

На правах рукописи

ГРАБОВОЙ Андрей Валерьевич

О сложности моделей и данных
в современных моделях глубокого обучения

1.2.1 — Искусственный интеллект и машинное обучение

Диссертация на соискание ученой степени
доктора физико-математических наук

Научный консультант:
профессор РАН К. В. Воронцов

Москва — 2025

Оглавление

	Стр.
Введение	5
Глава 1. Теория обучаемости моделей машинного обучения	15
1.1. Основные понятия статистического обучения	15
1.2. Классические меры сложности и их ограничения	17
1.2.1. VC-размерность и ростовая функция	19
1.2.2. PAC-обучаемость: реализуемый и агностический случаи	20
1.2.3. Радемахеровская сложность и локальные оценки	21
1.2.4. Комбинаторный подход Воронцова к снижению оценок Вапника–Червоненкиса	22
1.2.5. Общие ограничения классических мер сложности	24
1.3. Современные подходы к анализу сложности нейросетей	25
1.3.1. Анализ ландшафта функции потерь и матрицы Гессе	27
1.4. Эмпирические оценки и методы снижения сложности моделей	29
1.4.1. Эмпирические законы масштабирования	29
1.4.2. Методы снижения сложности моделей: дистилляция и регуляризация	30
1.5. Проблемы существующих подходов и вклад диссертации	31
Глава 2. Сложность моделей и данных	34
2.1. Оценка сложности моделей и данных	35
2.2. Достаточный объем выборки как мера сложности данных	40
2.3. Сходимость ландшафта оптимизационной задачи как мера сложности модели	42
2.3.1. Полносвязная нейросетевая модель глубокого обучения	48
2.3.2. Сверточные модели глубокого обучения	52
2.3.3. Трансформер модели глубокого обучения	54
2.4. Результаты вычислительных экспериментов	57
2.4.1. Полносвязная нейросетевая модель глубокого обучения	57
2.4.2. Сверточные модели глубокого обучения	62
2.4.3. Трансформер модели глубокого обучения	63
2.5. Заключение по главе	63

Глава 3. Матрицы Гессе нейросетевых моделей глубокого обучения	70
3.1. Полносвязная нейросетевая модель глубокого обучения	72
3.1.1. Спектральная оценка матрицы Гессе	75
3.2. Матричные модели глубокого обучения	82
3.2.1. Матричная факторизация матрицы Гессе	84
3.2.2. Оценка спектральных норм матрицы Гессе	86
3.3. Матрица Гессе для трансформерной модели глубокого обучения . . .	97
3.3.1. Матрица Гессе для слоя самовнимания	98
3.3.2. Матрица Гессе для LayerNorm слоя	106
3.3.3. Матрица Гессе для нелинейности ReLU	112
3.3.4. Матрица Гессе для трансформера	114
3.3.5. Спектральные оценки матрицы Гессе для трансформера	128
3.4. Результаты вычислительных экспериментов	132
3.5. Заключение по главе	137
 Глава 4. Достаточный объем выборки моделей	140
4.1. Статистические методы определения достаточного размера выборки	143
4.2. Эвристические методы определения достаточного размера выборки .	147
4.3. Байесовские методы определения достаточного размера выборки .	149
4.4. Метод определения достаточного размера выборки на основе сэмплирования эмпирической функции ошибки	150
4.5. Метод определения достаточного размера выборки на основе близости апостериорных распределений	154
4.6. Результаты вычислительных экспериментов	164
4.6.1. Определения достаточного размера выборки на основе статистических методов	164
4.6.2. Определение достаточного размера выборки на основе сэмплирования эмпирической функции ошибки	169
4.6.3. Определение достаточного размера выборки на основе близости апостериорных распределений	172
4.7. Заключение по главе	177
 Глава 5. Методы снижения сложности моделей глубокого обучения	181
5.1. Удаление параметров моделей глубокого обучения	181
5.2. Дистилляция моделей глубокого обучения на многодоменных данных	185
5.3. Антидистилляция моделей глубокого обучения	188

5.4. Результаты вычислительных экспериментов	191
5.4.1. Удаление параметров моделей глубокого обучения	191
5.4.2. Дистилляция моделей глубокого обучения на многодоменных данных	196
5.4.3. Антидистилляция моделей глубокого обучения	204
5.5. Заключение по главе	208
 Глава 6. Применение теоретических оценок в прикладных задачах	211
6.1. Сложность моделей в многозадачном обучении	212
6.1.1. Радемахеровская сложность моделей глубокого обучения	212
6.1.2. Радемахеровская сложность многозадачного обучения в LoRA-адаптерах	216
6.2. Снижение сложности данных в задаче декодирования фМРТ-снимков	222
6.3. Качество данных в задаче детекции машинно-генерированного контента	226
6.4. Результаты вычислительных экспериментов	229
6.4.1. Сложность моделей в многозадачном обучении	229
6.4.2. Снижение сложности данных в задаче декодирования фМРТ-снимков	236
6.4.3. Качество данных в задаче детекции машинно-генерированного контента	240
6.5. Заключение по главе	245
Заключение	248
Общие свойства и определения	250
Дополнительные Леммы и утверждения	254
Список иллюстраций	258
Список таблиц	267
Список литературы	272

Введение

Диссертационная работа решает проблему отсутствия строгого теоретического аппарата для оценки сложности моделей и данных в моделях глубокого обучения, что препятствует рациональному проектированию архитектур и эффективному использованию вычислительных ресурсов при масштабировании моделей. В работе разработан новый теоретический подход, основанный на анализе ландшафта оптимизационной задачи, который обеспечивает получение асимптотических оценок сложности различных семейств моделей глубокого обучения и устанавливает формальную связь между сложностью модели и сложностью данных, необходимых для ее обучения.

Актуальность темы. С начала XXI века наблюдается экспоненциальный рост производительности вычислительных систем, измеряемой в операциях с плавающей запятой (англ. FLOPS): от терафлопсов в начале столетия до экзафлопсов в настоящий момент. Параллельно происходит сопоставимый рост сложности моделей глубокого обучения, выраженный в увеличении количества обучаемых параметров на несколько порядков — от тысяч в начале нулевых годов до миллиардов и сотен миллиардов в двадцатых, с прогнозируемым переходом к триллионам параметров в ближайшем десятилетии. Современные исследования анализа сложности таких моделей преимущественно опираются на эмпирические корреляции, связывающие эффективность моделей с количественными метриками — числом параметров и вычислительными затратами на обучение или применение. Отсутствие строгой теоретической основы для предсказания поведения моделей при масштабировании делает процесс разработки экономически и энергетически неэффективным и зачастую приводит к получению необоснованных и противоречивых результатов.

Особую остроту проблема приобретает при разработке больших языковых моделей (англ. LLM), обучение которых требует значительных вычислительных, энергетических и финансовых ресурсов. Для снижения непредсказуемости процесса обучения на предварительной стадии (англ. pretrain) эмпирически подбираются оптимальные соотношения между размером модели в параметрах и объемом обучающих данных в токенах при заданном вычислительном ресурсе. Однако эмпирические оценки, полученные для одной архитектуры, не переносятся на другие модели, что делает подобные подходы несостоятельными и приводит к непредвиденным результатам при полномасштабном обучении. В связи с этим разработка теоретических оценок сложности моделей становится

критически важной, поскольку такие оценки позволяют связать сложность модели со сложностью выборки еще на этапе проектирования архитектуры, обеспечивая рациональный выбор параметров модели и объема данных. Отметим, что глубоким теоретическим анализом выбора и порождения моделей на этапе проектирования архитектур моделей глубокого обучения занимается Вадим Викторович Стрижов [1].

Классические подходы к оценке сложности моделей машинного обучения, разработанные для традиционных методов, находят ограниченное применение при анализе моделей глубокого обучения. Фундаментальный вклад в теорию оценивания сложности моделей внесли работы Владимира Наумовича Вапника и Алексея Яковлевича Червоненкиса, заложившие основы статистической теории обучения [2], а также Лесли Вэлианта, предложившего подход приближенно правильного обучения (англ. PAC-Learning) [3]. Современные подходы к определению сложности основаны на Радемахеровской сложности, предложенной Владимиром Кольчинским и Дмитрием Панченко [4], и комбинаторном подходе к оценке обучаемости алгоритмов Константина Вячеславовича Воронцова [5], в котором разработан математический аппарат на основе комбинаторных оценок вероятности переобучения моделей. Однако современные исследования в области теории машинного обучения редко рассматривают нейросетевые модели глубокого обучения ввиду их высокой сложности и трудности получения адекватных оценок сложности таких моделей. Значительная часть современных исследований на ведущих конференциях посвящена классическим методам машинного обучения и улучшению оценок для известных методов, так как текущие оценки сложности даже для классических моделей являются сильно завышенными.

Отдельным направлением в исследованиях являются оценки репрезентативной способности моделей глубокого обучения к аппроксимации по прецедентам. Основополагающий результат принадлежит Джорджу Цибенко [6], который доказал, что нейронные сети аппроксимируют непрерывные функции с некоторыми ограничениями сколь угодно большим качеством. Йохану Хестад [7] принадлежат первые оценки, указывающие на рост аппроксимирующей способности моделей глубокого обучения с глубиной модели. В более современных работах Яна Лекуна [8], Йошуа Бенджио [9], Надава Коэна [10] эти оценки получены с более мягкими ограничениями на класс рассматриваемых функций. В целом все эти оценки наряду с работой Охада Шамира [11] указывают на экспоненциальное увеличение аппроксимирующей способности моделей глубокого обучения с

ростом числа слоев.

В настоящей работе предлагается теоретический анализ нейросетевых моделей, опирающийся на анализ их матриц Гессе. Матрица Гессе функции потерь по параметрам модели содержит информацию о локальной кривизне оптимизационного ландшафта и используется для анализа сложности моделей и данных, а также для оценки важности параметров в задачах прореживания нейросетевых моделей.

Степень разработанности темы диссертационного исследования. Классические подходы к оценке сложности моделей машинного обучения, основанные на VC-размерности, PAC-обучаемости и радемахеровской сложности, разработаны для моделей с ограниченным числом параметров и ориентированы на worst-case анализ, что делает их оценки слишком консервативными для пере-параметризованных нейронных сетей [12]. Эти меры не учитывают специфику архитектур глубокого обучения, такие как сверточные фильтры, остаточные связи и механизмы внимания, а также не отражают влияние регуляризации, ранней остановки и особенностей процесса оптимизации на обобщающую способность моделей. Комбинаторный подход К.В. Воронцова [5] позволяет снизить оценки Вапника–Червоненкиса, но остается применимым лишь к ограниченным классам моделей.

Что касается оценки сложности данных, существующие подходы преимущественно сводятся к анализу размера выборки для обучения [13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26], что является неполным, так как сложность выборки также определяется сложностью каждого объекта. Современные методы оценки сложности объектов выборки опираются на исследования сложности многообразий, аппроксимирующих элементы выборки. Для больших языковых моделей используются эмпирические законы масштабирования [27, 28], полученные Джаредом Капланом и Джорданом Хоффманом [29], однако эти законы не имеют строгого теоретического обоснования и не объясняют механизмы, лежащие в основе зависимостей между параметрами, данными и качеством модели.

В области снижения сложности моделей существуют два основных направления: квантизация параметров моделей и дистилляция знаний [30, 31, 32]. Однако существующие методы не опираются на строгие теоретические оценки сложности моделей и данных, что ограничивает их эффективность и предсказуемость.

Таким образом, на текущий момент в исследованиях отсутствует единый

теоретический аппарат для описания сложности моделей и данных, позволяющий получить асимптотические оценки и установить формальные критерии соответствия между сложностью модели и сложностью данных. Разработка такого аппарата позволит проводить сравнительный анализ различных нейросетевых архитектур для выбора оптимального решения для заданной задачи, характеризуемой выборкой определенной сложности.

В настоящей работе разработан комплексный теоретический подход к оценке сложности моделей и данных. Предложен новый математический аппарат для анализа сложности на основе анализа ландшафта функции потерь нейросетевых моделей через матрицы Гессе. Получены теоретические оценки матриц Гессе для различных архитектур моделей глубокого обучения, необходимые для оценки сложности модели. Разработан анализ связи объема и сложности выборки со сложностью модели, введены частные случаи общей теории сложности, позволяющие получать практические оценки для прикладных задач.

Объектом исследования являются параметрические семейства функций, задаваемые суперпозициями линейных и нелинейных преобразований, а также конечные выборки данных, применяемые для оценки параметров указанных семейств в рамках задачи минимизации эмпирического риска.

Предмет исследования: разработка теоретического аппарата для оценки и анализа сложности моделей глубокого обучения и сложности данных, а также установление формальных критериев соответствия между сложностью модели и сложностью выборки, необходимой для ее обучения.

Цель и задачи исследования. Целью исследования является построение единого теоретического аппарата для оценки сложности моделей глубокого обучения и сложности данных, а также установление формальных критериев соответствия между сложностью модели и сложностью выборки, необходимой для ее обучения.

Для достижения цели были поставлены и решены следующие задачи:

- введение формальных определений мер сложности моделей и данных в рамках теории мер и установление критерия обучаемости модели на выборке;
- получение теоретических оценок ландшафтных мер на основе матриц Гессе для полносвязных, сверточных и трансформерных архитектур моделей глубокого обучения;
- построение ландшафтной меры сложности модели на основе анализа матриц Гессе и установление ее связи с условной сложностью выборки;

- построение методов оценки достаточного объема выборки на основе анализа стабильности функции потерь и близости апостериорных распределений параметров;
- построение методов снижения сложности моделей глубокого обучения на основе анализа матриц Гессе и методов дистилляции знаний;
- демонстрация практического применения построенного теоретического аппарата в задачах многозадачного обучения, нейровизуализации и детекции машинно-генерированного контента.

Методы исследования. Для решения поставленных задач в диссертации используются методы теории мер, линейной алгебры, матричного анализа, включая матричное дифференцирование и спектральный анализ матриц, методы теории вероятностей, математической статистики, включая байесовский анализ, анализ сходимости случайных процессов, методы теории оптимизации, анализа функций многих переменных, методы статистической теории обучения.

Научная новизна. Научной новизной проведенного исследования является построение единого теоретического аппарата для оценки сложности моделей глубокого обучения и сложности данных на основе теории мер и анализа ландшафта оптимизационной задачи. Впервые введены формальные определения меры сложности выборки и меры сложности модели в рамках теории мер, установлен критерий обучаемости модели на выборке, определяющий необходимое условие предотвращения переобучения. Введена ландшафтная мера сложности модели, определяемая через спектральные свойства матриц Гессе функции потерь, и установлена ее связь с условной сложностью выборки. Получены строгие теоретические оценки ландшафтных мер на основе оценок матриц Гессе для полносвязных, сверточных и трансформерных архитектур моделей глубокого обучения. Построены методы оценки достаточного объема выборки на основе анализа стабильности функции потерь и близости апостериорных распределений параметров.

Теоретическая значимость работы. Диссертационная работа представляет собой фундаментальный теоретический вклад в теорию статистического обучения, расширяющий классические подходы к оценке сложности моделей на случай перепараметризованных нейронных сетей. В работе рассматриваются фундаментальные вопросы о сложности моделей машинного обучения, для которых получены строгие теоретические результаты, связывающие сложность модели со сложностью данных в рамках единого математического аппарата. Полученные оценки матриц Гессе и ландшафтной меры сложности открывают

новые направления исследований в теории выбора моделей машинного обучения, теории оптимизации нейронных сетей и анализе обобщающей способности моделей глубокого обучения.

Практическая значимость работы. Разработанный теоретический аппарат применен к решению прикладных задач машинного обучения: многозадачного обучения, декодирования фМРТ-изображений и детекции машинно-генерированного контента. Полученные методы оценки и управления сложностью моделей и данных экспериментально подтверждены на реальных задачах компьютерного зрения, обработки естественного языка и классификации.

Положения, выносимые на защиту:

1. Разработанный единый теоретический аппарат оценки сложности моделей глубокого обучения и сложности данных на основе теории мер и анализа ландшафта оптимизационной задачи, включающий формальные определения мер сложности и критерий обучаемости модели на выборке.
2. Введенная ландшафтная мера сложности модели, определяемая через спектральные свойства матриц Гессе функции потерь, задает количественную связь между архитектурными характеристиками модели и условной сложностью выборки.
3. Полученные теоретические оценки ландшафтной меры сложности на основе анализа спектральных норм матриц Гессе для полносвязных, сверточных раскрывают характер зависимости сложности моделей от их глубины, ширины и иных структурных параметров.
4. Получены теоретические оценки сходимости методов определения достаточного объема выборки, для которых установлена связь со сложностью моделей
5. Предложенные методы снижения сложности моделей глубокого обучения на основе анализа ковариационной матрицы градиентов, дистилляции на многодоменных данных и анти-дистилляции обеспечивают эффективное сокращение числа параметров и передачу знаний между моделями различной сложности и доменами данных.

Степень достоверности результатов. Достоверность научных результатов работы подтверждается непротиворечивостью и согласованностью с известными фактами и исследованиями в рассматриваемой области, высокой степенью сходимости теоретических результатов с данными экспериментов и определяется применением теоретических и методологических основ разработок ве-

дущих ученых в области обработки естественного языка, корректным и обоснованным использованием математического аппарата, экспериментальными исследованиями разработанных моделей и методов.

Соответствие диссертации паспорту специальности. Тема и основные результаты диссертации соответствуют следующим областям исследований паспорта специальности 1.2.1 — Искусственный интеллект и машинное обучение.

2 Исследования в области оценки качества и эффективности алгоритмических и программных решений для систем искусственного интеллекта и машинного обучения. Методики сравнения и выбора алгоритмических и программных решений при многих критериях.

4 Разработка методов, алгоритмов и создание систем искусственного интеллекта и машинного обучения для обработки и анализа текстов на естественном языке, для изображений, речи, биомедицины и других специальных видов данных.

16 Исследования в области специальных методов оптимизации, проблем сложность и элиминации перебора, снижения размерности.

17 Исследования в области многослойных алгоритмических конструкций, в том числе – многослойных нейросетей.

Апробация результатов диссертации. Основные результаты работы докладывались и обсуждались на Всероссийской конференции с международным участием «Математические методы распознавания образов» (Москва, 2019, Москва, 2021, Муром, 2025), Международной конференции «Интеллектуализация обработки информации» (Гаэта, 2018, Москва, 2020, Москва, 2022), Всероссийской научной конференции МФТИ (Москва, 2018, 2019, 2020, 2021, 2023, 2024, 2025), Ivannikov Ispras Open Conference (Москва, 2021, 2022, 2023, 2024), Ivannikov Memorial Workshop (Казань, 2022), Iberian Languages Evaluation Forum co-located with the Conference of the Spanish Society for Natural Language Processing (Андалусия, 2023, 2024), 35th Conference of Open Innovations Association (Тампере, 2024), Fourth Workshop on Scholarly Document Processing (Бангкок, 2024), 1st Workshop on GenAI Content Detection (GenAIDetect) (Абу-Даби 2025), 19th International Workshop on Semantic Evaluation (Вена, 2025).

Публикации. По теме диссертации опубликовано 56 научных работ, из которых 17 статей в научно-технических журналах, входящих в перечень ВАК, 32 – в изданиях, входящих в международные научометрические базы Scopus и Web of Science. В трудах российских и международных конференций опубли-

ковано 39 работ. Также на основе работ автора зарегистрировано 13 программ для ЭВМ.

Личный вклад соискателя. Все выносимые на защиту результаты и положения, составляющие основное содержание диссертационного исследования, разработаны и получены лично автором или при его непосредственном участии вместе с учениками. В работах, опубликованных в соавторстве, соискателю принадлежит определяющая роль в построении теоретических методов и направлении исследований.

Структура и объем работы. Диссертация состоит из оглавления, введения, шести разделов, заключения, списка иллюстраций, списка таблиц, перечня основных обозначений и списка литературы из 176 наименований. Основной текст занимает 289 страницы.

