иголки равна 2. Студент «случайно» и независимо бросил погнутую иголку 1 000 000 раз и посчитал суммарное число пересечений, учитывая кратность. Помогите студенту оценить число π: а) с помощью неравенства Чебышёва; б*) с помощью з.б.ч. (закона больших чисел) и неравенств о вероятностях больших уклонений; в) с помощью ц.п.т. (центральной предельной теоремы) и оценок скорости сходимости в ц.п.т., например, с помощью неравенства Берри – Эссена или более точных аппроксимаций.

Задача (метод Монте-Карло). Предложите эффективный способ вычисления с заданной точностью ε и с заданной доверительной вероятностью γ интеграла:

$$J = \int\limits_{[0,\,1]^m} f\left(\vec{x}\right) d\vec{x}.$$

Считайте, что $|f\left(\vec{x}\right)| \leqslant 1$ для всех $\vec{x} \in [0,\,1]^m$.

Пояснение. Введём случайный m-вектор $\vec{X} \in R\big([0,\,1]^m\big)$ и с.в.

$$\xi = f\big(\vec{X}\big).$$

Тогда $M\xi = \int\limits_{[0,1]^m} f\left(\vec{x}\right) d\vec{x} = J$. Поэтому получаем оценку интеграла $\bar{J}_n = \dfrac{1}{n}\sum\limits_{k=1}^{n} f\left(\vec{x}^k\right)$, где $\vec{x}^k$, $k = 1,\ldots,n$ — повторная выборка значений случайного вектора $\vec{X}$ (т. е. все $\vec{x}^k$, $k = 1,\ldots,n$ — независимы и одинаково распределены, так же как и вектор $\vec{X}$). В задаче требуется оценить сверху число n ($n \gg m$), начиная с которого $P\left(\left|J - \bar{J}_n\right| \leqslant \varepsilon\right) \geqslant \gamma$.

Задача (теорема Шеннона – Макмиллан или основная теорема теории кодирования). Пусть буква X — дискретная с.в., принимающая значения из алфавита $(x_1, \ldots, x_m)$ с вероятностями $(p_1, \ldots, p_m)$. Имеется случайный текст из $n \gg 1$ букв X (предполагается, что буквы в тексте не зависимы друг от друга). Общее количество таких текстов $2^{n \log m}$. Поэтому можно закодировать все эти слова, используя $n \log m$ бит. Однако, используя то обстоятельство, что $(p_1, \ldots, p_m)$ — в общем случае неравномерное распределение, предложите лучший способ кодирования, основанный на усиленном законе больших чисел.

Указание. Пусть $\Omega = \{\omega : \omega = (X_1, X_2, \ldots, X_n), X_i \in 1, 2, \ldots, m\}$ — пространство элементарных исходов. Вероятность появления слова $\omega = (X_1, X_2, \ldots, X_n)$ равна $p(\omega) = p_{X_1} \cdot \ldots \cdot p_{X_n}$. Покажите, что по закону больших чисел (простое следствие неравенства Чебышёва):

$$-\frac{1}{n}\log p(\omega) = -\frac{1}{n}\sum_{i=1}^{n}\log p_{X_i} \xrightarrow[n\to\infty]{P} -\sum_{i=1}^{m} p_i \log p_i \stackrel{\text{def}}{=} H(\vec{p}).$$

Текст ω будем называть *δ-типичным*, если

$$2^{-n(H(\vec{p})+\delta)} < p(\omega) < 2^{-n(H(\vec{p})-\delta)}.$$

Покажите, что

1. Существует не более $2^{n(H(\vec{p})+\delta)}$ типичных текстов.
2. Для $n > n(\varepsilon, \delta)$ существует по крайней мере $(1-\varepsilon)\, 2^{n(H(\vec{p})-\delta)}$ типичных текстов.
3. Множество нетипичных текстов имеет вероятность $\leqslant \varepsilon$.

Таким образом, можно осуществить эффективное кодирование данных, используя все двоичные последовательности длины $n(H(\vec{p}) + \delta)$, чтобы закодировать все δ-типичные тексты и отбросить нетипичные. Вероятность ошибки при таком кодировании будет не больше ε. Обратно, любой код, использующий двоичные последовательности длины $n(H(\vec{p}) - \delta)$, имеет асимптотически не исчезающую вероятность ошибки, стремящуюся к единице при $n \to \infty$.

Функцию $H(\vec{p})$, которую можно проинтерпретировать как меру количества информации (в битах на передаваемый символ) в случайном тексте, называют *энтропией*. Величина $nH(\vec{p})$ характеризует меру неопределённости случайного текста. Ясно, что для $(p_1, \ldots, p_m) = (1/m, \ldots, 1/m)$ энтропия максимальна, $H(\vec{p}) = \log m$, и эффективное кодирование невозможно.

Линейность математического ожидания

Задача (о лифте). На первом этаже семнадцатиэтажного общежития в лифт вошли десять человек. Предполагая, что каждый из вошедших (независимо от остальных) может с равной вероятностью жить на любом из шестнадцати этажей (со 2-го по 17-й), найдите математическое ожидание числа остановок лифта.

Указание. Попробовать ввести случайные величины, принимающие значения 1 и 0 (то есть индикатор некоторого события), так чтобы вопрос задачи сводился к нахождению математического ожидания суммы эти случайных величин.

Задача (о посадке космического корабля). Поверхность некоторой шарообразной планеты состоит из океана и суши (множество мелких островков). Суша (измеримое по Лебегу множество) занимает больше половины площади планеты. На планету хочет совершить посадку космический корабль, сконструированный так, что концы всех шести его ножек лежат на поверхности планеты. Посадка окажется успешной, если не меньше четырёх ножек из шести окажутся на суше. Возможна ли успешная посадка корабля на планету?

Указание. Обратим внимание на то, что ответ не зависит от того, какое именно множество представляет собой суша, и как располагаются шесть ножек корабля, на которые он совершает посадку. Эта задача, на первый взгляд, не имеет ничего общего с теорией вероятностей. Однако метод её решения базируется на введении вероятностных объектов и использовании двух простых фактов: 1) линейности математического ожидания (независимость слагаемых не нужна), 2) если математическое ожидание с.в. больше какого-то числа, то существует исход (точнее говоря, исходы, вероятностная мера которых больше нуля) такой, что с.в. принимает на этом исходе значение больше упомянутого числа).

Задача (парадокс бросания симметричной монетки). Симметричную монету независимо бросили n раз. Результат бросания записали в виде последовательности нулей и единиц. Покажите, что с вероятностью, стремящейся к единице при $n \to \infty$, длина максимальной подпоследовательности из подряд идущих единиц лежит в промежутке

$$\left(\log \sqrt{n},\, \log n^2\right).$$

Парадоксы

Задача (о теннисных матчах). Юноша собирается сыграть три теннисных матча со своими родителями, и он должен победить два раза подряд. Порядок матчей может быть следующим отец-мать-отец, мать-отец-мать. Юноше нужно решить, какой порядок для него предпочтительней, учитывая, что отец играет лучше матери.

Задача (парадокс транзитивности). Будем говорить, что случайная величина X больше по вероятности случайной величины Y, если $P\left(X > Y\right) > P\left(X \leqslant Y\right)$. Пусть известно, что для случайных величин X, Y, Z, Wвыполнена следующая цепочка равенств:

$$P\left(X > Y\right) = P\left(Y > Z\right) = P\left(Z > W\right) = \alpha > 1/2.$$

Верно ли, что X больше по вероятности W и почему?

Задача (парадокс С. Банаха). В двух спичечных коробках имеется по n спичек. На каждом шаге наугад выбирается коробок, и из него удаляется (используется) одна спичка. Найти вероятность того, что в момент, когда один из коробков опустеет, в другом останется k спичек.

Задача (игра У. Пенни, 1969). Алиса и Билл играют в игру: они бросают монету до тех пор, пока не встретится РРО или РОО (Р — решка, О — орёл). Если первой появится последовательность РРО, выигрывает Алиса, если РОО — Билл. Покажите, что игра не будет честной. Алиса будет выигрывать примерно вдвое чаще Билла!

Задача «сто заключённых». В коридоре находятся 100 человек, у каждого свой номер (от 1 до 100). Их по одному заводят в комнату, в которой находится комод со 100 выдвижными ящиками. В ящики случайным образом разложены карточки с номерами (от 1 до 100). Каждому разрешается заглянуть в не более чем 50 ящиков. Цель каждого — определить, в каком ящике находится его номер. Общаться и передавать друг другу информацию запрещается. Предложите стратегию, которая с вероятностью не меньшей 0.3 (в предположении, что все 100! способов распределения карточек по ящикам равновероятны) приведёт к выигрышу всей команды. Команда выигрывает, если все 100 участников верно определили ящик с карточкой своего номера.

Стратегия. Каждый человек первым открывает ящик под его номером, вторым — под номером, который указан на карточке, лежащей в ящике, открытом перед этим и т. д. Среднее число циклов длины r в случайной перестановке есть $1/r$ (покажите, используя, например, задачу «про предельные меры»). Тогда среднее число циклов длины большей $n/2$ есть $\sum\limits_{k=n/2}^{n} 1/k$. Это и есть вероятность существования цикла длины большей $n/2$. Поэтому вероятность успеха команды есть $1 - \sum\limits_{k=51}^{100} 1/k \approx 0.31$. Если же просто произвольно открывать ящики, то вероятность успеха будет $2^{-100} \approx 8 \cdot 10^{-31}$. В случае, когда карточки разложены не случайным образом, то следует сделать случайной нумерацию ящиков, и далее следовать старой стратегии.

С деталями можно познакомиться, например, здесь [45, p. 176].

Задача (о шляпах; Тод Эберт, 1998). А) Трёх игроков отводят в комнату, где на них надевают (случайно и независимо) белые и чёрные шляпы. Каждый видит цвет других шляп и должен написать на бумажке одно из трёх слов: «белый», «чёрный», «пас» (не советуясь с другими и не показывая им свою бумажку). Команда выигрывает, если хотя бы один из игроков назвал правильный цвет своей шляпы и ни один не назвал неправильного. Как им сговориться, чтобы увеличить шансы. Б*) Решите эту же задачу, если игроков $n = 2^m - 1$ $(m \in \mathrm{N})$.

