Федеральное государственное образовательное учреждение высшего образования
«Московский физико-технический институт (государственный университет)»

Федеральное государственное бюджетное учреждение науки
Институт проблем передачи информации им. А.А. Харкевича
Российской академии наук (ИППИ РАН)

На правах рукописи

Животовский Никита Кириллович

Минимаксные оценки риска в задачах статистического обучения

05.13.17 – Теоретические основы информатики

ДИССЕРТАЦИЯ
на соискание ученой степени
кандидата физико-математических наук

Научный руководитель
д. ф.-м. н., проф.
Воронцов Константин Вячеславович

Москва – 2017

Оглавление

Введение

Актуальность темы исследования. Вопросы точности восстановления закономерностей по эмпирическим данным являются ключевыми одновременно в теории машинного обучения и математической статистике. В последние годы интерес к этим вопросом мотивирован развитием методов машинного обучения и методов обнаружения новых знаний. Первые фундаментальные результаты в этой области были заложены в уже ставшей классической книге Вапника и Червоненкиса [1]. В серии своих работ они показали, что для некоторых семейств классификаторов малая ошибка на обучающей выборке влечет и малую ошибку на контрольной выборке. Таким образом, Вапник и Червоненкис первыми смогли обосновать применение популярного алгоритма обучения: алгоритма минимизации эмпирического риска. Так как их основные результаты представляются в виде статистических гарантий на предсказательную (обобщающую) способность алгоритмов обучения, то за теоретической частью машинного обучения закрепилось название *Теория статистического обучения*. В настоящее время результаты теории статистического обучения значительно расширены на практически произвольные классы обучаемых функций и различные функции потерь (см. монографии Энтони и Бартлетта [2] и Шалев-Шварца и Бен-Давида [3]).

Одной из проблем классических оценок предсказательного риска является тот факт, что скорости сходимости в них очень медленные, а именно обратно пропорциональны квадратному корню из длины выборки. Большое количество работ посвящено получению так называемых *быстрых порядков сходимости*, а именно, оценок на обобщающую способность, которые в некоторых случаях обратно пропорциональны размеру выборки. Данное направление развивается в работах Бартлетта и др. [4], Массара [5], Цыбакова [6], Колчинского [7], Жине и Колчинского [8]. Настоящая диссертационная работа продолжает исследование в этой области.

Для доказательства верхних оценок на предсказательный риск минимизато-

ра эмпирического риска используются результаты из теории эмпирических процессов. Стандартные версии равномерных законов больших чисел не позволяют получать порядки сходимости обратно пропорциональные длине выборке, поэтому для получения быстрых порядков сходимости вводятся так называемые *относительные равномерные уклонения частот от ожидаемых значений* (см. Главу 12 в книге Вапника и Червоненкиса). Подробные конструкции появляются при исследовании так называемых *локализованных* мер сложности и исследуются в работах Бартлетта и др. и Жине и Колчинского. В данной диссертационной работе доказываются новые оценки для равномерных уклонений, которые затем применяются для получения различных оценок обобщающей способности с быстрыми порядками сходимости. Другим схожим направлением является анализ трансдуктивного обучения. Данная постановка была введена Вапником [9]. В этой постановке предполагается, что обучающая и тестовая выборки реализуются с помощью равновероятных разбиений генеральной совокупности объектов на два множества. Существенным отличием от стандартных статистических постановок является тот факт, что элементы выборки более не являются независимыми. В настоящий момент основные оценки в этой области получены в работах Вапника, Дербеко и др. [10], Кортес и Мори [11] и других. Тем не менее вопросы точности имеющихся результатов также до конца не изучены. В настоящей работе улучшаются существующие верхние оценки, анализируются версии равномерных уклонений в трансдуктивной постановке и благодаря этому вводится новая меры сложности, так называемая перестановочную Радемахеровскую сложность.

Следующим естественным направлением, развиваемым в данной диссертации, является получение нижних оценок обобщающей способности. Эти вопросы актуальны в свете доказываемых верхних оценок и отвечают на вопросы точности последних. В книге Вапника и Червоненкиса даются базовые нижние оценки в некоторых специальных случаях. Также нижние оценки доказывается в работах Цыбакова [6], Рагинского и Рахлина [12], Массара [5]. Недостатком существующих результатов является тот факт, что они доказаны для некоторого фиксиро-

ванного специального семейства классификаторов, в то время как нижние оценки, совпадающие с верхними оценками и верные для произвольного семейства классификаторов, во многих случаях до сих пор не были доказаны. В данной работе для задачи классификации доказывается универсальная нижняя оценка, верная для практически произвольного семейства классификаторов. Стоит также отметить, что в задачах непараметрической регрессии подобные универсальные нижние оценки появляются в недавней работе Лекуэ и Мендельсона [13] и работе Мендельсона [14].

Последим направлением исследования диссертации является анализ задач, для которых минимизаторы эмпирического риска не являются оптимальной стратегий, в том смысле, что существуют другие алгоритмы обучения со строго лучшими теоретическими гарантиями. Это направление тесно связано с так называемой PAC-теорией, развиваемой в работах Хаусслера и др. [15], Флойд и Вармута [16], Зимона [17], Ауэра и Ортнера [18]. В данной диссертационной работе доказываются новые оптимальные оценки для схем сжатий выборок. Благодаря ним, в частности, впервые в PAC постановке удается построить алгоритм с полиномиальной сложностью и с практически оптимальным предсказательным риском.

Цели и задачи диссертационной работы:

1. Произвести локализованный анализ относительных равномерных уклонений частот от ожидаемых значений и проверить их применимость для получения верхних оценок на предсказательный риск.

2. Исследовать возможность доказательства нижних оценок, верных одновременно для произвольных семейств классификаторов.

3. Математически исследовать статистические свойства схем сжатия выборок и схем голосования классификаторов.

4. Проанализировать возможность улучшения существующих оценок риска в трансдуктивном обучении.

Для достижения поставленных целей ставятся следующие **задачи** исследования

1. Доказать экспоненциальные верхние оценки для вероятности относительных равномерных уклонений частот от ожидаемых значений, зависящих от локальных мер сложности классов. Применить полученные оценки для доказательства быстрых порядков сходимости предсказательного риска.

2. В задаче бинарной классификации в условиях шума Массара доказать минимаксные нижние оценок, верные для произвольного семейства классификаторов.

3. С помощью схем сжатия выборок и схем голосования классификаторов построить полиномиальный алгоритм в задаче линейной классификации с оптимальными гарантиями на предсказательный риск.

4. Обобщить понятие равномерных уклонений частот от математических ожиданий на трансдуктивный случай. С помощью этого обобщения доказать оценки предсказательного риска в трансдуктивном обучении.

Научная новизна. Все основные результаты диссертации являются новыми. В диссертации получены новые верхние оценки для относительных равномерных уклонений. Будучи примененными к задаче классификации в условиях малого шума, предложенные оценки позволяет получить оптимальные порядки сходимости.

Следующим направлением, развиваемым в данной диссертации, является получение оптимальных оценок в случаях, когда произвольный минимизатор эмпирического риска не является оптимальным решающим правилом. В частности, развивается подход, связанный со схемами сжатия выборок, который в предложенных случаях позволяет найти практически оптимальные порядки предсказательного риска. Благодаря понятиям устойчивости и анализу схем голосования впервые в PAC постановке удается построить алгоритм с полиномиальной сложностью с практически оптимальным предсказательным риском.

Последним направлением является анализ трансдуктивного обучения. В работе введена новая мера сложности, для которой показано качественное превосходство по сравнению с ранее вводимыми мерами сложности в трансдуктивной постановке. Предложены верхние оценки для вводимой меры сложности.

Теоретическая и практическая значимость. Полученные в диссертации результаты имеют широкий спектр применений. В работе показана связь полученных общих результатов с большим количеством конкретных задач теории статистического обучения и математической статистики. В частности, кроме нижеизложенных основных положений, выносимых на защиту, удается найти применение полученных оценок для получения улучшенных результатов в теории агрегации и неточных оракульных неравенств. Также удается применить оценки для получения точных теоретических гарантий для задачи превращения онлайнового алгоритма в алгоритм, работающий с независимой конечной выборкой.

Положения, выносимые на защиту:

1. Получены новые односторонние оценки для равномерных относительных законов больших чисел, выраженные в терминах локальной скобочной энтропии и локальной эмпирической энтропии.

2. Полученные общие оценки применены в задаче бинарной классификации при условии шума Массара. С помощью них впервые в данных условиях получены общие верхние оценки риска, для которых в тексте диссертации для произвольного класса с конечной размерностью Вапника-Червоненкиса также доказаны совпадающие с точностью до констант нижние оценки. Показаны примеры неоптимальности ранее получавшихся верхних и нижних оценок.

3. Для задачи линейной классификации с двумя классами впервые в PAC постановке получен практически минимаксно оптимальный результат для полиномиального алгоритма обучения для произвольного распределения данных. Оптимальные результаты для минимизатора эмпирического риска по-

лучены также в общих условиях на шум при условии лог-вогнутых распределений данных, а также при условии конечности локальных метрических энтропий в задачах непараметрической регрессии.

4. Для задач трансдуктивного обучения введена новая мера сложности, названная перестановочной Радемахеровской сложностью. Доказано превосходство получаемых статистических гарантий над ранними результатами в данной области. Получены верхние оценки для введенной меры сложности в случае конечных классов.

Степень достоверности и апробация результатов. Достоверность результатов обеспечивается математическими доказательствами теорем и утверждений. Результаты диссертационной работы докладывались и обсуждались на следующих конференциях и научных семинарах:

1. Доклад на семинаре группы Комбинаторной Геометрии, Ecole Polytechnique Federale de Lausanne, Лозанна, Швейцария, август 2017.

2. Международная конференция “Conference on Learning Theory (COLT)„ Амстердам, Нидерланды, июль 2017.

3. Доклад на Городском семинаре по теории вероятностей ПОМИ РАН, Санкт-Петербург, май 2017.

4. Доклад на семинаре исследовательской группы “Stochastic Algorithms and Nonparametric Statistics„ Weierstrass Institute for Applied Analysis and Stochastics, Берлин, Германия, апрель 2017.

5. Доклад на семинаре сектора Интеллектуальные системы ВЦ РАН, Москва, апрель 2017.

6. Доклад на семинаре Добрушинской математической лаборатории ИППИ РАН, Москва, март 2017.

7. Международная конференция “Algorithmic Learning Theory (ALT)„ Бари, Италия, октябрь 2016.

8. Выступление на конференции
   “Workshop on Modern Statistics and Optimization„ ИППИ РАН, Москва, февраль 2016.

9. Международная конференция “Algorithmic Learning Theory (ALT)„ Банф, Канада, октябрь 2015;

10. Выступление на научном семинаре Max Planck Institute for Intelligent Systems, Тюбинген, Германия, Май 2015.

11. Доклады на семинаре НМУ “Стохастический анализ в задачах„ Москва, октябрь 2014, март 2015.

Публикации из списка ВАК по теме диссертации. Основные результаты по теме диссертации изложены в 4 печатных работах, из которых 4 изданы в журналах, рекомендованных ВАК [19], [20], [21], [22].

Личный вклад автора. Содержание диссертации и основные положения, выносимые на защиту, отражают персональный вклад автора в опубликованные работы. В статьях с соавторами подготовка к публикации полученных результатов проводилась совместно с соавторами, причем, за исключением отдельно оговоренного в тексте диссертации результата в главе 5, вклад диссертанта был определяющим.

Структура и объем диссертации. Диссертация состоит из введения, обзора литературы, четырех глав, заключения и библиографии. Общий объем диссертации 110 страниц, из них 102 страница текста. Библиография включает 67 наименований на пяти страницах.

Глава 1

Обзор базовых результатов теории статистического обучения

1.1. Вероятностная постановка

Предположим, что существует множество объектов $\mathcal{X}$ (объекты принято отождествлять с их признаковыми описаниями) и множество ответов $\mathcal{Y}$. Последнее, например, в случае задачи классификации может состоять всего из двух элементов (классы 1 и -1) или в случае задачи регрессии совпадать со множеством действительных чисел. Далее предполагается, что нам дана *обучающая* выборка из n пар $(X_i, Y_i)_{i=1}^n$ из $(\mathcal{X} \times \mathcal{Y})^n$.

Говоря неформально, цель статистического обучения заключается в том чтобы на основании имеющейся обучающей выборки построить некоторое правило, которое бы смогло предсказать ответ Y на основании нового объекта X. О природе данных делаются следующие предположения:

- На $\mathcal{X} \times \mathcal{Y}$ задано вероятностное пространство с неизвестной нам вероятностной мерой P.
- Все пары (X_i, Y_i) из обучающей выборки получены независимо согласно этой мере (вероятностному распределению).
- Любая новая пара (X, Y) получается согласно тому же самому распределению и независимо от остальных.

Предположим, что на основании обучающей выборки нам удалось построить функцию $\hat{f} : \mathcal{X} \to \mathcal{Y}$. В этом случае, говорят, что был использован некоторый *алгоритм* или *метод* обучения, а процесс его применения мы в дальнейшем будем называть *обучением*. Заметим, что наличие взаимосвязи между X и Y как-то

характеризуется самой вероятностной мерой P. Для того чтобы делать какие-то предсказания логично предположить, что P не является произведением мер по X и Y, то есть объекты и, например, их классы вовсе не независимые случайные величины. Одновременно слишком сильное предположение заключается и в существовании строгой функциональной зависимости между X и Y. Поэтому P такова, что предполагается существование достаточно хорошей (в некотором смысле) связи между объектами и ответами. Для того чтобы формализовать эту идею нужно ввести *функцию ошибок*. Функция ошибок (функция потерь) — это некоторая неотрицательная функция $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, которая характеризует потери при отнесении объекта X к ответу $\hat{f}(X)$ в сравнении с его реальным ответом Y. Не вводя требований, которые обычно предъявляются к функциям потерь, перейдем сразу к типичным примерам:

- Бинарные потери $\ell(\hat{f}(X), Y) = \mathbb{1}\{\hat{f}(X) \neq Y\}$.
- В задачах регрессии $\ell(\hat{f}(X), Y) = (\hat{f}(X) - Y)^2$.
- Или $\ell(\hat{f}(X), Y) = |\hat{f}(X) - Y|$.
- Так называемый hinge loss $\ell(\hat{f}(X), Y) = \max\{0, 1 - Y\hat{f}(X)\}$.

Разумной характеристикой решающего правила была бы его ожидаемая ошибка по отношению к обучающей выборке, на основании которого оно построено:

$$R(\hat{f}) = \mathbb{E}\left[\ell(\hat{f}(X), Y) \mid (X_i, Y_i)_{i=1}^n\right]$$

Важно понимать, что математическое ожидание берется по новой паре (X, Y), в то время как решающее правило $\hat{f}$ само строится по случайной обучающей выборке $(X_i, Y_i)_{i=1}^n$. Если считать $\hat{f}$ не случайным объектом, то $R(\hat{f}) = \mathbb{E}\left[\ell(\hat{f}(X), Y)\right]$ называют *риском* правила $\hat{f}$. Для того чтобы избавиться от зависимости от случайной реализации определим уже неслучайную величину, называемую в дальнейшем *ожидаемым риском*:

$$\mathbb{E}R(\hat{f}) := \mathbb{E}\left[\mathbb{E}\left[\ell(\hat{f}(X), Y) \mid (X_i, Y_i)_{i=1}^n\right]\right]$$

Данная терминология несколько отличается от принятой в математической статистике, однако, для удобства мы будем использовать введенные определения. Средний риск зависит только от меры P и способа выбора $\hat{f}$ и дает разумный критерий выбора способа построения $\hat{f}$: выбирается та оценка, которая доставляет минимальный средний риск. Напомним, что в общем случае $\hat{f}$ – это случайное отображение, которое строится на основании обучающей выборки.

Если бы распределение P было известно, то задача поиска оптимального правила $\hat{f}$ была бы лишь задачей оптимизации.

Пример 1. *Пусть мы имеем дело с задачей классификации $\mathcal{Y} = \{1, -1\}$ с бинарной функцией потерь. В этом случае риск равен $P(\hat{f}(X) \neq Y)$. Среди всевозможных выборов $\hat{f}$ его минимизирует так называемое* байесовское решающее правило $g(x) = sign(\eta(x))$, *где $\eta(x) = \mathbb{E}(Y|X = x)$. Отметим, что байесовское решающее правило зависит не от обучающей выборки, а от неизвестной меры P, поэтому одним из способов приближенного построения байесовского решающего правила являются так называемые plug-in правила, основанные на построении по наблюдаемой выборке эмпирического аналога $g(x)$.*

В теории статистического обучения не принято задавать модель данных в явном виде или предполагать зависимость между X и Y. Наше априорное знание о задаче должно быть представлено в основном не ограничением на меру P, а априорно заданным семейством отображений $\mathcal{F}$, каждое из которых отображает X в Y. Алгоритмы в результате обучения выбирают решающее правило, принадлежащее $\mathcal{F}$. Такие алгоритмы обычно называют *proper learning* алгоритмами. В литературе, однако, часто и семейство решающих правил $\mathcal{F}$ называется моделью, а выбор оптимального для задачи класса $\mathcal{F}$ — называется задачей выбора модели. В качестве семейства решающих правил могут выступать, например, семейство полупространств (афинные подпространства) в случае линейных клас-

сификаторов, семейства функций с определенными свойствами гладкости и так далее. Рассмотрим некоторые примеры:

Пример 2. *1.* Бесшумный случай *(в литературе часто называется* function learning*): существует некоторая функция $T : \mathcal{X} \to \mathcal{Y}$, такая, что $Y = T(X)$. При этом не делается предположений о том, что $T \in \mathcal{F}$. Очевидно, что, например, для квадратичной функции ошибки в этом случае функция T является еще и байесовским решающим правилом.*

2. Непараметрическая регрессия*: Зависимость $Y = f^*(X) + \varepsilon$, где ε — центрированная случайная величина с конечной дисперсией, не зависящая от X. Обозначение f^* выбрано не случайно. Действительно, данная функция является байесовским решающим правилом относительно квадратичной функции потерь. Данный тип зависимостей обычно изучается в математической статистике, где часто предполагается, что $f^* \in \mathcal{F}$.*

*3. В более общей задаче статистического обучения никаких функциональных связей между X и Y не предполагается, а рассматривается отдельно лишь хорошо специфицированный (well-specified) случай, когда байесовское решающее правило $f^* \in \mathcal{F}$, и агностический случай (*agnostic case*), когда не делается предположений о принадлежности байесовского решающего правила семейству $\mathcal{F}$. В обоих случаях анализ проводится на основании свойств семейства $\mathcal{F}$.*

Аппроксимация и оценивание

Ясно, что байесовское решающее правило, то есть, правило, минимизирующее риск, является в некотором смысле эталоном для любого метода обучения. Таким образом, возникает разумный вопрос о построении таких алгоритмов обучения, используя которые, мы получили бы риск сколь угодно близкий к байесовскому. Обозначая отображения из $\mathcal{X}$ в $\mathcal{Y}$ как $\mathcal{Y}^{\mathcal{X}}$, рассмотрим разницу риска

оценки $\hat{f}$ и байесовского решающего правила, то есть:

$$\mathbb{E}R(\hat{f}) - \inf_{f \in \mathcal{Y}^{\mathcal{X}}} R(f),$$

где $\inf\limits_{f \in \mathcal{Y}^{\mathcal{X}}} R(f) = R(f^*)$. Однако, в общем случае в задачах статистического обучения $f^* \notin \mathcal{F}$, поэтому перепишем предыдущую разность в следующем виде:

$$\left(\mathbb{E}R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f)\right) + \left(\inf_{f \in \mathcal{F}} R(f) - \inf_{f \in \mathcal{Y}^{\mathcal{X}}} R(f)\right).$$

Левое слагаемое называют *ошибкой оценивания* (*estimation error*), а правое называется *ошибкой аппроксимации* (*approximation error*). Очевидно, что чем больше $\mathcal{F}$, тем меньше ошибка аппроксимации, но одновременно больше ошибка оценивания. Действительно, так как алгоритм выбирает правило из $\mathcal{F}$, то в лучшем случае его приближает именно $\inf\limits_{f \in \mathcal{F}} R(f)$. Действительно, если $\mathcal{F}$ состоит всего из одной функции, то ошибка оценивания равна нулю. Заметим также, что в бесшумном случае ошибка аппроксимации равна нулю. Компонента

$$R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f)$$

называется *избыточным риском*. Таким образом, ошибка оценивания в наших обозначениях является математическим ожиданием избыточного риска.

Обучаемость

Введем классическое понятие *Probably Approximately Correct* – обучаемости. Ограничимся задачей классификации $\mathcal{Y} = \{1, -1\}$, бинарной функцией потерь и примем бесшумную модель, то есть для некоторой функции $T : \mathcal{X} \to \{1, -1\}$ имеет место $Y = T(X)$. Также предполагается, что $T \in \mathcal{F}$.

Определение 1 (PAC-обучаемость [3]). *Семейство $\mathcal{F}$ называется PAC–обучаемым, если существует функция $n_{\mathcal{F}} : (0,1)^2 \to \mathbb{N}$ и некоторый обучающий алгоритм, такие что для всех $\varepsilon, \delta \in (0,1)$ при любом вероятностном распределении на $\mathcal{X}$ и любой целевой функции $T \in \mathcal{F}$, если алгоритм на простой выборке из*

хотя бы $n \geq n_{\mathcal{F}}(\varepsilon, \delta)$ *объектов выдает классификатор* $\hat{f}$*, то с вероятностью не меньшей чем* $1 - \delta$ *(по отношению к обучающей выборке) имеет место неравенство*

$$R(\hat{f}) \leq \varepsilon.$$

Обучаемость означает, что для достаточно большой выборки алгоритм выдает с большой вероятностью решение, обладающее маленькой вероятностью ошибки. Среди функций $n_{\mathcal{F}}$ ту, что принимает наименьшие значения, принято называть *выборочной сложностью* (sample complexity). Она показывает сколько нужно объектов выборки, чтобы обучиться с заданной точностью. Обобщим введенное понятие на агностический случай:

Определение 2 (Агностическая PAC-обучаемость [3])**.** *Семейство* $\mathcal{F}$ *называется агностически PAC–обучаемым, если существует функция* $n_{\mathcal{F}} : (0,1)^2 \to \mathbb{N}$ *и некоторый обучающий алгоритм, такие что для всех* $\varepsilon, \delta \in (0,1)$ *при любом вероятностном распределении на* $\mathcal{X} \times \mathcal{Y}$*, если алгоритм на простой выборке из хотя бы* $n \geq n_{\mathcal{F}}(\varepsilon, \delta)$ *объектов выдает классификатор* $\hat{f}$*, то с вероятностью не меньшей чем* $1 - \delta$ *(по отношению к обучающей выборке) имеет место неравенство*

$$R(\hat{f}) \leq \inf_{f \in \mathcal{F}} R(f) + \varepsilon.$$

1.2. Принцип равномерной сходимости

Введем понятие *класса потерь*:

$$\ell \circ \mathcal{F} = \{(x, y) \to \ell(f(x), y) : f \in \mathcal{F}\}.$$

Рассмотрим некоторую функцию $g : X \to \mathbb{R}_+$. Введем два стандартных обозначения: $Pg = \mathbb{E}g$ и $P_n g = \frac{1}{n} \sum\limits_{i=1}^{n} g(x_i)$, где математическое ожидание берется по распределению на X, которое также обозначается P, а суммирование ведется по реализации независимых X_i, распределенных согласно P.

Говорят, что для класса функций $\mathcal{G}$ выполняется *не зависящий от распределения усиленный равномерный закон больших чисел* (*distribution free uniform strong law of large numbers*), если для любого $\varepsilon > 0$

$$\lim_{n\to\infty} \sup_P \mathbb{P}\left\{\sup_{m\geq n} \sup_{g\in\mathcal{G}} |P_m g - Pg| > \varepsilon\right\} = 0;$$

Классы, для которых выполнено данное свойство называются *равномерными классами Гливенко–Кантелли.*

Утверждение 1 (Принцип равномерной сходимости [3])**.** *Для того чтобы класс $\mathcal{F}$ был агностически PAC-обучаемым достаточно, чтобы соответсвующий класс потерь $\ell \circ \mathcal{F}$ был равномерным классом Гливенко–Кантелли.*

Доказательство. Докажем, что обучаемость достигается с помощью метода минимизации эмпирического риска. Обозначим $f^*_{\mathcal{F}} = \arg\inf_{f\in\mathcal{F}} R(f)$. Для минимизатора эмпирического риска $\hat{f}$:

$$\begin{aligned} R(\hat{f}) - R(f^*_{\mathcal{F}}) &= R(\hat{f}) - R_n(\hat{f}) + R_n(f^*_{\mathcal{F}}) - R(f^*_{\mathcal{F}}) + R_n(\hat{f}) - R_n(f^*_{\mathcal{F}}) \\ &\leq R(\hat{f}) - R_n(\hat{f}) + R_n(f^*_{\mathcal{F}}) - L(f^*_{\mathcal{F}}) \\ &\leq 2\sup_{f\in\mathcal{F}} |R(f) - R_n(f)| \\ &= 2\sup_{g\in\ell\circ\mathcal{F}} |Pg - P_n g|. \end{aligned}$$

□

Условие теоремы вовсе не является необходимым для обучаемости. Фактически оно является лишь достаточным для того, чтобы можно было агностически обучиться с помощью метода минимизации эмпирического риска. Оказывается, что для задач классификации с бинарной функцией потерь варианты равномерных законов больших чисел являются также и необходимыми для обучаемости с помощью метода минимизации эмпирического риска. Тем не менее в общем случае нужно согласовывать это со следующим элементарным результатом [1].

Пример 3 (обучаемость без равномерной сходимости)**.** *Возьмем произвольный класс $\mathcal{F}$ и предположим, что к нему можно добавить такую $f^{(1)}$, что*

$$\ell(f^{(1)}(X), Y) < \inf_{f \in \mathcal{F}} \ell(f(X), Y)$$

с вероятностью единица. Тогда очевидно, что минимизатор эмпирического риска с вероятностью единица будет выбирать $f^{(1)}$, который доставляет и минимальный риск в новом классе. Заметим, что для класса $\mathcal{F}$ в этом случае может не выполняться никаких версий равномерного закона больших чисел.

1.3. Радемахеровский процесс

Ранее мы показали, что

$$\mathbb{E}(R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f)) \leq 2\mathbb{E} \sup_{f \in \mathcal{F}} |R(f) - R_n(f)|.$$

Обозначим $R'_n(f)$ – эмпирическое среднее по независимой копии обучающей выборки. Соответсвующее ей математическое ожидание будем обозначать $\mathbb{E}'$. С помощью неравенства Йенсена имеем

$$\begin{aligned}
\mathbb{E} \sup_{f \in \mathcal{F}} |R(f) - R_n(f)| &= \mathbb{E} \sup_{f \in \mathcal{F}} |\mathbb{E}' R'_n(f) - R_n(f)| \\
&\leq \mathbb{E}\mathbb{E}' \sup_{f \in \mathcal{F}} |R'_n(f) - R_n(f)| \\
&= \frac{1}{n} \, \mathbb{E}\mathbb{E}' \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} (\ell(f(X'_i), Y'_i) - \ell(f(X_i), Y_i)) \right|.
\end{aligned}$$

Введем *Радемахеровские случайные величины*, то есть независимые в совокупности (и от X_i, Y_i) случайные величины ε_i, принимающие равновероятно значения 1 и -1. Легко видеть, что для всех i распределения случайных величин $(\ell(f(X'_i), Y'_i) - \ell(f(X_i), Y_i))$ и $\varepsilon_i(\ell(f(X'_i), Y'_i) - \ell(f(X_i), Y_i))$ одинаковы. Данный прием принято называть *симметризацией* [7]. Обозначая математическое ожида-

ние по всем ε_i как $\mathbb{E}_\varepsilon$, получаем:

$$\begin{aligned}
&\frac{1}{n}\,\mathbb{E}\mathbb{E}'\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}(\ell(f(X_i'),Y_i')-\ell(f(X_i),Y_i))\right| \\
&=\frac{1}{n}\,\mathbb{E}\mathbb{E}'\mathbb{E}_\varepsilon\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\varepsilon_i(\ell(f(X_i'),Y_i')-\ell(f(X_i),Y_i))\right| \\
&\leq\frac{2}{n}\,\mathbb{E}\mathbb{E}_\varepsilon\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\varepsilon_i\ell(f(X_i),Y_i)\right|.
\end{aligned}$$

Введем для фиксированной выборки $(X_i,Y_i)_{i=1}^n$ *условную Радемахеровскую сложность*:

$$\mathcal{R}_n(\ell\circ\mathcal{F})=\frac{1}{n}\,\mathbb{E}_\varepsilon\sup_{g\in\ell\circ\mathcal{F}}\left|\sum_{i=1}^{n}\varepsilon_i g(X_i,Y_i)\right|=\frac{1}{n}\,\mathbb{E}_\varepsilon\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\varepsilon_i\ell(f(X_i),Y_i)\right|$$

и просто *Радемахеровскую сложность*

$$\mathcal{R}(\ell\circ\mathcal{F})=\mathbb{E}\mathcal{R}_n(\ell\circ\mathcal{F})=\frac{1}{n}\,\mathbb{E}\sup_{g\in\ell\circ\mathcal{F}}\left|\sum_{i=1}^{n}\varepsilon_i g(X_i,Y_i)\right|=\frac{1}{n}\,\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\varepsilon_i\ell(f(X_i),Y_i)\right|.$$

Таким образом, мы получили, что

$$\mathbb{E}\sup_{f\in\mathcal{F}}|R(f)-R_n(f)|\leq 2\mathcal{R}(\ell\circ\mathcal{F}).$$

Радемахеровскую сложность можно рассматривать как величину, описывающую сложность класса решающих правил. Чем больше Радемахеровская сложность, тем лучше ошибки $\mathcal{F}$ могут кореллировать со случайным шумом ε_i. Как только мы зафиксировали выборку $(X_i,Y_i)_{i=1}^n$ условную Радемахеровскую сложность можно рассматривать как *Радемахеровское среднее*, связанное со множеством $A\subset\mathbb{R}^n$:

$$\mathcal{R}_n(A)=\frac{1}{n}\,\mathbb{E}_\varepsilon\sup_{a\in A}\left|\sum_{i=1}^{n}\varepsilon_i a_i\right|,$$

где множество A является множеством векторов ошибок $\mathcal{F}$ на $(X_i,Y_i)_{i=1}^n$.

Рассмотрим простые свойства Радемахеровских средних [23]. Если A, B – ограниченные множества в $\mathbb{R}^n$, $c\in\mathbb{R}$, то

1. $\mathcal{R}_n(A\cup B)\leq\mathcal{R}_n(A)+\mathcal{R}_n(B)$.

2. $\mathcal{R}_n(cA) = |c|\mathcal{R}_n(A)$.

3. $\mathcal{R}_n(A \oplus B) \leq \mathcal{R}_n(A) + \mathcal{R}_n(B)$.

4. Если $A = \{a^{(1)}, \ldots, a^{(N)}\}$, то $\mathcal{R}_n(A) \leq \max\limits_j ||a^{(j)}||_2 \ \frac{\sqrt{2\log(2N)}}{n}$.

5. (Contraction inequality [24, 25]) Если $\phi : \mathbb{R} \to \mathbb{R}$ Липшицева с константой L, причем $\phi(0) = 0$, то $\mathcal{R}_n(\phi(A)) \leq L\mathcal{R}_n(A)$, где ϕ действует на векторы A покомпонентно.

6. $\mathcal{R}_n(A) = \mathcal{R}_n(conv(A))$.

Имеет место следующая общая теорема.

Теорема 1 (см. обзорную статью [23])**.** *С вероятностью не меньшей* $1 - \delta$ *для функций потерь, принимающих значения в отрезке* $[0, 1]$ *выполнено:*

$$\sup_{f \in \mathcal{F}} |R(f) - R_n(f)| \leq 2\mathcal{R}(\ell \circ \mathcal{F}) + \sqrt{\frac{2\log(\frac{1}{\delta})}{n}}.$$

Также

$$\sup_{f \in \mathcal{F}} |R(f) - R_n(f)| \leq 2\mathcal{R}_n(\ell \circ \mathcal{F}) + 3\sqrt{\frac{2\log(\frac{2}{\delta})}{n}}$$

1.4. Верхние оценки на Радемахеровский процесс

Рассмотрим задачу классификации с бинарной функцией потерь. Будем говорить, что множество $\{x_1, \ldots, x_k\} \in \mathcal{X}^k$ разбивается $\mathcal{F}$, если существуют 2^k различных классификаций $\{x_1, \ldots, x_k\}$ с помощью классификаторов из $\mathcal{F}$. Размерность Вапника–Червоненкиса класса $\mathcal{F}$ — это наибольшее натуральное число d такое, что существует подмножество $\{x_1, \ldots, x_d\}$, разбиваемое $\mathcal{F}$ [26]. Мы определим функцию роста $\mathcal{S}_{\mathcal{F}}(n)$ как наибольшее возможное число различных классификаций множества из n точек, реализуемое с помощью классификатора $\mathcal{F}$ (максимизированное по выбору n точек).

Пример 4. *Одномерное семейство пороговых решающих правил*

$$\mathcal{F} = \{f_\theta(x) = \mathbb{1}\{x \leq \theta\} : \theta \in [0, 1]\}$$

имеет размерность Вапника-Червоненкиса, равную единице.

Пример 5. *Семейство классификаторов, представляющее собой семейство разделюящих k-мерных гиперплоскостей имеет размерность Вапника–Червоненкиса, равную $d = k + 1$. Данное утверждение связано с теоремой Радона.*

Утверждение 2 (теорема Радона [27])**.** *Произвольное подмножество из $k + 2$ или более точек k-мерного евклидова пространства может быть разделено на два непересекающихся подмножества, чьи выпуклые оболочки имеют непустое пересечение.*

Пример 6. *Семейство классификаторов*

$$\mathcal{F} = \{x \to sgn(\sin(tx)) : t \in \mathbb{R}\}$$

имеет размерность равную ∞, даже несмотря на то, что параметризуется лишь одним параметром.

Семейство классификаторов, обладающее конечной размерностью обладает замечательным свойством:

Лемма 1 (Вапник-Червоненкис [1])**.** *Для любого семейства классификаторов с размерностью Вапника–Червоненкиса d для $n \geq d$:*

$$S_{\mathcal{F}}(n) \leq \sum_{i=0}^{d} \binom{n}{i}$$

Особенность Радемахеровского процесса заключается в том, что его можно анализировать с помощью мощных средств теории эмпирических процессов. Действительно, можно рассматривать процесс $\sup\limits_{a \in A} \left| \sum\limits_{i=1}^{n} \varepsilon_i a_i \right|$ как верхнюю оценку эмпирического процесса $\sup\limits_{f \in \mathcal{F}} |R_n(f) - R(f)|$ со множеством состояний A, где A

— проекция класса потерь на конечную выборку. В этом случае Радемахеровское среднее есть не что иное, как ожидаемый супремум этого процесса. Теория эмпирических процессов показывает, что во многих случаях поведение процесса зависит от метрических свойств пространства состояний. В нашем случае – это метрические свойства множества A. Условно по обучающей выборке множество $A = A\left((X_i, Y_i)_{i=1}^n\right)$ можно представить себе как набор из не более чем $\sum_{i=0}^{d} \binom{n}{i}$ различных булевых векторов. Введем на паре векторов метрику ρ:

$$\rho(a, b) = \sqrt{\frac{1}{n} d_H(a, b)},$$

где d_H – метрика Хэмминга. Будем говорить, что множества $B \subset \{1, -1\}^n$ является ε-покрытием множества A, если объединение замкнутых ε-шаров (по введенной метрике) с центрами в точках B содержат A. Обозначим $\mathcal{N}(\varepsilon, A)$ – *число покрытия*, равное мощности минимального ε-покрытия множества A.

