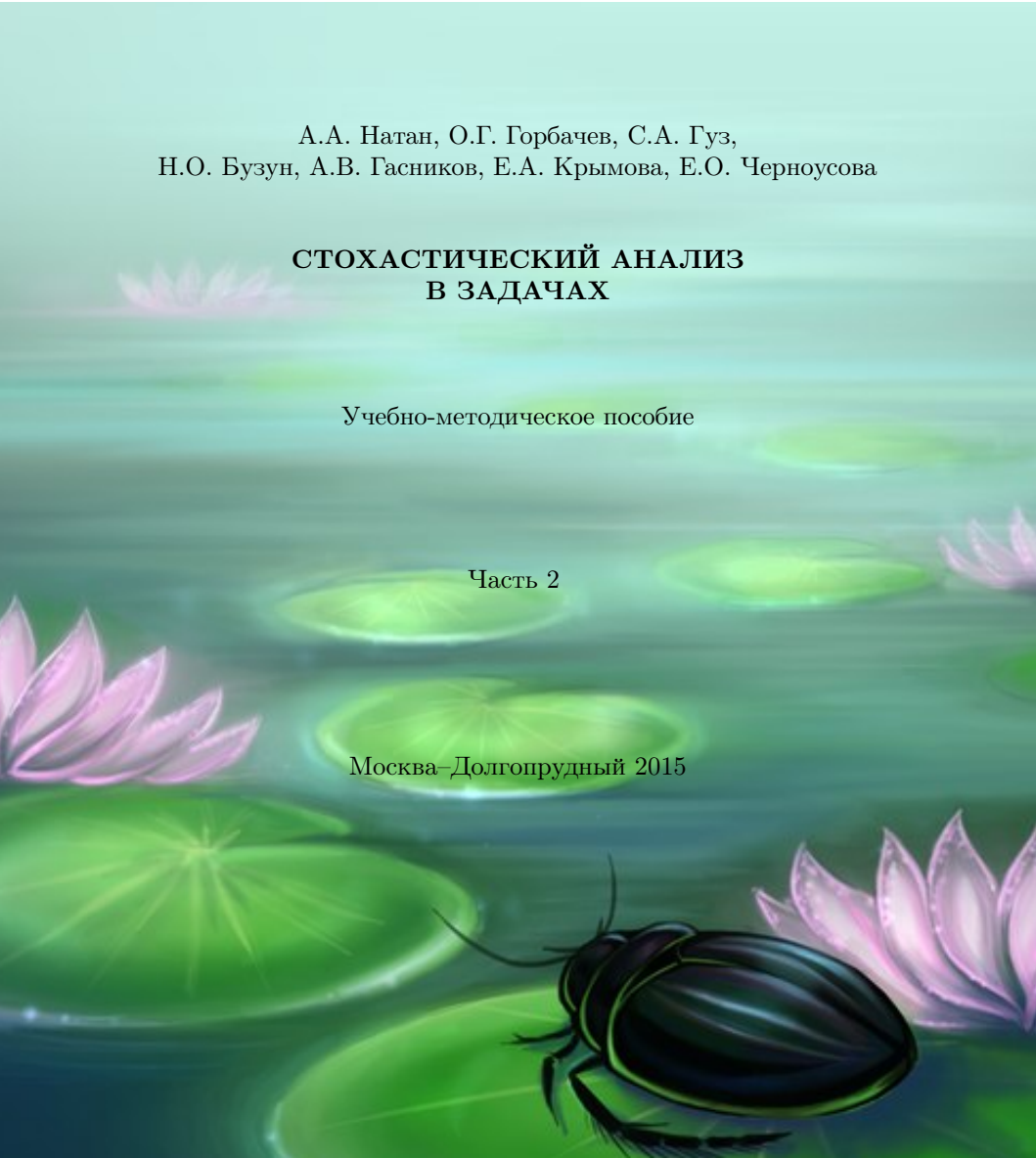
МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Московский физико-технический институт
(государственный университет)



А.А. Натан, О.Г. Горбачев, С.А. Гуз,
Н.О. Бузун, А.В. Гасников, Е.А. Крымова, Е.О. Черноусова

СТОХАСТИЧЕСКИЙ АНАЛИЗ В ЗАДАЧАХ

Учебно-методическое пособие

Часть 2

Москва–Долгопрудный 2015

УДК 519.21

А.А. Натан, О.Г. Горбачев, С.А. Гуз, Н.О. Бузун, А.В. Гасников, Е.А. Крымова, Е.О. Черноусова. Стохастический анализ в задачах: Учебно-методическое пособие. Часть 1 / МФТИ. М.–Д., 2015, изд. 3-е, доп.

Аннотация

Содержит программу, список литературы и задачи одноименного курса, читаемого сотрудниками кафедры Математических основ управления студентам факультета управления и прикладной математики Московского физико-технического института. Задачи могут быть использованы в качестве упражнений на семинарских занятиях, сдачах заданий, экзаменах, а также при самостоятельном освоении курса. В новое издание добавлен ряд задач, отражающих также опыт преподавания вероятностных дисциплин в Независимом московском университете и опыт наших коллег из ПреМоЛаб МФТИ, преподающих вероятностные дисциплины в ведущих научных центрах Германии и Франции. Данный сборник задач разрабатывался с учетом направлений подготовки: студентов МФТИ, студентов НМУ, магистров факультета Компьютерных наук ВШЭ и магистров факультета Прикладной математики БФУ им. И. Канта.

рецензенты: д.ф.-м.н. А.В. Колесников
проф. Факультета математики НИУ ВШЭ
д.ф.-м.н. А.Н. Соболевский
проф. Независимого московского университета,
зам. директора ИППИ РАН

ОБЯЗАТЕЛЬНАЯ ЧАСТЬ ПРОГРАММЫ УЧЕБНОГО КУРСА «Теория вероятностей»

Интуитивные предпосылки теории вероятностей. Множество элементарных исходов опыта, событие. Классическое и статистическое определение вероятности. Математическое определение вероятности. Алгебра и сигма-алгебра событий, минимальная сигма алгебра. Аксиомы теории вероятностей и следствия из них. Вероятностное пространство.

Теорема непрерывности вероятности. Теорема сложения вероятностей. Зависимые и независимые события. Условная вероятность события. Формула полной вероятности. Формула Байеса. Леммы Бореля–Кантелли. Закон "0-1" Колмогорова.

Случайная величина как измеримая функция. Функция распределения случайной величины. Дискретные и непрерывные случайные величины. Плотность распределения вероятностей. Формула включений-исключений.

Конкретные распределения случайных величин. Схема Бернулли, геометрическое и биномиальное распределение. Простейший поток событий и распределение Пуассона. Показательное, равномерное, нормальное, log-нормальное и отрицательно-биномиальное распределения. Бета-распределение и гамма-распределение.

Случайный вектор. Функция распределения случайного вектора. Зависимые и независимые случайные величины, условные законы распределения. Функции случайных величин. Невырожденное функциональное преобразование случайного вектора. Интеграл Стилтьеса. Математическое ожидание и дисперсия случайной величины. Моменты случайной величины. Условное математическое ожидание. Корреляционная матрица случайного вектора. Коэффициент корреляции двух случайных величин.

Характеристическая функция и ее свойства. Связь моментов случайной величины с ее характеристической функцией. Разложение характеристической функции в ряд. Сходимость последовательностей случайных величин с вероятностью единица (почти наверное), порядка p (в среднем квадратичном), по вероятности, по распределению. Соотношение между различными типами сходимости.

Неравенство Чебышева. Закон больших чисел. Критерий Колмогорова. Теоремы Хинчина и Чебышева. Усиленный закон больших чисел. Теорема Колмогорова и Бореля. Оценивание скорости сходимости частоты к вероятности в схеме Бернулли. Неравенство Бернштейна.

Интегральная и локальная теоремы Муавра–Лапласа. Дискретная поправка. Теорема Линдберга. Центральная предельная теорема для одинаково распределенных случайных величин. Центральная предельная теорема в форме Натана. Условие Ляпунова. Теорема Гливенко.

1. Введение

В основу предлагаемого сборника задач по теории вероятностей положены задачи (в том числе повышенной сложности), предлагавшиеся в разные годы студентам факультета управления и прикладной математики (ФУПМ) МФТИ и Независимого московского университета на семинарах, сдачах заданий и экзаменах. Главными отличительными особенностями пособия являются: а) широкий спектр представленного материала, б) отражение ряда современных направлений развития теории вероятностей и в) нацеленность на приложения.

По замыслу авторов, предлагаемый сборник задач отчасти демонстрирует роль курса как "вероятностного фундамента" для ряда других дисциплин: прикладной статистики, стохастических дифференциальных уравнений, эффективных алгоритмов, экспериментальной экономики (финансовой математики) и др. Несмотря на широкий спектр представленных тем, основной акцент делается на формирование у читателей геометрической интуиции, восходящей к Пуанкаре, которая позволяет с единых позиций понять многообразие асимптотических результатов стохастической теории как проявление одного общего принципа концентрации меры.

Большое внимание в сборнике уделяется различным приемам доказательства предельных теорем – асимптотических результатов теории вероятностей. Для этого прежде всего используется аппарат производящих функций и теории функций комплексного переменного.

Более половины представленных в сборнике задач не являются стандартными. Для таких задач даны указания, комментарии (замечания), ссылки на публикации.

Желающих более глубоко изучить представленные темы отсылаем к списку литературы, приведенному в конце сборника задач.

Следует отметить, что за последние десять лет заметно возросло значение для выпускников Физтеха глубоких знаний вероятностных дисциплин. Это обусловлено множеством причин.

Прежде всего, это вызвано широким распространением задач анализа больших массивов данных (machine learning, data mining). Подтверждением служит взаимная востребованность студентов ФУПМ и Школы анализа данных компании Яндекс, большая популярность среди студентов ФУПМ кафедры интеллектуальных

систем (заведующий кафедрой член-корреспондент РАН К.В. Рудаков, Вычислительный центр РАН) и кафедры предсказательного моделирования (заведующий кафедрой академик РАН А.П. Кулешов, Институт проблем передачи информации РАН).

Другая не менее важная причина – разработка эффективных (приближенных, рандомизированных) алгоритмов решения сложных задач. Сложно, например, представить себе современного специалиста по моделированию, который бы не использовал методы Монте-Карло. Также сложно представить себе специалиста в области computer science, которому не приходилось бы применять рандомизированные алгоритмы и подвергать алгоритмы вероятностному анализу (например, для оценки сложности в среднем).

Еще одна причина связана с тем, что приложение вероятностных методов к анализу и разработке экономических моделей для части студентов ФУПМ является фундаментом в работе на базовых кафедрах. Например, заведующие базовыми кафедрами ФУПМ член-коррерспондент РАН И.Г. Поспелов (Вычислительный центр РАН) и член-коррерспондент РАН Ю.С. Попков (Институт системного анализа РАН) активно работают в направлении разработки вероятностных моделей экономических агентов и вероятностного анализа агломерационных моделей.

Наконец, можно заметить, что задачи анализа больших компьютерных, социальных, транспортных сетей в последнее время выходят на передний план во многих приложениях. Огромную роль в изучении таких сетей играют вероятностные модели, некоторые из которых будут приведены в предлагаемом сборнике задач.

Важную роль в подготовке настоящего сборника задач сыграла Лаборатория структурных методов анализа данных в предсказательном моделировании (ПреМоЛаб), открытая на базе ФУПМ МФТИ в 2011 году. В частности, благодаря этой лаборатории, у студентов есть возможность посмотреть на сайте www.mathnet.ru, www.premolab.ru (семинары: Стохастический анализ в задачах, Математический кружок и др.) видеозаписи выступлений ведущих ученых, посвященные ряду нестандартных задач из этого сборника.

Мы благодарны нашим коллегам Д.В. Беломестному, Я.И. Белопольской, М.Л. Бланку, Н.Д. Введенской, В.В. Веденяпину, Д.П. Ветрову, А.М. Вершику, К.В. Воронцову, В.В. Вьюгину, В.В. Высоцкому, А.В. Калинкину, М.Н. Вялому, М.С. Гельфанду, Э.Х. Гимади, Г.К. Голубеву, А.Б. Дайняку, П.Е. Двуреченскому, Н.Х. Ибрагимову,

М.И. Исаеву, Г.А. Кабатянскому, А.В. Колесникову, А.В. Леонизову, Г. Лугоши, Ю.В. Максимову, В.А. Малышеву, В.Д. Мильману, В.В. Моттлю, Т.А. Нагапетяну, А.В. Назину, Ю.Е. Нестерову, А.С. Немировскому, В.И. Опойцеву, Ф.В. Петрову, С.А. Пирогову, Б.Т. Поляку, И.Г. Поспелову, А.М. Райгородскому, В.Н. Разжевайкину, М.А. Раскину, В.Г. Редько, А.Е. Ромащенко, А.В. Савватееву, А.Н. Соболевскому, А. Содину, В.Г. Спокойному, Й. Стоянову, У. Сэндхольму, С.П. Тарасову, И.О. Толстихину, М.Ю. Хачаю, О.С. Федько, Ю.А. Флёрову, А.Х. Шеню, оказавшим заметное влияние на формирование различных разделов этого учебного пособия, и А.А. Шананину, во многом способствовавшему развитию на ФУПМ базового цикла вероятностных дисциплин. Также отметим большую помощь старшекурсников, аспирантов ФУПМ и участников нашего стохастичексого семинара в НМУ и Физтехе в вычитке этого сборника задач (в особенности, Д. Бабичева, А. Балицкого, Ф. Гончарова, Ю. Дорна, Е. Клочкова, А. Макарова, Е. Молчанова, Н. Животовского, М. Панова, Л. Прохоренкову(Остроумову), А. Суворикову, Д. Петрашко, М. Широбокова).

Список обозначений

$\langle \Omega, \mathcal{F}, \mathbb{P} \rangle$ – вероятностное пространство (Ω – множество исходов, $\mathcal{F}$ – σ-алгебра, $\mathbb{P}$ – вероятностная мера);
$p(x),\ f(x),\ f_X(x)$ – плотность распределения случайной величины X; $\mathbb{E}X$ – математическое ожидание случайной величины X;
$\mathbb{E}_p X$ – математическое ожидание X с плотностью распределения p;
$\mathbb{E}_{\mathbb{P}} X$ – математическое ожидание X по мере $\mathbb{P}$;
$\mathbb{D}\, X$ – дисперсия случайной величины X;
$\mathcal{N}(m, \sigma^2)$ – нормальное распределение;
$\mathrm{Po}(\lambda)$ – распределение Пуассона;
$\mathrm{Dir}(\alpha_1, ..., \alpha_n)$ – распределение Дирихле;
$\mathrm{Beta}(\alpha, \beta)$ – бета-распределение;
$\mathrm{Be}(p)$ – распределение Бернулли;
$R[a, b], [a, b]$ – равномерное распределение;
$\Phi(x)$ – функция стандартного нормального распределения $\mathcal{N}(0, 1)$;
$\mathrm{Exp}(x)$ – показательное распределение;
$\xrightarrow{d}$ $\left(\xrightarrow[n\to\infty]{d}\right)$ – сходимость по распределению, в ряде случаев $n \to \infty$ опущено во избежание громоздких обозначений;
$\xrightarrow{p}$ – сходимость по вероятности;
$\xrightarrow{\text{п.н.}}$ – сходимость с вероятностью 1;
$[x^n]\varphi(x)$ – коэффициент при x^n в разложении в степенной ряд функции $\varphi(x)$;
с.в. – случайная величина;
з.б.ч. – закон больших чисел;
х.ф. – характеристическая функция;
ц.п.т. – центральная предельная теорема;
ПФ – производящая функция;
ЭПФ – экспоненциальная производящая функция;
$\langle \cdot, \cdot \rangle$ – скалярное произведение;
Индикаторная функция:

$$\mathrm{I}(\text{true}) = [\text{true}] = 1, \quad \mathrm{I}(\text{false}) = [\text{false}] = 0;$$

Простая выборка с плотностью распределения p:

$$X_1, \ldots, X_n \sim p(X);$$

В схожих ситуациях используется символ $\in$ вместо $\sim$ (например, $X \in \mathcal{N}(m, \sigma^2)$), если подразумевается принадлежность случайной

величины семейству распределений, $\sim$ также обозначает пропорциональность;
Ненормированная плотность распределения:

$$p(x) \propto g(x), \quad p(x) = \frac{g(x)}{\int g(x)dx};$$

Математическое ожидание по заданной переменной:

$$\mathbb{E}_X h(X,Y) = \int_{-\infty}^{+\infty} h(x,Y)dF(x);$$

Дивергенция Кульбака–Лейблера для распределений $\mathbb{P}_1$ и $\mathbb{P}_2$ с общим носителем Ω (соответствующие плотности распределений обозначены как p_1 и p_2):

$$\mathcal{KL}(\mathbb{P}_1 \| \mathbb{P}_2) = \int_\Omega \log\left(\frac{p_1(x)}{p_2(x)}\right) p_1(x)dx;$$

Дивергенция Кульбака–Лейблера для распределения $\mathbb{P}_\theta$, зависящего от параметра (соответствующая плотность распределения обозначена как $p(x|\theta)$):

$$\mathcal{KL}(\theta_1 \| \theta_2) = \mathcal{KL}(\mathbb{P}_{\theta_1} \| \mathbb{P}_{\theta_2});$$

$A \triangle B = A \setminus B + B \setminus A$ – симметрическая разность;
$\log = \ln$;
$\mathcal{B}$ – борелевская сигма алгебра;

Содержание

Часть 1

Часть 2

2. Байесовские методы

1. Охранная система представлена в виде Байесовской модели, изображенной на Рис. 1. Пусть t обозначает факт срабатывания

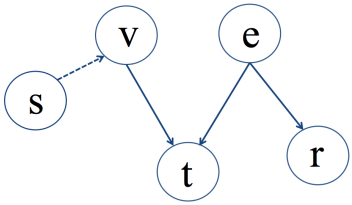


Рис. 1: Модель охранной системы.

тревоги, v – наличие вора $\in \mathrm{Be}(s)$, e – было ли землетрясение, r – наличие радиосообщения. Величины t, v, e, r имеют распределение Бернулли. Переменная s – статистика по криминогенной активности – принимает значения из отрезка $[0, 1]$. Значения $\mathbb{P}(t = 1|v, e)$, $\mathbb{P}(r = 1|e)$, $\mathbb{P}(e = 1)$ и $\mathbb{P}(v = 1)$ заданы, причем $\mathbb{P}(t = 1|v = 0, e = 0) = 0$, $\mathbb{P}(r = 1|e = 0) = 0$. Проведите расчет $\mathbb{P}(v = 1|t = 1, s = s_0)$ и $\mathbb{P}(v = 1|t = 1, s = s_0, r = 1)$.

Замечание. Генеративные вероятностные модели часто представляются в виде интуитивно понятных *графических моделей*. Графическая модель – это граф, определяющий зависимости между случайными величинами. Вершинам соответствуют случайные величины, ребрам – зависимости между ними.

Наблюдаемые величины (значения которых известны) закрашиваются, скрытые (значения которых надо найти) – остаются незакрашенными. Стрелка из вершины A в B обозначает зависимость B от A (часто под такого рода зависимостью подразумевается, что A является параметрами распределения с.в. B).

Для обозначения повторяющихся величин с одинаковым распределением используются прямоугольники, которые могут быть вложенными. В одном из углов прямоугольника обычно указывается количество повторений случайных величин, расположенных внутри него.

2. Радиоактивный источник излучает за одну секунду $n \in \mathrm{Po}(s)$ частиц, где s – неизвестная интенсивность излучения. Прибор, ре-

гистрирующий частицы, имеет погрешность $\theta = 0.9$, т.е. количество зарегистрированных частиц за одну секунду $c \in \mathrm{Bin}(\theta, n)$. Покажите, что

$$\mathbb{P}(c|\theta, s) = \mathrm{Po}(s\theta),$$

$$\mathbb{P}(n - c|c, \theta, s) = \mathrm{Po}(s(1 - \theta)).$$

Предположим, что за первую секунду прибор зарегистрировал $c_1 = 10$ частиц, а за вторую – $c_2 = 16$ частиц (см. Рис. 2). Докажите формулы для условного распределения и математического ожидания n_1:

$$\mathbb{P}(n_1|c_1, c_2, \theta, s) = \int_0^\infty p(n_1|c_1, \theta, s)p(s|c_1, c_2, \theta)ds =$$

$$= C_{n_1+c_2}^{c_1+c_2} \left(\frac{2\theta}{1+\theta}\right)^{c_1+c_2+1} \left(\frac{1-\theta}{1+\theta}\right)^{n_1-c_1},$$

$$\mathbb{E}(n_1|c_1, c_2, \theta, s) = \frac{c_1 + c_2 + 1 - \theta}{2\theta} + \frac{c_1 - c_2}{2} = 131.5.$$

Найдите 90% доверительный интервал для $n_1|c_1, c_2, \theta, s$, воспользовавшись соотношением

$$\mathbb{P}(|X - \mathbb{E}X| > t\sqrt{\mathbb{D}\,X}) \leq \frac{1}{t^2}.$$

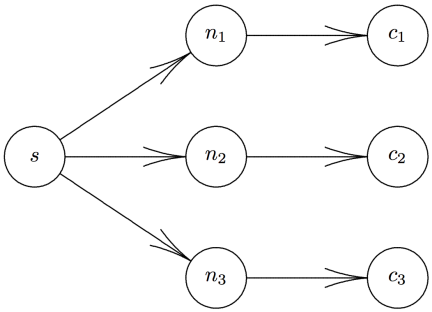


Рис. 2: Модель радиоактивного источника.

3 (Распределение Дирихле)**.** Ознакомьтесь с распределениями Дирихле, бета и гамма, приведенными в замечании. Покажите, что если $(X_1, \ldots, X_k)$, $X_i \in \Gamma(\gamma_i, 1)$ – гамма распределенные независимые с.в. и $S = \sum\limits_{i=1}^{k} X_i$, то

$$(X_1/S, \ldots, X_k/S) \in \mathrm{Dir}(\gamma_1, \ldots, \gamma_k). \tag{1}$$

Предложите способ генерации вектора с распределением Дирихле ($\gamma_i \in \mathbb{N}$), имея в распоряжении генератор с.в. с равномерным распределением на отрезке $[0, 1]$.

Покажите, что каждая компонента вектора с распределением Дирихле имеет бета-распределение $X_i/S \in \mathrm{Beta}(\gamma_i, \sum\limits_{j \neq i} \gamma_j)$.

Положим, что вектор $(\theta_1, \ldots, \theta_k)$ априорно имеет распределение (1), выборка $Y_1, \ldots, Y_n$ сгенерирована из дискретного распределения $\mathrm{Cat}(\theta_1, \ldots, \theta_k)$, т.е. $\mathbb{P}(Y_i = m | \theta_1, \ldots, \theta_k) = \theta_m$. Докажите справедливость формулы для плотности апостериорного распределения $\theta_1, \ldots, \theta_k$:

$$p(\theta_1, \ldots, \theta_k | Y_1, \ldots, Y_n, \gamma_1, \ldots, \gamma_k) =$$
$$= \mathrm{Dir}\left(\gamma_1 + \sum_{i=1}^{n} \delta_1(Y_i), \ldots, \gamma_k + \sum_{i=1}^{n} \delta_k(Y_i)\right),$$

где $\delta_k(x) = [x = k]$. Убедитесь, что данная функция принимает максимальное значение при

$$\theta_m \propto \gamma_m - 1 + \sum_{i=1}^{n} \delta_m(Y_i). \tag{2}$$

Замечание. Говорят, что вектор X принадлежит распределению *Дирихле* $\mathrm{Dir}(\gamma_1, \ldots, \gamma_k)$, если:

$$f_X(x_1, \ldots, x_k) = \frac{\Gamma(\gamma_1 + \ldots + \gamma_k)}{\Gamma(\gamma_1) \ldots \Gamma(\gamma_k)} x_1^{\gamma_1 - 1} \ldots x_k^{\gamma_k - 1},$$

где $(x_1, \ldots, x_k) \in \{x_i \geq 0,\ \sum x_i = 1\}$, $\gamma_i \geq 0$, Γ – гамма функция.

Говорят, что вектор X принадлежит *бета* распределению $\mathrm{Beta}(\gamma_1, \gamma_2)$, если:

$$f_X(x) = \tfrac{\Gamma(\gamma_1 + \gamma_2)}{\Gamma(\gamma_1)\Gamma(\gamma_2)} x^{\gamma_1 - 1}(1 - x)^{\gamma_2 - 1}, \text{ где } x \in [0, 1].$$

Допускается также обозначение следующего вида $X \in \mathrm{Dir}(\gamma_1, \ldots, \gamma_{k+1})$, если

$$f_X(x_1, \ldots, x_k) = \frac{\Gamma(\gamma_1 + \ldots + \gamma_{k+1})}{\Gamma(\gamma_1) \ldots \Gamma(\gamma_{k+1})} x_1^{\gamma_1 - 1} \ldots x_k^{\gamma_k - 1} (1 - x_1 - \ldots - x_k)^{\gamma_{k+1} - 1},$$

где $(x_1, \ldots, x_k) \in \{x_i \geq 0,\ \sum x_i \leq 1\}$.

Говорят, что с.в. X принадлежит *гамма* распределению $\Gamma(\alpha, \lambda)$, если:

$$f_X(x) = \frac{x^{\alpha - 1} e^{-x/\lambda}}{\lambda^\alpha \Gamma(\alpha)}, \text{ где } x \geq 0.$$

Распределение Дирихле $\mathrm{Dir}(\gamma_1, \ldots, \gamma_k)$ может быть интерпретировано следующим образом. Его функция плотности возвращает вероятность того, что вероятности $A_1, \ldots, A_K$ – несовместных событий равны соответственно $x_1, \ldots, x_K$, $\sum_i x_i = 1$ при условии, что i-е событие наблюдалось $\gamma_i - 1$ раз.

Отметим два свойства распределения Дирихле, делающего его удобным для использования в качестве априорного, при оценке параметров дискретного распределения (см. выражение (2)). Первое, семейство распределений Дирихле является сравнительно большим среди всех возможных распределений на симплексе, что является своего рода страховкой от неправильного выбора типа априорного распределения. Второе, имея выборку, сгенерированную из дискретного распределения, апостериорное распределение допускает простое аналитическое вычисление.

Обратите внимание, что в выражении (2) при $\gamma_i < 1$, $i = \overline{1, k}$ априорное распределение способствует разреживанию (увеличению количества нулевых компонент) $\theta_1, \ldots, \theta_k$, при $\gamma_i = 1$, $i = \overline{1, k}$ априорное распределение является равномерным, а при $\gamma_i > 1$, $i = \overline{1, k}$ распределение способствует сглаживанию $\theta_1, \ldots, \theta_k$.

Распределение Дирихле приобрело широкое практическое применение в области тематического моделирования текстов (см. задачу 14), машинного перевода, построения моделей в экологии, байесовской проверки гипотез, точечного оценивания и оценки доверительных интервалов.

4 (Упорядоченное распределение Дирихле)**.** Пусть $X_{(k)}$ является k-ой порядковой статистикой в наборе $X_1, \ldots, X_n$ (*) случайных величин с функцией распределения $F_X(x)$. Пусть $U_k = F_X(X_{(k)})$. Покажите, что $U \in \mathrm{Dir}^*(1, \ldots, 1)$ (см. замечание).

Сравните доказанное утверждение с задачей **??** раздела **??**. Покажите, что если $\mathbb{P}(X \in [X_{(i-1)}, X_{(i)}]) = W_i = U_i - U_{i-1}$, где $X_{(0)} = -\infty$, $X_{(n+1)} = +\infty$, то $W \in \mathrm{Dir}(1, \ldots, 1)$.

Имея в распоряжении набор векторов (*), где $X_i \in \mathbb{R}^m$, эмпирическая оценка плотности распределения в точке x может быть найдена следующим образом:

$$\widehat{f}_X(x) = \frac{\mathbb{E}U_k}{V_{r_k}},$$

где r_k – расстояние от x до $X_{(k)}$, V_{r_k} – объем шара радиуса r_k в пространстве $\mathbb{R}^m$. Докажите, что

$$U_k \in \mathrm{Beta}(k, n-k+1),$$

$$\widehat{f}_X(x) \xrightarrow{p} f_X(x), \quad n \to \infty.$$

Замечание. Говорят, что вектор Y принадлежит *упорядоченному распределению Дирихле* $\mathrm{Dir}^*(\gamma_1, \ldots, \gamma_{k+1})$, если:

$$f_Y(y_1, \ldots, y_k) =$$

$$\frac{\Gamma(\gamma_1 + \ldots + \gamma_{k+1})}{\Gamma(\gamma_1) \ldots \Gamma(\gamma_{k+1})} y_1^{\gamma_1 - 1} (y_2 - y_1)^{\gamma_2 - 1} \ldots (y_k - y_{k-1})^{\gamma_k - 1} (1 - y_k)^{\gamma_{k+1} - 1},$$

где $(y_1, \ldots, y_k) \in \{0 \le y_1 \le \ldots \le y_k \le 1\}$, $\gamma_i \ge 0$.

Причем, если $X \in \mathrm{Dir}(\gamma_1, \ldots, \gamma_{k+1})$ и $Y_i = \sum\limits_{j=1}^{i} X_j$, $i = \overline{1,k}$ то

$$Y \in \mathrm{Dir}^*(\gamma_1, \ldots, \gamma_{k+1}).$$

5. Распределение Пуассона–Дирихле получается из упорядоченного вектора с распределением Дирихле предельным переходом, описанным в замечании. Докажите, что величины $\theta_{(i)}$ с вероятностью 1 убывают с экспоненциальной скоростью:

$$\theta_{(i)} \propto e^{-i/\lambda}.$$

Такое распределение часто возникает при моделировании разделения большого числа элементов из S на большое число видов. Предположим, что из S, была произведена выборка размера n. Докажите, что вероятность того, что все элементы выборки имеют один и тот же вид равна

$$\frac{\lambda \Gamma(\lambda) \Gamma(n)}{\Gamma(\lambda + n)}.$$

Замечание. Пусть $\theta_1, \ldots, \theta_n \in \mathrm{Dir}(\gamma_1, \ldots, \gamma_n)$, при этом выполнены условия

$$\max_i \gamma_i \to 0, \quad n \to \infty,$$

$$\lambda_n = \sum_{i=1}^{n} \gamma_i \to \lambda.$$

Введем обозначения

$$s_0 = 0, \quad s_j = \gamma_1 + \ldots + \gamma_j, \quad j = \overline{1, n}.$$

Распределение Дирихле может быть сконструировано из $Y_i \in \Gamma(\gamma_i, 1)$ по правилу $\theta_i = Y_i / \sum_j Y_j$. Известно, что Y_i безгранично делимы, т.е. могут быть разбиты на сколь угодно большое число гамма распределенных случайных величин $Y_{i1}, \ldots, Y_{ip}$. Частичные суммы величин $\sum_j Y_j$ приближаются в пределе к случайному процессу с независимыми приращениями g:

$$g(s_m) = \sum_{i=1}^{m} Y_i, \quad \theta_j = \frac{g(s_j) - g(s_{j-1})}{g(s_n)}.$$

Из вида характеристической функции

$$\phi_{g(s)}(\mu) = \exp\left(-s \int_0^\infty \left(1 - e^{-\mu z}\right) \frac{e^{-z}}{z} dz\right)$$

$$\gamma(dz) = \frac{e^{-z}}{z} dz$$

заключаем, что $g(s)$ является субординатором на множестве $s \in [0, \lambda]$ (см. задачу 12 из раздела 6). Величины скачков процесса $g(s)$ образуют пуассоновский процесс $J(t)$ на множестве $t \in (0, \infty)$ с переменной интенсивностью

$$\lambda(z) = \lambda \frac{e^{-z}}{z}.$$

Обозначим за $J_{(1)} \geq J_{(2)} \geq J_{(3)} \geq \ldots$ упорядоченные величины скачков. Последовательность $\theta_{(1)} \geq \theta_{(2)} \geq \theta_{(3)} \geq \ldots$, где

$$\theta_{(i)} = \frac{J_{(i)}}{g(\lambda)},$$

имеет распределение *Пуассона–Дирихле.*

6 (Сопряженные распределения)**.** Сформулируем задачу байесовского вывода. Пусть известно распределение $p(X|\theta)$ (*правдоподобие выборки* X). Требуется найти:

а) $p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}$ – апостериорное распределение;

б) $p(x'|X) = \int_\Theta p(x'|\theta)p(\theta|X)d\theta$ – предсказание нового x'.

Ключевым моментом является выбор наиболее разумного $p(\theta)$ – *априорного распределения*. Выбор осуществляется, исходя из дополнительной информации о параметре θ и необходимости введения регуляризации (отдание предпочтения более простым функциям $p(X|\theta)$, которые могут быть построены более точно при ограниченном размере выборки $|X|$).

Решение задач Байесовского вывода существенно упрощается при использовании *сопряженных семейств* распределений (см. М. Де Гроот. Оптимальные статистические решения. – М.: Мир, 1974). Семейство распределений $\{p(\theta|\alpha),\ \alpha \in \Omega\}$ называется *сопряженным* к семейству правдоподобий $\{p(X|\theta),\ \theta \in \Theta\}$, если $\forall\alpha\ \exists\alpha'$ такой, что апостериорное распределение $p(\theta|X,\alpha) = p(\theta|\alpha')$ (т.е. апостериорное распределение имеет такой же вид, что и априорное, но с другими параметрами). Именно сопряженное распределение разумно взять в качестве априорного, что, в частности, упростит вычисление нормировочного интеграла:

$$p(X) = \int\limits_\theta p(X|\theta)p(\theta)d\theta.$$

Рассмотрим несколько примеров сопряженных семейств. Покажите, что:

а) Сопряженным к распределению Бернулли является *бета* распределение (см. замечание к задаче 3). Если $x_1, \ldots, x_n$ – реализация выборки из распределения Бернулли, априорное распределение есть $\mathrm{Beta}(\alpha, \beta)$, то апостериорное распределение будет иметь параметры:

$$\alpha + \sum_{i=1}^{n} x_i,\quad \beta + n - \sum_{i=1}^{n} x_i.$$

б) Пусть $x_1, \ldots, x_n$ – реализация одного элемента выборки из мультиномиального распределения с неизвестным вектором параметров θ длины n. Допустим, что априорное распределение θ есть распределение Дирихле (см. замечание к задаче 3) $\mathrm{Dir}(\alpha_1, \ldots, \alpha_n)$. Тогда апостериорное распределение есть $\mathrm{Dir}(\alpha_1 + x_1, \ldots, \alpha_n + x_n)$.

в) Если $x_1, \ldots, x_n$ – реализация выборки из распределения Пуассона с неизвестным значением среднего θ, априорное распределение есть гамма-распределение $\Gamma(\alpha, \beta)$, то апостериорное распределение будет иметь параметры:

$$\alpha + \sum_{i=1}^{n} x_i, \quad \beta + n.$$

г) Если $x_1, \ldots, x_n$ – реализация выборки из экспоненциального распределения с неизвестным параметром θ, априорное распределение есть гамма-распределение $\Gamma(\alpha, \beta)$, то апостериорное распределение будет иметь параметры:

$$\alpha + n, \quad \beta + \sum_{i=1}^{n} x_i.$$

д) Если $x_1, \ldots, x_n$ – реализация выборки из нормального распределения с неизвестным значением среднего m и заданной мерой точности $r = 1/\sigma^2$, априорное распределение m есть $N(\mu, 1/\tau)$, то апостериорное распределение есть $N(\mu', 1/\tau')$, где:

$$\mu' = \frac{\tau\mu + nr\overline{x}}{\tau + nr}, \quad \tau' = \tau + nr.$$

е) Если $x_1, \ldots, x_n$ – реализация выборки из нормального распределения с заданным значением среднего m и неизвестной мерой точности $r = 1/\sigma^2$. Пусть априорное распределение r есть гамма-распределение $\Gamma(\alpha, \beta)$. Тогда апостериорное распределение будет иметь параметры:

$$\alpha + \frac{n}{2}, \quad \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - m)^2.$$

ж) Если $x_1, \ldots, x_n$ – реализация выборки из нормального распределения с неизвестным значением среднего m и неизвестной мерой точности $r = 1/\sigma^2$ (см. Рис. 3). Пусть условное априорное распределение m при фиксированном r есть $N(\mu, 1/(\tau r))$; априорное распределение r есть гамма-распределение $\Gamma(\alpha, \beta)$. Тогда апостериорное совместное распределение m и r имеет следующий вид:

$$m|r \in N\left(\frac{\tau\mu + n\overline{x}}{\tau + n}, (\tau + n)r\right),$$

$$r \in \Gamma\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \overline{x})^2 + \frac{1}{2}\frac{\tau n(\overline{x} - \mu)^2}{\tau + n}\right),$$

$$p(m, r|X, \mu, \tau, \alpha, \beta) = \text{const} \cdot p(r|\alpha, \beta)p(m|r, \mu, \tau)p(X|m, r) =$$

$$p(m|r, X, \mu, \tau)p(r|X, \mu, \tau, \alpha, \beta).$$

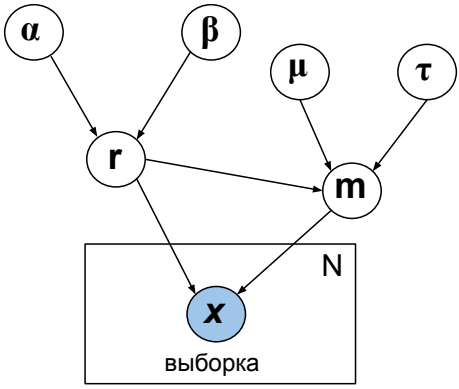


Рис. 3: Графическая модель генерации выборки из нормального распределения с неизвестными параметрами.

7 (Экспоненциальное семейство распределений)**.** Многие стандартные распределения принадлежат экспоненциальному семейству (см. замечание), например, нормальное, гамма, бета, Бернулли, Дирихле. Пусть с.в. $Y \in EF$. Докажите следующие свойства EF:

$$\varphi_{T(x)}(t) = \mathbb{E}e^{\langle t, T(x)\rangle} = \frac{Z(\theta + t)}{Z(\theta)};$$

$$\nabla \log Z(\theta) = \mathbb{E}T(x); \quad \nabla^2 \log Z(\theta) = \operatorname{cov}(T(x), T(x));$$

для выборки $X_1, \ldots, X_n \in p(x|\theta)$ оценка максимума правдоподобия может быть представлена как

$$\nabla \log Z(\widehat{\theta}_{ML}) = \frac{1}{n}\sum_{i=1}^{n} T(X_i);$$

$$\mathcal{KL}(\nu_2 \| \nu_1) = \nabla d(\nu_2)(\nu_2 - \nu_1) - [d(\nu_2) - d(\nu_1)];$$

сопряженное распределение (см. задачу 6) на параметры θ может быть вычислено по формуле

$$p(\theta|\alpha_1, \alpha_2) \propto e^{\langle \theta, \alpha_1\rangle - \alpha_2 \log Z(\theta)}.$$

Замечание. С.в. $X \in EF$, если ее плотность распределения имеет следующий вид (canonical parametrization):

$$p(x|\nu) = h(x)e^{\langle x, \nu\rangle - d(\nu)}.$$

Возможно также другое представление:

$$p(x|\theta) = \frac{h(x)}{Z(\theta)}e^{\langle \theta, T(x)\rangle},$$

где $T(x)$ – *достаточная статистика* распределения $p(x|\theta)$.