Замечание. Как ни странно, вероятность выигрыша можно сделать достаточно большой. Скажем, для семи игроков есть стратегия, успешная в семи из восьми случаев (и связано это с так называемым кодом Хэмминга).

Задача (парадокс времени ожидания). Автобусы прибывают на остановку в соответствии с пуассоновским процессом с параметром $\lambda > 0$. Вы приходите на остановку в фиксированный момент времени (скажем, в полдень). Каково математическое ожидание времени, в течение которого вы ждёте автобуса?

* * * * * *

Большое число вероятностных парадоксов собрано в книге [29].

Условные вероятности

Задача (о самолёте). В самолёте n мест. Есть n пассажиров, выстроившихся друг за другом в очередь. Во главе очереди — «заяц». У всех кроме «зайца» есть билет, на котором указан номер посадочного места. Так как «заяц» входит первым, он случайным образом занимает некоторое место. Каждый следующий пассажир, входящий в салон самолёта, действует по такому принципу: если его место свободно, то садится на него, если занято, то занимает с равной вероятностью любое свободное. Найдите вероятность того, что последний пассажир сядет на своё место.

Указание. Попробовать решить задачу для $n = 2, 3, 4, \ldots$ и понять, что происходит для любого n.

Задача (о несимметричной монетке). 1) Имеется несимметричная монетка. Несимметричность монетки заключается в том, что либо орёл выпадает в два раза чаще решки; либо наоборот (априорно (до проведения опытов) оба варианта считаются равновероятными). Монетку бросили 10 раз. Орёл выпал 7 раз. Определите апостериорную вероятность того, что орёл выпадает в два раза чаще решки (апостериорная вероятность считается с учётом проведённых опытов (иначе говоря, это просто условная вероятность)).

2) Определите апостериорную вероятность того, что орёл выпадает <u>не менее</u>, чем в два раза чаще решки. Если несимметричность монетки заключается в том, что либо орёл выпадает <u>не менее</u>, чем в два раза чаще решки; либо наоборот (априорно оба варианта считаются равновероятными).

Указание. Условие 2) задачи можно понимать, например, следующим образом. Рассмотрим два события

$$A = \{p \in R\,[0,\, 1/3]\} \quad \text{и} \quad \bar{A} = \{p \in R\,[2/3,\, 1]\}\,,$$

где p — вероятность выпадения орла, запись $p \in R\,[0,\, 1/3]$ означает, что с.в. p имеет равномерное распределение на отрезке $[0,\, 1/3]$. По условию $P(A) = P(\bar{A}) = 1/2$. Нужно найти $P(\bar{A} \mid r_{10} = 7)$, где $r_{10} \in Bi\,(p,\, 10)$. Заметим, что

$$P(r_{10} = 7 \mid \bar{A}) = \int\limits_{2/3}^{1} P(r_{10} = 7 \mid p) P(dp \mid \bar{A}) = 3 \int\limits_{2/3}^{1} P(r_{10} = 7 \mid p)\, dp =$$

$$= 3C_{10}^{7} \int\limits_{2/3}^{1} p^7 (1-p)^3\, dp \approx 0.144.$$

Вероятностный метод

Задача (о турнире). На турнир приехало n игроков. Каждая пара игроков, согласно регламенту турнира, должна провести одну встречу (ничьих быть не может). Пусть

$$C_n^k \cdot \left(1 - 2^{-k}\right)^{n-k} < 1.$$

Докажите, что тогда игроки могли сыграть так, что для каждого множества из k игроков найдётся игрок, который побеждает их всех.

Указание. Введите на множестве всех турниров равномерную меру, т. е. считайте, что все $2^{C_n^2}$ турниров равновероятны. Введите событие A_K, состоящее в том, что не существует игрока, побеждающего всех игроков из множества K. Докажите, что $P\Bigl(\bigcup\limits_{K \subset \{1,\ldots,n\},\, |K|=k} A_K\Bigr) \leqslant C_n^k \cdot \left(1 - 2^{-k}\right)^{n-k}$.

Следовательно, $P\Bigl(\overline{\bigcup\limits_{K \subset \{1,\ldots,n\}, |K|=k} A_K}\Bigr) > 0.$

Задача (о числах Рамсея). Покажите, что можно так раскрасить в два цвета ребра полного графа с n вершинами (т. е. графа без петель, в котором любые две различные вершины соединены одним ребром), что любой его полный подграф с m вершинами, где $2C_n^m\left(\frac{1}{2}\right)^{C_m^2} < 1$, имеет рёбра разного цвета.

Задача (теорема Эрдёша – Ко – Радо). Семейство множеств Φ называется пересекающимся, если для любых двух множеств $A,B \in \Phi$ выполняется условие $A \cap B \neq \varnothing$. Пусть $n \geqslant 2k$ и семейство Φ является пересекающимся семейством k-элементных подмножеств множества $\{0, 1, 2, \ldots, n-1\}$. Покажите, что число множеств в семействе Φ удовлетворяет неравенству

$$|\Phi| \leqslant C_{n-1}^{k-1} = \frac{(n-1)!}{(k-1)!\,(n-k)!}\,.$$

Задача (о лампочках). Рассмотрим матрицу $n\times n$, составленную из лампочек, каждая из которых любо включена ($a_{ij} = 1$), либо выключена ($a_{ij} = -1$). Предположим, что для каждой строки и каждого столбца имеется переключатель, поворот которого ($x_i = -1$ для строки i и $y_j = -1$ для столбца j) переключает все лампочки в соответствующей линии: с «вкл.» на «выкл.» и с «выкл.» на «вкл.». Тогда для любой начальной конфигурации лампочек можно установить такое положение переключателей, что разность между числом включённых и выключенных лампочек будет не меньше $\left(\sqrt{2/\pi} + o\,(1)\right)n^{3/2}$.

Указание. Пусть $y_1, \ldots, y_n$ — независимые одинаково распределённые с.в. с законом распределения

$$y_j = \begin{cases} \ \ 1, & p = 1/2, \\ -1, & p = 1/2. \end{cases}$$

Введите с.в. $R_i = \sum\limits_{j=1}^{n} a_{ij}y_j$, $i = 1, \ldots, n$ и $R = \sum\limits_{j=1}^{n} |R_i|$. Покажите, что с.в. R_i, $i = 1, \ldots, n$ распределены также, как с.в. $S_n = \sum\limits_{j=1}^{n} y_j$. Покажите далее, что $E(|R_i|) = E(|S_n|) = \left(\sqrt{2\pi} + o\,(1)\right)\sqrt{n}$ (можно получить точную формулу для $E(|S_n|)$ из комбинаторных соображений, а затем воспользоваться формулой Стирлинга, однако, более простым вариантом является применение ц.п.т. к $S_n = \sum\limits_{j=1}^{n} y_j$). Далее следует выбирать с.в. $x_1, \ldots, x_n$ так, что $x_iR_i \overset{\text{п.н.}}{=} |R_i|$, $i = 1, \ldots, n$.

* * * * * *

Для дополнительного знакомства с вероятностным методом можно рекомендовать книги [1, 28, 35].

Марковские процессы

Задача (игра в орлянку). Пусть имеются два натуральных числа A и B. Есть два игрока, у которых начальные капиталы равны соответственно A и B рублей. Они играют в орлянку: если выпал орёл (с вероятностью p) второй игрок платит первому рубль, если решка — наоборот. Найдите вероятности разорения игроков. Найдите среднюю продолжительность игры.

Задача (об очереди за булочками). В буфете за булочками выстроилась очередь из $2n$ человек. Каждый человек в очереди хочет купить ровно одну булочку. Одна булочка стоит 10 рублей 50 копеек. У половины людей в очереди имеется монетки достоинством 50 копеек, а у половины не имеется. У продавца изначально нет копеечной сдачи. Какова вероятность, что продавец всегда сможет дать сдачу? Считайте, что все способы расстановки людей в очереди равновероятны.

Задача о разборчивой невесте (Гарднер – Дынкин, 1965). В аудитории находится невеста, которая хочет выбрать себе жениха. За дверью выстроилась очередь из $N \gg 1$ женихов. Относительно любых двух женихов невеста может сделать вывод, какой из них для неё предпочтительнее. Таким образом, невеста задаёт на множестве женихов отношение порядка (естественно считать, что если A предпочтительнее B, а B предпочтительнее C, то A предпочтительнее C). Предположим, что все $N!$ вариантов очередей равновероятны и невеста об этом знает (равно, как и число N). Женихи запускаются в аудиторию по очереди. Невеста видит каждого из них в первый раз! Если на каком-то женихе невеста остановится (сделает свой выбор), то оставшаяся очередь расходиться. Невеста хочет выбрать наилучшего жениха (исследуя k-го по очереди жениха, невеста лишь может сравнить его со всеми предыдущими, которых она уже просмотрела и пропустила).

Оцените (при $N \to \infty$) вероятность того, что невесте удастся выбрать наилучшего жениха, если она придерживается следующей стратегии: просмотреть (пропустить) первых по очереди $\lfloor N/e \rfloor$ кандидатов и затем выбрать первого кандидата, который лучше всех предыдущих (впрочем, такого кандидата может и не оказаться, тогда, очевидно, невеста не смогла выбрать наилучшего жениха).

Эта задача уже неоднократно предлагалась школьникам, см., например, брошюру [13].

Отметим, также брошюру [31], вышедшую в этой же серии и посвящённую тому, как «с помощью случайных блужданий» можно численно решать задачу Дирихле для уравнения Лапласа (и других полуэллиптических уравнений).

Задача (модель Эренфестов)*. Рядом стоят две собаки с номерами *1* и *2*. На собаках как-то расположились $M = 2n \gg 1$ блох. Скажем, в начальный момент все блохи собрались на собаке с номером *1*. На каждом шаге случайно и независимо от предыстории определяется блоха (с вероятностью $1/M$ будет выбрана любая из блох), которая перепрыгивает на другую собаку. Микросостояние системы есть способ распределения M различных блох по двум различным собакам. Макросостояние системы есть способ распределения M одинаковых блох между двумя различными собаками. Микросостояний будет 2^M, а макросостояний $M + 1$. Очевидно, макросостояние можно задавать числом блох на первой собаке.