Теорема 2 ([28]). *Для задачи классификации с бинарной функцией потерь*

$$\mathcal{R}_n(\ell \circ \mathcal{F}) \leq \frac{12}{\sqrt{n}} \sup_{(X_i, Y_i)_{i=1}^n} \int_0^1 \sqrt{\log(2\mathcal{N}(\varepsilon, A))} d\varepsilon,$$

где $A = \ell \circ \mathcal{F}_{(X_i, Y_i)_{i=1}^n}$.

Важность данного результата связна с использованием следующей Леммы.

Лемма 2 (Хаусслер [29]). *Если множество булевых векторов A состоит из различных векторов ошибок семейства классификаторов с размерностью Вапника-Червоненкиса равной d, то для $0 \leq \varepsilon \leq 1$:*

$$\mathcal{N}(\varepsilon, A) \leq e(d + 1) \left(\frac{2e}{\varepsilon^2}\right)^d.$$

Применяя данную Лемму можно получить, что для некоторой абсолютной константы C для задачи классификации с бинарной функцией потерь

$$\mathcal{R}_n(\ell \circ \mathcal{F}) \leq C\sqrt{\frac{d}{n}}, \tag{1.1}$$

что вместе с Теоремой 1 дает классическую оценку для минимизаторов эмпирического риска в классах с конечной размерностью Вапника-Червоненкиса. С вероятностью не меньшей чем $1 - \delta$ выполнено

$$R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f) \leq C \left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(\frac{1}{\delta})}{n}} \right). \tag{1.2}$$

Глава 2

Оценки, не зависящие от распределения P_X

2.1. Быстрые порядки сходимости

В 1968 году Вапник и Червоненкис [26] ввели комбинаторное свойство классификаторов, так называемую VC размерность (размерность Вапника-Червоненкиса), играющую ключевую роль не только в математической статистике, но и в других разделах математики. В настоящий момент считается, что VC размерность характеризует свойства минимизаторов эмпирического риска. Например, если никакие ограничения не делаются на распределения данных, можно показать, что для любого фиксированного класса классификаторов предсказательный риск минимизатора эмпирического риска близок к риску лучшего классификатора в данном классе с точностью до члена порядка $\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(\frac{1}{\delta})}{n}}$ с вероятностью $1-\delta$, где d — это VC размерность класса, а n — размер выборки (см. неравенство (1.2)). Можно также показать минимаксную нижнюю оценку (верную для любой обучающей процедуры), которая с точностью до констант совпадает с последней оценкой. Но тот факт, что одна только VC размерность действительно описывает сложность класса оказывается верным только в самом худшем случае, когда не делаются никакие предположения о зависимости между объектами и ответами. В литературе замечено, что если рассматривать условия малого шума, то одна размерность Вапника–Червоненкиса не является правильной характеристикой сложности для метода минимизации эмпирического риска [5, 12, 30]. До сих пор правильная мера сложности была найдена только для некоторых специальных классов. В данной диссертационной работе, в частности, мы предлагаем универсальные меры сложности, дающие практически (а иногда и точно) совпадающие оценки для риска минимизатора эмпирического риска.

В последние 20 лет много усилий было предпринято для того, чтобы понять

условия, которые влекут так называемую быструю сходимость риска к нулю (порядка $\frac{1}{n}$), вместо $\frac{1}{\sqrt{n}}$. В настоящее время эти условия хорошо поняты [31]. В начале 2000-ых годов в теорию статистического обучения были введены так называемые локализованные сложности ([4], [25]). Эти меры сложности активно используются для доказательства порядков сходимости $\frac{1}{n}$. Но в дополнение к быстрым порядкам сходимости локализация означает также и то, что только небольшая окрестность лучшего классификатора влияет на порядок сходимости риска классифкаторов к нулю. Существующие подходы, основанные на локализации (в основном на локальной Радемахеровской сложности) обычно сложны для оценивания, а более простые следствия из этих оценок используют локализацию для получения порядков $\frac{1}{n}$, а не для локализации класса. Более того в литературе до сих пор нет общих минимаксных нижних оценок, основанных на локализованных процессах.

Тем не менее существует серия результатов, дающая одновременно быструю скорость сходимости и связанная с локализацией класса классификаторов. В работе Массара и Неделек [5] доказана оценка порядка $\frac{d}{nh}\log(\frac{nh^2}{d})+\frac{\log(\frac{1}{\delta})}{nh}$ при условиях малого шума Массара, где h — параметр отвечающий за степень шума. Для получения этой оценки использовали локальный анализ для получения улучшенных порядков. Однако эта их оценка не учитывает локализацию в терминах сложности самого класса: в данном случае члена $d\log(\frac{nh^2}{d})$. Жине и Колчинский [8] улучшили эту оценку, и получили оценку порядка $\frac{d}{nh}\log(\tau(\frac{d}{nh^2}))+\frac{\log(\frac{1}{\delta})}{nh}$ для минимизатора эмпирического риска, где τ – зависящая от распределения мера сложности, называемая емкостью Александера (из работы Александера [32]). Совсем недавно Ханнеке и Янг [33] ввели новый комбинаторный параметр $\mathbf{s}$, называемый числом связности, который дает точные и не зависящие от распределения оценки $\tau(\frac{d}{nh^2})$ и который в общем случае не контролируется VC размерностью. Таким образом, как следствие результата Жине и Колчинского получается оценка $\frac{d}{nh}\log(\mathbf{s}\wedge\frac{nh^2}{d})+\frac{\log(\frac{1}{\delta})}{nh}$. Однако, в некоторых случаях эта оценка неоптимальна. В этой диссертационной работе мы строго ее улучшим.

Одной из целей данной диссертационной работы является получение точных

и не зависящих от распределения локализованных оценок для VC классов с помощью введения правильной меры сложности, не зависящей от распределения. Таким образом, будет устранен существующий зазор между верхними и нижними оценками. Мерой сложности будет локализованная локальная метрическая энтропия или ее неподвижная точка. Большинство результатов будут получены по вероятности и по математическому ожиданию.

2.2. Обозначения и некоторые результаты

Мы вводим пространство объектов $\mathcal{X}$ и пространство меток $\mathcal{Y} = \{1, -1\}$. Мы считаем, что множество $\mathcal{X} \times \mathcal{Y}$ задано вместе с σ-алгеброй подмножеств и вероятностной мерой P на измеримых подмножествах. Также считаем, что нам дано множество измеримых функций, называемых в дальнейшем классификаторами, $\mathcal{F}$, отображающих $\mathcal{X}$ в $\mathcal{Y}$. Таким образом, P задает совместное распределение (X, Y). Риск классификатора f определяется его вероятностью ошибки $R(f) = P(f(X) \neq Y)$. Известно, что среди всех классификаторов минимум риска достигается так называемым байесовским классификатором $f^*(x) = \text{sign}(\eta(x))$, где $\eta(x) = \mathbb{E}[Y|X = x]$ [34]. Символ $\wedge$ будет обозначать минимум из двух действительных чисел, $\vee$ — максимум из двух действительных чисел и $\mathbb{1}[A]$ — индикатор события A. Для любого подмножества $B \subseteq \mathcal{F}$ определим область рассогласования $\text{DIS}(B) = \{x \in \mathcal{X} |\ \exists f, g \in B$ такие что $f(x) \neq g(x)\}$. Также мы будем рассматривать абстрактные действительнозначные функциональные классы, которые мы обычно будем обозначать символом $\mathcal{G}$. Мы немного перегрузим обозначения и будем считать, что $\log(x)$ обозначает $\ln(\max(x, e))$. Обозначение $f(n) \lesssim g(n)$ или $g(n) \gtrsim f(n)$ означает, что существует универсальная константа $c > 0$ такая что $f(n) \leq cg(n)$ для всех $n \in \mathbb{N}$. Аналогично мы считаем, что $f(n) \simeq g(n)$ эквивалентно $g(n) \lesssim f(n) \lesssim g(n)$.

Мы наблюдаем $(X_1, Y_1), \ldots, (X_n, Y_n)$ независимую, распределенную согласно P^n выборку. Обозначим $Z_i = (X_i, Y_i)$ и $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. С помощью P_n мы обозна-

чим математическое ожидание по отношению к эмпирической мере (эмпирическое среднее), индуцированное данной выборкой. Минимизатор эмпирического риска — любой алгоритм определяемый следующим образом: для любой выборки он выдает классификатор $\hat{f}$ который минимизирует $R_n(f) = P_n \mathbb{1}[f(X) \neq Y]$ среди всех $f \in \mathcal{F}$. Вопросы существования минимизаторов эмпирического риска подробно освещены в работе в данной постановке [5] и в дальнейшем мы будем предполагать их существование. Также мы будем использовать *независимые выборки* (ghost samples), которые также составлены из независимых элементов, распределенных согласно P. Эмпирическое среднее по отношению к этим выборкам мы будем обозначать P'_n.

Определение 3 (Массар и Неделек [5]). *Пара* $(P, \mathcal{F})$ *удовлетворяет условиям малого шума Массара, если* $f^* \in \mathcal{F}$ *и для некоторого* $h \in [0, 1]$ *выполнено* $|\eta(X)| \geq h$ *с вероятностью единица* 1.

Для всякого класса $\mathcal{F}$ множество всех таких распределений будет обозначаться $\mathcal{P}(h, \mathcal{F})$. Случай $h = 1$ соответствует классификации без шума, $Y = f^*(X)$ с вероятностью единица, а случай $h = 0$ относится к случаю $f^* \in \mathcal{F}$ без ограничений на распределения. Промежуточные значения h могут соответствовать задаче, в которой случайный шум изменяет значение фиксированной целевой функции на противоположное с некоторой вероятностью. Пусть $\mathcal{F}$ — это класс с VC размерностью d. Как мы знаем из неравенства (1.2) для любого минимизатора эмпирического риска $\hat{f}$ и для любого распределения $P \in \mathcal{P}(0, \mathcal{F})$, с вероятностью не менее $1 - \delta$ выполнено

$$R(\hat{f}) - R(f^*) \lesssim \sqrt{\frac{d}{n}} + \sqrt{\frac{\log(\frac{1}{\delta})}{n}}.$$

Более того, имеет место нижняя оценка, выполненная для классификатора $\tilde{f}$, выдаваемого любым алгоритмом, основанным на наблюдении n точек. Для него

существует $P \in \mathcal{P}(0, \mathcal{F})$ такое что, с вероятностью не меньше чем $1 - \delta$,

$$R(\tilde{f}) - R(f^*) \gtrsim \left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(\frac{1}{\delta})}{n}} \right) \wedge 1.$$

Таким образом, мы знаем, что VC размерность характеризует риск минимизатора эмпирического риска, когда никаких ограничений не делается на распределение данных. Интересно, что это не верно в случае, когда $h > 0$. В данной диссертационной работы мы описываем эту неизвестную меру сложности, в случая, когда h отделен от 0 и 1. Сначала мы представим улучшение ранее описанной оценки Жине и Колчинского в случае, когда $h > 0$ [8]. Рассмотрим следующее определение.

Определение 4. *Для $\varepsilon_0 > 0$ рассмотрим множество $\mathcal{F}_{\varepsilon_0} = \big\{ f \in \mathcal{F} : P_X\big(f(X) \neq f^*(X)\big) \leq \varepsilon_0 \big\}$. Для $\varepsilon \in (0, 1]$ определим*

$$\tau(\varepsilon) = \sup_{\varepsilon_0 \geq \varepsilon} \frac{P_X(\{x \in \mathcal{X} : \exists f \in \mathcal{F}_{\varepsilon_0} \text{ такая что } f(x) \neq f^*(x)\})}{\varepsilon_0} \vee 1.$$

Эта величина была введена (фактически[1]) в теорию эмпирических процессов Александером [32], и называется *емкостью Александера* [8]. Аналогичная величина была введена независимо в литературе по активному обучению, в котором она называется коэффициент рассогласования [35, 36]. $\tau(\varepsilon)$ является зависящей от распределения локальной мерой сложности. Жине и Колчинский [8] дают следующую верхнюю оценку. Пусть $\mathcal{F}$ — класс VC размерности d, и $\hat{f}$ является минимизатором эмпирического риска по n обучающим объектам. Для любого распределения $P \in \mathcal{P}(h, \mathcal{F})$, с вероятностью не менее $1 - \delta$,

$$R(\hat{f}) - R(f^*) \lesssim \frac{d}{nh} \log\left(\tau\left(\frac{d}{nh^2} \right) \right) + \frac{\log(\frac{1}{\delta})}{nh}. \tag{2.1}$$

Доказательство этой оценки основано на анализе локальной Радемахеровской сложности. В недавней работе [33] была введена мера сложности класса, назы-

[1] Оригинальное определение не включало в себя супремум по ε_0, вместо этого выбиралось значение $\varepsilon_0 = \varepsilon$. Однако результаты были доказаны при ограничительных предположениях о монотонности.

ваемая числом связности (*star number*), которое в точности равно максимально допустимому значению емкости Александера.

Определение 5. *Число связности* $\mathbf{s}$ *это наибольшее натуральное число такое, что существуют* $x_1, \ldots, x_{\mathbf{s}} \in \mathcal{X}$ *и* $f_0, f_1, \ldots, f_{\mathbf{s}} \in \mathcal{F}$ *такие что для всех* $i \in \{1, \ldots, \mathbf{s}\}$, $DIS(\{f_0, f_i\}) \cap \{x_1, \ldots, x_{\mathbf{s}}\} = \{x_i\}$.

По аналогии с емкостью Александера число связности есть некоторая мера локального разнообразия класса $\mathcal{F}$. В терминах так называемого графа один-включений , изученного Хаусслером, Литтлстоуном и Вармутом [15], число связности может быть определено как наибольшая возможная степень вершины в этом графе. Легко показать, что для VC размерности d имеет место неравенство $d \leq \mathbf{s}$. Однако зазор между двумя величинами может быть бесконечным. Мы отсылаем к работе [33] для дополнительных примеров и более детального обсуждения связности. Работа Ханнеке и Янг содержит результат

$$\sup_{f^* \in \mathcal{F}} \sup_{P_X} \tau(\varepsilon) = \mathbf{s} \wedge \frac{1}{\varepsilon}. \tag{2.2}$$

Прямым следствием является то, что для $P \in \mathcal{P}(h, \mathcal{F})$, с вероятностью не меньшей $1 - \delta$,

$$R(\hat{f}) - R(f^*) \lesssim \frac{d}{nh} \log\left(\frac{nh^2}{d} \wedge \mathbf{s}\right) + \frac{\log(\frac{1}{\delta})}{nh}. \tag{2.3}$$

В частности, в бесшумном случае, когда $h = 1$, с вероятностью не менее $1 - \delta$,

$$R(\hat{f}) \lesssim \frac{d}{n} \log\left(\frac{n}{d} \wedge \mathbf{s}\right) + \frac{\log(\frac{1}{\delta})}{n}.$$

Так как $\mathbf{s}$ контролирует емкость Александера точным равенством, при рассмотрении случая без ограничений на распределение P_X у нас нет возможности напрямую улучшить оценку Жине и Колчинского. Следующая оценка в бесшумном случае была доказана в работе [30]. Для любого распределения $P \in \mathcal{P}(1, \mathcal{F})$, с вероятностью не меньше $1 - \delta$,

$$R(\hat{f}) \lesssim \frac{d}{n} \log\left(\frac{n}{d} \wedge \frac{\mathbf{s}}{d}\right) + \frac{\log(\frac{1}{\delta})}{n}. \tag{2.4}$$

Даже это небольшое улучшение показывает неоптимальность оценки (2.3). В этой диссертационной работе мы улучшим эту оценку и покажем, что пара $d, \mathbf{s}$ не является исчерпывающей характеристикой поведения минимизатора эмпирического риска в случае, когда h отделено от нуля.

2.3. Некоторые факты из теории эмпирических процессов

Пусть нам дан функциональный класс $\mathcal{G}$, отображающий $\mathcal{Z}$ в $\mathbb{R}$, можно рассмотреть супремум эмпирического процесса:

$$\sup_{g \in \mathcal{G}} (P - P_n) g.$$

Начиная с работы Вапника и Червоненкиса [26], анализ алгоритмов обучения обычно осуществляется с помощью равномерного контроля за процессом $(P - P_n) g$ для специальных классов функций. Поведение супремума эмпирического процесса тесно связано с супремумом Радемахеровского процесса (введенного нами ранее):

$$\frac{1}{n} \mathbb{E}_\varepsilon \sup_{g \in \mathcal{G}} \left(\sum_{i=1}^{n} \varepsilon_i g_i \right),$$

где g_i обозначает $g(Z_i)$, а ε_i являются независимыми Радемахеровскими случайными величинами, принимающими равновероятно значения ± 1. Такой подход, однако, часто ведет к неоптимальным верхним и нижним оценкам, которые не позволяют получить как быстрые порядки сходимости, так и улучшения за счет локализации классов. Вместо этого мы рассмотрим другие величины, а именно супремум *смещенного процесса*. Для $c > 0$ мы рассмотрим

$$\sup_{g \in \mathcal{G}} (P - (1 + c) P_n) g.$$

Другой важной величиной является ожидаемый супремум так называемого офф-сет–процесса, введенный в работе Рахлина, Лианга, Сридхарана [37]:

$$\frac{1}{n} \mathbb{E}_\varepsilon \sup_{g \in \mathcal{G}} \left(\sum_{i=1}^{n} \varepsilon_i g_i - c' g_i^2 \right).$$

Последняя величина была введена для анализа специальной процедуры агрегации при квадратичных потерях и до сих пор не была отнесена к смещенным процессам. Следующая лемма, доказанная в более общем виде [37], будет нам полезна.

Лемма 3 (Рахлин, Лианг, Сридхаран [37]). *Пусть $V \subset \{-1, 0, 1\}^n$ — конечное множество из N векторов. Тогда для любого $c > 0$,*

$$\frac{1}{n}\mathbb{E}_\varepsilon \max_{v \in V} \left(\sum_{i=1}^n \varepsilon_i v_i - c|v_i| \right) \leq \frac{1}{2c} \frac{\log(N)}{n}.$$

Следующая простая Лемма дает *симметризацию* для смещенного процесса

Лемма 4 (Симметризация для смещенного процесса в среднем). *Пусть $\mathcal{G}$ — функциональный класс и $c \geq 0$ — некоторая абсолютная константа. Тогда*

$$\mathbb{E} \sup_{g \in \mathcal{G}} ((P - (1+c)P_n)g) \leq \frac{c+2}{n} \mathbb{E}\mathbb{E}_\varepsilon \sup_{g \in \mathcal{G}} \left(\sum_{i=1}^n \varepsilon_i g(Z_i) - \frac{c}{c+2} g(Z_i) \right).$$

Доказательство. Используя стандартный прием симметризации и неравенство Йенсена, мы получаем

$$\begin{aligned}
&\mathbb{E} \sup_{g \in \mathcal{G}} ((P - (1+c)P_n)g) \\
&\leq \mathbb{E} \sup_{g \in \mathcal{G}} (P_n' g - (1+c)P_n g) \\
&= \mathbb{E} \sup_{g \in \mathcal{G}} ((1+c/2)(P_n' g - P_n g) - cP_n' g/2 - cP_n g/2) \\
&\leq 2\mathbb{E}\mathbb{E}_\varepsilon \sup_{g \in \mathcal{G}} \left(\frac{1+c/2}{n} \sum_{i=1}^n \varepsilon_i g(Z_i) - cP_n g/2 \right) \\
&= 2(1+c/2)\mathbb{E}\mathbb{E}_\varepsilon \sup_{g \in \mathcal{G}} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i g(Z_i) - \frac{c/2}{1+c/2} P_n g \right).
\end{aligned}$$

□

Интересно, что подставляя $c = 0$, мы мгновенно получаем стандартное симметризационное неравенство. Следующая лемма, которая дает новый симметризационный результат для смещенного процесса по вероятности, требует следующего определения. Говорят, что функциональный класс $\mathcal{G}$ является (B, β)-*классом Бернштейна*, если для всех $g \in \mathcal{G}$ мы имеем $Pg^2 \leq B\,(Pg)^\beta$.

Лемма 5 (Симметризация для смещенного процесса по вероятности)**.** *Пусть* $\mathcal{G}$ — $(B,1)$*-класс Бернштейна, такой что для всех* $g \in \mathcal{G}$ *выполнено* $Pg \geq 0$*. Пусть* $c_1 > c_2 > 0$ — *некоторые произвольные константы. Если* $nt \geq \frac{B(1+c_2)^2}{c_2}$*, то*

$$P\left(\sup_{g\in\mathcal{G}}(P-(1+c_1)P_n)g \geq t\right) \leq 2P\left(\sup_{g\in\mathcal{G}}((1+c_2)P'_n-(1+c_1)P_n)g \geq t/2\right).$$

Доказательство. Обозначим с помощью $\tilde{g}$ — случайную функцию, доставляющую супремум.

$$\begin{aligned}&\mathbb{1}[(P-(1+c_1)P_n)\tilde{g} > t]\mathbb{1}[(P-(1+c_2)P'_n)\tilde{g} < t/2]\\ &\leq \mathbb{1}[((1+c_2)P'_n-(1+c_1)P_n)\tilde{g} > t/2].\end{aligned}$$

Беря математическое ожидание по отношению ко второй выборке, получаем

$$\begin{aligned}&\mathbb{1}[(P-(1+c_1)P_n)\tilde{g} > t]P'[(P-(1+c_2)P'_n)\tilde{g} < t/2] \leq\\ &P'[((1+c_2)P'_n-(1+c_1)P_n)\tilde{g} > t/2].\end{aligned}$$

Далее мы имеем

$$P'\Big[(P-(1+c_2)P'_n)\tilde{g} \geq t/2\Big] = P'\Big[(P-P'_n)\tilde{g} \geq \frac{t/2+c_2P\tilde{g}}{1+c_2}\Big].$$

Используя неравенство Чебышева вместе с $4ab \leq (a+b)^2$, получаем

$$P'\Big[(P-P'_n)\tilde{g} \geq \frac{t/2+c_2P\tilde{g}}{1+c_2}\Big] \leq \frac{P\tilde{g}^2(1+c_2)^2}{n(t/2+c_2P\tilde{g})^2} \leq \frac{BP\tilde{g}(1+c_2)^2}{2ntc_2P\tilde{g}} = \frac{B(1+c_2)^2}{2ntc_2}.$$

В итоге, получаем, что если $\frac{ntc_2}{B(1+c_2)^2} \geq 1$, то $P'[(P-(1+c_2)P'_n)\tilde{g} < t/2] \geq \frac{1}{2}$. Доказательство завершается с помощью взятия математического отношения по отношению к первой выборке. □

Следствие 1. *В условиях предыдущей леммы*

$$\begin{aligned}&P\left(\sup_{g\in\mathcal{G}}(P-(1+c_1)P_n)g \geq t\right)\\ &\leq 4P\left(\sup_{g\in\mathcal{G}}\left(\frac{1+c'/2}{n}\sum_{i=1}^{n}\varepsilon_i g(Z_i) - c'P_ng/2\right) \geq t'/2\right),\end{aligned}$$

где $c' = \frac{c_1-c_2}{1+c_2}$ *и* $t' = \frac{t}{2(1+c_2)}$.

Доказательство. Используя те же обозначения, мы переписываем результат Леммы 5 в следующей форме

$$P\left(\sup_{g\in\mathcal{G}}(P-(1+c_1)P_n)g\geq t\right)\leq 2P\left(\sup_{g\in\mathcal{G}}(P_n'-(1+c')P_n)g\geq t'\right).$$

Вводим Радемахеровские случайные величины и замечаем, что $\sup_{g\in\mathcal{G}}(P_n'-(1+c')P_n)g$ имеет такое же распределение как и

$$\sup_{g\in\mathcal{G}}\left(\frac{1+c'/2}{n}\sum_{i=1}^{n}\varepsilon_i(g_i-g_i')-c'P_ng/2-c'P_n'g/2\right).$$

В итоге,

$$\begin{aligned}&\sup_{g\in\mathcal{G}}\left(\frac{1+c'/2}{n}\sum_{i=1}^{n}\varepsilon_i(g_i-g_i')-c'P_ng/2-c'P_n'g/2\right)\\&\leq\sup_{g\in\mathcal{G}}\left(\frac{1+c'/2}{n}\sum_{i=1}^{n}\varepsilon_ig_i-c'P_ng/2\right)+\sup_{g\in\mathcal{G}}\left(-\frac{1+c'/2}{n}\sum_{i=1}^{n}\varepsilon_ig_i'-c'P_n'g/2\right).\end{aligned}$$

Оба слагаемых опять же имеют одно и то же распределение. □

Пусть $\mathbf{s}$ обозначает число связности класса бинарных классификаторов $\mathcal{F}$. В работе Ханнеке [30] доказывается, что

$$\mathbb{E}P_X(\mathrm{DIS}(\mathcal{V}_n))\leq\frac{\mathbf{s}}{n+1},\tag{2.5}$$

где $\mathcal{V}_n=\{f\in\mathcal{F}|P_n\mathbb{1}[f(X)\neq f^*(X)]=0\}$ — множество минимизаторов эмпирического риска. Эта же работа устанавливает, что с вероятностью не меньшей $1-\delta$ выполнено

$$P_X(\mathrm{DIS}(\mathcal{V}_n))\leq\frac{21\mathbf{s}}{n}+\frac{16\log(\frac{3}{\delta})}{n}.\tag{2.6}$$

Этот результат в случае ограниченной связности означает, что в бесшумном случае ожидаемая мера рассогласования множества минимизаторов эмпирического риска имеет порядок $\frac{\mathbf{s}}{n}$. Читатель, знакомый с работой Хаусслера, Литтлстоуна, Вармута [15], может вспомнить, что для некоторых алгоритмов обучения качество классификации может контролироваться максимальной исходящей степенью вершины в графе один-исключений. Более того всегда существует ориентация такого

графа, при которой максимальная исходящая степень не превосходит величину VC размерности. Этот результат в некотором смысле похож на результаты [15], однако вместо степени ориентированного графа один исключений, мера рассогласования множества минимизаторов эмпирического риска контролируется максимальной степенью неориентированного графа. Так как в бесшумном случае любой минимизатор эмпирического риска, как и истинный классификатор находятся во множестве $\mathcal{V}_n$, простым следствием этого результата в случае $\mathbf{s} \approx d$, будет то, что любой минимизатор эмпирического риска будет достигать оптимальный порядка риска d/n. Более того в работе [30] оценка (2.6) используется для получения верхней оценки для минимизатора эмпирического риска $\frac{d}{n}\log\frac{n\wedge\mathbf{s}}{d}$ по математическому ожиданию, и оценки $\frac{d}{n}\log\frac{n\wedge\mathbf{s}}{d}+\frac{1}{n}\log\frac{1}{\delta}$ с вероятностью $1-\delta$. Мы докажем новый вариант подобной оценки

Теорема 3. *Пусть* $\mathbf{s}$ *— число связности семейства бинарных классификаторов* $\mathcal{F}$*. В бесшумном случае для любого минимизатора эмпирического риска* $\hat{f}$ *выполнено*

$$\mathbb{E}R(\hat{f}) \lesssim \frac{\log\left(\mathcal{S}_{\mathcal{F}}\left(\mathbf{s}\wedge n\right)\right)}{n}.$$

Более того, с вероятностью не меньшей $1-\delta$*,*

$$R(\hat{f}) \lesssim \frac{\log\left(\mathcal{S}_{\mathcal{F}}\left(\mathbf{s}\wedge n\right)\right)}{n} + \frac{\log(\frac{1}{\delta})}{n}.$$

Можно легко показать (используя оценку Вапника и Червоненкиса на функцию роста [26]), что это неравенство является альтернативным способом доказательства верхней оценки $\frac{d\log\left(\frac{\mathbf{s}\wedge n}{d}\right)}{n}$, которая в свою очередь является улучшением оценки Жине и Колчинского для бесшумного случая.

Пример 7. *Теорема 3 позволяет получать простые примеры неоптимальности оценки* (2.3) *в бесшумном случае. Например, пусть* $\mathcal{X} = \{x_1, \ldots, x_{\mathbf{s}}\}$*, определим класс* $\mathcal{F}_1$*, как класс состоящий из всех классификаторов, определенных на* $\mathcal{X}$ *классифицирующих не более* d *точек как* $+1$*, и класс* $\mathcal{F}_2$ *всех классификаторов, имеющих не более* $d-1$ *точек относимых к* $+1$ *среди* $\{x_1, \ldots, x_{d-1}\}$ *и не более*

одной точки, классифицируемой $+1$ *среди* $\{x_d, \ldots, x_{\mathbf{s}}\}$. *Для классов* $\mathcal{F}_1$ *и* $\mathcal{F}_2$, *размерность Вапника Червоненкиса равна* d, *а число связности равна* $\mathbf{s}$. *Однако, для* $\mathcal{F}_1$ *Теорема 3 дает оценку порядка* $\frac{d\log\left(\frac{\mathbf{s}\wedge n}{d}\right)}{n}$, *а для* $\mathcal{F}_2$ *дает оценку порядка* $\frac{d+\log(\mathbf{s}\wedge n)}{n}$. *В обоих случаях можно показать, что правильный порядок для минимизатора эмпирического риска совпадает с полученными оценками [15, 30]. Тем не менее, можно привести примеры, когда предложенная оценка не оптимальна.*

2.4. Оценки в терминах глобальных упаковок

Главная цель этого раздела доказать простую оценку в терминах неподвижных точек глобальных чисел покрытия. В следующих главах мы существенно улучшим этот результат, поэтому для простоты будем рассматривать только бесшумный случай. Отметим, что подобные результаты можно получить с помощью равномерных относительных уклонений (см. раздел 19.6 в [2]). Зафиксируем n-элементную выборку и для любых двух классификаторов $f, g \in \mathcal{F}$ определим $\rho_H(f,g) = |\{i \in \{1,\ldots,n\} : f(x_i) \neq g(x_i)\}|$. Обозначим

$$\mathcal{M}_1^*(\mathcal{F},\gamma,n) = \max_{x_1,\ldots,x_n\in\mathcal{X}} \mathcal{M}_1(\mathcal{F}(\{x_1,\ldots,x_n\}),\gamma),$$

где $\mathcal{M}_1(\mathcal{H},\varepsilon)$ обозначает размер максимальной ε-упаковки $\mathcal{H}$ по расстоянию ρ_H (для данных точек $x_1,\ldots,x_n$) и где $\mathcal{F}(\{x_1,\ldots,x_n\})$ — множество проекций $\mathcal{F}$ на $\{x_1,\ldots,x_n\}$.

Во многих статистических постановках оптимальные порядки риска получаются с помощью правильной балансировки радиуса и логарифма числа покрытия, соответствующего данному радиусу (см., например работу Янга и Баррона [38]). Мы покажем, что в наших оценках можно выбирать γ такое, что $c\gamma \approx \log(\mathcal{M}_1^*(\mathcal{F},\gamma,n))$ для некоторого $c \in [0,1]$. Определим

$$\gamma_c^*(n,\mathcal{F}) = \max\{\gamma \in \mathbb{N} : c\gamma \leq \log\left(\mathcal{M}_1^*(\mathcal{F},\gamma,n)\right)\}.$$

Значение $\gamma_c^*(n, \mathcal{F})$ будет называться *стационарной (неподвижной) точкой глобальной энтропии*. Когда класс $\mathcal{F}$ понятен из контекста, мы будем писать $\gamma_c^*(n)$ вместо $\gamma_c^*(n, \mathcal{F})$. Заметим, что $\gamma_c^*(n, \mathcal{F})$ является корректно определенной неотрицательной величиной.