8 (Процесс Дирихле)**.** Рассмотрим графическую модель разделения смеси распределений (mixture model). Предполагается, что каждый элемент выборки $X = \{x_1, \ldots, x_n\}$ имеет скрытую компоненту из вектора $Z = \{z_1, \ldots, z_n\}$, обозначающую компоненту смеси, к которой относится элемент. Процесс генерации выборки устроен таким образом (см. Рис. 4): генерируется априорное дискретное распределение компонент смеси $(\pi_1, \ldots, \pi_K) \in \operatorname{Dir}(\alpha/K, \ldots, \alpha/K)$; в таком случае $\mathbb{P}(z_i = m) = \pi_m$; j-я ($j \in \overline{1, K}$) компонента смеси имеет параметры θ_j, которые сгенерированы из произвольного распределения $F(\lambda)$; x_i генерируется из распределения i-й компоненты с параметрами θ_i.

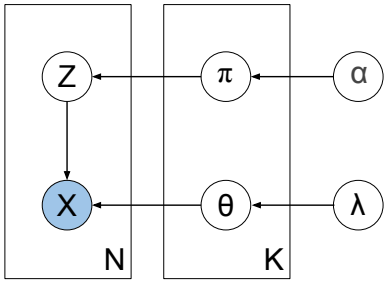


Рис. 4: Dirichlet Process Mixture Model.

Как правило, в таких задачах требуется оценить скрытые переменные Z, π, θ по известной выборке $X = \{x_1, \ldots, x_n\}$ и заданных значениях параметров α, λ.

В случае, когда K – число компонент дискретного распределения не задано, вместо распределения Дирихле для конечного K уместно воспользоваться процессом Дирихле DP, где $K = \infty$ (см. замечание).

Пусть $\theta_1, \ldots, \theta_{n+1}$ – независимые случайные величины, у которых распределение является случайной мерой $G \in \mathrm{DP}(H, \alpha)$, $A \subset \Theta$. Докажите, что

$$\mathbb{P}(\theta_{n+1} \in A | \theta_1, \ldots, \theta_n) = \frac{1}{n+\alpha}\left(\alpha H(A) + \sum_{i=1}^{n} \delta_{\theta_i}(A)\right),$$

Пусть m – число уникальных величин среди $\theta_1, \ldots, \theta_n$. Докажите следующие выражения:

$$\mathbb{E}(m|n) = \alpha(\psi(n+\alpha) - \psi(\alpha)) \simeq \alpha \ln\left(1 + \frac{n}{\alpha}\right) \quad \text{при } \alpha, n \gg 0,$$

$$\mathbb{D}(m|n) = \alpha(\psi(n+\alpha) - \psi(\alpha)) + \alpha^2(\psi'(n+\alpha) - \psi'(\alpha)) \simeq$$
$$\alpha \ln\left(1 + \frac{n}{\alpha}\right) \quad \text{при } n > \alpha \gg 0,$$

где $\psi(x) = \frac{\Gamma\text{'(x)}}{\Gamma\text{(x)}}$ – digamma функция.

Замечание. Пусть H – некоторое распределение с носителем Θ, α – положительное число. G, случайная вероятностная мера (распределение), называется *процессом Дирихле* $\mathrm{DP}(H, \alpha)$ с базовым распределением H и параметром концентрации α, если для любого конечного измеримого разбиения пространства элементарных исходов Θ: $A_1, \ldots, A_r$ выполнено

$$(G(A_1), \ldots, G(A_r)) \in \mathrm{Dir}(\alpha H(A_1), \ldots, \alpha H(A_r)).$$

Процесс Дирихле обладает следующими свойствами:

а) $\forall A \subset \Theta$: $\mathbb{E}\left[G(A)\right] = H(A)$;

б) $\forall A \subset \Theta$: $\mathbb{D}\left[G(A)\right] = \frac{H(A)(1-H(A))}{\alpha+1}$;

в) Пусть $\theta_1, \ldots, \theta_n$ – независимые случайные величины со случайным распределением $G \in \mathrm{DP}(H, \alpha)$, тогда

$$\mathbb{P}(G(A_1), \ldots, G(A_r)|\theta_1, \ldots, \theta_n) \in \mathrm{Dir}(\alpha H(A_1) + n_1, \ldots, \alpha H(A_r) + n_r),$$

где $n_j = |i : \theta_i \in A_j|$. Что также можно записать в другом виде:

$$G|\theta_1, \ldots, \theta_n \in \mathrm{DP}\left(\alpha + n, \frac{\alpha}{\alpha+n} H + \frac{n}{\alpha+n} \cdot \frac{\sum \delta_{\theta_i}}{n}\right),$$

где $\frac{\sum \delta_{\theta_i}}{n}$ – смесь распределений $\delta_{\theta_1}, \ldots, \delta_{\theta_n}$, в котором $\mathbb{P}(\theta = \theta_i) = 1$ при $\theta \in \delta_{\theta_i}$.

9 (Stick-breaking construction)**.** Рассмотрим альтернативное определение процесса Дирихле (см. замечание к задаче 8). Пусть J_k – k-й по счету скачек гамма процесса $g(s)$, $s \in [0, \alpha]$ (см. замечание к задаче 5), нормированное значение скачка обозначим как $\pi_k = J_k / g(\alpha)$. Введем также последовательность с.в. $v_k : \Theta \to \Theta$ с вероятностной мерой $H : \mathcal{B}(\Theta) \to [0, 1]$, где $\mathcal{B}(\Theta)$ – все измеримые подмножества Θ. Докажите, что случайная мера $G : \mathcal{B}(\Theta) \to [0, 1]$, такая что

$$\forall A \in \mathcal{B}(\Theta) : \; G(A) = \sum_{k=1}^{\infty} \pi_k \delta_{v_k}(A),$$

является процессом Дирихле $DP(H, \alpha)$.

Для программной реализации процесса Дирихле, как правило, используется генеративная модель “ломания палки”:

$$\beta_k \in \mathrm{Beta}(\alpha, 1), \quad \pi_k = \beta_k \prod_{l=1}^{k-1}(1-\beta_l), \quad \theta_k \in F(\lambda) = H,$$

$$G \in \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}.$$

Докажите, что $G \in \mathrm{DP}(H, \alpha)$.

Указание. См. статью T. Ferguson. A Bayesian analysis of some nonparametric problems. 1973.

10. Для процесса Дирихле $G \in DP(H, \alpha)$ (см. альтернативное определение в задаче 9) и измеримой функции $Z : \Theta \to \mathbb{R}$ докажите справедливость следующего утверждения. Если $\int_\Theta |Z(\theta)| dH(\theta) < \infty$, то с вероятностью единица $\int_\Theta |Z(\theta)| dG(\theta) < \infty$ и

$$\mathbb{E} \int_\Theta Z(\theta) dG(\theta) = \int_\Theta Z(\theta) d\mathbb{E}G(\theta) = \int_\Theta Z(\theta) dH(\theta).$$

11. Требуется по выборке $X_1, \ldots, X_n \in \mathbb{R}$ оценить функцию распределения $F_X(t)$. Априорно предполагается, что F_X соответствует вероятностной мере $H : \mathbb{R} \to [0, 1]$. Предлагается оценивать функцию F_X как

$$\widehat{F}_X(t) = \mathbb{E}G((-\infty, t)|X_1, \ldots, X_n),$$

где $G((-\infty, t)|X_1, \ldots, X_n)$ – с.в., соответствующая условной случайной мере $G|X_1, \ldots, X_n$ (см. замечание к задаче 8). Покажите, что

$$\widehat{F}_X(t) = \frac{\alpha}{\alpha + n} H((-\infty, t)) + \frac{n}{\alpha + n} \frac{1}{n} \sum_{i=1}^{n} [X_i < t].$$

Аналогично, для оценки $\mathbb{E}X$ предлагается использовать оценку

$$\widehat{m}_X = \mathbb{E} \int_{-\infty}^{+\infty} x dG(x).$$

Покажите, что

$$\widehat{m}_X = \frac{\alpha}{\alpha + n} \mathbb{E}_H X + \frac{n}{\alpha + n} \frac{1}{n} \sum_{i=1}^{n} X_i.$$

12. Требуется по выборкам $X_1, \ldots, X_n \in \mathbb{R}$ и $Y_1, \ldots, Y_m \in \mathbb{R}$ оценить вероятность

$$\triangle = \mathbb{P}(X_1 < Y_1) = \int_{-\infty}^{+\infty} F_X(t) dF_Y(t).$$

Априорно предполагается, что F_X соответствует вероятностной мере $H_X : \mathbb{R} \to [0,1]$, F_Y – вероятностной мере H_Y. Предлагается оценивать $\triangle$ как (см. задачу 11)

$$\widehat{\triangle} = \int_{-\infty}^{+\infty} \widehat{F}_X(t) d\widehat{F}_Y(t).$$

Покажите, что

$$\widehat{\triangle} = p_x p_y \triangle_0 + p_x(1-p_y)\frac{1}{m}\sum_{i=1}^{m} H_X(Y_i) + (1-p_x)p_y\frac{1}{n}\sum_{i=1}^{n} H_Y(X_i) +$$

$$+(1-p_x)(1-p_y)\frac{1}{mn}\sum_{i,j}[X_i < Y_i],$$

где

$$p_x = \frac{\alpha_x}{\alpha_x + n}, \quad p_y = \frac{\alpha_y}{\alpha_y + m}, \quad \triangle_0 = \int_{-\infty}^{+\infty} H_X(t) dH_Y(t).$$

13 (Вариационный вывод для DP)**.** Для поиска максимума функции правдоподобия $\log p(X|Z, \theta, \pi, \alpha, \lambda)$ в задаче 8 можно воспользоваться вариационным выводом (см. задачу 14 из раздела 5), введя более простую параметрическую модель распределения скрытых переменных $q(Z, \theta, \pi)$ и приближая $q(Z, \theta, \pi)$ к $p(Z, \theta, \pi|X, \alpha, \lambda)$ в смысле расстояния $\mathcal{KL}(q\|p)$.

Приведем генеративную модель из задачи 8, в которой в качестве процесса Дирихле используется Stick-breaking construction (см. задачу 9), а также конкретный пример распределения $F(\lambda) = \mathrm{Dir}(\lambda_1, \ldots, \lambda_M)$:

$$\beta_k \in \mathrm{Beta}(\alpha, 1),\ \pi_k = \beta_k \prod_{l=1}^{k-1}(1-\beta_l),\ z_i \in \mathrm{Cat}(\pi),\ \theta_k \in \mathrm{Dir}(\lambda_1, \ldots, \lambda_M),$$

$$\beta = \|\beta_k\|_{K\times 1}, \quad Z = \|z_i\|_{n\times 1}, \quad \theta = \|\theta_{km}\|_{K\times M}.$$

В качестве примера распределения x в каждой из компонент смеси расмматривается дискретное распределение с параметрами θ_k, т.е.

$$x_i|z_i \in \mathrm{Cat}(\theta_{z_i}) \quad \Leftrightarrow \quad \mathbb{P}(x_i = m|z_i = k) = \theta_{km},$$
$$m = \overline{1, M}, \quad k = \overline{1, K}, \quad i = \overline{1, n}.$$

Будем искать $q(Z, \beta, \theta)$ в следующем факторизованном виде

$$q(Z, \beta, \theta) = \prod_{k=1}^{K} q(\beta_k|\gamma_k) \prod_{k=1}^{K} q(\theta_k|\tau_k) \prod_{i=1}^{n} q(z_i|\phi_i),$$

γ_k – Beta параметры для распределений с.в. β_k, τ_k – параметры распределения F аналогичные λ, ϕ_i – параметры дискретного распределения с.в. z_i, $K \sim \alpha \ln n$. Покажите, что параметры q, соответствующие минимуму $\mathcal{KL}(q\|p)$, удовлетворяет следующей системе уравнений:

$$\gamma_{k1} = \alpha + \sum_{i=1}^{n} q(z_i = k), \quad \gamma_{k2} = 1 + \sum_{i=1}^{n} q(z_i > k),$$

$$\tau_{km} = \lambda_m + \sum_{i=1}^{n} [x_i = m] q(z_i = k),$$

$$\phi_{ik} \propto \exp\left(\mathbb{E}_q \log(\beta_k) + \sum_{l<k} \mathbb{E}_q \log(1 - \beta_l) + \sum_{i=1}^{n} \mathbb{E}_q \log(\theta_{kx_i}) \right),$$

где

$$q(z_i = k) = \phi_{i,k}, \quad q(z_i > k) = \sum_{j=k+1}^{K} \phi_{i,j},$$
$$\mathbb{E}_q \log(\beta_k) = \psi(\gamma_{k1}) - \psi(\gamma_{k1} + \gamma_{k2}),$$
$$\mathbb{E}_q \log(1 - \beta_k) = \psi(\gamma_{k2}) - \psi(\gamma_{k1} + \gamma_{k2}),$$
$$\mathbb{E}_q \log(\theta_{km}) = \psi(\tau_{km}) - \psi\left(\sum_j \tau_{kj} \right).$$

14 (Тематическое моделирование)**.** Вероятностная тематическая модель представляет собой средство для выявления тем в коллекции текстовых документов, описывая каждую тему $z \in Z$ дискретным распределением на множестве терминов W, а каждый документ $d \in D$ — дискретным распределением на множестве тем. Предполагается, что порядок терминов в документах не важен, и коллекция является выборкой из дискретного распределения $\mathbb{P}(d, w, z)$, где $\Omega = D \times W \times Z$, z – скрытая переменная темы термина w в документе d. Коллекция представляется матрицей частот $F = \|\widehat{p}_{wd}\|_{W \times D}$. Предполагается также, что

$$\mathbb{P}(d, w, z) = \mathbb{P}(w|d, z)\mathbb{P}(z|d)\mathbb{P}(d) = \mathbb{P}(w|z)\mathbb{P}(z|d)\mathbb{P}(d) = \phi_{wz}\theta_{zd}\,\mathbb{P}(d),$$

$$\mathbb{P}(d, w) = \mathbb{P}(d)\sum_z \phi_{wz}\theta_{zd}.$$

Обозначим за n_{dw} число слов w в документе d. Поиск максимума правдоподобия модели для заданной выборки $X = \{(d, w, n_{dw}) : d \in D, w \in W\}$

$$L(\Phi, \Theta) = \log \prod_{d,w} \Big[\mathbb{P}(d, w)\Big]^{n_{dw}}$$

равносилен минимизации расстояния $\mathcal{KL}(F\|\Phi\Theta)$, где стохастическое матричное разложение $\Phi\Theta$ не единственно и определено с точностью до невырожденного преобразования $\Phi S^{-1} S\Theta$. Ввиду наличия неопределенности в выборе решения, наряду с правдоподобием имеет смысл максимизировать одновременно набор критериев $R_i(\Phi, \Theta)$, выражающих априорные предположения на счет ϕ_z и θ_d как столбцов матриц Φ, Θ^T и называемых *регуляризаторами*. Предлагается в качестве решения задачи многокритериальной оптимизации максимизировать

$$L(\Phi, \Theta) + R(\Phi, \Theta) = L(\Phi, \Theta) + \sum_i \tau_i R_i(\Phi, \Theta).$$

Покажите, что шагами EM-алгоритма (см. задачу 12 в разделе 5) для решения данной задачи будут

Expectation step: вычислить распределение по темам для каждого термина в документе

$$q_{dw}(z) = \frac{\phi_{wz}\theta_{zd}}{\sum_s \phi_{ws}\theta_{sd}};$$

Maximization step: найти оптимальные Φ, Θ

$$\phi_{wz} \propto \left(\sum_d n_{dw} q_{dw}(z) + \phi_{wz} \frac{\partial R}{\partial \phi_{wz}} \right)_+, \quad \theta_{zd} \propto \left(\sum_w n_{dw} q_{dw}(z) + \theta_{zd} \frac{\partial R}{\partial \theta_{zd}} \right)_+.$$

Докажите, что регуляризатор вида

$$R(\Theta) = -\sum_{d=1}^{D} \sum_{k=1}^{K} \alpha_k \log(\theta_{dk})$$

соответствует следующему априорному предположению на счет распределения θ_d

$$\theta_d \in \mathrm{Dir}(\alpha_1 + 1, \ldots, \alpha_K + 1), \quad d \in \overline{1, D},\ K = |T|.$$

При каком значении $\alpha_1, \ldots, \alpha_K$ данный регуляризатор будет способствовать увеличению количества нулевых компонент θ_d?

15 (PAC-Байесовские неравенства)**.** Ознакомьтесь с постановкой задачи классификации. Для правила (алгоритма) классификации $h \in H$ и набора классифицируемых объектов $X_1, \ldots, X_n$ пусть выборка $X_1^h, \ldots, X_n^h$ есть значения функции потерь с $X_i^h \in [0, 1]$ и $\mathbb{E}[X_1^h] = \mu^h$, $\mathbb{D}[X_1^h] = (\sigma^h)^2$.

Воспользовавшись неравенством Хефдинга и указанием к задаче, докажите следующее утверждение. Для любого не зависящего от выборок распределения p на H ($h \sim p$), $\lambda \geq 0$, $\delta \in (0, 1)$ и произвольного априорного распределения q на H с вероятностью не меньше $1 - \delta$ справедливо:

$$\mathbb{E}_q \left(\frac{1}{n} \sum_{i=1}^{n} X_i^h - \mu^h \right) \leq \frac{\mathcal{KL}(q\|p) + \ln \frac{1}{\delta}}{n\lambda} + \frac{\lambda}{8}. \tag{1}$$

Воспользовавшись неравенством Бернштейна, докажите утверждение, аналогичное предыдущему.

$$\mathbb{E}_q \left(\frac{1}{n} \sum_{i=1}^{n} X_i^h - \mu^h \right) \leq \frac{\mathcal{KL}(q\|p) + \ln \frac{1}{\delta}}{n\lambda} + \lambda(e-2)\mathbb{E}_p(\sigma^h)^2. \tag{2}$$

Указание. Докажите следующие утверждения.

Для любого измеримого отображения $f : H \to \mathbb{R}$ и любых пар распределений q и p на H справедливо:

$$\mathbb{E}_q f(h) \leq \mathcal{KL}(q\|p) + \ln\left(\mathbb{E}_p e^{f(h)}\right).$$

Случайная величина X определена на множестве D. Рассмотрим с.в. $h \sim p$, $h \in H$, не зависящую от X, измеримое отображения $f : H \times D \to \mathbb{R}$. Тогда $\forall \delta \in (0,1)$ и произвольного распределения q на H с вероятностью не меньше $1-\delta$ справедливо следующее:

$$\mathbb{E}_q f(h, X) \leq \mathcal{KL}(q\|p) + \ln\frac{1}{\delta} + \ln\left(\mathbb{E}_p \mathbb{E}_X e^{f(h,X)}\right).$$

Замечание. Минимизация по λ, фигурирующем в выражениях (1) и (2), приводит к зависимости вида $\lambda = \lambda(q)$. Для получения оценки, выполняющейся одновременно для всех q, рассмотрим конечный набор значений λ_i, и для каждого q будем брать i, соответствующее минимуму в выражениях (1) и (2). Т.о., если определить λ_i как $c^i\lambda_{\min}$, где $\lambda_{\min}$ соответствует $\mathcal{KL}(q\|p) = 0$, то получим следующие *PAC-Байесовские неравенства* Хефдинга и Бернштейна.

$$\mathbb{E}_q\left(\frac{1}{n}\sum_{i=1}^{n} X_i^h - \mu^h\right) \leq \frac{1+c}{2}\sqrt{\frac{\mathcal{KL}(q\|p) + \ln\frac{1}{\delta} + \varepsilon(q)}{2n}},$$

$$\varepsilon(q) = \frac{\ln 2}{2\ln c}\left(1 + \frac{\mathcal{KL}(q\|p)}{\ln(1/\delta)}\right);$$

при условии $\mathcal{KL}(q\|p) \leq n(e-2)\mathbb{E}_p(\sigma^h)^2 - \ln\frac{\nu}{\delta}$ выполнено неравенство

$$\mathbb{E}_q\left(\frac{1}{n}\sum_{i=1}^{n} X_i^h - \mu^h\right) \leq (1+c)\sqrt{\frac{(e-2)\mathbb{E}_p(\sigma^h)^2(\mathcal{KL}(q\|p) + \ln\frac{\nu}{\delta})}{n}},$$

$$\nu = \left\lceil \frac{1}{\ln c}\ln\left(\frac{(e-2)n}{4\ln(1/\delta)}\right)\right\rceil + 1.$$

Данные неравенства позволяют с хорошей точностью оценить отличие среднего риска от эмпирического риска, по сравнению с VC-оцениванием. Оптимизация по q позволяет настраивать гиперпараметры алгоритма классификации.

3. Неравенства концентрации меры и вероятности больших уклонений

1 (Концентрация площади сферы и объема шара)**.**

а) Рассмотрим шар $B^n(r)$ радиуса r в евклидовом пространстве $\mathbb{R}^n$ большой размерности, пусть в шаре задана равномерная мера. Необходимо убедиться в том, что мера сконцентрирована в малой окрестности границы шара.

б) Рассмотрим сферу $S^{n-1}(r)$ в евклидовом пространстве $\mathbb{R}^r$ с радиусом в начале координат. Необходимо убедиться в том, что выбранные наугад два единичных вектора в пространстве $\mathbb{R}^n$ большой размерности с большой вероятностью окажутся почти ортогональными, если на сфере задано равномерное распределение.

Замечание. Во втором пункте достаточно доказать, что для всякого сколь угодно малого $\delta > 0$ проекция второго вектора на ось x_1 с вероятностью, близкой к единице, лежит в промежутке $[-\delta, \delta]$ при $n \to \infty$. Это равносильно тому, что доля от площади всей сферы $S^{n-1}(r)$, которую занимает сферический слой $S_\delta^{n-1}(r)$, проектирующийся в отрезок $[-\delta, \delta]$ оси x_1, может быть сделана сколь угодно близкой к 1 при $n \to \infty$.

Перейдя к n-мерным сферическим координатам и обратно, покажите, что мера сферического слоя $S_\delta^{n-1}(r)$ равна

$$\mu_{n-1} S_\delta^{n-1}(r) = C r^{n-1} \int\limits_{-\delta}^{\delta} \Big(1 - (x/r)^2\Big)^{(n-3)/2} dx,$$

тогда вероятность попадания в данный слой $S_\delta^{n-1}(r)$ равна

$$\mathbb{P}[-\delta, \delta] = \frac{\int\limits_{-\delta}^{\delta} \Big(1 - (x/r)^2\Big)^{(n-3)/2} dx}{\int\limits_{-r}^{r} \Big(1 - (x/r)^2\Big)^{(n-3)/2} dx}.$$

Данное отношение на зависит от r, поэтому можно считать $r = 1$.

Для нахождения асимптотики имеющихся интегралов ($n \to \infty$) используйте классические результаты относительно асимптотики

интеграла Лапласа $F(\lambda) = \int_a^b f(x) e^{\lambda S(x)}\, dx$ при $\lambda \to +\infty$ (см. указание к задаче **??** из раздела **??**).

Для решения этой и следующей задачи рекомендуется ознакомиться с книгой Зорич В.А. Математический анализ задач естествознания,– М.: МЦНМО, 2008.

2 (Изопериметрическое неравенство и принцип концентрации меры; П. Леви, 1919)**.** Число μ_f называют медианой функции f, если $\mu\left(\vec{x} \in S_1^n : \ f\left(\vec{x}\right) \geq \mu_f\right) \geq 1/2$ и $\mu\left(\vec{x} \in S_1^n : \ f\left(\vec{x}\right) \leq \mu_f\right) \geq 1/2$, где $\mu\left(d\vec{x}\right)$ – равномерное 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 (расстояние определяется по геодезический) множества A на сфере S_1^n. Предположим теперь, что в некотором царстве, расположенном на S_1^n, 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 на плоскости – "круглый огород", это можно обобщить на наш случай. 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 $\mu\left(A\right) \geq 1/2$, то

$$\mu\left(A_\delta\right) \geq 1 - \sqrt{\frac{\pi}{8}} \exp\left(\frac{-\delta^2 n}{2}\right).$$

Пусть теперь на S_1^n задана функция с модулем непрерывности

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

Покажите, что тогда

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

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) исследовать плотную концентрацию около среднего значения

различные функций на случайных графах: число независимости, хроматическое число и т.п.

Замечание. Изопериметрические неравенства на сфере были обобщены в начале 80-х М.Л. Громовым на римановы многообразия (см. подробнее в [39]).

Пусть (X, g) — компактное связное гладкое риманово многообразие размерности $n > 2$ со строго положительной кривизной Риччи и римановой метрикой g, наделенное элементом объема $d\mu = \frac{dv}{V}$, где V — полный объем X. Кривизна Риччи — способ описания изменения многообразия по отношению к евлидовой мере, то есть степени отличия мнообразия от евклидова пространства. Обозначим за $c(X)$ точную нижнюю грань тензора кривизны Риччи по всем единичным касательным векторам. Пусть $c(X) > 0$. Тогда

$$\mathcal{P}_\mu \geq \mathcal{P}_{\sigma^n_{\mathbb{R}}},$$

где σ^n_R – равномерная инвариантная мера, $\mathcal{P}_\mu$ – изопериметрическая функция, т.е. наибольшая функция на $[0, \mu(X)]$, такая что

$$\mu^+(A) \geq \mathcal{P}_\mu(\mu(A)) \quad \text{при } \mu^+(A) = \lim_{t\to\infty} \inf \frac{1}{t}\mu(A_t/A),$$

где $A_t = \{x \in X; g(x, A) < t\}$.

Здесь $R > 0$ таково, что

$$c(S^n_R) = \frac{n-1}{R^2} = c(X),$$

где S^n_R — n-сфера с радиусом R, снабженная нормированной равномерной инвариантной мерой σ^n_R. В частности для (X, g, μ) верно

$$\alpha(X; r) \leq \mathrm{e}^{-cr^2/2}, \quad r > 0,$$

где концентрационная функция $\alpha(X; r)$ определяется следующим образом

$$\alpha(X; r) = 1 - \inf\left\{\mu(A_r) |\, A \subset X, \mu(A) \geq \frac{1}{2}\right\}.$$

3. (Концентрация на дискретном кубе) Пусть $E = \{-1, 1\}^n$ — дискретный n-мерный куб, на множестве вершин которого задана равномерная вероятностная мера μ. Введем на вершинах куба стандартную Хэммингову метрику. Тогда для всякой 1-липшицевой (по

введенной метрике) действительнозначной функции f, заданной на E, с медианой M для $\varepsilon \geq 0$ имеет место неравенство

$$\mathbb{P}\{|f(X) - M| \geq \varepsilon\} \leq 2\exp\left(\frac{-\varepsilon^2}{2n}\right).$$

а) Пусть $\mathbb{E}f(X)$ – математическое ожидание функции. Докажите, что

$$|\mathbb{E}f(X) - M| \leq \sqrt{2\pi n}.$$

б) Покажите, что фактор $\frac{1}{n}$ под экспонентой не может быть 'улучшен'. В частности, неравенство концентрации явно зависит от размерности куба (при условии сохранения гауссовского хвоста по ε).

Замечание. Используйте тот факт, что для случайной величины $Y > 0$ имеет место равенство $\mathbb{E}Y = \int\limits_0^\infty \mathbb{P}\{Y \geq x\}dx$.

См.[22] и Barvinok A. Math 710: Measure Concentration. Lecture notes, 2005. http://www.math.lsa.umich.edu/ barvinok/total710.pdf.

4. (Неравенство Талаграна для дискретного куба) Пусть $E = \{-1, 1\}^n$ — дискретный n-мерный куб, на множестве вершин которого задана равномерная вероятностная мера μ. Введем на вершинах куба стандартную евклидову метрику. Тогда для всякой выпуклой 1-липшицевой действительнозначной функции f, заданной на E, с медианой M для $\varepsilon \geq 0$ имеет место неравенство

$$\mathsf{P}\{|f(X) - M| \geq \varepsilon\} \leq 2\exp\left(\frac{-\varepsilon^2}{2n}\right).$$

а) Покажите, что условие выпуклости функции f нельзя опустить в данном неравенстве.

б) Получите неравенство, аналогичное неравенству Талаграна для концентрации около математического ожидания.

Замечание. См. Talagrand M. An Isoperimetric Theorem on the Cube and the Kinitchine-Kahane Inequalities, Proceedings of the American Mathematical Society, 1988. Pp. 905-909.

В отличие от предыдущей теоремы в данном неравенстве нет зависимости от размерности куба. Одновременно накладываются два дополнительных требования, во-первых, липшицевость по меньшей

– евклидовой метрике, во-вторых, дополнительно на функцию накладывается условие выпуклости.

Существуют аналогичные неравенства концентрации вокруг математического ожидания с гораздо более точными константами. Тем не менее, даже этот простой способ позволяет получить нужный нам гауссовский хвост, не зависящий от размерности.

5. (Концентрация на сечениях куба) Пусть $E = \{-1, 1\}^n$ — дискретный n-мерный куб. Рассмотрим его сечение, состоящее из всех его вершин, содержащих ровно k координат, равных 1. Введем на вершинах сечения нормированную метрику d, равную половине Хэмминговой метрики. При этом расстояние между вершинами, координаты которых отличаются лишь перестановкой пары координат равно единице. На множестве вершин сечения зададим равномерную вероятностную меру μ. Введем также понятие длины дискретного градиента $|\nabla f(X)|$. Обозначим

$$|\nabla f(X)|^2 = \sum_{Y : d(X,Y)=1} |f(X) - f(Y)|^2,$$

где суммирование ведется по всем $k(n-k)$ вершинам сечения, удаленным от данной вершины на единицу.

Тогда для всякой действительнозначной функции f, заданной на описанном сечении, длина дискретного градиента которой в точках сечения ограничена числом σ, для $\varepsilon \geq 0$ имеет место неравенство

$$\mathbb{P}\{f(X) - \mathbb{E}f(X) \geq \varepsilon\} \leq \exp\left(\frac{-(n+2)\varepsilon^2}{4\sigma^2}\right).$$

Пусть случайна величина Y имеет гипергеометрическое распределение с параметрами N, K, n, то есть для неотрицательного целого k

$$\mathbb{P}\{Y = k\} = \frac{C_K^k C_{N-K}^{n-k}}{C_N^n}.$$

Получите экспоненциальную оценку на величину $\mathbb{P}\{Y - \mathbb{E}Y \geq \varepsilon\}$.

Замечание. Вспомните интерпретацию гипергеометрического распределения и свяжите ее с равномерным распределением на вершинах сечения дискретного куба. См. Bobkov S. G. Concentration of normalized sums and a central limit theorem for noncorrelated random

variables, Annals of probability, No. 4, Pp. 2884-2907, 2004 и Bobkov S. G., Tetali, P. Modified logarithmic Sobolev inequalities in discrete settings, Journal of Theoretical Probability, No. 2, Pp. 289–336, 2006.

6. * 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 ℓ длину наикратчайшего из таких путей (кратчайший гамильтонов путь). Покажите, что найдется такая константа $c > 0$, не зависящая от n, что

$$\mathbb{P}(\ell - \mathbb{E}\ell \geq t) \leq \exp\left(-\frac{t^2}{4c}\right).$$

Можно показать, что $\mathbb{E}\ell \sim \beta\sqrt{n}$, где β также не зависит от n.

Замечание. См. книгу Dubhashi D. P., Panconesi A. «Concentration of measure for the analysis of randomized algorithms», Cambridge University Press, 2009.

Приведем несколько неравенств Талаграна. Пусть заданы множества Ω_i, $i = 1, \ldots, n$, элементарных исходов. На этих множествах заданы вероятностные меры $\mathbb{P}_i$, $i = 1, \ldots, n$. Положим

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

Введем взвешенную метрику Хэмминга:

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

и определим $d_\alpha(x, A) = \min_{y \in A} d_\alpha(x, y)$, $\rho(x, A) = \sup_{\alpha \in \mathbb{R}^n} d_\alpha(x, A)$. Пусть $A \in \sigma(\Omega)$. Определим t-окрестность ($t \geq 0$) множества A по формуле

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

Тогда справедливо неравенство Талаграна:

$$\mathbb{P}(A)(1-\mathbb{P}(A_t)) \le \exp\left(-\frac{t^2}{4}\right).$$

Следствие. В качестве приложения неравенства Талаграна рассмотрим функцию $h:\ \Omega \to \mathrm{R}$, $\Omega = \mathbb{R}^n$, удовлетворяющую условию Липшица с константой σ. Функцию h будем называть проверяемой со сложностью f, если при $h(x) \ge s$ существует такое множество $I \subseteq \{1,\dots,n\}$ с $|I| \le f(s)$, что для всех $y \in \Omega$, совпадающих с x в координатах из I, выполняется $h(y) \ge s$. Грубо говоря, из выполнения неравенства $h(x) \ge s$ следует, что существует сравнительно небольшое количество координат, обеспечивающих выполнение данного неравенства. Тогда из неравенства Талаграна можно получить следующее соотношение о плотной концентрации случайной величины h: для всех b и t справедливо

$$\mathbb{P}\left[h \le b - t\sqrt{f(b)}\right]\mathbb{P}\left[h \ge b\right] \le e^{-\frac{t^2}{4}}.$$

Выбирая либо в качестве b, либо в качестве $b - t\sqrt{f(b)}$ медиану случайной величины h, получаем результат о плотной концентрации h вокруг своей медианы

$$\mathbb{P}(|h(X) - M| > t) \le 4\mathrm{e}^{-t^2/8\sigma^8}.$$

Неравенство Талаграна для независимых случайных величин. Пусть $X_1,\dots,X_n$ независимые случайные величины в S. Для любого класса функций $\mathcal{F}$ на S равномерно ограниченного на S константой U для всех $t > 0$ выполнено

$$\mathbb{P}\left\{\left\|\sum_{i=1}^n f(X_i)\right\|_{\mathcal{F}} - \mathbb{E}\left\|\sum_{i=1}^n f(X_i)\right\|_{\mathcal{F}} \ge t\right\} \le K\exp\left\{-\frac{t}{UK}\log\left(1+\frac{tU}{V}\right)\right\},$$

где K - универсальная константа и V удовлетворяет условию

$$V \ge \mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^n f^2(X_i)$$

и использованы обозначения $\|Y\|_{\mathcal{F}} = \sup_{\mathcal{F}}|Y(f)|$, где $Y : \mathcal{F} \to \mathbb{R}$.

7 (Семейства Леви)**.** Пусть (X_n, ρ_n, μ_n) – метрическое вероятностное пространство, X_n – компактное множество с метрикой ρ_n, $\operatorname{diam} X_n \geq 1$ и заданной вероятностной мерой μ_n. Семейство (X_n, ρ_n, μ_n) метрических вероятностных пространств называется семейством Леви, если для любого $\varepsilon > 0$, $\alpha(X_n, \varepsilon \cdot \operatorname{diam} X_n) \to 0$ для $n \to \infty$ (см. определение концентрационной функции в замечании к задаче 2). Семейство называется нормальным семейством Леви с константами (c_1, c_2), если

$$\alpha(X_n; \varepsilon) \leq c_1 \exp(-c_2 \varepsilon^2 n),$$

Докажите, что следующие семейства будут нормальными семействами Леви.

а) Пусть на $E^n = \{-1, 1\}^n$ задана Хэммингова метрика

$$d(s,t) = \frac{1}{n} |\{i : s_i \neq t_i\}|$$

и нормированная считающая мера μ, т.е. $\mu(A) = |A|/2^n$. Докажите неравенство

$$\alpha(F_2^n; \varepsilon) \leq \frac{1}{2} \exp(-2\varepsilon^2 n).$$

б) Задана группа Π_n перестановок $\{1, \ldots, n\}$ с заданной нормированной метрикой Хэмминга

$$d(\pi_1, \pi_2) = \frac{1}{n} |\{i : \pi_1(i) \neq \pi_2(i)\}|$$

и нормированной считающей мерой (см. пункт а)). Докажите неравенство

$$\alpha(\Pi_n; \varepsilon) \leq \exp\Big(- \varepsilon^2 n/64 \Big).$$

Замечание. Примеры приведены в статье Milman V.D. The heritage of P. Levy in geometrical functional analysis // Asterisques. 1988. V. 157-158. P. 273-302, см. также книгу M. Ledoux «The Concentration of Measure Phenomenon», American Mathematical Soc., 2005, а также теорему Талаграна в замечании к предыдущей задаче.

8. (Кац, Секей) Рассмотрим полином с вещественными коэффициентами

$$a_0 + a_1 t + \ldots + a_{n-1} t^{n-1},$$

где $(a_0, a_1, \ldots, a_{n-1})$ точка на единичной сфере $S_n(1)$. Среднее число вещественных корней полинома определим следующим образом:

$$\mathbb{E}N = \frac{1}{|S_n(1)|} \int_{S_n(1)} N(a) d\sigma,$$

где $N(a)$ число вещественных корней полинома, $d\sigma$ — элемент поверхности единичной сферы площадью

$$|S_n(1)| = \frac{(2\pi)^{n/2}}{\Gamma\left(\frac{n}{2}\right)}.$$

а) Покажите, что среднее число вещественных корней полинома с коэффициентами на единичной сфере равно среднему числу вещественных корней полинома, коэффициенты которого независимы и распределены стандартно нормально, то есть

$$\mathbb{E}N = (2\pi)^{-n/2} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} N(a) \exp\left(-\frac{1}{2}\|a\|^2\right) da_0, \ldots, da_{n-1}.$$

б) Пусть $\mathbb{E}N^{(1)}$, $\mathbb{E}N^{(2)}$, $\mathbb{E}N^{(3)}$, $\mathbb{E}N^{(4)}$ — средние значения числа вещественных корней, заключенных в интервалах $(-\infty, -1)$, $(-1, 0)$, $(0, 1)$, $(1, \infty)$ соответственно. Покажите, что

$$\mathbb{E}N^{(1)} = \mathbb{E}N^{(2)} = \mathbb{E}N^{(3)} = \mathbb{E}N^{(4)}$$

и

$$\mathbb{E}N^{(i)} \sim (2\pi)^{-1} \ln n, \quad n \to \infty.$$

Замечание. Н.Б. Маслова в статье О распределении числа вещественных корней случайных полиномов // ТВП, 1974, Т. 19, В. 3, с. 488—500 доказала следующую теорему: если коэффициенты X_i случайного алгебраического уравнения

$$\sum_{j=1}^{n} X_j z^j = 0$$

являются независимыми одинаково распределенными случайными величиными с нулевыми математическими ожиданиями и

$$\mathbb{E}\left(|X_j|^{2+\epsilon}\right) < \infty$$

для некоторого положительного ϵ, то число действительных корней этого уравнения распределено нормально с математическим ожиданием $\frac{2}{\pi}\ln n$ и стандартным отклонением $2\sqrt{\pi^{-1}(1-2\pi^{-1})\ln n}$.

9. ** Дан случайный граф (модель Эрдеша–Реньи, см. задачу **??** раздела **??**) $G(n,\ p)$ с n вершинами и вероятностью появления каждого ребра p. Пусть $p \geq \sqrt{\frac{2\ln n}{n}}$, причем длины ребер r_{ij} являются независимыми случайными величинами, имеющими равномерное распределение на отрезке $[0,\ 2r]$. Покажите, что тогда почти наверное граф $G(n,\ p)$ имеет гамильтонов цикл, причём длина почти всех гамильтоновых циклов стабилизируется около nr.

10. ** Дан случайный граф (модель Эрдеша–Реньи, см. задачу **??.??**) $G(n,\ p)$ с n вершинами и вероятностью появления каждого ребра $p \in [\varepsilon, 1]$, $\varepsilon > 0$. Вес каждого появившегося случайного ребра разыгрывается независимо согласно равномерному распределению на отрезке $[0, 2]$. Источник и сток выбираются случайно. Обозначим через S_n значение максимального потока для полученного случайного взвешенного графа. Покажите, что $S_n/pn \xrightarrow{p} 1$.