Покажите, что существует такое $T = O(M)$, что для любого $m \geqslant T$

$$P\left(\frac{|n_1(m) - n_2(m)|}{M} \leqslant \frac{3}{\sqrt{M}}\right) \geqslant 0.99,$$

где $n_1(m)$ — число блох на первой собаке на шаге m, а $n_2(m)$ — на второй (случайные величины). Т. е. относительная разность числа блох на собаках будет иметь порядок малости $O(1/\sqrt{M})$ на больших временах ($T \geqslant 2M$).

Замечание. Обозначим через

$$\tau(k) = \inf\{m \in \mathbb{Z}, m \geqslant 0 : n_1(m) = k\},$$
$$\sigma(k) = \inf\{m \in \mathbb{Z}, m > 0 : n_1(m) = k,\ n_1(0) = k\}$$

времена соответственно первого попадания и первого возвращения в состояние k. Тогда

а) $E\sigma(k) = 2^M \dfrac{k!\,(M-k)!}{M!}$, и, в частности, среднее время возвращения в нулевое состояние $E\sigma(0) = 2^M$, где $E\sigma(k)$ — математическое ожидание времени первого возвращения в состояние k, если $n_1(0) = k$, $k = 0, \ldots, n$;

б) $E_n\tau(0) = \dfrac{1}{M}2^M(1 + o(M))$, где $E_n\tau(0)$ — математическое ожидание времени первого попадания в состояние 0, если $n_1(0) = n$;

в) $E_0\tau(n) = n\ln n + n + O(1)$, где $E_0\tau(n)$ — математическое ожидание времени первого попадания в состояние n, если $n_1(0) = 0$.

На примере этой модели можно говорить о том, что в макросистемах возврат к неравновесным макросостояниям вполне допустим, но происходить это может только через очень большое время (*циклы Пуанкаре*), так что нам может не хватить отведённого времени, чтобы это заметить (*парадокс Цермело*). Напомним, что описанный выше случайный процесс обратим во времени. Однако наблюдается необратимая динамика относительной разности числа блох на собаках (*парадокс Лошмидта*). Но в таком случае можно удивляться также и тому, что газ, собранный в начальный момент в одной половине сосуда, с течением времени равномерно распределится по сосуду [16].

Задача (Вильфредо Парето, «Кинетика социального неравенства»)*. В городе живёт $M = 2n \gg 1$ (например, 10 000) пронумерованных жителей. У каждого i-го жителя есть в начальный (нулевой) момент времени целое (неотрицательное) количество рублей $s_i(0)$ (монетками, достоинством в один рубль). Со временем пронумерованные жители (количество которых не изменяется, также как и суммарное количество рублей) случайно разыгрывают своё имущество. В каждый момент времени $t = 1, 2, 3, \ldots$ случайно и независимо от предыстории формируются n пар (все $M!/2^n$ возможных наборов пар равновероятны). В каждой паре с вероятностью $1/2$ житель с большим номером отдаёт 1 рубль (если, конечно, он не банкрот) жителю с меньшим номером, и с вероятностью $1/2$ наоборот. Пусть $c_s(t)$ — доля жителей города, имеющих ровно s рублей в момент времени t (заметим, что $c_s(t)$ — случайная величина). Пусть $S = \sum\limits_{i=1}^{M} s_i(0)$, $\bar{s} = S/M$. Покажите, что для любого $0 \leqslant q \leqslant S$ найдутся такие λ_q, $T_q = O(M)$, что для любого $t \geqslant T_q$

$$P\left(\left|\frac{c_s(t)}{Ce^{-s/\bar{s}}} - 1\right| \leqslant \frac{\lambda_q}{\sqrt{M}}\,,\ s = 0, \ldots, q\right) \geqslant 0.99,$$

где C определяется из условия нормировки: $\sum\limits_{s=0}^{S} Ce^{-s/\bar{s}} = 1$, т. е. $C \approx \bar{s}^{-1}$.

Замечание. Скорость сходимости оценивается сверху, исходя из оценок в доказательстве эргодической теоремы для однородных марковских цепей с конечным числом состояний. Как показывают численные эксперименты, оценка $\mathrm{O}(M)$ точная. Так, если в городе 10 000 жителей и единица времени — день, то при начальном «социальном равенстве» с вероятностью, близкой к единице, через 20–30 лет установится «социальное неравенство». Оказывается (см. также модель Эренфестов), что оценка скорости сходимости $\mathrm{O}(\mathrm{poly}(M))$ характерна для большинства макросистем. Отмеченное выше обстоятельство хорошо известно специалистам по имитационному моделированию, как Markov chain Monte Carlo revolution.

Эта задача в упрощённом виде предлагалась школьникам в [6].

Задача (Markov Chain Monte Carlo Revolution и состоятельность оценок максимального правдоподобия [42])*. В руки опытных криптографов попалось закодированное письмо (10 000 символов). Чтобы это письмо прочитать нужно его декодировать. Для этого берётся стохастическая матрица переходных вероятностей $P = \|p_{ij}\|$ (линейный размер которой определяется числом возможных символов — букв, знаков препинания и т. п. в языке, на котором до шифрования было написано письмо — этот язык известен и далее будет называться базовым), в которой p_{ij} отвечает за вероятность появления символа с номером

j сразу после символа под номером i. Такая матрица может быть идентифицирована с помощью статистического анализа какого-нибудь большого текста, скажем, «Войны и мира» Л. Н. Толстого.

Пусть способ (де)шифрования (подстановочный шифр) определяется некоторой неизвестной дешифрующей функцией $\bar{f}$ — преобразование (перестановка) множества кодовых букв во множество символов базового языка.

В качестве «начального приближения» выбирается какая-то функция f, например, полученная исходя из легко осуществимого частотного анализа. Далее рассчитывается вероятность выпадения полученного закодированного текста $\vec{x}$, сгенерированного при заданной функции f (функция правдоподобия):

$$L(\vec{x}; f) = \prod_k p_{f(x_k), f(x_{k+1})}. \qquad (*)$$

Случайно выбираются два аргумента у функции f и значения функции при этих аргументах меняются местами. Если в результате получилась такая f^*, что $L(\vec{x}; f^*) \geqslant L(\vec{x}; f)$, то $f := f^*$, иначе независимо бросается монетка с вероятностью выпадения орла $p = L(\vec{x}; f^*)/L(\vec{x}; f)$, и если выпадает орёл, то $f := f^*$, иначе $f := f$. Далее процедура повторяется (в качестве f выбирается функция, полученная на предыдущем шаге).

Объясните, почему предложенный алгоритм после некоторого числа итераций с большой вероятностью и с хорошей точностью восстанавливает дешифрующую функцию $\bar{f}$? Почему сходимость оказывается такой быстрой (0.01 сек. на современном PC)?

Замечание. Приведенный алгоритм является частным случаем более общего алгоритма Метрополиса: для того чтобы сгенерировать требуемое дискретное распределение вероятностей (нормировка распределения на единицу несущественна), строится неразложимая марковская цепь с сильно разреженной матрицей переходных вероятностей, (единственным) стационарным распределением которой будет требуемое распределение. По эргодической теореме для марковских цепей, это означает, что после некоторого числа шагов построенной марковской цепи «получится» (независимо от начального распределения) распределение вероятностей близкое к стационарному, т. е. требуемому. Поскольку матрица переходных вероятностей сильно разрежена, то эволюция, согласно описанной марковской динамике, вычислительно малозатратна. Осталось заметить (см., например, предыдущие две задачи), что при довольно естественных предположениях марковская цепь может крайне быстро «выходить» на своё стационарное распределение [48]. Собственно, одним из революционных направлений последних десятилетий при построении эффективных алгоритмов стало использование только что отмеченного факта (возможности потенциально быстрой сходимости в эргодической теореме для марковских процессов). В качестве другого нетривиального и интересного примера укажем вероятностные алгоритмы (работающие быстрее известных детерминированных) поиска центра тяжести выпуклого множества и его объёма [41] (одна из работ в этом направлении была удостоена премии Фалкерсона — аналога Нобелевской премии в области Computer Science).

Обратим также внимание, что школьники могли познакомиться с этой темой на примере задачи тасования карт (сколько раз надо «перемешивать» колоду, чтобы все возможные перестановки были практически равновероятны) из выступления Андрея Окунькова в ЛШСМ-2010 [25].

Для лучшего понимания происходящего в условиях задачи, отметим, что одним из самых универсальных способов получения асимптотически наилучших оценок неизвестных параметров по выборке является метод наибольшего правдоподобия. Напомним вкратце в чем он заключается. Пусть имеется выборка из распределения, зависящего от неизвестного параметра — в нашем случае выборкой $\vec{x}$ из 10 000 элементов будет письмо, а неизвестным «параметром» будет функция f (хотя наш параметр может принимать много значений, но все-таки конечное число). Далее считается вероятность (или плотность вероятности в случае непрерывных распределений) $L\left(\vec{x}; f\right)$ того что выпадет данный $\vec{x}$ при условии, что значения параметра f. Если посмотреть на распределение $L\left(\vec{x}; f\right)$, как на распределение в пространстве параметров ($\vec{x}$ — зафиксирован), то при большом объёме выборки (размерности $\vec{x}$) при естественных условиях это распределение концентрируется в малой окрестности наиболее вероятного значения

$$f\left(\vec{x}\right) = \underset{f}{\mathrm{Argmax}}\, L\left(\vec{x}; f\right),$$

которое «асимптотически» совпадает с искомым значением $\bar{f}$.

Задача (ветвящийся процесс). В колонию зайцев внесли зайца с необычным геном. Обозначим через p_k вероятность того, что в потомстве этого зайца ровно k зайчат унаследуют этот ген ($k = 0, 1, 2, \ldots$). Это же распределение вероятностей характеризует всех последующих потомков, унаследовавших необычный ген. Будем считать, что каждый заяц даёт потомство один раз в жизни в возрасте одного года (как раз в этом возрасте находился самый первый заяц с необычным геном в момент попадания в колонию).