Утверждение 3. *Пусть некоторый класс $\mathcal{F}$ имеет VC размерность d. Если $P \in \mathcal{P}(1, \mathcal{F})$ (бесшумный случай), то для любого минимизатора эмпирического риска $\hat{f}$,*

$$\mathbb{E}R(\hat{f}) \lesssim \frac{\gamma_{\frac{1}{2}}^*(n)}{n}.$$

Более того, с вероятностью не менее $1 - \delta$,

$$R(\hat{f}) \lesssim \frac{\gamma_{\frac{1}{2}}^*(n)}{n} + \frac{\log \frac{1}{\delta}}{n},$$

и

$$\gamma_{\frac{1}{2}}^*(n) \lesssim d \log(n/d). \tag{2.7}$$

Для доказательства этого утверждения нам понадобится следующая техническая лемма, которую можно рассматривать как модификацию Леммы 6 из [37]

Лемма 6. *Пусть $\mathcal{G}$ — множество функций принимающих бинарные значения и пусть $c \in [0, 1]$ некоторая константа. Пусть также $\varepsilon_1, \ldots, \varepsilon_n$ независимые Радемахеровские случайные величины. Тогда*

$$\frac{1}{n}\mathbb{E}_\varepsilon \max_{g \in \mathcal{G}} \left(\sum_{i=1}^n \varepsilon_i g(X_i) - c g(X_i) \right) \leq \frac{7\gamma_c^*(n)}{n}.$$

Доказательство. Зафиксируем точки $X_1, \ldots, X_n$, положим $V = \{(g(X_1), \ldots, g(X_n)) : g \in \mathcal{G}\}$. Как и ранее для γ фиксируем γ-покрытие $\mathcal{N}_\gamma \subseteq V$, для $v \in V$, $p(v)$ будет

обозначать ближайший вектор к v в $\mathcal{N}_\gamma$. Рассмотрим следующую декомпозицию

$$\begin{aligned}
&\mathbb{E}_\varepsilon \max_{v\in V}\left(\sum_{i=1}^n \varepsilon_i v_i - cv_i\right)\\
&\le \mathbb{E}_\varepsilon \max_{v\in V}\left(\sum_{i=1}^n \varepsilon_i \left(v_i - p(v)_i\right)\right) + \max_{v\in V}\left(\sum_{i=1}^n \frac{c}{4}p(v)_i - cv_i\right)\\
&+ \mathbb{E}_\varepsilon \max_{v\in V}\left(\sum_{i=1}^n \varepsilon_i p(v)_i - \frac{c}{4}p(v)_i\right).
\end{aligned}$$

Так как $p(v)$ находится на расстоянии не более γ от v, то мы знаем, что $\sum_{i=1}^n p(v)_i \le \gamma + \sum_{i=1}^n v_i$. Таким образом, второе слагаемое в последнем выражении не превосходит

$$\max_{v\in V}\left(\frac{c}{4}\gamma - \frac{3c}{4}\sum_{i=1}^n v_i\right) \le \frac{c}{4}\gamma.$$

Третий член ограничен $\frac{2}{c}\log(\mathcal{N}_1(\mathcal{F},\gamma))$ с помощью Леммы 3, а первый член ограничен γ. Используя стандартное отношение между максимальными упаковками и минимальными покрытиями [39], мы получаем

$$\frac{1}{n}\mathbb{E}_\varepsilon \max_{v\in V}\left(\sum_{i=1}^n \varepsilon_i v_i - cv_i\right) \le \frac{(1+c/4)\gamma}{n} + \frac{2}{c}\frac{\log(\mathcal{M}_1(V,\gamma))}{n}.$$

Выбирая $\gamma = \gamma_c^*(n) + 1$, мы получаем

$$\frac{(1+c/4)\gamma}{n} + \frac{2}{c}\frac{\log(\mathcal{M}_1(V,\gamma))}{n} \le \frac{(1+c/4)(\gamma_c^*(n)+1)}{n} + \frac{2(\gamma_c^*(n)+1)}{n} \le \frac{7\gamma_c^*(n)}{n}.$$

□

Утверждение 3. Введем *класс потерь* $\mathcal{G}_{f^*} = \{x \to \mathbb{1}[f(x) \ne f^*(x)]$ для $f \in \mathcal{F}\}$. Пусть $\hat{f}$ — любой минимизатор эмпирического риска и $\hat{g}$ — соответствующая функция в классе потерь $\mathcal{G}_{f^*}$. Очевидно, что мы имеем $\mathbb{E}R(\hat{f}) = P\hat{g}$ и $P_n\hat{g} = 0$. Тогда для любого $c > 0$

$$\mathbb{E}R(\hat{f}) = \mathbb{E}(R(\hat{f}) - (1+c)R_n(\hat{f})) \le \mathbb{E}\sup_{g\in\mathcal{G}_{f^*}}(Pg - (1+c)P_n g).$$

С помощью Леммы 4 мы получаем

$$\mathbb{E}\sup_{g\in\mathcal{G}_{f^*}}(Pg - (1+c)P_n g) \le \frac{c+2}{n}\mathbb{E}\mathbb{E}_\varepsilon \sup_{g\in\mathcal{G}_y}\left(\sum_{i=1}^n \varepsilon_i g(X_i) - \frac{c}{c+2}g(X_i)\right)$$

Применяя Лемму 6 и фиксируя $c = 2$, мы заканчиваем доказательство оценки по математическому ожиданию. □

Пример 8. *Рассмотрим класс пороговых классификаторов — это класс* $\mathcal{F} = \{x \to 2\mathbb{1}[x \leq t] - 1 : t \in \mathbb{R}\}$. *Используя определение числа связности легко показать, что оно равно* 2. *В этом случае Теорема 3 дает оптимальный порядок* $\frac{1}{n}$ *для минмимизатора эмпирического риска. Одновременно, в худшем случае числа упаковки* $\mathcal{M}_1^*(\mathcal{F}, \gamma, n)$ *имеют порядок* $\frac{n}{\gamma}$. *Простой анализ дает нам неподвижную точку порядка* $\gamma^*_{\frac{1}{2}}(n) \simeq \log(n)$, *а значит утверждение 3 дает неоптимальную* $\frac{\log n}{n}$ *оценку. Хотя мы правильно уловили, что порядок сходимости быстрее* $\frac{1}{\sqrt{n}}$, *наш анализ неоптимален. В следующем разделе будет предложен общий подход, позволяющий получить правильную скорость сходимости.*

2.5. Локальная эмпирическая метрическая энтропия

Введем *локальную метрическую энтропию* следующим образом. Возьмем n точек, зафиксируем некоторый классификатор $f \in \mathcal{F}$ и строим Хэммингов шар радиуса γ. Определим $\mathcal{B}_H(f, \gamma, \{x_1, \ldots, x_n\}) = \{g \in \mathcal{F} | \rho_H(f, g) \leq \gamma\}$ и введем

$$\mathcal{M}_1^{\text{loc}}(\mathcal{F}, \gamma, n, h) = \max_{x_1,\ldots,x_n} \max_{f \in \mathcal{F}} \max_{\varepsilon \geq \gamma} \mathcal{M}_1(\mathcal{B}_H(f, \varepsilon/h, \{x_1, \ldots, x_n\}), \varepsilon/2), \tag{2.8}$$

где как и ранее $\mathcal{M}_1(\mathcal{H}, \varepsilon)$ обозначает размер наибольшей ε-упаковки $\mathcal{H}$ с расстоянием ρ_H (для данных точек $x_1, \ldots, x_n$). Фиксируем $h, h' \in (0, 1]$ и определяем

$$\gamma^{\text{loc}}_{h,h'}(n, \mathcal{F}) = \max\{\gamma \in \mathbb{N} : h\gamma \leq \log(\mathcal{M}_1^{\text{loc}}(\mathcal{F}, \gamma, n, h'))\}.$$

Когда ясно, о каком классе $\mathcal{F}$ идет речь, мы будем писать $\gamma^{\text{loc}}_{h,h'}(n)$ вместо $\gamma^{\text{loc}}_{h,h'}(n, \mathcal{F})$. Величина $\gamma^{\text{loc}}_{h,h'}(n)$ определяет стационарную точку *локальной эмпирической энтропии*. Заметим, что так как $1 \leq d < \infty$ при условии $h, h' > 0$ множество в правой части определения конечно и не пусто.Таким образом, $\gamma^{\text{loc}}_{h,h'}(n)$ корректно определенная строго неотрицательная величина. Действительно, для любых

$h, h' \in (0,1]$, величина $\gamma = \lfloor \frac{1}{h} \rfloor$ удовлетворяет $h\gamma \leq 1$ (из нашего переопределения логарифма $\log(\cdot)$). Это влечет $h\gamma^{\mathrm{loc}}_{h,h'}(n, \mathcal{F}) \geq h\lfloor \frac{1}{h} \rfloor \geq \frac{1}{2}$.

Следующая теорема — наша главная верхняя оценка этого раздела.

Теорема 4. *Для любого класса $\mathcal{F}$ с VC размерностью d и числом связности $\mathbf{s}$ зафиксируем $h \in \left(\sqrt{\frac{d}{n}}, 1\right]$. Если $P \in \mathcal{P}(h, \mathcal{F})$, тогда для любого минимизатора эмпирического риска $\hat{f}$,*

$$\mathbb{E}(R(\hat{f}) - R(f^*)) \lesssim \frac{\gamma^{\mathrm{loc}}_{h,h}(n)}{n}. \tag{2.9}$$

Также с вероятностью не менее $1 - \delta$,

$$R(\hat{f}) - R(f^*) \lesssim \frac{\gamma^{\mathrm{loc}}_{h,h}(n)}{n} + \frac{\log(\frac{1}{\delta})}{nh}. \tag{2.10}$$

Также выполнено

$$\frac{d + \log\left(nh^2 \wedge \mathbf{s}\right)}{h} \lesssim \gamma^{\mathrm{loc}}_{h,h}(n) \lesssim \frac{d \log\left(\frac{nh^2}{d} \wedge \mathbf{s}\right)}{h} + \frac{d \log\left(\frac{1}{h}\right)}{h}. \tag{2.11}$$

Наша мера сложности (2.11) не хуже, чем независящая от распределения оценка (2.3) в случае, когда h отделен от 0 константой. В дальнейшем мы обсудим возможную неоптимальность оценки в случае, когда h мало, вызванную членом $\frac{d \log\left(\frac{1}{h}\right)}{h}$ в (2.11). Другим интересным свойством является тот факт, что оценки (2.9) и (2.10) не включают ни VC размерность, ни число связности явным образом. С другой стороны оба параметра позволяют контролировать нашу меру сложности сверху и снизу. Для любого $f \in \mathcal{F}$, обозначим $g_f(x, y) = \mathbb{1}[f(x) \neq y] - \mathbb{1}[f^*(x) \neq y]$. Рассмотрим *класс избыточных потерь* $\mathcal{G}_{\mathcal{Y}} = \{g_f | f \in \mathcal{F}\}$, класс $\mathcal{G}_{f^*} = \{x \to \mathbb{1}[f(x) \neq f^*(x)] \mid f \in \mathcal{F}\}$ и класс $\mathcal{F}^* = \frac{1}{2}(\mathcal{F} - f^*)$. Последний класс состоит из функций вида $\frac{1}{2}(f - f^*)$ for $f \in \mathcal{F}$. Следующие свойства хорошо известны

1. Для любой функции $g_f \in \mathcal{G}_{\mathcal{Y}}$ выполнено $g_f^2(x, y) = \mathbb{1}[f(x) \neq f^*(x)] = \frac{1}{2}|f(x) - f^*(x)| = \frac{1}{4}(f(x) - f^*(x))^2$.

2. Для любой функции $g_f \in \mathcal{G}_{\mathcal{Y}}$ выполнено $g_f(x, y) = \frac{y(f^*(x) - f(x))}{2}$.

3. Для любого распределения $P \in \mathcal{P}(h, \mathcal{F})$ класс $\mathcal{G}_\mathcal{Y}$ является $(\frac{1}{h}, 1)$-классом Бернштейна [23] и $R(f^*) \le \frac{1}{2}(1-h)$ [34].

Лемма 7 (Сжатие)**.** *Пусть $\mathcal{G}_\mathcal{Y}$ — класс избыточных потерь, ассоциированный с $\mathcal{F}$. Зафиксируем $h \in [0,1]$. Для любого фиксированного $c \in [0,1]$ и любого распределения $P \in \mathcal{P}(h, \mathcal{F})$ условно по $X_1, \dots, X_n$*

$$\begin{aligned}
&\mathbb{E}_{Y|X}\mathbb{E}_\varepsilon \sup_{g\in\mathcal{G}_\mathcal{Y}} \left(\sum_{i=1}^n \varepsilon_i g(X_i, Y_i) - c g(X_i, Y_i) \right) \\
&\le \mathbb{E}_\varepsilon \sup_{f'\in\mathcal{F}^*} \left(\sum_{i=1}^n \varepsilon_i f'(X_i) - \frac{1}{2} hc |f'(X_i)| \right) \\
&\quad + \frac{3c}{2} \mathbb{E}_\xi \sup_{g'\in\mathcal{G}_{f^*}} \left(\sum_{i=1}^n \xi_i g'(X_i) - \frac{1}{3} h g'(X_i) \right)
\end{aligned}$$

где $\xi_1, \dots, \xi_n$ — случайные величины условно независимые (при фиксированных $X_1, \dots, X_n$), для которых выполнено $\mathbb{E}[\xi_i | X_1, \dots, X_n] = 0$ и $\mathbb{E}[\exp(\lambda \xi_i) | X_1, \dots, X_n] \le \exp(\frac{\lambda^2}{2})$ для всех λ. Более того, для любого $x > 0$

$$\begin{aligned}
&P_{Y|X,\varepsilon} \left(\sup_{g\in\mathcal{G}_\mathcal{Y}} \left(\sum_{i=1}^n \varepsilon_i g(X_i, Y_i) - c g(X_i, Y_i) \right) \ge x \right) \\
&\le P_\varepsilon \left(\sup_{f'\in\mathcal{F}^*} \left(\sum_{i=1}^n \varepsilon_i f'(X_i) - \frac{1}{2} hc |f'(X_i)| \right) \ge \frac{x}{2} \right) \\
&\quad + P_\xi \left(\sup_{g'\in\mathcal{G}_{f^*}} \left(\sum_{i=1}^n \xi_i g'(X_i) - \frac{1}{3} h g'(X_i) \right) \ge \frac{x}{3c} \right)
\end{aligned}$$

Доказательство. Обратим внимание, что $g \in \mathcal{G}_\mathcal{Y}$ определяется некоторым клас-

сификатором $f \in \mathcal{F}$.

$$\begin{aligned}
&\mathbb{E}_{Y|X}\mathbb{E}_{\varepsilon} \sup_{g\in\mathcal{G}_{\mathcal{Y}}} \left(\sum_{i=1}^{n} \varepsilon_i g(X_i, Y_i) - c g(X_i, Y_i) \right) \\
&= \mathbb{E}_{Y|X}\mathbb{E}_{\varepsilon} \sup_{f\in\mathcal{F}} \left(\sum_{i=1}^{n} \frac{1}{2}\varepsilon_i Y_i (f(X_i) - f^*(X_i)) - c g_f(X_i, Y_i) \right) \\
&= \mathbb{E}_{Y|X}\mathbb{E}_{\varepsilon} \sup_{f\in\mathcal{F}} \left(\sum_{i=1}^{n} \frac{1}{2}\varepsilon_i (f(X_i) - f^*(X_i)) - c g_f(X_i, Y_i) \right) \\
&= \frac{1}{2}\mathbb{E}_{Y|X}\mathbb{E}_{\varepsilon} \sup_{f\in\mathcal{F}} \left(\sum_{i=1}^{n} \varepsilon_i (f(X_i) - f^*(X_i)) - 2c g_f(X_i, Y_i) \right).
\end{aligned}$$

Рассмотрим член $-\sum_{i=1}^{n} g(X_i, Y_i)$. Обозначая $h'_i = 1 - 2P(f^*(X_i) \neq Y_i | X_i)$ (зависящую X_i-случайную величину), мы знаем, что $1 \geq h'_i \geq h$ с вероятностью 1. Далее $f^*(X_i) \neq Y_i$ имеет условную вероятность (при условии X_i) равную $\frac{1}{2}(1 - h'_i)$, и на этом событии $\frac{1}{2}|f(X_i) - f^*(X_i)| = -g(X_i, Y_i)$. Аналогично, событие $f^*(X_i) = Y_i$ происходит с условной вероятностью равной $\frac{1}{2}(1 + h'_i)$, на этом событии $\frac{1}{2}|f(X_i) - f^*(X_i)| = g(X_i, Y_i)$. Таким образом, обозначая $\xi_i^{(h')} = h'_i + \mathbb{1}[f^*(X_i) \neq Y_i] - \mathbb{1}[f^*(X_i) = Y_i]$, случайные величины $\xi_1^{(h')}, \ldots, \xi_n^{(h')}$ условно независимы при фиксированных $X_1, \ldots, X_n$, для которых выполнено $\mathbb{E}[\xi_i^{(h')} | X_1, \ldots, X_n] = 0$. В частности, если $h'_i = 0$ для всех i, эти случайные величины суть Радемахеровские случайные величины, в то время, как если $h'_i = 1$ эти случайные величины равны 0 с вероятностью 1. Далее

$$\begin{aligned}
-\sum_{i=1}^{n} g(X_i, Y_i) &= -\sum_{i=1}^{n} \frac{h'_i}{2} |f(X_i) - f^*(X_i)| + \sum_{i=1}^{n} \frac{\xi_i^{(h')}}{2} |f(X_i) - f^*(X_i)| \\
&\leq -(\min_i h'_i) \sum_{i=1}^{n} \frac{1}{2} |f(X_i) - f^*(X_i)| + \sum_{i=1}^{n} \frac{\xi_i^{(h')}}{2} |f(X_i) - f^*(X_i)|.
\end{aligned}$$

Используя, $h \le h'_i$ с вероятностью 1, мы получаем

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}_{Y|X}\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}}\left(\sum_{i=1}^n \varepsilon_i(f(X_i)-f^*(X_i)) - 2cg_f(X_i,Y_i)\right)\\
&\le \mathbb{E}_\xi\mathbb{E}_\varepsilon \sup_{f'\in\mathcal{F}^*}\left(\sum_{i=1}^n \varepsilon_i f'(X_i) + c\xi_i^{(h')}|f'(X_i)| - hc|f'(X_i)|\right)\\
&\le \mathbb{E}_\varepsilon \sup_{f'\in\mathcal{F}^*}\left(\sum_{i=1}^n \varepsilon_i f'(X_i) - \frac{1}{2}hc|f'(X_i)|\right)\\
&\quad + c\mathbb{E}_\xi \sup_{f'\in\mathcal{F}^*}\left(\sum_{i=1}^n \xi_i^{(h')}|f'(X_i)| - \frac{1}{2}h|f'(X_i)|\right).
\end{aligned}$$

В итоге, так как $-1 \le \xi_i^{(h_i)} \le 2$, Лемма Хеффдинга ([34] Лемма 8.1) дает $\mathbb{E}[\exp(\lambda\xi_i^{(h_i)})|X_1,\dots,X_n] \le \exp(9\lambda^2/8)$. Первое утверждение Леммы легко следует, если взять $\xi_i = \frac{2}{3}\xi_i^{(h_i)}$.

Для доказательства второго утверждения мы повторяем схожие шаги. Заметим, что $\sup_{g\in\mathcal{G}_\mathcal{Y}}\left(\sum_{i=1}^n \varepsilon_i g(X_i,Y_i) - cg(X_i,Y_i)\right)$ имеют то же распределение (при фиксированных $X_1,\dots,X_n$), что и $\frac{1}{2}\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^n \varepsilon_i(f(X_i)-f^*(X_i)) - 2cg_f(X_i,Y_i)\right)$. В итоге, используя определение ξ_i мы получаем с вероятностью 1 (опять же при фиксированных $X_1,\dots,X_n$)

$$\begin{aligned}
&\frac{1}{2}\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^n \varepsilon_i(f(X_i)-f^*(X_i)) - 2cg_f(X_i,Y_i)\right)\\
&\le \sup_{f'\in\mathcal{F}^*}\left(\sum_{i=1}^n \varepsilon_i f'(X_i) - \frac{1}{2}hc|f'(X_i)|\right)\\
&\quad + c\mathbb{E}_\xi \sup_{f'\in\mathcal{F}^*}\left(\sum_{i=1}^n \xi_i^{(h')}|f'(X_i)| - \frac{1}{2}h|f'(X_i)|\right).
\end{aligned}$$

□

Напомним, что $\mathcal{G}_{f^*} = \{x \to \mathbb{1}[f(x) \ne f^*(x)] \mid f \in \mathcal{F}\}$ и $\mathcal{F}^* = \frac{1}{2}(\mathcal{F} - f^*)$.

Лемма 8 (Локализация)**.** *Определим для $\mathcal{F}$ класс $\mathcal{G} = \mathcal{F}^*$ или $\mathcal{G} = \mathcal{G}_{f^*}$ и пусть $c \in [0, \frac{1}{4}]$ — некоторая фиксированная константа. Пусть $\xi_1,\dots,\xi_n$ — случайные*

величины, условно независимые при данных $X_1, \dots, X_n$, *для которых* $|\xi_i| \lesssim 1$, $\mathbb{E}[\xi_i | X_1, \dots, X_n] = 0$ *и* $\mathbb{E}[\exp(\lambda \xi_i) | X_1, \dots, X_n] \le \exp(\frac{\lambda^2}{2})$ *для всех* λ. *Тогда, если* $\mathcal{G}$ *содержит нулевую функцию, то*

$$\frac{1}{n}\mathbb{E}_\xi \sup_{g \in \mathcal{G}} \left(\sum_{i=1}^n \xi_i g(X_i) - 4c|g(X_i)| \right) \lesssim \frac{\gamma^{\mathrm{loc}}_{c,c}(n, \mathcal{F})}{n}.$$

Доказательство этого утверждения будет приведено ниже. Теперь мы можем перейти к доказательству Теоремы 4.

Теоремы 4. Пусть $\hat{f}$ — произвольный минимизатор эмпирического риска, $\hat{g}$ — соответствующая функция в классе избыточных потерь $\mathcal{G}_\mathcal{Y}$. Очевидно, что $\mathbb{E}(R(\hat{f}) - R(f^*)) = \mathbb{E}P\hat{g}$ и $P_n\hat{g} \le 0$. Тогда для любого $c > 0$,

$$\mathbb{E}(R(\hat{f}) - R(f^*)) \le \mathbb{E}(P\hat{g} - (1+c)P_n\hat{g}) \le \mathbb{E} \sup_{g \in \mathcal{G}_\mathcal{Y}} (Pg - (1+c)P_n g).$$

Используя лемму симметризации (Лемма 4), получаем

$$\mathbb{E} \sup_{g \in \mathcal{G}_\mathcal{Y}} (Pg - (1+c)P_n g) \le \frac{c+2}{n}\mathbb{E}\mathbb{E}_\varepsilon \sup_{g \in \mathcal{G}_\mathcal{Y}} \left(\sum_{i=1}^n \varepsilon_i g(X_i, Y_i) - \frac{c}{c+2} g(X_i, Y_i) \right).$$

Применяя лемму сжатия (Лемма 7), получаем

$$\begin{aligned}
&\frac{c+2}{n}\mathbb{E}\mathbb{E}_\varepsilon \sup_{g \in \mathcal{G}_\mathcal{Y}} \left(\sum_{i=1}^n \varepsilon_i g(X_i, Y_i) - \frac{c}{c+2} g(X_i, Y_i) \right) \\
&\le \frac{(c+2)}{n}\mathbb{E}\mathbb{E}_\varepsilon \sup_{f' \in \mathcal{F}^*} \left(\sum_{i=1}^n \varepsilon_i f'(X_i) - \frac{hc}{2(c+2)} |f'(X_i)| \right) \\
&\quad + \frac{3c}{2n}\mathbb{E}\mathbb{E}_\xi \sup_{f' \in \mathcal{F}^*} \left(\sum_{i=1}^n \xi_i |f'(X_i)| - \frac{1}{3} h |f'(X_i)| \right).
\end{aligned}$$

Используем (Лемму 8). Условия на ξ_i и ε_i, требуемые Леммой 8, следуют из Леммы 7, и все функции класса $\mathcal{F}^*$ принимают только значения $\{-1, 0, 1\}$. Таким образом, для фиксированного c,

$$\frac{(c+2)}{n}\mathbb{E}\mathbb{E}_\varepsilon \sup_{f' \in \mathcal{F}^*} \left(\sum_{i=1}^n \varepsilon_i f'(X_i) - \frac{hc}{2(c+2)} |f'(X_i)| \right) \lesssim \frac{\gamma^{\mathrm{loc}}_{h,h}(n)}{n}.$$

Такая же оценка выполнена и для $\frac{3c}{2n}\mathbb{E}\mathbb{E}_\xi \sup_{f'\in\mathcal{F}^*}\left(\sum_{i=1}^n \xi_i|f'(X_i)| - \frac{1}{3}h|f'(X_i)|\right)$. Доказательство оценки с большой вероятностью аналогично и будет предложено позже. Оценка на величину $\gamma^{\text{loc}}_{h,h}(n)$ дается в Утверждении 4 ниже.

□

Следующее Утверждение завершает доказательство Теоремы 4.

Утверждение 4. *В условиях Теоремы 4 для всех $h \in (0,1]$ выполнено*

$$\frac{d+\log\left(nh^2\wedge \mathbf{s}\right)}{h}\wedge\sqrt{dn} \lesssim \gamma^{\text{loc}}_{h,h}(n) \lesssim \frac{d\log\left(\frac{nh^2}{d}\wedge\mathbf{s}\right)}{h} + \frac{d\log(\frac{1}{h})}{h}.$$

Доказательство. Первая часть доказательства схожа с доказательством Теоремы 17 из [33], с модификациями, необходимыми для получения верхней оценки на $\mathcal{M}^{\text{loc}}_1(\mathcal{F},\gamma,n,h)$. Пусть супремум в определении локальной энтропии достигается на некотором множестве $\{x_1,\dots,x_n\}$, функции $f\in\mathcal{F}$, и некотором $\varepsilon\in[\gamma,n]$. Обозначая $r=\varepsilon/n$ и $\mathcal{M}_r$ как максимальную $(rn/2)$-упаковку (с расстоянием ρ_H) класса $\mathcal{B}_H(f,rn/h,\{x_1,\dots,x_n\})$, так что $|\mathcal{M}_r| = \mathcal{M}^{\text{loc}}_1(\mathcal{F},\gamma,n,h)$. Также введем равномерное распределение P_X на $\{x_1,\dots,x_n\}$ и зафиксируем $m = \left\lceil \frac{4}{r}\log(|\mathcal{M}_r|)\right\rceil$. Пусть $X_1,\dots,X_m$ — m независимых P_X-распределенных случайных величин, и пусть A обозначает событие, что для всех $g,g'\in\mathcal{M}_r$ с $g\neq g'$, существует $i\in\{1,\dots,n\}$ такое что $g(X_i)\neq g'(X_i)$. Для данной пары различных функций $g,g'\in\mathcal{M}_r$, они отличаются на некотором X_i с вероятностью

$$1-(1-P_X(g(X)\neq g'(X)))^m > 1-\exp(-rm/2) \geq 1-\frac{1}{|\mathcal{M}_r|^2}.$$

Используя неравенство Буля для всех неупорядоченных пар $g,g'\in\mathcal{M}_r$, мы получаем $\mathbb{P}(A)>\frac{1}{2}$. На событии A, функции в $\mathcal{M}_r$ доставляют различные классификации $X_1,\dots,X_m$. Для всех

$$X_i \notin \text{DIS}(\mathcal{B}_H(f,rn/h,\{x_1,\dots,x_n\}),$$

все классификаторы $\mathcal{M}_r$ совпадают. Таким образом, $|\mathcal{M}_r|$ ограничено числом различных классификаций $\{X_1,\dots,X_m\}\cap\text{DIS}(\mathcal{B}_H(f,rn/h))$, реализуемых классифи-

каторами из $\mathcal{F}$. С помощью оценки Чернова на событии B с $\mathbb{P}(B) \geq \frac{1}{2}$ мы имеем $|\{X_1, \ldots, X_m\} \cap \mathrm{DIS}(\mathcal{B}_H(f, rn/h))| \leq 1 + 2eP_X(\mathrm{DIS}(\mathcal{B}_H(f, rn/h))m$. Используя определение емкости Александера $\tau(\cdot)$ (Определение 4) мы получаем

$$1 + 2eP_X(\mathrm{DIS}(\mathcal{B}_H(f, rn/h)))m \leq 1 + 2e\tau\left(\frac{r}{h}\right)\frac{r}{h}m \leq 11e\tau\left(\frac{r}{h}\right)\frac{\log(|\mathcal{M}_r|)}{h}.$$

С вероятностью не менее $\frac{1}{2}$,

$$|\{X_1, \ldots, X_m\} \cap \mathrm{DIS}(\mathcal{B}_H(f, rn/h))| \leq 11e\tau\left(\frac{r}{h}\right)\frac{\log(|\mathcal{M}_r|)}{h}.$$

Используя неравенство Буля, мы получаем с вероятностью не менее нуля, что существует последовательность из не менее чем $11e\tau\left(\frac{r}{h}\right)\frac{\log(|\mathcal{M}_r|)}{h}$ элементов, такая что все функции из $\mathcal{M}_r$ классифицируют эту последовательность различными способами. С помощью леммы Вапника и Червоненкиса мы получаем

$$|\mathcal{M}_r| \leq \left(\frac{11e^2\tau\left(\frac{r}{h}\right)\frac{\log(|\mathcal{M}_r|)}{h}}{d}\right)^d.$$

Используя следствие 4.1 из [40], мы получаем

$$\log(|\mathcal{M}_r|) \leq 2d\log\left(11e^2\tau\left(\frac{r}{h}\right)\frac{1}{h}\right).$$

Используя $\tau\left(\frac{r}{h}\right) \leq \mathbf{s} \wedge \frac{h}{r} \leq \mathbf{s} \wedge \frac{nh}{\gamma}$ (Теорема 10 в [33]), мы в итоге получаем

$$\log(\mathcal{M}_1^{\mathrm{loc}}(\mathcal{F}, \gamma, n, h)) \leq 2d\log\left(11e^2\left(\frac{n}{\gamma} \wedge \frac{\mathbf{s}}{h}\right)\right).$$

Теперь мы ограничиваем сверху $\gamma_{h,h}^{\mathrm{loc}}(n)$, зная что

$$h\gamma_{h,h}^{\mathrm{loc}}(n) \leq 2d\log\left(11e^2\left(\frac{n}{\gamma_{h,h}^{\mathrm{loc}}(n)} \wedge \frac{\mathbf{s}}{h}\right)\right).$$

Мы очевидно имеем $\gamma_{h,h}^{\mathrm{loc}}(n) \leq \frac{2d\log\left(11e^2\frac{\mathbf{s}}{h}\right)}{h}$. Для $\gamma = \frac{2d\log\left(11e^2\frac{nh}{d}\right)}{h}$ мы получаем $h\gamma = 2d\log\left(11e^2\frac{nh}{d}\right)$, но $2d\log\left(11e^2\frac{n}{\gamma}\right) \leq 2d\log\left(11e^2\frac{nh}{d}\right)$, если $h > \frac{d}{11en}$. В итоге

$$\gamma_{h,h}^{\mathrm{loc}}(n) \leq \frac{2d\log\left(11e^2\left(\frac{nh}{d} \wedge \frac{\mathbf{s}}{h}\right)\right)}{h}.$$

Теперь мы доказываем нижнюю оценку. Из оценки(2.9), установленной выше, мы знаем, что $\frac{\gamma^{\mathrm{loc}}_{h,h}(n)}{n}$ с точностью до констант является верхней оценкой для $\mathbb{E}(R(\hat{f}) - R(f^*))$, выполненной для всех минимизаторов эмпирического риска $\hat{f}$. Тогда любая нижняя оценка на $\sup\limits_{P\in\mathcal{P}(h,\mathcal{F})} \mathbb{E}(R(\hat{f})-R(f^*))$, выполненная для всех минимизаторов эмпирического риска, также является и нижней оценкой для $\frac{\gamma^{\mathrm{loc}}_{h,h}(n)}{n}$. В частности, известно, что [5, 30] для любой процедуры $\tilde{f}$, если $h \geq \sqrt{\frac{d}{n}}$, то $\sup\limits_{P\in\mathcal{P}(h,\mathcal{F})} \mathbb{E}(R(\tilde{f}) - R(f^*)) \gtrsim \frac{d+(1-h)\log(nh^2\wedge \mathbf{s})}{nh}$, а если $h < \sqrt{\frac{d}{n}}$, то $\sup\limits_{P\in\mathcal{P}(h,\mathcal{F})} \mathbb{E}(R(\tilde{f}) - R(f^*)) \gtrsim \sqrt{\frac{d}{n}}$. Далее, для минимизаторов эмпирического риска в работе [30] доказано, что любая верхняя оценка на $\sup\limits_{P\in\mathcal{P}(1,\mathcal{F})} \mathbb{E}(R(\hat{f}) - R(f^*))$, выполненная для всех минимизаторов эмпирического риска $\hat{f}$, должна быть по порядку не меньше $\frac{\log(n\wedge \mathbf{s})}{n}$. Вместе эти нижние оценки дают $\gamma^{\mathrm{loc}}_{h,h}(n) \gtrsim \frac{d+\log(nh^2\wedge \mathbf{s})}{h} \wedge \sqrt{dn}$. □

2.6. Минимаксные нижние оценки

В этом разделе мы доказываем, что для условий малого шума Массара, неподвижные точки локальной метрической энтропии появляются также и в минимаксных нижних оценках. Результаты построены с помощью классических техник для получения нижних оценок, [5, 12, 38], используемых ранее только для специальных классов. Нам потребуется следующее определение, которое будет мотивировано ниже.

Определение 6. *Фиксируем класс классификаторов $\mathcal{F}$. Предположим, что имеется константа $c \geq 1$, такая что для любого N в определении $\mathcal{M}^{\mathrm{loc}}_1(\mathcal{F}, \gamma^{\mathrm{loc}}_{h,1}(N), N, 1)$ супремум по радиусу шара достигается на радиусе $\varepsilon_h(N) \leq c\gamma^{\mathrm{loc}}_{h,1}(N)$. Такой класс будем называть* c-псевдовыпуклым.

Теорема 5. *Пусть $\tilde{f}$ — любой алгоритм классификации, построенный по n наблюдениям. Зафиксируем любой $c_{\mathcal{F}}$-псевдовыпуклый класс $\mathcal{F}$ и любое h, удовле-*

творяющее $\sqrt{\frac{d}{n}} \leq h \leq 1$. *Тогда существует* $P \in \mathcal{P}(h, \mathcal{F})$, *такое что*

$$\mathbb{E}(R(\tilde{f}) - R(f^*)) \gtrsim \frac{d}{nh} + \frac{1}{c_{\mathcal{F}}} \frac{(1-h)\gamma_{h,1}^{\mathrm{loc}}\left(\lceil \frac{nc_{\mathcal{F}}h}{(1-h)} \rceil\right)}{n}. \tag{2.12}$$

Условия, включающие константу $c_{\mathcal{F}}$ могут быть ослаблены. Из нашего доказательства будет понятно, что мы можем убрать условие псевдовыпуклости с помощью переопределения локальной метрической энтропии, убрав из него супремум по радиусу. Альтернативно, можно убрать супремум, наложив условия монотонности. Заметим, что условия монотонности неявно используются в предыдущих работах [8, 12]. В обоих случаях наша нижняя оценка будет выполнена с константой $c_{\mathcal{F}} = 1$. Более того, оценка (2.12) верна для любого класса $\mathcal{F}$, так как мы всегда можем рассматривать $c_{\mathcal{F}}(N)$ вместо $c_{\mathcal{F}}$, что является минимальным натуральным числом, удовлетворяющим $\varepsilon_h(N) \leq c_{\mathcal{F}}(N)\gamma_{h,1}^{\mathrm{loc}}(N)$. Заметим также, что проблемы монотонности не возникают для выпуклых классов, как показано Мендельсоном в работе [14]. Это и есть наша мотивация для названия условия: локальная энтропия ведет себя почти также, как если бы находились в выпуклом случае. В следующем разделе мы приведем примеры псевдовыпуклых классов.

Следующая лемма дана в [41] (Следствие 2.18).

Лемма 9 (Бирдж)**.** *Пусть* $\{P_i\}_{i=0}^N$ *— конечное семейство распределений, определенных на одном и том же измеримом пространстве и* $\{A_i\}_{i=0}^N$ *есть семейство непересекающихся событий. Тогда*

$$\min_{0 \leq i \leq N} P_i(A_i) \leq 0.71 \vee \frac{\sum_{i=1}^N KL(P_i \| P_0)}{N \log(N+1)}.$$

Теорема 5. Рассмотрим сначала значение $\mathcal{M}_1^{\mathrm{loc}}(\mathcal{F}, \gamma_{h,1}^{\mathrm{loc}}(N), N, 1)$. Вспомним, что определение этой величины включает супремум по $f \in \mathcal{F}$ и по N-элементным подмножествам $\mathcal{X}^n$. Без потерь общности мы предположим, что этот супремум достигается на некотором классификаторе $g \in \mathcal{F}$, некотором $\varepsilon_h(N) \in [\gamma_{h,1}^{\mathrm{loc}}(N), N]$ и на множестве $\mathcal{X}_N = \{x_1, \ldots, x_N\}$. Пусть k_i определяет число копий x_i в $\mathcal{X}_N$.