11 (Стохастическое агрегирование; В.И. Опойцев)**.** а) На рынке имеется n продавцов, каждый из которых может продать свой товар в объеме x_k для k-го товара. Спрос на товары обеспечивают покупатели. Пусть y_k – спрос на товар k. Общий объем сделок $L_n = \sum\limits_{k=1}^{n} \min\{x_k, y_k\}$. Предложите такой естественный способ определения вероятностной меры на двух симплексах

$$S_X = \left\{ \vec{x} \geq \vec{0} :\ \sum_{k=1}^{n} x_k = X \right\}, \quad S_Y = \left\{ \vec{y} \geq \vec{0} :\ \sum_{k=1}^{n} y_k = Y \right\},$$

чтобы нашлась функция $f(X, Y)$, удовлетворяющая

$$L_n/n \xrightarrow{p} f(X, Y).$$

б) Пусть $\vec{y} = A\vec{x}$, $A = \|a_{ij}\|_{i,j=1,1}^{l,n}$, $X = \langle \vec{p}, \vec{x} \rangle$, $Y = \langle \vec{q}, \vec{y} \rangle$. Матрицы и векторы предполагаются положительными. Легко проверить, что

$$Y = \lambda(\vec{x})\, X, \quad \lambda(\vec{x}) = \sum_{i,j} \frac{q_i a_{ij}}{p_j} \frac{p_j x_j}{X} \stackrel{def}{=} \sum_{i,j} b_{ij} z_j.$$

Считая $z_j > 0$ независимыми одинаково распределенными с.в.: $\mathbb{E}z_j = m_j$ ($\sum_j m_j = 1$), $\mathbb{D}\,z_j = \sigma_j^2$ (например, $z_j \in \left[0, 2n^{-1}\right]$), покажите, что если выражение

$$\frac{\max\limits_j \sum\limits_i b_{ij} \cdot \max\limits_j \sigma_j}{\min\limits_j \sum\limits_i b_{ij} \cdot \min\limits_j m_j}$$

равномерно ограничено с ростом n, то существует такое число $\bar{\lambda}$, что с вероятностью стремящейся к единице $Y \simeq \bar{\lambda} X$.

12. (Устойчивые системы большой размерности; В.И. Опойцев) 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 $x^{k+1} = Ax^k + b$ (СОДУ $\dot{x} = -x + Ax + b$), вне зависимости от точки старта x^0, сходится к единственному решению уравнения $x^* = Ax^* + b$. Скажем, если $\|A\| = \max\limits_i \sum\limits_j |a_{ij}| < 1$, то и $\rho(A) < 1$ (обратное, конечно, не верно). Предположим, что существует такое $\varepsilon > 0$, что

$$\frac{1}{n} \sum_{i,j=1}^{n} |a_{ij}| < 1 - \varepsilon. \quad (S)$$

Очевидно, что отсюда 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).

Указание. 1. Покажите, что достаточно рассматривать матрицы с неотрицательными элементами.

2. Покажите, что достаточно доказать утверждение задачи на множестве матриц, удовлетворяющих условию

$$\frac{1}{n} \sum_{i,j=1}^{n} a_{ij} = 1 - \varepsilon. \quad (SE)$$

3. Далее положим $a_{ij} \in \mathrm{Exp}\big(n/(1-\varepsilon)\big)$ — независимые одинаково распределенные случайные величины. Покажите, что при $n \to \infty$

распределение элементов случайной матрицы $A = \|a_{ij}\|_{i,j=1}^n$ будет сходиться к равномерному распределению на множестве матриц, удовлетворяющих (SE).

4. Введя обозначения $P_n = \mathbb{P}(\|A\| \geq 1) \geq \mathbb{P}(\rho(A) \geq 1)$, воспользуйтесь неравенством Чебышёва

$$P_n \leq n\mathbb{P}\Big(\sum_{j=1}^n a_{1j} \geq 1\Big) = n\mathbb{P}\Big(X \geq 1\Big) \leq n\mathbb{P}\Big(|X - (1-\varepsilon)| \geq \varepsilon\Big) =$$

$$= n\mathbb{P}\Big(|X - \mathbb{E}X| \geq \varepsilon\Big) \leq \frac{n}{\varepsilon^4}\mathbb{E}(X - \mathbb{E}X)^4 = O\Big(\frac{1}{n}\Big) \xrightarrow{n\to\infty} 0.$$

13 (Красносельский–Крейн)**.** Рассмотрим систему линейных алгебраических уравнений

$$x = Ax + b,$$

где A положительно определенная неособенная матрица, собственные числа $\lambda_1 \leq \cdots \leq \lambda_n$ которой меньше единицы, b — заданный и x — искомый векторы n-мерного подпространства. Пусть эта система решается при помощи итерационного процесса

$$x_{m+1} = Ax_m + b, \quad m = 1, 2, \ldots,$$

который заканчивается на p-м шаге, если вектор–невязка $\delta_p = x_{p+1} - x_p$ попадает в шар радиуса α с центром в нуле. Ошибка итерационного процесса $\varepsilon_m = x^* - x_m$, где x^* истинное решение системы уравнений, связана с невязкой (проверить)

$$\varepsilon_m = (I - A)^{-1}\delta_m,$$

поэтому исходя из значений вектора невязки можно определить вероятностное распределение ошибок.

Оказывается, что наиболее вероятными ошибками являются максимальные. А именно, пусть начальная ошибка равномерно распределена в шаре T радиуса R, тогда для любого $\eta < 1$ вероятность выполнения следующего неравенства стремится к единице при $R \to \infty$

$$\eta\frac{\alpha\lambda_n}{1-\lambda_n} \leq \|\varepsilon\| \leq \frac{\alpha}{1-\lambda_n}.$$

Проверьте это для случая $n = 2$.

Замечание. Пусть итерационный процесс заканчивается на шаге p, если вектор-невязка $\delta_p = x_{p+1} - x_p$ попадает в окрестность G нуля. Пусть G — шар радиуса α. Тогда ошибка ε_p попадет в множество $G_0 = (I - A)^{-1}G$, которое является эллипсоидом с полуосями длины $\frac{\alpha}{1-\lambda_1}$ и $\frac{\alpha}{1-\lambda_2}$. Обозначим $G_{-1} = AG_0$ — эллипсоид с полуосями $\frac{\alpha\lambda_i}{1-\lambda_i}$, $i = 1, 2$.

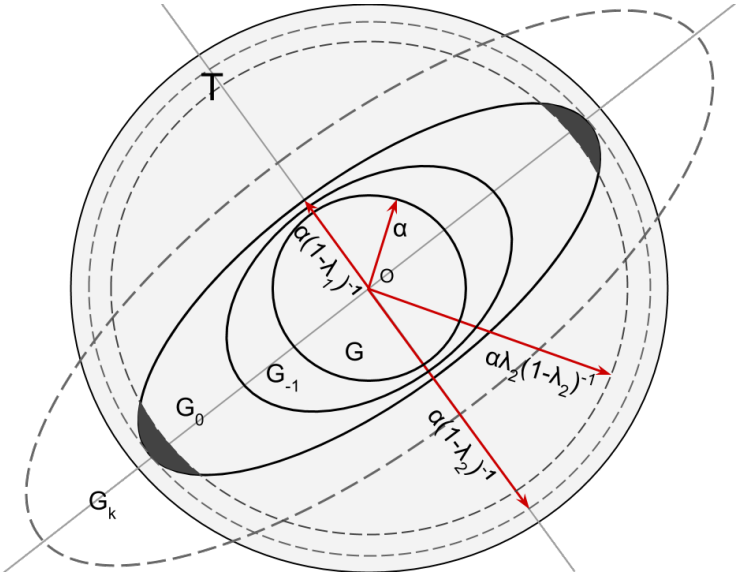


Рис. 5: Множества ошибок в случае матрицы 2×2. Закрашенная темно-серым цветом часть соответствует области из утверждения задачи при η близком к 1.

Обозначим $G_m = A^{-m}G_0$. Процесс заканчивается на p-м шаге, если $\varepsilon_0 \in G_p$ и $\varepsilon_0 \notin G_p$, те когда ε_0 находится в слое $G_p - G_{p-1}$. При этом окончательная ошибка будет находиться в слое $A^p(G_p - G_{p-1}) = G_0 - G_{-1}$.Таким образом окончательная ошибка принадлежит G_{-1} только в случае, если ε_0 принадлежит G_0. Вероятность нахождения окончательной ошибки в элементе слоя $G_0 - G_{-1}$ равна сумме вероятностей нахождения начальной ошибки в элементах $\Delta = A^{-m}\Delta_0$ $(m = 0, 1, \dots)$.

Предположим, что начальная ошибка ε_0 равномерно распределена в шаре T достаточно большого радиуса. В этом случае вероятность $P(\Delta_m)$ нахождения начальной ошибки в элементе $\Delta_m \in T$ пропорциональна объему этого элемента. Подсчитайте вероятность $P(\Delta_0)$ попадания окончательной ошибки в элемент Δ_0 и убедитесь, что вероятность попадания окончательной ошибки в точку слоя $G_0 - G_{-1}$ для данной точки тем больше, чем позже эта точка выйдет из T при последовательном применении к ней операции A^{-1}.

Пусть e_i — собственные векторы матрицы A, соответствующие собственным числам λ_i, $i = 1, 2$. Для доказательства утверждения задачи оцените вероятность того, что окончательная ошибка $\varepsilon = \sum_{i=1}^{n} \xi_i e_i$ не удовлетворяет неравенству (утверждению задачи). Перейдя к пределу по $R \to \infty$ получите, что вероятность невыполнения утверждения задачи стремится к нулю.

14 (Физическая интерпретация концентрации меры на сфере)**.** Имеется n частиц массы m со скоростями v_i, $i = 1, \ldots, n$. Известно, что вектор скоростей молекул идеального газа равномерно распределен по поверхности постоянной энергии. Суммарная кинетическая энергия E_n растет пропорционально n, то есть

$$\frac{1}{2}mv_1^2 + \cdots + \frac{1}{2}mv_n^2 = E_n; \quad \sum_{i=1}^{n} v_i^2 = \frac{2E_n}{m} \asymp n.$$

Получите закон распределения Максвелла скоростей частиц одномерного идеального газа.

Указание. В решении задачи 1 этого раздела перейдите к термодинамическому пределу, когда $n \to \infty$, $r = \sigma n^{1/2}$, чтобы получить закон распределения Максвелла скоростей частиц.

Замечание. Равномерное распределение на поверхности постоянной энергии возникло из-за того, что инвариантной (и предельной, то есть возникающей на больших временах, по эргодической гипотезе) мерой для гамильтоновой системы будет как раз равномерная мера по теореме Лиувилля (фазовый объем сохраняется). Поскольку выполняется закон сохранения энергии, то система "живет" на поверхности постоянной энергии. Следовательно, носитель инвариантной меры сосредоточен именно на этой поверхности.

Приведем для справки результат, полученный для выпуклых тел с заданной на них равномерной мерой (теорема Б. Кластарга в статье

Milman V.D. Geometrization of Probability // Progress in Mathematics. 2008. V. 265, http://www.math.tau.ac.il/ milman/, также см. статью Klartag B. A central limit theorem for convex sets // Inventiones mathematicae April 2007, Volume 168, Issue 1, pp. 91–131). Существует последовательность $\epsilon_n \to 0$, $n \to \infty$ для которой выполнено: пусть $K \in \mathbb{R}^n$ выпуклое компактное множество с непустой внутренностью, случайный вектор X распределен равномерно в K, тогда существуют вектор $\theta \in \mathbb{R}^n$, $t_0 \in \mathbb{R}$ и $\sigma > 0$, что выполнено

$$\sup_{A \in \mathbb{R}} \left| \mathbb{P}\left\{ \sum_{i=1}^{n} X_i \theta_i \in A \right\} - \frac{1}{\sqrt{2\pi}\sigma} \int_A \mathrm{e}^{-\frac{(t-t_0)^2}{2\sigma^2}}\, dt \right| \le \epsilon_n,$$

где супремум берется по всем измеримым множествами $A \in \mathbb{R}$. Если $\mathbb{E}X_i = 1$, $\mathbb{E}X_iX_j = \delta_{ij}$, то $t_0 = 0$, $\sigma = 1$, то можно найти такой единичный вектор θ, что выполняется приведенное неравенство.

15 (Лемма Пуанкаре)**.** Пусть X_n – случайный вектор с равномерным распределением на единичной сфере в $\mathbb{R}^n$. Равномерное распределение характеризуется тем, что оно инвариантно относительно группы ортогональных преобразований. Пусть Y_n обозначает первую координату X_n. Докажите, что $\sqrt{n}\, Y_n \xrightarrow{d} N(0,1)$ при $n \to \infty$. Заметим, что в статистической физике с помощью утверждения этой задачи получался закон распределения Максвелла скоростей частиц одномерного идеального газа.

Указание. См. предыдущую задачу. Решение задачи содержит в себе способ генерирования равномерного распределения. Пусть $\xi_1, \ldots, \xi_n$ — независимые в совокупности с.в., имеющие одинаковое распределение $N(0,1)$. Рассмотрим случайный вектор $Z_n = (\xi_1, \xi_2, \ldots, \xi_n)$. Тогда $Z_n \in N(0, I_n)$, I_n — единичная матрица размера n.

Показажите, что Z_n инвариантно относительно группы ортогональных преобразований. Заметим, что распределения

$$X_n \quad \text{и} \quad \frac{Z_n}{\|Z_n\|_{\mathbb{R}^n}} \quad \text{совпадают.}$$

Поэтому имеет место равенство по распределению с.в.

$$Y_n = \frac{\xi_1}{\sqrt{\xi_1^2 + \ldots + \xi_n^2}}$$

$$\Rightarrow \quad \sqrt{n}Y_n = \frac{\xi_1}{\sqrt{(\xi_1^2 + \ldots + \xi_n^2)/n}}.$$

Применить теорему Колмогорова у.з.б.ч. для $\frac{\xi_1^2+\ldots+\xi_n^2}{n}$.

16 (Геометрическая интерпретация закона больших чисел)**.** Рассмотрим куб $C^n = [-1,1]^n$ в евклидовом пространстве $\mathbb{R}^n$. Пусть ξ_i, $i = 1, \ldots, n$ независимые случайные величины с равномерным распределением на $[-1,1]$. Приведите геометрическую интерпретацию закона больших чисел.

Указание. Рассмотреть объем следующего множества — пусть $\mathcal{H}$ часть гиперплоскости, содержащаяся в кубе и перпендикулярная главной диагонали куба, т.е. $\sum_{i=1}^n x_i = 0$. Необходимо подсчитать объем $\varepsilon\sqrt{n}$-окрестности $\mathcal{H}$.

17 (В.И. Опойцев)**.** а) Пусть имеются абсолютно непрерывные (имеющие плотность) независимые случайные величины $X_1, \ldots, X_n$ и пусть $Y_n = G_n(X_1, \ldots, X_n)$ — также случайная величина. Докажите следующее неравенство, описывающее нелинейный закон больших чисел:

$$\mathbb{D}\, Y_n \leqslant \int_{\mathbb{R}^n} \sum_{i=1}^n \Big(\frac{\partial G_n(x_1, \ldots, x_n)}{\partial x_i} \Big)^2 f_i^*(x_i) \prod_{j \neq i} f_j(x_j)\, dx_1 \ldots dx_n,$$

где сопряженные плотности $f_i^*(x_i)$ существуют и определяются следующим образом

$$f_i^*(x_i) = \mu_i(\infty) \int\limits_{-\infty}^{x_i} f_i(t)\, dt - \mu_i(x_i), \quad \mu_i(x) = \int\limits_{-\infty}^{x} t f_i(t)\, dt.$$

б) Пусть независимые одинаково распределенные случайные величины $X_1, \ldots, X_n$ имеют равномерное распределение на отрезке $[0,1]$ (часто пишут $X_i \in R[0,1]$), а

$$\max_{x \in [0,1]^n} \|\nabla G_n(x_1, \ldots, x_n)\| \xrightarrow{n \to \infty} 0.$$

Тогда $Y_n \xrightarrow{P} \mathbb{E}Y_n$ при $n \to \infty$.

в) Пусть независимые одинаково распределенные случайные величины $X_1, \ldots, X_n$ имеют равномерное распределение на отрезке $[0,1]$, а $Y_n = G_n(X_1, \ldots, X_n) = \max\limits_{i=1,\ldots,n} X_i$. Тогда $Y_n \xrightarrow{p} \mathbb{E}Y_n$ при $n \to \infty$.

Замечание. См. В. Босс Лекции по математике. Т.4: Вероятность, информация, статистика, 2005.

18 (Неравенство Чернова)**.** Докажите, что неравенство Чернова для неотрицательной случайной величины X

$$\mathbb{P}\{X > t\} \leq \inf_{s>0} \mathbb{E}\exp(sX - st)$$

дает более завышенную границу по сравнению с моментной границей

$$\mathbb{P}\{X > t\} \leq \min_{q>0} \mathbb{E}[X^q]t^{-q},$$

то есть

$$\min_{q>0} \mathbb{E}[X^q]t^{-q} \leq \inf_{s>0} \mathbb{E}\left[\mathrm{e}^{s(X-t)}\right].$$

Указание. См. [41]. Использовать следствие из неравенства Маркова: для монотонной возрастающей неотрицательной функции $\phi(\cdot)$ и произвольной случайной величины X верно

$$\mathbb{P}\{\phi(X) \geq \phi(t)\} \leq \frac{\mathbb{E}\phi(X)}{\phi(t)}.$$

19 (Лемма Хёфдинга)**.** Пусть X— случайная величина, такая что $\mathbb{E}X = 0$, $a \leq X \leq b$. Покажите, что для $s > 0$ верно

$$\mathbb{E}\exp(sX) \leq \exp\left[\frac{s^2(b-a)^2}{8}\right].$$

Указание. Используя выпуклость экспоненты, для $a \leq x \leq b$

$$\mathrm{e}^{sx} \leq \frac{x-a}{b-a}\mathrm{e}^{sb} + \frac{b-x}{b-a}\mathrm{e}^{sa},$$

получите

$$\mathbb{E}\mathrm{e}^{sx} \leq \mathrm{e}^{\phi(u)},$$

где $u = s(b-a)$, $\phi(u) = -pu + \log(1-p+pe^u)$, $p = -a/(b-a)$. Найдите $\phi''(u)$, $\phi(0)$, $\phi'(0)$. Покажите, что

$$\phi''(u) \le \frac{1}{4}.$$

Используя формулу Тейлора, получите

$$\phi(u) \le \frac{u^2}{8} \le \frac{s^2(b-a)^2}{8}.$$

20 (Теорема Хёфдинга)**.** Пусть ξ_t, $t \in T$ – независимые случайные величины, такие что $\xi_t \in [a, b]$. Докажите, что

$$\mathbb{P}\left\{ \left| \frac{1}{n} \sum_{t \in T} \big(\xi_t - \mathbb{E}\xi_t\big) \right| \ge x \right\} \le 2 \exp\left\{ - \frac{2nx^2}{(b-a)^2} \right\}.$$

Указание. Введите случайную величину $\xi = \frac{1}{n}\sum_{i=1}^n (\xi_i - \mathbb{E}\xi_i)$. Воспользуйтесь неравеством Чернова и леммой Хёфдинга для ξ и получите

$$\mathbb{P}\{\xi > x\} \le \exp\left\{ \min_{\lambda} \left[- \lambda x + \frac{\lambda^2}{8} \frac{(b-a)^2}{n} \right] \right\}.$$

Затем найдите оптимальное λ. Аналогичное неравенство справедливо для $-\xi$.

21 (Неравенство Беннетта)**.** Пусть $X_1, \ldots, X_n$ независимые центрированные ограниченные случайные величины, такие, что с вероятностью 1 выполнено $|X_i| \le c$. Пусть $\sigma^2 = \sum_{i=1}^n \mathbb{D}\{X_i\}$. Покажите, что для любого $t > 0$

$$\mathbb{P}\left\{ \sum_{i=1}^n X_i > t \right\} \le \exp\left(- \frac{n\sigma^2}{c^2} h\left(\frac{ct}{n\sigma^2} \right) \right),$$

где $h(u) = (1+u)\log(1+u) - u$ для $u \ge 0$.

Указание. Введем $\sigma_i^2 = \mathbb{E}[X_i^2]$ и $F_i = \sum_{r=2}^{\infty} \frac{s^{r-2}\mathbb{E}[X_i^r]}{r!\sigma_i^2}$. Используя разложение для ряда Тейлора $\exp(sX)$, показать, что

$$\mathbb{E}[e^{sX_i}] \le \exp(s^2\sigma_i^2 F_i).$$

Из ограниченности X_i получите оценку

$$F_i \le \frac{\exp(sc) - 1 - sc}{(sc)^2}.$$

Далее воспользуйтесь неравенством Чернова для X_i и минимизируйте правую часть в неравенстве Чернова по s.

22 (Неравенство Бернштейна [41])**.** Докажите, что при выполнении условий предыдущей задачи для любого $\varepsilon > 0$ верно следующее неравенство

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i > \varepsilon\right\} \le \exp\left(-\frac{n\varepsilon^2}{2\sigma^2 + 2c\varepsilon/3}\right).$$

Указание. Покажите, что верно элементарное неравенство

$$h(u) \ge \frac{u^2}{2 + 2u/3}$$

и используйте неравенство Беннетта.

23 (Неравенство Азумы–Хёфдинга)**.** Пусть $\{X_i\}_{i=0}^{\infty}$ последовательность со следующим свойством (см. замечание к задаче 3.8)

$$\mathbb{E}(X_n | X_1, \ldots, X_{n-1}) = X_{n-1},$$

и пусть $Y_i = X_i - X_{i-1}$ соответствующая последовательность приращений (мартингальная разность). Покажите, что если существуют такие $c_i > 0$, что $|Y_i| \le c_i$ для всех i, то

$$\mathbb{P}\left\{\sum_{i=1}^{m} Y_i \ge t\right\} \le 2\exp\left\{\frac{-t^2}{2\sum_{i=1}^{m} c_i^2}\right\}.$$

Указание. Вначале необходимо докажите следующее утверждение. Пусть Y случайная величина, $Y \in [-1, +1]$ и $\mathbb{E}[Y] = 0$. Тогда для любого $t \ge 0$ верно

$$\mathbb{E}[\exp(tY)] \le \exp(t^2/2).$$

Для этого необходимо использовать выпуклость $\exp(tx)$, а именно для $x \in [-1, 1]$ верно

$$\mathrm{e}^{tx} \le \frac{1}{2}(1 + x)\mathrm{e}^{t} + \frac{1}{2}(1 - x)\mathrm{e}^{-t}.$$

Подсчитайте оценку математического ожидания $\mathbb{E}[\mathrm{e}^{tY}]$ используя разложение экспоненты в ряд Тейлора и элементарный факт $(2n)! > 2^n n!$. Затем покажите, что

$$\mathbb{E}\exp\biggl(s\sum_{j=1}^{m}Y_j\biggr)=\mathbb{E}\biggl[\exp\biggl(s\sum_{j=1}^{m-1}Y_j\biggr)\mathbb{E}[\exp(sY_m)|X_1,\dots,X_{m-1}]\biggr].$$

Используйте неравенство Чернова

$$\mathbb{P}[Y_1+\dots+Y_m>t]\le\exp\biggl[-st+\sum_{i=1}^{m}c_i^2s^2/2\biggr].$$

Остается минимизировать правую часть неравенства по s.

24. Пусть независимые одинаково распределенные невырожденные ($\not\equiv$ const) случайные величины $\xi_1,\xi_2,\dots$ с математическим ожиданием m удовлетворяют *условию Крамера*, то есть существует такая окрестность нуля, что для любого λ из этой окрестности

$$\mathbb{E}e^{\lambda\xi}<\infty.$$

Пусть

$$S_n=\xi_1+\dots+\xi_n,\quad \psi(\lambda)=\ln\mathbb{E}\exp(\lambda\xi)$$

и

$$H(a)=\sup_{\lambda}[a\lambda-\psi(\lambda)],\quad a\in\mathbb{R},$$

Покажите, что верно следующее неравенство Дуба

$$\mathbb{P}\biggl\{\sup_{k\ge n}\biggl|\frac{S_k}{k}-m\biggr|>\varepsilon\biggr\}\le 2\exp\biggl(-\min\Bigl[H(m-\varepsilon),H(m+\varepsilon)\Bigr]n\biggr).$$

Указание. Зафиксируйте $n\ge 1$ и положите

$$\kappa=\inf\biggl\{k\ge n:\frac{S_k}{k}>a\biggr\},$$

считая $\kappa = \infty$, если $\frac{S_k}{k} \le a$, $k \ge n$. Пусть $\lambda > 0$ и $\lambda a - \psi(\lambda) \ge 0$. Показать, что

$$\mathbb{P}\Big\{\sup_{k\ge n}\frac{S_k}{k} > a\Big\} = \mathbb{P}\Big\{\frac{S_\kappa}{\kappa} > a,\, \kappa < \infty\Big\}$$
$$\le \mathbb{P}\Big\{\exp\Big(\lambda S_\kappa - \kappa\psi(\lambda)\Big) > \exp\Big(n\lambda - n\psi(\lambda)\Big), \kappa < \infty\Big\}$$
$$\le \mathbb{P}\Big\{\sup_{k\ge n}\exp\Big(\lambda S_k - k\psi(\lambda)\Big) > \exp\Big(n\lambda - n\psi(\lambda)\Big)\Big\}.$$

Затем, необходимо воспользоваться тем фактом, что последовательность случайных величин $Y_k = \exp(\lambda S_k - k\psi(\lambda))$, $k \ge 1$ является мартингалом (см. задачу **??** раздела **??**).

Для момента остановки $\tau = \inf\{k \le n : Y_k \ge \lambda\}$, $\tau = n$ если $\max_{k\le n} Y_k < \lambda$ верна теорема Дуба (см. задачу **??** раздела **??**). Тогда из неравенства Маркова для любого $x > 0$

$$x \cdot \mathbb{P}\Big\{\sup_{k\ge n} Y_k \ge x\Big\} \le \mathbb{E}Y_n.$$

Отсюда

$$\mathbb{P}\Big\{\sup_{k\ge n}\frac{S_k}{k} > a\Big\} \le \exp\Big\{-n\Big(\lambda a - \psi(\lambda)\Big)\Big\}.$$

Рассмотрите случаи $a > m$ и $a < m$.

25. Докажите, что для последовательности независимых одинаково распределенных с.в. $\xi_1, \xi_2, \dots$ с математическими ожиданиями $m = \mathbb{E}\xi_i$, дисперсиями $\mathbb{D}\,\xi_i = d$ и функцией распределения $F(x)$ верна следующая оценка вероятности больших отклонений

$$\mathbb{P}\Big\{\Big|\sum_{i=1}^n \xi_i - nm\Big| \ge n\varepsilon\Big\} \le B_n(\phi(t_0))^n \exp\big(-t_0 n\varepsilon\big),$$

где $\lim_{n\to\infty} B_n = \frac{1}{2}$, $\phi(t) = \int_{-\infty}^{\infty} \mathrm{e}^{tx}\, dF(x)$, $m(t) = \frac{\phi'(t)}{\phi(t)}$, $r(\lambda_0) = \mathrm{e}^{-\lambda_0 c} R(\lambda_0)$, значение t_0 удовлетворяет условию $m(t_0) = \varepsilon$.

Замечание. Если использовать для оценки $\mathbb{P}\Big\{\Big|\sum_{i=1}^n \xi_i - nm\Big| \ge s\Big\}$ неравенство Чебышева, то при $s = \varepsilon\sqrt{n}$ и $s = \varepsilon n$, где ε — некоторая

постоянная, получается разный порядок сходимости вероятности. В отличие от первого случая, оценка при $s = \varepsilon n$ является очень грубой.

Доказательство существования, единственности, а также положительности t_0 может быть найдено в учебнике Коралова–Синая.

Воспользуемся *методом Крамера* вычисления асимптотики вероятностей. Покажите, что верно

$$\mathbb{P}\Big\{\Big|\sum_{i=1}^{n}\xi_i - nm\Big| \ge n\varepsilon\Big\} \le (\phi(t_0)\mathrm{e}^{-t_0\varepsilon})^n \int\limits_{\sum_{i=1}^n x_i > \varepsilon n} \cdots \int dF_{t_0}(x_1)\dots dF_{t_0}(x_n),$$

где $F_t(x) = \frac{1}{\phi(t)}\int_{-\infty}^{x}\mathrm{e}^{tu}dF(u)$ – функция распределения (проверьте это). Для того, чтобы оценить интеграл, рассмотрите случайные величины $\tilde\xi_1, \dots, \tilde\xi_n$ с распределением F_{t_0}, воспользуйтесь центральной предельной теоремой, чтобы показать, что $B_n \to 1/2$ при $n \to \infty$.

26. Задана последовательность независимых одинаково распределенных случайных величин $\xi_1, \xi_2, \dots$ и $S_n = \xi_1 + \dots + \xi_n$. Показать, что в бернуллиевском случае ($\mathbb{P}\{\xi_1 = 1\} = p$, $\mathbb{P}\{\xi_1 = 0\} = 1 - p$)

а) при $p < x < 1$

$$\lim \frac{1}{n}\ln \mathbb{P}\{S_n \ge nx\} = -H(x),$$

$$H(x) = x\ln\frac{x}{p} + (1-x)\ln\frac{1-x}{1-p};$$

б) при $x_n = n(x-p)$ и при $p < x < 1$

$$\mathbb{P}\{S_n \ge np + x_n\} = \exp\left\{-nH\left(p + \frac{x_n}{n}\right)(1 + o(1))\right\};$$

в) при $x_n = a_n\sqrt{np(1-p)}$ с $a_n \to \infty$, $\frac{a_n}{\sqrt{n}} \to 0$

$$\mathbb{P}\{S_n \ge np + x_n\} = \exp\left\{-\frac{x_n^2}{2np(1-p)}(1 + o(1))\right\}.$$

г) Обобщите предыдущие три пункта на случай, когда вместо $\frac{S_n}{n}$ под знаком вероятности будет стоять $\sup\limits_{k \ge n}\frac{S_k}{k}$.

Указание. Для доказательства первого пункта используйте следующий результат (теорема Чернова). Пусть $S_n = \sum_{i=1}^n \xi_i$, где ξ_i, $i = 1, \ldots, n$ независимые одинаково распределенные простые случайные величины с

$$\mathbb{E}\xi_1 \leq 0 \quad \text{и} \quad \mathbb{P}\{\xi_1 > 0\} > 0,$$

$$\inf_\lambda \phi(\lambda) = \rho, \quad 0 < \rho < 1, \quad \phi(\lambda) = \mathbb{E}\mathrm{e}^{\lambda\xi_1}.$$

Тогда

$$\lim \frac{1}{n} \ln \mathbb{P}\{S_n \geq 0\} \to \ln \rho \quad \text{при } n \to \infty.$$

27. а) Сравнить оценки вероятности отклонений выборочного среднего от теоретического среднего для последовательности одинаково распределенных случайных величин $\xi_1, \ldots, \xi_n$ с распределением Бернулли с вероятностью успеха p (не предполагая равномерную отделимость p от нуля, в частности, допуская $p = \lambda/n$, где λ постоянна), получаемые с помощью неравенства Хёфдинга, Бернштейна, Спокойного, Буске и неравенства больших уклонений из предыдущих задач.

б) 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 5% и с вероятностью не меньшей 0.99. Какой способ решения является более точным: с помощью использования центральной предельной теоремы и неравенства Берри-Эссеена или полученный с использованием неравенств концентрации меры, больших уклонений (см. задачи выше в разделе, неравенства Спокойного, Буске, Бернштейна)?

Замечание. *Неравенство О. Буске.* Пусть $X_1, \ldots, X_n$ – последовательность независимых случайных величин с распределением $\mathbb{P}$, принимающих значения из полного сепарабельного метрического пространства $\mathcal{X}$. Пусть $Z = f(X_1, \ldots, X_n)$, где $f : \mathcal{X}^n \to \mathbb{R}$ измеримая функция, обозначим $Z_i = f(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$.

Пусть Z удовлетворяет

$$\sum_{i=1}^k (Z - Z_k) \leq Z$$

и существуют такие случайные величины Y_i, что

$$Y_i \le Z - Z_i \le 1, \quad Y_i \le a \text{ п.н.}$$

для некоторого $a > 0$ и $\mathbb{E}_i Y_i \ge 0$, где $\mathbb{E}_i X = \mathbb{E}_i[X|X_1, \dots, X_{i-1}, X_{i+1}, X_n]$. Также пусть существует σ, такое что п.н. $\sigma^2 \ge \frac{1}{n}\mathbb{E}_i Y_i^2$.

Тогда для всех $x \ge 0$ верно

$$\mathbb{P}[Z \ge \mathbb{E}[Z] + x] \le \exp\left(-vh\left(\frac{x}{v}\right)\right)$$

при $v = n\sigma^2 + (1+a)\mathbb{E}[Z]$ и $h(x) = (1+x)\ln(1+x) - x$, а также

$$\mathbb{P}\left[Z \ge \mathbb{E}[Z] + \sqrt{2vx} + \frac{x}{3}\right] \le \exp(-x).$$

См. также книгу Stéphane Boucheron, Gábor Lugosi, Pascal Massart «Concentration Inequalities: A Nonasymptotic Theory of Independence», Oxford University Press, 2013.

Неравенство В. Г. Спокойного. Пусть Y_i одинаково распределенные случайные величины с распределением $\mathbb{P}_{u^*}$ которое принадлежит экспоненциальному семейству следующего вида

$$p_\nu(y) = p(y)\exp\{y\nu - d(\nu)\},$$

где $d(\nu)$ заданная выпуклая функция на множестве параметров $\Theta \subset \mathbb{R}$, $p(y)$ — неотрицательная функция на множестве значений случайной величины, $d(\nu)$ дважды непрерывно дифференцируема и для всех u верно $d''(\nu) > 0$. Тогда для всех $x > 0$ верно следующее неравенство

$$\mathbb{P}_{v^*}(L(\tilde{v}, v^*) > x) = \mathbb{P}_{v^*}\left\{n\mathcal{KL}(\tilde{\nu}, \nu^*) > x\right\} \le 2\exp(-x),$$

где $\mathcal{KL}(\nu, \nu^*)$ — расстояние Кульбака–Лейблера (см. также задачу 14 раздела 2)

$$\mathcal{KL}(\nu, \nu^*) = \int \ln \frac{d\mathbb{P}_\nu(y)}{d\mathbb{P}_{\nu^*}} d\mathbb{P}_\nu(y),$$

где $L(\nu, \nu^*)$ — log-отношение правдоподобий моделей с параметрами ν и ν^*, которое определяется следующим образом:

$$L(\nu, \nu^*) = L(\nu) - L(\nu^*),$$

где $L(\nu)$ — log-правдоподобие модели с параметрами u

$$L(\nu) = n^{-1} \sum_{i=1}^{n} \log f_\nu(Y_i),$$

здесь в случае непрерывного распределения $f_\nu(Y_i)$ — плотность распределения случайной величины Y_i, в случае дискретного распределения вероятность получить наблюдаемое Y_i; а $\tilde{\nu}$ — оценка параметров методом максимального правдоподобия, т.е.

$$\tilde{\nu} = \arg\max_{\nu \in \Theta} L(\nu).$$

Расстояние Кульбака–Лейблера $\mathcal{KL}(P\|Q)$ характеризует "вложенность" распределения P в Q.

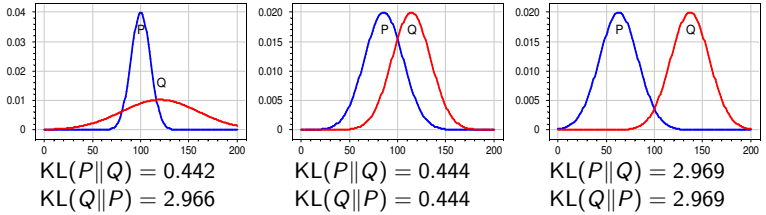


Рис. 6: Иллюстрация свойства "вложенность" метрики $\mathcal{KL}$.

Теорема Берри–Эссеена. (В.В. Сенатов) Пусть $\xi_1, \xi_2 \dots$ независимые одинаково распределенные с.в., причем $\mathbb{E}\xi_i = m$ $\mu^3 = \mathbb{E}|\xi_i - \mathbb{E}\xi_i|^3 < \infty$, $\sigma^2 = \mathbb{D}\xi_i$. Близость с.в. $\frac{\sum_{i=1}^n \xi_i - nm}{\sigma\sqrt{n}}$ к стандартной нормально распределенной с.в. (согласно ц.п.т.) в смысле близости их функций распределения определяется неравенством Берри–Эссена

$$\sup_x \left| \mathbb{P}\Big(\frac{\sum_{i=1}^n \xi_i - nm}{\sigma\sqrt{n}} < x \Big) - \Phi(x) \right| \le \frac{C_0 \mu^3}{\sigma^3 \sqrt{N}},$$

где $0.4 < C_0 < 0.7056$, $\Phi(x) = \int_{-\infty}^{x} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\, dt$.

Неравенство Берри-Эссеена дает неулучшаемый в общем случае результат.

Теперь пусть $X_1, X_2 \dots$ — независимые случайные величины, имеющие одно и то же решетчатое распределение с шагом $h > 0$. Очевидно, что распределение их суммы будет решетчатым с тем же шагом h а функция распределения нормированной суммы

$$\frac{X_1 + \dots + X_n - na}{\sigma\sqrt{n}}$$

будет решетчатым с шагом $h_n = h/(\sigma\sqrt{n})$. Обозначим решетку, на которой сосредоточено распределение нормированной суммы, через D_n. Ясно, что на полуинтервале $[-1, 1)$ находится не более $2/h_n$ точек из решетки D_n. В силу центральной предельной теоремы для эмпирического распределения верно

$$F_n(1) - F_n(-1) \to \Phi(1) - \Phi(-1) = 0.6826\dots$$

при $n \to \infty$, поэтому, начиная с некоторого n, сумма скачков F_n на полуинтервале $[-1, 1)$ будет не меньше 0.5. Отсюда сразу следует, что при таких n максимальный скачок будет не меньше $0.25h/(\sigma\sqrt{n})$. Так как нормальная функция распределения Φ непрерывна, а приблизить разрывную функцию F_n непрерывной функцией с точностью, превосходящей половину максимального скачка, невозможно, то

$$\sup_x |F_n(x) - \Phi(x)| \geq 0.125\frac{h}{\sigma\sqrt{n}}.$$

Эти рассуждения показывают, что порядок по n оценки теоремы Берри-Эссеена является правильным. См. В.В. Сенатов Центральная предельная теорема. Точность аппроксимации и асимптотическое разложение. М. Книжный дом ЛИБРОКОМ, 2009.

28 (В.Г. Спокойный)**.** Пусть ξ – стандартный нормальный вектор в $\mathbb{R}^p$. Тогда для любого $u > 0$ выполнено

$$\mathbb{P}(\|\xi\|^2 > p + u) \leq \exp\bigl\{-(p/2)\psi(u/p)\bigr\},$$

где

$$\psi(t) = t - \ln(1 + t).$$

Пусть $\psi^{-1}(\cdot)$ обратная функция к $\psi(\cdot)$.

а) Покажите, что для любого x верно

$$\mathbb{P}(\|\xi\|^2 > p + \psi^{-1}(2x/p)) \leq \exp(-x).$$

И, в частности, при $\kappa = 6.6$

$$\mathbb{P}(\|\xi\|^2 > p + \max(\sqrt{\kappa x p}, \kappa x)) \le \exp(-x).$$

Можно ли уменьшить константу κ?

