Федеральное государственное бюджетное учреждение науки
Вычислительный центр имени А. А. Дородницына Росскийской академии наук

На правах рукописи

Толстихин Илья Олегович

Неравенства концентрации вероятностной меры в трансдуктивном обучении и PAC-Байесовском анализе

Специальность 05.13.17 —
«Теоретические основы информатики»

Диссертация на соискание учёной степени
кандидата физико-математических наук

Научный руководитель:
д. ф.-м. н.
Воронцов К. В.

Москва – 2014

Содержание

Введение

Диссертационная работа посвящена теории статистического обучения, изучающей свойства процедур обучения в рамках строгого математического формализма.

Актуальность темы. Задачи поиска закономерностей или восстановления функциональных зависимостей в наблюдаемых данных сегодня играют ключевую роль во многих прикладных областях. Методы машинного обучения, позволяющие во многих случаях эффективно решать задачи распознавания образов, классификации, восстановления регрессии, оценивания неизвестной плотности и другие задачи предсказания, стали неотъемлемой частью различных аспектов современной жизни. С теоретической точки зрения, важным вопросом является выявление факторов, влияющих на качество работы найденных на основе обучающей выборки закономерностей на новых данных, что позволило бы разрабатывать новые и более качественные алгоритмы обучения.

Теория статистического обучения (или *VC-теория*), предложенная в работах В. Н. Вапника и А. Я. Червоненкиса в конце 1960-х годов и позже получившая мировую известность, впервые позволила строго описать соотношение между необходимым для успешного обучения числом наблюдаемых данных и сложностью используемого класса отображений. Подобные результаты формулируются обычно в виде верхних границ (или *оценок*) на вероятность того, что найденное на основе обучающей выборки отображение даст ошибочный ответ на новых данных. Проблемой первых оценок была их сильная завышенность, обусловленная попыткой получения результатов, справедливых в чересчур общих постановках. Дальнейшее развитие VC-теории было связано с попытками улучшения точности оценок на основе учета различных свойств рассматриваемых задач [18]. Среди предложенных за последние 45 лет подходов VC-теории можно выделить результаты, основанные на покрытиях класса отображений [8, 45], на учете отступов объектов [49, 84], на понятии стабильности процедуры обучения [22], на глобальной Радемахеровской сложности класса [7, 52], на локальных мерах сложности класса [5, 47, 52, 66] и на изучении рандомизированных отображений [57, 73, 85, 90].

Несмотря на многочисленные попытки улучшения точности оценок, которые продолжаются до настоящего времени [78], она остается по-прежнему не достаточной для применения оценок на практике. Поэтому актуальной проблемой является получение более точных оценок, с одной стороны достаточно общих, но с другой стороны учитывающих специфику решаемой прикладной задачи.

Цель диссертационной работы. Улучшение точности существующих оценок в теории статистического обучения на основе современных результатов теории неравенств концентрации вероятностной меры. Получение новых неравенств концентрации для выборок случайных величин без возвращения, учитывающих их дисперсии.

Методы исследования. В первой части настоящей работы используется *энтропийный* подход в *теории неравенств концентрации вероятностной меры* [19], предложенный М. Леду и развитый П. Массаром, С. Бушроном, Г. Лугоши и О. Буске в работах [19,58,66,67]. Данный подход позволяет получать неравенства концентрации, сравнимые по точности с более сильными результатами *индуктивного подхода* [93] М. Талаграна, избегая при этом чересчур громоздких доказательств. В частности, ключевую роль будут играть субгауссовское неравенство концентрации для функций, определенных на срезах Булева куба, полученное С. Г. Бобковым в работе [17], и неравенство Талаграна для супремумов эмпирических процессов, полученное в [94] и позже усиленное О. Буске на основе энтропийного подхода в работе [21].

Вторая часть работы будет использовать подход в *теории статистического обучения* [18,104], существенно основанный на результатах теории неравенств концентрации вероятностной меры и *теории эмпирических процессов* [100,101]. Предложенный впервые в конце 60-х годов в работах В. Н. Вапника и А. Я. Червоненкиса [104–106], данный подход продолжает активно развиваться и на сегодняшний день. В частности, ряд важных результатов настоящей работы будет основан на так называемом *локальном* подходе, развитом в начале 2000-х годов В. Колчинским, Д. Панченко, П. Массаром, П. Бартлетом, О. Буске, Ш. Мендельсоном, Г. Лугоши и рядом других авторов в серии работ [5, 47, 52, 66]. В отличие от большинства других подходов теории Вапника-Червоненкиса, локальный подход позволяет эффективно учитывать свойства конкретной решаемой задачи при оценке качества процедуры обучения, что часто ведет к существенно более точным результатам.

Третья часть работы основана на *комбинаторной теории переобучения*, предложенной К. В. Воронцовым [109–111, 116].

Наконец, четвертая часть работы использует *PAC-Байесовский анализ* — относительно новый подход в теории статистического обучения, предложенный Д. МакАллистером и Дж. Шоу-Тейлором в работах [73, 74, 90] и далее развитый Дж. Лэнгфордом, М. Зигером и рядом других авторов в работах [57, 72, 85]. Известно, что оценки PAC-Байесовского анализа в ряде прикладных задач ведут к наиболее точным на сегодняшний день результатам.

Основные положения, выносимые на защиту:

1. Получено два новых неравенства концентрации типа Талаграна для супремумов эмпирических процессов и выборок без возвращений, которые учитывают дисперсии случайных величин.

2. Получена новая оценка избыточного риска в трансдуктивной постановке теории статистического обучения, основанная на локальных мерах сложности рассматриваемого класса отображений и впервые в трансдуктивном подходе ведущая к быстрой скорости сходимости в общих предположениях.

3. В рамках теоретико-группового подхода комбинатонной теории переобучения предложено учитывать орбиты множества выборок при вычислении точного значения вероятности переобучения. На основе этого подхода получены новые точные (не завышенные) оценки вероятности переобучения для трех модельных семейств отображений, бинарные векторы ошибок которых являются различными подмножествами шара в Булевом кубе.

4. В рамках PAC-Байесовского анализа теории статистического обучения получено новое PAC-Байесовское эмпирическое неравенство Бернштейна, полностью вычислимое на основе обучающей выборки и во многих случаях ведущее к существенно более точным оценкам по сравнению с известными ранее результатами.

Научная новизна. В диссертационной работе впервые доказано, что в трансдуктивной постановке теории статистического обучения быстрая скорость сходимости может достигаться в достаточно общих предположениях. В частности, продемонстрировано, что избыточный риск при использовании метода минимизации эмпирических потерь определяется величиной неподвижной точки модуля непрерывности эмпирического процесса в окрестностях оптимального на генеральной выборке отображения. Подобные результаты до этого были известны в задачах М-оценивания в теории эмпирических процессов и позже в индуктивной постановке

теории статистического обучения. Однако, все они были основаны на неравенстве Талаграна для эмпирических процессов, которое, в свою очередь, существенно опирается на предположение о независимости случайных величин и, следовательно, не может быть использовано в трансдуктивной постановке теории статистического обучения.

Для преодоления этой трудности в диссертационной работе были впервые получены аналоги неравенства Талаграна, которые справедливы для случайных величин, выбранных равномерно без возвращений из произвольного конечного множества. До этого в литературе были известны лишь неравенства типа МакДиармида для супремумов эмпирических процессов и выборок без возвращений, не учитывающие дисперсии случайных величин.

Результаты, полученные в рамках теоретико-группового подхода в комбинаторной теории переобуения, основаны на новой идее учета при вычислении вероятности переобучения *орбит разбиений* генеральной выборки.

Новое PAC-Байесовское эмпирическое неравенство Бернштейна является первым примером PAC-Байесовских неравенств, одновременно учитывающих дисперсии потерь и вычислимых на основе обучающей выборки.

Теоретическая значимость. Полученные в диссертационной работе неравенства концентрации типа Талаграна являются достаточно общими и могут быть использованы в теоретическом анализе большого числа современных прикладных задач (в том числе выходящих за рамки теории обучения), где важную роль играют выборки без возвращений. Одним из примеров таких задач является неасимптотический анализ свойств процедуры скользящего контроля, широко применяемой на практике.

Новые оценки избыточного риска и обобщающей способности, полученные в диссертационной работе, улучшают точность известных ранее в теории статистического обучения результатов, давая более глубокое понимание процесса обучения на основе эмпирических данных в трансдуктивной постановке. В частности, новые оценки позволяют заключить, что сложность задач трансдуктивного обучения, по крайней мере, не превосходит сложность задач индуктивного обучения. Более того, они показывают, что свойства задач трансдуктивного обучения в ряде случаев могут выгодно отличаться от свойств задач индуктивного обучения.

Новые результаты комбинаторного подхода, полученные в диссертационной работе, расширяют класс задач и семейств отображений, для которых возможно эффективное (полиномиальное по длине выборки) вычисление вероятности переобучения.

Практическая значимость. PAC-Байесовского эмпирическое неравенство Бернштейна, полученное в настоящей работе, во многих случаях ведет к существенно более точным оценкам обобщающей способности по сравнению с известными ранее результатами PAC-Байесовского анализа. Кроме того, новая оценка полностью вычислима на основе наблюдаемых данных. Это дает возможность эффективно применять ее на практике при решении задач обучения по прецедентам для оценивания качества получаемых решений или настройки гиперпараметров, избегая при этом больших вычислительных затрат процедуры скользящего контроля. Наконец, минимизация полученной оценки может вести к новым более точным методам обучения, имеющим гарантированную обобщающую способность.

Полученные в диссертационной работе оценки избыточного риска для трансдуктивного обучения могут вести к применимым на практике методам выбора моделей, основанным на использовании всех объектов генеральной совокупности. В частности, вместе со Следствием 15 (стр. 111) они могут вести к новым алгоритмам выбора ядер, поскольку собственные значения матрицы Грамма спрямляющего ядра, определяющие скорость сходимости метода минимизации эмпирического риска, в этом случае могут быть вычислены на основе наблюдаемых данных.

Степень достоверности. Достоверность результатов обеспечивается математическими доказательствами теорем и серией подробно описанных вычислительных экспериментов, результаты которых согласуются с теоретическими результатами настоящей работы.

Апробация работы. Результаты диссертационной работы неоднократно докладывались и обсуждались на следующих конференциях и научных семинарах:

1. Международная конференция «Ломоносов-2010», 2010 г. [120];
2. Международная конференция «Интеллектуализация обработки информации», 2010 г. [119];
3. Международная конференция «Интеллектуализация обработки информации», 2012 г. [121];
4. Международная конференция “Neural Information Processing Systems (NIPS)”, озеро Тахо, США, Декабрь 2013 г. [95];
5. Научный семинар группы проф. F. Laviolette и M. Marchand, Лавальский Университет, Квебек, Канада, Декабрь 2013 г.;

6. Три доклада на совместном НМУ–МФТИ семинаре «Стохастический анализ в задачах», Москва, Декабрь 2013 г. и Апрель 2014 г.;

7. Научный семинар группы профессора B. Schoelkopf, Max Planck Institute for Intelligent Systems, Тюбинген, Германия, Май 2014 г.;

8. Научный семинар Лаборатории 7 Института Проблем Управления РАН, Москва, Июнь 2014 г.;

9. Международная конференция “Conference on Learning Theory (COLT)”, Барселона, Испания, Июнь 2014 г. [96];

10. Научные семинары отдела Интеллектуальных систем Вычислительного Центра им. А. А. Дородницына РАН.

Публикации. Основные результаты настоящей диссертационной работы опубликованы в 7 работых [37, 95, 96, 119–121, 124], 3 из которых входят в список изданий, рекомендованных ВАК [95, 96, 124].

Личный вклад диссертанта заключается в выполнении основного объёма теоретических и экспериментальных исследований, изложенных в диссертационной работе. Все результаты, приведенные в настоящей работе, относятся к личному вкладу диссертанта, за исключением отдельно оговоренных случаев.

Подготовка к публикации полученных в работах [37,95,96,124] результатов проводилась совместно с соавторами. Все экспериментальные и основная часть теоретических результатов работы [95] получены лично автором. В работах [37,124] к личному вкладу автора относится разработка техники учета орбит разбиений при вычислении вероятности переобучения, использовавшаяся в доказательстве всех основных результатов данных работ, а также теоремы о вероятности переобучения центрального слоя Хэммингова шара и монотонного роста вероятности переобучения множеств бинарных векторов ошибок, лежащих в одном слое Булева куба.

Объем и структура работы Диссертация состоит из оглавления, введения, шести глав, заключения, списка иллюстраций (13 п.), списка таблиц (1 п.), списка литературы (124 п.) и списка обозначений. Общий объём работы составляет 201 страницу.

Краткое содержание работы по главам В Главе 1 приводится подробный обзор классических и современных результатов теории неравенств концентрации вероятностной меры для *независимых случайных величин*. В Разделе 1.1 рассматриваются суммы независимых и ограниченных случайных величин и приводятся неравенства Хефдинга, Бернштейна и Беннетта. В Разделе 1.2 рассматриваются функции с ограниченными разностями и приводятся неравенства Азумы-Хефдинга и МакДиармида. В Разделе 1.3 формулируются основные результаты энтропийного метода, включая неравенства для самоограничивающих функций, а также приводятся эмпирическое неравенство Бернштейна и неравенство Талаграна для супремумов эмпирических процессов.

Глава 2 посвящена результатам теории неравенств концентрации вероятностной меры для случайных величин, выбранных *без возвращений*. Первая часть главы содержит подробный обзор известных результатов: в Разделе 2.1 рассмотриваются суммы случайных величин и приводятся неравенство Стирлинга и метод редукции Хефдинга; в Разделе 2.2 рассматриваются функции, определенные на множестве разбиений генеральной выборки, и приводятся неравенство Эль-Янива-Печиони и субгауссовское неравенство С. Г. Бобкова. Вторая часть главы посвящена новым результатам для супремумов эмпирических процессов. В Разделе 2.3 получено два новых неравенства типа Талаграна (Теоремы 20 и 22), первое из которых основано на неравенстве С. Г. Бобкова, а второе — на методе редукции Хефдинга и неравенстве Талаграна.

В Главе 3 приведен подробный обзор ряда классических и современных результатов *теории статистического обучения*. Раздел 3.1 посвящен введению определений и обсуждению различных постановок задач. В Разделе 3.2 приводится подробный обзор и сравнение ряда подходов к получению оценок избыточных потерь и обобщающей способности. Сначала в Разделе 3.2.1 рассмотрены подходы, существенно опирающиеся на применении *неравенства Буля*, включая оценку Вапника-Червоненкиса, оценку «бритвы Оккама» и оценку, основанную на покрытиях множества отображений. Затем в Разделе 3.2.2 рассматриваются оценки, основанные на *Радемахеровской сложности*, а также приводятся неравенства симметризации и сжатия. Наконец, в Разделе 3.2.3 приводится обзор так называемого *локального анализа*, включая обсуждение условий ограниченного шума Маммена-Цыбакова и Массара, быстрых скоростей сходимости и локальных Радемахеровских сложностей.

В Главе 4 рассматривается трансдуктивная постановка теории статистического обучения. В Разделе 4.1 приводится формальная постановка задачи. В Разделе 4.2 на основе результатов Главы 2 получены новые оценки избыточных потерь (Теоремы 41 и 42, Следствия 13 и 14), опирающиеся на локальные меры сложности семейств отображений и впервые

в трансдуктивном обучении ведущие к быстрым скоростям сходимости в общих предположениях. Также получены новые оценки обобщающей способности, учитывающие дисперсию потерь (Теоремы 39 и 40). Подробные доказательства этих оценок приводятся в Разделе 4.3.

В Главе 5 рассматривается *комбинаторная теория переобучения*, тесно связанная с трансдуктивной постановкой. В Разделе 5.1 вводятся определения, ставится формальная постановка задачи и приводится ее сравнение с постановкой трансдуктивного обучения. В Разделе 5.2 рассматривается *теоретико-групповой подход* в комбинаторной теории. Первая часть этого раздела посвящена обзору известных результатов, который приводится в Разделе 5.2.1. Затем приводятся новые результаты теоретико-группового подхода: в Разделе 5.2.2 получена новая формула вычисления вероятности переобучения (Теорема 44), основанная на *орбитах множества разбиений* генеральной выборки на обучающую и контрольную подвыборки; на основе нее в Разделе 5.2.4 получены новые точные (не завышенные) оценки вероятности переобучения для трех модельных семейств отображений (Теоремы 45, 47 и 48).

Наконец, Глава 6 посвящена *PAC-Байесовскому анализу* в теории статистического обучения. В Разделах 6.1 и 6.2 описывается общий подход к получению PAC-Байесовских неравенств и приводится достаточно подробный обзор известных результатов, включая неравенство МакАллистера (также известное как PAC-Байесовское неравенство Хефдинга), PAC-Байесовское неравенство Бернштейна и PAC-Байесовское kl-неравенство. Также в этих разделах приводится подробное сравнение трех описанных неравенств и обсуждаются способы их использования в теории статистического обучения. В Разделе 6.3 получено новое PAC-Байесовское неравенство для неизвестной усредненной дисперсии потерь отображений (Теорема 54), основанное на результатах первой главы. Вместе с PAC-Байесовским неравенством Бернштейна оно ведет к новому *PAC-Байесовскому эмпирическому неравенству Бернштейна* (Теорема 55) — мощному и полностью вычислимому на основе обучающей выборки неравенству, во многих случаях ведущему к существенно более точным оценкам по сравнению с известными ранее результатами. В конце раздела приводятся численные эксперименты с модельными выборками и реальными данными из репозитория UCI, демонстрирующие превосходство полученного неравенства над известными ранее аналогами.

Благодарности Автор выражает благодарности своему научному руководителю Константину Вячеславовичу Воронцову за постановку задачи, а также жене и родителям за терпение и поддержку.

1 Неравенства концентрации для независимых случайных величин

В данной главе приводится обзор *неравенств концентрации вероятностной меры* — мощных математических инструментов, на которых будет основано большинство результатов настоящей работы.

Рассмотрим функцию $g\colon \mathcal{X}^n \to \mathbb{R}$, зависящую от *большого числа* аргументов, принимающих значения в некотором множестве $\mathcal{X}$. Рассмотрим также *большое число* случайных величин $\{X_1, \dots, X_n\} \subset \mathcal{X}$, принимающих значения в $\mathcal{X}$. Мы хотим получить возможность контролировать отклонения случайных значений $Q = g(X_1, \dots, X_n)$ от математического ожидания $\mathbb{E}[g(Q)]$. При этом нам будет важно получить *неассимптотические* результаты, которые, в отличие от закона больших чисел и других классических предельных теорем, оставались бы применимыми при конечных размерах выборок n.

Поставленной цели можно добиваться двумя эквивалентными способами. С одной стороны, можно попытаться оценивать вероятности больших уклонений

$$\mathsf{P}\{Q - \mathbb{E}[Q] \geq t\}, \quad \mathsf{P}\{\mathbb{E}[Q] - Q \geq t\}$$

для положительных $t \geq 0$. С другой стороны, можно пытаться получать верхние оценки для случайных отклонений

$$Q - \mathbb{E}[Q], \quad \mathbb{E}[Q] - Q,$$

которые бы выполнялись с большой вероятностью относительно случайной реализации выборки $\{X_1, \dots, X_n\}$.

На сегодняшней день достаточно подробно изучен специальный случай, когда случайные величины $\{X_1, \dots, X_n\}$ *независимы* [19]. Первые исследования были связаны с суммами случайных величин, для которых начиная с 1950-х годов С. Н. Бернштейном, Дж. Беннеттом, В. Хефдингом и другими авторами были получены верхние оценки вероятности больших уклонений, *экспоненциально убывающие* с ростом размера выборки n [11, 12, 43]. Позже пришло понимание, что похожими свойствами обладают более общие функции g. Так, развитие в

1970-х годах метода мартингалов показало, что если функция g не зависит «слишком сильно» ни от одного из своих аргументов, то ее значения Q также концентрируются вокруг математического ожидания $\mathbb{E}[Q]$. Одним из наиболее плодотворных в этом направлении результатов, по-видимому, оказался *метод ограниченных разностей* К. МакДиармида. Так, неравенство МакДиармида и по сегодняшний день является одним из наиболее часто применяемых неравенств концентрации в самых разных областях математики. Исследованию свойств функций g, ведущих к концентрации их значений, было посвящено множество других подходов. Среди прочих стоит упомянуть *транспортный метод* К. Мартон [63], основанный на методах теории информации и по технике сильно отличающийся от всех других. Важной вехой в теории концентрации стало появление *индуктивного метода*, развитого М. Талаграном в середине 1990-х годов [93]. В частности, на основе индуктивного метода М. Талаграном было получено функциональное обобщение неравенства Бернштейна, известное сегодня как неравенство Талаграна для эмпирических процессов. К сожалению, метод М. Талаграна был основан на нетривиальной технике индукции и доказательства результатов отличались своей технической сложностью. Позже на пути поиска метода, позволяющего упростить получение сравнимых с индуктивным методом результатов, М. Леду [58] был предложен *энтропийный метод*, который позже был существенно развит П. Массаром, С. Бушроном, Г. Лугоши и О. Буске в работах [19, 20, 66, 67].

В настоящей главе мы приведем достаточно подробный обзор ряда классических и современных результатов теории концентрации вероятностной меры для независимых случайных величин, упомянутых выше. Подробный обзор может быть найден в книге [19].

1.1 Суммы независимых случайных величин

Мы начнем обзор с рассмотрения наиболее классического объекта изучения теории вероятностей — суммы независимых случайных величин, принимающих действительные значения. На этом примере мы познакомимся с основными подходами к получению неравенств концентрации, которые лягут в основу более общих результатов следующих разделов. Поскольку *простые* выборки, то есть выборки независимых и одинаково распределенных случайных величин, часто встречаются в теории машинного обучения, результаты настоящего раздела будут активно использоваться на протяжении всех следующих глав.

1.1.1 Неравенства Маркова, Чернова и метод Чернова

Одним из самых простых и в то же время полезных неравенств является следующее неравенство Маркова, которое ограничивает вероятность больших значений *неотрицательных* случайных величин:

Теорема 1 (неравенство Маркова)**.** *Для любой неотрицательной случайной величины ξ и произвольного $\epsilon > 0$:*

$$\mathbb{P}\{\xi \geqslant \epsilon\} \leqslant \frac{\mathbb{E}[\xi]}{\epsilon}.$$

Само по себе неравенство Маркова будет редко использоваться нами. Больший интерес представляет очевидное следствие из него: если $\phi\colon \mathbb{R} \to \mathbb{R}^+$ — произвольная неубывающая и неотрицательная функция, тогда для *любой* случайной величины ξ и $\epsilon > 0$ справедливо:

$$\mathbb{P}\{\xi \geq \epsilon\} \leq \mathbb{P}\{\phi(\xi) \geq \phi(\epsilon)\} \leq \frac{\mathbb{E}[\phi(\xi)]}{\phi(\epsilon)}.$$

В частности, если мы положим $\phi(\xi) = \xi^2$, то получим следующее неравенство Чебышева:

Теорема 2 (Неравенство Чебышева)**.** *Для любой случайной величины ξ и любого $\epsilon > 0$ выполнено:*

$$\mathbb{P}\{|\xi - \mathbb{E}[\xi]| \geqslant \epsilon\} \leqslant \frac{\mathbb{D}[\xi]}{\epsilon^2}.$$

Это первый пример неравенства концентрации, который мы приводим в настоящей главе. С ростом отклонения ϵ вероятность убывает, и скорость убывания контролируется дисперсией случайной величины. Из неравенств концентрации легко получить верхние (или нижние) оценки случайных величин, которые выполняются с большой вероятностью. Подобный переход основан на *обращении вероятности:*

Лемма 1. *Пусть для случайной величины ξ и произвольного $\epsilon \geq 0$ справедливо неравенство концентрации:*

$$\mathbb{P}\{\xi - \mathbb{E}[\xi] \geqslant \epsilon\} \leqslant B_\xi(\epsilon),$$

для некоторой непрерывной и монотонно убывающей неотрицательной функции $B_\xi\colon \mathbb{R}^+ \to \mathbb{R}^+$. Тогда для любой $\delta \in [0, 1]$ следующая оценка справедлива с вероятностью не меньше $1 - \delta$:

$$\xi \leq \xi + B_\xi^{-1}(\delta).$$

Правая часть неравенства Чебышева имеет порядок $1/\epsilon^2$. В дальнейшем нас будут интересовать неравенства, правая часть которых убывает как $\exp(-\epsilon)$ — то есть *экспоненциальные* неравенства концентрации. Одним из стандартных методов получения подобных неравенств является так называемый *метод Чернова*, заключающийся в применении неравенства

Маркова:

$$\mathsf{P}\{\xi - \mathbb{E}[\xi] \geq \epsilon\} = \mathsf{P}\left\{e^{\lambda(\xi - \mathbb{E}[\xi])} \geq e^{\lambda\epsilon}\right\} \leq \frac{\mathbb{E}\left[e^{\lambda(\xi - \mathbb{E}[\xi])}\right]}{e^{\lambda\epsilon}} \tag{1.1}$$

для неотрицательного действительного числа $\lambda \geq 0$, получении верхней оценки $F(\lambda)$ на производящую функцию моментов случайной величины $\xi - \mathbb{E}[\xi]$:

$$\mathbb{E}\left[e^{\lambda(\xi - \mathbb{E}[\xi])}\right] \leq F(\lambda) \tag{1.2}$$

и последующей минимизации полученной оценки по $\lambda \geq 0$:

$$\mathsf{P}\{\xi - \mathbb{E}[\xi] \geq \epsilon\} \leq \min_{\lambda \geq 0} \frac{F(\lambda)}{e^{\lambda\epsilon}}.$$

Теорема 3 (Метод Чернова)**.** *Для любой случайной величины ξ и любого $\epsilon > 0$ выполнено:*

$$\mathbb{P}\{\xi \geq \epsilon\} \leq \min_{\lambda \geq 0} \frac{\mathbb{E}\left[e^{\lambda\xi}\right]}{e^{\lambda\epsilon}}.$$

На методе Чернова будут основаны все без исключения результаты, приведенные в данной главе. Функция $\psi_\xi(\lambda) = \mathbb{E}\left[e^{\lambda\xi}\right]$ для $\lambda \geq 0$ называется производящей функцией моментов случайной величины ξ, поскольку, как несложно проверить, $\frac{\partial^n}{\partial\lambda^n}\psi_\xi(\lambda)\big|_{\lambda=0} = \mathbb{E}[\xi^n]$. Метод Чернова позволяет свести изучение поведения случайной величины к изучению ее производящей функции. В большинстве случаев производящую функцию в числителе заменяют на ее верхнюю оценку и уже затем проводят оптимизацию по λ.

Обычно наиболее сложным шагом описанного подхода является построение верхней оценки для производящей функции моментов $\psi_\xi(\lambda)$. В дальнейшем для наглядного обсуждения результатов нам будет полезно помнить, что для центрированной нормально распределенной случайной величины $\xi \sim \mathcal{N}(0, \sigma^2)$ с дисперсией σ^2 и для любого $\lambda \geq 0$ справедливо следующее тождество:

$$\mathbb{E}\left[e^{\lambda\xi}\right] = \exp\left(\frac{\lambda^2\sigma^2}{2}\right). \tag{1.3}$$

В частности, применение описанного выше метода Чернова дает нам следующее неравенство концентрации для нормальной случайной величины:

$$\mathbb{P}\{\xi \geq \epsilon\} \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right). \tag{1.4}$$

В следующем параграфе мы увидим, что в тех случаях, когда ξ — сумма независимых случайных величин, метод Чернова очень простым образом ведет к экспоненциальным неравенствам концентрации для $\xi - \mathbb{E}[\xi]$.

1.1.2 Неравенство Хефдинга

Пусть $\xi_1, \dots, \xi_n$ — последовательность независимых случайных величин, таких что $\xi_i \in [a_i, b_i]$ с вероятностью 1 для некоторых $a_i, b_i \in \mathbb{R}$, $i = 1, \dots, n$. Введем обозначение $S_n = \sum_{i=1}^n \xi_i$. Мы хотим изучать отклонение случайной величины S_n от ее среднего значения $\mathbb{E}[S_n]$. То есть получить неравенство концентрации для $\xi = S_n - \mathbb{E}[S_n]$. Воспользовавшись для этого методом Чернова, получим, что для любого $\lambda \geq 0$:

$$\begin{aligned}\mathbb{P}\{S_n - \mathbb{E}[S_n] \geqslant \epsilon\} = \mathbb{P}\{e^{\lambda(S_n - \mathbb{E}[S_n])} \geqslant e^{\lambda\epsilon}\} &\leqslant \frac{\mathbb{E}\left[e^{\lambda(S_n - \mathbb{E}[S_n])}\right]}{e^{\lambda\epsilon}} \\ = \frac{\mathbb{E}\left[e^{\lambda\sum_{i=1}^n(\xi_i - \mathbb{E}[\xi_i])}\right]}{e^{\lambda\epsilon}} &= \frac{\mathbb{E}\left[\prod_{i=1}^n e^{\lambda(\xi_i - \mathbb{E}[\xi_i])}\right]}{e^{\lambda\epsilon}}. \end{aligned} \tag{1.5}$$

На этом примере мы продемонстрируем главное достоинство работы с суммами независимых случайных величин. Поскольку для независимых случайных величин ξ' и ξ'' справедливо $\mathbb{E}[\xi'\,\xi''] = \mathbb{E}[\xi']\,\mathbb{E}[\xi'']$ мы можем продолжить цепочку равенств (1.5):

$$\mathbb{P}\{S_n - \mathbb{E}[S_n] \geqslant \epsilon\} \leq \frac{\mathbb{E}\left[\prod_{i=1}^n e^{\lambda(\xi_i - \mathbb{E}[\xi_i])}\right]}{e^{\lambda\epsilon}} = \frac{\prod_{i=1}^n \mathbb{E}\left[e^{\lambda(\xi_i - \mathbb{E}[\xi_i])}\right]}{e^{\lambda\epsilon}}. \tag{1.6}$$

Нам остается построить верхние оценки для производящих функций $\psi_{\xi_i}(\lambda)$. Следующий результат дает нам такие оценки в тех случаях, когда случайные величины ξ_i принимают значения из ограниченных интервалов.

Лемма 2 (Лемма Хефдинга)**.** *Для любой случайной величины ξ, такой что $\mathbb{E}[\xi] = 0$ и $\xi \in [a, b]$ с вероятностью 1, для любого $\lambda > 0$ справедливо:*

$$\mathbb{E}\left[e^{\lambda\xi}\right] \leqslant e^{\frac{\lambda^2(b-a)^2}{8}}.$$

Доказательство основано на выпуклости экспоненты и может быть найдено в [43].

Замечание 1. *Обратим внимание, что для случайной величины ξ, ограниченной в интервале $[a, b]$, справедлива следующая цепочка неравенств:*

$$\begin{aligned}\mathbb{D}[\xi] = \mathbb{D}[\xi - a] &= (b-a)^2\mathbb{D}\left[\frac{\xi - a}{b-a}\right] \\ &= (b-a)^2\left(\mathbb{E}\left[\left(\frac{\xi - a}{b-a}\right)^2\right] - \left(\mathbb{E}\left[\frac{\xi - a}{b-a}\right]\right)^2\right) \\ &\leq (b-a)^2\left(\mathbb{E}\left[\frac{\xi - a}{b-a}\right] - \left(\mathbb{E}\left[\frac{\xi - a}{b-a}\right]\right)^2\right) \\ &\leq \frac{(b-a)^2}{4}. \end{aligned} \tag{1.7}$$

Сравним верхнюю оценку Леммы 2 с производящей функцией моментов нормальной случайной величины (1.3). Грубо говоря, лемма Хефдинга дает тот же результат, используя вместо дисперсии случайной величины $\mathbb{D}[\xi]$ ее верхнюю оценку. Случайные величины,

производящие функции которых ограничены сверху выражением $e^{\lambda^2 v}$ *для* $v \in \mathbb{R}^+$ *принято называть* субгауссовскими.

Применив Лемму 2 в неравенстве (1.6), мы получим:

$$\mathbb{P}\{S_n - \mathbb{E}[S_n] \geqslant \epsilon\} \leq \frac{\prod_{i=1}^n \mathbb{E}\left[e^{\lambda(\xi_i - \mathbb{E}[\xi_i])}\right]}{e^{\lambda\epsilon}} \leqslant \frac{\prod_{i=1}^n e^{\lambda^2(b_i-a_i)^2/8}}{e^{\lambda\epsilon}} = \frac{e^{\lambda^2 \sum_{i=1}^n (b_i-a_i)^2/8}}{e^{\lambda\epsilon}}. \tag{1.8}$$

Мы применили лемму к случайным величинам $\xi_i - \mathbb{E}[\xi_i]$, которые с вероятностью 1 лежат в интервалах $\left[a_i - \mathbb{E}[\xi_i], b_i - \mathbb{E}[\xi_i]\right]$.

Нам остается минимизировать правую часть неравенства (1.6) по $\lambda \geq 0$. Выбор

$$\lambda = \frac{4\epsilon}{\sum_{i=1}^n (b_i - a_i)^2}$$

дает нам следующий результат:

Теорема 4 (Неравенство Хефдинга)**.** *Пусть* $\xi_1, \ldots, \xi_n$ *— последовательность ограниченных и независимых случайных величин, таких что* $\xi_i \in [a_i, b_i]$, $a_i, b_i \in \mathbb{R}$, $i = 1, \ldots, n$ *с вероятностью 1. Тогда для любого* $\lambda \geq 0$ *справедливо:*

$$\mathbb{E}\left[e^{\lambda(S_n - \mathbb{E}[S_n])}\right] \leq \exp\left(\lambda^2 \frac{\sum_{i=1}^n (b_i - a_i)^2}{8}\right)$$

Также для любого $\epsilon > 0$ *справедливо:*

$$\mathbb{P}\{S_n - \mathbb{E}[S_n] \geqslant \epsilon\} \leqslant \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{1.9}$$

Аналогичное неравенство справедливо для $\mathbb{P}\{\mathbb{E}[S_n] - S_n \geqslant \epsilon\}$*, поскольку условия теоремы инварианты относительно замены знака слагаемых.*

Замечание 2. *На следующем примере мы продемонстрируем стандартный способ применения неравенства Буля, которым мы будем часто пользоваться в дальнейших главах:*

$$\begin{aligned}\mathbb{P}\left\{\left|S_n - \mathbb{E}[S_n]\right| \geqslant \epsilon\right\} &\leqslant \mathbb{P}\{S_n - \mathbb{E}[S_n] \geqslant \epsilon\} + \mathbb{P}\{\mathbb{E}[S_n] - S_n \geqslant \epsilon\} \\ &\leq 2\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).\end{aligned}$$

Если слагаемые ξ_i одинаково распределены, то простым следствием прошлой теоремы является следующее неравенство:

Следствие 1. *Пусть* $\xi_1, \ldots, \xi_n$ *— последовательность независимых и одинаково распределенных случайных величин, таких что* $\xi_i \in [a, b]$, $a_i, b_i \in \mathbb{R}$, $i = 1, \ldots, n$ *с вероятностью 1. Пусть* $\mathbb{E}[\xi_1] = \mu$*. Тогда для любого* $\epsilon > 0$ *справедливо:*

$$\mathbb{P}\left\{\frac{1}{n}S_n - \mu \geqslant \epsilon\right\} \leqslant \exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right). \tag{1.10}$$

Аналогичное неравенство справедливо для $\mathbb{P}\left\{\mu - \frac{1}{n}S_n \geqslant \epsilon\right\}$. *Кроме того, для любого* $t \geq 0$ *с вероятностью не меньше* $1 - e^{-t}$ *справедливо:*

$$\frac{1}{n}S_n \leq \mu + (b-a)\sqrt{\frac{t}{2n}}. \tag{1.11}$$

Последняя оценка получена применением Леммы 1.

Итак, мы получили первый пример экспоненциального неравенства концентрации для суммы независимых и ограниченных случайных величин. Рассмотрим подробнее случай одинаково распределенных случайных величин $\xi_1, \ldots, \xi_n$ с $\mathbb{E}[\xi_1] = \mu$ и $\mathbb{D}[\xi_1] = \sigma^2$. В отличие от центральной передельной теоремы, которая дает следующий *асимптотический* результат:

$$\lim_{n\to\infty} \mathbb{P}\left\{\frac{1}{n}S_n - \mu \geqslant \frac{\epsilon}{\sqrt{n}}\right\} \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right) \tag{1.12}$$

(где мы воспользовались оценкой (1.4) на правый хвост нормального распределения), неравенство Хефдинга дает *неасимптотический* результат, справедливый для выборок любых длин n. При этом, как было отмечено в замечании 1, оценка неравенства Хефдинга воспроизводит верхнюю оценку центральной предельной теоремы (1.12) с точностью до замены дисперсии σ^2 ее верхней оценкой. При росте n среднее выборочное $\frac{1}{n}S_n$ сходится к математическому ожиданию μ со скоростью $n^{-1/2}$. В следующих главах мы увидим, что эта скорость является типичной во многих случаях.

Неравенство Хефдинга учитывает лишь ширину интервалов значений слагаемых ξ_i. Оказывается, в ряде случаев этого недостаточно. Рассмотрим две случайных величины:

(а) $\xi^b := \frac{1}{n}S_n^b$, где S_n^b — сумма n независимых случайных величин Бернулли, принимающих значения из $\{0, 1\}$ с вероятностями $1/2$ каждое;

(б) $\xi^u := \frac{1}{n}S_n^u$, где S_n^u — сумма n независимых и равномерно распределенных на отрезке $[0, 1]$ случайных величин.

Математические ожидания обеих сумм совпадают и равны 0.5. На Рисунке 1.1 приведены эмпирические оценки вероятностей $\mathbb{P}\{\xi - \mathbb{E}[\xi] \geq \epsilon\}$ для ξ^b и ξ^u при $n = 100$, каждая из которых посчитана по 10 000 случайных реализаций ξ. Мы видим, что случайная величина S_n^u значительно *сильнее сконцентрирована* вокруг ее среднего значения 0.5, чем S_n^b. Однако, неравенство Хефдинга (1.10) дает одинаковые верхние оценки (розовая пунктирная линия), поскольку интервалы значений слагаемых ξ_i совпадают в обоих случаях.

Разница в поведении этих случайных величин объясняется различием их дисперсий: дисперсия случайной величины Бернулли равна 0.25, в то время как равномерно распределенная случайная величина имеет дисперсию $\frac{1}{12}$. Отметим, что 0.25 — наибольшее возможное

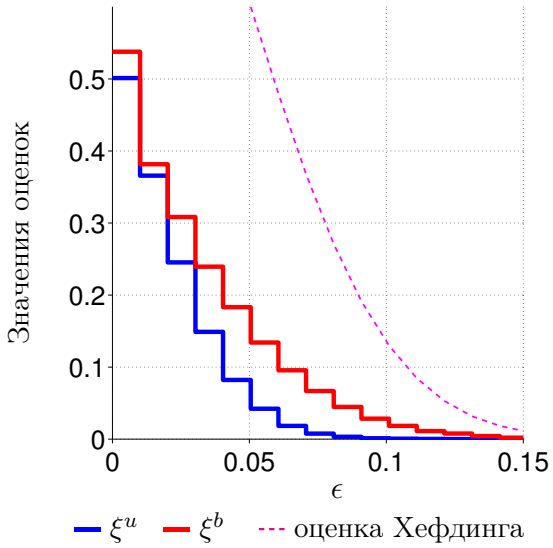


Рисунок 1.1: Эмпирические оценки $\mathbb{P}\{\xi - \mathbb{E}[\xi] \geq \epsilon\}$ для ξ^b и ξ^u (жирные линии) вместе с оценкой Хефдинга (пунктиром).

значение дисперсии случайной величины, ограниченной на интервале $[0, 1]$. Мы уже видели на примере неравенства Чебышева, что дисперсия случайной величины должна играть важную роль при изучении ее концентрации вокруг математического ожидания. В следующем разделе мы приведем экспоненциальные неравенства концентрации для сумм независимых случайных величин, учитывающие их дисперсии.

1.1.3 Неравенства Беннета и Бернштейна

Продолжим рассмотрение суммы независимых случайных величин $S_n = \sum_{i=1}^n \xi_i$. На этот раз мы будем предполагать, что случайные величины ограничены сверху $\xi_i \leq 1$ и центрированы: $\mathbb{E}[\xi_i] = 0$. Для таких случайных величин справедлива следующая верхняя оценка производящих функций:

Лемма 3 (Лемма Беннетта)**.** *Для любой случайной величины ξ, такой что $\mathbb{E}[\xi] = 0$ и $\xi \leq 1$ с вероятностью 1, для любого $\lambda \geq 0$ справедливо:*

$$\mathbb{E}\left[e^{\lambda\xi}\right] \leq \exp\left((e^\lambda - \lambda - 1)\mathbb{E}[\xi^2]\right).$$

Кроме того, поскольку $e^\lambda - \lambda - 1 \leq (e-2)\lambda^2$ для $\lambda \in [0, 1]$, для $\lambda \in [0, 1]$ справедливо:

$$\mathbb{E}\left[e^{\lambda\xi}\right] \leq \exp\left((e-2)\lambda^2\mathbb{E}[\xi^2]\right).$$

Доказательство этого результата может быть найдено в [19].

Воспользуемся этим результатом, вернувшись к неравенству (1.6). Тогда мы получим:

$$\mathbb{P}\{S_n \geqslant \epsilon\} \leq \frac{\prod_{i=1}^n \mathbb{E}\left[e^{\lambda\xi_i}\right]}{e^{\lambda\epsilon}} \leq \frac{e^{(e^\lambda-\lambda-1)\sum_{i=1}^n \mathbb{E}[\xi_i^2]}}{e^{\lambda\epsilon}} = \frac{e^{(e^\lambda-\lambda-1)\sum_{i=1}^n \mathbb{D}[\xi_i]}}{e^{\lambda\epsilon}} = \frac{e^{(e^\lambda-\lambda-1)\mathbb{D}[S_n]}}{e^{\lambda\epsilon}}.$$

Правая часть последнего неравенства достигает своего минимума при

$$\lambda = \ln\left(1 + \frac{\epsilon}{\mathbb{D}[S_n]}\right).$$

Мы получили следующий результат:

Теорема 5 (Неравенство Беннета)**.** *Пусть $\xi_1, \ldots, \xi_n$ — последовательность независимых случайных величин, таких что $\mathbb{E}[\xi_i] = 0$ и $\xi_i \leq 1$ (с вероятностью 1) для $i = 1, \ldots, n$. Положим*

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^n \mathbb{D}[\xi_i].$$

Тогда для любых $\lambda \geq 0$ справедливо:

$$\mathbb{E}\left[e^{\lambda S_n}\right] \leq e^{(e^\lambda-\lambda-1)n\sigma^2}.$$

Также для любых $\epsilon \geq 0$ справедливо:

$$\mathbb{P}\left\{S_n \geq \epsilon\right\} \leq \exp\left(-n\sigma^2 h\left(\frac{\epsilon}{n\sigma^2}\right)\right),$$

где $h(u) = (1+u)\ln(1+u) - u$ для $u \geq 0$.

Применив элементарное неравенство $h(u) \geq u^2/(2 + 2u/3)$, справедливое для $u \geq 0$, мы немедленно получаем следующую теорему:

Теорема 6 (Неравенство Бернштейна)**.** *В условиях теоремы 5 для $\epsilon \geq 0$ справедливо:*

$$\mathbb{P}\left\{S_n > \epsilon\right\} \leq \exp\left(-\frac{\epsilon^2}{2n(\sigma^2 + \frac{\epsilon}{3n})}\right).$$

Кроме того для любого $t \geq 0$ с вероятностью не меньше $1 - e^{-t}$ справедливо[1]*:*

$$S_n \leq \sqrt{2n\sigma^2 t} + \frac{2t}{3}.$$

Последнее неравенство теоремы вытекает из первого после применения Леммы 1 и неравенства $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$.

Также для удобства приведем следующее следствие:

[1] Позже мы покажем, что с помощью более аккуратного анализа можно избавиться от множителя 2 в последнем слагаемом.

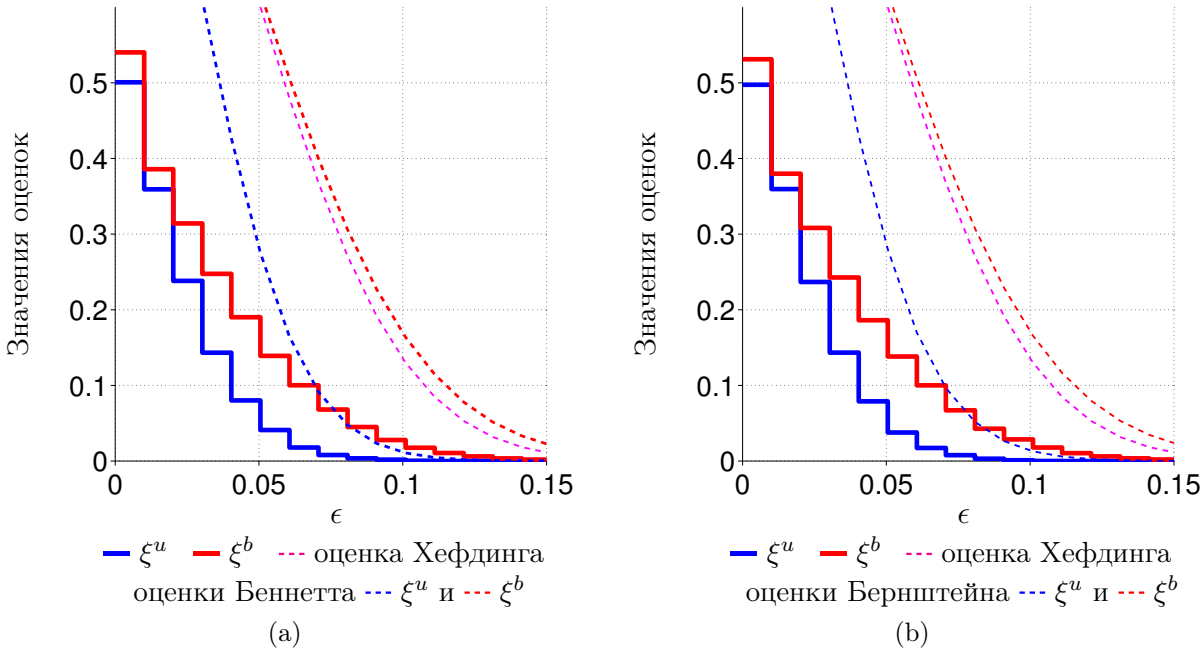


Рисунок 1.2: Эмпирические оценки $\mathbb{P}\{\xi - \mathbb{E}[\xi] \geq \epsilon\}$ для ξ^b и ξ^u (жирные линии) вместе с соответствующими верхними оценками (a) Беннета и (b) Бернштейна.

Теорема 7. *В тех же условиях, что две последних теоремы, для любого* $\epsilon \geq 0$ *справедливо:*

$$\mathbb{P}\left\{\frac{1}{n}S_n > \epsilon\right\} \leq \exp\left(-n\sigma^2 h\left(\frac{\epsilon}{\sigma^2}\right)\right); \tag{1.13}$$

$$\mathbb{P}\left\{\frac{1}{n}S_n > \epsilon\right\} \leq \exp\left(-\frac{n\epsilon^2}{2(\sigma^2 + \epsilon/3)}\right). \tag{1.14}$$

Отметим интересное поведение оценки Бернштейна. При отклонении ϵ существенно большем, чем дисперсия случайных величин ($\epsilon \gg \sigma^2$), оценку (1.14) можно грубо заменить на $\exp(-3n\epsilon/2)$. С другой стороны, при $\epsilon \ll \sigma^2$ оценка превращается в $\exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right)$. Таким образом, у оценки Бернштейна существует два «режима»: для маленьких отклонений ϵ она имеет субгауссовское поведение $e^{-n\epsilon^2}$, а для больших — лапласовское $e^{-n\epsilon}$.

На Рисунке 1.2 приведены верхние оценки (1.13) и (1.14). Мы видим, что учет дисперсии существенно улучшил результаты — по крайней мере для равномерно распределенных случайных величин. Тот факт, что для случайных величин Бернулли неравенства Бернштейна и Беннетта дают худшие по сравнению с неравенством Хефдинга результаты можно объяснить следующими размышлениями. Дисперсия случайной величины Бернулли ξ равна $\frac{1}{4}$, поэтому неравенства (1.7) обращаются в равенство и верхняя оценка леммы Хефдинга принимает вид $\exp(\lambda^2\mathbb{D}[\xi]/2)$. Лемма Беннетта по сравнению с этим дает более завышенную оценку.

1.2 Неравенства Азумы-Хефдинга и МакДиармида

До настоящего момента мы имели дело лишь с одним типом функций $Z = f(\xi_1, \dots, \xi_n)$, зависящих от последовательности независимых случайных величин $\xi_1, \dots, \xi_n$, а именно — с суммами $Z = \sum_{i=1}^n \xi_i$. На примере сумм мы продемонстрировали, что применение метода Чернова ведет к экспоненциальным неравенствам концентрации. Оказывается, экспоненциальные неравенства концентрации, ограничивающие вероятность отклонения случайных величин от их математических ожиданий, подобные рассмотренным выше, справедливы для более широкого класса функций f. Изучению свойств функций f, достаточных для концентрации $Z = f(\xi_1, \dots, \xi_n)$ вблизи $\mathbb{E}[Z]$, посвящена теория неравенств концентрации меры [19]. В этом разделе мы приведем результаты одного из классических подходов к данной задаче, предложенного МакДиармидом в [76] и основанного на *методе мартингалов*.

Рассмотрим последовательность независимых случайных величин $\xi_1, \dots, \xi_n$, принимающих значения в некотором пространстве $\mathcal{X}$, и произвольную функцию $f\colon \mathcal{X}^n \to \mathbb{R}$. Нас как и раньше интересует, насколько случайная величина $Z = f(\xi_1, \dots, \xi_n)$ сосредоточена вокруг своего математического ожидания $\mathbb{E}[Z]$.

Введем следующее удобное обозначение для условного математического ожидания:

$$\mathbb{E}_i[\cdot] = \mathbb{E}[\cdot|\xi_1, \dots, \xi_i].$$

Тогда $\mathbb{E}_0[Z] = \mathbb{E}[Z]$ и $\mathbb{E}_n[Z] = Z$. Обозначим

$$\Delta_i = \mathbb{E}_i[Z] - \mathbb{E}_{i-1}[Z].$$

Тогда мы можем представить $Z - \mathbb{E}[Z]$ в следующем виде:

$$Z - \mathbb{E}[Z] = \mathbb{E}_n[Z] - \mathbb{E}_{n-1}[Z] + \mathbb{E}_{n-1}[Z] - \mathbb{E}_{n-2}[Z] + \dots + \mathbb{E}_1[Z] - \mathbb{E}_0[Z] = \sum_{i=1}^n \Delta_i.$$

Подобная запись известная в литературе как *мартингал Дуба*. Рассмотрим особый случай, когда для всех i с вероятностью 1 выполнено $|\Delta_i| \leqslant c_i$. В этом случае несложно заметить, что

$$\begin{aligned} \mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] &= \mathbb{E}\left[e^{\lambda(\sum_{i=1}^n \Delta_i)}\right] = \mathbb{E}\left[\mathbb{E}_{n-1}\left[e^{\lambda(\sum_{i=1}^n \Delta_i)}\right]\right] \\ &= \mathbb{E}\left[e^{\lambda(\sum_{i=1}^{n-1} \Delta_i)}\mathbb{E}_{n-1}\left[e^{\lambda\Delta_n}\right]\right] \leqslant \mathbb{E}\left[e^{\lambda(\sum_{i=1}^{n-1} \Delta_i)}e^{\lambda^2 c_n^2/2}\right] \\ &= e^{\lambda^2 c_n^2/2}\mathbb{E}\left[e^{\lambda(\sum_{i=1}^{n-1} \Delta_i)}\right] \leqslant \dots \leqslant e^{\lambda^2\left(\sum_{i=1}^n c_i^2\right)/2}, \end{aligned} \tag{1.15}$$

где мы последовательно применяем Лемму Хефдинга 2, пока не избавимся от всех математических ожиданий. Мы доказали следующую теорему:

Теорема 8 (Неравенство Хефдинга–Азумы)**.** *Пусть $\xi_1, \dots, \xi_n$ — последовательность независимых случайных величин, принимающих значения в некотором множестве $\mathcal{X}$, и пусть $Z = f(\xi_1, \dots, \xi_n)$ для некоторой функции $f\colon \mathcal{X}^n \to \mathbb{R}$. Пусть $|\Delta_i| \leqslant c_i,\ i = 1, \dots, n$ с вероятностью 1. Тогда для любого $\lambda > 0$ справедливо:*

$$\mathbb{E}\left[e^{\lambda(Z - \mathbb{E}[Z])}\right] \leqslant e^{\lambda^2\left(\sum_{i=1}^n c_i^2\right)/2}. \tag{1.16}$$

Кроме того для любого $\epsilon \geq 0$:

$$\mathbb{P}\{Z - \mathbb{E}[Z] \geqslant \epsilon\} \leqslant \exp\left\{-\frac{\epsilon^2}{2\sum_{i=1}^n c_i^2}\right\}. \tag{1.17}$$

Отметим, что неравенство концентрации (1.17) получается из верхней оценки на производящую функцию (1.16) применением метода Чернова (Теорема 2), описанного ранее. В данном случае нам достаточно выбрать $\lambda = \frac{\epsilon}{\sum_{i=1}^n c_i^2}$.

Неравенство Хефдинга-Азумы является чрезвычайно простым и в то же время мощным инструментом изучения концентрации для целого класса функций f. Сейчас мы рассмотрим один из примеров его применения, ведущий к одному из наиболее широко используемых неравенств концентрации — неравенству *ограниченных разностей* или неравенству МакДиармида. Введем следующее определение.

Определение 1 (Функция с ограниченными разностями)**.** *Мы будем говорить, что функция $f\colon \mathcal{X}^n \to \mathbb{R}$ удовлетворяет* условию ограниченных разностей*, если существуют такие числа $c_1, \dots, c_n \in \mathbb{R}^+$, что для всех $i = 1, \dots, n$ выполнено следующее:*

$$\sup_{x_1,\dots,x_n,x_i' \in \mathcal{X}} |f(x_1, \dots, x_n) - f(x_1, \dots, x_{i-1}, x_i', x_{i+1} \dots, x_n)| \leqslant c_i. \tag{1.18}$$

Грубо говоря, данное условие говорит нам о том, что значение функции не сильно меняется при изменении одного из ее аргументов. Можно интерпретировать его и по-другому: значение функции не зависит слишком сильно ни от одного из ее аргументов.

Заметим, что справедливо следующее:

$$\begin{aligned} a_i &= \inf_{\xi_i} \mathbb{E}[Z|\xi_1, \dots, \xi_i] - \mathbb{E}[Z|\xi_1, \dots, \xi_{i-1}] \\ &\leqslant \Delta_i = \mathbb{E}[Z|\xi_1, \dots, \xi_i] - \mathbb{E}[Z|\xi_1, \dots, \xi_{i-1}] \\ &\leqslant \sup_{\xi_i} \mathbb{E}[Z|\xi_1, \dots, \xi_i] - \mathbb{E}[Z|\xi_1, \dots, \xi_{i-1}] = b_i, \end{aligned} \tag{1.19}$$

для $i = 1, \dots, n$. Кроме того, поскольку случайные величины $\xi_1, \dots, \xi_n$ независимы, для их *независимых копий* $x_1, \dots, x_n$ справедливо следующее:

$$\begin{aligned} \mathbb{E}[Z|\xi_1, \dots, \xi_i] &= \mathbb{E}\left[f(\xi_1, \dots, \xi_n)|\xi_1, \dots, \xi_i\right] \\ &= \int f(\xi_1, \dots, \xi_i, x_{i+1}, \dots, x_n) dP(x_{i+1}, \dots, x_n), \quad i = 1, \dots, n. \end{aligned}$$

Таким образом, с учетом (1.19) мы получаем:

$$\begin{aligned} b_i - a_i &= \sup_{\xi_i} \mathbb{E}[Z|\xi_1, \ldots, \xi_i] - \inf_{\xi_i} \mathbb{E}[Z|\xi_1, \ldots, \xi_i] \\ &= \sup_{x,y} \int \Big(f(\xi_1, \ldots, x, x_{i+1}, \ldots, x_n) - f(\xi_1, \ldots, y, x_{i+1}, \ldots, x_n) \Big) dP(x_{i+1}, \ldots, x_n) \\ &\leq c_i, \end{aligned}$$

где в последнем неравенстве мы воспользовались определением ограниченных разностей. Возвращаясь к цепочке неравенств (1.15) и применив Лемму Хефдинга 2, мы получим:

$$\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] \leq e^{\lambda^2\left(\sum_{i=1}^n c_i^2\right)/8},$$

что после последующего применения метода Чернова ведет к следующей теореме:

Теорема 9 (Неравенство МакДиармида)**.** *Пусть функция f удовлетворяет условию ограниченных разностей с константами $c_1, \ldots, c_n$. Пусть, кроме того, $\xi_1, \ldots, \xi_n$ — последовательность независимых случайных величин. Тогда для случайной величины $Z = f(\xi_1, \ldots, \xi_n)$ и любого $\lambda \geq 0$ справедливо:*

$$\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] \leq e^{\lambda^2\left(\sum_{i=1}^n c_i^2\right)/8}.$$

Также для любого $\epsilon \geq 0$ справедливо:

$$\mathbb{P}\{Z - \mathbb{E}[Z] \geqslant \epsilon\} \leqslant \exp\left\{-2\frac{\epsilon^2}{\sum_{i=1}^n c_i^2}\right\}. \tag{1.20}$$

Аналогичное неравенство справедливо для $\mathbb{P}\{\mathbb{E}[Z] - Z \geqslant \epsilon\}$, поскольку условия теоремы инвариантны относительно замены знака функции f.

Неравенство МакДиармида, как мы убедимся далее, является чрезвычайно простым в применении: для этого достаточно проверить свойство 1.18, что во многих случаях не составляет труда.

Примером функции, удовлетворяющей условию ограниченных разностей 1.18, является сумма $f(\xi_1, \ldots, \xi_n) = \sum_{i=1}^n \xi_i$ ограниченных слагаемых $\xi_i \in [0, 1]$. Неравенство МакДиармида в том случае в точности воспроизводит неравенство Хефдинга Теоремы 4. Оба этих неравенства учитывают исключительно ограниченность рассматриваемых случайных величин. Можно считать, что неравенство МакДиармида является обобщением неравенства Хефдинга для более общего класса функций f.

Учитывая упомянутую аналогию неравенства МакДиармида с неравенством Хефдинга, встает вопрос: нельзя ли получить экспоненциальные неравенства для достаточно общего

класса функций f, учитывающие дисперсию случайной величины $f(\xi_1, \ldots, \xi_n)$? Такие неравенства, возможно, уточняли бы оценки неравенства МакДиармида подобно тому, как неравенства Бернштейна и Беннета уточняют результаты неравенства Хефдинга. Оказывается, такие неравенства получить можно. Данной теме посвящен следующий раздел.

1.3 Энтропийный метод Леду

Одним из наиболее плодотворных подходов в теории концентрации меры последних лет является так называемый *индуктивный метод*, развитый М. Талаграном в серии работ [93, 94]. В частности, на основе индуктивного метода М. Талаграну удалось получить функциональное обобщение неравенства Беннетта.

Несмотря на все достоинства индуктивного метода, его активное применение ограничено одним его общепризнанным недостатком — высоким уровнем сложности и громоздкостью его выкладок. Относительно недавно рядом авторов во главе с М. Леду был развит альтернативный подход, дающий сравнимые с индуктивным методом результаты, имеющие при этом гораздо более простые и короткие доказательства. Этот подход получил название «энтропийный метод», и в настоящем разделе мы вкратце рассмотрим его некоторые его результаты, которые понадобятся нам в дальнейшем.

Пусть $\boldsymbol{X} = (\xi_1, \ldots, \xi_n)$ — последовательность независимых случайных величин, принимающих значения в некотором множестве $\mathcal{X}$. Для $k = 1, \ldots, n$ и $x \in \mathcal{X}$ обозначим с помощью $\boldsymbol{X}_{k,x}$ последовательность, полученную из $\boldsymbol{X}$ заменой ξ_k на x. Введем следующее определение:

Определение 2 (Слабо самоограничивающая функция)**.** *Мы будем говорить, что функция $f\colon \mathcal{X}^n \to \mathbb{R}$ является* слабо самоограничивающей*, если существует $a \geq 1$, такое что выполнены следующие условия:*

$$\begin{aligned} &f(\boldsymbol{X}) - \inf_{x \in \mathcal{X}} f(\boldsymbol{X}_{k,x}) \leq 1, \quad k = 1, \ldots, n; \\ &\sum_{k=1}^{n} \left(f(\boldsymbol{X}) - \inf_{x \in \mathcal{X}} f(\boldsymbol{X}_{k,x}) \right)^2 \leq a f(\boldsymbol{X}). \end{aligned} \tag{1.21}$$

Примером функции, удовлетворяющей данным условиям с $a = 1$, является сумма $\sum_{i=1}^{n}$ ограниченных случайных величин $\xi_i \in [0, 1]$, $i = 1, \ldots, n$.

Оказывается, этих условий достаточно для получения нетривиальных неравенств концентрации. Следующий результат представлен в [70, Теорема 13].

Теорема 10 (А. Маурер, [70])**.** *Пусть f — слабо самоограничивающая функция с параметром $a \geq 1$. Пусть, кроме того, $\xi_1, \ldots, \xi_n$ — последовательность независимых случайных*

величин. Тогда для случайной величины $Z = f(\xi_1, \ldots, \xi_n)$ *и любого* $\lambda \geq 0$ *справедливо:*

$$\mathbb{E}\left[e^{\lambda(\mathbb{E}[Z]-Z)}\right] \leq \exp\left(\lambda^2 \frac{a\mathbb{E}[Z]}{2}\right). \tag{1.22}$$

Кроме того, для любого $\lambda \in [0, 2/a]$ *справедливо:*

$$\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] \leq \exp\left(\frac{\lambda^2}{1-a\lambda/2}\frac{a\mathbb{E}[Z]}{2}\right).$$

Метод Чернова ведет к следующей оценке на левый хвост случайной величины Z*, справедливой для любого* $\epsilon \geq 0$*:*

$$\mathsf{P}\left\{\mathbb{E}[Z] - Z \geq \epsilon\right\} \leq \exp\left(-\frac{\epsilon^2}{2a\mathbb{E}[Z]}\right).$$

Также для любого $t \geq 0$ *с вероятностью не меньше* $1 - e^{-t}$ *справедливо:*

$$\mathbb{E}[Z] \leq Z + \sqrt{2a\mathbb{E}[Z]t}.$$

Замечание 3. *Последнее неравенство получено последовательным применением метода Чернова с использованием оценки* (1.22) *и Леммы 1. Подобным же образом на основе второго неравенства теоремы может быть получена верхняя оценка для* $Z - \mathbb{E}[Z]$*.*

В отличие от неравенства МакДиармида условия последней теоремы не инвариантны относительно замены знака функции f. Это ведет к различным оценкам на правый и левые хвосты, получаемым из приведенных неравенств с помощью метода Чернова. Кроме того, условия (1.21) интерпретировать сложнее, чем условия ограниченных разностей. Здесь мы приведем лишь их краткое обсуждение. Оказывается, сумма квадратов, фигурирующая во втором неравенстве условий (1.21), непосредственным образом связана с дисперсией случайной величины $f(\boldsymbol{X})$. Пользуясь результатами Раздела 3.1 [19] (неравенством Эфрона-Стайна) можно показать, что справедливо следующее:

$$\mathbb{D}[f(\boldsymbol{X})] \leq \mathbb{E}\left[\sum_{k=1}^{n}\left(f(\boldsymbol{X}) - \inf_{x\in\mathcal{X}} f(\boldsymbol{X}_{k,x})\right)^2\right].$$

Таким образом, условия (1.21) влекут за собой требование $\mathbb{D}[f(\boldsymbol{X})] \leq a\mathbb{E}[f(\boldsymbol{X})]$. Вновь сравним верхнюю оценку (1.22) с производящей функцией моментов нормальной случайной величины (1.3). Мы видим, что верхняя оценка А. Маурера (1.22) дает тот же результат, с точностью до замены дисперсии $\mathbb{D}[f(\boldsymbol{X})]$ на ее верхнюю оценку $a\mathbb{E}[f(\boldsymbol{X})]$. Подобное свойство является типичным для результатов энтропийного подхода.

1.3.1 Эмпирическое неравенство Бернштейна

В Разделе 1.1.3 мы рассмотрели неравенства Бернштейна и Беннетта, ограничивающие отклонение суммы независимых и ограниченных сверху случайных величин от их средних значений. Было показано, что за счет учета дисперсии случайных величин эти результаты дают гораздо более точные оценки, чем неравенство Хефдинга Теоремы 4. Однако, для применения этих неравенств нам необходимо знать дисперсию случайных величин, которая в большинстве случаев неизвестна. В ряде случаев мы можем заменить дисперсию на ее неслучайную (не зависящую от случайных выборок) верхнюю оценку, однако такая оценка тоже не всегда доступна, а также она может быть сильно завышена. В данном разделе, пользуясь неравенствами МакДиармида и А. Маурера, мы приведем ряд результатов, позволяющих ограничить дисперсию случайной величины с помощью ее выборочной оценки.

Напомним, что для независимых и одинаково распределенных случайных величин $\xi_1, \dots, \xi_n$ *несмещенной выборочной дисперсией* называется следующая случайная величина:

$$\mathbb{D}_n = \mathbb{D}_n(\xi_1, \dots, \xi_n) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\xi_i - \frac{1}{n} \sum_{j=1}^{n} \xi_j \right)^2 .$$

Для случайной величины $\mathbb{D}_n$ справедливо тождество:

$$\mathbb{E}[\mathbb{D}_n] = \mathbb{D}[\xi_1].$$

Возникает вопрос: возможно ли построить верхнюю оценку для дисперсии $\mathbb{D}[\xi_1]$, основанную на $\mathbb{D}_n$ и справедливую с большой вероятностью? Мы приходим к первому примеру практического применения приведенных ранее неравенств концентрации. Действительно, неравенства концентрации дают нам возможность контролировать отклонение значений случайных величин от их математических ожиданий. В данном случае в качестве случайной величины выступает $\mathbb{D}_n$, и мы будем контролировать ее отклонение от $\mathbb{D}[\xi_1]$.

Первый подход: Начнем с более простого подхода: применим неравенство МакДиармида для случайной величины $\mathbb{D}_n$. Для этого нам надо убедиться, что условия ограниченных разностей (1.18) выполнены для функции $\mathbb{D}_n \colon \mathbb{R}^n \to \mathbb{R}$. Нам потребуется следующее вспомогательное утверждение:

Лемма 4. *Для любой конечной последовательности действительных чисел* $\{x_1, \dots, x_n\}$ *выполнено следующее:*

$$\frac{1}{n(n-1)} \sum_{1 \le i < j \le n} (x_i - x_j)^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \frac{1}{n} \sum_{j=1}^{n} x_j \right)^2 .$$

Доказательство.

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\frac{1}{n}\sum_{j=1}^{n}x_j\right)^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2-\frac{2}{n}x_i\sum_{j=1}^{n}x_j+\frac{1}{n^2}\left(\sum_{j=1}^{n}x_j\right)^2\right)=$$

$$=\frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-\frac{2}{n}\sum_{i=1}^{n}x_i\sum_{j=1}^{n}x_j+\frac{1}{n^2}\sum_{i=1}^{n}\left(\sum_{j=1}^{n}x_j\right)^2\right)=\frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{j=1}^{n}x_j\right)^2\right)=$$

$$=\frac{1}{n(n-1)}\left((n-1)\sum_{i=1}^{n}x_i^2-2\sum_{1\le i<j\le n}x_ix_j\right)=\frac{1}{n(n-1)}\sum_{1\le i<j\le n}(x_i-x_j)^2.$$

■

Лемма 5. *Несмещенная выборочная дисперсия $\mathbb{D}_n(\xi_1,\dots,\xi_n)$ для ограниченных случайных величин $\xi_i\in[0,1]$, $i=1,\dots,n$, удовлетворяет условию ограниченных разностей с параметрами $c_i=\frac{1}{n}$.*

Доказательство. Из Леммы 4, следует, что справедливо следующее тождество:

$$\mathbb{D}_n=\frac{1}{n(n-1)}\sum_{1\le i<j\le n}(\xi_i-\xi_j)^2.$$

Таким образом, для $x,y\in[0,1]$ справедливо следующее:

$$\begin{aligned}&\left|\mathbb{D}_n(\xi_1,\dots,\xi_{i-1},x,\xi_{i+1},\dots,\xi_n)-\mathbb{D}_n(\xi_1,\dots,\xi_{i-1},y,\xi_{i+1},\dots,\xi_n)\right|\\&=\left|\frac{1}{n(n-1)}\left(\sum_{j\neq i}(x-\xi_j)^2-\sum_{j\neq i}(y-\xi_j)^2\right)\right|\\&=\left|\frac{1}{n(n-1)}\sum_{j\neq i}\left((x-\xi_j)^2-(y-\xi_j)^2\right)\right|\\&\le\frac{1}{n}.\end{aligned}$$

■

Применяя неравенство МакДиармида Теоремы 9 для случайной величины $\mathbb{D}_n$ мы получаем следующее следствие:

Следствие 2. *Рассмотрим независимые одинаково распределенные случайные величины $\xi_1,\dots,\xi_n$, ограниченные в интервале $[0,1]$. Тогда для любого $t\ge0$ с вероятностью не меньше $1-e^{-t}$ справедливо следующее:*

$$\mathbb{D}[\xi_1]\le\mathbb{D}_n+\sqrt{\frac{t}{2n}}.$$

Также для любого $t\ge0$ с вероятностью не меньше $1-e^{-t}$ справедливо:

$$\mathbb{D}_n\le\mathbb{D}[\xi_1]+\sqrt{\frac{t}{2n}}.$$

Обратим внимание, что оценки последнего следствия в точности совпадают с неравенством Хефдинга (1.11).

Второй подход: Теперь мы применим неравенство А. Маурера Теоремы 10. Следующий результат представлен в [71, Теорема 10]:

Лемма 6 (А. Маурер и М. Понтил, [71]). *Функция $n \cdot \mathbb{D}_n(\xi_1, \dots, \xi_n)$ для ограниченных случайных величин $\xi_i \in [0,1]$, $i = 1, \dots, n$, является слабо самоограничивающей (удовлетворяет условиям (1.21)) с параметром $a = \frac{n}{n-1}$.*

Теперь мы можем применить неравенство А. Маурера Теоремы 10 к случайной величине $n \cdot \mathbb{D}_n$ и получить следующее следствие:

Следствие 3 (А. Маурер и М. Понтил, [71]). *Рассмотрим независимые одинаково распределенные случайные величины $\xi_1, \dots, \xi_n$, ограниченные в интервале $[0,1]$. Тогда для любого $t \geq 0$ с вероятностью не меньше $1 - e^{-t}$ справедливо:*

$$\sqrt{\mathbb{D}[\xi_1]} \leq \sqrt{\mathbb{D}_n} + \sqrt{\frac{2t}{n-1}}. \tag{1.23}$$

Кроме того, для любого $t \geq 0$ с вероятностью не меньше $1 - e^{-t}$ справедливо:

$$\mathbb{D}[\xi_1] \leq \mathbb{D}_n + 2\sqrt{\frac{\mathbb{D}_n t}{n-1}} + \frac{2t}{n-1}. \tag{1.24}$$

Доказательство. Применив неравенство А. Маурера мы получаем, что для любого $t \geq 0$ с вероятностью не меньше $1 - e^{-t}$ справедливо следующее:

$$\mathbb{D}[\xi_1] \leq \mathbb{D}_n + \sqrt{\frac{2\mathbb{D}[\xi_1] t}{n-1}}.$$

Применив неравенство $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, мы получаем первое неравенство следствия. ■

Сравивая неравенство (1.24) с оценками Следствия 2, мы приходим к выводу, что последнее следствие может давать существенно лучшие результаты при малых значениях $\mathbb{D}_n$. В противном случае оба следствия дают сравнимые результаты.