Мы определим $P_{\mathcal{X}_N}(\{x_i\}) = \frac{k_i}{N}$. Если все элементы различны, то это распределение — есть равномерное распределение на $\mathcal{X}_N$. Введем естественную параметризацию: любой классификатор представляется N-мерным бинарным вектором и два вектора (для классификаторов g, f) различаются только на множестве, соответствующем $\mathrm{DIS}(\{g, f\}) \cap \mathcal{X}_N$. Множество бинарных векторов, соответствующих классификаторам из $\mathcal{F}$, будет обозначаться $\mathcal{B}$. Для данного бинарного вектора b определим $P_b = P_{\mathcal{X}_\mathcal{N}} \times P^b_{Y|X}$, где $P^b_{Y=1|X_i} = \frac{1+(2b_i-1)h}{2}$. Пусть $\tilde{f}_b$ обозначает классификатор $\tilde{f}$, получаемый с помощью алгоритма обучения, когда P_b является распределением данных. Пусть также $\tilde{b}$ обозначает бинарный вектор, соответствующий $\tilde{f}_b$; таким образом, $\tilde{b}$ является случайным вектором, который зависит от параметра b посредством n точек, имеющих распределение P_b. Известно [34], что $R(\tilde{f}) - R(f^*) = \mathbb{E}(|\eta(X)|\mathbb{1}[\tilde{f}(X) \neq f^*(X)]|\tilde{f}) \geq hP(\tilde{f}(X) \neq f^*(X)|\tilde{f})$, когда $P \in \mathcal{P}(h, \mathcal{F})$. Более того, $P_b(\tilde{f}_b(X) \neq f^*(X)|\tilde{f}_b) = \frac{\rho_H(\tilde{b},b)}{N}$. Таким образом,

$$\sup_{P\in\mathcal{P}(h,\mathcal{F})} \mathbb{E}(R(\tilde{f}) - R(f^*)) \geq \max_{b\in\mathcal{B}} \mathbb{E}\left(hP_b((x,y) : \tilde{f}_b(x) \neq f^*(x))\right)$$
$$\geq \frac{h}{N} \max_{b\in\mathcal{B}} \mathbb{E}(\rho_H(\tilde{b}, b)).$$

Пусть b^* — бинарный вектор $\mathcal{B}$, соответствующий классификатору g, определенному выше. Выберем максимальное подмножество $\mathcal{B}^{\mathrm{loc}} \subset \mathcal{B}$ удовлетворяющее свойству, что для всех $b' \in \mathcal{B}^{\mathrm{loc}}$ выполнено $\rho_H(b', b^*) \leq \varepsilon_h(N)$ и для любой пары векторов $b', b'' \in \mathcal{B}^{\mathrm{loc}}$ мы имеем $\rho_H(b', b'') > \varepsilon_h(N)/2$. Далее определим $\breve{b}$ как минимзатор $\rho_H(\breve{b}, \tilde{b})$ среди $\mathcal{B}^{\mathrm{loc}}$. В частности, если $b \in \mathcal{B}^{\mathrm{loc}}$, мы получаем $\rho_H(\breve{b}, \tilde{b}) \leq \rho_H(b, \tilde{b})$, поэтому $\rho_H(\breve{b}, b) \leq \rho_H(\breve{b}, \tilde{b}) + \rho_H(\tilde{b}, b) \leq 2\rho_H(\tilde{b}, b)$. Таким образом,

$$\frac{h}{N} \max_{b\in\mathcal{B}} \mathbb{E}(\rho_H(\tilde{b}, b)) \geq \frac{h}{N} \max_{b\in\mathcal{B}^{\mathrm{loc}}} \mathbb{E}(\rho_H(\tilde{b}, b)) \geq \frac{h}{2N} \max_{b\in\mathcal{B}^{\mathrm{loc}}} \mathbb{E}(\rho_H(\breve{b}, b)).$$

Вспоминаем, что $\breve{b}$ строго определяется по классификатору $\tilde{f}$, который в свою очередь есть функция от n точек, поэтому мы можем определить несовместные события A_b of $(\mathcal{X} \times \mathcal{Y})^n$ для $b \in \mathcal{B}^{\mathrm{loc}}$, где A_b, относится к выборкам, дающим $\breve{b} = b$. Теперь, используя неравенство Маркова и факт, что векторы $\mathcal{B}^{\mathrm{loc}}$ являются

$\frac{\varepsilon_h(N)}{2}$-отделенными, мы получаем $\mathbb{E}(\rho_H(\breve{b}, b)) \geq \frac{\varepsilon_h(N)}{2} P(\breve{b} \neq b) = \frac{\varepsilon_h(N)}{2}(1 - P_b^n(A_b))$. Таким образом, мы получаем

$$\frac{h}{2N} \max_{b \in \mathcal{B}^{\mathrm{loc}}} \mathbb{E}(\rho_H(\breve{b}, b)) \geq \frac{h\varepsilon_h(N)}{4N} \left(1 - \min_{b \in \mathcal{B}^{\mathrm{loc}}} P_b^n(A_b)\right).$$

Мы используем Лемму 9 чтобы ограничить $\min\limits_{b \in \mathcal{B}^{\mathrm{loc}}} P_b^n(A_b)$. Для этого заметим, что для всех $b', b'' \in \mathcal{B}^{\mathrm{loc}}$, стандартные выкладки дают

$$\mathrm{KL}(P_{b'}^n \| P_{b''}^n) = \frac{n}{N} h \ln\left(\frac{1+h}{1-h}\right) \rho_h(b', b'').$$

Так как для $x > 0$ выполнено $\ln(x+1) \leq x$, мы получаем $h \ln\left(\frac{1+h}{1-h}\right) \leq \frac{2h^2}{1-h}$. Далее для всех $b', b'' \in \mathcal{B}^{\mathrm{loc}}$ мы имеем $\rho_H(b', b'') \leq 2\varepsilon_h(N)$. Таким образом,

$$\mathrm{KL}(P_{b'}^n \| P_{b''}^n) \leq \frac{4nh^2\varepsilon_h(N)}{N(1-h)}.$$

С помощью Леммы 9,

$$\min_{b \in \mathcal{B}^{\mathrm{loc}}} P_b^n(A_b) \leq 0.71 \vee \frac{\frac{4nh^2\varepsilon_h(N)}{N(1-h)}}{\log(|\mathcal{B}^{\mathrm{loc}}|)}. \tag{2.13}$$

Обратим внимание, что

$$\log(|\mathcal{B}^{\mathrm{loc}}|) = \log(\mathcal{M}_1^{\mathrm{loc}}(\mathcal{F}, \varepsilon_h(N), N, 1)) \geq h\gamma_{h,1}^{\mathrm{loc}}(N) \geq h\varepsilon_h(N)/c_{\mathcal{F}}.$$

Выбирая $N = \left\lceil \frac{6nc_{\mathcal{F}}h}{(1-h)} \right\rceil$, получаем

$$\frac{4nh^2\varepsilon_h(N)}{N(1-h)} \leq \frac{2h\varepsilon_h(N)}{3c_{\mathcal{F}}} \leq \frac{2}{3}\log(|\mathcal{B}^{\mathrm{loc}}|),$$

так что выражение в правой части (2.13) есть 0.71. Мы получаем для $h < 1$,

$$\begin{aligned}\sup_{P \in \mathcal{P}(h,\mathcal{F})} \mathbb{E}(R(\tilde{f}) - R(f^*)) &\geq 0.29 \frac{h\varepsilon_h(N)}{4N} \\ &\geq \frac{0.29}{48c_{\mathcal{F}}} \frac{(1-h)\varepsilon_h(N)}{n} \geq \frac{0.29}{48c_{\mathcal{F}}} \frac{(1-h)\gamma_{h,1}^{\mathrm{loc}}(N)}{n}.\end{aligned}$$

Член $\frac{d}{nh}$ для $h > \sqrt{\frac{d}{n}}$ есть часть стандартной нижней оценки [5]. □

Следствие 2. *Пусть* $0 < C_0 \leq C_1 < 1$. *Тогда, если параметр* h *такой, что* $C_0 \vee \sqrt{\frac{d}{n}} \leq h \leq C_1$, *то для любого VC класса* $\mathcal{F}$ *верхняя оценка для минимизатора эмпирического риска* (2.9) *и нижняя оценка* (2.12) *совпадают с точностью до абсолютных констант (также появляющихся в аргументах стационарной точки), которые могут зависеть только от* C_0, C_1 *и* $c_{\mathcal{F}}$.

Известно [5], что минимизатор эмпирического риска оптимален в режиме $0 \leq h < \sqrt{\frac{d}{n}}$. Интересно, что наше следствие точно неверно в случае $C_1 = 1$. Оптимальная оценка в случае бесшумной классификации имеет порядок $\frac{d}{n} + \frac{\log(\frac{1}{\delta})}{n}$ [42], но минимизатор эмпирического риска не может иметь этот порядок сходимости [15, 17, 18]. Этот факт отлично отражается в нижней оценке. Когда параметр h близок к 1, член $\frac{(1-h)\gamma_{h,1}^{\mathrm{loc}}}{nc_{\mathcal{F}}}$ пропадает и мы получаем стандартную $\frac{d}{n}$ нижнюю оценку [43].

2.7. Оценивание неподвижной точки для некоторых специальных классов

В этом разделе мы предложим два примера точного оценивания неподвижных точек локальных энтропий. Сначала мы рассмотрим пороговые классификторы, введенные ранее в примере 8. Для этого специального класса $d = 1$ и $\mathbf{s} = 2$. Из Теоремы 4 мы получаем $\frac{1}{h} \lesssim \gamma_{h,h}^{\mathrm{loc}}(n) \lesssim \frac{\log(\frac{1}{h})}{h}$ и значит для этого класса $\gamma_{h,h}^{\mathrm{loc}}(n) \simeq \frac{\log(\frac{1}{h})}{h}$. В частности, в бесшумном случае $\gamma_{1,1}^{\mathrm{loc}}(n) \simeq 1$.

Другой пример — класс линейных разделителей в $\mathbb{R}^k$ для $k \geq 2$. Известно, что класс обладает VC размерностью $d = k + 1$. Легко проверить, что для этого класса $\mathbf{s} = \infty$ [33].

Утверждение 5. *Для множества линейных разделителей* $\mathcal{F}$ *в* $\mathbb{R}^d$, *если* $d \geq 2$,*то для всех* $h > \sqrt{\frac{d}{n}}$

$$\frac{d \log\left(\frac{nh^2}{d}\right)}{h} \lesssim \gamma_{h,h}^{\mathrm{loc}}(n) \lesssim \frac{d \log\left(\frac{nh}{d}\right)}{h}.$$

В частности, $\gamma^{\text{loc}}_{1,1}(n) \simeq d\log(\frac{n}{d})$.

Доказательство. Верхняя оценка напрямую следует из Теоремы 4. Выберем некоторое специальное множество точек $x_1, \ldots, x_n \in \mathbb{R}^d$. Известно (Теорема 6.5 в [27]), что в $\mathbb{R}^d$ существует так называемый *циклический политоп* с n вершинами, такой что он имеет ровно $\binom{n}{k}$ $(k-1)$-мерных граней для любого $k \leq \lfloor \frac{d}{2} \rfloor$. Мы выбираем $x_1, \ldots, x_n$, таких что x_i есть вершина циклического политопа. Мы фиксируем любой линейный разделитель f_1 такой что $x_i, \ldots, x_n$ находятся в одной и той же полуплоскости относительно этого линейного разделителя. Без ограничения общности мы можем предположить, что $f_1(x_1) = \ldots = f_1(x_n) = -1$. Использую свойство циклического политопа, мы получаем, что $\mathcal{F}$ содержит все классификаторы с не более чем $\lfloor \frac{d}{2} \rfloor$ единицами на заданных точках. Мы обозначаем это подмножество как $\mathcal{F}_{d/2}$. Анализ этого класса дает (Теорема 5 в [5]) нижнюю оценку $\frac{(1-h)d\log(\frac{nh^2}{d})}{nh}$ для $R(\hat{f}) - R(f^*)$ при условии, что $h > \sqrt{\frac{d}{n}}$. Из Теоремы 4 мы знаем, что эта нижняя оценка также и нижняя оценка для$\gamma^{\text{loc}}_{h,h}(n)$. Таким образом, $\frac{(1-h)d\log(\frac{nh^2}{d})}{h} \lesssim \gamma^{\text{loc}}_{h,h}(n)$. Одновременно, мы получаем $\gamma^{\text{loc}}_{1,1}(n) \leq \gamma^{\text{loc}}_{h,h}(n)$. Поэтому достаточно получить нижнюю оценку для $\gamma^{\text{loc}}_{1,1}(n)$, которая может быть получена как нижняя оценка для минимизаторов эмпирического риска в бесшумном случае. Известно (Теорема 6 в [17] или Теорема 5 в [18]) что для этого класса $\mathcal{F}_{d/2}$ в бесшумном случае существует минимизатор эмпирического риска, такой что с вероятностью $\frac{1}{2}$ имеет место $\frac{d\log(\frac{n}{d})}{n} \lesssim R(\hat{f})$. Это влечет $\frac{d\log(\frac{n}{d})}{n} \lesssim \mathbb{E}R(\hat{f})$, а значит $d\log(\frac{n}{d}) \lesssim \gamma^{\text{loc}}_{1,1}(n)$. В итоге мы получаем $d\log(\frac{n}{d}) \vee \frac{(1-h)d\log(\frac{nh^2}{d})}{h} \lesssim \gamma^{\text{loc}}_{h,h}(n)$. Мы завершаем доказательство, заметив что $\frac{d\log\left(\frac{nh^2}{d}\right)}{h} \lesssim d\log(\frac{n}{d}) \vee \frac{(1-h)d\log(\frac{nh^2}{d})}{h}$.

□

Заметим, что нижняя оценка (2.12) может быть применена к обоим классам.

2.8. Обсуждение и некоторые открытые вопросы

Локальные энтропии хорошо известны в работах по статистике, начиная с работ Ле Кама [44]. Обычно локальные энтропии возникают в минимаксных нижних оценках [13, 14, 38] а также являются необходимым и достаточным условием состоятельности минимизатора эмпирического риска в непараметрической регрессии [45]. Одновременно верхние оценки чаще выражаются глобальными энтропиями. Интересно, что иногда можно получить правильные порядки для риска, рассматривая лишь глобальные упаковки [38, 46]. В общем случае эмпирические числа покрытия в статистике имеют два типа поведения. Существуют параметрические и VC классы, для которых логарифм числа покрытия имеет порядок $\log(\frac{1}{\varepsilon})$ и непараметрические классы, где они имеют порядок ε^{-p} для некоторых $p > 0$. В работе [38] показано, что для этих непараметрических классов локальные и глобальные энтропии имеют один и тот же порядок. Таким образом, локализация для таких классов не дает существенных преимуществ и минимаксные порядки достигаются с помощью глобальных энтропий [46]. Случай параметрических и особенно VC классов более тонкий. Наши результаты показывают, что локализация нужна для VC классов, но не всегда. Некоторые параметрические классы обладают свойствами непараметрических классов: их локальные энтропии имеют тот же порядок, что и глобальные. Таким примером является класс $\mathcal{F}_{d/2}$, введенный ранее. В то же время, например, классы пороговых классификаторов, не могут быть правильно проанализированы, основываясь только на глобальной энтропии.

Заметим, что зависящие от распределения локальные энтропии появлялись ранее в верхних оценках для задач классификации под названием *размерности удвоения*. Для класса $\mathcal{F}$ и распределения P_X определим размерность удвоения

$$D(\mathcal{F}, \gamma) = \max_{f \in \mathcal{F}} \max_{\varepsilon \geq \gamma} \log(\mathcal{N}(\mathcal{B}_{P_X}(f, \varepsilon), \varepsilon/2)), \tag{2.14}$$

где $\mathcal{B}_{P_X}(f, \varepsilon) = \{g \in \mathcal{F} | P_X(f(X) \neq g(X)) \leq \varepsilon\}$ и $\mathcal{N}(\mathcal{G}, \varepsilon)$ есть ε-число покрытия $\mathcal{G}$ по расстоянию $P_X(g(X) \neq g'(X))$. В работе [47] доказано, что в бесшумном

случае для любого $\varepsilon > 0$, если

$$n \gtrsim \frac{d + D(\mathcal{F}, \varepsilon_0)}{\varepsilon} \sqrt{\log\left(\frac{1}{\varepsilon}\right)} + \frac{\log(\frac{1}{\delta})}{\varepsilon},$$

то с вероятностью хотя бы $1 - \delta$, для любого минимизатора эмпирического риска $\hat{f}$ мы получам $R(\hat{f}) \leq \varepsilon$. Здесь $\varepsilon_0 = \varepsilon \exp\left(-\sqrt{\log(\frac{1}{\varepsilon})}\right)$. Легко показать, что когда рассматривается независящая от распределений постановка, последняя оценка слабее из-за присутствия квадратного корня из дополнительного логарифмического фактора. Следующее простое неравенство сравнивает размерность удвоения с локальной метрической энтропией. Для любого $\gamma \in \mathbb{N}$,

$$\log(\mathcal{M}_1^{\text{loc}}(\mathcal{F}, \gamma, n, 1)) \leq 2 \sup_{P_X} D(\mathcal{F}, \gamma/n). \tag{2.15}$$

Для доказательства этого неравенства рассмотрим равномерное распределение P_X на n точках, максимизирующих размер локальной упаковки в левой части выражения. Константный член 2 появляется из-за рассмотрения упаковок (а не покрытий) в определении локальной метрической энтропии. В работе [47] также исследуется процедура отличная от минимизатора эмпирического риска в бесшумном случае. Авторами для специальной процедуры, доказано, что ошибка классификации в бесшумном случае ограничена сверху величиной порядка $\varepsilon \approx \frac{D(\mathcal{F}, \varepsilon/4)}{n} + \frac{\log(\frac{1}{\delta})}{n}$, с вероятностью $1 - \delta$. В свете (2.15) мы видим, что в худшем случае данная оценка не лучше чем оценка, полученная в Теореме 4 для минимизатора эмпирического риска.

Мы сравнили наши оценки с лучшим ослаблением оценок, полученных на основании локальных Радемахеровских сложностей. Однако, мы легко можем произвести сравнение и с самим локальными Радемахеровскими оценками. Для этого нам понадобится следующая Теорема.

Теорема 6 (Нижняя оценка Судакова для процессов Бернулли [48]). *Пусть $V \subset \mathbb{R}^n$ — конечное множество, такое что для любых $v_1, v_2 \in V$, если $v_1 \neq v_2$, то $\|v_1 - v_2\|_2 \geq a$ для некоторого $a > 0$ и для любых $v \in V$ выполнено $\|v\|_\infty \leq b$ для*

некоторого $b > 0$. *Тогда*

$$\mathbb{E}_\varepsilon \sup_{v \in V} \sum_{i=1}^{n} \varepsilon_i v_i \gtrsim a\sqrt{\log |V|} \wedge \frac{a^2}{b}. \tag{2.16}$$

Для простоты мы рассмотрим лишь бесшумный случай и постановку, не зависящую от распределений. Зафиксируем $x_1, \ldots, x_n$. Применяя следствие 5.1 из [4], мы получаем

$$\mathbb{E}R(\hat{f}) \lesssim \sup_{x_1,\ldots,x_n} r^*,$$

где r^* — неподвижная точка локальной Радемахеровской сложности, которая является решением следующего уравнения

$$\frac{1}{n}\mathbb{E}_\varepsilon \sup_{g \in \mathrm{star}(\mathcal{G}_{f^*}), P_n g \le 2r} \sum_{i=1}^{n} \varepsilon_i g(x_i) = r,$$

где $\mathrm{star}(\mathcal{G})$ обозначает *звездную оболочку* класса $\mathcal{G}$: она состоит из функций вида αg, где $g \in \mathcal{G}$ и $\alpha \in [0, 1]$. Так как $\mathrm{star}(\mathcal{G}_{f^*})$ звездное, можно легко показать, что локальные энтропии не убывают по радиусу. Используя этот факт вместе с (2.16), можно показать, что

$$\mathbb{E}_\varepsilon \sup_{g \in \mathrm{star}(\mathcal{G}_{f^*}), P_n g \le \frac{2\gamma}{n}} \sum_{i=1}^{n} \varepsilon_i g(x_i) \gtrsim \sqrt{\gamma}\sqrt{\log(\mathcal{M}_1^{\mathrm{loc}}(\mathcal{F}, \gamma, n, 1))} \wedge \gamma.$$

Отсюда следует, что $\frac{\gamma_{1,1}^{\mathrm{loc}}(n)}{n} \lesssim r^*$. Таким образом, наши оценки не хуже, чем оценки основанные только на локальной Радемахеровской сложности. Мы ищем неподвижные точки в левой части выражения (2.16), в то время как Радемахеровский анализ напрямую работает со стационарными точками супремума локализованного процесса. Следующие два вопроса мы оставляем открытыми:

1. Интересно улучшить наши оценки в режимах, когда h близко к единице. Известно [5], что если $h < \sqrt{\frac{d}{n}}$, то контроль Радемахеровского процесса, основаный на интеграле Дадли [49], дает минимаксно оптимальный порядок $\sqrt{\frac{d}{n}}$. Более того, известно, что оценки, основанные лишь на одном покрытии субоптимальны в этом случае. Если зафиксировать $h = \sqrt{\frac{d}{n}}$, то оценка

Жине и Колчинского (2.1) (также основанная на интеграле Дадли) дает оптимальный порядок $\sqrt{\frac{d}{n}}$ по математическому ожиданию. Одновременно мы знаем, что их оценка неоптимальна в режиме, когда h близко к 1. Из-за дополнительного члена $\frac{d\log\left(\frac{1}{h}\right)}{h}$ в (2.11) наша оценка (2.9) может гарантировать только субоптимальный порядок $\sqrt{\frac{d}{n}}\log\left(\frac{n}{d}\right)$, когда $h=\sqrt{\frac{d}{n}}$, но мы знаем, что для других значений h наша оценка существенно лучше. Тем не менее, разумно предположить, что существует непрерывный переход в терминах параметра h, из режима Дадли интеграла к режиму, когда $h<\sqrt{\frac{d}{n}}$ и локальная энтропия дает правильный порядок риска.

2. Мы обсуждали, что минимизатор риска неоптимален, когда $h=1$. Тем не менее интересно точно охарактеризовать риск минимизатора эмпирического риска в этом режиме. Напомним, что случай, когда h отделен от нуля и единицы покрыт нашим следствием 2. Мы предполагаем, что в бесшумном случае и в случае, когда h близок к 1 наша оценка (2.10) является наилучшей возможной оценкой для произвольного минимизатора эмпирического риска. Как минимум эта гипотеза выполнена для классов в разделе 2.7. Частичный анализ сложности минимизаторов эмпирического риска был произведен в работе [30]. В ней показано, что риск некоторого минимизатора эмпирического риска находится между верхней оценкой (2.3), (2.4) и нижней оценкой

$$R(\hat{f})-R(f^*)\gtrsim\frac{d}{nh}+\frac{\log(nh^2\wedge\mathbf{s})}{nh}+\frac{\log(\frac{1}{\delta})}{nh}, \tag{2.17}$$

с вероятностью не менее $1-\delta$ для некоторого $P\in\mathcal{P}(h,\mathcal{F})$ (и некоторого минимизатора эмпирического риска).

2.9. Доказательства

Утверждение 6 (Неравенства Чернова)**.** *Пусть Z имеет биномиальное распределение с параметрами p,n. Тогда для любого $\eta\in(0,1)$*

$$P\left[Z>(1+\eta)\mathbb{E}Z\right]\leq\exp\left(-\eta^2pn/3\right),P\left[Z<(1-\eta)\mathbb{E}Z\right]\leq\exp\left(-\eta^2pn/2\right)$$

и для специального $\eta = \frac{1}{2}$

$$P\left[Z < \mathbb{E}Z/2\right] \leq \exp\left(-pn/8\right), P\left[Z > 3\mathbb{E}Z/2\right] \leq \exp\left(-pn/8\right).$$

Теорема 3. Пусть DIS_0 является множество рассогласования множества минимизаторов эмпирического риска, построенных по первым $\lfloor n/2 \rfloor$ объектам обучающей выборки. Мы обозначим случайное множество ошибок $E_1 = \{x \in \mathcal{X} | \hat{f}(x) \neq f^*(x)\}$. Используя Лемму 4 и Лемму 3, мы получаем для любого $c > 0$

$$\mathbb{E}P(E_1) = \mathbb{E}R(\hat{f}) \leq \mathbb{E} \sup_{g \in \mathcal{G}_{f^*}} \left(Pg - (1+c)P_n g\right) \leq \frac{2\left(1+\frac{c}{2}\right)^2}{c} \frac{\log\left(\mathcal{S}_{\mathcal{F}}\left(n\right)\right)}{n}.$$

Мы фиксируем $c = 2$ и доказываем, что для любого распределения $\mathbb{E}P(E_1) \leq \frac{4\log(\mathcal{S}_{\mathcal{F}}(n))}{n}$. Заметим, что $E_1 \subseteq \mathrm{DIS}_0$. Мы используем это чтобы получить $R(\hat{f}) = P(E_1|\mathrm{DIS}_0)P(\mathrm{DIS}_0)$. Пусть $\xi = |\mathrm{DIS}_0 \cap \{X_{\lfloor n/2 \rfloor + 1}, \ldots, X_n\}|$. Условно по первым $\lfloor n/2 \rfloor$ объектам случайная величина ξ имеет биномиальное распределение. Математические ожидания для первой и второй выборок мы обозначим соответственно $\mathbb{E}$ и $\mathbb{E}'$. Условно по $\{X_1, \ldots, X_{\lfloor n/2 \rfloor}\}$ мы вводим два события

$$A_1 : \xi < \frac{nP(\mathrm{DIS}_0)}{4},$$
$$A_2 : \xi > \frac{3nP(\mathrm{DIS}_0)}{4}.$$

С помощью неравенства Чернова $P(A_j) \leq \exp\left(-\frac{nP(\mathrm{DIS}_0)}{16}\right)$, $j = 1, 2$. Обозначим $A = A_1 \cup A_2$. Тогда

$$\mathbb{E}'P(E_1|\mathrm{DIS}_0) = \mathbb{E}'\left[P(E_1|\mathrm{DIS}_0)\middle|\overline{A}\right]P(\overline{A}) + \mathbb{E}'\left[P(E_1|\mathrm{DIS}_0)\middle|A\right]P(A).$$

Для первого члена

$$\mathbb{E}'\left[P(E_1|\mathrm{DIS}_0)\middle|\overline{A}\right]P(\overline{A}) \leq \mathbb{E}'\left[P(E_1|\mathrm{DIS}_0)\middle|\overline{A}\right] \leq \frac{16\log\left(\mathcal{S}_{\mathcal{F}}\left(\frac{3nP(\mathrm{DIS}_0)}{4}\right)\right)}{nP(\mathrm{DIS}_0)}.$$

Для второго члена, умноженного на $P(\mathrm{DIS}_0)$, мы получаем

$$\mathbb{E}'\left[P(E_1|\mathrm{DIS}_0)\middle|A\right]P(\mathrm{DIS}_0)P(A) \leq 2\mathbb{E}'P(\mathrm{DIS}_0)\exp\left(-\frac{nP(\mathrm{DIS}_0)}{16}\right)$$
$$= 2P(\mathrm{DIS}_0)\exp\left(-\frac{nP(\mathrm{DIS}_0)}{16}\right) \leq \frac{12}{n}.$$

Комбинируя предыдущие результаты,

$$\mathbb{E}'P(E_1|\mathrm{DIS}_0)P(\mathrm{DIS}_0) \le \frac{16\log\left(\mathcal{S}_{\mathcal{F}}\left(\frac{3nP(\mathrm{DIS}_0)}{4}\right)\right)}{n} + \frac{12}{n}.$$

Легко видеть, что для $k, r \in \mathbb{N}$

$$(\mathcal{S}_{\mathcal{F}}(kr))^{\frac{1}{r}} \le \mathcal{S}_{\mathcal{F}}(k).$$

Мы получаем

$$\begin{aligned}
\mathbb{E}R(\hat{f}) &\le \mathbb{E}\left(\frac{16\log\left(\mathcal{S}_{\mathcal{F}}\left(\frac{3nP(\mathrm{DIS}_0)}{4}\right)\right)}{n} + \frac{12}{n}\right) \\
&\le \mathbb{E}\frac{16\log\left(\mathcal{S}_{\mathcal{F}}\left(\mathbf{s}\max\left\{1, \frac{3nP(\mathrm{DIS}_0)}{4\mathbf{s}}\right\}\right)\right)}{n} + \frac{12}{n} \\
&\le \frac{16\mathbb{E}\max\left\{1, \frac{3nP(\mathrm{DIS}_0)}{4\mathbf{s}}\right\}\log\left(\mathcal{S}_{\mathcal{F}}\left(\mathbf{s}\right)\right)}{n} + \frac{12}{n} \\
&\le \frac{16\left(1+\frac{3}{2}\right)\log\left(\mathcal{S}_{\mathcal{F}}\left(\mathbf{s}\right)\right)}{n} + \frac{12}{n} = \frac{40\log\left(\mathcal{S}_{\mathcal{F}}\left(\mathbf{s}\right)\right)}{n} + \frac{12}{n}.
\end{aligned}$$

Доказательство оценки с большой вероятностью аналогично. □

Докажем версию Утверждения 3 с большой вероятностью.

Утверждение 3. Доказательство основано на Лемме симметризации и Следствии 1. Заметим, что $\mathcal{G}_{f^*}$ является $(1,1)$-классом Бернштейна. Для фиксированных c_1, c_2, таких что $0 < c_2 < c_1$ и $t \ge \frac{(1+c_2)^2}{nc_2}$, достаточно оценить сверху

$$P\left(\sup_{g\in\mathcal{G}_{f^*}}\left(\frac{1+c'/2}{n}\sum_{i=1}^{n}\varepsilon_i g_i - c'P_n g/2\right) \ge y\right)$$

для $y > 0$. Используя то же разложение, что и в Лемме 6 и оценку Чернова [50]

для фиксированного $\lambda > 0$, $x > 0$ и $c'' = (1 + c'/2)(1 + \frac{c'}{4(1+c'/2)})$ мы получаем

$$
\begin{aligned}
&P\left(\sup_{g\in\mathcal{G}_{f^*}}\left(\frac{1+c'/2}{n}\sum_{i=1}^{n}\varepsilon_i g_i - c'P_n g/2\right) \geq x + \frac{\gamma c''}{n}\right)\\
&\leq P\left(\frac{\gamma c''}{n} + \sup_{g\in\mathcal{G}_{f^*}}\left(\frac{1+c'/2}{n}\sum_{i=1}^{n}\varepsilon_i p(g)_i - c'P_n p(g)/2\right) \geq x + \frac{\gamma c''}{n}\right)\\
&\leq \exp(-\lambda x n)\mathbb{E}\mathbb{E}_\varepsilon \exp\left(\lambda(1+c'/2)\sup_{g\in\mathcal{G}_{f^*}}\left(\sum_{i=1}^{n}\varepsilon_i p(g)_i - \frac{c'}{c'+2}p(g)_i\right)\right),
\end{aligned}
$$

где, как и в Лемме 6, оператор p обозначает ближайший элемент в γ-покрытии. Обозначая $c''' = \frac{c'}{c'+2}$ и $\lambda' = \lambda(1 + c'/2)$, мы получаем

$$
\begin{aligned}
&\mathbb{E}_\varepsilon \exp\left(\lambda' \sup_{g\in\mathcal{G}_{f^*}}\left(\sum_{i=1}^{n}\varepsilon_i p(g)_i - c'''P_n p(g)\right)\right)\\
&\leq \mathcal{M}_1^*(\mathcal{F},\gamma,n)\exp\left(\sum_{i=1}^{n}\left(\frac{(\lambda')^2}{2}p(g)_i - \lambda' c''' p(g)_i\right)\right).
\end{aligned}
$$

Обозначая $\lambda' = 2c''$, мы получаем

$$
\begin{aligned}
&P\left(\sup_{g\in\mathcal{G}_{f^*}}\left(\frac{1+c'/2}{n}\sum_{i=1}^{n}\varepsilon_i g_i - c'P_n g/2\right) \geq x + \frac{\gamma c''}{n}\right)\\
&\leq \exp\left(-\frac{4c'xn}{(2+c')^2}\right)\mathcal{M}_1^*(\mathcal{F},\gamma,n).
\end{aligned}
$$

Пусть $x = \frac{(2+c')^2}{4c'}\left(\frac{\log(\mathcal{M}_1^*(\mathcal{F},\gamma,n))}{n} + \frac{\log(\frac{4}{\delta})}{n}\right)$. Выберем $c_1 = 3$ и $c_2 = 1$. Тогда с вероятностью $1 - \delta$,

$$
\sup_{g\in\mathcal{G}_{f^*}}(P - (1+c_1)P_n)g \lesssim \frac{\gamma}{n} + \frac{\log(\mathcal{M}_1^*(\mathcal{F},\gamma,n))}{n} + \frac{\log(\frac{1}{\delta})}{n}.
$$

Завершаем доказательство, положив $\gamma = \gamma_{\frac{1}{2}}^*(n) + 1$. Верхняя оценка (2.7) легко следует из общей верхней оценки для чисел упаковки VC классов [29]. □

Переходим к доказательству Леммы о локализации 8.

Лемма 8. Для фиксированных $X_1, \ldots, X_n$ пусть $V = \{(g(X_1), \ldots, g(X_n)) : g \in \mathcal{G}\}$. Как и ранее для фиксированного γ и фиксированного минимального γ-покрытия

подмножества $\mathcal{N}_\gamma \subseteq V$, для каждого $v \in V$ обозначим как $p(v)$ ближайший к v вектор в $\mathcal{N}_\gamma$. Мы обозначим $\mathbb{E}_\xi$ как условное математическое ожидания по ξ_i при условии $X_1, \ldots, X_n$. Заметим, что

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_\xi \max_{v\in V}\left(\sum_{i=1}^n \xi_i v_i - c|v_i|\right)\\
&\leq \frac{1}{n}\mathbb{E}_\xi \max_{v\in V}\left(\sum_{i=1}^n \xi_i (v_i - |p(v)_i|)\right)\\
&+ \frac{1}{n}\mathbb{E}_\xi \max_{v\in V}\left(\sum_{i=1}^n \frac{c}{4}|p(v)_i| - c|v_i|\right)\\
&+ \frac{1}{n}\mathbb{E}_\xi \max_{v\in V}\left(\sum_{i=1}^n \xi_i p(v)_i - \frac{c}{4}|p(v)_i|\right).
\end{aligned}$$

Для первого члена имеет место неравенство $\lesssim \frac{\gamma}{n}$ из-за свойств γ-покрытия и того, что $|\xi_i| \lesssim 1$. Далее как и в доказательстве Леммы 6, второй член не больше чем $\frac{c}{4}\frac{\gamma}{n}$. Проанализируем третий член. Для множества W мы определим $W[a,b] = \{w \in W | a \leq \rho_H(w,0) < b\}$.