Краткое содержание работы по главам.

В главе 1 рассматривается текущее состояние теории обучаемости моделей глубокого обучения.

В главе 2 вводятся общие определения меры сложности моделей и данных, а также понятие обучаемости модели, основанное на сравнении меры сложности модели и меры сложности данных. Рассматривается условная сложность выборки, частным случаем которой при простой генеральной совокупности является достаточный объем выборки для обучения модели глубокого обучения; методы оценки достаточного объема выборки рассматриваются в главе 4. Вводится понятие условной сложности моделей глубокого обучения, частным случаем которой является ландшафтная мера, анализирующая изменение ландшафта функции потерь при вариации обучающей выборки. Ландшафтная мера основывается на исследовании матриц Гессе в окрестности локального минимума; подробные результаты по оценке матриц Гессе для специальных семейств моделей глубокого обучения рассматриваются в главе 3. В главе получена связь между достаточным объемом выборки, являющимся частным случаем условной сложности выборки, и ландшафтной мерой модели глубокого обучения. Из полученных общих теоретических оценок получены оценки для полносвязных, сверточных и трансформерных архитектур моделей глубокого обучения. Полученные оценки имеют асимптотический характер и указывают на экспоненциальный рост ландшафтной меры при увеличении числа слоев сети и полиномиальный рост при увеличении числа параметров внутри слоя.

В главе 3 описываются ключевые теоремы об оценках матриц Гессе для различных семейств моделей глубокого обучения. Матрица Гессе функции ошибки

по параметрам модели глубокого обучения содержит информацию о поведении функции вокруг некоторой точки и служит основой для введенной в главе 2 ландшафтной меры. В рамках данной главы получены выражения для матриц Гессе и оценки их спектральных норм для полносвязных, сверточных и трансформерных архитектур моделей глубокого обучения. Несмотря на технический характер главы, используемый в других главах диссертации, полученные оценки представляют самостоятельный научный результат, применимый в других направлениях исследования моделей глубокого обучения, в частности в теории оптимизации.

В главе 4 предлагаются методы оценки достаточного объема выборки для линейных и нейросетевых моделей глубокого обучения. В главе сравниваются классические и байесовские методы оценки достаточного объема выборки. Предложен новый метод оценки достаточного объема выборки на основе семплирования эмпирической функции ошибки, для которого указана сходимость для линейной модели и эмпирически подтверждена работа для моделей глубокого обучения. Также предложены методы определения достаточного размера выборки, основанные на близости апостериорных распределений близких выборок, с теоретическими оценками сходимости для любых параметрических семейств с нормальным апостериорным распределением параметров.

В главе 5 рассматриваются методы снижения сложности моделей глубокого обучения. Предложены методы удаления параметров на основе анализа ковариационной матрицы градиентов функции ошибки по параметрам модели. Рассматриваются методы дистилляции моделей глубокого обучения, в частности предложен метод дистилляции моделей на многодоменных данных. Также предложен метод анти-дистилляции для наращивания сложности модели, при котором использование информации из обученной небольшой модели учителя при обучении большой модели ученика обеспечивает повышенную устойчивость к шуму и более высокую точность аппроксимации.

В главе 6 рассматриваются прикладные применения теории сложности моделей глубокого обучения. Рассматривается оценка радемахеровской сложности в задаче многозадачного обучения; установлено, что использование LoRA-адаптеров снижает радемахеровскую сложность. Приведен пример снижения размерности признакового описания без снижения сложности данных и качества аппроксимации модели на примере высокоразмерных фМРТ-снимков. Рассмотрено качество данных в задаче детекции машинно-генерированного контента; с использованием введенных метрик продемонстрировано, что качество

данных непосредственно влияет на оценку детекторов.

Благодарности. Автор благодарен научной школе академиков РАН Константина Владимировича Рудакова и Юрия Ивановича Журавлева, в частности своим учителям доктору физико-математических наук Вадиму Викторовичу Стрижову и профессору РАН Константину Вячеславовичу Воронцову за интерес к работе, советам и поддержку в исследованиях, связанных с данной докторской диссертацией.

Также автор благодарен компании АО «Антиплагиат», где под руководством Юрия Викторовича Чеховича и Александра Сергеевича Кильдякова в отделе исследований проведено большое количество прикладных экспериментов, которые снизили общую сложность моделей глубокого обучения в сервисах компании.

Отдельно, автор признателен своим ученикам, аспирантам и студентам Герману Грицаю, Никите Киселеву, Данилу Дорину, Ильдару Хабутдинову, Владиславу Мешкову, Игорю Игнашину, Анастасии Вознюк, Анне Зверевой, Камилу Баязитову, Анне Ремизовой и Егору Петрову за обсуждение результатов, экспериментальную работу и за общее развитие теории моделей глубокого обучения и в частности их теорию сложности.

Глава 1

Теория обучаемости моделей машинного обучения

1.1. Основные понятия статистического обучения

В теории машинного обучения базовой является стохастическая постановка, в рамках которой данные рассматриваются как выборка из неизвестного распределения. Пусть задано вероятностное пространство $(\Omega, \mathcal{F}, \mathbb{P})$ и случайная пара (\mathbf{X}, \mathbf{Y}) , принимающая значения в декартовом произведении пространств объектов и ответов $\mathcal{X} \times \mathcal{Y}$, с неизвестным совместным распределением $\mathbb{P}_{\mathbf{X}, \mathbf{Y}}$. Генеральная совокупность Γ задает множество всех возможных объектов, а кольцо выборок

$$\mathfrak{D} = \{D_\Gamma^i : D_\Gamma^i \subset \Gamma\}$$

состоит из всех конечных подмножеств, доступных для анализа. В дальнейшем под выборкой будем понимать произвольный элемент $D \in \mathfrak{D}$ вида

$$D = \{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_m, \mathbf{y}_m)\},$$

который рассматривается как реализация независимых одинаково распределенных случайных величин $(\mathbf{X}_i, \mathbf{Y}_i) \sim \mathbb{P}_{\mathbf{X}, \mathbf{Y}}$.

Моделью (или гипотезой) называется отображение $f: \mathcal{X} \rightarrow \mathcal{Y}$, выбранное из фиксированного класса параметрических функций \mathfrak{F} . Для формализации качества модели вводится *функция потерь* $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$, измеряющая несоответствие предсказания $f(\mathbf{x})$ и истинного ответа \mathbf{y} . В задачах классификации часто используют 0–1-потерю

$$\ell_{0-1}(f(\mathbf{x}), \mathbf{y}) = \mathbb{I}\{f(\mathbf{x}) \neq \mathbf{y}\},$$

а в задачах регрессии — квадратичную или абсолютную потерю.

Определение 1 (Риск модели). *Риском модели $f \in \mathfrak{F}$ называют математическое ожидание потерь относительно истинного распределения данных:*

$$R(f) = \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim \mathbb{P}_{\mathbf{X}, \mathbf{Y}}} [\ell(f(\mathbf{X}), \mathbf{Y})].$$

Так как распределение $\mathbb{P}_{\mathbf{X}, \mathbf{Y}}$ неизвестно, прямой расчет $R(f)$ обычно невозможен. Вместо этого вводят *эмпирический риск* по выборке D :

$$\hat{R}_m(f) = \frac{1}{m} \sum_{i=1}^m \ell(f(\mathbf{x}_i), \mathbf{y}_i),$$

который рассматривается как статистическая оценка истинного риска. Связь между $R(f)$ и $\hat{R}_m(f)$ является центральным объектом исследования в теории обучаемости.

Замечание 1. В главах 2, 3 и далее для эмпирического риска параметризованных моделей используется обозначение $\mathcal{L}_m(\boldsymbol{\theta})$, которое совпадает с $\hat{R}_m(f)$ при параметризации $f_{\boldsymbol{\theta}}$. В настоящей главе оба обозначения рассматриваются как взаимозаменяемые.

Определение 2 (Алгоритм минимизации эмпирического риска). Алгоритмом минимизации эмпирического риска называют отображение, которое по выборке $D \in \mathfrak{D}$ возвращает модель

$$\hat{f}_{\text{ERM}} \in \arg \min_{f \in \mathfrak{F}} \hat{R}_m(f).$$

Даже если класс моделей достаточно богат, чтобы аппроксимировать зависимость между \mathbf{X} и \mathbf{Y} , для конечной выборки D может возникать *переобучение*: алгоритм подстраивается под случайные шумы в данных, достигая малого эмпирического риска при высоком истинном риске. Поэтому важной задачей является построение оценок отклонения $|R(f) - \hat{R}_m(f)|$ для различных классов \mathfrak{F} и схем выбора моделей.

Определение 3 (Наилучшая модель в классе и избыточный риск). Модель f^* , минимизирующая риск в фиксированном классе \mathfrak{F} ,

$$f^* \in \arg \min_{f \in \mathfrak{F}} R(f),$$

называется наилучшей в классе. Величина

$$R(f) - R(f^*)$$

называется избыточным риском модели f .

В идеальном случае теоретические оценки должны гарантировать, что избыточный риск $(R(\hat{f}_{\text{ERM}}) - R(f^*))$ мал с высокой вероятностью при разумном объеме выборки t . При этом ключевую роль играет мера сложности класса моделей \mathfrak{F} , определяющая скорость сходимости эмпирического риска к истинному.

Определение 4 (Обучаемость класса моделей). Пусть $\mathfrak{D}_m = \{D \subset \Gamma : |D| = m\}$ обозначает множество всех выборок размера m . Класс моделей \mathfrak{F} называется обучаемым в заданной постановке, если существует алгоритм

$A: \bigcup_{m=1}^{\infty} \mathfrak{D}_m \rightarrow \mathfrak{F}$, такой что для любого распределения $\mathbb{P}_{\mathbf{X}, \mathbf{Y}}$ и любых $\varepsilon > 0$, $\delta \in (0, 1)$ найдется такое число наблюдений $m(\varepsilon, \delta) \in \mathbb{N}$, что для любой выборки D размера $m \geq m(\varepsilon, \delta)$ выполняется

$$\mathbb{P}\{R(A(D)) - R(f^*) \leq \varepsilon\} \geq 1 - \delta,$$

где $f^* = \arg \min_{f \in \mathfrak{F}} R(f)$ — наилучшая модель в классе \mathfrak{F} .

Дальнейший анализ в рамках классических подходов концентрируется на поиске верхних оценок $m(\varepsilon, \delta)$ через различные характеристики сложности класса \mathfrak{F} : VC-размерность, ростовую функцию, радемахеровскую сложность и другие. Эти меры играют фундаментальную роль в теории обучаемости и подробно рассматриваются в следующем разделе.

1.2. Классические меры сложности и их ограничения

Первые строгие представления о сложности моделей сформировались в рамках статистической теории распознавания [2, глава 5], где ключевую роль сыграли понятия VC-размерности и структурного риска. Пусть задано пространство объектов \mathcal{X} и класс булевых моделей $\mathfrak{F} \subset \{0, 1\}^{\mathcal{X}}$.

Определение 5 (Разделимость выборки). *Набор точек $D = \{\mathbf{x}_1, \dots, \mathbf{x}_m\} \subset \mathcal{X}$ называется разделимым классом \mathfrak{F} , если для любого разметочного вектора $(\mathbf{y}_1, \dots, \mathbf{y}_m) \in \{0, 1\}^m$ существует модель $f \in \mathfrak{F}$ такая, что $f(\mathbf{x}_i) = \mathbf{y}_i$ для всех $i = 1, \dots, m$.*

Определение 6 (VC-размерность). *VC-размерностью класса \mathfrak{F} называется величина*

$$\text{VC}(\mathfrak{F}) = \sup \{m \in \mathbb{N} : \exists D \subset \mathcal{X}, |D| = m, D \text{ разделим классом } \mathfrak{F}\}.$$

Если множество в правой части пусто, полагаем $\text{VC}(\mathfrak{F}) = 0$. Если супремум не достигается, полагаем $\text{VC}(\mathfrak{F}) = +\infty$.

На основе введенных понятий VC-размерности и разделимости удалось связать обобщающую способность алгоритмов с ростовой функцией $\Pi_{\mathfrak{F}}(m)$ и получить верхние оценки разности между эмпирическим риском $\hat{R}_m(f)$ и истинным риском $R(f)$ вида

$$R(f) \leq \hat{R}_m(f) + C \sqrt{\frac{\text{VC}(\mathfrak{F}) \log m}{m}},$$

где $C > 0$ некоторая универсальная константа. Однако такие оценки используют грубые верхние границы, мало чувствительные к структуре конкретной задачи и распределению данных.

Развитие PAC-подхода дополнило картину формальными критериями обучаемости [3, раздел 2].

Определение 7 (PAC-обучаемость). Класс моделей \mathfrak{F} называется PAC-обучаемым (англ. *Probably Approximately Correct learnable*), если существует алгоритм $A: \bigcup_{m=1}^{\infty} \mathfrak{D}_m \rightarrow \mathfrak{F}$ и функция $m: (0, 1) \times (0, 1) \rightarrow \mathbb{N}$, такие что для любого распределения $\mathbb{P}_{\mathbf{X}, \mathbf{Y}}$ на $\mathcal{X} \times \{0, 1\}$ и любых $\varepsilon > 0, \delta \in (0, 1)$ для любой выборки D размера $t \geq m(\varepsilon, \delta)$ выполняется

$$\mathbb{P}\{R(A(D)) - R(f^*) \leq \varepsilon\} \geq 1 - \delta,$$

где $f^* = \arg \min_{f \in \mathfrak{F}} R(f)$. При этом функция $m(\varepsilon, \delta)$ должна быть полиномиальной по $1/\varepsilon$ и $\log(1/\delta)$.

Для класса \mathfrak{F} с конечной VC-размерностью $d = \text{VC}(\mathfrak{F}) < \infty$ известна оценка [3, раздел 2]

$$m(\varepsilon, \delta) \geq C \frac{1}{\varepsilon^2} \left(d \log \frac{1}{\varepsilon} + \log \frac{1}{\delta} \right),$$

где $C > 0$ — универсальная константа. Для реальных данных высокой размерности классические PAC-оценки остаются слабыми: они требуют огромных выборок даже для относительно простых архитектур и не учитывают внутреннюю структуру параметров, регуляризацию и специфику процесса оптимизации.

Более тонким инструментом анализа является радемахеровская сложность, связывающая способность класса функций к переобучению с эмпирическими оценками [4, раздел 2].

Определение 8 (Эмпирическая радемахеровская сложность). Пусть \mathcal{F} — класс вещественных функций на \mathcal{X} , $D = \{\mathbf{x}_1, \dots, \mathbf{x}_m\} \subset \mathcal{X}$ — конечная выборка, и $\sigma_1, \dots, \sigma_m$ — независимые случайные переменные Радемахера, то есть $\mathbb{P}\{\sigma_i = +1\} = \mathbb{P}\{\sigma_i = -1\} = 1/2$ для всех $i = 1, \dots, m$. Эмпирическая радемахеровская сложность класса \mathcal{F} на выборке D определяется как

$$\hat{\mathfrak{R}}_D(\mathcal{F}) = \mathbb{E}_{\sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i f(\mathbf{x}_i) \right],$$

где математическое ожидание берется по распределению случайных переменных Радемахера.

Определение 9 (Ожидаемая радемахеровская сложность). Ожидаемая радемахеровская сложность класса \mathcal{F} для выборок размера t определяется как

$$\mathfrak{R}_m(\mathcal{F}) = \mathbb{E}_{D \sim \mathbb{P}^m} [\hat{\mathfrak{R}}_D(\mathcal{F})],$$

где математическое ожидание берется по распределению выборок D размера m , порожденных распределением \mathbb{P} на \mathcal{X} .

Соответствующие неравенства связи с обобщающей способностью имеют вид

$$R(f) \leq \hat{R}_m(f) + 2\hat{\mathfrak{R}}_D(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2m}},$$

что позволяет учитывать адаптивность класса к конкретной выборке. Однако даже эти оценки остаются в основном асимптотическими и практически не отражают поведение современных нейронных сетей, имеющих миллионы параметров и сложные регуляризационные схемы [12]. Недостаток адаптивности классических мер к глубине, нелинейностям и особенностям оптимизации создает разрыв между теорией и практикой, который необходимо закрыть в дальнейших разделах обзора.

1.2.1. VC-размерность и ростовая функция

Одним из центральных понятий классической теории обучаемости является *ростовая функция* (англ. growth function) класса моделей.

Определение 10 (Индуктированные разметки). Для конечной выборки $D = \{\mathbf{x}_1, \dots, \mathbf{x}_m\} \subset \mathcal{X}$ множество индуцированных разметок класса \mathfrak{F} на выборке D определяется как

$$\mathfrak{F}|_D = \{(f(\mathbf{x}_1), \dots, f(\mathbf{x}_m)) : f \in \mathfrak{F}\} \subseteq \{0, 1\}^m.$$

Определение 11 (Ростовая функция). Ростовая функция класса \mathfrak{F} определяется как

$$\Pi_{\mathfrak{F}}(m) = \max_{D \subset \mathcal{X}, |D|=m} |\mathfrak{F}|_D|,$$

где максимум берется по всем возможным выборкам D размера m из пространства \mathcal{X} .

Тривиальная верхняя оценка имеет вид $\Pi_{\mathfrak{F}}(m) \leq 2^m$. Если VC-размерность класса конечна и равна $d = \text{VC}(\mathfrak{F}) < \infty$, то лемма Сауэра [2, глава 5, лемма 5.1] обобщает эту грубую оценку и дает полиномиальный рост числа реализаций:

$$\Pi_{\mathfrak{F}}(m) \leq \sum_{k=0}^d \binom{m}{k} \leq \left(\frac{em}{d}\right)^d, \quad \text{для всех } m \geq d.$$

В сочетании с неравенствами концентрации это приводит к классическим оценкам равномерной сходимости эмпирического риска к истинному риску, которые лежат в основе принципа минимизации структурного риска.

Ключевым следствием конечности VC-размерности является возможность построения иерархии вложенных классов $\mathfrak{F}_1 \subset \mathfrak{F}_2 \subset \dots$ с возрастающей сложностью, что позволяет формализовать баланс между смещением и дисперсией модели. Однако на практике вычисление или даже оценка VC-размерности для сложных архитектур оказывается нетривиальной задачей: для линейных классификаторов в \mathbb{R}^d верно $VC(\mathfrak{F}) = d + 1$, но для многослойных нейронных сетей VC-размерность оценивается как степенная функция от числа параметров и может быть астрономически велика, что делает полученные границы очень слабыми.

Кроме того, классические оценки через $VC(\mathfrak{F})$ не отражают влияние таких практических аспектов, как ранняя остановка, стохастическая оптимизация, нормализация и другие регуляризаторы, которые существенно улучшают обобщающую способность, но никак не уменьшают формальную VC-размерность класса.

1.2.2. PAC-обучаемость: реализуемый и агностический случаи

В рамках PAC-подхода (англ. Probably Approximately Correct) различают два основных случая обучаемости: реализуемый и агностический. Для 0–1-потерь в бинарной классификации PAC-обучаемость сводится к поиску гипотезы с малой ошибкой обобщения при ограниченном числе прецедентов.

В реализуемой постановке предполагается, что существует модель $f^* \in \mathfrak{F}$, для которой $R(f^*) = 0$, то есть данные порождаются без шума в рамках выбранного класса. В этой ситуации удается получить более сильные гарантии сходимости алгоритмов минимизации эмпирического риска:

$$\Pr \left\{ \sup_{f \in \mathfrak{F}} |R(f) - \hat{R}_m(f)| > \varepsilon \right\} \leq \delta,$$

при числе наблюдений $m(\varepsilon, \delta)$, логарифмически зависящем от $1/\delta$ и линейно зависящем от $VC(\mathfrak{F})$.