Далее нам потребуются следующий результат, приведенный в [21, Раздел 2.7.2]:

Лемма 7. *Пусть для случайной величины ξ, констант $A, B > 0$ и любого $\lambda \in [0, 1/B)$ выполнено:*

$$\mathbb{E}\left[e^{\lambda \xi}\right] \leq \exp\left(\frac{A\lambda^2}{2(1 - B\lambda)}\right).$$

Тогда для любого $t \geq 0$ с вероятностью не меньше $1 - e^{-t}$ справедливо:

$$\xi \leq \mathbb{E}[\xi] + \sqrt{2At} + Bt.$$

Воспользовавшись последним следствиями и неравенствами Бернштейна и Беннетта мы получаем следующий результат:

Теорема 11 (Эмпирическое неравенство Бернштейна, [71]). *Рассмотрим независимые одинаково распределенные случайные величины $\xi_1, \dots, \xi_n$, ограниченные в интервале $[0, 1]$. Обозначим $S_n = \sum_{i=1}^n \xi_i$. Тогда для любого $t \geq 0$ с вероятностью не меньше $1-2e^{-t}$ справедливо:*

$$\frac{1}{n}S_n \leq \frac{1}{n}\mathbb{E}[S_n] + \sqrt{\frac{2\mathbb{D}_n t}{n}} + \frac{7t}{3(n-1)}.$$

Кроме того для любого $t \geq 0$ с вероятностью не меньше $1 - 2e^{-t}$ справедливо:

$$\frac{1}{n}\mathbb{E}[S_n] \leq \frac{1}{n}S_n + \sqrt{\frac{2\mathbb{D}_n t}{n}} + \frac{7t}{3(n-1)}.$$

Доказательство. Воспользовавшись Теоремой 5, мы имеем:

$$\mathbb{E}\left[e^{\lambda(S_n - \mathbb{E}[S_n])}\right] \leq e^{(e^\lambda - \lambda - 1)n\sigma^2}.$$

Кроме того, для $\lambda \in [0, 1/3)$ справедливо $e^\lambda - \lambda - 1 \leq \frac{\lambda^2}{2(1-\lambda/3)}$, в чем легко убедиться, раскладывая функции в ряд Тейлора. Таким образом, для $\lambda \in [0, 1/3)$ справедливо следующее:

$$\mathbb{E}\left[e^{\lambda(S_n - \mathbb{E}[S_n])}\right] \leq \exp\left(\frac{n\sigma^2\lambda^2}{2(1-\lambda/3)}\right).$$

Из этой верхней оценки на производящую функцию моментов Лемма 7, приведенная в секции дополнительных материалов, позволяет заключить, что для любого $t \geq 0$ с вероятностью не меньше $1 - e^{-t}$ справедливо[2]:

$$S_n - \mathbb{E}[S_n] \leq \sqrt{2n\mathbb{D}[\xi_1]t} + \frac{t}{3}.$$

С той же вероятностью выполнено неравенство (1.23). Воспользовавшись неравенством Буля мы получаем, что оба неравенства выполнены одновременно с вероятностью не меньше $1 - 2e^{-t}$:

$$\begin{aligned} S_n - \mathbb{E}[S_n] &\leq \sqrt{2nt}\left(\sqrt{\mathbb{D}_n} + \sqrt{\frac{2t}{n-1}}\right) + \frac{t}{3} \\ &\leq \sqrt{2n\mathbb{D}_n t} + t\left(\frac{2n}{n-1} + \frac{n}{3(n-1)}\right) \\ &= \sqrt{2n\mathbb{D}_n t} + \frac{7nt}{3(n-1)}. \end{aligned}$$

Все рассуждения доказательства можно повторить для случайной величины $\mathbb{E}[S_n] - S_n$, что даст нам второе неравенство теоремы. ■

[2] Заметим, что мы избавились от лишнего множителя в последнем слагаемом по сравнению с неравенством Бернштейна.

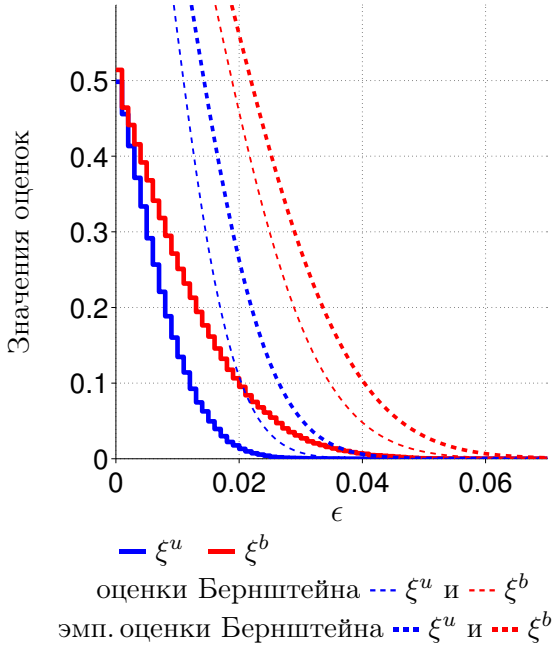


Рисунок 1.3: Эмпирические оценки $\mathbb{P}\{\xi - \mathbb{E}[\xi] \geq \epsilon\}$ для ξ^b и ξ^u (жирные линии) вместе с оценками неравенства Бернштейна (тонкие пунктирные) и эмпирического неравенства Бернштейна (толстые пунктирные линии).

В отличие от неравенств Беннетта и Бернштейна, включающих в себя неизвестную дисперсию $\mathbb{D}[\xi_1]$ случайных величин, все величины, входящие в эмпирическое неравенство Бернштейна, могут быть вычислены на основе выборки. На Рисунке 1.3 представлены оценки неравенства Бернштейна вместе с оценками эмпирического неравенства Бернштейна для случайных величин ξ^u и ξ^b (на этот раз для $n = 1000$). По мере роста размера выборки n эмпирические оценки дисперсий будут сходиться к самим дисперсиям и разница между двумя оценками исчезает.

1.3.2 Неравенство Буске для эмпирических процессов

В этом разделе мы рассмотрим еще одно применение энтропийного метода. Рассмотрим выборку независимых выборку независимых и одинаково распределенных случайных величин $\xi_1, \dots, \xi_n$, принимающих значения в некотором множестве $\mathcal{X}$. Рассмотрим *счетное* множество отображений $\mathcal{F} = \{f \colon \mathcal{X} \to \mathbb{R}\}$, таких что $\mathbb{E}[f(\xi_1)] = 0$ и $f(x) \in [-1, 1]$ для всех $f \in \mathcal{F}$ и $x \in \mathcal{X}$. *Супремумом эмпирического процесса*[3] называется следующая случайная

[3] В настоящей работе мы будем предполагать, что все величины, с которыми мы работаем, измеримы. Вопрос об измеримости случайной величины Q_n обсуждается, например, в начале Главы 2 работы [50].

величина:

$$Q_n = \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f(\xi_i).$$

Сейчас мы не будем приводить примеры задач, в которых естественным образом возникает случайная величина Q_n. Однако, в следующих главах мы убедимся, что она играет центральную роль в теории статистического обучения. В настоящем разделе мы приведем два неравенства концентрации для Q_m.

Первый результат основан на простом применении неравенства МакДиармида Теоремы 9.

Теорема 12. *Для любого $\epsilon \geq 0$ справедливо:*

$$\mathsf{P}\{Q_n - \mathbb{E}[Q_n] \geq \epsilon\} \leq \exp\left(-\frac{\epsilon^2}{2n}\right).$$

Аналогичное неравенство справедливо для $\mathsf{P}\{\mathbb{E}[Q_n] - Q_n \geq \epsilon\}$.

Доказательство. Нам остаточно заметить, что отображение $f(\xi_1, \ldots, \xi_n) = Q_n$ удовлетворяет условию ограниченных разностей с параметрами $c_i = 2$. ■

Этот результат можно интерпретировать следующим образом: он контролирует отклонение средних выборочных от математических ожиданий равномерно по счетному классу выборок. Поэтому мы будем говорить, что он является равномерным по классу функций аналогом неравенства Хефдинга.

Вопрос возможности получения равномерного аналога неравенства Беннетта является более сложным. Положительный ответ на него был дан в работе [94] М. Талаграна лишь в 1996 году, где автору удалось получить его с помощью индуктивного метода. Неравенство Талаграна позже было усилено О. Буске и рядом других авторов в [19, 21] на основе энтропийного метода. Помимо получения более точных констант, авторам удалось существенно упростить доказательство. Здесь мы приведем без доказательства версию О. Буске с оптимальными константами.

Теорема 13 (О. Буске, [21][4])**.** *Положим $\sigma^2 = \sup_{f \in \mathcal{F}} \mathbb{D}[f(X_1)]$, $v = n\sigma^2 + 2\mathbb{E}[Q_n]$ и для $u \geq -1$ определим $\phi(u) = e^u - u - 1$, $h(u) = (1+u)\log(1+u) - u$. Тогда для $\lambda \geq 0$ справедливо:*

$$\mathbb{E}\left[e^{\lambda(Q_n - \mathbb{E}[Q_n])}\right] \leq e^{v\phi(\lambda)}. \tag{1.25}$$

Метод Чернова дает для любых $\epsilon \geq 0$ следующее неравенство:

$$\mathsf{P}\left\{Q_n - \mathbb{E}[Q_n] \geq \epsilon\right\} \leq e^{-vh(\epsilon/v)}. \tag{1.26}$$

Кроме того, для любого $t \geq 0$ *с вероятностью не меньше* $1 - e^{-t}$ *справедливо:*

$$Q_n \leq \mathbb{E}[Q_n] + \sqrt{2vt} + \frac{t}{3}. \tag{1.27}$$

Снова воспользовавшись неравенством $h(u) \geq \frac{u^2}{2(1+u/3)}$ для $u > 0$, мы можем получить следующий более наглядный результат:

$$\mathsf{P}\left\{Q_n - \mathbb{E}[Q_n] \geq \epsilon\right\} \leq \exp\left(-\frac{\epsilon^2}{2(v + \epsilon/3)}\right). \tag{1.28}$$

Важно отметить, что если класс $\mathcal{F} = \{f_0\}$ состоит из одного элемента, неравенство (1.28) в точности воспроизводит неравенство Бернштейна Теоремы 6 для суммы независимых случайных величин, а неравенство (1.25) — неравенство Беннетта. Таким образом, данный результат является равномерным по классу функций $\mathcal{F}$ аналогом неравенств Беннетта и Бернштейна.

Замечание 4. *Отметим, что результаты настоящего раздела могут быть без труда обобщены на случай* несчетного *множества* $\mathcal{F}$ *в том случае, когда эмпирический процесс* сепарабелен, *то есть множество* $\mathcal{F}$ *содержит счетное и всюду плотное подмножество. Подробности могут быть найдены в [19, стр. 314] или [21, стр. 72].*

Коротко подведем итоги настоящей главы:

- Неравенства концентрации дают возможность контролировать флуктуации значений $Q = g(X_1, \ldots, X_n)$ вокруг их математических ожиданий $\mathbb{E}[Q]$, обусловленные случайностью аргументов $X_1, \ldots, X_n$. Особенно много результатов известно для случая, когда случайные величины $X_1, \ldots, X_n$ *независимы*.

- Суммы $Q = \frac{1}{n}\sum_{i=1}^{n} X_i$ являются наиболее хорошо изученными объектами. Типичная величина их отклонения имеет порядок $n^{-1/2}$, но может быть меньше, если дисперсии $\mathbb{D}[X_i]$ малы.

- Похожими свойствами обладают функции, не зависящие слишком сильно ни от одного из своих аргументов. Такие функции называются функциями с ограниченными разностями и для них справедливо чрезвычайно простое в применении неравенство МакДиармида.

[4] Далее мы будем называть Теорему 13 попеременно неравенством Талаграна или неравенством Буске.

- Супремумы эмпирических процессов $Q = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(X_i)$ являются примером функций с ограниченными разностями, откуда мы немедленно получаем оценки порядка $n^{-1/2}$ для них. Однако, более сильные результаты (неравенство Талаграна) показывают, что порядок может быть меньше, если дисперсии $\mathbb{D}[f(X_i)]$ малы.

- Теорема 10 является чрезвычайно мощным средством получения новых неравенств концентрации. Для ее применения достаточно проверить сравнительно несложные условия, накладываемые на функцию g.

2 Неравенства концентрации для выборок без возвращений

В прошлом разделе мы подробно рассмотрели неравенства концентрации, справедливые для выборок независимых случайных величин. Однако, часто предположение о независимости может оказаться необоснованным. В таких случаях рассматривают более общие модели случайных последовательностей: мартингалы [76], последовательности с перемешиванием [112] и другие. В данном разделе мы рассмотрим конкретную модель зависимых случайных величин, которая, тем не менее, окажется чрезвычайно полезной в дальнейших главах. Мы рассмотрим неравенства концентрации, контролирующие отклонение значений функций $f(\eta_1, \dots, \eta_n)$ от математических ожиданий $\mathbb{E}[f(\eta_1, \dots, \eta_n)]$, где случайные величины $\eta_1, \dots, \eta_n$ выбраны *без возвращений* из конечного множества мощностью $N > n$.

Примером задач, в которых естественным образом возникают выборки без возвращений, является получение оценок надёжности восстановления зависимостей по конечным выборкам данных. Первые оценки равномерного уклонения частот ошибок в двух подвыборках были получены в [106,113]. Позже эти идеи получили развитие в трансдуктивном подходе теории статистического обучения [23,104] и в комбинаторной теории переобучения [109–111,116]. Еще одним интересным примером можно считать теоретический анализ процедуры кросс-валидации.

Данный раздел состоит из трех частей. В Разделе 2.1 будут рассмотрены известные результаты для сумм случайных величин, выбранных без возвращений. На этом примере будет наглядно продемонстрировано, что выборки без возвращений ведут к более сильной концентрации по сравнению с выборками независимых случайных величин. В Разделе 2.2 будет приведено два результата, справедливых для более широкого класса функций, определенных на случайных разбиениях конечного множества действительных чисел. Первый, полученный Р. Эль-Янивом, Д. Печиони и другими в [25, 34], является аналогом неравенства МакДиармида и усиливает его при росте размера выборки n к размеру генеральной совокупности N. Второй получен С. Г. Бобковым в [17] и устанавливает субгауссовское поведение

таких функций, позволяя, в отличие от неравенства МакДиармида, учитывать дисперсии случайных величин. Наконец, в Разделе 2.3 на основе описанных подходов будет получено два новых неравенства концентрации для супремумов эмпирических процессов для выборок без возвращений, одно из которых является непосредственным обобщением неравенства Талаграна, представленного в Разделе 1.3.2.

2.1 Суммы случайных величин

В том случае, когда случайные величины независимы и ограничены, неравенства Хефдинга, Бернштейна и Беннета, представленные в Разделе 1.1, дают оптимальные неасимптотические оценки отклонения сумм от их математических ожиданий. В этом разделе будут приведены известные результаты для случая, когда слагаемые выбраны без возвращений из конечного множества действительных чисел.

Рассмотрим конечное множество $\mathcal{C} = \{c_1, \dots, c_N\}$ ограниченных действительных чисел $c_i \in [0,1]$, $i = 1 \dots, N$, с возможными повторениями. Для произвольного натурального числа $n \leqslant N$ всюду далее с помощью $\{\eta_1, \dots, \eta_n\}$ и $\{\xi_1, \dots, \xi_n\}$ будем обозначать последовательности случайных величин, выбранных равномерно из $\mathcal{C}$ без возвращений и с возвращениями соответственно. Еще раз отметим, что случайные величины $\xi_1, \dots, \xi_n$ являются независимыми, в то время как $\eta_1, \dots, \eta_n$ таковыми не являются. Введем обозначения $S'_n = \frac{1}{n}\sum_{i=1}^n \eta_i$ и $S_n = \frac{1}{n}\sum_{i=1}^n \xi_i$.

Для получения неравенств концентрации для S'_n в литературе принято использовать метод Чернова, описанный в прошлом разделе. Для получения верхних оценок на производящие функции моментов существует два подхода, описанных далее. Первый основан на классическом результате В. Хефдинга, который позволяет свести анализ схемы выборки без возвращений к выборке с возвращениями. Второй, впервые примененный в такой постановке Серфлингом в [89], использует метод мартингалов для непосредственной оценки производящей функции моментов.

2.1.1 Метод Хефдинга

Первый подход основан на слеующем результате В. Хефдинга:

Теорема 14 (В. Хефдинг, [43]). [1] *Пусть* $\{U_1, \dots, U_n\}$ *и* $\{W_1, \dots, W_n\}$ *выбраны равномерно из конечного множества* d*-мерных векторов* $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_N\} \subset \mathbb{R}^d$ *с и без возвращений соответственно. Тогда для любой выпуклой функции* $f \colon \mathbb{R}^d \to \mathbb{R}$ *справедливо:*

$$\mathbb{E}\left[f\left(\sum_{i=1}^{n} W_i\right)\right] \leqslant \mathbb{E}\left[f\left(\sum_{i=1}^{n} U_i\right)\right].$$

Как будет показано далее в работе, Теорема 14 является чрезвычайно удобным средством переноса результатов концентрации, справедливых для функций независимых случайных величин, на выборки без возвращений. Положив $f(x) = e^{\lambda x}$, мы немедленно получаем следующее неравенство:

Следствие 4. *Для любого* $\lambda \geqslant 0$ *справедливо:*

$$\mathbb{E}\Big[\exp\big(\lambda\left(S'_n - \mathbb{E}[S'_n]\right)\big)\Big] \leqslant \mathbb{E}\Big[\exp\big(\lambda\left(S_n - \mathbb{E}[S_n]\right)\big)\Big].$$

Доказательство. Положив в Теореме 14 в качестве $f(x) = \exp\left(\frac{\lambda}{n}x\right)$ для произвольной $\lambda > 0$, мы получим:

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^{n}\left(\eta_i - \mathbb{E}[\eta_i]\right)\right)\right] &= \mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^{n}\eta_i\right)\right]\exp\left(-\frac{\lambda}{n}\sum_{i=1}^{n}\mathbb{E}[\eta_i]\right) \\
&\leqslant \mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^{n}\xi_i\right)\right]\exp\left(-\frac{\lambda}{n}\sum_{i=1}^{n}\mathbb{E}[\eta_i]\right) \\
&= \mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^{n}\left(\xi_i - \mathbb{E}[\xi_i]\right)\right)\right].
\end{aligned}$$

■

Замечание 5. *В последнем доказательстве мы также использовали следующий очевидный факт, который мы приводим без доказательства:*

$$\mathbb{E}[S'_n] = \mathbb{E}[S_n] = \frac{1}{N}\sum_{i=1}^{N} c_i.$$

Применив метод Чернова для S'_n и ограничив производящую функцию моментов с помощью Следствия 4, мы приходим к выводу, что все неравенства концентрации для S_n, полученные с помощью метода Чернова, также справедливы для S'_n. В частности, справедливы следующие неравенства Хефдинга и Бернштейна:

[1] Хотя в статье В. Хефдинга в явном виде не указано, что результат справедлив для случайных векторов, все доказательства остаются справедливы и для этого случая. См., например, [40] Раздел D.

Теорема 15 (Н-во Хефдинга для выборок без возвращений)**.** *Пусть* $\bar{c} = \frac{1}{N}\sum_{i=1}^{N} c_i$*. Тогда для любого* $\varepsilon \geqslant 0$ *справедливо:*

$$\begin{aligned}\mathbb{P}\left\{S'_n - \mathbb{E}[S'_n] \geqslant \varepsilon\right\} &\leqslant \exp\bigl(-n\,\mathrm{kl}(\bar{c}+\varepsilon\|\bar{c})\bigr)\\ &\leqslant e^{-2n\varepsilon^2},\end{aligned}$$

где мы обозначили дивергенцию Кульбака-Лейблера между двумя распределениями Бернулли с параметрами $0 \leqslant p \leqslant 1$ *и* $0 \leqslant q \leqslant 1$ *с помощью* $\mathrm{kl}(p\|q) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}$*. Аналогичные неравенства справедливы для* $\mathbb{P}\left\{\mathbb{E}[S'_n] - S'_n \geqslant \varepsilon\right\}$*.*

Теорема 16 (Н-во Бернштейна для выборок без возвращений)**.** *Пусть* $\bar{c} = \frac{1}{N}\sum_{i=1}^{N} c_i$ *и*

$$\sigma_N^2 = \frac{1}{N}\sum_{i=1}^{N}(c_i - \bar{c})^2.$$

Обозначим $h(u) = (1+u)\log(1+u) - u$ *для* $u \geqslant 0$*. Тогда для любого* $\varepsilon \geqslant 0$ *справедливо:*

$$\begin{aligned}\mathbb{P}\left\{S'_n - \mathbb{E}[S'_n] \geqslant \varepsilon\right\} &\leqslant \exp\left(-n\,\sigma_N^2 h\left(\frac{\varepsilon}{\sigma^2}\right)\right)\\ &\leqslant \exp\left(-\frac{n\varepsilon^2}{2(\sigma_N^2 + \varepsilon/3)}\right).\end{aligned} \tag{2.1}$$

Аналогичные неравенства справедливы для $\mathbb{P}\left\{\mathbb{E}[S'_n] - S'_n \geqslant \varepsilon\right\}$*. Кроме того, для любого* $t > 0$ *с вероятностью не меньше* $1 - e^{-t}$ *выполнено:*

$$\mathbb{E}[S'_n] \leqslant S'_n + \sqrt{\frac{2\sigma_N^2 t}{n}} + \frac{2t}{3n}.$$

Обратим внимание, что дисперсии сумм S'_n и S_n, в отличие от их математических ожиданий, не совпадают. Справедливо следующее [43]:

$$\mathbb{D}[S'_n] = \frac{N-n}{N-1}\frac{\sigma_N^2}{n} = \frac{N-n}{N-1}\,\mathbb{D}[S_n],$$

где σ_N^2, определенная в Теореме 16, — дисперсия равномерно распределенной на $\mathcal{C}$ случайной величины. Таким образом, дисперсия S'_n убывает по сравнению с дисперсией S_n по мере роста n, пока не будут выбраны все $n = N$ элементов множества $\mathcal{C}$: в этом случае S'_n вырождается в константу и $\mathbb{D}[S'_n] = 0$. Этот факт (наряду с Теоремой 14) демонстрирует, что S'_n концентрируется сильнее S_n. Авторы [29] дают следующее интуитивное объяснение этого эффекта: последовательное уменьшение размера множества, из которого мы выбираем очередную случайную величину η_i без возвращения, ведет к уменьшению ее разброса по сравнению со случаем выборки с возвращениями.

2.1.2 Неравенство Серфлинга

Второй подход основан на непосредственной оценке производящей функции моментов $\mathbb{E}\left[e^{\lambda(\xi-\mathbb{E}[\xi])}\right]$ случайной величины S'_n без использования Теоремы 14. В работе [89] с помощью метода мартингалов получен следующий результат, который всегда точнее второй верхней оценки Теоремы 15:

Теорема 17 (Серфлинг [89]). *Обозначим* $\Delta(n) = \frac{n-1}{N}$. *Тогда для любого* $\varepsilon \geqslant 0$ *справедливо:*

$$\mathbb{P}\left\{S'_n - \mathbb{E}[S'_n] \geqslant \varepsilon\right\} \leqslant \exp\left(-\frac{2n\varepsilon^2}{1-\Delta(n)}\right).$$

Аналогичные неравенства справедливы для $\mathbb{P}\left\{\mathbb{E}[S'_n] - S'_n \geqslant \varepsilon\right\}$.

По мере роста $n \to N$ знаменатель показателя экспоненты убывает к $1/N$, и неравенство существенно уточняет Теорему 15. С другой стороны, при $n = o(N)$ — случай, когда схема выборки без возвращений приближается к схеме с возвращениями, — знаменатель стремится к 1 и неравенство совпадает с Теоремой 15.

Замечание 6. *Как мы отмечали ранее, при* $n = N$ *случайная величина* S'_n *вырождается в константу, что влечет за собой тождество* $\mathbb{P}\left\{S'_n - \mathbb{E}[S'_n] \geqslant \varepsilon\right\} = 0$ *для произвольного* $\varepsilon > 0$. *Однако, при* $n = N$ *в правой части Теоремы 17 мы получаем* $e^{-2nN\varepsilon^2} > 0$. *В [4, Утверждение 4] предложен результат, совпадающий с неравенством Теоремы 17 с точностью до замены* $1-(n-1)/N$ *в знаменателе показателя экспоненты на* $1-n/N$. *Кроме того, авторы [4] развивают подход, предложенный Серфлингом, и предлагают неравенства «бернштейновского» типа для выборок без возвращений, которые, подобно тому как Теорема 17 улучшает Теорему 15, уточняют неравенство Теоремы 16 при* $n \to N$.

На примере сумм мы показали, что, хотя применение Теоремы 14 является удобным способом получения неравенств концентрации для выборок без возвращений, непосредственная оценка производящей функции моментов $\mathbb{E}\left[e^{\lambda(\xi-\mathbb{E}[\xi])}\right]$ рассматриваемой случайной величины может вести к существенно лучшим результатам. Кроме того было показано, что суммы для выборок без возвращений концентрируются сильнее сумм независимых случайных величин. Результаты следующего раздела показывают, что подобный эффект наблюдается и для более общих функций случайных величин.

2.2 Функции, определенные на разбиениях

Заметим, что случайную величину S'_n из прошлого раздела можно эквивалентно определить с помощью случайных перестановок. Рассмотрим случайную перестановку, выбран-

ную равномерно из симметрической группы перестановок множества $\{1,\dots,N\}$. Такая перестановка может быть выражена N-мерным вектором $\boldsymbol{\pi}$ с натуральными координатами, полученными перестановкой множества $\{1,\dots,N\}$. Тогда S'_n можно определить как

$$S'_n = S'_n(\boldsymbol{\pi}) = \frac{1}{n}\sum_{i=1}^{n} c_{\pi_i},$$

где π_i — i-ая координата $\boldsymbol{\pi}$, а с помощью $\mathcal{C} = \{c_1,\dots,c_N\}$ мы по-прежнему обозначаем конечное множество ограниченных действительных чисел $c_i \in [0,1]$.

Существует также третий способ определения функции S'_n, основанный на разбиениях. Пусть $(\mathcal{U}^n,\mathcal{U}^u)$ — разбиение множества $\{1,\dots,N\} = \mathcal{U}^n \cup \mathcal{U}^u$ на два непересекающихся подмножества мощностей n и $u = N-n$ соответственно. Мы будем выбирать $(\mathcal{U}^n,\mathcal{U}^u)$ равномерно из множества всех таких разбиений, которых всего $C_N^n = \frac{N!}{n!(N-n)!}$. Тогда

$$S'_n = S'_n(\mathcal{U}^n,\mathcal{U}^u) = \frac{1}{n}\sum_{i\in\mathcal{U}^n} c_i.$$

Очевидно, не все функции $f(\boldsymbol{\pi})$ случайных перестановок $\boldsymbol{\pi}$ могут быть эквивалентно представлены с помощью разбиений: например, подобное представление невозможно для функции $f(\boldsymbol{\pi}) = c_{\pi_1}+c^2_{\pi_2}+c^3_{\pi_3}$. Для суммы случайных величин S'_n это возможно благодаря ее симметричности относительно перестановок слагаемых. Оказывается, для функций, допускающих представление с помощью разбиений, справедлив ряд нетривиальных неравенств концентрации, обзор которых будут приведен в данном разделе. Часть из них первоначально формулировалась в терминах случайных перестановок $\boldsymbol{\pi}$, а часть — в терминах случайных разбиений $\mathcal{U}^n \cup \mathcal{U}^u$. Для удобства и однообразия мы будем формулировать все результаты в терминах перестановок.

Сначала будет рассмотрен аналог неравенства МакДиармида Теоремы 9 для выборок без возвращений. Данный результат, подобно неравенству Серфлинга, получен авторами [34] на основе непосредственной работы с производящей функцией моментов с помощью метода мартингалов. Затем будет приведено субгауссовское неравенство С. Г. Бобкова [17], непосредственно связанное с энтропийным подходом М. Леду.

2.2.1 Неравенство МакДиармида для выборок без возвращений

Для обобщения неравенства МакДиармида на выборки без возвращений в работе [34] вводится понятие (n,u)-*симметричных относительно перестановок* функций:

Определение 3. *Функция* $f\colon \boldsymbol{\pi} \to \mathbb{R}$, *заданная на симметрической группе множества* $\{1,\dots,N\}$, *называется* (n,u)-симметричной относительно перестановок, *если она не меняет*

своего значения при замене порядка первых n и/или последних $u = N - n$ координат $\boldsymbol{\pi}$. Для краткости такие функции мы будем называть просто (n, u)-симметричными.

Очевидно, любая (n, u)-симметричная функция $f(\boldsymbol{\pi})$ может быть определена в терминах случайных разбиений. И, наоборот, любая функция, определенная на множестве разбиений $\mathcal{U}^n \cup \mathcal{U}^u$, допускает представление с помощью (n, u)-симметричной функции. Все результаты настоящего раздела будут формулироваться в терминах (n, u)-симметричных функций.

Авторы [34] приводят следующий результат, как и неравенство Серфлинга Теоремы 17 основанный на методе мартингалов:

Теорема 18 (Эль-Янив, Печиони [34])**.** *Пусть $\boldsymbol{\pi}$ — вектор случайной перестановки, выбранной равномерно из симметрической группы перестановок множества $\{1, \ldots, N\}$. Пусть $f(\boldsymbol{\pi})$ — (n, u)-симметричная функция, для которой существует константа $\beta > 0$, такая что $|f(\boldsymbol{\pi}) - f(\boldsymbol{\pi}^{i,j})| \leqslant \beta$ для всех $\boldsymbol{\pi}$, $i \in \{1, \ldots, n\}$ и $j \in \{n+1, \ldots, N\}$, где перестановка $\boldsymbol{\pi}^{i,j}$ получена из $\boldsymbol{\pi}$ транспозицией ее i-й и j-й координат. Тогда для любого $\varepsilon \geqslant 0$:*

$$\mathbb{P}\left\{f(\boldsymbol{\pi}) - \mathbb{E}[f(\boldsymbol{\pi})] \geqslant \varepsilon\right\} \leqslant \exp\left\{-\frac{2\varepsilon^2}{n\beta^2}\left(\frac{N - 1/2}{N - n}\right)\left(1 - \frac{1}{2\max(n, N-n)}\right)\right\}. \tag{2.2}$$

Аналогичное неравенство справедливо и для $\mathbb{P}\left\{\mathbb{E}[f(\boldsymbol{\pi})] - f(\boldsymbol{\pi}) \geqslant \varepsilon\right\}$, поскольку предположения Теоремы инвариантны относительно замены знака функции f.

Теорема 18 является аналогом неравенства МакДиармида для выборок без возвращений, а ее предположения непосредственно связаны с условием ограниченных разностей (1.18). Грубое сравнение двух неравенств показывает, что они совпадают с точностью до отсутствия выражения $\left(\frac{N-1/2}{N-n}\right)\left(1 - \frac{1}{2\max(n,N-n)}\right)$ в показателе экспоненты неравенства МакДиармида. Пренебрегая вторым множителем, который близок к 1 для больших N, можно заключить, что при $n \to N$ оценка Теоремы 18 точнее неравенства МакДиармида.

Заметив, что сумма S'_n из прошлого раздела удовлетворяет условию последней Теоремы с $\beta = \frac{1}{n}$, мы немедленно получаем следствие:

Следствие 5. *Для любого $\varepsilon \geqslant 0$ справедливо:*

$$\mathbb{P}\left\{S'_n - \mathbb{E}[S'_n] \geqslant \varepsilon\right\} \leqslant \exp\left\{-2n\varepsilon^2\left(\frac{N - 1/2}{N - n}\right)\left(1 - \frac{1}{2\max(n, N-n)}\right)\right\}.$$

Аналогичное неравенство справедливо для $\mathbb{P}\left\{\mathbb{E}[S'_n] - S'_n \geqslant \varepsilon\right\}$.

При больших N последнее неравенство имеет тот же порядок, что неравенство Серфлинга Теоремы 17.

2.2.2 Неравенство Бобкова

Следующий результат основан на энтропийном подходе. Рассмотрим случайное разбиение $(\mathcal{U}^n, \mathcal{U}^u)$ множества $\{1, \dots, N\} = \mathcal{U}^n \cup \mathcal{U}^u$ на два непересекающихся подмножества мощностей n и $u = N - n$ соответственно, равномерно распределенное на множестве из всех $\frac{N!}{n!(N-n)!}$ таких разбиений. Далее разбиение будет удобно представлять вектором перестановки $\boldsymbol{\pi}$, подразумевая, что перестановка задает разбиение на подмножества $\{\pi_i\}_{i \in I} \cup \{\pi_j\}_{j \in J}$, где $I = \{1, \dots, n\}$ и $J = \{n+1, \dots, N\}$. *Соседними* естественно считать разбиения, которые *могут быть заданы* перестановками $\boldsymbol{\pi}_1$ и $\boldsymbol{\pi}_2$, отличающимися ровно на одну транспозицию: $\boldsymbol{\pi}_1 = \boldsymbol{\pi}_2^{i,j}$ для некоторых $i \in I$ и $j \in J$ (обратим внимание на то, что каждое разбиение может быть задано $n(N-n)$ разными перестановками). Каждое разбиение, таким образом, имеет ровно $n(N-n)$ соседних разбиений.

Как отмечалось ранее, любая (n, u)-симметричная функция f фактически задается на множестве разбиений. Дискретный градиент $\nabla f(\boldsymbol{\pi})$ такой функции является вещественным вектором размерностью $n(N-n)$ и квадрат его длины выражается следующим образом:

$$V^f(\boldsymbol{\pi}) = |\nabla f(\boldsymbol{\pi})|^2 = \sum_{i \in I} \sum_{j \in J} \big(f(\boldsymbol{\pi}) - f(\boldsymbol{\pi}^{i,j})\big)^2.$$

Следующий результат получен в [17, Теорема 2.1]:

Теорема 19 (С. Г. Бобков, [17])**.** *Пусть $\boldsymbol{\pi}$ — вектор случайной перестановки, выбранной равномерно из симметрической группы перестановок множества $\{1, \dots, N\}$. Пусть $f(\boldsymbol{\pi})$ — (n, u)-симметричная функция и $\Sigma^2 \geqslant 0$ — действительное число, такое что $V^f(\boldsymbol{\pi}) \leqslant \Sigma^2$ для всех $\boldsymbol{\pi}$. Тогда для любого $\varepsilon \geqslant 0$ справедливо:*

$$\mathbb{P}\left\{f(\boldsymbol{\pi}) - \mathbb{E}[f(\boldsymbol{\pi})] \geqslant \varepsilon\right\} \leqslant \exp\left\{-\frac{(N+2)\varepsilon^2}{4\Sigma^2}\right\}.$$

Аналогичное неравенство справедливо для $\mathbb{P}\left\{\mathbb{E}[f(\boldsymbol{\pi})] - f(\boldsymbol{\pi}) \geqslant \varepsilon\right\}$, поскольку предположения Теоремы инвариантны относительно замены знака функции f.

Замечание 7. *Приведенная здесь формулировка теоремы отличается от первоначальной версии С. Г. Бобкова, формулировавшейся в терминах случайных разбиений $(\mathcal{U}^n, \mathcal{U}^u)$. Она является следствием того, что для (n, u)-симметричных функций f случайные величины $f(\mathcal{U}^n, \mathcal{U}^u)$ и $f(\boldsymbol{\pi})$, как несложно убедиться, одинаково распределены.*

Предположим, что (n, u)-симметричная относительно перестановок функция $f(\boldsymbol{\pi})$ удовлетворяет предположениям Теоремы 18, то есть существует значение β, такое что $|f(\boldsymbol{\pi}) -$

$f(\boldsymbol{\pi}^{i,j})| \leqslant \beta$ для всех $\boldsymbol{\pi}$, $i \in I$ и $j \in J$. Отсюда очевидным образом следует, что она также удовлетворяет предположениям Теоремы 19 с параметром $\sigma^2 = n(N-n)\beta^2$, поскольку:

$$V^f(\boldsymbol{\pi}) = \sum_{i\in I}\sum_{j\in J}\big(f(\boldsymbol{\pi}) - f(\boldsymbol{\pi}^{i,j})\big)^2 \leqslant \sum_{i\in I}\sum_{j\in J}\beta^2 = n(N-n)\beta^2.$$

В этом случае Теорема 19 дает более слабую версию Теоремы 18: показатель экспоненты уменьшается в 8 раз. Тем не менее, предположения Теоремы 19 представляются менее строгими, и известно множество приложений (см. замечания к Теореме 6.7 [19]), для которых Теорема 18 дает лишь тривиальные неравенства, в то время как Теорема 19 позволяет получить достаточно сильные результаты. Один из таких примеров будет приведен в Разделе 2.3 настоящей работы.

Попробуем применить Теорему 19 для получения неравенств концентрации для (n,u)-симметричной функции $S'_n = \frac{1}{n}\sum_{i=1}^n c_{\pi_i}$. Для этого нам понадобится верхняя оценка на $V^f(\boldsymbol{\pi})$ для $f(\boldsymbol{\pi}) = S'_n(\boldsymbol{\pi})$. Заметим, что для любых $\boldsymbol{\pi}$, $i \in I$ и $j \in J$ справедливо:

$$S'_n(\boldsymbol{\pi}) - S'_n(\boldsymbol{\pi}^{i,j}) = \frac{1}{n}(c_{\pi_i} - c_{\pi_j}),$$

где, как и ранее, перестановка $\boldsymbol{\pi}^{i,j}$ получена из $\boldsymbol{\pi}$ транспозицией ее i-й и j-й координат. Мы получаем

$$V^f(\boldsymbol{\pi}) = \frac{1}{n^2}\sum_{i\in I}\sum_{j\in J}(c_i - c_j)^2.$$

Очевидная оценка $(c_i - c_j)^2 \leqslant 1$ дает нам $V^f(\boldsymbol{\pi}) \leqslant \frac{(N-n)}{n}$, что с учетом Теоремы 19 дает более слабую версию Следствия 5. Мы также можем воспользоваться следующей более точной оценкой.

Лемма 8. *Для $f(\boldsymbol{\pi}) = S'_n = \frac{1}{n}\sum_{i=1}^n c_{\pi_i}$ справедливо:*

$$\sup_{\boldsymbol{\pi}} V^f(\boldsymbol{\pi}) \leqslant \left(\frac{N}{n}\right)^2 \sigma_N^2, \tag{2.3}$$

где с помощью σ_N^2 обозначена дисперсия случайной величины, равномерно распределенной на множестве $\mathcal{C}$.

Доказательство.

$$\begin{aligned} V^f(\boldsymbol{\pi}) &= \frac{1}{n^2}\sum_{i\in I}\sum_{j\in J}(c_{\pi_i} - c_{\pi_j})^2 \leqslant \frac{1}{n^2}\sum_{1\leqslant i<j\leqslant N}(c_i - c_j)^2 \\ &= \left(\frac{N(N-1)}{n^2}\right)\frac{1}{N(N-1)}\sum_{1\leqslant i<j\leqslant N}(c_i - c_j)^2 \qquad (2.4) \\ &= \left(\frac{N(N-1)}{n^2}\right)\frac{1}{N-1}\sum_{i=1}^N\left(c_i - \frac{1}{N}\sum_{i=1}^N c_i\right)^2 = \left(\frac{N}{n}\right)^2\sigma_N^2, \end{aligned}$$

где в (2.4) мы воспользовались Леммой 4. ■

Теорема 19 вместе с Леммой 8 дают следующее следствие:

Следствие 6. *Для любого $\varepsilon \geqslant 0$ справедливо:*

$$\mathbb{P}\left\{S'_n - \mathbb{E}[S'_n] \geqslant \varepsilon\right\} \leqslant \exp\left\{-\frac{(N+2)m^2\varepsilon^2}{4N^2\sigma_N^2}\right\} \leqslant \exp\left\{-\frac{n^2\varepsilon^2}{4N\sigma_N^2}\right\}. \tag{2.5}$$

Аналогичные неравенства справедливы и для $\mathbb{P}\left\{\mathbb{E}[S'_n] - S'_n \geqslant \varepsilon\right\}$. Кроме того, для любого $t > 0$ с вероятностью не меньше $1 - e^{-t}$ справедливо:

$$\mathbb{E}[S'_n] \leqslant S'_n + 2\sqrt{\sigma_N^2\left(\frac{N}{n}\right)\frac{t}{n}}.$$

Замечание 8. *Результат Следствия 6 не является новым. Неравенство (2.5) без множителя 4 было впервые получено в [68, Лемма 3.1] для случая $N = m \cdot n$, $n \in \mathbb{N}$, на основе совершенно другого подхода. Автор моделирует реализацию выборки без возвращений с помощью двух последовательных шагов: (1) разбиение множества $\mathcal{C}$ на n непересекающихся подмножеств $\mathcal{C}_1, \ldots, \mathcal{C}_n$; (2) случайный выбор одного элемента из каждого из подмножеств $\mathcal{C}_1, \ldots, \mathcal{C}_n$. При фиксированном разбиении, случайные величины, выбираемые на втором шаге, являются независимыми с математическими ожиданиями $\bar{c}_1, \ldots, \bar{c}_n$ соответственно, где $\bar{c}_i$ — среднее значение подмножества $\mathcal{C}_i$. Совмещая эту модель с методом Чернова и пользуясь неравенством Хефдинга для оценок производящих функций, мы приходим к утверждению теоремы.*

Важно отметить, что силами Теоремы 18 (фактически являющейся аналогом неравенства МакДиармида) невозможно получить подобный результат, учитывающий дисперсию случайной величины. Сравнивая полученный результат с неравенством Бернштейна (а именно — последним неравенством Теоремы 16), мы видим, что член порядка $1/n$ исчез: теперь у оценки лишь один субгауссовский «режим», который описывает и малые и большие отклонения. В то же время у члена порядка $1/\sqrt{n}$ появился дополнительный множитель $\sqrt{2N/n}$, из-за которого оценка Следствия 6 может вырождаться при больших N и $n = o(N)$.

Возникает вопрос: возможно ли силами Теоремы 19, заменив (2.3) более точной верхней оценкой, улучшить результат Следствия 6? Следующая лемма дает отрицательный ответ на этот вопрос.

Утверждение 1. *Верхняя оценка Леммы 8 является неулучшаемой.*

Доказательство. Мы приведем пример множества $\mathcal{C}$, для которого неравенство Леммы 8 обращается в равенство. Рассмотрим случай, когда $\mathcal{C} = \{c_1, \ldots, c_N\}$, где $c_1, \ldots, c_n = r$,

$c_{n+1}, \dots, c_N = v$ для двух различных действительных чисел $r, v \in [0,1]$. В этом случае, как легко проверить, супремум $\sup_{\boldsymbol{\pi}} V^f(\boldsymbol{\pi})$ достигается на тождественной перестановке $\boldsymbol{\pi} = (1, 2, \dots, N)$. Действительно, в этом случае все $n(N-n)$ слагаемых суммы

$$\sum_{i \in I} \sum_{j} (c_{\pi_i} - c_{\pi_j})^2$$

обращаются в $(r-v)^2$, в то время как для любой другой перестановки $\boldsymbol{\pi}$ некоторые из них будут обращаться в ноль. Таким образом, для такого множества $\mathcal{C}$ спарведливо:

$$\sup_{\boldsymbol{\pi}} V^f(\boldsymbol{\pi}) = \frac{n(N-n)}{n^2}(r-v)^2 = \frac{N-n}{n}(r-v)^2.$$

В то же время, легко проверить, что в этом случае

$$\sigma_N^2 = \frac{n(N-n)}{N^2}(r-v)^2.$$

Это завершает доказательство утверждения. ■

Оказывается, для выборок без возвращений концентрируются сильнее не только суммы, но и ряд более общих функций. Это наглядно продемонстрировано на примере неравенства Эль-Янива, равномерно улучшающего неравенство МакДиармида при росте размера выборки n к N. Также мы привели результат Бобкова, позволяющий получать нетривиальные неравенства концентрации, учитывающие дисперсии случайных величин. В следующем разделе мы рассмотрим супремум эмпирических процессов, и на основе описанных в прошлых разделах результатов получим для него два новых неравенства концентрации.

2.3 Супремумы эмпирических процессов для выборок без возвращений

Результаты настоящего раздела являются новыми и опубликованы в работах [96, 121].

В этом разделе мы вновь обратимся к супремумам эмпирических процессов, ранее рассмотренным в Разделе 1.3.2. Однако, на этот раз — для выборок без возвращений. Поскольку неравенство Талаграна Теоремы 13 формулируется для выборки независимых случайных величин, оно перестает выполняться для выборок без возвращений. В данном разделе будет получено два новых неравенства концентрации для супремумов эмпирических процессов для выборок без возарвщаений. Первое основано на применении неравенства С. Г. Бобкова Теоремы 19. Второе — на методе Хефдинга, описанном в Разделе 2.1.1.

Введем необходимые обозначения и определения. Пусть $\mathcal{C} = \{c_1, \dots, c_N\}$ — некоторое конечное множество. Для $n \leqslant N$ рассмотрим последовательности случайных величин $\{\eta_1, \dots, \eta_n\}$ и $\{\xi_1, \dots, \xi_n\}$, выбранные равномерно из $\mathcal{C}$ без возвращений и с возвращениями соответственно. Пусть $\mathcal{F}$ — *счетное*[2] множество отображений $f\colon \mathcal{C} \to \mathbb{R}$, таких что $\mathbb{E}[f(\xi_1)] = 0$ и $f(x) \in [-1, 1]$ для всех $f \in \mathcal{F}$ и $x \in \mathcal{C}$. Супремумы эмпирического процесса для выборок с и без возвращений соответственно введем следующим образом:

$$Q_n = \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} f(\xi_i), \quad Q'_n = \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} f(\eta_i).$$

Со случайной величиной Q_n мы уже сталкивались в Разделе 1.3.2, где мы привели неравенства типа МакДиармида и Беннетта для нее.

Заметим, что случайная величина Q'_n может быть эквивалентно определена с помощью случайных перестановок, так же как рассмотренная в Разделе 2.1 случайная величина S'_n:

$$Q'_n = Q'_n(\boldsymbol{\pi}) = \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} f(c_{\pi_i}). \tag{2.6}$$

Кроме того, функция $Q'_n(\boldsymbol{\pi})$, очевидно, (n, u)-симметрична относительно перестановок. Применяя Теорему 18 к функции Q'_n, мы получим наиболее точное из существующих для Q'_n неравенств концентрации:

Следствие 7 (Эль-Янив, Печиони [34][3])**.** *Для любого $\varepsilon \geqslant 0$ справедливо:*

$$\mathbb{P}\left\{Q'_n - \mathbb{E}[Q'_n] \geqslant \varepsilon\right\} \leqslant \exp\left\{-\frac{\varepsilon^2}{2n}\left(\frac{N-1/2}{N-n}\right)\left(1 - \frac{1}{2\max(n, N-n)}\right)\right\}.$$

Аналогичное неравенство справедливо и для $\mathbb{P}\left\{\mathbb{E}[Q'_n] - Q'_n \geqslant \varepsilon\right\}$.

Данный результат можно считать аналогом Теоремы 12 типа МакДиармида для выборок без возвращений. Неравенств же типа Беннетта для случайной величины Q'_n в литературе представлено до сих пор не было.

Новые неравенства. Далее будет получено два новых неравенства концентрации для случайной величины Q'_n. Первый результат основан на Теореме 19 и описывает субгауссовское поведение случайной величины Q'_n:

Теорема 20. *Введем обозначение $\sigma^2_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \mathbb{D}[f(\xi_1)]$. Тогда для любого $\varepsilon \geqslant 0$ справедливо:*

$$\mathbb{P}\left\{Q'_n - \mathbb{E}[Q'_n] \geqslant \varepsilon\right\} \leqslant \exp\left(-\frac{(N+2)\varepsilon^2}{8N^2\sigma^2_{\mathcal{F}}}\right). \tag{2.7}$$

[2] Замечание 4 также остается в силе для результатов настоящего раздела.

[3] Это неравенство не было приведено в [25, 34] в явном виде. Однако, оно является простым следствием применения [34, Леммы 2] для Q'_m при $\beta = 2$.

Аналогичное неравенство справедливо и для $\mathbb{P}\{\mathbb{E}[Q'_n] - Q'_n \geqslant \varepsilon\}$. *Кроме того, для любого* $t \geqslant 0$ *с вероятностью не меньше* $1 - e^{-t}$ *справедливо:*

$$Q'_n \leqslant \mathbb{E}[Q'_n] + 2\sqrt{2N\sigma_{\mathcal{F}}^2 t}. \tag{2.8}$$

Доказательство Теоремы 20 основано на применении Теоремы 19 к функции $Q'_n(\boldsymbol{\pi})$. Для этого нам необходимо получить верхнюю оценку для $V^{Q'_n}(\boldsymbol{\pi})$, что является непростой задачей. Вместо этого мы получим новую версию Теоремы 19, которая упростит нашу работу с дискретным градиентом. Для любой (n, u)-симметричной функции $f\colon \boldsymbol{\pi} \to \mathbb{R}$ определим величину, связанную с $V^f(\boldsymbol{\pi})$:

$$V_+^f(\boldsymbol{\pi}) := \sum_{i\in I}\sum_{j\in J}\big(f(\boldsymbol{\pi}) - f(\boldsymbol{\pi}^{i,j})\big)^2 \mathbb{1}\{f(\boldsymbol{\pi}) \geqslant f(\boldsymbol{\pi}^{i,j})\}.$$

Оказывается, справедлива следующая модификация Теоремы 19:

Теорема 21. *Пусть* $\boldsymbol{\pi}$ *— вектор случайной перестановки, выбранной равномерно из симметрической группы перестановок множества* $\{1, \dots, N\}$. *Пусть* $f(\boldsymbol{\pi})$ *—* (n, u)*-симметричная функция и* $\Sigma^2 \geqslant 0$ *— действительное число, такое что* $V_+^f(\boldsymbol{\pi}) \leqslant \Sigma^2$ *для всех* $\boldsymbol{\pi}$. *Тогда справедливо* [4]*:*

$$\mathbb{P}\{f(\boldsymbol{\pi}) - \mathbb{E}[f(\boldsymbol{\pi})] \geqslant \varepsilon\} \leqslant \exp\left\{-\frac{(N+2)\varepsilon^2}{8\Sigma^2}\right\}. \tag{2.9}$$

Аналогичное неравенство справедливо и для $\mathbb{P}\{\mathbb{E}[f(\boldsymbol{\pi})] - f(\boldsymbol{\pi}) \geqslant \varepsilon\}$.

Доказательство. Мы будем следовать шагам доказательства Теоремы 19, представленным в [17] (см. первое неравенство Теоремы 2.1). Там показано, что для любой (n, u)-симметричной функции $g(\boldsymbol{\pi})$ справедливо:

$$\begin{aligned}&(N+2)\big(\mathbb{E}\big[e^{g(\boldsymbol{\pi})}\log e^{g(\boldsymbol{\pi})}\big] - \mathbb{E}\big[e^{g(\boldsymbol{\pi})}\big]\mathbb{E}\big[\log e^{g(\boldsymbol{\pi})}\big]\big)\\ &\leqslant \mathbb{E}\left[\sum_{i\in I}\sum_{j\in J}\big(g(\boldsymbol{\pi}) - g(\boldsymbol{\pi}^{i,j})\big)\big(e^{g(\boldsymbol{\pi})} - e^{g(\boldsymbol{\pi}^{i,j})}\big)\right],\end{aligned} \tag{2.10}$$

где, как и раньше, $I = \{1, \dots, n\}$ и $J = \{n+1, \dots, N\}$. Заметим также, что для любых $a, b \in \mathbb{R}$:

$$(a-b)(e^a - e^b) \leqslant \frac{e^a + e^b}{2}(a-b)^2. \tag{2.11}$$

[4] Более слабая версия этого результата была представлена в [121].

Обозначим симметрическую группу перестановок множества $\{1,\dots,N\}$ с помощью Π_N. Перепишем правую часть неравенства (2.10) следующим образом:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i\in I}\sum_{j\in J}\bigl(g(\boldsymbol{\pi})-g(\boldsymbol{\pi}^{i,j})\bigr)\bigl(e^{g(\boldsymbol{\pi})}-e^{g(\boldsymbol{\pi}^{i,j})}\bigr)\right]\\
&=\frac{1}{N!}\sum_{\boldsymbol{\pi}\in\Pi_N}\left[\sum_{i\in I}\sum_{j\in J}\bigl(g(\boldsymbol{\pi})-g(\boldsymbol{\pi}^{i,j})\bigr)\bigl(e^{g(\boldsymbol{\pi})}-e^{g(\boldsymbol{\pi}^{i,j})}\bigr)\right]\\
&=\frac{2}{N!}\sum_{\boldsymbol{\pi}\in\Pi_N}\sum_{i\in I}\sum_{j\in J}\bigl(g(\boldsymbol{\pi})-g(\boldsymbol{\pi}^{i,j})\bigr)\bigl(e^{g(\boldsymbol{\pi})}-e^{g(\boldsymbol{\pi}^{i,j})}\bigr)\mathbb{1}\{g(\boldsymbol{\pi})\geqslant g(\boldsymbol{\pi}^{i,j})\}.
\end{aligned}$$

Воспользовавшись неравенством (2.11), мы получаем:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i\in I}\sum_{j\in J}\bigl(g(\boldsymbol{\pi})-g(\boldsymbol{\pi}^{i,j})\bigr)\bigl(e^{g(\boldsymbol{\pi})}-e^{g(\boldsymbol{\pi}^{i,j})}\bigr)\right]\\
&\leqslant\frac{2}{N!}\sum_{\boldsymbol{\pi}\in\Pi_N}\sum_{i\in I}\sum_{j\in J}\frac{\bigl(e^{g(\boldsymbol{\pi})}+e^{g(\boldsymbol{\pi}^{i,j})}\bigr)}{2}\bigl(g(\boldsymbol{\pi})-g(\boldsymbol{\pi}^{i,j})\bigr)^2\mathbb{1}\{g(\boldsymbol{\pi})\geqslant g(\boldsymbol{\pi}^{i,j})\}\\
&\leqslant\frac{2}{N!}\sum_{\boldsymbol{\pi}\in\Pi_N}\sum_{i\in I}\sum_{j\in J}e^{g(\boldsymbol{\pi})}\bigl(g(\boldsymbol{\pi})-g(\boldsymbol{\pi}^{i,j})\bigr)^2\mathbb{1}\{g(\boldsymbol{\pi})\geqslant g(\boldsymbol{\pi}^{i,j})\}\\
&=2\mathbb{E}\left[V_+^g(\boldsymbol{\pi})e^{g(\boldsymbol{\pi})}\right].
\end{aligned}$$

Таким образом, справедливо следующее:

$$(N+2)\bigl(\mathbb{E}\bigl[e^{g(\boldsymbol{\pi})}\log e^{g(\boldsymbol{\pi})}\bigr]-\mathbb{E}\bigl[e^{g(\boldsymbol{\pi})}\bigr]\mathbb{E}\bigl[\log e^{g(\boldsymbol{\pi})}\bigr]\bigr)\leqslant 2\mathbb{E}\left[V_+^g(\boldsymbol{\pi})e^{g(\boldsymbol{\pi})}\right].$$

Применяя это неравенство к функции λf для произвольного $\lambda\in\mathbb{R}$, мы получаем:

$$(N+2)\bigl(\mathbb{E}\bigl[e^{\lambda f(\boldsymbol{\pi})}\log e^{\lambda f(\boldsymbol{\pi})}\bigr]-\mathbb{E}\bigl[e^{\lambda f(\boldsymbol{\pi})}\bigr]\mathbb{E}\bigl[\log e^{\lambda f(\boldsymbol{\pi})}\bigr]\bigr)\leqslant 2\mathbb{E}\left[V_+^{\lambda f}(\boldsymbol{\pi})e^{\lambda f(\boldsymbol{\pi})}\right]\leqslant 2\Sigma^2\lambda^2\mathbb{E}\left[e^{\lambda f(\boldsymbol{\pi})}\right]. \tag{2.12}$$

В доказательстве Теоремы 19 в [17] отмечено, что неравенство (2.12) влечет за собой следующую верхнюю оценку на производящую функцию моментов:

$$\mathbb{E}\left[e^{\lambda(f-\mathbb{E}[f])}\right]\leqslant e^{\frac{2\Sigma^2\lambda^2}{N+2}}. \tag{2.13}$$

Этот факт известен в литературе как «метод Хербста» и лежит в основе энтропийного подхода. Теперь мы применяем метод Чернова, описанный ранее, который дает нам для любого $\lambda,\varepsilon\geqslant 0$:

$$\mathbb{P}\left\{f(\boldsymbol{\pi})-\mathbb{E}[f(\boldsymbol{\pi})]\geqslant\varepsilon\right\}\leqslant\frac{\mathbb{E}\left[e^{\lambda(f-\mathbb{E}[f])}\right]}{e^{\lambda\varepsilon}}\leqslant e^{\frac{2\Sigma^2\lambda^2}{N+2}-\lambda\varepsilon}.$$

Нам остается минимизировать правую часть последнего неравенства по λ, что достигается при $\lambda=\frac{\varepsilon(N+2)}{4\Sigma^2}$, и мы получаем (2.9).

Неравенство для левого хвоста распределения может быть получено аналогичным образом, используя (2.13) при $\lambda < 0$:

$$\mathbb{P}\left\{\mathbb{E}[f(\boldsymbol{\pi})] - f(\boldsymbol{\pi}) \geqslant \varepsilon\right\} = \mathbb{P}\left\{\lambda\big(f(\boldsymbol{\pi}) - \mathbb{E}[f(\boldsymbol{\pi})]\big) \geqslant -\lambda\varepsilon\right\} \leqslant \frac{\mathbb{E}\left[e^{\lambda(f-\mathbb{E}[f])}\right]}{e^{-\lambda\varepsilon}} \leqslant e^{\frac{2\Sigma^2\lambda^2}{N+2}+\lambda\varepsilon}.$$

На этот раз мы полагаем $\lambda = -\frac{\varepsilon(N+2)}{4\Sigma^2}$, что завершает доказательство. ■

Замечание 9. *Использованная в доказательстве идея отдельной оценки положительных и отрицательных приращений часто используется в литературе (например в [19, Теорема 6.16]).*

Доказательство Теоремы 20. Рассмотрим две функции $f, g\colon \mathcal{A} \to \mathbb{R}$, определенных на произвольном множестве $\mathcal{A}$, и предположим, что $\sup_{a\in\mathcal{A}} f(a) = f(\bar{a})$ для некоторого $\bar{a} \in \mathcal{A}$. Тогда справедливо:

$$\left(\sup_{a\in\mathcal{A}} f(a) - \sup_{a\in\mathcal{A}} g(a)\right)^2 \mathbb{1}\left\{\sup_{a\in\mathcal{A}} f(a) \geqslant \sup_{a\in\mathcal{A}} g(a)\right\} \leqslant \big(f(\bar{a}) - g(\bar{a})\big)^2. \tag{2.14}$$

Предположим, что супремум в определении (2.6) достигается на функции $\bar{f} \in \mathcal{F}$. Тогда из (2.14) следует:

$$\begin{aligned}
V_+^{Q'_n}(\boldsymbol{\pi}) &= \sum_{i\in I}\sum_{j\in J}\big(Q'_n(\boldsymbol{\pi}) - Q'_n(\boldsymbol{\pi}^{i,j})\big)^2\,\mathbb{1}\left\{Q'_n(\boldsymbol{\pi}) \geqslant Q'_n(\boldsymbol{\pi}^{i,j})\right\} \\
&\leqslant \sum_{i\in I}\sum_{j\in J}\left(\sum_{k=1}^{n}\bar{f}(c_{\pi_k}) - \sum_{k=1}^{n}\bar{f}\big(c_{\pi_k^{i,j}}\big)\right)^2 \\
&= \sum_{i\in I}\sum_{j\in J}\Big(\bar{f}(c_{\pi_i}) - \bar{f}(c_{\pi_j})\Big)^2 \\
&\leqslant \sum_{1\leqslant i<j\leqslant N}\Big(\bar{f}(c_i) - \bar{f}(c_j)\Big)^2 = N^2\mathbb{D}[\bar{f}(\xi_1)],
\end{aligned}$$

где мы воспользовались Леммой 4. Поскольку $\bar{f}$ зависит от выбора $\boldsymbol{\pi}$, мы получаем:

$$V_+^{Q'_n}(\boldsymbol{\pi}) \leqslant N^2 \sup_{f\in\mathcal{F}} \mathbb{D}[f(\xi_1)].$$

Применение Теоремы 21 завершает доказательство. ■

Следующий результат основан на непосредственном применении метода Хефдинга, описанного в Разделе 2.1.1, и дает неравенство типа Беннетта для случайной величины Q'_n:

Теорема 22. *Положим* $\sigma_{\mathcal{F}}^2 = \sup_{f\in\mathcal{F}} \mathbb{D}[f(\xi_1)]$, $v = n\sigma_{\mathcal{F}}^2 + 2\mathbb{E}[Q_n]$ *и для* $u \geqslant -1$ *определим* $\phi(u) = e^u - u - 1$, $h(u) = (1+u)\log(1+u) - u$. *Тогда для* $\varepsilon \geqslant \mathbb{E}[Q_n] - \mathbb{E}[Q'_n] \geqslant 0$ *справедливо:*

$$\mathbb{P}\left\{Q'_n - \mathbb{E}[Q'_n] \geqslant \varepsilon\right\} \leqslant \exp\left(-vh\left(\frac{\varepsilon - \mathbb{E}[Q_n] + \mathbb{E}[Q'_n]}{v}\right)\right) \tag{2.15}$$

$$\leqslant \exp\left(-\frac{(\varepsilon - \mathbb{E}[Q_n] + \mathbb{E}[Q'_n])^2}{2v + \frac{2}{3}(\varepsilon - \mathbb{E}[Q_n] + \mathbb{E}[Q'_n])}\right). \tag{2.16}$$

Кроме того, для любого $t \geqslant 0$ с вероятностью не менее $1 - e^{-t}$ справедливо:

$$Q'_n \leqslant \mathbb{E}[Q_n] + \sqrt{2vt} + \frac{t}{3}. \tag{2.17}$$

Мы приведем доказательство для конечного множества $\mathcal{F} = \{f_1, \ldots, f_m\}$. Как отмечено в доказательстве Теоремы 2.11 [17], случай счетного $\mathcal{F}$ доказывается простым переходом к пределу при $m \to \infty$. Для доказательства нам понадобится следующая техническая лемма.

Лемма 9. *Пусть $\boldsymbol{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$. Тогда для всех $\lambda > 0$ функция*

$$F(\boldsymbol{x}) = \exp\left(\lambda \sup_{i=1,\ldots,d} x_i\right)$$

является выпуклой.

Доказательство. Для начала покажем, что если $g\colon \mathbb{R} \to \mathbb{R}$ — выпуклая и неубывающая функция, а $f\colon \mathbb{R}^d \to \mathbb{R}$ выпукла, то $g\big(f(\boldsymbol{x})\big)$ тоже выпукла. Действительно, для $\alpha \in [0,1]$ и $\boldsymbol{x}', \boldsymbol{x}'' \in \mathbb{R}^d$:

$$g\Big(f\big(\alpha\boldsymbol{x}' + (1-\alpha)\boldsymbol{x}''\big)\Big) \leqslant g\big(\alpha f(\boldsymbol{x}') + (1-\alpha) f(\boldsymbol{x}'')\big) \leqslant \alpha g\big(f(\boldsymbol{x}')\big) + (1-\alpha)\alpha g\big(f(\boldsymbol{x}'')\big).$$

С учетом того, что $g(y) = e^{\lambda y}$ — выпукла и возрастает для $\lambda > 0$, нам остается показать, что $f(\boldsymbol{x}) = \sup_{i=1,\ldots,d}(\xi_i)$ выпукла. Но для любого $\alpha \in [0,1]$ и $\boldsymbol{x}', \boldsymbol{x}'' \in \mathbb{R}^d$ справедливо:

$$\sup_{i=1,\ldots,d}\big(\alpha x'_i + (1-\alpha)x''_i\big) \leqslant \alpha \sup_{i=1,\ldots,d} x'_i + (1-\alpha) \sup_{i=1,\ldots,d} x''_i.$$

Это завершает доказательство леммы. ■

Доказательство Теоремы 22. Пусть последовательности $\{U_1, \ldots, U_n\}$ и $\{W_1, \ldots, W_n\}$ выбраны равномерно с и без возвращений соответственно из конечного множества m-мерных векторов $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N\} \subset \mathbb{R}^m$, где $\boldsymbol{v}_j = \big(f_1(c_j), \ldots, f_m(c_j)\big)^{\mathrm{T}}$. Пользуясь Леммой 9 и Теоремой 14, мы получаем для любого $\lambda > 0$:

$$\mathbb{E}\left[e^{\lambda Q'_n}\right] = \mathbb{E}\left[\exp\left(\lambda \sup_{j=1,\ldots,m}\left(\sum_{i=1}^{n} W_i\right)_j\right)\right] \leqslant \mathbb{E}\left[\exp\left(\lambda \sup_{j=1,\ldots,m}\left(\sum_{i=1}^{n} U_i\right)_j\right)\right] = \mathbb{E}\left[e^{\lambda Q_n}\right], \tag{2.18}$$

где нижним индексом j мы обозначили j-ую координату вектора. Воспользовавшись оценкой (1.25), мы ограничиваем производящую функцию моментов, стоящую в правой части неравенства (2.18):

$$\mathbb{E}\left[e^{\lambda Q'_n}\right] \leqslant \mathbb{E}\left[e^{\lambda Q_n}\right] \leqslant e^{\lambda\mathbb{E}[Q_n] + v\phi(\lambda)},$$

или

$$\mathbb{E}\left[e^{\lambda(Q'_n - \mathbb{E}[f'_n])}\right] \leqslant e^{\lambda(\mathbb{E}[Q_n] - \mathbb{E}[Q'_n]) + v\phi(\lambda)}.$$

Воспользовавшись методом Чернова, мы получаем для $\varepsilon \geqslant 0$ и $\lambda > 0$:

$$\mathbb{P}\left\{Q'_n - \mathbb{E}[Q'_n] \geqslant \varepsilon\right\} \leqslant \frac{\mathbb{E}\left[e^{\lambda(Q'_n - \mathbb{E}[Q'_n])}\right]}{e^{\lambda\varepsilon}} \leqslant \exp\bigl(\lambda(\mathbb{E}[Q_n] - \mathbb{E}[Q'_n]) + v\phi(\lambda) - \lambda\varepsilon\bigr). \tag{2.19}$$

Правая часть последнего неравенства выпукла и достигает своего минимума при

$$\lambda = \log\left(\frac{v + \varepsilon - \mathbb{E}[Q_n] + \mathbb{E}[Q'_n]}{v}\right), \tag{2.20}$$

откуда и берется техническое условие $\varepsilon \geqslant \mathbb{E}[Q_n] - \mathbb{E}[Q'_n]$. В противном случае мы полагаем $\lambda = 0$ и получаем тривиальную оценку 1. Кроме того, $\mathbb{E}[Q_n] \geqslant \mathbb{E}[Q'_n]$ следует также из Теоремы 14 с учетом Леммы 9. Подстановка (2.20) в (2.19) завершает доказательство первого неравенства Теоремы 22. Неравенства (2.16) и (2.17) вытекают из (2.15) так же, как неравенства (1.28) и (1.27) из (1.26). ■

Замечание 10. *Следует отметить, что в работе [46] получена верхняя оценка на* $\mathbb{E}\left[e^{-\lambda(Q_m - \mathbb{E}[Q_m])}\right]$ *для* $\lambda \geqslant 0$. *Эта оценка вместе с методом Чернова ведет к верхней оценки на* $\mathbb{P}\left\{\mathbb{E}[Q_m] - Q_m \geqslant \varepsilon\right\}$. *Однако, подход, использованный в доказательстве Теоремы 22, не может быть применен в этом случае, поскольку Теорема 14 не применима для отрецательных* λ.

Мы видим, что применение метода Хефдинга привело к появлению технического условия на параметр ε. Таким образом, Теорема 22 дает верхнюю оценку отклонения случайной величины Q'_n не от своего математического ожидания $\mathbb{E}[Q'_n]$, а от превосходящего его математического ожидания $\mathbb{E}[Q_n]$. При рассмотрении сумм в Следствии 4 такого эффекта не наблюдалось благодаря совпадению математических ожиданий. В общем случае, очевидно, $\mathbb{E}[Q_n]$ и $\mathbb{E}[Q'_n]$ не равны. Однако, справедлив следующий результат, показывающий, что в ряде случаев условие на ε в Теореме 22 не является строгим:

Лемма 10. *Справедливо следующее:*

$$0 \leqslant \mathbb{E}[Q_n] - \mathbb{E}[Q'_n] \leqslant 2\frac{n^3}{N}.$$

Доказательство. Доказательство первого неравенства было приведено ранее в доказательстве Теоремы 22. Перейдем к доказательству второго. Воспользовавшись определением, запишем:

$$\mathbb{E}[Q_n] - \mathbb{E}[Q'_n] = \frac{1}{N^n}\sum_{x_1,\ldots,x_n}\sup_{f\in\mathcal{F}}\sum_{i=1}^{n} f(x_i) + \left(\frac{1}{N^n} - \frac{(N-n)!}{N!}\right)\sum_{z_1,\ldots,z_n}\sup_{f\in\mathcal{F}}\sum_{i=1}^{n} f(z_i),$$

где первая сумма берется по всем упорядоченным последовательностям $(x_1, \ldots, x_n)$, в которых есть повторяющиеся члены, а вторая — по упорядоченным последовательностям

$(z_1, \ldots, z_n)$ без повторяющихся членов. Несложно убедиться, что во второй сумме всего $n! \cdot C_N^n$ слагаемых, а в первой, следовательно, $N^n - n! \cdot C_N^n$ слагаемых. С учетом того, что $\frac{1}{N^n} \leqslant \frac{(N-n)!}{N!}$ и $f(x) \in [-1, 1]$ для всех x, мы получаем:

$$\begin{aligned}
\mathbb{E}[Q_n] - \mathbb{E}[Q'_n] &\leqslant n\left(\frac{N^n - n! \cdot C_N^n}{N^n}\right) + n\left(\frac{(N-n)!}{N!} - \frac{1}{N^n}\right) n! \cdot C_N^n \\
&= 2n\left(\frac{N^n - n! \cdot C_N^n}{N^n}\right) \\
&= 2n - 2n\left(1 \cdot \left(1 - \frac{1}{N}\right) \cdots \left(1 - \frac{n-1}{N}\right)\right).
\end{aligned}$$

Оценивая все множители второго слагаемого снизу, мы получаем:

$$\begin{aligned}
\mathbb{E}[Q_n] - \mathbb{E}[Q'_n] &\leqslant 2n - 2n\left(1 - \frac{n-1}{N}\right)^n \\
&= 2n\left(\frac{n-1}{N}\right)\left(1 + \left(1 - \frac{n-1}{N}\right) + \cdots + \left(1 - \frac{n-1}{N}\right)^{n-1}\right) \\
&\leqslant 2n\left(\frac{n-1}{N}\right) n \\
&\leqslant 2\frac{n^3}{N}.
\end{aligned}$$

■

Обсуждение неравенств. Приведем короткий анализ и сравнение трех неравенств концентрации для Q'_n, представленных в настоящем разделе. Сначала мы сравним новые результаты Теорем 20 и 22 с неравенством Следствия 7. Затем приведем их сравнение с неравенством Буске Теоремы 13 для случайной величины Q_n.

Неравенство Следствия 7, в отличие от результатов Теорем 20 и 22, не учитывает дисперсий случайных величин и основано лишь на предположении об ограниченности функций из $\mathcal{F}$. Тем не менее, применение неравенства МакДиармида Теоремы 9 для случайной величины Q_n (для выборок *с возвращениями*) дает более слабый результат, в особенности для размеров выборки n, близких к N. Этот эффект, уже обсуждавшийся в прошлом разделе, ведет к более сильной концентрации случайной величины Q'_n по сравнению с Q_n.

Неравенства Теорем 20 и 22 учитывают дисперсию случайных величин. Первое является субгауссовским неравенством и дает совпадающие верхние оценки для $Q'_n - \mathbb{E}[Q'_n]$ и $\mathbb{E}[Q'_n] - Q'_n$. Сравним результаты Следствия 7 и Теоремы 20, немного переписав соответствующие неравенства:

$$\begin{aligned}
\mathbb{P}\left\{Q'_n - \mathbb{E}[Q'_n] \geqslant \varepsilon\right\} &\leqslant \exp\left\{-\frac{\varepsilon^2}{2n}\left(\frac{N - 1/2}{N - n}\right)\right\}; \\
\mathbb{P}\left\{Q'_n - \mathbb{E}[Q'_n] \geqslant \varepsilon\right\} &\leqslant \exp\left\{-\frac{\varepsilon^2}{8n\sigma_{\mathcal{F}}^2} \cdot \frac{n}{N}\right\}.
\end{aligned}$$

Для $n = o(N)$ (случай, который можно интерпретировать как переход к выборкам *с возвращениями*) первое неравенство, очевидно, дает более точный результат. Однако, в случае $n = O(N)$ ситуация меняется. Так, при $n = N/2$, результат сравнения зависит от соотношения величин $-\varepsilon^2/n$ и $-\varepsilon^2/(16\, n\sigma_{\mathcal{F}}^2)$, и неравенство Теоремы 20 оказывается точнее уже при $\sigma_{\mathcal{F}}^2 < 1/16$.

Теорема 22 является непосредственным аналогом неравенства Буске Теоремы 13. Она утверждает, что верхняя оценка для Q_m, вытекающая из неравенства Буске, также справедлива для Q'_m. Сравним Теоремы 20 и 22. Для начала заметим, что субгауссовское неравенство (2.8) Теоремы 20, в отличие от Теоремы 22, не содержит под корнем члена $2\mathbb{E}[Q_m] \geqslant 0$, что, как мы увидим в дальнейших главах, может вести к лучшим константам. Кроме того, Теорема 20 предоставляет верхние оценки для обоих отклонений $Q'_m - \mathbb{E}[Q'_m]$ и $\mathbb{E}[Q'_m] - Q'_m$, в то время как Теорема 22 позволяет оценивать сверху лишь правый хвост. Наконец, Теорема 22 контролирует отклонение случайной величины не от своего математического ожидания $\mathbb{E}[Q'_n]$, а от превосходящей его величины $\mathbb{E}[Q_n]$. Главным же недостатком Теоремы 20, как объясняется дальше, является наличие множителя N в показателе экспоненты.