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_\xi \max_{v\in V}\left(\sum_{i=1}^n \xi_i p(v)_i - \frac{c}{4}|p(v)_i|\right)\\
&= \frac{1}{n}\mathbb{E}_\xi \max_{v\in \mathcal{N}_\gamma}\left(\sum_{i=1}^n \xi_i v_i - \frac{c}{4}|v_i|\right)\\
&\leq \frac{1}{n}\mathbb{E}_\xi \max_{v\in \mathcal{N}_\gamma[0,2\gamma/c]}\left(\sum_{i=1}^n \xi_i v_i - \frac{c}{4}|v_i|\right)\\
&+ \frac{1}{n}\sum_{k=1}^\infty \mathbb{E}_\xi \max_{\mathcal{N}_\gamma[2^k\gamma/c, 2^{k+1}\gamma/c]}\left(\sum_{i=1}^n \xi_i v_i - \frac{c}{4}|v_i|\right)_+
\end{aligned}$$

Первый член ограничен $\frac{2\log(\mathcal{M}_1^{\mathrm{loc}}(\mathcal{F},\gamma,n,c))}{cn}$ с помощью Леммы 3 и того, что

$$|\mathcal{N}_\gamma[0, 2\gamma/c]| \leq \mathcal{M}_1(\mathcal{B}_H(0,(2\gamma)/c, \{X_1,\ldots,X_n\}), (2\gamma)/2) \leq \mathcal{M}_1^{\mathrm{loc}}(\mathcal{F},\gamma,n,c).$$

Ограничим второй член. Начнем с произвольного слагаемого. Для $\lambda = \frac{c}{8}$ получаем

$$\begin{aligned}
&\mathbb{E}_\xi \max_{v\in\{0\}\cup\mathcal{N}_\gamma[2^k\gamma/c,2^{k+1}\gamma/c]} \left(\sum_{i=1}^n \xi_i v_i - \frac{c}{4}|v_i|\right) \\
&\le \frac{1}{\lambda}\ln \mathbb{E}_\xi \max_{v\in\{0\}\cup\mathcal{N}_\gamma[2^k\gamma/c,2^{k+1}\gamma/c]} \exp\left\{\sum_{i=1}^n \lambda\xi_i v_i - \frac{\lambda c}{4}|v_i|\right\} \\
&\le \frac{1}{\lambda}\ln\left(\sum_{v\in\mathcal{N}_\gamma[2^k\gamma/c,2^{k+1}\gamma/c]} \mathbb{E}_\xi \exp\left\{\sum_{i=1}^n \lambda\xi_i v_i - \frac{\lambda c}{4}|v_i|\right\} + 1\right) \\
&\le \frac{1}{\lambda}\ln\left(\left|\mathcal{N}_\gamma\left[2^k\gamma/c, 2^{k+1}\gamma/c\right]\right| \exp\left\{2^{k-2}\gamma(4\lambda^2 - \lambda c)/c\right\} + 1\right) \\
&\le \frac{1}{\lambda}\ln\left(\left|\mathcal{N}_\gamma\left[0, 2^{k+1}\gamma/c\right]\right| \exp\left\{2^{k-2}\gamma(4\lambda^2 - \lambda c)/c\right\} + 1\right) \\
&\le \frac{1}{\lambda}\ln\left(\left(\mathcal{M}_1^{\mathrm{loc}}(\mathcal{F}, 2\gamma, n, c)\right)^{2^{k+1}} \exp\left\{2^{k-2}\gamma(4\lambda^2 - \lambda c)/c\right\} + 1\right).
\end{aligned}$$

Здесь мы используем, что минимальное покрытие также является упаковкой и $\left|\mathcal{M}_\gamma\left[0, 2^{k+1}\gamma/c\right]\right| \le \left|\mathcal{M}_1^{\mathrm{loc}}(\mathcal{F}, 2\gamma, n, c)\right|^{2^{k+1}}$, где $\mathcal{M}_\gamma$ есть γ-упаковка (пользуясь Леммой 2.2 в [14] и монотонностью локальной энтропии 2.8 по отношению к радиусу). Фиксируем $\gamma = K\gamma_{c,c}^{\mathrm{loc}}(n)$ для некоторого $K > 2$. Заметим, что локальная энтропия не возрастает и $K\gamma_{c,c}^{\mathrm{loc}}(n) > 2\gamma_{c,c}^{\mathrm{loc}}(n) \ge \gamma_{c,c}^{\mathrm{loc}}(n) + 1$. Таким образом,

$$\begin{aligned}
&\frac{1}{\lambda}\ln\left(\exp\left\{2^{k+1}\log\left(\mathcal{M}_1^{\mathrm{loc}}(V, 2K\gamma_{c,c}^{\mathrm{loc}}(n), n, c)\right) + 2^{k-2}K\gamma_{c,c}^{\mathrm{loc}}(n)(4\lambda^2 - \lambda c)/c\right\} + 1\right) \\
&\le \frac{1}{\lambda}\ln\left(\exp\left\{2^{k+1}c(\gamma_{c,c}^{\mathrm{loc}}(n) + 1) + 2^{k-2}K\gamma_{c,c}^{\mathrm{loc}}(n)(4\lambda^2 - \lambda c)/c\right\} + 1\right).
\end{aligned}$$

Получаем, что

$$\begin{aligned}
&\sum_{k=1}^\infty \frac{8}{c}\ln\left(\exp\left(2^{k+1}\log\left(\mathcal{M}_1^{\mathrm{loc}}(\mathcal{G}, 2K\gamma_{c,c}^{\mathrm{loc}}(n), n)\right)\right)\exp\left(-2^{k-6}Kc\gamma_{c,c}^{\mathrm{loc}}(n)\right) + 1\right) \\
&\le \sum_{k=1}^\infty \frac{8}{c}\ln\left(\exp\left(2^{k+2}c\gamma_{c,c}^{\mathrm{loc}}(n) - 2^{k-6}Kc\gamma_{c,c}^{\mathrm{loc}}(n)\right) + 1\right).
\end{aligned}$$

Задаем $K = 2^9$ и получаем $\sum\limits_{k=1}^\infty \ln\left(\exp\left(2^{k+2}c\gamma_{c,c}^{\mathrm{loc}}(n) - 2^{k-6}Kc\gamma_{c,c}^{\mathrm{loc}}(n)\right) + 1\right) \le C$, где $C > 0$ некоторая абсолютная константа. Здесь мы используем, что $\ln(x+1) \le x$ для $x > 0$ и $c\gamma_{c,c}^{\mathrm{loc}} \gtrsim 1$. В итоге,

$$\frac{1}{n}\mathbb{E}_\xi \max_{v\in V}\left(\sum_{i=1}^n \xi_i v_i - c|v_i|\right) \lesssim \frac{\gamma_{c,c}^{\mathrm{loc}}(n)}{n} + \frac{\log(\mathcal{M}_1^{\mathrm{loc}}(\mathcal{F}, \gamma_{c,c}^{\mathrm{loc}}(n), n, c))}{cn} + \frac{1}{cn} \lesssim \frac{\gamma_{c,c}^{\mathrm{loc}}(n)}{n}.$$

□

Теперь мы доказываем версию Теоремы 4 с большой вероятностью.

Теорема 4 с большой вероятностью. Мы дадим детальный план доказательства. Техника полностью повторяет аргументы из предыдущих результатов. Константные факторы будут обозначаться c_i для $i \in \mathbb{N}$. Идея заключается в комбинации техники, используемой для доказательства Теоремы 4 по математическому ожиданию вместе с Леммой 5. Как и ранее, пусть $\hat{f}$ является любым минимизатором эмпирического риска, а $\hat{g}$ — соответствующая функция в классе избыточных потерь $\mathcal{G}_{\mathcal{Y}}$. Легко видеть, что $R(\hat{f}) - R(f^*) = P\hat{g}$ и $P_n\hat{g} \leq 0$. Тогда для любого $c > 0$

$$R(\hat{f}) - R(f^*) \leq P\hat{g} - (1+c)P_n\hat{g} \leq \sup_{g \in \mathcal{G}_{\mathcal{Y}}} (Pg - (1+c)P_n g).$$

Теперь из-за того, что $\mathcal{G}_{\mathcal{Y}}$ является $(\frac{1}{h}, 1)$-классом Бернштейна с помощью Леммы 5 мы получаем,

$$P\left(\sup_{g \in \mathcal{G}_{\mathcal{Y}}} (P - (1+c_1)P_n)g \geq t\right) \leq 2P\left(\sup_{g \in \mathcal{G}_{\mathcal{Y}}} ((1+c_2)P_n' - (1+c_1)P_n)g \geq t/2\right),$$

при условии, что $0 < c_2 < c_1$ и $t \geq \frac{1}{nh}\frac{(1+c_2)^2}{c_2}$. Теперь мы используем те же аргументы, что и в доказательстве Утверждения 3. Для того чтобы контролировать отклонения $\sup_{g \in \mathcal{G}_{\mathcal{Y}}} (P_n' - (1+c_3)P_n)g$ достаточно контролировать отклонения

$$P_\varepsilon\left(\sup_{f' \in \mathcal{F}^*}\left(\sum_{i=1}^n \varepsilon_i f'(X_i) - \frac{1}{2}hc_4|f'(X_i)|\right) \geq \frac{x}{2}\right)$$
$$+ P_\xi\left(\sup_{g' \in \mathcal{G}_{f^*}}\left(\sum_{i=1}^n \xi_i g'(X_i) - \frac{1}{3}hg'(X_i)\right) \geq \frac{x}{3c_4}\right).$$

Оба слагаемых анализируются аналогично. Мы рассмотрим второе слагаемое. Зафиксируем $\gamma \in \mathbb{N}$ и используем разложение как в начале доказательства Леммы

8. Теперь проблема сводится к анализу γ-покрытий

$$\sup_{g' \in \mathcal{G}_{f^*}} \left(\sum_{i=1}^{n} \xi_i g'(X_i) - \frac{h}{3} g'(X_i) \right)$$
$$\leq \gamma \left(1 + \frac{h}{12} \right) + \sup_{g' \in \mathcal{G}_{f^*}} \left(\sum_{i=1}^{n} \xi_i p(g'(X_i)) - \frac{h}{12} p(g'(X_i)) \right).$$

Первый член не является случайным. Концентрация последнего члена дается комбинацией оценки Чернова (как в Утверждении 3) и верхней оценки на экспоненциальные моменты

$$\sup_{g' \in \mathcal{G}_{f^*}} \left(\sum_{i=1}^{n} \xi_i p(g'(X_i)) - \frac{h}{12} p(g'(X_i)) \right)$$

из Леммы 8. □

Глава 3

Меры сложности, зависящие от распределения данных

3.1. Равномерные односторонние уклонения

Одним из самых важных понятий теории статистического обучения является понятие сложности классов. В предыдущих разделах мы рассматривали меры сложности, которые не зависят от распределений. В первоначальных работах Вапника и Червоненкиса [26] рассматривалась идея построения теории статистического обучения, не зависящая от распределения. В более поздних работах стало понятно, что можно рассматривать специальные условия на распределение $Y|X$ [5, 6]. Однако вместе с тем не используются конкретные преимущества для специальных распределений X. Некоторые последние работы [30, 51] показывают, что иногда, зная свойства распределения случайной величины X, возможно получать существенные улучшения в различных статистических постановках. Наш подход не будет основан на неравенствах симметризации, которые обычно используются в литературе для получения не зависящих от распределения верхних оценок. Введем некоторые дополнительные обозначения.

Теперь будем полагать, что $\mathcal{Y} \subseteq \mathbb{R}$. Введем функцию потерь $\ell : \mathbb{R}^2 \to \mathbb{R}_+$, которая измеряет штраф за предсказание $\hat{Y}$ вместо Y. Далее предположим, что для всех $y \in \mathbb{R}$ выполнено $\ell(y, y) = 0$. Риск f обозначим $R(f) = \mathbb{E}\ell(f(X), Y)$. Функция $f^* \in \mathcal{F}$ будет обозначать минимизатор $R(f)$. Минимизатор эмпирического риска минимизирует величину $R_n(f) = P_n\ell(f(X), Y)$ среди всех $f \in \mathcal{F}$. Для класса $\mathcal{G} \subseteq L_p(P)$ и $f, g \in \mathcal{G}$ обозначим $\|f - g\|_{L_p(P)} = \left(\int\limits_{\mathcal{Z}} |f(z) - g(z)|^p dP(z)\right)^{\frac{1}{p}}$ для $p > 0$. В частности, если $p = 1$, то $\|f - g\|_{L_1(P)} = \mathbb{E}|f(X) - g(X)|$, $\|f - g\|_{L_1(P_n)} = \frac{1}{n}\sum\limits_{i=1}^{n} |f(Z_i) - g(Z_i)|$, а если $p = \infty$, то мы получаем стандартную $\| \ \|_{L_\infty} = \| \ \|_\infty$ норму.

В специальном случае, когда $\mathcal{Y} = \{1, -1\}$, мы будем рассматривать бинарную функцию потерь $\ell(Y, \hat{Y}) = \mathbb{1}[Y \neq \hat{Y}]$.

В этом разделе мы получим простые оценки, выраженные с помощью $L_1(P)$ энтропий. Рассмотрим *класс избыточных потерь* $\mathcal{L}_\mathcal{Y} = \{(x, y) \to \ell(f(x), y) - \ell(f^*(x), y) \text{ for } f \in \mathcal{F}\}$ и *класс потерь* $\mathcal{G}_\mathcal{Y} = \{(x, y) \to \ell(f(x), y) \text{ for } f \in \mathcal{F}\}$. Напомним, что класс $\mathcal{G}$ является (β, B) классом Бернштейна, если для всех $g \in \mathcal{G}$

$$Pg^2 \leq B(Pg)^\beta$$

для некоторого $\beta \in [0, 1]$ и $B > 1$. Это условие естественным образом обобщает условия шума Массара [5] и Цыбакова [6]. Оно выполнено в условиях выпуклости в регрессионных задачах с квадратичной потерей и выпуклым классом $\mathcal{F}$ [52, 53]. Также мы рассмотрим другое условие, выполняемые в некоторых случаях:

$$P|g| \leq B(Pg)^\beta \tag{3.1}$$

для некоторого $\beta \in [0, 1]$ и $B > 1$. Оно эквивалентно, например, условию Бернштейна в случае бинарной функции потерь. Более того оно также выполнено для классов потерь, состоящих из неотрицательных ограниченных функций. Оба условия можно соотнести с помощью неравенства $P|g| \leq (Pg^2)^{\frac{1}{2}} \leq B^{\frac{1}{2}}(Pg)^{\frac{\beta}{2}}$.

При условии, что нам дан класс $\mathcal{G} \subset L_1(P)$, определим числа покрытия $\mathcal{N}(\mathcal{G}, \varepsilon)$ класса $\mathcal{G}$ как минимальное число функций $g_1, \ldots, g_N \in \mathcal{G}$, таких что для любого $g \in \mathcal{G}$ существует $j \in \{1, \ldots, N\}$, такое что $\|g - g_j\|_{L_1(P)} \leq \varepsilon$. Пусть $\mathcal{B}_{L_1}(g, \varepsilon)$ есть шар по норме $L_1(P)$ радиуса ε с центром g. Для $\beta \in [0, 1]$ и $B > 1$ введем

$$\mathcal{D}^{\text{loc}}(\mathcal{G}, \varepsilon, B, \beta) = \sup_{\gamma \geq \varepsilon} \sup_{g \in \mathcal{G}} \log \left(\mathcal{N}(\mathcal{G} \cap \mathcal{B}_{L_1}(g, 2B\gamma^\beta), \gamma) \right), \tag{3.2}$$

которое будем называть локальной энтропией. Другой важной для нас величиной будет локальная скобочная энтропия. Пусть $f_1, f_2 \in L_1(P)$ и $f_1 \leq f_2$ с вероятностью единица. Если $\|f_1 - f_2\|_{L_1(P)} \leq \varepsilon$, то *$\varepsilon$-скобка* состоит из всех функций f, таких что $f \in L_1(P)$ и $f_1 \leq f \leq f_2$. Для класса $\mathcal{G} \subset L_1(P)$ определим скобочную энтропию $\mathcal{N}_{[\,]}(\mathcal{G}, \varepsilon)$ как минимальное число ε-скобок $B_1, \ldots, B_N$, таких что

$\mathcal{G} \subseteq \cup_{i=1}^{N} B_i$. Аналогично определим *локальную скобочную энтропию*:

$$\mathcal{D}^{\mathrm{loc}}_{[\,]}(\mathcal{G}, \varepsilon, B, \beta) = \sup_{\gamma \geq \varepsilon} \sup_{g \in \mathcal{G}} \log \left(\mathcal{N}_{[\,]}(\mathcal{G} \cap \mathcal{B}_{L_1}(g, 2B\gamma^{\beta}), \gamma) \right), \tag{3.3}$$

Введем также следующие неподвижные точки

$$\gamma(\mathcal{G}, k, \beta, B) = \inf\{\varepsilon > 0 : k\mathcal{D}^{\mathrm{loc}}\left(\mathcal{G}, \varepsilon^{\frac{1}{2-\beta}}, B, \beta\right) \leq \varepsilon\} \tag{3.4}$$

и

$$\gamma_{[\,]}(\mathcal{G}, k, \beta, B) = \inf\{\varepsilon > 0 : k\mathcal{D}^{\mathrm{loc}}_{[\,]}\left(\mathcal{G}, \varepsilon^{\frac{1}{2-\beta}}, B, \beta\right) \leq \varepsilon\} \tag{3.5}$$

Для упрощения обозначений иногда мы не будем писать β или B как аргумент в $\gamma_{[\,]}(\ )$ и $\mathcal{D}^{\mathrm{loc}}(\ )$. Следующая простая теорема дает оценки для минимизатора эмпирического риска при условии (3.1) для класса избыточных потерь.

Далее мы предполагаем, что функция потерь ограничена 1. Наши результаты легко обобщаются на неограниченные потери, так как единственным используемым неравенством концентрации будет неравенство Бернштейна, версии которого известны для неограниченных случайных величин [50, 54, 55].

Теорема 7. *Пусть функция потерь ограничена единицей, а для класса избыточных потерь $\mathcal{L}_{\mathcal{Y}}$ выполнено условие $P|g| \leq B(Pg)^{\beta}$. Тогда с вероятностью не меньше $1 - \delta$ для минимизатора эмпирического риска $\hat{f}$ выполнено*

$$R(\hat{f}) - R(f^*) \lesssim \left(\gamma_{[\,]}\left(\mathcal{G}_{\mathcal{Y}}, \frac{B}{n}, \beta, B\right) + \frac{B\log(\frac{1}{\delta})}{n} \right)^{\frac{1}{2-\beta}}.$$

Заметим, что оценка не требует условий выпуклости или звездности, которые требуются в похожих результатах в литературе [4, 52]. Однако оценка имеет некоторые ограничения. В частности, условия на скобочную энтропию в общем случае достаточно сложно проверить.

Далее нам нужны некоторые вспомогательные результаты. Версии следующей леммы известны для нескобочных энтропий (см. Лемму 2 в [47] или Лемму 2.2 в [14]). Обозначим $\mathcal{N}(\rho, \varepsilon) = \sup_{g \in \mathcal{G}} \mathcal{N}(\mathcal{G} \cap \mathcal{B}_{L_1}(g, \rho), \varepsilon)$ и $\mathcal{N}_{[\,]}(\delta, \varepsilon) = \sup_{g \in \mathcal{G}} \mathcal{N}_{[\,]}(\mathcal{G} \cap \mathcal{B}_{L_1}(g, \rho), \varepsilon)$.

Лемма 10. *Для всех* $B > 1, \beta \in [0,1], \varepsilon \in [0,1]$ *и* $\delta > 1$

$$\log(\mathcal{N}_{[\,]}(2\delta B\varepsilon^{\beta}, \varepsilon)) \leq \log_4(16\delta)\, \mathcal{D}^{loc}_{[\,]}(\mathcal{G}, \varepsilon, \beta, B)$$

и

$$\log(\mathcal{N}(2\delta B\varepsilon^{\beta}, \varepsilon)) \leq \log_2(4\delta)\mathcal{D}^{loc}(\mathcal{G}, \varepsilon, \beta, B).$$

Следующая лемма дает верхнюю оценку на $\sup_{g\in\mathcal{G}}(Pg - (1+c)P_n g)$ в терминах локальной скобочной энтропии.

Лемма 11. *Пусть* $\mathcal{G} \subset L_1(P)$ *— класс функций, таких что* $0 \in \mathcal{G}$, $Pg \geq 0$ *и* $\|g\|_\infty \leq 1$ *для всех* $g \in \mathcal{G}$, *а также выполнено* $P|g| \leq B(Pg)^{\beta}$ *для некоторых констант* $B \geq 1$ *и* $\beta \in [0,1]$. *Тогда для любого* $c \geq 1$ *с вероятностью не меньше* $1-\delta$ *выполнено*

$$\sup_{g\in\mathcal{G}}(Pg - (1+c)P_n g) \lesssim \left(\gamma_{[\,]}\left(\mathcal{G}, \frac{c'B}{n}, \beta, B\right) + \frac{c'B\log(\frac{1}{\delta})}{n}\right)^{\frac{1}{2-\beta}},$$

где $c' = 64(1+c)^2$.

Доказательство. Зафиксируем $\varepsilon > 0$. Для класса $\mathcal{G}$ и распределения P построим $\varepsilon^{\frac{1}{2-\beta}}$-скобочное покрытие класса (по $L_1(P)$ метрике). Пусть p обозначает проекцию на наименьшую функцию в скобке, а $p[\mathcal{G}]$ обозначает множество всех проекций, состоящее из функций вида $\{p[g] | g \in \mathcal{G}\}$. Далее предполагаем, что $0 \in p[\mathcal{G}]$. Тогда, так как $p[g] \leq g$ с вероятностью 1 мы получаем

$$\sup_{g\in\mathcal{G}}(Pg - (1+c)P_n g) \leq \sup_{g\in\mathcal{G}}(Pg - Pp[g] + Pp[g] - (1+c)P_n p[g]) \tag{3.6}$$

$$\leq \varepsilon^{\frac{1}{2-\beta}} + \sup_{g\in\mathcal{G}}(Pp[g] - (1+c)P_n p[g]), \tag{3.7}$$

Обозначим $\mathcal{G}_0 = p[\mathcal{G}] \cap \mathcal{B}_{L_1}(0, 2B\varepsilon^{\frac{\beta}{2-\beta}})$ и $\mathcal{G}_1 = \{0\} \cup (p[\mathcal{G}] \setminus \mathcal{B}_{L_1}(0, 2B\varepsilon^{\frac{\beta}{2-\beta}}))$. Очевидно, что $\mathcal{G}_0 \cup \mathcal{G}_1 = p[\mathcal{G}]$. Перепишем последнее слагаемое как

$$\sup_{g\in\mathcal{G}}(Pp[g] - (1+c)P_n p[g]) \leq \sup_{g\in\mathcal{G}_0}(Pg - (1+c)P_n g) + \sup_{g\in\mathcal{G}_1}(Pg - (1+c)P_n g).$$

Шаг 1. Начнем с $\sup_{g\in\mathcal{G}_1}(Pg-(1+c)P_ng)$. Мы оценим следующую величину

$$
\begin{aligned}
&P(\exists g\in\mathcal{G}_1 : P_ng < \frac{1}{1+c}Pg)\\
&\le \sum_{j=1}^{\infty} P(\exists g\in p[\mathcal{G}] : P|g| \in [2^jB\varepsilon^{\frac{\beta}{2-\beta}}, 2^{j+1}B\varepsilon^{\frac{\beta}{2-\beta}}] \ \cap \ P_ng < \frac{1}{1+c}Pg)
\end{aligned}
$$

Для функции $g\in p[\mathcal{G}]$, для которой $P|g|\in[2^jB\varepsilon^{\frac{\beta}{2-\beta}}, 2^{j+1}B\varepsilon^{\frac{\beta}{2-\beta}}]$, рассмотрим

$$
P(Pg-(1+c)P_ng>0) = P\left(Pg-P_ng > \frac{cPg}{1+c}\right).
$$

Используя неравенство Бернштейна [50], мы получаем (с учетом $Pg>0$)

$$
\begin{aligned}
P\left(Pg-P_ng>\frac{cPg}{1+c}\right) &\le \exp\left(-\frac{nc^2(Pg)^2}{(1+c)^2(2Pg^2+\frac{2cPg}{3(1+c)})}\right)\\
&\le \exp\left(-\frac{nc^2}{4(1+c)}\left(\frac{(Pg)^2}{(1+c)Pg^2}\wedge\frac{3Pg}{c}\right)\right).
\end{aligned}
$$

Пусть $g'\in\mathcal{G}$ — любая функция, такая что $p[g']=g$ для $g\in p[\mathcal{G}]$, для которой $P|g|\ge 2B\varepsilon^{\frac{\beta}{2-\beta}}$. Без ограничения общности предположим, что $\|g\|_\infty\le 1$ и с вероятностью единица $|g(Z)-g'(Z)|\le 2$. Используя предположение (3.1), мы получаем

$$
\begin{aligned}
P|g| &\le P|g'| + P|g-g'| \le P|g'| + \varepsilon^{\frac{1}{2-\beta}}\\
&\le B(Pg')^\beta + \varepsilon^{\frac{1}{2-\beta}} \le 2B(Pg)^\beta.
\end{aligned}
$$

Далее (используем $(Pg)^{2-\beta}\le Pg$, $B\ge 1$ и $Pg^2\le P|g|$)

$$
\begin{aligned}
P\left(Pg-P_ng>\frac{cPg}{1+c}\right) &\le \exp\left(-\frac{nc^2}{4(1+c)}\left(\frac{(Pg)^{2-\beta}}{2B(1+c)}\wedge\frac{3Pg}{c}\right)\right)\\
&= \exp\left(-\frac{nc^2(Pg)^{2-\beta}}{8B(1+c)^2}\right)\\
&\le \exp\left(-\frac{nc^2(P|g|)^{\frac{2-\beta}{\beta}}}{8B^{\frac{2}{\beta}}(1+c)^2}\right)\\
&\le \exp\left(-\frac{nc^2 2^{j\frac{(2-\beta)}{\beta}}\varepsilon}{8B(1+c)^2}\right).
\end{aligned}
$$

Мы хотим оценить число функций $g \in p[\mathcal{G}]$, для которых $P|g| \in [2^j B\varepsilon^{\frac{\beta}{2-\beta}}, 2^{j+1}B\varepsilon^{\frac{\beta}{2-\beta}}]$. Это легко с помощью Леммы 10. Небольшой технический момент в том, что мы не можем гарантировать, что глобальное минимальное покрытие также останется минимальным при проецировании на подмножество. Однако, оно почти минимальное в том смысле, что достаточно рассматривать $\mathcal{D}^{\text{loc}}_{[\ ]}\left(\mathcal{G}, \varepsilon^{\frac{1}{2-\beta}}/2\right)$ вместо $\mathcal{D}^{\text{loc}}_{[\ ]}\left(\mathcal{G}, \varepsilon^{\frac{1}{2-\beta}}\right)$. Этот шаг стандартный и основан на соотношении между максимальными упаковками и минимальными покрытиями. Далее

$$\begin{aligned}
&\sum_{j=1}^{\infty} P(\exists g \in p[\mathcal{G}] : P|g| \in [2^j B\varepsilon^{\frac{\beta}{2-\beta}}, 2^{j+1}B\varepsilon^{\frac{\beta}{2-\beta}}] \ \cap \ P_n g < \frac{1}{1+c} Pg) \\
&\leq \sum_{j=1}^{\infty} (2^{j+5})^{\mathcal{D}^{\text{loc}}_{[\ ]}\left(\mathcal{G}, \varepsilon^{\frac{1}{2-\beta}}\right)/\log(4)} \exp\left(-\frac{nc^2 2^j \varepsilon}{8B(1+c)^2}\right) \\
&\leq \sum_{j=1}^{\infty} \exp\left(\frac{(j+5)\mathcal{D}^{\text{loc}}_{[\ ]}\left(\mathcal{G}, \varepsilon^{\frac{1}{2-\beta}}/2\right)}{\log(4)} - \frac{nc^2(j+5)\varepsilon}{24B(1+c)^2}\right).
\end{aligned}$$

При условии, что $n \geq \frac{24B(1+c)^2}{c^2\log(4)}\left(\frac{\mathcal{D}^{\text{loc}}_{[\ ]}\left(\mathcal{G}, \varepsilon^{\frac{1}{2-\beta}}/2\right)}{\varepsilon} + \frac{\log(\frac{1}{\delta})}{\varepsilon}\right)$ последний член ограничен сверху $\frac{\delta}{2}$. Таким образом, с вероятностью не меньшей $1 - \frac{\delta}{2}$ для всех $g \in \mathcal{G}_1$ выполнено $Pg - (1+c)P_n g \leq 0$. Значит при выполнении этого события $\sup_{g\in\mathcal{G}_1}(Pg - (1+c)P_n g) = 0$.

Шаг 2. Рассмотрим $\sup_{g\in\mathcal{G}_0}(Pg-(1+c)P_n g)$. Для построения верхней оценки мы используем неравенство Бернштейна и неравенство Буля , учитывая что $|\mathcal{G}_0| \leq$

$\exp\left(\mathcal{D}^{\text{loc}}_{[\ ]}\left(\mathcal{G}, \varepsilon^{\frac{1}{2-\beta}}/2\right)\right)$ и что $P|g| \le P|g'| + P|g-g'| \le P|g'| + \varepsilon^{\frac{1}{2-\beta}} \le 2B\varepsilon^{\frac{\beta}{2-\beta}}$

$$
\begin{aligned}
&P(\sup_{g\in\mathcal{G}_0}(Pg - (1+c)P_n g) \ge \varepsilon^{\frac{1}{2-\beta}}) \\
&\le \exp\left(\mathcal{D}^{\text{loc}}_{[\ ]}\left(\mathcal{G}, \varepsilon^{\frac{1}{2-\beta}}/2\right) - \frac{n}{4(1+c)}\left(\frac{(\varepsilon^{\frac{1}{2-\beta}} + cPg)^2}{(1+c)Pg^2} \wedge 3(\varepsilon^{\frac{1}{2-\beta}} + cPg)\right)\right) \\
&\le \exp\left(\mathcal{D}^{\text{loc}}_{[\ ]}\left(\mathcal{G}, \varepsilon^{\frac{1}{2-\beta}}/2\right) - \frac{n}{4(1+c)}\left(\frac{(\varepsilon^{\frac{1}{2-\beta}} + cPg)^2}{2(1+c)B\varepsilon^{\frac{\beta}{2-\beta}}} \wedge 3(\varepsilon^{\frac{1}{2-\beta}} + cPg)\right)\right) \\
&\le \exp\left(\mathcal{D}^{\text{loc}}_{[\ ]}\left(\mathcal{G}, \varepsilon^{\frac{1}{2-\beta}}/2\right) - \frac{n}{4(1+c)}\left(\frac{\varepsilon}{2(1+c)B} \wedge 3\varepsilon^{\frac{1}{2-\beta}}\right)\right) \\
&\le \exp\left(\mathcal{D}^{\text{loc}}_{[\ ]}\left(\mathcal{G}, \varepsilon^{\frac{1}{2-\beta}}/2\right) - \frac{n}{4(1+c)}\left(\frac{\varepsilon}{2(1+c)B}\right)\right).
\end{aligned}
$$

При условии $n \ge 8B(1+c)^2\left(\frac{\mathcal{D}^{\text{loc}}_{[\ ]}\left(\mathcal{G}, \varepsilon^{\frac{1}{2-\beta}}/2\right)}{\varepsilon} + \frac{\log(\frac{2}{\delta})}{\varepsilon}\right)$ мы получаем с вероятностью не меньшей $1 - \frac{\delta}{2}$, что $\sup\limits_{g\in\mathcal{G}_0}(Pg - (1+c)P_n g) \le \varepsilon^{\frac{1}{2-\beta}}$.

Шаг 3. Используя неравенство Буля для событий из предыдущих шагов, мы получаем с вероятностью не меньшей $1-\delta$, что при условии $n \ge 24B(1+c)^2\left(\frac{\mathcal{D}^{\text{loc}}_{[\ ]}\left(\mathcal{G}, \varepsilon^{\frac{1}{2-\beta}}/2\right)}{\varepsilon} + \frac{\log(\frac{2}{\delta})}{\varepsilon}\right)$ выполнено

$$\sup_{g\in\mathcal{G}}(Pg - (1+c)P_n g) \le 2\varepsilon^{\frac{1}{2-\beta}}$$

Обозначим $c' = 64(1+c)^2$. Выбирая $\gamma^*(\mathcal{G}, k, \beta, \delta) = \inf\{\varepsilon > 0 : k(\mathcal{D}^{\text{loc}}_{[\ ]}\left(\mathcal{G}, \varepsilon^{\frac{1}{2-\beta}}/2\right) + \log(\frac{2}{\delta})) \le \varepsilon\}$, получаем, что с вероятностью не меньшей $1-\delta$ для данного n выполнено $\sup\limits_{g\in\mathcal{G}}(Pg - (1+c)P_n g) \le 2(\gamma^*(\mathcal{G}, c'B/n, \beta, \delta))^{\frac{1}{2-\beta}}$. Однако, если взять $\gamma_{[\ ]}(\mathcal{G}, k, \beta) = \inf\{\varepsilon > 0 : k\mathcal{D}^{\text{loc}}_{[\ ]}\left(\mathcal{G}, \varepsilon^{\frac{1}{2-\beta}}/2\right) \le \varepsilon\}$, то легко видеть, что

$$\gamma^*(\mathcal{G}, c'B/n, \beta, \delta) \le \gamma_{[\ ]}(\mathcal{G}, c'B/n, \beta) + \frac{c'B\log(\frac{2}{\delta})}{n}.$$

□

Теорема 7. С помощью Леммы 11 доказательство следует напрямую. Для минимизатора эмпирического риска $\hat{f}$ обозначим соответствующую функцию в $\mathcal{L}_\mathcal{Y}$ как

$\hat{g}$. Получаем $P\hat{g} = R(\hat{f}) - R(f^*)$ и $P_n\hat{g} \leq 0$. Тогда для любого $c > 0$ (можно взять $c = 1$) мы получаем

$$P\hat{g} \leq P\hat{g} - (1+c)P_n\hat{g} \leq \sup_{g \in \mathcal{L}_\mathcal{Y}} (Pg - (1+c)P_n g)$$

Финальный шаг заключается в том чтобы понять, что метрические свойства $\mathcal{L}_\mathcal{Y}$ совпадают с метрическими свойствами $\mathcal{G}_\mathcal{Y}$. Это означает, что, например,

$$\gamma_{[\ ]}\left(\mathcal{L}_\mathcal{Y}, \frac{B}{n}, \beta, B\right) = \gamma_{[\ ]}\left(\mathcal{G}_\mathcal{Y}, \frac{B}{n}, \beta, B\right).$$

Лемма 11 для $\mathcal{L}_\mathcal{Y}$ завершает доказательство. □

3.2. Минимизатор риска на ε-сетях

Следующим важным вопросом является возможность оценивания локальных скобочных энтропий. В некоторых случаях это легко. Для многих непараметрических классов известные оценки на энтропии и скобочные энтропии имеют один и тот же порядок (см [5, 6, 56]). Более того, из работы Янга и Бэррона [38] следует, что для этих непараметрических классов локальные энтропии имеют тот же порядок, что и глобальные. В общем случае вопрос оценивания локальных скобочных энтропий вовсе не тривиальный. Оценки известны для некоторых классов плотностей [57]. Для общих классов Вапника-Червоненкиса известна лишь конечность скобочной энтропии [58].

Когда невозможно гарантировать, что скобочные энтропии близки к энтропиям без скобок можно использовать следующий прием, который будет использован для того чтобы избавиться от скобочных энтропий. Техника основана на так называемых оценках на сетках: это алгоритмы, которые представляют из себя минимизаторы риска на ε-сетях исходных классов. Версии этих алгоритмов часто анализируются в литературе [6, 14, 34, 38, 46, 47].