б) Обобщите результаты предыдущего пункта на случай, если компоненты вектора являются независимыми субгауссовскими случайными величинами с параметром C, то есть для любого $\lambda > 0$, $i = 1, \dots, p$ выполнено

$$\mathbb{E} \exp(\xi_i \lambda) \le \exp(C\lambda^2/2).$$

Указание. Показать, что

$$\ln \mathbb{E} \exp(\mu \|\xi\|^2/2) = -0.5p \ln(1-\mu).$$

Из неравенства Чернова получите

$$\mathbb{P}(\|\xi\|^2 > p + u) \le \exp\bigl\{-\mu(p+u)/2 - (p/2)\ln(1-\mu)\bigr\}.$$

Минимизируйте правую часть по μ. Затем используйте $x - \ln(1+x) \ge a_0 x^2$ при $x \le 1$ и $x - \ln(1+x) \ge a_0 x$ при $x > 1$ и $a_0 = 1 - \ln 2 \ge 0.3$.

Замечание. См. также Spokoiny V. Basics of Modern Parametric Statistics. 2012, http://premolab.ru/sites/default/files/stat.pdf.

29. Пусть $\xi_0, \dots, \xi_k$ — независимые одинаково распределенные случайные величины. Обозначим за $\xi_{[i]}$ совокупность случайных величин $\xi_0, \dots, \xi_{i-1}$

а) Пусть $\Delta_i = \Delta_i(\xi_{[i]})$ неслучайная измеримая функция от $\xi_{[i]}$, такая, что

$$\mathbb{E}\left[\exp\left(\frac{\Delta_i^2}{\sigma^2}\right) \mid \xi_{[i-1]}\right] \le \exp(1).$$

Покажите, что для любого $k \ge 0$ и $\Omega > 0$ верно

$$\mathbb{P}\left(\sum_{i=0}^{k} c_i \Delta_i^2 \ge (1+\Omega) \sum_{i=0}^{k} c_i \sigma^2\right) \le \exp(-\Omega),$$

где $c_0, \dots, c_k$ — последовательность положительных коэффициентов.

б) Пусть Γ_k и η_k неслучайные измеримые функции от $\xi_{[k]}$ такие что

- $\mathbb{E}[\Gamma_i|\xi_{[i-1]}] = 0$,
- $|\Gamma_i| \le c_i\eta_i$, где c_i положительная неслучайная константа,
- $\mathbb{E}\left[\exp\left(\frac{\eta_i^2}{\sigma^2}\right)|\xi_{[i-1]}\right] \le \exp(1)$.

Покажите, что для всех $k \ge 0$ и $\Omega \ge 0$ верно

$$\mathbb{P}\left(\sum_{i=0}^{k}\Gamma_i \ge \sqrt{3\Omega}\sigma\sqrt{\sum_{i=0}^{k}c_i^2}\right) \le \exp(-\Omega).$$

Замечание. При доказательстве пункта а) необходимо использовать выпуклость экспоненты и линейность математического ожидания, затем неравенство Маркова. Пункт б) является следствием леммы 2 из статьи Lan G., Nemirovski A. and Shapiro A. Validation analysis of mirror descent stochastic approximation method // Mathem. Programming Serie A. 2012. V. 134(2). P. 425–458.

30 (Неравенства Эфрона-Стайна и МакДиармида)**.** а) Пусть X_i, $i = 1, \dots, n$ произвольные независимые (не обязательно одинаково распределенные) случайные величины, принимающие значения из $\mathcal{X}$ и пусть $g : \mathcal{X}^n \to \mathbb{R}$ измеримая функция n переменных. Покажите, что для случайной величины $Z = g(X_1, \dots, X_n)$ верно

$$\mathbb{D}(Z) \le \sum_{i=1}^{n}\mathbb{E}\big[(Z - \mathbb{E}_i Z)^2\big],$$

где $\mathbb{E}_i Z = \mathbb{E}[Z|X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n]$.

б) Неравенство Эфрона-Стайна. Пусть $X_1', \dots, X_n'$ — независимые копии $X_1, \dots, X_n$ и

$$Z_i' = g(X_1, \dots, X_i', \dots, X_n).$$

Покажите, что верно неравенство

$$\mathbb{D}(Z) \le \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}[(Z - Z_i')^2].$$

в) Неравенство Эфрона-Стайна в случае функций с ограниченными разностями. Функция $g : \mathcal{X}^n \to \mathbb{R}$ является функцией с ограниченными разностями, если для некоторых $c_1, \ldots, c_n$ выполнено

$$\sup_{x_1,\ldots,x_n;\, x_i' \in \mathcal{X}} |g(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \le c_i,$$
$$1 \le i \le n.$$

Выпишите неравенство Эфрона-Стайна для случая функций с ограниченными разностями.

г) Докажите неравенство МакДиармида для функции g с ограниченными разностями (см. предыдущий пункт), а именно, что для любого $\varepsilon > 0$ верно

$$\mathbb{P}\Big\{\big|g(X_1, \ldots, X_n) - \mathbb{E}g(X_1, \ldots, X_n)\big| > \varepsilon\Big\} \le 2\exp\left\{-\frac{2\varepsilon^2}{nC^2}\right\},$$

где $C^2 = \sum_{i=1}^n c_i^2$.

31 (Лемма Джонсона-Линденштаусса)**.** Лемма гласит, что если задан произвольный набор из n точек в многомерном (D-мерном) евклидовом пространстве, то существует линейное вложение этих точек в d-мерное евклидово пространство, такое что все попарные расстояния сохраняются с точностью до множителя $1 \pm \varepsilon$, если d пропорционально $(\log n)/\varepsilon^2$.

Пусть A — конечное подмножество $\mathbb{R}^D$ размерности n. И для некоторого $v \ge 0$, случайные величины $X_{i,j}$, $i = 1, \ldots, n$, $j = 1, \ldots, D$ независимы, одинаково распределены и являются субгауссовскими с параметром v (см. задачу 28), причем $\mathbb{E}X_{i,j} = 0$, $\mathbb{E}X_{i,j}^2 = 1$. При заданном $\varepsilon \in (0, 1)$ отображение $f : \mathbb{R}^D \to R^d$ называется ϵ-изометрией на A если для каждой пары $a, a' \in A$ выполняется

$$(1 - \varepsilon)\|a - a'\|^2 \le \|f(a) - f(a')\|^2 \le (1 + \varepsilon)\|a - a'\|^2.$$

Пусть $d \ge 32v\varepsilon^{-2}\log(n/\sqrt{\delta})$, где $\delta \in (0, 1)$. Покажите, что тогда с вероятностью не меньшей $1 - \delta$, отображение $W : \mathbb{R}^D \to \mathbb{R}^d$, где $W_i(\alpha) = \frac{1}{\sqrt{d}} \sum_{j=1}^D \alpha_j X_{i,j}$ для всех $\alpha \in \mathbb{R}^D$, $i \in \{1, \ldots, d\}$, является ε-изометрией на A.

Замечание. Замечательным фактом является то, что результат не зависит от размерности D, которая может быть даже бесконечной!

Указание. Идея доказательства заключается в использовании специальной случайной линейной вектор-функции $W(\alpha)$ и проверке ее на ε-изометрию. Основные шаги доказательства

а) Проверьте, что $\mathbb{E}[\|W(\alpha)\|^2] = \|\alpha\|^2$, где $\alpha \in \mathbb{R}^D$.

б) Убедитесь, что доказательство того, что W является ε-изометрией эквивалентно тому, что с вероятностью не меньшей $1-\delta$

$$\sup_{\alpha \in T} \big| \|W(\alpha)\|^2 - 1| \leq \varepsilon,$$

где T —подмножество единичной сферы S в $\mathbb{R}^D$ следующего вида

$$T = \left\{ \frac{a - a'}{\|a - a'\|} : a, a' \in A, a \neq a' \right\}.$$

в) Для того, чтобы это показать, докажите, что $\sqrt{d} W_i(\alpha)$ — субгауссовская случайная величина (см. задачу 28) с параметром v.

г) Воспользуйтесь следующим фактом (см. теорему 2.1 из Boucheron S., Lugosi G., Massart P. Concentration Inequalities: A Nonasymptotic Theory of Independence, Oxford University Press, 2013): если случайная величина X суб-гауссовская с параметром $4C$, тогда $\mathbb{E}[X^{2q}] \leq q! C^q$ для $q > 1$. Получите, что при $q \geq 2$

$$\mathbb{E}[W_i(\alpha)^{2q}] \leq \frac{q!}{2} (4v)^q.$$

д) С помощью неравенства Бернштейна (задача 22) получите

$$\mathbf{P} \left\{ \sup_{\alpha \in T} | \|W(\alpha)\|^2 - 1| \geq 4 \sqrt{\frac{v \log(n/\sqrt{\delta})}{d}} + \frac{8v \log(n/\sqrt{\delta})}{d} \right\} \leq \delta.$$

е) Подставьте в правую часть неравенства в выражении вероятности условие на d из формулировки леммы и получите утверждение леммы.

32 (Теорема Дворецкого)**.** См. также формулировку задачи 16 раздела 6. Для каждого натурального k и любого $\varepsilon > 0$ найдется такое n, что всякое n-мерное нормированное пространство X имеет

k-мерное подпространство, расстояние от которого до l_2^k по метрике Банаха-Мазура не превосходит $1+\varepsilon$, то есть можно найти векторы $x_1, \ldots, x_k \in X$ такие, что

$$\left(\sum_{i=1} k|a_i|^2\right)^{1/2} \leq \|\sum_{i=1}^{k} a_i x_i\|_2 \leq (1+\varepsilon)\left(\sum_{i=1}^{k} |a_i|^2\right)^{1/2}$$

для любой последовательности скаляров $a_1, \ldots, a_k$.

Указание. Подход состоит в том, чтобы выбрать k-мерное подпространство X случайным образом. Прежде этим надо выбрать подходящую вероятностную меру. Что может быть сделано с помощью теоремы Фрица Джона. Последняя уверждает, что если существует базис $x_1, \ldots, x_n$ пространства X, который не слишком далек от ортонормированного в смысле, что

$$\left(\sum_{i=1}^{n} |a_i|^2\right)^{1/2} \leq \left\|\sum_{i=1}^{n} a_i x_i\right\|_2 \leq \sqrt{n}\left(\sum_{i=1}^{n} |a_i|^2\right)^{1/2}$$

для всякой последовательности скаляров $a_1, \ldots, a_n$. Тогда берется естественная мера грассманиана $G_{n,k}$ относительно этого базиса. Также необходимо следствие неравенства Леви: пусть $f : S^n \to \mathbb{R}$ — функция со средним значением M и пусть $A \in S^n$ — множество всех точек x, для которых $f(x) \leq M$, тогда вероятность того, что случайно выбранная точка S^n удалена от A более, чем на ε не превосходит $\sqrt{\pi/2}\exp(-\varepsilon^2 n/2)$. Заменив f на $-f$ найдем также, что почти каждая точка y близка x с $f(x) \geq M$. Положим теперь $f(a_1, \ldots, a_m) = \|\sum_{i=1}^{n} a_i x_i\|_2$. Поскольку f в достаточной мере непрерывна и почти каждая точка y близка некоторой точке $x \in A$, заключаем, что $f(y)$ не намного больше M. Точно так же $f(y)$ не намного меньше M для большинства точек y.

См. В.Д. Мильман, Новое доказательство теоремы А. Дворецкого о сечениях выпуклых тел, Функц. анализ, Т.5, № 4, 1971, С.28–37.

Замечание. Еще одна формулировка теоремы: каждое n-мерное симметричное выпуклое тело имеет k-мерное центральное сечение, которое содержит k-мерный эллипсоид B и содержится в $(1+\epsilon)B$, то есть само является почти эллипсоидальным.

Compressed sensing. «Сжатие измерений» понимается как метод экономного восстановления неизвестной функции, заданной на конечном множестве мощности m, то есть вектора $u \in \mathbb{R}^m$ по информации, полученной измерениями скалярных произведений (u, ϕ_j), $\phi_j \in \mathbb{R}^m$, $j = 1, \ldots, n$, причем $n \ll m$. Пусть Φ — матрица со строками $\phi_j \in \mathbb{R}^m$, $j = 1, \ldots, n$. Предполагается, что о разреженности вектора известно, что $\|u\|_0 = |\{i : u_i \leq 0\}| \leq t$. Целью является

а) построение алгоритма аппроксимации функции u по информации $y = ((u, \phi_1), \ldots, (u, \phi_n)) \in \mathbb{R}^n$, то есть

$$\min \|u\|_0 \text{ при условии, что } \Phi v = y;$$

б) построение измеряющего множества векторов $\phi_j \in \mathbb{R}^m$, $j = 1, \ldots, n$, то есть описание матриц Φ.

Первую задачу было предложено (D. L. Donoho, M. Elad, V. N. Temlyakov, “Stable recovery of sparse overcomplete repre- sentations in the presence of noise”, IEEE Trans. Inform. Theory, 52:1 (2006), 6–18.) решать с помощью релаксации к выпуклой задаче

$$\min \|u\|_1 \text{ при условии, что } \Phi v = y.$$

В 1977 году Кашиным было доказано, что для любой пары (m, n), где $m \geq n$, существует такое подпространство V размерности большей или равной $m - n$, такое, что для любого $x \in V$:

$$\|x\|_2 \leq C \left(\frac{1 + \log(m/n)}{n} \right)^{1/2} \|x\|_1$$

(см. Кашин Б.С., Изв. АН СССР, серия матем., Т.41, 1977, 334–351). См. также Кашин Б.С., Темляков В.Н. Замечание о задаче сжатого измерения // Математические заметки. 2007. Т. 82, № 6, С. 829–837, E.J. Candes, T. Tao, “Decoding by linear programming”, IEEE Trans. Inform. Theory, Vol.51, № 12, 2005.

4. Теория информации и кодирование

1. Имеется 12 монет, из них ровно одна фальшивая, которая может быть как легче, так и тяжелее настоящих. Предложите алгоритм взвешиваний на чашечных весах без гирь, выявляющий за 3 взвешивания подделку, а также определяющий является ли фальшивая монетка тяжелее или легче настоящих.

Указание. Любая из 12 монет может быть фальшивой, при этом быть легче, либо тяжелее настоящих. Для решения требуется с помощью взвешиваний получить $\log_2 4$ бит информации (по Хартли), так как каждый исходов может быть закодирован двоичным словом длины 24. С другой стороны, каждое взвешивание имеет три возможных исхода: выше левая чаша весов, выше правая или они в равновесии. То есть каждое взвешивание дает не более $\log 3$ бит информации. Соответственно, требуемое число взвешеваний не может быть меньше $\frac{\log 24}{\log 3} = 3$.

Для решения этой задачи и последующих полезно ознакомиться с книгой Н.К. Верещагин, Е.В. Щепин Информация, кодирование и предсказание. – М.: МЦНМО, 2012, а также см. метод Дайсона в статье Г. Шестопала "Как обнаружить фальшивую монетку" в журнале Квант 1979, номер 10.

2. Патриций решил устроить праздник и для этого приготовил 240 бочек вина. Однако к нему пробрался недоброжелатель, который подсыпал яд в одну из бочек. Про яд известно, что человек, его выпивший, умирает в течение (не «через»!) 24 часов. До праздника осталось два дня, то есть 48 часов. У патриция есть пять собак, которыми он готов пожертвовать, чтобы узнать в какой именно бочке яд. Как патрицию вычислить отравленную бочку?

3 (Цена информации)**.** Имеется неизвестное число от 1 до n, $n \geq 2$. Разрешается задавать любые вопросы с ответами ДА/НЕТ. При этом при ответе ДА игрок платит 1 рубль, а при ответе НЕТ - 2 рубля. Сколько необходимо и достаточно заплатить для отгадывания числа?

4 (Аксиоматическое определение энтропии)**.** Рассмотрим множество функций, заданных на единичном симплексе. Докажите, что существует единственная (с точностью до множителя) функция, удовлетворяющая нижеперечисленным требованиям. Она имеет вид

$H(P) = -\sum_{i=1}^{n} p_i \log p_i$ и используется в качестве количественной характеристики меры неопределенности.

а) Значение функции $H(P)$ не меняется при перестановке чисел $p_{a_1}, ..., p_{a_n}$,

б) $H(P)$ непрерывная функция,

в) выполняется равенство

$$H(p_1, ..., p_n) = H(p_1 + p_2, p_3, ..., p_n) + (p_1 + p_2)H(\tfrac{p_1}{p_1+p_2}, \tfrac{p_2}{p_1+p_2}).$$

То есть неопределенность в исходе опыта не зависит от того, осуществляется ли выбор среди всех возможных альтернатив одномоментно или в несколько этапов.

5. а) Ф.М. Достоевский решил изменить своим привычкам и отправился на скачки. У него есть предварительные (априорные) данные о том, какие шансы на победу имеет каждая из восьми лошадей-участниц: $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})$. Оцените энтропию, которая содержится в таких данных.

б) Сравните результат со случаем, когда все исходы равновероятны. Какое из двух респределений содержит больше информации?

в) Докажите в общем случае, что из всех дискретных распределений на конечном множестве A, наибольшей энтропией обладает равномерное.

6. Пусть для некоторого населенного пункта вероятность того, что 15 июня будет дождь, равна 0.4, а вероятность того, что дождя не будет, равна 0.6. Для этого же пункта вероятность дождя 15 октября равна 0.8, а вероятность отсутствия осадков равна 0.2. Предположим, что определенный метод прогноза погоды 15 июня оказывается верным в 3/5 всех тех случаев, когда предсказывается дождь и в 4/5 тех случаев, в которых прогнозируется отсутствие осадков. Применительно к погоде на 15 октября этот метод оказывается правильным в 9 из 10 случаев, когда предсказывается дождь, и в половине случаев, когда предсказывается его отсутствие. В какой из указанных двух дней прогноз дает нам больше информации о реальной погоде?

7 (Задача о шляпах. Тодд Эберт, 1998)**.** Трех игроков отводят в комнату, где на них надевают (случайно и независимо) белые и черные шляпы. Каждый видит цвет других шляп и должен написать на бумажке одно из трех слов: «белый», «черный», «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? Оптимальна ли предложенная Вами стратегия? Решите эту же задачу, если игроков $n = 2^m - 1$ ($m \in \mathbb{N}$).

Указание. Воспользуйтесь кодом Хэмминга. Докажем для случая трех игроков, что вероятность выигрыша не может превышать $3/4$.

Единственная информация, которой владеет i-й игрок — это цвета шляп двух других. Поэтому стратегия для i-го игрока должна зависеть только от этих двух цветов. В каждом случае имеется три варианта ответа для игрока: 0, 1 или «пас», т.е. всего 3^{12} различных стратегий. Поскольку есть 8 вариантов расположения шляп на игроках, более выгодная стратегия должна обеспечивать выигрыш в 7 вариантах. Тогда один из игроков должен угадать свой цвет в 3 ситуациях. Значит, имеются для него ответы $\alpha_{i_1 j_1}$, $\alpha_{i_2 j_2}$, не являющиеся пасами. Но тогда в ситуациях $\overline{\alpha_{i_1 j_1}} i_1 j_1$ и $\overline{\alpha_{i_2 j_2}} i_2 j_2$ он ошибется, что противоречит предположению о 7 выигрышных ситуациях.

Таким образом, максимальная вероятность выигрыша не превышает $3/4$. Обобщите это рассуждение на случай $n = 2^m - 1$.

8 (Граница Эдгара Гилберта)**.** Для обеспечения помехоустойчивости кода при передачи информации, вместо исходного k-буквенного сообщения, передается n-буквенное ($n > k$). Возникает вопрос, при каких значениях параметров $q = |\Sigma|$ - размер алфавита, k, n, l существует код $F : \Sigma^k \to \Sigma^n$, исправляющий l ошибок, и как его построить? Достаточное условие существования кода дает так называемая *граница Гилберта*, которая описана ниже.

Пространство кодовых слов содержит всего q^n элементов. Назовем шаром радиуса e с центром в слове $x \in \Sigma^n$ множество слов, отличающихся от x не более чем в l позициях. Будем выбирать кодовые слова одно за другим произвольным образом, следя за тем, чтобы расстояния (по Хэммингу) между ними были больше $2l$. В какой-то момент пространство окажется полностью покрытым шарами, каждый из которых состоит из $V_q(2l, n)$ элементов, где $V_q(2l, n)$ -

объем шара радиуса $2l$. Тогда при выполнении условия:

$$(q^k - 1)V_q(2e, n) < q^n$$

существует код с параметрами q, k, n, l. Используя формулу Стирлинга ($k! \approx (k/)^k$), оцените число элементов в шаре $V_q(2l, n)$. В случае $q = 2$ последняя оценка легко получается применением оценок больших уклонений биномиальной случайно величины.

С помощью идеи *случайного кодирования* можно посторить код, с точностью до двух битов, реализующий границу Гилберта с большой вероятностью. Для этого случайно и независимо выбираются N кодовых слов $\xi_1, \ldots, \xi_N$ в пространстве Σ^n.

Найдите достаточное условие на параметры q, N, n, e, при котором среди сгенерированных кодовых слов менее половины в среднем (усреднение по выбору случайного кода) будут иметь "ближайшего соседа" на расстоянии, не превышающем $2l$.

Замечание. В контексте это задачи полезно познакомиться с брошюрой Ромащенко А., Румянцев А., Шень А. Заметки по теории кодирования. – М.: МЦНМО, 2011.

9. Пусть $\{X_i\}_{i=1}^n$ – независимые в совокупности одинаково распределенные случайные величины с распределением $P = \{p_a\}$, на конечном множестве $A = \{a\}$. Доказать, что

$$-\tfrac{1}{n}\sum_{i=1}^n \log\Big(P(X_i)\Big) \xrightarrow[n\to\infty]{\mathbb{P}} H(P)$$

Или, другими словами, для любых $\delta, \varepsilon > 0$ найдется n_0 такое что для всех $n \geq n_0$:

$$\mathbb{P}(|-\tfrac{1}{n}\sum_{i=1}^n \log P(X_i) - H(P)| < \delta) > 1 - \varepsilon$$

Указание. Воспользуйтесь тем, что $-\mathbb{E}\log P(X) = H(P)$.

Замечание. При достаточно больших значениях n можно определить множество *типичных последовательностей* или *слов*, энтропия которых близка к истинной энтропии распределения P. Вероятность появления слова w

$$p_w = p_{x_1}...p_{x_n} = 2^{-n(-\frac{1}{n}\sum_{i=1}^n \log P(X_i))}.$$

Множеством δ-типичных n-буквенных слов в модели, где буквы появляются независимо, назовем $T_\delta^{(n)}$:

$$T_\delta^{(n)} = \{w : 2^{-n(H(P)+\delta)} < p_w < 2^{-n(H(P)-\delta)}\}$$

10 (Асимптотическая равнораспределенность)**.** Докажите, что:

а) мощность множества типичных слов ограничено:

$$|T_\delta^{(n)}| \le 2^{n(H(X)+\delta)};$$

б) $|T_\delta^{(n)}| \ge (1-\varepsilon)2^{n(H(X)+\delta)}$ для достаточно больших n;

в) вероятность нетипичности w: $\mathbb{P}\{w \notin T_\delta^{(n)}\} \to 0,\ n \to \infty$.

Замечание. Идея о типичных последовательностях лежит в основе кодирования. Например, δ-типичные n-буквенные слова кодируются при помощи двоичных последовательностей длины $n(H(X)+\delta)$, нетипичные отбрасываются или представляются одним и тем же добавочным символом. Очевидно, что при декодировании (восстановлении) вероятность ошибки не превысит ε.

11. Рассмотрим n-буквенные слова, порождаемые следующей моделью: появление каждой буквы $x_i \in A$ ($|A| = m$ - алфавит) не зависит от контекста и подчиняется закону распределения $P = \{p_a, a \in A\}$. Всего существует $2^{n \log m}$ таких слов, поэтому каждое из них может быть закодировано при помощи $n \log m$ бит информации. Если же предполагать, что распределение букв P не является равномерным, то найдется лучший способ кодирования. Предложите такой способ.

Замечание. Во многих приложениях код должен быть не только однозначно декодируем, но и *оптимален* в смысле минимальности средней длины. Необходимым (а в случае префиксных кодов и достаточным) условием для выполнения первого требования является *неравенство Крафта – Макмиллана*: пусть $l(a)$ - длина кодового слова для буквы $a \in A$, тогда $\sum_{a\in A} 2^{-l(a)} \le 1$.

Среди всех кодов, удовлетворяющих неравенству Крафта – Макмиллана найдется код минимальной средней длины: $\sum_{a\in A} p_a l(a) \to \min$, при условии $\sum_{a\in A} 2^{-l(a)} \le 1$. Для этого можно, например, воспользоваться методом множителей Лагранжа:

$$\sum_{a\in A} p_a l(a) + \lambda\Big(\sum_{a\in A} 2^{-l(a)} - 1\Big) \to \min\,.$$

Решением является набор $\{l_{opt}(a) = -\log p(a), a \in A\}$, а величина $-\log p(a)$ также называется *собственной информацией*. Тогда минимальная средняя длина кода $-\sum_{a\in A} p(a)\log p(a) = H(P)$ это ни что иное, как энтропия.

12 (Оценка энтропии марковской цепи)**.** Приведем еще одну модель генерирования слов над алфавитом $A = \{a_1, ..., a_m\}$. Текст моделируется стационарной конечной цепью Маркова, порождающей слова вида $X_1, \ldots, X_n \in \{A\}^n$. Вероятность появления j-й буквы зависит только от того, какая буква стоит перед ней: $\mathbb{P}(X_k = a_j | X_{k-1} = a_i) = p_{ij} > 0$. Стационарность означает, что $\forall k:\ P(X_k = a_j) = p_j$, причем, пользуясь формулой условной вероятности, p_j можно представить как $p_j = \sum_{i=1}^m p_i p_{ij}$.

Энтропия X_k при фиксированной $(k-1)$-й букве определяется как $H(X_k|X_{k-1} = a_i) = -\sum_{j=1}^m p_{ij}\log p_{ij}$, а условная энтропия X_k при условии, что X_{k-1} станет известным перед генерацией X_{k-1}, определяется как $H = H(X_k|X_{k-1}) = -\sum_{i=1}^m p_i \sum_{j=1}^m p_{ij}\log p_{ij}$. Энтропия всей цепочки случайных величин в силу марковского свойства равна

$$H(X_1, ..., X_n) = H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_{n-1}) \sim nH.$$

Пусть $W_n = (X_1, ..., X_n)$ – некоторое слово, порождаемое описанной моделью, докажите что:

а) $\frac{-\log \mathbb{P}(W_n)}{n} \xrightarrow{p} H$, т.е. все слова W_n могут быть разбиты на два множества: для первого множества *типичных* слов $|\frac{-\log P(W_n)}{n} - H| < \delta(n)$, для второго – сумма вероятностей элементов сходится к 0 при $n \to \infty$;

б) Обозначим через $M_\alpha(n)$ максимальное количество значений W_n с суммарной вероятностью не более α. Докажите, что

$$\forall \alpha:\ \frac{M_\alpha(n)}{n} \to H, \quad n \to \infty.$$

Замечание. Пусть длина алфавита $m = 2^N$, а длина слова – n. Пункт б) утверждает, что найдется код, с помощью которого с высокой вероятностью исходное сообщение W_n может быть передано в $N/H \geq 1$ раз более коротким сообщением, чем при кодировании при помощи двоичных слов, когда каждому слову W_n ставится в однозначное соответствие двоичная цепочка длины $2^n = 2^{nN}$.

13 (Неравенство Пинскера)**.** *Относительной энтропией* двух распределений P и Q на множестве A (или *расстоянием Кульбака–Лейблера* между ними) называется $\mathcal{KL}(P||Q) = \sum_{a\in A} p_a \log \frac{p_a}{q_a}$. *Расстоянием по вариации* между двумя распределениями называется $||\mathbb{P}_1 - \mathbb{P}_2||_1 = \sum_{a\in A} |\mathbb{P}_1(a) - \mathbb{P}_2(a)|$. Доказать, что между ним и расстоянием Кульбака–Лейблера справедливо следующее соотношение:

$$\mathcal{KL}(\mathbb{P}_1||\mathbb{P}_2) \geq \frac{1}{2\ln 2}||\mathbb{P}_1 - \mathbb{P}_2||_1^2.$$

14. В уездном городе M хотят опросить население с целью восстановления матрицы трудовых корреспонденций. Город разделен на $l \gg 1$ районов. Таким образом, число различных пар (место жительство)–(место работы) равно $m = l^2$. Именно эти пропорции (какая доля людей в городе p_k соответствует корреспонденции с номером k): p_k, $k = 1, ..., m$ и нужно определить, опрашивая случайно выбранных жителей города о том где они живут и работают. Используя формулу Стирлинга (для мультиномиального распределения) и неравенство Пинскера, определите какое количество людей достаточно опросить (постарайтесь оценить это число как можно точнее – опросы стоят денег), чтобы имело место неравенство

$$\mathbb{P}\left(\sum_{k=1}^{m}\left|\frac{n_k}{n} - p_k\right| \geq 0.05\right) = \mathbb{P}\left(\frac{1}{n}\|\vec{n} - n\vec{p}\|_1 \geq 0.05\right) \leq 0.05,$$

Покажите, что справедливо аналогичное неравенство для l_2 нормы:

$$\mathbb{P}\left(\frac{1}{n}\|\vec{n} - n\vec{p}\|_2 \geq \sqrt{\frac{8x}{n}}\right) \leq e^{-x}.$$

Указание. В первом неравенстве воспользуйтесь неравенствами концентрации меры из раздела 3. Для второго неравенства докажите неравенство Хефдинга в Гильбертовом пространстве: пусть $X_1, \ldots, X_n$ – независимые случайные вектора, причем $\mathbb{E}X_i = 0$, $\|X_i\|_2 < c/2$, $v = nc^2/4$, тогда при $t \geq \sqrt{v}$

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} X_i\right\| > t\right) \leq e^{-(t-\sqrt{v})/2v}.$$

Воспользуйтесь неравенством МакДиармида и следующим свойством нормы

$$\mathbb{E}\left\|\sum_{i=1}^{n} X_i\right\| \le \sqrt{\sum_{i=1}^{n} \mathbb{E}\|X_i\|^2}.$$

15. Для двух монеток, симметричной ($p = \frac{1}{2}$) и неправильной ($q = \frac{1+\varepsilon}{2}$), требуется выяснить какая из них является симметричной. Покажите, что для выявления симметричной монетки потребуется $\Omega(1/\varepsilon^2)$ бросаний. Рассмотрите также случай $p = \varepsilon$, $q = 2\varepsilon$. В каком случае потребуется больше бросаний?

Указание. Пусть $f(X_1, \ldots, X_n) \in [0, M]$ является некоторой статистикой, которую можно подсчитать для каждой из монеток и по разности значений идентифицировать монетки. Покажите, что

$$\left|\mathbb{E}_{\mathbb{Q}}[f(X_1, \ldots, X_n)] - \mathbb{E}_{\mathbb{P}}[f(Y_1, \ldots, Y_n)]\right| \le M\|\mathbb{Q} - \mathbb{P}\|_1,$$

где $(X_1, \ldots, X_n) \in \mathbb{Q}$, $(Y_1, \ldots, Y_n) \in \mathbb{P}$. Далее, для оценки $\|\mathbb{Q} - \mathbb{P}\|_1$ можно воспользоваться неравенством Пинскера (см. задачу 13), а также цепным правилом для $\mathcal{KL}$ дивергенции (см. указание к задаче 19).

16. * Есть N ручек, с каждой из которых связана вероятность успеха p_k (дергая k-ю ручку с вероятностью p_k мы получим 1, а с вероятностью $1 - p_k$ ничего). Таким образом, выигрыш игрока при выборе k-й ручки есть $r_k \in \mathrm{Be}(p_k)$. Вероятности p_k не известны игроку. Игрок намеревается выполнить $T \gg N$ дерганий ручек. При выборе ручки на новом шаге можно использовать всю предысторию. Предложите стратегию, "максимально близкую" в среднем по размеру выигрыша к величине $T \max\limits_k p_k$.

Замечание. См. монографию Lugoshi G., Cesa-Bianchi N. Prediction, learning and games. New York: Cambridge University Press, 2006, а также S. Bubeck, N. Cesa-Bianchi. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. In Foundations and Trends in Machine Learning, Vol 5, 2012.

17 (Многорукие бандиты)**.** В модели *стохастического многорукого бандита* имеется K распределений $f_1, \ldots, f_K$ на множестве

$\Omega = [0,1]$, каждое из которых является распределением выигрыша в зависимости от выбора действия $a \in \{1,\dots,K\}$. Игра повторяется $T \gg K$ раундов, причем на t-м раунде

а) игрок выбирает действие $A(t)$, исходя из результатов предыдущих раундов;

б) среда генерирует выигрыш r_t из распределения $f_{A(t)}$, независимо от предыдущих раундов.

Пусть $m_k = \mathbb{E}_{f_k}(r)$, $m^* = \max_k m_k$, $\triangle_k = m^* - m_k$ Цель игрока состоит в минимизации следующей функции потерь

$$\overline{R}_T = Tm^* - \sum_{t=1}^{T} \mathbb{E}m_{A(t)} = \sum_{k=1}^{K} \triangle_k \mathbb{E}A_k(T),$$

где $A_k(T)$ – количество действий с номером k за T раундов. Опишем используемую стратегию игры (то есть алгоритм выбора действий A). В каждом раунде будем вычислять доверительные интервалы для оценок $\widehat{m}_1,\dots,\widehat{m}_K$ и выбирать в следующем раунде действие, соответствующее максимальному значению верхней границы доверительного интервала.

Используя неравенство Маркова в экспоненциальной форме (метод Чернова), покажите что согласно описанной стратегии

$$A(t+1) \in \underset{k\in\{1,\dots,K\}}{\arg\max} \left[\widehat{m}_{k,A_k(t)} + g^{-1}\left(\frac{\alpha \ln t}{A(t)}\right)\right],$$

при уровне значимости для доверительного интервала $1/t^\alpha$, где g определяется из условия

$$\mathbb{E}e^{\lambda|r-\mathbb{E}_{f_k}(r)|} \le \psi(\lambda), \quad g(x) = \psi^*(x) = \sup_\lambda(\lambda x - \psi(\lambda)).$$

18. Докажите верхнюю оценку для стратегии, предложенной в задаче 17:

$$\overline{R}_T \le \sum_{k:\triangle_k>0} \left(\frac{\alpha\triangle_k}{g(\triangle_k/2)}\ln(T) + \frac{\alpha}{\alpha-2}\right).$$

Также установите для случая $f_k = \mathrm{Be}(p_k)$, что для любой стратегии A, при выполнении условия $\forall k : \triangle_k > 0 \to \mathbb{E}A_k(T) = o(T^\gamma)$, $\gamma > 0$,

справедлива нижняя оценка

$$\varliminf_{T\to\infty} \frac{\overline{R}_T}{\ln(T)} \geq \sum_{k:\triangle_k>0} \frac{\triangle_k}{\mathcal{KL}(m_k, m^*)}.$$

Указание. См. S. Bubeck, N. Cesa-Bianchi. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. In Foundations and Trends in Machine Learning, Vol 5, 2012.

Замечание. Последняя оценка говорит об асимптотической неулучшаемости стратегии из задачи 17. Для сравнения приведенных оценок можно воспользоваться неравенством

$$\mathcal{KL}(m_k, m^*) = \mathcal{KL}(p_k, p^*) \leq \frac{(p_k - p^*)^2}{p^*(1-p^*)} = \frac{\triangle_k^2}{p^*(1-p^*)}.$$

19. Пусть параметры распределений $f_k = \mathrm{Be}(p_k)$: $p_1, \ldots, p_K$ в модели из задачи 17 принадлежат распределению F. Докажите, что сушествует такое F, что для любого детерминированного алгоритма игрока A выполнено неравенство

$$T \max_k p_k - \sum_{t=1}^{T} \mathbb{E}(r_t|A) \geq \frac{1}{20} \min(\sqrt{KT}, T) = \Omega(\sqrt{KT}).$$

Указание. В качестве F возьмите следуюшее распределение: все $p_k = 0.5$, за исключением одного $p_i = 0.5 + \varepsilon$, i выбрано случайно равновероятно из множества $\{1, \ldots, K\}$. Установите тождество

$$\mathbb{E}(r_t|A) = \frac{1}{2} + \frac{\varepsilon}{K} \sum_{j=1}^{K} \sum_{r} \big[A(t,r) = j\big] \mathbb{P}(r|A, j = i),$$

где $r \in \{0,1\}^T$, $[\cdot]$ – индикаторная функция. Далее, воспользуйтесь соотношением

$$\sum_r \big[A(t,r) = j\big] \mathbb{P}_i(r) \leq \sum_r \big[A(t,r) = j\big] \mathbb{P}_u(r) + \|\mathbb{P}_i - \mathbb{P}_u\|_1,$$

где $\mathbb{P}_i(r) = \mathbb{P}(r|A, j = i)$, $\mathbb{P}_u(r)$ – распределение выигрышей при $p_1 = \ldots = p_K = 0.5$. Воспользуйтесь неравенством Пинскера и цепным разложением величины

$$\mathcal{KL}(p(x,y), q(x,y)) = \mathcal{KL}(p(x), q(x)) + \sum_x p(x) \mathcal{KL}(p(y|x), q(y|x)).$$

20. Пусть казино делает n бросаний, используя распределение вероятностей на бинарных словах длины n – $p(x)$, где $x \in \{0,1\}^n$, известное игроку. При этом казино производит выплаты так, как если бы оно использовало распределение $q(x)$ (то есть выигранная ставка на 0, после выпадения последовательности исходов x увеличивается в $\frac{q(x)}{q(x:+0)}$ раз, выигранная ставка на 1 – увеличивается в $\frac{q(x)}{q(x:+1)}$ раз). Докажите, что у игрока есть стратегия, логарифм значения капитала которой равен расстоянию Кульбака–Лейблера (см. задачу 13) между распределениями p и q.

21 (Теорема Шеннона о пропускной способности канала с шумом)**.** Канал (поток) связи с шумом описывается матрицей переходных вероятностей $p(Y|X)$, где X и Y – случайные величины с распределениями P (на множестве A – входной алфавит) и Q (на множестве B – выходной алфавит) соответственно. Другими словами, $p(y|x)$ – вероятность прочитать символ y из потока при условии, что в него был записан символ x. Определим *Шенноновское количество информации* как $I(P,Q) = H(P)+H(Q)-H(P,Q)$. Пропускной способностью такого канала называется величина $C = \max_{\{p_x\}} I(P;Q)$. Пусть по каналу передаются слова $w_1, ..., w_N$ длины n. Разобьем множество кодовых слов в алфавите Y на непересекающиеся области $V_0, ..., V_N$. Если принятое слово $y \in V_j$, $j = \overline{1,N}$, то принимается решение о том, что было послано слово w_j. Если $y \in V_0$ то никакое определенное решение не принято. Введем *среднюю вероятность ошибки*

$$\overline{P}_\varepsilon(W,V) = \frac{1}{N}\sum_{i=1}^{n}(1 - p(V_i|w_i)).$$

Пусть $p_\varepsilon(n,N) = \min_{W,V} \overline{P}_\varepsilon(W,V)$, докажите что:

а) $p_\varepsilon(n, 2^{nR}) \to 0$, $R < C$;

б) $p_\varepsilon(n, 2^{nR}) \not\to 0$, $R > C$;

в) $p_\varepsilon(n, 2^{nR}) \to 1$, $R > C$,

где $R = \frac{\log N}{n}$ – скорость передачи.

Замечание. Величина $H(P,Q) = -\sum_{a\in A} p(a) \sum_{b\in B} p(a,b) \log p(a,b)$ называется *совместной энтропией* двух случайных величин.