Типичным порядком величины $\mathbb{E}[Q_n]$ является $\sqrt{n}$ (например, для случая конечного множества $\mathcal{F} = \{f_1, \ldots, f_m\}$ это вытекает из Теорем 2.1 и 3.5 работы [48]). Кроме того, справедлива очевидная оценка $\sigma_{\mathcal{F}}^2 \leqslant 1/4$. Наконец, последовательно воспользовавшись элементарными неравенствами $\sqrt{a+b} \leqslant \sqrt{a}+\sqrt{b}$ и $\sqrt{ab} \leqslant (a+b)/2$, правую часть неравенства (2.17) можно оценить сверху с помощью выражения $2\mathbb{E}[Q_n] + \sqrt{2n\sigma_{\mathcal{F}}^2 t} + Ct$ для некоторой константы C. Объединяя три этих шага, мы приходим к выводу, что Теорема 22 ведет к оценке порядка $\sqrt{n}$. В то же время, учитывая неравенство $\mathbb{E}[Q'_n] \leqslant \mathbb{E}[Q_n]$, Теорема 20 дает оценку, порядок которой не превосходит $\sqrt{n}+\sqrt{N}$. Для $n = o(N)$ результат Теоремы 22 является более предпочтительным. В противном случае, Теорема 20 может вести к лучшим результатам (в особенности при большом «зазоре» между величинами $\mathbb{E}[Q_n]$ и $\mathbb{E}[Q'_n]$).

Теперь вкратце обсудим связь новых результатов для Q'_n с неравенством Буске Теоремы 13 для случайной величины Q_n. Лемма 10 показывает, что для $n = o(N^{2/5})$ порядок разности $\mathbb{E}[Q_m] - \mathbb{E}[Q'_m]$ не превосходит $\sqrt{n}$. Следовательно, по крайней мере в этом случае необходимость использовать $\mathbb{E}[Q_n]$ для оценки сверху величины Q'_n в Теореме 22 не ведет к ограничению области ее применимости: новая теорема дает оценки отклонения Q'_n от $\mathbb{E}[Q'_n]$ в точности совпадающие с неравенством Буске. Таким образом, если $n = O(N)$ или $n = o(N^{2/5})$, результаты настоящего параграфа дают возможность контролировать отклонения Q'_m от $\mathbb{E}[Q'_m]$ с помощью неравенств, идентичных существующим неравенствам для Q_n.

Возможность получения неравенств того же порядка для других значений n и N является открытым вопросом.

Подводя итог, отметим, что для больших N и $m = o(N)$, Теорема 22 и Следствие 7 (в зависимости от максимальной дисперсии $\sigma_{\mathcal{F}}^2$ и порядка величины $\mathbb{E}[Q_n] - \mathbb{E}[Q'_n]$) могут давать более точные результаты, чем Теорема 20. Однако, для $n = O(N)$, Теорема 20 является более предпочтительной.

Замечание 11. *Все результаты, приведенные в настоящем разделе, могут быть обобщены на случайную величину* $\sup_{f \in \mathcal{F}} |\sum_{i=1}^{n} f(\eta_i)|$ *на основе тех же подходов.*

Замечание 12. *К сожалению, неравенства для* Q'_n *Теорем 20 и 22, в отличие от Следствия 7, не ведут к более точным оценкам по сравнению с известными результатами для* Q_n *при росте* $n \to N$*. Возможность получения неравенства концентрации для выборок без возвращений, равномерно улучшающего неравенство Талаграна при росте размера выборки* $n \to N$*, является интересным направлением дальнейших исследований. Этот вопрос, по-видимому, является непростым и ведет к другому более общему вопросу о возможности применения энтропийного метода М. Леду для выборок без возвращений.*

Применение полученных неравенств. Неравенства Теорем 20 и 22, полученные в прошлом разделе, могут играть важную роль в различных задачах машинного обучения и теории статистического обучения. Одним примером является *трансдуктивная постановка* теории статистического обчения, которую мы рассмотрим в Главе 4.

Другим чрезвычайно важным примером является неасимптотический анализ процедуры *скользящего контроля* [92], где обучающие и контрольные подвыборки выбираются без возвращений из конечного множества объектов. Другими приложениями, где эмпирические процессы для выборок без возвращений могут представлять интерес, являются следующие: анализ процедур представления матриц в виде произведения множителей с низким рангом; анализ метода Нистрема [54], ведущего к эффективной процедуре аппроксимации матриц малого ранга в задачах с большими объемами данных; анализ стохастических алгоритмов оптимизации, таких как метод стохастического градиента, в практических реализациях которых часто применяются выборки без возвращений [82]. Наконец, интересным направлением является попытка обобщения матричных неравенств Бернштейна [97] на случай выборки без возвращений на основе подходов, изложенных в настоящей главе. Подобные результаты могли бы применяться в задачах восстановлени матриц малого ранга [53]. Однако, все упо-

мянутые выше задачи выходят за рамки настоящей работы и являются предметом будущих исследований.

Коротко подведем итоги настоящей главы:

- Неравенства концентрации можно получать не только для независимых случайных величин, но также и для зависимых. В частности — для случайных величин, выбранных веб возвращений из произвольных конечных множеств.
- Все результаты для сумм независимых случайных величин (включая неравенства Хефдинга, Беннетта и Бернштейна) также справедливы для сумм случайных величин, выбранных без возвращений.
- Метод Хефдинга дает простой подход к получению неравенств концентрации для выборок без возарвщений, позволяя свести задачу к рассмотрению выборок с возвращениями. Однако, отказ от метода Хефдинга может вести к более точным результатам.
- Неравенства Стирлинга и Эль-Янива -Печиони показывают, что при выборе случайных величин без возвращений наблюдается более сильная концентрация, чем при выборе с возвращениями.

В настоящей главе получены следующие новые результаты:

- Два новых неравенства концентрации для супремумов эмпирических процессов и выборок без возвращений (Теоремы 20 и 20). Также получена новая модицифированная версия неравенства С. Г. Бобкова (Теорема 21).

3 Теория статистического обучения

В настоящей главе мы приведем достаточно подробный обзор классических и современных результатов *теории статистического обучения* [18, 104].

Теория статистического обучения изучает свойства процедуры поиска закономерностей в наблюдаемых данных в рамках строгого математического подхода. Впервые вероятностная постановка задачи обучения была предложена и изучена в конеце 1960-х годов в работах В. Н. Вапника и А. Я. Червоненкиса [104–106, 108]. Подход, развитый в этих работах, позже получил название *VC-теория* и продолжает активно развиваться и на сегодняшний день[1].

Важной отличительной чертой VC-теории является существенное использование при анализе процедур обучения результатов теорий эмпирических процессов [102] и неравенств концентрации вероятностной меры. Естественным образом вводя в рассмотрение равномерные отклонения (супремумы эмпирических процессов), VC-теория получает возможность использовать богатый и хорошо развитый математический аппарат. Позже бурное развитие теорий эмпирических процессов и открытие новых мощных неравенств концентрации получило отражение в новых плодотворных подходах VC-теории. Фактически, три этих области были с самого начала чрезвычайно тесно связаны между собой, и прорывы в одной из них, как правило, вели к новым результатам в двух других.

Далее мы приведем достаточно подробный обзор ключевых результатов VC-теории. Приведенный обзор не в коем случае не является исчерпывающим, и ряд интересных подходов останутся за рамками рассмотрения настоящей работы. В частности, мы не будем рассматривать обширную область оценок, основанных на отступах объектов (*large margin bounds*), включающую результаты работ [49, 84].

[1] Стоит отметить, что позже в середине 1980-х годов Л. Валиантом была предложена альтернативная и тесно связанная с VC-теорией постановка, получившее название *PAC-обучение* [99]. Эта постановка также получила большую популярность. Сегодня термины VC-теория и PAC-обучение (с точностью до небольших отличий) используются в литературе как синонимы.

3.1 Определения и постановки задач

Рассмотрим множество *объектов* $\mathcal{X}$ и множество *ответов* $\mathcal{Y}$. Пусть на декартовом произведении $\mathcal{X} \times \mathcal{Y}$ задано неизвестное нам вероятностное распределение P. Обозначим (X, Y) случайную пару из распределения P. Пусть нам дана *обучающая выборка* $S = \{(X_i, Y_i)\}_{i=1}^n$, состоящая из n *независимых* наблюдений случайной пары (X, Y). Всюду далее X будет играть роль наблюдаемого объекта, в то время как Y будет описывать некоторое интересующее нас свойство этого объекта. Нашей главной задачей будет предсказание значения Y по объекту X. Типичными примерами являются задачи *регрессии*, когда $\mathcal{Y} = \mathbb{R}$, и задачи *классификации*, когда множество Y конечно. *Бинарной классификацией* называется задача классификации с двумя классами $\mathcal{Y} = \{0, 1\}$ (иногда удобнее будет полагать $\mathcal{Y} = \{-1, +1\}$).

Заметим, что частным случаем описанной *вероятностной* постановки является ситуация, когда между объектами X и ответами Y существует строгая функциональная зависимость: $Y = F(X)$ с вероятностью 1 для некоторой функции $F\colon \mathcal{X} \to \mathcal{Y}$. В этом случае условное распределение $P(Y|X)$ целиком концентрируется в одной точке: $P\big(Y = f(x)|X = x\big) = 1$ почти наверное. Однако, при решении прикладных задач все измеримые величины так или иначе подвержены случайному шуму. Кроме того, в большинстве современных прикладных задач однозначного соответствия между наблюдаемыми признаками X объекта и интересующим нас свойством Y этого объекта может вообще не существовать. Примерами таких задач могут служить задачи предсказания возможности рецидива после операции по ряду медицинских характеристик, собранных у пациента, и задачи предсказания перехода абонента к другому оператору мобильной связи по его абонентской истории (предсказание оттока клиентов, *churn prediction*). Очевидно, что однозначного ответа на вопрос в обоих примерах нет. Тем не менее, в подобных случаях можно пытаться предсказывать значение Y, допуская при этом как можно меньше ошибок. По этой причине в настоящей главе мы будем рассматривать общую вероятностную постановку.

В отличие от математической статистики, классической задачей которой является оценивание функции условного распределения $P(Y|X = x)$, теория статистического обучения занимается поиском на основе обучающей выборки S отображения[2] $h_n\colon \mathcal{X} \to \mathcal{Y}$, пригодного для предсказания ответа Y на новом объекте X. Эту задачу принято называть задачей *обучения с учителем*. Критерий качества найденного отображения h_n мы определим с помощью неотрицательной *функции потерь* $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$. Значение $\ell(y, y')$ для $y, y' \in \mathcal{Y}$

[2] Всюду далее нижним индексом n мы будем подчеркивать факт зависимости рассматриваемой величины от обучающей выборки S.

выражает величину потерь (или «штрафа») при замене правильного ответа y на ответ y'. Потери при использовании отображения h для предсказания ответа на паре (X, Y) мы будем обозначать $\ell_h(X, Y) = \ell\big(Y, h(X)\big)$.

Минимизация среднего риска. *Средним риском* отображеня h мы будем называть математическое ожидание потерь этого отображения:

$$L(h) = \mathbb{E}_{(X,Y)\sim P}\left[\ell_h(X, Y)\right].$$

В зависимости от множества $\mathcal{Y}$ часто используют следующие функции потерь:

- **В задачах классификации** часто используют *бинарную функцию потерь:*

  $$\ell(y, y') = \mathbb{1}\{y \neq y'\},$$

  штрафующую единицей неправильную классификацию объекта. В этом случае, как несложно убедиться, средний риск $L(h)$ отображения h равен вероятности ошибки этого отображения $L(h) = P(h(X) \neq Y)$. Заметим, что эта функция потерь симметрична $\ell(y, y') = \ell(y', y)$, $y, y' \in \mathcal{Y}$. Кроме того, для любых $y, y', y'' \in \mathcal{Y}$, таких что $y' \neq y$ и $y'' \neq y$, справедливо $\ell(y, y') = \ell(y, y'')$. В некоторых случаях ошибочное отнесение объекта класса y к классу y' может оказаться менее предпочтительным, чем к классу y''. Тогда целесообразно задавать функцию потерь с помощью (в общем случае несимметричной) квадратной матрицы из $\mathbb{R}^{M\times M}$.

  Другими примерами функций потерь, часто используемых в задачах бинарной классификации $\mathcal{Y} = \{-1, +1\}$, являются *экспоненциальная* функция потерь $\ell(y, y') = e^{-yy'}$ (используемая в методе *бустинга* [36, 84]), *логистическая* функция потерь $\ell(y, y') = \log_2(1+e^{-yy'})$ (в *логистической регрессии* [41]) и функция потерь $\ell(y, y') = \max(0, 1-yy')$ (в методе опорных векторов [26,28,41]). Обратим внимание на то, что все перечисленные функции являются верхними оценками бинарной функции потерь.

- **В задачах регрессии** бинарная функция потерь может вести к необоснованно жестким критериям. Вместо этого принято использовать *квадратичную функцию потерь*:

  $$\ell(y, y') = (y - y')^2$$

  либо одну из функций потерь, задаваемых параметром $q \geqslant 1$:

  $$\ell(y, y') = |y - y'|^q.$$

Имея критерий, описанный выше, основную задачу теории статистического обучения можно сформулировать как *поиск отображений с малым средним риском на основе обучающей выборки* S. В частности, можно заняться поиском оптимального отображения g^*, имеющего *минимальный средний риск* $L(g^*)$ среди всех измеримых отображений $\mathcal{X} \to \mathcal{Y}$ (заметим, что такое отображение g^* может оказаться не единственным). Такое отображение g^* принято называть *Байесовским*. Средний риск Байесовского отображения называется *Байесовским риском*, и мы будем обозначать его $L^* = L(g^*)$. В некоторых случаях Байесовское отображение может быть найдено аналитически, например:

- **В задачах регрессии** с квадратичной функцией потерь справедливо следующее тождество:

$$L(h) = \mathbb{E}_{(X,Y)\sim P}\left[\big(Y - h(X)\big)^2\right] = \mathbb{E}\left[\big(Y - \eta(X)\big)^2\right] + \mathbb{E}\left[\big(h(X) - \eta(X)\big)^2\right],$$

  где $\eta(x) = \mathbb{E}[Y|X = x]$ — *функция регрессии*. Поскольку первое слагаемое в правой части последнего тождества не зависит от выбора отображения h, минимум среднего квадратичного риска достигается на функции регрессии и $L^* = L(\eta)$.

- **В задачах бинарной классификации** с бинарной функцией потерь можно показать (см. [31, Теорема 2.1]), что минимум среднего риска достигается на функции

$$\eta^*(x) = \mathbb{1}\left\{P(Y = 1|X = x) > 1/2\right\}.$$

Эти примеры показывают, что задача минимизации среднего риска является менее общей, чем восстановление функции условного распределения $P(Y|X = x)$, поскольку знание последней позволяет немедленно получить отображения η и η^*. В то же время к минимизации среднего риска сводятся многие важные прикладные задачи, включая задачи распознавания образов и восстановления регрессии, задачи оценивания плотности распределения (которые относятся к классу задач обучения *без учителя*) и многие другие. Возникает вопрос: зачем в качестве промежуточного шага решать более сложную задачу восстановления условного распределения $P(Y|X = x)$? Эта идея является частным случаем следующего методологического принципа, приведенного В. Н. Вапником в [104] и играющего важную роль в теории статистического обучения:

> При решении задачи в условиях ограниченного количества доступной информации, ни в коем случае не следует решать более общую задачу в качестве промежуточного шага. Доступной информации может оказаться достаточно для решения исходной задачи, но недостаточно для решения более общей задачи.

Минимизация эмпирического риска. На практике минимизация среднего риска по классу всех измеримых отображений $\mathcal{X} \to \mathcal{Y}$ часто по ряду причин, описанных далее, не представляется возможной. Вместо этого мы будем оптимизировать средний риск на заранее выбранном (не зависящем от обучающей выборки S) достаточно большом классе отображений $\mathcal{H} = \{h\colon \mathcal{X} \to \mathcal{Y}\}$:

$$L(h) \to \min_{h \in \mathcal{H}}. \tag{3.1}$$

Всюду далее мы будем предполагать, что минимум достигается, и решение задачи (3.1) мы будем обозначать с помощью h^*: $L(h^*) = \min_{h \in \mathcal{H}} L(h)$. В том случае, когда задача (3.1) имеет множество решений, с помощью h^* будем обозначать любое отображение, такое что $L(h^*) \leqslant L(h)$ для всех $h \in \mathcal{H}$.

Распределение P нам неизвестно, а значит мы не можем вычислять средний риск $L(h)$. Одним из естественных способов оценки величины $L(h)$ является использование *эмпирического риска* $L_n(h) = \frac{1}{n}\sum_{i=1}^{n} \ell_h(X_i, Y_i)$. Действительно, если отображение h не зависит от обучающей выборки S, то случайные величины $\ell_h(X_i, Y_i)$ для $i = 1, \ldots, n$ независимы и одинаково распределены, и справедливо тождество $\mathbb{E}_{S \sim P^n}[L_n(h)] = L(h)$[3]. Тогда, согласно усиленному закону больших чисел [35], эмпирический риск L_n сходится к среднему риску $L(h)$ почти наверное. По этой причине часто задачу (3.1) заменяют на задачу *минимизации эмпирического риска* (МЭР):

$$L_n(h) \to \min_{h \in \mathcal{H}}, \tag{3.2}$$

решение которой мы будем обозначать $\hat{h}_n$[4].

Поскольку мы заменили интересующую нас задачу минимизации среднего риска другой, то возникает следующий вопрос: насколько хорошо решение $\hat{h}_n$ задачи (3.2) приближает оптимальное решение h^* изначальной задачи (3.1)? При этом нас будет интересовать близость отображения $\hat{h}_n$ к h^* исключительно в смысле близости значений среднего риска этих отображений. Этому вопросу посвящена большая часть теории статистического обучения, и достаточно полные обзоры известных результатов могут быть найдены в книгах [31, 104], работах [18, 23, 55, 65, 115, 117] и лекциях [48, 64].

[3] Здесь и далее с помощью P^n мы будем обозначать закон распределения выборки, состоящей из n независимых и одинаково распределенных в соответствие с P случайных величин.

[4] Поскольку чаще всего оптимизируемый в МЭР функционал не является выпуклым, поиск отображения $\hat{h}_n$ может оказаться сложной задачей. Большая часть результатов машинного обучения посвящена именно этому вопросу [14, 41].

Выбор класса $\mathcal{H}$. Вернемся к поставленной изначально задаче поиска Байесовского отображения g^*. Мы можем записать следующее тождество:

$$L(\hat{h}_n) - L^* = \Big(L(\hat{h}_n) - \inf_{h\in\mathcal{H}} L(h)\Big) + \Big(\inf_{h\in\mathcal{H}} L(h) - L^*\Big). \tag{3.3}$$

Первое слагаемое последнего тождества принято называть *ошибкой оценивания* (estimation error), а второе — *ошибкой аппроксимации* (approximation error). Чем больше класс отображений $\mathcal{H}$, тем меньше ошибка аппроксимации (величина $L(h^*)$ приближается к Байесовскому риску L^*), и тем больше ошибка оценивания (отображение h^* становится все сложнее оценить по обучающей выборке). И наоборот. При слишком «большом» классе $\mathcal{H}$ ошибка аппроксимации может почти исчезнуть, однако мы можем столкнуться с проблемой *переобучения* (overfitting), когда отображение $\hat{h}_n$ с малым значением эмпирического риска имеет большой средний риск $L(\hat{h}_n)$. Эту ситуацию традиционно принято иллюстрировать следующим примером. Рассмотрим задачу бинарной классификации $\mathcal{Y} = \{0, 1\}$ и класс всех бинарных отображений $\mathcal{H}$. Отображение $\hat{h}_n$, не допускающее ошибок на объектах $X_1, \ldots, X_n$ обучающей выборки и возвращающее 0 на всех остальных объектах X, очевидно, имеет наименьший возможный эмпирический риск $L_n(\hat{h}_n) = 0$. Однако, такое отображение непригодно для предсказания ответов на новых объектах.

Как правило, выбор класса $\mathcal{H}$ осуществляется на основе здравого смысла и априорной информации о специфике решаемой задаче.

Вероятностные оценки. Изучение второго слагаемого в разложении (3.3) как правило требует введения дополнительных предположений о рассматриваемом классе задач, в том числе о распределении P и Байесовском отображении g^*. Мы будем по возможности избегать подобных ограничений и сосредоточимся в настоящей главе на ошибке оценивания — величине $L(\hat{h}_n) - L(h^*)$.

Получив решение $\hat{h}_n$ задачи минимизации эмпирического риска (3.2), мы хотели бы оценить средний риск найденного решения $L(\hat{h}_n)$, а также понять, насколько эта величина превосходит оптимальное значение $L(h^*)$. Поскольку отображение $\hat{h}_n$ зависит от обучающей выборки S и $L(\hat{h}_n) = \mathbb{E}\big[\ell_{\hat{h}_n}(X, Y)|S\big]$, величина $L(\hat{h}_n)$ (как и $L_n(\hat{h}_n)$) является случайной. Поэтому большинство утверждений об отображении $\hat{h}_n$, представленных в настоящей главе, будут нести вероятностный характер. В частности, нас будут интересовать результаты следующего вида:

- **Оценки обобщающей способности.** Величину $L(\hat{h}_n) - L_n(\hat{h}_n)$ принято называть *обобщающей способностью* (generalization error bound, error bound), поскольку она из-

меряет способность метода минимизации эмпирического риска выбирать на основе обучающей выборки S действительно хорошие отображения. Нас будут интересовать неравенства следующего вида:

$$\begin{aligned} \mathbb{P}\left\{L(\hat{h}_n) - L_n(\hat{h}_n) \geqslant \varepsilon\right\} &\leqslant B_1(\varepsilon, n, \mathcal{H}); \\ \mathbb{P}\left\{|L(\hat{h}_n) - L_n(\hat{h}_n)| \geqslant \varepsilon\right\} &\leqslant B_1(\varepsilon, n, \mathcal{H}) \end{aligned} \tag{3.4}$$

для неотрицательных отклонений $\varepsilon \geqslant 0$ и некоторой неотрицательной, непрерывной и монотонно убывающей по первому аргументу функции B_1.

- **Оценки избыточного риска.** Избыточным риском отображения $h \in \mathcal{H}$ мы будем называть величину $\mathcal{E}(h, \mathcal{H}) = L(h) - \inf_{g \in \mathcal{H}} L(g)$. Избыточный риск показывает, насколько средний риск отображения h превосходит оптимальное по классу $\mathcal{H}$ значение $L(h^*)$. Для краткости всюду далее мы будем писать просто $\mathcal{E}(h)$, когда класс $\mathcal{H}$ понятен из контекста. Очевидно, $\mathcal{E}(h, \mathcal{H}) \geqslant 0$ для всех $h \in \mathcal{H}$. Нас будут интересовать неравенства следующего вида:

$$\mathbb{P}\left\{\mathcal{E}(\hat{h}_n) \geqslant \varepsilon\right\} \leqslant B_2(\varepsilon, n, \mathcal{H}) \tag{3.5}$$

для неотрицательных отклонений $\varepsilon \geqslant 0$ и некоторой неотрицательной, непрерывной и монотонно убывающей по первому аргументу функции B_2.

Мы увидим, что на основе оценок (3.4) и (3.5) мы можем получить функции $C_1(n, \mathcal{H})$ и $C_2(n, \mathcal{H})$, такие что следующие неравенства

$$L(\hat{h}_n) \leqslant L_n(\hat{h}_n) + C_1(n, \mathcal{H}), \quad L(\hat{h}_n) \leqslant L(h^*) + C_2(n, \mathcal{H}),$$

выполняются с большой вероятностью.

Величины $B(\varepsilon, n, \mathcal{H})$, стоящие в правых частях неравенств (3.4) и (3.5) (а также величины $C(n, \mathcal{H})$), мы будем часто называть просто *оценками*. Главной целью рассматриваемых далее результатов является получение как можно более точных оценок обобщающей способности и избыточного риска. В настоящей главе мы приведем короткий обзор известных результатов для *ограниченных функций потерь*. Приведенные далее оценки будут учитывать «геометрическую структуру» класса $\mathcal{H}$ и убывать к нулю при росте размера обучающей выборки $n \to \infty$. Кроме того, значения оценок, как правило, будут расти с ростом «сложности» класса $\mathcal{H}$. Мы увидим, что некоторые из оценок будут зависеть также от *неизвестных* свойств рассматриваемой задачи (в том числе от свойств распределения P). Такие оценки мы будем называть *невычислимыми*, поскольку их невозможно вычислить на основе наблюдаемых данных. Другие оценки, наоборот, будут *вычислимыми*: то есть зависеть только от наблюдаемых величин.

Задача выбора модели. Часто при решении задачи минимизации среднего риска на большом классе $\mathcal{H}$ для него вводят счетное (или даже конечное) семейство подклассов $\mathcal{H}_\alpha \subset \mathcal{H}$, $\alpha \in \mathcal{A}$, и рассматривают семейство задач МЭР:

$$L_n(h) \to \min_{h \in \mathcal{H}_\alpha}, \quad \alpha \in \mathcal{A}.$$

Подклассами могут быть как совершенно различные модели (полиномы, тригонометрические функции), так и похожие модели, отличающиеся своими структурными параметрами (классы полиномов разных степеней). Часто рассматривают монотонные семейства подклассов $\{\mathcal{H}_k,\ k = 1, 2, \dots\}$, такие что $\mathcal{H}_k \subset \mathcal{H}_{k+1}$, $k \geqslant 1$ и $\mathcal{H} = \bigcup_{j \geqslant 1} \mathcal{H}_j$. Решение соответствующей задачи МЭР мы будем обозначать $\hat{h}_{n,\alpha}$, так что $L_n(\hat{h}_{n,\alpha}) = \min_{h \in \mathcal{H}_\alpha} L_n(h)$.

Задача выбора модели заключается в построении вычислимого по обучающей выборке индекса модели $\hat{\alpha} = \hat{\alpha}(S) \in \mathcal{A}$, такого что отображение $\hat{h}_{n,\hat{\alpha}}$ имеет близкий к оптимальному избыточный риск $\mathcal{E}(\hat{h}_{n,\hat{\alpha}}, \mathcal{H})$. Одним из известных подходов к решению этой задачи является метод *структурной минимизации риска* (structural risk minimization, SRM, [104, Глава 6]). Его основная идея заключается в поиске компромисса между минимальным эмпирическим риском, достигаемым на подклассе, и размером этого подкласса:

$$L_n(\hat{h}_{n,\alpha}) + C(\alpha, n) \to \min_{\alpha \in \mathcal{A}}, \tag{3.6}$$

где функция C измеряет «сложность» модели $\mathcal{H}_\alpha$ (и не зависит от обучающей выборки S). Чем богаче класс $\mathcal{H}_\alpha$, тем меньше эмпирический риск $L_n(\hat{h}_{n,\alpha})$, и тем больше *сложностное* слагаемое C. В частности, для построения функции $C(\alpha, n)$ можно использовать оценки обобщающей способности и избыточного риска (3.4) и (3.5). Функция $C(\alpha, n)$ может зависеть от таких структурных характеристик класса $\mathcal{H}_\alpha$, как его VC-размерность, Радемахеровская сложность или метрическая энтропия (определения будут даны далее).

Задача выбора модели выходит за рамки рассмотрения настоящей работы, и подробные обзоры способа построения функций $C(\alpha, n)$, индекса $\hat{\alpha}$ и анализа свойств отображения $\hat{h}_{n,\hat{\alpha}}$ могут быть найдены в лекциях [48] и [64].

Использование оценок. Наконец, вкратце обозначим возможные способы использования оценок обобщающей способности и избыточного риска:

- **Измерение качества решения.** Оценки (в особенности измеримые) дают возможность оценивать качество найденного в результате обучения отображения h_n, основываясь на его эмпирическом риске $L_n(h_n)$, что часто может оказаться полезным при

решении прикладных задач. Хотя большинство известных оценок обобщающей способности и избыточного риска сильно завышены, целое направление в теории статистического обучения последних лет посвящено их уточнению. Так, точность оценок некоторых современных подходов (например, некоторых *PAC-Байесовских* оценок, которые мы рассмотрим в Главе 6) вполне позволяет использовать их численные значения на практике.

- **Исследование состоятельности метода обучения.** Заметим, что для избыточного риска $\mathcal{E}(\hat{h}_n)$ справедливы следующие неравенства:

$$\begin{aligned}\mathcal{E}(\hat{h}_n) &= \mathcal{E}(\hat{h}_n) + L_n(\hat{h}_n) - L_n(h^*) - \big(L_n(\hat{h}_n) - L_n(h^*)\big) \\ &\leqslant \mathcal{E}(\hat{h}_n) - \big(L_n(\hat{h}_n) - L_n(h^*)\big) \\ &= L(\hat{h}_n) - L_n(\hat{h}_n) + L_n(h^*) - L(h^*),\end{aligned} \tag{3.7}$$

где мы воспользовались определением $\hat{h}_n$. Кроме того, избыточный риск неотрицателен: $\mathcal{E}(\hat{h}_n) \geqslant 0$. Поскольку h^* не зависит от обучающей выборки, по закону больших чисел последовательность случайных величин $L_n(h^*)$ сходится по вероятности к $L(h^*)$. Если оценка $B_1(\varepsilon, n, \mathcal{H})$ в (3.4) для любого $t > 0$ убывает к нулю при $n \to \infty$, то мы немедленно получаем сходимость по вероятности последовательности случайных величин $L_n(\hat{h}_n)$ к $L(h^*)$, а значит и случайных величин $L(\hat{h}_n)$ к $L(h^*)$ (это следует, например, из [102, Теоремы 2.3 и 2.7]). Таким образом, в этом случае при неограниченном росте размера обучающей выборки n метод МЭР ведет к оптимальному решению h^*. Оценка $B_1(\varepsilon, n, \mathcal{H})$ также дает скорость этой сходимости.

  Описанное свойство *метода обучения* — отображения, ставящего в соответствие обучающей выборке S длины n и классу $\mathcal{H}$ отображение h_n из него, — принято называть его *состоятельностью*. Обзор различных определений состоятельности может быть найден, например, в [21, Раздел 1.2.3]. Другим определением состоятельности может служить сходимость последовательности $L(\hat{h}_n)$ к Байесовскому риску L^*.

- **Использование для выбора модели.** Мы также видели, что оценки обобщающей способности и избыточного риска могут использоваться при выборе модели. Например, если мы положим в качестве штрафующего слагаемого $C(\alpha, n)$ в (3.6) оценку $B_1(\varepsilon, n, \mathcal{H})$, мы будем выбирать модель $\mathcal{H}_\alpha$ с минимальной среди всех прочих верхней оценкой среднего риска $L(\hat{h}_{n,\alpha})$.

- **Построение новых методов обучения.** В настоящей главе изучается метод МЭР и все представленные оценки связаны с характеризующей его величиной $L(\hat{h}_n)$. Тем

не менее, большинство подходов, используемых в настоящей главе, могут быть также применены для изучения свойств других методов обучения. В некоторых случаях сами вероятностные оценки могут нетривиальным образом вести к новым методам обучения (хорошими примерами служат результаты Главы 6 и работы [71]).

3.2 Обзор известных результатов

Для фиксированного (не зависящего от обучающей выборки) отображения h вопрос об отклонении эмпирического риска $L_n(h)$ от среднего риска $L(h)$ изучать достаточно легко. Действительно, поскольку справедливо[5] $\mathbb{E}[L_n(h)] = L(h)$, а $\ell_h(X_1, Y_1), \ldots, \ell_h(X_n, Y_n)$ — независимые и одинаково распределенные случайные величины, то вопрос сводится к анализу отклонения математического ожидания случайной величины от ее среднего выборочного значения, посчитанного по простой выборке. В первой главе был приведен обширный обзор результатов, позволяющих изучать данный вопрос. В частности, неравенства концентрации для сумм (неравенства Хефдинга или Бернштейна), приведенные в разделе 1.1, дают оценки величины $L(h) - L_n(h)$, убывающие со скоростью порядка $O(n^{-1/2})$ при росте размера обучающей выборки n (точные формулировки и доказательства мы приведем ниже).

Однако, нас интересует отображение $\hat{h}_n$, на котором достигается минимум эмпирического риска $L_n(h)$. Если класс $\mathcal{H}$ включает более одного отображения, то минимизатор эмпирического риска $\hat{h}_n$ — случайное отображение, которое зависит от обучающей выборки S. Этот факт, в частности, ведет к тому, что потери отображения $\hat{h}_n$ на обучающей выборке S перестают быть независимыми. Таким образом, неравенства концентрации, основанные на предположении о независимости случайных величин, уже неприменимы.

Классический подход, развитый В. Н. Вапником и А. Я. Червоненкисом [106, 107], основан на следующем неравенстве:

$$\forall h \in \mathcal{H}: \quad L(h) - L_n(h) \leqslant \sup_{g \in \mathcal{H}} L(g) - L_n(g). \tag{3.8}$$

Переходя к рассмотрению *равномерного по классу отображений* $\mathcal{H}$ отклонения средних выборочных от математических ожиданий, мы получаем возможность контролировать в том числе и обобщающую способность $L(\hat{h}_n) - L_n(\hat{h}_n)$, поскольку, очевидно, следующее неравенство выполнено с вероятностью 1:

$$L(\hat{h}_n) - L_n(\hat{h}_n) \leqslant \sup_{h \in \mathcal{H}} L(h) - L_n(h). \tag{3.9}$$

[5] Далее мы будем опускать нижний индекс математического ожидания всюду, где мера, по которой оно вычисляется, понятна из контекста.

Кроме того, из приведенной ранее цепочки неравенств (3.7) следует следующее неравенство[6] для избыточного риска $\mathcal{E}(\hat{h}_n)$:

$$\mathcal{E}(\hat{h}_n) \leqslant \sup_{h\in\mathcal{H}} L(h) - L_n(h) + L_n(h^*) - L(h^*). \tag{3.10}$$

Поскольку h^* не зависит от обучающей выборки S, слагаемое $L_n(h^*) - L(h^*)$ мы можем оценить с помощью неравенств Хефдинга или Бернштейна. Таким образом, задачи построения оценок обобщающей способности и избыточного риска сводятся к изучению *супремума эмпирического процесса:*

$$\sup_{h\in\mathcal{H}} L(h) - L_n(h)$$

— случайной величины, уже рассматривавшейся в прошлых главах. А именно, нашей целью является построение верхних оценок для этой случайной величины.

Замечание 13. *Задача изучения асимптотического поведения случайной величины* $\sup_{h\in\mathcal{H}} L(h) - L_n(h)$ *тесно связана с теоремой Гливенко–Кантелли [102]. Классы функций, для которых* $\sup_{h\in\mathcal{H}} L(h) - L_n(h)$ *стремится к нулю почти наверное, принято называть* классами Гливенко–Кантелли *и говорят, что для них выполнен* равномерный закон больших чисел. *Поиск необходимых и достаточных условий принадлежности множества отображений* $\mathcal{H}$ *к классу Гливенко–Кантелли является одной из центральных задач теории эмпирических процессов [101].*

Удобно выделять три группы подходов к построению верхних оценок для супремума эмпирического процесса, на которых основаны все приведенные в настоящей главе результаты.

Неравенство Буля. Исторически первый подход совмещает оценки индивидуальных отклонений $L(h) - L_n(h)$ для $h \in \mathcal{H}$ с неравенством Буля. Поскольку неравенство Буля справедливо для не более чем счетного класса событий, основная сложность подхода заключается в переходе от (в общем случае) несчетного класса отображений $\mathcal{H}$ к его счетному (или даже конечному) подмножеству. Этот шаг, как правило, использует различные модификации так называемого *неравенства симметризации* (см. Лумма 11) вместе с анализом метрических энтропий класса отображений $\mathcal{H}$.

[6] Заранее отметим, что приведенная цепочка неравенств часто ведет к очень грубым оценкам избыточного риска. Более точный подход, известный как *локализация оценок*, приведен далее.

Рademахеровская сложность. Второй подход, развитый позже в 2000-х в серии работ [7,9,51,52], состоит из двух шагов. Сначала с помощью неравенств концентрации для супремума эмпирических процессов (например, неравенства МакДиармида Теоремы 12) оценивается его отклонение от своего математического ожидания $\mathbb{E}\left[\sup_{h\in\mathcal{H}}|L(h)-L_n(h)|\right]$. Затем с помощью одной из версий неравенства симметризации математическое ожидание супремума эмпирического процесса ограничивается сверху математическим ожиданием супремума другого случайного процесса, называемого *Радемахеровским* (определения будут даны далее). Последняя величина известна в литературе как *Радемахеровская сложность* класса $\mathcal{H}$ и примечательна тем, что из теории эмпирических процессов [59,100,101] известно большое число полезных технических результатов для работы с ней. Кроме того, как мы увидим, использование Радемахеровской сложности ведет к универсальному и простому способу получения вычислимых оценок обобщающей способности и избыточного риска.

Локализация оценок. Последний третий подход был развит относительно недавно в работах [5,47,52,66]. Отправной точкой данного подхода служит следующее наблюдение: неравенство (3.8) является чересчур грубым. Действительно, если класс $\mathcal{H}$ достаточно велик, то для конкретной задачи (конкретного распределения P) минимизатор эмпирического риска $\hat{h}_n$ будет, как правило, принадлежать некоторому малому подмножеству класса $\mathcal{H}$ — подмножеству отображений, «пригодных» для решения данной задачи. Временно обозначим это подмножество $\mathcal{H}_0\subset\mathcal{H}$. При этом может оказаться, что супремум в (3.8) достигается на отображениях из $\mathcal{H}\setminus\mathcal{H}_0$, то есть той части класса $\mathcal{H}$, которой мы фактически «не пользуемся». Очевидно, справедливо $\sup_{h\in\mathcal{H}_0}|L(h)-L_n(h)|\leqslant\sup_{h\in\mathcal{H}}|L(h)-L_n(h)|$, и чем меньше подмножество $\mathcal{H}_0$, тем более грубым становится это неравенство. Оказывается, при правильном выборе подмножества $\mathcal{H}_0$ можно добиться порой очень существенных улучшений оценок. Наиболее сложной задачей при этом является построение как можно более узкого подмножества класса $\mathcal{H}$, в котором с большой вероятностью содержится отображение $\hat{h}_n$. Решение этой задачи, как правило, основано на использовании неравенства Талаграна, приведенного в разделе 1.3.2. Оценки, основанные на подобных идеях, принято называть *локальными*.

В настоящем разделе мы приведем краткий обзор основных результатов описанных выше подходов. Мы не претендуем на полноту обзора и, в частности, не будем рассматривать такие важные вопросы, как *нижние* и *минимаксные* оценки обобщающей способности и избыточного риска. Нашей главной целью будет описать возможные подходы к получению верхних оценок, привести примеры типичных результатов, к которым эти подходы ведут,

а также грубо сравнить эти результаты между собой. Более подробные обзоры могут быть найдены в книгах [31, 104], работах [18, 23, 55, 65, 115, 117] и лекциях [48, 64].

Всюду далее мы будем предполагать, что функция потерь ограничена: $\ell\colon \mathcal{Y} \times \mathcal{Y} \to [0, 1]$. Мы не теряем общности, ограничив функцию потерь единицей, поскольку в противном случае все результаты настоящего раздела справедливы после нормировки функции потерь.

3.2.1 Оценки, существенно опирающиеся на неравенство Буля

Мы отдельно рассмотрим случаи конечного, счетного и несчетного класса $\mathcal{H}$. Каждый последующий результат будет опираться на доказательство предыдущего.

Класс $\mathcal{H}$ содержит один алгоритм. Мы начнем с рассмотрения элементарного случая, когда класс $\mathcal{H}$ состоит из одного отображения: $\mathcal{H} = \{h_0\}$. В этом случае, очевидно, минимизатор эмпирического риска совпадает с ним: $\hat{h}_n = h_0$. Следующий результат является простым следствием неравенства Хефдинга:

Теорема 23 (Класс состоит из одного отображения)**.** *Пусть класс $\mathcal{H} = \{h_0\}$ содержит лишь одно отображение и функция потерь ограничена в отрезке $[0, 1]$. Тогда для любого $\varepsilon > 0$ справедливо:*

$$\mathbb{P}\left\{\sup_{h\in\mathcal{H}} L(h) - L_n(h) \geqslant \varepsilon\right\} = \mathbb{P}\left\{L(h_0) - L_n(h_0) \geqslant \varepsilon\right\} \leqslant e^{-2n\varepsilon^2}.$$

Кроме того, для любого $\delta \geqslant 0$ с вероятностью не меньше $1 - \delta$ справедливо:

$$L(h_0) - L_n(h_0) \leqslant \sqrt{\frac{\log\frac{1}{\delta}}{2n}}. \tag{3.11}$$

Доказательство. Заметим, что в этом случае $\sup_{h\in\mathcal{H}} L(h) - L_n(h) = L(h_0) - L_n(h_0)$. Эмпирический риск $L_n(h_0)$ является суммой независимых и одинаково распределенных случайных величин $\{\ell_{h_0}(X_i, Y_i)\}_{i=1}^n$. Кроме того, $\mathbb{E}[L_n(h_0)] = L(h_0)$. Утверждение вытекает из Следствия 1. ■

В этом случае, как мы видим, достаточно использовать неравенство Хефдинга для суммы независимых случайных величин. Мы также могли бы использовать неравенство Бернштейна вместо него.

Замечание 14. *Еще раз отметим, что неравенства (3.9) и (3.10) позволяют нам получить оценки обобщающей способности и избыточного риска на основе неравенств для супремума эмпирических процессов $\sup_{h\in\mathcal{H}} L(h) - L_n(h)$. Кроме того, последний и все последующие*

результаты настоящего параграфа (если это отдельно не оговорено) также справедливы при рассмотрении величины $\sup_{h\in\mathcal{H}} L_n(h) - L(h)$ *вместо* $\sup_{h\in\mathcal{H}} L(h) - L_n(h)$. *Это следует из того факта, что для ограниченных в отрезке* $[0,1]$ *случайных величин неравенства Хефдинга и Бернштейна дают одинаковые оценки обоих хвостов распределения. Наконец, применив неравенство Буля, мы можем получить оценки и для* $\sup_{h\in\mathcal{H}} |L(h) - L_n(h)|$.

Конечный класс $\mathcal{H}$. Пусть теперь класс $\mathcal{H}$ состоит из N отображений: $\mathcal{H} = \{h_1, \ldots, h_N\}$. В этом случае минимизатор эмпирического риска $\hat{h}_n$ принимает разные значения в зависимости от обучающей выборки S. Как было отмечено во введении к настоящему разделу, в этом случае случайные величины $\{\ell_{\hat{h}_n}(X_i, Y_i)\}_{i=1}^n$ перестают быть независимыми, и мы более не можем применять неравенство Хефдинга (или Бернштейна). Перейдя к рассмотрению равномерных по классу уклонений, мы получаем следующий результат:

Теорема 24 (Конечный класс отображений)**.** *Пусть класс* $\mathcal{H} = \{h_1, \ldots, h_N\}$ *состоит из* N *отображений и функция потерь ограничена в отрезке* $[0,1]$. *Тогда для любого* $\varepsilon > 0$ *справедливо:*

$$\mathbb{P}\left\{\sup_{h\in\mathcal{H}} L(h) - L_n(h) \geqslant \varepsilon\right\} \leqslant Ne^{-2n\varepsilon^2}.$$

Кроме того, для любого $\delta \geqslant 0$ *с вероятностью не меньше* $1-\delta$ *справедливо:*

$$\forall h \in \mathcal{H}, \quad L(h) - L_n(h) \leqslant \sqrt{\frac{\log N + \log\frac{1}{\delta}}{2n}}. \tag{3.12}$$

Наконец, для любого $\delta \geqslant 0$ *с вероятностью не меньше* $1-\delta$ *справедливо:*

$$\mathcal{E}(\hat{h}_n) \leqslant \sqrt{\frac{\log N + \log\frac{2}{\delta}}{2n}} + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

Доказательство. Воспользуемся следующей цепочкой неравенств:

$$\begin{aligned}
\mathbb{P}\left\{\sup_{h\in\mathcal{H}} L(h) - L_n(h) \geqslant \varepsilon\right\} &= \mathbb{P}\left\{\exists g \in \mathcal{H}\colon L(g) - L_n(g) \geqslant \varepsilon\right\} \\
&= \mathbb{P}\left\{\bigcup_{j=1,\ldots,N} L(h_j) - L_n(h_j) \geqslant \varepsilon\right\} \\
&\leqslant \sum_{j=1}^N \mathbb{P}\left\{L(h_j) - L_n(h_j) \geqslant \varepsilon\right\} \\
&\leqslant Ne^{-2nt^2},
\end{aligned}$$

где мы последовательно воспользовались неравенствами Буля и Хефдинга. Неравенство (3.12) мы получаем, решая уравнение $\delta = Ne^{-2nt^2}$ и обращая вероятность (пользуясь Леммой 1). Наконец, оценку избыточного риска мы получаем, ограничив в неравенстве (3.7) последнее слагаемое с помощью неравенства Хефдинга и применив неравенство Буля. ■

Обратим внимание, что для конечного класса отображений мы получили результат, отличающийся от неравенства Хефдинга дополнительным слагаемым $\log(|\mathcal{H}|)$. Это «штраф», который мы платим за требование равномерного по классу $\mathcal{H}$ контроля отклонений $L(h)$ от $L_n(h)$.

Счетный класс $\mathcal{H}$. Рассмотрим более общий случай, когда класс $\mathcal{H}$ состоит из счетного числа отображений $\mathcal{H} = \{h_1, h_2, \dots\}$. Различные версии следующего результата были получены в работах [16, 55, 75].

Теорема 25 (Оценка «Бритва Оккама»). *Пусть класс $\mathcal{H} = \{h_1, h_2, \dots\}$ состоит из счетного числа отображений и функция потерь ограничена в отрезке $[0, 1]$. Рассмотрим произвольную ограниченную весовую функцию $p\colon \mathcal{H} \to [0, 1]$, не зависящую от обучающей выборки S. Пусть $\sum_{i=1}^{\infty} p(h_i) = 1$. Тогда для любого $\delta \geqslant 0$ с вероятностью не меньше $1 - \delta$ справедливо:*

$$\forall h \in \mathcal{H}, \quad L(h) - L_n(h) \leqslant \sqrt{\frac{\log \frac{1}{p(h)} + \log \frac{1}{\delta}}{2n}}. \tag{3.13}$$

Кроме того, для любого $\delta \geqslant 0$ с вероятностью не меньше $1 - \delta$ справедливо:

$$\mathcal{E}(\hat{h}_n) \leqslant \sqrt{\frac{\log \frac{1}{p(\hat{h}_n)} + \log \frac{2}{\delta}}{2n}} + \sqrt{\frac{\log \frac{2}{\delta}}{2n}}.$$

Доказательство. Доказательство фактически повторяет шаги доказательства Теоремы 24. Воспользуемся следующей цепочкой неравенств:

$$\begin{aligned}
\mathbb{P}\left\{\exists h \in \mathcal{H}\colon L(h) - L_n(h) \geqslant \sqrt{\frac{\log \frac{1}{p(h)\cdot\delta}}{2n}}\right\} &= \mathbb{P}\left\{\bigcup_{j=1,2,\dots} L(h_j) - L_n(h_j) \geqslant \sqrt{\frac{\log \frac{1}{p(h)\cdot\delta}}{2n}}\right\} \\
&\leqslant \sum_{j=1}^{\infty} \mathbb{P}\left\{L(h_j) - L_n(h_j) \geqslant \sqrt{\frac{\log \frac{1}{p(h)\cdot\delta}}{2n}}\right\} \\
&\leqslant \sum_{j=1}^{\infty} \delta \cdot p(h_j) = \delta,
\end{aligned}$$

где мы вновь последовательно воспользовались неравенствами Буля и Хефдинга. Оценка избыточного риска вытекает из неравенств (3.7), (3.13), Хефдинга и Буля. ■

Обсудим смысл последнего утверждения. Весовую функцию p, фигурирующую в оценке Бритвы Оккама, часто называют *априорным распределением* (поскольку она не зависит от обучающей выборки, то есть должна быть выбрана *до наблюдения данных*). В отличие от прошлых оценок, последний результат дает оценки отклонения для $L(h) - L_n(h)$, зависящие от самого́ отображения h. При этом чем ближе к единице значение $p(h)$ в оценке (3.13)

для некоторого отображения $h \in \mathcal{H}$, тем ближе к нулю слагаемое $\log \frac{1}{p(h)}$, и тем точнее эта оценка. Заметим, что правая часть оценки Бритвы Оккама (3.13) ограничена снизу величиной $\sqrt{\frac{\log(1/\delta)}{2n}}$, совпадающей с правой частью оценки (3.11). Кроме того, если мы выберем равномерное априорное распределение $p(h_i) = \frac{1}{N}, i = 1, \ldots, N$ на конечном классе функций $\mathcal{H} = \{h_1, \ldots, h_N\}$, то оценка Бритвы Оккама воспроизводит оценку (3.12), справедливую для конечного класса функций.

Преимущество оценки Бритвы Оккама заключается в том, что, выбирая различные априорные распределения p, мы можем контролировать, какие отображения $h \in \mathcal{H}$ нам интересны, а какие — нет. При удачном выборе априорного распределения p она может вести к достаточно точным оценкам отклонения $L(h) - L_n(h)$ для интересующих нас отображений $h \in \mathcal{H}$. Однако, с этим связана определенная сложность. Представим, что мы хотим получить из оценки Бритвы Оккама как можно более точную оценку обобщающей способности $L(\hat{h}_n) - L_n(\hat{h}_n)$. Для этого нам необходимо, чтобы величина $p(\hat{h}_n)$ оказалась как можно ближе к единице. Поскольку отображение $\hat{h}_n$ зависит от случайной обучающей выборки S, мы не можем знать наперед, куда нужно сосредоточить априорное распределение.

Подводя итог, отметим, что если мы располагаем некоторой априорной информацией о задаче, оценка Бритвы Оккама может вести к достаточно точной оценке обобщающей способности. В противном случае (при неудачном выборе априорного распределения p) она может привести к плохим результатам.

Бесконечный несчетный класс $\mathcal{H}$ и оценка Вапника–Червоненкиса. Наконец, рассмотрим общий случай бесконечного несчетного класса $\mathcal{H}$. В настоящем параграфе мы ограничимся рассмотрением задачи классификации с бинарной функцией потерь.

В дальнейших результатах нам пригодится ряд определений. С помощью $\mathcal{H}_S$ обозначим множество различных векторов ошибок отображений класса $\mathcal{H}$ на обучающей выборке S:

$$\mathcal{H}_S = \left\{ \big(\ell_h(X_1, Y_1), \ldots, \ell_h(X_n, Y_n)\big), h \in \mathcal{H} \right\}.$$

Величина $S_{\mathcal{H}}(n) = \sup_S |\mathcal{H}_S|$ известна как *функция роста*.

Главной проблемой рассматриваемого случая является следующий факт. Несмотря на бесконечное число отображений в классе $\mathcal{H}$, число различных значений эмпирического риска $L_n(h)$ для $h \in \mathcal{H}$ конечно, поскольку $|\mathcal{H}_S| \leqslant 2^n$. Однако, множество различных значений среднего риска $\{L(h), h \in \mathcal{H}\}$ может оказаться несчетным. В этом случае мы не можем повторять шаги предыдущих доказательств, поскольку неравенство Буля справедливо для не более чем счетного множества событий.

Следующий результат [106, Лемма 2], известный как *неравенство симметризации*, позволяет справиться с этой проблемой и лежит в корне подхода, развитого В. Н. Вапником и А. Я. Червоненкисом. Мы приведем его здесь без доказательства.

Лемма 11 (Неравенство симметризации)**.** *Рассмотрим выборку $S' = \{(X'_i, Y'_i))\}_{i=1}^n$ — независимую копию обучающей выборки S (такую, что $S \cup S'$ — выборка из $2n$ независимых и одинаково распределенных пар «объект-ответ»). Обозначим с помощью $L'_n(h)$ эмпирический риск отображения h, посчитанный по выборке S' и рассмотрим бинарную функцию потерь. Тогда для всех $\varepsilon > 0$, таких что $n\varepsilon^2 \geqslant 2$, справедливо:*

$$\mathbb{P}\left\{\sup_{h\in\mathcal{H}} L(h) - L_n(h) \geqslant \varepsilon\right\} \leqslant 2\mathbb{P}\left\{\sup_{h\in\mathcal{H}} L'_n(h) - L_n(h) \geqslant \varepsilon/2\right\}. \tag{3.14}$$

Утверждение также справедливо для $\sup_{h\in\mathcal{H}} L_n(h) - L(h)$.

Неравенство симметризации позволяет избавиться от «мешавшей» нам величины $L(h)$ и перейти к рассмотрению конечных выборок. В частности, поскольку $|\mathcal{H}_{S'}|$ (а значит и число различных значений $L'_n(h), h \in \mathcal{H}$) также не превосходит 2^n, мы получили возможность рассматривать конечное множество событий.

Справедлив следующий результат:

Теорема 26 (оценка Вапника–Червоненкиса, [106])**.** *Рассмотрим бесконечный несчетный класс $\mathcal{H}$ и бинарную функцию потерь. Тогда для любого $\varepsilon > 0$, такого что $n\varepsilon^2 \geqslant 2$, справедливо:*

$$\mathbb{P}\left\{\sup_{h\in\mathcal{H}} L(h) - L_n(h) \geqslant \varepsilon\right\} \leqslant 2S_{\mathcal{H}}(2n)e^{-n\varepsilon^2/8}.$$

Кроме того, для любого $\delta \in (0, 1]$ с вероятностью не меньше $1 - \delta$ справедливо:

$$\forall h \in \mathcal{H}, \quad L(h) - L_n(h) \leqslant 2\sqrt{2\frac{\log S_{\mathcal{H}}(2n) + \log\frac{2}{\delta}}{n}}.$$

Наконец, для любого $\delta \geqslant 0$ с вероятностью не меньше $1 - \delta$ справедливо:

$$\mathcal{E}(\hat{h}_n) \leqslant 2\sqrt{2\frac{\log S_{\mathcal{H}}(2n) + \log\frac{4}{\delta}}{n}} + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

Доказательство.

$$\begin{aligned}
\mathbb{P}\left\{\sup_{h\in\mathcal{H}} L(h) - L_n(h) \geqslant \varepsilon\right\} &\leqslant 2\mathbb{P}\left\{\sup_{h\in\mathcal{H}} L'_n(h) - L_n(h) \geqslant \varepsilon/2\right\} \\
&= 2\mathbb{P}\left\{\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \big(\ell_h(X'_i, Y'_i) - \ell_h(X_i, Y_i)\big) \geqslant \varepsilon/2\right\} \\
&\leqslant 2S_{\mathcal{H}}(2n)\exp\Big(-\frac{n\varepsilon^2}{8}\Big),
\end{aligned}$$

где мы последовательно воспользовались неравенством симметризации, неравенством Буля и неравенством Хефдинга, заметив, что случайные величины $\{\ell_h(X_i', Y_i') - \ell_h(X_i, Y_i)\}_{i=1}^n$ независимы, одинаково распределены и принадлежат отрезку $[-1, 1]$. Последнее утверждение теоремы мы получаем, решая уравнение $\delta = 2S_{\mathcal{H}}(2n)e^{-n\varepsilon^2/8}$ и обращая вероятность (пользуясь Леммой 1).

Оценка избыточного риска доказывается аналогично прошлым теоремам. ■

Сравнивая последний результат с Теоремой 24, мы видим, что в оценке Вапника–Червоненкиса слагаемое $\log N$ заменено на $\log S_{\mathcal{H}}(2n)$. Таким образом можно считать, что функция роста выражает *эффективный размер* класса отображений $\mathcal{H}$. Интересно изучить, в каких случаях оценка Вапника-Червоненккиса сходится к нулю с ростом n. Для этого достаточно, чтобы $\log S_{\mathcal{H}}(2n) = o(n)$. Тривиальная оценка $S_{\mathcal{H}}(2n) \leqslant 2^{2n}$ дает нам лишь верхнюю оценку $\log S_{\mathcal{H}}(2n) \leqslant 2n$. В изучении этого вопроса нам поможет следующая комбинаторная характеристика класса $\mathcal{H}$, известная как *размерность Вапника–Червоненкиса*:

Определение 4. *Размерностью Вапника–Червоненкиса* $\mathrm{VC}(\mathcal{H})$ *класса отображений* $\mathcal{H}$ *называется наибольшее целое число* n*, такое что выполнено*

$$S_{\mathcal{H}}(n) = 2^n.$$

Если тождество выполняется для всех размеров n*, мы положим* $\mathrm{VC}(\mathcal{H}) = \infty$.

Следующий результат, независимо полученный в работах [83, 91, 106], устанавливает полиномиальный рост величины $S_{\mathcal{H}}(n)$ для классов $\mathcal{H}$ с конечной размерностью Вапника–Червоненкиса:

Лемма 12. *Пусть* $\mathcal{H}$ *— класс с конечной VC-размерностью* h*:*

$$\mathrm{VC}(\mathcal{H}) = h < \infty.$$

Тогда для всех натуральных n *справедливо:*

$$S_{\mathcal{H}}(n) \leqslant \sum_{i=1}^{h} C_n^i,$$

и для всех $n \geqslant h$*:*

$$S_{\mathcal{H}}(n) \leqslant \left(\frac{e \cdot n}{h}\right)^h.$$

Мы немедленно получаем следующее следствие оценки Вапника–Червоненкиса:

Следствие 8. *Рассмотрим бесконечный несчетный класс $\mathcal{H}$ и бинарную функцию потерь. Пусть класс $\mathcal{H}$ имеет конечную размерность Вапника–Червоненкиса $\mathrm{VC}(\mathcal{H}) = h < \infty$. Тогда для любого $\delta \in (0, 1]$ с вероятностью не меньше $1 - \delta$ справедливо:*

$$\forall h \in \mathcal{H}, \quad L(h) - L_n(h) \leqslant 2\sqrt{2\frac{h \log \frac{2ne}{h} + \log \frac{2}{\delta}}{n}}.$$

Аналогичная оценка справедлива для $L_n(h) - L(h)$. Кроме того, для любого $\delta \geqslant 0$ с вероятностью не меньше $1 - \delta$ справедливо:

$$\mathcal{E}(\hat{h}_n) \leqslant 2\sqrt{2\frac{h \log \frac{2ne}{h} + \log \frac{4}{\delta}}{n}} + \sqrt{\frac{\log \frac{2}{\delta}}{2n}}.$$

Замечание 15. *Отметим, что подход, известный как* чейнинг *(chaining), позволяет избавиться от величины $\log n$, стоящей под корнем, и получить более оптимальную скорость сходимости [50, Раздел 3].*

Замечание 16. *Таким образом, достаточным условием сходимости равномерного по классу бинарных функций уклонения $\sup_{h \in \mathcal{H}} |L(h) - L_n(h)|$ по вероятности к нулю является конечность размерности Вапника–Червоненкиса класса $\mathcal{H}$. Оказывается, в этом случае также имеет место сходимость почти наверное, и для таких классов выполняется равномерный закон больших чисел. В работе [106] также приводится необходимое и достаточное условие равномерной сходимости в терминах энтропии множества бинарных векторов ошибок отображений из класса $\mathcal{H}$ на обучающей выборке. Позже этот критерий был обобщен на произвольные ограниченные функции потерь ℓ в работе [108].*

Отметим также, что упомянутые выше достаточные условия и критерии равномерной сходимости не всегда могут быть успешно применены на практике, поскольку энтропии классов зависят от неизвестного нам распределения P, а размерность Вапника–Червоненкиса не зависит от распределения P и, следовательно, игнорирует важные свойства задачи. Достаточные условия, опирающиеся на другие свойства рассматриваемых задач, могут быть найдены в работе [118].

Оценки на основе покрытий. Завершим настоящий раздел коротким описанием еще одного подхода, являющегося в некотором смысле обобщением результата прошлого параграфа на случай произвольных функций потерь.

Для произвольной реализации обучающей выборки S определим эмпирическую метрику на *классе потерь* $\mathcal{F} = \{\ell_h(X, Y) \colon h \in \mathcal{H}\}$ следующим образом:

$$\|f\|_{L_1(S)} = \frac{1}{n} \sum_{i=1}^{n} |f(X_i, Y_i)|, \quad f \in \mathcal{F}.$$

Для произвольного $\varepsilon > 0$ обозначим с помощью $N\big(\varepsilon, \mathcal{F}, L_1(S)\big)$ минимальную мощность подмножества $\mathcal{F}' \subseteq \mathcal{F}$, удовлетворяющего следующему условию: для любого элемента $f \in \mathcal{F}$ найдется $f' \in \mathcal{F}'$, такой что $\|f - f'\|_{L_1(S)} < \varepsilon$. Величина $N\big(\varepsilon, \mathcal{F}, L_1(S)\big)$ называется *мощностью ε-покрытия* множества $\mathcal{F}$ в метрике $L_1(S)$, а логарифм этой феличины $\log N\big(\varepsilon, \mathcal{F}, L_1(S)\big)$ известен в литературе как *метрическая энтропия*. В том случае, когда мощность покрытия множества $\mathcal{F}$ конечна, мы можем сначала связать поведение функций из $\mathcal{F}$ с конечным числом функций, задающих покрытие $\mathcal{F}' \subseteq \mathcal{F}$, а затем использовать неравенство Буля для конечного подмножества $\mathcal{F}'$. Это снова позволяет нам пользоваться техникой, основанной на неравенстве Буля и рассмотренной в предыдущих примерах, для бесконечных несчетных классов $\mathcal{H}$. Мы приведем без доказательства следующий результат, который может быть найден в [77, Теорема 2.3]:

Теорема 27 (Оценка на основе покрытий)**.** *Рассмотрим бесконечный несчетный класс $\mathcal{H}$ и произвольную ограниченную в интервале $[0, 1]$ функцию потерь. Тогда для любого $\varepsilon > 0$, такого что $n\varepsilon^2 \geqslant 8$, справедливо:*

$$\mathbb{P}\left\{\sup_{h \in \mathcal{H}} L(h) - L_n(h) \geqslant \varepsilon\right\} \leqslant 8\mathbb{E}_S\big[N\big(\varepsilon, \mathcal{F}, L_1(S)\big)\big]e^{-n\varepsilon^2/128}.$$

Кроме того, для любого $\delta \in (0, 1]$ с вероятностью не меньше $1 - \delta$ справедливо:

$$\forall h \in \mathcal{H}, \quad L(h) - L_n(h) \leqslant 8\sqrt{2\frac{\log \mathbb{E}_S\big[N\big(\varepsilon, \mathcal{F}, L_1(S)\big)\big] + \log\frac{8}{\delta}}{n}}.$$

Наконец, для любого $\delta \geqslant 0$ с вероятностью не меньше $1 - \delta$ справедливо:

$$\mathcal{E}(\hat{h}_n) \leqslant 8\sqrt{2\frac{\log \mathbb{E}_S\big[N\big(\varepsilon, \mathcal{F}, L_1(S)\big)\big] + \log\frac{16}{\delta}}{n}} + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

Задача оценки мощности покрытия является отдельной сложной задачей. Множество полезных результатов может быть найдено в работе [100] (см. также ссылки, приведенные в этой работе), включая связь мощности покрытия с размерностью Вапника-Червоненкиса. Однако, эти вопросы выходят за рамки настоящей работы.

В качестве мотивации для рассмотрения следующих подходов коротко перечислим некоторые из общих недостатков результатов настоящего параграфа:

1. Ни одна из оценок не зависит от распределения P, то есть эти оценки справедливы одновременно для всех возможных распределений P (задач). В частности, размерность Вапника–Червоненкиса зависит лишь от свойств класса $\mathcal{H}$, но не от распределения P. Отсюда следует, что для большинства задач эти оценки оказываются завышенными.

Одним из возможных способов учета распределения P в оценках является использование величины $\mathbb{E}[\log|\mathcal{H}_S|]$ вместо размерности $\mathrm{VC}(\mathcal{H})$ (см. например [23, Теорема 3]). Однако, оценка величины $\mathbb{E}[\log|\mathcal{H}_S|]$ часто оказывается отдельной сложной задачей.

2. Ни одна из приведенных выше оценок не зависит от обучающей выборки S, что также является упущением по названным выше причинам.

3. Все приведенные выше оценки используют неравенство Хефдинга, которое учитывает лишь факт ограниченности случайных величин и не зависит от их дисперсий. В первой главе было продемонстрировано, что в ряде случаев учет дисперсий случайных величин ведет к более сильным результатам.

4. Все приведенные выше оценки используют неравенство Буля в качестве важного шага своего доказательства. Неравенство Буля обращается в равенство только тогда, когда рассматриваемые события независимы в совокупности. Это означает, что чем сильнее зависимость между векторами ошибок $\mathcal{H}_S$ отображений класса $\mathcal{H}$ на обучающей выборке S, тем сильнее завышены все приведенные оценки. Оказывается, в большинстве используемых на практике семейств $\mathcal{H}$ подобная зависимость имеет место. В *комбинаторной теории переобучения* [111, 115, 116], которая будет рассмотрена в Главе 5, это свойство семейства отображений названо *свойством сходства* отображений семейства $\mathcal{H}$ и играет ключевую роль в получении численно точных оценок обобщающей способности.

5. Оценка Вапника–Червоненкиса существенно опирается на использовании бинарной функции потерь. Описанный подход перестает работать в общем случае ограниченной на отрезке $[0,1]$ функции потерь. В работах [8, 45] вводится понятие *псевдоразмерности* (fat-shattering dimension) класса $\mathcal{H}$, которое обобщает размерность Вапника–Червоненкиса на случай функций потерь, принимающих действительные значения из интервала $[0,1]$, и изучаются его свойства. Оценки, использующие псевдоразмерность класса, выводятся с помощью подхода, похожего на описанный в настоящем параграфе.

6. Часто задача вычисления (или оценки сверху) упомянутых сложностных (структурных) характеристик класса $\mathcal{H}$ — таких как его размерность Вапника–Червоненкиса, псевдоразмерность или метрическая энтропия — оказывается сама по себе очень сложной.

7. Наконец, как отмечалось выше, переход к рассмотрению равномерного по классу уклонения $\sup_{h\in\mathcal{H}} L(h) - L_n(h)$ часто ведет к завышенным и неоптимальным оценкам. С этой проблемой позволяют справляться приведенные ниже подходы, основанные на локальном анализе.

3.2.2 Оценки, основанные на Радемахеровской сложности

В прошлых главах мы продемонстрировали, что супремум эмпирического процесса $\sup_{h\in\mathcal{H}} L(h) - L_n(h)$ концентрируется вокруг своего математического ожидания $\mathbb{E}\left[\sup_{h\in\mathcal{H}} L(h) - L_n(h)\right]$. Этот факт может быть установлен с помощью неравенств МакДиармида и Талаграна, приведенных в Теоремах 12 и 13. Следовательно, мы можем перейти от изучения супремума эмпирического процесса к изучению его математического ожидания.

Пусть $\sigma_1, \ldots, \sigma_n$ — независимые и одинаково распределенные случайные величины, такие что $\mathbb{P}\{\sigma_i = 1\} = \mathbb{P}\{\sigma_i = -1\} = 1/2$. Такие случайные величины называются *Радемахеровскими*. Пусть, кроме того, $\sigma_1, \ldots, \sigma_n$ не зависят от обучающей выборки S. Для произвольного множества $\mathcal{G}$ отображений из $\mathcal{X} \times \mathcal{Y}$ в $\mathbb{R}$ рассмотрим следующую случайную величину:

$$R_n(\mathcal{G}) = \sup_{g\in\mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(X_i, Y_i).$$

Величины $\mathbb{E}[R_n(\mathcal{G})]$ и $\mathbb{E}[R_n(\mathcal{G})|S]$ принято называть *глобальной Радемахеровской сложностью*[7] и *глобальной эмпирической Радемахеровской сложностью* класса отображений $\mathcal{G}$ соответственно. В дальнейшем для простоты мы будем обозначать условное математическое ожидание $\mathbb{E}[R_n(\mathcal{G})|S]$ с помощью $\mathbb{E}_\sigma[R_n(\mathcal{G})]$.

Классом потерь $\mathcal{F}_\mathcal{H}$ мы будем называть следующее множество отображений:

$$\mathcal{F}_\mathcal{H} = \left\{(X, Y) \to \ell_h(X, Y) \colon h \in \mathcal{H}, (X, Y) \in \mathcal{X} \times \mathcal{Y}\right\}.$$

Всюду дальше, где класс отображений $\mathcal{H}$, индуцирующий класс потерь $\mathcal{F}_\mathcal{H}$, понятен из контекста, мы будем писать просто $\mathcal{F}$. Величину $R_n(\mathcal{F})$ можно интерпретировать как максимальную корреляцию векторов ошибок отображений класса $\mathcal{H}$ с вектором случайного шума. Чем больше величина $\mathbb{E}[R_n(\mathcal{F})]$, тем лучше класс $\mathcal{H}$ «подстраивается» под случайный шум[8] и тем больше его сложность. Таким образом, Радемахеровская сложность является еще одним

[7] Термин «глобальная» в данном случае противопоставляет описываемый подход более сложному подходу, основанному на *локальной* Радемахеровской сложности, который будет представлен далее.

[8] Работа [78] показывает, что измерение сложности класса на основе случайного *Радемахеровского* шума не всегда отражает свойства конкретной рассматриваемой задачи, которой соответствует специфический шум. Вместо этого автор предлагает использовать вектора с координатами $\sigma_i\big(Y_i - h^*(X_i)\big)$, $i = 1, \ldots, n$.

примером *сложностной* характеристики класса отображений. В отличие от размерности Вапника–Червоненкиса, Радемахеровская сложность зависит от неизвестного распределения P.

Оказывается, Радемахеровская сложность класса потерь $\mathcal{F}$ непосредственным образом связана с интересующей нас величиной $\mathbb{E}\left[\sup_{h\in\mathcal{H}} L(h) - L_n(h)\right]$. Следующий результат, доказательства которого могут быть найдены в работах [5,50], является еще одной модификацией неравенства симметризации. Для краткости мы будем обозначать

$$\hat{\mathbb{E}}_n[g] = \frac{1}{n}\sum_{i=1}^{n} g(X_i, Y_i)$$

для функций $g\colon \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$.

Лемма 13 (Неравенство симметризации)**.** *Для любого класса $\mathcal{G}$ отображений из $\mathcal{X}\times\mathcal{Y}$ в $\mathbb{R}$ справедливо:*

$$\frac{1}{2}\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\big(g(X_i,Y_i)-\mathbb{E}[g(X,Y)]\big)\right| \leqslant \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\mathbb{E}[g(X,Y)]-\hat{\mathbb{E}}_n[g]\right|\right]$$

и

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\mathbb{E}[g(X,Y)]-\hat{\mathbb{E}}_n[g]\right|\right] \leqslant 2\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(X_i,Y_i)\right|\right].$$

В частности, справедливо следующее неравенство:

$$\max\left(\mathbb{E}\left[\sup_{g\in\mathcal{G}}\mathbb{E}[g(X,Y)]-\hat{\mathbb{E}}_n[g]\right], \mathbb{E}\left[\sup_{g\in\mathcal{G}}\hat{\mathbb{E}}_n[g]-\mathbb{E}[g(X,Y)]\right]\right) \leqslant 2\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(X_i,Y_i)\right].$$

Применив эти неравенства для класса $\mathcal{G} = \{\ell_h(X,Y)\colon h\in\mathcal{H}\}$, мы немедленно получаем следующее соотношение:

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}} L(h) - L_n(h)\right] \leqslant 2\mathbb{E}[R_n(\mathcal{F})].$$

Радемахеровский случайный процесс

$$\left\{\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(X_i,Y_i),\ g\in\mathcal{G}\right\}$$

и в особенности его супремум $R_n(\mathcal{G})$ хорошо изучены в теории эмпирических процессов, и для них существует ряд полезных технических результатов (короткий перечень может быть найден в [18, Теорема 3.3]). Помимо неравенства симметризации далее нам понадобится следующий результат, доказательства которого могут быть найдены в [19, 50, 59]:

Теорема 28 (Неравенство сжатия, [19])**.** *Рассмотрим 1-липшицевы отображения $\phi_i\colon\mathbb{R}\to\mathbb{R}$, $i=1,\ldots,n$:*

$$|\phi_i(u)-\phi_i(v)| \leqslant |u-v|, \quad u,v\in\mathbb{R},$$

для которых выполнено $\phi_i(0) = 0$. *Рассмотрим множество* n*-мерных векторов* $\mathcal{G} \subset \mathbb{R}^n$. *Тогда справедливо следующее:*

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \cdot \phi_i(g_i)\right] \leqslant \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \cdot g_i\right],$$

где g_i *— i-ая координата вектора* $g = (g_1, \ldots, g_n) \in \mathcal{G}$.

Неравенство сжатия является чрезвычайно полезным инструментом. Оно позволяет переходить от рассмотрения Радемахеровских сложностей более сложных производных классов к изучению свойств базовых классов. В частности, если функция потерь Липшицева, то этот результат может позволить нам перейти от рассмотрения Радемахеровской сложности класса потерь $\mathcal{F}$ к изучению геометрических и структурных свойств самого класса отображений $\mathcal{H}$.

Воспользовавшись неравенством МакДиармида вместе с неравенством симметризации, приведенным выше, мы немедленно получаем следующий результат:

Теорема 29. *Рассмотрим произвольный класс* $\mathcal{H}$ *и функцию потерь, ограниченную в отрезке* $[0, 1]$. *Тогда для любого* $\delta \geqslant 0$ *с вероятностью не меньше* $1 - \delta$ *справедливо:*

$$\forall h \in \mathcal{H}, \quad L(h) - L_n(h) \leqslant 2\mathbb{E}[R_n(\mathcal{F}_\mathcal{H})] + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}. \tag{3.15}$$

Кроме того, для любого $\delta \geqslant 0$ *с вероятностью не меньше* $1 - \delta$ *справедливо:*

$$\forall h \in \mathcal{H}, \quad L(h) - L_n(h) \leqslant 2\mathbb{E}_\sigma[R_n(\mathcal{F}_\mathcal{H})] + 3\sqrt{\frac{2\log\frac{2}{\delta}}{n}}. \tag{3.16}$$

Аналогичные неравенства справедливы для $L_n(h) - L(h)$. *Кроме того, для любого* $\delta \geqslant 0$ *с вероятностью не меньше* $1 - \delta$ *справедливо:*

$$\mathcal{E}(\hat{h}_n) \leqslant 2\mathbb{E}_\sigma[R_n(\mathcal{F}_\mathcal{H})] + 4\sqrt{\frac{2\log\frac{4}{\delta}}{n}}.$$

Доказательство. Первое неравенство немедленно следует из применения неравенства МакДиармида Теоремы 9 к случайной величине $Z = \sup_{h\in\mathcal{H}} L(h) - L_n(h)$, которая удовлетворяет условиям теоремы с параметрами $c_i = 2/n$, и последующего применения Леммы 13.