Следствие 3 (Оценка для минимизатора эмпирического риска на ε-сетях)**.** *Пусть для любого $\eta \in [0,1]$ возможно выбрать $f_1, \ldots, f_{N_\eta} \in \mathcal{F}$ такие, что функции*

$\ell(f_1(X), Y), \ldots, \ell(f_{N_\eta}(X), Y)$ *образуют минимальное* $L_1(P)$ η*-покрытие класса потерь* $\mathcal{G}_\mathcal{Y}$. *Определим* $\hat{f}_\eta = \arg \min_{f \in \{f_1, \ldots, f_{N_\eta}\}} R_n(f)$. *Если*

$$\eta \simeq \left(\gamma \left(\mathcal{G}_\mathcal{Y}, \frac{B}{n}, \beta, B \right) + \frac{B \log(\frac{1}{\delta})}{n} \right)^{\frac{1}{2-\beta}},$$

то в условиях Теоремы 7 с вероятностью не меньшей $1 - \delta$ *выполнено*

$$R(\hat{f}_\eta) - R(f^*) \lesssim \left(\gamma \left(\mathcal{G}_\mathcal{Y}, \frac{B}{n}, \beta, B \right) + \frac{B \log(\frac{1}{\delta})}{n} \right)^{\frac{1}{2-\beta}}.$$

Если кроме того для всех $g \in \mathcal{L}_\mathcal{Y}$ *выполнено*

$$P|g| \simeq B(Pg)^\beta, \tag{3.8}$$

и $\eta \simeq B \left(\gamma^* \left(\mathcal{G}_\mathcal{Y}, \frac{B}{n}, \beta, B \right) + \frac{B \log(\frac{1}{\delta})}{n} \right)^{\frac{\beta}{2-\beta}}$, *то с вероятностью не меньшей* $1 - \delta$ *выполнено*

$$R(\hat{f}_\eta) - R(f^*) \lesssim \left(\gamma^* \left(\mathcal{G}_\mathcal{Y}, \frac{B}{n}, \beta, B \right) + \frac{B \log(\frac{1}{\delta})}{n} \right)^{\frac{1}{2-\beta}}, \tag{3.9}$$

где

$$\gamma^*(\mathcal{G}, k, \beta, B) = \inf\{\varepsilon > 0 : k\mathcal{D}^{loc}\left(\mathcal{G}, B\varepsilon^{\frac{\beta}{2-\beta}}, 1, 1\right) \leq \varepsilon\}.$$

Доказательство будет приведено ниже. Условия этого следствия выполнены, например, для бинарной функции потерь, так как в этом случае

$$|\mathbb{1}[f(X) \neq Y] - \mathbb{1}[g(X) \neq Y]| = |f(X) - g(X)|/2. \tag{3.10}$$

Таким образом, если требуется покрыть класс потерь, то достаточно покрыть исходный класс $\mathcal{F}$ Следующее утверждение показывает конкретные примеры, когда выполнено условие (3.8).

Утверждение 7. *Рассмотрим бинарную классификацию с классами* $\mathcal{Y} = \{1, -1\}$ *и функцией* $f^* \in \mathcal{F}$, *определенной* $f^*(x) = sign(\eta(x))$, *где* $\eta(X) = \mathbb{E}[Y|X]$. *Тогда, если существуют* $h_1, h_2 \in [0, 1]$, *такие что* $h_1 \leq |\eta(X)| \leq h_2$, *то для любого класса бинарных классификаторов* $\mathcal{F}$ *и всех* $g \in \mathcal{L}_\mathcal{Y}$ *выполнено*

$$h_1 P|g| \leq Pg \leq h_2 P|g|.$$

Иначе, если для некоторых констант $c_1, c_2 > 0$, $\beta \in [0,1)$ *и любого* $t \in [0,1]$ *выполнено* $P(|\eta(X)| \leq t) \leq c_1 t^{\frac{\beta}{1-\beta}}$ *и для* $t_0 = c_2 \inf_{f \in \mathcal{F},\ f \neq f^*} P(f(X) \neq f^*(X))^{\frac{1-\beta}{\beta}}$ *выполнено* $P(|\eta(X)| \geq t_0) \leq c_3 t_0^{\frac{1}{1-\beta}}$, *то для всех* $g \in \mathcal{L}_{\mathcal{Y}}$ *выполнено* $P|g| \simeq (Pg)^{\beta}$, *где константные факторы зависят только от* c_1, c_2, c_3, β.

Доказательство. Для первой части нам нужен стандартный результат, что (см. [23]) для любой функции $f \in \mathcal{F}$ выполнено $R(f) - R(f^*) = \mathbb{E}(|\eta(X)| \mathbb{1}[f(X) \neq f^*])$. Тогда, если $h_1 \leq |\eta(X)| \leq h_2$, то $h_1 P|g| \leq Pg \leq h_2 P|g|$ для всех $g \in \mathcal{L}_{\mathcal{Y}}$. Для второй части мы используем тот факт, что $P(|\eta(X)| \leq t) \leq c_1 t^{\frac{\beta}{1-\beta}}$ влечет $P|g| \lesssim (Pg)^{\beta}$ (Утвержение 1 в [6]). Тогда для всех $t \geq t_0$

$$\begin{aligned} R(f) - R(f^*) &= \mathbb{E}(|\eta(X)| \mathbb{1}[f(X) \neq f^*(X)]) \\ &\leq tP(f(X) \neq f^*(X)) + \mathbb{E}(|\eta(X)| \mathbb{1}(|\eta(X)| > t)) \\ &\leq tP(f(X) \neq f^*(X)) + P(|\eta(X)| > t) \\ &\leq tP(f(X) \neq f^*(X)) + c_3 t^{\frac{1}{1-\beta}}. \end{aligned}$$

Выбирая $t = c_2 P(f(X) \neq f^*(X))^{\frac{1-\beta}{\beta}}$, мы доказываем утверждение. □

Первая часть утверждения есть специальный случай условий малого шума Массара [5], а вторая часть является специальным случаем более общего условия Цыбакова [6]. Стоит отметить, что в обоих случаях $|\eta(X)|$ отделена от 1 с большой вероятностью, что соответствует задачам с более высоким уровнем шума. Действительно, бесшумная классификация относится к случаю, когда $|\eta(X)| = 1$ с вероятностью единица.

3.3. Примеры оценок

Глобальные условия на энтропии

Классические порядки для глобальных скобочных $L_1(P)$ энтропий, полученные Цыбаковым при условии бинарных потерь (см. Теорему 1 в [6]), практически достигаются нашей верхней оценкой. При условии $\log(\mathcal{N}_{[\,]}(\mathcal{G}_{\mathcal{Y}}, \varepsilon)) \simeq \varepsilon^{-r}$ для

$\varepsilon \in [0,1)$ и $r > 0$ мы получаем при (β, B) условии Бернштейна на класс (что эквивалентно (3.1) в нашем случае) оценку порядка

$$\left(\left(\frac{1}{n}\right)^{\frac{2-\beta}{2-\beta+r}} + \frac{\log(\frac{1}{\delta})}{n}\right)^{\frac{1}{2-\beta}}, \tag{3.11}$$

Важно, что результат дается для минимизатора риска, а не для минимизатора риска на сети, как в работе [6]. Также, при условии (3.8) и $\log(\mathcal{N}(\mathcal{G}_{\mathcal{Y}}, \varepsilon)) \simeq \varepsilon^{-r}$ оценка (3.9) из Следствия 3 дает точно такой же порядок $\left(\left(\frac{1}{n}\right)^{\frac{2-\beta}{2-\beta+\beta r}} + \frac{\log(\frac{1}{\delta})}{n}\right)^{\frac{1}{2-\beta}}$, как и в работе [6], который ранее получался при условии $\log(\mathcal{N}_{[\,]}(\mathcal{G}_{\mathcal{Y}}, \varepsilon)) \simeq \varepsilon^{-r}$

Однородные линейные разделители и изотропные лог-вогнутые распределения

Этому примеру посвящены несколько работ (см. [30, 47, 51, 59]) и он является одной из наших мотиваций для рассмотрения мер сложности, зависящих от распределений. Из недавнего результата (см. 5.1 в [30]) следует, что для класса $\mathcal{F}$ однородных линейных разделителей (проходящих через начало координат) в $\mathbb{R}^d$ для изотропных лог-вогнутых распределений X выполнено $\mathcal{D}^{\text{loc}}(\mathcal{G}_{\mathcal{Y}}, \varepsilon, 1, 1) \lesssim d$ для бинарной функции потерь. Таким обрзом, используя минимизатор риска по сети при условии (3.8) мы получаем порядок $R(\hat{f}_\eta) - R(f^*) \lesssim \left(\frac{Bd}{n} + \frac{B\log(\frac{1}{\delta})}{n}\right)^{\frac{1}{2-\beta}}$. получен в работе [47] только для бесшумного случая (в частности, в этом случае $\beta = B = 1$ и условие (3.8) выполнено) и этот порядок строго лучше, чем порядок, полученный [1] в Теореме 19 в [30]. В случае $\beta = 1$ наша оценка имеет порядок $R(\hat{f}_\eta) - R(f^*) \lesssim \frac{Bd}{n} + \frac{B\log(\frac{1}{\delta})}{n}$ и мы докажем зависящую от B совпадающую нижнюю оценку, тем самым показывая, что член $\frac{Bd}{n}$ является неизбежным.

Утверждение 8 (Нижняя оценка для лог-вогнутых распределений)**.** *Рассмотрим задачу обучения линейных однородных разделителей $\mathcal{F}$ в $\mathbb{R}^d$ с бинарной функцией потерь. Пусть $\tilde{f}$ получен произвольным обучающим алгоритмом. Тогда*

[1] Стоит отметить, что предыдущая оценка получена другим способом, с помощью алгоритма, основанного на идеях активного обучения.

для любого $B \le \sqrt{\frac{n}{d}}$ *существует распределение* $P_{X,Y}$*, такое что класс избыточных потерь* $\mathcal{L}_{\mathcal{Y}}$ *является* $(1, B)$*-классом Бернштейна,* P_X *является изотропным лог-вогнутым распределением и* $\mathbb{E}(R(\tilde{f}) - R(f^*)) \gtrsim \frac{Bd}{n}$.

Доказательство мы приведем ниже.

Неточные оракульные неравенства в теории агрегации

Неточными оракульными неравенствами называются верхние оценки на величину $R(\hat{f}) - (1 + a)R(f^*)$ для некоторого $a > 0$. Известно, что член $1 + a$, стоящий вместо 1, позволяет при слабых ограничениях давать такие же порядки обучения, как если бы условие Бернштейна было выполнено $\mathcal{L}_{\mathcal{Y}}$. Как отмечено в работе [53] оценки на смещенные процессы достаточны для доказательства подобных результатов. Можно показать, что (см. [53]) для любого минимизатора эмпирического риска $\hat{f}$ выполнено $R(\hat{f}) - (1+2c)R(f^*) \le \sup_{g \in \mathcal{G}_{\mathcal{Y}}} (Pg - (1 + c)P_n g) + (1 + c) \sup_{g \in \mathcal{G}_{\mathcal{Y}}} \left(P_n g - \frac{1+2c}{1+c} Pg\right)$, где как и ранее $\mathcal{G}_{\mathcal{Y}}$ обозначает класс потерь. Однако наш способ оценивания этого процесса будет использовать Лемму 11 и ее простое обобщение для $\sup_{g \in \mathcal{G}_{\mathcal{Y}}} \left(P_n g - \frac{1+2c}{1+c} Pg\right)$. Важно, что здесь условие (3.1) требуется не для класса избыточных потерь, а для класса потерь, которое в случае ограниченных потерь выполнено очевидным образом. Таким образом, для любой функции потерь, ограниченной 1 выполнено с вероятностью не менее $1 - \delta$ для всех минимизаторов эмпирического риска $\hat{f}$:

$$R(\hat{f}) - 2R(f^*) \lesssim \gamma_{[\,]}\left(\mathcal{G}_{\mathcal{Y}} \cup \{0\}, \frac{1}{n}, 1\right) + \frac{\log(\frac{1}{\delta})}{n}. \tag{3.12}$$

Важным случаем является случай малого $R(f^*)$, имеющего тот же порядок величины, что и правая часть неравенства. В этом случае мы получаем те же гарантии для избыточного риска, как если бы для $\mathcal{L}_{\mathcal{Y}}$ было выполнено условие (3.1) с параметрами $B = 1, \beta = 1$. Ранее в литературе для получения подобных порядков использовались специальные агрегационные процедуры (см., например, Теорему 5 в [46]).

Локальные $L_2(P)$ энтропии в регрессионных моделях

До сих пор мы рассматривали $L_1(P)$ энтропии. Однако, для непараметрических классов и квадратичных потерь анализ обычно производится с помощью $L_2(P)$ расстояний. Рассмотрим ограниченную регрессионную модель с квадратичной функцией потерь с дополнительным условием, что функция $f^*(X) = \mathbb{E}[Y|X]$ лежит в классе, но без дополнительных ограничений на выпуклость класса $\mathcal{F}$. Например, это включает в себя модель $Y = f^*(X) + \xi$, гдеξ независимый ограниченный шум с нулевым средним и $f^* \in \mathcal{F}$. Этот случай интересен тем, что в нем существует полезное соотношение между избыточным риском и расстоянием между функциями, а именно $R(f) - R(f^*) = \|f - f^*\|_{L_2}$. Это условие похоже на то, что влечет условие (3.8).

Определим

$$\mathcal{D}_{L_2}^{\text{loc}}(\mathcal{G}, \varepsilon) = \sup_{\gamma \geq \varepsilon} \sup_{g \in \mathcal{G}} \log \left(\mathcal{N}_{L_2}(\mathcal{G} \cap \mathcal{B}_{L_2}(g, 2\gamma), \gamma) \right),$$

где $\mathcal{N}_{L_2}$ обозначает число покрытия с помощью расстояния $L_2(P)$. Далее

$$\zeta(\mathcal{G}, k) = \inf\{\varepsilon > 0 : k\mathcal{D}_{L_2}^{\text{loc}}(\mathcal{G}, \varepsilon) \leq \varepsilon^2\}$$

Следствие 4. *Рассмотрим вышеописанную регрессионную модель. Для $\eta \in [0, 1]$ выбираем $f_1, \ldots, f_{N_\eta} \in \mathcal{F}$, такие что они образуют минимальное η-покрытие класса $\mathcal{F}$ по расстоянию $L_2(P)$. Определим $\hat{f}_\eta = \arg \min_{f \in \{f_1, \ldots, f_{N_\eta}\}} R_n(f)$. Тогда, если $\eta \simeq \zeta\left(\mathcal{F}, \frac{1}{n}\right) + \sqrt{\frac{\log(\frac{1}{\delta})}{n}}$, то с вероятностью не менее $1 - \delta$ выполнено*

$$R(\hat{f}_\eta) - R(f^*) \lesssim \left(\zeta\left(\mathcal{F}, \frac{1}{n}\right)\right)^2 + \frac{\log(\frac{1}{\delta})}{n}.$$

Ранее похожие порядки были получены с помощью глобальных эмпирических энтропий в работе [46], используя так называемую *skeleton aggregation* или процедуру *aggregation of leaders*. Однако в нашем случае используется более простая процедура, принимающая значения в исходном классе. Более того наша мера сложности локализована. Похожая $L_2(P)$ локальная энтропия (см. Теорему

4.5 в [14]) появляется в общих нижних оценках в неограниченном случае для выпуклых классов $\mathcal{F}$. Наш результат выполнен даже для очень больших непараметрических классов, таких, для которых $\log(\mathcal{N}_{L_2}(G,\varepsilon)) \simeq \varepsilon^{-r}$, для $r > 2$ и не требует выпуклости или звездности классов $\mathcal{F}$ или $\mathcal{L}_{\mathcal{Y}}$.

3.4. Доказательства

Лемма 10. Обозначим для $\delta > 4$

$$\mathcal{N}(2\delta B\gamma^{\beta},\gamma) = \sup_{g\in\mathcal{G}} \mathcal{N}(\mathcal{G}\cap B_P(g, 2\delta B\gamma^{\beta}),\gamma)$$

и

$$\mathcal{N}_{[\,]}(2\delta B\gamma^{\beta},\gamma) = \sup_{g\in\mathcal{G}} \mathcal{N}_{[\,]}(\mathcal{G}\cap B_P(g, 2\delta B\gamma^{\beta}),\gamma)$$

Пусть $t_1,\dots,t_N$ являются центрами минимального покрытия $2\delta B\gamma^{\beta}$-шара, пересеченного с $\mathcal{G}$ с помощью L_1 шаров с радиусом $\delta B\gamma^{\beta}/2$. Общее их число ограничено $\mathcal{N}(2\delta B\gamma^{\beta}, \delta B\gamma^{\beta}/2)$. Теперь для фиксированного i мы хотим покрыть множество $\mathcal{B}_{L_1}(t_i, \delta B\gamma^{\beta}/2)\cap\mathcal{G}$ с помощью γ-скобок. Очевидно, так как $t_i \in \mathcal{G}$ для всех i минимальное число скобок ограничено $\mathcal{N}_{[\,]}(\delta B\gamma^{\beta}/2,\gamma)$. В итоге

$$\mathcal{N}_{[\,]}(2\delta B\varepsilon^{\beta},\varepsilon) \le \mathcal{N}(2\delta B\varepsilon^{\beta}, \delta B\varepsilon^{\beta}/2)\mathcal{N}_{[\,]}(\delta B\varepsilon^{\beta}/2,\gamma).$$

Используя стандартную оценку (см. [56]), мы получаем $\mathcal{N}(\rho,\rho/4) \le \mathcal{N}_{[\,]}(\rho,\rho/2)$. Используя определение локальной скобочной энтропии, мы запишем

$$\mathcal{N}_{[\,]}(2\delta B\gamma^{\beta},\gamma) \le \exp(\mathcal{D}^{\mathrm{loc}}_{[\,]}(\mathcal{G},\gamma,B,\beta))\mathcal{N}_{[\,]}(\delta B\gamma^{\beta}/2,\gamma). \tag{3.13}$$

Мы продолжим анализ члена $\mathcal{N}_{[\,]}(\delta B\gamma^{\beta}/2,\gamma)$ таким же образом. Если $\delta/16 > 1$, то мы используем то же представление 3.13. Иначе, если $\delta/16 \le 1$

$$\begin{aligned}\mathcal{N}_{[\,]}(\delta B\gamma^{\beta}/2,\gamma) &\le \mathcal{N}_{[\,]}(8B\gamma^{\beta},\gamma) \le \mathcal{N}(8B\gamma^{\beta}, 2B\gamma^{\beta})\mathcal{N}_{[\,]}(2B\gamma^{\beta},\gamma)\\ &\le \mathcal{N}_{[\,]}(8B\gamma^{\beta}, 4B\gamma^{\beta})\exp(\mathcal{D}^{\mathrm{loc}}_{[\,]}(\mathcal{G},\gamma)) \le \exp(2\mathcal{D}^{\mathrm{loc}}_{[\,]}(\mathcal{G},\gamma)).\end{aligned}$$

Продолжая аналогично, мы получаем

$$\mathcal{N}_{[\,]}(\delta,\gamma) \leq \exp(\mathcal{D}^{\text{loc}}_{[\,]}(\mathcal{G},\gamma))^{(\log_4(\delta)+2)}.$$

□

Следствие 3. Определим $f^*_\eta = \arg\min_{f\in\{f_1,\ldots,f_{N_\eta}\}} R(f)$. Так как $R_n(\hat{f}_\eta) - R_n(f^*_\eta) \leq 0$ для $c \geq 1$

$$\begin{aligned}
&R(\hat{f}_\eta) - R(f^*) \\
&\leq R(\hat{f}_\eta) - R(f^*) - (1+c)(R_n(\hat{f}_\eta) - R_n(f^*_\eta)) \\
&= R(\hat{f}_\eta) - R(f^*) - (1+c)(R_n(\hat{f}_\eta) - R_n(f^*)) + (1+c)(R_n(f^*_\eta) - R_n(f^*)) \\
&\leq \sup_{f\in\{f_1,\ldots,f_{N_\eta}\}} (R(f) - R(f^*) - (1+c)(R_n(f) - R_n(f^*))) + (1+c)(R_n(f^*_\eta) - R_n(f^*)) \\
&= \sup_{g\in\{g_1,\ldots,g_{N_\eta}\}} (Pg - (1+c)P_n g) + (1+c)(R_n(f^*_\eta) - R_n(f^*)),
\end{aligned}$$

где $g_1,\ldots,g_{N_\eta} \in \mathcal{L}_\mathcal{Y}$ соответствуют $f_1,\ldots,f_{N_\eta}$. Проанализируем второй член отдельно. Используя неравенство Бернштейна и тот факт, что $0 \leq R(f^*_\eta) - R(f^*) \leq \eta \leq 1$, получаем

$$\begin{aligned}
&P(R_n(f^*_\eta) - R_n(f^*) \geq R(f^*_\eta) - R(f^*) + \eta) \\
&= P\left(R_n(f^*_\eta) - R_n(f^*) - (R(f^*_\eta) - R(f^*)) \geq \eta\right) \\
&\leq \exp\left(-\frac{n}{4}\left(\frac{\eta^{2-\beta}}{B} \wedge 3\eta\right)\right) \\
&\leq \exp\left(-\frac{n\eta^{2-\beta}}{4B}\right).
\end{aligned}$$

Для $n > \frac{B\log(\frac{2}{\delta})}{4\eta^{2-\beta}}$ последний член ограничен $\frac{\delta}{2}$. С вероятностью не менее $1-\frac{\delta}{2}$ мы получаем $(1+c)(R_n(f^*_\eta) - R_n(f^*)) \leq 2(1+c)\eta$. Обозначая $\varepsilon^{\frac{1}{2-\beta}} = \eta$, мы получаем (так как условие на моменты выполнено для $g_1,\ldots,g_{N_\eta}$), что с вероятностью не меньшей $1-\frac{\delta}{2}$ выполнено (см. шаги Леммы 11)

$$\sup_{g\in\{g_1,\ldots,g_{N_\eta}\}} (Pg - (1+c)P_n g) \leq \varepsilon^{\frac{1}{2-\beta}},$$

при условии, что $n \gtrsim B(1+c)^2 \left(\frac{\mathcal{D}^{\text{loc}}\left(\mathcal{G}, \varepsilon^{\frac{1}{2-\beta}}/2,\ \beta,\ B\right)}{\varepsilon} + \frac{\log(\frac{2}{\delta})}{\varepsilon} \right)$. Используя неравенство Буля, получаем, что с вероятностью не меньшей $1-\delta$ при тех же условиях на n выполнено $R(\hat{f}_\eta) - R(f^*) \leq (3+2c)\varepsilon^{\frac{1}{2-\beta}}$. Первая часть утверждения следует отсюда по аналогии с доказательством Леммы 11.

Теперь рассмотрим случай, когда выполнено (3.8). В этом случае для $\eta_1 > 0$

$$B(R(f^*_{\eta_1}) - R(f^*))^\beta \leq C\mathbb{E}|\ell(f^*_{\eta_1}(X), Y) - \ell(f^*(X), Y)| \leq \eta_1.$$

для некоторой абсолютной константы $C > 1$. Фиксируем $\eta_1 = CB\varepsilon^{\frac{\beta}{2-\beta}}$ и получаем $R(f^*_{\eta_1}) - R(f^*) \leq \varepsilon^{\frac{1}{2-\beta}}$. Повторяя шаги Леммы 11, мы получаем с вероятностью не меньшей $1 - \frac{\delta}{2}$

$$\sup_{g \in \{g_1, \ldots, g_{N_{\eta_1}}\}} (Pg - (1+c)P_n g) \leq \varepsilon^{\frac{1}{2-\beta}},$$

при условии $n \gtrsim B(1+c)^2 \left(\frac{\mathcal{D}^{\text{loc}}\left(\mathcal{G}_{\mathcal{Y}}, B\varepsilon^{\frac{\beta}{2-\beta}},\ 1,\ 1\right)}{\varepsilon} + \frac{\log(\frac{2}{\delta})}{\varepsilon} \right)$. Доказательство завершается как и ранее.

□

Утверждение 8. Доказательство основано на Теореме 6 из работы [5]. Пусть $h \in [0,1]$. Предположим, что X имеет равномерное распределение на единичном шаре, которое является изотропным и лог-вогнутым распределением [51]. Для $f \in \mathcal{F}$ распределение $Y|X$ будет определено следующим образом: $P^f_{Y=1|X} = \frac{1+f(X)h}{2}$. Известно [5], что это распределение $Y|X$ порождает класс $\mathcal{L}_{\mathcal{Y}}$, который является $(1, \frac{1}{h})$ классом Бернштейна при любом выборе P_X. Для $\varepsilon = \varepsilon_d \in [0,1]$ мы хотим построить подмножество $\mathcal{F}' \subset \mathcal{F}$, такое что оно образует ε-упаковку $\mathcal{F}$ пересеченного с $L_1(P)$ шаром радиуса 2ε.

Заметим, что $\sup\limits_{\varepsilon \in [0,1]} \sup\limits_{f \in \mathcal{F}} \log\left(\mathcal{N}(\mathcal{F} \cap \mathcal{B}_{L_1}(f, 2\varepsilon), \varepsilon)\right) \gtrsim d$, так как иначе в бесшумном случае наша оценка противоречила бы нижней оценке [59]. Однако из симметрии единичного шара и того факта, что распределение равномерное, легко видеть, что для любого $\varepsilon \in [0,1]$ и любого $f \in \mathcal{F}$ выполнено $\log\left(\mathcal{N}(\mathcal{F} \cap \mathcal{B}_{L_1}(f, 2\varepsilon), \varepsilon)\right) \gtrsim d$.

Используя доказательство упомянутой Теоремы 6, мы получаем, что если $\varepsilon \in [0,1]$ и $8n\frac{h^2}{1-h}\varepsilon \leq 0.71\log(|\mathcal{F}'|)$, то $\inf\limits_{\tilde{f}} \sup\limits_{f^*\in\mathcal{F}} \mathbb{E}(R(\tilde{f}) - R(f^*)) \geq \frac{0.29\varepsilon h}{4}$. Но так как $\log(|\mathcal{F}'|) \gtrsim d$, выбирая $\varepsilon \simeq \frac{d(1-h)}{nh^2}$ мы получаем $\inf\limits_{\tilde{f}} \sup\limits_{f^*\in\mathcal{F}} \mathbb{E}(R(\tilde{f}) - R(f^*)) \gtrsim \frac{d(1-h)}{nh}$, при условии что $\varepsilon \leq 1$. Последнее условие выполнено, если $h \gtrsim \sqrt{\frac{d}{n}}$. Комбинируя эту оценку с нижней оценкой $\frac{d}{n}$ для бесшумного случая [59], мы получаем $\frac{d}{nh}$. Доказательство завершается выбором $h = \frac{1}{B}$. □

Следствие 4. Известно, что в этом случае для всех $g \in \mathcal{G}_Y$ выполнено $Pg = \|f - f^*\|^2_{L_2(P)}$ (см., например, [46]), где f — функция, соответствующая g. Для всех $g_1, g_2 \in \mathcal{G}_Y$ выполнено

$$\begin{aligned}\|g_1 - g_2\|_{L_2(P)} &= \sqrt{P((f_1(X)-Y)^2 - (f_2(X)-Y)^2)^2} \\ &\leq 2\sqrt{P((f_1(X)-f_2(X))^2} \\ &= 2\|f_1 - f_2\|_{L_2(P)}.\end{aligned}$$

Мы повторяем те же шаги, которые мы использовали в Следствии 3.

$$\begin{aligned}&R(\hat{f}_\eta) - R(f^*) \\ &\leq \sup_{g\in\{g_1,\dots,g_{N_\eta}\}} (Pg - (1+c)P_n g) + (1+c)(R_n(f^*_\eta) - R_n(f^*)),\end{aligned}$$

Заметим, что для всех $g \in \mathcal{L}_\mathcal{Y}$ условие Бернштейна выполнено, так как $Pg^2 = P((f(X)-Y)^2 - (f^*(X)-Y)^2)^2 \leq 4P(f(X)-f^*(X))^2 = 4Pg$. Далее $g_1, \dots, g_{N_\eta}$ образуют 2η покрытие класса $\mathcal{L}_\mathcal{Y}$ (так как $\|g_1 - g_2\|_{L_2(P)} \leq 2\|f_1 - f_2\|_{L_2(P)}$). Мы повторяем шаги Леммы 11, но по отношению к $L_2(P)$ расстоянию. Далее мы будем отмечать только различия с доказательствами Леммы 11. Мы определим $\mathcal{G}_\eta = \{0, g_1, \dots, g_{N_\eta}\}$, $\mathcal{G}_0 = \mathcal{G}_\eta \cap \mathcal{B}_{L_2}(0, 4\eta)$ и $\mathcal{G}_1 = \{0\} \cup (\mathcal{G}_\eta \setminus \mathcal{B}_{L_2}(0, 4\eta))$. Добавляя нулевую функцию наши числа покрытия изменятся только на небольшой константный фактор. Далее

$$\sup_{g\in\mathcal{G}_\eta}(Pp[g] - (1+c)P_n p[g]) \leq \sup_{g\in\mathcal{G}_0}(Pg - (1+c)P_n g) + \sup_{g\in\mathcal{G}_1}(Pg - (1+c)P_n g).$$

Как и ранее мы анализируем эти члены по очереди. Используя условие Бернштейна мы получаем для всех $g \in \mathcal{G}_\eta$ $P\left(Pg - P_n g > \frac{cPg}{1+c}\right) \leq \exp\left(-\frac{nc^2 Pg}{32(1+c)^2}\right)$. Так как $Pg^2 \leq 4Pg$ мы получаем с учетом $\sqrt{Pg^2} \geq 2^j \eta$

$$P\left(Pg - P_n g > \frac{cPg}{1+c}\right) \leq \exp\left(-\frac{nc^2 2^{2j}\eta^2}{128(1+c)^2}\right).$$

Далее мы должны контролировать размер подмножества $\mathcal{G}_\eta$, состоящего из функций, для которых $2^j\eta \leq \sqrt{Pg^2} \leq 2^{j+1}\eta$. Как и в Лемме 11 мы показываем, что для фиксированного c, если $n \gtrsim \frac{\mathcal{D}_{L_2}^{\mathrm{loc}}(\mathcal{F},\eta)}{\eta^2} + \frac{\log(\frac{1}{\delta})}{\eta^2}$, то с вероятностью не меньшей $1-\delta/3$ выполнено $\sup_{g\in\mathcal{G}_1}(Pg-(1+c)P_n g) = 0$. Далее мы анализируем $\sup_{g\in\mathcal{G}_0}(Pg-(1+c)P_n g)$. Как и ранее мы получаем $P(Pg - P_n g \geq \eta^2) \leq \exp\left(-\frac{nc\eta^2}{32(1+c)^2}\right)$. Для данного $n \gtrsim \frac{\mathcal{D}_{L_2}^{\mathrm{loc}}(\mathcal{F},\eta)}{\eta^2} + \frac{\log(\frac{1}{\delta})}{\eta^2}$ мы получаем с вероятностью не меньшей $1-\delta/3$, что $\sup_{g\in\mathcal{G}_1}(Pg - (1+c)P_n g) \leq \eta^2$. В итоге, так как $0 \leq R(f_\eta^*) - R(f^*) \leq \eta^2 \leq 1$

$$P(R_n(f_\eta^*) - R_n(f^*) \geq R(f_\eta^*) - R(f^*) + \eta^2) \leq \exp\left(-\frac{n\eta^2}{16}\right)$$

и для $n \gtrsim \frac{\log(\frac{1}{\delta})}{\eta^2}$ последняя вероятность ограничена сверху $\delta/3$. Доказательство заканчивается как и ранее. □

Глава 4

Устойчивость и схемы сжатия выборок

4.1. Основная теорема

Подход, основанный на локальной энтропии, очевидно, не является панацеей. В некоторых случаях эта техника не может дать достаточно точных статистических гарантий. Самым простым таким примером является бинарная классификация, когда не делается никаких предположений о шуме. Также в непараметрическом случае, когда $\beta \neq 1$ наша верхняя оценка несколько хуже чем оценка, получаемая в работах [5, 6]. Другим важным случаем является бесшумная классификация, определенная выше. Оптимальной верхней оценкой для риска является $\frac{d}{n} + \frac{\log(\frac{1}{\delta})}{n}$ [15, 42], где d – VC размерность класса. Стоит отметить, что этот порядок не достигается с помощью минимизаторов эмпирического риска [18]. В некоторых случаях оценка для минимизатора риска в точности имеет порядок $\frac{d\log(\frac{n}{d})}{n} + \frac{\log(\frac{1}{\delta})}{n}$. Более того, так как локальные энтропии связаны именно с равномерной сходимостью и минимизацией эмпирического риска, они не появляются в нижних минимаксных оценках.

Есть еще два принципа, гарантирующие обобщающую способность: схемы сжатия выборок и устойчивость. Мы дадим несколько формальных определений.

Определение 7 (Схемы сжатия выборок [16])**.** *Определим последовательность функций, независящих от перестановок аргументов* $\kappa_n : (\mathcal{X} \times \mathcal{Y})^n \to \cup_{i=1}^k (\mathcal{X} \times \mathcal{Y})^i$*. Эти функции будем называть* функциями сжатия*. Функции восстановления* $\rho : \cup_{i=1}^k (\mathcal{X} \times \mathcal{Y})^i \to \mathcal{Y}^{\mathcal{X}}$*. Функции* κ_n *и* ρ *определяют схему сжатия размера* k*, если для всех* $n \in \mathbb{N}$ *и* $f \in \mathcal{F}$ *и любой выборки* $(x_i, f(x_i))_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ *выполнено* $\kappa_n((x_i, f(x_i))_{i=1}^n) \subseteq (x_i, f(x_i))_{i=1}^n$*, то, обозначая* $\hat{f} = \rho(\kappa_n((x_i, f(x_i))_{i=1}^n))$*, выполнено* $f(x_i) = \hat{f}(x_i)$ *для всех* $i = 1, \ldots, n$*.*

Лемма 12 (Флойд и Вармут [16]). *Пусть ρ — функция восстановления для некоторой схемы сжатия размера k. Для $f \in \mathcal{F}$ и простой выборки $(X_i, f(X_i))_{i=1}^n$ выполнено с вероятностью не менее $1-\delta$ одновременно для всех множеств $A \subset (X_i, f(X_i))_{i=1}^n$, для которых $|A| \leq k$ и $(\rho(A))(X_i) = f(X_i)$ для $i = 1, \ldots, n$,*

$$P((\rho(A))(X) \neq f(X)) \leq \frac{k \log(\frac{en}{k})}{n-k} + \frac{\log(\frac{1}{\delta})}{n-k}. \tag{4.1}$$

Вопрос существование схем сжатия размера $O(d)$ является хорошо известной открытой проблемой [16]. Однако, если и можно построить схему сжатия размера $O(d)$, то известно [16], что порядок $\frac{d \log(\frac{n}{d})}{n} + \frac{\log(\frac{1}{\delta})}{n}$ оценки 4.1 не может быть улучшен для некоторых схем сжатия. Одновременно оценки, основанные на локальных энтропиях, всегда не хуже [20]. Тем не менее при некоторых дополнительных предположениях схемы сжатия дают оптимальные порядки.

Определение 8 (Устойчивая схема сжатия выборок). *Схема сжатия (κ_n, ρ) называется устойчивой, если для всех n, $f \in \mathcal{F}$, любой выборки $(x_i, f(x_i))_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ и любой $(x, y) \in (x_i, f(x_i))_{i=1}^n \backslash \kappa_n((x_i, f(x_i))_{i=1}^n)$ выполнено $\kappa_{n-1}((x_i, f(x_i))_{i=1}^n \backslash (x, y)) = \kappa_n((x_i, f(x_i))_{i=1}^n)$.*

Следующее общее определение мотивировано схожим свойством, появлявшимся при анализе классов, замкнутых относительно пересечений [18].