5. Вероятностные методы в Computer Science

1. Алгоритм быстрой сортировки основан на парадигме "разделяй и властвуй". Выбирается из элементов массива опорный элемент, относительно которого переупорядочиваются все остальные элементы. Желательно выбрать опорный элемент близким к значению медианы, чтобы он разбивал список на две примерно равные части. Переупорядочивание элементов относительно опорного происходит так, что все переставленные элементы, лежащие левее опорного, меньше его, а те, что правее – больше или равны опорному. Далее процедура быстрой сортировки рекурсивно применяется к левому и правому списку для их упорядочивания по отдельности.

Наихудшие входные данные для описанного алгоритма быстрой сортировки (предполагается, что в качестве опорного элемента выбирается последний элемент обрабатываемого массива) – элементы уже упорядоченные по возрастанию. Откуда следует, что асимптотика времени работы быстрой сортировки в худшем случае $\Theta(n^2)$.

Оценить время работы алгоритма быстрой сортировки в среднем.

Указание. Получить рекуррентное соотношение для математического ожидания времени работы, введя индикаторную функцию позиции опорного элемента. Воспользоваться соотношением:

$$\sum_{k=1}^{n-1} k \log k \le \log \tfrac{n}{2} \sum_{k=1}^{\lceil \frac{n}{2} \rceil - 1} k + \log n \sum_{k=\lceil \frac{n}{2} \rceil}^{n-1} k =$$
$$= \tfrac{n(n-1)}{2} \log n - \tfrac{\lceil \frac{n}{2} \rceil (\lceil \frac{n}{2} \rceil - 1)}{2} \le \tfrac{1}{2} n^2 \log n - \tfrac{n^2}{8}.$$

Показать неулучшаемость оценки для произвольного алгоритма сортировки. Привести способ сортировки с асимптотикой $O(n \log n)$ в худшем случае.

2 (Задача поиска k-ой порядковой статистики)**.** Рекурсивное применение процедуры, основанной на методе быстрой сортировки, позволяет быстро (в среднем) находить k-ую порядковую статистику. Задача вычисления порядковых статистик состоит в следующем: дан список (массив) из n чисел, необходимо найти значение, которое стоит в k-ой позиции в отсортированном в возрастающем порядке списке.

Модифицируем алгоритм быстрой сортировки:

Выбираем опорный элемент. Делим список на две группы. В первой — элементы меньше опорного, во второй — больше либо равны.

Если размер (число элементов) первой группы больше либо равен k, то к ней снова применяется эта процедура. Иначе — вызывается процедура для второй группы.

Покажите, используя ту же технику, что и при анализе в среднем алгоритма быстрой сортировки, что среднее время работы такого алгоритма линейно.

Указание. Покажите, что выполняется оценка среднего времени работы алгоритма:

$$\mathbb{E}[T(n)] \leq \mathbb{E}\left\{\sum_{k=1}^{n} T\left(\max(k-1, n-k)\right) + O(n)\right\}$$

$$\leq \frac{2}{n} \sum_{k=\lfloor \frac{n}{2} \rfloor}^{n-1} \mathbb{E}[T(k)] + O(n).$$

3 (Задача о рюкзаке)**.** Рассмотрим NP-трудную задачу

$$\sum_{j=1}^{n} x_j \to \max, x_j \in \{0,\ 1\}, j = 1, ..., n;$$

$$\sum_{j=1}^{n} a_{ij} x_j \leq 1, i = 1, \ldots, m, (*)$$

где $a_{ij} \in \{0,\ 1\}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$.

Булев вектор $\vec{x}$ длины n будем называть допустимыми, 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–ой 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 (математическое ожидание времени работы алгоритма), т.е. $\mathrm{O}\left(\mathbb{E}\left(|T\left(n\right)|\right)mn\right)$, если с.в. $\left\{a_{ij}\right\}_{i,j=1}^{m,n}$ – независимые и одинаково распределенные по закону Бернулли $\mathrm{Be}\left(p\right)$ $(mp^2 \geq \ln n)$.

Указание. Пусть $k > 0$. Положим: $\vec{x}_{j_1,\ldots,j_k}$ - вектор с k единицами (на позициях $\{j_1, ..., j_k\}$) и $n-k$ нулями; p_{ki} - вероятность выполнения i-го неравенства системы (*) для $\vec{x}_{j_1,\ldots,j_k}$; P_k – вероятность того, что $\vec{x}^k$ – допустимое решение (покажите, что p_{ki} и P_k не зависят от набора $\{j_1, ..., j_k\}$). Докажите, что $p_{ki} \leq \left(1-p^2\right)^{k-1} \leq e^{-p^2(k-1)}$, $P_k \leq e^{-mp^2(k-1)}$ и $\mathbb{E}\left(|T\left(n\right)|\right) = \sum_{k=0}^{n} C_n^k P_k < 1+n+n\sum_{k=2}^{n} e^{(k-1)\left(\ln n - mp^2\right)}$.

4. Даны три матрицы A, B, C размера $n \times n$. Требуется проверить равенство $AB = C$.

Простой детерминированный алгоритм перемножает матрицы A, B и сравнивает результат с C. Время работы такого алгоритма при использовании обычного перемножения матриц составляет $O(n^3)$, при использовании быстрого - $O(n^{2,376})$. Вероятностный алгоритм Фрейвалда с односторонней ошибкой проверяет равенство за время $O(n^2)$.

Описание вероятностного алгоритма:

а) взять случайный вектор $x \in \{0,1\}^n$;

б) вычислить $y = Bx$;

в) вычислить $z = Ay$;

г) вычислить $t = Cx$;

д) если $z = t$ вернуть «да», иначе «нет».

Покажите, что для предъявленного алгоритма выполняется

$$\mathbb{P}\left\{z = t | AB = C\right\} = 1,$$
$$\mathbb{P}\left\{z \neq t | AB \neq C\right\} \geq 1/2.$$

Замечание. (амплификация) Оцените вероятность ошибочного ответа на одной ненулевой строке матрицы $D = AB - C$. Как можно добиться того, чтобы вероятность ошибочного ответа стала меньше 0.01?

5. Рассмотрим задачу из класса *NP-трудных* задач – *максимальная выполнимость (MAX-SAT)*: даны m скобок *конъюнктивной нормальной формы* (КНФ) с n переменными, нужно найти значения переменных, максимизирующее число выполненных скобок.

а) Для *приближенного* решения задачи MAX-SAT воспользуемся простейшим вероятностным алгоритмом, выбирая значения каждой переменной (0 или 1) независимо и равновероятно. Покажите, что такой алгоритм гарантирует точность 1/2: для всех входов I

$$\frac{\mathbb{E}m_A(I)}{m_0(I)} \geq \frac{1}{2},$$

где $m_0(I)$ - оптимум, $m_A(I)$ - случайное значение, найденное алгоритмом.

б) Алгоритм с лучшими оценками точности строится на основе *метода вероятностного округления*. Для начала переформулируем задачу MAX-SAT в терминах задачи целочисленного линейного программирования (ЦЛП). Каждой скобке C_j поставим в соответствие булеву переменную $z_j \in \{0,1\}$, которая равна 1, если скобка C_j выполнена; каждой входной переменной x_i сопоставляем переменную y_i, которая равна 1, если $x_i = 1$, и равна 0 в противном случае. Обозначим C_j^+ индексы переменных в скобке C_j, которые входят в нее без отрицания, а через C_j^- - множество индексов переменных, которые входят в скобку с отрицанием. Тогда задача MAX-SAT эквивалентна следующей задаче ЦЛП:

$$\begin{array}{l} \sum_{j=1}^m z_j \to \max\limits_{z,y} \\ \sum_{i\in C_j^+} y_i + \sum_{i\in C_j^+}(1-y_i) \geq z_j, \quad j = 1,...,m \\ y_i, z_j \in \{0,1\}, \quad i = 1,...,n, j = 1,...,m \end{array}$$

Рассмотрим и решим задачу *линейной релаксации* целочисленной программы:

$$\begin{array}{l} \sum_{j=1}^m \hat{z}_j \to \max\limits_{\hat{y},\hat{z}} \\ \sum_{i\in C_j^+} \hat{y}_i + \sum_{i\in C_j^+}(1-\hat{y}_i) \geq \hat{z}_j, \quad j = 1,...,m \\ \hat{y}_i, \hat{z}_j \in \{0,1\}, \quad i = 1,...,n, j = 1,...,m \end{array}$$

Пусть $\hat{y}_i$, $\hat{z}_j$ - решения задачи линейной релаксации. Ясно, что $\sum_{j=1}^{m} \hat{z}_j$ является верхней оценкой числа выполненных скобок для данной КНФ.

Рассмотрим вероятностный алгоритм решения задачи максимальной выполнимости, где каждая переменная y_i независимо принимает значения 0 или 1 уже не с равными вероятностями, а с вероятностью $\hat{y}_i$ принимает значение 1 (и 0 с вероятностью $1 - \hat{y}_i$). Такой метод называется *вероятностным округлением.*

Докажите, что если в скобке C_j имеется k литералов, то вероятность того, что она выполнена при вероятностном округлении, не менее

$$\left(1 - \left(1 - \frac{1}{k}\right)^k\right) \hat{z}_j.$$

Замечание. Тогда из того, что

$$1 - \left(1 - \frac{1}{k}\right)^k \geq 1 - \frac{1}{e} > 0.63$$

для всех положительных целых k, получаем, что для произвольной КНФ среднее число скобок, выполненное при вероятностном округлении, не меньше $1 - \frac{1}{e} > 0.63$ от максимально возможного числа выполненных скобок.

в) Для получения детерминированного алгоритма приближенного решения задачи MAX-SAT воспользуемся один из способов *дерандомизации – методом условных математических ожиданий.* Введем случайную величину $Z(x)$, где в булевом векторе $x = (x_1, ..., x_n)$ компоненты - суть значения переменных в КНФ, назначенных вероятностным алгоритмом - являются независимыми случайными величинами, причем $\mathbb{P}\{x_i = 1\} = p_i$, $\mathbb{P}\{x_i = 0\} = 1 - p_i$; $Z(x)$ - число невыполненных скобок. Требуется найти булев вектор $\hat{x}$, для которого выполнено неравенство $Z(\hat{x}) \leq \mathbb{E}Z$. Обозначим через $Z(x|x_1 = d_1, ..., x_k = d_k)$ новую случайную величину, которая получена из Z фиксированием значений первых k булевых переменных.

Рассмотрим покомпонентную стратегию определения искомого вектора $\hat{x}$. Для определения его первой компоненты вычисляем значения $f_0 = \mathbb{E}Z(x|x_1 = 0)$ и $f_1 = \mathbb{E}Z(x|x_1 = 1)$. Если $f_0 < f_1$ полагаем

$x_1 = 0$, иначе полагаем $x_1 = 1$. При определенной таким образом первой компоненты (обозначим ее d_1) вычисляем значение функционала $f_0 = \mathbb{E}Z(x|x_1 = d_1, x_2 = 0)$ и $f_1 = \mathbb{E}Z(x|x_1 = d_1, x_2 = 1)$. Если $f_0 < f_1$ полагаем $x_2 = 0$, иначе полагаем $x_2 = 1$. Фиксируем вторую координату (обозначая ее d_2) и продолжаем описанный процесс до тех пор, пока не определится последняя компонента решения.

Покажите, что найденный вектор $(x_1 = d_1, \ldots, x_n = d_n)$ будет удовлетворять требованию минимизации оценки математического ожидания:

1) для этого докажите неравенство $\mathbb{E}Z \geq \mathbb{E}Z(x|x_1 = d_1)$;

2) рекуррентно получите неравенство $\mathbb{E}Z \geq \mathbb{E}Z(x|x_1 = d_1, \ldots, x_n = d_n)$;

3) заметьте, что $\mathbb{E}Z(x|x_1 = d_1, \ldots, x_n = d_n) = Z(x|x_1 = d_1, \ldots, x_n = d_n)$.

Покажите справедливость формул:

1) $\mathbb{E}Z = \sum_{j=1}^{m} \mathbb{P}_j$, где $\mathbb{P}_j = \mathbb{P}\left\{\sum_{i \in C_j^+} x_i + \sum_{i \in C_j^-} (1 - x_i) = 0\right\}$;

2) в предположении, что значения первых k переменных уже определены и I_0 — множество индексов тех переменных, значения которых равно 0, а I_1 — множество индексов тех переменных, значения которых равно 1:

$$\mathbb{E}Z(x|x_1 = d_1, \ldots, x_k = d_k) = \sum_{j=1}^{m} \mathbb{P}_j^k,$$

$$\text{где} \mathbb{P}_j^k = \mathbb{P}\left\{\sum_{i \in C_j^+} x_i + \sum_{i \in C_j^-} (1 - x_i) = 0 | x_1 = d_1, \ldots, x_k = d_k\right\} =$$

$$\begin{cases} 0, & \text{при } (I_0 \cap C_j^-) \cup (I_1 \cap C_j^+) \neq \emptyset; \\ \prod_{i \in C_j^+ \setminus I_0} (1 - p_i) \prod_{i \in C_j^- \setminus I_1} p_i, & \text{иначе} \end{cases}$$

Здесь C_j^+ — множество индексов переменных в скобке C_j, которые входят в нее без отрицания, C_j^- — множество индексов переменных, которые входят в скобку с отрицанием.

Замечание. Детали см. в Кузюрин Н.Н., Фомин С.А. Эффективные алгоритмы и сложность вычислений: Учебное пособие. – М.: МФТИ, 2007.

6 (Коммуникационная сложность, хэширование)**.** Требуется сравнить ("достаточно достоверно") две битовые строки a, b, осуществив как можно меньше по битовых сравнений. Основная идея – сравнивать не сами строки, а функции от них. Так сравниваются $a \mod p$ и $b \mod p$, для некоторого простого числа p. Для этого требуется передать $2 \log p$ бит информации.

Описание алгоритма сравнения строк:

а) Пусть $|a| = |b| = n$, $N = n^2 \log n^2$;

б) Выбираем случайное простое число pиз интервала $[2..N]$;

в) Выдать «да», если $a \mod p = b \mod p \Leftrightarrow (a - b) \equiv 0 \mod p$, иначе выдать «нет».

Обоснуйте выбор именно простого числа на шаге 2 и предложите способ его генерации.
Покажите что,

$$\mathbb{P}\{(a - b) \equiv 0, \mod (p) | a = b\} = 1,$$
$$\mathbb{P}\{(a - b) \equiv 0, \mod (p) | a \neq b\} = O(1/n),$$

При этом необходимое количество переданных бит равно $O(\log n)$.

Указание. Воспользоваться асимптотическим законом распределения простых чисел:

$$\lim_{n \to \infty} \frac{\pi(n)}{n/\ln n} = 1,$$

где $\pi(n)$ - функция распределения простых чисел, равная количеству простых чисел, не превосходящих n.

Замечание. В приведенной задаче требуется проверка простоты числа. Согласно малой теореме Ферма, если N - простое число и целое a не делится на N, то

$$a^{N-1} \equiv 1 \mod N. \quad (*)$$

Отсюда следует, что если при каком-то a сравнение $(*)$ нарушается, то можно утверждать, что N - составное. К сожалению, простой

вариант подбора a не всегда позволяет эффективно выявить составное число. Имеются составные числа N, обладающие свойством $(*)$ для любого целого a с условием $(a, N) = 1$ (a и N - взаимно простые). Такие числа называются числами Кармайкла.

В 1976 г. Миллер предложил заменить проверку $(*)$ проверкой несколько иного условия. Если N - простое число, то $N - 1 = 2^s t$, где t нечетно, то согласно малой теоремы Ферма для каждого a с условием $(a, N) = 1$ хотя бы одна из скобок в произведении

$$\left(a^t - 1\right)\left(a^t + 1\right)\left(a^{2t} + 1\right)\ldots\left(a^{2^{s-1}t} + 1\right) = a^{N-1} - 1$$

делится на N.

Пусть N - нечетное составное число, $N - 1 = 2^s t$, где t нечетно. Назовем целое число a, $1 < a < N$ «выявляющим» для N, если нарушается одно из двух условий:

I) N не делится на a

II) $a^t \equiv 1 \mod N$ или существует целое k, $0 \leq k < s$ такое, что $a^{2^k t} \equiv -1 \mod N$.

Если N составное число, то согласно теореме Рабина существует не менее $\frac{3}{4}(N - 1)$ выявляющих чисел.

7 (Интерактивные доказательства)**.** а) (изоморфизм графов). Авдотье известен изоморфизм ϕ графов G_0 и G_1. Но она посылает Евлампию граф $H = \psi(G_0)$, либо $H = \psi(G_1)$, где ψ – некоторый другой изоморфизм, не равный ϕ. Евлампий бросает симметричную монетку и просит изоморфизм либо $H : G_0$, либо $H : G_1$. В первом случае Авдотья посылает ψ, во втором – $\psi\phi^{-1}$. Таких партий разыгрывается N штук. Заметим, что в каждой новой партии Авдотья придумывает новую перестановку ψ вершин графа G_0. Если ϕ – действительно изоморфизм $G_0 : G_1$, то все проверки Евлампия будут положительны. Покажите, что если ϕ – блеф, то с вероятностью $1-2^{-N}$ хотя бы одна проверка обнаружит это (та проверка, в которой Евлампий попросил $H : G_1$).

Замечание. Этот пример поучителен с точки зрения “криптографического фокуса” – Авдотья убедила Евлампия в $G_0 : G_1$ так и не огласив самого изоморфизма ϕ. Если ϕ — пароль, то диалог можно вести даже в открытую, что служит примером криптосистемы с нулевым разглашением.

Обоснуйте справедливость данного замечания (Евлампий не получает никакой информации об изоморфизме ϕ).

б) (неизоморфизм графов). Теперь наделим Авдотью сверхъестественными вычислительными способностями. Для удобства переименуем игроков: “Prover” и “Verifier”. Verifier выбирает случайно (равновероятно) $i \in \{0, 1\}$ и некоторую перестановку π вершин графа G_i, затем посылает граф $H = \psi(G_i)$ и требует, чтобы Prover определил i. Таких партий разыгрывается N штук. Аналогично предыдущему примеру, π в каждой новой партии свое. Если графы неизоморфны, то Prover всегда верно определит индекс i. Все тесты будут пройдены. Покажите, что иначе с вероятностью $1 - 2^{-N}$ Prover ошибется хотя бы один раз (хотя бы в одной партии).

Замечание. Заинтересовавшимся в этой теме, можно также порекомендовать посмотреть про цифровую подпись (протокол аутентификации) и электронную систему голосвания в книге Введение в криптографию под ред. В.В. Ященко. М.: МЦНМО, 2013.

8 (Хеш функции)**.** Для компактного хранения данных, индексированных, как правило, ключами целочисленного или строкового форматов, используются хеш-функции, при помощи которых вычисляются номера ячеек в таблице в зависимости от значения ключа. Пусть хэш-функция задана на множестве ключей K

$$h : K \to \{1, \ldots, m\}.$$

Заведем таблицу размера m, в которой элемент с ключом k будет размещаться по адресу $h(k)$ (во многих случаях $m \ll |K|$). Ключи с одинаковым значением $h(\cdot)$ образуют список, начинающийся с ячейки таблицы. Если всего имеется n элементов, то желательно чтобы количество ключей с одинаковым хеш-значением не превышало значительно $\alpha = n/m$ (в таком случае длины списков будут ограничены $O(1 + \alpha)$). Формально данное свойство может быть сформулировано в виде двух требований:

а) $h(k)$ является с.в. с равномерным распределением;

б) $\forall k_1, k_2 \in K$ $h(k_1)$ и $h(k_2)$ независимы.

Покажите, что в случае выполнения требований среднее время поиска отсутствующего ключа в таблице составляет $(1 + \alpha)$, а также

среднее время поиска присутствующего ключа в таблице составляет $(1 + \alpha/2)$ (где усреднение ведется как по $h(k)$, так и по множеству других присутствующих в таблице ключей).

Замечание. Так как хеш-функция является детерминированной, то случайность, присутствующая в требованиях, может быть реализована посредством выбора функции h из некоторого семейства H. H называется *универсальным* для множества K, если $\forall k_1, k_2 \in K$ вероятность выбрать функцию из H, в которой возникает коллизия, не превышает $1/m$. Для универсального семейства справедливы приведенные в задаче оценки времени поиска ключа в таблице. Примером универсального семейства H для целых чисел может служить

$$\{h_{ab} : h_{ab}(k) = ((ak + b) \bmod p) \bmod m,\ a, b \in \{1, \ldots, p-1\}\},$$

где p – простое число, большее m.

Другой пример универсального семейства, используемого при хешировании IP адресов вида (x_1, x_2, x_3, x_4), $x_i \in \{0, \ldots, 255\}$ представляет собой множество хеш-функций

$$\{h_a(x_1, x_2, x_3, x_4) = (a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4) \bmod p\}.$$

где p – также простое число, большее m.

9 (Хеширование без коллизий)**.** При известном заранее множестве ключей $K, |K| = n$ хеширование без коллизий может быть реализовано при помощи двухуровневой схемы. На первом уровне используется таблица размера $m = n$, для которой хеш функция выбирается случайно из универсального семейства (см. замечание к задаче 8). На втором уровне осуществляется хеширование без коллизий (вместо создания списка ключей будем использется вторичная хеш-таблица T_i, хранящая все ключи, хешированные функцией h в ячейку i, со своей функцией h_i). Покажите, что при выборе на втором уровне таблицы размером n_i^2 (n_i – количество ключей в i-й ячейке первого уровня) вероятность возникновения коллизий менее $1/2$ (поэтому для данной ячейки всегда найдется функция h_i свободная от коллизий). Докажите, что существует хеш функция h для первого уровня, при которой затраты памяти составят суммарно $O(n)$.

Указание. Получите оценку $\mathbb{E}\left(\sum_i n_i^2\right) < 2n$, воспользуйтесь неравенством Маркова.

10 (Bloom filter)**.** Вероятностная структура данных Bloom filter используется для проверки принадлежности элемента множеству S. Она представляет собой массив A длиной m бит и k различных хэш-функций $h_1, \ldots, h_k$, равновероятно отображающих элементы множества $D \supseteq S$ в позиции массива A ($h_i : S \to \{1, \ldots, m\}$). Интерфейс данной структуры включает следующие операции:

`add_element(s)`: устанавливает биты $A[h_1(s)], \ldots, A[h_k(s)]$ равными 1 (изначально все биты равны 0);

`contains(s)`: если биты $A[h_1(s)], \ldots, A[h_k(s)]$ равны 1, то возвращает `true`, иначе `false`;

Последовательно добавив все элементы множества S при помощи операции `add_element`, получим объект структуры данных, который при вызове метода `contains(s)` всегда возвращает `true` при $s \in S$, но не всегда возвращает `false` при $s \notin S$, так как соответствующие биты могли быть установлены равными 1 за счет других элементов из S.

`approx_size`: возвращает приближенное количество добавленных элементов

$$|S| \approx -m\frac{\ln(1 - \mathbb{I}/m)}{k},$$

где $\mathbb{I}$ – сумма элементов A.

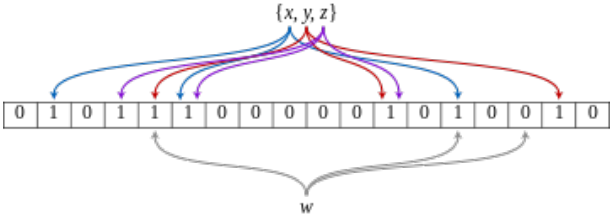


Рис. 7: Пример структуры данных Bloom filter ($k = 3$, $m = 18$): x, y, z – принадлежат множеству, w – не принадлежит.

Считая события $A[i] = 1$ и $A[j] = 1$ при добавлении n элементов в пустой объект структуры Bloom filter независимыми, ввиду независимости значений хэш-функций, покажите, что:

$$\mathbb{P}(\text{contains(s)} = \text{true} \mid s \text{ not in } S) \approx p = \left(1 - e^{kn/m}\right)^k.$$

Откуда следует, что оптимальные значения m и k могут быть вычислены по формулам:

$$k = \frac{m}{n} \ln 2, \quad m = -n \frac{\ln p}{(\ln 2)^2}.$$

Замечание. Bloom filter обладает следующими отличительными характеристиками:

а) При вероятности ошибки $p = 0.01$ требуется всего лишь 9.6 бит на один элемент множества.

б) Сложность операций `contains` и `add_element` составляет $O(k)$, вне зависимости от $|S|$. Стоит также учесть, что k операций над элементами массива могут быть выполнены параллельно.

в) Операции объединения и пересечения двух множеств выполняются побитно (за сравнительно короткое время):

$$|S_1 \cup S_2| \approx -m \frac{\ln(1 - \mathbb{I}_{12}/m)}{k},$$

$$|S_1 \cap S_2| = |S_1| + |S_2| - |S_1 \cup S_2|,$$

где $\mathbb{I}_{12}$ – скалярное произведение A_1 и A_2.

г) Сборщик мусора, ровно как и сериализатор могут осуществлять сравнительно быстрые операции чтения/записи/удаления структуры данных Bloom filter, т. к. A представляет собой единичный объект, состоящий из элементарных типов данных.

Области применения: снижение количества обращений к базе данных, хранящейся на диске в случае отсутствия запрашиваемых данных (Apache Cassandra); локальная проверка url-адресов на принадлежность списку, хранящемуся на удаленном сервере (Google Chrome); выявление содержимого архива (Venti archival storage system);

11 (Jaccard similarity)**.** Для измерения близости (веса связи) двух множеств часто используется коэффициент Жаккара:

$$J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}.$$

Типичными представителями таких множеств являются списки смежных вершин в графе, изображения, профили пользователей социальных сетей и прочие web-страницы. Вычисление коэффициента Жаккара позволяет выявлять дубликаты, проводить кластеризацию, восстанавливать фрагменты перечисленных выше объектов.

Для получения попарной близости $S_1, \ldots, S_N$ множеств больших размеров ($|S_i| = s > 500$, $i \in \overline{1, N}$) возникает потребность в приближенном вычислении коэффициента Жаккара, снижающем нагрузку на сеть вычислительного кластера и временную сложность вычислений. Для этой цели достаточно для каждого множества $|S_i|$ найти минимумы k хэш-функций и далее при попарном сравнении оперировать только с набором подмножеств из минимумов. В итоге, временная сложность окажется равной $O(kN^2 + ksN \log N)$, в то время как в исходной задаче сравнения без аппроксимации меры $O(s \log sN^2)$.

Рассмотрим два варианта приближенного вычисления, базирующиеся на следующем свойстве: пусть h – хэш-функция, иньективно и равновероятно отображающая элементы множества $\cup_i S_i$ в $\mathbb{N}$, тогда

$$\mathbb{P}(\min_{s \in S_1} h(s) = \min_{s \in S_2} h(s)) = J(S_1, S_2).$$

1 вар. Задействуем k различных хэш-функций. В этом случае оценкой J будет являться доля функций, у которых совпадают минимальные значения на обоих множествах.

2 вар. Обозначим за $h_{(k)}(S)$ подмножество S из k элементов с наименьшими значениями h. Тогда J можно оценить как $h_{(k)}(S_1) \cap h_{(k)}(S_2) \cap h_{(k)}(S_1 \cup S_2)$.

Сравните временные сложности двух вариантов. Используя метод Чернова (см. задачу 26 из раздела 3), покажите, что порядок ошибки аппроксимации в обоих вариантах $O(1/\sqrt{k})$.

12 (EM-алгоритм)**.** Рассмотрим задачу поиска неизвестных параметров распределения при помощи метода максимального правдоподобия. В ряде случаев, где функция правдоподобия имеет вид, не допускающий удобных аналитических методов исследования, для ее упрощения удобно ввести дополнительные "скрытые" (латентные) переменные и воспользоваться EM-алгоритмом. Пусть требуется найти максимум правдоподобия $L(X, \theta) = \log p(X|\theta)$ в вероят-

ностной модели со скрытыми переменными Z

$$p(X|\theta) = \int p(X, Z|\theta) dZ.$$

Покажите, что $L(X, \theta)$ можно представить как

$$L(X, \theta) = \int \log \frac{p(X, Z|\theta)}{q(Z)} q(Z) dZ - \int \log \frac{p(Z|X, \theta)}{q(Z)} q(Z) dZ =$$

$$= l(X, \theta, q) + \mathcal{KL}(q \| p(Z|X, \theta)),$$

где $q(Z)$ – произвольное распределение над скрытыми переменными.

Итерационная схема EM-флгоритма состоит в фиксации на шаге t некоторого значения $\theta^{(t)}$ и аппроксимации в этой точке правдоподобия с помощью его нижней оценки $l(\cdot)$:

$$q(Z) = p(Z|X, \theta^{(t)}), \quad l(X, \theta, q) \to \max_{\theta}.$$

Expectation step: фиксируется значение параметров $\theta^{(t)}$. Оценивается распределение на скрытые переменные

$$q(Z) = p(Z|X, \theta^{(t)}) = \frac{p(X, Z|\theta^{(t)})}{p(X|\theta^{(t)})};$$

Maximization step: фиксируется распределение $p(Z|X, \theta)$ и выполняется поиск новых параметров

$$\theta^{(t+1)} = \arg\max_{\theta} \mathbb{E}_q \log p(X, Z|\theta).$$

Покажите, что

$$L(X, \theta^{(t+1)}) \geq L(X, \theta^{(t)}),$$

а также при гладкости L

$$\left.\frac{\partial L}{\partial \theta}\right|_{\theta^{(\infty)}} = 0.$$

Указание. Воспользуйтесь конструкцией

$$\theta^{(t+1)} = \arg\max_{\theta} \{Q(\theta) - b_t \mathcal{R}(\theta, \theta^{(t)})\},$$

где $\mathcal{R}(\theta, \theta^{(t)}) \geq 0$ – регуляризатор, в нашем случае равный $\mathcal{KL}(\cdot\|\cdot)$.

13 (Разделение смеси распределений)**.** Целью разделения смеси является как восстановление плотности наблюдаемых данных X, так и покомпонентная их категоризация (каждый элемент выборки X принадлежит одному из распределений, входящих в осотав смеси). Допустим, что плотность распределения в z-й компоненте смеси равна $p(x|z) = p(x|z, \alpha_z)$, т.е. известна с точностью до параметра α_z. Тогда плотность $x \in X$ можно аппроксимировать смесью

$$p_\theta(x) = \sum_{z=1}^{K} w_i \; p(x|z),$$

где $\theta = (w_1, \ldots, w_K, z_1, \ldots, z_K)$, $w_i \geq 0$, $\sum_i w_i = 1$, K – количество компонент смеси. Правдоподобие X задается формулой

$$\log L(\theta, X) = \sum_{j=1}^{n} \left(\log \sum_{z=1}^{K} w_i p(x_j|z) \right).$$

Непосредственный поиск точки максимума данной функции весьма затруднителен. Для упрощения вычислений применим EM-алгоритм (см. задачу 12), сопоставив каждому x_j скрытую переменную z_j – номер компоненты смеси, породившей x_j, т. е. $x_j \sim p(x|z_j)$. Проверьте, что при такой модификации данных, функция правдоподобия примет вид

$$\log L(\theta, X, Z) = \sum_{j=1}^{n} \log w_{z_j} + \sum_{j=1}^{n} \log p(x_j|z_j).$$

Покажите, что шагами EM-алгоритма в этом случае будут

1) Expectation step:

$$q_j^{(t)}(z) \propto w_z^{(t)} \; p(x_j|z, \alpha_z^{(t)});$$

2) Maximization step:

$$w_z^{(t+1)} = \frac{1}{n} \sum_{j=1}^{n} q_j^{(t)}(z),$$

$$\alpha_z^{(t+1)} = \arg\max_{\alpha} \sum_{j=1}^{n} q_j^{(t)}(z) \; p(x_j|z, \alpha).$$

Замечание. Для более глубокого ознакомления с модификациями EM алгоритма – медианные модификации, SEM, CEM, MCEM, SAEM, выбор начального приближения, определение числа компонент и типа смеси – рекомендуем ознакомится с работой В.Ю. Королева EM-алгоритм, его модификации и их применение к задаче разделения смесей вероятностных распределений.

14 (Вариационный вывод)**.** Рассмотрим задачу оценки некоторой статистики $T(X)$ для распределения с плотностью $p(X)$, т.е. величины

$$\mathbb{E}_p T(X) = \int T(X)p(X)dX.$$

Предполагается, что $p(X)$ известно с точностью до нормировочной константы ($S_p = \int \tilde{p}(X)dX$ является недоступным):

$$p(X) = \frac{1}{S_p}\tilde{p}(X) \propto \tilde{p}(X).$$

Например, для получения оценки скрытых параметров Z распределения $p(X, Z)$ можно подсчитать $\mathbb{E}_{p(Z|X)}Z$. В этом случае в качестве недоступной для вычисления нормировочной константы выступает $p(X)$. В задачах, решаемых при помощи EM-алгоритма (см. задачу 12), в качестве ненормированного распределения $\tilde{p}$ выступает совместное распределение $p(X, Z|\theta)$, недоступной нормировочной константой является неполное правдоподобие $p(X|\theta)$, а искомой статистикой $T(X, Z)$ – величина $\log p(X, Z|\theta)$.

Одним из вариантов решения задач такого рода является нахождение приближения $q(x)$ в некотором простом семействе распределений и последующая оценка статистики как $\mathbb{E}_p T(X) \approx \mathbb{E}_q T(X)$. Для нахождения $q(X)$ решается следующая оптимизационная задача

$$\mathcal{KL}(q\|p) \to \min_q \Leftrightarrow \int q(X) \log \frac{\tilde{p}(X)}{q(X)} dX \to \max_q .$$

Получите для семейства полностью факторизованных распределений $q(X) = \prod_i q_i(x_i)$, $X = (x_1, \ldots, x_n)$ итерационный метод вычисления оптимальных q_i:

$$q_i(x_i) \propto \exp\left(\int \log\left[\tilde{p}(X)\right] \prod_{j \neq i} q_j(x_j)dx_j \right), \quad i = \overline{1, n}.$$

Рассмотрите случай экспоненциального распределения

$$p(x_i|X_{-i}) = h(x_i)e^{\langle\theta, f(x_i)\rangle - d(\theta)}, \quad \theta = \theta(X_{-i}).$$

Докажите, что для этого случая справедливо упрощенное соотношение для q_i:

$$q_i(x_i) = h(x_i)e^{\langle\mathbb{E}\theta, f(x_i)\rangle - d(\mathbb{E}\theta)}, \quad i = \overline{1,n}.$$

Замечание. Сравним между собой два метода приближенной оценки статистик $T(X)$: вариационный вывод и методы MCMC. В методе MCMC оценка $\mathbb{E}_p T(X)$ является тем точнее, чем больше выборок X генерируется, а в пределе является точной. В вариационном выводе нет никаких гарантий на близость между $\mathbb{E}_p T(X)$ и $\mathbb{E}_q T(X)$. В итерациях вариационного вывода происходит максимизация функционала $L(q)$, являющегося нижней оценкой для $\log S_p$, что, как правило, обеспечивает достаточно точную оценку на значение нормировочной константы даже при существенно ограниченном семействе распределений q. Время работы одной итерации вариационного вывода и одной итерации схемы MCMC, обычно, очень близки. Однако, для сходимости вариационного вывода часто достаточно несколько десятков итераций, в то время как для надежной оценки статистик MCMC требует несколько тысяч итераций.

Для того чтобы описать следующий цикл задач на метод *зеркального спуска* А.С. Немировского, нам потребуется ввести ряд определений и сформулировать некоторые необходимые в дальнейшем результаты (см. раздел Вспомогательные материалы). Отметим, что одну задачу, которая может быть отнесена к задачам этого цикла мы уже встречали (см. задачу 17 раздела 4). Далее при изложении мы будем опираться в основном на работы

Lugosi G., Cesa-Bianchi N. Prediction, learning and games. – New York: Cambridge University Press, 2006.

Вьюгин В.В. Математические основы теории машинного обучения и прогнозирования. – М.: МЦНМО, 2013.

Гасников А.В., Нестеров Ю.Е., Спокойный В.Г. Об эффективности одного метода рандомизации зеркального спуска в задачах онлайн оптимизации. – М.: Автоматика и телемеханика. 2014.

15. Рассмотрим задачу взвешивание экспертных решений. Имеется n различных Экспертов. Каждый Эксперт играет на рынке. Игра

повторяется $N \gg 1$ раз (это число может быть заранее неизвестно). Пусть l_i^k – проигрыш Эксперта i на шаге k ($\left|l_i^k\right| \le M$). На каждом шаге k мы распределяем один доллар между Экспертами, согласно вектору $x^k \in S_n\,(1)$. Потери, которые мы при этом несем, рассчитываются по потерям экспертов $\left\langle l^k, x^k \right\rangle$. Целью является таким образом организовать процедуру распределения доллара на каждом шаге, чтобы наши суммарные потери были бы минимальны. Допускается, что потери экспертов l^k могут зависеть еще и от текущего хода x^k. Легко проверить, что для данной постановки применима теорема 1 в детерминированном варианте с функциями

$$f_k\left(x; \xi^k\right) \equiv f_k\left(x\right) = \left\langle l^k, x \right\rangle.$$

Покажите, что оценка, даваемая теоремой 1, имеет вид

$$\mathrm{O}\left(M\sqrt{\frac{\ln n}{N}}\right).$$

Докажите, что данную оценку нельзя улучшить.

Указание. См. S. Bubeck, N. Cesa-Bianchi. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. In Foundations and Trends in Machine Learning, Vol 5, 2012.

16. В условиях предыдущей задачи предположим, что на k-м шаге i-й эксперт использует стратегию $\zeta_i^k \in \Delta$ (множество Δ – выпуклое), дающую потери $\lambda\left(\omega^k, \zeta_i^k\right)$, где ω^k – "ход", возможно, враждебной "Природы", знающей, в том числе, и нашу текущую стратегию. Функция $\lambda\left(\,\cdot\,\right)$ – выпуклая по второму аргументу и $\left|\lambda\left(\,\cdot\,\right)\right| \le M$. На каждом шаге мы должны выбирать свою стратегию

$$x \stackrel{def}{=} \sum_{i=1}^{n} x_i \cdot \zeta_i^k \in \Delta,$$

дающую потери $\lambda\left(\omega^k, x\right)$ так, чтобы наши суммарные потери были минимальны. Для данной постановки также применима теорема 1 в детерминированном варианте с

$$f_k\left(x; \xi^k\right) \equiv f_k\left(x\right) = \sum_{i=1}^{n} x_i \lambda\left(\omega^k, \zeta_i^k\right) \ge \lambda\left(\omega^k, x\right).$$

Покажите, что оценка, даваемая теоремой 1, имеет вид

$$\mathrm{O}\left(M\sqrt{\frac{\ln n}{N}}\right).$$

Отметим, что эта оценка для данного класса задач.

Указание. Чтобы применить теорему заметим, что функция $\lambda\left(\omega^k,\zeta\right)$ – выпуклая по ζ для любого ω^k, поэтому

$$\sum_{k=1}^{N}\lambda\left(\omega^k,x^k\right)-\min_{i=1,\ldots,n}\sum_{k=1}^{N}\lambda\left(\omega^k,\zeta_i^k\right)\le\sum_{k=1}^{N}f_k\left(x^k\right)-\min_{x\in S_n(1)}\sum_{k=1}^{N}f_k\left(x\right).$$

17. Предположим, что в условиях предыдущей задачи мы не можем гарантировать выпуклость $\lambda(\,\cdot\,)$ – по второму аргументу. Тогда мы выбираем стратегию – распределение вероятностей на множестве стратегий Экспертов, и разыгрываем случайную величину согласно этому распределению вероятностей. Другими словами мы просто пользуемся МЗС2 с $f_k\left(x;\xi^k\right)\equiv f_k\left(x\right)=\sum\limits_{i=1}^{n}x_i\lambda\left(\omega^k,\zeta_i^k\right)$, применимость которого обосновывается теоремой 2. Получите оценки

$\mathrm{O}\left(M\sqrt{\frac{\ln n}{N}}\right)$ – в среднем;

$\mathrm{O}\left(M\sqrt{\frac{\ln(n/\sigma)}{N}}\right)$ – с вероятностью $\ge 1-\sigma$.