Обозначим через $G\left(z\right)$ — производящую функцию распределения

$$p_k,\ k = 0, 1, 2, \ldots,\ \text{т. е. } G\left(z\right) = \sum_{k=0}^{\infty} p_k z^k.$$

Пусть X_n — количество зайцев в возрасте одного года с необычным геном спустя n лет после попадания в колонию первого такого зайца. Производящую функцию с.в. X_n обозначим $\Pi_n(z) = M\left(z^{X_n}\right)$.

1. Получите уравнение, связывающее $\Pi_{n+1}(z)$ с $\Pi_n(z)$ посредством $G(z)$.

   Указание. Покажите, что $M\left(z^{X_{n+1}} \mid X_n\right) = [G(z)]^{X_n}$. Затем возьмите математическое ожидание от обеих частей равенства.

2. Покажите, что вероятность вырождения гена $q_n = P(X_k = 0;\ k \geqslant n)$ равна $\Pi_n(0)$. Существует ли предел $q = \lim\limits_{n\to\infty} q_n$? Если существует, то найдите его.

   Указание. Легко видеть, что функция $G(z)$ выпуклая. Уравнение $z = G(z)$ имеет два корня: один в любом случае равен 1, другой $q \leqslant 1$. Если $\nu = G'(1) > 1$, то $q < 1$. Если $\nu \leqslant 1$, то $q = 1$.

Подробнее с ветвящимися процессами можно познакомиться, например, по лекционным курсам [4, 8].

Задача (замкнутая сеть, теорема Гордона – Ньюэлла)*. Рассматривается (следуя [5]) транспортная сеть, в которой между N станциями курсируют M такси. Клиенты прибывают в i-й узел в соответствии с пуассоновским потоком с параметром $\lambda_i > 0$ $(i = 1, \ldots, N)$. Если в момент прибытия в i-й узел там есть такси, клиент забирает его и с вероятностью $p_{ij} \geqslant 0$ направляется в j-й узел, по прибытии в который покидает сеть. Такси остаётся ждать в узле прибытия нового клиента. Времена перемещений из узла в узел — независимые случайные величины, имеющие показательное распределение с параметром $\nu_{ij} > 0$ для пары узлов (i, j). Если в момент прихода клиента в узел там нет такси, клиент сразу покидает узел. Считая $p_{ij} = N^{-1}$, $\lambda_i = \lambda$, $\nu_{ij} = \nu$, покажите, что вероятность того, что клиент, поступившей в узел (в установившемся (стационарном) режиме работы сети), получит отказ, равна

$$p_{\text{отказа}}(N, M) = \sum_{k=0}^{M} \frac{C_{N-2+k}^{k}\rho^{M-k}}{(M-k)!} \Bigg/ \sum_{k=0}^{M} \frac{C_{N-1+k}^{k}\rho^{M-k}}{(M-k)!}, \quad \rho = N\lambda/\nu.$$

Методом перевала покажите справедливость следующей асимптотики при $N \to \infty$:

$$p_{\text{отказа}}(N, rN) = 1 - \frac{2r}{\lambda/\nu + r + 1 + \sqrt{(\lambda/\nu + r + 1)^2 - 4\lambda r/\nu}} + O\left(\frac{1}{N}\right).$$

Метод перевала — очень полезный инструмент исследования асимптотик интегралов по параметру. Подробное изложение метода перевала имеется, например, в [32].

* * * * * *

Недавно вышел современный учебник по марковским случайным процессам, который мы рекомендуем заинтересовавшимся в этой теме читателям [18].

Вероятностные алгоритмы и вероятностный анализ алгоритмов

Задача (о сложности в среднем быстрой сортировки массива). Дан случайным образом приготовленный массив длины n (все $n!$ возможных порядков равновероятны). Покажите, что сложность в среднем алгоритма сортировки этого массива Quicksort есть $\mathrm{O}\,(n\ln n)$. Верно ли, что сложность в худшем случае при этом есть $\mathrm{O}\left(n^2\right)$?

Задача (об упаковке). Рассмотрим (следуя [21]) NP-полную задачу

$$\sum_{j=1}^{n} x_j \to \max;$$
$$x_j \in \{0,\,1\},\ j=1,\dots,n;\quad \sum_{j=1}^{n} a_{ij}x_j \leqslant 1,\ i=1,\dots,m \tag{*}$$
$$(a_{ij}\in\{0,\,1\}\,,\ i=1,\dots,m,\ j=1,\dots,n).$$

Булев вектор $\vec{x}$ длины n будем называть допустимым, если он удовлетворяет системе $(*)$. Обозначим через $T\,(j)$ множество всех допустимых булевых векторов для системы $(*)$ с $n-j$ нулевыми последними компонентами и через $\vec{e}_j$ — вектор длины n с единичной j-й компонентой и с остальными нулевыми компонентами.

Рассмотрим алгоритм: 1) строим множество допустимых решений $T\,(j)$ на основе множества $T\,(j-1)$, пытаясь добавить вектор $\vec{e}_j$ ко всем булевым векторам $T\,(j-1)$; 2) среди $|T\,(n)|$ допустимых булевых векторов ищем «наилучший».

а) Покажите, что сложность описанного алгоритма составляет $O(|T(n)mn|)$. При каких $a_{ij}\in\{0,\,1\}$ алгоритм будет работать экспоненциально долго?

б) Оцените сложность в среднем $E(|T(n)|)mn$, т.е. математическое ожидание времени работы алгоритма, если с.в. $\{a_{ij}\}_{i,j=1}^{m,n}$ — независимые и одинаково распределенные по закону Бернулли $Be\,(p)$ $(mp^2\geqslant \ln n)$.

Указание к б). Пусть $k>0$. Положим: $\vec{x}_{j_1,\dots,j_k}$ — вектор с единицами на позициях $\{j_1,\dots,j_k\})$ и $n-k$ нулями; p_{ki} — вероятность выполнения i-го неравенства системы $(*)$ для $\vec{x}_{j_1,\dots,j_k}$; P_k — вероятность того, что $\vec{x}^k$ — допустимое решение (покажите, что p_{ki} и P_k не зависят от набора $\{j_1,\dots,j_k\})$.

Докажите, что $p_{ki} \leqslant \left(1-p^2\right)^{k-1} \leqslant e^{-p^2(k-1)}$, $P_k \leqslant e^{-mp^2(k-1)}$ и

$$E\left(|T\left(n\right)|\right) = \sum_{k=0}^{n} C_n^k P_k < 1 + n + n\sum_{k=2}^{n} e^{(k-1)\left(\ln n - mp^2\right)}.$$

Задача (сравнение строк)*. Требуется сравнить две битовые строки a, b, затратив как можно меньше операций. Основная идея — сравнивать не сами строки, а функции от них; скажем, сравнивать $a \bmod p$ и $b \bmod p$, для некоторого простого числа p.

Описание алгоритма сравнения строк:

1. Пусть $|a| = |b| = n$, $N = n^2 \log_2 n^2$.
2. Выбираем случайное простое число p из интервала $[2, \ldots, N]$.
3. Выдать «да», если $a \bmod p = b \bmod p$ (т. е. $(a-b) \equiv 0 \pmod p$), иначе выдать «нет».

Покажите, что этот алгоритм является вероятностным алгоритмом с односторонней ошибкой и вероятностью ошибки $O\left(1/n\right)$. То есть

$$P\left\{\text{«да»} \mid a = b\right\} = 1,$$
$$P\left\{\text{«да»} \mid a \neq b\right\} = O(1/n).$$

При этом необходимое количество сравнений равно $O\left(\log_2 n\right)$.

Задача (интерактивные доказательства, изоморфизм графов). Алисе известен изоморфизм φ графов G_0 и G_1. Она хочет убедить Боба в том, что графы G_0 и G_1 изоморфны, не сообщив Бобу «никакой информации» об изоморфизме φ.[1)] Алиса посылает Бобу граф, который получается некоторой случайной перестановкой вершин ψ из графов G_0, G_1, т. е. $H = \psi(G_a)$, где $a \in \{0, 1\}$ выбирается равновероятно. Боб просит сообщить ему изоморфизм между G_b и H, где $b \in \{0, 1\}$ также выбирается равновероятно. При $a = b$ Алиса посылает ψ, иначе — $\psi\varphi^{1-2b}$. Таких партий разыгрывается N штук, случайные величины в разных партиях независимы (ψ в каждой партии генерируется заново). Если φ — действительно изоморфизм $G_0 \sim G_1$, то все проверки Боба будут положительны. Покажите, что если φ не является изоморфизмом G_0 и G_1, то с вероятностью 2^{-N} хотя бы одна проверка обнаружит это, т. е. Боб получит перестановку, которая не является изоморфизмом H и G_b.

Замечание. Этот задача поучительна с точки зрения «криптографического фокуса»: Алиса убедила Боба в $G_0 \sim G_1$ так и не огласив самого изоморфизма φ. Если φ — это пароль, диалог можно вести даже в открытую, что служит примером системы с нулевым разглашением.

[1)] Мы не приводим точных определений — в задаче они не используются.

Обратим внимание, что школьники могли познакомиться с тематикой нескольких последних задач из выступления А. А. Разборова в ЛШСМ-2011 [27].

Задача (адаптивные правила [46])*. Имеется два «одноруких бандита» (так называют игровые автоматы с ручкой, дёргая за которую получаем случайный выигрыш). Вероятность выиграть на первом автомате $p_1 > 0$, а на втором $p_2 > 0$. Обе вероятности неизвестны. Игрок может в любом порядке $n \gg 1$ раз дёргать за ручки «одноруких бандитов». Стратегией игрока является выбор ручки на каждом шаге, в зависимости от результатов всех предыдущих шагов, так чтобы суммарный выигрыш был бы максимальным. Приведите асимптотически оптимальную стратегию игрока.

Решите предыдущую задачу, если выигрыш есть случайная величина с распределением из экспоненциального семейства, зависящего от неизвестного параметра θ (определение см. в [22]). Хотя значения θ_1 и θ_2 (для 1-го и 2-го игрового автомата) неизвестны, но, не ограничивая общности, считайте, что $\theta_1 \neq \theta_2$. Выигрыш является суммой выигрышей во всех розыгрышах.