Для доказательства второго неравенства заметим, что случайная величина

$$Z = \sup_{h\in\mathcal{H}} L(h) - L_n(h) - 2\mathbb{E}_\sigma[R_n(\mathcal{F})]$$

удовлетворяет условиям ограниченных разностей с параметрами $c_i = 6/n$. Применяя для нее неравенство МакДиармида мы получим, что для $\delta \geqslant 0$ с вероятностью не меньше $1 - \delta$ справедливо:

$$Z - \mathbb{E}[Z] \leqslant 3\sqrt{\frac{2\ln\frac{1}{\delta}}{n}}.$$

Также из неравенства симметризации вытекает:

$$E[Z] = \mathbb{E}\left[\sup_{h\in\mathcal{H}} L(h) - L_n(h)\right] - 2\mathbb{E}[R_n(\mathcal{F})] \leqslant 0.$$

Мы завершаем доказательство второго неравенства применением неравенства Буля. Оценка избыточного риска вновь следует из неравенств (3.7) и Бернштейна. ■

Мы получили первую вычислимую оценку обобщающей способности и избыточного риска, зависящую от обучающей выборки. Действительно, правая часть неравенства (3.16) не зависит от неизвестного нам распределения P и может быть вычислена на основе данных. Также отметим, что мы могли бы пойти на шаг дальше и воспользоваться неравенством Мак-Диармида для случайной величины $R_n(\mathcal{F})$, которая также концентрируется вокруг своего математического ожидания $\mathbb{E}_\sigma[R_n(\mathcal{F})]$. В результате мы получили бы оценку, не содержащую взятия математического ожидания по Радемахеровским случайным величинам, справедливую с большой вероятностью для любой реализации $\sigma_1, \ldots, \sigma_n$.

Задача оценивания величины $\sup_{h\in\mathcal{H}} L(h) - L_n(h)$ свелась, таким образом, к оцениванию величин $E[R_n(\mathcal{F})]$ и $E_\sigma[R_n(\mathcal{F})]$. Далее мы приведем оценки этих величин лишь для простейшего случая, когда класс $\mathcal{H}$ содержит конечное число отображений. Подробный обзор других результатов может быть найден в [50, Раздел 3]. Нам понадобится следующий результат:

Лемма 14. *Пусть функция потерь ℓ ограничена в отрезке $[0,1]$. Для конечного класса $\mathcal{H} = \{h_1, \ldots, h_N\}$ справедливо:*

$$\mathbb{E}_\sigma[R_n(\mathcal{F}_\mathcal{H})] \leqslant \max_{j=1,\ldots,N} \left(\sum_{i=1}^n \big(\ell_{h_j}(X_i, Y_i)\big)^2\right)^{1/2} \frac{\sqrt{2\log N}}{n} \leqslant \sqrt{\frac{2\log N}{n}}.$$

Доказательство. Для любого $s > 0$ справедливо:

$$\mathbb{E}_\sigma[R_n(\mathcal{F})] = \frac{1}{s}\mathbb{E}\left[\log e^{s\cdot R_n(\mathcal{F})}\right] \leqslant \frac{1}{s}\log \mathbb{E}_\sigma\left[e^{s\cdot R_n(\mathcal{F})}\right],$$

где мы воспользовались неравенством Йенсена. Таким образом:

$$\begin{aligned}
\mathbb{E}_\sigma[R_n(\mathcal{F})] &\leqslant \frac{1}{s}\log \mathbb{E}_\sigma\left[\exp\left(s\cdot \max_{j=1,\ldots,N} \frac{1}{n}\sum_{i=1}^n \sigma_i \ell_{h_j}(X_i, Y_i)\right)\right] \\
&\leqslant \frac{1}{s}\log\left\{\sum_{j=1}^N \mathbb{E}_\sigma\left[\exp\left(s\cdot \frac{1}{n}\sum_{i=1}^n \sigma_i \ell_{h_j}(X_i, Y_i)\right)\right]\right\} \\
&= \frac{1}{s}\log\left\{\sum_{j=1}^N \prod_{i=1}^n \mathbb{E}_\sigma\left[\exp\left(s\cdot \frac{1}{n}\sigma_i \ell_{h_j}(X_i, Y_i)\right)\right]\right\} \\
&\leqslant \frac{1}{s}\log\left\{\sum_{j=1}^N \prod_{i=1}^n \exp\left(s^2 \frac{\big(\ell_{h_j}(X_i, Y_i)\big)^2}{2n^2}\right)\right\} \\
&\leqslant \frac{1}{s}\log\left\{N \max_{j=1,\ldots,N} \exp\left(s^2 \frac{\sum_{i=1}^n \big(\ell_{h_j}(X_i, Y_i)\big)^2}{2n^2}\right)\right\},
\end{aligned}$$

где мы воспользовались Леммой 2. Доказательство первого неравенства завершается минимизацией последнего выражения по s. Второе неравенство следует из ограниченности функции потерь сверху единицей. ■

Таким образом, Теорема 29 вместе с Леммой 14 снова дают нам оценки порядка $O(n^{-1/2})$. В частности, рассматривая задачу классификации $\mathcal{Y} = \{0, 1\}$ с бинарной функцией потерь и класс отображений $\mathcal{H}$ с конечной VC-размерностью $\mathrm{VC}(\mathcal{H}) = h$, мы немедленно получим следующую оценку:

$$\mathbb{E}[R_n(\mathcal{H})] = \mathbb{E}[\mathbb{E}_\sigma[R_n(\mathcal{F})]] \leqslant \mathbb{E}\left[\sqrt{\frac{2\log S_{\mathcal{F}}(n)}{n}}\right] \leqslant \sqrt{\frac{2h\log\frac{en}{h}}{n}},$$

которая вместе с Теоремой 29 воспроизводит оценку Вапника–Червоненкиса Следствия 8.

3.2.3 Оценки, основанные на локальных мерах сложности, и быстрые скорости сходимости

Все оценки, приведенные выше, давали скорость сходимости $\sup_{h\in\mathcal{H}} |L(h) - L_n(h)|$ к нулю порядка $O(n^{-1/2})$. Во многих случаях при отказе от дополнительных предположений о распределении P этот порядок оказывается оптимальным [30, 65]. Однако, дополнительные ограничения на рассматриваемые задачи могут вести к оценкам порядка $o(n^{-1/2})$ — так называемым *быстрым скоростям* сходимости.

Приведенные выше результаты основаны на неравенствах Хефдинга и МакДиармида, которые ведут себя очень похожим образом и оба учитывают лишь ограниченность случайных величин. В большинстве случаев получение оценок порядка $o(n^{-1/2})$ основано на использовании неравенств типа Беннетта, принципиальным отличием которых является учет дисперсии случайных величин. Учет дисперсии, как мы увидим, будет играть ключевую роль во многих приведенных в настоящем разделе результатах.

Задачи без шума. Для начала мы рассмотрим идеализированный случай задачи без шума (то есть Байесовский риск L^* равен нулю) и приведем первый пример оценок порядка $O(n^{-1})$.

Теорема 30. *Пусть класс $\mathcal{H} = \{h_1, \ldots, h_N\}$ состоит из N отображений и функция потерь ограничена в отрезке $[0, 1]$. Тогда для любого $\delta \geqslant 0$ с вероятностью не меньше $1 - \delta$ справедливо:*

$$\forall h \in \mathcal{H}: \quad L(h) - L_n(h) \leqslant \sqrt{2L_n(h)\frac{\log N + \log\frac{1}{\delta}}{n}} + 4\frac{\log N + \log\frac{1}{\delta}}{n}.$$

Также для любого $\delta \geqslant 0$ с вероятностью не меньше $1-\delta$ справедливо:

$$\mathcal{E}(\hat{h}_n) \leqslant 3\sqrt{2L(h^*)\frac{\log N + \log\frac{2}{\delta}}{n}} + 7\frac{\log N + \log\frac{2}{\delta}}{n}.$$

Пусть, кроме того, Байесовское отображение g^* принадлежит классу $\mathcal{H}$ и $Y = g^*(X)$ с вероятностью 1, то есть $L^* = L(g^*) = L(h^*) = 0$. Тогда для любого $\delta \geqslant 0$ с вероятностью не меньше $1-\delta$ справедливо:

$$\mathcal{E}(\hat{h}_n) = L(\hat{h}_n) - L_n(\hat{h}_n) \leqslant 4\frac{\log N + \log\frac{1}{\delta}}{n}. \tag{3.17}$$

Для доказательства теоремы нам потребуется следующий результат, приведенный в [21, Раздел 2.7.2]:

Лемма 15. *Пусть $A, B, C \geqslant 0$ и справедливо:*

$$A \leqslant B + C\sqrt{A}.$$

Тогда справедливо следующее неравенство:

$$A \leqslant B + C^2 + \sqrt{B}C.$$

Доказательство. Нам достаточно показать, что справедливо неравенство

$$\sqrt{A} \leqslant C + \sqrt{B}.$$

Из условия леммы следует, что $\sqrt{A}$ не превосходит бóльшего корня квадратного уравнения $x^2 - Cx - B = 0$, то есть

$$\sqrt{A} \leqslant \frac{C + \sqrt{C^2 + 4B}}{2} \leqslant C + \sqrt{B},$$

где мы воспользовались неравенством $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$. ■

Доказательство Теоремы 30. Для каждого отображения $h \in \mathcal{H}$ и любого $\delta \geqslant 0$ в силу неравенства Бернштейна Теоремы 6 следующее неравенство справедливо с вероятностью не меньше $1-\delta$:

$$L(h) - L_n(h) \leqslant \sqrt{\frac{2\mathbb{D}[\ell_h(X,Y)]\log\frac{1}{\delta}}{n}} + \frac{2\log\frac{1}{\delta}}{3n}. \tag{3.18}$$

Заметим, что также справедливо:

$$\mathbb{D}[\ell_h(X,Y)] = \mathbb{E}[\ell_h^2(X,Y)] - (\mathbb{E}[\ell_h(X,Y)])^2 \leqslant \mathbb{E}[\ell_h^2(X,Y)] \leqslant L(h),$$

в силу ограниченности функции потерь единицей. Тогда с помощью неравенства (3.18) и неравенства Буля мы получаем:

$$\begin{aligned}
&\mathbb{P}\left\{\exists h \in \mathcal{H}\colon L(h) - L_n(h) \geqslant \sqrt{\frac{2L(h)\log\frac{N}{\delta}}{n}} + \frac{2\log\frac{N}{\delta}}{3n}\right\} \\
&\leqslant \mathbb{P}\left\{\exists h \in \mathcal{H}\colon L(h) - L_n(h) \geqslant \sqrt{\frac{2\mathbb{D}[\ell_h(X,Y)]\log\frac{N}{\delta}}{n}} + \frac{2\log\frac{N}{\delta}}{3n}\right\} \\
&= \mathbb{P}\left\{\bigcup_{j=1,2,\ldots} L(h_j) - L_n(h_j) \geqslant \sqrt{\frac{2\mathbb{D}[\ell_{h_j}(X,Y)]\log\frac{N}{\delta}}{n}} + \frac{2\log\frac{N}{\delta}}{3n}\right\} \\
&\leqslant \sum_{j=1}^{N} \mathbb{P}\left\{L(h_j) - L_n(h_j) \geqslant \sqrt{\frac{2\mathbb{D}[\ell_{h_j}(X,Y)]\log\frac{N}{\delta}}{n}} + \frac{2\log\frac{N}{\delta}}{3n}\right\} \\
&\leqslant N \cdot \frac{\delta}{N} = \delta.
\end{aligned}$$

Таким образом, с вероятностью не менее $1-\delta$ выполнено:

$$L(\hat{h}_n) - L_n(\hat{h}_n) \leqslant \sqrt{\frac{2L(\hat{h}_n)\log\frac{N}{\delta}}{n}} + \frac{2\log\frac{N}{\delta}}{3n}.$$

Из Леммы 15 и неравенства $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$, а также из простых вычислений следует, что также выполнено:

$$L(\hat{h}_n) - L_n(\hat{h}_n) \leqslant \sqrt{2L_n(\hat{h}_n)\frac{\log\frac{N}{\delta}}{n}} + 4\frac{\log\frac{N}{\delta}}{n}.$$

Кроме того, из условий теоремы следует, что с вероятностью 1 выполнено $L_n(\hat{h}_n) = 0$. Это завершает доказательство первого и последнего неравенств теоремы.

Для доказательства оценки избыточного риска заметим, что справедливо:

$$L_n(\hat{h}_n) \leqslant L_n(h^*) = L_n(h^*) - L(h^*) + L(h^*). \tag{3.19}$$

С учетом первого неравенства теоремы, неравенства Бернштейна и неравенства Буля мы получаем:

$$\mathcal{E}(\hat{h}_n) \leqslant \sqrt{2L_n(\hat{h}_n)\frac{\log\frac{2N}{\delta}}{n}} + 4\frac{\log\frac{2N}{\delta}}{n} + \sqrt{\frac{2L(h^*)\log\frac{2}{\delta}}{n}} + \frac{2\log\frac{2}{\delta}}{3n}.$$

Вновь воспользовавшись неравенством (3.19), мы получаем:

$$\mathcal{E}(\hat{h}_n) \leqslant \sqrt{2\left(L(h^*) + \sqrt{\frac{2L(h^*)\log\frac{2}{\delta}}{n}} + \frac{2\log\frac{2}{\delta}}{3n}\right)\frac{\log\frac{2N}{\delta}}{n}} + \sqrt{\frac{2L(h^*)\log\frac{2}{\delta}}{n}} + 5\frac{\log\frac{2N}{\delta}}{n}$$

или

$$\mathcal{E}(\hat{h}_n) \leqslant \sqrt{2L(h^*)\frac{\log\frac{2N}{\delta}}{n} + 2\left(\sqrt{\frac{2L(h^*)\log\frac{2}{\delta}}{n}} + \frac{2\log\frac{2}{\delta}}{3n}\right)\frac{\log\frac{2N}{\delta}}{n}} + \sqrt{\frac{2L(h^*)\log\frac{2}{\delta}}{n}} + 5\frac{\log\frac{2N}{\delta}}{n}.$$

Последовательно применяя неравенства $\sqrt{a+b} \leqslant \sqrt{a}+\sqrt{b}$ и $\sqrt{ab} \leqslant \frac{a+b}{2}$ к первому слагаемому, мы получаем:

$$\mathcal{E}(\hat{h}_n) \leqslant 3\sqrt{2L(h^*)\frac{\log\frac{2N}{\delta}}{n}} + 7\frac{\log\frac{2N}{\delta}}{n}.$$

■

Оказывается, последний результат можно обобщить на более общий случай бесконечных классов $\mathcal{H}$. В частности, справедлива следующая односторонняя оценка (см. [107, Глава 12] и [18]), которую мы приводим без доказательства:

Теорема 31. *Рассмотрим бесконечный несчетный класс $\mathcal{H}$ и бинарную функцию потерь. Пусть класс $\mathcal{H}$ имеет конечную размерность Вапника–Червоненкиса $\mathrm{VC}(\mathcal{H}) = h < \infty$. Тогда для любого $\delta \in (0,1]$ с вероятностью не меньше $1-\delta$ справедливо:*

$$\forall h \in \mathcal{H}, \quad L(h) - L_n(h) \leqslant 2\sqrt{L_n(h)\frac{2h\log(n+1)+\log\frac{4}{\delta}}{n}} + 4\frac{2h\log(n+1)+\log\frac{4}{\delta}}{n}.$$

Также существует универсальная константа $C > 0$, такая что для любого $\delta \in (0,1]$ с вероятностью не меньше $1-\delta$ справедливо:

$$\mathcal{E}(\hat{h}_n) \leqslant C\left(\sqrt{L(h^*)\frac{h\log n+\log\frac{1}{\delta}}{n}} + \frac{h\log n+\log\frac{1}{\delta}}{n}\right).$$

Пусть, кроме того, Байесовское отображение g^* принадлежит классу $\mathcal{H}$ и $Y = g^*(X)$ с вероятностью 1, то есть $L^* = L(g^*) = L(h^*) = 0$. Тогда для любого $\delta \geqslant 0$ с вероятностью не меньше $1-\delta$ справедливо:

$$\mathcal{E}(\hat{h}_n) \leqslant 4\frac{2h\log(n+1)+\log\frac{4}{\delta}}{n}.$$

Еще раз подчеркнем, что появление множителя $L(h^*)$ перед слагаемым порядка $O(n^{-1/2})$ в оценке Теоремы 30 обусловлено именно неравенством Бернштейна. Ключевую роль при этом играло соотношение между средним риском отображения и дисперсией его потерь — неравенство $\mathbb{D}[\ell_h(X,Y)] \leqslant L(h)$, справедливое для класса неотрицательных и ограниченных функций. Теорема 31 основана на другой технике, развитой в [104] и рассматривающей *взвешенные эмпирические процессы* вида

$$\sup_{h\in\mathcal{H}} \frac{L(h)-L_n(h)}{\sqrt{L(h)}}.$$

Предположения, использованные в двух последних результатах — а именно, предположения о принадлежности Байесовского отображения рассматриваемому классу функций

$g^* \in \mathcal{H}$ и отсутствии шума в задаче $L^* = 0$, — являются чрезвычайно строгими и на практике никогда не выполняются. Однако, они ведут к первым наглядным примерам оценок, имеющих быструю скорость сходимости $O(n^{-1})$. В случае $L(h^*) > 0$ оценки имеют порядок $O(n^{-1/2})$. При этом прочие значения показателя степени n из интервала $[-1, -1/2]$ приведенные оценки получать не позволяют.

Класс избыточных потерь. Проблема приведенных результатов, оказывается, заключается в том, что величина $L(h^*)$ — не лучший выбор множителя перед слагаемым порядка $O(n^{-1/2})$, поскольку с ростом обучающей выборки эта величина в общих предположениях не стремится к нулю. В результате мы вынуждены ограничиваться рассмотрением двух крайних случаев $L(h^*) > 0$ и $L(h^*) = 0$, второй из которых ведет к быстрым скоростям сходимости, но, к сожалению, не является реалистичным.

Основной идеей дальнейших результатов, позволяющих получать быстрые скорости сходимости в более общих задачах с шумом (то есть при условии $L^* > 0$), является замена множителя $L(h^*)$ на величину $L(\hat{h}_n) - L(h^*)$, которая обычно убывает к нулю с ростом размера обучающей выборки n. Для этого мы перейдем от рассмотрения класса потерь $\{\ell_h \colon h \in \mathcal{H}\}$, с которым мы познакомились в прошлом параграфе, к классу *избыточных потерь* (excess loss class) $\{\ell_h - \ell_{h^*} \colon h \in \mathcal{H}\}$. Для дальнейших результатов нам пригодятся следующие короткие обозначения двух этих классов:

$$\mathcal{F}_{\mathcal{H}} = \{\ell_h \colon h \in \mathcal{H}\}, \quad \mathcal{F}^*_{\mathcal{H}} = \{\ell_h - \ell_{h^*} \colon h \in \mathcal{H}\}.$$

Всюду далее, где класс отображений $\mathcal{H}$ понятен из контекста, мы будем опускать нижний индекс в двух этих определениях и писать просто $\mathcal{F}$ и $\mathcal{F}^*$.

Обратим внимание, что для $f \in \mathcal{F}^*$ справедливо $\mathbb{E}\,[f(X,Y)] = L(h) - L(h^*)$. Кроме того, для минимизатора эмпирического риска $\hat{h}_n$ на классе $\mathcal{H}$ также справедливо:

$$L_n(\hat{h}_n) - L_n(h^*) = \min_{h \in \mathcal{H}} L_n(h) - L_n(h^*) \leqslant 0.$$

На основе двух этих наблюдений мы могли бы попытаться воспользоваться подходами, примененными в двух последний результатах, для класса избыточных потерь $\mathcal{F}^*$ и получить множитель $L(\hat{h}_n) - L(h^*)$ вместо множителей $L(h^*)$ и $L(\hat{h}_n)$. Однако, при попытке повторить те же шаги мы столкнулись бы с проблемой: функции из множества $\mathcal{F}^*$, в отличие от функций из $\mathcal{F}$, не являются положительными. Их математические ожидания $L(h) - L(h^*)$, $h \in \mathcal{H}$, являются положительными (что следует из определения h^*), но функции могут принимать отрицательные значения с ненулевой вероятностью. Это ведет к тому, что мы больше не мо-

жем воспользоваться элементарным неравенством $\mathbb{E}[\xi^2] \leqslant a \cdot \mathbb{E}[\xi]$, справедливым для произвольной константы $a > 0$ и случайных величин ξ, принимающих значения в отрезке $[0, a]$. В то же время именно это неравенство позволило нам установить в Теореме 30 связь между дисперсией $\mathbb{D}[\ell_h(X,Y)]$ и средним риском $L(h)$.

Оказывается, соотношения вида

$$\mathbb{D}\left[\ell_h(X,Y) - \ell_{h^*}(X,Y)\right] \leqslant \mathbb{E}\left[\left(\ell_h(X,Y) - \ell_{h^*}(X,Y)\right)^2\right] \leqslant c\big(L(h) - L(h^*)\big)^\alpha = c\cdot\big(\mathcal{E}(h)\big)^\alpha, \quad (3.20)$$

где $c > 0$ и $\alpha \in (0, 1]$ — некоторые константы, во многих случаях вытекают из достаточно мягких дополнительных предположений о распределении P и позволяют получать оценки избыточного риска, имеющие быструю скорость сходимости.

Задачи с шумом: конечный класс $\mathcal{H}$. Для начала на простом примере конечного класса отображений $\mathcal{H} = \{h_1, \ldots, h_N\}$ покажем, что условия вида (3.20) действительно ведут к оценкам быстрых порядков $o(n^{-1/2})$ без дополнительных ограничений на рассматриваемые распределения P.

Теорема 32. *Пусть класс $\mathcal{H} = \{h_1, \ldots, h_N\}$ состоит из N отображений и функция потерь ограничена в отрезке $[0, 1]$. Пусть, кроме того, существуют константы $\alpha \in (0, 1]$ и $c > 0$, такие что соотношение (3.20) справедливо для всех $h \in \mathcal{H}$. Тогда для любого $\delta \geqslant 0$ с вероятностью не меньше $1 - \delta$ справедливо:*

$$\mathcal{E}(\hat{h}_n) = O\left(\left(\frac{\log N + \log\frac{1}{\delta}}{n}\right)^{\frac{1}{2-\alpha}}\right).$$

Доказательство. Повторяя первые шаги доказательства Теоремы 30 мы получим, что для любого $\delta \geqslant 0$ с вероятностью не меньше $1 - \delta$:

$$\forall h \in \mathcal{H}: \quad L(h) - L(h^*) \leqslant L_n(h) - L_n(h^*) + \sqrt{\frac{2\mathbb{D}\left[\ell_h(X,Y) - \ell_{h^*}(X,Y)\right]\log\frac{N}{\delta}}{n}} + \frac{2\log\frac{N}{\delta}}{3n}.$$

Воспользовавшись условием (3.20), мы получим:

$$\forall h \in \mathcal{H}: \quad \mathcal{E}(h) \leqslant L_n(h) - L_n(h^*) + \sqrt{\frac{2c\big(\mathcal{E}(h)\big)^\alpha \log\frac{N}{\delta}}{n}} + \frac{2\log\frac{N}{\delta}}{3n}.$$

Поскольку $L_n(\hat{h}_n) - L_n(h^*) \leqslant 0$, мы, в частности, получаем:

$$\mathcal{E}(\hat{h}_n) \leqslant \sqrt{\frac{2c\big(\mathcal{E}(\hat{h}_n)\big)^\alpha \log\frac{N}{\delta}}{n}} + \frac{2\log\frac{N}{\delta}}{3n}.$$

Тривиальная оценка $\mathcal{E}(\hat{h}_n) \leqslant 1$ показывает, что $\mathcal{E}(\hat{h}_n) = O(n^{-1/2})$. Подставляя эту оценку вновь в последнее неравенство мы видим, что порядок $\mathcal{E}(\hat{h}_n)$ в действительности может оказаться меньше $O(n^{-1/2})$. На примере этих рассуждений становится понятно, что порядок

избыточного риска связан с неподвижной точкой x^* отображения $x \to \sqrt{\frac{2cx^\alpha \log\frac{N}{\delta}}{n}} + \frac{2\log\frac{N}{\delta}}{3n}$. В дальнейших параграфах мы еще ни раз столкнемся с подобными рассуждениями.

Предположим, что для любого $C > 0$ и любого $n_0 \in \mathbb{N}$ найдется номер $n_1 \geqslant n_0$, такой что выполнено $\mathcal{E}(\hat{h}_n) \geqslant C/n_1$ (в противном случае $\mathcal{E}(\hat{h}_n) = O(n^{-1})$ и мы немедленно завершаем доказательство). Поскольку $\alpha \in (0, 1]$, отсюда следует, что:

$$\frac{\log\frac{N}{\delta}}{n} = O\left(\sqrt{\frac{\left(\mathcal{E}(\hat{h}_n)\right)^\alpha \log\frac{N}{\delta}}{n}}\right).$$

Следовательно, найдется константа C_2, такая что для достаточно больших n справедливо:

$$\mathcal{E}(\hat{h}_n) \leqslant C_2\sqrt{\frac{\left(\mathcal{E}(\hat{h}_n)\right)^\alpha \log\frac{N}{\delta}}{n}},$$

что завершает доказательство теоремы. ■

Обратим внимание, что при $\alpha = 1$ мы получаем оценку избыточного риска порядка $O(n^{-1})$, как в случае задачи без шума $L^* = 0$. Также во всех случаях мы получаем скорость сходимости быстрее $O(n^{-1/2})$. Для этого нам потребовалось лишь условие ограниченности сверху дисперсии относительных потерь избыточным риском. Далее мы рассмотрим некоторые случаи, когда подобное соотношение выполнено.

Задачи классификации: условия ограниченного шума . Рассмотрим задачу классификации $\mathcal{Y} = \{-1, +1\}$ с бинарной функцией потерь $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}$. Ранее было показано, что для задач без шума (когда $L^* = L(g^*) = 0$) скорость сходимости имеет порядок $O(n^{-1})$, а если мы не накладываем вообще никакие ограничения на уровень шума — $O(n^{-1/2})$. Возникает вопрос: что происходит в промежуточных ситуациях? Для этого нам понадобится способ измерения уровня шума в задаче.

Напомним, что Байесовский классификатор g^* имеет вид

$$g^*(x) = \operatorname{sgn}\left\{\eta(x) - 1/2\right\} = \operatorname{sgn}\left\{P(Y = 1|X = x) - 1/2\right\},$$

где $\eta(x)$ — функция апостериорной вероятности. Мы приведем два условия на уровень шума, часто используемые в теории статистического обучения. Первое условие, известное как *условие Массара* и использованное в [65], требует существование положительной константы $h > 0$, такой что с вероятностью 1 выполнено:

$$|2\eta(X) - 1| > h. \tag{3.21}$$

Это требование гарантирует, что значения функции $\eta(x)$ всегда удалены от критического значения $1/2$, соответствующего максимальному шуму. Обратим внимание, что для задач без шума выполнено условие $|2\eta(X) - 1| = 1$, то есть $P(Y = 1|X = x) \in \{0, 1\}$.

Второе условие, которое в литературе принято называть *условием Маммена–Цыбакова* [62, 98], может быть выражено несколькими эквивалентными способами. Мы приведем здесь один из них. Пусть существуют константы $\alpha \in [0, 1]$ и $B > 0$, такие что для всех $t \geqslant 0$ выполнено:

$$\mathbb{P}\{|2\eta(X) - 1| \leqslant t\} \leqslant Bt^{\frac{\alpha}{1-\alpha}}. \tag{3.22}$$

Это условие гарантирует, что значения функции $\eta(X)$ не слишком часто близки к критической точке $1/2$. В частности, случай $\alpha = 1$ соответствует жестким условиям (3.21).

Следующий результат, который может быть найден в [18], устанавливает справедливость соотношений между дисперсией относительных потерь и избыточным риском вида (3.20) в задачах классификации с ограниченным шумом, когда Байесовский классификатор принадлежит множеству $\mathcal{H}$:

Теорема 33. *Рассмотрим задачу классификации $\mathcal{Y} = \{-1, 1\}$ с бинарной функцией потерь. Пусть $g^* \in \mathcal{H}$ (и не обязательно $L^* = 0$). Тогда, если выполнено условие (3.21), то для всех $h \in \mathcal{H}$ справедливо:*

$$\mathbb{D}\left[\ell_h(X, Y) - \ell_{g^*}(X, Y)\right] \leqslant \mathbb{E}\left[\left(\ell_h(X, Y) - \ell_{g^*}(X, Y)\right)^2\right] \leqslant \frac{1}{h}\big(L(h) - L(g^*)\big).$$

Если же выполнено условие (3.22), то существует константа $c > 0$, такая что для всех $h \in \mathcal{H}$ справедливо:

$$\mathbb{D}\left[\ell_h(X, Y) - \ell_{g^*}(X, Y)\right] \leqslant \mathbb{E}\left[\left(\ell_h(X, Y) - \ell_{g^*}(X, Y)\right)^2\right] \leqslant c\big(L(h) - L(g^*)\big)^{\alpha}.$$

В качестве следствия Теорем 33 и 32 мы немедленно получаем следующий результат:

Следствие 9. *Рассмотрим задачу классификации $\mathcal{Y} = \{-1, 1\}$ с бинарной функцией потерь и конечным классом отображений $\mathcal{H} = \{h_1, \ldots, h_N\}$. Пусть $g^* \in \mathcal{H}$ (и не обязательно $L^* = 0$) и выполнено условие Массара (3.21). Тогда для любого $\delta \geqslant 0$ с вероятностью не меньше $1 - \delta$ справедливо:*

$$\mathcal{E}(\hat{h}_n) = O\left(\frac{\log N + \log \frac{1}{\delta}}{n}\right).$$

Задачи регрессии с выпуклым классом $\mathcal{H}$. Оказывается, соотношения вида (3.20) выполняются и для задач регрессии с квадратичной функцией потерь, если множество отображений $\mathcal{H}$ выпукло. Этот факт впервые был использован в работе [60] и далее обобщен в работах [6, 79].

Теорема 34. *Рассмотрим задачу регрессии с ограниченными ответами* $\mathcal{Y} \subseteq [0,1]$*, квадратичной функцией потерь* $\ell\colon (y', y'') \to (y' - y'')^2$ *и* выпуклым *множеством равномерно ограниченных отображений* $\mathcal{H}$*, таких что* $h(x) \in [0,1]$ *для всех* $x \in \mathcal{X}, h \in \mathcal{H}$*. Тогда для всех* $h \in \mathcal{H}$ *справедливо:*

$$\mathbb{D}\left[\ell_h(X,Y) - \ell_{h^*}(X,Y)\right] \leqslant \mathbb{E}\left[\big(\ell_h(X,Y) - \ell_{h^*}(X,Y)\big)^2\right] \leqslant 8\big(L(h) - L(h^*)\big).$$

Доказательство. Для начала заметим, что функция $\ell(y,\cdot)$ в данном случае — Липшицева на отрезке $[0,1]$ с константой 2. Действительно, для любых $a, b, x \in [0,1]$ справедливо:

$$|(a-x)^2 - (b-x)^2| = |(a-b)(a+b-2x)| \leqslant 2|a-b|.$$

Таким образом, справедливо:

$$\begin{aligned}
\mathbb{D}\left[\ell_h(X,Y) - \ell_{h^*}(X,Y)\right] &\leqslant \mathbb{E}\left[\big(\ell_h(X,Y) - \ell_{h^*}(X,Y)\big)^2\right] \\
&= \mathbb{E}\left[\Big(\big(Y - h(X)\big)^2 - \big(Y - h^*(X)\big)^2\Big)^2\right] \\
&\leqslant 4\mathbb{E}\left[\big(h(X) - h^*(X)\big)^2\right].
\end{aligned}$$

Далее воспользуемся тождеством

$$\frac{\big(Y - h(X)\big)^2 + \big(Y - h^*(X)\big)^2}{2} = \left(\frac{h(X) + h^*(X) - 2Y}{2}\right)^2 + \frac{1}{4}\big(h(X) - h^*(X)\big)^2.$$

Взяв математические ожидания от обеих сторон, получаем:

$$\frac{L(h) + L(h^*)}{2} = L\left(\frac{h(X) + h^*(X)}{2}\right) + \mathbb{E}\left[\frac{1}{4}\big(h(X) - h^*(X)\big)^2\right].$$

Поскольку $\mathcal{H}$ — выпуклое множество и h^* минимизирует средний риск на $\mathcal{H}$, то мы получаем

$$\frac{L(h) + L(h^*)}{2} \geqslant L(h^*) + \mathbb{E}\left[\frac{1}{4}\big(h(X) - h^*(X)\big)^2\right]$$

или

$$\mathbb{E}\left[\big(h(X) - h^*(X)\big)^2\right] \leqslant 2\big(L(h) - L(h^*)\big),$$

что завершает доказательство. ■

Замечание 17. *Последний результат был обобщен на более общий класс* L_q *потерь* $\ell(y', y'') = (y' - y'')^q$ *с* $q \geqslant 2$ *в работе [79].*

Однако, мы не можем применить последнюю теоремы вместе с Теоремой 32, поскольку для конечного класса $\mathcal{H}$ не будет выполнено требование выпуклости множества $\mathcal{H}$. Более того, известно, что для конечного класса $\mathcal{H}$ без каких либо ограничений на распределение P или Байесовский риск L^* порядок $O(n^{-1/2})$ является оптимальным [44].

Далее мы покажем, как обобщить описанные результаты на более общий случай бесконечных классов $\mathcal{H}$.

Задачи с шумом: бесконечный класс $\mathcal{H}$. В настоящем разделе мы рассмотрим наиболее общий случай задач с бесконечным классом $\mathcal{H}$, шумом $L^* > 0$ и без требований принадлежности Байесовского отображения g^* классу $\mathcal{H}$. Результаты, приведенные в настоящем параграфе, основаны на работах [5, 47, 52, 61, 66].

Мы будем называть *субкоренным* (subroot) неубывающее и неотрицательное отображение $\psi\colon [0,+\infty) \to [0,+\infty)$, такое что отображение $r \to \psi(r)/\sqrt{r}$ — невозрастающее для $r > 0$. Легко показать, что любое субкоренное отображение имеет единственную неподвижную точку. Следующий общий результат получен в работе [5]:

Теорема 35 ([5]). *Рассмотрим задачу с ограниченной в $[0,1]$ функцией потерь. Обозначим множество функций вида $\{(X,Y) \to \ell_h(X,Y) - \ell_{h^*}(X,Y)\colon h \in \mathcal{H}, (X,Y) \in \mathcal{X} \times \mathcal{Y}\}$ с помощью $\mathcal{F}^*_{\mathcal{H}}$. Пусть существует константа $B > 0$ и функционал $T\colon \mathcal{F}^*_{\mathcal{H}} \to \mathbb{R}^+$, такие что для всех $f \in \mathcal{F}^*_{\mathcal{H}}$ справедливо:*

$$\mathbb{D}[f(X,Y)] \leqslant T(f) \leqslant B \cdot \mathbb{E}[f(x)] = B \cdot \big(L(h) - L(h^*)\big). \tag{3.23}$$

Пусть, кроме того, существует субкоренное отображение $\psi_n(r)$, такое что справедливо:

$$B \cdot \mathbb{E}\left[\sup_{\substack{f \in \mathcal{F}^*_{\mathcal{H}}:\\ T(f) \leqslant r}} \mathbb{E}[f(X,Y)] - \frac{1}{n}\sum_{i=1}^{n} f(X_i,Y_i)\right] \leqslant \psi_n(r).$$

Пусть r^*_n — неподвижная точка отображения $\psi_n(r)$. Тогда для любого $\delta > 0$ с вероятностью не меньше $1-\delta$ справедливо:

$$\mathcal{E}(\hat{h}_n) \leqslant 151\frac{r^*_n}{B} + \frac{(6+7B)\log\frac{1}{\delta}}{n}.$$

Последний результат дает свободу выбора функционала T. Далее мы подробно рассмотрим частный случай последней теоремы, соответствующий специфическому выбору этого функционала. Поскольку, очевидно, справедливо следующее:

$$\mathbb{D}[\ell_h(X,Y) - \ell_{h^*}(X,Y)] \leqslant \mathbb{E}\left[\big(\ell_h(X,Y) - \ell_{h^*}(X,Y)\big)^2\right],$$

положив в прошлой теореме $T(f) = \mathbb{E}\left[\big(f(X,Y)\big)^2\right]$, мы тем самым удовлетворим первое неравенство в (3.23). Введем также следующее обозначение:

$$\mathcal{H}(r) = \left\{h \in \mathcal{H}\colon \mathbb{E}\left[\big(\ell_h(X,Y) - \ell_{h^*}(X,Y)\big)^2\right] \leqslant r\right\}. \tag{3.24}$$

Учитывая последнее соотношение, мы заключаем, что во множество $\mathcal{H}(r)$ входят отображения из $\mathcal{H}$ с малыми дисперсиями избыточных потерь. С учетом этих ремарок, следующий результат является следствием Теоремы 35:

Теорема 36. *Рассмотрим задачу с ограниченной в* $[0,1]$ *функцией потерь. Пусть существует константа* $B > 0$*, такая что для всех* $h \in \mathcal{H}$ *справедливо:*

$$\mathbb{E}\left[\big(\ell_h(X,Y) - \ell_{h^*}(X,Y)\big)^2\right] \leqslant B\big(L(h) - L(h^*)\big). \tag{3.25}$$

Пусть, кроме того, существует субкоренное отображение $\psi_n(r)$*, такое что справедливо:*

$$B \cdot \mathbb{E}\left[\sup_{h \in \mathcal{H}(r)} \big(L(h) - L_n(h)\big) - \big(L(h^*) - L_n(h^*)\big)\right] \leqslant \psi_n(r). \tag{3.26}$$

Пусть r_n^* *— неподвижная точка отображения* $\psi_n(r)$*. Тогда для любого* $\delta > 0$ *с вероятностью не меньше* $1 - \delta$ *справедливо:*

$$\mathcal{E}(\hat{h}_n) \leqslant 151\frac{r_n^*}{B} + \frac{(6 + 7B)\log\frac{1}{\delta}}{n}.$$

Прежде чем представить основные шаги доказательства мы приведем короткое обсуждение последнего результата.

Теорема 36 справедлива в наиболее общей постановке и не требует ни конечности класса отображений $\mathcal{H}$, ни отсутствия шума в задаче (условия $L^* = 0$), ни принадлежности Байесовского отображения g^* классу $\mathcal{H}$. Условие (3.25) мы уже рассматривали в предыдущих разделах и приводили ряд примеров, когда оно выполнено.

Величина, фигурирующая в левой части неравенства (3.26), называется *модулем непрерывности* эмпирического процесса в точке h^*. Теорема утверждает, что порядок убывания оценок избыточного риска в теории статистического обучения определяется величиной неподвижной точки модуля непрерывности эмпирического процесса, посчитанного вокруг лучшего в классе $\mathcal{H}$ отображения h^*. Похожие результаты были известны и использовались до этого в задачах, связанных с М-оценками [100].

Способ поиска субкоренной функции $\psi_n(r)$ и оценки ее неподвижной точки r_n^* мы вкратце обсудим далее. Главным здесь является тот факт, что во многих интересных случаях величина r_n^* имеет порядок $o(n^{-1/2})$, что дает нам оценку избыточного риска с быстрой скоростью сходимости.

Константы, фигурирующие в последней теореме, не являются оптимальными и, возможно, другой подход позволил бы их существенно уменьшить. Тем не менее, нашей главной целью было продемонстрировать, что быстрые скорости сходимости достигаются не только в частных случаях, рассмотренных ранее, но и в общих постановках задач минимизации риска.

Замечание 18. *Дальнейшие обсуждения последней теоремы и ее интерпретаций может быть найдено на Странице 132.*

Далее мы приведем основные идеи доказательства Теоремы 36. Теорема 35 доказывается аналогичным образом. Нам пригодится следующее короткое обозначение:

$$\hat{\mathbb{E}}_n[g] = \frac{1}{n}\sum_{i=1}^{n} g(X_i, Y_i) \tag{3.27}$$

для функций $g\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

Идея доказательства Теоремы 36: Мы не будем приводить здесь подробное доказательство данного результата и ограничимся лишь описанием основных шагов. Подробное доказательство может быть найдено в работе [5] или в Главе 4, где мы приведем с доказательством результат, воспроизводящий данный в *трансдуктивной постановке* теории статистического обучения.

Напомним также определения класса потерь $\mathcal{F}$ и класса избыточных потерь $\mathcal{F}^*$:

$$\mathcal{F} = \{\ell_h\colon h \in \mathcal{H}\}, \quad \mathcal{F}^* = \{\ell_h - \ell_{h^*}\colon h \in \mathcal{H}\},$$

где h^* — минимизатор среднего риска на множестве $\mathcal{H}$.

Шаг 1 Для начала мы зафиксируем произвольную $r > 0$ и введем следующую нормированную версию класса избыточных потерь:

$$\mathcal{G}_r = \left\{ \frac{r}{\Delta(r, f)} f,\ f \in \mathcal{F}^* \right\},$$

где нормировочный множитель $\Delta(r, f)$ выбран так, чтобы дисперсии функций множества $\mathcal{G}_r$ не превосходили r.

Шаг 2 Затем мы можем воспользоваться неравенством Талаграна Теоремы 13 и получить следующую верхнюю оценку для случайной величины

$$V_r = \sup_{g \in \mathcal{G}_r} (\mathbb{E}[g] - \hat{\mathbb{E}}_n[g]),$$

которая выполняется для любого $\delta > 0$ с вероятностью не меньше $1 - \delta$:

$$V_r \leqslant 2\mathbb{E}[V_r] + \sqrt{\frac{2r \log \frac{1}{\delta}}{n}} + \frac{8 \log \frac{1}{\delta}}{3n}.$$

Шаг 3 Далее мы воспользуемся техникой, известной в литературе как *пилинг* (peeling), которая заключается в «нарезании» множества $\mathcal{F}^*$ на «слои», каждый из которых состоит из функций с ограниченными в некотором интервале дисперсиями. Техника пилинга позволяет

нам получить неравенство $\mathbb{E}[V_r] \leqslant 5\psi_n(r)/B$. Также, воспользовавшись свойством субкоренных функций, мы получаем, что для любых $r \geqslant r_n^*$ справедливо $\psi_n(r) \leqslant \sqrt{rr_n^*}$. Откуда мы получаем, что для $r \geqslant r_n^*$:

$$V_r \leqslant \sqrt{r}\left(\frac{10\sqrt{r_n^*}}{B} + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}\right) + \frac{8\log\frac{1}{\delta}}{3n}.$$

Шаг 4 Теперь мы можем показать, что, выбрав определенным образом значение $r_0 > r_n^*$, мы получим, что для любой константы $K > 1$ справедливо:

$$V_{r_0} \leqslant \frac{r_0}{KB}.$$

Наконец, воспользовавшись определением V_r, мы получаем, что для любого $\delta > 0$ с вероятностью не менее $1 - \delta$ выполнено следующее:

$$\forall f \in \mathcal{F}^*,\ \forall K > 1: \quad \mathbb{E}[f] - \hat{\mathbb{E}}_n[f] \leqslant \frac{\Delta(r_0, f)}{r_0}\frac{r_0}{KB} = \frac{\Delta(r_0, f)}{KB}.$$

Шаг 5 Нам остается рассмотреть два случая $\mathbb{E}[f^2] > r_0$ и $\mathbb{E}[f^2] \leqslant r_0$, воспользоваться неравенством (3.25) и получить:

$$\forall f \in \mathcal{F}^*: \quad \mathbb{E}[f] - \hat{\mathbb{E}}_n[f] \leqslant 151\frac{r_n^*}{B} + \frac{(6+7B)\log\frac{1}{\delta}}{n}.$$

Наконец, заметив, что для $\hat{f}(X) = \ell_{\hat{h}_n}(X) - \ell_{h^*}(X)$ выполнено $\hat{\mathbb{E}}_n\hat{f} \leqslant 0$, мы завершаем доказательство. ■

Выбор субкоренных функций $\psi_n(r)$ и оценка их неподвижных точек r_n^*. Перейдем к очень короткому обсуждению выбора субкоренной функции $\psi_n(r)$, фигурирующей в последней теореме, и оценки ее неподвижной точки r_n^*.

Для начала обратим внимание, что с помощью неравенства симметризации Леммы 13 мы можем получить следующую верхнюю оценку модуля непрерывности, стоящего в правой части неравенства (3.26):

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}(r)}\big(L(h) - L_n(h)\big) - \big(L(h^*) - L_n(h^*)\big)\right] \leqslant 2\mathbb{E}\Big[R_n\Big(\{\ell_h - \ell_{h^*} : h \in \mathcal{H}(r)\}\Big)\Big],$$

где множество $\mathcal{H}(r)$ было определено в (3.24). Пользуясь обозначениями класса избыточных потерь $\mathcal{F}^*$ и (3.27), мы можем записать это неравенство короче:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}^*:\,\mathbb{E}[f^2]\leqslant r} \mathbb{E}[f] - \hat{\mathbb{E}}_n[f]\right] \leqslant 2\mathbb{E}\Big[R_n\Big(f \in \mathcal{F}^* : \mathbb{E}[f^2] \leqslant r\}\Big)\Big].$$

Величину, стоящую в правой части последнего неравенства, принято называть *локальной Радемахеровской сложностью* класса избыточных потерь $\mathcal{F}^*$. Сама по себе локальная Радемахеровская сложность не обязательно является субкоренным выражением. Большинство подходов к построению субкоренных отображений $\psi_n(r)$ (например, см. [52]) заключается в применении известных верхних оценок на локальную Радемахеровскую сложность. Эти оценки часто основаны на метрических энтропиях и других сложностных характеристиках класса отображений $\mathcal{H}$.

Мы рассмотрим два примера: (1) задачи с бинарной функцией потерь и классом $\mathcal{H}$ с конечной VC-размерностью и (2) задачи с Липшицевыми функциями потерь и классом $\mathcal{H}$, являющимся шаром в Гильбертовом пространстве, соответствующем воспроизводящему[9] ядру k.

Классы с конечной VC-размерностью. Следующий результат получен в [66, Раздел 4.3.1] и устанавливает связь между модулем непрерывности эмпирического процесса в точке h^* класса функций $\mathcal{H}$ с конечной VC-размерностью и функцией роста этого класса. Напомним, что в Разделе 3.2.1 мы ввели следующее определение функции роста $S_{\mathcal{H}}(n)$ класса отображений $\mathcal{H}$ при использовании бинарной функции ошибок ℓ:

$$S_{\mathcal{H}}(n) = \sup_{S} \Big| \big\{ \big(\ell_h(X_1, Y_1), \dots, \ell_h(X_n, Y_n)\big) \colon h \in \mathcal{H} \big\} \Big|.$$

Теорема 37 ([66]). *Рассмотрим задачу с бинарной функцией потерь ℓ. Пусть, кроме того, класс отображений $\mathcal{H}$ имеет конечную VC-размерность $h < \infty$. Тогда справедливо:*

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}(r)} \big(L(h) - L_n(h)\big) - \big(L(h^*) - L_n(h^*)\big) \right] \leqslant 24 \sqrt{\frac{r \log S_{\mathcal{H}}(n)}{n}} \leqslant 24 \sqrt{\frac{rh \log\left(\frac{en}{h}\right)}{n}} = \psi_n(r).$$

Кроме того, функция $\psi_n(r)$, очевидно, является субкоренной, и существует константа $C > 0$, такая что для ее неподвижной точки r_n^* справедливо:

$$r_n^* \leqslant C \frac{h \log\left(\frac{en}{h}\right)}{n}.$$

Мы немедленно получаем следующее следствие из Теорем 37, 33 и 36:

Следствие 10. *Рассмотрим задачу с бинарной функцией потерь ℓ и классом отображений $\mathcal{H}$, имеющим конечную VC-размерность $h < \infty$. Пусть, кроме того, Байесовское отображение g^* принадлежит классу $\mathcal{H}$, и выполнено условие на шум Массара (3.21). Тогда*

[9] Пространство, имеющее в машинном обучении непереводимое название *Reproducing Kernel Hilbert Space, RKHS*

для любого $\delta > 0$ *с вероятностью не меньше* $1 - \delta$ *справедливо:*

$$\mathcal{E}(\hat{h}_n) = O\left(\frac{h \log\left(\frac{en}{h}\right)}{n}\right).$$

Обратим внимание, что в предположениях последнего следствия нижняя оценка на обобщающую способность имеет порядок $O(1/n^{-1/2})$.

Классы в RKHS Теперь рассмотрим задачи с Липшицевыми функциями потерь и классом $\mathcal{H}$, определенным следующим образом. Рассмотрим *функцию ядра* $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ — симметричную положительно определенную функцию. Функция ядра порождает пространство функций $\{f(\cdot) = \sum_{i=1}^{m} \alpha_i k(x_i, \cdot)\colon m \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_i \in \mathcal{X}\}$. Вводя на этом множестве скалярное произведение $\left\langle \sum_{i=1}^{m_1} \alpha_i k(x'_i, \cdot), \sum_{j=1}^{m_2} \beta_j k(x''_j, \cdot) \right\rangle = \sum_{i=1}^{m_1} \sum_{i=j}^{m_2} \alpha_i \beta_j k(x'_i, x''_j)$, мы можем построить его замыкание и получим Гильбертово пространство $\mathcal{C}_k$. Норму в пространстве $\mathcal{C}_k$ мы будем обозначим $\|\cdot\|_k$. Интересной особенностью полученного пространства $\mathcal{C}_k$ является тот факт, что для всех $f \in \mathcal{C}_k$ и $x \in \mathcal{X}$ справедливо $f(x) = \langle f, k(x, \cdot)\rangle$. Положим

$$\mathcal{H}_k = \{f \in \mathcal{C}_k\colon \|f\|_k \leqslant 1\}.$$

Наконец, рассмотрим линейный интегральный оператор A_k, определяемый с помощью функции ядра k и распределения P на $\mathcal{X}$:

$$A_k f(x) = \int_{\mathcal{X}} K(x, z) f(z) dP(z),$$

и его собственные значения $\{\lambda_1, \lambda_2, \dots\}$, отсортированные в порядке убывания.

Следующий результат получен в [80, Теорема 2.1]:

Теорема 38 ([80]). *Для любого* $r > 0$ *выполнено:*

$$\mathbb{E}\left[\sup_{\substack{h \in \mathcal{H}_k: \\ \mathbb{E}[h^2(X)] \leqslant r}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(X_i)\right] \leqslant \left(\frac{2}{n} \sum_{i=1}^{\infty} \min(r, \lambda_i)\right)^{1/2}.$$

Воспользовавшись этим результатом вместе с Теоремой 35, мы получаем следующее следствие:

Следствие 11. *Рассмотрим задачу с ограниченно функцией потерь* $\ell \in [0, 1]$ *и классом отображений* $\mathcal{H} = \mathcal{H}_k$. *Пусть, кроме того, функция потерь* L*-Липшицева по первому аргументу и для некоторой положительной константы* B *и всех* $h \in \mathcal{H}$ *удовлетворяет условию:*

$$\mathbb{E}\left[\big(h(X) - h^*(X)\big)^2\right] \leqslant B\big(L(h) - L(h^*)\big).$$

Тогда для любого $\delta > 0$ с вероятностью не меньше $1 - \delta$ справедливо:

$$\mathcal{E}(\hat{h}_n) = O\left(\min_{q=0,1,\ldots}\left\{\frac{q}{n} + \sqrt{\frac{1}{n}\sum_{i>q}\lambda_i}\right\}\right).$$

Доказательство. Поскольку функция потерь ℓ L-Липшицева по своему первому аргументу, с учетом предположений следствия, для всех $h \in \mathcal{H}_k$ справедливо следующее неравенство:

$$\begin{aligned}\mathbb{D}[\ell_h(X,Y) - \ell_{h^*}(X,Y)] &\leqslant \mathbb{E}\left[\big(\ell_h(X,Y) - \ell_{h^*}(X,Y)\big)^2\right] \\ &\leqslant L^2\mathbb{E}\left[\big(h(X) - h^*(X)\big)^2\right] \\ &\leqslant BL^2\big(L(h) - L(h^*)\big).\end{aligned}$$

Мы собираемся воспользоваться Теоремой 35 с функционалом

$$T(f_h) = L^2\,\mathbb{E}\left[\big(h(X) - h^*(X)\big)^2\right],$$

где $f_h = \ell_h - \ell_{h^*}$. Обозначим

$$\mathcal{H}'(r) = \left\{h \in \mathcal{H}_k \colon L^2\,\mathbb{E}\left[\big(h(X) - h^*(X)\big)^2\right] \leqslant r\right\}.$$

Заметим, что в силу симметричности Радемахеровских случайных величин справедливо следующее:

$$\begin{aligned}\mathbb{E}\Big[R_n\Big(\{\ell_h - \ell_{h^*} \colon h \in \mathcal{H}'(r)\}\Big)\Big] &= \mathbb{E}\left[\sup_{h\in\mathcal{H}'(r)}\frac{1}{n}\sum_{i=1}^n \sigma_i\Big(\ell\big(h(X_i),Y_i\big) - \ell\big(h^*(X_i),Y_i\big)\Big)\right] \\ &= \mathbb{E}\left[\sup_{h\in\mathcal{H}'(r)}\frac{1}{n}\sum_{i=1}^n \sigma_i\ell\big(h(X_i),Y_i\big)\right] - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \sigma_i\ell\big(h^*(X_i),Y_i\big)\right] \\ &= \mathbb{E}\left[\sup_{h\in\mathcal{H}'(r)}\frac{1}{n}\sum_{i=1}^n \sigma_i\ell\big(h(X_i),Y_i\big)\right] \\ &= \mathbb{E}\Big[R_n\Big(\{\ell_h \colon h \in \mathcal{H}'(r)\}\Big)\Big].\end{aligned}$$

Положим $\phi_i(a) = \ell(a, Y_i) - \ell(0, Y_i)$ для $a \in \mathbb{R}$. Поскольку $\ell(\cdot, Y_i)$ — L-Липшицева функция, а прибавление константы не меняет Липшицевости, мы приходим к выводу, что все функции $\phi_i(a)$ L-Липшицевы. С учетом неравенства сжатия Теоремы 28, используя функции $\phi_i(a)$, мы получаем:

$$\begin{aligned}\mathbb{E}\Big[R_n\Big(\{\ell_h \colon h \in \mathcal{H}'(r)\}\Big)\Big] &= \mathbb{E}\left[\sup_{h\in\mathcal{H}'(r)}\frac{1}{n}\sum_{i=1}^n \sigma_i\ell\big(h(X_i),Y_i\big)\right] \\ &= \mathbb{E}\left[\sup_{h\in\mathcal{H}'(r)}\frac{1}{n}\sum_{i=1}^n \sigma_i\Big(\ell\big(h(X_i),Y_i\big) - \ell\big(0,Y_i\big)\Big)\right] \\ &\leqslant L\mathbb{E}\left[\sup_{h\in\mathcal{H}'(r)}\frac{1}{n}\sum_{i=1}^n \sigma_i h(X_i)\right].\end{aligned}$$

Мы получаем следующее неравенство:

$$\mathbb{E}\Big[R_n\Big(\{\ell_h - \ell_{h^*} \colon h \in \mathcal{H}'(r)\}\Big)\Big] \leqslant L\mathbb{E}\left[\sup_{h\in\mathcal{H}'(r)} \frac{1}{n}\sum_{i=1}^n \sigma_i h(X_i)\right].$$

С учетом неравенства симметризации справедливо следующее:

$$\begin{aligned} B\cdot\mathbb{E}\left[\sup_{h\in\mathcal{H}'(r)} \big(L(h) - L_n(h)\big) - \big(L(h^*) - L_n(h^*)\big)\right] &\leqslant 2B\mathbb{E}\Big[R_n\Big(\{\ell_h - \ell_{h^*} \colon h \in \mathcal{H}'(r)\}\Big)\Big] \\ &\leqslant 2BL\,\mathbb{E}\left[\sup_{h\in\mathcal{H}'(r)} \frac{1}{n}\sum_{i=1}^n \sigma_i h(X_i)\right]. \end{aligned}$$

В силу симметричности Радемахеровских случайны величин, мы получаем:

$$\begin{aligned} &\mathbb{E}\left[\sup_{h\in\mathcal{H}'(r)} \frac{1}{n}\sum_{i=1}^n \sigma_i h(X_i)\right] = \mathbb{E}\left[\sup_{h\in\mathcal{H}'(r)} \frac{1}{n}\sum_{i=1}^n \sigma_i\big(h(X_i) - h^*(X_i)\big)\right] \\ &= \mathbb{E}\left[\sup\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i\big(h(X_i) - h^*(X_i)\big) \colon h\in\mathcal{H}_k, L^2\,\mathbb{E}\left[\big(h(X) - h^*(X)\big)^2\right] \leqslant r\right\}\right] \\ &\leqslant \mathbb{E}\left[\sup\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i\big(h(X_i) - g(X_i)\big) \colon h,g\in\mathcal{H}_k, L^2\,\mathbb{E}\left[\big(h(X) - g(X)\big)^2\right] \leqslant r\right\}\right]. \end{aligned}$$

Поскольку класс $\mathcal{H}_k$ является выпуклым и симметричным, то для $f, g \in \mathcal{H}_k$ отображения $-g$ и $\frac{f+g}{2}$ также принадлежат классу $\mathcal{H}_k$. Мы можем продолжить цепочку неравенств:

$$\begin{aligned} &\mathbb{E}\left[\sup_{h\in\mathcal{H}'(r)} \frac{1}{n}\sum_{i=1}^n \sigma_i h(X_i)\right] \\ &\leqslant \mathbb{E}\left[\sup\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i\cdot 2\cdot\left(\frac{h(X_i)}{2} + \frac{-g(X_i)}{2}\right) : h,g\in\mathcal{H}_k, L^2\,\mathbb{E}\left[\big(h(X) - g(X)\big)^2\right] \leqslant r\right\}\right] \\ &= 2\mathbb{E}\left[\sup\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i h(X_i) \colon h\in\mathcal{H}_k, L^2\,\mathbb{E}\left[h^2(X)\right] \leqslant r/4\right\}\right]. \end{aligned}$$

Воспользовавшись Теоремой 38 мы получаем:

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}'(r)} \frac{1}{n}\sum_{i=1}^n \sigma_i h(X_i)\right] \leqslant \left(\frac{2}{n}\sum_{i=1}^\infty \min\left(\frac{r}{4L^2}, \lambda_i\right)\right)^{1/2}$$

или

$$B\cdot\mathbb{E}\left[\sup_{h\in\mathcal{H}'(r)} \big(L(h) - L_n(h)\big) - \big(L(h^*) - L_n(h^*)\big)\right] \leqslant 2BL\left(\frac{2}{n}\sum_{i=1}^\infty \min\left(\frac{r}{4L^2}, \lambda_i\right)\right)^{1/2}.$$

Обозначим правую часть последнего неравенства $\psi(r)$. Несложно проверить, что $\psi(r)$ является субкоренной функцией и мы можем использовать ее в Теореме 35. Нам остается оценить сверху неподвижную точку r^* функции $\psi(r)$:

$$r^* = 2BL\left(\frac{2}{n}\sum_{i=1}^\infty \min\left(\frac{r^*}{4L^2}, \lambda_i\right)\right)^{1/2}. \tag{3.28}$$

Поскольку собственные значения отсортированы в порядке убывания, легко заметить, что

$$\sum_{i=1}^{\infty} \min\left(\frac{r}{4L^2}, \lambda_i\right) = \min_{W \subseteq \{1,2,\dots\}} \left\{ \sum_{i \in W} \frac{r}{4L^2} + \sum_{i \notin W} \lambda_i \right\}$$
$$= \min_{q=0,1,\dots} \left\{ \frac{qr}{4L^2} + \sum_{i>q} \lambda_i \right\}.$$

Подставляя это тождество в (3.28), мы получаем:

$$r^* = 2BL \left(\frac{2}{n} \min_{q=0,1,\dots} \left\{ \frac{qr^*}{4L^2} + \sum_{i>q} \lambda_i \right\} \right)^{1/2}.$$

Решая неравенство для разных значений q и выбрав минимальное из найденных решений, мы получаем, что для некоторой константы $C_{B,L}$, зависящей лишь от L и B, справедливо:

$$r^* \leqslant C_{L,B} \min_{q=0,1,\dots} \left\{ \frac{q}{n} + \sqrt{\frac{1}{n} \sum_{i>q} \lambda_i} \right\}.$$

■

Замечание 19. *Условия на функцию потерь в последнем следствии не являются чрезвычайно жесткими. Например, мы видели, что они выполняются для квадратичных потерь (см. доказательство Теоремы 34).*

Оценки, вычислимые по данным. Мы завершаем настоящую главу следующим коротким замечанием. Все оценки, приведенные в последнем параграфе, не являются вычислимыми по данным, поскольку зависят от неизвестных параметров распределения. Действительно, оценки в Теоремах 35 и 36 зависят от верхней оценки на модуль непрерывности эмпирического процесса, зависящий от неизвестного распределения P, а последняя оценка для классов в RKHS зависит от собственных значений интегрального оператора, определенного для маргинального распределения $P_{\mathcal{X}}$. Оказывается, переходя к рассмотрению локальных Радемахеровских сложностей, для большинства результатов становится возможным получить аналоги, полностью вычислимые по данным. Так, подмножества $\left\{h \in \mathcal{H} \colon \mathbb{E}\left[\left(\ell_h(X,Y) - \ell_{h^*}(X,Y)\right)^2\right]\right\}$, на которых определены локальные сложности последних результатов, могут быть заменены на их эмпирические аналоги $\left\{h \in \mathcal{H} \colon \hat{\mathbb{E}}_n\left[\left(\ell_h(X,Y) - \ell_{\hat{h}_n}(X,Y)\right)^2\right]\right\}$. Собственные же значения интегрального оператора, фигурирующие в Следствии 11, могут быть заменены на собственные значения нормированной матрицы Грамма $\frac{1}{n}\big(k(X_i,X_j)\big)_{i,j=1,\dots,n}$, посчитанной по обучающей выборке. В настоящей работе мы не будем рассматривать эти результаты. Подробности могут быть найдены в работах [5, 48, 50].

Коротко подведем итоги настоящей главы:

- Теория статистического обучения позволяет исследовать свойства процедуры обучения на основе наблюдаемых данных в рамках строгой математической вероятностной формулировки.

- Теория Вапника–Червоненкиса (или VC-теория) основана на переходе к рассмотрению равномерных (по классу отображений) уклонений средних выборочных значений потерь от их математического ожидания. Это позволяет использовать богатый и хорошо известный математический аппарат теорий эмпирических процессов и неравенств концентрации вероятностной меры.

- Результаты, основанные на применении неравенств Хефдинга и МакДиармида, как правило, ведут к оценкам медленного порядка $O(n^{-1/2})$. Этот же порядок мы получим, если не будем накладывать на класс рассматриваемых задач никаких дополнительных ограничений, помимо ограниченности функции потерь.

- К результатам быстрого порядка $o(n^{-1/2})$ ведут дополнительные предположения, накладываемые на классы отображений и неизвестное распределение, учет которых осуществляется применением неравенств Бернштейна и Талаграна.

4 Трансдуктивное обучение

В прошлой главе существенную роль играло предположение о том, что объекты обучающей выборки *независимы*. Этот ключевой факт позволил воспользоваться многими известными результатами из теории неравенств концентрации вероятностной меры и получить достаточно сильные результаты, в особенности результаты Раздела 3.2.3, основанные на неравенстве Талаграна. Однако, во многих случаях это предположение может не выполняться. Одним примером может служить ситуация, когда между объектами существуют некие временные зависимости. Представим задачи из области финансов, связанные с акциями: цены разных акций в один и тот же день или одной и той же акции в различные дни могут быть взаимосвязаны. В таких случаях неравенство Талаграна, к сожалению, перестает быть применимым, и нам необходимо заменить его другими результатами.

Другим примером, когда предположение о независимости не выполняется, является *трансдуктивная* постановка теории статистического обучения. В данной постановке предполагается, что объекты обучающей выборки выбраны равномерно *без возвращений* из конечной генеральной совокупности. Алгоритм обучения получает доступ к объектам обучающей выборки с ответами на них и объектам контрольной выборки. Задача в этом случае заключается в предсказании ответов на объектах контрольной выборки. Эта постановка естественным образом возникает практически во всех современных приложениях, включая анализ текста, вычислительную биологию, рекомендательные системы, компьютерное зрение и многие другие. В качестве примера рассмотрим задачу категоризации изображений. В этом случае мы имеем дело со *всеми* изображениями, хранящимися в интернете, и лишь небольшое число из них имеют метки классов (возможно, проставленные вручную экспертами). Задача заключается в предсказании меток классов всех неразмеченных изображений, что позволило бы в будущем предоставить доступ к этим изображениям через поисковые системы.

Однако, с теоретической точки зрения, трансдуктивное обучение изучено менее подробно, чем более стандартная индуктивная постановка, рассмотренная в прошлой главе. Ряд оценок обобщающей способности для трансдуктивной постановки был получен в серии работ [15, 24, 25, 29, 34, 103, 104], включая первые результаты, основанные на глобальной Ра-

демахеровской сложности [34]. Однако, в отличие от индуктивного обучения, ни один из известных в трансдуктивном обучении результатов не ведет к оценкам с быстрой скоростью сходимости, справедливых в общих предположениях[1].

В настоящей главе мы рассмотрим трансдуктивную постановку теории статистического обучения с произвольной ограниченной функцией потерь. Нашими основными инструментами будут неравенства концентрации вероятностной меры, полученные в Главе 2 и справедливые для выборок без возвращений. В особенности часто мы будем пользоваться неравенствами типа Беннетта для супремумов эмпирических процессов Теорем 20 и 22. Мы получим оценки избыточного риска, основанные на *локальных* мерах сложности рассматриваемого класса отображений. Эти оценки справедливы в достаточно общих предположениях и могут считаться аналогами Теорем 35 и 36 для трансдуктивной постановки. Кроме того, они справедливы для многих часто используемых на практике функций потерь ℓ (включая квадратичную) и классов отображений $\mathcal{H}$. В частности, применив полученные оценки к классам в RKHS, мы получим первые в трансдуктивной постановке оценки избыточного риска, зависящие от суммы собственных значений интегрального оператора, соответствующего рассматриваемому ядру. Также мы приведем новые оценки обобщающей способности для трансдуктивной постановки, учитывающие дисперсии потерь отображений из рассматриваемого класса $\mathcal{H}$.

Все результаты настоящей главы являются новыми и опубликованы в работе [96]. К личному вкладу автора не относится формулировка и доказательство Следствия 15, принадлежащие одному из соавторов.

4.1 Постановка задачи и обзор известных результатов

С теоретической точки зрения, основным отличием между трансдуктивной и индуктивной постановками теории статистического обучения является способ получения обучающей выборки S. Как мы видели в прошлой главе, объекты обучающей выборки в индуктивной постановке независимы и одинаково распределены в соответствии с неизвестным вероятностным распределением P на декартовом произведении $\mathcal{X} \times \mathcal{Y}$ пространств объектов $\mathcal{X}$ и ответов $\mathcal{Y}$. На основе обучающей выборки S процедура обучения выбирает отображение h из некоторого заранее выбранного класса отображений $\mathcal{H}$. Целью является поиск отображения

[1] Исключениями являются работы [15,24], которые, однако, рассматривают задачи без шума и предполагают принадлежность Байесовского классификатора рассматриваемому классу отображений. Как обсуждалось в прошлой главе, это предположение является чересчур сильным.

с наименьшим средним риском $\mathbb{E}_{(X,Y)\sim P}[\ell\big(h(X),Y\big)] \to \min_{h\in\mathcal{H}}$, где ℓ — заранее выбранная, ограниченная и неотрицательная функция потерь $\ell\colon \mathcal{Y}^2 \to [0,1]$.

В настоящей главе мы будем пользоваться первой из двух постановок[2] трансдуктивного обучения, сформулированных в [104], которая часто рассматривается в теории обучения (например, в [29, 34]). Пусть нам дано произвольное множество $\boldsymbol{X}_N$ из N объектов (мы не будем делать никакие предположения о способе получения множества $\boldsymbol{X}_N$). Затем мы выбираем $m \leqslant N$ объектов[3] $\boldsymbol{X}_m \subseteq \boldsymbol{X}_N$ равномерно без возвращений из множества $\boldsymbol{X}_N$. Как обсуждалось ранее, объекты $\boldsymbol{X}_m$ в этом случае не являются независимыми. Наконец, мы получаем ответы $\boldsymbol{Y}_m$ для объектов $\boldsymbol{X}_m$, выбирая ответ Y для каждого объекта $X \in \boldsymbol{X}_m$ из неизвестного нам условного распределения $P(Y|X)$. Мы обозначим обучающую выборку с помозью $S_m = (\boldsymbol{X}_m, \boldsymbol{Y}_m)$. Оставшиеся объекты без ответов на них $\boldsymbol{X}_u = \boldsymbol{X}_N \setminus \boldsymbol{X}_m$, $u = N - m$, формируют контрольную выборку. В настоящей главе мы рассмотрим частный случай, часто рассматривающийся в литературе (включая работы [29] и [34]), когда ответы получаются применением неизвестной, но неслучайной целевой функции $\phi\colon \mathcal{X} \to \mathcal{Y}$. В этом случае, очевидно, $P\big(Y = \phi(X)|X\big) = 1$ с вероятностью 1. Процедура обучения на основе обучающей выборки S_m и объектов контрольной выборки $\boldsymbol{X}_u$ выбирает отображение h из некоторого заранее выбранного класса отображений $\mathcal{H}$ (не обязательно содержащего целевую функцию ϕ). Для краткости мы будем пользоваться обозначением $\ell_h(X) = \ell\big(h(X), \phi(X)\big)$. Определим соответственно ошибку на обучении (эмпирический риск) и ошибку на контроле отображения h следующим образом:

$$\hat{L}_m(h) = \frac{1}{m}\sum_{X\in\boldsymbol{X}_m} \ell_h(X), \quad L_u(h) = \frac{1}{u}\sum_{X\in\boldsymbol{X}_u} \ell_h(X).$$

Крышечка в данном случае подчеркивает тот факт, что эмпирический риск может быть вычислен по имеющимся данным. В доказательствах нам также потребуется рассматривать ошибку отображения h на полной выборке $L_N(h) = \frac{1}{N}\sum_{X\in\boldsymbol{X}_N} \ell_h(X)$. Заметим, что для отображения h, не зависящего от обучающей выборки, величина $L_N(h)$ не является случайной. Главной задачей процедуры обучения является поиск отображения с минимальной ошибкой на контроле $L_u(h) \to \inf_{h\in\mathcal{H}}$, которое мы обозначим с помощью h^*_u.

[2] Вторая постановка предполагает, что обучающая и контрольная выборки выбраны независимо из одного и того же неизвестного распределения. Процедура обучения получает на вход объекты обучающей выборки с ответами на них, а также объекты контрольной выборки. В работе [104] показано, что любая верхняя оценка на величину $L_u(h) - \hat{L}_m(h)$, справедливая в первой постановке, также справедлива и во второй.

[3] В настоящей главе размер обучающей выборки обозначается с помощью m (а не n, как в прошлой главе) исключительно во избежание путаницы между двумя рассматриваемыми в работе постановками.

Поскольку ответы объектов контрольной выборки неизвестны, мы не можем вычислить $L_u(h)$ и должны оценить эту величину на основе обучающей выборки S_m. Мы будем пользоваться тем же способом, который рассматривался в прошлой главе, и оценим ошибку на контроле с помощью эмпирического риска, переходя к рассмотрению процедуры *минимизации эмпирического риска* $\hat{L}_m(h) \to \min_{h\in\mathcal{H}}$. Мы будем обозначать решение задачи минимизации эмпирического риска (МЭР) с помощью $\hat{h}_m$ и использовать его в качестве приближения отображения h_u^*. Для отображения $h \in \mathcal{H}$ определим его *избыточный риск*:

$$\mathcal{E}_u(h) = L_u(h) - \inf_{g\in\mathcal{H}} L_u(g) = L_u(h) - L_u(h_u^*).$$

Естественным вопросом является следующий: насколько хорошо отображение $\hat{h}_m$, выбранное процедурой МЭР, приближает оптимальное отображение h_u^*? Для ответа на этот вопрос мы будем изучать свойства величины $\mathcal{E}_u(\hat{h}_m)$. Поскольку величина $\mathcal{E}_u(\hat{h}_m)$ случайна, нашей главной задачей в настоящей главе будет получение ее верхних оценок, справедливых с большой вероятностью относительно случайно реализации обучающей выборки. Как и в индуктивной постановке мы можем также рассматривать оценки обобщающей способности, ограничивающие сверху разность $L_u(\hat{h}_m) - \hat{L}_m(\hat{h}_m)$. Еще раз обратим внимание на то, что оба отображения $\hat{h}_m$ и h_u^* случайны, поскольку зависят от разбиения генеральной совокупности на обучающую и контрольную подвыборки. Более того, в отличие от индуктивной постановки (где избыточный риск отображения $h \in \mathcal{H}$, не зависящего от обучающей выборки, не был случайным), в трансдуктивном обучении величина $\mathcal{E}_u(h)$ является случайной.