18 (Антагонистические матричные игры)**.** Пусть есть два игрока А и Б. Задана матрица игры $A=\|a_{ij}\|$, где $|a_{ij}|\le M$, a_{ij} – выигрыш игрока А (проигрыш игрока Б) в случае когда игрок А выбрал стратегию i, а игрок Б стратегию j. Отождествим себя с игроком Б. И предположим, что игра повторяется $N\gg 1$ раз (это число может быть заранее неизвестно). Мы находимся в условиях предыдущей задачи с $\lambda\left(\omega^k,\zeta_j^k\right)=\sum\limits_{i=1}^{n}\omega_i^k a_{ij}$, то есть

$$f_k\left(x\right)=\left\langle\omega^k,Ax\right\rangle,\quad x\in S_n\left(1\right),$$

где ω^k – вектор[1] со всеми компонентами равными 0, кроме одной компоненты, соответствующей ходу А на шаге k, равной 1. Хотя

[1]Вообще говоря, зависящий от всей истории игры до текущего момента включительно, в частности, как-то зависящий и от текущей стратегии (не хода) игрока Б, заданной распределением вероятностей (результат текущего разыгрывания (ход Б) игроку А не известен).

функция $f_k(x)$ определена на единичном симплексе, по "правилам игры" вектор x^k имеет ровно одну единичную компоненту, соответствующую ходу Б на шаге k, остальные компоненты равны нулю. Обозначим цену игры

$C = \max\limits_{\omega\in S_n(1)} \min\limits_{x\in S_n(1)} \langle \omega, Ax\rangle = \min\limits_{x\in S_n(1)} \max\limits_{\omega\in S_n(1)} \langle \omega, Ax\rangle$. (теорема фон Неймана о минимаксе)

Пару векторов (ω, x), доставляющих решение этой минимаксной задачи, назовем равновесием Нэша. По определению (это неравенство восходит к Ханнану)

$$\min_{x\in S_n(1)} \frac{1}{N}\sum_{k=1}^{N} f_k(x) \le C.$$

Тогда, если мы (игрок Б) будем придерживаться рандомизированной стратегии МЗС2, то по теореме 2 с вероятностью $\ge 1-\sigma$ (в случае когда N заранее известно оценку можно уточнить)

$$\frac{1}{N}\sum_{k=1}^{N} f_k\left(x^k\right) - \min_{x\in S_n(1)} \frac{1}{N}\sum_{k=1}^{N} f_k(x) \le \frac{2M}{\sqrt{N}}\left(\sqrt{\ln n} + \sqrt{2\ln\left(\sigma^{-1}\right)}\right),$$

т.е. с вероятностью $\ge 1-\sigma$ наши потери ограничены

$$\frac{1}{N}\sum_{k=1}^{N} f_k\left(x^k\right) \le C + \frac{2M}{\sqrt{N}}\left(\sqrt{\ln n} + \sqrt{2\ln\left(\sigma^{-1}\right)}\right).$$

Самый плохой для нас случай (с точки зрения такой оценки) – это когда игрок А тоже "знает" теорему 2, и действует согласно МЗС2 (точнее версии МЗС2 для максимизации вогнутых функций на симплексе). Очевидно, что если и А и Б будут придерживаться МЗС2, то они сойдутся к равновесию Нэша, причем чрезвычайно быстро:

$\frac{8M\left(\ln n + 2\ln\left(\sigma^{-1}\right)\right)}{\varepsilon^2}$ – итераций;

$\mathrm{O}\left(n + M\frac{s\ln n\left(\ln n + \ln\left(\sigma^{-1}\right)\right)}{\varepsilon^2}\right)$ – общая сложность вычислений,

где $s \le n$ – среднее число элементов в строках и столбцах матрицы A. Докажите эти оценки.

6. Геометрические вероятности

1. Три бабочки капустницы садятся на круглый кочан капусты радиуса 1 случайным образом (имеется в виду, что место положение каждой бабочки – с.в., равномерно распределенная на сфере) и независимо друг от друга. Если между двумя бабочками (геодезическое) расстояние оказывается меньше $\pi/2$, то обе улетают. Найдите вероятность того, что на капусте сидят все три бабочки.

2. Выбирается случайно и равномерно n точек $P_1, \ldots, P_n$ на единичной окружности. Какова вероятность того, что начало координат (цент окружности) окажется внутри выпуклой оболочки этих точек.

Указание. Выберите n случайных пар диаметрально противоположных точек $Q_1, Q_{n+1} = -Q_1, Q_2, Q_{n+2} = -Q_2, \ldots, Q_n, Q_{2n} = -Q_n$ в соответствии с равномерным распределением. Ясно, что с вероятность 1 все пары различны. В качестве точки P_i равновероятно выбирается либо точка Q_i, либо диаметрально противоположная ей $Q_{n+i} = -Q_i$. Покажите, что такая процедура эквивалентна случайному выбору точек P_i. Покажите, что вероятность того, что начало координат не окажется внутри выпуклой оболочки точек $P_1, \ldots, P_n$, при заданных различных точках $Q_1, \ldots, Q_n, Q_{n+1}, \ldots, Q_{2n}$ равна $\frac{2n}{2^n}$, так как нужные точки $P_1, \ldots, P_n$ могут давать только подмножества вида $\left\{\tilde{Q}_i, \ldots, \tilde{Q}_{i+n-1}\right\}$(суммирование в индексах берется по модулю $2n$), где $\tilde{Q}_1, \ldots, \tilde{Q}_{2n}$ перенумерованные, например по часовой стрелке, точки $Q_1, \ldots, Q_{2n}$.

3. Найти среднюю длину секущих трехмерного куба с единичной длиной.

4 (Парадокс Бертрана)**.** Рассмотрим окружность, описанную около равностороннего треугольника ABC. Какова вероятность того, что случайным образом проведенная хорда будет иметь длину большую, чем длина стороны треугольника ABC?

Указание. Предложите разные способы генерации хорды (не менее трех). Зависит ли ответ на задачу от способа генерации?

5. Пусть в пространстве $\mathbb{R}^n$ с евклидовой нормой задан n-мерный шар единичного радиуса. Внутри него имеются две случайные точки с радиус-векторами $\mathbf{r}_1$ и $\mathbf{r}_2$ соответственно, имеющие равномерное

пространственное распределение внутри шара. Найти распределение случайной величины, являющейся расстоянием между этими двумя точками $r = |\mathbf{r}_1 - \mathbf{r}_2|$.

6. На плоскости проведены параллельные прямые на единичном расстоянии друг от друга, и на плоскость наугад бросается иголка длиной $L < 1$. Угол между прямыми и иголкой и расстояние от середины иглы до ближайшей прямой являются независимыми с.в., равномерно распределенными на $(0, 2\pi)$ и $(-1/2, 1/2)$ соответственно. С помощью серии таких опытов вычислить число π с заданной точностью $\delta = 1\%$ и с вероятностью ошибки не больше $\varepsilon = 5\%$. Решите аналогичную задачу для случая погнутой иглы длиной менее 1.

Указание. Рассмотрим окружность диаметра 1, т.е. длины π. Такая окружность с вероятностью 1 пересекает дважды одну из прямых. Тогда, исходя из линейности математического ожидания числа попаданий иглы на прямую относительно длины иглы, для иглы длиной $L < 1$ имеем $\mathbb{E}\xi_L = 2L/\pi$.

7. Покажите, что средняя площадь ортогональной проекции куба с ребром единица на случайную плоскость равна $3/2$.

Указание. Покажите, что средняя площадь ортогональной проекции всякого измеримого тела линейно зависит от площади его границы. Рассмотрим вспомогательное (см. предыдущую задачу) тело, у которого легко вычисляется средняя площадь ортогональной проекции.

Замечание. Обозначим через S_k k-мерный объем ортогональной проекции рассматриваемой области V в $\mathbb{R}^n$ на случайную k-мерную плоскость. Имеет место следующая формула для объема h-окрестности данной области:

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

где V_0 – объем области; V_1 – $(n-1)$-мерный объем границы области, пропорциональный среднему значению от числа 1; число V_k пропорционально S_k и выражается через средние значения (усредненным по поверхности рассматриваемой области) 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 пропорционален интегралу от средней кривизны по всей поверхности, а V_3 – от гауссовской:

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

Например, для сферы радиуса R

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

Здесь $k_1 + k_2 = \frac{2}{R}$, $k_1 k_2 = \frac{1}{R^2}$,

$$\int (k_1 + k_2) dS = 8\pi R,$$

Формула Гаусса-Бонне:

$$\iint (k_1 k_2)\, dS = 4\pi.$$

8. Приведем геометрическую интерпретацию пуассоновского процесса (см. задачу **??**). Пуассоновским процессом Π в пространстве $S \subset \mathbb{R}^m$ называется счетное множество точек, случайно разбросанных по S, но подчиняющихся следующему правилу: существует мера $\mu : S \to [0, \infty]$, соответствующая процессу Π, такая что для любых непересекающихся измеримых множеств $A_1, \ldots, A_n \subset S$ случайные величины

$$N(A_i) = \#\{A_i \cap \Pi\} \sim \mathrm{Po}(\mu(A_i)), \quad i = \overline{1, n},$$

порожденные случайным попаданием точек в множества $A_1, \ldots, A_n$, независимы и имеют распределение $\mathrm{Po}(\mu)$.

К примеру стандартный пуассоновский процесс из задачи **??** раздела **??** определен в пространстве $S = \mathbb{R}_+$, имеет меру $\mu\big((t_1, t_2]\big) = \lambda(t_2 - t_1)$, $N((t_1, t_2]) = K(t_2) - K(t_1) \sim \mathrm{Po}\big(\lambda(t_2 - t_1)\big)$.

Отметим, что определение Π требует, чтобы мера μ была *неатомической* (значение на любом счетном множестве равно 0), а также представимой в следующем виде

$$\mu = \sum_{k=1}^{\infty} \mu_k, \quad \mu_k(S) < \infty.$$

Пусть $X_1, \ldots, X_n$ – независимые случайные величины, распределенные по $A \subset S$, $\mu(A) < \infty$ в соответствии с вероятностной мерой $p(\cdot) = \mu(\cdot)/\mu(A)$, Обозначим за $N(B)$ количество $X_i \in B$. Покажите, что для непересекающихся $A_1, \ldots, A_k \subset A$, $A = A_1 \cup \ldots \cup A_k$ выполнено

$$\mathbb{P}_n(N(A_1) = n_1, \ldots, N(A_k) = n_k) = \frac{n!}{n_1!, \ldots, n_k!} p(A_1)^{n_1} \ldots p(A_k)^{n_k}.$$

Докажите, что если $X_1, \ldots, X_n$ – точки пуассоновского процесса с мерой μ внутри множества A, то справедливо равенство

$$\mathbb{P}(\cdot | N(A) = n) = \mathbb{P}_n(\cdot).$$

Замечание. Последнее выражение утверждает, что при фиксированном числе точек $N(A)$ внутри множества A, сами точки пуассоновского процесса выглядят как $N(A)$ независимых случайных величин с плотностью распределения $p(\cdot) = \mu(\cdot)/\mu(A)$. Таким образом, пуассоновский процесс является неизбежным следствием моделирования системы с большим числом независимых точек в пространстве $\mathbb{R}^m$.

9. Однородный пуассоновский процесс Π определен в пространстве $S = \mathbb{R}^2$ и имеет интенсивность λ, т.е. $\mu(A) = \int_A \lambda dx dy$. Осуществим переход к полярным координатам (r, θ) при помощи преобразования

$$f(x, y) = \left((x^2 + y^2)^{1/2},\ \arctan(y/x) \right).$$

Покажите, что образы точек Π образуют пуассоновский процесс в полосе

$$\{(r, \theta) :\ r > 0,\ 0 \leq \theta < 2\pi\},$$

который имеет интенсивность $\lambda^*(r) = \lambda r$. Покажите, что значения r, соответствующие Π, образуют пуассоновский процесс в $(0, \infty)$ с интенсивностью $2\pi\lambda r$.

Замечание. Сформулируем правило отображения Π при помощи произвольного правила преобразования координат $f : S \to T$: если мера μ процесса Π является σ-конечной (S можно представить в виде $\cup_n S_n$, где $\mu(S_n) < \infty$), помимо того мера на множестве образов T

$$\mu^*(B) = \mu(f^{-1}(B))$$

является неатомической, то $f(\Pi)$ – пуассоновский процесс в пространстве T с мерой интенсивности μ^*.

10. Используя результат предыдущей задачи, покажите, что плотность распределения упорядоченных расстояний $r_{(1)}, r_{(2)}, \ldots$ имеет следующий вид

$$f_{r_{(s)}}(r) = \frac{2(\lambda\pi)^s r^{2s-1} e^{-\lambda\pi r^2}}{(s-1)!}.$$

В таком случае, $2\lambda\pi r_{(s)}^2$ распределено как χ_{2s}^2. Данный результат может быть использован для оценки плотности точек на плоскости путем выбора случайных точек и измерения расстояния до 1-го ближайшего соседа. Пусть $X_1, \ldots, X_n$ – n реализаций $r_{(1)}^2$. Покажите, что $2\lambda\pi n\overline{X}$ распределено как χ_{2n}^2, предложите оценку для λ и вычислите ее дисперсию.

Замечание. Аналогичный подход с измерением переменной плотности точек $\lambda(\cdot)$ продемонстрирован в задаче **??** раздела 2 с использованием упорядоченного распределения Дирихле.

11. Сопоставим каждой точке $X \in S$ случайного множества Π (пуассоновский процесс) случайную величину m_X (метку), принимающую значения из множества M. Распределение $m_X \sim p(X, m)$ может зависеть от X, но не зависит от других точек Π и их меток. Докажите, что случайное множество

$$\Pi^* = \{(X, m_X) : \ X \in \Pi\}$$

является пуассоновским процессом на множестве $S \times M$ с мерой интенсивности

$$\mu^*(C) = \int\int\limits_{(x,m)\in C} \mu(dx)p(x,m)dm.$$

Используя замечание к задаче 9, убедитесь что метки m_X образуют пуассоновский процесс на M с мерой интенсивности

$$\mu_m(B) = \int_S \int_B \mu(dx)p(x,m)dm.$$

Указание. Докажем, что Π однозначно задается набором характеристических функционалов вида

$$\mathbb{E}(e^{\Sigma_f}) = \exp\left(-\int_S (1 - e^{-f(x)})\mu(dx)\right),$$

где

$$\Sigma_f = \sum_{X \in \Pi} f(X), \quad f \in \mathcal{F},$$

$\mathcal{F}$ – класс функций, содержащий все индикаторные функции для измеримых множеств из S. Выбрав набор непересекающихся множеств $A_1, \ldots, A_k$, сопоставим каждому множеству свое значение f_i. Тогда

$$\Sigma_f = \sum_{i=1}^{k} f_i N(A_i), \quad \mathbb{E}(e^{\Sigma_f}) = \exp\left(-\sum_{i=1}^{k} (1 - e^{-f_i}) \mu(A_i) \right).$$

Сделав замену $z_i = e^{-f_i}$, получаем

$$\mathbb{E}\left(z_1^{N(A_i)} \ldots z_k^{N(A_k)} \right) = \prod_{i=1}^{k} e^{\mu(A_i)(z-1)},$$

что свидетельствует о независимости $N(A_i)$ и принадлежности распределению Пуассона.

12. Случайный процесс $\varphi(t)$ называется *субординатором* (процесс Леви с положительными приращениями), если он удовлетворяет следующим свойствам: приращения процесса независимы, положительны и зависят только от приращения аргумента t. Докажите, что справедливо следующее представление субординатора

$$\varphi(t) = \beta t + \sum_{\tau} \{ z : \ (\tau, z) \in \Pi, \ 0 < \tau < t \},$$

где суммируются значения второй координаты пуассоновского процесса Π с мерой интенсивности

$$\mu(d\tau, dz) = d\tau \gamma(dz).$$

Мера $\gamma(dz)$ соответствует скачкам процесса $\varphi(t)$, образующем пуассоновский процесс по координате z, и может быть найдена из выражения (представление Леви-Хинчина)

$$\mathbb{E} e^{-s\varphi(t)} = \exp\left(-t \left[s\beta - \int_0^{\infty} (1 - e^{-sz}) \gamma(dz) \right] \right).$$

Указание. Используя результат предыдущей задачи, покажите, что для чисто атомической меры на S, определенной как

$$\Psi(A) = \sum_{\tau} \{z : \ (\tau, z) \in \Pi, \ \tau \in A\}$$

характеристическая функция имеет вид

$$\mathbb{E}e^{-s\Psi(A)} = \exp\left(-\int_0^\infty (1 - e^{-sz})\gamma(A, dz)\right),$$

где $\gamma(\cdot, \cdot)$ – мера процесса Π.

Покажите, что для случайной неатомической меры Φ с независимыми значениями на непересекающихся множествах $\Phi(A)$ есть безгранично делимая с.в. и для нее справедливо представление Леви-Хинчина:

$$\mathbb{E}e^{-s\Phi(A)} = \exp\left(-s\beta(A) + \int_0^\infty (1 - e^{-sz})\gamma(A, dz)\right).$$

Далее, $\varphi(t)$ может быть представлен как случайная мера $\Phi(0, t]$ при значениях $t \geq 0$. Убедитесь, что такая мера является неатомической, используя свойство инвариантности Φ относительно сдвига.

13. Рассмотрим звезды, находящиеся на расстоянии, не превышающем R от наблюдателя. Для простоты будем считать, что все звезды имеют одинаковый диаметр δ и равномерное пространственное распределение с количеством звезд λ на единицу объема. Показать, что при $R \to \infty$ любой участок неба будет полностью светящимся.

Замечание. В действительности такое явление не наблюдается. В связи с конечным возрастом вселенной ($14 \cdot 10^9$ лет) ее радиус ограничен величиной ct.

14 (Формула Крофтона)**.** Пусть N точек независимо распределены в области D n-мерного пространства, P – вероятность того, что фигура F, образованная N точками, обладает определенным свойством, зависящим только от взаимного расположения точек. Область D является измеримой по Лебегу и ее мара равна V. Обозначим как P_1 вероятность того, что F обладает требуемым свойством для случайных точек в области $D_1 \supset D$. Докажите следующее соотношение для малых приращений δV:

$$\delta P = N(P_1 - P)V^{-1}\delta V.$$

15. Вычислите распределение расстояния между точками, случайно взятыми внутри круга радиуса R, воспользовавшись формулой Крофтона.

Указание. Пусть $f(x,R)dx$ есть вероятность того, что расстояние между точками A и B принадлежит интервалу $(x, x+dx)$. Для случая, когда A лежит на границе

$$f(x,R) = \frac{2\theta x}{\pi R^2}, \quad \theta = \arccos(x/2R).$$

Тогда уравнение Крофтона примет вид

$$\frac{df}{d\theta} + 4f \operatorname{tg}\theta = \frac{32\theta}{\pi x}\sin\theta\cos\theta.$$

16 (Теорема Дворецкого)**.** Доказать, что для любого $\epsilon > 0$ и $k \in \mathbb{N}$ существует $N = N(k,\epsilon) < \exp(C\frac{\log\epsilon}{\epsilon^2}k)$ такое, что любое конечномерное банахово пространство $(X, \|\cdot\|)$, где $\dim X > N$, содержит k-мерное подпространство E, являющееся ϵ-евклидовым, т.е. в нем можно задать такую норму $|\cdot|$, что $\|x\| \leqslant |x| \leqslant (1+\epsilon)\|x\|$ $\forall x \in E$.

Замечание. См. работу В. Д. Мильман, "Новое доказательство теоремы А. Дворецкого о сечениях выпуклых тел", Функц. анализ и его прил., 5:4 (1971), 28–37.

17. Выберем наугад (равновероятно) k вершин m-мерного куба $[0,1]^m$. Обозначим как X выпуклую оболочку выбранных вершин. Пусть p_{km} - вероятность того, что все вершины многогранника попарно смежны. Докажите справедливость следующей оценки при $m > 3$:

$$p_{km} > 1 - \frac{k^4 \cdot 5^m}{4 \cdot 8^m}$$

Замечание. См. монографию Бондаренко В.А., Максименко А.Н. Геометрические конструкции, сложность в комбинаторной оптимизации. – М.: УРСС, 2008.

18. На плоскости нарисована выпуклая фигура, ограниченная кривой длины L. Докажите, что ее диаметр, т.е. максимальное расстояние между двумя ее точками, не меньше $\frac{L}{\pi}$.

Указание. Проведите в случайном направлении прямую. Покажите, что математическое ожидание длины проекции фигуры на случайное направление равно $\frac{L}{\pi}$.

19. В московском метро можно провозить коробки, у которых сумма измерений (длины, ширины и высоты) не превосходит некоторой границы. Можно ли перехитрить правила, поместив одну коробку в другую (сумма измерений внутренней коробки больше суммы измерений внешней)?

Указание. Спроектируйте коробку на случайно выбранное (в пространстве) направление. Длина проекции коробки складывается из проекций отрезков, идущих по ее высоте, длине и ширине. Проекция внутренней коробки не превосходит проекции внешней.

20. Несамопересекающаяся кривая длины 22 находится внутри круга радиуса 1. Докажите, что найдется прямая, имеющая с этой кривой по крайней мере 8 общих точек.

21. Известно, что более половины поверхности Земли занимают океаны. Докажите, что можно найти две диаметрально противоположные точки, обе попавшие в океан.

22. На плоскости расположено $2n$ векторов, выходящих из начала координат и длиной не более 1. Доказать, что существует угол α такой, что при повороте каждого из векторов на угол $\pm\alpha$, их векторная сумма окажется не большей 1.

23 (случайный поворот куба, мера Хаара)**.** 10% поверхности шара (по площади) выкрашено в черный цвет, остальные 90% – белые. Докажите, что можно вписать в шар куб таким образом, чтобы все вершины куба попали в белые точки.

24 (Теорема Уилкса)**.** Для выборки $X \sim f(\theta, X)$, $\dim X = n$ осуществляется проверка гипотезы $H_0 : \theta = \theta_0$, для чего используется обобщенный критерий отношения правдоподобия. Определим статистику для критерия как

$$W(\theta_0) = \log f(\widehat{\theta}, X) - \log f(\theta_0, X), \quad \widehat{\theta} = \arg\max_{\theta\in\Theta} f(\theta, X).$$

Теорема Уилкса утверждает, что в случае истинности гипотезы H_0 и при асимптотической нормальности оценки $\widehat{\theta}$, статистика $2W(\theta_0)$

сходится по распределению к χ^2_p, где $p = |\Theta|$. Тогда критерием отклонения гипотезы H_0 будет $[2W(\theta_0) > x]$, $\mathbb{P}(\chi^2_p > x) < \alpha$.

Существует также обобщение теоремы Уилкса для случая, когда $\widehat{\theta}$ не является асимптотически нормальной. При довольно слабых дополнительных ограничениях для выполнения теоремы Уилкса достаточно, чтобы поверхности уровня $S_w = \{\theta : W(\theta) = w\}$ в пределе при $n \to \infty$ имели форму

$$S_w \approx \widehat{\theta} + a_n w^r S$$

где $a_n \to 0$ при $n \to \infty$, $S = \{\theta : h(\theta) = 1\}$, $h(t\theta) = t^{1/r} h(\theta)$. В частности, при асимптотической нормальности оценки $\widehat{\theta}$ поверхности уровня являются эллиптическими:

$$S_w \approx \widehat{\theta} + \sqrt{\frac{2w}{n}} \theta^T D \theta.$$

Докажите при дополнительных ограничениях, $W(\widehat{\theta})$ строго локально вогнута и $\inf\{h(\theta) : \|\theta\| = 1\} > 0$, что

$$W(\theta_0) \xrightarrow{d} \Gamma(rp, 1),$$

где плотность гамма $\Gamma(\alpha, \lambda)$ распределения задается как $\frac{x^{\alpha-1} e^{-x/\lambda}}{\lambda^\alpha \Gamma(\alpha)}$.

Указание. См. работу Fan et al. Geometric Understanding of Likelihood Ratio Statistics. Department of Statistics, UCLA 1998.

7. Вероятностные основы математической статистики

1. В байесовской теории оценивания оптимальная оценка $\widehat{\theta}$ для функции потерь λ, выборки X и параметрического распределения $\mathbb{P}(X|\theta)$ определяется по следующему правилу

$$\widehat{\theta} = \arg\min_{\theta} \mathbb{E}_{\mathbb{P}(\theta^*|X)} \lambda(\theta, \theta^*).$$

Докажите соответствие функций потерь и оптимальных оценок, приведенных в таблице

функция потерь $\lambda(\theta, \theta^*)$	оптимальная оценка
$\sum_i [\theta_i \neq \theta_i^*]$	$\widehat{\theta_i} = \arg\max_{\theta_i} \mathbb{P}(\theta_i \mid X)$
$\sum_i (\theta_i - \theta_i^*)^2$	$\widehat{\theta_i} = \mathbb{E}_{\mathbb{P}(\theta_i \mid X)} \theta_i$
$[\theta \neq \theta^*]$	$\widehat{\theta} = \arg\max_{\theta} \mathbb{P}(\theta \mid X)$

2 (Эмпирическая функция распределения)**.** Пусть $\mathrm{x}_k,\ k = 1, ..., n$ – независимые одинаково распределенные с.в. с непрерывной функцией распределения $F(x)$. Введём

$$F\left(x; \{\vec{\mathrm{x}}\}_n\right) = \frac{1}{n} \sum_{k=1}^{n} I\left(\mathrm{x}_k < x\right), I\left(\mathrm{x}_k < x\right) = \begin{cases} 1, & \mathrm{x}_k < x \\ 0, & \mathrm{x}_k \geq x \end{cases} .$$

а) Покажите, что $F\left(x; \{\vec{\mathrm{x}}\}_n\right) \xrightarrow[n\to\infty]{\text{п.н.}} F(x)\ \forall x \in \mathbb{R}$.

б) (теорема Гливенко) Покажите, что

$$\max_{x \in \mathbb{R}} \left| F\left(x; \{\vec{\mathrm{x}}\}_n\right) - F(x) \right| \xrightarrow[n\to\infty]{\text{п.н.}} 0.$$

Замечание. (Неравенство Дворецкого–Кифера–Вольфовица)

Имеет место более тонкий результат:

$$P\left(\sqrt{n} \max_{x \in \mathbb{R}} \left| F\left(x; \{\vec{\mathrm{x}}\}_n\right) - F(x) \right| \geq \varepsilon \right) \leq 2e^{-2\varepsilon^2}.$$

Этот замечание, равно как и идеи следующих трех задач, взяты из конспекта лекций Г.К. Голубева "Введение в математическую статистику".

в) Какие из трех приведенных фактов справедливы без условия непрерывности функции $F(x)$?

3 (Бутстреп, оценка дисперсии)**.** $T = T(X_1, \ldots, X_n)$ некоторая статистика i.i.d. выборки из распределения F. Для оценки дисперсии $\mathbb{D}_F(T)$ можно воспользоваться выборочной функцией распределения $\widehat{F}_n$:

$$\mathbb{D}_{\widehat{F}}(T) = \int (T - \mathbb{E}T)^2 d\widehat{F}_n(X_1) \ldots d\widehat{F}_n(X_n).$$

Выражение для $\mathbb{D}_{\widehat{F}}(T)$ в некоторых случаях удается подсчитать в явном виде (например при $T = \overline{X}$, $\mathbb{D}_{\widehat{F}}(T) = \frac{1}{n}\sum(X - \overline{X})^2$), но в общем случае используется метод Монте–Карло для приближенного вычисления интеграла:

а) Выполнить B раз генерацию выборки $X_1^*, \ldots, X_n^* \sim \widehat{F}_n$;

б) Вычислить значения $T_1^*, \ldots, T_B^*$;

в) Найти оценку $\mathbb{D}_{\widehat{F}}(T)$ по формуле

$$V_{\text{boot}}(T) = \frac{1}{B}\sum_{b=1}^{B}\left(T_b^* - \overline{T}^*\right)^2.$$

Существует также альтернативный метод оценки $\mathbb{D}_F(T)$ – *метод складного ножа* – где выборка $X_1^*, \ldots, X_{n-1}^*$ генерируется посредством поочередного выбрасывания одного из элементов выборки. Обозначим за T_{-i} значение статистики, подсчитанное на основе подвыборки $X_1, \ldots, X_{i-1}, X_{i+1} \ldots, X_n$, тогда оценка равна

$$V_{\text{jack}}(T) = \frac{n-1}{n}\sum_{i=1}^{n}\left(T_{-i} - \overline{T}\right)^2.$$

Покажите, что оценка квантилей распределения F посредством V_{boot} является состоятельной ($V_{\text{boot}}(X_{(k)}) \to \mathbb{D}\,X_{(k)}$ при $n \to \infty$), в то время как $V_{\text{jack}}(X_{(k)})$ не является состоятельной, однако

$$\frac{V_{\text{jack}}(X_{(k)})}{\mathbb{D}\,X_{(k)}} \xrightarrow{p} 1.$$

4 (Бутстреп, доверительные интервалы)**.** Существует несколько методов оценки доверительных интервалов на основе бутстрепа (к примеру, нормальный интервал, центральный интервал, интервал на основе процентилей). Рассмотрим способ нахождения центрального интервала.

Обозначим через $T_1^*, \ldots, T_B^*$ повторную выборку значений статистики $T(X_1^*, \ldots, X_n^*)$ на основе бутстрепа, а через $F_\triangle$ распределение случайной величины $\triangle_n = T(X_1, \ldots, X_n) - \mathbb{E}T$. Определим доверительный интервал (a, b) по формуле

$$a = T(X_1, \ldots, X_n) - F_\triangle^{-1}\left(1 - \frac{\alpha}{2}\right), \quad b = T(X_1, \ldots, X_n) - F_\triangle^{-1}\left(\frac{\alpha}{2}\right).$$

Убедитесь, что $\mathbb{P}(a \le \mathbb{E}T \le b) = 1 - \alpha$, где α – заданный уровень значимости. Неизвестное распределение $F_\triangle$ можно оценить как

$$\widehat{F}_\triangle(x) = \frac{1}{B}\sum_{b=1}^{B}[T_b^* - T_n < x].$$

Предложите способ нахождения $\widehat{F}_\triangle^{-1}(x)$, например с использованием квантилей выборки $T_1^* - T_n, \ldots, T_n^* - T_n$.

5 (Оценка максимума правдоподобия)**.** Докажите приведенные ниже свойства ОМП оценок

$$\widehat{\theta} = \arg\max L(X, \theta) = \arg\max \sum_{i=1}^{n} \log f(X_i, \theta).$$

а) ОМП состоятельная, то есть $\widehat{\theta} \to \theta^*$;

б) ОМП не зависит от параметризации, то есть при замене параметра $\eta = \eta(\theta)$, $\widehat{\eta} = \eta(\widehat{\theta})$;

в) ОМП асимптотически нормальна

$$\frac{\widehat{\theta} - \theta^*}{\sqrt{S}} \xrightarrow{d} \mathcal{N}(0, 1), \quad S = \widehat{\mathbb{D}}(\widehat{\theta});$$

г) ОМП асимптотически оптимальна (имеет наименьшую дисперсию).

Указание. Введем функцию

$$\mathcal{KL}_n(\theta) = \frac{1}{n}\sum_i \log\frac{f(X_i,\theta)}{f(X_i,\theta^*)}.$$

Потребуем, чтобы имела место сходимость по вероятности равномерно по θ

$$\mathbb{E}_X\mathcal{KL}_n(\theta) \to -\mathcal{KL}(\theta^*,\theta) = -\int f(x,\theta^*)\log\frac{f(x,\theta^*)}{f(x,\theta)}dx.$$

Докажите тождество

$$\mathbb{P}(\mathcal{KL}(\theta^*,\widehat{\theta}) > \mathcal{KL}(\theta^*,\theta^*) + \delta) \to 0,$$

из которого в случае локальной вогнутости $\mathcal{KL}(\theta^*,\theta)$ в окрестности нуля будет следовать состоятельность ОМП.

Для доказательства асимптотической нормальности потребуем разложимость $L(\theta)$ по формуле Тейлора в точке θ^* до второго члена, тогда справедливо соотношение

$$0 = \nabla L(\widehat{\theta}) \approx \nabla L(\theta^*) + \nabla^2 L(\theta^*)(\widehat{\theta} - \theta^*),$$

из которого следует сходимость

$$\sqrt{n}(\widehat{\theta} - \theta^*) \to \frac{\frac{1}{\sqrt{n}}\nabla L(\theta^*)}{\frac{1}{n}\nabla^2 L(\theta^*)}.$$

Введем обозначение $Y_i = \nabla \log f(X_i,\theta)$. Покажите, что

$$\mathbb{E}(Y_i) = 0, \quad \mathbb{D}(Y_i) = I(\theta) = -\int \big(\nabla^2 \log f(x,\theta)\big) f(x,\theta)dx.$$

Тогда из ЦПТ следует, что $\sqrt{n}\overline{Y} \to \mathcal{N}(0, I(\theta))$.

6 (Точность ОМП)**.** Ввиду заложенной в наблюдаемые данные X случайности ОМП оценка $\widehat{\theta}(X) = \arg\max L(X,\theta)$ является случайной величиной, сходящейся в пределе к истинному значению параметра θ^*. Отметим, что для выборки X, сгенерированной при помощи модели P отличной от $L(X,\theta)$, под "истинным" значением параметра θ^* подразумевается

$$\theta^* = \arg\max \mathbb{E}L(X,\theta)$$

Такой выбор θ^* соответствует наиболее близкому распределнию из параметрического семейства $L(X,\theta)$ к распределению P.

Для оценки отклонения $\widehat{\theta}$ от θ^* исследуем *квадратичную* модель $L(Y,\theta)$ на примере регрессии

$$Y = X^T\theta + \varepsilon, \quad \dim\theta = p, \varepsilon \in \mathcal{N}(0,S).$$

Предположим, что выборка Y была сгенерирована из распределения с параметрами $\mathbb{E}Y = M_Y$, $\mathbb{D}Y = S_0$, где S_0 может не совпадать с S. Докажите справедливость следующих выражений

$$D_0(\widehat{\theta} - \theta^*) = \xi \sim \mathcal{N}(0, I),$$

$$L(Y,\widehat{\theta}) - L(Y,\theta^*) = \frac{\|\xi\|^2}{2} \sim \frac{\chi^2_p}{2},$$

где

$$D_0^2 = -\nabla^2\mathbb{E}L(Y,\theta)\Big|_{\theta^*} = XS^{-1}X^T,$$

$$\xi = D_0^{-1}\nabla L(\theta^*) = D_0^{-1}XS^{-1}\varepsilon,$$

$$\theta^* = (XS^{-1}X^T)^{-1}XS^{-1}M_Y, \quad \widehat{\theta} = (XS^{-1}X^T)^{-1}XS^{-1}Y.$$

Замечание. Отметим, что вне зависимости от распределения Y для нахождения истинного значения θ^* достаточно знать $\mathbb{E}Y$. Также стоит заметить, что вектор ξ зависит от неизвестного параметра θ^* и может быть оценен с погрешностью в точке $\widehat{\theta}$ или при помощи метода бутстреп (см. задачу **??**).

7. Случайный вектор $\xi \in \mathbb{R}^p$ имеет следующую суб-гауссовскую верхнюю границу характеристической функции

$$\mathbb{E}\exp\left(\lambda\frac{\gamma^T\xi}{\|V\gamma\|}\right) \le \exp\left(\frac{c^2\lambda^2}{2}\right), \quad V^2 = \mathbb{D}\xi.$$

Докажите следующие неравенства и найдите константы c_1, c_2.

$$\forall\mu < 1: \ \exp\left(\frac{\mu\|\xi\|^2}{2}\right) \le c_1\int\exp\left(\gamma^T\xi - \frac{\|\gamma\|^2}{2\mu}\right)d\gamma,$$

$$\mathbb{P}(\|\xi\|^2 - \mathbb{E}\|\xi\|^2 > xc_2) \le 2e^{-x}.$$

8 (Приближение Лапласа)**.** Чтобы исследовать отклонение ОМП оценки от истинного значения ($\widehat{\theta} - \theta^*$, а также $L(Y, \widehat{\theta}) - L(Y, \theta^*)$) для произвольной модели $L(Y, \theta)$, необходимо наложить ряд условий на функцию правдоподобия, обеспечивающих приближение $L(Y, \theta)$ квадратичной формой $\mathbb{L}(Y, \theta)$ в эллиптической окрестности точки θ^*:

$$\theta \in \Theta(r) = \{\theta : \|D_0(\theta - \theta^*)\| \le r\}.$$

Наложим ограничения на первую и вторую производные $L(Y, \theta)$ в окрестности $\Theta(r)$:

а) $D^2(\theta) = -\nabla^2 \mathbb{E}_Y L(Y, \theta)$ непрерывна и

$$\|D_0^{-1} D^2(\theta) D_0^{-1} - I\| \le \delta(r), \quad D_0(\theta) = D(\theta^*);$$

б) $\nabla \overset{o}{L}(Y, \theta)$ является суб-гауссовской с.в., т.е. существуют такие константы ω и c, что $\forall \gamma_1, \gamma_2, \lambda: \ \|\gamma\| \le 1, \ \|\gamma_2\| \le 1$

$$\mathbb{E}_Y \exp\left(\frac{\lambda}{\omega} \gamma_1^T D_0^{-1} \nabla^2 \overset{o}{L}(Y, \theta) D_0^{-1} \gamma_2\right) \le \exp\left(\frac{c^2 \lambda^2}{2}\right).$$

Покажите, что при наложенных ограничениях справедливы неасимптотические варианты теорем Фишера и Вилкса, с вероятностью не менее $1 - e^{-x}$ в окрестности $\Theta(r)$

$$\|D_0^{-1} (\nabla L(\theta) - \nabla L(\theta^*)) - D_0(\theta - \theta^*)\| \le \diamondsuit(r, x),$$

$$\left| \left(L(\theta) - L(\theta^*) \right) - \frac{(D_0(\theta - \theta^*))^2}{2} \right| \le r \diamondsuit(r, x),$$

где

$$\diamondsuit(r, x) = r\delta(r) + 6\omega c r \sqrt{2p + 2x},$$

для простой выборки характерны следующие значения констант

$$\omega \sim \frac{1}{\sqrt{n}}, \quad \delta(r) \sim \frac{r}{\sqrt{n}}, \quad r^2 \sim (p + x).$$

Указание. Воспользуйтесь теоремой для оценки отклонения центрированного случайного процесса $u(Y, s)$ от нуля в окрестности $\{s : d(s, s_o) < r\}$. Пусть выполнено условие

$$\mathbb{E}_Y \exp\left(\lambda \frac{u(Y, s) - u(Y, s_o)}{d(s, s_o)}\right) \le \exp\left(\frac{c^2 \lambda^2}{2}\right),$$

тогда с вероятностью не менее $1-e^{-x}$ в окрестности $\{s : d(s, s_o) < r\}$

$$\frac{1}{3cr}|u(Y,s) - u(Y,s_0)| \le \sqrt{Q+2x}, \quad Q = 2p \text{ при } s \in \mathbb{R}^p,$$

в более общем случае Q представляет собой энтропию пространства значений параметра s.