Задача (устойчивые системы большой размерности; В.И.Опойцев, 1985 [26]). Из курсов функционального анализа и вычислительной математики хорошо известно, что если спектральный радиус матрицы $A = \|a_{ij}\|_{i,j=1}^{n}$ меньше единицы: $\rho(A) < 1$, то итерационный процесс $\vec{x}^{n+1} = A\vec{x}^{n} + \vec{b}$ независимо от начальной точки $\vec{x}^{0}$, сходится к единственному решению уравнения $\vec{x}^{*} = A\vec{x}^{*} + \vec{b}$.

Если $\|A\| = \max\limits_{i}\sum\limits_{j}|a_{ij}| < 1$, то и $\rho(A) < 1$. Обратное, конечно же, неверно. Предположим, что существует такое маленькое $\varepsilon > 0$, что

$$\frac{1}{n}\sum_{i,j}|a_{ij}| < 1 - \varepsilon. \tag{S}$$

Очевидно, что отсюда тем более не следует $\rho(A) < 1$. Тем не менее, введя на множестве матриц, удовлетворяющих условию (S), равномерную меру, покажите, что относительная мера тех матриц (удовлетворяющих условию (S)), для которых спектральный радиус не меньше единицы, стремится к нулю с ростом n (ε — фиксировано и от n не зависит).

Указания. 1. Покажите, что при доказательстве можно ограничиться матрицами с неотрицательными элементами.

2. Покажите, что, не ограничивая общности, можно также считать, что в определении множества S стоит не неравенство, а равенство. Так определённое множество матриц будем называть SE.

3. Положите, например, $a_{ij} \in \mathrm{Exp}\,(n/(1-\varepsilon))$ — независимые одинаково распределённые с.в., и покажите, что при $n \to \infty$ распределение элементов случайной матрицы $A = \|a_{ij}\|_{i,j=1}^{n}$ будет сходиться к равномерному распределению на SE.

4. Покажите, введя обозначение $P_n = P\,(\|A\| \geqslant 1) \geqslant P\,(\rho\,(A) \geqslant 1)$ и используя неравенство Чебышёва, что

$$P_n \leqslant nP\Big(\sum_j a_{1j} \geqslant 1\Big) \leqslant \frac{n}{\varepsilon^4} M\left[\Big(\sum_j a_{1j} - (1-\varepsilon)\Big)^4\right] = O\big(\frac{1}{n}\big) \underset{n\to\infty}{\longrightarrow} 0.$$

* * * * * *

В качестве дополнительной литературы по вероятностным алгоритмам и вероятностному анализу алгоритмов можно рекомендовать книги [20,49].

Геометрические вероятности

Задача (средняя площадь поверхности и интегральная геометрия). Покажите, что средняя площадь ортогональной проекции куба с ребром единица на случайную плоскость равна 3/2.

Указание. Обозначим через S_k величину k-мерного объёма ортогональной проекции рассматриваемой области в $\mathbb{R}^n$ на случайную k-мерную плоскость, или, что то же самое, среднее значение (усреднённое по всем k-мерным плоскостям, предполагаемым равновероятными) площади ортогональной k-мерным проекции области. Оказывается, что S_k также равны средним значениям (усреднённым по поверхности рассматриваемой области) симметрических функций от главных кривизн поверхности, и участвуют в (удивительной) формуле для объёма h-окрестности этой области:

$$V\,(h) = V_0 + V_1 h + V_2 h^2 + \ldots + V_n h^n,$$

где V_0 — объем области; V_1 — $(n-1)$-мерный объем границы области; число V_k пропорционально S_k и выражается через средние значения от произведений k главных кривизн. В случае $n = 3$, из главных кривизн k_1 и k_2 в каждой точке можно составить *среднюю кривизну* $k_1 + k_2$ и *гауссову кривизну* $K = k_1 k_2$. В этом случае объем h-окрестности получается $V\,(h) = V_0 + V_1 h + V_2 h^2 + V_3 h^3$, где V_2 пропорционально интегралу от средней кривизны по всей поверхности, а V_3 — от гауссовой:

$$V_3 = \frac{4}{3}\pi \iint K\,dS.$$

Например, для сферы радиуса R

$$V(h) = \frac{4}{3}\pi \cdot (R+h)^3 = \frac{4}{3}\pi R^3 + h \cdot (4\pi R^2) + h^2(4\pi R) + \frac{4}{3}\pi h^3.$$

Здесь

$$k_1 = k_2 = 1/R, \quad k_1 + k_2 = 2/R, \quad k_1 k_2 = 1/R^2, \quad \iint (k_1 + k_2)\, dS = 8\pi R,$$

$$\iint (k_1 k_2) dS = 4\pi \quad \text{(формула Гаусса – Бонне).}$$

Коэффициент V_3 не зависит от деталей области, а зависит только от *эйлеровой характеристики* поверхности рассматриваемой области. Это обстоятельство привело Г. Вейля к созданию теории характеристических классов и чисел, обобщающих формулу Гаусса – Бонне.

По-видимому, первым эту задачу решился предложить школьникам В. И. Арнольд [2].

Задача (принцип концентрации площади сферы; А. Пуанкаре, 1911). Покажите, что если в многомерном шаре задано равномерное распределение вероятностей и согласно этому распределению вероятностей сгенерированы два случайных вектора, то с вероятностью, близкой к единице, концы этих векторов будут лежать почти на границе шара и эти два случайных вектора будут почти ортогональны.

Указание. Нетривиально второе утверждение (про ортогональность). Для того чтобы его установить, покажите, что доля от площади всей сферы S_r^n (радиуса r), которую занимает площадь сегмента, проектирующегося в отрезок $[a, b]$, скажем, оси x_1, равна

$$P[a, b] = \int_a^b \left(1 - (x/r)^2\right)^{\frac{n-3}{2}} dx \Bigg/ \int_{-r}^{r} \left(1 - (x/r)^2\right)^{\frac{n-3}{2}} dx \ .$$

Фиксируя $r = 1$ и устремляя n к бесконечности, получите,

$$P\left[-\delta, \delta\right] \sim 1 - \sqrt{\pi/2} \exp\left(-\delta^2 n/2\right).$$

В статистической физике $\sum\limits_{i=1}^{n} V_i^2 = \dfrac{2E_n}{m} \sim n$. Поэтому если известно, что вектор скоростей молекул газа равномерно распределён по поверхности постоянной энергии, то для того чтобы найти (следуя Максвеллу) распределение компонент вектора скорости, скажем V_1, нужно осуществить термодинамический скейлинг $n \to \infty$, $r = \sigma n^{1/2} \to \infty$

$$P[a, b] = \frac{\int\limits_a^b \exp\left(-\dfrac{x^2}{2\sigma^2}\right) dx}{\int\limits_{-\infty}^{\infty} \exp\left(-\dfrac{x^2}{2\sigma^2}\right) dx} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_a^b \exp\left(-\frac{x^2}{2\sigma^2}\right) dx.$$

Таким образом, получаем нормальный закон распределения Максвелла в статистической физике.

Задача (изопериметрическое неравенство и принцип концентрации меры; П. Леви, 1919)*. Число μ_f называют медианой функции f, если

$$\mu\left(\vec{x} \in S_1^n : f(\vec{x}) \geqslant \mu_f\right) \geqslant 1/2 \text{ и } \mu\left(\vec{x} \in S_1^n : f(\vec{x}) \leqslant \mu_f\right) \geqslant 1/2,$$

где $\mu(d\vec{x})$ — равномерная мера на единичной сфере S_1^n в $\mathbb{R}^n$. Пусть A — измеримое (борелевское) множество на сфере S_1^n. Через A_δ будем обозначать δ-окрестность множества A на сфере S_1^n. Предположим теперь, что в некотором царстве, расположенном на S_1^3, царь предложил царице Дидоне построить огород с заданной длиной забора. Царица хочет, чтобы её огород при заданном периметре имел наибольшую площадь. Таким образом, царице надо решить изопериметрическую задачу (такие задачи обычно рассматриваются в курсах вариационного исчисления). Решение этой задачи хорошо известно — «круглый огород». Для нас же полезно, рассмотрение двойственной задачи, имеющей такое же решение: при заданной площади огорода спроектировать его так, чтобы он имел наименьшую длину забора, его ограждающего. Используя решение этой задачи, покажите, что если $\mu(A) \geqslant 1/2$, то

$$\mu(A_\delta) \geqslant 1 - \sqrt{\pi/2}\exp\left(-\delta^2 n/2\right).$$

Пусть теперь на S_1^n задана функция с модулем непрерывности

$$\omega_f(\delta) = \sup\left\{|f(\vec{x}) - f(\vec{y})| : \rho(\vec{x}, \vec{y}) \leqslant \delta,\ \vec{x}, \vec{y} \in S_1^n\right\}.$$

Тогда

$$\mu\left(\vec{x} \in S_1^n : |f(\vec{x}) - \mu_f| \geqslant \omega_f(\delta)\right) \leqslant \sqrt{\pi/2}\exp\left(-\delta^2 n/2\right).$$

Можно показать, что при весьма естественных условиях медиана асимптотически близка к среднему значению (математическому ожиданию). Аналогичное неравенство можно получить (М. Талагран, 1994), например, для модели случайных графов (Эрдёша – Реньи) и исследовать плотную концентрацию около среднего значения различные функций на случайных графов: число независимости, хроматическое число и т. п.