Переход к рассмотрению равномерных по классу $\mathcal{H}$ откланений, описанный в прошлой главе, применим и в данном случае:

$$L_u(\hat{h}_m) - \hat{L}_m(\hat{h}_m) \leqslant \sup_{h\in\mathcal{H}} L_u(h) - \hat{L}_m(h). \tag{4.1}$$

В правой части неравенства, таким образом, мы получаем супремум эмпирического процесса для выборок без возвращений. Эти рассуждения демонстрируют, что в анализе трансдуктивной постановки важную роль должны играть верхние оценки на функции $f(Z_1, \ldots, Z_m)$, справедливые с большой вероятностью, где $\{Z_1, \ldots, Z_m\}$ — случайные величины, выбранные *без возвращений* из конечного множества. В частности, нас могут интересовать неравенства концентрации для супремумов эмпирических процессов для выборок без возвращений, рассмотренные в Разделе 2.3.

Обзор известных результатов. Здесь мы приведем короткий обзор известных оценок обобщающей способности и избыточного риска для трансдуктивного обучения. Заинтересо-

ванный читатель может найти исчерпывающий обзор существующих результатов в работе [81].

Первые оценки для бинарных потерь, представленные в работе [103], были *неявными*, в том смысле что вмето аналитической формулы они предоставляли описание вычислительной процедуры, дающей на выходе численное значение оценки. Несколько улучшенная версия этих оценок была представлена позднее в [15]. Как было показано в прошлой главе, жесткие ограничения на рассматриваемые задачи могут вести к оценкам с быстрой скоростью сходимости (порядка $o(1/m^{-1/2})$). Например, в работе [15] получена оценка порядка $\frac{1}{\min(u,m)}$ для задач без шума, в которых отображение ϕ принадлежит рассматриваемому классу $\mathcal{H}$. Однако, эти предположения являются мало реалистичными.

Авторы [24] рассматривают задачи регрессии в трансдуктивной постановке с ограниченными квадратичными потерями и получают оценку обобщающей способности порядка $\sqrt{\hat{L}_m(\hat{h}_m)\frac{\log N}{\min(m,u)}}$, который также в общем случае не является быстрым. В работах [15,29] были представлены несколько PAC-Байесовских оценок, которые, однако, существенно зависят от *априорного распределения*, выбираемого процедурой обучения *до получения обучающей выборки* (подробнее PAC-Байесовские оценки будут рассмотрены в Главе 6). Трансдуктивные оценки, основанные на определении алгоритмической стабильности [22], были получены для задач классификации в [33] и регрессии в [25]. Однако, все они имеют *медленный* порядок $\min(u,m)^{-1/2}$. Наконец, в работе [34] приведены первые результаты трансдуктивной постановки, основанные на Радемахеровских сложностях. Однако, эти результаты были основаны на *глобальной* Радемахеровской сложности, которая, как мы знаем из прошлой главы, идет в паре с неравенством МакДиармида и ведет к медленным скоростям сходимости.

4.2 Трансдуктивные оценки избыточного риска и локальные меры сложности

Мы начнем с доказательства достаточно простых оценок обобщающей способности, на примере которых мы продемонстрируем способы применения результатов Раздела 2.3 в рассматриваемой постановке.

Оценки обобщающей способности. Заметим, что неравенство (4.1) может быть переписано в следующем виде:

$$L_u(\hat{h}_m) - \hat{L}_m(\hat{h}_m) \leqslant \sup_{h\in\mathcal{H}} L_u(h) - \hat{L}_m(h) = \frac{N}{u}\cdot\sup_{h\in\mathcal{H}} L_N(h) - \hat{L}_m(h),$$

где мы воспользовались тем фактом, что $N \cdot L_N(h) = m \cdot \hat{L}_m(h) + u \cdot L_u(h)$. Также заметим, что для любого $h \in \mathcal{H}$ справедливо $L_N(h) - \hat{L}_m(h) = \frac{1}{m} \sum_{X \in \boldsymbol{X}_m} \big(L_N(h) - \ell_h(X)\big)$, где объекты $\boldsymbol{X}_m$ выбраны равномерно без возвращений из $\boldsymbol{X}_N$. Обратим внимание, что $L_N(h) - \ell_h(X) \in [-1, 1]$ и $\mathbb{E}[L_N(h) - \ell_h(X)] = L_N(h) - \mathbb{E}[\ell_h(X)] = 0$. Таким образом, мы можем воспользоваться результатами Раздела 2.3, где в качестве $\mathcal{C}$ будет выступать $\boldsymbol{X}_N$, а в качестве класса отображений — множество $\mathcal{F}_\mathcal{H} = \{f_h \colon f_h(X) = L_N(h) - \ell_h(X), h \in \mathcal{H}\}$, определяемое классом $\mathcal{H}$. Мы можем получить оценки для $\sup_{f_h \in \mathcal{F}_\mathcal{H}} \sum_{X \in \boldsymbol{X}_m} f_h(X) = m \cdot \sup_{h \in \mathcal{H}} \big(L_N(h) - \hat{L}_m(h)\big)$, справедливые с большой вероятностью. Не будем забывать, что в Разделе 2.3 мы имели дело с ненормированными суммами, и поэтому в предыдущем тождестве возникает множитель m. Как уже было замечено, для отображения $h \in \mathcal{H}$, не зависящего от обучающей выборки, величина $L_N(h)$ не является случайной. Также прибавление константы к случайной величине не меняет ее дисперсии. С учетом этих наблюдений, определим:

$$\sigma^2_\mathcal{H} = \sup_{f_h \in \mathcal{F}_\mathcal{H}} \mathbb{D}[f_h(X)] = \sup_{h \in \mathcal{H}} \mathbb{D}[\ell_h(X)] = \sup_{h \in \mathcal{H}} \left(\frac{1}{N} \sum_{X \in \boldsymbol{X}_N} \big(\ell_h(X) - L_N(h)\big)^2 \right). \tag{4.2}$$

С помощью Теорем 20 и 22 мы можем получить следующие оценки обобщающей способности, справедливые *без каких-либо дальнейших предположений* о свойствах рассматриваемых задач, кроме ограниченности функции потерь в интервале $[0, 1]$. Первый результат немедленно вытекает из Теоремы 20:

Теорема 39. *Для любого $\delta > 0$ с вероятностью не меньше $1 - \delta$ справедливо:*

$$\forall h \in \mathcal{H} : \quad L_N(h) - \hat{L}_m(h) \leqslant \mathbb{E}\left[\sup_{h \in \mathcal{H}} \big(L_N(h) - \hat{L}_m(h)\big)\right] + 2\sqrt{2\left(\frac{N}{m^2}\right)\sigma^2_\mathcal{H} \log \frac{1}{\delta}},$$

где величина $\sigma^2_\mathcal{H}$ определена в (4.2).

Пусть $\{\xi_1, \ldots, \xi_m\}$ — случайные величины, выбранные равномерно *с возвращениями* из множества $\boldsymbol{X}_N$. Введем следующее обозначение:

$$E_m = \mathbb{E}\left[\sup_{h \in \mathcal{H}} \left(L_N(h) - \frac{1}{m} \sum_{i=1}^{m} \ell_h(\xi_i) \right)\right].$$

Следующий результат следует из Теоремы 22 и несложных вычислений. Подробное доказательство будет приведено в Разделе 4.3.

Теорема 40. *Для любого $\delta > 0$ с вероятностью не меньше $1 - \delta$ справедливо:*

$$\forall h \in \mathcal{H} : \quad L_N(h) - \hat{L}_m(h) \leqslant 2E_m + \sqrt{\frac{2\sigma^2_\mathcal{H} \log \frac{1}{\delta}}{m}} + \frac{4 \log \frac{1}{\delta}}{3m},$$

где величина $\sigma^2_\mathcal{H}$ определена в (4.2).

Замечание 20. *Обратим внимание, что E_m — математическое ожидание супремум-нормы «обычного» эмпирического процесса, фигурировавшего в прошлой главе при исследовании индуктивной постановки. Пользуясь неравенством симметризации (см. Лемму 13), мы можем оценить его сверху с помощью супремум-нормы Радемахеровского процесса. Применив эти размышления к последней теореме, мы получим оценку величины $\sup_{h\in\mathcal{H}}\big(L_N(h) - \hat{L}_m(h)\big)$, в точности совпадающую с оценкой Теоремы 2.1 работы [5] (с параметрами $\alpha = 1$ и $(b-a) = 1$).*

Приведем короткое обсуждение двух последних оценок обобщающей способности. Поскольку $\sigma^2_{\mathcal{H}}$ — дисперсия случайной величины, ограниченной в интервале $[0,1]$, то справедливо $\sigma^2_{\mathcal{H}} \leqslant 1/4$. Таким образом, оценка Теоремы 40 имеет порядок $m^{-1/2}$, поскольку типичным порядком[4] величины E_m является $m^{-1/2}$. Повторяя доказательство Леммы 10, мы немедленно получим следующее следствие:

Следствие 12. *Пусть случайные величины $\{\xi_1, \ldots, \xi_m\}$ выбраны равномерно* с возвращениями *из множества $\boldsymbol{X}_N$. Тогда для любого счетного множества $\mathcal{F}$ отображений, определенных на $\boldsymbol{X}_N$, справедливо следующее:*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\mathbb{E}[f(X)] - \frac{1}{m}\sum_{X\in\boldsymbol{X}_m} f(X)\right] \leqslant \mathbb{E}\left[\sup_{f\in\mathcal{F}}\mathbb{E}[f(X)] - \frac{1}{m}\sum_{i=1}^{m} f(\xi_i)\right].$$

Следствие показывает, что для $m = \Omega(N)$ оценка Теоремы 39 также имеет порядок $m^{-1/2}$. Однако, при $m = o(N)$ скорость ее сходимости замедляется. При $m = o(N^{1/2})$ оценка и вовсе расходится.

Замечание 21. *Последнее следствие позволяет нам использовать в трансдуктивном обучении хорошо известные результаты, связанные с индуктивными Радемахеровскими процессами, включая неравенства симметризации и сжатия, приведенные в прошлой главе. Позже в данном разделе эти рассуждения позволят нам получить оценки избыточного риска для классов $\mathcal{H}$ в RKHS, зависящие от суммы собственных значений интегрального оператора, соответствующего спрямляющему ядру k. Тем не менее, мы подчеркиваем, что между величинами $\mathbb{E}[Q_m]$ и $\mathbb{E}[Q'_m]$ (см. определения Раздела 2.3) может существовать значительный «зазор», и в этом случае такие рассуждения могут привести к плохим результатам.*

Оценки избыточного риска. Главной задачей настоящего раздела является воспроизведение хорошо известных в индуктивной постановке теории статистического обучения ре-

[4] Например, если множество $\mathcal{F}$ конечно, это следует из Теорем 2.1 и 3.5 работы [48].

зультатов *локального анализа*, развитого в работах [5, 47, 52, 66] и описанного ранее в Разделе 3.2.3, также в трансдуктивной постановке. Напомним, что эти результаты устанавливают связь между величиной избыточного риска и неподвижной точкой модуля непрерывности эмпирического процесса, соответствующего рассматриваемому классу $\mathcal{H}$. Нашими основными инструментами снова будут неравенства концентрации для супремумов эмпирических процессов и выборок без возвращений Теорем 20 и 22.

Далее мы будем пользоваться следующими операторами, отображающими функции f, определенные на $\boldsymbol{X}_N$, в $\mathbb{R}$:

$$Ef = \frac{1}{N}\sum_{X\in\boldsymbol{X}_N} f(X), \quad \hat{E}_m f = \frac{1}{m}\sum_{X\in\boldsymbol{X}_m} f(X).$$

В частности, в терминах введенных операторов, справедливо $L_N(h) = E\ell_h$ и $\hat{L}_m(h) = \hat{E}_m\ell_h$.

Напомним определение *класса избыточных потерь* $\mathcal{F}^* = \{\ell_h - \ell_{h_N^*}, h \in \mathcal{H}\}$. На протяжении всего раздела мы будем предполагать, что функция потерь ℓ и класс $\mathcal{H}$ удовлетворяют следующим предположениям:

Предположения 1.

1. *Существует отображение* $h_N^* \in \mathcal{H}$, *такое что* $L_N(h_N^*) = \inf_{h\in\mathcal{H}} L_N(h)$.
2. *Существует константа* $B > 0$, *такая что для всех* $f \in \mathcal{F}^*$ *выполнено*
$$Ef^2 \leqslant B \cdot Ef.$$
3. *Функция потерь* ℓ *ограничена в интервале* $[0, 1]$.

Коротко обсудим эти предположения. Предположение 1.1 широко используется в теории обучения и не является жестким. Предположение 1.2, устанавливающее определенную связь между дисперсией избыточных потерь и избыточным риском, знакомо нам из прошлой главы. Напомним, что оно выполнено, например, для квадратичной функции потерь ℓ и равномерно ограниченного выпуклого класса $\mathcal{H}$ (другие примеры могут быть найдены в Разделе 5.2 работы [5] и Разделе 2.1 работы [10]). Предположение 1.3, используемое нами на протяжении всей настоящей работы, возможно, может быть ослаблено применением аналогов Теорем 20 и 22, справедливых для неограниченных функций [5].

[5] В работе [1] приводятся версии неравенства Талаграна для неограниченных функций в случае независимых и одинаково распределенных случайных величин.

Далее мы приводим главные результаты настоящей главы, которые являются трансдуктивными аналогами[6] Теоремы 36, полученной в работе [5]. Все результаты идут парами, обусловленными применением Теоремы 20 или 22 в доказательстве. Напомним определение субкоренной функции. *Субкоренной* мы будем называть неубывающую и неотрицательную функцию $\psi\colon [0,\infty) \to [0,\infty)$, такую что отображение $r \to \psi(r)/\sqrt{r}$ является невозрастающим для $r > 0$. Можно показать, что у любой субкоренной функции существует единственная положительная неподвижная точка.

Теорема 41. *Пусть $\mathcal{H}$ и ℓ удовлетворяют Предположениям 1. Пусть существует субкоренная функция $\psi_m(r)$, такая что*

$$B\,\mathbb{E}\left[\sup_{f\in\mathcal{F}^*:Ef^2\leqslant r}(Ef-\hat{E}_m f)\right]\leqslant\psi_m(r). \tag{4.3}$$

Обозначим с помощью r_m^* неподвижную точку функции $\psi_m(r)$. Тогда для любого $\delta > 0$ с вероятностью не меньше $1-\delta$ справедливо:

$$L_N(\hat{h}_m)-L_N(h_N^*)\leqslant 51\frac{r_m^*}{B}+17B\left(\frac{N}{m^2}\right)\log\frac{1}{\delta}.$$

Обратим внимание на то, что константы последнего результата немного меньше констант Теоремы 36. Этот результат основан на Теореме 20 и имеет все ее недостатки: оценка не сходится для $m = o(N^{-1/2})$. Пользуясь в доказательстве Теоремой 22 вместо Теоремы 20, мы можем заменить множитель N/m^2, фигурирующий в прошлой оценке, на множитель $1/m$ (ценой незначительного увеличения констант):

Теорема 42. *Пусть $\mathcal{H}$ и ℓ удовлетворяют Предположениям 1. Пусть случайные величины $\{\xi_1,\dots,\xi_m\}$ выбраны равномерно* с возвращениями *из $\boldsymbol{X}_N$. Пусть существует субкоренная функция $\psi_m(r)$, такая что:*

$$B\cdot\mathbb{E}\left[\sup_{f\in\mathcal{F}^*:Ef^2\leqslant r}\left(Ef-\frac{1}{m}\sum_{i=1}^m f(\xi_i)\right)\right]\leqslant\psi_m(r). \tag{4.4}$$

Обозначим с помощью r_m^* неподвижную точку функции $\psi_m(r)$. Тогда для любого $\delta > 0$ с вероятностью не меньше $1-\delta$ справедливо:

$$L_N(\hat{h}_m)-L_N(h_N^*)\leqslant 901\frac{r_m^*}{B}+\frac{\log\frac{1}{\delta}(16+25B)}{3m}.$$

Также заметим, что в Теореме 42 модуль непрерывности эмпирического процесса для выборок без возвращений, фигурирующий в левой части неравенства (4.3), заменен на его

[6] Повторяя рассуждения, приведенные в настоящем разделе, мы можем получить также трансдуктивные аналоги более общей Теоремы 35.

индуктивный аналог. Как показывает Следствие 12, неподвижная точка r_m^* Теоремы 41 может оказаться меньше неподвижной точки Теоремы 42. Поэтому для больших N и $m = \Omega(N)$ первый результат может вести к более точным оценкам. Если же $m = o(N)$, результат Теоремы 42 может быть более предпочтительным.

Доказательства двух последних теорем основаны на технике *пилинга* и приведены в Разделе 4.3.

Прошлые результаты ограничивали сверху величину $L_N(\hat{h}_m) - L_N(h_N^*)$. Далее мы приводим оценки избыточного риска — оценки величины $\mathcal{E}_u(\hat{h}_m)$. Первая из них основана на Теореме 41:

Следствие 13. *Пусть выполнены предположения Теоремы 41. Тогда для любого $\delta > 0$ с вероятностью не меньше $1 - \delta$ справедливо:*

$$\mathcal{E}_u(\hat{h}_m) \leqslant \frac{N}{u}\left(51\frac{r_m^*}{B} + 17B\frac{N}{m^2}\log\frac{2}{\delta}\right) + \frac{N}{m}\left(51\frac{r_u^*}{B} + 17B\frac{N}{u^2}\log\frac{2}{\delta}\right).$$

Следующая версия основана на Теореме 42 и позволяет заменить множители N/m^2 и N/u^2 на $1/m$ и $1/u$, соответственно:

Следствие 14. *Пусть выполнены предположения Теоремы 42. Тогда для любого $\delta > 0$ с вероятностью не меньше $1 - \delta$ справедливо:*

$$\mathcal{E}_u(\hat{h}_m) \leqslant \frac{N}{u}\left(\frac{901}{B}r_m^* + \frac{\log\frac{2}{\delta}(16 + 25B)}{3m}\right) + \frac{N}{m}\left(\frac{901}{B}r_u^* + \frac{\log\frac{2}{\delta}(16 + 25B)}{3u}\right).$$

Доказательство двух последних результатов основаны на следующем наблюдении: отображение h_u^* так же как и $\hat{h}_m$ минимизирует средний выборочный риск (только посчитанный по контрольной выборке). Поэтому, повторяя доказательства Теоремы 41, мы немедленно получаем те же оценки для h_u^*, с точностью до замены r_m^* и N/m^2 на r_u^* и N/u^2 соответственно. Эти рассуждения показывают, что потери на всей выборке $L_N(\hat{h}_m)$ и $L_N(h_u^*)$ близки друг к другу с большой вероятностью. После получения этого факта нам остается применить несложные вычисления. Подробные доказательства приведены в Разделе 4.3.

Как и в прошлой главе мы изучим поведение величин r_m^* и r_u^*, фигурирующих в Следствии 14, в одном частном случае, когда класс $\mathcal{H}$ является подмножеством RKHS, соответствующего спрямляющему ядру k. Мы будем повторять рассуждения, приведенные для индуктивной постановки в работе [5] и описанные ранее в Разделе 3.2.3.

Применяя Следствие 12 к левой части неравенства (4.3), мы также приходим к выводу, что приводимые далее верхние оценки на *индуктивные* величины r_m^*, r_u^* (то есть фигурирующие в Следствии 14) также справедливы для их трансдуктивных аналогов из Следствия 13. Однако, не стоит забывать, что при этом мы теряем возможность воспользоваться

всеми упомянутыми выше превосходствами использования Теоремы 41/Следствия 13 над Теоремой 42/Следствием 14. В частности, как было отмечено в Замечании 21, схема выборки без возвращений может вести к улучшенным результатам (по крайней мере для случая $m = \Omega(N)$). Вопрос о возможности получения более точных верхних оценок при непосредственном рассмотрении трансдуктивной неподвижной точки (4.3) является открытым и оставлен для будущих исследований.

Следствие 15. *Пусть k — положительная неотрицательно определенная функция ядра, определенная на $\mathcal{X}$, такая что $\sup_{x\in\mathcal{X}} k(x,x) \leqslant 1$. Пусть $\mathcal{C}_k$ — Гильбертово пространство (RKHS), соответствующее спрямляющему ядру k; $\mathcal{H} := \{f \in \mathcal{C}_k : \|f\| \leqslant 1\}$ и $\mathcal{F}^*$ — соответствующий класс избыточных потерь. Пусть выполнены Предположения 1. Предположим также, что функция потерь ℓ является L-Липшицевой по своему первому аргументу. Пусть $(K_N)_{ij} := \frac{1}{N}k(X_i, X_j)$ — нормированная матрица Грамма ядра k, где $\mathbf{X}_N = (X_1, \ldots, X_N)$. Обозначим с помощью $\lambda_{1,N} \geqslant \ldots \geqslant \lambda_{N,N}$ невозрастающую последовательность ее собственных значений. Тогда для $k = u$ и $k = m$ справедливо:*

$$r_k^* \leqslant c_N \min_{0\leqslant q\leqslant k} \left(\frac{q}{k} + \sqrt{\frac{1}{k}\sum_{i\geqslant q} \lambda_{i,N}} \right) ,$$

где c_N — константа, зависящая только от L.

Эта оценка является прямым следствием применения результатов Раздела 6.3 работы [5] и результатов работы [80], сформулированных также ранее в Теореме 38. Единственное важное отличие заключается в том, что в качестве распределения мы берем равномерное распределение на множестве $\mathbf{X}_N$. Повторяя рассуждения, приведенные в упомянутых работах, мы приходим к выводу, что величины r_m^* и r_u^* имею порядки, не превосходящие $1/\sqrt{m}$ и $1/\sqrt{u}$ соответственно. В случае, когда собственные значения убывают достаточно быстро (например, экспоненциально быстро), скорость сходимости может увеличиться (вплоть до быстрых порядков).

Замечание 22. *Мы закончим настоящий раздел следующим замечанием. Вопрос о* скорости сходимости в трансдуктивной постановке *в действительности является неоднозначным. Действительно, все результаты, приведенные выше, предполагали, что множество $\mathbf{X}_N$ зафиксировано. Например, это предположение привело к появлению в Следствии 15 собственных значений матрицы Грамма, посчитанной на множестве $\mathbf{X}_N$. Для того, чтобы в точности определить смысл* скорости сходимости, *нам необходимо описать законы изменения множества $\mathbf{X}_N$ с ростом размера популяции N. Одним из естественных подходов к*

данному вопросу является рассмотрение второй постановки задачи трансдуктивного обучения, приведенной в [104], где объекты $\mathbf{X}_N$ выбираются независимо из одного и того же неизвестного рспределения. Мы считаем, что в этом случае на основе методов, изложенных в [5], может быть установлена строгая связь между величиной $r_m^*(N)$, зависящей от N, и ее асимптотическим аналогом (при $N \to \infty$). Однако, этот вопрос выходит за рамки настоящей работы.

4.3 Доказательства результатов Раздела 4.2

Доказательство Теоремы 40: Применяя Теорему 22, мы получаем, что с вероятностью не меньше $1-\delta$ выполнено:

$$\sup_{h\in\mathcal{H}}\big(L_N(h) - \hat{L}_m(h)\big) \leqslant E_m + \sqrt{2\left(\sigma_{\mathcal{H}}^2 + 2E_m\right)\frac{\log\frac{1}{\delta}}{m}} + \frac{\log\frac{1}{\delta}}{3m}.$$

Мы можем упростить полученное неравенство, пользуясь элементарными неравенствами $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$ и $\sqrt{ab} \leqslant \frac{a+b}{2}$:

$$\begin{aligned}\sup_{h\in\mathcal{H}}\big(L_N(h) - \hat{L}_m(h)\big) &\leqslant E_m + \sqrt{\frac{2\sigma_{\mathcal{H}}^2\log\frac{1}{\delta}}{m}} + 2\sqrt{\frac{E_m\log\frac{1}{\delta}}{m}} + \frac{\log\frac{1}{\delta}}{3m} \\ &\leqslant 2E_m + \sqrt{\frac{2\sigma_{\mathcal{H}}^2\log\frac{1}{\delta}}{m}} + \frac{4\log\frac{1}{\delta}}{3m}.\end{aligned}$$

■

Доказательство Теоремы 41 основано на следующем промежуточном результате, фигурирующем в доказательстве Теоремы 3.3 работы [5]. Мы приводим его в виде леммы:

Лемма 16. *(Пилинг на основе Теоремы 20) Пусть выполнены условия Теоремы 41. Выберем произвольную $\lambda > 1$. Для $w(r,f) = \min\{r\lambda^k \colon k \in \mathbb{N}, r\lambda^k \geqslant Ef^2\}$ определим следующую* взвешенную *версию класса избыточных потерь:*

$$\mathcal{G}_r = \left\{\frac{r}{w(r,f)}f \colon f \in \mathcal{F}^*\right\}.$$

Тогда для всех $r > r_m^*$ и $t > 0$ с вероятностью не меньше $1-\delta$ справедливо следующее:

$$\sup_{g\in\mathcal{G}_r} Eg - \hat{E}_m g \leqslant \sqrt{r}\left(5\frac{\sqrt{r_m^*}}{B} + 2\sqrt{2\frac{N}{m^2}\log\frac{1}{\delta}}\right). \tag{4.5}$$

Доказательство. Мы будем в точности повторять шаги доказательства первой части Теоремы 3.3, представленные в работе [5] (см. страницы 15–16), но вместо неравенства Талаграна будем пользоваться Теоремой 20.

Очевидно, для всех $f \in \mathcal{F}^*$ справедливо:

$$\mathbb{D}[f(X)] = Ef^2 - (Ef)^2 \leqslant Ef^2. \tag{4.6}$$

Выберем произвольные $\lambda > 1$ и $r > 0$ и рассмотрим следующее множество:

$$\mathcal{G}_r = \left\{ \frac{r}{w(r,f)} f \colon f \in \mathcal{F}^* \right\},$$

где $w(r,f) = \min\{r\lambda^k \colon k \in \mathbb{N}, r\lambda^k \geqslant Ef^2\}$. Рассмотрим сначала функции $f \in \mathcal{F}^*$, такие что $Ef^2 \leqslant r$, что означает $w(r,f) = r$. Отображения $g \in \mathcal{G}_r$, соответствующие этим функциям, удовлетворяют $g = f$ и поэтому $\mathbb{D}[g(X)] = \mathbb{D}[f(X)] \leqslant Ef^2 \leqslant r$. В противном случае, если $Ef^2 > r$, тогда $w(r,f) = \lambda^k r$, и отображения $g \in \mathcal{G}_r$, соответствующие этим функциям удовлетворяют $g = f/\lambda^k$ and $Ef^2 \in (r\lambda^{k-1}, r\lambda^k]$. Поэтому мы получаем $\mathbb{D}[g] = \mathbb{D}[f]/\lambda^{2k} \leqslant Ef^2/\lambda^{2k} \leqslant r$. Мы приходим к выводу, что для всех $g \in \mathcal{G}_r$ справедливо$\mathbb{D}[g(x)] \leqslant r$.

Далее мы будем получать верхнюю оценку для величины

$$V_r = \sup_{g \in \mathcal{G}_r} Eg - \hat{E}_m g.$$

Заметим, что для всех $f \in \mathcal{F}^*$ справедливо $f(X) \in [-1,1]$ и, следовательно, все отображения $g \in \mathcal{G}_r$ также удовлетворяют условию $g(X) \in [-1,1]$. Справедливо также следующее:

$$\frac{1}{2}\left(Eg - \hat{E}_m g\right) = \frac{1}{m} \sum_{X \in \boldsymbol{X}_m} \frac{Eg - g(X)}{2}.$$

Также справедливо $(Eg - g(X))/2 \in [-1,1]$ и $\mathbb{E}\left[Eg - g(X)\right] = 0$. Поскольку величина Eg неслучайна, пользуясь (4.6), мы получаем:

$$\mathbb{D}[(Eg - g(X))/2] = \mathbb{D}[g(x)]/4 \leqslant r/4$$

для всех $g \in \mathcal{G}_r$. Мы теперь можем применить Теорему 20 либо Теорему 22 для следующего класса функций: $\{(Eg - g(X))/2, g \in \mathcal{G}_r\}$. Здесь мы приводим доказательство, основанное на Теореме 20. Применив его, мы получаем, что для всех $\delta > 0$ с вероятностью не меньше $1 - \delta$ справедливо:

$$\begin{aligned}\frac{1}{2}\sup_{g \in \mathcal{G}_r} Eg - \hat{E}_m g &\leqslant \frac{1}{2}\mathbb{E}\left[\sup_{g \in \mathcal{G}_r} Eg - \hat{E}_m g\right] + 2\sqrt{2\left(\frac{N}{m^2}\right)\frac{1}{4}\sup_{g \in \mathcal{G}_r}\mathbb{D}[g(X)]\log\frac{1}{\delta}} \\ &\leqslant \frac{1}{2}\mathbb{E}\left[\sup_{g \in \mathcal{G}_r} Eg - \hat{E}_m g\right] + \sqrt{2\left(\frac{N}{m^2}\right) r \log\frac{1}{\delta}},\end{aligned}$$

что эквивалентно:

$$V_r \leqslant \mathbb{E}[V_r] + 2\sqrt{2\left(\frac{N}{m^2}\right) r \log\frac{1}{\delta}}. \tag{4.7}$$

Положим $\mathcal{F}^*(x,y) = \{f \in \mathcal{F}^* \colon x \leqslant Ef^2 \leqslant y\}$. Из условий леммы следует, что для $f \in \mathcal{F}^*$ справедливо $\mathbb{D}[f(X)] \leqslant Ef^2 \leqslant B \cdot Ef \leqslant B$. Обозначим с помощью k наименьшее положительное целое число, такое что $r\lambda^{k+1} \geqslant B$. Заметим, что для любых множеств A и B выполняется следующее:

$$\mathbb{E}\left[\sup_{g\in A\cup B} Eg - \hat{E}_m g\right] \leqslant \mathbb{E}\left[\sup_{g\in B} Eg - \hat{E}_m g\right] + \mathbb{E}\left[\sup_{g\in A} Eg - \hat{E}_m g\right].$$

Действительно, поскольку супремум является выпуклой функцией, мы можем воспользоваться неравенством Йенсена и показать, что оба слагаемых в правой части неравенства неотрицательны. Мы получаем:

$$\begin{aligned}
\mathbb{E}[V_r] &= \mathbb{E}\left[\sup_{g\in\mathcal{G}_r} Eg - \hat{E}_m g\right] \\
&\leqslant \mathbb{E}\left[\sup_{f\in\mathcal{F}^*(0,r)} Ef - \hat{E}_m f\right] + \mathbb{E}\left[\sup_{f\in\mathcal{F}^*(r,B)} \frac{r}{w(r,f)}(Ef - \hat{E}_m f)\right] \\
&\leqslant \mathbb{E}\left[\sup_{f\in\mathcal{F}^*(0,r)} Ef - \hat{E}_m f\right] + \sum_{j=0}^{k} \mathbb{E}\left[\sup_{f\in\mathcal{F}^*(r\lambda^j, r\lambda^{j+1})} \frac{r}{w(r,f)}(Ef - \hat{E}_m f)\right] \\
&\leqslant \mathbb{E}\left[\sup_{f\in\mathcal{F}^*(0,r)} Ef - \hat{E}_m f\right] + \sum_{j=0}^{k} \lambda^{-j}\mathbb{E}\left[\sup_{f\in\mathcal{F}^*(r\lambda^j, r\lambda^{j+1})} (Ef - \hat{E}_m f)\right] \\
&\leqslant \frac{\psi_m(r)}{B} + \frac{1}{B}\sum_{j=0}^{k} \lambda^{-j}\psi(r\lambda^{j+1})
\end{aligned}$$

где в последнем неравенстве мы воспользовались условием леммы. Теперь, поскольку отображение ψ_m является субкоренным, для всех $\beta \geqslant 1$ справедливо $\psi_m(\beta r) \leqslant \sqrt{\beta}\psi_m(r)$. Поэтому:

$$\mathbb{E}[V_r] \leqslant \frac{\psi_m(r)}{B}\left(1 + \sqrt{\lambda}\sum_{j=0}^{k} \lambda^{-j/2}\right).$$

Выбрав $\lambda = 4$, мы ограничиваем правую часть сверху с помощью $5\psi_m(r)/B$. Наконец, заметим, что для всех $r \geqslant r_m^*$ выполнено $\psi_m(r) \leqslant \sqrt{r/r_m^*}\psi_m(r_m^*) = \sqrt{rr_m^*}$ и поэтому:

$$\mathbb{E}[V_r] \leqslant \frac{5}{B}\sqrt{rr_m^*}.$$

Подставив эту верхнюю оценку в (4.7), мы завершаем доказательство. ■

Доказательство Теоремы 41: Пользуясь Леммой 16, мы получаем, что для любых $r > r_m^*$, $\delta > 0$, и $\lambda > 1$, с вероятностью не меньше $1 - \delta$ справедливо:

$$\sup_{g\in\mathcal{G}_r} Eg - \hat{E}_m g \leqslant \sqrt{r}\left(5\frac{\sqrt{r_m^*}}{B} + 2\sqrt{2\frac{N}{m^2}\log\frac{1}{\delta}}\right), \tag{4.8}$$

где мы ввели в рассмотрение взвешенный класс избыточных потерь:

$$\mathcal{G}_r = \left\{\frac{r}{w(r,f)}f \colon f \in \mathcal{F}^*\right\},$$

а $w(r,f) = \min\{r\lambda^k \colon k \in \mathbb{N}, r\lambda^k \geqslant Ef^2\}$. Теперь мы хотим выбрать $r_0 > r_m^*$ таким образом, чтобы верхняя оценка (4.8) приняла вид $r_0/(\lambda BK)$. Для этого достаточно положить:

$$r_0 = K^2\lambda^2\left(5\sqrt{r_m^*} + 2B\sqrt{2\frac{N}{m^2}\log\frac{1}{\delta}}\right)^2 > r_m^*.$$

Подставив $r = r_0$ в (4.8), мы получаем:

$$\sup_{g\in\mathcal{G}_{r_0}} Eg - \hat{E}_m g \leqslant \frac{r_0}{\lambda BK}. \tag{4.9}$$

Более того, пользуясь неравенством $(u+v)^2 \leqslant 2(u^2+v^2)$, мы получаем:

$$r_0 \leqslant 50K^2\lambda^2 r_m^* + 16K^2\lambda^2B^2t\left(\frac{N}{m^2}\right). \tag{4.10}$$

Вспомним, что для всех $r > 0$ и всех $g \in \mathcal{G}_r$ следующее выполнено с вероятностью 1:

$$Eg - \hat{E}_m g \leqslant \sup_{g\in\mathcal{G}_r} Eg - \hat{E}_m g.$$

Пользуясь определением $\mathcal{G}_r$, мы получаем, что для всех $f \in \mathcal{F}^*$ следующее выполняется с вероятностью 1:

$$E\left(\frac{r}{w(r,f)}f\right) - \hat{E}_m\left(\frac{r}{w(r,f)}f\right) \leqslant \sup_{g\in\mathcal{G}_r} Eg - \hat{E}_m g,$$

что эквивалентно:

$$Ef - \hat{E}_m f \leqslant \frac{w(r,f)}{r}\sup_{g\in\mathcal{G}_r} Eg - \hat{E}_m g.$$

Положив $r = r_0$ и воспользовавшись (4.9), мы получаем, что с вероятностью не меньше $1-\delta$

$$\forall f \in \mathcal{F}^*, \forall K > 1: \quad Ef - \hat{E}_m f \leqslant \frac{w(r_0,f)}{r_0}\frac{r_0}{\lambda KB} = \frac{w(r_0,f)}{\lambda KB}.$$

Теперь мы воспользуемся Предположением 1.2. Если для $f \in \mathcal{F}^*$ справедливо $Ef^2 \leqslant r_0$, тогда $w(r_0,f) = r_0$ и, пользуясь (4.10), мы получаем:

$$Ef - \hat{E}_m f \leqslant \frac{w(r_0,f)}{\lambda KB} = \frac{r_0}{\lambda KB} \leqslant 50\frac{K}{B}\lambda r_m^* + 16\lambda KBt\left(\frac{N}{m^2}\right),$$

что эквивалентно:

$$Ef \leqslant \hat{E}_m f + 50\frac{K}{B}\lambda r_m^* + 16\lambda KBt\left(\frac{N}{m^2}\right). \tag{4.11}$$

Если же $Ef^2 > r_0$, тогда $w(r_0,f) = \lambda^i r_0$ для некоторого целого числа $i > 0$, а также $Ef^2 \in (r_0\lambda^{i-1}, r_0\lambda^i]$. Тогд мы получаем:

$$Ef - \hat{E}_m f \leqslant \frac{w(r_0,f)}{\lambda KB} = \frac{\lambda^i r_0}{\lambda KB} = \frac{\lambda\cdot(\lambda^{i-1}r_0)}{\lambda KB} \leqslant \frac{Ef^2}{KB} \leqslant \frac{Ef}{K}.$$

Поэтому

$$Ef \leqslant \frac{K}{K-1}\hat{E}_m f. \tag{4.12}$$

Совмещая (4.11) и (4.12), мы наконец полчаем, что с вероятностью не меньше $1-\delta$ справедливо следующее:

$$\forall f \in \mathcal{F}^*, \forall K > 1: \quad Ef \leqslant \inf_{K>1} \frac{K}{K-1}\hat{E}_m f + 50\frac{K}{B}\lambda r_m^* + 16\lambda K B\left(\frac{N}{m^2}\right)\log\frac{1}{\delta}. \tag{4.13}$$

Нам остается вспомнить определение $\mathcal{F}^*$ и положить $\hat{f}_m = \ell_{\hat{h}_m} - \ell_{h_N^*}$. Заметив, что

$$\begin{aligned}\hat{E}_m \hat{f}_m &= \hat{E}_m \ell_{\hat{h}_m} - \hat{E}_m \ell_{h_N^*} \\ &= \hat{L}_m(\hat{h}_m) - \hat{L}_m(h_N^*) \leqslant 0\end{aligned}$$

и

$$E\hat{f}_m = L_N(\hat{h}_m) - L_N(h_N^*),$$

мы завершаем доказательство. ■

Пусть случайные величины $\{\xi_1, \ldots, \xi_m\}$ выбраны равномерно *с возвращениями* из множества $\boldsymbol{X}_N$. Обозначим

$$E_{r,m} = \mathbb{E}\left[\sup_{f\in\mathcal{F}^*: Ef^2\leqslant r}\left(Ef - \frac{1}{m}\sum_{i=1}^m f(\xi_i)\right)\right]. \tag{4.14}$$

Повторяя доказательство Леммы пилинга 16 и пользуясь Теоремой 22 вместо Теоремы 20, мы немедленно получаем следующий результат:

Лемма 17. *(Пилинг на основе Теоремы 22) Пусть $\mathcal{H}$ и ℓ удовлетворяют Предположениям 1. Пусть для субкоренной функции $\psi_m(r)$ справедливо:*

$$B \cdot E_{r,m} \leqslant \psi_m(r),$$

где величина $E_{r,m}$ определена в (4.14). Обозначим с помощью r_m^* неподвижную точку отображения $\psi_m(r)$.

Выбирем произвольную $\lambda > 1$. Тогда для любых $r > r_m^*$ и $\delta > 0$ следующее выполняется с вероятностью не меньше $1-\delta$:

$$\sup_{g\in\mathcal{G}_r} Eg - \hat{E}_m g \leqslant \sqrt{r}\left(15\frac{\sqrt{r_m^*}}{B} + \sqrt{\frac{2\log\frac{1}{\delta}}{m}}\right) + \frac{8\log\frac{1}{\delta}}{3m}, \tag{4.15}$$

где для $w(r,f) = \min\{r\lambda^k : k \in \mathbb{N}, r\lambda^k \geqslant Ef^2\}$ мы определили следующий взвешенный класс избыточных потерь:

$$\mathcal{G}_r = \left\{\frac{r}{w(r,f)} f : f \in \mathcal{F}^*\right\}.$$

Доказательство Теоремы 42 основано на Лемме пилинга 17 и повторяет шаги доказательства Теоремы 41.

Доказательство Следствия 13: Заметим, что поскольку h_u^* также (как $\hat{h}_m$) минимизирует средний выборочный риск (только посчитанный по контрольной выборке), результаты Теорем 41 и 42 также справедливы для h_u^*, с точностью до замены всех m на u. Также заметим, что следующее справедливо с вероятностью 1:

$$\begin{aligned} 0 &\leqslant L_N(\hat{h}_m) - L_N(h_N^*) \\ &= L_N(\hat{h}_m) - L_N(h_N^*) - \hat{L}_m(\hat{h}_m) + \hat{L}_m(h_N^*) + \hat{L}_m(\hat{h}_m) - \hat{L}_m(h_N^*) \\ &\leqslant L_N(\hat{h}_m) - L_N(h_N^*) - \hat{L}_m(\hat{h}_m) + \hat{L}_m(h_N^*) \\ &= \frac{u}{N}\left(L_u(\hat{h}_m) - L_u(h_N^*) - \hat{L}_m(\hat{h}_m) + \hat{L}_m(h_N^*)\right) \end{aligned} \tag{4.16}$$

и

$$\begin{aligned} 0 &\leqslant L_N(h_u^*) - L_N(h_N^*) \\ &= L_N(h_u^*) - L_N(h_N^*) - L_u(h_u^*) + L_u(h_N^*) + L_u(h_u^*) - L_u(h_N^*) \\ &\leqslant L_N(h_u^*) - L_N(h_N^*) - L_u(h_u^*) + L_u(h_N^*) \\ &= \frac{m}{N}\left(\hat{L}_m(h_u^*) - \hat{L}_m(h_N^*) - L_u(h_u^*) + L_u(h_N^*)\right), \end{aligned}$$

где последние неравенства в обоих случаях используют следующий факт:

$$N \cdot L_N(h) = m \cdot \hat{L}_m(h) + u \cdot L_u(h).$$

Мы собираемся воспользоваться неравенством (4.13), полученным в доказательстве Теоремы 41. Пользуясь последним тождеством в (4.16) и применяя затем (4.13) для $f = \ell_{\hat{h}_m} - \ell_{h_N^*}$, где мы вычли $\hat{E}_m f$ из обеих частей (4.13), мы получаем:

$$\begin{aligned} 0 &\leqslant L_u(\hat{h}_m) - L_u(h_N^*) - \hat{L}_m(\hat{h}_m) + \hat{L}_m(h_N^*) \\ &\leqslant \frac{N}{u}\left(\inf_{K>1}\frac{1}{K-1}\underbrace{\hat{L}_m(\hat{h}_m - h_N^*)}_{\leqslant 0} + 50\frac{K}{B}\lambda r_m^* + 16\lambda K B\frac{N}{m^2}\log\frac{1}{\delta}\right), \end{aligned}$$

что справедливо с вероятностью не меньше $1-\delta$. Как было замечено выше, та же цепочка рассуждений применима для h_u^*, что ведет к следующему результату:

$$\begin{aligned} 0 &\leqslant \hat{L}_m(h_u^*) - \hat{L}_m(h_N^*) - L_u(h_u^*) + L_u(h_N^*) \\ &\leqslant \frac{N}{m}\left(\inf_{K>1}\frac{1}{K-1}\underbrace{L_u(h_u^* - h_N^*)}_{\leqslant 0} + 50\frac{K}{B}\lambda r_u^* + 16\lambda K B\frac{N}{u^2}\log\frac{1}{\delta}\right), \end{aligned}$$

также справедливому с вероятностью не меньше $1-\delta$. Воспользовавшись неравенством Буля и заменив δ на $\delta/2$ мы получаем, что с вероятностью не меньше $1-\delta$ следующие результаты справедливы одновременно:

$$0 \leqslant L_u(\hat{h}_m) - L_u(h_N^*) - \hat{L}_m(\hat{h}_m) + \hat{L}_m(h_N^*) \leqslant \frac{N}{u}\left(50K\lambda\frac{r_m^*}{B} + 16\lambda KB\frac{N}{m^2}\log\frac{2}{\delta}\right)$$

и

$$0 \leqslant \hat{L}_m(h_u^*) - \hat{L}_m(h_N^*) - L_u(h_u^*) + L_u(h_N^*) \leqslant \frac{N}{m}\left(50K\lambda\frac{r_u^*}{B} + 16\lambda KB\frac{N}{u^2}\log\frac{2}{\delta}\right).$$

Складывая два неравенства, мы получаем:

$$\begin{aligned}0 &\leqslant L_u(\hat{h}_m) - L_u(h_u^*) - \hat{L}_m(\hat{h}_m) + \hat{L}_m(h_u^*)\\ &\leqslant \frac{N}{u}\left(50\lambda K\frac{r_m^*}{B} + 16\lambda KB\frac{N}{m^2}\log\frac{2}{\delta}\right) + \frac{N}{m}\left(50\lambda K\frac{r_u^*}{B} + 16\lambda KB\frac{N}{u^2}\log\frac{2}{\delta}\right).\end{aligned}$$

Воспользовавшись тем фактом, что $\hat{h}_m$ и h_u^* оба минимизируют средний выборочный риск (на обучающей и контрольной выборках соответственно), мы получаем:

$$\begin{aligned}0 &\leqslant L_u(\hat{h}_m) - L_u(h_u^*)\\ &\leqslant \hat{L}_m(\hat{h}_m) - \hat{L}_m(h_u^*) + \frac{N}{u}\left(50\lambda K\frac{r_m^*}{B} + 16\lambda KBt\frac{N}{m^2}\right) + \frac{N}{m}\left(50\lambda K\frac{r_u^*}{B} + 16\lambda KBt\frac{N}{u^2}\right)\\ &\leqslant \frac{N}{u}\left(50\lambda K\frac{r_m^*}{B} + 16\lambda KBt\frac{N}{m^2}\right) + \frac{N}{m}\left(50\lambda K\frac{r_u^*}{B} + 16\lambda KBt\frac{N}{u^2}\right).\end{aligned}$$

■

Доказательство Следствия 14 повторяет шаги последнего доказательства и использует Теорему 42 вместо Теоремы 41.

Мы также приводим следующий промежуточный результат:

Следствие 16. *В условиях Теоремы 41, для всех $\delta > 0$ и $K > 1$, с вероятностью не меньше $1-\delta$ справедливо:*

$$\begin{aligned}|L_N(\hat{h}_m) - L_N(h_u^*)| &\leqslant \max\left(2K\frac{r_m^*}{B} + 16KB\left(\frac{N}{m^2}\right)\log\frac{2}{\delta}, 2K\frac{r_u^*}{B} + 16KB\left(\frac{N}{u^2}\right)\log\frac{2}{\delta}\right)\\ &\leqslant 2K\frac{r_m^* + r_u^*}{B} + 16KBN\left(\frac{1}{m^2} + \frac{1}{u^2}\right)\log\frac{2}{\delta}.\end{aligned}$$

Доказательство. Заметим, что $L_N(h_u^*) - L_N(h_N^*) \geqslant 0$, а также $L_N(\hat{h}_m) - L_N(h_N^*) \geqslant 0$. Нам остается совместить Теорему 41 с ее аналогом для отображения h_u^* с помощью неравенства Буля. ■

Коротко подведем итоги настоящей главы:

- Трансдуктивное обучение, в отличие от индуктивного, пытается избежать решение более сложной задачи на пути к решению поставленной цели. Вместо получения ответов на точках контрольной выборки с помощью отображения, найденного в процессе обучения, трансдуктивное обучение занимается поиском ответов *только на заданных контрольных точках*.

- В отличие от индуктивной постановки, где объекты обучающей выборки независимы и одинаково распределены, объекты обучающей выборки в трансдуктивном обучении выбраны равномерно без возвращений из конечного генерального множества.

- Трансдуктивное обучение улучшает качество процедуры обучения, учитывая в ней признаковые описания объектов контрольной выборки.

- В трансдуктивном обучении известны оценки с медленной скоростью сходимости. Они основаны на неравенствах Стирлинга и Эль-Янива-Печиони. Быстрая скорость сходимости достигалась в известных результатах только при наложении чрезвычайно ограничивающих предположений на класс рассматриваемых задач.

В настоящей главе получены следующие новые результаты:

- Получены первые оценки обобщающей способности, учитывающие дисперсию потерь и справедливые для произвольных классов отображений и ограниченных функций потерь (Теоремы 39 и 40).

- Показано (Следствие 12), что математическое ожидание супремума эмпирических процессов в трансдуктивном обучении (для выборок *без возвращений*) может быть оценено сверху его индуктивным аналогом (для выборок *с возвращениями*). Этот результат позволяет применять для анализа сложности класса отображений в трансдуктивной постановке, так же как и в индуктивной, все известные результаты из теории эмпирических процессов, включая неравенства симметризации и сжатия.

- Получены новые оценки избыточного риска в трансдуктивном обучении, основанные на локальных мерах сложности семейств отображений и справедливые при общих предположениях на класс рассматриваемых задач (Теоремы 41 и 42 и Следствия 13 и 14). Эти результаты демонстрируют, что быстрые скорости сходимости в трансдуктивном обучении, как и в индуктивном обучении, могут быть достигнуты в очень общих предположениях.

5 Комбинаторная теория переобучения

В прошлых разделах мы видели, что дополнительные предположения, накладываемые на свойства рассматриваемой задачи, могут вести к получению существенно более точных оценок по сравнению с рассмотрением наиболее общего случая. Так, в Разделе 3.2.3 предположение об отсутствии шума в задаче и принадлежности Байесовского отображения рассматриваемому классу $\mathcal{H}$ позволило нам получить оценки порядка n^{-1}. Конечно, эти предположения являются чересчур строгими и на практике никогда не будут выполнены. Однако, позже мы изучили примеры более мягких и реалистичных предположений на уровень шума в задаче — условия малого шума Массара и Маммена-Цыбакова, которых оказалось достаточно для получения оценок быстрых скоростей сходимости $n^{-1/(2-\alpha)}$ для $\alpha \in (0, 1]$.

В то же время, оптимальные оценки в случае отказа от наложения каких-либо дополнительных (помимо принадлежности значений функции потерь отрезку $[0, 1]$) ограничений на рассматриваемые задачи имеют медленный порядок $n^{-1/2}$. Доказательство этой нижней оценки является конструктивным и заключается в построении примера задачи классификации с бинарной функцией потерь (то есть конкретного класса $\mathcal{H}$ и распределения P на $\mathcal{X} \times \mathcal{Y}$), для которой обобщающую способность метода минимизации эмпирического риска удается ограничить снизу величиной порядка $n^{-1/2}$. Оказывается, в качестве такого примера в доказательстве используется весьма специфический с прикладной точки зрения класс задач, для которых величина $P(Y = 1|X)$ очень близка к «критическому» значению $1/2$ с большой вероятностью. Это условие фактически означает, что ответы являются случайными. Целесообразность решения таких задач на практике, вообще говоря, является спорным вопросом.

Эти размышления показывают, что попытки изучения чересчур общих постановок и применения полученных таким образом оценок в прикладных задачах может вести к сильно завышенным результатам. И причина завышенности будет лежать не только в использовании ряда грубых шагов в доказательствах. Действительно, в приведенном примере наибо-

лее общей постановки порядок оценки Вапника–Червоненкиса, существенно опирающейся на весьма неточные неравенства Хефдинга и Буля, с точностью до логарифма совпадает с порядком нижней оценки, приведенной в [30]. Что важнее, к завышенным на практике результатам будет вести отказ от учета важных отличительных особенностей прикладных задач.

История развития теории статистического обучения (и не только) показывает, что учет новых дополнительных ограничений, накладываемых на рассматриваемые задачи, как правило, основан на использовании все более и более сложных и глубоких теоретических результатов. Так, в полной мере учесть ограничения на шум в рассматриваемых задачах удалось лишь с помощью неравенства Талаграна, основанного на, вероятно, самом сложном подходе в теории концентрации меры. Эти *локальные* результаты (приведенные выше в Теоремах 35 и 36 для индуктивной и в Теоремах 41 и 42 для трансдуктивной постановок) уже позволяют во многих случаях успешно бороться с завышенностью оценок. Однако, они по-прежнему имеют ряд недостатков, включая сложность вычисления значений оценок в конкретных прикладных задачах. Стоит ожидать, что учет еще более «тонких» свойств задач, которые могут оказаться ключевыми для окончательного преодоления проблемы завышенности оценок, будет требовать либо новых и еще более глубоких фундаментальных математических результатов, либо кардинально нового подхода к задаче.

Начиная с появления теории Вапника–Червоненкиса в теории статистического обучения было развито множество новых подходов. Даже короткий обзор всех существующих подходов потребовал бы отдельной обзорной главы. В настоящей работе мы рассматриваем лишь их подмножество. В Главах 3 и 4 была подробно рассмотрена теория Вапника–Червоненкиса и, в частности, *локальные подходы*. В Главе 6 будет рассмотрен PAC-Байесовский подход, объединяющий в себе преимущества теории Вапника-Червоненкиса и Байесовского обучения. Все эти подходы *существенно* опираются на использование неравенств концентрации.

В настоящем разделе мы рассмотрим комбинаторную теорию переобучения [109–111, 116], во многих отношениях тесно связанную с рассмотренной в прошлой главе трансдуктивной постановкой теории статистического обучения. Главное методологическое отличие комбинаторной теории переобучения к анализу процедуры обучения от всех перечисленных выше заключается в *принципиальном отказе* от рассмотрения чересчур общих постановок задач в пользу детального учета свойств конкретной решаемой задачи. В частности, это влечет за собой отказ от использования завышенных неравенств концентрации вероятностной меры (справедливых, как правило, для достаточно широких классов распределений и, значит, часто завышенных для конкретных задач).

Мы начнем с описания постановки основной задачи комбинаторной теории переобучения. Поскольку эта постановка похожа на постановку трансдуктивного обучения, мы продолжим пользоваться в настоящем разделе всеми обозначениями, введенными в прошлой главе. Затем мы приведем краткое сравнение постановок комбинаторного и трансдуктивного подходов. Наконец, мы приведем ряд новых результатов, полученных на основе теоретико-группового подхода, развитого в работах [37, 119, 122–124]. Подробный обзор результатов комбинаторной теории переобучения может быть найден в работе [115].

5.1 Обозначения и постановка задачи

Как было отмечено ранее, в настоящем разделе мы продолжим пользоваться обозначениями, введенными в прошлой главе.

В комбинаторной теории переобучения рассматриваются задачи с *бинарной функцией потерь*. Поскольку на практике количество наблюдаемой исследователем информации ограничено (например, временем его жизни или скоростью вычислений компьютера), в комбинаторной теории переобучения *не рассматриваются* бесконечные пространства $\mathcal{X} \times \mathcal{Y}$. Ограничение размера пространства $\mathcal{X} \times \mathcal{Y}$ также позволяет рассматривать лишь конечные дискретные распределения и избежать применение сложных результатов из теории мер. Генеральную выборку, состоящую из N пар «объект-ответ», мы будем обозначать $\mathbb{S} = \{(X_i, Y_i)\}_{i=1}^N$. При этом множество $\mathbb{S}$ пожет содержать как повторяющиеся несколько раз пары, так и несколько пар с одинаковыми объектами $X \in \mathcal{X}$ и разными ответами $Y \in \mathcal{Y}$.

Конечное множество *различных* векторов ошибок отображений из $\mathcal{H}$ на генеральной выборке $\mathbb{S}$ мы будем обозначать $\mathbb{A}$:

$$\mathbb{A} = \left\{ \left(\mathbb{1}_{\{h(X_1) \neq Y_1\}}, \ldots, \mathbb{1}_{\{h(X_1) \neq Y_1\}} \right) : h \in \mathcal{H} \right\} \subseteq \{0,1\}^N.$$

Очевидно, $|\mathbb{A}| \leqslant 2^N$. Однако, мощность может оказаться гораздо меньше. Таким образом, множество отображений $\mathcal{H}$ разбивается на классы эквивалентности. Два отображения считаются эквивалентными, если их векторы ошибок на генеральной выборке $\mathbb{S}$ совпадают. Все результаты, рассматриваемые далее, могут формулироваться эквивалентно как в терминах классов $\mathcal{H}$ и функций потерь ℓ, так и непосредственно в терминах бинарных векторов ошибок $\mathbb{A}$. Поскольку все результаты комбинаторной теории переобучения основаны на детальном анализе структуры конечного множества бинарных векторов $\mathbb{A}$, гораздо удобнее будет пользоваться бинарными векторами ошибок, «забыв» об их связи с реальной структурой класса $\mathcal{H}$ и функцией потерь ℓ. Единственное, что нам потребуется для этого, — это следующее опре-

деление *индикатора ошибки* вектора $a \in \mathbb{A}$ на паре $(X, Y) \in \mathbb{S}$: $I_a(X, Y) = \mathbb{1}_{\{h_a(X) \neq Y\}}$, где h_a — произвольное отображение из $\mathcal{H}$, вектор ошибок на генеральной выборке которого равен a.

Из генеральной выборки $\mathbb{S}$ *равномерно без возвращений* выбирается $n \leqslant N$ пар «объект-ответ» $S_n \subseteq \mathbb{S}$, которые образуют обучающую выборку, и становятся доступны алгоритму обучения. Оставшиеся $u = N - n$ пар S_u остаются неизвестными. Обратим внимание, что описанная схема допускает рассмотрение задач с шумом, когда объект $X \in \mathcal{X}$ может иметь несколько разных ответов $Y \in \mathcal{Y}$ с ненулевыми вероятностями. Для этого достаточно, чтобы в генеральной выборке $\mathbb{S}$ содержалось несколько пар «объект-ответ» с одинаковыми объектами и разными ответами.

Задача процедуры обучения заключается в поиске на основе обучающей выборки S_n вектора ошибок $a \in \mathbb{A}$, имеющего минимальную частоту ошибок на контрольной выборке:

$$L(a, S_u) \to \min_{a \in \mathbb{A}},$$

где для произвольной подвыборки $S \subseteq \mathbb{S}$ мы положили по определению:

$$L(a, S) := \frac{1}{|S|} \sum_{(X,Y) \in S} I_a(X, Y).$$

Также для простоты число ошибок вектора a на подвыборке $S \in \mathbb{S}$ будем обозначать

$$n(a, S) = |S| \cdot L(a, S).$$

Обозначим с помощью $[\mathbb{S}]^n$ всевозможные подвыборки $\mathbb{S}$, состоящие в точности из n пар «объект-ответ». В комбинаторной теории переобучения, вообще говоря, рассматриваются различные *методы обучения* — отображения $\mu\colon [\mathbb{S}]^n \to \mathbb{A}$, ставящие в соответствие обучающим выборкам векторы ошибок из $\mathbb{A}$. Однако, в настоящем разделе мы ограничимся рассмотрением знакомого нам МЭР, ставящего в соответствие выборке S_n вектор $a \in \mathbb{A}$ с минимальным эмпирическим риском $L(a, S_n)$. Для выборки S обозначим с помощью $A(S)$ множество векторов в $\mathbb{A}$, на которых достигается минимум среднего риска на этой выборке:

$$A(S) = \operatorname{Arg}\min_{a \in \mathbb{A}} L(a, S).$$

Обратим внимание, что в предыдущих главах мы не конкретизировали, как именно МЭР решает неопределенность в случае $|A(S_n)| > 1$. В работах [122,123] был введен *рандомизированный метод МЭР* (РМЭР), выбирающий в этих случаях случайный вектор из равномерного распределения на $A(S_n)$. Всюду далее мы ограничимся рассмотрением РМЭР.

Основная задача комбинаторной теории переобучения заключается в получении *точных* (не завышенных) оценок *вероятности переобучения*:

$$Q_{\mu,\varepsilon}(\mathbb{A}) = \mathbb{P}\left\{L\big(\mu(S_n), S_u\big) - L\big(\mu(S_n), S_n\big) \geqslant \varepsilon\right\}.$$

Ранее в Главе 2 мы убедились, что выбор n элементов равномерно без возвращений из конечной генеральной выборки мощностью $N \geqslant n$ эквивалентен равномерному выбору *разбиения* генеральной выборки на две не пересекающиеся между собой подвыборки из множества всевозможных таких разбиений. Также вспомним, что вероятность события равна математическому ожиданию индикатора этого события. Поэтому в случае детерминированного метода обучения μ вероятность переобучения может быть записана в следующем виде:

$$Q_{\mu,\varepsilon}(\mathbb{A}) = \frac{1}{C_N^n} \sum_{S \in [\mathbb{S}]^n} \mathbb{1}_{\{L(\mu(S),\mathbb{S}\setminus S) - L(\mu(S),S) \geqslant \varepsilon\}}.$$

В случае же использования рандомизированного МЭР нам необходимо дополнительно взять математическое ожидание по случайному выбору вектора из $A(S_n)$, что дает нам следующую формулу [122]:

$$Q_{\mu,\varepsilon}(\mathbb{A}) = \frac{1}{C_N^n} \sum_{S \in [\mathbb{S}]^n} \frac{1}{|A(S)|} \sum_{a \in A(S)} \mathbb{1}_{\{L(a,\mathbb{S}\setminus S) - L(a,S) \geqslant \varepsilon\}}. \tag{5.1}$$

Сходства и различия с трансдуктивной постановкой. Вкратце перечислим основные различия трансдуктивной и комбинаторной постановок.

1. Главным сходством двух постановок является способ получения обучающей выборки: в обоих случаях она выбирается равномерно без возвращений из некоторого конечного генерального множества.

2. В прошлой главе мы рассмотрели частный *детерминированный* случай трансдуктивной постановки, когда ответы на объектах получаются некоторой неизвестной но детерминированной функцией $\varphi\colon \mathcal{X} \to \mathcal{Y}$. Поскольку мы также предположили, что генеральное множество объектов $\boldsymbol{X}_N$ конечно, то отсюда следует, что множество всевозможных пар «объект-ответ», которые мы могли бы наблюдать, конечно — также как и в комбинаторной теории переобучения.

3. Однако, если мы рассмотрим более общую постановку трансдуктивного обучения *с шумом*, то ответ Y на объекте $X \in \mathcal{X}$ будет принимать различные значения из $\mathcal{Y}$ с ненулевыми вероятностями. Поскольку никаких ограничений на множество $\mathcal{Y}$ мы не накладывали, в этом случае мы можем иметь дело с бесконечным множеством возможных пар «объект-ответ» (хоть множество различных объектов и конечно).

4. Предположения о конечности множества объектов (или пар «объект-ответ»), так или иначе представленные в каждой из двух постановок, имеют различную природу. В комбинаторной теории переобучения, как было отмечено выше, конечность генеральной выборки является следствием ограниченности времени вычислений и самого исследования прикладной задачи. В трансдуктивной постановке она обусловлена отказом от построения индуктивного правила — то есть отображения, способного давать ответы на всевозможных будущих объектах. Поскольку целью является предсказание ответов на конкретной и *конечной* контрольной выборке, нас не интересуют другие точки пространства объектов $\mathcal{X}$.

5. В отличие от трансдуктивной постановки, процедура обучения в комбинаторной теории переобучения не получает доступ к объектам контрольной выборки S_u. Это также связано с прошлым замечанием. В этом смысле комбинаторная теория ближе по духу к индуктивной постановке, рассмотренной в Главе 3.

6. Наконец, в отличие от трансдуктивной постановки, методы, используемые в комбинаторной теории переобучения, накладывают существенное ограничение на множество рассматриваемых функций потерь ℓ, не давая возможность работать лишь с бинарными функциями потерь.

Учитывая все выше сказанное, мы можем заключить, что новые результаты для детерминированной постановки трансдуктивного обучения, представленные в Разделе 4.2 прошлой главы, без изменений переносятся на постановку комбинаторной теории переобучения. Сравнение локальных результатов Теорем 41 и 42 с известными результатами комбинатоной теории выходят за рамки настоящей работы и являются интересным направлением дальнейших исследований.

5.2 Теоретико-групповой подход

Теоретико-групповой подход был впервые развит в работах [122, 123] и позволяет существенно облегчить вычисление точных значений вероятности переобучения множества векторов ошибок $\mathbb{A}$ в тех случаях, когда оно наделено определенными симметриями.

Сначла мы введем определения и обозначения, используемые в теоретико-групповом подходе, и сформулируем его основные результаты в Разделе 5.2.1. Затем в Разделе 5.2.2 мы приведем новые результаты, полученные в работах [119, 120] и развитые позже в [37, 124], которые позволяют в ряде случаев существенно облегчить вычисления теоретико-группового

подхода. В частности, они позволяют получить новые точные оценки вероятности переобучения для трех нетривиальных модельных семейств $\mathbb{A}$, являющихся подмножествами шара в Булевом кубе, которые мы рассмотрим в Разделе 5.2.4.

5.2.1 Обзор известных результатов

Определим *вклад вектора ошибок* $a \in \mathbb{A}$ в вероятность переобучения следующим образом:

$$Q_{\mu,\varepsilon}(\mathbb{A}, a) = \frac{1}{C_N^n} \sum_{\substack{S \in [\mathbb{S}]^n: \\ a \in A(S)}} \frac{\mathbb{1}_{\{L(a,\mathbb{S}\setminus S) - L(a,S) \geqslant \varepsilon\}}}{|A(S)|}.$$

Тогда, переставив в (5.1) местами знаки суммирования, мы получаем разложение вероятности переобучения по индивидуальным вкладам векторов ошибок $a \in \mathbb{A}$:

$$Q_{\mu,\varepsilon}(\mathbb{A}) = \sum_{a \in \mathbb{A}} Q_{\mu,\varepsilon}(\mathbb{A}, a). \tag{5.2}$$

В общем случае в последней сумме содержится $|\mathbb{A}|$ *разных* слагаемых. Однако, оказывается, если множество векторов $\mathbb{A}$ обладает некими симметриями, то многие слагаемые начинают совпадать между собой. Следовательно, с приведенной выше формулой разложения становится легче работать аналитически. На этой идее основан *теоретико-групповой подход* в комбинаторной теории переобучения, впервые предложенный в работах [122, 123]. Далее мы приведем строгое описание данного метода.

Группа симметрий множества $\mathbb{A}$. Напомним, что генеральная выборка $\mathbb{S} = \{(X_i, Y_i)\}_{i=1}^N$ представляет собой конечное множество занумерованных пар из $\mathcal{X} \times \mathcal{Y}$. Рассмотрим симметрическую группу перестановок Π_N. Элементы группы Π_N очевидным образом действует на генеральной выборке, переставляя местами занумерованные пары, составляющие ее. Для пары $(X, Y) \in \mathbb{S}$ и $\pi \in \Pi_N$ с помощью $\pi(X, Y)$ обозначим ту пару «объект-ответ», в которую пара (X, Y) переходит под действием перестановки π. На основе этого можно определить действие элементов $\pi \in \Pi_N$ на множестве подвыборок $S \subseteq \mathbb{S}$:

$$\pi(S) = \{\pi(X, Y) \colon (X, Y) \in S\},$$

на множестве бинарных векторов ошибок $a \in \mathbb{A}$:

$$\pi(a) = \big(I_{\pi(a)}(X_i, Y_i))\big)_{i=1}^N = \Big(I_a\big(\pi^{-1}(X_i, Y_i)\big)\Big)_{i=1}^N$$

и на всевозможных подмножествах $A \subseteq \mathbb{A}$:

$$\pi(A) = \{\pi(a) \colon a \in \mathbb{A}\}.$$

Все эти отображения биективны. Очевидно, действие элементов Π_n на подвыборку $S \in \mathbb{S}$ не меняет числа объектов в ней $|S| = |\pi(S)|$. При введенных определениях справедлив следующий результат, полученный в [123]:

Лемма 18 ([123]). *Для любых $\pi \in \Pi_N, a \in \mathbb{A}$ и $S \in \mathbb{S}$ справедливо:*

$$L(a, S) = L\big(\pi(a), \pi(S)\big).$$

Группой симметрий $S_{\mathbb{A}}$ множества векторов ошибок $\mathbb{A} \subseteq \{0,1\}^N$ будем называть его стационарную подгруппу:

$$S_{\mathbb{A}} = \{\pi \in \Pi_N : \pi(\mathbb{A}) = \mathbb{A}\}.$$

Орбитой действия группы симметрий $S_{\mathbb{A}}$ на вектор ошибок $a \in \mathbb{A}$ будем называть множество $\{\pi(a) : \pi \in S_{\mathbb{A}}\}$. Из теории групп известно, что орбиты разных элементов множества либо не пересекаются, либо совпадают. Таким образом, множество бинарных векторов $\mathbb{A}$ разбивается на классы эквивалентности — непересекающиеся орбиты действия группы $S_{\mathbb{A}}$. Обозначим множество орбит действия группы $S_{\mathbb{A}}$ на $\mathbb{A}$ с помощью $\Omega(\mathbb{A})$. Векторы ошибок из одной орбиты договоримся называть *идентичными*. Обратим внимание, что это определение не стоит путать с понятием *эквивалентные отображения из* $\mathcal{H}$, введенным ранее.

В [122, 123] показано, что идентичные векторы ошибок вносят равные вклады в вероятность переобучения. А именно, справедлив следующий результат:

Лемма 19 ([123]). *Для всех $a \in \mathbb{A}$ и $\pi \in S_{\mathbb{A}}$ справедливо:*

$$Q_{\mu,\varepsilon}(\mathbb{A}, a) = Q_{\mu,\varepsilon}\big(\mathbb{A}, \pi(a)\big).$$

Объединив этот результат с формулой разложения (5.2) мы немедленно получаем следующую формулу разложения вероятности переобучения по орбитам векторов ошибок:

Теорема 43 ([123]). *Для рандомизированного МЭР и произвольного множества попарно различных векторов ошибок $\mathbb{A} \in \{0,1\}^N$ справедливо:*

$$\begin{aligned} Q_{\mu,\varepsilon}(\mathbb{A}) &= \sum_{\omega \in \Omega(\mathbb{A})} |\omega| \frac{1}{C_N^n} \sum_{\substack{S \in [\mathbb{S}]^n: \\ a_\omega \in A(S)}} \frac{\mathbb{1}_{\{L(a_\omega, \mathbb{S}\setminus S) - L(a_\omega, S) \geqslant \varepsilon\}}}{|A(S)|} \\ &= \sum_{\omega \in \Omega(\mathbb{A})} |\omega| \, \mathbb{E}\left[\frac{\mathbb{1}_{\{L(a_\omega, \mathbb{S}\setminus S_n) - L(a_\omega, S_n) \geqslant \varepsilon\}} \mathbb{1}_{a_\omega \in A(S_n)}}{|A(S)|} \right], \end{aligned}$$

где a_ω — произвольный вектор ошибок из орбиты $\omega \in \Omega(\mathbb{A})$.

Как упоминалось ранее, этот результат существенно упрощает получение точных оценок вероятности переобучения в тех случаях, когда семейство векторов обладает симметрией. Если симметрий нет, — то есть подгруппа $S_{\mathbb{A}}$ состоит из одной тождественной перестановки — то последняя сумма состоит из $|\mathbb{A}|$ различных слагаемых, поскольку в этом случае $|\Omega(\mathbb{A})| = |A|$. Однако, по мере уменьшения числа орбит число слагаемых уменьшается.

Описанные результаты предлагают следующий двухшаговый подход к получения точных оценок вероятности переобучения:

1. Для заданного множества $\mathbb{A}$ находим его группу симметрий $S_{\mathbb{A}}$.
2. Вычисляем индивидуальные вклады векторов a_ω из различных орбит $\omega \in \Omega(\mathbb{A})$ в вероятность переобучения.

5.2.2 Новые результаты теоретико-группового подхода

В настоящем разделе будут приведены новые результаты в рамках теоретико-группового подхода, полученные в настоящей работе, описанные в [119, 120] и позже использованные в [37, 124].

Подгруппы симметрий. Часто поиск группы симметрий $S_{\mathbb{A}}$ для заданного множества бинарных векторов ошибок $\mathbb{A}$ оказывается сложной задачей. В этом случае в настоящей работе предлагается ограничиться поиском *подгруппы* группы симметрий $S_{\mathbb{A}}$. Следующий результат дает возможность работать с подгруппами группы симметрий $S_{\mathbb{A}}$:

Лемма 20. *Пусть* Π *— подгруппа группы симметрий* $S_{\mathbb{A}}$. *Тогда для всех* $a \in \mathbb{A}$ *и* $\pi \in \Pi$ *справедливо:*

$$Q_{\mu,\varepsilon}(\mathbb{A}, a) = Q_{\mu,\varepsilon}\big(\mathbb{A}, \pi(a)\big).$$

Доказательство. Достаточно заметить, что орбиты векторов ошибок подгруппы группы симметрий $S_{\mathbb{A}}$ являются подмножествами соответствующих орбит самой группы симметрии $S_{\mathbb{A}}$. С учетом этого утверждение леммы следует из Леммы 19. ■

Орбиты выборок. Ранее мы определили действие элементов симметрической группы Π_N на множестве всевозможных подвыборок $S \in \mathbb{S}$. Поскольку для всех $S \in \mathbb{S}$ и $\pi \in \Pi_N$ справедливо $|S| = |\pi(S)|$, то для любых $S \in [\mathbb{S}]^n, \pi \in \Pi_N$ также справедливо $\pi(S) \in [\mathbb{S}]^n$. *Орбитой действия* группы симметрий $S_{\mathbb{A}}$ на подвыборку $S \in [\mathbb{S}]^n$ мы будем называть множество

$\{\pi(S)\colon \pi \in S_{\mathbb{A}}\}$. Множество всех орбит действия группы симметрий $S_{\mathbb{A}}$ на $[\mathbb{S}]^n$ будем обозначать $\Omega\left([\mathbb{S}]^n\right)$. Выборки из одной орбиты действия группы $S_{\mathbb{A}}$ на множестве выборок $[\mathbb{S}]^n$ будем называть *идентичными*.