9 (Критическая размерность)**.** В регрессионной модели, представленной ниже, требуется найти только первую компоненту параметра $\theta \in \mathbb{R}^p$.

$$X_i = \begin{pmatrix} \theta_1 + \|\theta\|^2 \\ \theta_2 \\ \vdots \\ \theta_{p(n)} \end{pmatrix} + \varepsilon_i, \quad \varepsilon_i \in N(0, I_p), \quad i = \overline{1,p}$$

Положим значение истинного параметра θ^* равным нулю. Покажите, что оценка максимума правдоподобия $\widehat{\theta}_1$ сходится с ростом n к нулю со скоростью $n^{-1/2}$ только при $p(n) = o(\sqrt{n})$.

10. Исследуем сепарабельный процесс $U(v)$, $v \in \Upsilon$, характеризуемый суб-экспоненциальным ограничением характеристической функции

$$\mathbb{E}\exp\left(\lambda \frac{U(v) - U(v')}{d(v,v')}\right) \le e^{\nu_0^2 \lambda^2/2}, \quad |\lambda| \le g,\ d(v,v') \le r_0,\ \nu_0 \ge 1.$$

На множестве Υ задана σ-конечная мера π. Введем обозначения

$$B_r(v) = \{u \in \Upsilon : d(v,u) \le r\}, \quad r_k = 2^{-k} r_0, \quad \pi_k(v) = \int_{B_k(u)} \pi(du),$$

$$M_k = \max_{v \in \Upsilon} \frac{\pi(\Upsilon)}{\pi_k(u)},$$

$$H_1 = \sum_{k=0}^{\infty} c_k \sqrt{2\log(2M_k)}, \quad H_2 = 2\sum_{k=0}^{\infty} c_k \log(2M_k), \quad c_0 = \frac{1}{3},\ c_k = \frac{2^{1-k}}{3}.$$

Докажите, что с вероятностью не менее $1 - e^{-x}$ имеет место неравенство

$$\frac{1}{3\nu_0 r_0} \sup_{v \in B_{r_0}(v_0)} \{U(v) - U(v_0)\} \le H(x) = H_1 + \sqrt{2x} + \frac{g^{-2}x + 1}{g} H_2.$$

Указание. Введем замену $U(\cdot) = \nu_0^{-1} U(\cdot)$, что равносильно случаю $\nu_0 = 1$, $g_0 = g\nu_0$, а также определим оператор S_k по правилу

$$S_k f(v_0) = \frac{1}{\pi_k(v_0)} \int_{B_k(v_0)} f(v)\pi(dv), \quad k \geq 0,$$

$$S_{-1}U(v) = U(v_0) = S_k S_{k-1} \dots S_{-1}U(v).$$

$$|U(v) - U(v_0)| = \lim_{k\to\infty} |S_k U(v) - S_k S_{k-1} \dots S_{-1}U(v)| \leq$$

$$\leq \lim_{k\to\infty} \sum_{i=0}^{k} |S_k \dots S_i (I - S_{i-1})U(v)| \leq \sum_{i=0}^{\infty} \xi_i^*,$$

где

$$\xi_i^* = \sup_{v\in B_{r_0}(v_0)} \xi_i(v), \quad \xi_0(v) = |S_0 U(v) - U(v_0)|, \quad \xi_i(v) = |S_i(I - S_{i-1})U(v)|.$$

Докажите свойство

$$\mathbb{E} e^{\lambda \xi_i^* / r_{i-1}} \leq 2M_i e^{\lambda^2/2},$$

воспользуйтесь результатом задачи **??** из раздела **??**.

11. В контексте задачи 10 рассмотрите частный случай

$$\Upsilon = B(r_0, v_0) = \{v \in \mathbb{R}^p : \|D(v - v_0)\| \leq r_0\},$$

где

$$d(v, v_0) \leq \|D(v - v_0)\|.$$

Докажите верхнюю оценку

$$H_2 \leq 4p$$

при $p \geq 2$, а также соотношение $H_1 \leq \sqrt{H_2}$.

Указание. Не теряя общности, можно выполнить расчеты для случая

$$v_0 = 0, \quad D = I_p, \quad r_0 = 1.$$

Чтобы найти верхнюю границу M_k, оценим снизу величину

$$\pi(B(1, 0) \cap B(r, v))$$

, значение которой достигает минимума при $\|v\| = 1$. Рассмотрите шар $B(\rho, u)$, где $\rho = r - r^2/2$, $u = (1 - r^2/2)v$. Покажите, что

$$B(r, v) \supseteq B(\rho, u), \quad \pi(B(1, 0) \cap B(\rho, u)) \geq \pi(B(\rho, u))/2,$$

откуда будет следовать, что $2M_k \leq 2^{2+kp}(1 - 2^{-k-1})^{-p}$ при $r_k = 2^{-k}$.

Замечание. При наличии регуляризатора $\|Gv\|^2$, при котором появляется дополнительное ограничение на значения параметра v:

$$B(r, v_0) = \{v \in \mathbb{R}^p : \|D(v - v_0)\|^2 + \|Gv\|^2 \leq r_0^2\}$$

в результате *эффективная* размерность параметра v уменьшается согласно формулам

$$H_2 \leq 1 + \frac{8}{3}\mathrm{tr}[B^{-1}], \quad H_1 \leq 1 + 2\mathrm{tr}^{1/2}[B^{-2}\log^2(B^2)],$$

где $B^2 = I_p + D^{-1}G^2D^{-1}$.

12. Докажите, что из ограничения на градиент случайного процесса $U(v)$ вида

$$\log \mathbb{E}\exp\left\{\lambda \frac{\gamma^T \nabla U(v)}{\|D(v)\gamma\|}\right\} \leq \frac{\nu_0^2\lambda^2}{2},$$

следует ограничение

$$\log \mathbb{E}\exp\left\{\lambda \frac{U(v) - U(v_0)}{\|D(v - v_0)\|}\right\} \leq \frac{\nu_0^2\lambda^2}{2},$$

где

$$\nu_0 \geq 1, \ |\lambda| \leq g, \ \gamma \in \mathbb{R}^p, D \succcurlyeq D(v).$$

Указание. Представьте приращение процесса U в виде

$$U(v, v_0) = \delta\gamma^T \int_0^1 \nabla U(v_0 - t\delta\gamma)dt, \quad \delta = \|v - v_0\|, \ \gamma = (v - v_0)/\delta.$$

13 (Локализация ОМП)**.** Чтобы определить границы локальной области параметра v – $B_r(v_0)$, вне которой с малой вероятностью выполнено неравенство

$$\mathcal{U}(v) - \mathcal{U}(v_0) \geq 0,$$

разделим разности сепарабельного процесса $\mathcal{U}(v) - \mathcal{U}(v_0)$ на стохастическую $U(v, v_0)$ и детерминированную $(-f(v, v_0, \rho))$ части. Предположим, что для $U(v, v_0)$ выполнены условия из задачи 10, а также

$$f(v, v_0, \rho) \geq 3\nu_0 r H\left(x + \log\left(\frac{\rho^{-1} d(v, v_0)}{r_0}\right)\right), \quad r_0 \leq d(v, v_0) \leq r^*.$$

Докажите, что с вероятностью менее $\frac{\rho}{1-\rho} e^{-x}$ имеет место неравенство

$$\sup_{v \in B_{r^*}(v_0) \setminus B_{r_0}(v_0)} \{U(v, v_0) - f(v, v_0, \rho)\} \geq 0, \quad 0 < \rho \leq 1.$$

Указание. Разбейте область $B_{r^*}(v_0) \setminus B_{r_0}(v_0)$ на слои с радиусами $r_k = r_0 \rho^{-k}$, примените для каждого слоя результат задачи 10.

Замечание. Найдем область, в которой $L(\theta) - L(\theta^*) < 0$ с большой вероятностью. Данная область является дополнением к области локализации ОМП. Пусть для $\zeta(\theta) = L(\theta, Y) - \mathbb{E}_Y L(\theta, Y)$ выполнено условие

$$\mathbb{E}_Y \exp\left(\frac{\lambda}{\omega} \gamma_1^T D_0^{-1} \nabla^2 \zeta(\theta) D_0^{-1} \gamma_2\right) \leq \exp\left(\frac{\nu_0^2 \lambda^2}{2}\right), \quad \|\gamma_1\| = \|\gamma_2\| = 1.$$

Тогда из результатов задач 10 и 12 следует, что в локальной области $\|D_0(\theta - \theta^*)\| \leq r$ выполнено

$$|\zeta(\theta, \theta^*) - (\theta - \theta^*)^T \nabla \zeta(\theta^*)| \leq 6\nu_0 H(x) w r.$$

Из результата данной задачи получаем неравенство

$$|L(\theta, \theta^*) - \mathbb{E} L(\theta, \theta^*) - (\theta - \theta^*)^T \nabla L(\theta^*)| \leq$$

$$\leq 6\nu_0 w r H\left(x + \log\left(\frac{2r}{r_0}\right)\right) = \rho(r, x) r.$$

Так как

$$|(\theta - \theta^*)^T \nabla L(\theta^*)| \leq r\|\xi\|,$$

то из неравенств

$$-2\mathbb{E} L(\theta, \theta^*) \geq \|D_0(\theta, \theta^*)\|^2 b, \quad r^* b(r^*) \geq 2(\|\xi\| + \rho(r^*, x))$$

получаем возможность найти радиус локальной области ОМП r^*.

14. Пусть $y(v) : \mathbb{R}^p \to \mathbb{R}^q$ – гладкий случайный процесс, причем $\mathbb{E}y(v) = 0$, $y(v_0) = 0$. Без ограничения общности можно положить $v_0 = 0$. Предположим, что для любых $\|\gamma\| = \|\alpha\| = 1$

$$\log \mathbb{E}\exp\left(\lambda \gamma^T \nabla y(v) \alpha\right) \le \frac{\nu_0^2 \lambda^2}{2}, \quad |\lambda| \le g.$$

Докажите неравенство

$$\mathbb{P}\left(\sup_{\|v - v_0\| \le r} \|y(v)\| > 6\nu_0 r H(x)\right) \le e^{-x},$$

где

$$H(x) = H_1 + \sqrt{2x} + \frac{g^{-2}x + 1}{g} H_2, \quad H_1 = \sqrt{4(p+q)},\ H_2 = 4(p+q).$$

Указание. Используя утверждение задачи 12, получите неравенство

$$\log \mathbb{E}\exp\left(\frac{\lambda}{r}\gamma^T y(v)\right) \le \frac{\nu_0^2 \lambda^2 \|v - v_0\|^2}{2r^2}.$$

Представьте норму вектора в виде

$$\|y(v)\| = \sup_{\|u\| \le r} \frac{1}{r} u^T y(v).$$

Получите неравенство

$$\log \mathbb{E}\exp\left(\frac{\lambda}{2r}(\gamma, \alpha)^T \nabla [u^T y(v)]\right) \le \frac{\nu_0^2 \lambda^2}{2}.$$

15 (Теорема Уилкса)**.** Функция правдоподобия $L(\theta, Y)$ достигает максимального значения при параметре $\widehat{\theta}$, θ^* – истинные значения параметра. Введем также набор параметров $\widehat{\theta}_m$, при котором достигается минимум функции L при условии, что $\forall i \in [1, m] :\ \widehat{\theta}_m(i) = \theta^*(i)$. Определим статистику

$$W(\theta, Y) = L(\widehat{\theta}, Y) - L(\widehat{\theta}_m, Y)$$

Предполагая возможность квадратичной аппроксимации Лапласа (см. задачу **??**), докажите слабую сходимость $2W(\theta, Y)$ к распределению хи-квадрат со степенями свободы m при неограниченном увеличении выборки.

16. Для следующих примеров выборок, проверьте асимптотическую ненормальность ОМП оценок. Исходя из результата задачи **??** из раздела **??**, докажите, что, тем не менее, имеет место сходимость статистики $W(\theta, Y)$ к распределению хи-квадрат.

а) $Y_1, \ldots, Y_n \sim N(\theta^3, I_p)$, $\widehat{\theta} = \overline{Y}^{1/3}$;

б) $Y_i \sim \theta + \varepsilon_i$, $\mathbb{P}(\varepsilon_{ik} > x) = e^{-x}$, $\widehat{\theta}_k = \min\{Y_{1k}, \ldots, Y_{nk}\}$;

в) $Y_1, \ldots, Y_n$ выборка с заданной совместной плотностью распределения

$$p(Y_1, \ldots, Y_n, y, \theta) = ce^{\left(-n\|y-\theta\|^\gamma - \sum_{i=1}^n (Y_i - y)^2\right)}.$$

Указание. В приведенных примерах поверхности уровня определяются как

а) $S_w = \{\theta : n\|\overline{Y} - \theta^3\|^2 = 2w\} \approx \overline{Y}^{1/3} - (3\overline{Y}^{2/3})^{-1}\sqrt{2w/n}S^{1/3}$, S – единичная сфера;

б) $S_w = \{\theta : n\sum_i(\widehat{\theta}_i - \theta_i) = w,\ \widehat{\theta} > \theta\} = \widehat{\theta} + (w/n)S$, S – единичный симплекс;

в) $S_w = \overline{Y} + (w/n)^{1/\gamma}S$, S – единичная сфера.

17 (Неравенство Ван Трисса)**.** Пусть x_k, $k = 1, ..., n$ – независимые одинаково распределенные с.в. с плотностью распределения $p_{\vec{\mathrm{x}}}\left(\vec{x}, \vec{\theta}\right)$, зависящей от случайного вектора $\vec{\theta}$ (носитель распределения $\vec{\mathrm{x}}$ предполагается не зависящим от $\vec{\theta}$, т.е. $p_{\vec{\mathrm{x}}}\left(\vec{x}, \vec{\theta}\right)$ - регулярное семейство), который имеет плотность распределения $\pi(\vec{\theta})$:

$$\lim_{\|\vec{\theta}\| \to \infty} \left(\left\| \vec{\theta} \right\| \pi\left(\vec{\theta}\right) \right) = 0.$$

Покажите (для скалярного случая когда $\vec{\theta} \in \mathbb{R}$), что для любой измеримой вектор-функции $\vec{\widehat{\theta}}\,(\vec{\mathrm{x}})$:

$$D_{\vec{\mathrm{x}},\vec{\theta}}\left[\vec{\widehat{\theta}}\,(\vec{\mathrm{x}})\right] < \infty$$

имеет место неравенство:

$$\mathbb{E}_{\vec{\mathrm{x}},\vec{\theta}}\left[\left(\vec{\tilde{\theta}}\left(\vec{\mathrm{x}}\right)-\vec{\theta}\right)\left(\vec{\tilde{\theta}}\left(\vec{\mathrm{x}}\right)-\vec{\theta}\right)^{T}\right]\succ\left[I_{p,n}+I_{\pi}\right]^{-1},$$

т.е.

$$\mathbb{E}_{\vec{\mathrm{x}},\vec{\theta}}\left[\left(\vec{\tilde{\theta}}\left(\vec{\mathrm{x}}\right)-\vec{\theta}\right)\left(\vec{\tilde{\theta}}\left(\vec{\mathrm{x}}\right)-\vec{\theta}\right)^{T}\right]-\left[I_{p}+I_{\pi}\right]^{-1}$$

– неотрицательно определенная матрица.

Здесь информационные матрицы (Фишера) рассчитываются по формулам:

$$I_{p,n}\overset{def}{=}E_{\vec{\mathrm{x}},\vec{\theta}}\left[\frac{\partial\ln p_{\vec{\mathrm{x}}}\left(\vec{\mathrm{x}},\vec{\theta}\right)}{\partial\vec{\theta}}\left(\frac{\partial\ln p_{\vec{\mathrm{x}}}\left(\vec{\mathrm{x}},\vec{\theta}\right)}{\partial\vec{\theta}}\right)^{T}\right]=nI_{p,1}<\infty,$$

$$I_{\pi}\overset{def}{=}E_{\vec{\theta}}\left[\frac{\partial\ln\pi\left(\vec{\theta}\right)}{\partial\vec{\theta}}\left(\frac{\partial\ln\pi\left(\vec{\theta}\right)}{\partial\vec{\theta}}\right)^{T}\right]<\infty.$$

Замечание. В случае неинформативного $\pi(\vec{\theta})$ (несобственное равномерное распределение на всем пространстве, которое получается предельным переходом из равномерного на шаре, при стремлении радиуса шара к бесконечности) и $E_{\vec{\mathrm{x}}}\left[\vec{\tilde{\theta}}\left(\vec{\mathrm{x}}\right)\right]\equiv\vec{\theta}$, неравенство Ван Трисса переходит в намного более известное по классическим стохастическим курсам *неравенство Рао–Крамера*, которое мы приводим (для наглядности) в скалярном случае:

$$D_{\vec{\mathrm{x}}}\left[\tilde{\theta}\left(\vec{\mathrm{x}}\right)\right]\geq n^{-1}E_{\vec{\mathrm{x}}}\left[\left(\partial\ln p_{\mathrm{x}}\left(x,\theta\right)/\partial\theta\right)^{2}\right]^{-1}.$$

В классе оценок, для которых смещение $b(\vec{\theta})=E[\tilde{\theta}-\vec{\theta}]\not\equiv 0$ ненеравенство Рао-Крамера имеет вид

$$D_{\vec{\mathrm{x}}}\left[\tilde{\theta}\left(\vec{\mathrm{x}}\right)\right]\geq\left(1+b'(\vec{\theta})\right)^{2}n^{-1}E_{\vec{\mathrm{x}}}\left[\left(\partial\ln p_{\mathrm{x}}\left(x,\theta\right)/\partial\theta\right)^{2}\right]^{-1}.$$

Замечание. Несобственная плотность распределения отличается от классического определения плотности распределение тем, что интеграл по области определения расходится. Примером несобственной плотности распределения может служить константная функция, заданная на действительной оси (обобщение равномерной плотности с бесконечным интервалом определения).

Указание. Для скалярного случая $\vec{\theta} = \theta$ воспользуйтесь неравенством Коши–(Шварца)–Буняковского: $\langle a, b\rangle^2 \leq \langle a, a\rangle \langle b, b\rangle$, рассмотрев случай, когда $\langle a, b\rangle \stackrel{def}{=} E(ab)$, где $a = \tilde{\theta}(\vec{\mathrm{x}}) - \theta$, $b = \frac{\partial(\ln p_{\vec{\mathrm{x}}}(\vec{x},\theta)\pi(\theta))}{\partial\theta}$.

18. $F(x)$ – произвольная функция распределения с нулевым средним и конечным стандартным отклонением, $Y_i \in F_\theta$, где $F_\theta(x) = F(x-\theta)$, $\theta \in \mathbb{R}$. Приведите пример $F(x)$ для которого оценка $(\min Y_i + \max Y_i)/2$ имеет меньшую дисперсию (равную $O(1/n^2)$) нежели $\overline{Y}$. Не противоречит ли приведенный пример неравенству Рао-Крамера?

Замечание. Оценка максимального правдоподобия параметра сдвига θ зачастую принимает следующий вид

$$\widehat{\theta} = \sum_{i=1}^{n} a_i Y_{(i)},$$

где $\sum a_i = 1$, $Y_{(i)}$ – i-я порядковая статистика. В этом случае если $\widehat{\theta}$ отлично от $\overline{Y}$, то не более двух коэффициентов среди a_i отличны от нуля: это либо a_1 и a_n и в этом случае элементы выборки распределены равномерно, либо a_i и a_{i+1}, либо только один коэффициент не нулевой.

19 (Байесовская оценка с квадратичным штрафом)**.** В условия предыдущей задачи введем штраф: $I(\vec{\theta'}(\,\cdot\,), \vec{\theta})$. Оценка $\vec{\tilde{\theta}}(\vec{\mathrm{x}})$ вектора неизвестных параметров $\vec{\theta}$ называется *байесовской*, если для любого $\vec{x}$:

$$\vec{\tilde{\theta}}(\vec{x}) = \arg\min_{\vec{\theta'}} \int I\left(\vec{\theta'}, \vec{\theta}\right) p_{\vec{\mathrm{x}}}\left(\vec{x}, \vec{\theta}\right) \pi\left(\vec{\theta}\right) d\vec{\theta}.$$

а) (Уравнение Винера–Хопфа) Покажите, что если $I\left(\vec{\theta'}(\,\cdot\,), \vec{\theta}\right) = \left\|\vec{\theta'}(\vec{\mathrm{x}}) - \vec{\theta}\right\|_2^2$ и $\left(\vec{\mathrm{x}}^T, \vec{\theta}^T\right)^T$ – нормальный случайный вектор, то $\vec{\tilde{\theta}}(\vec{x}) = A\vec{x} + \vec{b}$ (линейная регрессия).

б) Рассмотрим следующую схему эксперимента $\mathrm{x}_i = \theta + \varepsilon_i$, где $\varepsilon_i \in N(0, \sigma^2)$, а $\theta \in N(0, \delta^2)$, причем с.в. ε_i, $i = 1, ..., n$ и θ независимы в совокупности. Постройте с помощью п. а) байесовскую оценку неизвестного параметра θ (функция штрафа квадратичная).

С помощью неравенства Ван Трисса исследуйте качество этой оценки. Сравните эту байесовскую оценку с байесовской оценкой в случае неинформативного априорного распределения ($\theta \in N\left(0, \delta^2\right)$, $\delta^2 \to \infty$) – это будет оценка метода наименьших квадратов, хорошо известная из курса лабораторных работ по физике.

в) (Байесовская регуляризация) Пусть нужно решить систему уравнений $\vec{X} = A\vec{\theta}$ относительно $\vec{\theta}$ (m – размер вектора $\vec{\theta}$ может быть много меньше n – размера вектора $\vec{X}$). Однако из-за ошибок округления, всевозможных шумов и неточностей измерений в действительности приходится решать систему $\vec{X} = A\vec{\theta} + \vec{\varepsilon}$, где $\vec{\varepsilon} \in N\left(\vec{0}, \sigma^2 I_n\right)$. Считая, что $\vec{\theta} \in N\left(0, \delta^2 I_m\right)$, постройте байесовскую оценку (с квадратичным штрафом) неизвестного вектора $\vec{\theta}$. Рассмотрите два случая: $\delta^2 < \infty$ – информация о локализации искомого вектора есть; $\delta^2 = \infty$ – информации нет.

Замечание. В лабораторных работах по физике: $y_i = kx_i + b + \varepsilon_i$, т.е.

$$\vec{X} = \vec{y}, \vec{\theta} = (k, b)^T, A = \begin{pmatrix} x_1 & \dots & x_n \\ 1 & \dots & 1 \end{pmatrix}^T, \delta^2 = \infty.$$

Отметим, что задача поиска байесовской оценки может быть проинтерпретирована, как поиск минимума регуляризованного (по Тихонову) функционала метода наименьших квадратов

$$\bar{\theta}\left(\vec{x}\right) = \arg\min_{\vec{\theta}} \left\{ \left\| \vec{x} - A\vec{\theta} \right\|_2^2 + \left(\sigma/\delta\right)^2 \left\| \vec{\theta} \right\|_2^2 \right\}$$

Регуляризация крайне важна, например, в случае, когда положительно определенная матрица $A^T A$ плохо обусловлена (в частности, это означает, что матрица $\left(A^T A\right)^{-1}$ может содержать очень большие элементы). Отметим, также, что эта оценка может быть получена при помощи метода наибольшего правдоподобия Фишера:

$$\vec{\theta}\left(\vec{x}\right) = \arg\max_{\vec{\theta}} \ln\left(p_{\vec{\mathrm{x}}}\left(\vec{x} \mid \vec{\theta}\right)\pi(\vec{\theta})\right) = \arg\max_{\vec{\theta}} p_{\vec{\mathrm{x}}}\left(\vec{x} \mid \vec{\theta}\right)\pi(\vec{\theta}).$$

20 (Робастное оценивание или l_1–оптимизация)**.** Рассмотрим следующую схему эксперимента $\mathrm{x}_i = \theta + \varepsilon_i$, $i = 1, ..., n$, где θ – неизвестный

параметр, ε_i – независимые одинаково распределенные с.в. с нулевым математическим ожиданием (если математичекесое ожидние не существует, то считаем, что с.в. имеют симметричное распределение относительно 0).

а) (Робастное оценивание) Положим
$\breve{\theta}(\vec{x}) = \arg\min_{\theta} \left\| \vec{x} - (\theta, ..., \theta)^T \right\|_1 = \arg\min_{\theta} \sum_{k=1}^{n} |x_k - \theta| \approx x_{(n/2)}$ – медиана. Покажите, что

$$\sqrt{n} \cdot \left(\breve{\theta}(\vec{\mathrm{x}}) - \theta \right) \xrightarrow[n\to\infty]{d} N\left(0, \left(4\left(p_\varepsilon(0)\right)^2\right)^{-1}\right).$$

Замечание. Отметим, что в этом пункте с.в. ε_i могут иметь, например, распределение Коши, для которого не существует математического ожидания. Тем не менее, среднеквадратичное отклонение $\breve{\theta}(\vec{\mathrm{x}})$ от математического ожидания θ имеет порядок $\sim n^{-1/2}$. Используя, например, неравенство Чебышёва отсюда можно заключить, что истинное значение θ лежит в интервале порядка $\sim n^{-1/2}$ с центром в $\breve{\theta}(\vec{\mathrm{x}})$. Таким образом, качество оценки неизвестного параметра вполне естественно характеризовать его дисперсией. Чем дисперсия меньше, тем оценка лучше. Это обстоятельство отчасти объясняет выбор квадратичной функции штрафа (например, в байесовском оценивании). Из неравенства Рао—Крамера следует, что для регулярных случаев такой порядок убывания дисперсии $\sim n^{-1}$ с ростом объема выборки n является типичным, и "борьба идет", как правило, за константу при n^{-1}.

б) Будем считать, что с.в. $\varepsilon_i \in N\left(0, \sigma^2\right)$. Покажите, что оценка метода наименьших квадратов:

$$\bar{\theta}(\vec{x}) = \arg\min_{\theta} \left\| \vec{x} - (\theta, ..., \theta)^T \right\|_2^2 = \arg\min_{\theta} \sum_{k=1}^{n} (x_k - \theta)^2 = \frac{1}{n} \sum_{k=1}^{n} x_k$$

доставляет равенство в неравенстве Рао–Крамера.

в) (Нерегулярное семейство) Нерегулярность семейства означает, что носитель распределения вектора $\vec{x}$ зависит от параметра, это дает возможность существования в нерегулряной модели лучших, но не робастных оценок. Пусть $\varepsilon_i \in R\left[-\sqrt{3}\sigma, \sqrt{3}\sigma\right]$. Положим,

$\tilde{\theta}(\vec{x}) = \frac{1}{2}\left(x_{(1)} + x_{(n)}\right)$, где $x_{(1)} = \min\limits_{k=1,\ldots,n} x_{(k)}$, $x_{(n)} = \max\limits_{k=1,\ldots,n} x_{(k)}$. Покажите, что

$$n\left(\tilde{\theta}(\vec{\mathrm{x}}) - \theta\right) \xrightarrow[n\to\infty]{P} \sigma(e_1 - e_2),$$

где e_1, e_2 – независимые с.в., имеющие распределение Лапласа (показательное) с параметром равным 1. Покажите, что если мы ошиблись в предположении, что $\varepsilon_i \in R\left[-\sqrt{3}\sigma, \sqrt{3}\sigma\right]$, и на самом деле $|\varepsilon_i|$ имеет, скажем, распределение Лапласа (показательное) с параметром равным 1, то

$$\tilde{\theta}(\vec{\mathrm{x}}) - \theta \xrightarrow[n\to\infty]{d} \ln e_1 - \ln e_2 + \mathrm{O}\left(\frac{1}{\sqrt{n}}\right).$$

Замечание. Все приведенные выше оценки могут быть получены методом наибольшего правдоподобия, когда:

1) $|\varepsilon_i|$ имеет распределения Лапласа;
2) $\varepsilon_i \in N\left(0, \sigma^2\right)$;
3) $\varepsilon_i \in R\left[-\sqrt{3}\sigma, \sqrt{3}\sigma\right]$.

Метод наибольшего правдоподобия также можно понимать, как способ поиска такого значения параметра(-ов) распределения, при котором это распределение наиболее близко (в смысле расстояния Кульбака–Лейблера) к эмпирическому распределению, построенному по имеющейся выборке (данным) $\vec{\mathrm{x}}$ (реализации x_i с.в. x_i считаются известными экспериментатору).

21 (Суперэффективные оценки)**.** Оценка T_n некоторого параметра θ называется *суперэффективной* если

$$\lim_{n\to\infty} (\sqrt{n}(T_n - \theta))^2 \le I^{-1}(\theta)$$

и хотя бы в одной точке θ имеет место строгое неравенство.

Для простой выборки $X_1 \ldots X_n \in N(\theta, 1)$ определим две оценки: $\widehat{\theta}_n = \overline{X}$ (оценка максимума правдоподобия) и

$$T_n = \begin{cases} \widehat{\theta}_n, & |\widehat{\theta}_n| > n^{-1/4} \\ \alpha\widehat{\theta}_n, & |\widehat{\theta}_n| \le n^{-1/4}. \end{cases}$$

где $|\alpha| < 1$. Является ли T_n суперэффективной с точкой суперэффективности $\theta = 0$? Для исследования качества оценки T_n в окрестности точки $\theta = 0$ рассмотрим последовательность $\{\theta_n = c/\sqrt{n}\}$, сходящуюся к 0. Покажите, что

$$\lim_{n\to\infty} (\sqrt{n}(T_n - \theta_n))^2 > 1,$$

в то время как

$$\lim_{n\to\infty} (\sqrt{n}(\widehat{\theta}_n - \theta_n))^2 \leq 1.$$

Замечание. Для проверки эффективности оценки в некотором множестве W можно воспользоваться следующим правилом: для набора экспериментов $\langle \Omega_\varepsilon, \mathcal{F}_\varepsilon, \mathbb{P}_\varepsilon \rangle$ оценка $\widehat{\theta}_\varepsilon$ будет *асимптотически эффективной* по отношению к функциям потерь λ_ε, если равномерно по $u \in W$ существует предел

$$\lim_{\varepsilon\to 0} \mathbb{E}\lambda_\varepsilon(\widehat{\theta}_\varepsilon - u) = L(u),$$

а также для любой оценки T_ε и непустого открытого множества $U \subseteq W$ выполнено

$$\varliminf_{\varepsilon\to 0} \sup_{u\in U} \mathbb{E}\lambda_\varepsilon(T_\varepsilon - u) \geq \sup_{u\in U} L(u).$$

В частном случае, когда эксперимент *регулярный* (достаточным условием регулярности есть существование $I(\theta)$) для любой T_ε и не слишком быстро возрастающей функции λ справедливо неравенство

$$\lim_{\delta\to 0} \varliminf_{\varepsilon\to 0} \sup_{|u-\theta|<\delta} \mathbb{E}\lambda(c_{\theta,\varepsilon}|T_\varepsilon - u|) \geq \frac{1}{\sqrt{2\pi}} \int \lambda(|y|) e^{-y^2/2} dy.$$

С другой стороны, оценки максимального правдоподобия и байесовские оценки достигают равенства в последнем выражении.

22. Пусть $\widehat{t}_{\varepsilon,q}$ – байесовская оценка параметра $\theta \in \mathbb{R}^k$ относительно априорной плотности q и функции потерь λ_ε для набора экспериментов $\langle \Omega_\varepsilon, \mathcal{F}_\varepsilon, \mathbb{P}_\varepsilon \rangle$. Допустим, что $\forall u \in W, \forall q : q(u) > 0$, выполнено соотношение

$$\lim_{\varepsilon\to 0} \sup_{u\in U} \mathbb{E}\lambda_\varepsilon(\widehat{t}_{\varepsilon,q} - u) = L(u).$$

Докажите $\forall U \subseteq W, \forall T_\varepsilon$ справедливость неравенства

$$\varliminf_{\varepsilon\to 0} \sup_{u\in U} \mathbb{E}\lambda_\varepsilon(T_\varepsilon - u) \geq \int\limits_U L(u)q(u)du.$$

23 (Оракульное неравенство)**.** а) Пусть $\mathrm{x}_k,\ k = 1, ..., n$ – независимые одинаково распределенные с.в. $\mathrm{x}_k \in N\left(0, \sigma^2\right)$. Покажите, что $E\left[\max\limits_{k=1,...,n} \mathrm{x}_k\right] \leq \sqrt{2\sigma^2 \ln n}$.

б) Покажите, что

$$P\left[\max_{k=1,...,n} |\mathrm{x}_k| \geq \sigma(\sqrt{2\ln n} + u)\right] \leq \frac{1}{\sqrt{\pi \ln n}} e^{-u^2/2}$$

в) (Sparsity, экспрессия генов) $y_k = \theta_k + \varepsilon_k,\ \varepsilon_k \in N\left(0, \sigma^2\right)$, $k = 1, ..., n$ – независимые с.в. Результаты измерений y_k – известны, параметры θ_k – неизвестны. Однако известно, что большинство компонент (правда, не известно какие именно) вектора $\vec{\theta}$ – нулевые. Предложите выбор такого порога $\tau > 0$, чтобы оценка неизвестных параметров $\tilde{\theta}_k = y_k I\left(|y_k| > \tau\right)$, где $I(\cdot)$ - индикаторная функция была бы "наиболее разумной".

Указание. а) 1. $\max\limits_{k=1,...,n} \mathrm{x}_k = \lambda^{-1} \ln\left[\max\limits_{k=1,...,n} e^{\mathrm{x}_k}\right] \leq \lambda^{-1} \ln\left[\sum\limits_{k=1}^{n} e^{\mathrm{x}_k}\right]$. 2.(Неравенство Йенсена) Для вогнутой функции $Ef\left(\xi\right) \leq f\left(E\xi\right)$. 3. Воспользовавшись 1 и 2, оптимально подберите λ. В действительности, для доказательства этой оценки не требуется независимость, а нормальность можно заменить на субгауссовость. б) Воспользуйтесь *неравенством Буля*: $P\left(\bigcup U_k\right) \leq \sum P\left(U_k\right)$.

Замечание. (Оракульное неравенство) Можно показать, что существует такая константа $C > 0$, что при $\tau = \sigma\sqrt{2\ln n}$ и числе ненулевых компонент вектора $\vec{\theta}$ равным $m \ll n$ имеет место следующее (с точностью до константы не улучшаемое) неравенство:

$$E\left\|\tilde{\vec{\theta}} - \vec{\theta}\right\|_2^2 \leq C\sigma^2 \frac{m \ln n}{n}.$$

Кроме того, если $\min\limits_{k=1,...,n\ \theta_k \neq 0} \theta_k > 2\tau$, то с вероятностью не меньшей $1 - 1\Big/\sqrt{\pi \ln n}$ выполняется $\tilde{\theta}_k > 0 \Leftrightarrow \theta_k > 0$.

Можно немного "поднять" порог τ, тогда существенно улучшится скорость сходимости.

24 (Теорема Бернштейна–фон-Мизеса)**.** Пусть $\mathbb{Y} = (Y_1, ..., Y_n)$ – независимые в совокупности, одинаково распределенные случаные

величины, подчиняющиеся закону $\mathrm{Be}(p)$, причем параметр p так же является случайно величиной: $p \in \mathrm{Beta}(1,1)$. Докажите, что апостериорное распределение параметра p ассимптотически нормальное:

$$p|\mathbb{Y} \to N\left(\overline{p}, \frac{1}{n} I_{p^*}^{-1}\right),$$

где $p^* \in (0,1)$ – истиное значение параметра, I_{p^*} – информационная матрица Фишера, $\overline{p}$ – средневыборочная оценка.

Рассмотрим следующую последовательность экспериментов:

$$\begin{gathered} Y_1^1 \\ Y_2^1, Y_2^2 \\ \dots \\ Y_n^1, \dots, Y_n^n, \end{gathered}$$

где $Y_k^j \in \mathrm{Po}(\frac{p}{k})$ независимые в совокупности одинаково распределенные случайные величины. Докажите, что:

$$p|\mathbb{Y} \not\to N\left(\overline{p}, \frac{1}{n} I_{\frac{p^*}{n}}^{-1}\right),$$

$$[p]|\mathbb{Y} \to \mathrm{Po}(p^*).$$

Замечание. Приведем более общую формулировку теоремы БфМ. Ведем вспомогательные обозначения:

$$p^* = \arg\max_{p \in \Theta} \mathbb{E}L(p), \quad \tilde{p} = \arg\max_{p \in \Theta} L(p), \quad \mathbb{P}(p|\mathbb{Y}) \propto e^{L(p)}\pi(p).$$

$$D_0^2 = -\bigtriangledown^2 \mathbb{E}L(p^*), \quad V_0^2 = \mathbb{D}[\bigtriangledown L(p^*)].$$

Теорема 1. Пусть $\mathbb{P}_p$ – некоторое семейство распределений с фиксированным носителем. Ковариационная функция $k_p(y, y')$ трижды непрерывно дифференцируема по p, а соответствующие ковариационные матрицы K, $K_p = \{k_p(y_i, y_j)\}$ удовлетворяют следующим условиям: собственные числа $0 < \lambda_{min} < \lambda < \lambda_{max} < \infty$; $\left\| \frac{\partial K_p}{\partial p_i} \right\|_2 < \lambda_1 < \infty$, $\left\| \frac{\partial^2 K_p}{\partial p_i \partial p_j} \right\|_2 < \lambda_2 < \infty$, $\left\| \frac{\partial^3 K_p}{\partial p_i \partial p_j \partial p_k} \right\|_2 < \lambda_3 < \infty$. Минимальное собственное число матрицы $\frac{1}{n} D_0^2 > d_0 > 0$. Вектор p^* существует. $\exists r : \forall p \in \{\|V_0(p - p^*)\|_2 > r\}$ выполнено $\mathbb{E}L(p) - \mathbb{E}L(p^*) \neq 0$. Тогда $\exists \tau, x$, такие что с вероятностью $1 - ce^{-x}$ выполнено:

$$\left\| D_0(\overline{p} - \tilde{p}) \right\|_2 \leq c\tau(\dim p + x),$$

$$\left\|E_{\dim p} - D_0\sigma^2 D_0\right\|_\infty \leq c\tau(\dim p + x),$$

где значение $\tau(\dim p + x)$ мало и уменьшается с ростом выборки, $E_{\dim p}$ – единичная матрица, $c \gg \tau$ – некоторая константа,

$$\overline{p} = \mathbb{E}(p|\mathbb{Y}), \quad \sigma^2 = \mathbb{E}((p-\overline{p})(p-\overline{p})^T|\mathbb{Y}).$$

Кроме того, $\forall\lambda : \left\|\lambda\right\|_2 \leq (\dim p + x)$ выполнено:

$$\left|\ln \mathbb{E}\left[exp\{\lambda^T\sigma^{-1}(p-\overline{p})\}|\mathbb{Y}\right] - \frac{\left\|\lambda\right\|_2}{2}\right| \leq c\tau(\dim p + x),$$

что описывает близость апостериорного распределения p к нормальному распределению.

Стоит отметить, что истинное распределение может не лежать в семействе: $\mathbb{P} \notin (\mathbb{P}_p)$ – модель данных является ошибочной. В таком случае $\mathbb{P}_{p^*}$ – ближайшее распределение по метрике $\mathcal{KL}$.

25 (Парадокс Байеса)**.** Пусть $X_1, X_2, \ldots$ – независимые с.в. из распределения с неизвестным параметром $p \in [a, b]$. Можно ли ожидать, что последовательность апостериорных распределений (при равномерном априорном распределении) все более и более концентрируются около истинного значения p? Оказывается, что это не всегда верно. Приведите пример такого распределения.

Указание. В качестве примера можно рассмотреть следующую с.в. X:

$$\mathbb{P}(X = k) = c(1-p)p^k, \quad k = \overline{0, f(p)},$$

где $f(p) \in \mathbb{N}$, $f(1/4) = f(3/4) = \infty$, $p \in [1/8, 7/8]$, $1/4$ – истинное значение p, $3/4$ – точка концентрации апостериорных вероятностей.