Примером служат линейные классификаторы в \mathbb{R}^d : для гиперплоскостей VC-размерность равна $d + 1$, и PAC-обучаемость достигается при $m = O((d/\varepsilon) \log(1/\varepsilon) + (1/\varepsilon) \log(1/\delta))$, что согласуется с персентроном Розенблата. Связь с VC следует из леммы Vapnik–Chervonenkis [2, глава 5, теорема 5.1]: если $VC(\mathfrak{F}) = d < \infty$, то класс PAC-обучаем, с оценкой $m \geq (1/\varepsilon)(4d \log(13/\varepsilon) + \log(2/\delta))$.

В агностической постановке шум допускается, и оптимальная модель f^* достигает ненулевого риска $R(f^*) > 0$. В этом случае стандартной целью является достижение неравенства вида

$$R(f) \leq R(f^*) + \varepsilon$$

с вероятностью не менее $1 - \delta$. Соответствующие оценки на число наблюдений имеют тот же порядок, что и в реализуемом случае, но константы и зависимости от параметров модели оказываются менее благоприятными. Следует подчеркнуть, что классические PAC-оценки делают акцент на наихудшем распределении данных и не используют информацию о структуре реальных задач.

Следует отметить, что PAC-анализ, как правило, не учитывает вычислительную сторону задачи: существование алгоритма, удовлетворяющего PAC-критериям, не гарантирует его практическую реализуемость с полиномиальной сложностью. Для глубоких нейронных сетей это особенно важно, так как реальное обучение основано на приближенных численных процедурах, которые не обязательно находят глобальный минимум риска.

1.2.3. Радемахеровская сложность и локальные оценки

Переход к радемахеровской сложности был мотивирован стремлением получить более точные, зависящие от данных оценки сложности класса функций. В отличие от VC-размерности, которая является глобальной характеристикой \mathfrak{F} , величина $\hat{\mathfrak{R}}_D(\mathcal{F})$ зависит от конкретной выборки D и потому позволяет учитывать геометрию данных. Через технику симметризации и цепочечные неравенства могут быть получены оценки общего вида

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}} |R(f) - \hat{R}_m(f)| \right] \leq C \mathfrak{R}_m(\mathcal{F}),$$

где $\mathfrak{R}_m(\mathcal{F})$ — ожидаемая радемахеровская сложность, определенная выше.

Для типичных классов гладких функций могут быть получены явные верхние границы на $\mathfrak{R}_m(\mathcal{F})$ через нормы параметров и радиус области определения. Например, для линейных классификаторов

$$\mathcal{F} = \{x \mapsto \langle w, x \rangle : \|w\|_2 \leq B\},$$

при ограничении $\|x_i\|_2 \leq R$ для всех i верна оценка

$$\hat{\mathfrak{R}}_D(\mathcal{F}) \leq \frac{BR}{\sqrt{m}}.$$

Для нейронных сетей аналогичные оценки строятся через произведения норм весов по слоям, что уже лучше согласуется с эмпирическими наблюдениями о роли регуляризации весов.

Однако и здесь сохраняется ряд принципиальных ограничений. Во-первых, получение достаточно точных верхних границ для глубоких архитектур приводит к сильно завышенным оценкам, поскольку приходится применять грубые неравенства при переходе от нелинейных слоев к линейным аппроксимациям. Во-вторых, стандартные радемахеровские оценки учитывают только статический класс функций и не отражают динамику обучения, в частности, влияние траектории градиентного спуска и стохастичности минибатчей. В-третьих, даже будучи зависящими от данных, эти оценки по-прежнему нацелены на наихудший случай и плохо согласуются с наблюдаемой в практике устойчивостью крупных моделей к шуму в данных.

1.2.4. Комбинаторный подход Воронцова к снижению оценок Вапника–Червоненкиса

Значительный вклад в развитие комбинаторной теории надежности обучения по прецедентам внес К.В. Воронцов, развивший идеи В.Н. Вапника и А.Я. Червоненкиса в русле комбинаторных границ для статистического обучения [5]. В его докторской диссертации [33, раздел 2] предложен комбинаторный подход к снижению оценок Вапника–Червоненкиса, основанный на учете структуры данных и специфики класса гипотез. Ключевая идея заключается в том, что классические VC-оценки используют worst-case анализ и не учитывают комбинаторные свойства конкретной выборки, что приводит к завышенным границам на вероятность переобучения.

В рамках комбинаторного подхода Воронцов вводит меры, такие как α -расширение класса гипотез и энтропийные характеристики, которые позволяют получить более точные оценки вероятности переобучения. Для класса \mathfrak{F} с конечным VC-размером d комбинаторные оценки дают вероятность ошибки обобщения вида [33, раздел 1.5]

$$\Pr\{R(f) - \hat{R}_m(f) > \varepsilon\} \leq 8\Pi_{\mathfrak{F}}(2m)e^{-m\varepsilon^2/32},$$

где $\Pi_{\mathfrak{F}}(2m)$ контролируется комбинаторными свойствами прецедентов, а не только глобальной VC-размерностью. В отличие от классических оценок, которые зависят только от $VC(\mathfrak{F})$ и размера выборки m , комбинаторный подход

учитывает структуру данных через энтропийные характеристики, что позволяет получить более точные границы для конкретных задач.

Ключевым преимуществом комбинаторного подхода является возможность существенного снижения оценок по сравнению со стандартными границами Вапника–Червоненкиса. Это достигается за счет учета специфических свойств выборки, таких как коррелированность объектов, наличие кластерной структуры или дискретная природа данных. Воронцов также предложил алгоритмы для эмпирического расчета таких границ, что делает подход применимым к реальным задачам прецедентного обучения, в особенности в задачах распознавания образов, где данные имеют дискретную структуру.

Отдельно стоит подчеркнуть, что в диссертации Воронцова [33] установлена формальная связь комбинаторного подхода с радемахеровской сложностью [4]. В работе [33, раздел 2.2.1] показано, что комбинаторные оценки могут быть выражены через эмпирическую радемахеровскую сложность. Для класса функций \mathfrak{F} и выборки D верна оценка

$$\hat{\mathfrak{R}}_D(\mathfrak{F}) \leq \sqrt{\frac{2 \log \Pi_{\mathfrak{F}}(m)}{m}},$$

где $\Pi_{\mathfrak{F}}(m)$ — ростовая функция класса \mathfrak{F} , контролируемая комбинаторными свойствами прецедентов. Обратно, комбинаторные границы на вероятность переобучения могут быть переформулированы через радемахеровскую сложность [33, раздел 2.2]:

$$\Pr\{R(f) - \hat{R}_m(f) > \varepsilon\} \leq \exp\left(-\frac{m\varepsilon^2}{8\hat{\mathfrak{R}}_D^2(\mathfrak{F}) + 2\varepsilon}\right),$$

что позволяет объединить преимущества обоих подходов: адаптивность к данным, характерную для радемахеровской сложности, и комбинаторную структуру, учитывающую специфику прецедентного обучения. Эта формальная связь демонстрирует, что комбинаторный подход не является изолированным методом, а естественным образом дополняет и развивает классические меры сложности, включая VC-размерность и радемахеровскую сложность.

Комбинаторный подход дополняет классический VC-анализ более адаптивными метриками, которые могут быть существенно ниже стандартных оценок, что особенно важно для практических приложений, где классические worst-case границы оказываются слишком консервативными. Этот подход подчеркивает роль комбинаторной оптимизации в контроле сложности и демонстрирует, что

учет структуры данных позволяет получать более реалистичные оценки обобщающей способности моделей.

1.2.5. Общие ограничения классических мер сложности

Суммируя рассмотренные подходы, выделим следующие ключевые ограничения классических мер сложности в контексте современных моделей глубокого обучения.

Во-первых, большинство оценок строится в предположении фиксированного, конечномерного пространства признаков и не учитывает появление дополнительных структур, таких как сверточные фильтры, остаточные связи и механизмы внимания, которые существенно меняют эффективную сложность класса моделей.

Во-вторых, как VC-размерность, так и радемахеровская сложность ориентированы на worst-case анализ и не используют специфические свойства реальных распределений данных, такие как низкая размерность многообразия, коррелированность признаков или наличие сильных инвариантностей. В результате получаемые границы оказываются слишком консервативными и часто на несколько порядков хуже эмпирически наблюдаемых характеристик.

В-третьих, классические меры, как правило, рассматривают модель и данные раздельно: сложность гипотезы оценивается независимо от сложностей выборки, а влияние распределения на обучение учитывается лишь через общие вероятностные предположения. Для глубоких нейросетевых архитектур, где наблюдается сложное взаимодействие между параметрами, структурой данных и алгоритмом оптимизации, такое раздельное рассмотрение оказывается недостаточным для объяснения эмпирических фактов.

Четвертое ограничение заключается в том, что классические результаты в основном фокусируются на асимптотических режимах $t \rightarrow \infty$ и не описывают поведение моделей при конечных, но больших размерах выборок, характерных для реальных задач. Это особенно заметно при анализе крупных языковых и мультимодальных моделей, для которых наблюдаются устойчивые законы масштабирования качества от размера данных и модели, выходящие за рамки традиционных теоретических предсказаний.

1.3. Современные подходы к анализу сложности нейросетей

Развитие теории сложности нейросетей началось с изучения их аппроксимирующих свойств. В фундаментальной теореме Дж. Цибенко (англ. Cybenko theorem) показано, что многослойный перцептрон с регулируемым числом нейронов в скрытом слое способен аппроксимировать любую непрерывную функцию на компактном подмножестве \mathbb{R}^n с произвольной точностью [6, теорема 1]. Этот результат формально выражается следующим образом: для любой непрерывной функции $g: [0, 1]^n \rightarrow \mathbb{R}$ и любого $\varepsilon > 0$ существует сеть вида

$$f(\mathbf{x}) = \sum_{j=1}^N \alpha_j \sigma(\mathbf{w}_j^\top \mathbf{x} + b_j),$$

где σ — сигмоидальная нелинейность, такая что $\|f - g\|_\infty < \varepsilon$. Теорема Цибенко предоставила первое строгое обоснование репрезентативной способности нейронных сетей, однако не дала конструктивных оценок на число нейронов N и не объяснила, как глубина влияет на эффективность представления функций.

Вскоре после этого Й. Хестад показал, что переход к многоуровневым архитектурам позволяет экспоненциально сократить число нейронов, необходимое для аппроксимации некоторых классов функций [7, глава 4]. Формально, для любой фиксированной глубины L существует семейство булевых функций F_L , для которого любая сеть глубины L требует $\Omega(\exp(n^{1/(L-1)}))$ нейронов, тогда как сеть глубины $L+1$ реализует те же функции полиномиально малым числом параметров. Этот результат свидетельствует о принципиальной роли глубины: существуют функции, которые требуют экспоненциального числа нейронов в сети фиксированной глубины, но допускают компактное представление при добавлении дополнительных слоев. Дальнейшее развитие этих идей представлено в работах Й. Бенджио и Я. Лекуна [8, 9], где обсуждалась экспоненциальная эффективность глубоких архитектур.

В более современных исследованиях Н. Коэна, А. Эльдана и др. показано, что глубина определяет не только число параметров, но и тип функций, которые допускают эффективное представление [10, теорема 3.1] [11, теорема 1]. В частности, теорема Коэна устанавливает, что для тензорных разложений полиномов степени d в сети с ReLU-нелинейностями минимальное число параметров p_L на глубине L удовлетворяет $p_L = \tilde{\Theta}(d^{L/2})$, что подчеркивает экспоненциальный рост выразительной силы с глубиной. Эльдан и Шамир в своей теореме доказывают существование функций, аппроксимируемых трехслойными сетя-

ми с $O(d^2)$ нейронами, но требующих $\Omega(2^d)$ нейронов в двухслойных сетях: формально, для класса полиномов P_d на $[0, 1]^n$ верна оценка

$$\inf\{\|f - g\|_\infty : f \text{ двухслойная сеть}\} \geq c \cdot 2^{-d},$$

в то время как трехслойная сеть достигает $\|f - g\|_\infty \leq C \cdot d^{-1/2}$. Эти результаты не только подтверждают преимущество глубины, но и связывают ее с геометрией параметрического пространства: в узких сетях оптимизация сталкивается с “проклятием размерности”, тогда как широкие сети позволяют эффективно захватывать нелинейные зависимости. Такие теоремы мотивируют анализ Hessian-спектра для оценки “гладкости” ландшафта и роли глубины в контроле сложности.

Работа Н. Коэна и А. Эльдана также указывает на экспоненциальную зависимость сложности сети от числа слоев: для определенного класса полиномов степени d минимальное число скрытых блоков N_L в сети глубины L удовлетворяет условию

$$N_L \geq c \cdot \exp\left(\gamma \frac{d}{L}\right),$$

где $c, \gamma > 0$ зависят от нормировки весов и выбора нелинейностей. При увеличении глубины до $L + 1$ тот же класс функций достигается при

$$N_{L+1} \leq C \cdot d^\alpha,$$

то есть сложность падает до полиномиальной. Это означает, что для фиксированного уровня точности гибрид “ширина/глубина” должен подчиняться экспоненциальной зависимости: слишком малое число слоев приводит к взрывному росту числа параметров, тогда как рост глубины позволяет удерживать сложность в полиномиальных границах.

Параллельно исследовались подходы к количественной оценке сложности обучаемых представлений. Теоремы PAC-Bayes [34] связывают обобщающую способность моделей с дивергенциями между апостериорным и априорным распределениями параметров. Для нейросетей эти подходы применялись в контексте анализа плотности спектров Гессиана и флатнеса локальных минимумов. Несмотря на то, что PAC-Bayes оценки предоставляют гибкие вероятностные ограничения, на практике получение информативных границ требует сложного подбора априорного распределения и зачастую приводит к слабым числам.

Альтернативным направлением исследования является изучение динамики обучения через линейные аппроксимации, такие как нейронно-тангентное ядро

(англ. NTK). Работы Артура Жако и коллег [35] показали, что в пределе бесконечной ширины динамика градиентного спуска описывается линейным ядерным методом, что позволяет получить строгие результаты по обучаемости и обобщению в этом режиме. Подходы на основе NTK демонстрируют, что некоторые архитектуры обладают явным «линейным» режимом, однако не объясняют поведение конечных сетей и не учитывают ограниченные вычислительные ресурсы.

1.3.1. Анализ ландшафта функции потерь и матрицы Гессе

Активное развитие получило направление анализа ландшафта функции потерь (англ. loss landscape) и спектральных свойств матриц Гессе как меры сложности нейросетевых моделей, начиная с работ 2017 года. Фундаментальные работы Л. Сагуна [36] установили, что матрицы Гессе перепараметризованных нейронных сетей обладают характерной структурой: большинство собственных значений сосредоточены около нуля, а небольшое число больших собственных значений определяет кривизну ландшафта. Это наблюдение привело к пониманию, что эффективная размерность пространства параметров может быть существенно меньше формального числа параметров.

Ключевым результатом является декомпозиция матрицы Гессе на G-компоненту и H-компоненту, предложенная в работах [36, 37]. G-компонента отражает кривизну функции потерь относительно выходов сети, тогда как H-компонента описывает кривизну самой нейронной сети. Эта декомпозиция позволяет анализировать вклад различных факторов в общую сложность оптимизационного ландшафта и связывает свойства Гессиана с обобщающей способностью моделей.

Эмпирические исследования показали, что минимумы с малым спектральным радиусом Гессиана коррелируют с лучшей обобщающей способностью [38, 39]. Работа Н. Кескара [38] установила связь между “острыми” минимумами и худшей обобщающей способностью при обучении большими батчами. Л. Динь [39] показали, что “плоские” минимумы могут обобщаться лучше, хотя эта связь не является универсальной и зависит от архитектуры и данных.

Визуализация ландшафта функции потерь, предложенная Х. Ли [40], позволила эмпирически исследовать геометрию оптимизационного пространства. Методы визуализации через фильтрацию случайных направлений и анализ одномерных сечений показали, что успешно обученные сети находятся в “широких

долинах” с плавным ландшафтом, тогда как плохо обученные модели характеризуются “острыми” минимумами с высокой кривизной.

Структурный анализ матриц Гессе на больших масштабах был проведен в работах В. Папяна [41] и Б. Горбани [42], которые исследовали динамику спектра Гессиана в процессе обучения и его зависимость от размера выборки. Было установлено, что распределение собственных значений Гессиана следует степенным законам, причем “хвост” распределения содержит информацию о критических направлениях в пространстве параметров.

Для сверточных нейронных сетей С.П. Сингх и коллеги [43] получили теоретические оценки структуры матриц Гессе, показав, что сверточная структура приводит к блочно-циркулянтным свойствам Гессиана. Для трансформеров В. Орманиец и коллеги [44] провели теоретический анализ Гессиана, связав его структуру с механизмами внимания и показав, как архитектурные особенности отражаются в спектральных свойствах.

Вычислительные методы для анализа Гессиана включают быстрый алгоритм умножения на Гессиан, предложенный Б. Перлмуттером [45], и библиотеку PyHessian [46], позволяющую эффективно вычислять собственные значения и след матрицы Гессе для больших моделей. Эти инструменты сделали возможным эмпирический анализ ландшафта функции потерь для современных архитектур.

Теоретические результаты о структуре ландшафта были получены в работах Дж. Пеннингтона [47, 48], который исследовал спектр матрицы Фишера и показал возникновение спектральной универсальности в глубоких сетях. Г. Гур-Ари и коллеги [49] доказали, что градиентный спуск находит глобальные минимумы для достаточно широких сетей, что связано с “плоскостью” ландшафта в пределе бесконечной ширины.

Методы регуляризации, основанные на анализе Гессиана, включают Entropy-SGD [50], который смещает градиентный спуск в “широкие долины” ландшафта, и методы, использующие информацию о кривизне для адаптивной настройки шага обучения. С. Фор и С. Ястжебски [51, 52] исследовали крупномасштабную структуру ландшафта, показав наличие иерархии минимумов и связь между геометрией ландшафта и обобщающей способностью.

Несмотря на значительный прогресс, большинство результатов остаются эмпирическими и зависят от режима оптимизации, архитектуры и данных. Строгие теоретические связи между спектральными свойствами Гессиана и обобщающей способностью пока установлены лишь для ограниченных классов мо-

делей, что мотивирует разработку новых теоретических подходов, способных объединить анализ ландшафта с мерами сложности данных и моделей.

1.4. Эмпирические оценки и методы снижения сложности моделей

1.4.1. Эмпирические законы масштабирования

С обоснованием потребности в больших данных связан анализ эмпирических законов масштабирования. Фундаментальная работа Дж. Каплана и коллег [28] установила, что функция потерь языковых моделей подчиняется степенным законам относительно числа параметров N , объема обучающих данных D и вычислительного бюджета C . Эмпирически наблюдается зависимость вида

$$L(N, D) = \left(\frac{N_c}{N}\right)^{\alpha_N} + \left(\frac{D_c}{D}\right)^{\alpha_D} + L_0,$$

где N_c, D_c — критические значения параметров и данных, $\alpha_N \approx 0.076$, $\alpha_D \approx 0.095$ — показатели степенных законов, а L_0 — минимально достижимая потеря. Эти зависимости показывают, что для достижения заданного уровня качества необходимо синхронно масштабировать как модель, так и данные, причем влияние данных оказывается несколько сильнее.

Ключевым следствием является то, что при фиксированном вычислительном бюджете C существует оптимальное распределение ресурсов между числом параметров и объемом данных. Работа Дж. Хоффмана и коллег [27] формализовала эту задачу оптимизации. В рамках модели, где вычислительная стоимость обучения пропорциональна $C \propto N \cdot D$, авторы показали, что оптимальное соотношение между числом параметров N^* и числом токенов D^* подчиняется закону

$$D^* = 20 \cdot N^*,$$

то есть для оптимального использования вычислительных ресурсов объем данных должен быть примерно в 20 раз больше числа параметров. Это противоречит распространенной практике обучения очень больших моделей на относительно малых датасетах: например, модель GPT-3 с $N = 175 \times 10^9$ параметров обучалась на $D \approx 300 \times 10^9$ токенах, тогда как согласно методу Chinchilla оптимальным было бы $D^* \approx 3.5 \times 10^{12}$ токенов.