Справедлив следующий результат:

Лемма 21. *Для любых двух идентичных подвыборок $S', S'' \subseteq \mathbb{S}$ справедливо:*

$$|A\left(S'\right)| = |A\left(S''\right)|.$$

Доказательство. Докажем, что если $a_0 \in A(S)$ и $\pi \in S_{\mathbb{A}}$ для некоторой подвыборки $S \in \mathbb{S}$, то $\pi(a_0) \in A\left(\pi(S)\right)$. Утверждение леммы будет немедленно следовать из этого более общего утверждения.

Пользуясь Леммой 18, запишем:

$$L\left(a_0, S\right) = \min_{a\in\mathbb{A}} L(a, S) = \min_{a\in\mathbb{A}} L\big(\pi(a), \pi(S)\big).$$

Поскольку $\pi(\mathbb{A}) = \mathbb{A}$, то

$$\min_{a\in\mathbb{A}} L\big(\pi(a), \pi(S)\big) = \min_{a\in\mathbb{A}} L\big(a, \pi(S)\big).$$

А так как

$$L\big(\pi(a_0), \pi(S)\big) = L\left(a_0, S\right) = \min_{a\in\mathbb{A}} L\big(a, \pi(S)\big),$$

то $\pi(a_0) \in A\big(\pi(S)\big)$. ■

Лемма 22. *Если $S', S'' \subseteq \mathbb{S}$ — идентичные подвыборки и существует бинарный вектор ошибок $a \in A\left(S'\right) \cap A\left(S''\right)$, то для него верно:*

$$L(a, \mathbb{S} \setminus S') - L(a, S') = L(a, \mathbb{S} \setminus S'') - L(a, S'').$$

Доказательство. Пусть $S'' = \pi\left(S'\right)$ для некоторой перестановки $\pi \in S_{\mathbb{A}}$. Поскольку $a \in A\left(S'\right)$, то, пользуясь доказательством предыдущей леммы, мы имеем:

$$\pi(a) \in A\big(\pi\left(S'\right)\big) = A\left(S''\right).$$

Но вектор a также принадлежит множеству $A\left(S''\right)$. Следовательно

$$L\left(a, S''\right) = L\big(\pi(a), S''\big) = L\big(\pi(a), \pi\left(S'\right)\big) = L\left(a, S'\right).$$

Отсюда следует утверждение леммы, поскольку для любых $S \in \mathbb{S}$ и $a \in \mathbb{A}$ выполнено:

$$L(a, \mathbb{S} \setminus S) = \frac{NL(a, \mathbb{S}) - |S|L(a, S)}{N - |S|}.$$

■

Эти результаты позволяют получить следующую формулу разложения вероятности переобучения по орбитам множества векторов ошибок и орбитам множества подвыборок:

Теорема 44. *Для рандомизированного МЭР и произвольного множества попарно различных векторов ошибок $\mathbb{A} \in \{0,1\}^N$ справедливо:*

$$Q_{\mu,\varepsilon}(\mathbb{A}) = \sum_{\omega \in \Omega(\mathbb{A})} \frac{|\omega|}{C_N^n} \sum_{\tau \in \Omega([\mathbb{S}]^n)} \frac{|\{S \in \tau \colon a_\omega \in A(S)\}|}{|A(S_\tau)|} \mathbb{1}_{\{L(a_\omega, \mathbb{S}\setminus S_\tau) - L(a_\omega, S_\tau) \geqslant \varepsilon\}}, \tag{5.3}$$

где a_ω — произвольный вектор ошибок из орбиты $\omega \in \Omega(\mathbb{A})$, а S_τ — произвольная выборка из орбиты $\Omega([\mathbb{S}]^n)$.

Результат остается справедлив, если мы заменим множества $\Omega(\mathbb{A})$ и $\Omega([\mathbb{S}]^n)$ соответствующими множествами орбит произвольной подгруппы группы симметрий $S_{\mathbb{A}}$

Доказательство. Доказательство следует из Теоремы 43 и Лемм 21, 22 и 20. ■

Далее на основе последнего результата мы получим точные значения вероятности переобучения для трех модельных семейств векторов ошибок, являющихся различными подмножествами *шара* в Булевом кубе $\{0,1\}^N$, определяемого метрикой Хэмминга. Но перед этим для наглядности изложения проиллюстрируем все введенные выше понятия на примере простого множества $\mathbb{A}$.

Пример одного множества $\mathbb{A}$. Рассмотрим следующее множество векторов ошибок $\mathbb{A}$:

	(X_1, Y_1)	(X_2, Y_2)	(X_3, Y_3)	(X_4, Y_4)
a_1	1	1	0	0
a_2	0	1	1	0
a_3	0	0	1	1

Здесь на пересечении столбца (X, Y) и строки a представлено значение индикатора ошибки $I_a(X, Y)$. Итак, множество $\mathbb{A}$ состоит из трех различных векторов. Несложно убедиться, что группа симметрий $S_{\mathbb{A}}$ в данном случае состоит всего лишь из двух перестановок $\pi_1 = \left(\begin{smallmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{smallmatrix}\right)$ и $\pi_2 = \left(\begin{smallmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{smallmatrix}\right)$. У группы $S_{\mathbb{A}}$ есть две подгруппы, каждая из которых состоит из одного элемента. Множество $\Omega(\mathbb{A})$ состоит из двух орбит:

$$\omega_1 = \{a_1, a_3\}, \quad \omega_2 = \{a_2\}.$$

Множество $\Omega(\Omega([\mathbb{S}]^n))$ состоит из четырех орбит:

$$\tau_1 = \big\{\{(X_1, Y_1), (X_2, Y_2)\}, \{(X_3, Y_3), (X_4, Y_4)\}\big\};$$
$$\tau_2 = \big\{\{(X_1, Y_1), (X_3, Y_3)\}, \{(X_2, Y_2), (X_4, Y_4)\}\big\};$$
$$\tau_3 = \big\{\{(X_1, Y_1), (X_4, Y_4)\}\big\};$$
$$\tau_4 = \big\{\{(X_2, Y_2), (X_3, Y_3)\}\big\}.$$

5.2.3 Свойства сходства и расслоения множества векторов ошибок

В работах [114,115] было экспериментально установлено, что получение точных или, по крайней мере, несильно завышенных оценок вероятности переобучения невозможно без учета двух важных свойств множества векторов ошибок $\mathbb{A}$, которые были названы свойствами *расслоения* и *сходства*:

- **Расслоение:** поскольку на практике зачастую используются *универсальные* семейства отображений $\mathcal{H}$, способные справиться со многими задачами, то при решении конкретной задачи «пригодным» будет лишь малое подмножество используемого семейства $\mathcal{H}$. Это подмножество имеет векторы ошибок $a \in \mathbb{A}$ с относительно малыми значениями риска $L(a, \mathbb{S})$ на генеральной выборке. Остальные векторы ошибок будут иметь достаточно много ошибок на генеральной выборке и, как следствие, выбираться методом обучения (МЭР) достаточно редко (с малой вероятностью). Такие векторы ошибок, фактически, не будут принимать участия в процессе обучения. Поэтому оценки, основанные на равномерных отклонениях по всему множеству $\mathcal{H}$, будут сильно завышены для конкретных задач.

- **Сходство:** во множествах векторов ошибок $\mathbb{A}$, соответствующих используемым на практике семействам отображений $\mathcal{H}$, часто содержится большое число похожих между собой (в смысле метрики Хэмминга) векторов ошибок. Например, для линейных классификаторов это можно объяснить непрерывностью изменения гиперплоскостей при изменении ее параметров. Чем больше похожих векторов во множестве $\mathbb{A}$, тем сильнее будет завышено неравенство Буля, и, соответственно, все оценки, использующие его.

Более того, в тех же работах было показано, что учитывать два этих свойства следует *одновременно*: отдельный учет лишь одного из свойств не гарантирует хорошего результата.

В работе [115] приведен обширный обзор существующих в теории статистического обучения подходов и оценок и показано, что большинство из них либо вовсе не учитывают

описанных выше свойств множества векторов ошибок $\mathbb{A}$, либо учитывают лишь одно из них. Например, оценка Бритвы Оккама, приведенная в Теореме 25, совершает попытку учета свойства расслоения посредством введения априорного распределения на множестве $\mathcal{H}$, но не учитывает сходств между элементами класса потерь $\mathcal{F} = \{\ell_h(X,Y)\colon h \in \mathcal{H}\}$. Оценка же Теоремы 27, основанная на покрытиях класса $\mathcal{F} = \{\ell_h(X,Y)\colon h \in \mathcal{H}\}$, наоборот, учитывает геометрическую структуру множества векторов ошибок, но по-прежнему опирается на равномерные отклонения по всему классу отображений $\mathcal{H}$.

Мы кратко дополним обсуждение, представленное в [115], рассмотрев в этом контексте результаты локального анализа, описанного в Разделе 3.2.3. Рассмотрим главный результат локального анализа — Теорему 36. Для функций $f \in \mathcal{F}$ из класса потерь определим следующую метрику:

$$\|f\|_{L_2(P)} = \big(\mathbb{E}_{(X,Y)\sim P}[f^2(X,Y)]\big)^{1/2}.$$

- Множество $\mathcal{H}(r)$, определенное в (3.24), является подмножеством отображений класса $\mathcal{H}$, потери которых близки в смысле метрики $L_2(P)$ к потерям оптимального отображения $\ell_{h^*}(X,Y)$.

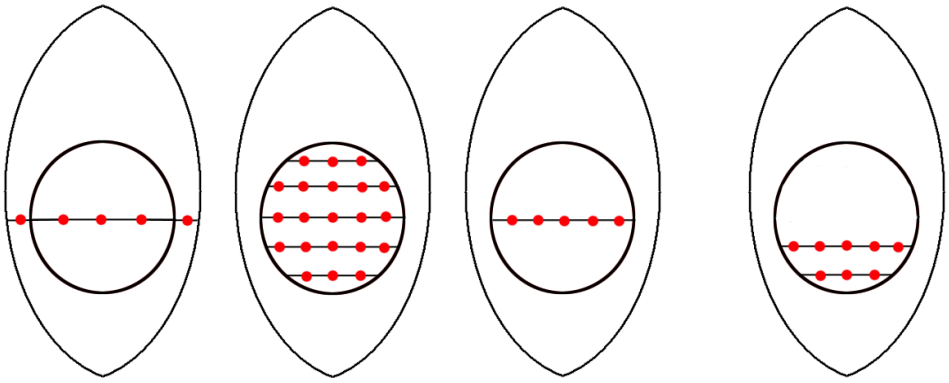
- Теперь обратимся к условию (3.25) теоремы. Оно может быть переписано в следующем виде:

$$\forall h \in \mathcal{H}: \quad \|\ell_h - \ell_{h^*}\|^2_{L_2(P)} \leqslant B \cdot \mathcal{E}(h) = B \cdot \big(L(h) - L(h^*)\big).$$

  Это условие можно интерпретировать следующим образом: расстояния от всех функций $\mathcal{F}$ до потерь оптимального отображения $\ell_{h^*}(X,Y)$ в смысле метрики $L_2(P)$ ограничены сверху значениями избыточных рисков этих функций. В частности, отсюда следует, что потери «хороших» отображений $h \in \mathcal{H}$ с малыми значениями избыточного риска $\mathcal{E}(h)$ близки по метрике $L_2(P)$ к потерям оптимального отображения $h^* \in \mathcal{H}$ (и поэтому можно надеяться, что они будут существенно пересекаться со множеством $\mathcal{H}(r)$).

- Наконец, обратим внимание на то, что теорема ограничивает избыточный риск сверху, грубо говоря, равномерными по подмножеству $\mathcal{H}(r)$ отклонениями.

Вместе эти наблюдения позволяют заключить, что Теорема 36 *одновременно учитывает* и сходство векторов потерь класса $\mathcal{H}$ между собой и расслоение класса $\mathcal{H}$ по уровням среднего риска $L(h)$.



1) слой 2) шар 3) слой шара 4) нижние слои шара

Рисунок 5.1: Иллюстрация четырех модельных множеств $\mathbb{A}$, рассматриваемых в настоящем параграфе.

5.2.4 Три подмножества шара в Булевом кубе

В настоящем разделе мы приведем три новых оценки вероятности переобучения для модельных семейств, полученные в настоящей работе, опубликованные в [119, 120] и позже частично использованные в [37, 124].

Множество всех бинарных векторов ошибок (Булев куб) длины N мы буем обозначать $\mathbb{B}^N$. Договоримся называть m-м *слоем* множества векторов ошибок $\mathbb{A} \subseteq \mathbb{B}^N$ (и обозначать $\mathbb{A}_m$) подмножество векторов $a \in \mathbb{A}$, допускающих одинаковое число ошибок на генеральной выборке: $\mathbb{A}_m = \{a \in \mathbb{A} \colon n(a, \mathbb{S}) = m\}$. При этом m-й слой всего Булева куба $\mathbb{B}^N$ мы будем обозначать $\mathbb{B}^N_m$. Кроме того, обозначим $L_1(P)$-расстояние (в данном случае — расстояние Хэмминга) между векторами ошибок $a', a'' \in \mathbb{A}$ следующим образом:

$$d(a', a'') = \sum_{(X,Y)\in\mathbb{S}} |I_{a'}(X,Y) - I_{a''}(X,Y)|.$$

В настоящем параграфе мы продолжим изучать эффект свойств сходства и расслоения на вероятность переобучения на примере четырех модельных семейств, геометрическая иллюстрация которых представлена на Рисунке 5.1:

1. Семейство, состоящее из случайных и не связанных между собой векторов ошибок m-го слоя Булева куба. Это множество $\mathbb{A}$ не обладает ни свойством расслоения, ни свойством сходства.

2. Шар радиуса r с центром в векторе $a_0 \in \{0,1\}^N$ — это следующее множество векторов ошибок:

$$B_r(a_0) = \big\{a \in \{0,1\}^N \colon d(a, a_0) \leqslant r\big\}.$$

Это множество, в некотором смысле, наиболее «компактно» среди всех множеств векторов ошибок, имеющих свойство расслоения.

3. Семейство, состоящее из центрального слоя шара:

$$B_r^c(a_0) = \mathbb{B}_m^N \cap B_r(a_0),$$

где $m = n(a_0, \mathbb{S})$. Это множество не обладает эффектом расслоения, но обладает эффектом сходства, причём является в некотором смысле «наиболее компактным» среди всех таких множеств.

4. Наконец, d нижних слоёв хэммингова шара — это следующее множество:

$$B_r(a_0, d) = \bigcup_{j=1,\ldots,d} \left(\mathbb{B}_{m_j}^N \cap B_r(a_0) \right),$$

где $m_j = \max\{n(a_0, \mathbb{S}) - r, 0\} + (d-1)$. Это семейство позволяет исследовать эффект расслоения внутри шара.

Поскольку шары в Булевом кубе обладают определенными симметриями, для вычисления вероятностей переобучения перечисленных семейств естественно использовать теоретико-групповой подход, описанный выше.

Свойства сходства: центральный слой шара. Начнём изучение влияния свойства сходства векторов во множестве $\mathbb{A}$ на вероятность переобучения с простого эксперимента.

Для произвольного заданного множества $\mathbb{A}$ вероятность переобучения Q_ε нетрудно оценить эмпирически методом Монте-Карло. Для этого достаточно заменить вероятность, определяемую как долю всех разбиений генеральной выборки $\mathbb{S}$, на долю разбиений из заданного случайного подмножества разбиений. В данном эксперименте бралась тысяча случайных разбиений. Сравнивались два множества, лежащих в m-м слое $\mathbb{B}_m^N$ для $m = 10$. Первое множество состояло из D *случайных* представителей m-го слоя $\mathbb{B}_m^N$. Второе множество состояло из D случайных представителей семейства $B_r^c(a_0)$, такого что $n(a_0, \mathbb{S}) = m$, для $r = 2$. Несложные вычисления показывают, что во втором семействе содержится в точности 1901 вектор. Векторы из второго множесва брались последовательно в порядке увеличения расстояния Хэмминга от центра шара a_0. При этом использовались параметры $n = u = 100$, $\varepsilon = 0.05$. Для наглядности эксперимента мы также вычислили значения оценок Теоремы 24, очевидно, совпадающие для обоих множеств (поскольку они зависят лишь от мощности семейства $\mathbb{A}$). Зависимости эмпирических оценок вероятности переобучения Q_ε, а также значений оценки Теоремы 24 от числа векторов D в семействе представлены на Рисунке 5.2.

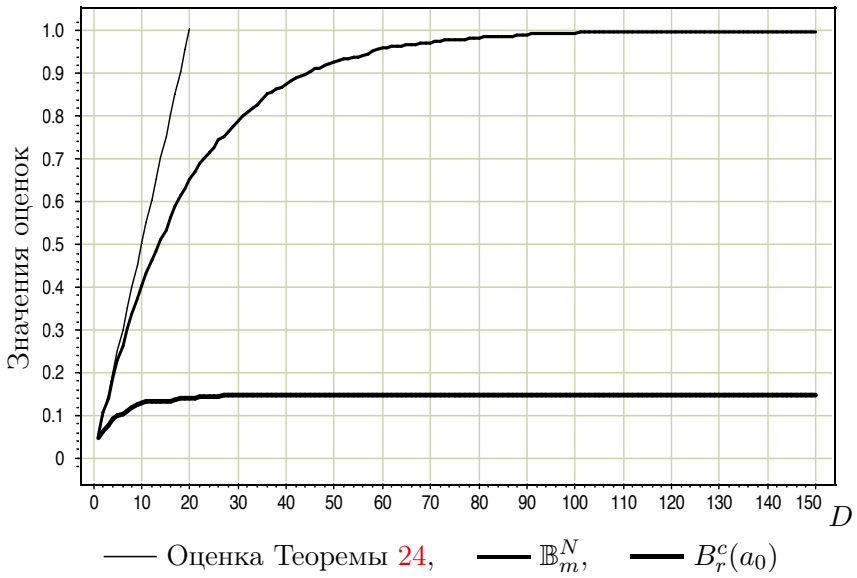


Рисунок 5.2: Оценки для различных подмножеств слоя Булева куба $\mathbb{B}_m^N$.

Заметим:

- Вероятность переобучения для более «плотного» второго семейства проходит значительно ниже, чем для «разреженного» первого семейства. Получение точных оценок вероятности переобучения для шара и его слоёв представляет значительный теоретический интерес, поскольку шар является в некотором смысле наиболее «плотным» множеством бинарных векторов.

- Для первого семейства вероятность переобучения Q_ε достигает единицы уже при $D \approx 100$. Это означает, что для оценивания вероятности переобучения «плотных» семейств не обязательно брать все алгоритмы семейства. Для получения достаточно точных нижних оценок можно брать относительно небольшое «разреженное» подмножество, состоящее из несхожих векторов. В частности, можно отбирать их случайным образом.

Далее мы получим точное значение вероятности переобучения $Q_{\mu,\varepsilon}(\mathbb{A})$ для множества $\mathbb{A} = B_r^c(a_0)$, где $n(a_0, \mathbb{S}) = m$.

Положим без ограничения общности, что вектор a_0 ошибается на m первых парах генеральной выборки $\mathbb{S}$. В дальнейшем множество, состоящее из первых m пар генеральной выборки, будем обозначать S^m. Множество, состоящее из последних $N - m$ объектов, будем обозначать S^{N-m}. Введём обозначение $b = m - \lfloor r/2 \rfloor$.

Пусть Π^m — подгруппа группы Π_N, состоящая из *всех* перестановок $\pi \in \Pi_N$, действие которых на множестве $\{m+1, \ldots, N\}$ является тождественным. Таким образом, элементы

Π^m переставляют пары подвыборки S^m, отображая при этом остальные пары генеральной выборки $\mathbb{S}$ в самих себя. Аналогично определим подгруппу Π^{N-m}, состоящую из всех перестановок $\pi \in \Pi_N$, действие которых на множестве $\{1, \ldots, m\}$ является тождественным. Определим стандартным образом декартово произведение $\Pi^m \times \Pi^{N-m}$. Действие элемента (π, φ) группы $\Pi^m \times \Pi^{N-m}$, где $\pi \in \Pi^m$ и $\varphi \in \Pi^{N-m}$, на векторы ошибок $a \in \mathbb{A}$ определяется последовательным действием перестановок π и φ. При этом, как несложно убедиться, для всех $a \in \mathbb{A}$ выполнено $\pi\big(\varphi(a)\big) = \varphi\big(\pi(a)\big)$.

Лемма 23. *Группа $\Pi^m \times \Pi^{N-m}$ является подгруппой группы симметрий множества векторов $\mathbb{A}$.*

Доказательство. Достаточно показать, что если $a \in \mathbb{A}$, то и $\pi(a) \in \mathbb{A}$ для $\pi \in \Pi^m \times \Pi^{N-m}$.

Поскольку действие элементов симметрической группы Π_N на векторы из $\mathbb{A}$ не меняет их числа ошибок, то $n\big(\pi(a), \mathbb{S}\big) = n(a, \mathbb{S}) = m$. Очевидно, для $\pi \in \Pi^m \times \Pi^{N-m}$ справедливы следующие соотношения: $n(a, S^m) = n\big(\pi(a), S^m\big)$ и $n(a, S^{N-m}) = n\big(\pi(a), S^{N-m}\big)$. Из них получаем:

$$\begin{aligned} d(a, a_0) &= m - n(a, S^m) + n(a, S^{N-m}) = \\ &= m - n\big(\pi(a), S^m\big) + n\big(\pi(a), S^{N-m}\big) \\ &= d\big(\pi(a), a_0\big). \end{aligned}$$

Итак, вектор $\pi(a)$ допускает ровно m ошибок на генеральной выборке $\mathbb{S}$, и расстояние Хэмминга $d\big(\pi(a), a_0\big) = d(a, a_0) \leqslant r$. ■

Лемма 24. *Орбитами действия группы $\Pi^m \times \Pi^{N-m}$ на множестве $\mathbb{A}$ являются пересечения m-го слоя Булева куба со сферами (в метрике Хэминга) радиусов $2k$, $k = 0, \ldots, \lfloor r/2 \rfloor$ с центрами в a_0.*

Доказательство. Пусть вектор a допускает m ошибок на генеральной выборке и принадлежит сфере радиуса r_1: $d(a, a_0) = r_1$. В доказательстве предыдущей леммы мы установили, что если $\pi \in \Pi^m \times \Pi^{N-m}$, то $d\big(\pi(a), a_0\big) = r_1$ и $\pi(a)$ также принадлежит m-му слою. Очевидно также, что расстояние Хэмминга между двумя векторами, допускающими одинаковое число ошибок на генеральной выборке $\mathbb{S}$ — величина четная.

Осталось доказать, что для любых векторов $a_1, a_2 \in \mathbb{B}_m^N$ таких, что $d(a_1, a_0) = d(a_2, a_0) = r_1$, найдется перестановка $\pi \in \Pi^m \times \Pi^{N-m}$, для которой $\pi(a_1) = a_2$. Но это следует из того, что $n(a_1, S^m) = n(a_2, S^m)$ и $n(a_1, S^{N-m}) = n(a_2, S^{N-m})$. Последний факт легко установить, выразив число ошибок, допускаемых векторами a_1 и a_2 на выборке S^m, через m и r_1. ■

Таким образом, всего имеется $\lfloor r/2 \rfloor + 1$ орбит. Пронумеруем их индексами $h = b, \ldots, m$ по числу ошибок, допускаемых векторами из них на множестве S^m. Тогда в h-й орбите, которую мы будем обозначать $\mathrm{Orb}(h)$, ровно $C_m^h C_{N-m}^{m-h}$ векторов.

В процессе доказательства мы будем пользоваться формулой (5.3). Для этого нам необходимо исследовать условия попадания векторов ошибок во множество $A(S)$.

Лемма 25. *Для всех $S \in [\mathbb{S}]^n$ и $a \in \mathbb{A} \backslash \mathrm{Orb}(m - \lfloor r/2 \rfloor)$ условия $a \in A(S)$ и $n(a, S) = 0$ эквивалентны.*

Доказательство. Достаточность очевидна. Докажем необходимость.

Пусть вектор $a^h \in \mathrm{Orb}(h)$ для $h \neq b$ попал во множество $A(S)$. Нам требуется доказать, что он не допускает ошибок на выборке S. Введем следующие обозначения: $S^{n_1} = S^m \cap S$, $S^{n_2} = S^{N-m} \cap S$; $n_1 = |S^{n_1}|$, $n_2 = |S^{n_2}|$.

Допустим, $n_1 \leqslant \lfloor r/2 \rfloor$. Покажем, что для любой выборки $S \in [\mathbb{S}]^n$ найдётся вектор $a \in \mathbb{A}$ такой, что $n(a, S) = 0$. Рассмотрим $\mathrm{Orb}(m - n_1)$. Все элементы из этой орбиты допускают ровно $m - n_1$ ошибок на множестве S^m и n_1 ошибок на множестве S^{N-m}. Пусть A_1 — те из них, которые не допускает ни одной ошибки на множестве S^{n_1}. Множество A_1, очевидно, непусто. Векторы из A_1 не ошибаются на $N - m - n_1$ объектах множества S^{N-m}. Поскольку $n_1 + n_2 = n \leqslant N - m$, то $N - m - n_1 \geqslant n_2$. Таким образом, во множестве A_1 существует вектор a', для которого $n(a', S^{n_2}) = 0$. По определению множества A_1 мы имеем $n(a', S) = 0$.

Рассмотрим теперь случай, когда $n_1 > \lfloor r/2 \rfloor$. Векторы из $\mathrm{Orb}(m - \lfloor r/2 \rfloor)$ допускают ровно $\lfloor r/2 \rfloor$ ошибок на S^{N-m}, а правильные ответы дают на $N - m - \lfloor r/2 \rfloor$ парах этого множества. Таким образом, множество $A_2 = \{a \in \mathrm{Orb}(m - \lfloor r/2 \rfloor) \colon n(a, S^{n_2}) = 0\}$ непусто, поскольку $n_2 = n - n_1 < n - \lfloor r/2 \rfloor \leqslant N - m - \lfloor r/2 \rfloor$. Векторы из указанной орбиты допускают $m - \lfloor r/2 \rfloor$ ошибок на множестве S^m. Во множестве A_2 найдётся вектор a'', для которого $n\left(a'', S^{n_1}\right) = n_1 - \lfloor r/2 \rfloor$.

Итак, $n\left(a'', S^{n_2}\right) = 0$ и $n\left(a'', S^{n_1}\right) = n_1 - \lfloor r/2 \rfloor$. Но $n\left(a^h, S^{n_1}\right) \geqslant n_1 - (m - h) > n_1 - m + (m - \lfloor r/2 \rfloor) = n_1 - \lfloor r/2 \rfloor = n\left(a'', S^{n_1}\right)$, что противоречит тому, что $a^h \in A(S)$. ■

Следствие 17. *В ходе предыдущего доказательства также установлено, что для $S \in [\mathbb{S}]^n$ при $|S \cap S^m| > \lfloor r/2 \rfloor$ векторы из орбит, отличных от b-ой, не могут попасть во множество $A(S)$.*

Лемма 26. *Для произвольного вектора $a^b \in \mathrm{Orb}(m - \lfloor r/2 \rfloor)$ и $S \in [\mathbb{S}]^n$ справедливо следующее:*

- *если S удовлетворяет условию $|S \cap S^m| \leqslant \lfloor r/2 \rfloor$, то следующие утверждения эквивалентны: $a^b \in A(S)$ и $n\left(a^b, S\right) = 0$;*

- *если же S удовлетворяет условию $|S\cap S^m| > \lfloor r/2\rfloor$, то условие $a^b \in A(S)$ эквивалентно следующему:*

$$\begin{cases} n\left(a^b, S\cap S^{N-m}\right) = 0; \\ n\left(a^b, S\cap S^m\right) = |S\cap S^m| - \lfloor r/2\rfloor. \end{cases}$$

Доказательство. Начнём со случая, когда $n_1 \leqslant \lfloor r/2\rfloor$. В этом случае рассуждения полностью повторяют предыдущие (первая часть доказательства Леммы 25), и мы приходим к выводу, что $a^b \in A(S)$ эквивалентно $n\left(a^b, S\right) = 0$.

Для $n_1 > \lfloor r/2\rfloor$ ситуация меняется. Предположим, что $a^b \in \mathrm{Orb}(m-\lfloor r/2\rfloor)$, $a^b \in A(S)$, и $|S\cap S^m| > \lfloor r/2\rfloor$. Нашей ближайшей целью будет показать, что в этом случае $n\left(a^b, S^{n_2}\right) = 0$, а $n\left(a^b, S^{n_1}\right) = n_1 - \lfloor r/2\rfloor$.

Предположим, что $n\left(a^b, S^{n_2}\right) \neq 0$. Поскольку $n_2 = n-n_1 < n-\lfloor r/2\rfloor \leqslant N-m-\lfloor r/2\rfloor$, в $\mathrm{Orb}(m-\lfloor r/2\rfloor)$ найдётся вектор a^b_0, ошибки которого на множестве X^m совпадают с ошибками вектора a^b, в то время как $n\left(a^b_0, S^{n_2}\right) = 0$. Это противоречит тому, что $a^b \in A(S)$.

Равенство $n\left(a^b, S^{n_1}\right) = n_1 - \lfloor r/2\rfloor$ следует из второй части доказательства Леммы 25.

На этом доказательство утверждения заканчивается, поскольку для выборок S, таких что $|S\cap S^m| > \lfloor r/2\rfloor$, не существует векторов из множества $\mathbb{A}$, допускающих на S меньше $n_1 - \lfloor r/2\rfloor$ ошибок. ■

Лемма 27. *Орбитами действия группы $\Pi^m \times \Pi^{N-m}$ на множестве выборок $[\mathbb{S}]^n$ являются множества $\{S \in [\mathbb{S}]^n : |S\cap S^m| = i\}$, где $i = \max(0, m-u), \ldots, \min(n, m)$.*

Доказательство. Лемма очевидным образом следует из определений орбиты и действия группы перестановок на множестве разбиений. ■

Для формулировки следующего результата нам потребуется определение *гипергеометрической функции распределения* $H_N^{n,m}(z)$:

$$H_N^{n,m}(z) = \sum_{s=\max(0,m-u)}^{\lfloor z\rfloor} \frac{C_m^s C_{N-m}^{n-s}}{C_N^n}.$$

Теорема 45 (Точная оценка для слоя шара)**.** *Пусть $n(a_0, \mathbb{S}) = m$. Рассмотрим множество $\mathbb{A} = B_r^c(a_0)$. Пусть $n \leqslant N-m$. Тогда для РМЭР вероятность переобучения может быть записана в виде:*

$$Q_{\mu,\varepsilon}(\mathbb{A}) = \begin{cases} H_N^{n,m}\big(s(\varepsilon) + r/2\big), & \text{если } s(\varepsilon) \geqslant 0; \\ 0, & \text{если } s(\varepsilon) < 0, \end{cases}$$

где $s(\varepsilon) = \frac{n}{N}(m - \varepsilon u)$.

Доказательство. Лемма 23 дает возможность воспользоваться формулой (5.3):

$$Q_{\mu,\varepsilon}(\mathbb{A}) = \sum_{\omega\in\Omega(\mathbb{A})} \frac{|\omega|}{C_N^n} \sum_{\tau\in\Omega([\mathbb{S}]^n)} \frac{|\{S\in\tau\colon a_\omega\in A(S)\}|}{|A(S_\tau)|} \mathbb{1}_{\{L(a_\omega,\mathbb{S}\setminus S_\tau)-L(a_\omega,S_\tau)\geqslant\varepsilon\}}.$$

С учетом Лемм 24 и 27

$$Q_{\mu,\varepsilon}(\mathbb{A}) = \sum_{h=m-\lfloor r/2\rfloor}^{m} \frac{C_m^h C_{N-m}^{m-h}}{C_N^n} \sum_{i=0}^{\min(n,m)} \underbrace{|\{S\in[\mathbb{S}]^n : |S\cap S^m| = i, a^h\in A(S)\}|}_{S(i,h)} \frac{\mathbb{1}_{\{L(a^h,\mathbb{S}\setminus S_i)-L(a^h,S_i)\geqslant\varepsilon\}}}{|A(S_i)|},$$

где a^h — произвольный вектор орбиты с индексом h, а $S_i\in[\mathbb{S}]^n$ — произвольная подвыборка, в которую входят ровно i объектов из S^m. Всюду далее будем пользоваться следующим обозначением: $\rho = \lfloor r/2\rfloor$.

Разобьем сумму по i на два слагаемых:

$$\begin{aligned} Q_{\mu,\varepsilon}(\mathbb{A}) = &\sum_{h=m-\rho}^{m} \frac{C_m^h C_{N-m}^{m-h}}{C_N^n} \sum_{i=0}^{\rho} S(i,h) \frac{\mathbb{1}_{\{L(a_\omega,\mathbb{S}\setminus S_\tau)-L(a_\omega,S_\tau)\geqslant\varepsilon\}}}{|A(X_i)|} \\ &+ \sum_{h=m-\rho}^{m} \frac{C_m^h C_{N-m}^{m-h}}{C_N^n} \sum_{i=\rho+1}^{\min(n,m)} S(i,h) \frac{\mathbb{1}_{\{L(a_\omega,\mathbb{S}\setminus S_\tau)-L(a_\omega,S_\tau)\geqslant\varepsilon\}}}{|A(X_i)|}. \end{aligned}$$

Из Леммы 25 следует, что первое слагаемое соответствует случаю выбора векторов, не допускающих ошибок на выборке. Из Леммы 26 и из Следствия 17 следует, что во втором слагаемом сумму по орбитам векторов ошибок можно опустить, поскольку при данных разбиениях генеральной выборки выбираться будут только векторы из орбиты $m-\rho$, допускающие $i-\rho$ ошибок на S_i. Учитывая эти факты, имеем:

$$\begin{aligned} Q_{\mu,\varepsilon}(\mathbb{A}) = \mathbb{1}_{\{s(\varepsilon)\geqslant 0\}} &\sum_{h=m-\rho}^{m} \frac{C_m^h C_{N-m}^{m-h}}{C_N^n} \sum_{i=0}^{\rho} \frac{S(i,h)}{|A(S_i)|} \\ &+ \frac{C_m^\rho C_{N-m}^\rho}{C_N^n} \sum_{i=\rho+1}^{\min(n,m)} \frac{S(i,h)}{|A(S_i)|} \mathbb{1}_{\{i\leqslant s(\varepsilon)+r/2\}}. \end{aligned} \tag{5.4}$$

Значения $|A(S_i)|$ легко вычисляются на основе Лемм 25 и 26:

$$|A(S_i)| = \begin{cases} \sum\limits_{j=m-\rho}^{m} C_{m-i}^j C_{u-m+i}^{m-j}, & i\leqslant\rho; \\ C_i^\rho C_{u-m+i}^\rho, & i>\rho. \end{cases}$$

Значения $S(i,h)$ также легко вычисляются на основе Лемм 25 и 26:

$$S(i,h) = \begin{cases} C_{m-h}^i C_{N-2m+h}^{n-i}, & i\leqslant\rho; \\ C_{m-\rho}^{m-i} C_{N-m-\rho}^{n-i}, & i>\rho. \end{cases}$$

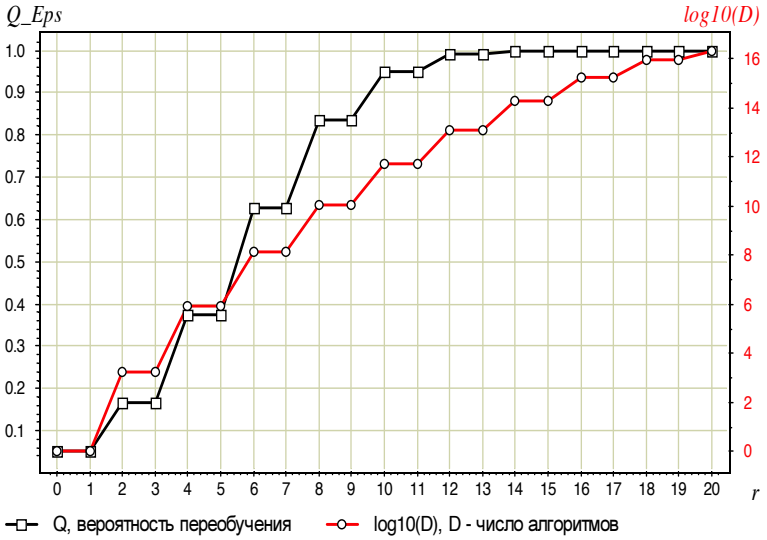


Рисунок 5.3: Зависимость вероятности переобучения Q_ε и $\log_{10}|\mathbb{A}|$ слоя шара $\mathbb{A} = B_r^c(a_0)$ от радиуса шара r, при $n = u = 100,\ n(a_0, \mathbb{S}) = m = 10,\ \varepsilon = 0.05$.

Подставим полученные значения в (5.4):

$$Q_{\mu,\varepsilon}(\mathbb{A}) = \mathbb{1}_{\{s(\varepsilon)\geqslant 0\}} \sum_{h=m-\rho}^{m} \frac{C_m^h C_{N-m}^{m-h}}{C_N^n} \sum_{i=0}^{\rho} \frac{C_{m-h}^i C_{N-2m+h}^{n-i}}{\sum\limits_{j=m-\rho}^{m} C_{m-i}^j C_{u-m+i}^{m-j}} + \frac{C_m^\rho C_{N-m}^\rho}{C_N^n} \sum_{i=\rho+1}^{\min(n,m)} \mathbb{1}_{\{i\leqslant s(\varepsilon)+r/2\}} \frac{C_{m-\rho}^{m-i} C_{N-m-\rho}^{n-i}}{C_i^\rho C_{u-m+i}^\rho}. \tag{5.5}$$

Сократим вид второго слагаемого с помощью комбинаторного тождества

$$\frac{C_{n-u}^{m-u}}{C_n^m} = \frac{C_m^u}{C_n^u}.$$

Заметим, что

$$\frac{C_{m-\rho}^{m-i}}{C_i^\rho} = \frac{C_{i-(i-m+\rho)}^{\rho-(i-m+\rho)}}{C_i^\rho} = \frac{C_\rho^{i-m+\rho}}{C_i^{i-m+\rho}} = \frac{C_\rho^{m-i}}{C_i^{m-\rho}}.$$

Далее,

$$\frac{C_{N-m-\rho}^{n-i}}{C_{N-m-n+i}^\rho} = \frac{C_{N-m-n+i-(i-n+\rho)}^{\rho-(i-n+\rho)}}{C_{N-m-n+i}^\rho} = \frac{C_\rho^{n-i}}{C_{N-m-n+i}^{i-n+\rho}}.$$

Теперь воспользуемся аналогичным тождеством $C_n^m C_m^u = C_n^u C_{n-u}^{m-u}$ и получим:

$$C_m^\rho C_\rho^{m-i} = C_m^i C_i^{m-\rho},$$
$$C_{N-m}^\rho C_\rho^{i-n+\rho} = C_{N-m}^{n-i} C_{N-m-n+i}^{\rho-n+i}.$$

Подставив эти результаты во второе слагаемое (5.5) и обозначив $h_N^{n,m}(i) = \frac{C_m^i C_{N-m}^{n-i}}{C_N^n}$, мы получаем:

$$\sum_{i=\rho+1}^{\min(n,m)} h_N^{n,m}(i) \mathbb{1}_{\{i-\rho\leqslant s(\varepsilon)\}} = \sum_{i=\rho+1}^{\lfloor r/2+s(\varepsilon)\rfloor} h_N^{l,m}(i). \tag{5.6}$$

Теперь займемся первым слагаемым. С учетом

$$C_m^h C_{m-h}^i = C_m^{m-h} C_{m-h}^i = C_m^i C_{m-i}^h,$$

$$C_{N-m}^{m-h} C_{N-2m+h}^{n-i} = C_{N-m}^{N-2m+h} C_{N-2m+h}^{n-i} = C_{N-m}^{n-i} C_{u-m+i}^{m-h}.$$

первое слагаемое (5.5) переписывается в виде:

$$\begin{aligned}
&\mathbb{1}_{\{0\leqslant s(\varepsilon)\}} \sum_{h=m-\rho}^{m} \sum_{i=0}^{\rho} h_N^{n,m}(i) \frac{C_{m-i}^h C_{u-m+i}^{m-h}}{\sum\limits_{j=m-\rho}^{m} C_{m-i}^j C_{u-m+i}^{m-j}} \\
&= \mathbb{1}_{\{0\leqslant s(\varepsilon)\}} \sum_{h=m-\rho}^{m} \sum_{i=0}^{\rho} h_N^{n,m}(i) \frac{h_u^{m,m-i}(h)}{\sum\limits_{j=m-\rho}^{m} h_u^{m,m-i}(j)} \\
&= \mathbb{1}_{\{0\leqslant s(\varepsilon)\}} \sum_{i=0}^{\rho} h_N^{n,m}(i) \frac{\sum\limits_{h=m-\rho}^{m} h_u^{m,m-i}(h)}{\sum\limits_{j=m-\rho}^{m} h_u^{m,m-i}(j)} \\
&= \mathbb{1}_{\{0\leqslant s(\varepsilon)\}} \sum_{i=0}^{\rho} h_N^{n,m}(i). \qquad (5.7)
\end{aligned}$$

Подстановка (5.6) и (5.7) в (5.5) завершает доказательство. ■

На Рисунке 5.3 представлена зависимость точной оценки вероятности переобучения Q_ε для слоя шара $\mathbb{A} = B_r^c(a_0)$, а также числа векторов в данном множестве, от радиуса шара r. Видно, что за счёт значительной «плотности» данного семейства вероятность переобучения может оставаться на приемлемо низком уровне при мощности семейства порядка тысяч и относительно небольшой длине выборки $n = u = 100$. Если мы попробуем применить оценку Теоремы 24 в данном случае, мы получим существенно завышенный результат.

Теорема 46. *Пусть $n(a_0, \mathbb{S}) = m$. Рассмотрим множество $\mathbb{A} = B_r^c(a_0)$. Пусть $n > N - m$. Тогда для РМЭР вероятность переобучения может быть записана в виде:*

$$Q_{\mu,\varepsilon}(\mathbb{A}) = \begin{cases} H_N^{n,m}\big(s(\varepsilon) + r/2\big), & \text{если } s(\varepsilon) \geqslant m - u; \\ 0, & \text{если } s(\varepsilon) < m - k, \end{cases}$$

где $s(\varepsilon) = \frac{n}{N}(m - \varepsilon u)$.

Доказательство этой теоремы полностью повторяет доказательство Теоремы 45. Отличие случая $u < m$ от предыдущего заключается в новых условиях попадания векторов во множество $A(S)$, описываемых следующими леммами.

Лемма 28. *Для всех $S \in [\mathbb{S}]^n$ и $a \in \mathbb{A} \backslash \mathrm{Orb}(m - \lfloor r/2 \rfloor)$ следующие условия эквивалентны: $a \in A(S)$ и $n(a, X) = m - u$.*

Лемма 29. *Для произвольного вектора* $a^b \in \mathrm{Orb}(m - \lfloor r/2 \rfloor)$ *и* $S \in [\mathbb{S}]^n$ *справедливо следующее:*

- *если* S *удовлетворяет условию* $|S \cap S^m| \leqslant \lfloor r/2 \rfloor + m - u$, *следующие условия являются эквивалентными:* $a^b \in A(S)$ *и* $n\left(a^b, S\right) = m - u$;
- *если* S *удовлетворяет условию* $|S \cap S^m| > \lfloor r/2 \rfloor + m - u$, *то условие* $a^b \in A(S)$ *эквивалентно следующему:*

$$\begin{cases} n\left(a^b, S \cap S^{N-m}\right) = 0; \\ n\left(a^b, S \cap X^m\right) = |S \cap S^m| - \lfloor r/2 \rfloor. \end{cases}$$

Обратим внимание, что обе полученные в настоящем параграфе оценки не зависят от центра шара a_0 и зависят лишь от номера слоя m, в котором лежит центр шара.

Замечание 23. *Позже в работах А. Фрея было предложено существенно более простое доказательство Теорем 45 и 46, также основанноjt на формуле разложения вероятности переобучения по орбитам выборок. Здесь мы приводим изначальный более длинный вариант доказательства.*

Приведем несколько интересных следствий:

Следствие 18 ([114]). *Вероятность переобучения множества* $\mathbb{A} = \{a\}$, *состоящего из одного вектора* a: $n(a, \mathbb{S}) = m$, *представляется следующим образом:*

$$Q_{\mu,\varepsilon}(\mathbb{A}) = H_N^{n,m}\big(s(\varepsilon)\big).$$

Доказательство. Утверждение следует из Теорем 45 и 46. ■

Следствие 19 ([115]). *Вероятность переобучения множества* $\mathbb{A} = \mathbb{B}_m^N$, *равна:*

$$Q_{\mu,\varepsilon}(\mathbb{A}) = \begin{cases} 1, & \text{при } \varepsilon u \leqslant m \leqslant N - \varepsilon n; \\ 0, & \text{иначе.} \end{cases}$$

Доказательство. Утверждение следует из теорем 45 и 46. ■

Свойства сходства и расслоения: шар векторов. В данном параграфе исследуется совместное влияние свойств сходства и расслоения множества векторов $\mathbb{A}$ на вероятность переобучения. Пусть $\mathbb{A} = B_{r_0}(a_0)$, где $n(a_0, \mathbb{S}) = m$. Пусть, без ограничения общности, вектор a_0 допускает ошибки на первых m парах генеральной выборки $\mathbb{S}$. В дальнейшем множество,

состоящее из первых m пар генеральной выборки, мы будем обозначать S^m. Множество, состоящее из последних $N-m$ пар, будем обозначать S^{N-m}.

Вновь определим симметрические группы Π^m и Π^{N-m}, переставляющие элементы подвыборок S^m и S^{N-m} соответственно.

Лемма 30. *Группа $\Pi^m \times \Pi^{N-m}$ является подгруппой группы симметрии множества векторов $\mathbb{A}$.*

Доказательство. При доказательстве Леммы 23 было установлено, что действие элементов группы $\Pi^m \times \Pi^{N-m}$ на вектора ошибок не меняет расстояние до центра шара a_0, то есть для любых $a \in \mathbb{A}$ и $\pi \in \Pi^m \times \Pi^{N-m}$ справедливо:

$$d(a, a_0) = d\big(\pi(a), a_0\big).$$

Тогда, если $a \in B_{r_0}(a_0)$, то и $\pi(a) \in B_{r_0}(a_0)$. Поскольку действие элементов симметрической группы на векторы ошибок инъективно, приходим к выводу, что $B_{r_0}(a_0) = \pi(B_{r_0}(a_0))$. ■

Лемма 31. *Орбитами действия группы $\Pi^m \times \Pi^{N-m}$ на множестве векторов $\mathbb{A}$ являются пересечения слоев Булева куба $\mathbb{B}_j^N$ для $j = m - r_0, \ldots, m + r_0$ со сферами радиусов $0, 1, \ldots, r_0$ и центрами в векторе a_0.*

Доказательство. Рассмотрим вектор $a \in \mathbb{B}_p^N$, такой что $d(a, a_0) = r_1$. Мы установили, что в этом случае $d\big(\pi(a), a_0\big) = d(a, a_0) = r_1$. Также мы знаем, что действие перестановки на вектор не меняет числа его ошибок на $\mathbb{S}$. Таким образом, $\pi(a)$ также принадлежит пересечению слоя $\mathbb{B}_p^N$ со сферой радиуса r_1.

Осталось показать, что для любых $a_1, a_2 \in \mathbb{B}_p^N$, таких что $d(a_1, a_0) = d(a_2, a_0) = r_1$, найдётся перестановка $\pi \in \Pi^m \times \Pi^{N-m}$, для которой $\pi(a_1) = a_2$. Это уже было доказано в Лемме 24. ■

На Рисунке 5.4 представлено по одному вектору из каждой орбиты семейства A. Пронумеруем орбиты двумя целочисленными индексами следующим образом: векторы из орбиты $\mathrm{Orb}(r, v)$, $r = 0, \ldots, r_0$, $v = 0, \ldots, r$, принадлежат пересечению сферы радиуса r с центром в a_0 со слоем $r + m - 2v$. Обратим внимание на то, что вектор $a_{(r,v)}$ из орбиты $\mathrm{Orb}(r, v)$ имеет $m - v$ единиц и v нулей на множестве S^m и $r - v$ единиц, $N - m - r + v$ нулей на множестве S^{N-m}. Всего в орбите $\mathrm{Orb}(r, v)$ содержится $C_m^v C_{N-m}^{r-v}$ векторов.

Лемма 32. *Для любых $S \in [\mathbb{S}]^n$ и $a \in \mathbb{A} \backslash \mathrm{Orb}(r_0, r_0)$ следующие утверждения эквивалентны: $a \in A(S)$ и $n(a, S) = 0$.*

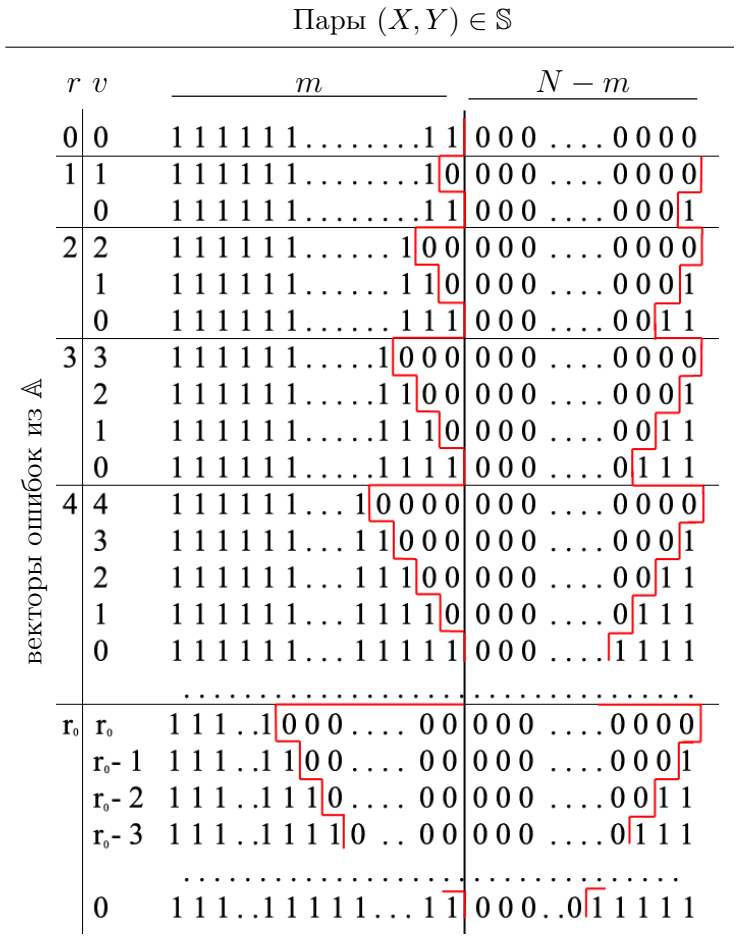


Рисунок 5.4: Векторы из орбит множества $\mathbb{A} = B_{r_0}(a_0)$.

Доказательство. Достаточность очевидна. Докажем необходимость.

Вновь введём обозначения: $S^{n_1} = S^m \cap S$, $S^{n_2} = S^{N-m} \cap S$, $n_1 = |S^{n_1}|$, $n_2 = |S^{n_2}|$.

Пусть $a \in A(S)$ и a принадлежит орбите $\mathrm{Orb}(r, v)$, отличной от $\mathrm{Orb}(r_0, r_0)$. Докажем, что вектор a не допускает ошибок на выборке S.

Начнем с рассмотрения случая $n_1 \leqslant r_0$. Векторы из орбиты $\mathrm{Orb}(n_1, n_1)$ имеют ровно n_1 нулей на множестве S^m и не допускают ни одной ошибки на множестве S^{N-m}. Очевидно, существует $a^{n_1} \in \mathrm{Orb}(n_1, n_1)$, такой что $n\left(a^{n_1}, S\right) = 0$.

В случае $n_1 > r_0$ в орбите $\mathrm{Orb}(r_0, r_0)$ существует вектор a^{r_0}, такой что $n\left(a^{r_0}, S^{N-m}\right) = 0$ и $n\left(a^{r_0}, S^m\right) = n_1 - r_0$. Поскольку $n\left(a, S^m\right) \geqslant n_1 - v \geqslant n_1 - r_0 = n\left(a^{r_0}, S^m\right)$, то мы приходим к противоречию с тем, что $a \in A(S)$. ■

Следствие 20. *Рассмотрим $S \in [\mathbb{S}]^n$. В ходе доказательства прошлой леммы также установлено, что при $|S \cap S^m| > r_0$ векторы из орбит, отличных от $\mathrm{Orb}(r_0, r_0)$, не могут попасть во множество $A(S)$.*

Лемма 33. *Для произвольного вектора $a \in \mathrm{Orb}(r_0, r_0)$ и $S \in [\mathbb{S}]^n$ справедливо следующее:*

- *если* $S \in [\mathbb{S}]^n$ *удовлетворяет условию* $|S \cap S^m| \leqslant r_0$, *то следующие утверждения эквивалентны:* $a \in A(S)$ *и* $n(a, S) = 0$.

- *если* $S \in [\mathbb{S}]^n$ *удовлетворяет условию* $|S \cap S^m| > r_0$, *то утверждение* $a \in A(S)$ *эквивалентно следующему:*

$$n\left(a, S \cap S^m\right) = |S \cap S^m| - r_0.$$

Доказательство. Случай $|S \cap S^m| \leqslant r_0$ повторяет первую часть доказательства Леммы 32. Рассмотрим случай $|S \cap S^m| > r_0$.

Поскольку в этом случае ни один вектор из множества $\mathbb{A}$ не допускает меньше $|S \cap S^m| - r_0$ ошибок на выборке S, а векторы из $\mathrm{Orb}(r_0, r_0)$ не ошибаются на парах множества S^{N-m}, то достаточность очевидна. Необходимость вытекает из того факта, что существует вектор $\exists a \in \mathrm{Orb}(r_0, r_0)$, такой что $n(a, S) = n(a, S^m) = |S \cap S^m| - r_0$. ■

Теорема 47 (Точная оценка для шара)**.** *Пусть* $n(a_0, \mathbb{S}) = m$. *Рассмотрим шар* $\mathbb{A} = B_{r_0}(a_0)$, *такой что* $r \leqslant \min(m, N - m)$. *Тогда для РМЭР вероятность переобучения может быть записана в виде:*

$$Q_{\mu,\varepsilon}(\mathbb{A}) = \sum_{i=\max(0,m-u)}^{r_0} h_N^{n,m}(i) \frac{\sum\limits_{r_1=0}^{r_0} \sum\limits_{v_1=0}^{r_1} \Delta(v_1, r_1, i) \mathbb{1}_{\{m+r_1-2v_1 \geqslant \varepsilon u\}}}{\sum\limits_{r_2=0}^{r_0} \sum\limits_{v_2=0}^{r_2} \Delta(v_2, r_2, i)} + \sum_{i=r_0+1}^{\lfloor s'(\varepsilon) \rfloor} h_N^{n,m}(i),$$

где $h_N^{n,m}(i) = \frac{C_m^i C_{N-m}^{n-i}}{C_N^n}$, $\Delta\left(v, r, i\right) = C_{m-i}^{v-i} C_{u-m+i}^{r-v}$, $s'(\varepsilon) = \frac{n}{N}(m - \varepsilon u) + \frac{r_0 u}{N}$.

Доказательство. Воспользуемся формулой (5.3) с учетом Лемм 30, 31 и 27:

$$Q_{\mu,\varepsilon}(\mathbb{A}) =$$
$$= \sum_{r=0}^{r_0} \sum_{v=0}^{r} \frac{C_m^v C_{N-m}^{r-v}}{C_N^n} \sum_{i=\max(0,m-u)}^{\min(n,m)} \underbrace{\left|\{S \in [\mathbb{S}]^n \colon |S \cap S^m| = i,\ a_{(r,v)} \in A(S)\}\right|}_{S(i,r,n)} \frac{\mathbb{1}_{\{L(a_{(r,v)}, \mathbb{S} \setminus S_i) - L(a_{(r,v)}, S_i) \geqslant \varepsilon\}}}{|A(S_i)|},$$

где $a(r, v)$ — векторы, представленные на Рисунке 5.4, а $S_i \in [\mathbb{S}]^n$ — произвольная подвыборка, в которую входят ровно i пар из S^m.

Разобьем суммирование по орбитам разбиений (по i) на два слагаемых:

$$Q_{\mu,\varepsilon}(\mathbb{A}) = \sum_{r=0}^{r_0} \sum_{v=0}^{r} \frac{C_m^v C_{N-m}^{r-v}}{C_N^n} \sum_{i=\max(0,m-u)}^{r_0} S\left(i, r, v\right) \frac{\mathbb{1}_{\{L(a_{(r,v)}, \mathbb{S} \setminus S_i) - L(a_{(r,v)}, S_i) \geqslant \varepsilon\}}}{|A(S_i)|}$$
$$+ \sum_{r=0}^{r_0} \sum_{v=0}^{r} \frac{C_m^v C_{N-m}^{r-v}}{C_N^n} \sum_{i=r_0+1}^{\min(n,m)} S\left(i, r, v\right) \frac{\mathbb{1}_{\{L(a_{(r,v)}, \mathbb{S} \setminus S_i) - L(a_{(r,v)}, S_i) \geqslant \varepsilon\}}}{|A(X_i)|}.$$

Из Лемм 32 и 33, а также из Следствия 20 следует, что первое слагаемое соответствует случаю выбора вектора, не допускающего ошибок на выборке S. Следствие 20 позволяет

опустить суммирование по орбитам во втором слагаемом, поскольку при $i > r_0$ во множество $A(S_i)$ попадают только векторы из $\mathrm{Orb}(r_0, r_0)$, допускающие в соответствии с Леммой 33 ровно $i - r_0$ ошибок на S. С учетом этого:

$$Q_{\mu,\varepsilon}(\mathbb{A}) = \sum_{r=0}^{r_0}\sum_{v=0}^{r}\frac{C_m^v C_{N-m}^{r-v}}{C_N^n}\sum_{i=\max(0,m-u)}^{r_0} S\left(i,r,v\right)\frac{\mathbb{1}_{\{r+m-2n\geqslant\varepsilon u\}}}{|A(S_i)|} + \frac{C_m^{r_0}C_{N-m}^0}{C_N^n}\sum_{i=r_0+1}^{\min(n,m)} S\left(i,r,v\right)\frac{\mathbb{1}_{\{i\leqslant s(\varepsilon)+\frac{r_0 u}{N}\}}}{|A(S_i)|}. \tag{5.8}$$

Вычислим значение $S\left(i,r,v\right)$. В случае $i \leqslant r_0$ — это число способов выбрать i из v объектов множества S^m и $n-i$ из $N-m-r+v$ объектов множества S^{N-m}, на которых не ошибается вектор $a_{(r,v)}$. В случае $i > r_0$— число способов выбрать $i - r_0$ объектов из множества S^m, на которых вектор $a_{(r,v)}$ ошибается, и $n-i$ произвольных объектов из S^{N-m}. Итого:

$$S\left(i,r,v\right) = \begin{cases} C_v^i C_{N-m-r+v}^{n-i}, & \text{если } i \leqslant r_0; \\ C_{m-r_0}^{i-r_0} C_{N-m}^{n-i}, & \text{если } i > r_0. \end{cases}$$

Найдём значения $|A(S_i)|$. В случае $i > r_0$ в $A(S_i)$ попадают те векторы из $\mathrm{Orb}(r_0, r_0)$, которые ошибаются на множестве S^m $i - r_0$ раз. При $i \leqslant r_0$ из каждой орбиты во множество $A(S_i)$ попадают векторы, не ошибающиеся ни на одной паре из множеств S^m и S^{N-m}. Получаем:

$$|A(S_i)| = \begin{cases} \sum\limits_{r_1=0}^{r_0}\sum\limits_{v_1=0}^{r_1} C_{m-i}^{v_1-i} C_{u-m+i}^{r_1-v_1}, & \text{если } i \leqslant r_0; \\ C_i^{r_0}, & \text{если } i > r_0. \end{cases}$$

Подстановка полученных результатов в (5.8) дает:

$$Q_{\mu,\varepsilon}(\mathbb{A}) = \sum_{r=0}^{r_0}\sum_{v=0}^{r}\frac{C_m^v C_{N-m}^{r-v}}{C_N^n}\sum_{i=\max(0,m-u)}^{r_0} C_v^i C_{N-m-r+v}^{n-i}\frac{\mathbb{1}_{\{r+m-2v\geqslant\varepsilon u\}}}{\sum\limits_{r_1=0}^{r_0}\sum\limits_{v_1=0}^{r_1} C_{m-i}^{v_1-i} C_{u-m+i}^{r_1-v_1}} + \frac{C_m^{r_0}}{C_N^n}\sum_{i=r_0+1}^{\min(n,m)} C_{m-r_0}^{i-r_0} C_{N-m}^{n-i}\frac{\mathbb{1}_{\{i\leqslant s(\varepsilon)+\frac{r_0 u}{N}\}}}{C_i^{r_0}}. \tag{5.9}$$

Во втором слагаемом:

$$\frac{C_{m-r_0}^{i-r_0}}{C_i^{r_0}} = \frac{C_{m-r_0}^{m-i}}{C_i^{r_0}} = \frac{C_{i-(i-m+r_0)}^{r_0-(i-m+r_0)}}{C_i^{r_0}} = \frac{C_{r_0}^{i-m+r_0}}{C_i^{i-m+r_0}} = \frac{C_{r_0}^{m-i}}{C_i^{m-r_0}}.$$

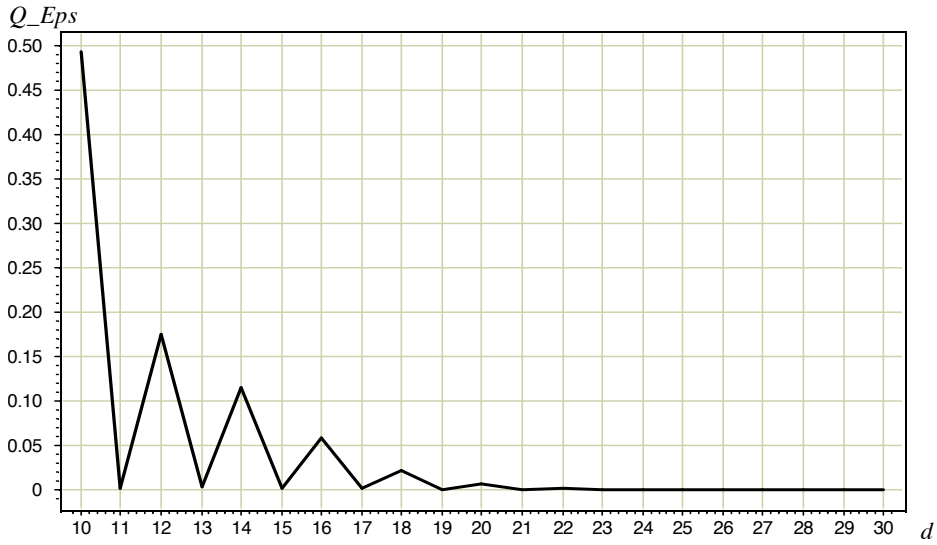


Рисунок 5.5: Зависимость вкладов слоёв шара $\mathbb{A}_d$ при $\mathbb{A} = B_{r_0}(a_0)$ в вероятность переобучения Q_ε от числа ошибок d при $n = u = 100,\ n(a_0, \mathbb{S}) = m = 20,\ r_0 = 10,\ \varepsilon = 0.05$.

Далее, $C_m^{r_0} C_{r_0}^{m-i} = C_m^i C_i^{r_0-m+i} = C_m^i C_i^{m-r_0}$. Подстановка полученных формул во второе слагаемое (5.9) дает:

$$\frac{C_m^{r_0}}{C_N^n} \sum_{i=r_0+1}^{\min(n,m)} C_{m-r_0}^{i-r_0} C_{N-m}^{n-i} \frac{\mathbb{1}_{\{i \leqslant s(\varepsilon)+\frac{r_0 u}{N}\}}}{C_i^{r_0}}$$
$$= \sum_{i=r_0+1}^{\min(n,m)} h_N^{n,m}(i) \mathbb{1}_{\{i \leqslant s(\varepsilon)+\frac{r_0 u}{N}\}} = \sum_{i=r_0+1}^{\lfloor s'(\varepsilon) \rfloor} h_N^{n,m}(i). \tag{5.10}$$

В первом слагаемом $C_m^v C_v^i = C_m^i C_{m-i}^{v-i}$, а $C_{N-m}^{r-v} C_{N-m-r+v}^{n-i} = C_{N-m}^{N-m-r+v} C_{N-m-r+v}^{n-i} = C_{N-m}^{n-i} C_{u-m+i}^{r-v}$. Заменив порядок суммирования, получим:

$$\sum_{r=0}^{r_0} \sum_{v=0}^{r} \frac{C_m^v C_{N-m}^{r-v}}{C_N^n} \sum_{i=\max(0,m-u)}^{r_0} C_v^i C_{N-m-r+v}^{n-i} \frac{\mathbb{1}_{\{r+m-2v \geqslant \varepsilon u\}}}{\sum\limits_{r_1=0}^{r_0} \sum\limits_{v_1=0}^{r_1} C_{m-i}^{v_1-i} C_{u-m+i}^{r_1-v_1}}$$
$$= \sum_{i=\max(0,m-u)}^{r_0} h_N^{n,m}(i) \frac{\sum\limits_{r=0}^{r_0} \sum\limits_{v=0}^{r} C_{m-i}^{v-i} C_{k-m+i}^{r-v} \mathbb{1}_{\{r+m-2v \geqslant \varepsilon u\}}}{\sum\limits_{r_1=0}^{r_0} \sum\limits_{v_1=0}^{r_1} C_{m-i}^{v_1-i} C_{u-m+i}^{r_1-v_1}} \tag{5.11}$$

Подстановка (5.10) и (5.11) в (5.9) завершает доказательство. ■

Обратим внимание на то, что полученная оценка снова не зависят от центра шара a_0, а зависит лишь от номера слоя m, в котором лежит центр шара.

Следствие 21. *Пусть $n(a_1, \mathbb{S}) = n(a_2, \mathbb{S}) = m$ и $u \leqslant n$. Тогда для $\varepsilon \in \left[0; \frac{m-r}{u}\right]$ вероятность переобучения множества $\mathbb{A} = B_r(a_1)$ не превосходит вероятности переобучения множества $\mathbb{A} = B_r(a_2) \cap \mathbb{B}_m^N$.*

Доказательство. Доказательство напрямую следует из Теорем 45, 46 и 47:

$$
\begin{aligned}
Q_{\mu,\varepsilon}\big(B_r(a_1)\big) &= \sum_{i=\max(0,m-u)}^{r_0} h_N^{n,m}(i)\, \frac{\sum\limits_{r_1=0}^{r}\sum\limits_{v_1=0}^{r_1} \Delta\left(v_1, r_1, i\right) \mathbb{1}_{\{m+r_1-2v_1 \geqslant \varepsilon u\}}}{\sum\limits_{r_2=0}^{r}\sum\limits_{v_2=0}^{r_2} \Delta\left(v_2, r_2, i\right)} + \sum_{i=r+1}^{\lfloor s'(\varepsilon)\rfloor} h_N^{n,m}(i) \\
&= \sum_{i=\max(0,m-u)}^{r_0} h_N^{n,m}(i) + \sum_{i=r+1}^{\lfloor s'(\varepsilon)\rfloor} h_N^{n,m}(i) \\
&= H_N^{n,m}\left(s(\varepsilon) + \frac{ru}{N}\right) \\
&\leqslant H_N^{n,m}\left(s(\varepsilon) + r/2\right) \\
&= Q_{\mu,\varepsilon}\big(B_r(a_2) \cap \mathbb{B}_m^N\big).
\end{aligned}
$$

■

Замечание 24. *В прошлом доказательстве нам удалось получить гипергеометрическую функцию распределения из двух слагаемых именно благодаря тому, что* $\varepsilon \in \left[0; \frac{m-r}{u}\right]$. *При этих* ε *значение* $\lfloor s'(\varepsilon)\rfloor \geqslant r$, *что и дает возможность «слить» две суммы в одну. В противном случае* $\lfloor s'(\varepsilon)\rfloor$ *строго меньше* r, *и значение гипергеометрической функции распределения следует брать не в точке* $s'(\varepsilon)$, *а в точке* r. *Также интересно отметить, что при* $\varepsilon \in \left[0; \frac{m-r}{u}\right]$ *первое слагаемое в оценке для шара в точности равняется* $H_N^{n,m}(r)$.

На Рисунке 5.5 представлены точные значения вкладов слоев шара в его вероятность переобучения. Видно, что несколько нижних слоев шара дают бóльшую часть вероятности переобучения. Возникает вопрос: нельзя ли приближать оценку шара оценкой для t его нижних слоев?

Замечание 25. *Поясним, почему график на Рисунке 5.5 имеет вид «гармошки». Понимание эффектов сходства и расслоения позволяют выдвинуть следующую гипотезу:* вероятность переобучения расслоенного и связного множества векторов может быть аппроксимирована вероятностью переобучения его подмножества, состоящего из существенно различных векторов нижних слоев. *Из рисунка 5.4 видно, что на внешней сфере шара представлены вектора не из всех слоев: слои чередуются через один. Поскольку «существенно различные» вектора лежат именно на внешней сфере, бóльшие вклады в вероятность переобучения дают те слои шара, которые имеют непустое пересечение с его внешней сферой, а именно — слои с количеством ошибок* $m - r_0, m - r_0 + 2, \ldots$.

Нижние слои шара векторов. Наконец, рассмотрим множество $\mathbb{A}$, состоящее из нескольких нижних слоев шара $B_{r_0}(a_0)$.

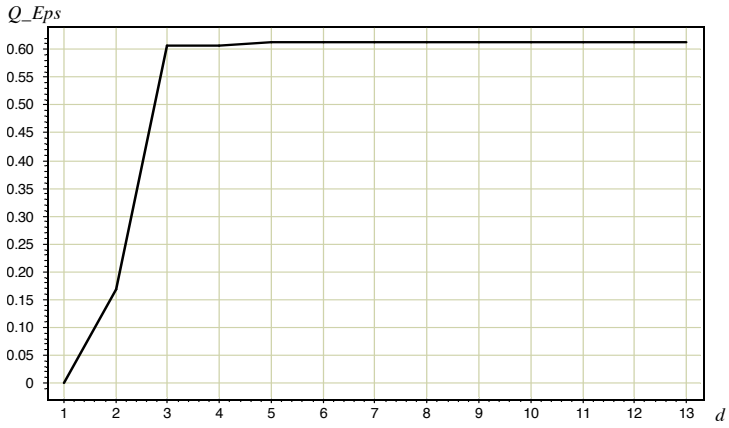


Рисунок 5.6: Зависимость вероятности переобучения Q_ε множества $\mathbb{A} = B_{r_0}(a_0, d)$ от числа d нижних слоев шара при $u = n = 100,\ L(a_0, \mathbb{S}) = m = 10,\ r_0 = 6,\ \varepsilon = 0.05$.

Теорема 48 (Точная оценка для нижних слоев шара)**.** *Пусть* $n(a_0, \mathbb{X}) = m$*. Рассмотрим* $d \geqslant 0$ *нижних слоев шара:* $\mathbb{A} = B_{r_0}(a_0, d)$*. Пусть, кроме того, выполнено* $r \leqslant \min(m, N - m)$*. Тогда для РМЭР вероятность переобучения может быть записана в виде:*

$$Q_{\mu,\varepsilon}(\mathbb{A}) =$$

$$= \sum_{i=\max(0,m-u)}^{r_0} h_N^{n,m}(i) \frac{\sum\limits_{r_1=0}^{r_0} \sum\limits_{v_1=0}^{r_1} \Delta'(v_1, r_1, i)\, \mathbb{1}_{\{m+r_1-2v_1 \geqslant \varepsilon u\}}}{\sum\limits_{r_2=0}^{r_0} \sum\limits_{v_2=0}^{r_2} \Delta'(v_2, r_2, i)} + \mathbb{1}_{\{d \geqslant 1\}} \sum_{i=r_0+1}^{\lfloor s'(\varepsilon) \rfloor} h_N^{n,m}(i),$$

где $h_N^{n,m}(i) = \frac{C_m^i C_{N-m}^{n-i}}{C_N^n}$, $\Delta'(v, r, i) = C_{m-i}^{v-i} C_{u-m+i}^{r-v} \mathbb{1}_{\{m+r-2v \leqslant m-r_0+d-1\}}$, $s'(\varepsilon) = \frac{n}{N}(m - \varepsilon u) + \frac{r_0 u}{N}$.

Доказательство. Доказательство аналогично Теореме 47. Леммы 30, 32 и 33 остаются справедливы и для этого множества векторов. В Лемме 31 множество слоев, в пересечении с которыми сферы дают орбиты векторов, изменяется на $m - r_0 \ldots m - r_0 + d - 1$.

Основное отличие в ходе доказательства — при использовании формулы (5.3) в начале доказательства множество суммируемых орбит векторов сокращается добавлением индикаторного множителя $\mathbb{1}_{\{m+r-2v \leqslant m-r_0+d-1\}}$ после знаков суммирования по индексам r и v. ■

На Рисунке 5.6 представлена зависимость точной оценки вероятности переобучения для d нижних слоев шара от параметра d. Мы снова видим, что существенные скачки происходят лишь на первых нескольких слоях.