Задача (изопериметрическое неравенство Талаграна и его приложения; М. Талагран, 1996 [43])*. а) Пусть заданы множества Ω_i, $i = 1, \ldots, n$, элементарных исходов. На этих множествах заданы вероятностные меры P_i, $i = 1, \ldots, n$. Положим

$$\Omega = \prod_{i=1}^{n} \Omega_i, \quad P = \prod_{i=1}^{n} P_i.$$

Введём взвешенную метрику Хэмминга:

$$d_\alpha(\vec{x},\vec{y}) = \sum_{x_i \neq y_i} \alpha_i \Bigg/ \sqrt{\sum_{i=1}^{n} \alpha_i^2}$$

и определим $d_\alpha(\vec{x}, A) = \min\limits_{\vec{y}\in A} d_\alpha(\vec{x},\vec{y})$, $\rho(\vec{x}, A) = \sup\limits_{\vec{\alpha}\in \mathrm{R}^n} d_\alpha(\vec{x}, A)$. Пусть $A \subset \Omega$. Определим t-окрестность ($t \geqslant 0$) множества A по формуле

$$A_t = \{\vec{x} \in \Omega :\ \rho(\vec{x}, A) \leqslant t\}.$$

Покажите, что тогда справедливо следующее неравенство:

$$P(A)(1 - P(A_t)) \leqslant \exp\left(-t^2/4\right).$$

б) Пусть в сельском районе, имеющем форму квадрата со стороной 1, находится n домов ($n \gg 1$), размерами которых можно пренебречь по сравнению с линейным размером района. Будем считать, что при строительстве домов застройщик случайно (согласно равномерному распределению $R[0,1]^2$) и независимо выбирал их местоположения. Почтальону необходимо обойти все n домов ровно по одному разу (от любого дома почтальон может направиться к любому другому по прямой). Обозначим через TSP длину кратчайшего из таких путей (кратчайший гамильтонов путь). Используя п. а), покажите, что найдутся такие постоянные $c > 0$ и $\beta > 0$, не зависящие от n, что

$$P(|TSP - E[TSP]| \geqslant t) \leqslant \exp\left(-t^2/(4c)\right), \quad \text{где } E[TSP] \sim \beta\sqrt{n}.$$

в) Пусть в условиях п. б) требуется построить систему дорог минимальной суммарной длины $SteinerTree$, по которой можно было бы добраться из любого дома в любой другой (дерево Штейнера с минимальной суммарной длиной рёбер). Получите неравенство о плотной концентрации с.в. $SteinerTree$ в окрестности своего математического ожидания, аналогичное неравенству п. б). Как себя асимптотически ведёт $E[SteinerTree]$ при $n \to \infty$?

Задача (перколяция, [37])*. В квадратном пруду (со стороной равной 1) выросли (случайным образом) $N \gg 1$ цветков лотоса, имеющих форму круга радиуса $r > 0$. Назовём r_N *радиусом перколяции*, если с вероятностью не меньшей 0.99 не любящий воду жук сможет переползти по цветкам лотоса с северного берега на южный, не замочившись.

Покажите, что $r_N \sim C \big/ \sqrt{N}$. Оцените C.

* * * * * *

Для более глубокого погружения в геометрическую теорию вероятностей можно рекомендовать следующие книги [14, 19, 47].

Применение вероятностных методов в теории чисел

Задача (теорема Гаусса – Гильдена – Вимана – Кузьмина [3, 17, 44]). Каждое число из промежутка $\Omega = [0,\, 1)$ может быть разложено в цепную дробь (вообще говоря, бесконечную). Цепные дроби играют важную роль, например, в различных вычислениях (поскольку позволяют строить в определённом смысле наилучшие приближения иррациональных чисел рациональными), в теории динамических систем (КАМ теории). Для рациональных чисел такие дроби конечны, для квадратичных иррациональностей — периодические (см. пример ниже, в котором период равен 1):

$$\frac{\sqrt{5}-1}{2} = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \ldots}}} = \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \ldots}}}.$$

Покажите, что, несмотря на приведённый выше пример, для почти всех (в равномерной мере) точек $\omega \in [0,\, 1)$ и любого натурального m

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} I_m\left(a_k\left(\omega\right)\right) = \frac{1}{\ln 2} \ln\left(1 + \frac{1}{m\left(m+2\right)}\right), \quad I_m(a_k) = \begin{cases} 1, & a_k = m, \\ 0, & a_k \neq m. \end{cases}$$

Указание. Покажите, что преобразование $T\colon [0,\, 1) \to [0,\, 1)$

$$T\omega = \begin{cases} \left\{\frac{1}{\omega}\right\}, & \omega \in \left(0,\, 1\right), \\ 0, & \omega = 0, \end{cases}$$

где $\{5.8\} = 0.8$ — дробная часть числа, сохраняет меру Гаусса

$$P(A) = \frac{1}{\ln 2} \int\limits_{A} \frac{dx}{1+x}\,,$$ где $A \in \Xi$ — σ-алгебре борелевских множеств на Ω;

т. е. покажите, что $P\left(T^{-1}A\right) = P\left(A\right)$.

Рассмотрите случайный процесс (в дискретном времени) $X_k\left(\omega\right) = x_m\left(T^k\omega\right)$, где

$$x_m\left(\omega\right) = \begin{cases} 1, & \omega^{-1} \in \left[m,\, m+1\right), \\ 0, & \omega^{-1} \in \left(1,\, m\right) \cup \left[m+1,\, \infty\right). \end{cases}$$

Покажите, что случайный процесс X_k — стационарный в узком смысле. В предположении, что этот процесс эргодичен по математическому ожиданию:

$$\frac{1}{n} \sum_{k=1}^{n} X_k\left(\omega\right) \xrightarrow[n\to\infty]{\text{п.н.}} MX_k\left(\omega\right) = \text{const},$$

найдите искомый предел.

Задача (КАМ теория [30]). Число α из отрезка $[0,\,1]$ назовём нормально приближаемым рациональными числами, если найдутся c, $\varepsilon > 0$ такие, что при любом натуральном q

$$\min_{p\in\mathbb{Z}}\left|\alpha - \frac{p}{q}\right| \geqslant \frac{c}{q^{2+\varepsilon}}\,.$$

Используя лемму Бореля – Кантелли докажите, что множество нормально приближаемых чисел на отрезка $[0,\,1]$ имеет лебегову меру 1.

Указание. Зафиксируем c, $\varepsilon > 0$ и рассмотрим множество

$$A_q = \left\{\alpha \in [0,\,1] : \min_{p\in\mathbb{Z}}\left|\alpha - \frac{p}{q}\right| < \frac{c}{q^{2+\varepsilon}}\right\}.$$

Покажите, что $\mu\left(A_q\right) \leqslant 2c/q^{1+\varepsilon}$. Таким образом, ряд $\sum\limits_q \mu\left(A_q\right)$ сходится. В силу леммы Бореля – Кантелли отсюда следует нужное утверждение.

Заметим, что эта задача пришла из теории динамических систем на двумерном торе. Подобного же рода задачи возникают и в КАМ теории.

В связи с полученным результатом, будет интересно заметить [33], что существует такая бесконечная последовательность q_k и соответствующая ей последовательность p_k, что

$$\left|\alpha - \frac{p_k}{q_k}\right| < \frac{1}{\sqrt{5}}\,\frac{1}{q_k^2}\,.$$

В теории цепных дробей показывается, что последовательность p_k/q_k будет подпоследовательностью последовательности подходящих дробей для числа α. Заметим также, что константу $1/\sqrt{5}$ в неравенстве уменьшить нельзя.

Задача (статистика теоретико-числовых функций). Довольно часто вероятностные соображения (например, независимость) используются в теории чисел не совсем строго, но зато весьма часто они позволяют угадать правильный ответ. Поясним сказанное, пожалуй, наиболее популярным примером из книги [17].

Пусть A — некоторое множество положительных целых чисел. Обозначим через $A(n)$ количество тех его элементов, которые содержатся среди первых n чисел натурального ряда. Если существует предел $\lim\limits_{n\to\infty} A(n)/n = P(A)$, то он называется плотностью A. К сожалению, вероятностная мера $P(A)$ не является вполне аддитивной (счётно-аддитивной).

Рассмотрим целые числа, делящиеся на простое число p. Плотность множества таких чисел, очевидно, равна $1/p$. Возьмём теперь множество целых чисел, которые делятся одновременно на p и q (q — другое простое число). Делимость на p и q эквивалентна делимости на pq, и, следовательно, плотность нового множества равна $1/pq$.Так как $1/pq = (1/p)\cdot(1/q)$, то мы можем истолковать это так: «события», заключающиеся в делимости на p и q, независимы. Это, конечно, выполняется для любого количества простых чисел.

Поставим теперь задачу посчитать долю несократимых дробей или, другими словами, «вероятность» несократимости дроби (фиксируется знаменатель дроби n, а затем случайно, с равной вероятностью $1/n$ выбирается любое число от 1 до n в качестве числителя, и подсчитывается доля случаев, в которых полученная дробь оказывалась несократимой) в следующем смысле (здесь и далее индекс p может пробегать только простые числа):

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\frac{\#\{k<n:\text{НОД}(n,k)=1\}}{n}=\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\frac{\varphi(n)}{n}=$$
$$=\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\prod_{p}\left(1-\frac{\rho_p(n)}{p}\right),$$

где $\varphi(n)$ — функция Эйлера, $\rho_p(n)=\begin{cases}1, & n \text{ делится на } p,\\ 0, & \text{иначе.}\end{cases}$

Согласно введённому выше определению плотности:

$$M\left\{\prod_{p\leqslant p_k}\left(1-\frac{\rho_p\left(n\right)}{p}\right)\right\}=\prod_{p\leqslant p_k}M\left\{\left(1-\frac{\rho_p\left(n\right)}{p}\right)\right\}=\prod_{p\leqslant p_k}\left(1-\frac{1}{p^2}\right).$$

С учётом этого хочется написать следующее:

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\frac{\varphi(n)}{n}=M\left\{\frac{\varphi(n)}{n}\right\}=M\left\{\prod_{p}\left(1-\frac{\rho_p\left(n\right)}{p}\right)\right\}\stackrel{?}{=}$$
$$\stackrel{?}{=}\prod_{p}M\left\{\left(1-\frac{\rho_p\left(n\right)}{p}\right)\right\}=\prod_{p}\left(1-\frac{1}{p^2}\right)=\frac{1}{\zeta(2)}=\frac{6}{\pi^2}.$$

Будь введённая вероятностная мера, по которой считается это математическое ожидание, счётно-аддитивной, то можно было бы поставить точку, получив ответ. Однако, это не так. Хотя ответ мы и получили правильный, но приведённое выше рассуждение не может считаться доказательством. Впрочем, часто вероятностные рассуждения удаётся пополнить, используя их в качестве основы. Так в разобранном нами примере все сводится к обоснованию равенства «?».

Легко понять, что полученный ответ несёт определённую информацию о статистических свойствах функции Эйлера.