Экспериментальная валидация метода Chinchilla на моделях размером от 70 миллионов до 16 миллиардов параметров подтвердила, что при соблюдении соотношения $D = 20N$ достигается лучшее качество при том же вычислительном

бюджете. В частности, модель Chinchilla-70B [27] (70 миллиардов параметров, обученная на 1.4×10^{12} токенах) превзошла модель Gopher-280B [29] (280 миллиардов параметров, обученную на меньшем объеме данных) по большинству метрик, несмотря на в 4 раза меньшее число параметров. Это демонстрирует критическую важность баланса между сложностью модели и объемом данных, что напрямую связано с понятиями условной сложности данных и достаточного размера выборки, развивамыми в настоящей диссертации.

Несмотря на практическую значимость, эти законы масштабирования остаются эмпирическими и не имеют строгого теоретического обоснования. Они подстраиваются под конкретную архитектуру, функцию потерь и метрики качества, и их применимость к другим типам моделей требует дополнительной валидации. Кроме того, метод Chinchilla предполагает фиксированную архитектуру и не учитывает адаптивные стратегии обучения, такие как обучение по учебному плану (англ. curriculum learning) или динамическое изменение размера модели в процессе обучения.

1.4.2. Методы снижения сложности моделей: дистилляция и регуляризация

Влияние регуляризации и структурных ограничений параметров анализируется через теории дистилляции, привилегированного обучения и методов снижения размерности. Классический подход к снижению сложности моделей — дистилляция знаний (англ. knowledge distillation), предложенная Дж. Хинтоном и коллегами [30]. В рамках данного подхода “учитель” передает знания “ученику” через “мягкие” вероятностные распределения на выходе, а не только через жесткие метки классов. Дистилляция интерпретируется как способ переноса сложности от сложной модели к компактной, позволяя сохранить качество при существенном уменьшении числа параметров.

Формально, задача дистилляции формулируется как минимизация функции потерь вида

$$\mathcal{L}_{\text{distill}} = \alpha \mathcal{L}_{\text{hard}}(f_{\text{student}}, \mathbf{y}) + (1 - \alpha) \mathcal{L}_{\text{soft}}(f_{\text{student}}, f_{\text{teacher}}),$$

где $\mathcal{L}_{\text{hard}}$ — стандартная функция потерь между предсказаниями ученика и истинными метками, $\mathcal{L}_{\text{soft}}$ — функция потерь между “мягкими” выходами ученика и учителя, а $\alpha \in [0, 1]$ — параметр баланса. Температурный параметр

$T > 1$ используется для “размягчения” распределений вероятностей, что позволяет ученику лучше усваивать структуру знаний учителя.

Д. Лопес-Пас и коллеги [31] предложили унифицированную теорию дистилляции и привилегированного обучения, показав, что дистилляция может быть интерпретирована как частный случай обучения с дополнительной информацией. Байесовский подход к дистилляции, развитый в работах [32], позволяет формализовать перенос знаний через дивергенции между апостериорными распределениями параметров учителя и ученика, что дает более строгие теоретические гарантии.

Строгие границы на избыточный риск для дистилляции пока существуют лишь для простых семей моделей, таких как линейные классификаторы или мелкие нейронные сети. Для глубоких архитектур большинство результатов остаются эмпирическими, хотя наблюдается устойчивая закономерность: правильно выполненная дистилляция позволяет достичь качества, близкого к учителю, при существенно меньшем числе параметров.

Помимо дистилляции, к методам снижения сложности относятся квантизация параметров, вырезание несущественных параметров на основе анализа важности, и методы структурной регуляризации, ограничивающие выразительную силу модели через архитектурные ограничения. Эти методы часто комбинируются с дистилляцией для достижения максимального сжатия моделей при сохранении качества.

1.5. Проблемы существующих подходов и вклад диссертации

Современная теория сложности нейросетей развивается в нескольких направлениях: (1) аппроксимативные теоремы, описывающие роль глубины и ширины; (2) вероятностные оценки обобщения (англ. PAC-Bayes [34], англ. NTK [35]), связывающие динамику оптимизации с геометрией параметрического пространства; (3) ландшафтные методы, изучающие спектральные свойства матриц Гессе и устойчивость к возмущениям; (4) эмпирические законы масштабирования [28, 27], связывающие качество с ресурсами; (5) методы снижения сложности, включая дистилляцию и регуляризацию. Несмотря на существенный прогресс, эти подходы по-прежнему дают неполную картину и мотивируют разработку новых мер сложности, способных совместить свойства данных и модели в единой теории.

Анализ представленных в настоящей главе направлений исследования слож-

ности моделей глубокого обучения выявляет ряд фундаментальных проблем, ограничивающих их практическую применимость и теоретическую строгость, решение которых является целью настоящей диссертации.

Во-первых, классические меры сложности, основанные на VC-размерности, PAC-обучаемости и радемахеровской сложности, разработаны для моделей с ограниченным числом параметров и не учитывают специфику перепараметризованных нейронных сетей. Эти меры дают слишком пессимистичные оценки для глубоких архитектур, не отражая их реальную обобщающую способность. Комбинаторный подход Воронцова [33] позволяет снизить оценки Вапника–Червоненкиса, но остается применимым лишь к ограниченным классам моделей.

Во-вторых, существующие подходы рассматривают сложность модели и сложность данных изолированно, не устанавливая формальных критериев их соответствия. Эмпирические законы масштабирования (англ. scaling laws) [28, 27] выявляют корреляции между числом параметров, объемом данных и качеством модели, но не имеют строгого теоретического обоснования и не объясняют механизмы, лежащие в основе этих зависимостей. Отсутствует формальный критерий обучаемости, связывающий свойства модели с характеристиками данных.

В-третьих, анализ матриц Гессе и ландшафта функции потерь [36, 38, 39] выявил важные эмпирические закономерности, однако большинство результатов остаются эмпирическими и зависят от режима оптимизации, архитектуры и данных. Строгие теоретические связи между спектральными свойствами Гессиана и обобщающей способностью установлены лишь для ограниченных классов моделей. Существующие теоретические результаты либо слишком общие, либо применимы только к специфическим режимам. Отсутствуют точные оценки сложности для конкретных архитектур, учитывающие их структурные особенности и связывающие гиперпараметры архитектуры с мерой сложности.

В-четвертых, для моделей с миллионами и миллиардами параметров прямое вычисление и анализ матриц Гессе становится непрактичным из-за квадратичной сложности по памяти и вычислениям. Существующие методы аппроксимации [46] не дают теоретических гарантий точности и не позволяют получить аналитические оценки сложности. Законы масштабирования [28, 27] подстраиваются под конкретную архитектуру, функцию потерь и метрики качества, их применимость к другим типам моделей требует дополнительной валидации. Кроме того, они не учитывают адаптивные стратегии обучения и не объясняют,

почему определенные соотношения между параметрами и данными являются оптимальными.

Таким образом, несмотря на существенный прогресс в различных направлениях исследования сложности моделей глубокого обучения, отсутствует единый формальный аппарат, способный связать сложность модели и данных в рамках строгой теоретической основы, применимой к широкому классу архитектур нейронных сетей и обеспечивающей вычислительно осуществимые методы оценки сложности.

Глава 2

Сложность моделей и данных

Как было продемонстрировано в предыдущей главе, существующие подходы к анализу сложности моделей глубокого обучения характеризуются существенными ограничениями. Во-первых, классические меры сложности не учитывают специфику перепараметризованных нейронных сетей. Во-вторых, сложность модели и данных рассматриваются изолированно без формальных критериев их соответствия. В-третьих, анализ матриц Гессе остается в основном эмпирическим и не обеспечивает строгих теоретических связей для конкретных архитектур. В результате отсутствует единый формальный аппарат, способный связать сложность модели и данных в рамках строгой теоретической основы, применимой к широкому классу архитектур нейронных сетей.

Для преодоления указанных ограничений в настоящей главе разрабатывается единый теоретический формализм, устанавливающий формальное соотношение между сложностью модели и сложностью данных. Предлагаемый подход основан на введении мер сложности в рамках теории мер, что позволяет формализовать критерий обучаемости модели на выборке и установить строгие теоретические связи между свойствами архитектуры модели и характеристиками данных.

Ключевая идея настоящей главы заключается в установлении формального соотношения между мерой сложности модели $\mu_f(f)$ и мерой сложности данных $\mu_D(D)$, определяемого через условие обучаемости:

$$\mu_f(f) \leq \mu_D(D),$$

а также получение частных случаев, которые имеют более подробный практический и теоретический анализ.

Основным инструментом анализа в предложенном подходе служит изучение изменения функции потерь при непрерывном изменении выборки. Данный подход обеспечивает установление строгих теоретических связей между спектральными свойствами матриц Гессе и обобщающей способностью моделей, что восполняет пробел, выявленный в обзоре литературы. В разделе 2.3. показывается, как абсолютное изменение функции потерь оценивается через спектральную норму матрицы Гессе:

$$|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})| \leq \frac{M_\ell}{k+1} + \frac{1}{k+1} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 \left\| \mathbf{H}_{k+1}(\boldsymbol{\theta}^*) - \frac{1}{k} \sum_{i=1}^k \mathbf{H}_i(\boldsymbol{\theta}^*) \right\|_2$$

Результат, основанный на теоретических выкладках из главы 3, обеспечивает формализацию понятия *условной сложности выборки* $\mu_D(D_i|f)$ и установление критериев достаточности объема данных для обучения конкретной модели. Предлагаемый формализм обеспечивает вычислительно осуществимые методы оценки сложности, основанные на аналитических оценках спектральных свойств матриц Гессе. Это решает проблему непрактичности прямого вычисления матриц Гессе для крупных моделей.

Предлагаемый формализм не только углубляет теоретическое понимание процессов обучения глубоких нейронных сетей, но и обладает практической значимостью для разработки эффективных стратегий обучения, выбора архитектур моделей и планирования экспериментов. Результаты настоящей главы создают мост между теоретическим анализом оптимизационных свойств моделей и практическими аспектами их применения к реальным данным, открывая новые возможности для систематического подхода к проектированию и обучению сложных нейросетевых архитектур.

Структура главы организована следующим образом. В разделе 2.1. вводятся формальные определения меры сложности выборки и меры сложности модели в рамках теории мер, а также устанавливается критерий обучаемости модели на выборке. В разделе 2.2. рассматривается частный случай меры сложности данных — достаточный объем выборки, и вводится понятие условной сложности выборки. В разделе 2.3. разрабатывается ландшафтная мера сложности модели на основе анализа сходимости функции потерь, и получены теоретические оценки для полносвязных, сверточных и трансформерных моделей. В разделе 2.4. представлены результаты вычислительных экспериментов, подтверждающие теоретические оценки.

2.1. Оценка сложности моделей и данных

Как было указано во введении, существующие подходы рассматривают сложность модели и сложность данных изолированно, не устанавливая формальных критериев их соответствия. В настоящем разделе производится формализация мер сложности выборки и модели в рамках теории мер, что позволяет установить строгий критерий обучаемости, связывающий свойства модели с характеристиками данных.

Определение 12 (Генеральная совокупность данных). *Генеральной совокупностью данных Γ назовем произвольное множество объектов, которые исследу-*

ются в рамках той или иной задачи. В общем случае нет никаких ограничений на счетность множества генеральной совокупности.

Определение 12 позволяет работать как с однородными, так и с многородовыми генеральными совокупностями.

Определение 13 (Однородная и многородная генеральная совокупность). Генеральную совокупность Γ назовем однородной, если все объекты генеральной совокупности порождаются из одного распределения. В противном случае генеральную совокупность назовем k -родной, где k является числом распределений, на основе которых была сгенерирована генеральная совокупность.

В определении 13 примером двуродной генеральной совокупности служит выборка, состоящая из текстов и изображений в качестве объектов исследования. Современные большие языковые модели, которые одновременно обрабатывают тексты и изображения и называются многомодальными моделями, представляют собой пример работы с многородовыми генеральными совокупностями.

Пусть задана генеральная совокупность данных Γ . Множество всех подмножеств объектов, образующих кольцо выборок, обозначим как:

$$\mathfrak{D} = \{D_\Gamma^i\}, \quad D_\Gamma^i \subset \Gamma.$$

Определение 14 (Мера сложности выборки). Мерой сложности выборки назовем отображение μ_D , такое, что:

$$\mu_D(D_i) : \mathfrak{D}_\Gamma \rightarrow \mathbb{R}_+,$$

удовлетворяющее свойству:

$$\mu_D(D_i \cup D_j) \leq \mu_D(D_i) + \mu_D(D_j),$$

где равенство достигается при условии $D_i \cap D_j = \emptyset$.

Определение 14 является классическим определением из теории меры, удовлетворяющим свойству конечной аддитивности. Конечной аддитивности достаточно, так как в исследованиях предполагается конечное число объектов при обучении моделей глубокого обучения. Предполагается сравнение выборок только из одной генеральной совокупности, однако этим никак не ограничивается сама генеральная совокупность, и в определении возможны мультимодальные генеральные совокупности.

Лемма 1 (*Грабовой*, Монотонность меры сложности выборки). *Мера сложности выборки μ_D , определенная в определении 14, обладает свойством монотонности: для любых выборок $D_1, D_2 \in \mathfrak{D}$ таких, что $D_1 \subseteq D_2$, выполняется неравенство*

$$\mu_D(D_1) \leq \mu_D(D_2).$$

Доказательство. Пусть $D_1 \subseteq D_2$. Представим D_2 в виде объединения $D_2 = D_1 \cup (D_2 \setminus D_1)$, где $D_1 \cap (D_2 \setminus D_1) = \emptyset$. По свойству субаддитивности меры сложности выборки из определения 14 имеем:

$$\mu_D(D_2) = \mu_D(D_1 \cup (D_2 \setminus D_1)) \leq \mu_D(D_1) + \mu_D(D_2 \setminus D_1).$$

Учитывая, что $D_1 \cap (D_2 \setminus D_1) = \emptyset$, по определению 14 равенство в свойстве субаддитивности достигается:

$$\mu_D(D_2) = \mu_D(D_1 \cup (D_2 \setminus D_1)) = \mu_D(D_1) + \mu_D(D_2 \setminus D_1).$$

Поскольку мера сложности выборки принимает неотрицательные значения ($\mu_D : \mathfrak{D}_\Gamma \rightarrow \mathbb{R}_+$), имеем $\mu_D(D_2 \setminus D_1) \geq 0$. Следовательно,

$$\mu_D(D_2) = \mu_D(D_1) + \mu_D(D_2 \setminus D_1) \geq \mu_D(D_1) + 0 = \mu_D(D_1),$$

что и требовалось доказать. □

Замечание 2. *Лемма 1 устанавливает монотонность меры сложности выборки как следствие субаддитивности. Данное свойство является общим свойством мер конечных множеств в теории мер. Однако в случае бесконечных выборок монотонность может не выполняться автоматически и требует более строгих формулировок в определении меры выборок, таких как счетная аддитивность или непрерывность меры снизу. В рамках настоящей работы рассматриваются только конечные выборки, что соответствует практическим задачам обучения моделей глубокого обучения.*

Пусть задано множество параметрических аппроксимирующих моделей

$$\mathfrak{F} = \{f_i\},$$

где каждое f_i является некоторым множеством параметрических функций. В определении 15 вводится характеристика параметрического семейства функций f , которые в дальнейшем рассматриваются в качестве моделей глубокого обучения.

Определение 15 (Мера сложности модели). *Мерой сложности модели f назовем отображение $\mu_f(f_i)$:*

$$\mu_f(f_i) : \mathfrak{F} \rightarrow \mathbb{R}_+.$$

Определение меры сложности модели f не является определением меры в общем случае, так как множество \mathfrak{F} не является кольцом. Поэтому данная мера представляет собой некоторое отображение, которое является характеристикой сложности. В качестве примера можно указать, что число параметров модели удовлетворяет определению 15.

После введения определений меры сложности как для выборки, так и для модели, переходим к определению обучаемости модели на выборке, которое сформулировано в определении 16.

Определение 16 (Обучаемость модели на выборке). *Назовем модель $f \in \mathfrak{F}$ обучаемой на выборке $D \in \mathfrak{D}$, если*

$$\mu_f(f) \leq \mu_D(D).$$

В определении 16 не вводится никакого ограничения на качество аппроксимации модели после обучения. Более подробно это будет определено для частных случаев мер в следующих разделах.

Определение 16 имеет важную эмпирическую интерпретацию: сложность модели не должна превышать сложности данных, на которых она обучается. В противном случае возникает проблема переобучения, когда модель запоминает шум в данных вместо выявления значимых закономерностей. Указанный критерий формализует интуитивное представление о балансе между выразительной способностью модели и информационной емкостью данных.

В диссертационной работе предполагается исследовать частный случай меры сложности модели на основе оценок ландшафта оптимизационных задач, используя матрицы Гессе для различных нейросетевых архитектур, которые описаны в главе 3. Подробнее о частном случае меры сложности см. раздел 2.3..

Теорема 2 (Необходимое условие дообучаемости модели). *Если для исходной выборки $D \in \mathfrak{D}$ выполняется условие $\mu_f(f) \leq \mu_D(D)$, тогда для новой выборки $D' \in \mathfrak{D}$ необходимое условие дообучаемости модели на обединенной выборке $D \cup D'$ имеет вид:*

$$\mu_f(f) - \mu_D(D) \leq \mu_D(D').$$

Доказательство. Доказательство основано на свойствах мер сложности и условии обучаемости модели.

По определению обучаемости модели на выборке D (определение 16) имеем:

$$\mu_f(f) \leq \mu_D(D).$$

При добавлении новых данных D' к исходной выборке D по лемме 1 о монотонности меры сложности выборки получаем:

$$\mu_D(D) \leq \mu_D(D \cup D').$$

Из свойства субаддитивности меры сложности выборки (определение 14) получаем:

$$\mu_D(D \cup D') \leq \mu_D(D) + \mu_D(D'),$$

где равенство достигается при условии $D \cap D' = \emptyset$. Объединяя эти три неравенства, получаем цепочку:

$$\mu_f(f) \leq \mu_D(D) \leq \mu_D(D \cup D') \leq \mu_D(D) + \mu_D(D'),$$

откуда, перенося $\mu_D(D)$ в левую часть, получаем окончательное неравенство:

$$\mu_f(f) - \mu_D(D) \leq \mu_D(D').$$

Для того чтобы модель была обучаемой на объединенной выборке $D \cup D'$, необходимо выполнение условия $\mu_f(f) \leq \mu_D(D \cup D')$. Полученное неравенство $\mu_f(f) - \mu_D(D) \leq \mu_D(D')$ является необходимым для выполнения этого условия, что завершает доказательство. \square

Данное неравенство демонстрирует, что оставшаяся емкость модели, определяемая как разность $\mu_f(f) - \mu_D(D)$ между сложностью модели и сложностью исходных данных, не превосходит сложности новых данных D' . Указанное условие является необходимым для успешного дообучения модели на новых данных: если оставшаяся емкость превышает сложность новых данных, модель не сможет эффективно адаптироваться к расширенной выборке. Достаточность данного условия зависит от конкретного выбора мер сложности и может требовать дополнительных предположений о структуре данных и модели.

Введение формальных мер сложности моделей и данных создает теоретическую основу для решения практических задач проектирования архитектур нейронных сетей, планирования экспериментов и оптимизации процессов обучения.

2.2. Достаточный объем выборки как мера сложности данных

В предыдущем разделе было введено определение обучаемости модели на выборке через условие $\mu_f(f) \leq \mu_D(D)$. Однако данное условие не учитывает важный аспект: одна и та же выборка данных может представлять различную сложность для разных архитектур моделей.

Для учета этой зависимости в настоящем разделе вводится понятие *условной сложности выборки*, которое позволяет формализовать зависимость сложности данных от выбранной модели.