26 (Парадокс Стейна)**.** Пусть $\mathbb{Y} = (Y_1, ..., Y_n)$, где $Y_i \in N(\theta, \sigma^2 E_k)$, $\theta \geq 0$, E_k – единичная матрица. $\hat{\theta}$ – произвольная оценка вектора параметров θ. Для квадратичной функции потерь $\|\theta - \hat{\theta}\|^2 = \sum\limits_{i=1}^{k}(\theta_i - \hat{\theta}_i)^2$ обозначим через $r_{\hat{\theta}}(\theta)$ функцию риска оценки $\hat{\theta}$, то есть $r_{\hat{\theta}}(\theta) = \mathbb{E}_\theta\|\theta - \hat{\theta}\|^2$. Покажите, что $\hat{\theta} = \overline{Y}$ является допустимой оценкой (см. замечание) только при $k \leq 2$.

Указание. Сравните значения функции потерь для оценок $\overline{Y}$ и

$$\max\left(0, 1 - \frac{(k-2)^2}{\|\overline{Y}\|^2}\right)\overline{Y},$$

$$\left(1 - \frac{(k-2)^2}{\|\overline{Y}\|^2}\right)\overline{Y}.$$

Пусть $Y_i \in 0.91N(\theta, 1) + 0.09N(\theta, 9)$. Покажите, что в этом случае медиана выборки оценивает параметр лучше, чем $\overline{Y}$.

Замечание. Оценку параметра $\hat{\theta}_1$ называют допустимой (при заданной функции потерь), если не существует такой оценки $\hat{\theta}_2$, что при любом значении параметра θ $r_{\hat{\theta}_1}(\theta) \leq r_{\hat{\theta}_2}(\theta)$ (и хотя бы при одном значении параметра неравенство строгое).

27. Найдите оценку максимального правдоподобия параметра θ для выборки $\mathbb{Y} = (Y_1, ..., Y_n)$, где $Y_i \in N(\theta, c\theta^2)$, где $c > 0$.

Замечание. Данный пример демонстрирует не единственность локального максимума функции правдоподобия.

28. Выборка $X_1, \ldots, X_n$ состоит из независимых с.в. из распределения $[\theta, 2\theta]$. Покажите, что оценкой максимального правдоподобия θ является $\max(X_i)/2$.

$$\widehat{\theta} = \frac{2n+2}{2n+1}\max(X_i)/2$$

– есть несмещенная оценка θ с дисперсией $1/(4n^2)$. Найдите дисперсию более эффективной оценки оценки:

$$\frac{n+1}{5n+4}(\min(X_i) + 2\max(X_i)).$$

Замечание. Пара $\min(X_i)$, $\max(X_i)$ в совокупности образуют *достаточную статистику* в отличие от $\max(X_i)$: т.е. при заданных значениях $\min(X_i)$ и $\max(X_i)$ совместное распределение $X_1, \ldots, X_n$ не зависит от θ.

29. Проведем сравнение оценок параметра θ согласно критерию:

$$\mathbb{P}(|\widehat{\theta}_1 - \theta| < |\widehat{\theta}_2 - \theta|) \overset{?}{>} \frac{1}{2}.$$

а) Покажите, что X_1 является более эффективной оценкой, нежели $(X_1 + X_2)/2$, где X_1, X_2 имеют симметричную плотность распределения $f(x-\theta)$,

$$f(x) = \frac{3}{2} \cdot \frac{1}{(1+|x|)^4}.$$

б) Для простой выборки $X_1, \ldots, X_n$ из $N(\theta, 1)$ покажите, что оценка математического ожидания $\overline{X}$ менее эффективна нежели:

$$\widehat{\theta} = \begin{cases} \overline{X} - \frac{1}{2\sqrt{n}} \min\{\sqrt{n}\overline{X}, \Phi(-\sqrt{n}\overline{X})\}, & \overline{X} \geq 0; \\ \overline{X} + \frac{1}{2\sqrt{n}} \min\{\sqrt{n}\overline{X}, \Phi(-\sqrt{n}\overline{X})\}, & \overline{X} \leq 0, \end{cases}$$

где Φ обозначает функцию стандартного нормального распределения.

30 (Критическая размерность)**.** Согласно теореме Фишера, ОМП $\tilde{\theta}$ близка к значению θ^* при условии $\dim\theta = p_n = o(n^{1/3})$. Покажем на примере, что данное условие нельзя ослабить. Пусть искомое распределение $\mathbb{P} = \mathrm{Po}(\exp\theta^*)$, $\mathbb{Y} = (Y_1, ..., Y_n)$, где $Y_i \in \mathbb{P}$. Определим параметрическое семейство $\mathbb{P}_\theta$, в котором будем искать элемент наиболее близкий к $\mathbb{P}$ по метрике $\mathcal{KL}$: $Y_i \in \mathrm{Po}(v_j)$ при $j \in \mathcal{I}_j = \{i : \lceil ip_n/n \rceil = j\}$ (будем считать, что $n/p_n \in \mathbb{N}$). Оцениваемый параметр определим как

$$\theta = \frac{1}{p_n} \sum_{k=1}^{p_n} \ln(v_j).$$

Покажите, что ОМП:

$$\tilde{\theta} = \frac{1}{p_n} \sum_{k=1}^{p_n} \ln\left(\frac{S_k}{n/p_n}\right), \quad S_k = \sum_{i \in \mathcal{I}_k} Y_i.$$

Используя свойства экспоненциального семейства распределений (см. замечание к задаче 27 из раздела 3), проверьте следующее неравенство:

$$\mathbb{P}(\tilde{\vec{v}} \in \Theta_0(r)) \geq 1 - 4e^{-x},$$

где $\Theta_0(r) = \{\vec{v} : \mathcal{KL}(v_j, v_j^*)n/p_n \leq r, j \in \overline{1, p_n}\}$, $r = x + \ln(p_n)$.

Данное неравенство позволяет ограничится областью $\Theta_0(r)$ при исследовании разности значений $\sup L(\theta)$ и $L(\theta^*)$.

Покажите, что:

а) при $\beta_n \to 0$, $p_n \to \infty$: $D_0(\tilde{\theta} - \theta^*) \to N(0,1)$;

б) при $\beta_n = \beta > 0$: $D_0(\tilde{\theta} - \theta^*) \to N(\beta/2, 1)$;

в) при $\beta_n \to \infty$, $\beta_n^2/\sqrt{p_n} \to 0$: $D_0(\tilde{\theta} - \theta^*) \to -\infty$,

где $\beta_n = \sqrt{p_n^3/n}$, $D_0^2 = -\bigtriangledown^2 \mathbb{E}L(\theta^*) = p_n^2\beta_n^{-2}$.

Указание. В пунктах а) б) в) воспользуйтесь разложением по формуле Тейлора:

$$p_n\beta_n^{-1}(\tilde{\theta} - \theta^*) = \beta_n^{-1}\sum_{j=1}^{p_n} \ln\left(\frac{S_j}{v^*n/p_n}\right) = \beta_n^{-1}\sum_{j=1}^{p_n} \ln\left(1 + \frac{\beta_n}{\sqrt{p_n}}\gamma_j\right) =$$

$$\frac{1}{\sqrt{p_n}}\sum_{j=1}^{p_n}\gamma_j - \frac{\beta_n}{2p_n}\sum_{j=1}^{p_n}\gamma_j^2 + o(1),$$

применимость которой следует из неравенства, доказываемого в задаче:

$$\ln\left(\frac{S_j}{v^*n/p_n}\right) \le \sqrt{r/p_n}.$$

Замечание. Классическая постановка задачи *обучения распознаванию образов* с двумя классами объектов. Изучается некоторое множество объектов $\omega \in \Omega$, каждый из которых обладает n измеряемыми свойствами, выраженными действительными числами $x_i(\omega) \in \mathbb{R}$, $i = 1, ..., n$. Совокупность результатов этих измерений будем называть вектором действительных признаков объекта $x(\omega) = (x_1(\omega) \cdots x_n(\omega))^T \in \mathbb{R}^n$.

Допустим, что все множество объектов $\omega \in \Omega$ разбито на два класса индикаторной функцией $y(\omega) : \Omega \to \{-1,\ 1\}$, вообще говоря, неизвестной наблюдателю. Целью наблюдателя является определение класса предъявленного объекта $y(\omega) \in \{-1,\ 1\}$, зная лишь доступный для непосредственного наблюдения вектор признаков $x(\omega) \in \mathbb{R}^n$. Иными словами, желание наблюдателя сводится к построению дискриминантной функции $\hat{y}(x) : \mathbb{R}^n \to \{-1,\ 1\}$. В качестве исходной информации для выбора дискриминантной функции будем рассматривать обучающую совокупность объектов, представленных и векторам их признаков $x_j = x(\omega_j) \in \mathbb{R}^n$, и фактическими значениями индикаторной функции класса $y_j = y(\omega_j) \in \{-1,\ 1\}$. Таким образом, обучающая совокупность представлена конечным множеством пар $(X, Y) = \{(x_j, y_j),\ j = 1, ..., N\}$.

31 (Байесовская интерпретация Метода опорных векторов SVM)**.** Рассмотрим следующую модель наблюдения. Пусть в $\mathbb{R}^n$ определена некоторая гиперплоскость $a^T x + b$ с направляющим вектором $a \in \mathbb{R}^n$ и параметром сдвига $b \in \mathbb{R}$, а также пара несобственных плотностей распределения вероятностей $\phi(x|y;\ a, b, c)$, $x, a \in \mathbb{R}^n$, $b, c \in \mathbb{R}$, $y = \pm 1$ (см. замечание к задаче 17), сконцентрированных преимущественно по разные стороны от этой гиперплоскости (параметр c считается известным):

$$\phi(x|y;\ a, b, c) = \exp\left[-c\left(1 - y(a^T x + b)\right)_+\right],$$

где $(x)_+ = \max(0, x)$.

Направляющий вектор $a \in \mathbb{R}^n$ будем рассматривать как случайный, распределенный с некоторой известной плотностью $\Psi(a)$. Никаких априорных предположений о значении случайного сдвига гиперплоскости $b \in \mathbb{R}$ приниматься будет, так что совместное распределение $\Psi(a, b)$ будет рассматриваться как несобственное:

$$\Psi(a, b) \propto \Psi(a).$$

Далее, пусть обучающая совокупность

$$(X, Y) = \{(x_j, y_j),\ j = 1, ..., N\}$$

есть результат многократных случайных независимых реализаций распределений $\phi(x|y = 1;\ a, b, c)$ и $\phi(x|y = -1;\ a, b, c)$, всякий раз с известным индексом $y = \pm 1$ принадлежности очередного объекта к одному их классов.

а) Запишите апостериорное распределение параметров разделяющей гиперплоскости после наблюдения обучающей совокупности (согласно формуле Байеса). Покажите, что с точностью до множителя, не зависящего от параметров гиперплоскости

$$\mathbb{P}(a, b \,|\, X, Y) \propto \Psi(a)\Phi(X \,|\, Y; a, b),$$

$$\Phi(X|Y;\ a, b, c) = \prod_{j=1}^{N} \phi(x_j|y_j;\ a, b, c).$$

Обучение естественно понимать как вычисление байесовской оценки параметров разделяющей гиперплоскости:

$$(\hat{a}, \hat{b} \,|\, X, Y; c) = \underset{a \in \mathbb{R}^n,\ b \in \mathbb{R}}{\arg\max}\ \mathbb{P}(a, b \,|\, X, Y).$$

Покажите, что критерий обучения можно записать через следующую задачу оптимизации:

$$-\ln \Psi(a) + c \sum_j \left(1 - y_i(a^T x_i + b)\right)_+ \to \min_{a,b},$$

или как задачу квадратичного программирования

$$\begin{cases} -\ln \Psi(a) + c \sum_{j=1}^N \delta_j \to \min(a, b, \delta_1, , ..., \delta_N), \\ y_j(a^T x_j + b) \geq 1 - \delta_j,\ \delta_j \geq 0,\ j = 1, ..., N. \end{cases}$$

б) Приняв дополнительное предположение об априорных вероятностях принадлежности случайно появляющегося объекта одному либо другому классу:

$$q(1) = P\left(y(\omega) = 1\right), q(-1) = P\left(y(\omega) = -1\right), q(1) + q(-1) = 1,$$

запишите апостериорную вероятность принадлежности объекта $\omega \in \Omega$ с вектором признаков $x \in \mathbb{R}$ классу $y = 1$ и $y = -1$: $p(y = 1 \,|\, x;\ a, b, c)$ и $p(y = -1 \,|\, x;\ a, b, c)$.

в) Приняв дополнительное предположение, что априорной плотности распределения направляющего вектора разделяющей гиперплоскости является нормальным с нулевым математическим ожиданием и независимыми компонентами $a = (a_1, ..., a_n)$, характеризующимися одинаковыми дисперсиями $r_1 = ... = r_n = r$:

$$\Psi(a) = \prod_{i=1}^n \frac{1}{r^{1/2}(2\pi)^{1/2}} \exp\left(-\frac{1}{2r} a_i^2\right),$$

запишите критерий обучения.

г) Классическая детерминированная постановка задачи SVM имеет наглядное объяснение для выбора параметров гиперплоскости. Пусть поступила обучающая совокупность $\{(x_j, y_j),\ j = 1, ..., N\}$. Представляется естественной эвристическая идея выбрать такую разделяющую гиперплоскость (a, b), которая правильно разделяет объекты двух классов: $y_j(a^T x_j + b) > 0$ для всех $j = 1, ..., N$.

Допустим, что для предъявленной обучающей совокупности разделяющая гиперплоскость существует. Но в этом случае существует

континуум разделяющих гиперплоскостей. Идея В.Н. Вапника заключается в выборе той из них, которая обеспечивает наибольший «зазор» между гиперплоскостью и ближайшими точками обучающей совокупности как одного, таки другого класса $y_j(a^T x_j + b) \geq \varepsilon > 0$. Правда, величина зазора ε условна, и определяется еще и нормой направляющего вектора, поэтому задача формулируется в виде задачи условной оптимизации:

$$y_j(a^T x_j + b) \geq \varepsilon \to \max_{a,b},\ j = 1, \ldots, N,\ a^T a = 1.$$

Такая концепция обучения названа концепцией оптимальной разделяющей гиперплоскости.

Или эквивалентная формулировка задачи поиска оптимальной разделяющей гиперплоскости:

$$\begin{cases} a^T a \to \min, \\ y_j(a^T x_j + b) \geq 1,\ j = 1, ..., N. \end{cases}$$

Однако предъявленная обучающая совокупность может оказаться линейно неразделимой, и поставленная оптимизационная задача не будет иметь решения. В качестве еще одной эвристики В.Н. Вапник предложил в качестве компромисса «разрешить» некоторым точкам обучающей совокупности располагаться с «неправильной» стороны разделяющей гиперплоскости, потребовав, чтобы такой дефект был минимальным:

$$\begin{cases} J(a, b, \delta_1, ..., \delta_N) = a^T a + C \sum_{j=1}^{N} \delta_j \to \min, \\ y_j(a^T x_j + b) \geq 1 - \delta_j,\ \delta_j \geq 0,\ j = 1, ..., N, \end{cases}$$

где $C > 0$ – некоторый коэффициент, согласующий два, вообще говоря, взаимно противоречивых требования – обеспечить как можно меньшее значение нормы направляющего вектора и как можно меньшую ошибку классификации в пределах обучающей совокупности. Сравните классическую постановку задачи SVM с вероятностной (из п. в), чему соответствует в вероятностной интерпретации задачи параметр C.

Записав двойственную задачу оптимизации и найдя множители Лагранжа, выпишите значение направляющего вектора оптимальной гиперплоскости. Обратите внимание, что направляющий вектор оптимальной гиперплоскости выражается как линейная комбинация

векторов признаков только части объектов обучения, для которых множитель Лагранжа ненулевой, т.е. лежащих за границей поверхности разделяющей полосы, образованной разделяющей гиперплоскостью и обладающей шириной обратно пропорциональной $\|a\|$ (это подмножество объектов обучения и называется «опорным» откуда и происходит название метода обучения – «метод опорных векторов» или «support vector machine»).

32. Выбирая алгоритм классификации $a : \mathbb{X} \to \{0, 1\}, a \in A$ (A - некоторое семейство алгоритмов) при помощи обучающей (заранее известной) выборки X ($|X| = l$), исследователей инетересует оценка частоты ошибок a на будущих данных $\overline{X}$: $\nu(a, \overline{X})$, причем $\mathbb{X} = X \sqcup \overline{X}$ Для того, чтобы было возможно получить эту оценку, необходимо сделать предположение о том, что элементы генеральной выборки $\mathbb{X}$ ($X \in \mathbb{X}$, $\overline{X} \in \mathbb{X}$) появляются в случайном порядке. Причем все $L!$ перестановок ($|\mathbb{X}| = L$) равновозможны. Иногда условие ослабляют и равновозможными считаются все C_L^l разбиений. Пусть алгоритм a допускает на генеральной совокупности m ошибок: $n(a, \mathbb{X}) = m$, докажите, что в этом случае число ошибок в наблюдаемой подвыборке $n(a, X)$ подчиняется гипергеометрическому распределению:

$$P(n(a, X) = s) = \frac{C_m^s C_{L-m}^{l-s}}{C_L^l},$$

где $s \in \{s_0 = \max(0, m - k), ..., \min(l, m)\}$.

33. В условиях предыдущей задачи определим *функционалы обобщающей способности* алгоритма:

а) вероятность переобучения $Q_\epsilon(a, \mathbb{X}) = P(\nu(a, \overline{X}) - \nu(a, X) \geq \epsilon)$;

б) вероятность высокой частоты ошибок на контрольной выборке $R_\epsilon(a, \mathbb{X}) = P(\nu(a.\overline{X}) \geq \epsilon)$;

в) средняя частота ошибок на скользящем контроле $C(a, \mathbb{X}) = \mathbb{E}\nu(a, \overline{X})$.

Докажите, что если $n(a, \mathbb{X}) = m$, то $\forall \epsilon \in [0, 1]$: $C(a, \mathbb{X}) = \frac{m}{L}$, $Q_\epsilon(a, \mathbb{X}) = H_L^{l,m}$, $R_\epsilon(a, \mathbb{X}) = H_L^{l,m}$, где $H_L^{l,m}(z) = \sum_{s=s_0}^{\lfloor z \rfloor} h_L^{l,m}(s)$, а $h_L^{l,m}(s) = \frac{C_m^s C_{L-m}^{l-s}}{C_L^l}$.

34. Пусть μ - метод выбора алгоритма по обучающей выборке X (метод обучения), $I : \mathbb{X} \to \{0, 1\}$ - индикатор ошибки для заданного алгоритма, A_ϵ - множество векторов ошибок, порождаемых множеством алгоритмов A. Доказать, что $\forall \epsilon \in [0, 1]$:

$$Q(\mu, \mathbb{X}) \leq |A_\epsilon| \max_{m=1,\ldots,L} H_L^{l,m}(s_m(\epsilon)), \text{ где } s_m(\epsilon) = \frac{l}{L}(m - k\epsilon).$$

Указание. Для получения верхних оценок вероятности переобучения, не зависящих от метода, часто используется *принцип равномерной сходимости*

$$Q(\mu, \mathbb{X}) \leq P(\max_{a \in A}(\nu(a, \overline{X}) - \nu(a, X)) \geq \epsilon).$$

35 (Оценка вероятности переобучения)**.** В статистической теории переобучения центральным объектом анализа является задача минимизации математического ожидания функции штрафа:

$$M(\alpha) = \mathbb{E}\lambda(x, \alpha) = \int \lambda(x, \alpha) dF(x) \to \min, \tag{1}$$

где $\alpha \in \Omega$ – набор параметров метода обучения, $F(x)$ – функция распределения выборки, $0 \leq \lambda(x, \alpha) \leq \Lambda$ – некоторая функция, измеримая $\forall \alpha \in \Omega$ относительно меры $F(x)$.

Ввиду того, что в большинстве практических задач $F(x)$ неизвестна, $M(\alpha)$ приближается эмпирическим риском:

$$M_l(\alpha) = \frac{1}{l} \sum_{i=1}^{l} \lambda(X_i, \alpha) \to \min, \tag{2}$$

где $\{X_i\}_{i=1}^l$ – выборка из распределения $F(x)$.

В задаче классификации обычно в качестве $M(\alpha)$ берется вероятность неправильной классификации с помощью алгоритма $a(f, \alpha) : \mathbb{F} \to \mathbb{Y}$, где $x = (f, y)$, f – множество признаков, y – индекс класса, при этом

$$\lambda(x, \alpha) = I[a(f, \alpha) \neq y],$$

$$M(\alpha) = \mathbb{P}(A_\alpha) = \mathbb{P}(a(f, \alpha) \neq y).$$

В свою очередь $M_l(\alpha)$ равна частоте события A_α при заданной выборке.

В качестве меры близости между оптимальными значениями параметров α_{min}, α^* в задачах (1) и (2) естественно взять

$$M(\alpha^*) - M(\alpha_{min}) \leq 2 \sup_{\alpha} |M(\alpha) - M_l(\alpha)|.$$

Таким образом, возникает вопрос: имеет ли место равномерная сходимость $M_l(\alpha)$ к $M(\alpha)$ по заданной системе событий S (или же по параметру α задающему систему событий). В случае задачи классификации $S = \{A_\alpha\}$ и близость оптимальных параметров означает близость частот к вероятностям системы S.

Применив ц.п.т., покажите, что для конечной системы S $M_l(\alpha)$ равномерно сходится к $M(\alpha)$.

Основная идея, на которой строятся условия равномерной сходимости для бесконечной системы S, состоит в разбиении S на конечное число классов эквивалентности так, что в каждом классе события неотличимы относительно выборки. Для понимания применимости данной замены, проверьте, что близость $M_l(\alpha)$ к $M(\alpha)$ равносильна сходимости $M_l(\alpha)$ на обучающей и тестовой выборках, а именно:

$$\mathbb{P}\left\{ \sup_{\alpha\in\Omega} |M(\alpha) - M_l(\alpha)| > \varepsilon \right\} \leq 2\mathbb{P}\left\{ \sup_{\alpha\in\Omega} |M_l(\alpha) - M_{l,2l}(\alpha)| > \frac{\varepsilon}{2} \right\}$$

при $l > 2/\varepsilon$.

Рассмотрим систему событий S более общего вида $A(\alpha, c) = \{x : \lambda(x, \alpha) \geq c\}$ для всевозможных значений $\alpha \in \Omega$ и c. Обозначим за $\Delta^S(x_1, ..., x_l)$ – число классов эквивалентности системы S. Введем функцию роста $m^S(l) = \max \Delta^S(x_1, ..., x_l)$, где максимум берется по всем последовательностям $(x_1, ..., x_l)$ длины l. Покажите, что

$$\mathbb{P}\left\{ \sup_{\alpha\in\Omega} |M(\alpha) - M_l(\alpha)| > \varepsilon \right\} \leq 6m^S(2l) \exp\left\{ -\frac{\varepsilon^2(l-1)}{4\Lambda^2} \right\}. \qquad (3)$$

При помощи данной оценки докажите теорему Гливенко (см. задачу **??**), взяв $S = \{x : x \leq \alpha\}$, а также ее обобщение на n-мерный случай, где $S = \{x : \langle x, \alpha\rangle \geq 0\}$, $\alpha \neq 0$.

Замечание. Для любой системы событий S имеет место

$$m^S(l) = 2^l \text{ или}$$

$$m^S(l) \leq \sum_{i=0}^{n-1} C_l^i \text{ т.е. } \exists n_0 \in \mathbb{N} : m^S(l) = O(l^{n_0}).$$

Минимально возможное значение n_0 принято называть *размерностью Вапника-Червоненкиса* (VC-размерность). Однако А.Я.Червоненкис предлагает называть её *комбинаторной размерностью* S. Так, например, для множества всевозможных решающих правил в пространстве размерности n комбинаторная размерность $n_0 = n+1$. 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 объема выборки $M_l(\alpha)$ к $M(\alpha)$, является то, что Ω – компакт, $\lambda(x,\alpha)$ непрерывна по α, $|\lambda(x,\alpha)| < K(x)$, где $\int K(x)dx < \infty$.

Наряду с достаточным условием (3) критерием равномерной сходимости является условие

$$\lim_{l\to\infty} \frac{H^S(l)}{l} = 0,$$

где $H^S(l) = \mathbb{E}\log_2 \Delta^S(x_1, ..., x_l)$ – энтропия системы S относительно выборок длины l.

Отметим, что оценка (3) в большинстве случаев является чрезмерно завышенной (на несколько порядков) и поэтому не может быть использована для подсчета достаточного размера обучающей выборки на практике.

36. Дана *обучающая выборка* $X = \{(X_i, Y_i)\}_{i=1}^l$, состоящая из l независимых пар (X_i, Y_i) из распределения $\mathbb{P}$ и некоторая *функция потерь* $\lambda\colon \mathbb{Y}^2 \to [0,1]$, которая характеризует величину потерь при отнесении объекта класса $y \in \mathbb{Y}$ к классу $y' \in \mathbb{Y}$. Рассматривается задача поиска алгоритма $a(x)$, минимизирующего *средний риск*:

$$M(a) \equiv \mathbb{E}_{\mathbb{P}}\big[\lambda\big(Y, a(X)\big)\big] \to \min_{a\in A}. \tag{1}$$

Поскольку распределение $\mathbb{P}$ неизвестно, задачу (1) часто заменяют задачей минимизации *эмпирического риска*:

$$M_l(a) \equiv \frac{1}{l}\sum_{i=1}^{l} \lambda\big(Y_i, h(X_i)\big) \to \min_{a\in A}. \tag{2}$$

Таким образом встает вопрос о соотношении величин $M(a^*)$ и $M_l(a^*)$, где a^* – решение задачи (2).

Пусть $\mathbb{Y} = \{-1, +1\}$, $\lambda(y, y') = \mathrm{Int}(y \neq y')$ и $A = \{a_1, \ldots, a_N\}$. Докажите, что $\forall \delta > 0$ выполнено:

$$\mathbb{P}\left(M(a^*) \leqslant M_l(a^*) + \sqrt{\frac{\log_2 \frac{N}{\delta}}{2l}} \right) \geq 1 - \delta.$$

Указание. Получите оценку для $\sup_{a \in A}\big(M(a) - M_l(a)\big)$ с помощью неравенства Хефдинга из раздела 3 и неравенства Буля: $\mathbb{P}\{A \cup B\} \leqslant \mathbb{P}\{A\} + \mathbb{P}\{B\}$.

Замечание. Эта оценка без изменений обобщается на произвольное множество ответов $\mathbb{Y}$ и любую функцию потерь $\lambda\colon \mathbb{Y}^2 \to [0, 1]$. В том числе она может использоваться в задачах регрессии ($\mathbb{Y} = \mathbb{R}$) с квадратичной функцией потерь $\lambda(y, y') = (y - y')^2$.

37. В постановке задачи 36 предположим дополнительно, что существует алгоритм $\hat{a}\colon \mathbb{X} \to \mathbb{Y} \in A$, такой что $\mathbb{P}\{Y = \hat{a}(X)\} = 1$. Такую упрощенную постановку принято называть реализуемым случаем без шума (noise-free realizable setting).

Докажите, что $\forall \delta > 0$ выполнено:

$$\mathbb{P}\left(M(a^*) \leqslant M_l(a^*) + C \cdot \frac{\log \frac{N}{\delta}}{l} \right) \geq 1 - \delta,$$

где C — некоторая универсальная константа.

Указание. Покажите, что $\forall \delta > 0$ и $\forall a \in A$ выполнено:

$$\mathbb{P}\left(M(a) \leqslant M_l(a) + \sqrt{\frac{2M(a)\log \frac{N}{\delta}}{l}} + \frac{2\log \frac{N}{\delta}}{3l} \right) \geq 1 - \delta.$$

Для этого воспользуйтесь неравенством Бернштейна из раздела 3 вместе с неравенством Буля.

38 (Рademахеровская сложность)**.** *Рademахеровской сложностью* и *условной рademахеровской сложностью* множества A при фиксированной функции потерь λ назовем соответственно

$$\mathcal{R}(\lambda, A) = \mathbb{E}\left[\sup_{a \in A} \frac{1}{l} \sum_{i=1}^{l} \sigma_i \lambda\big(Y_i, a(X_i)\big) \right],$$

$$\mathcal{R}_l(\lambda, A) = \mathbb{E}\left[\sup_{a\in A} \frac{1}{l}\sum_{i=1}^{l} \sigma_i \lambda\big(Y_i, a(X_i)\big)\Big| X\right],$$

где $\sigma_1, \ldots, \sigma_l$ — последовательность независимых рademахеровских случайных величин, принимающих значения $+1$ и -1 с вероятностями $1/2$. Математические ожидания берутся по всем случайным величинам.

Докажите, что $\forall \delta > 0$ выполнено:

$$\mathbb{P}\left(M(a^*) \leqslant M_l(a^*) + 2\mathcal{R}(\lambda, A) + \sqrt{\frac{\log\frac{1}{\delta}}{2l}}\right) \geq 1 - \delta, \tag{3}$$

$$\mathbb{P}\left(M(a^*) \leqslant M_l(a^*) + 2\mathcal{R}_n(\lambda, A) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2n}}\right) \geq 1 - \delta. \tag{4}$$

Указание. а) Докажите неравенство *симметризации*:

$$\mathbb{E}\left[\sup_{a\in A}\big(M(a) - M_l(a)\big)\right] \leqslant 2\,\mathcal{R}(\lambda, A).$$

Для этого введите независимую копию обучающей выборки $\{(X_i', Y_i')\}_{i=1}^{l}$;

б) Воспользуйтесь два раза неравенством ограниченных разностей для случайных величин $\sup_{a\in A}\big(M(a) - M_l(a)\big)$ и $\mathcal{R}_l(\lambda, A)$;

в) Объедините все результаты с помощью неравенства Буля.

Замечание. Обратим внимание, что оценки (3) и (4) справедливы для любого класса A, в том числе несчетного. Случай задачи классификации и бинарной функции потерь был изучен ранее, и для него справедлива оценка Вапника–Червоненкиса, которая использует другую комбинаторную меру сложности семейства A, известную как *размерность Вапника–Червоненкиса*. Помимо того, оценка (4), в отличие от оценки (3), полностью вычислима по обучающей выборке.

39 (Неравенство Талаграна)**.** На декартовом произведении $\mathbb{X} \times \mathbb{Y}$ задана вероятностная мера $\mathbb{P}$, $\{(X_i, Y_i)\}_{i=1}^{n}$ — i.i.d обучающая выборка из $\mathbb{P}$, *минимизатор эмпирического риска* $a^* = \arg\min_{a\in A} M_l(a)$.

Наша цель – оценить отличие среднего риска алгоритма a^* от минимального среднего риска. Для этого введем понятие *избыточного риска*:

$$\mathcal{E}(a^*) = M(a^*) - \min_{a \in A} M(a).$$

Получите следующую верхнюю границу, с которой работать удобнее нежели с самим избыточным риском:

$$\begin{aligned}\mathcal{E}(a^*) \leqslant \sup_{a_1,a_2 \in A} \big((M - M_l)(a_1 - a_2)\big) \leqslant \\ \leqslant 2 \sup_{a \in A} \big((M(a) - M_l(a)\big). \end{aligned} \tag{1}$$

Этим путем были получены ключевые в Теории Статистического Обучения (ТСО) оценки Вапника–Червоненкиса, позже — оценки, основанные на Радемахеровских сложностях.

Проверьте, что если множество алгоритмов A, в некотором смысле, не слишком "сложно" (например, имеет конечную VC-размерность), то выражение в (1) имеет порядок $O(1/\sqrt{n})$.

Для конкретного распределения $\mathbb{P}$, скорее всего, только очень маленькая часть A подходит для решения задачи минимизации $M(a)$. По этой причине оптимальнее брать sup по множеству:

$$A(\delta) = \{a \in A \colon \mathcal{E}(a) \leqslant \delta\}.$$

Приходим к следующему результату:

$$\delta^* = \mathcal{E}(a^*) \leqslant \sup_{a_1,a_2 \in A(\delta^*)} \big((M - M_l)(a_1 - a_2)\big).$$

Используя неравенство Буске (одна из версий неравенства Талаграна), докажите, что с вероятностью не меньше $1 - e^{-t}$ справедливо следующее:

$$\sup_{a_1,a_2 \in A(\delta)} \big((M - M_l)(a_1 - a_2)\big) \leqslant \phi_l(\delta) + \sqrt{2\frac{t}{l}\big(D^2(\delta) + 2\phi_l(\delta)\big)} + \frac{t}{2l}, \tag{2}$$

где

$$D(\delta) = \sup_{a_1,a_2 \in A(\delta)} \sqrt{M\big((a_1 - a_2)^2\big)},$$

$$\phi_l(\delta) = \mathbb{E}\left[\sup_{a_1,a_2 \in A(\delta)} |(M - M_l)(a_1 - a_2)|\right].$$

Замечание. Неравенство (2) дает оценку порядка $o(1/\sqrt{n})$. Установление точного порядка зависит от конкретного выбора функции потерь и условий, накладываемых на распределение $\mathbb{P}$. В частности, часто в качестве такого условия берут Tsybakov's low noise condition.

Литература

1. Колмогоров А.Н. Основные понятия теории вероятностей. – М.: Наука, 1974.

2. Феллер В. Введение в теорию вероятностей и ее приложения. Т. 1, 2. – М.: Мир, 1984.

3. Боровков А.А. Теория вероятностей. – М.: Наука, 1986.

4. Гнеденко Б.В. Курс теории вероятностей. – М: Наука, 1988.

5. Натан А.А., Горбачев О.Г., Гуз С.А. Теория вероятностей: Учеб. пособие. – М.: МЗ Пресс – МФТИ, 2007.

6. Натан А.А., Горбачев О.Г., Гуз С.А. Основы теории случайных процессов: Учеб. пособие. – М.: МЗ Пресс – МФТИ, 2003.

7. Натан А.А., Горбачев О.Г., Гуз С.А. Математическая статистика: Учеб. пособие. – М.: МЗ Пресс – МФТИ, 2005.

8. Розанов Ю.А. Лекции по теории вероятностей. – М.: Долгопрудный: Издательский дом “Интеллект”, 2008.

9. Чеботарев А.М. Введение в теорию вероятностей и математическую статистику для физиков. – М.: МФТИ, 2009.

10. 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

11. Durrett R. Probability: Theory and Examples. – M.: Cambridge Univ. Press, 2010.

12. Ширяев А.Н. Вероятность 1, 2. – М.: МЦНМО, 2011.

13. Шень А. Вероятность: примеры и задачи. – М.: МЦНМО, 2012.

14. Босс В. Лекции по математике: Вероятность, информация, статистика. Т. 4 (см. также Т. 10, 12) – М.: УРСС, 2013.

15. Коралов Л.Б., Синай Я.Г. Теория вероятностей. Случайные процессы. – М.: МЦНМО, 2013.

16. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. – М.: Высшая школа, 1979.

17. Прохоров А.В., Ушаков В.Г., Ушаков Н.Г. Задачи по теории вероятностей. Основные понятия. Предельные теоремы. Случайные процессы. – М.: Наука, 1986.

18. Зубков А.М., Севастьянов Б.А., Чистяков В.П. Сборник задач по теории вероятностей. – М.: Наука, 1989.

19. Кельберт М.Я., Сухов Ю.М. Вероятность и статистика в примерах и задачах. 1 Основные понятия теории вероятностей и математической статистики. – М.: МЦНМО, 2007.

20. Кельберт М.Я., Сухов Ю.М. Вероятность и статистика в примерах и задачах. 2 Марковские цепи как отправная точка теории случайных процессов. – М.: МЦНМО, 2010.

21. Кельберт М.Я., Сухов Ю.М. Вероятность и статистика в примерах и задачах. 3 Теория информации и кодирования. – М.: МЦНМО, 2014.

22. Ширяев А.Н. Задачи по теории вероятностей. – М.: МЦНМО, 2011.

23. Ширяев А.Н., Эрлих И.Г., Яськов П.А. Вероятность в теоремах и задачах. – М.: МЦНМО, 2013.

24. Кац М. Вероятность и смежные вопросы в физике. – М.: Мир, 1965.

25. Секей Г. Парадоксы в теории вероятностей и математической статистике. – М.: РХД, 2003.

26. Стоянов Й. Контрпримеры в теории вероятностей. – М.: МЦНМО, 2012.

27. Кнут Д., Грэхем Р., Паташник О. Конкретная математика. Основание информатики. — М.: Мир; Бином. Лаборатория знаний, 2009.

28. Ландо С.К. Лекции о производящих функциях. - М.: МЦНМО, 2007.

29. Кингман Дж. Пуассоновские процессы. – М.: МЦНМО, 2007.

30. DasGupta A. Asymptotic theory of statistic and probability. - Springer, 2008.

31. Flajolet P., Sedgewick R. Analytic combinatorics. – M.: Cambr. Univ. Press, 2009. http://algo.inria.fr/flajolet/Publications/book.pdf

32. Гардинер К.В. Стохастические модели в естественных науках. – М.: Мир, 1986.

33. Ethier N.S., Kurtz T.G. Markov processes. – Wiley Series in Probability and Mathematical Statistics. New York, 2005.

34. Sandholm W. Population games and Evolutionary dynamics. Economic Learning and Social Evolution. – MIT Press. Cambridge, 2010.

35. Михайлов Г.А., Войтишек А.В Численное статистическое моделирование. Методы Монте-Карло. – М.: Академия, 2006.

36. Levin D.A., Peres Y., Wilmer E.L. Markov chain and mixing times. – AMS, 2009.

37. Алон Н., Спенсер Дж. Вероятностный метод. – М.: Бином, 2006.

38. Motwani R., Raghavan P. Randomized algorithms. – M.: Cambridge Univ. Press, 1995.

39. Ledoux M. Concentration of measure phenomenon. – M.: Amer. Math. Soc., Math. Surv. Mon. V. 89, 2005.

40. Hopcroft J., Kannan R. Computer Science Theory for the Information Age. - 2012. http://www.cs.cmu.edu/ venkatg/teaching/CStheory-infoage/

41. Boucheron S., Lugosi G., Massart P. Concentration inequalities: A nonasymptotic theory of independence. - Oxford University Press, 2013.

42. Janes E.T. Probability theory. The logic of science. - Cambridge University Press, 2003.

43. Cover T.M., Thomas J.A. Elements of Information theory. – M.: Wiley-Interscience, 2006.

44. Верещагин Н.К., Щепин Е.В. Информация, кодирование и предсказание. - М.: МЦНМО, 2012.

45. Motwani R., Raghavan P. Randomized algorithms. – M.: Cambridge Univ. Press, 1995.

46. Dubhashi D.P., Panconesi A. Concentration of measure for the analysis of randomized algorithms. - Cambridge University Press, 2009.

47. Кендалл М., Моран П. Геометрические вероятности. – М.: Наука, 1972.

48. Сантало Л. Интегральная геометрия и геометрическая вероятности. - М.: Наука, 1983

49. Lugosi G., Cesa-Bianchi N. Prediction, learning and games. - New York: Cambridge University Press, 2006.

50. Rakhlin A., Sridharan K. Statistical Learning Theory and Sequential Prediction. - STAT928, 2014. http://www-stat.wharton.upenn.edu/ rakhlin/

51. Bishop C.M. Pattern Recognition and Machine Learning. – M.: Springer, Information Science and Statistics, 2006.

52. Лагутин М.Б. Наглядная математическая статистика. - М.: Бином, 2009.

53. Голубев Г.К., А.Н. Соболевский, Спокойный В.Г.; Пособия по теории вероятностей и математической статистике. Электронные версии доступны здесь. http://premolab.ru/content/books