На Рисунке 5.7 представлены результаты приближения оценки шара d его нижними слоями. Черным цветом изображена оценка шара. Красным, зеленым и синим — оценки 1, 2 и 3 его нижних слоев соответственно. Падение к нулю оценок первых слоев шара объясняется уменьшением числа ошибок, допускаемых векторами из нижнего слоя шара, с ростом

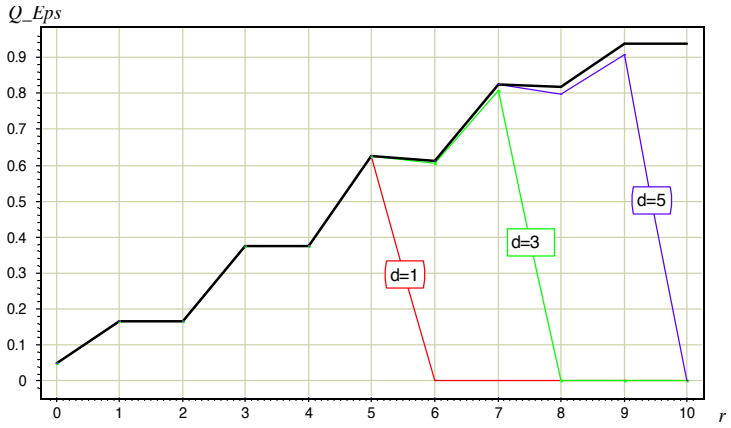


Рисунок 5.7: Зависимость вероятности переобучения Q_ε для $\mathbb{A} = B_r(a_0)$ (верхняя кривая) и $\mathbb{A} = B_r(a_0, d)$ от радиуса шара r, при $u = n = 100,\ \ L(a_0, \mathbb{S}) = m = 10,\ \ \varepsilon = 0.05$.

радиуса шара. В определенный момент количество ошибок, допускаемое векторами нижнего слоя шара, становится меньше εu. В этом случае, очевидно, переобучение невозможно.

Выводы. Итак, в последних параграфах были исследованы множества $\mathbb{A}$, представляющие из себя шар в Булевом кубе $\mathbb{B}^N$, центральный слой этого шара и сечение шара несколькими его нижними слоями. На примере центрального слоя шара мы еще раз продемонстрировали влияние сходств между векторами ошибок множества $\mathbb{A}$ на вероятность переобучения семейства. Как мы видели на Рисунке 5.2, вероятность переобучения множества, состоящего из случайных векторов с фиксированным числом ошибок на генеральнйо выборке, чрезвычайно быстро растёт к единице с ростом его мощности. Вероятность же переобучения центрального слоя шара остается на достаточно низком уровне даже при мощности более десяток тысяч.

Также на примере шара и его нижних слоев удалось показать возможность приближения вероятности переобучения множеств $\mathbb{A}$, расслоенных по уровню ошибок $L(a, \mathbb{S})$, несколькими их нижними слоями $\mathbb{A}_d$.

Коротко подведем итоги настоящей главы:

- Комбинаторная теория переобучения отказывается от рассмотрения чересчур общих постановок задач в пользу детального учета свойств конкретных решаемых задач.
- В частности, комбинаторная теория отказывается от использования часто завышенных неравенств концентрации.

- Все результаты комбинаторной теории переобучения основаны на работе с конечным множеством бинарных векторов, являющихся векторами потерь рассматриваемого класса отображений.

- При получении оценок вероятности переобучения необходимо учитывать два геометрических свойства множества бинарных векторов: расслоения векторов по числу ошибок на генеральной выборке и «структуры» множества векторов, выражающейся в наличии большого числа похожих между собой векторов ошибок. Отказ от учета любого из описанных эффектов ведет к существенно завышенным оценкам.

- Теоретико-групповой подход позволяет облегчить вычисление вероятности переобучения для множеств бинарных векторов, наделенных существенными симметриями.

В настоящей главе получены следующие новые результаты:

- В теоретико-групповом подходе предложено рассматривать орбиты разбиений генеральной выборки (Леммы 21 и 22), что ведет к формуле разложения вероятности переобучения по орбитам разбиений (Теорема 44).

- На основе этих результатов получены новые точные (не завышенные) оценки вероятности переобучения для трех модельных семейств (Теоремы 45, 47 и 48), являющихся различными подмножествами шара в Булевом кубе: центральный слой шара как пример семейства без расслоения, имеющего наибольшую степень похожести векторов между собой; шар как пример расслоенного семейства, имеющего наибольшую степень похожести векторов между собой; нижние слои шара, как модель для изучения возможности приближения вероятности переобучения расслоенного семейства, обладающего свойствами похожести, векторами из его нижних слоев.

6 PAC-Байесовский анализ

PAC-Байесовский анализ — общий и действенный подход к анализу алгоритмов машинного обучения, основанный на эмпирических данных. К настоящему моменту он был применен в таких разнообразных задачах, как обучение с учителем [55, 57, 72, 86], обучение без учителя [42, 86] и обучение с подкреплением [87]. PAC-Байесовский анализ совмещает в себе преимущества PAC обучения[1] и Байесовских подходов к обучению и позволяет: (1) получать строгие оценки обобщающей способности (как теория Вапника–Червоненкиса), (2) учитывать априорные знания (как Байесовские подходы) и (3) получать оценки обобщающей способности, основанные на данных (как результаты, основанные на Радемахеровских сложностях).

PAC-Байесовский анализ позволяет получить неравенства концентрации для разности среднего и эмпирического риска рандомизированных отображений. Рандомизированное отображение определяется множеством отображений $\mathcal{H}$ и распределением ρ на нем. Для получения ответа на каждом новом объекте $X \in \mathcal{X}$ рандомизированное отображение выбирает отображение $h \in \mathcal{H}$ из распределения ρ $\mathcal{H}$ и возвращает ответ $h(X)$. Если ρ — дельта-распределение, то рандомизированное отображение совпадает к обыкновенным отображением $h \in \mathcal{H}$. В противном случае работа рандомизированного отображения схожа с Байесовским прогнозированием, основанным на апостериорном[2] распределении, с той только разницей, что используемое апостериорное распределение не обязано быть Байесовским. Важно отметить, что многие PAC-Байесовские неравенства выполняются *одновременно для всех апостериорных распределений* ρ (с высокой вероятностью относительно случайной реализации обучающей выборки). По этой причине они могут использоваться двумя способами. С одной стороны, минимизируя PAC-Байесовские оценки по распределению ρ, мы можем получать новые методы обучения с гарантированной обобщающей способностью. С другой, пользу-

[1] PAC — аббревиатура от Probably Approximately Correct. Этим термином принято называть подход в теории статистического обучения, предложенный в работе [99] и во многом похожий на подход Вапника–Червоненкиса, ранее рассмотренный в Главе 3.

[2] Всюду далее словом *апостериорное* мы будем подчеркивать зависимость распределения от обучающей выборки.

ясь PAC-Байесовскими неравенствами, мы можем оценивать средний риск апостериорных распределений ρ, которые были получены другими методами обучения: минимизацией эмпирического риска, минимизацией эмпирического риска с регуляризацией, Байесовскими методами и так далее. В подобных случаях PAC-Байесовский подход может использоваться для получения оценок обобщающей способности различных методов обучения, а также может служить заменой процедуры скользящего контроля при настройке гиперпараметров. Поскольку PAC-Байесовские неравенства выполнены одновременно для всех апостериорных распределений ρ, мы можем применять их для оценки сразу большого числа распределений ρ, не подвергаясь переобучению, как в случае множественного использования процедуры скользящего контроля.

В Разделе 6.2 мы приведем обзор известных результатов PAC-Байесовского анализа и их подробное сравнение между собой. В Разделе 6.3 мы получим новое *PAC-Байесовское эмпирическое неравенства Бернштейна* [95]. На примере теоретических рассуждений и экспериментов (с использованием модельных данных и реальных выборок из репозитория UCI [3]) мы продемонстрируем, что новое неравенство в ряде интересных случаев ведет к существенно более точным оценкам по сравнению с известными ранее результатами.

6.1 Определения и постановка задачи

В отличие от общепринятого способа изложения, мы сформулируем задачу PAC-Байесовского анализа в достаточно общем виде, не ограничиваясь рассмотрением теории статистического обучения и оценки среднего риска отображений.

Постановка задачи. Рассмотрим семейство выборок случайных величин $\{X_1^h, \ldots, X_n^h\}$, где выборки индексированы некоторым (возможно несчетным) множеством: $h \in \mathcal{H}$. Мы будем предполагать, что внутри одной выборки $\{X_1^h, \ldots, X_n^h\}$ случайные величины независимы, но случайные величины из разных выборок могут быть зависимыми. Кроме того, предположим, что все случайные величины ограничены: $X_i^h \in [0, 1]$ для всех $i = 1, \ldots, n$ и $h \in \mathcal{H}$. Обозначим множество всевозможных вероятностных распределений ρ на индексном множестве $\mathcal{H}$ с помощью $\mathcal{P}(\mathcal{H})$.

Основная задача PAC-Байесовского анализа заключается в получении верхних оценок вида

$$\mathbb{E}_{h\sim\rho}\big[\mathbb{E}[X_1^h]\big] \leqslant \mathbb{E}_{h\sim\rho}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^h\right] + C(n, \rho, \mathcal{H})$$

на средние (по мере $\rho \in \mathcal{P}(\mathcal{H})$) значения математических ожиданий выборок, справедливые с большой вероятностью (относительно реализации случайных выборок $\{X_i^h\}_{i=1}^n, h \in \mathcal{H}$) *одновременно* для всех вероятностных распределений $\rho \in \mathcal{P}(\mathcal{H})$. Нас, конечно, будут интересовать такие результаты, где остаточное слагаемое $C(n, \rho, \mathcal{H})$ неотрицательно и убывает к нулю с ростом размера выборок $n \to \infty$.

Статистическое обучение и рандомизированные алгоритмы. Проиллюстрируем поставленную выше задачу на примере приложений в теории статистического обучения. Мы будем пользоваться обозначениями, введенными в Главе 3. Положим в качестве индексного множества $\mathcal{H}$ используемый в задаче минимизации среднего риска класс отображений $\mathcal{H} = \{h\colon \mathcal{X} \to \mathcal{Y}\}$. Для $h \in \mathcal{H}$ в качестве выборки случайных величин $\{X_1^h, \ldots, X_n^h\}$ возьмем потери отображения h на обучающей выборке: $\{\ell_h(X_1, Y_1), \ldots, \ell_h(X_n, Y_n)\}$. Напомним, что для $h \in \mathcal{H}$ и $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ мы используем короткое обозначение $\ell_h(X, Y) = \ell\big(h(X), Y\big)$ для некоторой ограниченной в $[0, 1]$ функции потерь ℓ. Очевидно, при этом выполнены все условия в постановке задачи PAC-Байесовского анализа с параметрами $a = 0, b = 1$.

Теперь рассмотрим произвольное вероятностное распределение $\rho \in \mathcal{P}(\mathcal{H})$ на множестве отображений $\mathcal{H}$. Если отображения из $h = h(\alpha) \in \mathcal{H}$ задаются своими параметрами α (например, $\alpha \in \mathbb{R}^d$), то можно перейти к рассмотрению вероятностного распределения ρ на пространстве этих параметров. Так, рассматривая задачу классификации и множество всех гиперплоскостей в пространстве $\mathbb{R}^d$, мы можем задавать распределение ρ на d-мерном вещественном пространстве всевозможных векторов нормалей, задающих гиперплоскости.

Распределению ρ мы будем ставить в соответствие *рандомизированный алгоритм* прогнозирования, который мы будем обозначать G_ρ (аббревиатура от общепринятого в литературе названия *отображение Гиббса*). Для получения ответа на новом объекте X рандомизированный алгоритм выбирает случайное отображение из распределения $h \sim \rho$ (независимо от объекта X) и возвращает ответ $h(X)$. Средний и эмпирический риск рандомизированного отображения G_ρ мы будем определять как $L(G_\rho) = \mathbb{E}_{h\sim\rho}[L(h)]$ и $L_n(G_\rho) = \mathbb{E}_{h\sim\rho}[L_n(h)]$ соответственно.

Тогда задача PAC-Байесовского анализа, сформулированная выше, сводится к получению оценок средних рисков рандомизированных отображений G_ρ на основе их эмпирических рисков:

$$L(G_\rho) \leqslant L_n(G_\rho) + C(n, \rho, \mathcal{H}),$$

справедливых с большой вероятностью *одновременно для всех рандомизированных отображений* G_ρ.

6.2 Обзор известных результатов

В настоящем разделе мы продемонстрируем основные идеи и технические средства, использующиеся в PAC-Байесовском анализе, на примере PAC-Байесовских неравенств Хефдинга [73,74] (также известного, как неравенство МакАллистера) и Бернштейна [88], а также PAC-Байесовского kl-неравенства [69,85]. Для наглядности и простоты изложения все результаты, описываемые в настоящем разделе, мы будем сопоставлять с приведенными ранее в Главе 1 неравенствами концентрации.

6.2.1 PAC-Байесовская лемма

Мы начнем с описания общей процедуры получения PAC-Байесовских неравенств. Ключевым ингредиентом во всех доказательствах будет следующая лемма, часто называемая в литературе *леммой замены распределений*:

Лемма 34 (Донскер и Варадхан, [32])**.** *Для любого измеримого отображения $f\colon \mathcal{H} \to \mathbb{R}$ и любых двух вероятностных распределений ρ и π на $\mathcal{H}$ справедливо:*

$$\mathbb{E}_{h\sim\rho}[f(h)] \leqslant \mathrm{KL}(\rho\|\pi) + \ln \mathbb{E}_{h\sim\pi}\left[e^{f(h)}\right],$$

где $\mathrm{KL}(\rho,\pi) = \mathbb{E}_{h\sim\rho}\left[\log \frac{\rho(h)}{\pi(h)}\right]$ *— дивергенция Кульбака-Лейблера (относительная энтропия).*

Обратим внимание, что естественным требованием на рассматриваемые распределения ρ и π в свете последнего результата является условие *вложенности носителя распределения ρ в носитель распределения π*.

Доказательство. Справедлива следующая цепочка неравенств:

$$\begin{aligned}
\mathbb{E}_{h\sim\rho}[f(h)] &= \mathbb{E}_{h\sim\rho}\left[\ln\left(\frac{\rho(h)}{\pi(h)}e^{f(h)}\frac{\pi(h)}{\rho(h)}\right)\right] \\
&= \mathrm{KL}(\rho\|\pi) + \mathbb{E}_{h\sim\rho}\left[\ln\left(e^{f(h)}\frac{\pi(h)}{\rho(h)}\right)\right] \\
&\leqslant \mathrm{KL}(\rho\|\pi) + \ln \mathbb{E}_{h\sim\rho}\left[e^{f(h)}\frac{\pi(h)}{\rho(h)}\right] \\
&= \mathrm{KL}(\rho\|\pi) + \ln \mathbb{E}_{h\sim\pi}\left[e^{f(h)}\right],
\end{aligned}$$

где мы воспользовались вогнутостью логарифма и неравенством Йенсена. ■

Замечание 26. *Предыдущий результат известен в теории информации как вариационное определение относительной энтропии Донскера–Варадхана, так как он является одним из*

следствий следующего более общего факта, показанного в работе [32]:

$$\mathrm{KL}(\rho\|\pi) = \sup_f \left(\mathbb{E}_{h\sim\rho}[f(h)] + \log \mathbb{E}_{h\sim\pi}\left[e^{f(h)}\right]\right),$$

где супремум берется по всевозможным измеримым отображениям $f\colon \mathcal{H} \to \mathbb{R}$.

Рассмотрим теперь выборку случайных величин $S_m = \{\xi_1, \ldots, \xi_m\}$, принимающих значения в некотором множестве $\mathcal{S}$, и измеримое отображение $f_m\colon \mathcal{H} \times \mathcal{S}^m \to \mathbb{R}$. Лемма 6.2.1, очевидно, выполняется и в этом случае. Мы получаем, что для любых двух вероятностных распределений ρ и π на $\mathcal{H}$ следующее выполнено с вероятностью 1 (относительно реализации выборки S_m):

$$\mathbb{E}_{h\sim\rho}[f(h, S_m)] \leqslant \mathrm{KL}(\rho\|\pi) + \ln \mathbb{E}_{h\sim\pi}\left[e^{f(h,S_m)}\right]. \tag{6.1}$$

Напомним, что неравенство Маркова утверждает, что для любой неотрицательной случайной величины η и $t > 0$ справедливо $\mathbb{P}\{\eta \geqslant t\} \leqslant \mathbb{E}[\eta]/t$. Это эквивалентно утверждению, что для любого $\delta > 0$ с вероятностью не меньше $1 - \delta$ выполнено $\eta \leqslant \mathbb{E}[\eta]/\delta$. Обозначим совместное распределение случайных величин в S_m с помощью $\mathcal{D}$. В случае, если распределение π не зависит от выборки S_m, мы можем воспользоваться следующей цепочкой неравенств:

$$\begin{aligned}
\mathbb{E}_{h\sim\rho}[f_n(h, S_m)] &\leqslant \mathrm{KL}(\rho\|\pi) + \ln \mathbb{E}_{h\sim\pi}\left[e^{f_n(h,S_m)}\right] \\
&\leqslant \mathrm{KL}(\rho\|\pi) + \ln\left(\frac{1}{\delta}\mathbb{E}_{S_m\sim\mathcal{D}}\left[\mathbb{E}_{h\sim\pi}\left[e^{f_n(h,S_m)}\right]\right]\right) \\
&= \mathrm{KL}(\rho\|\pi) + \ln\left(\frac{1}{\delta}\mathbb{E}_{h\sim\pi}\left[\mathbb{E}_{S_m\sim\mathcal{D}}\left[e^{f_n(h,S_m)}\right]\right]\right),
\end{aligned}$$

справедливых с вероятностью не меньше $1 - \delta$ *одновременно для всех распределений* ρ, где мы воспользовались неравенством (6.1) и неравенством Маркова. Подчеркнем, что распределение ρ в последней цепочке неравенств может зависеть от выборки S_m. Мы получили следующую основную лемму PAC-Байесовского анализа:

Лемма 35 (Основная PAC-Байесовская лемма)**.** *Пусть выборка случайных величин* $S_m = \{\xi_1, \ldots, \xi_m\} \subset \mathcal{S}$ *имеет совместное распределение* $\mathcal{D}$*. Для любого измеримого отображения* $f_m\colon \mathcal{H} \times \mathcal{S}^m \to \mathbb{R}$*, любого не зависящего от* S_m *распределения* π *на* $\mathcal{H}$ *и любого* $\delta > 0$ *с вероятностью не меньше* $1 - \delta$ *(относительно случайной реализации* S_m*) следующее*

$$\mathbb{E}_{h\sim\rho}[f_m(h, S_m)] \leqslant \mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta} + \ln\left(\mathbb{E}_{h\sim\pi}\left[\mathbb{E}_{S_m\sim\mathcal{D}}\left[e^{f_m(h,S_m)}\right]\right]\right)$$

справедливо одновременно для всех распределений ρ *на* $\mathcal{H}$*.*

Все PAC-Байесовские неравенства, приводимые далее, основаны на последней лемме. Идея доказательств заключается в выборе «подходящего» отображения f_m, чтобы в левой части неравенства леммы возникало выражение, которое мы хотим оценить сверху. Затем нам

остается получить верхнюю оценку на производящую функцию моментов $\mathbb{E}_{S_m\sim\mathcal{D}}\left[e^{f_m(h,S_m)}\right]$. Отметим сходство описанного подхода с методом Чернова, описанным в Разделе 1.1. Как и раньше, основная сложность при выводе PAC-Байесовских неравенств будет заключаться именно в получении верхней оценки на $\mathbb{E}_{S_m\sim\mathcal{D}}\left[e^{f_m(h,S_m)}\right]$.

6.2.2 Основные PAC-Байесовские неравенства

PAC-Байесовское неравенство Хефдинга. Мы начнем с рассмотрения наиболее простого PAC-Байесовского неравенства Хефдинга, также известного как неравенство МакАллистера. Это неравенство основано на лемме Хефдинга и может считаться PAC-Байесовским аналогом неравенства Хефдинга Теоремы 4.

Лемма 36 (МакАллистер, [73, 74])**.** *Пусть для каждой $h \in \mathcal{H}$ случайные величины $\{X_1^h, \ldots, X_n^h\}$ независимы и одинаково распределены. Пусть, кроме того, $\mathbb{E}[X_1^h] = \mu^h$ и $X_i^h \in [0,1]$ для $i = 1, \ldots, n$ и $h \in \mathcal{H}$. Для любого не зависящего от выборок распределения π на $\mathcal{H}$, любых $\lambda \geqslant 0$ и $\delta > 0$ с вероятностью не меньше $1-\delta$ (относительно реализации случайных выборок $\{X_1^h, \ldots, X_n^h\}$ для $h \in \mathcal{H}$) следующее:*

$$\mathbb{E}_{h\sim\rho}\left[\mu^h - \frac{1}{n}\sum_{i=1}^n X_i^h\right] \leqslant \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{n\lambda} + \frac{\lambda}{8}$$

справедливо одновременно для всех распределений ρ на $\mathcal{H}$.

Доказательство. Положим в Лемме 35 в качестве f_m следующую функцию:

$$f_m = \lambda\sum_{i=1}^n(\mu_i^h - X_i^h) - \frac{\lambda^2}{8}n.$$

Тогда, пользуясь Леммой 2, мы получаем, что с вероятностью не меньше $1-\delta$ следующее

$$\mathbb{E}_{h\sim\rho}\left[\lambda\sum_{i=1}^n(\mu_i^h - X_i^h) - \frac{\lambda^2}{8}n\right] \leqslant \mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta} + \ln\mathbb{E}_{h\sim\pi}[1] = \mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}$$

выполнено одновременно для всех распределений ρ на $\mathcal{H}$. Действительно:

$$\begin{aligned}\mathbb{E}\left[\exp\left\{\lambda\sum_{i=1}^n(\mu_i^h - X_i^h) - \frac{\lambda^2}{8}n\right\}\right] &= \mathbb{E}\left[\exp\left\{\lambda\sum_{i=1}^n(\mu_i^h - X_i^h)\right\}\right]e^{-\frac{\lambda^2}{8}n}\\ &= \left(\prod_{i=1}^n\mathbb{E}\left[e^{\lambda(\mu_i^h - X_i^h)}\right]\right)e^{-\frac{\lambda^2}{8}n}\\ &\leqslant \left(\prod_{i=1}^n e^{\frac{\lambda^2}{8}}\right)e^{-\frac{\lambda^2}{8}n} = 1,\end{aligned}$$

где математическое ожидание $\mathbb{E}$ бралось по совместному распределению выборок $\{X_i^h\}_{i=1}^n$ для $h \in \mathcal{H}$. ■

Обратим внимание, что мы хотели бы минимизировать правую часть последнего результата по λ и получить наиболее точную оценку сверху. Однако, для этого нам требуется, чтобы оценка леммы выполнялась с большой вероятностью *одновременно для всех* значений λ из некоторого множества Λ, по которому мы хотим оптимизировать оценку. Результат же последней леммы справедлив для *одного отдельно взятого* значения λ.

Замечание 27. *Для дальнейших пояснений приведем аналогию обсуждаемого вопроса с вопросом равномерной по классу функций сходимости средних выборочных значений к их математическим ожиданиям, рассмотренным ранее в Главе 3. Неравенство Хефдинга давало нам оценку $L(h) \leqslant L_n(h) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$, выполненную с вероятностью не меньше $1-\delta$ для произвольного отображения h из рассматриваемого класса отображений $\mathcal{H}$. Эта оценка говорит, что для отображения $h \in \mathcal{H}$ событие, на котором выполнено упомянутое выше неравенство, имеет большую вероятность. Однако, оценка ничего не говорит о взаимосвязи этих событий для разных отображений $h \in \mathcal{H}$. В результате, событие, на котором неравенство Хефдинга выполняется* одновременно *для всех $h \in \mathcal{H}$ (это событие является пересечением всех индивидуальных событий для $h \in \mathcal{H}$), может иметь очень малую вероятность. Если же мы хотим минимизировать эмпирический риск и получить таким образом отображение $\hat{h}_n$ с гарантированно малым значением среднего риска, нам необходимо воспользоваться равномерным по классу $\mathcal{H}$ аналогом неравенства Хефдинга. В простейшем случае, когда класс $\mathcal{H}$ состоит из N отображений, мы можем воспользоваться неравенством Буля и получить оценку $L(h) \leqslant L_n(h) + \sqrt{\frac{\log \frac{N}{\delta}}{2n}}$, справедливую с вероятностью не меньше $1-\delta$ одновременно для всех $h \in \mathcal{H}$. Теперь мы можем минимизировать правую часть этой оценки по $h \in \mathcal{H}$ и быть уверенными в результате.*

Рассматривая последовательность $\{\lambda_i\}_{i=1}^{\infty}$ значений $\lambda \geqslant 0$, образующую геометрическую прогрессию с множителем $c > 1$, и применяя для них Лемму 36 вместе с неравенством Буля, мы получим следующее PAC-Байесовское неравенство Хефдинга:

Теорема 49 (PAC-Байесовское неравенство Хефдинга, [73, 74])**.** *Пусть для каждой $h \in \mathcal{H}$ случайные величины $\{X_1^h, \ldots, X_n^h\}$ независимы и одинаково распределены. Пусть, кроме того, $\mathbb{E}[X_1^h] = \mu^h$ и $X_i^h \in [0,1]$ для $i = 1, \ldots, n$ и $h \in \mathcal{H}$. Для любого не зависящего от выборок распределения π на $\mathcal{H}$, любых $c > 1$ и $\delta > 0$ с вероятностью не меньше $1-\delta$ (относительно реализации случайных выборок $\{X_1^h, \ldots, X_n^h\}$ для $h \in \mathcal{H}$) следующее:*

$$\mathbb{E}_{h\sim\rho}\left[\mu^h - \frac{1}{n}\sum_{i=1}^{n} X_i^h\right] \leqslant \frac{1+c}{2}\sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln(1/\delta) + \varepsilon(\rho)}{2n}}$$

выполнено одновременно для всех распределений ρ на $\mathcal{H}$, где

$$\varepsilon(\rho) = \frac{\ln 2}{2 \ln c}\left(1 + \left(\frac{\mathrm{KL}(\rho\|\pi)}{\ln \frac{1}{\delta}}\right)\right).$$

Доказательство может быть найдено на странице 7 работы [88]. Обратим внимание, что повторяя те же шаги доказательства мы получим, что результат также справедлив для $\mathbb{E}_{h\sim\rho}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^h - \mu^h\right]$.

PAC-Байесовское неравенство Бернштейна. Прошлый результат, как и неравенство Хефдинга, учитывает лишь ограниченность случайных величин. Мы хорошо знаем, что учет дисперсий может часто приводить к более точным результатам. Если мы воспользуемся при оценке сверху производящей функции моментов в доказательстве прошлой теоремы Леммой Беннетта 3 вместо леммы Хефдинга, мы получим следующий результат, учитывающий дисперсию случайных величин:

Лемма 37 (Сельдин и др., [88])**.** *Пусть для каждой $h \in \mathcal{H}$ случайные величины $\{X_1^h, \ldots, X_n^h\}$ независимы и одинаково распределены. Пусть, кроме того, $\mathbb{E}[X_1^h] = \mu^h$, $\mathbb{D}[X_1^h] = V^h$ и $X_i^h \in [0,1]$ для $i = 1, \ldots, n$ и $h \in \mathcal{H}$. Для любого не зависящего от выборок распределения π на $\mathcal{H}$, любых $\lambda \in [0,1]$ и $\delta > 0$ с вероятностью не меньше $1-\delta$ (относительно реализации случайных выборок $\{X_1^h, \ldots, X_n^h\}$ для $h \in \mathcal{H}$) следующее:*

$$\mathbb{E}_{h\sim\rho}\left[\mu^h - \frac{1}{n}\sum_{i=1}^{n} X_i^h\right] \leqslant \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{n\lambda} + (e-2)\lambda\,\mathbb{E}_{h\sim\rho}\left[V^h\right]$$

справедливо одновременно для всех распределений ρ на $\mathcal{H}$.

Доказательство. Положим в Лемме 35 в качестве f_m следующую функцию:

$$f_m = \lambda\sum_{i=1}^{n}(\mu_i^h - X_i^h) - (e-2)\lambda^2 n V^h.$$

Тогда, пользуясь Леммой 3, мы получаем, что с вероятностью не меньше $1-\delta$ следующее

$$\mathbb{E}_{h\sim\rho}\left[\lambda\sum_{i=1}^{n}(\mu_i^h - X_i^h) - (e-2)\lambda^2 n V^h\right] \leqslant \mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta} + \ln \mathbb{E}_{h\sim\pi}[1] = \mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}$$

выполнено одновременно для всех распределений ρ на $\mathcal{H}$. Действительно:

$$\begin{aligned}\mathbb{E}\left[\exp\left\{\lambda\sum_{i=1}^{n}(\mu_i^h - X_i^h) - (e-2)\lambda^2 n V^h\right\}\right] &= \left(\prod_{i=1}^{n}\mathbb{E}\left[e^{\lambda(\mu_i^h - X_i^h)}\right]\right)e^{-(e-2)\lambda^2 n V^h} \\ &\leqslant \left(\prod_{i=1}^{n}\left[e^{(e-2)\lambda^2 V^h}\right]\right)e^{-(e-2)\lambda^2 n V^h} = 1,\end{aligned}$$

где математическое ожидание $\mathbb{E}$ бралось по совместному распределению выборок $\{X_i^h\}_{i=1}^n$ для $h \in \mathcal{H}$. ■

Замечание 28. *Мы воспользовались в доказательстве менее точной оценкой Беннетта для того, чтобы правую часть оценки леммы можно было легко оптимизировать по λ аналитически.*

Вновь воспользовавшись неравенством Булля по значениям λ из геометрической прогрессии на интервале $[0, 1]$, мы получаем следующее PAC-Байесовское неравенство Бернштейна:

Теорема 50 (PAC-Байесовское неравенство Бернштейна, [88])**.** *Пусть для каждой $h \in \mathcal{H}$ случайные величины $\{X_1^h, \ldots, X_n^h\}$ независимы и одинаково распределены. Пусть, кроме того, $\mathbb{E}[X_1^h] = \mu^h$, $\mathbb{D}[X_1^h] = V^h$ и $X_i^h \in [0, 1]$ для $i = 1, \ldots, n$ и $h \in \mathcal{H}$. Рассмотрим любое распределение π на $\mathcal{H}$, не зависящее от выборок. Тогда для любых $c > 1$ и $\delta > 0$ с вероятностью не меньше $1 - \delta$ (относительно реализации случайных выборок $\{X_1^h, \ldots, X_n^h\}$ для $h \in \mathcal{H}$) следующее:*

$$\mathbb{E}_{h\sim\rho}\left[\mu^h - \frac{1}{n}\sum_{i=1}^{n} X_i^h\right] \leqslant (1+c)\sqrt{\frac{(e-2)\mathbb{E}_{h\sim\rho}[V^h]\left(\mathrm{KL}(\rho\|\pi) + \ln(\nu/\delta)\right)}{n}},$$

выполнено одновременно для всех распределений ρ на $\mathcal{H}$, таких что

$$\sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln(\nu/\delta)}{(e-2)\mathbb{E}_{h\sim\rho}[V^h]}} \leqslant \sqrt{n},$$

где $\nu = \left\lceil \frac{1}{\ln c} \ln\left(\sqrt{\frac{(e-2)n}{4\ln(1/\delta)}}\right)\right\rceil + 1$. Для всех остальных распределений ρ справедливо:

$$\mathbb{E}_{h\sim\rho}\left[\mu^h - \frac{1}{n}\sum_{i=1}^{n} X_i^h\right] \leqslant 2\frac{\mathrm{KL}(\rho\|\pi) + \ln(\nu/\delta)}{n}.$$

Кроме того, результат остается справедливым, если мы заменим $\mathbb{E}_{h\sim\rho}\left[V^h\right]$ любой его верхней оценкой $\hat{\mathbb{D}}_n(\rho)$, такой что $\mathbb{E}_{h\sim\rho}\left[V^h\right] \leqslant \hat{\mathbb{D}}_n(\rho) \leqslant \frac{1}{4}$ для всех ρ.

Доказательство может быть найдено на странице 7 работы [88]. Обратим внимание, что повторяя те же шаги доказательства мы получим, что результат также справедлив для $\mathbb{E}_{h\sim\rho}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^h - \mu^h\right]$.

Грубо говоря, последний результат отличается от PAC-Байесовского неравенства Хефдинга появлением множителя $\sqrt{\mathbb{E}_{h\sim\rho}[V^h]}$ перед слагаемым порядка $1/\sqrt{n}$. Чем меньше дисперсия случайных величин, тем точнее оценка.

PAC-Байесовское kl-неравенство. Наконец, мы приведем еще один результат, полученный в работе [85] для случайных величин Бернулли и затем обобщенный в работе [69] на произвольные ограниченные случайные величины.

Нам понадобится следующее определение kl-*функции*. Для $0 \leqslant p, q \leqslant 1$ определим следующую функцию:

$$\mathrm{kl}(q\|p) = q \ln \frac{q}{p} + (1-q) \ln \frac{1-q}{1-p}.$$

kl-функция совпадает с относительной энтропией между двумя распределениями Бернулли с параметрами q и p соответственно. Следовательно, она неотрицательна и равна нулю тогда и только тогда, когда $p = q$. Кроме того, несложно показать, что kl-функция гладкая и выпукла по совокупности своих аргументов (что позволяет легко обращать ее численно). Наконец, из неравенства Пинскера [27] мы можем получить следующую нижнюю оценку:

$$\mathrm{kl}(q\|p) \geqslant 2(p-q)^2. \tag{6.2}$$

Таким образом, верхняя оценка на kl-функцию предоставляет верхнюю оценку на разность $|p-q|$. Приведенные далее результаты ограничивают откланения математических ожиданий от выборочных средних, получая верхние оценки на kl-функцию в двух этих точках. Учитывая описанный выше общий подход к получению PAC-Байесовских неравенств, нашей главной целью будет получение верхней оценки на производящую функцию моментов случайной величины

$$n \cdot \mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \middle\| \mu\right),$$

где случайные величины $\{X_1, \ldots, X_n\}$ независимы и одинаково распределены, и, кроме того, $\mathbb{E}[X_1] = \mu$, $X_i \in [0,1]$.

Изящный подход к этой задаче, состоящий из следующих трех шагов, был предложен А. Маурером в работе [69]:

Шаг 1 Для выпуклой функции $f\colon [0,1]^n \to \mathbb{R}$ величину $\mathbb{E}[f(X_1, \ldots, X_n)]$ оценим сверху математическим ожиданием $\mathbb{E}[f(Y_1, \ldots, Y_n)]$, где $\{Y_1, \ldots, Y_n\}$ — специальным образом выбранные слечайные величины Бернулли.

Шаг 2 Заметим, что отображение $f_0\colon \{x_1, \ldots, x_n\} \in [0,1]^n \to e^{n\,\mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^n x_i \| \mu\right)}$ выпукло.

Шаг 3 Нам остается оценить сверху величину $\mathbb{E}[f_0(Y_1, \ldots, Y_n)]$, что можно сделать, пользуясь известными комбинаторными результатами.

Первый шаг основан на следующей лемме:

Лемма 38 (Маурер, [69])**.** Пусть случайные величины $\{X_1, \ldots, X_n\}$ независимы и одинаково распределены, и, кроме того, $\mathbb{E}[X_1] = \mu$, $X_i \in [0,1]$. Рассмотрим независимые и одинаково

распределенные случайные величины Бернулли $\{Y_1, \dots, Y_n\}$ с параметром $\mathbb{E}[Y_1] = \mu$. Тогда для любой выпуклой функции $f\colon [0,1]^n \to \mathbb{R}$ справедливо:

$$\mathbb{E}[f(X_1, \dots, X_n)] \leqslant \mathbb{E}[f(Y_1, \dots, Y_n)].$$

Третий шаг основан на следующей верхней оценке на число сочетаний, которая может быть найдена в [27]:

$$C_n^k \leqslant \exp\left(-k \log \frac{k}{n} - (n-k) \log \frac{n-k}{n}\right).$$

Записав математическое ожидание дискретной случайной величины по определению в виде суммы и воспользовавшись приведенным неравенством, мы получим следующий результат:

Лемма 39 (Зигер, [85])**.** *Рассмотрим независимые и одинаково распределенные случайные величины Бернулли $\{Y_1, \dots, Y_n\}$ с параметром $\mathbb{E}[Y_1] = \mu$. Тогда:*

$$\mathbb{E}\left[e^{n \cdot \mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^n Y_i \| \mu\right)}\right] \leqslant n + 1.$$

В работе [69] Маурер приводит более точную верхнюю оценку $2\sqrt{n}$.

Совмещая описанные шаги, мы получаем следующую теорему:

Теорема 51. *Пусть случайные величины $\{X_1, \dots, X_n\}$ независимы и одинаково распределены, и, кроме того, $\mathbb{E}[X_1] = \mu$, $X_i \in [0,1]$. Тогда:*

$$\mathbb{E}\left[e^{n \cdot \mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^n X_i \| \mu\right)}\right] \leqslant n + 1.$$

Замечание 29. *Последняя теорема вместе с методом Чернова немедленно ведет к следующей верхней оценке:*

$$\mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^n X_i \middle\| \mu\right) \leqslant \frac{\log \frac{n+1}{\delta}}{n},$$

справедливой с вероятностью не меньше $1-\delta$. Пользуясь неравенством Пинскера (6.2), мы также получаем следующую оценку:

$$\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \leqslant \sqrt{\frac{\log \frac{n+1}{\delta}}{2n}},$$

похожую на неравенство Хефдинга. Однако, оказывается, неравенство Пинскера в данном случае является достаточно грубым и справедлива более точная оценка kl*-функции снизу. В действительности Теорема 51 ведет к гораздо более точной оценке на отклонение средних выборочных от математического ожидания. Мы вернемся к этому вопросу в Разделе 6.2.3.*

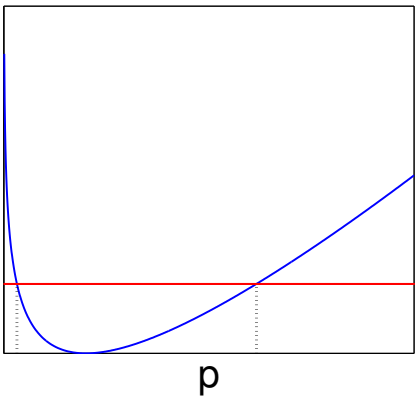


Рисунок 6.1: Применение PAC-Байесовского kl-неравенства.

Теорема 52 (PAC-Байесовское kl-неравенство, [56, 69, 85])**.** *Пусть для каждой $h \in \mathcal{H}$ случайные величины $\{X_1^h, \ldots, X_n^h\}$ независимы и одинаково распределены. Пусть, кроме того, $\mathbb{E}[X_1^h] = \mu^h$ и $X_i^h \in [0,1]$ для $i = 1, \ldots, n$ и $h \in \mathcal{H}$. Для любого не зависящего от выборок распределения π на $\mathcal{H}$, любой $\delta > 0$ с вероятностью не меньше $1-\delta$ (относительно реализации случайных выборок $\{X_1^h, \ldots, X_n^h\}$ для $h \in \mathcal{H}$) следующее:*

$$\mathrm{kl}\left(\mathbb{E}_{h\sim\rho}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^h\right] \middle\| \mathbb{E}_{h\sim\rho}\left[\mu^h\right]\right) \leqslant \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}}{n} \tag{6.3}$$

справедливо одновременно для всех распределений ρ на $\mathcal{H}$.

Доказательство. Положим в Лемме 35 в качестве f_m следующую функцию:

$$f_m = n \cdot \mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^h \middle\| \mu^h\right).$$

Доказательство повторяет шаги Леммы 37 и основано на Теореме 51. ■

Еще раз прокомментируем способ применения последнего результата. Предположим, что мы вычислили значение правой части неравенства, а также значение величины $\mathbb{E}_{h\sim\rho}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^h\right]$, и пусть эти значения равны некоторым константам C_1 и C_2 соответственно. Как получить искомую верхнюю оценку на $\mathbb{E}_{h\sim\rho}\left[\mu^h\right]$? Для краткости обозначим $\mathbb{E}_{h\sim\rho}\left[\mu^h\right] = p$. Мы хотим найти верхнюю оценку на $p \in [0,1]$, такую что выполнено следующее неравенство:

$$\mathrm{kl}\left(C_2\|p\right) \leqslant C_1.$$

Предположим, что $C_1 = C_2 = 0.02$. На Рисунке 6.1 синим изображен график функции $\mathrm{kl}(0.02, p)$ при $p \in [0,1]$. Красным изображена прямая $y = 0.02$. Рисунок демонстрирует, что мы одновременно получаем и верхнюю и нижние оценки на p, вычислив точки пересечения синей кривой с красной. Как отмечалось ранее, численное обращение kl-функции является несложной задачей благодаря хорошим свойствам этой функции.

6.2.3 Сравнение PAC-Байесовских неравенств

Теперь мы приведем короткое сравнение PAC-Байесовских неравенств, описанных в прошлом разделе. Для краткости мы будет использовать следующие аббревиатуры: PB-H, PB-B и PB-kl для PAC-Байесовских неравенств Хефдинга, Бернштейна и kl-неравенства соответственно.

Начнем со сравнения PAC-Байесовских неравенств Хефдинга и Бернштейна. Поскольку случайные величины ограничены в интервале $[0, 1]$, справедлива следующая тривиальная оценка: $\mathbb{E}_{h\sim\rho}[V^h] \leqslant 1/4$. Это показывает, что PB-B Теоремы 50 не может быть сильно хуже PB-H Теоремы 49. В тех случаях, когда дисперсия случайных величин мала, PB-B может быть существенно точнее PB-H.

Сравнение PAC-Байесовского неравенства Хефдинга с kl-неравенством также является простым. Пользуясь неравенством Пинскера (6.2), мы можем переписать kl-неравенство в следующем виде:

$$\left|\mathbb{E}_{h\sim\rho}[\mu^h] - \mathbb{E}_{h\sim\rho}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^h\right]\right| \leqslant \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}}{2n}}.$$

Мы заключаем, что PB-kl по крайней мере не хуже PB-H.

Сравнение PAC-Байесовского неравенства Бернштейна с kl-неравенством является менее тривиальным. Оказывается, справедлива следующая более точная чем неравенство Пинскера оценка снизу:

$$p \leqslant q + \sqrt{2q\mathrm{kl}(q\|p)} + 2\mathrm{kl}(q\|p), \quad p > q.$$

Пользуясь этой оценкой, мы можем переписать PB-kl в следующем виде:

$$\mathbb{E}_{h\sim\rho}\left[\mu^h - \frac{1}{n}\sum_{i=1}^{n} X_i^h\right] \leqslant \sqrt{2\mathbb{E}_{h\sim\rho}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^h\right]\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}}{n}} + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}}{n}. \tag{6.4}$$

Мы приходим к выводу, что сравнение между PB-B и PB-kl зависит от взаимоотношения средней дисперсии $\mathbb{E}_{h\sim\rho}[V^h]$ и выборочных средних $\mathbb{E}_{h\sim\rho}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^h\right]$. Чем меньше выборочные средние, тем точнее PB-kl неравенство. Заметим, что для случайной величины $\xi \in [0, 1]$ справедливо $\mathbb{D}[\xi] \leqslant \mathbb{E}[\xi]$. Вместе с тем фактом, что для больших n средние выборочные значения концентрируются вокруг математических ожиданий, мы приходим к выводу, что PB-B неравенство не сильно хуже PB-kl неравенства. Если в задаче доступна достаточно точная верхняя оценка на среднюю дисперсию $\mathbb{E}_{h\sim\rho}[V^h]$, PB-B может оказаться существенно точнее PB-kl. Однако, если мы вынуждены прибегнуть к тривиальной верхней оценке $1/4$, PB-kl неравенство является более предпочтительным.

6.2.4 Применение PAC-Байесовских неравенств в теории обучения

В данном разделе мы вкратце опишем способы применения результатов PAC-Байесовского анализа, описанных выше, в задачах теории статистического обучения. Мы будем пользоваться определениями и постановкой, рассмотренными в Главе 3.

Обучение рандомизированных предикторов. Ранее мы определили рандомизированный алгоритм Гиббса G_ρ, а также его средний и эмпирический риски $L(G_\rho)$, $L_n(G_\rho)$. Результаты, приведенные в прошлом разделе, позволяют ограничивать сверху средний риск рандомизированных отображений $L(G_\rho)$ на основе их эмпирического риска $L_n(G_\rho)$ *одновременно для всех распределений* ρ.

Мы можем ставить задачу поиска рандомизированного отображения G_ρ с наименьшим значением среднего риска $L(G_\rho)$ на основе обучающей выборки S. Конечно, использование такого случайного предиктора на практике может вызывать ряд сложностей. Тем не менее, PAC-Байесовские неравенства позволяют эффективно решать эту задачу. Действительно, поскольку все приведенные выше неравенства справедливы *одновременно* для всех распределений ρ, мы можем минимизировать верхние оценки на $L(G_\rho)$ и получать рандомизированные алгоритмы с гарантированно малыми значениями среднего риска. Например, PAC-Байесовское неравенство Бернштейна дает, грубо говоря, следующую верхнюю оценку на средний риск отображения Гиббса:

$$L(G_\rho) \leqslant L_n(G_\rho) + 2\sqrt{\frac{(e-2)\mathbb{E}_{h\sim\rho}\big[\mathbb{D}_{(X,Y)\sim P}[\ell_h(X,Y)]\big]\big(\mathrm{KL}(\rho\|\pi) + \log(1/\delta)\big)}{n}}.$$

На примере этой оценки видно, что минимизация PAC-Байесовских неравенств не всегда является простой задачей. В ряде случаев оценки могут вовсе не быть выпуклыми по ρ. Ряд примеров минимизации PAC-Байесовских оценок для различных семейств $\mathcal{H}$ и параметрических семейств распределений ρ могут быть найдены в работах [39, 57, 86].

Выбор априорного распределения π. Обратим внимание, что все PAC-Байесовские неравенства, приведенные выше, так или иначе ограничивают сверху средний риск $L(G_\rho)$ некоторой линейной комбинацией эмпирического риска $L_n(G_\rho)$ и «сложностного» слагаемого, зависящего от относительной энтропии $\mathrm{KL}(\rho,\pi)$. Пользуясь этим наблюдением мы можем прокомментировать выбор распределения π, часто называемого в литературе *априорным* (поскольку оно выбирается до наблюдения выборки S). Слагаемое, зависящее от $\mathrm{KL}(\rho,\pi)$, фактически является *регуляризатором*. Действительно, при существенном отклонении *апостериорного* распределения ρ от априорного π, член $\mathrm{KL}(\rho\|\pi)$ растет. В то же время, для

минимизации значения оценки нам необходимо, чтобы эмпирический риск $L_n(G_\rho)$ распределения ρ был мал. Таким образом, если априорное распределение π выбрано «плохо» и его вероятностная масса сосредоточена на отображениях $h \in \mathcal{H}$ с большими значениями средних рисков, то PAC-Байесовские оценки могут быть сильно завышенными, поскольку одновременная минимизация $\mathrm{KL}(\rho, \pi)$ и $L_n(G_\rho)$ становится невозможной. Эти размышления сильно перекликаются с обсуждением, представленным после неравенства Бритвы Оккама Теоремы 25. Оценка Бритвы Оккама может считаться первым зачаточным результатом PAC-Байесовского анализа.

Анализ классических задач с помощью PAC-Байесовских неравенств. Оказывается, в ряде случаев PAC-Байесовский анализ может позволить работать не только с рандомизированными алгоритмами, но и с обычными детерминистскими предикторами, рассмотренными в Главе 3. Следующие рассуждения и результаты приведены в работе [57].

Рассмотрим задачу бинарной классификации $\mathcal{Y} = \{-1, +1\}$. Наряду с рандомизированным отображением Гиббса G_ρ мы можем рассмотреть детерминистское отображение $B_\rho \colon \mathcal{X} \to \mathcal{Y}$, определяемое следующим соотношением: $B_\rho(X) = \operatorname{sgn}\{\mathbb{E}_{h\sim\rho}[h(X)]\}$, $X \in \mathcal{X}$. Отображение B_ρ является взвешенной композицией классификаторов из $\mathcal{H}$ с весами $\rho(h)$.

Рассмотрим бинарную функцию потерь, множеством объектов $\mathcal{X} \subseteq \mathbb{R}^d$ и множество классификаторов $\mathcal{H}$, состоящее из всех гиперплоскостей в пространстве $\mathbb{R}^d$, проходящих через начало координат. Мы можем параметризовать семейство $\mathcal{H}$ векторами нормалей $w \in \mathbb{R}^d$ следующим образом: $\mathcal{H} = \{h_w(X) = \operatorname{sgn}\langle w, X\rangle \colon w \in \mathbb{R}^d\}$. Возьмем в качестве априорного распределения π многомерное нормальное распределение в $\mathbb{R}^d$ с центром в начале координат и единичной ковариационной матрицей. Рассмотрим параметрическое семейство распределений $\rho = \rho(\boldsymbol{\mu})$, $\boldsymbol{\mu} \in \mathbb{R}^d$, состоящее из многомерных нормальных распределений с единичными ковариационными матрицами и центрами в точке $\boldsymbol{\mu}$.

Легко показать, что в этом случае для любых объектов $X \in \mathcal{X}$ справедливо тождество:

$$B_{\rho(\boldsymbol{\mu})}(X) = \operatorname{sgn}\{\langle X, \boldsymbol{\mu}\rangle\} = h_{\boldsymbol{\mu}}(X),$$

то есть ответы взвешенного классификатора B_ρ всегда совпадают с ответами одного отдельно взятого классификатора $h_{\boldsymbol{\mu}}$, нормаль гиперплоскости которого совпадает с центром $\boldsymbol{\mu}$ распределения $\rho(\boldsymbol{\mu})$. Кроме того, взвешенная композиция B_ρ ошибается на объекте $X \in \mathcal{X}$ тогда и только тогда, когда более половины (по мере ρ) классификаторов семейства $\mathcal{H}$ ошибается на объекте $X \in \mathcal{X}$. Пользуясь этим наблюдением можно легко показать, что для любых

распределений ρ (не обязательно нормальных) справедливо следующее неравенство:

$$L(B_\rho) \leqslant 2L(G_\rho). \tag{6.5}$$

Замечание 30. *Обратим внимание, что последнее неравенство может оказаться сильно завышенным. Рассмотрим, например, алгоритм бустинга, строящий взвешенную композицию простых классификаторов. Известно, что бустинг пороговых классификаторов часто ведет к композициям с очень малыми значениями среднего риска [41, 84]. В то же время пороговые классификаторы являются тривиальными моделями и зачастую имеют средний риск, близкий к* $1/2$*. В этом случае средний риск взвешенного классификатора* $L(B_\rho)$*, получаемого методом бустинга, может быть сколь угодно близок к нулю, в то время как средний риск классификатора Гиббса* $L(G_\rho) = \mathbb{E}_{h\sim\rho}[L(h)]$ *будет близок к значению* $1/2$*. Способы борьбы с этой проблемой рассматриваются в работах [38, 57].*

Таким образом, мы можем оценить средний риск классификатора $h_{\boldsymbol{\mu}}$ для произвольного $\boldsymbol{\mu} \in \mathbb{R}^d$ следующим образом:

$$L(h_{\boldsymbol{\mu}}) = L(B_{\rho(\boldsymbol{\mu})}) \leqslant 2L(G_{\rho(\boldsymbol{\mu})}).$$

Пользуясь PAC-Байесовскими неравенствами, мы можем далее оценить сверху средний риск рандомизированного отображения $L(G_{\rho(\boldsymbol{\mu})})$ в терминах наблюдаемых величин. Несложные вычисления показывают, что в рассматриваемом случае справедливы следующие соотношения:

Теорема 53 (Лэнгфорд и Шоу-Тейлор, [57])**.** *В описанной выше постановке выполнены следующие соотношения:*

$$\mathrm{KL}\big(\rho(\boldsymbol{\mu})\|\pi\big) = \frac{\|\boldsymbol{\mu}\|^2}{2};$$
$$L_n(G_{\rho(\boldsymbol{\mu})}) = \frac{1}{n}\sum_{i=1}^{n} \bar{F}\left(\frac{y_i\langle\boldsymbol{\mu}, \boldsymbol{x}_i\rangle}{\|\boldsymbol{x}\|}\right),$$

где $\bar{F}(\varepsilon) = 1 - \int_{-\infty}^{\varepsilon} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$ *— правый хвост стандартного нормального распределения.*

В частности, пользуясь последним результатом и Теоремой 52, мы получаем следующую оценку:

Следствие 22 (Лэнгфорд и Шоу-Тейлор, [57])**.** *В рассматриваемой постановке для любого* $\delta > 0$ *с вероятностью не меньше* $1 - \delta$ *(относительно случайной реализации обучающей выборки* S*) следующее:*

$$\mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^{n} \bar{F}\left(\frac{y_i\langle\boldsymbol{\mu}, \boldsymbol{x}_i\rangle}{\|\boldsymbol{x}\|}\right) \middle\| L(G_{\rho(\boldsymbol{\mu})})\right) \leqslant \frac{\frac{\|\boldsymbol{\mu}\|^2}{2} + \ln\frac{n+1}{\delta}}{n}$$

справедливо одновременно для всех $\boldsymbol{\mu} \in \mathbb{R}^d$*.*

Известно, что вместе с неравенством (6.5) последний результат ведет к *наиболее точным на сегодняшний день* оценкам для линейных классификаторов. Кроме того, верхнюю оценку Теоремы 22 можно оптимизировать по векторам нормали $\boldsymbol{\mu} \in \mathbb{R}^d$, что ведет к новому методу обучения с квадратичной регуляризацией. Подобные методы обучения линейных классификаторов подробно рассматриваются в работе [39].

6.3 PAC-Байесовское эмпирическое неравенство Бернштейна

Результаты настоящего раздела являются новыми и опубликованы в работе [95].

Как мы видели в прошлых разделах, PAC-Байесовское неравенство Бернштейна во многих случаях может оказаться наиболее точным из всех описанных неравенств. Однако, для использования этого неравенства на практике нам необходимо оценить сверху входящую в него неизвестную нам усредненную дисперсию $\mathbb{E}_{h\sim\rho}[V^h]$. Поскольку случайные величины X_i^h ограничены в интервале $[0,1]$, справедлива тривиальная оценка сверху $\mathbb{E}_{h\sim\rho}[V^h] \leqslant 1/4$. Однако, в этом случае, как мы видели, PAC-Байесовское неравенство Бернштейна сводится к PAC-Байесовскому неравенству Хефдинга и теряет все свои превосходства. В ряде случаев (в том числе в *сэмплировании по значимости*, на котором основаны последние результаты в *активном обучении* [2,13]) возможно получение достаточно точных и неслучайных верхних оценок усредненной дисперсии, которые могут оказаться существенно меньше усредненного эмпирического риска $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i^h\right]$ [87]. В этом случае, как мы знаем, PAC-Байесовское неравенство Бернштейна может давать существенно более точные оценки, чем kl-неравенство. Однако, в большинстве случаев нам недоступны нетривиальные и неслучайные оценки величины $\mathbb{E}_{h\sim\rho}[V^h]$ сверху.

В настоящем разделе мы получим новую *вычислимую по обучающей выборке* PAC-Байесовскую оценку для усредненной дисперсии $\mathbb{E}_{h\sim\rho}[V^h]$. Для этого мы совместим PAC-Байесовскую технику получения оценок, описанную выше, с *эмпирическим неравенством Бернштейна* — мощной верхней оценкой на производящую функцию моментов выборочной дисперсии, описанную в Разделе 1.3.1. Совместив полученную таким образом оценку с PAC-Байесовским неравенством Бернштейна Теоремы 50, мы получим новое PAC-Байесовское эмпирическое неравенство Бернштейна — полностью вычислимый по данным аналог PAC-Байесовского неравенства Бернштейна.

Новая PAC-Байесовская оценка для дисперсий ограничивает разность усредненных (относительно ρ) значений истинной и выборочной дисперсии риска алгоритмов из $\mathcal{H}$. В общем случае выборочная дисперсия ведет к более точным оценкам дисперсии, чем эмпирический риск, особенно в тех случаях, когда значение эмпирического риска относительно велико. В предположении, что потери ограничены в интервале [0,1], PAC-Байесовское эмпирическое неравенство Бернштейна точнее PAC-Байесовского kl-неравенства, грубо говоря, в тех случаях, когда эмпирический риск превышает значение 0.1, а выборочная дисперсия меньше 0.1. В противном случае PAC-Байесовское эмпирическое неравенство Бернштейна завышено немного больше (точное соотношение оценок также зависит от размера обучающей выборки и относительной энтропией между априорным и апостериорным распределениями и обсуждается в разделе 6.3.3). Простейшим примером ситуации, когда эмпирический риск велик, а выборочная дисперсия мала, являются задачи, где множество алгоритмов $\mathcal{H}$ не способно без ошибок справиться с распределением данных, а шум в задаче при этом относительно мал. В разделе 6.3.3 мы приводим несколько модельных выборок и выборок из репозитория UCI, на которых новая оценка существенно улучшает все предыдущие.

В настоящем разделе мы перейдем от обозначений общей постановки PAC-Байесовского анализа, введенных в Параграфе 6.1, к обозначениям, принятым в теории статистического обучения. Для краткости мы будем использовать следующие обозначения: $\mathbb{D}(h) = \mathbb{D}_{(X,Y)\sim P}[\ell_h(X,Y)]$.

6.3.1 PAC-Байесовское неравенство для дисперсии

В настоящем разделе мы получим верхнюю оценку на усредненную дисперсию $\mathbb{E}_{h\sim\rho}[\mathbb{D}(h)]$ в терминах ее выборочного аналога $\mathbb{E}_{h\sim\rho}\big[\mathbb{D}_n(h)\big]$, где:

$$\mathbb{D}_n(h) = \frac{1}{n-1}\sum_{i=1}^{n}\big(\ell_h(X_i,Y_i) - L_n(h)\big)^2 \tag{6.6}$$

— несмещенная выборочная оценка дисперсии $\mathbb{D}(h)$. Для этого мы воспользуемся общей стратегией получения PAC-Байесовских неравенств, описанной выше. Ранее мы рассмотрели три PAC-Байесовских неравенства, в доказательствах которых мы пользовались верхними оценками на производящие функции моментов сумм случайных величин (для PAC-Байесовских неравенства Бернштейна и Хефдинга) и kl-функции (в kl-неравенстве). Теперь нам необходима оценка для производящей функции моментов выборочной дисперсии $\mathbb{D}_n(h)$. В Разделе 1.3.1 было показано, что случайная величина $\mathbb{D}_n(h)$, будучи функцией обучающей выборки, удовлетворяет условию ограниченных разностей, описанному в Разделе 1.2. Таким обра-

зом мы можем воспользоваться верхней оценкой МакДиармида на производящую функцию моментов случайной величины $\mathbb{D}_n(h)$ Теоремы 9. Однако, этот простой подход ведет к сильно завышенным оценкам, и мы приведем более точный результат, основанный на энтропийном подходе.

Теорема 54. *Для любого распределения π на множестве отображений $\mathcal{H}$, не зависящего от обучающей выборки, для любого $\delta > 0$ и для любого $c > 1$ с вероятностью не меньше $1 - \delta$ (относительно случайной реализации обучающей выборки S) следующее справедливо одновременно для всех распределений ρ на $\mathcal{H}$:*

$$\mathbb{E}_{h\sim\rho}[\mathbb{D}(h)] \leqslant c\,\mathbb{E}_{h\sim\rho}[\mathbb{D}_n(h)] + 2c\sqrt{\big(\mathbb{E}_{h\sim\rho}[\mathbb{D}_n(h)] + \eta_n(\rho)\big)\eta_n(\rho)} + 2c\,\eta_n(\rho), \tag{6.7}$$

где

$$\eta_n(\rho) = \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{\nu}{\delta}}{2(n-1)}, \quad \nu = \left\lceil \frac{1}{\ln c} \ln\left(\frac{1}{2}\sqrt{\frac{n-1}{\ln 1/\delta} + 1} + \frac{1}{2}\right)\right\rceil.$$

Если мы применим неравенство $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$ к (6.7), оно становится похожим на релаксацию klнеравенства, представленную в (6.4). По мере уменьшения выборочной дисперсии $\mathbb{D}_n(h)$ член порядка $1/\sqrt{n}$ уменьшается, что увеличивает скорость сходимости оценки. Отметим, что вклад множителя c в $\eta_n(\rho)$ равен $\ln\ln c$, а значит c может быть выбран достаточно близко к 1 без значительного увеличения η_n.

Доказательство. Из Леммы 6 следует, что для случайной величины $n\cdot\mathbb{D}_n(h)$ и любого $\lambda > 0$ справедливо:

$$\mathbb{E}\left[e^{\lambda(n\mathbb{D}(h) - n\mathbb{D}_n(h))}\right] \leqslant e^{\frac{\lambda^2}{2}\frac{n^2}{n-1}\mathbb{D}(h)},$$

что эквивалентно:

$$\mathbb{E}\left[e^{\lambda(n\mathbb{D}(h) - n\mathbb{D}_n(h)) - \frac{\lambda^2}{2}\frac{n^2}{n-1}\mathbb{D}(h)}\right] \leqslant 1.$$

Последнее неравенство является верхней оценкой на производящую функцию моментов следующей случайной величины:

$$\Phi_\lambda(h) = \lambda n\left(1 - \frac{\lambda n}{2(n-1)}\right)\mathbb{D}(h) - \lambda n\mathbb{D}_n(h).$$

Положим в Лемме 35 в качестве f_m функцию $\Phi_\lambda(h)$. Тогда мы получим, что для любого распределения π на $\mathcal{H}$, не зависящего от обучающей выборки, с вероятностью не меньше $1 - \delta$ одновременно для всех распределений ρ справедливо:

$$\mathbb{E}_{h\sim\rho}[\Phi_\lambda(h)] \leqslant \mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}$$

или

$$\left(1 - \frac{\lambda n}{2(n-1)}\right) \mathbb{E}_{h\sim\rho}[\mathbb{D}(h)] \leqslant \mathbb{E}_{h\sim\rho}[\mathbb{D}_n(h)] + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{\lambda n}.$$

Наложив ограничение $\lambda \leqslant \frac{2(n-1)}{n}$ и поделив обе части неравенства на $1 - \frac{\lambda n}{2(n-1)}$, мы получаем:

$$\mathbb{E}_{h\sim\rho}[\mathbb{D}(h)] \leqslant \frac{\mathbb{E}_{h\sim\rho}[\mathbb{D}_n(h)]}{\left(1 - \frac{\lambda n}{2(n-1)}\right)} + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{\lambda n \left(1 - \frac{\lambda n}{2(n-1)}\right)}. \tag{6.8}$$

Как и при доказательстве всех прошлых PAC-Байесовских неравенств заметим, что правая часть неравенства (6.8) не может быть минимизирована одним значением λ одновременно для всех ρ. В доказательстве мы сначала найдем оптимальное значение λ, минимизирующее (6.8). Затем мы введем в рассмотрение последовательность значений λ, образующую геометрическую прогрессию, и выберем элемент из этой последовательности, ближайший к оптимальному значению λ. Этот технический прием заимствован из работы [88].

Для краткости введем следующие обозначения:

$$t = \frac{\lambda n}{2(n-1)}, \quad a = \mathbb{E}_{h\sim\rho}[\mathbb{D}_n(h)], \quad b = \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{2(n-1)}. \tag{6.9}$$

Используя их, мы можем переписать (6.8) в следующем виде:

$$\mathbb{E}_{h\sim\rho}[\mathbb{D}(h)] \leqslant F(t) = \frac{a}{1-t} + \frac{b}{t(1-t)}, \tag{6.10}$$

где $a, b \geqslant 0$ и $0 < t \leqslant 1$, поскольку мы наложили ограничения $\lambda \leqslant \frac{2(n-1)}{n}$.

Для $t \in (0, 1]$ справедливо:

$$\begin{aligned}
\frac{\partial F}{\partial t} &= \frac{a}{(1-t)^2} - \frac{(1-2t)b}{t^2(1-t)^2}, \\
\frac{\partial^2 F}{\partial t^2} &= \frac{2a}{(1-t)^3} + \frac{2bt^2(1-t)^2 + 2b(2t-1)^2(1-t)t}{t^4(1-t)^4} \geqslant 0.
\end{aligned}$$

Следовательно, функция $F(t)$ является выпуклой на интересующем нас интервале и достигает своего минимума в положительном решении уравнения

$$at^2 + 2bt - b = 0,$$

которое равно

$$t^* = \frac{\sqrt{b^2+ab} - b}{a} = \frac{\sqrt{b}(\sqrt{a+b} - \sqrt{b})}{(a+b) - b} = \frac{\sqrt{b}}{\sqrt{a+b} + \sqrt{b}} = \frac{1}{\sqrt{a/b + 1} + 1} \leqslant \frac{1}{2}. \tag{6.11}$$

Теперь мы покроем интересующий нас интервал последовательностью $(t_k)_{k\in\mathbb{N}^+}$ значений t, образующих геометрическую прогрессию. Ранее в (6.11) мы получили верхнюю границу

интересующего нас интервала. Для поиска нижней границы мы подставим значения a и b в (6.11) и получим:

$$t^* = \left(\sqrt{\frac{2(n-1)\mathbb{E}_{h\sim\rho}[\mathbb{D}_n(h)]}{\mathrm{KL}(\rho\|\pi)+\ln 1/\delta}+1}+1\right)^{-1}.$$

Поскольку $\mathrm{KL}(\rho\|\pi) \geqslant 0$ и $\mathbb{D}_n(h) \leqslant \frac{1}{2}$ (что является простым следствием вспомогательной Леммы 4), мы получаем:

$$t^* \geqslant \left(\sqrt{\frac{n-1}{\ln 1/\delta}+1}+1\right)^{-1}.$$

Поэтому нас интересует следующий интервал значений t:

$$t \in \left[\left(\sqrt{\frac{n-1}{\ln 1/\delta}+1}+1\right)^{-1}, \frac{1}{2}\right].$$

Мы покроем этот интервал последовательностью $t_i = c^i\left(\sqrt{\frac{n-1}{\ln 1/\delta}+1}+1\right)^{-1}$, $i = 0,\ldots,m-1$ для $c > 1$. Чтобы покрыть указанный выше интервал, достаточно взять

$$m = \left\lceil \frac{1}{\ln c}\ln\left(\frac{1}{2}\sqrt{\frac{n-1}{\ln 1/\delta}+1}+\frac{1}{2}\right)\right\rceil$$

первых членов последовательности. Значение t_{m-1} — последнее значение в последовательности, не превосходящее $\frac{1}{2}$. Для любой обучающей выборки мы можем найти значение t_{i^*}, $i^* \in \{0,\ldots,m-1\}$, такое что для него выполнено следующее:

$$t_{i^*} \leqslant t^* \leqslant t_{i^*+1} \leqslant ct_{i^*},$$

где t^* — оптимальное значение, минимизирующее правую часть неравенства (6.10) для конкретного распределения ρ. Пользуясь этим неравенством, мы получаем:

$$\begin{aligned}
F(t_{i^*}) &= \frac{a}{1-t_{i^*}} + \frac{b}{t_{i^*}(1-t_{i^*})} \\
&\leqslant \frac{a}{1-t^*} + \frac{b}{(t^*/c)(1-t^*)} \\
&= \frac{a}{1-\frac{\sqrt{b}}{\sqrt{b}+\sqrt{a+b}}} + \frac{cb}{\left(1-\frac{\sqrt{b}}{\sqrt{b}+\sqrt{a+b}}\right)\frac{\sqrt{b}}{\sqrt{b}+\sqrt{a+b}}} \\
&= \frac{a+cb+c\sqrt{b(a+b)}}{1-\frac{\sqrt{b}}{\sqrt{b}+\sqrt{a+b}}} \\
&= \frac{\left(a+cb+c\sqrt{b(a+b)}\right)\left(\sqrt{a+b}+\sqrt{b}\right)}{\sqrt{a+b}} \\
&= a+cb+c\sqrt{b(a+b)} + \frac{a\sqrt{b}}{\sqrt{a+b}} + \frac{cb\sqrt{b}}{\sqrt{a+b}} + cb \\
&\leqslant a+(1+c)\sqrt{ab}+4cb, \qquad (6.12)
\end{aligned}$$

где в последнем неравенстве мы воспользовались $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$. Подстановка значений a и b завершает доказательство. ■

Замечание 31. *Обратим внимание на то, что оценка (6.7) не состоятельна: ее значение не стремится к $\mathbb{E}_{h\sim\rho}[\mathbb{D}(h)]$ при $n \to \infty$ из-за множителя $c > 1$. Множитель c является следствием использованной в доказательстве техники, основанной на геомтерической прогресси. Вопрос о возможности устранения множителя c из оценки открыт. Тем не менее, для наших целей это не является проблемой: мы можем взять константу с достаточно близкой к единице. Например, когда $c = 1.1$, $n = 500$ и $\delta = 0.05$ мы получаем $\ln \frac{\nu}{\delta} \approx 6$.*

6.3.2 PAC-Байесовское эмпирическое неравенство Бернштейна

Теорема 54 позволяет контролировать усредненное значение дисперсии $\mathbb{E}_{h\sim\rho}[\mathbb{D}(h)]$ одновременно для всех апостериорных распределений ρ. Воспользовавшись неравенством Буля, мы получаем, что если в утверждениях Теорем 50 и 54 положить $\delta = \frac{\delta_1}{2}$, то они будут выполнены одновременно с вероятнсотью не меньше $1 - \delta_1$. Подстановка верхней оценки на $\mathbb{E}_{h\sim\rho}[\mathbb{D}(h)]$ Теоремы 54 в PAC-Байесовское неравенство Бернштейна Теоремы 50 ведет к PAC-Байесовскому эмпирическому неравенству Бернштейна — главному результату настоящего раздела, который ограничивает сверху средний риск стохастического предиктора $L(G_\rho) = \mathbb{E}_{h\sim\rho}[L(h)]$ одновременно для всех распределений ρ.

Теорема 55. *Для любого распределения π на $\mathcal{H}$, не зависящего от обучающей выборки, любых $\delta_1 > 0$ и $c_1, c_2 > 1$, обозначив $\hat{\mathbb{D}}_n(\rho)$ правую часть неравенства (6.7) (с $\delta = \frac{\delta_1}{2}$), и положив $\bar{\mathbb{D}}_n(\rho) = \min\left(\hat{\mathbb{D}}_n(\rho), \frac{1}{4}\right)$, мы получим, что с вероятностью не меньше $1-\delta_1$ (относительно случайной реализации обучающей выборки S) следующее:*

$$L(G_\rho) \leqslant L_n(G_\rho) + (1 + c_1)\sqrt{\frac{(e-2)\bar{\mathbb{D}}_n(\rho)\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{2\nu}{\delta_1}\right)}{n}} \tag{6.13}$$

выполнено одновременно для всех распределений ρ на $\mathcal{H}$, которые удовлетворяют условию

$$\sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{2\nu}{\delta_1}}{(e-2)\bar{\mathbb{D}}_n(\rho)}} \leqslant \sqrt{n},$$

где величина ν была определена в Теореме 50 (с $\delta = \frac{\delta_1}{2}$); для всех остальных распределений ρ справедливо:

$$L(G_\rho) \leqslant L_n(G_\rho) + 2\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{2\nu}{\delta_1}}{n}.$$

Отметим, что все величины, входящие в Теорему 55, могут быть вычислены по обучающей выборке.

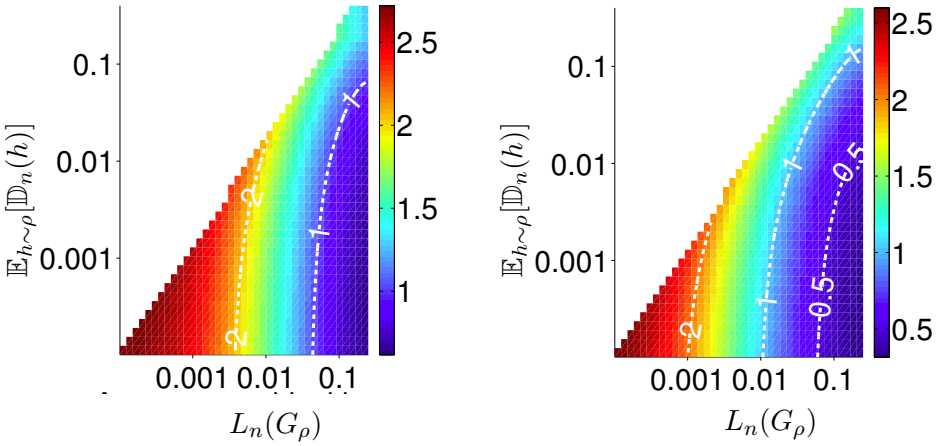
Сравнение главных членов порядка $O(1/\sqrt{n})$ в PAC-Байесовском эмпирическом неравенстве Бернштейна (в дальнейшем PB-EB для краткости) и релаксации PAC-Байесовского kl-неравенства (6.4) показывает, что PB-EB неравенство может оказаться точнее в тех случаях, когда $\mathbb{E}_{h\sim\rho}\left[\mathbb{D}_n(h)\right] \leqslant \big(1/2(e-2)\big)L_n(G_\rho) \approx 0.7L_n(G_\rho)$. Также отметим, что когда потери ограничены в интервале [0,1], выполнено следующее неравенство: $\mathbb{D}_n(h) \leqslant \big(n/(n-1)\big)L_n(h)$ (поскольку $\ell_h(Z)^2 \leqslant \ell_h(Z)$). Поэтому PB-EB оценка никогда не бывает сильно хуже PAC-Байесовского kl-неравенства, а в тех случаях, когда выборочная дисперсия мала по сравнению с эмпирическим риском, она может оказаться существенно точнее. Для бинарной функции потерь ($\ell(y,y') \in \{0,1\}$) выполнено $\mathbb{D}(h) = L(h)\big(1-L(h)\big)$, и в этом случае эмпирический риск не может существенно превосходить выборочную дисперсию, а значит PB-EB не улучшает PB-kl оценку. Также отметим, что PAC-Байесовское kl-неравенство Теоремы 52 в действительности имеет лучшее поведение члена порядка $O(1/n)$, и для достаточно малых размеров обучающих выборок PB-kl неравенство может оказаться точнее PB-EB даже в тех случаях, когда выборочная дисперсия мала. Подводя итоги вышесказанного, мы заключаем, что в тех случаях, когда $\mathbb{E}_{h\sim\rho}\left[\mathbb{D}_n(h)\right] \leqslant 0.7L_n(G_\rho)$, PB-EB неравенство может оказаться значительно точнее PB-kl, а в противном случае они сравнимы (исключая случай очень маленьких размеров выборок). В разделе 6.3.3 мы приводим более детальные численные сравнения этих двух неравенств.