Ключевым понятием становится *условная сложность выборки*:

$$\mu_D(D|f) : \mathfrak{D} \rightarrow \mathbb{R}_+, \quad (2.1)$$

которая характеризует сложность данных $D \in \mathfrak{D}$ относительно заданной параметрической модели f и степень трудности выборки D для обучения модели f .

Мера сложности модели $\mu_f(f)$ индуцирует меру сложности выборки следующим образом:

$$\mu_D(D|f) = \inf\{\mu_D(D') : D' \subseteq D, \mu_f(f) \leq \mu_D(D')\}, \quad (2.2)$$

то есть условная сложность выборки может быть задана как минимальная сложность данных, при которой модель f остается обучаемой.

Определение 17 (*Условной сложностью выборки*). *Условной сложностью выборки D относительно заданной параметрической модели f назовем выражение (2.1), определяемое выражением (2.2).*

Рассмотрим частный случай меры сложности данных μ_D , соответствующий определению достаточного объема выборки. Предположим, что генеральная совокупность Γ_C состоит из объектов одинаковой сложности C , то есть для каждого объекта $\gamma \in \Gamma_C$ выполняется:

$$\mu_D(\gamma) = C,$$

где $C \in \mathbb{R}_+$ — некоторая агрегированная сложность одного объекта выборки. Указанное предположение представляет собой сильное ограничение и может не выполняться на практике, поскольку в реальных задачах различные объекты могут обладать разной сложностью.

Замечание 3. Константа C представляет собой стоимость одного объекта выборки в единицах сложности. На практике C может зависеть от характеристик генеральной совокупности Γ и должна калиброваться экспериментально.

Определение 18 (Простая генеральная совокупность). Однородную генеральную совокупность Γ_C назовем простой, если она состоит из объектов одинаковой сложности C .

Теорема 3 (Грабовой, Мера сложности выборки из простой генеральной совокупности). Для простой генеральной совокупности Γ_C мера сложности любой выборки $D \subset \Gamma_C$ равна ее объему, умноженному на константу сложности:

$$\mu_D(D) = C \cdot |D|.$$

Доказательство. Доказательство состоит из двух частей: сначала покажем, что функция $\mu_D(D) = C \cdot |D|$ является мерой сложности выборки в смысле определения 14, затем покажем, что она единственным образом определяется условием $\mu_D(\gamma) = C$ для каждого объекта $\gamma \in \Gamma_C$.

Проверим, что функция $\mu_D(D) = C \cdot |D|$ удовлетворяет определению меры сложности выборки 14.

Поскольку $|D| \geq 0$ для любой выборки $D \subset \Gamma_C$, и константа $C \in \mathbb{R}_+$ по определению простой генеральной совокупности, имеем $\mu_D(D) = C|D| \geq 0$.

Для любых выборок $D_1, D_2 \subset \Gamma_C$ имеем:

$$\mu_D(D_1 \cup D_2) = C|D_1 \cup D_2|.$$

По свойству мощности множеств $|D_1 \cup D_2| \leq |D_1| + |D_2|$, причем равенство достигается при $D_1 \cap D_2 = \emptyset$. Следовательно,

$$\mu_D(D_1 \cup D_2) = C|D_1 \cup D_2| \leq C(|D_1| + |D_2|) = \mu_D(D_1) + \mu_D(D_2),$$

причем равенство достигается при $D_1 \cap D_2 = \emptyset$, что соответствует определению 14.

Покажем, что мера $\mu_D(D) = C \cdot |D|$ единственным образом определяется условием $\mu_D(\gamma) = C$ для каждого объекта $\gamma \in \Gamma_C$.

Для произвольной выборки $D = \{\gamma_1, \gamma_2, \dots, \gamma_n\} \subset \Gamma_C$ рассмотрим последовательность одноэлементных множеств $D_i = \{\gamma_i\}$ для $i = 1, \dots, n$. По условию простой генеральной совокупности для каждого объекта γ_i выполняется $\mu_D(\gamma_i) = C$, то есть $\mu_D(D_i) = C$ для всех $i = 1, \dots, n$.

Поскольку одноэлементные множества D_i попарно не пересекаются ($D_i \cap D_j = \emptyset$ при $i \neq j$), по свойству конечной аддитивности меры сложности выборки (равенство в определении 14 при непересекающихся множествах) имеем:

$$\mu_D(D) = \mu_D\left(\bigcup_{i=1}^n D_i\right) = \sum_{i=1}^n \mu_D(D_i) = \sum_{i=1}^n C = C \cdot n = C \cdot |D|,$$

что и требовалось доказать.

Таким образом, функция $\mu_D(D) = C \cdot |D|$ является единственной мерой сложности выборки, удовлетворяющей условию простой генеральной совокупности и определению 14. \square

Достаточный объем выборки представляет собой частный случай *условной сложности выборки* — минимальный объем данных из выборки D , необходимый для обучения модели f .

Определение 19 (Достаточный размер выборки). *Размер выборки m^* называется достаточным согласно критерию T , если T выполняется для всех $k \geq m^*$.*

Исследование достаточного объема выборки является частным случаем предложенного определения меры сложности данных. Подробные методы оценки достаточного объема выборки рассматриваются в главе 4.

2.3. Сходимость ландшафта оптимизационной задачи как мера сложности модели

В предыдущих разделах были введены общие определения мер сложности модели и данных, а также условной сложности выборки. Для практического применения этих определений необходимо построить конкретные вычислимые меры сложности. В настоящем разделе разрабатывается ландшафтная мера сложности модели, основанная на анализе сходимости функции потерь при увеличении объема выборки.

Рассмотрим выборку из простой генеральной совокупности Γ_C :

$$D = \{(\mathbf{x}_i, \mathbf{y}_i)\}, \quad i = 1, \dots, m, \quad \mathbf{x} \in \mathcal{X}, \quad \mathbf{y} \in \mathcal{Y}, \quad D \subset \Gamma_C.$$

Рассмотрим некоторое параметрическое отображение $f_{\boldsymbol{\theta}} : \mathcal{X} \rightarrow \mathcal{Y}$, которое аппроксимирует условное распределение целевой переменной для заданного признакового описания объекта $p(\mathbf{y}|\mathbf{x})$. Параметры $\boldsymbol{\theta}$ функции $f_{\boldsymbol{\theta}}$ принадлежат пространству \mathbb{R}^P , где P описывает число параметров отображения $f_{\boldsymbol{\theta}}$.

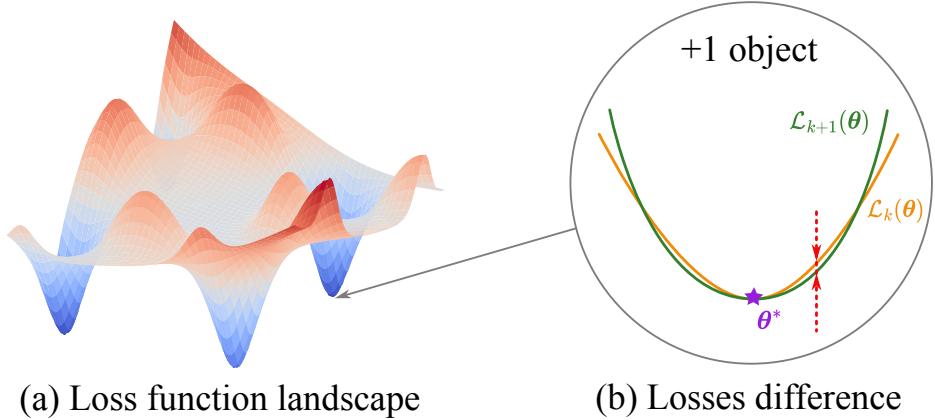


Рис. 2.1: Изменение функции потерь $\mathcal{L}_k(\boldsymbol{\theta})$ при добавлении нового объекта в выборку. Иллюстрация демонстрирует зависимость абсолютной разности $|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})|$ от размера выборки k , что является основой для анализа сходимости ландшафта и определения ландшафтной меры сложности модели.

Пусть для выбора оптимального вектора параметров $\hat{\boldsymbol{\theta}}$ используется подход минимизации эмпирического риска:

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} \mathcal{L}_m(\boldsymbol{\theta}),$$

где функция эмпирического риска для выборки размера $|D| = m$ задается в следующем виде:

$$\mathcal{L}_m(\boldsymbol{\theta}) = \frac{1}{m} \sum_{i=1}^m \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), \mathbf{y}_i) \approx \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim p(\mathbf{x}, \mathbf{y})} [\ell(f_{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{y})],$$

где функция $\ell(\mathbf{z}, \mathbf{y})$ описывает ошибку на одном объекте. Далее в качестве функции ℓ будут рассматриваться либо кросс-энтропийная функция потерь, либо средняя квадратическая ошибка, в зависимости от рассматриваемой задачи и архитектуры модели.

Функция эмпирического риска $\mathcal{L}_m(\boldsymbol{\theta})$ задает некоторую поверхность в пространстве параметров размерности P . Изучение изменения этой поверхности при добавлении новых объектов данных позволяет количественно оценить влияние объема выборки на оптимизационный ландшафт.

Изменение значения функции потерь при добавлении одного объекта вы-

числяется следующим образом:

$$\begin{aligned}
\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta}) &= \frac{1}{k+1} \sum_{i=1}^{k+1} \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), \mathbf{y}_i) - \frac{1}{k} \sum_{i=1}^k \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), \mathbf{y}_i) = \\
&= \frac{1}{k+1} \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_{k+1}), \mathbf{y}_{k+1}) - \sum_{i=1}^k \frac{1}{k(k+1)} \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), \mathbf{y}_i) = \\
&= \frac{1}{k+1} (\ell(f_{\boldsymbol{\theta}}(\mathbf{x}_{k+1}), \mathbf{y}_{k+1}) - \mathcal{L}_k(\boldsymbol{\theta})).
\end{aligned} \tag{2.3}$$

Дальнейшее исследование ландшафта нацелено на изучение данной разницы. Особый интерес представляют предельные свойства при стремлении размера выборки к бесконечности. Для дальнейших оценок данной разности вводится предположение 1, которое подтверждается на практике, однако является достаточно сильным, что упрощает дальнейшие выкладки.

Предположение 1 (Сохранение локальных минимумов). *Пусть $\boldsymbol{\theta}^*$ является локальным минимумом обеих эмпирических функций потерь $\mathcal{L}_k(\boldsymbol{\theta})$ и $\mathcal{L}_{k+1}(\boldsymbol{\theta})$, т.е.*

$$\nabla \mathcal{L}_k(\boldsymbol{\theta}^*) = \nabla \mathcal{L}_{k+1}(\boldsymbol{\theta}^*) = \mathbf{0}.$$

Теоретическое обоснование предположения 1 основано на нескольких фундаментальных принципах теории оптимизации и статистического обучения.

Во-первых, при достаточно большом объеме выборки k эмпирический риск $\mathcal{L}_k(\boldsymbol{\theta})$ сходится к истинному риску $R(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim p(\mathbf{x}, \mathbf{y})} [\ell(f_{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{y})]$ в соответствии с законом больших чисел. При условии равномерной сходимости градиентов эмпирического риска к градиентам истинного риска локальные минимумы эмпирического риска сходятся к локальным минимумам истинного риска [2].

Во-вторых, при добавлении одного объекта к выборке размера k изменение эмпирического риска имеет порядок $O(1/k)$, что следует из выражения (2.3). Для достаточно больших k указанное изменение становится пренебрежимо малым, и локальный минимум $\boldsymbol{\theta}^*$ функции $\mathcal{L}_k(\boldsymbol{\theta})$ остается в окрестности локального минимума функции $\mathcal{L}_{k+1}(\boldsymbol{\theta})$, при условии, что функция потерь является гладкой и локально выпуклой в окрестности $\boldsymbol{\theta}^*$.

В-третьих, работы по анализу ландшафта функции потерь нейронных сетей [40, 51] демонстрируют, что при достаточном объеме данных ландшафт функции потерь стабилизируется, и локальные минимумы становятся устойчивыми к малым возмущениям выборки. Указанное свойство особенно выражено для перепараметризованных моделей, где множество локальных минимумов образует связные многообразия [40].

При асимптотически большом объеме выборки указанное свойство не противоречит эмпирическим результатам, что подтверждается экспериментальной валидацией, представленной в разделе 2.4..

Воспользуемся квадратичным приближением Тейлора для указанных выше функций потерь в окрестности точки $\boldsymbol{\theta}^*$. Предполагается, что разложение до второго порядка является достаточным для изучения локального поведения. Член первого порядка обращается в ноль, поскольку градиенты $\nabla \mathcal{L}_k(\boldsymbol{\theta}^*)$ и $\nabla \mathcal{L}_{k+1}(\boldsymbol{\theta}^*)$ равны нулю:

$$\mathcal{L}_k(\boldsymbol{\theta}) \approx \mathcal{L}_k(\boldsymbol{\theta}^*) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \mathbf{H}^{(k)}(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*), \quad (2.4)$$

где введено обозначение гессиана функции $\mathcal{L}_k(\boldsymbol{\theta})$ по параметрам $\boldsymbol{\theta}$ в точке $\boldsymbol{\theta}^*$ как $\mathbf{H}^{(k)}(\boldsymbol{\theta}^*) \in \mathbb{R}^{P \times P}$. Полный гессиан может быть записан как среднее значение гессианов отдельных членов эмпирической функции потерь:

$$\mathbf{H}^{(k)}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}^2 \mathcal{L}_k(\boldsymbol{\theta}) = \frac{1}{k} \sum_{i=1}^k \nabla_{\boldsymbol{\theta}}^2 \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), \mathbf{y}_i) = \frac{1}{k} \sum_{i=1}^k \mathbf{H}_i(\boldsymbol{\theta}).$$

Используя полученное квадратичное приближение (2.4), формула для разности потерь (2.3) принимает вид:

$$\begin{aligned} \mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta}) &= \frac{1}{k+1} \left(\ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_{k+1}), \mathbf{y}_{k+1}) - \frac{1}{k} \sum_{i=1}^k \ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_i), \mathbf{y}_i) \right) + \\ &\quad + \frac{1}{k+1} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \left(\mathbf{H}_{k+1}(\boldsymbol{\theta}^*) - \frac{1}{k} \sum_{i=1}^k \mathbf{H}_i(\boldsymbol{\theta}^*) \right) (\boldsymbol{\theta} - \boldsymbol{\theta}^*), \end{aligned}$$

Используя неравенство треугольника, получаем следующую оценку:

$$\begin{aligned} |\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})| &\leqslant \frac{1}{k+1} \left| \ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_{k+1}), \mathbf{y}_{k+1}) - \frac{1}{k} \sum_{i=1}^k \ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_i), \mathbf{y}_i) \right| + \quad (2.5) \\ &\quad + \frac{1}{k+1} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 \left\| \mathbf{H}_{k+1}(\boldsymbol{\theta}^*) - \frac{1}{k} \sum_{i=1}^k \mathbf{H}_i(\boldsymbol{\theta}^*) \right\|_2. \end{aligned}$$

Первое слагаемое может быть легко ограничено константой, поскольку сама функция потерь принимает ограниченные значения. Выражение с гессианами требует более сложной оценки. Подробный анализ матриц Гессе для различных типов параметрических моделей глубокого обучения представлен в главе 3. Анализ локальной сходимости ландшафта функции потерь основан на ее матрице Гессе.

Получаем выражение для анализа, описывающее поведение ландшафта функции потерь:

$$|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})| \leq \frac{M_\ell}{k+1} + \frac{1}{k+1} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 \left\| \mathbf{H}_{k+1}(\boldsymbol{\theta}^*) - \frac{1}{k} \sum_{i=1}^k \mathbf{H}_i(\boldsymbol{\theta}^*) \right\|_2. \quad (2.6)$$

Установлено, что анализ сходимости ландшафта оптимизационной задачи сводится к анализу нормы матрицы Гессе, который подробно разобран в главе 3.

Оценка (2.6) задает некоторое свойство параметрического семейства функций f на заданной выборке D . Определим данное свойство как условную сложность модели f на выборке D :

$$\mu_f(f|D) : \mathfrak{F} \rightarrow \mathbb{R}_+, \quad (2.7)$$

Более подробно рассмотрим частный случай условной меры сложности параметрического семейства функций f вида:

$$\mu_f(f|D) = \mathsf{E}_{\mathbf{x}_i \in D} \|\mathbf{H}_i(\boldsymbol{\theta}^*) - \mathsf{E}_{\mathbf{x}_i \in D} \mathbf{H}_i(\boldsymbol{\theta}^*)\|_2. \quad (2.8)$$

Определение 20 (Условная сложность параметрической модели). *Условной сложностью параметрической модели f относительно заданной выборки D назовем отображение (2.7).*

Определение 21 (Ландшафтная мера сложности). *Ландшафтной мерой сложности параметрической функции f назовем условную сложность параметрической модели f , заданную выражением (2.8).*

Определение 20 описывает прикладный способ задания сложности на параметрических семействах функций в контексте оптимизации на заданных выборках. Условная сложность модели $\mu_f(f|D)$ характеризует сложность архитектуры модели f при ее обучении на выборке данных D и позволяет количественно оценить степень соответствия модели данным. При этом слишком простая модель может недообучаться, а слишком сложная — переобучаться.

Ландшафтная мера сложности 21 представляет собой явный вид условной сложности, основанный на анализе оптимизационного ландшафта функции потерь. Выражение (2.8) содержательно указывает на степень изменения кривизны функции потерь в окрестности оптимума при добавлении нового объекта данных.

Дальнейшее изложение в главе посвящено оценкам ландшафтной меры для различных параметрических моделей f . Все результаты основываются на анализе матриц Гессе, описанных в главе 3.

Используя выражение (2.6) и определение ландшафтной меры сложности, получаем следующую асимптотическую связь между этими оценками:

$$|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})| \leq \frac{M_\ell}{k+1} + \frac{\mu_f(f|D)}{k+1} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2. \quad (2.9)$$

Лемма 4 (*Грабовой*, Связь ландшафтной меры и условной сложности выборки). Пусть задан некоторый ε , описывающий допустимое изменение ландшафта при добавлении одного объекта в некоторой окрестности оптимума $\boldsymbol{\theta}^*$ радиуса R , причем выборка D принадлежит простой генеральной совокупности Γ_C .

Тогда верно следующее соотношение между ландшафтной мерой $\mu_f(f|D)$ и условной сложностью выборки $\mu_D(D|f)$:

$$\mu_f(f|D) \geq \mu_D(D|f) \frac{\varepsilon}{CR^2} - \frac{M_\ell}{R^2}.$$

Доказательство. Доказательство основано на связи между условной сложностью выборки и ландшафтной мерой сложности модели через анализ изменения функции потерь.

Рассмотрим произвольную подвыборку $D' \subseteq D$, удовлетворяющую условию $\mu_f(f) \leq \mu_D(D')$, и пусть $k = |D'|$. По определению условной сложности выборки (2.2) такая подвыборка существует, так как $\mu_D(D|f) = \inf\{\mu_D(D') : D' \subseteq D, \mu_f(f) \leq \mu_D(D')\}$.

Из выражения (2.9) для подвыборки D' размера k в окрестности оптимума $\boldsymbol{\theta}^*$ радиуса R имеем:

$$|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})| \leq \frac{M_\ell}{k+1} + \frac{\mu_f(f|D')}{k+1} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2.$$

По условию леммы, допустимое изменение ландшафта при добавлении одного объекта в окрестности оптимума радиуса R не превосходит ε . Следовательно, для любой точки $\boldsymbol{\theta}$ такой, что $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 \leq R^2$, выполняется:

$$\frac{M_\ell}{k+1} + \frac{\mu_f(f|D')}{k+1} R^2 \geq \varepsilon.$$

Переписывая это неравенство относительно $\mu_f(f|D')$, получаем:

$$\mu_f(f|D') \geq \frac{\varepsilon(k+1) - M_\ell}{R^2} = \frac{\varepsilon k}{R^2} + \frac{\varepsilon - M_\ell}{R^2}.$$

Поскольку выборка D принадлежит простой генеральной совокупности Γ_C , по определению простой генеральной совокупности (см. раздел 2.2.) для любой подвыборки D' выполняется $\mu_D(D') = C \cdot |D'| = C \cdot k$. Следовательно, $k = \frac{\mu_D(D')}{C}$.