В теории чисел такого типа задачи занимают крайне важное место. Достаточно сказать, что гипотеза Римана «на миллион» о распределении нетривиальных нулей дзета-функции Римана $\varsigma(z)$ равносильна следующему свойству функции Мёбиуса $\mu(n)$:

$$\left|\sum_{n=1}^{N}\mu(n)\right| \leqslant \sqrt{N} \quad \text{(Олдыжко – Риэль).}$$

Что в свою очередь (Х. М. Эдвардс) в определённом смысле «завязано» на случайность последовательности $\{\mu(n)\}_{n\in\mathbb{N}}$.

Используя указанный выше формализм, найдите долю чисел натурального ряда, свободных от квадратов, т. е. не делящихся на квадрат любого простого числа.

По-видимому, первым эту задачу решился предложить школьникам около десяти лет назад В. И. Арнольд. В заключение заметим, что применение вероятностных соображений в теории чисел продолжает привлекать ведущих математиков и по сей день (см., например, выступление Я. Г. Синая на семинаре «Глобус» [39]).

Задача (предельные меры; А. М. Вершик и др., 1977 [9, 11])*. В качестве множества элементарных исходов рассматривается группа всевозможных подстановок (перестановок) S_n (симметрическая группа), $n \gg 1$. В этой группе $n!$ элементов. Припишем каждой подстановке одинаковую вероятность $1/n!$.

А) Покажите, что математическое ожидание числа циклов есть $\approx \ln n$.

Б) В каком смысле нормированные длины циклов случайной подстановки убывают со скоростью геометрической прогрессии со знаменателем e^{-1}?

В) Положим $\rho_n(a) = |\{g \in S_n : n_{\max}(g) \leqslant an\}|/n!$, где $n_{\max}(g)$ — длина максимального цикла в подстановке g. Покажите, что $\rho_n(a)$ удовлетворяет *уравнению Дикмана – Гончарова* (40-е годы XX века):

$$\rho_n(a) = \int_0^a \rho_n\left(\frac{a}{1-t}\right) dt.$$

Г) Покажите, что, начиная с некоторого большого числа N, 99% натуральных чисел n, больших, чем N, обладают свойством

$$n^{0.99} < p_1 \cdot \ldots \cdot p_{11}, \quad n = p_1 \cdot \ldots \cdot p_{m(n)},\ p_1 \geqslant p_2 \geqslant \ldots \geqslant p_{m(n)},\ p_i \text{ — простые.}$$

Иначе говоря, у основной части (99%) натуральных чисел основная часть (99%) числа есть произведение наибольших простых делителей. Число 11 возникло из-за того, что мы выбрали 99% и 99%.

Указание. Решение задач всех пунктов сводится (технически весьма нетривиально!) к задаче о «ломании палки». Отрезок $[0, 1]$ делится («ломается») случайно с равномерной вероятностью. Левый отрезок фиксируем, а правый ломается аналогичным образом и т. д.

Обратим внимание, что школьники могли познакомиться с этой темой из выступления А. М. Вершика в ЛШСМ-2008 [10].

Элементы математической статистики

Задача (критерий согласия Колмогорова). Пусть $\vec{X} = (X_1, X_2, \ldots, X_n)$ — простая выборка объёма n из распределения $F(x)$. Покажите, что для непрерывных распределений $F(x)$ распределение статистики $D_n(\vec{X}) = \sup\limits_{x\in\mathbb{R}} \left|F_n\big(x;\vec{X}\big) - F(x)\right|$ не зависит от $F(x)$. Здесь $F_n\big(x;\vec{X}\big)$ обозначает эмпирическую функцию распределения:

$$F_n\big(x;\vec{X}\big) = \frac{\mu_n(x)}{n} = \frac{1}{n}\sum_{k=1}^{n} I_k(x),$$

где $\mu_n(x) = |\{j : X_j < x\}|$, $I_k(x) = \begin{cases} 1, & X_k < x, \\ 0, & X_k \geqslant x. \end{cases}$

В действительности, подобный результат будет справедлив и для широкого класса статистик $G(F_n, F)$ (т. е. измеримых функционалов) от $F_n\big(x;\vec{X}\big)$.

Указание. Положим по определению

$$F^{-1}(u) = \inf\{\xi : F(\xi) = u\} = \min\{\xi : F(\xi) = u\} \quad \text{для } u \in [0,\,1],$$

где последнее равенство имеет место в силу непрерывности $F(x)$. Понятно, что это отнюдь не единственный способ выбора однозначной функции из, вообще говоря, многозначного отображения $F^{-1}(u)$, однако именно такое определение окажется наиболее полезным в дальнейшем. Положим $u = F(x)$, тогда u пробегает как минимум все точки интервала $(0,\,1)$ (а как максимум — отрезка $[0,\,1]$), когда x пробегает $\mathbb{R}$. Делая замену $u = F(x)$ и используя определение $F^{-1}(u)$, получим

$$D_n(\vec{X}) = \max_{u\in[0,\,1]} \left|F_n\big(F^{-1}(u);\vec{X}\big) - u\right|.$$

Далее имеем

$$F_n\left(F^{-1}(u)\right) = \frac{1}{n}\sum_{k=1}^{n} \theta\left(F^{-1}(u) - X_k\right).$$

Остается показать эквивалентность событий

$$\left\{F^{-1}(u) - X_k \leqslant 0\right\} \sim \{u - F(X_k) \leqslant 0\}.$$

Сделайте это, завершив, тем самым, доказательство сформулированного утверждения.

Задача (распределение Коши). На плоскости на расстоянии $a > 0$ (неизвестный параметр) от детектирующей прямой располагается радиоактивный источник, который излучает вспышками равновероятно по любому направлению в этой плоскости. Пусть $\vec{x} = (x_1, \ldots, x_n)$ — вектор координат вспышек, регистрируемых детектором. Требуется построить по такой простой выборке состоятельную оценку координаты проекции источника на детектирующую прямую.

Задача (оценка параметра с помощью эргодической теоремы). В модели Блэка – Шоулса – Мертона эволюция цены акции описывается геометрическим броуновским движением

$$S(t) = S(0) \exp\left(at + \sigma W(t)\right),$$

где $W(t)$ — винеровский процесс ($\sigma > 0$). С помощью эргодической теоремы для случайных процессов оцените неизвестный параметр a, если известна реализация процесса $S(t)$ на достаточно длинном временном отрезке $[0, T]$. Предложите способ оценки неизвестного параметра σ.

Задача (критерий Неймана – Пирсона). В результате эксперимента получена простая выборка $\vec{x}$, относительно которой имеются две простые гипотезы:

$$H_0 : \vec{x} \in L(\vec{x} \mid H_0) \quad \text{и} \quad H_1 : \vec{x} \in L(\vec{x} \mid H_1).$$

С уровнем значимости $\alpha > 0$ постройте наиболее мощный критерий проверки гипотезы H_0 против альтернативы H_1, т. е. найдите такую функцию $\varphi(\vec{x})$ (решающее правило — вероятность, с которой следует принимать гипотезу H_1, если выпал $\vec{x}$), что

$$P(H_0 \mid H_1) = \int_\Omega (1 - \varphi(\vec{x})) L(\vec{x} \mid H_1)\, d\vec{x} \to \min_{0 \leqslant \varphi(\,\cdot\,) \leqslant 1}$$

$$P(H_1 \mid H_0) = \int_\Omega \varphi(\vec{x})\, L(\vec{x} \mid H_0)\, d\vec{x} = \alpha.$$

Покажите, что решение этой задачи — наиболее мощное решающее правило (Неймана – Пирсона) — имеет вид

$$\varphi(\vec{x}) = \begin{cases} 1, & \Lambda(\vec{x}) > \bar{\Lambda} \\ p(\vec{x}), & \Lambda(\vec{x}) = \bar{\Lambda} \\ 0, & \Lambda(\vec{x}) < \bar{\Lambda} \end{cases}, \quad \text{где } \Lambda(\vec{x}) = \frac{L(\vec{x} \mid H_1)}{L(\vec{x} \mid H_0)},$$

$\bar{\Lambda}$ и $0 \leqslant p(\vec{x}) \leqslant 1$ следует определять из условия (ошибкой первого рода):

$$P(H_1 \mid H_0) = \int\limits_{\{\vec{x}:\Lambda(\vec{x})>\bar{\Lambda}\}} L(\vec{x} \mid H_0)\, d\vec{x} + \int\limits_{\{\vec{x}:\Lambda(\vec{x})=\bar{\Lambda}\}} p(\vec{x}) L(\vec{x} \mid H_0)\, d\vec{x} = \alpha. \quad (*)$$

Причём $\bar{\Lambda}$ определяется единственным образом, а от того, как выбирать $p(\vec{x})$, удовлетворяющее $(*)$, не зависит ошибка второго рода:

$$P(H_0 \mid H_1) = \int\limits_{\{\vec{x}:\Lambda(\vec{x})<\bar{\Lambda}\}} L(\vec{x} \mid H_1)\, d\vec{x} + \int\limits_{\{\vec{x}:\Lambda(\vec{x})=\bar{\Lambda}\}} (1 - p(\vec{x})) L(\vec{x} \mid H_1)\, d\vec{x} = \beta.$$

Указание. Воспользуйтесь методом множителей Лагранжа [23].

Задача (о наиболее мощном критерии для оценки студентов). Опытный преподаватель математической статистики знает, что к нему на экзамен могут приходить два типа студентов: знающие предмет и не знающие предмет. Причём, ввиду всяких случайных факторов (не выспался, переволновался, воспользовался шпаргалкой) впечатление, которое студент производит на экзаменатора, — случайная величина. Для знающих студентов она имеет пуассоновское распределение с параметром 200, а для не знающих — с параметром 100. Преподаватель ставит только две оценки: «зачёт» и «незачёт». Философия преподавателя такова: «пусть лучше я поставлю «зачёт» студенту, который этого не заслуживает, нежели поставлю «незачёт» знающему материал студенту». Предложите, как следует действовать преподавателю (какую оценку ставить в зависимости от впечатления (целое неотрицательное число), которое на него производит студент), если он допускает, что с вероятностью не большей 0.1 можно ошибиться и поставить не знающему студенту «зачёт».