6.3.3 Эксперименты

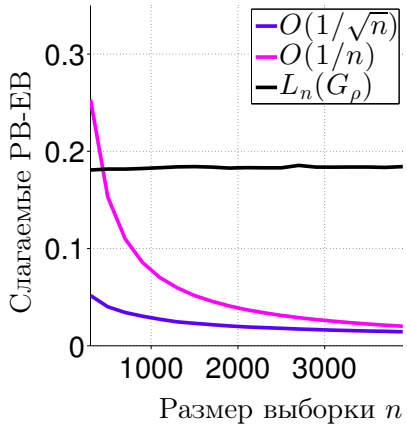
Перед тем как перейти к численным экспериментам, приведем общее сравнение поведения PB-EB и PB-kl оценок, как функций $L_n(G_\rho)$, $\mathbb{E}_{h\sim\rho}\left[\mathbb{D}_n(h)\right]$ и n. На Рисунке 6.2 приведено отношение

$$\frac{\text{PB-EB} - L_n(G_\rho)}{\text{PB-kl} - L_n(G_\rho)}$$

на плоскости $L_n(G_\rho) \times \mathbb{E}_{h\sim\rho}\left[\mathbb{D}_n(h)\right]$ для двух разных значений n. С помощью PB-EB мы обозначили значение PB-EB оценки из неравенства (6.13), а с помощью PB-kl — значение PB-kl оценки из неравенства (6.3). Мы приводим сравнение отношений сложностных членов в оценках (без слагаемого $L_n(G_\rho)$, которое совпадает в двух оценках). В этом примере мы зафиксировали $\mathrm{KL}(\rho\|\pi) = 18$. Во всех экспериментах, представленных в этом разделе, мы используем параметры $c_1 = c_2 = 1.15$ и $\delta = 0.05$. Как было отмечено ранее, PB-EB никогда не бывает существенно хуже PB-kl, а в случаях, когда $\mathbb{E}_{h\sim\rho}\left[\mathbb{D}_n(h)\right] \ll L_n(G_\rho)$, PB-EB может



(a) $n = 1000$ (b) $n = 4000$



(c) $d = 5$, $\dim(\boldsymbol{w}_0) = 2$

Рисунок 6.2: (c) Члены порядков $1/\sqrt{n}$ и $1/n$, входящие в PB-EB оценку, вместе с $L_n(G_\rho)$; (a), (b) отношение разности PB-EB и $L_n(G_\rho)$ к разности PB-kl и $L_n(G_\rho)$ для разных значений n, $\mathbb{E}_{h\sim\rho}[\mathbb{D}_n(h)]$ и $L_n(G_\rho)$. Оси представлены в логарифмических шкалах. PB-EB точнее PB-kl ниже пунктирной линии с меткой 1.

оказаться значительно точнее. Из иллюстрации на Рисунке 6.2 видно, что в худшем случае отношение немного превосходит 2.5, а в лучшем случае оно приближается к 0.3. Отметим, что по мере увеличения размера обучающей выборки максимальное значение отношения убывает (и в пределе стремится к 1.2), в то время как минимальное значение стремится к нулю.

Как уже отмечалось, преимущество PB-EB неравенства над PB-kl неравенством наиболее ощутимо в задачах восстановления регрессии (для задач классификации с бинарной функцией потерь PB-EB оценка сравнима с PB-kl оценкой). Ниже мы приводим численные эксперименты на примере задачи регрессии с абсолютными потерями $\ell(y', y'') = |y' - y''|$. Мы будем использовать модельные выборки и выборки из репозитория UCI [3]. Мы будем использовать PB-EB и PB-kl неравенства для получения оценок среднего риска отображений,

полученных методом минимизации эмпирического риска. Во всех экспериментах объекты X_i лежат в d-мерном шаре единичного радиуса с центром в начале координат ($\|X_i\|_2 \leqslant 1$), а ответы Y принимают значения из интервала $[-0.5, 0.5]$. Множество отображений $\mathcal{H}$ определим следующим образом:

$$\mathcal{H} = \Big\{ h_{\boldsymbol{w}}(x) = \langle \boldsymbol{w}, X \rangle \colon \boldsymbol{w} \in \mathbb{R}^d, \|\boldsymbol{w}\|_2 \leqslant 0.5 \Big\}.$$

Подобные построения гарантируют, что абсолютные потери принимают значения из интервала $[0, 1]$. Мы будем использовать равномерное априорное распределение на $\mathcal{H}$, определяемое как $\pi(\boldsymbol{w}) = \big(V(1/2, d)\big)^{-1}$, где $V(r, d)$ — объем d-мерного шара радиуса r, и равномерное апостериорное распределение $\rho_{\hat{\boldsymbol{w}}}$ на d-мерном шаре радиуса ε с центром в векторе весов $\hat{\boldsymbol{w}}$, где $\hat{\boldsymbol{w}}$ — решение следующей задачи оптимизации:

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \frac{1}{n} \sum_{i=1}^{n} |Y_i - \langle \boldsymbol{w}, X_i \rangle| + \lambda^* \|\boldsymbol{w}\|_2^2. \tag{6.14}$$

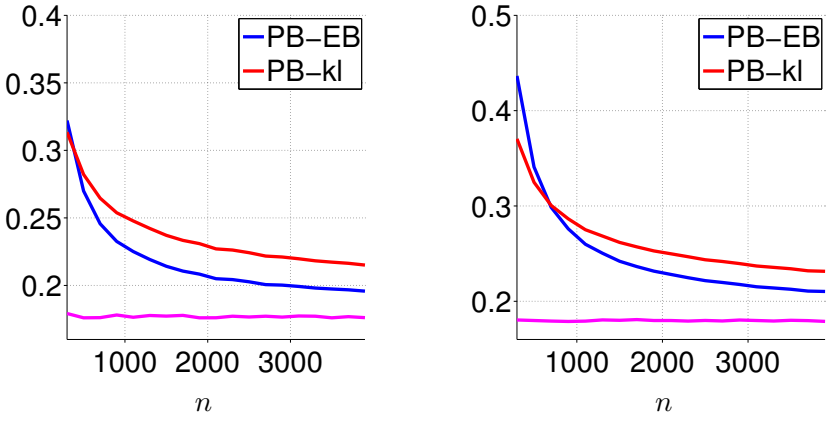
Заметим, что (6.14) — задача квадратичного программирования и может быть эффективно решена с помощью различных пакетов оптимизации (мы использовали Matlab quadprog). Регуляризатор $\lambda^* \|\boldsymbol{w}\|_2^2$ вводится для гарантии того, что апостериорное распределение имеет $\mathcal{H}$ в качестве своего носителя. С помощью бинарного поиска мы найдем минимальное (неотрицательное) значение λ^*, такое что апостериорное распределение $\rho_{\hat{\boldsymbol{w}}}$ имеет $\mathcal{H}$ в качестве своего носителя (что означает, что шар радиуса ε с центром в $\hat{\boldsymbol{w}}$ содержится в шаре радиуса 0.5 с центром в начале координат). Во всех дальнейших экспериментах мы использовали значение $\varepsilon = 0.05$. Все детали реализации описываемых в этом параграфе экспериментов приводятся в разделе 6.3.4

Модельная выборка. Модельные выборки будем генерировать следующим образом. Мы вытягиваем точки $X_1, \ldots, X_n$ из равномерного распределения на d-мерном шаре единичного радиуса ($\|X_i\|_2 \leqslant 1$) с центром в начале координат. Затем мы полагаем

$$Y_i = \sigma_0 \left(50 \cdot \langle \boldsymbol{w}_0, X_i \rangle\right) + \varepsilon_i,$$

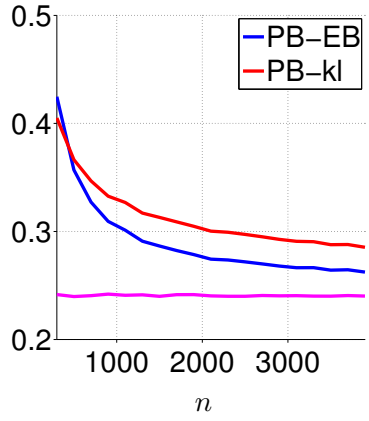
где весовой векторо $\boldsymbol{w}_0 \in \mathbb{R}^d$, отображение $\sigma_0(z) = \frac{1}{1+e^{-z}} - 0.5$ — центрированная сигмоидная функция, принимающая значения в $[-0.5, 0.5]$, а ε_i — равномерно распределенный на интервале $[-a_i, a_i]$ и независимый от X_i шум, где

$$a_i = \begin{cases} \min\big(0.1, 0.5 - \sigma_0(50 \cdot \langle \boldsymbol{w}_0, X_i \rangle)\big), & \text{для } \sigma_0(50 \cdot \langle \boldsymbol{w}_0, X_i \rangle) \geqslant 0; \\ \min\big(0.1, 0.5 + \sigma_0(50 \cdot \langle \boldsymbol{w}_0, X_i \rangle)\big), & \text{для } \sigma_0(50 \cdot \langle \boldsymbol{w}_0, X_i \rangle) < 0. \end{cases}$$

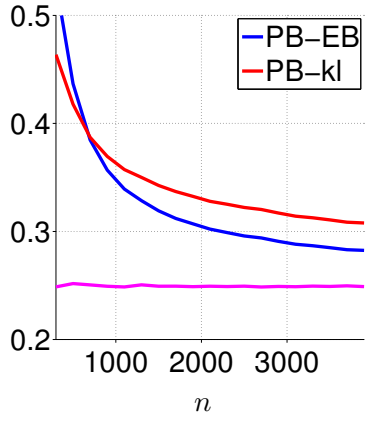


(a) $d = 2$, $\|\boldsymbol{w}_0\|_0 = 2$

(b) $d = 5$, $\|\boldsymbol{w}_0\|_0 = 2$



(c) $d = 3$, $\|\boldsymbol{w}_0\|_0 = 3$



(d) $d = 6$, $\|\boldsymbol{w}_0\|_0 = 3$

Рисунок 6.3: PB-kl оценка и PB-EB оценка вместе с ошибками на контроле (розовые линии) для модельных выборок. Приведенные результаты усреднены по 10 случайным выборкам.

Описанная схема гарантирует, что $Y_i \in [-0.5, 0.5]$. Обозначим j-ю координату вектора $\boldsymbol{u} \in \mathbb{R}^d$ с помощью $\boldsymbol{u}^j$, а число ненулевых координат $\boldsymbol{u}$ с помощью $\|\boldsymbol{u}\|_0$. Мы выбираем весовой вектор $\boldsymbol{w}_0$ так, чтобы у него было всего лишь несколько ненулевых координат. Рассмотрим два случая. В первом положим $d \in \{2, 5\}$, $\|\boldsymbol{w}_0\|_0 = 2$, $\boldsymbol{w}_0^1 = 0.12$ и $\boldsymbol{w}_0^2 = -0.04$. Во втором — $d \in \{3, 6\}$, $\|\boldsymbol{w}_0\|_0 = 3$, $\boldsymbol{w}_0^1 = -0.08$, $\boldsymbol{w}_0^2 = 0.05$ и $\boldsymbol{w}_0^3 = 0.2$. Выбор координат векторов произволен, и наши эксперименты показали, что результаты не чувствительны к нему. Сигмоидная функция делает решение задачи с помощью множества линейных отображений затруднительным. Вместе с относительно небольшим уровнем шума ($\varepsilon_i \leqslant 0.1$) это ведет к малым значениям выборочной дисперсии риска $\mathbb{D}_n(h)$ и относительно большим значениям эмпирического риска $L_n(h)$.

Для каждого размера обучающей выборки от 300 до 4000 мы усреднили результаты по 10 случайным выборкам. Результаты приведены на Рисунке 6.3. Мы видим, что за исключением случаев очень маленьких выборок ($n < 1000$) наша оценка лучше PB-kl оценки.

Таблица 6.1: Результаты для выборок из репозитория UCI

Выборка	n	d	**Обучение**	**Контроль**	**PB-kl оценка**	**PB-EB оценка**
winequality	6497	11	0.106 ± 0.0005	0.106 ± 0.0022	0.175 ± 0.0006	$\mathbf{0.162 \pm 0.0006}$
parkinsons	5875	16	0.188 ± 0.0014	0.188 ± 0.0055	0.266 ± 0.0013	$\mathbf{0.250 \pm 0.0012}$
concrete	1030	8	0.110 ± 0.0008	0.111 ± 0.0038	$\mathbf{0.242 \pm 0.0010}$	0.264 ± 0.0011

Из Рисунка 6.2.а видно, что более плохие результаты в случае очень маленьких выборок объясняются преобладанием члена порядка $O(1/n)$ в PB-EB оценке (6.13). Как только размерн выборки n становится достаточно большим, упомянутый член существенно уменьшается, и наша оценка становится точнее PB-kl оценки.

Выборки из репозитория UCI. Мы приводим сравнение PB-EB неравенства (6.13) с PB-kl неравенством (6.3) для трех относительно больших выборок (задач восстановления регрессии) из репозитория UCI: Wine Quality, Parkinsons Telemonitoring и Concrete Compressive Strength. Для каждой выборки мы нормируем и центрируем и признаковые описания объектов, и ответы, так чтобы $Y_i \in [-0.5, 0.5]$ и $\|X_i\| \leqslant 1$. В Таблице 6.1 приведены общие описания рассмотренных выборок вместе с результатами скользящего контроля с 5-ю разбиениями выборки на обучающую и контрольную подвыборки.

Замечание 32. *Результаты, представленные в данном разделе, открывают ряд интересных направлений дальнейших исследований. Наиболее интересным и важным из них является получение методов обучения, которые явно минимизируют нашу PB-EB оценку. Другим направлением является уменьшение члена порядка $O(1/n)$ в нашей оценке также, как это делается в PAC-Байесовском kl-неравенстве.*

6.3.4 Вспомогательные результаты

Здесь мы приводим доказательства результатов, необходимых для реализации численных экспериментов прошлого раздела, а также ряд других технических деталей. Нам понадобится следующий результат.

Лемма 40. *Рассмотрим случайную велтичину $\xi = \langle \boldsymbol{w}, \boldsymbol{v} \rangle$, где $\boldsymbol{w} \in \mathbb{R}^d$ распределен равномерно в d-мерном шаре радиуса ε с центром в начале координат, и $\boldsymbol{v} \in \mathbb{R}^d$ — фиксированный вектор отличной от нуля и конечной евклидовой нормой $0 < \|\boldsymbol{v}\|_2 \leqslant \infty$. Тогда случайная величина ξ имеет следующую плотность распределения с ограниченным носителем*

$\left[-\varepsilon\|\boldsymbol{v}\|_2, \varepsilon\|\boldsymbol{v}\|_2\right]$:

$$p_\xi(x) = \frac{\left(1 - \frac{x^2}{\varepsilon^2\|\boldsymbol{v}\|_2^2}\right)^{\frac{d-1}{2}}}{N(\boldsymbol{v}, \varepsilon, d)},$$

где

$$N(\boldsymbol{v}, \varepsilon, d) = 2\,\varepsilon\|\boldsymbol{v}\|_2 \int_0^{\frac{\pi}{2}} \cos^d(t)dt.$$

Также

$$\mathbb{E}[\xi] = 0, \quad \mathbb{D}[\xi] = \frac{(\varepsilon\|\boldsymbol{v}\|_2)^2}{d+2}.$$

Доказательство. Для начала заметим, что $\xi \in \left[-\varepsilon\|\boldsymbol{v}\|_2, \varepsilon\|\boldsymbol{v}\|_2\right]$, что является следствием неравенства Коши–Буняковского. Также в силу симметрии носителя $\boldsymbol{w}$ мы можем ограничиться рассмотрением случая, когда у вектора $\boldsymbol{w}$ лишь одна ненулевая координата. Без потери общности, пусть это будет первая координата.

Мы будем обозначать j-ю координату вектора $\boldsymbol{u} \in \mathbb{R}^d$, $j = 1, \ldots, d$, с помощью верхнего индекса $\boldsymbol{u}^j$. Тогда для любых значений C из носителя p_ξ условие $\xi = C$ эквивалентно $\boldsymbol{w}^1 = C/\boldsymbol{v}^1$ и ведет к тому, что $\boldsymbol{w}$ лежит в $(d-1)$-мерном шаре радиуса $\sqrt{\varepsilon^2 - (C/\|\boldsymbol{v}\|_2)^2}$ с центром в начале координат. Тогда очевидно, что

$$p_\xi(x) \propto \left(1 - \frac{x^2}{\varepsilon^2\|\boldsymbol{v}\|_2^2}\right)^{\frac{d-1}{2}}.$$

Нам достаточно найти нормализующий множитель, который мы обозначим $N(\boldsymbol{v}, \varepsilon, d)$:

$$N(\boldsymbol{v}, \varepsilon, d) = \int_{-\varepsilon\|\boldsymbol{v}\|_2}^{\varepsilon\|\boldsymbol{v}\|_2} \left(1 - \frac{x^2}{\varepsilon^2\|\boldsymbol{v}\|_2^2}\right)^{\frac{d-1}{2}} dx = 2\int_0^{\varepsilon\|\boldsymbol{v}\|_2} \left(1 - \frac{x^2}{\varepsilon^2\|\boldsymbol{v}\|_2^2}\right)^{\frac{d-1}{2}} dx.$$

Вводя замену переменных $\sin(t) = \frac{x}{\varepsilon\|\boldsymbol{v}\|_2}$, мы получаем:

$$N(\boldsymbol{v}, \varepsilon, d) = 2\,\varepsilon\|\boldsymbol{v}\|_2 \int_0^{\frac{\pi}{2}} \cos^d(t)dt,$$

что завершает доказательство первой части леммы.

Равенство $\mathbb{E}[\xi] = 0$ следует из факта, что распределение ξ симметрично. Наконец, мы пользуемся следующим выражением для вычисления дисперсии $\mathbb{D}[\xi]$. Оно утверждает, что для любого $m, n \in \mathbb{N}$ справедливо

$$\int \sin^m(t)\cos^n(t)dt = -\frac{\sin^{m-1}(t)\cos^{n+1}(t)}{m+n} + \frac{m-1}{m+n}\int \sin^{m-2}(t)\cos^n(t)dt. \tag{6.15}$$

Поскольку

$$\mathbb{D}[\xi] = \frac{2\int_0^{\varepsilon\|\boldsymbol{v}\|_2} x^2 \left(1 - \frac{x^2}{\varepsilon^2\|\boldsymbol{v}\|_2^2}\right)^{\frac{d-1}{2}} dx}{N(\boldsymbol{v}, \varepsilon, d)},$$

снова обозначив $\sin(t) = \frac{x}{\varepsilon\|\boldsymbol{v}\|_2}$, мы получаем

$$\mathbb{D}[\xi] = \frac{2(\varepsilon\|\boldsymbol{v}\|_2)^3}{N(\boldsymbol{v},\varepsilon,d)}\int_0^{\frac{\pi}{2}} \sin^2(t)\cos^d(t)dt.$$

Пользуясь выражением (6.15), мы получаем

$$\mathbb{D}[\xi] = \frac{2(\varepsilon\|\boldsymbol{v}\|_2)^3}{N(\boldsymbol{v},\varepsilon,d)}\frac{1}{d+2}\int_0^{\frac{\pi}{2}} \cos^d(t)dt = \frac{(\varepsilon\|\boldsymbol{v}\|_2)^2}{d+2}.$$

■

Отметим, что $N(\boldsymbol{v},\varepsilon,d)$ можно легко вычислить, пользуясь следующим выражением:

$$\int \cos^d(t)dt = \frac{1}{d}\cos^{d-1}(t)\sin(t) + \frac{d-1}{d}\int \cos^{d-2}(t)dt.$$

Теперь мы можем вывести все величины, входящие в PAC-Байесовских неравенства, для наших экспериментов. Начнем со следующего результата, который выполнен только для ограниченного множества радиусов ε апостериорного распределения.

Теорема 56. *Для апостериорного и априорного распределений $\rho_{\hat{\boldsymbol{w}}}$ и π, определенных в разделе 6.3.3, выберем радиус апостериорного распределения следующим образом: $\hat{\varepsilon} = \min_{i=1,\dots,n}|Y_i - \langle\hat{\boldsymbol{w}}, X_i\rangle|$, и предположим, что $\hat{\varepsilon} > 0$. Тогда справедливо следующее:*

$$\mathrm{KL}(\rho_{\hat{\boldsymbol{w}}}\|\pi) = d\ln\frac{2}{\hat{\varepsilon}}; \tag{6.16}$$

$$\mathbb{E}_{h\sim\rho_{\hat{\boldsymbol{w}}}}[L_n(h)] = L_n(h_{\hat{\boldsymbol{w}}}); \tag{6.17}$$

$$\mathbb{E}_{h\sim\rho_{\hat{\boldsymbol{w}}}}[\mathbb{D}_n(h)] = \frac{1}{n-1}\sum_{i=1}^n\left((Y_i - \langle\hat{\boldsymbol{w}}, X_i\rangle)^2 + \frac{\hat{\varepsilon}^2}{d+2}\|X_i\|_2^2\right) - \frac{n}{n-1}\big(L_n(h_{\hat{\boldsymbol{w}}})\big)^2 -$$
$$- \frac{\hat{\varepsilon}^2}{4n(n-1)(d+2)}\sum_{i=1}^n\sum_{j=1}^n\langle X_i, X_j\rangle\mathrm{sgn}\big\{(Y_i - \langle\hat{\boldsymbol{w}}, X_i\rangle)(Y_j - \langle\hat{\boldsymbol{w}}, X_j\rangle)\big\}. \tag{6.18}$$

Доказательство. Начнем с вывода (6.16):

$$\mathrm{KL}(\rho\|\pi) = \int_{\|\boldsymbol{w}\|\leqslant\frac{1}{2}}\rho_{\hat{\boldsymbol{w}}}(\boldsymbol{w})\ln\frac{\rho_{\hat{\boldsymbol{w}}}(\boldsymbol{w})}{\pi(w)}d\boldsymbol{w} =$$
$$= \int_{\|\boldsymbol{w}\|\leqslant\frac{1}{2}}\mathbb{1}\{\|\boldsymbol{w}-\hat{\boldsymbol{w}}\|_2\leqslant\hat{\varepsilon}\}\frac{1}{V(\hat{\varepsilon},d)}\ln\frac{V(1/2,d)}{V(\hat{\varepsilon},d)}d\boldsymbol{w} = d\ln\frac{2}{\hat{\varepsilon}},$$

где $V(\varepsilon,d)$ — объем d-мерного шара радиуса ε.

Вспомним определение:

$$\hat{\varepsilon} = \min_{i=1,\dots,n}(|Y_i - \langle\hat{\boldsymbol{w}}, X_i\rangle|).$$

Оно ведет к тому, что для всех $i = 1,\dots,n$ случайные величины $\xi_i = Y_i - \langle\boldsymbol{w}, X_i\rangle$ (где $\boldsymbol{w}\sim\rho_{\hat{\boldsymbol{w}}}$) не меняют своих знаков. Тогда равенство (6.17) следует непосредственно из определения:

$$\mathbb{E}_{h\sim\rho_{\hat{\boldsymbol{w}}}}[L_n(h)] = \frac{1}{n}\sum_{i=1}^n\mathbb{E}_{\boldsymbol{w}\sim\rho_{\hat{\boldsymbol{w}}}}[|Y_i - \langle\boldsymbol{w}, X_i\rangle|] = \frac{1}{n}\sum_{i=1}^n|Y_i - \langle\hat{\boldsymbol{w}}, X_i\rangle|] = L_n(\hat{h}).$$

Наконец, получим (6.18). Пользуясь Леммой 4, запишем

$$\mathbb{E}_{h\sim\rho_{\hat{\boldsymbol{w}}}}[\mathbb{D}_n(h)] = \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{w}\sim\rho_{\hat{w}}}\left[(Y_i - \langle \boldsymbol{w}, X_i\rangle)^2\right] - \frac{1}{n(n-1)}\mathbb{E}_{\boldsymbol{w}\sim\rho_{\hat{w}}}\left[\left(\sum_{i=1}^{n}|Y_i - \langle \boldsymbol{w}, X_i\rangle|\right)^2\right]. \tag{6.19}$$

Теперь заметим, что

$$\mathbb{E}_{\boldsymbol{w}\sim\rho_{\hat{w}}}\left[(Y_i - \langle \boldsymbol{w}, X_i\rangle)^2\right] = \mathbb{E}_{\boldsymbol{w}\sim\rho_{\hat{w}}}\left[(Y_i - \langle \hat{\boldsymbol{w}}, X_i\rangle + \langle \hat{\boldsymbol{w}} - \boldsymbol{w}, X_i\rangle)^2\right] =$$
$$= (Y_i - \langle \hat{\boldsymbol{w}}, X_i\rangle)^2 + 2\mathbb{E}_{\boldsymbol{w}\sim\rho_{\hat{w}}}\left[(Y_i - \langle \hat{\boldsymbol{w}}, X_i\rangle)(\langle \hat{\boldsymbol{w}} - \boldsymbol{w}, X_i\rangle)\right] + \mathbb{E}_{\boldsymbol{w}\sim\rho_{\hat{w}}}\left[(\langle \hat{\boldsymbol{w}} - \boldsymbol{w}, X_i\rangle)^2\right].$$

Второе слагаемое в последнем выражении равно нулю, поскольку $\mathbb{E}_{\boldsymbol{w}\sim\rho_{\hat{w}}}[\boldsymbol{w}] = \hat{w}$. По тем же причинам мы получаем, что

$$\mathbb{E}_{\boldsymbol{w}\sim\rho_{\hat{w}}}\left[(\langle \hat{\boldsymbol{w}} - \boldsymbol{w}, X_i\rangle)^2\right] = \mathbb{D}_{\boldsymbol{w}\sim\rho_{\hat{\boldsymbol{w}}}}[\langle \hat{\boldsymbol{w}} - \boldsymbol{w}, X_i\rangle].$$

Теперь заметим, что вектор $(\hat{\boldsymbol{w}} - \boldsymbol{w}) \in \mathbb{R}^d$ равномерно распределен в d-мерном шаре радиуса $\hat{\varepsilon}$ с центром в начале координат, а также что $\|X_i\|_2 \leqslant 1$. Мы применим Лемму 40 и получим:

$$\mathbb{E}_{\boldsymbol{w}\sim\rho_{\hat{w}}}\left[(\langle \hat{\boldsymbol{w}} - \boldsymbol{w}, X_i\rangle)^2\right] = \frac{(\hat{\varepsilon}\|X_i\|_2)^2}{d+2},$$

что влечет за собой

$$\mathbb{E}_{\boldsymbol{w}\sim\rho_{\hat{w}}}\left[(Y_i - \langle \boldsymbol{w}, X_i\rangle)^2\right] = (Y_i - \langle \hat{\boldsymbol{w}}, X_i\rangle)^2 + \frac{(\hat{\varepsilon}\|X_i\|_2)^2}{d+2}. \tag{6.20}$$

Наконец,

$$\mathbb{E}_{\boldsymbol{w}\sim\rho_{\hat{w}}}\left[\left(\sum_{i=1}^{n}|Y_i - \langle \boldsymbol{w}, X_i\rangle|\right)^2\right] = \mathbb{D}_{\boldsymbol{w}\sim\rho_{\hat{w}}}\left[\sum_{i=1}^{n}|Y_i - \langle \boldsymbol{w}, X_i\rangle|\right] + \left(\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{w}\sim\rho_{\hat{w}}}[|Y_i - \langle \boldsymbol{w}, X_i\rangle|]\right)^2 =$$
$$= \mathbb{D}_{\boldsymbol{w}\sim\rho_{\hat{w}}}\left[\sum_{i=1}^{n}|Y_i - \langle \boldsymbol{w}, X_i\rangle|\right] + \left(nL_n(\hat{h})\right)^2, \tag{6.21}$$

где мы воспользовались (6.17). Для любой последовательности случайных величин $\xi_1, \ldots, \xi_n$ справедливо следующее:

$$\mathbb{D}\left[\sum_{i=1}^{n}\xi_i\right] = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\Big[(\xi_i - \mathbb{E}[\xi_i])(\xi_j - \mathbb{E}[\xi_j])\Big].$$

Пользуясь

$$\xi_i = |Y_i - \langle \boldsymbol{w}, X_i\rangle| = |Y_i - \langle \hat{\boldsymbol{w}}, X_i\rangle + \langle \hat{\boldsymbol{w}} - \boldsymbol{w}, X_i\rangle|,$$

мы получаем

$$\mathbb{D}_{\boldsymbol{w}\sim\rho_{\hat{w}}}\left[\sum_{i=1}^{n}|Y_i - \langle \boldsymbol{w}, X_i\rangle|\right] =$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}_{\boldsymbol{w}\sim\rho_{\hat{w}}}\Big[\big(\xi_i - |Y_i - \langle \hat{\boldsymbol{w}}, X_i\rangle|\big)\big(\xi_j - |Y_j - \langle \hat{\boldsymbol{w}}, X_j\rangle|\big)\Big], \tag{6.22}$$

где мы снова воспользовались тем, что случайные величины $Y_i - \langle \boldsymbol{w}, X_i \rangle$, $i = 1, \ldots, n$, не меняют своих знаков. Поскольку для любых $a, b \in \mathbb{R}$, таких что $|b| \geqslant |a|$, справедливо $|b + a| - |b| = \operatorname{sgn}\{b\} \cdot a$, мы можем записать

$$\xi_i - |Y_i - \langle \hat{\boldsymbol{w}}, X_i \rangle| = \operatorname{sgn}\big\{Y_i - \langle \hat{\boldsymbol{w}}, X_i \rangle\big\} \langle \hat{\boldsymbol{w}} - \boldsymbol{w}, X_i \rangle.$$

Тогда мы получаем

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{w} \sim \rho_{\hat{w}}} \Big[\big(\xi_i - |Y_i - \langle \hat{\boldsymbol{w}}, X_i \rangle|\big) \big(\xi_j - |Y_j - \langle \hat{\boldsymbol{w}}, X_j \rangle|\big) \Big] = \\ &= \operatorname{sgn}\big\{(Y_i - \langle \hat{\boldsymbol{w}}, X_i \rangle)(Y_j - \langle \hat{\boldsymbol{w}}, X_j \rangle)\big\} \mathbb{E}_{\boldsymbol{w} \sim \rho_{\hat{w}}} \Big[\langle \hat{\boldsymbol{w}} - \boldsymbol{w}, X_i \rangle \langle \hat{\boldsymbol{w}} - \boldsymbol{w}, X_j \rangle \Big]. \end{aligned} \tag{6.23}$$

Воспользуемся тем, что

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{w} \sim \rho_{\hat{w}}} \Big[\langle \hat{\boldsymbol{w}} - \boldsymbol{w}, X_i \rangle \langle \hat{\boldsymbol{w}} - \boldsymbol{w}, X_j \rangle \Big] = \\ &= \frac{1}{4} \mathbb{E}_{\boldsymbol{w} \sim \rho_{\hat{w}}} \Big[\big(\langle \hat{\boldsymbol{w}} - \boldsymbol{w}, X_i + X_j \rangle\big)^2 - \big(\langle \hat{\boldsymbol{w}} - \boldsymbol{w}, X_i - X_j \rangle\big)^2 \Big] = \\ &= \frac{1}{4} \mathbb{D}_{\boldsymbol{w} \sim \rho_{\hat{w}}} \Big[\langle \hat{\boldsymbol{w}} - \boldsymbol{w}, X_i + X_j \rangle \Big] - \frac{1}{4} \mathbb{D}_{\boldsymbol{w} \sim \rho_{\hat{w}}} \Big[\langle \hat{\boldsymbol{w}} - \boldsymbol{w}, X_i - X_j \rangle \Big]. \end{aligned}$$

Снова заметив, что вектор $(\hat{\boldsymbol{w}} - \boldsymbol{w}) \in \mathbb{R}^d$ распределен равномерно в d-мерном шаре радиуса $\hat{\varepsilon}$ с центром в начале координат, а также что $\|X_i - X_j\|_2 \leqslant 1$ и $\|X_i + X_j\|_2 \leqslant 1$, мы можем применить Лемму 40 и получаем:

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{w} \sim \rho_{\hat{w}}} \Big[\langle \hat{\boldsymbol{w}} - \boldsymbol{w}, X_i \rangle \langle \hat{\boldsymbol{w}} - \boldsymbol{w}, X_j \rangle \Big] = \\ &= \frac{(\hat{\varepsilon} \|X_i + X_j\|_2)^2}{4(d+2)} - \frac{(\hat{\varepsilon} \|X_i - X_j\|_2)^2}{4(d+2)} = \frac{\hat{\varepsilon}^2 \langle X_i, X_j \rangle}{4(d+2)}. \end{aligned} \tag{6.24}$$

Объединяя (6.19)–(6.24) мы завершаем доказательство. ■

Заметим, что выбор $\hat{\varepsilon} = \min_{i=1,\ldots,n} |Y_i - \langle \hat{\boldsymbol{w}}, X_i \rangle|$ может привести к чересчур большим значениям $\mathrm{KL}(\rho_{\hat{\boldsymbol{w}}}, \pi)$ вследствие тождества (6.16). Мы можем справиться с этой сложностью с помощью следующей теоремы, которая позволяет нам выбрать произвольные значения ε.

Теорема 57. *Пусть n_ε — число точек, для которых $|Y_i - \langle \hat{\boldsymbol{w}}, X_i \rangle| < \varepsilon$. Тогда для апостериорного и априорного распределений $\rho_{\hat{\boldsymbol{w}}}$ и π, определенных в разделе 6.3.3, справедливо следу-*

ющее:

$$\mathrm{KL}(\rho_{\hat{\boldsymbol{w}}}\|\pi) = d\ln\frac{2}{\varepsilon};$$

$$\mathbb{E}_{h\sim\rho_{\hat{\boldsymbol{w}}}}[L_n(h)] \leqslant L_n(h_{\hat{\boldsymbol{w}}}) + \varepsilon\frac{n_\varepsilon}{n};$$

$$\mathbb{E}_{h\sim\rho_{\hat{\boldsymbol{w}}}}[\mathbb{D}_n(h)] \leqslant \frac{1}{n-1}\sum_{i=1}^{n}\left((Y_i - \langle\hat{\boldsymbol{w}}, X_i\rangle)^2 + \frac{\varepsilon^2}{d+2}\|X_i\|_2^2\right) -$$

$$- \frac{1}{n(n-1)}\left(\sum_{i=1}^{n}\mathbb{1}\{|Y_i - \langle\hat{\boldsymbol{w}}, X_i\rangle| \geqslant \varepsilon\}\big[|Y_i - \langle\hat{\boldsymbol{w}}, X_i\rangle|\big]\right)^2 - \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{i,j};$$

$$\gamma_{i,j} = \begin{cases} \mathrm{sgn}\big\{(Y_i - \langle\hat{\boldsymbol{w}}, X_i\rangle)(Y_j - \langle\hat{\boldsymbol{w}}, X_j\rangle)\big\}\frac{\varepsilon^2\langle X_i, X_j\rangle}{4(d+2)}, & \text{если } A_i \cap A_j; \\ -\frac{\varepsilon^2\|X_i\|_2}{\sqrt{d+2}}, & \text{если } A_i \cap A_j^c; \\ -\frac{\varepsilon^2\|X_j\|_2}{\sqrt{d+2}}, & \text{если } A_i^c \cap A_j; \\ -\varepsilon^2, & \text{если } A_i^c \cap A_j^c, \end{cases}$$

где мы определили события $A_i = \big\{|Y_i - \langle\hat{\boldsymbol{w}}, X_i\rangle| \geqslant \varepsilon\big\}$, *а* A^c *— дополнение события* A.

Доказательство. Доказательство полностью повторяет доказательство Теоремы 56 с точностью до незначительных изменений. Главное отличие заключается в том, что теперь для индексов i, для которых $|Y_i - \langle\hat{\boldsymbol{w}}, X_i\rangle| < \varepsilon$, случайные величины $\xi_i = Y_i - \langle\boldsymbol{w}, X_i\rangle$ меняют свой знак при варьировании $\boldsymbol{w}$. По этой причине для таких ξ_i математическое ожидание $\mathbb{E}[|\xi_i|]$ более не равно $|Y_i - \langle\hat{\boldsymbol{w}}, X_i\rangle|$, а имеет более сложный вид. Вместо вычисления точного значения $\mathbb{E}_{h\sim\rho_{\hat{\boldsymbol{w}}}}[L_n(h)]$ мы ограничим $\mathbb{E}[|\xi_i|]$ сверху для таких i, пользуясь следующими рассуждениями:

$$\mathbb{E}_{\boldsymbol{w}\sim\rho_{\hat{\boldsymbol{w}}}}[|Y_i - \langle\boldsymbol{w}, X_i\rangle|] = \mathbb{E}_{\boldsymbol{w}\sim\rho_{\hat{\boldsymbol{w}}}}[|Y_i - \langle\hat{\boldsymbol{w}}, X_i\rangle + \langle\hat{\boldsymbol{w}} - \boldsymbol{w}, X_i\rangle|] \leqslant$$

$$\leqslant |Y_i - \langle\hat{\boldsymbol{w}}, X_i\rangle| + \varepsilon\|X_i\|_2 \leqslant |Y_i - \langle\hat{\boldsymbol{w}}, X_i\rangle| + \varepsilon,$$

что завершает доказательство для $\mathbb{E}_{h\sim\rho_{\hat{\boldsymbol{w}}}}[L_n(h)]$.

Мы также получим верхнюю оценку для $\mathbb{E}_{h\sim\rho_{\hat{\boldsymbol{w}}}}[\mathbb{D}_n(h)]$. Для этого нам нужна нижняя оценка для второго члена в правой части (6.19) (первый член не меняется по сравнению с Теоремой 56). Мы будем использовать следующую нижнюю оценку для члена из (6.21):

$$\left(\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{w}\sim\rho_{\hat{w}}}[|Y_i - \langle\boldsymbol{w}, X_i\rangle|]\right)^2 \geqslant \left(\sum_{i=1}^{n}\mathbb{1}\{|Y_i - \langle\hat{\boldsymbol{w}}, X_i\rangle| \geqslant \varepsilon\}[|Y_i - \langle\hat{\boldsymbol{w}}, X_i\rangle|]\right)^2.$$

Также нам нужна следующая оценка снизу для дисперсии:

$$\mathbb{D}_{\boldsymbol{w}\sim\rho_{\hat{w}}}\left[\sum_{i=1}^{n}|Y_i - \langle\boldsymbol{w}, X_i\rangle|\right],$$

которую мы получаем с помощью нижних оценок ковариаций, входящих в (6.22), соответствующих тем парам (i,j), для которых выполнено хотя бы одно из двух условий $|Y_i - \langle \hat{\boldsymbol{w}}, X_i\rangle| < \varepsilon$, $|Y_j - \langle \hat{\boldsymbol{w}}, X_j\rangle| < \varepsilon$ (для всех остальных членов мы получили точные выражения в прошлом доказательстве). Известно, что для случайных величин ξ, η с конечными дисперсиями справедливо следующее:

$$\big|\mathbb{E}\big[(\xi - \mathbb{E}[\xi])(\eta - \mathbb{E}[\eta])\big]\big| \leqslant \sqrt{\mathbb{D}[\xi]\mathbb{D}[\eta]},$$

что ведет к

$$\mathbb{E}\big[(\xi - \mathbb{E}[\xi])(\eta - \mathbb{E}[\eta])\big] \geqslant -\sqrt{\mathbb{D}[\xi]\mathbb{D}[\eta]}.$$

Заметим, что если $|Y_i - \langle \hat{\boldsymbol{w}}, X_i\rangle| < \varepsilon$, то $|Y_i - \langle \boldsymbol{w}, X_i\rangle| \leqslant 2\varepsilon$ и мы получаем

$$\mathbb{D}_{\boldsymbol{w}\sim\rho_{\hat{\boldsymbol{w}}}}[|Y_i - \langle \boldsymbol{w}, X_i\rangle|] \leqslant \varepsilon^2,$$

где мы использовали тот факт, что для случайных величин $\xi \in [0,1]$ справедливо $\mathbb{D}[\xi] \leqslant 1/4$.

Если $|Y_i - \langle \hat{\boldsymbol{w}}, X_i\rangle| \geqslant \varepsilon$, мы получаем

$$\mathbb{D}_{\boldsymbol{w}\sim\rho_{\hat{\boldsymbol{w}}}}[|Y_i - \langle \boldsymbol{w}, X_i\rangle|] = \mathbb{D}_{\boldsymbol{w}\sim\rho_{\hat{\boldsymbol{w}}}}[\langle \hat{\boldsymbol{w}} - \boldsymbol{w}, X_i\rangle] = \frac{(\varepsilon\|X_i\|_2)^2}{d+2},$$

что завершает доказательство. ■

Замечание 33 (Комментарии к разделу 6.3.3)**.** *Заметим, что для апостериорных распределений $\rho_{\hat{\boldsymbol{w}}}$, определенных в разделе 6.3.3, очевидным образом выполнено $B(x, \rho_{\hat{\boldsymbol{w}}}) = \mathbb{E}_{\boldsymbol{w}\sim\rho_{\hat{\boldsymbol{w}}}}[h_{\boldsymbol{w}}(x)] = h_{\hat{\boldsymbol{w}}}(x)$, что ведет к тому, что взвешенная (Байесовская) функция регрессии совпадает с отображением $h_{\hat{\boldsymbol{w}}}$. Также заметим, что выпуклость L_1-потерь ведет к*

$$\mathbb{E}_{(X,Y)\sim\mathcal{D}}\big[|B(X,\rho) - Y|\big] \leqslant \mathbb{E}_{(X,Y)\sim\mathcal{D}}\mathbb{E}_{\boldsymbol{w}\sim\rho}\big[|h_{\boldsymbol{w}}(X) - Y|\big] = L(G_\rho)$$

для произвольных распределений ρ. Вместе два этих факта ведут к тому, что любая верхняя оценка среднего риска функции регрессии Гиббса, описываемой распределением $\rho_{\hat{\boldsymbol{w}}}$, одновременно является верхней оценкой среднего риска неслучайного отображения $h_{\hat{\boldsymbol{w}}}$. То же справедливо и в тех случаях, когда мы используем квадратичную или любую другую выпуклую функцию потерь.

Коротко подведем итоги настоящей главы:

- PAC-Байесовский анализ изучает рандомизированные отображения в рамках индуктивной постановки теории статистического обучения.

- Минимизация PAC-Байесовских оценок ведет к новым процедурам обучения.
- Часто рандомизированные отображения удается связать с обыкновенными детерминированными отображениями. В этом случае PAC-Байесовские неравенства дают возможность получать оценки обобщающей способности известных алгоритмов обучения.

В настоящей главе получены следующие новые результаты:

- Получено новое PAC-Байесовское неравенство для дисперсии (Теорема 54), которое позволяет оценить сверху неизвестное усредненное значение дисперсии потерь с помощью усредненного значения выборочной дисперсии. Новая оценка полностью вычислима на основе наблюдаемых данных.
- На основе нее получено новое PAC-Байесовское эмпирическое неравенство Бернштейна (Теорема 55), полностью вычислимое на основе наблюдаемых данных и во многих случаях ведущее к существенно более точным оценкам по сравнению с известными ранее результатами.
- Результаты серии экспериментов согласуются с представленным теоретическим анализом.

Заключение

Основные результаты диссертационной работы заключаются в следующем:

1. В Разделе 2.3 получено два новых неравенства концентрации для супремумов эмпирических процессов и выборок без возвращений (Теоремы 20 и 20). Оба неравенства учитывают дисперсию случайных величин, что отличает их от всех известных ранее результатов (неравенств типа МакДиармида работ [25,34]). Доказательство Теоремы 20 основано на неравенстве С. Г. Бобкова [17], а точнее — на его новой модификации, полученной в Теореме 21. Подробный теоретический анализ, представленный в конце Раздела 2.3, показывает, что во многих интересных случаях новые неравенства существенно превосходят точность предыдущих результатов.

2. В Главе 4 на основе новых неравенств концентрации Раздела 2.3 получены первые оценки для трансдуктивного обучения, основанные на локальных мерах сложности рассматриваемых семейств. В первой части Раздела 4.2 получены новые неравенства обобщающей способности (Теоремы 39 и 40), учитывающие дисперсии потерь и справедливые для произвольных классов отображений и огрониченных функций потерь. Важно отметить, что прошлые известные в литературе аналоги накладывали дополнительные ограничения на рассматриваемый класс задач, позволяя работать лишь с конечными (счетными) классами отображений или бинарными функциями потерь. Далее получен результат (Следствие 12), показывающий, что математическое ожидание супремума эмпирических процессов в трансдуктивном обучении (для выборок *без возвращений*) может быть оценено сверху его индуктивным аналогом (для выборок *с возвращениями*). Этот результат позволяет применять для анализа сложности класса отображений в трансдуктивной постановке, так же как и в индуктивной, все известные результаты из теории эмпирических процессов, включая неравесntва симметризации и сжатия. Наконец, в конце раздела получены первые оценки избыточного риска в трансдуктивном обучении (Теоремы 41 и 42 и Следствия 13 и 14), основанные на локальных мерах сложности и справедливые в очень общих предположениях на рассматриваемый класс отоб-

ражений. Эти результаты можно считать аналогами известных результатов работы [5] (в частности — Следствия 5.3 указанной работы). Новые результаты в ряде интересных случаев ведут к первым в трансдуктивном обучении оценкам, имеющим быструю скорость сходимости: для классов с конечной размерностью Вапника–Червоненкиса это следует из Теоремы 37 П. Массара, а для классов в RKHS, соответствующих спрямляющему ядру, — из Следствия 15.

3. В Главе 5 рассматривается комбинаторная теория переобучения. В Разделе 5.2.2 предложено учитывать в рамках теоретико-группового подхода орбиты действия группы симметрий множества векторов ошибок на подвыборки (разбиения) генеральной совокупности. Эта идея развита в Леммах 21 и 22, позволяющих получить новую формулу разложения вероятности переобучения по орбитам разбиений (Теорема 44). Кроме того, при поиске группы симметрий множества векторов ошибок предлагается ограничиться поиском произвольной ее подгруппы (Лемме 20). В отличие от известной до этого формулы разложения по орбитам векторов ошибок, предложенной А. Фреем в работах [122, 123], новая формула позволяет существенно упростить вычисление вероятности переобуения для ряда модельных множеств векторов ошибок. В частности, в Разделе 5.2.4 на ее основе получены новые точные (не заышенные) оценки вероятности переобучения для трех модельных семейств (Теоремы 45, 47 и 48), являющихся различными подмножествами шара в Булевом кубе. Новые оценки еще раз подтверждают хорошо известный в комбинаторной теории переобучения факт о том, что получение не сильно завышенных оценок невозможно без тщательного учета геометрических и структурных свойств множества бинарных векторов ошибок.

4. В Разделе 6.3 получено новое PAC-Байесовское эмпирическое неравенство Бернштейна (Теорема 55), полностью вычислимое на основе наблюдаемых данных и во многих случаях ведущее к существенно более точным оценкам по сравнению с известными ранее результатами. Новое неравенство основано на полученном в Разделе 6.3.1 новом PAC-Байесовском неравенстве для дисперсии (Теорема 54), которое позволяет оценить сверху неизвестное усредненное значение дисперсии потерь с помощью усредненного значения выборочной дисперсии равномерно по классу всех усредняющих распределений. В Разделе 6.3.3 приводятся результаты серии экспериментов, использующих модельные выборки и реальные выборки из репозитория UCI, которые согласуются с теоретически обоснованным выводом о превосходстве нового неравенства над известными аналогами.

Список рисунков

Список таблиц

Литература

1. Adamczak R. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains // Electronic Journal of Probability. 2008. Vol. 34, no. 13.

2. Alina Beygelzimer Sanjoy Dasgupta J. L. Importance weighted active learning // Proceedings of the 26th Annual International Conference on Machine Learning. 2009. P. pp. 49–56.

3. Asuncion A., Newman D. UCI Machine Learning Repository. 2007. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

4. Bardenet R., Maillard O.-A. Concentration inequalities for sampling without replacement. http://arxiv.org/abs/1309.4029. 2013.

5. Bartlett P., Bousquet O., Mendelson S. Local Rademacher Complexities // The Annals of Statistics. 2005. Vol. 33, no. 4. p. 1497–1537.

6. Bartlett P. L., Jordan M. I., McAuliffe J. D. Convexity, Classification, and Risk Bounds // Journal of the American Statistical Association. 2006. Vol. 101.

7. Bartlett P. L., Mendelson S. Rademacher and Gaussian complexities: Risk bounds and structural results // Proceedings of the International Conference on Computational Learning Theory (COLT). 2001.

8. Bartlett P. L., Long P. M., Williamson R. C. Fat-Shattering and the Learnability of Real-Valued Functions. // J. Comput. Syst. Sci. 1996. Vol. 52, no. 3. P. 434–452.

9. Bartlett P. L., Boucheron S., Lugosi G. Model selection and error estimation. // Machine Learning. 2001.

10. Bartlett P. L., Mendelson S., Phillips P. On the Optimality of Sample-Based Estimates of the Expectation of the Empirical Minimizer. // ESAIM: Probability and Statistics. 2010.

11. Bennett G. Probability inequalities for the sum of independent random variables // Journal of the American Statistical Association. 1962. Vol. 57.

12. Bernstein S. N. Probability Theory. 4^{th} edition. Moscow-Leningrad, 1946. In Russian.

13. Beygelzimer A., Hsu D., Langford J., Tong Z. Agnostic Active Learning Without Constraints // Advances in Neural Information Processing Systems 23. 2010. P. 199–207.

14. Bishop C. M. Pattern Recognition and Machine Learning. Springer, 2006.

15. Blum A., Langford J. PAC-MDL Bounds // Proceedings of the International Conference on Computational Learning Theory (COLT). 2003.

16. Blumer A., Ehrenfueucht A., Haussler D., Warmuth M. Occam's razor // Information Processing Letters. 1987. Vol. 24. P. 377–380.

17. Bobkov S. Concentration of normalized sums and a central limit theorem for noncorrelated random variables // Annals of Probability. 2004. Vol. 32.

18. Boucheron S., Lugosi G., Bousquet O. Theory of Classification: a Survey of Recent Advances. // ESAIM: Probability and Statistics. 2005.

19. Boucheron S., Lugosi G., Massart P. Concentration Inequalities: A Nonasymptotic Theory of Independence. Oxford University Press, 2013.

20. Bousquet O. A Bennett Concentration Inequality and Its Application to Suprema of Empirical Processes // C. R. Acad. Sci. Paris, Ser. I. 2002. Vol. 334. P. 495–500.

21. Bousquet O. Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms. Ph.D. thesis: Ecole Polytechnique. 2002.

22. Bousquet O., Elisseeff A. Stability and Generalization // Journal of Machine Learning Research. 2002. March. Vol. 2. P. 499–526.

23. Bousquet O., Boucheron S., Lugosi G. Introduction to statistical learning theory. // Lecture Notes in Artificial Intelligence. 2004.

24. Cortes C., Mohri M. On transductive regression // Advances in Neural Information Processing Systems (NIPS). 2006.

25. Cortes C., Mohri M., Pechyony D., Rastogi A. Stability Analysis and Learning Bounds for Transductive Regression Algorithms. http://arxiv.org/abs/0904.0814. 2009.

26. Cortes C., Vapnik V. Support-Vector Networks // Mach. Learn. Hingham, MA, USA, 1995. September. Vol. 20. P. 273–297.

27. Cover T. M., Thomas J. A. Elements of Information Theory. John Wiley & Sons, 1991.

28. Cristianini N., Shawe-Taylor J. An Introduction to Support Vector Machines. Cambridge University Press, 2000.

29. Derbeko P., El-Yaniv R., Meir R. Explicit Learning Curves for Transduction and Application to Clustering and Compression Algorithms // Journal of Artificial Intelligence Research. 2004. Vol. 22.

30. Devroye L., Lugosi G. Lower Bounds in Pattern Recognition and Learning // Pattern Recognition. 1995. Vol. 28.

31. Devroye L., Györfi L., Lugosi G. A Probabilistic Theory of Pattern Recognition. Springer, 1996.

32. Donsker M. D., Varadhan S. S. Asymptotic evaluation of certain Markov process expectations for large time. // Communications on Pure and Applied Mathematics. 1975. Vol. 28.

33. El-Yaniv R., Pechyony D. Stable Transductive Learning // Proceedings of the International Conference on Computational Learning Theory (COLT). 2006.

34. El-Yaniv R., Pechyony D. Transductive Rademacher Complexity and its Applications // Journal of Artificial Intelligence Research. 2009.

35. Feller. An Introduction to Probability Theory and its Applications. New York, 1971.

36. Freund Y., Schapire R. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting // Journal of Computer and System Sciences. 1997. Vol. 55.

37. Frey A., Tolstikhin I. Combinatorial bounds on probability of overfitting based on clustering and coverage of classifiers // Machine Learning and Data Analysis (JMLDA). 2013. Vol. 1, no. 6. P. 761–778.

38. Germain P., Lacasse A., Laviolette F., Marchand M. PAC-Bayes Risk Bounds for General Loss Functions // Advances in Neural Information Processing Systems (NIPS). 2006.

39. Germain P., Lacasse A., Laviolette F., Marchand M. PAC-Bayesian Learning of Linear Classifiers // Proceedings of the International Conference on Machine Learning (ICML). 2009.

40. Gross D., Nesme V. Note on sampling without replacing from a fnite collection of matrices. http://arxiv.org/abs/1001.2738v2. 2010.

41. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Springer, 2001.

42. Higgs M., Shawe-Taylor J. A PAC-Bayes Bound for Tailored Density Estimation // Proceedings of the International Conference on Algorithmic Learning Theory (ALT). 2010.

43. Hoeffding W. Probability inequalities for sums of bounded random variables // Journal of the American Statistical Association. 1963. Vol. 58, no. 301. P. 13–30.

44. Juditsky A., Rigollet P., Tsybakov A. Learning by Mirror Averaging // Annals of Statistics. 2008. Vol. 36, no. 5.

45. Kearns M. J., Schapire R. E. Efficient Distribution-Free Learning of Probabilistic Concepts (extended abstract) // Proceedings of the 31^{st} Symposium on the Foundations of Computer Science. 1990.

46. Klein T., Rio E. Concentration around the mean for maxima of empirical processes // The Annals of Probability. 2005. Vol. 33, no. 3. p. 1060–1077.

47. Koltchinskii V. Local Rademacher Complexities and Oracle Inequalities in Risk Minimization // The Annals of Statistics. 2006. Vol. 34, no. 6. p. 2593–2656.

48. Koltchinskii V. Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008. Ecole d'été de probabilités de Saint-Flour. Springer, 2011.

49. Koltchinskii V., Panchenko D. Empirical margin distributions and bounding the generalization error of combined classifiers // The Annals of Statistics. 2002. Vol. 30.

50. Koltchinskii V. Oracle inequalities in empirical risk minimization and sparse recovery problems. École d'été de probabilités de Saint-Flour XXXVIII-2008. Springer Verlag, 2011.

51. Koltchinskii V. Rademacher penalties and structural risk minimization // IEEE Transactions on Information Theory. 2001.

52. Koltchinskii V., Panchenko D. Rademacher processes and bounding the risk of function learning // High Dimensional Probability, II / Ed. by D. E. Gine, J.Wellner. Birkhauser, 1999. P. 443–457.

53. Koltchinskii V., Lounici K., Tsybakov A. B. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion // Annals of Statistics. 2011. Vol. 39. P. 2302–2329.

54. Kumar S., Mohri M., Talwalkar A. Sampling Methods for the NyströM Method // Journal of Machine Learning Research. 2012. Vol. 13, no. 1. P. 981–1006.

55. Langford J. Tutorial on Practical Prediction Theory for Classification // Journal of Machine Learning Research. 2005.

56. Langford J., Seeger M. Bounds for Averaging Classifiers: Tech. Rep.: : Carnegie Mellon, 2001.

57. Langford J., Shawe-Taylor J. PAC-Bayes & Margins // Advances in Neural Information Processing Systems (NIPS). 2002.

58. Ledoux M. On Talagrand's Deviation Inequalities for Product Measures // ESAIM: Probability and Statistics. 1996.

59. Ledoux M., Talagrand M. Probability in Banach Space. Springer-Verlag, 1991.

60. Lee W., Bartlett P. L., Williamson R. C. The Importance of Convexity in Learning with Squared Loss // IEEE Transactions on Information Theory. 1998. Vol. 44.

61. Lugosi G., Wegkamp M. Complexity regularization via localized random penalties // Annals of Statistics. 2004. Vol. 32, no. 4.

62. Mammen E., Tsybakov A. Smooth Discrimination Analisys // Annals of Statistics. 1999. Vol. 27. P. 1808–1829.

63. Marton K. A simple proof of the blowing-up lemma // IEEE Transactions on Information Theory. 1986. Vol. 32.

64. Massart P. Concentration Inequalities and Model Selection: École D'Été de Probabilités de Saint-Flour 2003. Ecole d'été de probabilités de Saint-Flour. Springer, 2007.

65. Massart P., Nédélec E. Risk Bounds for Statistical Learning // The Annals of Statistics. 2006. Vol. 34, no. 5. P. 2326–2366.

66. Massart P. Some applications of concentration inequalities to statistics // Ann. Fac. Sci. Toulouse Math. 2000. Vol. 9, no. 6. P. 245–303.

67. Massart P. About the constants in Talagrand's concentration inequalities for empiricsl processes. // Ann. Prob. 2000.

68. Massart P. Rates of convergence in the central limit theorem for empirical processes // Annales de l'institut Henri Poincaré (B) Probabiliés et Statistiques. 1986. Vol. 22, no. 4. P. 381–423.

69. Maurer A. A Note on the PAC-Bayesian Theorem. www.arxiv.org. 2004.

70. Maurer A. Concentration inequalities for functions of independent variables // Random Structures and Algorithms. 2006. Vol. 29, no. 2.

71. Maurer A., Pontil M. Empirical Bernstein Bounds and Sample Variance Penalization // Proceedings of the International Conference on Computational Learning Theory (COLT). 2009.

72. McAllester D. PAC-Bayesian Stochastic Model Selection // Machine Learning. 2003. Vol. 51, no. 1.

73. McAllester D. Some PAC-Bayesian Theorems // Proceedings of the International Conference on Computational Learning Theory (COLT). 1998.

74. McAllester D. Some PAC-Bayesian Theorems // Machine Learning. 1999. Vol. 37.

75. McAllester D. PAC-Bayesian Model Averaging // Proceedings of the International Conference on Computational Learning Theory (COLT). 1999.

76. McDiarmid C. On the method of bounded differences // Surveys in Combinatorics. Cambridge: Cambridge University Press, 1989. P. 148–188.

77. Mendelson S. A Few Notes on Statistical Learning Theory. // Lecture Notes in Computer Science. 2003.

78. Mendelson S. Learning without Concentration. http://arxiv.org/abs/1401.0304v1. 2014.

79. Mendelson S. Improving the Sample Complexity Using Global Data // IEEE Transactions on Information Theory. 2002. Vol. 48.

80. Mendelson S. On the performance of kernel classes // J. Mach. Learn. Res. 2003. December. Vol. 4. P. 759–771.

81. Pechyony D. Theory and Practice of Transductive Learning. Ph.D. thesis: Technion. 2008.

82. Recht B., Re C. Toward a Noncommutative Arithmetic-geometric Mean Inequality: Conjectures, Case-studies, and Consequences // COLT. 2012.

83. Sauer N. On the density of families of sets // Journal of Combinatorial Theory Series A. 1972. Vol. 13. P. 145–147.

84. Schapire R., Freund Y., Bartlett P., W. L. Boosting the Margin: a New Explanation for the Effectiveness of Voting Methods // Annals of Statistics. 1998. Vol. 26.

85. Seeger M. PAC-Bayesian Generalization Error Bounds for Gaussian Process Classification // Journal of Machine Learning Research. 2002.

86. Seldin Y., Tishby N. PAC-Bayesian Analysis of Co-clustering and Beyond // Journal of Machine Learning Research. 2010. Vol. 11.

87. Seldin Y., Auer P., Laviolette F., Shawe-Taylor J., Ortner R. PAC-Bayesian Analysis of Contextual Bandits // Advances in Neural Information Processing Systems (NIPS). 2011.

88. Seldin Y., Laviolette F., Cesa-Bianchi N., Shawe-Taylor J., Auer P. PAC-Bayesian Inequalities for Martingales // IEEE Transactions on Information Theory. 2012. Vol. 58.

89. Serfling R. J. Probability inequalities for the sum in sampling without replacement // The Annuals of Statistics. 1974. Vol. 2, no. 1. P. 39–48.

90. Shawe-Taylor J., Williamson R. C. A PAC analysis of a Bayesian estimator // Proceedings of the International Conference on Computational Learning Theory (COLT). 1997.

91. Shelah S. A combinatorial problem: Stability and order for models and theories in infinity languages // Pacific Journal of Mathematics. 1972. Vol. 41. P. 246–261.

92. Stone M. Cross-validatory choice and assessment of statistical predictors (with discussion) // Journal of the Royal Statistical Society. 1974. Vol. B36. P. 111–147.

93. Talagrand M. Concentration of measure and isoperimetric inequalities in product spaces // Publ. Math. I.H.E.S. 1995. Vol. 81. P. 73–203.

94. Talagrand M. New concentration inequalities in product spaces // Inventiones Mathematicae. 1996. Vol. 126.

95. Tolstikhin I., Seldin Y. PAC-Bayes-Empirical-Bernstein Inequality // Advances in Neural Information Processing Systems (NIPS). 2013.

96. Tolstikhin I., Blanchard G., Kloft M. Localized Complexities for Transductive Learning // Proceedings of the 27th Annual Conference on Learning Theory (COLT 2014), JMLR W&CP. 2014. P. 857–884.

97. Tropp J. A. User-Friendly Tail Bounds for Sums of Random Matrices // Foundations of Computational Mathematics. 2012. Vol. 12, no. 4. P. 389–434.

98. Tsybakov A. Optimal Aggregation of Classifiers in Statistical Learning // Annals of Statistics. 2004. Vol. 32. P. 135–166.

99. Valiant L. G. A theory of the learnable // Communications of the Association for Computing Machinery. 1984. Vol. 27, no. 11.

100. van de Geer S. Empirical Processes in M-Estimation. Cambridge University Press, 2000.

101. van der Vaart A. W., Wellner J. Weak Convergence and Empirical Processes: With Applications to Statistics. Springer, 2000.

102. van der Vaart A. Asymptotic statistics. Cambridge University Press, 1998.

103. Vapnik V. N. Estimation of Dependences Based on Empirical Data. Springer-Verlag New York, Inc., 1982.

104. Vapnik V. N. Statistical Learning Theory. John Wiley & Sons, 1998.

105. Vapnik V. N., Chervonenkis A. Y. On the uniform convergence of relative frequencies of events to their probabilities // Soviet Math. Dokl. 1968. Vol. 9.

106. Vapnik V. N., Chervonenkis A. Y. On the uniform convergence of relative frequencies of events to their probabilities // Theory of Probability and its Applications. 1971. Vol. 16, no. 2.

107. Vapnik V. N., Chervonenkis A. Y. Theory of Pattern Recognition // Nauka, Moscow (in Russian). 1974. German translation: W.N.Wapnik, A.Ya.Tschervonenkis (1979), Theorie der Zeichenerkennug, Akademia, Berlin.

108. Vapnik V. N., Chervonenkis A. Y. Necessary and Sufficient Conditions for the Uniform Convergence of Means to their Expectations // Theory of Probability and its Applications. 1981. Vol. 26, no. 3. P. 532–553.

109. Vorontsov K. V., Ivahnenko. A. A. Tight combinatorial generalization bounds for threshold conjunction rules // Lecture Notes in Computer Science, Springer-Verlag. 2011. P. 66–73.

110. Vorontsov K. V., Frey A. I., Sokolov E. A. Computable Combinatorial Overitting Bounds // Machine Learning and Data Analysis. 2013. Vol. 1, no. 6. P. 734–743.

111. Vorontsov K. V. Exact combinatorial bounds on the probability of overfitting for empirical risk minimization // Pattern Recognition and Image Analysis. 2010. Vol. 20, no. 3. P. 269–285.

112. Yu B. Rates of convergence for empirical processes of stationary mixing sequences // The Annals of Probability. 1994. Vol. 22, no. 1.

113. Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. Москва, Наука., 1974.

114. Воронцов К. В. Точные оценки вероятности переобучения // Доклады РАН. 2009. Т. 429, № 1. С. 15–18.

115. Воронцов К. В. Комбинаторная теория надежности обучения по прецедентам. Диссертация на соискание ученой степени д. ф.-м. н.: ВЦ РАН, 2010.

116. Воронцов К. В. Комбинаторная теория переобучения: результаты, приложения и открытые проблемы // Математические методы распознавания образов: 15-ая Всеросс. конф.: Докл. 2011.

117. Вьюгин В. В. Элементы Математической Теории Машинного Обучения. МФТИ — ИППИ РАН, 2012.

118. Пыткеев Е. Г., Хачай М. Ю. Сигма-компактность метрических булевых алгебр и равномерная сходимость частот к вероятностям // Тр. ИММ УрО РАН. 2010. Т. 16, № 1. С. 127–139.

119. Толстихин И. О. Вероятность переобучения плотных и разреженных семейств алгоритмов // Межд. конф. Интеллектуализация обработки информации ИОИ-8: Докл. 2010. С. 83–86.

120. Толстихин И. О. Точная оценка вероятности переобучения для одного специального семейства алгоритмов // Межд. науч. конф. студентов, аспирантов и молодых ученых «Ломоносов 2010»: Докл. 2010. С. 54–57.

121. Толстихин И. О. Локализация оценок избыточного риска в комбинаторной теории переобучения // Межд. конф. Интеллектуализация обработки информации ИОИ-9: Докл. 2012. С. 54–57.

122. Фрей А. И. Точные оценки вероятности переобучения для симметричных семейств алгоритмов // Математические методы распознавания образов: 14-ая Всеросс. конф.: Докл. 2009.

123. Фрей А. И. Точные оценки вероятности переобучения для симметричных семейств алгоритмов и рандомизированных методов обучения // Pattern Recognition and Image Analysis. 2010. Т. 20, № 3.

124. Фрей А. И., Толстихин И. О. Комбинаторные оценки вероятности переобучения на основе покрытий множества алгоритмов // Доклады РАН. 2014. Т. 455, № 3. С. 265–268.

Обозначения и символы

$\mathbb{E}[\cdot]$	Математическое ожидание
$\mathbb{D}[\cdot]$	Дисперсия
$\mathbb{P}\{\cdot\}$	Вероятность события
Q_n	Супремум эмпирического процесса (Стр. 33)
π, $\boldsymbol{\pi}$	Перестановки, векторы перестановок
$V^f(\boldsymbol{\pi})$	Дискретный градиент функции f в точке $\boldsymbol{\pi}$ (Стр. 43)
Q'_n	Супремум эмп. процесса для выборок без возвращений (Стр. 47)
Π_N	Симметрическая группа перестановок на множестве $\{1,\ldots,N\}$
$\mathbb{1}\{\cdot\}$ и $\mathbb{1}_{\{\cdot\}}$	Индикаторы событий
C_N^n	Число сочетаний $\frac{N!}{n!(N-n)!}$
$\mathcal{X}$ и $\mathcal{Y}$	Пространства объектов и ответов (Стр. 58)
P	Неизвестное вероятностное распределение на $\mathcal{X}\times\mathcal{Y}$ (Стр. 58)
S, S_n, S_m	Обучающие выборки $\{(X_i,Y_i)\}_{i=1}^n$ (Стр. 58)
S_u	Контрольные выборки
$\ell\colon \mathcal{Y}\times\mathcal{Y}\to[0,1]$	Функция потерь (Стр. 58)
$h\colon \mathcal{X}\to\mathcal{Y}$	Отображение, предиктор, классификатор (Стр. 58)
$L(h)$	Средний риск отображения h (Стр. 58)
g^*, L^*	Байесовское отображение и Байесовский риск $L^* = L(g^*)$ (Стр. 60)
$\mathcal{H}$	Класс отображений h (Стр. 61)
h^*	Отображение с минимальным в $\mathcal{H}$ средним риском $L(h)$ (Стр. 61)
$\ell_h(X,Y)$	Потери отображения h на паре (X,Y): $\ell_h(X,Y) = \ell\big(h(X),Y\big)$
$L_n(h)$	Эмпирический риск $L_n(h) = \frac{1}{n}\ell_h(X_i,Y_i)$ (Стр. 61)
$\hat{h}_n$	Минимизатор эмпирического риска (Стр. 61)
$\mathcal{E}(h), \mathcal{E}(h,\mathcal{H})$	Избыточный риск отображения h в классе $\mathcal{H}$ (Стр. 63)
$\mathcal{H}_S$	Потери отображений класса $\mathcal{H}$ на обучающей выборке S (Стр. 72)
$S_{\mathcal{H}}(n)$	Функция роста (Стр. 72)
$\mathrm{VC}(\mathcal{H})$	Размерность Вапника–Червоненкиса (Стр. 74)
$N(\varepsilon,\mathcal{F},\|\cdot\|)$	Мощность покрытия множества $\mathcal{F}$ (Стр. 76)
$\mathcal{F}, \mathcal{F}_{\mathcal{H}}$	Класс потерь $\mathcal{F} = \{\ell_h(X,Y)\colon h\in\mathcal{H}\}$ (Стр. 76)
$\mathcal{F}^*, \mathcal{F}^*_{\mathcal{H}}$	Класс избыточных потерь $\mathcal{F}^* = \{\ell_h(X,Y) - \ell_{h^*}(X,Y)\colon h\in\mathcal{H}\}$ (Стр. 86)
σ_i	Радемахеровские случайные величины (Стр. 78)
$R_n(\mathcal{G})$	Супремум Радемахеровского процесса (Стр. 78)
$\mathbb{E}[R_n(\mathcal{G})]$	Глобальная Радемахеровская сложность (Стр. 78)

$\mathbb{E}[R_n(\mathcal{G})\|S]$	Глобальная эмпирическая Рademахеровская сложность (Стр. 78)
$\hat{\mathbb{E}}[g]$	Среднее значение функции g на выборке
$k(X', X'')$	Функция спрямляющего ядра (Стр. 96)
$\boldsymbol{X}_N, \boldsymbol{X}_m, \boldsymbol{X}_u$	Объекты генеральной, обучающей и контрольных выборок (Стр. 103)
$\hat{L}_m(h), L_u(h), L_N(h)$	Риск на обуч., контрол. и генеральн. выборках (Стр. 103)
h^*_u, h^*_N	Отображения, оптимальные на контрол. и генеральн. выборках (Стр. 103)
$Ef, \hat{E}_m f$	Среднее значение f на генеральной и обучающей выборках (Стр. 108)
r^*_m	неподвижная точка субкоренного отображения ψ_m (Стр. 109)
$\mathbb{S}$	генеральная выборка $\mathbb{S} = \{(X_i, Y_i)\}_{i=1}^N$ (Стр. 122)
$\mathbb{A}$	Множество векторов ошибок (Стр. 122)
a	Вектор ошибок (Стр. 122)
$L(a, S)$	Частота ошибок вектора a на выборке S (Стр. 123)
$n(a, S)$	Число ошибок вектора a на выборке S (Стр. 123)
$A(S)$	Оптимальные на выборке S векторы (Стр. 123)
$[\mathbb{S}]^n$	Множество всех подвыборок $\mathbb{S}$ длины n (Стр. 123)
$\mu\colon [\mathbb{S}]^n \to \mathbb{A}$	Метод обучения (Стр. 123)
$Q_{\mu,\varepsilon}(\mathbb{A})$	Вероятность переобучения (Стр. 123)
$S_{\mathbb{A}}$	Группа симметрий множества $\mathbb{A}$ (Стр. 127)
$\Omega(\mathbb{A}), \Omega([\mathbb{S}]^n)$	Орбиты векторов и выборок (Стр. 127)
$\mathbb{B}^N$	Булев куб (Стр. 133)
$\mathbb{A}_m$	m-й слой множества векторов (Стр. 133)
$d(a', a'')$	Расстояние Хэминга (Стр. 133)
$B_r(a), B^c_r(a), B_r(a, d)$	Шар, слой шара и нижние слои шара (Стр. 133)
$H_N^{n,m}(z)$	Гипергеометрическая функция распределения (Стр. 138)
ρ, π	Апостериорное и априорное распределения на $\mathcal{H}$ (Стр. 153)
$\mathcal{P}(\mathcal{H})$	Множество вероятностных распределений на $\mathcal{H}$ (Стр. 153)
G_ρ	Рандомизированное (стохастическое) отображение (Стр. 154)
$L(G_\rho), L_n(G_\rho)$	Средний и эмпирический риск G_ρ (Стр. 154)
$\mathrm{KL}(\rho\|\pi)$	Относительная дивергенция между распределениями (Стр. 154)
μ^h	Математическое ожидание выборки с индексом h
V^h	Дисперсия выборки с индексом h
$\mathrm{kl}(q\|p)$	kl-функция (Стр. 161)
B_ρ	Взвешенный предиктор (Стр. 166)
$\mathbb{D}_n(h)$	Выборочная дисперсия потерь h (Стр. 169)