Подставляя $k = \frac{\mu_D(D')}{C}$ в полученное неравенство, имеем:

$$\mu_f(f|D') \geq \frac{\varepsilon \mu_D(D')}{CR^2} + \frac{\varepsilon - M_\ell}{R^2} = \mu_D(D') \frac{\varepsilon}{CR^2} + \frac{\varepsilon - M_\ell}{R^2}.$$

Поскольку данное неравенство выполняется для любой подвыборки $D' \subseteq D$, удовлетворяющей условию $\mu_f(f) \leq \mu_D(D')$, и учитывая, что $\mu_D(D|f) = \inf\{\mu_D(D') : D' \subseteq D, \mu_f(f) \leq \mu_D(D')\}$, для любой такой подвыборки D' имеем:

$$\mu_f(f|D') \geq \mu_D(D') \frac{\varepsilon}{CR^2} + \frac{\varepsilon - M_\ell}{R^2} \geq \mu_D(D|f) \frac{\varepsilon}{CR^2} + \frac{\varepsilon - M_\ell}{R^2},$$

где последнее неравенство следует из того, что $\mu_D(D') \geq \mu_D(D|f)$ по определению инфимума.

Рассмотрим теперь саму выборку D . Если $\mu_f(f) \leq \mu_D(D)$, то D входит в множество подвыборок, для которых выполняется указанное неравенство, и мы получаем:

$$\mu_f(f|D) \geq \mu_D(D) \frac{\varepsilon}{CR^2} + \frac{\varepsilon - M_\ell}{R^2} \geq \mu_D(D|f) \frac{\varepsilon}{CR^2} + \frac{\varepsilon - M_\ell}{R^2},$$

где последнее неравенство следует из того, что $\mu_D(D) \geq \mu_D(D|f)$ по определению инфимума.

Если же $\mu_f(f) > \mu_D(D)$, то по определению условной сложности выборки $\mu_D(D|f) = \inf\{\mu_D(D') : D' \subseteq D, \mu_f(f) \leq \mu_D(D')\}$ может быть больше $\mu_D(D)$ или бесконечным. В этом случае неравенство тривиально выполняется, так как правая часть может быть отрицательной или неограниченной, а левая часть $\mu_f(f|D)$ неотрицательна по определению.

Учитывая, что $\varepsilon - M_\ell \geq -M_\ell$, в обоих случаях получаем окончательное неравенство:

$$\mu_f(f|D) \geq \mu_D(D|f) \frac{\varepsilon}{CR^2} - \frac{M_\ell}{R^2},$$

что и требовалось доказать. \square

2.3.1. Полносвязная нейросетевая модель глубокого обучения

В настоящем подразделе получены оценки ландшафтной меры сложности для полносвязных нейронных сетей, являющихся базовой архитектурой глубокого обучения. Анализ основан на результатах главы 3, а именно на теореме 10, в которой показана асимптотика нормы матрицы Гессе от гиперпараметров полносвязной нейросетевой модели:

$$\|\mathbf{H}_i(\boldsymbol{\theta})\|_2 \propto L(hM)^{2L},$$

из которой следует, что спектральная норма матрицы Гессе имеет полиномиальную зависимость от размера слоя и экспоненциальную зависимость от числа слоев.

Теорема 5 (*Грабовой-Киселев*, Сходимость ландшафта функции потерь для полно связных сетей). *Пусть параметры $\boldsymbol{\theta}$ выбраны так, что $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 \leq R^2$ для некоторого $R > 0$. Если существует неотрицательная константа M_ℓ такая, что $|\ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_i), \mathbf{y}_i)| \leq M_\ell$ для всех объектов $i = 1, \dots, m$ в наборе данных, то при выполнении условий Теоремы 9 справедливо:*

$$|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})| \leq \frac{2}{k+1} \left(M_\ell + \left(L\sqrt{2}M_{\mathbf{x}}^2 M_{\mathbf{W}}^{2L} + \sqrt{2} \frac{M_{\mathbf{W}}^2(M_{\mathbf{W}}^{2L} - 1)}{M_{\mathbf{W}}^2 - 1} \right) R^2 \right),$$

причем выражение асимптотически стремится к 0, то есть

$$|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})| \rightarrow 0 \text{ при } k \rightarrow \infty.$$

Таким образом, имеет место следующая пропорциональность:

$$|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})| \propto \frac{L(hM)^{2L}R^2}{k}.$$

Доказательство. Доказательство основано на оценке разности функций потерь через выражение (2.5), которое получено из квадратичного приближения Тейлора и предположения 1.

Оценим первое слагаемое в выражении (2.5), используя неравенство треугольника для модуля разности:

$$\begin{aligned} & \left| \ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_{k+1}), \mathbf{y}_{k+1}) - \frac{1}{k} \sum_{i=1}^k \ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_i), \mathbf{y}_i) \right| \leq \\ & \leq |\ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_{k+1}), \mathbf{y}_{k+1})| + \left| \frac{1}{k} \sum_{i=1}^k \ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_i), \mathbf{y}_i) \right|. \end{aligned}$$

Применяя неравенство треугольника к модулю суммы и используя условие ограниченности функции потерь $|\ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_i), \mathbf{y}_i)| \leq M_\ell$ для всех $i = 1, \dots, m$, получаем:

$$\begin{aligned} & \leq |\ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_{k+1}), \mathbf{y}_{k+1})| + \frac{1}{k} \sum_{i=1}^k |\ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_i), \mathbf{y}_i)| \leq \\ & \leq M_\ell + \frac{1}{k} \sum_{i=1}^k M_\ell = M_\ell + M_\ell = 2M_\ell. \end{aligned}$$

Таким образом, первое слагаемое ограничено константой $2M_\ell = \mathcal{O}(1)$ при $k \rightarrow \infty$.

Оценим норму разности матриц Гессе во втором слагаемом, используя свойства спектральной нормы матриц. Применяя неравенство треугольника для спектральной нормы, получаем:

$$\begin{aligned} & \left\| \mathbf{H}_{k+1}(\boldsymbol{\theta}^*) - \frac{1}{k} \sum_{i=1}^k \mathbf{H}_i(\boldsymbol{\theta}^*) \right\|_2 \leqslant \\ & \leqslant \|\mathbf{H}_{k+1}(\boldsymbol{\theta}^*)\|_2 + \left\| \frac{1}{k} \sum_{i=1}^k \mathbf{H}_i(\boldsymbol{\theta}^*) \right\|_2. \end{aligned}$$

Используя свойство субмультипликативности спектральной нормы и неравенство треугольника для суммы матриц, имеем:

$$\leqslant \|\mathbf{H}_{k+1}(\boldsymbol{\theta}^*)\|_2 + \frac{1}{k} \sum_{i=1}^k \|\mathbf{H}_i(\boldsymbol{\theta}^*)\|_2.$$

По теореме 9 о норме матрицы Гессе для полносвязных нейронных сетей, при выполнении условий теоремы существует константа $M_{\mathbf{H}}$ такая, что $\|\mathbf{H}_i(\boldsymbol{\theta}^*)\|_2 \leqslant M_{\mathbf{H}}$ для всех $i = 1, \dots, k+1$, где

$$M_{\mathbf{H}} = L\sqrt{2}M_{\mathbf{x}}^2M_{\mathbf{W}}^{2L} + \sqrt{2}\frac{M_{\mathbf{W}}^2(M_{\mathbf{W}}^{2L}-1)}{M_{\mathbf{W}}^2-1}.$$

Следовательно,

$$\left\| \mathbf{H}_{k+1}(\boldsymbol{\theta}^*) - \frac{1}{k} \sum_{i=1}^k \mathbf{H}_i(\boldsymbol{\theta}^*) \right\|_2 \leqslant M_{\mathbf{H}} + \frac{1}{k} \sum_{i=1}^k M_{\mathbf{H}} = M_{\mathbf{H}} + M_{\mathbf{H}} = 2M_{\mathbf{H}}.$$

Таким образом, второе слагаемое ограничено константой $2M_{\mathbf{H}} = \mathcal{O}(1)$ при $k \rightarrow \infty$.

Подставляя полученные оценки в выражение (2.5), получаем:

$$|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})| \leqslant \frac{2M_\ell}{k+1} + \frac{2M_{\mathbf{H}}}{k+1} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2.$$

По условию теоремы параметры $\boldsymbol{\theta}$ находятся в R -окрестности оптимума, то есть $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 \leqslant R^2$. Следовательно,

$$|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})| \leqslant \frac{2}{k+1} (M_\ell + M_{\mathbf{H}}R^2).$$

Подставляя выражение для $M_{\mathbf{H}}$, получаем требуемую оценку:

$$|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})| \leq \frac{2}{k+1} \left(M_\ell + \left(L\sqrt{2}M_{\mathbf{x}}^2 M_{\mathbf{W}}^{2L} + \sqrt{2} \frac{M_{\mathbf{W}}^2(M_{\mathbf{W}}^{2L} - 1)}{M_{\mathbf{W}}^2 - 1} \right) R^2 \right).$$

Поскольку M_ℓ и $M_{\mathbf{H}}$ являются константами, не зависящими от k , имеем:

$$|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})| \leq \frac{2(M_\ell + M_{\mathbf{H}}R^2)}{k+1} = \mathcal{O}\left(\frac{1}{k}\right) \rightarrow 0 \text{ при } k \rightarrow \infty,$$

что доказывает асимптотическую сходимость к нулю.

Из выражения для $M_{\mathbf{H}}$ и асимптотики нормы матрицы Гессе $\|\mathbf{H}_i(\boldsymbol{\theta})\|_2 \propto L(hM)^{2L}$ (см. теорему 9) следует, что $M_{\mathbf{H}} \propto L(hM)^{2L}$, где h — размер скрытого слоя, а M — константа, ограничивающая параметры и данные. Следовательно,

$$|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})| \propto \frac{L(hM)^{2L}R^2}{k},$$

что завершает доказательство. \square

На основе теоремы 5 и использованных в ее доказательстве оценок нормы матрицы Гессе получается выражение для ландшафтной меры полносвязной нейросетевой модели глубокого обучения, которое сформулировано в следствии 1.

Следствие 1 (Асимптотика ландшафтной меры для полносвязных сетей). *Ландшафтная мера сложности параметрической функции f полносвязной нейросетевой модели глубокого обучения имеет асимптотику:*

$$\mu_f(f|D) \propto L(hM)^{2L},$$

где M — некоторая константа, ограничивающая параметры и данные.

Теорема 5 и следствие 1 устанавливают, что сходимость ландшафта функции потерь для полносвязных сетей происходит со скоростью $O(L(hM)^{2L}/k)$ при увеличении размера выборки k . Это означает, что для более глубоких сетей требуется больше данных для стабилизации ландшафта, что согласуется с интуитивным представлением о том, что сложные модели требуют большего объема данных для обучения.

В работе [11] показано, что существуют функции, которые могут быть эффективно представлены трехслойными сетями, но требуют экспоненциального числа нейронов в двухслойных сетях. Следствие 1 указывает на схожий результат: сложность модели увеличивается экспоненциально при увеличении числа

слоев ($\mu_f(f|D) \propto L(hM)^{2L}$), и, следовательно, при уменьшении числа слоев для сохранения заданной сложности модели потребуется экспоненциальный рост параметров. Это демонстрирует фундаментальный компромисс между глубиной и шириной нейронных сетей с точки зрения их выразительной способности.

2.3.2. Сверточные модели глубокого обучения

В настоящем подразделе получены оценки ландшафтной меры сложности для сверточных нейронных сетей, широко применяющихся в задачах обработки последовательностей и изображений. Анализ ландшафта основан на результатах о матрицах Гессе, представленных в главе 3.

Начнем с анализа 1D-сверточных сетей, применяющихся в обработке последовательностей, временных рядов и сигналов. Основные результаты для определения ландшафтной меры сложности сформулированы в теореме 6.

Теорема 6 (Грабовой-Киселев-Мешков, Сходимость ландшафта для 1D-сверточных сетей). *Пусть параметры $\boldsymbol{\theta}$ находятся в R -окрестности оптимума:*

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq R,$$

а функция потерь ограничена некоторой константой:

$$\exists M_\ell > 0 : \forall i |\ell_i| \leq M_\ell.$$

Пусть все объекты в наборе данных ограничены:

$$\exists M_{\mathbf{x}} > 0 \forall i \|\mathbf{x}_i\| \leq M_{\mathbf{x}}.$$

Тогда в условиях теоремы 13:

$$\begin{aligned} |\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})| &\leq \frac{2}{k+1} M_\ell + \\ &+ \frac{2}{k+1} R^2 \sqrt{2} d^2 M_{\mathbf{x}}^2 (L+1) (C^2 w^2 k d)^L, \end{aligned}$$

где $M_{\mathbf{x}}$ — константа, ограничивающая нормы объектов данных.

Доказательство. Доказательство проводится аналогично доказательству теоремы 5 с подстановкой оценок нормы матрицы Гессе из теоремы 13. \square

На основе теоремы 6 и использованных в ее доказательстве оценок нормы матрицы Гессе получается выражение для ландшафтной меры 1D-сверточной нейросетевой модели, которое сформулировано в следствии 2.

Следствие 2 (Асимптотика ландшафтной меры для 1D-сверточных сетей). *Ландшафтная мера сложности параметрической функции f 1D-сверточной модели глубокого обучения имеет асимптотику:*

$$\mu_f(f|D) \propto L(C^2 M^2 k d)^L,$$

где C — максимальное число каналов, d — длина входной последовательности, k — размер свертки, M — некоторая константа, ограничивающая параметры и данные.

Теорема 6 и следствие 2 показывают, что для 1D-сверточных сетей сходимость ландшафта происходит со скоростью $O(L(C^2 M^2 k d)^L/k)$. Полученные оценки демонстрируют, что сложность 1D-сверточных нейросетевых моделей экспоненциально зависит от глубины L и полиномиально — от остальных гиперпараметров архитектуры. Особенностью 1D-архитектур является линейная зависимость от длины последовательности d , что отражает специфику обработки последовательностей и отличает их от полно связанных сетей, где такая зависимость отсутствует.

Перейдем к анализу 2D-сверточных сетей, применяющихся в задачах обработки изображений.

Основные результаты для определения ландшафтной меры сложности 2D-сверточных сетей сформулированы в теореме 7.

Теорема 7 (Грабовой-Киселев-Мешков, Сходимость ландшафта для 2D-сверточных сетей). *Пусть параметры $\boldsymbol{\theta}$ находятся в R -окрестности оптимума:*

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq R.$$

Также функция потерь ограничена некоторой константой:

$$\exists M_\ell > 0 : \forall i |\ell_i| \leq M_\ell.$$

Пусть все объекты в наборе данных также ограничены:

$$\exists M_{\mathbf{x}} > 0 \forall i \|\mathbf{x}_i\| \leq M_{\mathbf{x}}.$$

Тогда при выполнении условий теоремы 14 справедливо:

$$\begin{aligned} |\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})| &\leq \frac{2}{k+1} M_\ell + \\ &+ \frac{2}{k+1} R^2 \sqrt{2} q^2 M_{\mathbf{x}}^2 (L+1) (C^2 k^2 w^2 m n)^L, \end{aligned}$$

где $q^2 = C^2 k^2 m n$, а $M_{\mathbf{x}}$ — константа, ограничивающая нормы объектов данных.

Доказательство. Доказательство проводится аналогично доказательству теоремы 5 с подстановкой оценок нормы матрицы Гессе из теоремы 14. \square

На основе теоремы 7 и использованных в ее доказательстве оценок нормы матрицы Гессе получается выражение для ландшафтной меры 2D-сверточной нейросетевой модели, которое сформулировано в следствии 3.

Следствие 3 (Асимптотика ландшафтной меры для 2D-сверточных сетей). *Ландшафтная мера сложности параметрической функции f 2D-сверточной модели глубокого обучения имеет асимптотику:*

$$\mu_f(f|D) \propto C^2 k^2 L (C^2 k^2 M^2 mn)^L,$$

где C — максимальное число каналов, m, n — размеры входного изображения, k — размер свертки, M — некоторая константа, ограничивающая параметры и данные.

Теорема 7 и следствие 3 устанавливают, что для 2D-сверточных сетей сходимость ландшафта происходит со скоростью $O(C^2 k^2 L (C^2 k^2 M^2 mn)^L / k)$. Для 2D-сверточных сетей наблюдается более быстрый рост сложности по сравнению с 1D-архитектурами, что обусловлено двумерной природой данных: зависимость от размеров изображения $m \times n$ является квадратичной, в отличие от линейной зависимости от длины последовательности в 1D-сверточных сетях.

2.3.3. Трансформер модели глубокого обучения

Архитектура трансформеров представляет собой один из наиболее значительных прорывов в области глубокого обучения последних лет. Указанные модели демонстрируют state-of-the-art результаты в задачах обработки естественного языка, компьютерного зрения и других областях. Особенностью трансформеров является механизм самовнимания, который позволяет модели учитывать глобальные зависимости в данных независимо от их положения. В настоящем подразделе получены оценки ландшафтной меры сложности для трансформерных моделей, основанные на анализе матриц Гессе компонентов трансформера.

Теорема 8 (Грабовой-Киселев-Петров, Сходимость ландшафта для трансформеров). *Для одного блока самовнимания и одного блока трансформера 3.10 при условии ограниченности функции потерь*

$$0 \leq \ell(\mathbf{f}_{\mathbf{w}^*}(\mathbf{x}_i), \mathbf{y}_i) \leq L,$$

и ограниченности норм матриц Гессе справедливо:

$$|\mathcal{L}_{k+1}(\mathbf{w}) - \mathcal{L}_k(\mathbf{w})| \leq \frac{2L}{k+1} + \frac{M \|\mathbf{w} - \mathbf{w}^*\|_2^2}{(k+1)},$$

где для блока самовнимания константа M может быть непосредственно вычислена из Теоремы 18, а для блока трансформера $M = M_{tr}$ вычисляется в соответствии с Теоремой 27.

Доказательство. Доказательство проводится в несколько этапов. На первом этапе оценивается разность значений функции потерь в оптимальной точке, на втором этапе оценивается разность гессианов. На третьем этапе, используя результаты обоих этапов, производится объединение оценок.