Задача (оценка вероятности переобучения). Пусть $x_1, \ldots, x_l$ — простая выборка из распределения с функцией распределения $F(x)$. Элементами этой выборки x_i могут быть, например, векторы. Пусть $\alpha \in \Omega$ — некоторый абстрактный параметр, $0 \leqslant R(x, \alpha) \leqslant a$ — некоторая функция, измеримая при всех $\alpha \in \Omega$ относительно меры $F(x)$. Далее

$$M(\alpha) = MR(x, \alpha) = \int R(x, \alpha)\, dF(x), \quad M_{\text{эмп}}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} R(x_i, \alpha).$$

Рассмотрим систему событий S вида $A(\alpha, c) = \{x : R(x, \alpha) \geqslant c\}$ для всевозможных значений $\alpha \in \Omega$ и c.

Обозначим через $\Delta^S(x_1, \ldots, x_l)$ число (бинарных) решающих правил класса S, по-разному классифицирующих объекты заданной выборки. Выборке $x_1, \ldots, x_l$ и конкретному $A(\alpha, c)$ ставится в соответствие последовательность нулей и единиц по правилу: $x_i \in A(\alpha, c) \Rightarrow 1$, $x_i \notin A(\alpha, c) \Rightarrow 0$.

Разным $A(\alpha, c)$ могут соответствовать как разные, так и одинаковые последовательности нулей и единиц. Очевидно, что $\Delta^S(x_1, \ldots, x_l)$ есть число различных последовательностей нулей и единиц, построенных по семейству $\{A(\alpha, c)\}_{\alpha\in\Omega,\, c}$. Очевидно также, что $\Delta^S(x_1, \ldots, x_l) \leqslant 2^l$.

Введём функцию роста $M^S(l) = \max \Delta^S(x_1, \ldots, x_l)$, где максимум берётся по всем последовательностям $(x_1, \ldots, x_l)$ длины l. Покажите, что

$$P\left\{\sup_{\alpha\in\Omega} |M(\alpha) - M_{\text{эмп}}(\alpha)| > \varepsilon\right\} \leqslant 6M^S(2l)\exp\left[-\frac{\varepsilon^2(l-1)}{4a^2}\right].$$

Замечание. Заметим, что для любой системы событий S имеет место:

$$M^S(l) = 2^l \text{ или } M^S(l) \leqslant \sum_{i=0}^{n-1} C_l^i,$$

т. е. $M^S(l) = O(l^{n_0})$ для некоторого $n_0 \in \mathbb{N}$.

Минимально возможное значение n_0 принято называть размерностью Вапника – Червоненкиса (VC-размерность). Однако А. Я. Червоненкис предлагает (в учебном пособии [34]) называть её комбинаторной размерностью S. Так, например, для множества всевозможных линейных решающих правил в пространстве размерности n комбинаторная размерность равна $n_0 = n+1$. Если $M^S(l) = 2^l$, то говорят, что комбинаторная размерность бесконечна. Для рассматриваемого в задаче случая достаточным условием конечности комбинаторной размерности, как следствие равномерной сходимости с ростом объёма выборки $M_{\text{эмп}}(\alpha)$ к $M(\alpha)$, является условие, что Ω — компакт, $R(x, \alpha)$ непрерывна по α, $|R(x, \alpha)| < K(x)$, где $\int K(x)\, dx < \infty$.

* * * * * *

Для более глубоко знакомства с методами математической статистики можно рекомендовать [7, 15, 22].

Список литературы

[1] Алон Н., Спенсер Дж. *Вероятностный метод*. М.: Бином, 2006.

[2] Арнольд В.И. *Математическое понимание природы*. М.: МЦНМО, 2011.

[3] Арнольд В.И. *Цепные дроби*. М.: МЦНМО, 2001.

[4] Афанасьев В.И. *Случайные блуждания и ветвящиеся процессы*. Лекционные курсы НОЦ, вып. 6. М.: МИАН, 2008. `http://www.mi.ras.ru/index.php?c=lectures`

[5] Афанасьева Л.Г. *Очерки исследования операций*. М.: Изд-во механико-математического факультета МГУ, 2007.

[6] Богданов К.Ю. *Прогулки с физикой*. Б-ка «Квант», вып. 98. М.: Бюро Квантум, 2006; глава 18.

[7] Боровков А.А. *Математическая статистика*. М.: ФИЗМАТЛИТ, 2007.

[8] Ватутин В.А. *Ветвящиеся процессы и их применение*. М.: МИАН, Лекционные курсы НОЦ, вып. 8, 2008.

[9] Вершик А.М. *Асимптотическое распределение разложений натуральных чисел на простые делители* // ДАН, 1986. Т. 289, №2. С. 269–272.

[10] Вершик А.М. *А что будет, если n очень большое*? Лекция 1. Дубна, 2008.
`http://www.mathnet.ru/PresentFiles/231/v231.pdf`,
`http://www.mathnet.ru/php/presentation.phtml?presentid=231`

[11] Вершик А.М., Шмидт А.А. *Предельные меры, возникающие в асимптотической теории симметрических групп* // ТВП, Т. 22. №1. 1977.С. 72-88; Т. 23. №1. 1978. С. 42-54.

[12] Гнеденко Б.В. *Курс теории вероятностей*. М.: Едиториал УРСС, 2005.

[13] Гусейн-Заде С.М. *Разборчивая невеста*. М.: МЦНМО, Б-ка «Математическое просвещение», вып. 25, 2003.

[14] Зорич А.В. *Математический анализ задач естествознания*. М.: МЦНМО, 2008. С. 48-56.

[15] Ивченко Г.И., Медведев Ю.И. *Введение в математическую статистику*. М.: Издательство ЛКИ, 2010.

[16] Кац М. *Вероятность и смежные вопросы в физике*. М.: Мир, 1965.

[17] Кац М. *Статистическая независимость в теории вероятностей, анализе и теории чисел*. М.: ИЛ, 1963.

[18] Кельберт М. Я., Сухов Ю. М. *Вероятность и статистика в примерах и задачах*. Т. 2. М.: МЦНМО, 2010.

[19] Кендалл М., Моран П. *Геометрические вероятности*. М.: Наука, 1972.

[20] Кормен Т., Лейзерсон Ч., Ривест Р., Штайн Ш. *Алгоритмы. Построение и анализ*. М.: Изд-во Вильямс, 2005.

[21] Кузюрин Н.Н., Фомин С.А. *Эффективные алгоритмы и сложность вычислений*. М.: МФТИ, 2007.

[22] Лагутин М.Б. *Наглядная математическая статистика*. М.: Бином, 2009.

[23] Магарил-Ильяев М.А., Тихомиров В.М. *Выпуклый анализ*. М.: Едиториал УРСС, 2011.

[24] Малышев В.А. *Кратчайшее введение в современные вероятностные модели*. М.: Изд-во мехмата МГУ, 2009.
`http://mech.math.msu.su/~malyshev/Malyshev/Lectures/course.pdf`

[25] А.Ю. Окуньков. *Легко ли заблудиться в группе*? Дубна, 2010.
`http://www.mathnet.ru/php/presentation.phtml?presentid=3491`

[26] Опойцев В.И. *Устойчивые системы большой размерности* // АиТ, №6. 1986. С. 43-49.

[27] Разборов А.А. *Теория сложности вычислений*. Лекция 3. Дубна, 2011.
`http://www.mathnet.ru/php/presentation.phtml?presentid=3645`

[28] Райгородский А.М. *Вероятность и алгебра в комбинаторике*. М.: МЦНМО, 2008.

[29] Секей Г. *Парадоксы в теории вероятностей и математической статистике*. Москва – Ижевск: РХД, 2002.

[30] Синай Я.Г. *Основы эргодической теории*. М.: ФАЗИС, 1996.

[31] Сосинский А.Б. *Мыльные плёнки и случайные блуждания*. М.: МЦНМО, Б-ка «Математическое просвещение», вып. 6, 2000.

[32] Федорюк М.В. *Метод перевала*. М.: УРСС, 2010.

[33] Хинчин А.Я. *Цепные дроби*. М.: УРСС, 2004.

[34] Червоненкис А.Я. *Компьютерный анализ данных*. М.: Яндекс, 2009.

[35] Шень А. *Вероятностные доказательства* // Квант, №6. 2009. С. 11–15.

[36] Шень А. *Вероятность: примеры и задачи*. М.: МЦНМО, 2008.
`http://www.mccme.ru/free-books/shen/shen-probability.pdf`

[37] Эфрос А.Л. *Физика и геометрия беспорядка*. Б-ка «Квант», вып. 19. М.: Наука, 1982.

[38] `http://dame.mipt.ru/studyandscience/stohanaliz.html`

[39] http://erb-files.narod.ru/#GLOBUS

[40] http://www.mou.mipt.ru/natan.html

[41] Annual IEEE Symposium on Foundations of Computer Science, 1 – 51. Например, http://theory.stanford.edu/focs2010/

[42] Diaconis P. *The Markov chain Monte Carlo revolution* // Bulletin (New Series) of the AMS. 2009. V. 49. №2. P. 179-205.
http://www-stat.stanford.edu/~cgates/PERSI/papers/MCMCRev.pdf

[43] Dubhashi D.P., Panconesi A. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, 2009.

[44] Durett R. *Probability*: *Theory and examples*. Pacific Grove CA: Wadsworth, 1991.
http://www.math.duke.edu/~rtd/PTE/PTE4_Jan2010.pdf

[45] Flajolet P., Sedgewick R. *Analytic combinatorics*. Cambridge University Press, 2008. http://algo.inria.fr/flajolet/Publications/book.pdf

[46] Lai T., Robbins H. *Asymptotically efficient adaptive allocation rules* // Advances in Applied Mathematics. V. 6. 1985. P. 41–22.

[47] Ledoux M. *Concentration of measure phenomenon*. Providence, RI, Amer. Math. Soc., 2001 (Math. Surveys Monogr. V. 89).

[48] Montenegro R., Tetali P. *Mathematical aspects of mixing times in Markov chains.* 2006.
http://people.math.gatech.edu/~tetali/PUBLIS/survey.pdf

[49] Motwani R., Raghavan P. *Randomized algorithms*. Cambridge Univ. Press, 1995.

А. Гасников, доцент МФТИ
Email: gasnikov@yandex.ru
Е. Черноусова, асс. МФТИ
Т. Нагапетян, асп. МФТИ
О. Федько, доцент МФТИ