Определение 9 (Однородные схемы сжатия). *Устойчивая схема сжатия (κ_n, ρ) называется однородной, если для всех n, $f \in \mathcal{F}$, любой выборки $(x_i, f(x_i))_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ и любой $(x, y) \in \kappa_n((x_i, f(x_i))_{i=1}^n$ выполнено $\kappa_n((x_i, f(x_i))_{i=1}^n) \setminus (x, y) \subseteq \kappa_{n-1}((x_i, f(x_i))_{i=1}^n \setminus (x, y))$.*

Мы докажем следующую Теорему.

Теорема 8. *Для устойчивой схемы сжатия (κ_n, ρ) размера k с вероятностью не меньшей $1-\delta$ выполнено*

$$\mathbb{E}R(\hat{f}) \leq \frac{k}{n+1}, \quad R(\hat{f}) \lesssim \frac{k \log(\frac{1}{\delta})}{n},$$

где $\hat{f} = \rho(\kappa_n((X_i, Y_i)_{i=1}^n))$. *Более того, если функция* ρ *принимает значения в классе с VC размерностью* $d \lesssim k$, *то существует эффективная модификация* [1] *схемы сжатия (выдающая функцию, которую мы обозначим* $\hat{g}$), *которая дает алгоритм с оценкой*

$$R(\hat{g}) \lesssim \frac{k \log(k)}{n} + \frac{\log(\frac{1}{\delta})}{n}. \tag{4.2}$$

Если (κ_n, ρ) *является однородной схемой сжатия, то*

$$R(\hat{f}) \lesssim \frac{k}{n} + \frac{\log(\frac{1}{\delta})}{n}.$$

Доказательство будет приведено ниже. Опишем модификацию схемы, используемую для получения 4.2. Для выборки S из $3n$ элементов обозначим $S_{1/3}, S_{2/3}$ соответственно первые n и первые $2n$ элементов выборки. Модификация схемы сжатия заключается в том, что новая точка X классифицируется согласно голосованию большинства трех функций, полученных с помощью применения схемы сжатия к выборкам $S_{1/3}, S_{2/3}, S$ (аналогично $\mathbf{L}_2$ алгоритму, введенному для общих минимизаторов эмпирического риска [17]). Как прямое следствие Леммы 8, когда размер множества имеет порядок $k = O(d)$, где d — VC размерность, однородные схемы сжатия дают оптимальный порядок обучения до константных факторов, совпадая с нижней оценкой [43], а все устойчивые схемы сжатия оптимальны по вероятности [1].

Оценки SVM и онлайновое обучение

Следствие 5. *Для класса* $\mathcal{F}$ *линейных разделителей в бесшумном случае существует обучающий алгоритм, обучение которого происходит за полиномиальное время, такой что для построенного им классификатора* $\hat{f}$ *с вероятностью не меньшей* $1 - \delta$

$$R(\hat{f}) \lesssim \frac{d \log(d)}{n} + \frac{\log(\frac{1}{\delta})}{n}. \tag{4.3}$$

[1] Хотя мы не рассматриваем вопросы оптимальности, под эффективностью мы будем иметь в виду тот факт, что схему сжатия выборок нужно запустить в точности три раза. Техника, которую мы будем использовать, основана на алгоритме голосования из работы [17].

Доказательство. Сначала нужно показать, что в бесшумном случае разделяющая гиперплоскость $\mathbb{R}^d$, построенная с помощью SVM, определяет устойчивую схему сжатия размера не больше чем $d+1$. Это следует из существования *необходимых* опорных векторов (см главу 14 в [1]). Принимая во внимание, что VC размерность класса равна d+1, используя Теорему 4.2 мы получаем порядок $\frac{d\log(\frac{1}{\delta})}{n}$ для SVM и порядок $\frac{d\log(d)}{n} + \frac{\log(\frac{1}{\delta})}{n}$ для его модификации. Также отметим, что в обоих случаях построение классификатора требует полиномиальное время. □

Последняя оценка практически совпадает с минимаксной нижней оценкой $\frac{d}{n} + \frac{\log(\frac{1}{\delta})}{n}$ [43] . Ранее полиномиальные алгоритмы для этой задачи, для которых также выполнены точные статистические гарантии, были известны только для классов однородных разделяющих правил и лог-вогнутых распределений X.

Построение классификаторов с помощью онлайновых алгоритмов

Пусть в бесшумном случае нам дан консервативный онлайновый алгоритм, допускающий не более k ошибок на любой выборке (см. [16] или [60] для описания постановки). Нашей целью будет использование онлайнового алгоритма для построения алгоритма в стандартной постановке. Предположим, что множество $\mathcal{X}$ упорядочено. Рассмотрим следующий классификатор, выдающий $\hat{f}$ на выборке $S = (X_i, Y_i)_{i=1}^n$. Для выборки S мы определим множество S^* как множество пар (X_i, Y_i), отсортированных согласно порядку на множестве $\mathcal{X}$ и множество $S^*_{\preceq x}$ как подмножество S^*, состоящее из пар (X_i, Y_i), таких что все X_i предшествуют x. Теперь для всех $x \in \mathcal{X}$

- Если существуют $j \in \{1, \ldots, n\}$, такие что $x = X_j$, то определим $\hat{f}(x) = Y_j$.
- Иначе определим $\hat{f}(x)$ как метра x, который мы получили, применяя последний классификатор, получаемый при применении крайнего классификатора, полученного при запуске онлайнового алгоритма на множестве $S^*_{\preceq x}$.

Легко показать, что $\hat{f}$ является выходным классификатором схемы сжатия размера k [16]. Более того, так как алгоритм консервативный, то и соответствующая схема сжатия также консервативная и это дает нам порядок $R(\hat{f}) \leq \frac{k\log(\frac{1}{\delta})}{n}$. Это дает улучшения по сравнению со стратегией выбора самого живучего классификатора, для которой известен порядок $\frac{k\log(\frac{k}{\delta})}{n}$ [60].

Однако в том случае, когда мы можем гарантировать, что алгоритм имеет VC размерность $d \lesssim k$ мы можем применить модификацию 4.2. Как и ранее мы строим три подмножества $S_{\frac{1}{3}}, S_{\frac{2}{3}}, S$ и соотвествующие упорядоченные множества $S^*_{\frac{1}{3},\preceq x}, S^*_{\frac{2}{3},\preceq x}, S^*_{\preceq x}$. Модифицированный классификатор $\hat{g}(x)$ определяется как голосование большинства по трем значениям, которые мы получим применением последнего классификатора при использовании онлайнового алгоритма на множествах $S^*_{\frac{1}{3},\preceq x}, S^*_{\frac{2}{3},\preceq x}, S^*_{\preceq x}$. Эта модификация дает порядок $R(\hat{g}) \leq \frac{k\log(k)}{n} + \frac{\log(\frac{1}{\delta})}{n}$, который практически совпадает с наилучшим известным порядком $\frac{k}{n} + \frac{\log(\frac{1}{\delta})}{n}$ [60]. Интересно, что последняя оценка достигается с помощью другой стратегии и использует подход, основанный на мартингалах.

4.2. Доказательства

Теоремы 8. Теорема по математическому ожиданию хорошо известна. С несколько другими обозначениями она следует из Леммы 2.2 в [15] или аналогичных выкладок в [1]. Далее для простой выборки $(X_i, Y_i)_{i=1}^n$ мы определим

$$\hat{f} = \rho(\kappa_n((X_i, Y_i)_{i=1}^n).$$

Мы продолжим с помощью метода моментов. Для любого $\varepsilon > 0$ и $p \in \mathbb{N}$ с помощью неравенства Маркова мы получаем $P(R(\hat{f}) \geq \varepsilon) \leq \frac{\mathbb{E}(R(\hat{f}))^p}{\varepsilon^p}$. Используя идею из [18] (Теорема 4) мы получаем, что $\mathbb{E}(R(\hat{f}))^p$ равно ожидаемой вероятности того, что $\hat{f}$ ошибется в точности на p независимых объектах. Используя симметризационный аргумент, так как все перестановки $n+p$ независимых точек равновероятны мы ограничиваем $\frac{\mathbb{E}(R(\hat{f}))^p}{\varepsilon^p}$ с помощью $\frac{\psi(n,p)}{\varepsilon^p\binom{n+p}{p}}$, где $\psi(n,p)$ — максимально возмож-

ное число способов (при условии $(X_i, Y_i)_{i=1}^{n+p}$) выбрать p из $n+p$ точек, так что $\hat{f}$, вычисленная на оставшихся n точках неправильно классифицирует эти p точек.

Теперь мы ограничим $\psi(n,p)$ для устойчивых схем сжатия. Мы докажем, что $\psi(n,p) \leq k^p$. Для $\psi(n,1)$ единственная точка, которая не в обучающей выборке должна быть точкой, попавшей в сжатую выборку для $n+1$ точки, так как иначе $\hat{f}$ правильно классифицирует оставшуюся точку. Иначе $\psi(n,1) \leq k$. Мы делаем то же самое $\psi(n,p)$ для общего p. Легко показать по индукции, что $\psi(n,p) \leq k^p$. Таким образом, мы получаем

$$P(R(\hat{f}) \geq \varepsilon) \leq \frac{k^p}{\varepsilon^p \binom{n+p}{p}} \leq \frac{(kp)^p}{(\varepsilon n)^p}.$$

Нам интересны два конкретных значения p. Первое из них $p = k$. Обозначая $\delta = \frac{(k^2)^k}{(\varepsilon n)^k}$ мы получаем, что с вероятностью не меньшей $1-\delta$ выполнено

$$R(\hat{f}) \leq \frac{k^2}{n\delta^{\frac{1}{k}}}. \tag{4.4}$$

Второе значение $p = \lceil \log(\frac{1}{\delta}) \rceil$. Для $\varepsilon = \frac{ed\log(\frac{1}{\delta})}{n}$ мы получаем, что с вероятностью не меньшей $1-\delta$ выполнено $R(\hat{f}) \leq \frac{ed\log(\frac{1}{\delta})}{n}$, таким образом первая часть утверждения установлена.

Теперь мы доказываем вторую часть утверждения. Без ограничения общности мы предполагаем, что нам дана выборка S из $3n$ элементов и определим $S_{1/3}, S_{2/3}$ как первые n и $2n$ элементов S. Определим $\hat{f}_1 = \rho(\kappa_n(S_{1/3}))$, $\hat{f}_2 = \rho(\kappa_{2n}(S_{2/3}))$, и $\hat{f}_3 = \rho(\kappa_{3n}(S))$, и $\hat{g}$ как голосование большинства $\hat{f}_1, \hat{f}_2, \hat{f}_3$, формально $\hat{g} = \text{sign}(\hat{f}_1 + \hat{f}_2 + \hat{f}_3)$. Определим $E_i = \{x \in \mathcal{X} : \hat{f}_i(x) \neq f^*(x)\}$. Используя ту же технику доказательства, что и в Теореме 5 в [17] мы получаем $P(\hat{g}(X) \neq f^*(X)) \leq 3 \max_{1 \leq i < j \leq 3} P(E_i \cap E_j)$. Достаточно контролировать $P(E_i \cap E_j)$, так как для оставшихся слагаемых доказательство будет таким же. Мы выберем множества E_1 и E_2. Запишем $P(E_1 \cap E_2) = P(E_2|E_1)P(E_1)$. Определим $N = \sum_{i=n+1}^{2n} \mathbb{1}[X_i \in E_1]$. Условно по $S_{1/3}$ случайная величина N имеет биномиальное распределение со средним $nP(E_1)$. Более того, (X_i, Y_i) для $i \in n+1, \ldots, 2n$ с

$X_i \in E_1$ (это элементы в $S_{2/3} \cap E_1$) являются условно независимыми при условии $S_{1/3}$.

Теперь покажем, что с вероятностью не меньшей $1-\delta$ выполнено

$$P(E_2|E_1) \lesssim \frac{k\log(N/k)}{N} + \frac{\log(\frac{1}{\delta})}{N}. \tag{4.5}$$

Заметим, что из-за наших предположений ρ выдает классификатор из класса $\mathcal{F}'$, размерность которого $d \lesssim k$. Таким образом, можно рассмотреть E_2 как множество ошибок минимизатора эмпирического риска на множестве $\mathcal{F}'$. Используя Теорему 2 из [17] мы получаем одновременно для всех минимизаторов эмпирического риска $\hat{h}$ в классе $\mathcal{F}'$ по отношению к выборке $S_{\frac{2}{3}}$, что с вероятностью не меньшей $1-\delta$ выполнено $P(E_2) \lesssim \frac{k\log(n/k)}{n} + \frac{\log(\frac{1}{\delta})}{n}$. Так как множества всех минимизаторов эмпирического риска на выборке $S_{\frac{2}{3}}$ является подмножеством всех минимизаторов эмпирического риска на выборке $S_{\frac{2}{3}} \cap E_1$, применяя Теорему 2 для обучающей выборки $S_{\frac{2}{3}} \cap E_1$ (при условии $S_{1/3}$) мы получаем 4.5.

Если $P(E_1) \geq C(\frac{k\log k}{n} + \frac{\log(\frac{1}{\delta})}{n})$ для достаточно большой константы C, то используя неравенство Чернова для N, с вероятностью не меньшей $1-\delta$ мы получаем $N \geq \frac{1}{2}P(E_1)n$ и с вероятностью как минимум $1-\delta$ мы получаем $N \leq 2P(E_1)n$. Если иначе $P(E_1) \leq C(\frac{k\log k}{n} + \frac{\log(\frac{1}{\delta})}{n})$, то мы получаем требуемую оценку, так как $P(E_1 \cap E_2) \leq P(E_1)$. В итоге, с помощью монотонности $\log(x)/x$ с вероятностью не меньшей $1-3\delta$

$$P(E_2|E_1)P(E_1) \lesssim \frac{k\log(P(E_1)n/k)}{n} + \frac{\log(\frac{1}{\delta})}{n}.$$

Используя 4.4, с вероятностью $1-\delta$ по выборке $S_{1/3}$ получаем, что $P(E_1) \lesssim \frac{k^2}{n\delta^{\frac{1}{k}}}$. Таким образом, с вероятностью $1-4\delta$

$$P(E_1 \cap E_2) \lesssim \frac{k\log(k/\delta^{\frac{1}{k}})}{n} + \frac{\log(\frac{1}{\delta})}{n} \lesssim \frac{k\log(k)}{n} + \frac{\log(\frac{1}{\delta})}{n}.$$

Неравенство 4.2 доказано.

Доказываем утверждение для однородных схем сжатия. Мы обобщаем общую технику из работы [18]. Как и ранее мы должны ограничить сверху величину $\psi(n,p)$. Предположим, что все $n+p$ элементов упорядочены и обозначим эту

упорядоченную выборку S_{n+p}. Рассмотрим функцию $\hat{f}$, которая неверно классифицирует p элементов и основана на n оставшихся элементах. Обозначим текущую обучающую выборку как S_n. Рассмотрим сжатие выборки из $n+p$ точек (это выборка $\kappa_{n+p}(S)$) и выберем первый элемент x_1 (первую компоненту пары (x_1, y_1)) в ней в соответствии с этим порядком. Для этого элемента есть два варианта:

1. Этот элемент x_1 неверно классифицируется $\hat{f}$. В этом случае мы закодируем этот элемент 1.

2. Этот элемент x_1 верно классифицирован $\hat{f}$. В этом случае x_1 находится в сжатой подвыборке выборки из n элементов. Мы закодируем этот элемент x_1 с помощью нуля 0.

Существует всего две возможные ситуации так как обучающая выборка из n элементов есть подмножество множества из $n+p$ элементов и из-за свойства однородности его множество сжатия содержит все элементы пересечения S_n с $\kappa(S_{n+p})$. Формально $\kappa(S_{n+p}) \cap S_n \subseteq \kappa(S_n)$. Тогда, так как объект x_1 правильно классифицирован, мы получаем $x_1 \in S_n$ и, таким образом, $x_1 \in \kappa(S_n)$.

Теперь после выбора первого элемента x_1 мы переходим к элементу x_2. Есть два варианта: если на первом шаге элемент x_1 был правильно классифицирован $\hat{f}$, то мы выбираем следующий элемент (по порядку) в множестве $\kappa_{n+p}(S_{n+p})$. Иначе, если x_1 был неправильно классифицирован, то мы рассмотрим множество $S_{n+p} \backslash x_1$ и его множество сжатия $\kappa_{n+p-1}(S_{n+p} \backslash x_1)$ и выбираем x_2 из его множества сжатия.

Как и ранее для первого элемента x_1 для второго элемента есть две опции x_2 в зависимости от того, как $\hat{f}$ классифицирует x_2. Мы кодируем x_2 с помощью 0 или 1 в зависимости от этого продолжаем аналогично для x_3. Для $\hat{f}$, после не более чем $s \leq k+p$ шагов мы закодируем множество $x_1, \ldots, x_s$, состоящее из элементов $\kappa_n(S_n)$ и p неверно классифицированных элементов. Заметим, что эта схема кодирования сопоставляет каждому $\hat{f}$, который делает в точности p ошибок, уникальную и не более чем $k+p$-элементную последовательность нулей и единиц.

Так как классификатор делает p ошибок, то всего будет не более $\binom{k+p}{p}$ этих упорядоченных последовательностей. Таким образом, $\psi(n,p) \leq \binom{k+p}{p}$ и мы получаем $P(R(\hat{f}) \geq \varepsilon) \leq \frac{\binom{k+p}{k}}{\varepsilon^p \binom{n+p}{p}}$. Выбирая $p = \lceil \log(\frac{1}{\delta}) \rceil$ получаем, что с вероятностью не менее $1-\delta$ выполнено $R(\hat{f}) \leq \frac{ek}{n} + \frac{e \log(\frac{1}{\delta})}{n}$. $\square$

Глава 5

Меры сложности в трансдуктивном обучении

Настоящая глава посвящена изучению трансдуктивного обучения [9]. Эта постановка несколько отличается от ранее изучавшейся постановки, где предполагалось, что данные приходят из неизвестного распределения. В трансдуктивном обучении мы наблюдаем m точек, которые имеют метки классов и u неразмеченных точек. Целью является как можно более точная классификация тестовых точек. Существуют две различные постановки трансдуктивного обучения, определенные в работе [9]. В первой постановке предполагается, что обучающая и тестовые выборки получены независимо из неизвестного распределения P. Во второй постановке предполагается, что обучающая и тестовая выборки реализуются с помощью равновероятных разбиений генеральной совокупности из $N = m + u$ объектов на два множества, размеры которых соответственно m и u. Вторая постановка более хорошо изучена (см [9], [10], [11], [61], [62] и [63]), во многом из-за того, что верхние оценки в этой постановке влекут оценки на риск и для первой постановки.

Важным отличием от стандартной постановки является тот факт, что m элементов обучающей выборки являются зависимыми, так как получены с помощью вытягивания без возвращений. Как следствие нужны другие вероятностные техники для анализа данной постановки.

Некоторые предыдущие результаты. В работе [61] вводятся так называемая Трансдуктивная Радемахеровская сложность (TPC). Эта мера сложности появляется в верхних оценках в трансдуктивном обучении, но имеет некоторые недостатки: например, зависит от меток неизвестной тестовой выборки. Для того чтобы избежать эту проблему авторы используют неравенство сжатия [24] (contraction), которое часто дает достаточно грубые результаты [64].

В работе [63], разрабатывается вариант неравенства Талаграна для семпли-

рования без возвращений, которое используется там для получения быстрых порядков сходимости.

5.1. Обозначения и ранние результаты

Для функции f среднее значение на конечном множестве S будет обозначаться $\bar{f}(S)$.

Определение 10 (Условная Радемахеровская сложность)**.** *Зафиксируем множество $\mathcal{Z}_m = \{Z_1, \ldots, Z_m\} \subseteq \mathcal{Z}$. Следующая случайная величина называется условной Радемахеровской сложностью:*

$$\hat{R}_m(F, \mathcal{Z}_m) = \mathbb{E}_\varepsilon \left[\frac{2}{m} \sup_{f \in F} \sum_{i=1}^{m} \varepsilon_i f(Z_i) \right],$$

где $\varepsilon = \{\varepsilon_i\}_{i=1}^m$ независимые Радемахеровские случайные величины, принимающие значения ± 1 с вероятностью $1/2$. Когда множество $\mathcal{Z}_m$ ясно из контекста, мы будем обозначать Радемахеровскую сложность символом $\hat{R}_m(F)$.

Для трансдуктивной постановки в работе [61] вводится следующее обозначение:

Определение 11 (Трансдуктивная Радемахеровская Сложность)**.** *Зафиксируем множество $\mathcal{Z}_N = \{Z_1, \ldots, Z_N\} \subseteq \mathcal{Z}$, неотрицательные целые числа m, u такие что $N = m + u$ и $p \in \left[0, \frac{1}{2}\right]$. Следующая величина называется Трансдуктивной Радемахеровской сложностью:*

$$\hat{R}^{td}_{m+u}(F, \mathcal{Z}_N, p) = \left(\frac{1}{m} + \frac{1}{u} \right) \mathbb{E}_\sigma \left[\sup_{f \in F} \sum_{i=1}^{N} \sigma_i f(Z_i) \right],$$

где $\sigma = \{\sigma_1\}_{i=1}^{m+u}$ являются независимыми случайными величинами, принимающими значения ± 1 с вероятностью p и значение 0 с вероятностью $1 - 2p$.

Важность обеих мер сложности можно продемонстрировать следующим результатом

Теорема 9 ([63], [61]). *Зафиксируем N-элементное подмножество $\mathcal{Z}_N \subseteq \mathcal{Z}$ и пусть $m < N$ элементы $\mathcal{Z}_m$ получены равновероятным выбором без возвращений из $\mathcal{Z}_N$. Пусть m элементов $\mathcal{X}_m$ получены равновероятно из $\mathcal{Z}_N$ с возвращениями. Обозначим $\mathcal{Z}_u = \mathcal{Z}_N \setminus \mathcal{Z}_m$ и $u = |\mathcal{Z}_u| = N - m$. Тогда*

$$\mathbb{E}_{\mathcal{Z}_m}\left[\sup_{f\in F}\left(\bar{f}(\mathcal{Z}_u) - \bar{f}(\mathcal{Z}_m)\right)\right] \leq \mathbb{E}_{\mathcal{X}_m}\left[\hat{R}_m(F, \mathcal{X}_m)\right]. \tag{5.1}$$

Следующая оценка в терминах TPC доказана в работе [61]. Пусть функции в F равномерно ограничены B. Тогда для $p_0 = \frac{mu}{N^2}$ и $c_0 < 5.05$:

$$\mathbb{E}_{\mathcal{Z}_m}\left[\sup_{f\in F}\left(\bar{f}(\mathcal{Z}_u) - \bar{f}(\mathcal{Z}_m)\right)\right] \leq \hat{R}^{td}_{m+u}(F, \mathcal{Z}_N, p_0) + c_0 B\frac{N\sqrt{\min(m,u)}}{mu}. \tag{5.2}$$

Теперь мы введем нашу новую меру сложности

Определение 12 (Перестановочная Радемахеровская сложность). *Пусть $\mathcal{Z}_m \subseteq \mathcal{Z}$ — любое множество из m элементов. Для всех $n \in \{1, \ldots, m-1\}$ следующее выражение будет называться перестановочной Радемахеровской сложностью (ПРС):*

$$\hat{Q}_{m,n}(F, \mathcal{Z}_m) = \mathbb{E}_{\mathcal{Z}_n}\left[\sup_{f\in F}\left(\bar{f}(\mathcal{Z}_k) - \bar{f}(\mathcal{Z}_n)\right)\right],$$

где $\mathcal{Z}_n$ есть случайное подмножество $\mathcal{Z}_m$, содержащее n элементов, выбранных равновероятно без возвращений и $\mathcal{Z}_k = \mathcal{Z}_m \setminus \mathcal{Z}_n$. Когда множество $\mathcal{Z}_m$ ясно из контекста мы будем использовать обозначение $\hat{Q}_{m,n}(F)$.

Название ПРС объясняется тем, что если m четное число, то $\hat{Q}_{m,m/2}(F)$ и $\hat{R}_m(F)$ очень похожи. Действительно, единственная разница в том, что математическое ожидание в ПРС берется по всем перестановкам, содержащим одинаковое число "-1" и "$+1$" в то время как в Радемахеровской сложности математическое ожидание берется по всем последовательностям знаков. Термин перестановочная сложность появлялся в работе [65], где он был использован для обозначения некоторой меры сложности для задачи выбора модели.

5.2. Симметризация и сравнения

Мы начнем с варианта неравенства симметризации для постановки, где точки получены равновероятным выбором без возвращений.

Теорема 10. *Зафиксируем N-элементное подмножество $\mathcal{Z}_N \subseteq \mathcal{Z}$ и пусть $m < N$ элементов $\mathcal{Z}_m$ были получены равновероятно без возвращений из $\mathcal{Z}_N$. Обозначим $\mathcal{Z}_u = \mathcal{Z}_N \setminus \mathcal{Z}_m$ с $u = |\mathcal{Z}_u| = N - m$. Если $m = u$ и m является четным, то для всех $n \in \{1, \ldots, m-1\}$:*

$$\frac{1}{2}\mathbb{E}_{\mathcal{Z}_m}\left[\hat{Q}_{m,m/2}(F, \mathcal{Z}_m)\right] \leq \mathbb{E}_{\mathcal{Z}_m}\left[\sup_{f \in F}\left(\bar{f}(\mathcal{Z}_u) - \bar{f}(\mathcal{Z}_m)\right)\right] \leq \mathbb{E}_{\mathcal{Z}_m}\left[\hat{Q}_{m,n}(F, \mathcal{Z}_m)\right].$$

Неравенства выполнены если мы добавим абсолютные значения в супремум.

Доказательство будет дано ниже. Это неравенство можно сравнить с предыдущими известными оценками Теоремы 9.

По сравнению с (5.1) и (5.2) новая оценка дает двухсторонний контроль и показывает, что в некотором смысле ПРС является правильной мерой сложности для данной постановки. Заметим, что нижняя оценка (называемая часто неравенством дессимметризации) не имеет аддитивных слагаемых, в отличие от известных подобных неравенств в стандартной постановке и не требуют ограниченности функций в классе F, что, однако, является необходимым условием в (5.2) и стандартном неравенстве диссметризации. Следующее простое утверждение, доказательство которого представлено в работе [22], позволяет получить верхние оценки для $\mathbb{E}_{\mathcal{Z}_m}\left[\hat{Q}_{m,m/2}(F, \mathcal{Z}_m)\right]$ в случае конечного класса F.

Утверждение 9. *Пусть F — конечный класс бинарных функций. Тогда*

$$\mathbb{E}_{\mathcal{Z}_m}\left[\hat{Q}_{m,m/2}(F, \mathcal{Z}_m)\right] \leq 2\sqrt{\frac{\log|F|}{m}}$$

Далее мы сравниваем ПРС с условной Радемахеровской сложностью:

Теорема 11. *Пусть $\mathcal{Z}_m \subseteq \mathcal{Z}$ является любым множеством из четного числа элементов m. Тогда:*

$$\hat{Q}_{m,m/2}(F, \mathcal{Z}_m) \leq \left(1 + \frac{2}{\sqrt{2\pi m} - 2}\right) \hat{R}_m(F, \mathcal{Z}_m). \tag{5.3}$$

Более того, если функции в F являются абсолютно ограниченными B, то

$$\left|\hat{Q}_{m,m/2}(F, \mathcal{Z}_m) - \hat{R}_m(F, \mathcal{Z}_m)\right| \leq \frac{2B}{\sqrt{m}}. \tag{5.4}$$

Доказательство. Доказательство основано на кауплинге между Радемахеровскими случайными величинами $\{\varepsilon_i\}_{i=1}^m$ и равновероятными перестановками $\{\eta_i\}_{i=1}^m$, содержащими по $m/2$ плюс единиц и минус единиц. Подробное доказательство будет дано ниже. □

Заметим, что Теорема 11 дает также улучшение по сравнению с Леммой 3 в [66]. Наш следующий результат показывает, что Теорема 11 дает практически оптимальные оценки.

Лемма 13. *Пусть $\mathcal{Z}_m \subseteq \mathcal{Z}$ и m является четным числом. Существуют два конечных класса F'_m и F''_m функций, отображающих $\mathcal{Z}$ в $\mathbb{R}$ и ограниченных* 1*, таких что:*

$$\hat{Q}_{m,m/2}(F'_m, \mathcal{Z}_m) = 0, \quad \frac{1}{\sqrt{2m}} \leq \hat{R}_m(F'_m, \mathcal{Z}_m) \leq \frac{2}{\sqrt{m}}; \tag{5.5}$$

$$\hat{Q}_{m,m/2}(F''_m, \mathcal{Z}_m) = 1, \quad 1 - \sqrt{\frac{2}{\pi m}} \leq \hat{R}_m(F''_m, \mathcal{Z}_m) \leq 1 - \frac{4}{5}\sqrt{\frac{2}{\pi m}}. \tag{5.6}$$

Доказательство будет дано ниже. Неравенства (5.5) показывают, что порядок $O(m^{-1/2})$ аддитивной оценки (5.4) не может быть улучшен. Более того, можно показать, что множитель, появляющийся в (5.3) не может быть улучшен до $1 + o(m^{-1/2})$. Следующая лемма сравнивает ПРС и Т врансдуктивную Радемахеровскую сложность:

Лемма 14. *Зафиксируем множество $\mathcal{Z}_N = \{Z_1, \ldots, Z_N\} \subseteq \mathcal{Z}$. Если $m = u$ и $N = m + u$:*

$$\hat{R}_N(F, \mathcal{Z}_N) \leq \hat{R}^{td}_{m+u}(F, \mathcal{Z}_N, 1/4) \leq 2\hat{R}_N(F, \mathcal{Z}_N).$$

Доказательство. Верхняя оценка была доказана в работе [61]. Для нижней оценки отметим, что если $p = 1/4$ случайные знаки σ_i из определения 11 имеют такое же распределение как и $\varepsilon_i \eta_i$, где ε_i являются независимыми Радемахеровскими случайными величинами, а η_i являются Бернулиевскими случайными величинами с параметром $1/2$. Таким образом, неравенство Йенсена дает

$$\hat{R}^{td}_{m+u}\left(F, \mathcal{Z}_N, 1/4\right) = \frac{4}{N}\mathbb{E}_{(\varepsilon,\eta)}\left[\sup_{f\in F}\sum_{i=1}^{m+u}\varepsilon_i\eta_i f(Z_i)\right] \geq \frac{4}{N}\mathbb{E}_{\varepsilon}\left[\sup_{f\in F}\sum_{i=1}^{m+u}\varepsilon_i\frac{1}{2}f(Z_i)\right].$$

□

Вместе с Теоремами 10 и 11 этот результат показывает, что если $m = u$, то ПРС не может быть существенно больше ТРС:

Следствие 6. *В обозначениях Теоремы 10 мы имеем:*

$$\mathbb{E}_{\mathcal{Z}_m}\left[\hat{Q}_{m,m/2}(F, \mathcal{Z}_m)\right] \leq \left(2 + \frac{4}{\sqrt{2\pi N} - 2}\right)\hat{R}^{td}_{m+u}(F, \mathcal{Z}_N, 1/4).$$

Если функции в F равномерно ограничены B, то мы также имеем нижнюю оценку:

$$\mathbb{E}_{\mathcal{Z}_m}\left[\hat{Q}_{m,m/2}(F, \mathcal{Z}_m)\right] \geq \frac{1}{2}\hat{R}^{td}_{m+u}(F, \mathcal{Z}_N, 1/4) + \frac{2B}{\sqrt{N}}.$$

Доказательство. Заметим, что $\mathbb{E}_{\mathcal{Z}_m}\left[\sup_{f\in F}\left(\bar{f}(\mathcal{Z}_u) - \bar{f}(\mathcal{Z}_m)\right)\right] = \hat{Q}_{N,m}(F, \mathcal{Z}_N)$. □

5.3. Трансдуктивные оценки риска

Далее мы используем результаты из раздела 5.2 чтобы получить новые трансдуктивные оценки риска. Рассмотрим свободную от распределения постановку трансдуктивного обучения, описанную ранее. Фиксируем произвольную последовательность пар $\mathcal{Z}_N = \{(x_i, y_i)\}_{i=1}^N \subseteq \mathcal{X} \times \mathcal{Y}$. На вход мы получаем размеченную обучающую выборку $\mathcal{Z}_m$ объектов, полученных равновероятно без возвращений из $\mathcal{Z}_N$. Оставшееся множество является тестовым $\mathcal{Z}_u = \mathcal{Z}_N \setminus \mathcal{Z}_m$ и дано нам без меток. Нашей целью будет построить классификатор из множества $\mathcal{H}$ на основании выборки $\mathcal{Z}_m$ и неразмеченных точек $\mathcal{X}_u$, такой что тестовая ошибка мала

для данной функции потерь $\ell\colon \mathcal{Y} \times \mathcal{Y} \to [0,1]$. Для $h \in \mathcal{H}$ и $(x,y) \in \mathcal{Z}_N$ обозначим $\ell_h(x,y) = \ell\big(h(x),y\big)$, а также обозначим класс потерь $L_{\mathcal{H}} = \{\ell_h\colon h \in \mathcal{H}\}$. Тогда тестовая и ошибка на обучении $h \in \mathcal{H}$ определяется как $\mathrm{err}_u(h) = \overline{\ell_h}(\mathcal{Z}_u)$ и $\mathrm{err}_m(h) := \overline{\ell_h}(\mathcal{Z}_m)$ соответственно.

Теорема 12 ([61]). *Если $m = u$, то с вероятностью не менее $1-\delta$ относительно выбора $\mathcal{Z}_m$ любая функция $h \in \mathcal{H}$ удовлетворяет:*

$$\mathrm{err}_u(h) \le \mathrm{err}_m(h) + \hat{R}^{td}_{m+u}(L_{\mathcal{H}}, \mathcal{Z}_N, 1/4) + 11\sqrt{\frac{2}{N}} + \sqrt{\frac{2N\log(1/\delta)}{(N-1/2)^2}}. \tag{5.7}$$

Используя результаты из раздела 5.2, мы получаем:

Теорема 13. *Если $m = u$ и $n \in \{1,\dots,m-1\}$, то с вероятностью не менее $1-\delta$ относительно выбора $\mathcal{Z}_m$ любая функция $h \in \mathcal{H}$ удовлетворяет:*

$$\mathrm{err}_u(h) \le \mathrm{err}_m(h) + \mathbb{E}_{\mathcal{S}_m}\left[\hat{Q}_{m,n}(L_{\mathcal{H}}, \mathcal{Z}_m)\right] + \sqrt{\frac{2N\log(1/\delta)}{(N-1/2)^2}}. \tag{5.8}$$

Более того, с вероятностью $1-\delta$ любая функция $h \in \mathcal{H}$ удовлетворяет:

$$\mathrm{err}_u(h) \le \mathrm{err}_m(h) + \hat{Q}_{m,n}(L_{\mathcal{H}}, \mathcal{Z}_m) + 2\sqrt{\frac{2N\log(2/\delta)}{(N-1/2)^2}}. \tag{5.9}$$

Доказательство будет дано ниже. Сравним результаты Теорем 13 и 12:

1. Заметим, что оценка (5.9) вычислима по выборке. Оценка (5.7) зависит от неизвестных меток тестовой выборки.

2. Отметим, что для бинарной функции потерь TPC (как и Радемахеровская сложность) не зависит от меток выборки. Однако, это неверно для ПРС, которая является чувствительной к меткам.