Рассмотрим разность эмпирических функций потерь при добавлении нового объекта. Используя разложение разности и свойства норм, получаем:

$$\begin{aligned} |\mathcal{L}_{k+1}(\mathbf{w}) - \mathcal{L}_k(\mathbf{w})| &\leq \frac{1}{k+1} \left| \ell(\mathbf{f}_{\mathbf{w}^*}(\mathbf{x}_{k+1}), \mathbf{y}_{k+1}) - \frac{1}{k} \sum_{i=1}^k \ell(\mathbf{f}_{\mathbf{w}^*}(\mathbf{x}_i), \mathbf{y}_i) \right| + \\ &+ \frac{1}{2(k+1)} \|\mathbf{w} - \mathbf{w}^*\|_2^2 \left\| \mathbf{H}_{k+1}(\mathbf{w}^*) - \frac{1}{k} \sum_{i=1}^k \mathbf{H}_i(\mathbf{w}^*) \right\|_2. \end{aligned}$$

Первое слагаемое характеризует изменение значения функции потерь в оптимальной точке параметров \mathbf{w}^* при добавлении нового объекта. Это разность между значением потерь на новом объекте и средним значением потерь на предыдущей выборке, до добавления нового объекта. Предположим, что функция потерь $\ell(\mathbf{f}_{\mathbf{w}^*}(\mathbf{x}_i), \mathbf{y}_i)$ ограничена сверху константой L для всех объектов выборки:

$$0 \leq \ell(\mathbf{f}_{\mathbf{w}^*}(\mathbf{x}_i), \mathbf{y}_i) \leq L.$$

Это предположение является естественным для большинства функций потерь, используемых в машинном обучении, таких как кросс-энтропия или среднеквадратичная ошибка. Тогда для нового объекта выполняется:

$$\ell(\mathbf{f}_{\mathbf{w}^*}(\mathbf{x}_{k+1}), \mathbf{y}_{k+1}) \leq L,$$

а для среднего значения по предыдущей выборке:

$$\frac{1}{k} \sum_{i=1}^k \ell(\mathbf{f}_{\mathbf{w}^*}(\mathbf{x}_i), \mathbf{y}_i) \leq \frac{1}{k} \sum_{i=1}^k L = L.$$

Используя неравенство треугольника для модуля разности, получаем:

$$\left| \ell(\mathbf{f}_{\mathbf{w}^*}(\mathbf{x}_{k+1}), \mathbf{y}_{k+1}) - \frac{1}{k} \sum_{i=1}^k \ell(\mathbf{f}_{\mathbf{w}^*}(\mathbf{x}_i), \mathbf{y}_i) \right| \leq L + L = 2L.$$

Таким образом, вклад первого слагаемого в общую оценку не превосходит:

$$\frac{1}{k+1} \left| \ell(\mathbf{f}_{\mathbf{w}^*}(\mathbf{x}_{k+1}), \mathbf{y}_{k+1}) - \frac{1}{k} \sum_{i=1}^k \ell(\mathbf{f}_{\mathbf{w}^*}(\mathbf{x}_i), \mathbf{y}_i) \right| \leq \frac{2L}{k+1}.$$

Второе слагаемое в оценке связано с изменением гессиана функции потерь.
Рассмотрим выражение:

$$\left\| \mathbf{H}_{k+1}(\mathbf{w}^*) - \frac{1}{k} \sum_{i=1}^k \mathbf{H}_i(\mathbf{w}^*) \right\|_2,$$

где $\mathbf{H}_{k+1}(\mathbf{w}^*) = \nabla_{\mathbf{w}}^2 \ell(\mathbf{f}_{\mathbf{w}^*}(\mathbf{x}_{k+1}), \mathbf{y}_{k+1})$ — матрица Гессе функции потерь для нового объекта, а $\frac{1}{k} \sum_{i=1}^k \mathbf{H}_i(\mathbf{w}^*) = \mathbf{H}_k(\mathbf{w}^*)$ — средняя матрица Гессе по всей предыдущей выборке. Перепишем это выражение в более удобной форме:

$$\begin{aligned} \mathbf{H}_k(\mathbf{w}^*) &= \frac{1}{k} \sum_{i=1}^k \mathbf{H}_i(\mathbf{w}^*), \\ \mathbf{H}_{k+1}(\mathbf{w}^*) - \mathbf{H}_k(\mathbf{w}^*) &= \mathbf{H}_{k+1}(\mathbf{w}^*) - \frac{1}{k} \sum_{i=1}^k \mathbf{H}_i(\mathbf{w}^*). \end{aligned}$$

Для оценки нормы этой разности используем неравенство треугольника:

$$\left\| \mathbf{H}_{k+1}(\mathbf{w}^*) - \frac{1}{k} \sum_{i=1}^k \mathbf{H}_i(\mathbf{w}^*) \right\|_2 \leq \|\mathbf{H}_{k+1}(\mathbf{w}^*)\|_2 + \left\| \frac{1}{k} \sum_{i=1}^k \mathbf{H}_i(\mathbf{w}^*) \right\|_2.$$

Предположим, что выполняется ограниченность следующих матриц Гессе:

$$\|\mathbf{H}_i(\mathbf{w}^*)\|_2 \leq M$$

для некоторой константы M . Тогда для гессиана нового объекта:

$$\|\mathbf{H}_{k+1}(\mathbf{w}^*)\|_2 \leq M,$$

а для суммы гессианов:

$$\left\| \sum_{i=1}^k \mathbf{H}_i(\mathbf{w}^*) \right\|_2 \leq \sum_{i=1}^k \|\mathbf{H}_i(\mathbf{w}^*)\|_2 \leq kM.$$

Следовательно:

$$\left\| \frac{1}{k} \sum_{i=1}^k \mathbf{H}_i(\mathbf{w}^*) \right\|_2 \leqslant \frac{1}{k} \cdot kM = M.$$

Объединяя полученные оценки, получаем:

$$\left\| \mathbf{H}_{k+1}(\mathbf{w}^*) - \frac{1}{k} \sum_{i=1}^k \mathbf{H}_i(\mathbf{w}^*) \right\|_2 \leqslant M + M = 2M.$$

Теперь оценим вклад второго слагаемого в общую разность функций потерь:

$$\begin{aligned} \frac{1}{2(k+1)} \|\mathbf{w} - \mathbf{w}^*\|_2^2 \|\mathbf{H}_{k+1}(\mathbf{w}^*) - \mathbf{H}_k(\mathbf{w}^*)\|_2 &\leqslant \frac{2M}{2(k+1)} \|\mathbf{w} - \mathbf{w}^*\|_2^2 = \\ &= \frac{M \|\mathbf{w} - \mathbf{w}^*\|_2^2}{k+1}. \end{aligned}$$

Комбинируя оценки для обоих слагаемых, получаем итоговую оценку:

$$|\mathcal{L}_{k+1}(\mathbf{w}) - \mathcal{L}_k(\mathbf{w})| \leqslant \frac{2L}{k+1} + \frac{M \|\mathbf{w} - \mathbf{w}^*\|_2^2}{k+1}.$$

□

2.4. Результаты вычислительных экспериментов

В настоящем разделе представлены результаты вычислительных экспериментов, направленных на эмпирическую валидацию теоретических результатов, полученных в предыдущих разделах главы.

На рис. 2.2 представлены соответствующие результаты, показывающие, что хотя предположение 1 может быть ослаблено, его выполнимость улучшается с увеличением длины последовательностей.

2.4.1. Полносвязная нейросетевая модель глубокого обучения

Для проверки полученных теоретических оценок проведен вычислительный эксперимент. В настоящем разделе представлены результаты обучения полносвязной нейронной сети для задачи классификации изображений. Основной целью экспериментов является эмпирическое подтверждение сходимости ландшафта функции потерь с увеличением размера выборки. Для достижения этой цели обучена полносвязная нейронная сеть на полном наборе данных и получены соответствующие параметры $\hat{\theta}$ как точка вблизи минимума.

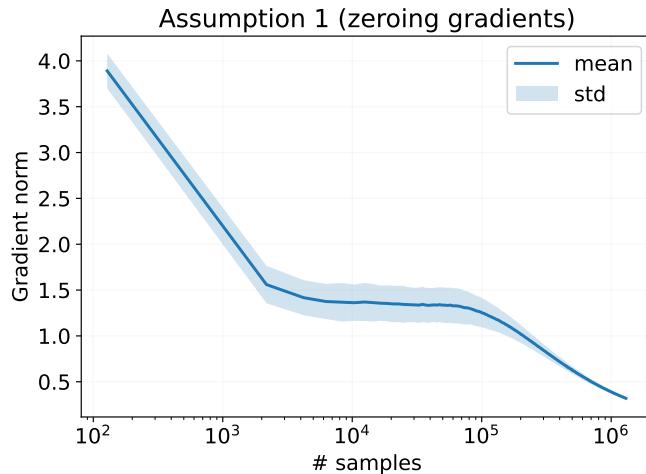


Рис. 2.2: Проверка предположения 1 о сходимости локальных минимумов при увеличении объема выборки. График демонстрирует выполнимость предположения о том, что локальный минимум $\boldsymbol{\theta}^*$ функции потерь $\mathcal{L}_k(\boldsymbol{\theta})$ остается локальным минимумом функции $\mathcal{L}_{k+1}(\boldsymbol{\theta})$ при добавлении нового объекта, причем выполнимость улучшается с увеличением длины последовательностей.

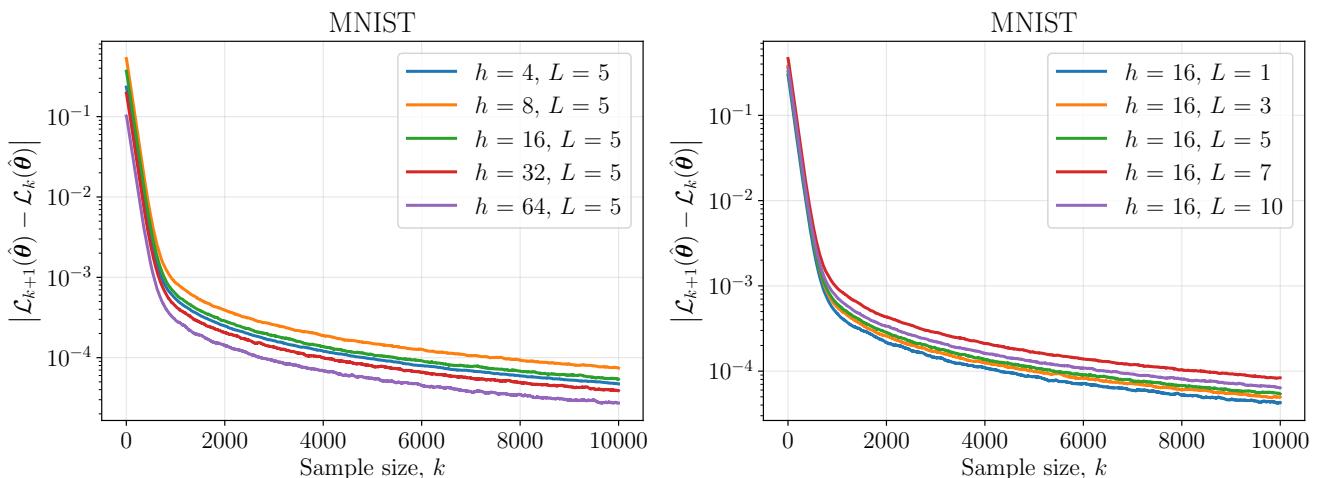


Рис. 2.3: Зависимость абсолютного значения разности функций потерь $|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})|$ от размера выборки k для полносвязной нейронной сети на наборе данных MNIST. Левый график: при $L = 5$ уменьшение разности с увеличением размера скрытого слоя h от 4 до 64; правый график: при $h = 16$ увеличение разности с увеличением количества слоев L от 1 до 10. Результаты подтверждают теорему 5.

Для вычислительного эксперимента использовалась библиотека `pytorch` [56] в качестве Python-фреймворка для обучения нейронных сетей. Архитектура сети была единообразной и состояла из нескольких линейных слоев с функцией

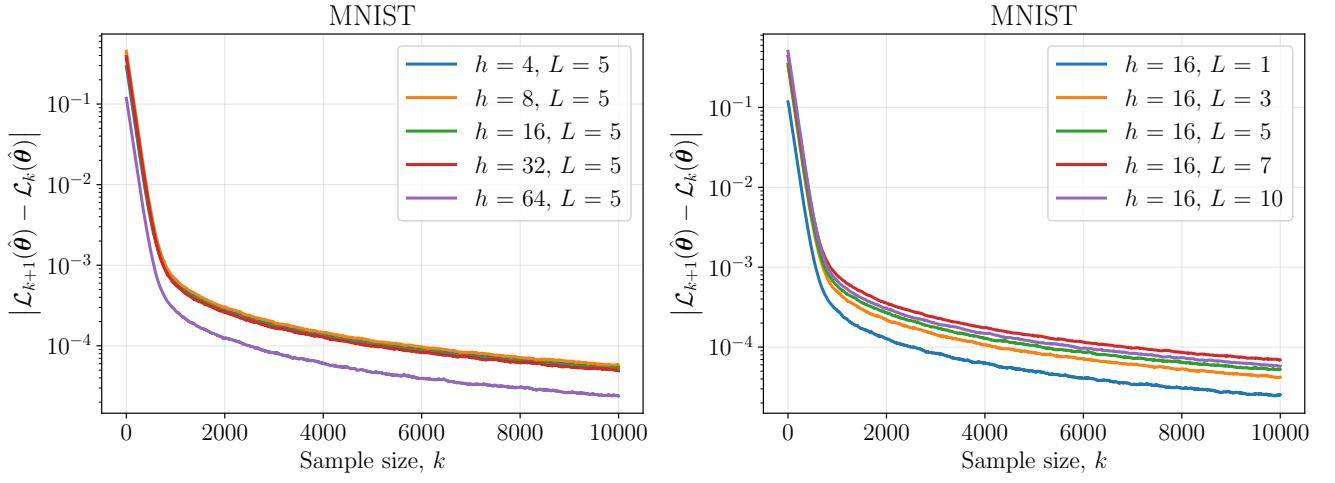


Рис. 2.4: Зависимость абсолютного значения разности функций потерь $|\mathcal{L}_{k+1}(\hat{\theta}) - \mathcal{L}_k(\hat{\theta})|$ от размера выборки k для полносвязной нейронной сети с предобученным экстрактором признаков (Vision Transformer, ViT) на наборе данных MNIST. Левый график: при $L = 5$ уменьшение разности с увеличением h от 4 до 64; правый график: при $h = 16$ увеличение разности с увеличением L от 1 до 10. Результаты подтверждают независимость сходимости от природы пространства исходных объектов и согласуются с теоремой 5.

Таблица 2.1: Описание наборов данных для классификации изображений, использованных в экспериментах по валидации теоретических оценок сходимости ландшафта функции потерь. Все наборы данных из библиотеки `torchvision` с нормализацией значений пикселей к диапазону $[-1, 1]$.

Название	Описание	Формат	Разрешение
MNIST [53]	Рукописные цифры	Оттенки серого	28×28
FashionMNIST [54]	Элементы одежды	Оттенки серого	28×28
CIFAR10 [55]	Различные объекты	RGB	32×32
CIFAR100 [55]	Различные объекты	RGB	32×32

активации ReLU после каждого слоя, за исключением последнего. Размер h был зафиксирован для всех скрытых слоев L . Обучение сети проводилось в течение нескольких эпох с использованием оптимизатора Adam [57] с постоянной скоростью обучения 10^{-3} . Для обучения использовались различные наборы данных для классификации изображений, доступные в библиотеке `torchvision`. Для процесса обучения был выбран размер батча (англ. batch size) 64. Экспе-

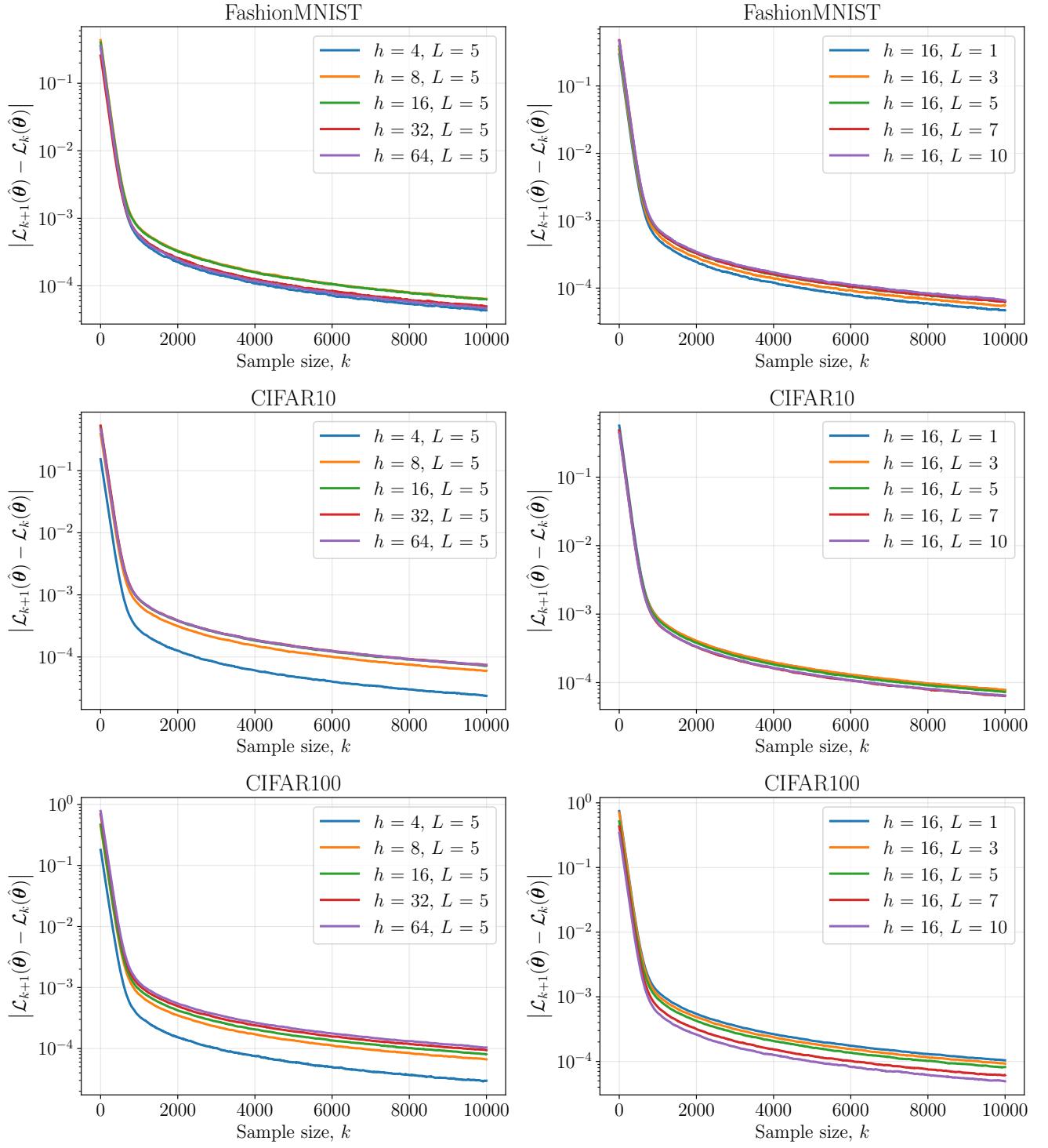


Рис. 2.5: Зависимость абсолютного значения разности функций потерь $|\mathcal{L}_{k+1}(\theta) - \mathcal{L}_k(\theta)|$ от размера выборки k для полносвязной нейронной сети при прямой классификации на наборах данных FashionMNIST, CIFAR10 и CIFAR100. Верхний ряд: при $L = 5$ уменьшение разности с увеличением h от 4 до 64; нижний ряд: при $h = 16$ увеличение разности с увеличением L от 1 до 10. Результаты подтверждают теорему 5 для различных наборов данных.

рименты проводились на GPU Tesla A100 80GB с 16 CPU ядрами и 243 GB оперативной памяти.

В первом эксперименте использовались значения пикселей изображений в качестве входных данных. На рис. 2.3 представлены результаты, полученные при анализе 10 000 объектов из набора данных MNIST [53], при этом сеть обучалась в течение 10 эпох. Соответствующий размер входа составляет 784, а размер выхода — 10. Графики слева были получены при фиксированном количестве слоев $L = 5$ в сети. Размер скрытого слоя во всех слоях варьировался от 4 до 64. График справа иллюстрирует поведение разности потерь при изменении количества скрытых слоев от 1 до 10, при фиксированном размере скрытого слоя $h = 16$. Последовательность была повторена 100 раз для усреднения. К полученным результатам было применено экспоненциальное скользящее среднее с коэффициентом сглаживания 0,99. Из наблюдаемых зависимостей следует, что, хотя изменение и не является значительным, добавление большего количества слоев приводит к большей разнице в функциях потерь. И наоборот, увеличение размера скрытого слоя приводит к меньшей разнице между функциями потерь. На практике константа M , ограничивающая величину весов, оказывается относительно небольшой.

Поскольку задача классификации набора данных MNIST считается относительно простой, неглубокая, но широкая нейронная сеть может давать хорошие результаты классификации. Наблюдалось, что значения функции потерь ниже для больших значений h , и поэтому их разница также меньше.