3. Еще одним применением концентрации можно упростить вычисление ПРС, вычисляя математическое ожидание в 12 с помощью единственного разбиения.

5.4. Доказательства

Доказетельство Теоремы 10

Лемма 15. *Для $0 < m \leq N$ пусть $\mathcal{S}_m = \{s_1, \ldots, s_m\}$ получены равновероятно без возвращений из конечного множества чисел $\mathcal{C} = \{c_1, \ldots, c_N\} \subset \mathbb{R}$. Тогда*

$$\mathbb{E}_{\mathcal{S}_m}\left[\frac{1}{m}\sum_{i=1}^{m} s_i\right] = \frac{1}{\binom{N}{m}}\sum_{\mathcal{S}_m \subseteq \mathcal{C}}\frac{1}{m}\sum_{z\in\mathcal{S}_m} z = \frac{1}{m\binom{N}{m}}\sum_{i=1}^{N}\binom{N-1}{m-1}c_i = \frac{1}{N}\sum_{i=1}^{N} c_i.$$

Теорема 10. Зафиксируем натуральные числа n и k, такие что $n + k = m$, что влечет $n < m$ and $k < m = u$. Заметим, что лемма 15 влечет:

$$\bar{f}(\mathcal{Z}_u) = \mathbb{E}_{\mathcal{S}_k}\left[\bar{f}(\mathcal{S}_k)\right], \quad \bar{f}(\mathcal{Z}_m) = \mathbb{E}_{\mathcal{S}_n}\left[\bar{f}(\mathcal{S}_n)\right],$$

где $\mathcal{S}_k$ и $\mathcal{S}_n$ получены равновероятно без возвращений из $\mathcal{Z}_u$ и $\mathcal{Z}_m$ соотвественно. Используя неравенство Йенсена, получаем

$$\begin{aligned}\mathbb{E}_{\mathcal{Z}_m}\left[\sup_{f\in F}\left(\bar{f}(\mathcal{Z}_u) - \bar{f}(\mathcal{Z}_m)\right)\right] &= \mathbb{E}_{\mathcal{Z}_m}\left[\sup_{f\in F}\left(\mathbb{E}_{\mathcal{S}_k}\left[\bar{f}(\mathcal{S}_k)\right] - \mathbb{E}_{\mathcal{S}_n}\left[\bar{f}(\mathcal{S}_n)\right]\right)\right] \\ &\leq \mathbb{E}_{(\mathcal{Z}_m,\mathcal{S}_k,\mathcal{S}_n)}\left[\sup_{f\in F}\left(\bar{f}(\mathcal{S}_k) - \bar{f}(\mathcal{S}_n)\right)\right]. \qquad (5.10)\end{aligned}$$

Легко видеть, что

$$\mathbb{E}_{(\mathcal{Z}_m,\mathcal{S}_k,\mathcal{S}_n)}\left[\sup_{f\in F}\left(\bar{f}(\mathcal{S}_k) - \bar{f}(\mathcal{S}_n)\right)\right] = \mathbb{E}_{\mathcal{Z}_m}\left[\mathbb{E}_{\mathcal{S}_n}\left[\sup_{f\in F}\left(\bar{f}(\mathcal{Z}_m \setminus \mathcal{S}_n) - \bar{f}(\mathcal{S}_n)\right)\middle|\mathcal{Z}_m\right]\right],$$

что доказывает верхнюю оценку. Мы показали, что для $n \in \{1, \ldots, m-1\}$ и $k = m - n$:

$$\mathbb{E}_{\mathcal{Z}_m}\left[\hat{Q}_{m,n}(F, \mathcal{Z}_m)\right] = \mathbb{E}_{(\mathcal{Z}_k,\mathcal{Z}_n)}\left[\sup_{f\in F}\left(\bar{f}(\mathcal{Z}_k) - \bar{f}(\mathcal{Z}_n)\right)\right], \qquad (5.11)$$

где $\mathcal{Z}_n$ и $\mathcal{Z}_k$ получены равновероятным выбором без возвращений из $\mathcal{Z}_N$ и $\mathcal{Z}_N \setminus \mathcal{Z}_n$ соответственно. Пусть $\mathcal{Z}_{m-n}$ получено равновероятным выбором без возвращений $\mathcal{Z}_N \setminus (\mathcal{Z}_n \cup \mathcal{Z}_k)$ и пусть $\mathcal{Z}_{u-k}$ — оставшиеся $u - k$ элементы $\mathcal{Z}_N$. Используя Лемму 15 мы получаем:

$$\mathbb{E}\left[\bar{f}(\mathcal{Z}_{m-n})\middle|(\mathcal{Z}_n, \mathcal{Z}_k)\right] = \mathbb{E}\left[\bar{f}(\mathcal{Z}_{u-k})\middle|(\mathcal{Z}_n, \mathcal{Z}_k)\right].$$

Правую часть выражения можно переписать (5.11) как:

$$\mathbb{E}_{(\mathcal{Z}_n,\mathcal{Z}_k)}\left[\sup_{f\in F}\left(\bar{f}(\mathcal{Z}_k)-\bar{f}(\mathcal{Z}_n)+\mathbb{E}\left[\bar{f}(\mathcal{Z}_{u-k})-\bar{f}(\mathcal{Z}_{m-n})\middle|(\mathcal{Z}_n,\mathcal{Z}_k)\right]\right)\right]$$
$$\leq \mathbb{E}\left[\sup_{f\in F}\left(\bar{f}(\mathcal{Z}_k)-\bar{f}(\mathcal{Z}_n)+\bar{f}(\mathcal{Z}_{u-k})-\bar{f}(\mathcal{Z}_{m-n})\right)\right],$$

где мы использовали неравенство Йенсена. Если мы возьмем $n^* = k^* = m/2$, то получим

$$\mathbb{E}_{\mathcal{Z}_m}\left[\hat{Q}_{m,m/2}(F,\mathcal{Z}_m)\right] \leq \mathbb{E}\left[\sup_{f\in F}\left(2\bar{f}(\mathcal{Z}_{k^*}\cup\mathcal{Z}_{u-k^*})-2\bar{f}(\mathcal{Z}_{n^*}\cup\mathcal{Z}_{m-n^*})\right)\right].$$

Остается заметить, что случайные подмножества $\mathcal{Z}_{k^*}\cup\mathcal{Z}_{u-k^*}$ и $\mathcal{Z}_{n^*}\cup\mathcal{Z}_{m-n^*}$ имеют такое же распределение как и $\mathcal{Z}_u$ и $\mathcal{Z}_m$. □

5.4.1. Доказательство Теоремы 11

[В этом доказательстве роль соавторов решающая.] Пусть $m = 2\cdot n$, $\varepsilon = \{\varepsilon_i\}_{i=1}^m$ независимые Радемахеровские случайные величины, и $\eta = \{\eta_i\}_{i=1}^m$ случайная перестановка, содержащая n плюсов и n минусов. Доказательство Теоремы 11 основано на кауплинге ε и η, что описано в Лемме 16. Рассмотрим бинарный куб $B_m = \{-1,+1\}^m$. Обозначим $S_m = \{v\in B_m\colon \sum_{i=1}^m v_i = 0\}$, что является подмножество B_m, в котором каждый элемент имеет одинаковое количество плюсов и минусов.

Для каждого $v\in B_m$ обозначим $\|v\|_1 = \sum_{i=1}^m |v_i|$ и рассмотрим множество

$$T(v) = \arg\min_{v'\in S_m}\|v-v'\|_1,$$

которое состоит из всех точек из S_m, ближайших к v по метрике Хэмминга. Для любого $v\in B_m$ пусть $t(v)$ является случайным элементом $T(v)$, выбранным равновероятно. Обозначим $t_i(v)$ — i-ую координату $t(v)$.

Если $v\in S_m$, то $T(v) = \{v\}$. Иначе $T(v)$ содержит более одного элемента S_m. Можно показать, что если для q выполнено $\sum_{i=1}^m v_i = q$, то q обязательно четное число и $T(v)$ состоит из всех векторов из S_m, которые можно получить заменой $q/2$ знаков $+1$ в v на -1. Таким образом в этом случае, $\operatorname{card}\big(T(v)\big) = \binom{(m+q)/2}{q/2}$.

Лемма 16 (Кауплинг)**.** *Пусть* $m = 2 \cdot n$*. Случайная последовательность* $t(\varepsilon)$ *имеет такое же распределение, что и* η*.*

Доказательство. Очевидно из симметрии. □

Лемма 17. *Пусть* $m = 2 \cdot n$*. Для всех* $q \in \{1, \dots, m\}$ *выполнено*

$$\mathbb{E}[\varepsilon_q | t(\varepsilon)] = \left(1 - 2^{-m}\binom{m}{n}\right) t_q(\varepsilon) \geq \left(1 - 2(2\pi m)^{-1/2}\right) t_q(\varepsilon).$$

Доказательство. Мы ограничим $P\{\varepsilon_q \neq t_q(\varepsilon) | t(\varepsilon) = \mathbf{e}\}$, где $\mathbf{e} = \{e_i\}_{i=1}^m$ последовательность из n единиц и n минус единиц.

$$\begin{aligned} P\{\varepsilon_q \neq t_q(\varepsilon) | t(\varepsilon) = \mathbf{e}\} &= \frac{P\{\varepsilon_q \neq t_q(\varepsilon) \cap t(\varepsilon) = \mathbf{e}\}}{P\{t(\varepsilon) = \mathbf{e}\}} \\ &= \binom{m}{n} P\{\varepsilon_q \neq t_q(\varepsilon) \cap t(\varepsilon) = \mathbf{e}\} \\ &= \binom{m}{n} 2^{-m} \sum_{\mathbf{s}} P\{\varepsilon_q \neq t_q(\varepsilon) \cap t(\varepsilon) = \mathbf{e} | \varepsilon = \mathbf{s}\}, \end{aligned} \tag{5.12}$$

где мы использовали Лемму 16. Для произвольного $\mathbf{s}$ обозначим $S(\mathbf{s}) = \sum_{j=1}^n s_j$ и рассмотрим члены в (5.12), соотвествующие $\mathbf{s}$ с $S(\mathbf{s}) = 0$, $S(\mathbf{s}) > 0$ и $S(\mathbf{s}) < 0$. Рассмотрим случаи

Случай 1: $S(\mathbf{s}) = 0$. Эти члены равны нулю так как $t(\mathbf{s}) = \mathbf{s}$.

Случай 2: $S(\mathbf{s}) > 0$. Если $s_q = -1$, то $t_q(\mathbf{s}) = s_q$ и соответствующие члены равны нулю. Если $s_q = 1$ и одновременно $e_q = 1$, то событие $\{\varepsilon_q \neq t_q(\varepsilon) \cap t(\varepsilon) = \mathbf{e}\}$ не может быть выполнено. Более того равенство $\mathbf{e} = t(\mathbf{s})$ выполнено только если $\mathbf{e} \in T(\mathbf{s})$, которое с необходимостью влечет

$$\big\{j \in \{1, \dots, m\} \colon s_j = -1\big\} \subseteq \big\{j \in \{1, \dots, m\} \colon e_j = -1\big\}. \tag{5.13}$$

Отсюда мы заключаем, что если $q \in \{1, \dots, n\}$, то все члены, относящиеся к $\mathbf{s}$ с $S(\mathbf{s}) > 0$ нулевые. Мы будем использовать $U_q(\mathbf{e})$ чтобы обозначить подмножество B_m, состоящее из последовательностей $\mathbf{s}$, таких что, во-первых, $S(\mathbf{s}) > 0$ и $s_q = 1$, а, во-вторых, для которых условие (5.13) выполнено. Если $\mathbf{s} \in U_q(\mathbf{e})$, то:

$$P\{\varepsilon_q \neq t_q(\varepsilon) \cap t(\varepsilon) = \mathbf{e} | \varepsilon = \mathbf{s}\} = \frac{1}{\binom{n + S(\mathbf{s})/2}{S(\mathbf{s})/2}}.$$

Это условие выполнено, так как $t(\varepsilon)$ принимает в точности одно из $\binom{n+S(\mathbf{s})/2}{S(\mathbf{s})/2}$ различных значений, когда всего одно из них равно $\mathbf{e}$.

Рассчитаем мощность множества $U_q(\mathbf{e})$ для $q \in \{n+1, \ldots, m\}$. Легко видеть, что условие $S(\mathbf{s}) = 2j$ для некоторого неотрицательного j влечет, что $\mathbf{s}$ имеет в точности $n - j$ знаков минус. Учитывая $s_q = 1$ для $\mathbf{s} \in U_q(\mathbf{e})$ мы получаем:

$$\left|U_q(\mathbf{e})\right| = \binom{n-1}{n-j}.$$

Комбинируя все вместе, мы получаем

$$\sum_{\mathbf{s}\colon S(\mathbf{s})>0} P\{\varepsilon_q \neq t_q(\varepsilon) \cap t(\varepsilon) = \mathbf{e} | \varepsilon = \mathbf{s}\} = \mathbb{1}\{q > n\} \sum_{j=1}^{n} \frac{\binom{n-1}{n-j}}{\binom{n+j}{j}}.$$

Далее легко доказать, что

$$\sum_{j=1}^{n} \frac{\binom{n-1}{n-j}}{\binom{n+j}{j}} = \frac{1}{2}.$$

Случай 3: $S(\mathbf{s}) < 0$. Также легко показать, что

$$\sum_{\mathbf{s}\colon S(\mathbf{s})<0} P\{\varepsilon_q \neq t_q(\varepsilon) \cap t(\varepsilon) = \mathbf{e} | \varepsilon = \mathbf{s}\} = \frac{1}{2}\mathbb{1}\{q \leq n\}.$$

Далее мы заключаем, что

$$P\{\varepsilon_q \neq t_q(\varepsilon) | t(\varepsilon) = \mathbf{e}\} = \frac{1}{2}\binom{m}{n} 2^{-m} \leq \frac{1}{\sqrt{2\pi m}},$$

где мы используем верхнюю оценку на биномиальный коэффициент из работы [67]. В итоге

$$\mathbb{E}[\varepsilon_q | t(\varepsilon)] = t_q(\varepsilon)\left(1 - 2P\{\varepsilon_q \neq t_q(\varepsilon) | t(\varepsilon)\}\right) \geq t_q(\varepsilon)\left(1 - 2(2\pi m)^{-1/2}\right).$$

□

Теорема 11. Сначала мы докажем (5.3). Пусть $\mathcal{Z}_m = \{z_1, \ldots, z_m\}$. Далее

$$\hat{Q}_{m,n}(F) = \mathbb{E}\left[\sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} t_i(\varepsilon) f(z_i)\right] \tag{5.14}$$

$$\leq \left(1 - 2(2\pi m)^{-1/2}\right)^{-1} \mathbb{E}\left[\sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} \mathbb{E}[\varepsilon_i | t(\varepsilon)] f(z_i)\right] \tag{5.15}$$

$$\leq \left(1 + \frac{2}{\sqrt{2\pi m} - 2}\right) \mathbb{E}\left[\sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} \varepsilon_i f(z_i)\right], \tag{5.16}$$

где мы используем Лемму 16 в (5.14), Лемму 17 в (5.15) и неравенство Йенсена в (5.16). Это завершает доказательство (5.3).

Далее мы докажем (5.3). Запишем

$$\left|\hat{Q}_{m,n}(F) - \hat{R}_m(F)\right| = \left|\mathbb{E}_\eta\left[\sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} \eta_i f(z_i)\right] - \mathbb{E}_\varepsilon\left[\sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} \varepsilon_i f(z_i)\right]\right|.$$

Используя Лемму 16 и неравенство Йенсена, мы получаем

$$\begin{aligned}
&\left|\hat{Q}_{m,n}(F) - \hat{R}_m(F)\right| \\
&= \left|\mathbb{E}_\varepsilon\left[\mathbb{E}_t\left[\sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} t_i(\varepsilon) f(z_i) \middle| \varepsilon\right]\right] - \mathbb{E}_\varepsilon\left[\sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} \varepsilon_i f(z_i)\right]\right| \\
&\leq \mathbb{E}_\varepsilon\left[\mathbb{E}_t\left[\left|\sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} t_i(\varepsilon) f(z_i) - \sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} \varepsilon_i f(z_i)\right| \middle| \varepsilon\right]\right].
\end{aligned} \tag{5.17}$$

Далее

$$\left|\sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} t_i(\varepsilon) f(z_i) - \sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} \varepsilon_i f(z_i)\right| \leq \left|\sup_{f\in F} \frac{4}{m}\sum_{i\in S(\varepsilon,t)} \varepsilon_i f(z_i)\right|, \tag{5.18}$$

где $S(\varepsilon, t) \subseteq \{1, \ldots, m\}$ является подмножеством индексов, таких что $\big(t(\varepsilon)\big)_i \neq \varepsilon_i$ тогда и только тогда, когда $i \in S(\varepsilon, t)$. Далее

$$\left|\sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} t_i(\varepsilon) f(z_i) - \sup_{f\in F} \frac{2}{m}\sum_{i=1}^{m} \varepsilon_i f(z_i)\right| \leq \frac{4}{m}\sup_{f\in F}\sum_{i\in S(\varepsilon,t)} |f(z_i)|. \tag{5.19}$$

Заметим, что так как функции в F ограничены B, то

$$\sup_{f\in F}\sum_{i\in S(\varepsilon,t)} |f(z_i)| \leq B \cdot |S(\varepsilon, t)|.$$

Возвращаясь к (5.17) и используя 5.4.1, мы получаем

$$\left|\hat{Q}_{m,n}(F) - 2\hat{R}_m(F)\right| \le \frac{4B}{m}\mathbb{E}_\varepsilon\left[\mathbb{E}_t\left[|S(\varepsilon,t)|\ \ |\varepsilon\right]\right] = \mathbb{E}_\varepsilon\left[\frac{1}{2}\left|\sum_{i=1}^m \varepsilon_i\right|\right].$$

Неравенство Хинчина [24] дает $\mathbb{E}_\varepsilon\left[\left|\sum_{i=1}^m \varepsilon_i\right|\right] \le \sqrt{m}$, что завершает доказательство (5.4). □

5.4.2. Доказательство Леммы 5.5

Доказательство. Пусть $\mathcal{Z}_m = \{z_1, \ldots, z_m\}$. Выберем F'_m как множество, состоящее из двух константных функций: $f_1(z) = 1$ и $f_2(z) = 0$ для всех $z \in \mathcal{Z}$. Очевидно, что $\hat{Q}_{m,n}(F'_m) = 0$. Одновременно

$$\mathbb{E}_\varepsilon\left[\sup_{f\in F'_m}\frac{2}{m}\sum_{i=1}^m \varepsilon_i f(z_i)\right] = \mathbb{E}_\varepsilon\left[\max\left\{0, \frac{2}{m}\sum_{i=1}^m \varepsilon_i\right\}\right] \le \mathbb{E}_\varepsilon\left[\left|\frac{2}{m}\sum_{i=1}^m \varepsilon_i\right|\right] \le \frac{2}{\sqrt{m}},$$

Далее с помощью неравенства Хинчина получаем

$$\mathbb{E}_\varepsilon\left[\max\left\{0, \frac{2}{m}\sum_{i=1}^m \varepsilon_i\right\}\right] = \frac{1}{2}\mathbb{E}_\varepsilon\left[\left|\frac{2}{m}\sum_{i=1}^m \varepsilon_i\right|\right] \ge \frac{1}{\sqrt{2m}}.$$

Далее пусть F''_m содержит $\binom{m}{m/2}$ функций, таких что их проекции на $\mathcal{Z}_m$ воспроизводят все возможные функции, принимающие значение 0 ровно в $m/2$ точка и 1 в оставшихся. Очевидно, что в этом случае $\hat{Q}_{m,n}(F''_m) = 1$. При этом легко показать, что $\hat{R}_m(F''_m) = 1 - 2^{-m}\binom{m}{n}$. □

5.4.3. Доказательство Теоремы 13

Следующая версия неравенства ограниченных разностей представлена в работе [61] и улучшена в работе [62]:

Теорема 14 ([61], [62]). *Пусть $\mathcal{Z}_m$ выбраны равновероятно без возвращений из фиксированного множества $\mathcal{Z}_{m+u} \subseteq \mathcal{Z}$, имеющего $m+u$ элементов. Пусть $g\colon \mathcal{Z}^m \to \mathbb{R}$ функцией, такой что для всех $i = 1, \ldots, m$ и всех $z_1, \ldots, z_m \in \mathcal{Z}$ и $z'_1, \ldots, z'_m \in \mathcal{Z}$,*

$$\left|g(z_1, \ldots, z_m) - g(z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_m)\right| \le c. \tag{5.20}$$

Тогда, если $m = u$, то с вероятностью не менее $1 - \delta$ выполнено

$$g \leq \mathbb{E}[g] + \sqrt{\frac{c^2 N^3 \log(1/\delta)}{8(N - 1/2)^2}}.$$

Заметим, что функция $\sup_{h \in \mathcal{H}} (\mathrm{err}_h(\mathcal{Z}_u) - \mathrm{err}_h(\mathcal{Z}_m))$ отображает $(\mathcal{X} \times \mathcal{Y})^m$ в $\mathbb{R}$. Легко показать, что эта функция удовлетворяет условиям ограниченной разности (5.20) с $c = \frac{1}{m} + \frac{1}{u}$. Тогда с вероятностью $1 - \delta$:

$$\sup_{h \in \mathcal{H}} (\mathrm{err}_u(h) - \mathrm{err}_m(h)) \leq \mathbb{E}_{\mathcal{S}_m} \left[\sup_{h \in \mathcal{H}} (\mathrm{err}_u(h) - \mathrm{err}_m(h)) \right] + \sqrt{\frac{2N \log(1/\delta)}{(N - 1/2)^2}}. \quad (5.21)$$

Используя верхнюю оценку Теоремы 10 с $L_{\mathcal{H}}$ вместо F мы завершаем доказательство (5.8). Рассмотрим функцию $\hat{Q}_{m,n}(L_{\mathcal{H}}, \mathcal{Z}_m)$, которая отображает $(\mathcal{X} \times \mathcal{Y})^m$ в $\mathbb{R}$. Можно показать, что она удовлетворяет условию (5.20) с $c = \frac{2}{m}$. Тогда с вероятностью не менее $1 - \delta$:

$$\mathbb{E}_{\mathcal{S}_m} \left[\hat{Q}_{m,n}(L_{\mathcal{H}}, \mathcal{Z}_m) \right] \leq \hat{Q}_{m,n}(L_{\mathcal{H}}, \mathcal{Z}_m) + \sqrt{\frac{2N \log(1/\delta)}{(N - 1/2)^2}}. \quad (5.22)$$

Используя это неравенство вместе с (5.8) получаем второе утверждение Теоремы.

Улучшенный вариант Леммы о масксимальном несоотвествии из [66]

Пусть μ — вероятностное распределение на $\mathcal{Z}$ и $\mathcal{X}_m = \{X_1, \ldots, X_m\}$ выбраны независимо согласно μ. Максимальное несоответствие (Maximal discrepancy) F определяется в [66] как

$$\hat{D}_m(F, \mathcal{X}_m) = \sup_{f \in F} \left(\frac{2}{m} \sum_{i=1}^{m/2} f(X_i) - \frac{2}{m} \sum_{i=m/2+1}^{m} f(X_i) \right).$$

В работе [66] показано, что если функции F равномерно ограничены 1, то

$$\frac{1}{2} \mathbb{E} \left[\hat{R}_m(F, \mathcal{X}_m) \right] - 2\sqrt{\frac{2}{m}} \leq \mathbb{E} \left[\hat{D}_m(F, \mathcal{X}_m) \right] \leq \mathbb{E} \left[\hat{R}_m(F, \mathcal{X}_m) \right] + 4\sqrt{\frac{2}{m}}. \quad (5.23)$$

Так как элементы в $\mathcal{X}_m$ независимы и распределение $\hat{D}_m$ инвариантно относительно перестановок, то $\mathbb{E} \left[\hat{D}_m(F, \mathcal{X}_m) \right] = \mathbb{E} \left[\hat{Q}_{m,m/2}(F, \mathcal{X}_m) \right]$. Используем Теорему 11

чтобы улучшить оценку (5.23):

$$\mathbb{E}\left[\hat{R}_m(F, \mathcal{X}_m)\right] - \frac{2}{\sqrt{m}} \leq \mathbb{E}\left[\hat{D}_m(F, \mathcal{X}_m)\right] \leq \left(1 + \frac{2}{\sqrt{2\pi m} - 2}\right) \mathbb{E}\left[\hat{R}_m(F, \mathcal{X}_m)\right].$$

Заключение

1. Получены новые односторонние оценки для равномерных относительных законов больших чисел, выраженные в терминах локальной скобочной энтропии и локальной эмпирической энтропии.

2. Полученные общие оценки применены в задаче бинарной классификации при условии шума Массара. С помощью них получены общие верхние оценки риска, для которых в тексте диссертации также доказаны совпадающие с точностью до констант нижние оценки. Показаны примеры неоптимальности ранее получавшихся верхних и нижних оценок.

3. Для задачи линейной классификации с двумя классами впервые в PAC постановке получен практически минимаксно оптимальный результат для полиномиального алгоритма обучения. Оптимальные результаты для минимизатора эмпирического риска получены при условии лог-вогнутых распределений данных, а также при условиях конечных локальных энтропий в задачах непараметрической регрессии.

4. Для задач трансдуктивного обучения введена новая мера сложности, названная перестановочной Радемахеровской сложностью. Показано, что оценки, основанные на ней, улучшают существующие оценки для трансдуктивной постановки. Предложены верхние оценки для вводимой меры сложности.

Список литературы

1. Vapnik V., Chervonenkis A. Theory of Pattern Recognition. Nauka, Moscow, 1974.
2. Anthony M., Bartlett P. L. Neural Network Learning: Theoretical Foundations. Cambridge University Press, 1999.
3. S. S.-S., S. B.-D. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 2014.
4. Bartlett P. L., Bousquet O., Mendelson S. Local Rademacher Complexities // The Annals of Statistics. 2005. Vol. 33(4). P. 1497–1537.
5. Massart P., Nédélec E. Risk bounds for statistical learning // Annals of Statistics. 2006.
6. Tsybakov A. B. Optimal aggregation of classifiers in statistical learning // The Annals of Statistics. 2004. Vol. 32, no. 1. P. 135–166.
7. Koltchinskii V. Oracle inequalities in empirical risk minimization and sparse recovery problems. Springer, New York, 2011.
8. Giné E., Koltchinskii V. Concentration inequalities and asymptotic results for ratio type empirical processes // The Annals of Probability. 2006. Vol. 34(3). P. 1143–1216.
9. Vapnik. V. Statistical Learning Theory. John Wiley & Sons, 1998.
10. Derbeko, P. E.-Y., R., Meir R. Explicit learning curves for transduction and application to clustering and compression algorithms // Journal of Artificial Intelligence Research. 2004. Vol. 22(1). P. 117–142.
11. C. Cortes M. M. On transductive regression // NIPS 2006. 2007. P. 305–312.
12. Raginsky M., Rakhlin A. Lower Bounds for Passive and Active Learning // Advances in Neural Information Processing Systems 24. 2011.
13. Lecue G., Mendelson S. Learning subgaussian classes: Upper and minimax bounds // `http://arxiv.org/abs/1305.4825`. 2013.
14. Mendelson S. 'Local' vs. 'global' parameters – breaking the Gaussian complexity barrier // Annals of statistics. 2017.
15. Haussler D., Littlestone N., Warmuth M. Predicting $\{0,1\}$-functions on randomly drawn points // Information and Computation. 1994. Vol. 115. P. 248–292.
16. Floyd S., Warmuth M. Sample Compression, learnability, and the Vapnik Chervonenkis Dimension // Machine Learning. 1995. Vol. 21. P. 269–304.
17. Simon H. An almost optimal PAC-algorithm // Proceedings of The 28th Confer-

ence on Learning Theory. 2015. P. 1552–1563.
18. Auer P., Ortner R. A new PAC bound for intersection-closed concept classes // Machine Learning. 2007. Vol. 66. P. 151–163.
19. Zhivotovskiy N. Optimal learning via local entropies and sample compression // Conference on Learning Theory, Proceedings of Machine Learning Research (formerly JMLR WCP). 2017. Vol. 65. P. 1–23.
20. Zhivotovskiy N., Hanneke S. Localization of VC Classes: Beyond Local Rademacher Complexities // Algorithmic Learning Theory, Lecture Notes in Computer Science, Springer. 2016. Vol. 9925. P. 18–33.
21. I.Tolstikhin, N.Zhivotovskiy, G.Blanchard. Permutational Rademacher Complexity: a New Complexity Measure for Transductive Learning // In Algorithmic Learning Theory, Lecture Notes in Computer Science. 2015. Vol. 9355. P. 209–223.
22. Н. Животовский. Комбинаторные оценки переобучения с сублогарифмическим темпом роста // Труды МФТИ. 2015. Т. 7, № 3. С. 42 – 54.
23. Boucheron S., Bousquet O., Lugosi G. Theory of classification: a survey of recent advances // ESAIM: Probability and Statistics. 2005. P. 323–375.
24. Ledoux M., Talagrand M. Probability in Banach Space. Springer-Verlag, 1991.
25. Koltchinskii V. Local Rademacher complexities and oracle inequalities in risk minimization // Annals of Statistics. 2006. Vol. 34(6). P. 2593–2656.
26. Vapnik V., Chervonenkis A. On the uniform convergence of relative frequencies of events to their probabilities // Proc. USSR Acad. Sci. 1968. Vol. 181, no. 4. P. 781–783.
27. Edelsbrunner H. Algorithms in Combinatorial Geometry. Springer, Berlin, 1987.
28. Talagrand M. Sharper bounds for Gaussian and empirical processes // The Annals of Probability. 1994. Vol. 22. P. 28–76.
29. Haussler D. Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension // J. Comb. Theory Ser. A. 1995. Vol. 69(2). P. 217–232.
30. Hanneke S. Refined error bounds for several learning algorithms // Journal of Machine Learning Research. 2016. Vol. 17 (135). P. 1–55.
31. van Erven T., Grünwald P., Mehta N., M. Reid R. W. Fast rates in statistical and online learning // Journal of Machine Learning Research. 2015. Vol. 16. P. 1793–1861.
32. Alexander K. S. Rates of growth and sample moduli for weighted empirical pro-

cesses indexed by sets // Probability Theory and Related Fields. 1987. no. 75. P. 379–423.
33. Hanneke S., Yang L. Minimax analysis of active learning // Journal of Machine Learning Research. 2015. Vol. 16 (12). P. 3487–3602.
34. Devroye L., Györfi L., Lugosi G. A Probabilistic Theory of Pattern Recognition. Springer–Verlag, New York, 1996.
35. Hanneke S. A bound on the label complexity of agnostic active learning // In Proceedings of the 24th Annual International Conference on Machine Learning. 2007.
36. Hanneke S. Theory of Disagreement-Based Active Learning // Foundations and Trends in Machine Learning. 2014. Vol. 7 (2-3). P. 131–309.
37. Liang T., Rakhlin A., Sridharan K. Learning with square loss: Localization through offset Rademacher complexity // Proceedings of The 28th Conference on Learning Theory. 2015.
38. Yang Y., Barron A. Information-theoretic determination of minimax rates of convergence // Annals of Statistics. 1999. Vol. 27. P. 1564–1599.
39. Devroye L., Lugosi G. Combinatorial Methods in Density Estimation. Springer, New York, 2001.
40. Vidyasagar M. Learning and Generalization with Applications to Neural Networks. Springer-Verlag, 2003.
41. Massart P. Concentration Inequalities and Model Selection. Springer, New York, 2003.
42. Hanneke S. The optimal sample complexity of PAC learning // Journal of Machine Learning Research. 2016. Vol. 17 (38). P. 1–15.
43. Ehrenfeucht A., Haussler D., Kearns M., Valiant L. A general lower bound on the number of examples needed for learning // Information and Computation. 1989. Vol. 82(3). P. 247–261.
44. Cam L. M. L. Convergence of estimates under dimensionality restrictions // Ann. Statist. 1973. Vol. 1. P. 38–53.
45. van de Geer S., Wegkamp M. Consistency for the least squares estimator in nonparametric regression // Annals of Statistics. 1996. Vol. 24, no. 6. P. 2513–2523.
46. Rakhlin A., Sridharan K., Tsybakov A. B. Empirical entropy, minimax regret and minimax risk // Bernoulli. 2017.
47. Bshouty N. H., Li Y., Long P. M. Using the doubling dimension to analyze the

generalization of learning algorithms // Journal of Computer and System Sciences. 2009.
48. Talagrand M. Upper and lower bounds for stochastic processes. Springer, Berlin, 2014.
49. Dudley R. Empirical processes. In Ecole de Probabilité de St. Flour 1982. Lecture Notes in Mathematics 1097, Springer Verlag, New York, 1984.
50. Boucheron S., Lugosi G., Massart P. Concentration inequalities: A nonasymptotic theory of independence. Cambridge, 2013.
51. Balcan M., Long P. M. Active and passive learning of linear separators under log-concave distributions // In Proceedings of the 26th Conference on Learning Theory. 2013.
52. Bartlett P. L., Mendelson S. Empirical minimization // Probability Theory Related Fields. 2006. Vol. 135(3). P. 311–334.
53. Lecué G. Interplay between concentration, complexity and geometry in learning theory with applications to high dimensional data analysis // Habilitation thesis, Université Paris-Est. 2011.
54. Adamczak R. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains // Electron. J. Probab. 2008. P. 1000–1034.
55. Lecué G., Mitchell C. Oracle inequalities for cross-validation type procedures // Electronic Journal of Statistics. 2012. Vol. 6. P. 1803–1837.
56. Van der Vaart A. W., Wellner J. Weak Convergence and Empirical Processes: With Applications to Statistics. Springer, 2000.
57. E. Gassiat R. v. H. The local geometry of finite mixtures // Trans. Amer. Math. Soc. 2014. Vol. 366. P. 1047–1072.
58. Adams T. M., Nobel A. B. Uniform approximation and bracketing properties of VC classes // Bernoulli. 2012. Vol. 18. P. 1310–1319.
59. Long. P. M. On the sample complexity of PAC learning halfspaces against the uniform distribution // IEEE Transactions on Neural Networks. 1995. Vol. 6(6). P. 1556–1559.
60. Littlestone N. From On-line to batch learning // In COLT. 1989.
61. El-Yaniv R., Pechyony D. Transductive rademacher complexity and its applications // Journal of Artificial Intelligence Research. 2009. Vol. 35(1). P. 193–234.
62. Cortes C., Mohri M., Pechyony D., Rastogi A. Stability analysis and learning bounds for transductive regression algorithms // CoRR **abs/0904.0814**. 2009.

63. Tolstikhin I., Blanchard G., Kloft M. Localized complexities for transductive learning // COLT 2014. 2014. P. 857–884.
64. Mendelson S. Learning without Concentration // Journal of ACM. 2015.
65. Magdon-Ismail M. Permutation complexity bound on out-sample error // Advances in Neural Information Processing Systems (NIPS 2010). 2010. P. 1531–1539.
66. Bartlett P., Mendelson S. Rademacher and Gaussian complexities: Risk bounds and structural results // Journal of Machine Learning Research. 2001. no. 3. P. 463–482.
67. Stanica P. Good lower and upper bounds on binomial coefficients // Journal of Inequalities in Pure and Applied Mathematics. 2001. Vol. 2(3).