В отличие от предыдущего эксперимента, использовался предварительно обученный экстрактор признаков изображений. Полносвязная сеть использовалась в качестве многоклассового классификатора. В качестве модели была выбрана модель Vision Transformer (ViT) [58] от Google. Аналогичным образом были выбраны случайным образом 10 000 объектов из набора данных MNIST и варьировались размер скрытого слоя и количество слоев. Результаты согласуются с наблюдениями при прямой классификации изображений. Согласованность подтверждает, что представленная сходимость не зависит от природы пространства исходных объектов \mathcal{X} . Для наблюдения сходимости ландшафта функции потерь достаточно ограниченности этого пространства.

Эксперимент подтверждает сходимость, доказанную в теореме 5. Верхняя оценка скорости этой сходимости остается верной. Изменение параметров нейронной сети, таких как количество слоев и размер слоя, приводит к изменению разности функций потерь, что согласуется с теоретическими предсказаниями.

Приведем расширенную версию проведенных экспериментов. В таблице 2.1 представлено описание используемых наборов данных. Были использованы четыре набора данных из библиотеки `torchvision`: MNIST [53], FashionMNIST [54], CIFAR10 и CIFAR100 [55]. Единственной предобработкой данных являлась нормализация для приведения значений к диапазону $[-1; 1]$.

На рис. 2.5 графики слева получены при фиксированном количестве слоев $L = 5$ в сети, при этом размер скрытого слоя варьировался на всех уровнях от 4 до 64. График справа демонстрирует поведение разности потерь при изменении количества скрытых слоев от 1 до 10, при сохранении размера $h = 16$ неизменным. Последовательность была повторена 100 раз для усреднения. Для полученных результатов применялось экспоненциальное скользящее среднее с коэффициентом сглаживания 0.99.

Аналогично рис. 2.4, на рис. 2.6 результаты подтверждают сходимость, доказанную в теореме 5. Верхняя оценка скорости этой сходимости также верна. Изменение параметров нейронной сети приводит к изменению разности функций потерь, что согласуется с теоретическими предсказаниями.

2.4.2. Сверточные модели глубокого обучения

В настоящем разделе представлены результаты обучения сверточных сетей с различными параметрами. Основной целью экспериментов является демонстрация зависимости ландшафта функции потерь от таких параметров, как количество слоев, размер ядра, количество каналов и позиции пулинга, а также наблюдение того, как скорость сходимости зависит от этих параметров.

Для достижения указанной цели обучались сверточные сети и получались параметры $\hat{\theta}$ вблизи оптимума. В качестве модели использовалась сверточная архитектура с функцией активации ReLU после каждого слоя. Для прослеживания влияния заданного параметра на сходимость фиксировались все другие параметры нейронной сети, а заданный параметр варьировался.

Исследовалась зависимость между средним абсолютным различием значений функции потерь и доступным размером выборки. Для каждой модели производилось усреднение разности потерь по перемешанным выборкам. Для улучшения визуализации использовалось экспоненциальное сглаживание с коэффициентом 0.995. Использовалось числовое представление пикселей изображений в качестве входных данных. Результаты получены на основе анализа выборок из баз данных MNIST [53], FashionMNIST [54] и CIFAR10 [55]. Во всех экс-

периментах использовались следующие гиперпараметры: постоянная скорость обучения 10^{-3} , оптимизатор Adam, мини-пакеты размером 64, обучение проводилось в течение 10 эпох на наборах данных MNIST и Fashion-MNIST и 15 эпох на наборе данных CIFAR-10. Если параметр не варьировался, он сохранялся одинаковым во всех слоях.

2.4.3. Трансформер модели глубокого обучения

Для более глубокого изучения зависимости между функцией потерь и ее гессианом проведен эксперимент, соответствующий теореме 8. Использовалась конфигурация модели на наборе данных CIFAR-100 [55]. По сравнению с моделью для набора данных MNIST, указанная модель имеет в 8 раз больше блоков трансформера, а также скрытые слои в 8 раз шире. Во время обучения модель обучалась в течение ряда эпох для достижения точности более 50% на валидационном наборе данных. Результаты представлены на рис. 2.11.

Настройка эксперимента организована следующим образом:

1. Модель обучается до сходимости и сохраняется вектор параметров \mathbf{w}^* .
2. Начиная с пустого набора данных, добавляются данные батч за батчем (англ. batch), и вычисляется среднее значение потерь по просмотренным батчам.
3. Вычисляется абсолютная разность в соответствии с выражением:

$$|\mathcal{L}_{k+1}(\mathbf{w}) - \mathcal{L}_k(\mathbf{w})|.$$

2.5. Заключение по главе

В настоящей главе разработан единый теоретический аппарат для формализации соотношения между сложностью модели и сложностью данных в контексте обучения глубоких нейронных сетей.

Основным результатом главы является введение формальных определений меры сложности выборки $\mu_D(D)$ и меры сложности модели $\mu_f(f)$ в рамках теории мер, а также установление критерия обучаемости модели на выборке: $\mu_f(f) \leq \mu_D(D)$. Указанный критерий обеспечивает необходимое условие для предотвращения переобучения и формализует соответствие между выражительной способностью модели и информационной емкостью данных.

В рамках предложенного формализма введены понятия условной сложности выборки $\mu_D(D|f)$ и условной сложности модели $\mu_f(f|D)$, устанавливающие

взаимосвязь между свойствами данных и архитектурными характеристиками модели. Доказана теорема 2, устанавливающая необходимое и достаточное условие дообучения модели на расширенной выборке данных.

Ключевым частным случаем условной сложности модели является ландшафтная мера сложности, определяемая через спектральные свойства матриц Гессе функции потерь. Установлено, что анализ сходимости ландшафта оптимизационной задачи при увеличении объема выборки сводится к анализу спектральной нормы матрицы Гессе, что позволяет количественно оценить влияние добавления новых объектов данных на локальную геометрию функции потерь в окрестности оптимума.

Для полносвязных нейронных сетей получены строгие теоретические оценки сходимости функции потерь при увеличении размера выборки. Теорема 5 устанавливает, что абсолютная разность между значениями функции потерь при добавлении нового объекта стремится к нулю со скоростью $O(L(hM)^{2L}/k)$ при $k \rightarrow \infty$, где L — число слоев, h — размер скрытого слоя, M — константа, ограничивающая параметры и данные. Следствие 1 определяет асимптотику ландшафтной меры сложности: $\mu_f(f|D) \propto L(hM)^{2L}$, что демонстрирует экспоненциальную зависимость сложности от глубины сети.

Для сверточных нейронных сетей установлены теоретические оценки ландшафтной меры сложности, демонстрирующие экспоненциальную зависимость от глубины и полиномиальную — от остальных гиперпараметров архитектуры. Теоремы 6 и 7 определяют асимптотики для 1D- и 2D-сверточных сетей соответственно: $\mu_f(f|D) \propto L(C^2 M^2 k d)^L$ и $\mu_f(f|D) \propto C^2 k^2 L(C^2 k^2 M^2 m n)^L$.

Для трансформерных моделей получены теоретические оценки сходимости ландшафта функции потерь, восполняющие пробел в анализе путем явного вывода якобианов и гессианов для компонентов LayerNorm и FFN. Теорема 8 устанавливает неравенство сходимости $|\mathcal{L}_{k+1}(\mathbf{w}) - \mathcal{L}_k(\mathbf{w})| \leq 2L/(k+1) + M\|\mathbf{w} - \mathbf{w}^*\|_2^2/(k+1)$ и демонстрирует гетерогенность гессиана по блокам трансформера.

Проведенные вычислительные эксперименты подтверждают теоретические оценки для всех рассмотренных архитектур. Эмпирические результаты демонстрируют согласованность с теоретическими предсказаниями как для полносвязных сетей, так и для сверточных и трансформерных моделей.

Полученные результаты создают теоретическую основу для формального анализа соответствия между сложностью модели и характеристиками данных, что имеет практическое значение для проектирования архитектур нейронных сетей, планирования экспериментов и оптимизации процессов обучения. Пред-

ложенный формализм открывает перспективы для разработки методов определения достаточного размера выборки и алгоритмов адаптивного обучения.

Основные ограничения исследования связаны с детерминистическим характером анализа, предположением о существовании единой точки минимума для последовательных размеров выборки, а также возможностью улучшения верхних оценок за счет учета специфики разреженных матриц. Полученные результаты вносят вклад в теорию анализа локальной геометрии ландшафтов функции потерь и создают основу для дальнейших исследований в области формализации сложности моделей глубокого обучения.

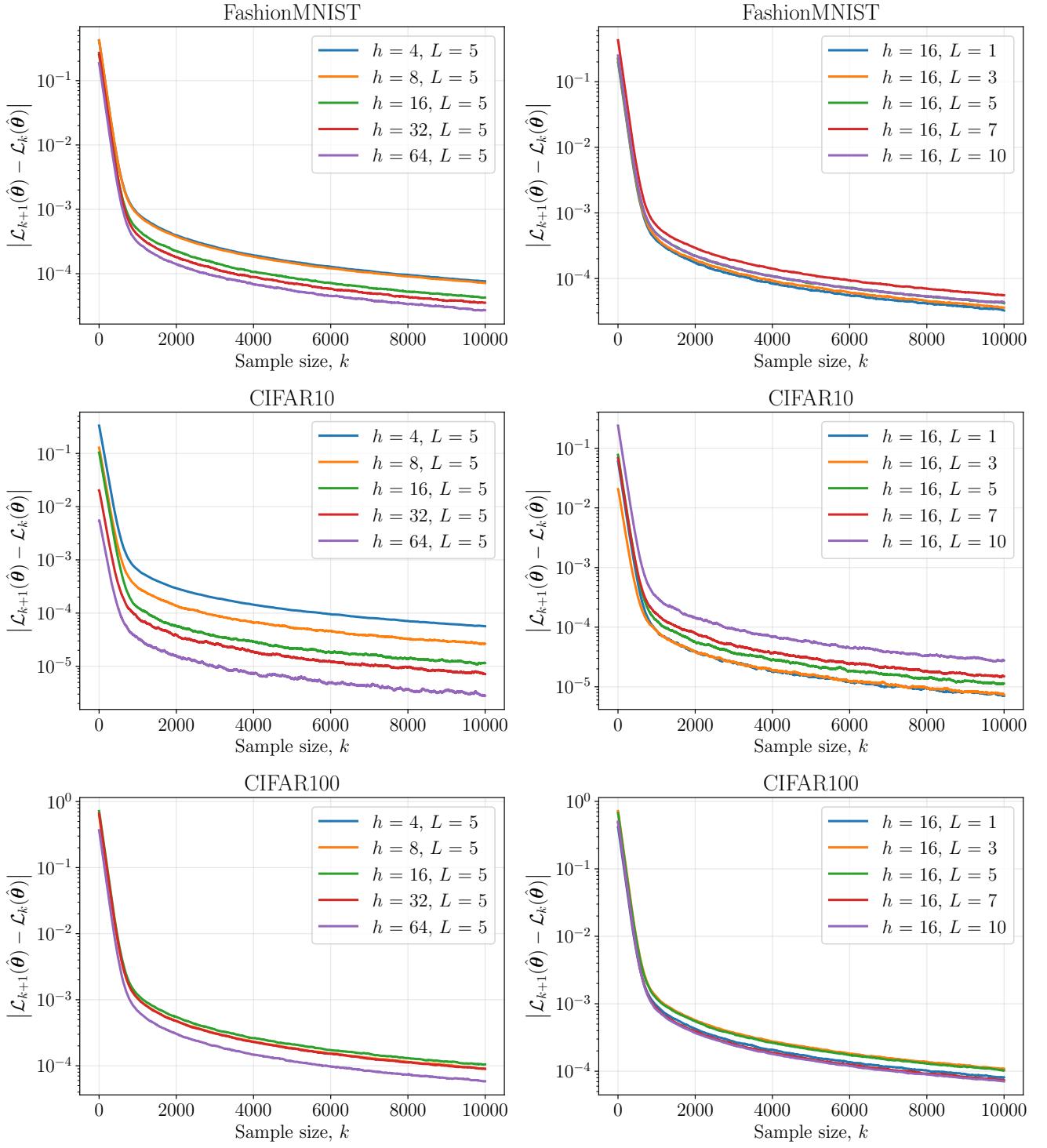


Рис. 2.6: Зависимость абсолютного значения разности функций потерь $|\mathcal{L}_{k+1}(\hat{\boldsymbol{\theta}}) - \mathcal{L}_k(\hat{\boldsymbol{\theta}})|$ от размера выборки k для полносвязной нейронной сети с предобученным экстрактором признаков (англ. Vision Transformer, ViT) на наборах данных FashionMNIST, CIFAR10 и CIFAR100. Верхний ряд: при $L = 5$ уменьшение разности с увеличением h от 4 до 64; нижний ряд: при $h = 16$ увеличение разности с увеличением L от 1 до 10. Результаты подтверждают независимость сходимости от природы пространства исходных объектов и согласуются с теоремой 5.

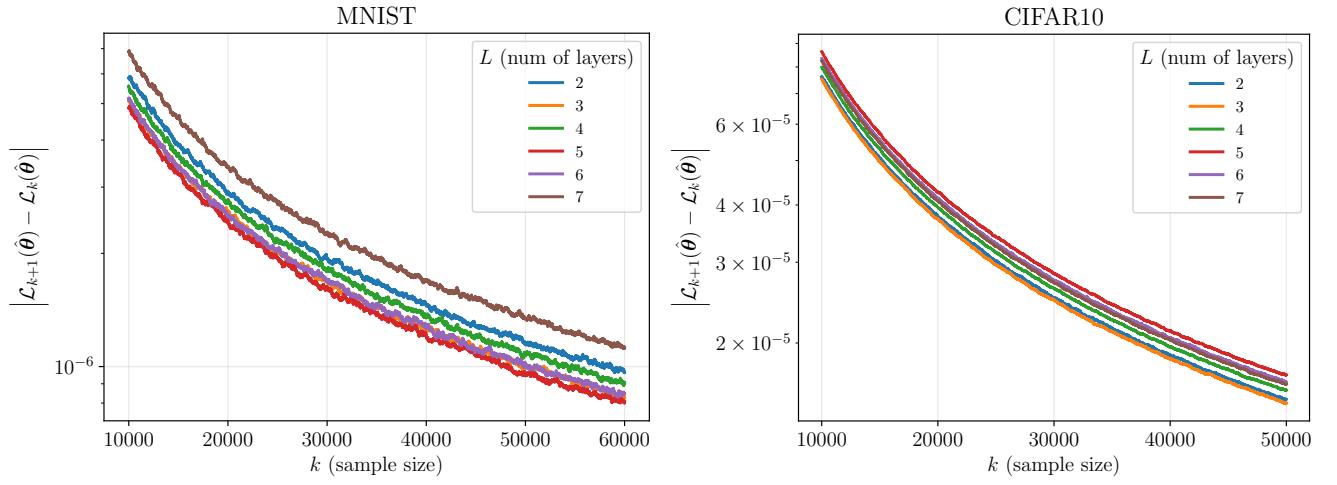


Рис. 2.7: Зависимость абсолютного значения разности функций потерь $|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})|$ от размера выборки k для сверточной нейронной сети при изменении количества сверточных слоев L при фиксированных размере ядра $k = 3$ и количестве каналов $C = 6$. Графики демонстрируют немонотонный характер зависимости потерь от количества слоев.

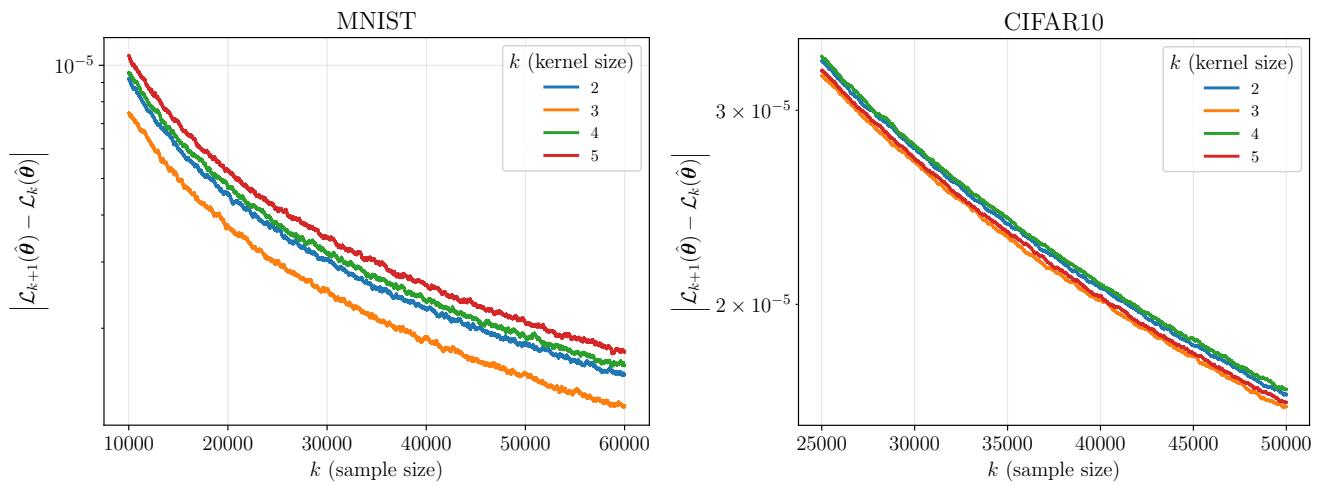


Рис. 2.8: Зависимость абсолютного значения разности функций потерь $|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})|$ от размера выборки k для сверточной нейронной сети при изменении размера ядра свертки k при фиксированных количестве слоев L и количестве каналов $C = 6$. Графики демонстрируют немонотонный характер зависимости потерь от размера ядра.

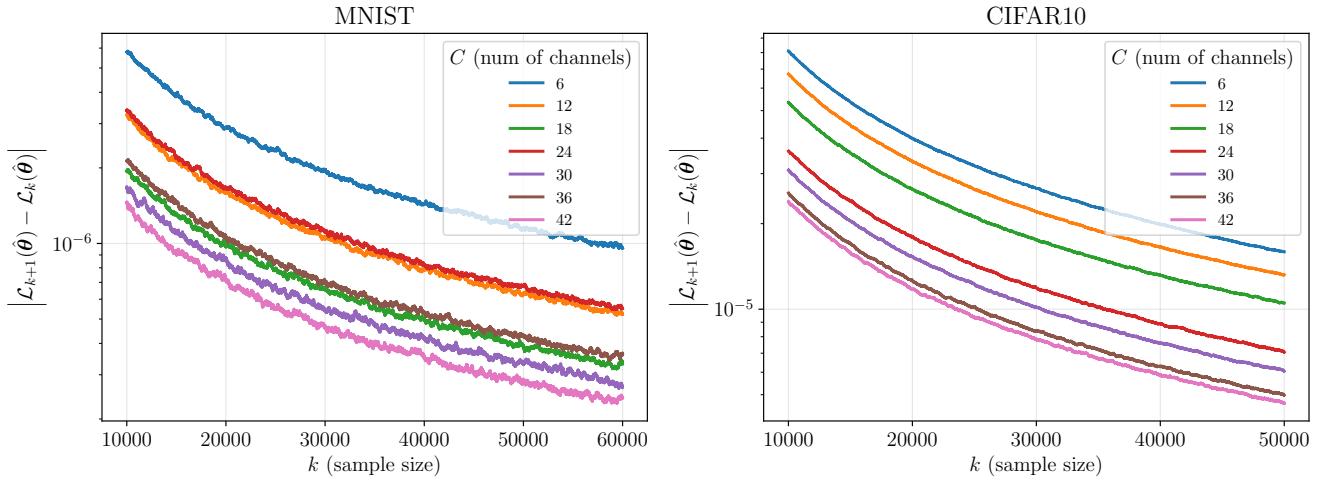


Рис. 2.9: Зависимость абсолютного значения разности функций потерь $|\mathcal{L}_{k+1}(\hat{\theta}) - \mathcal{L}_k(\hat{\theta})|$ от размера выборки k для сверточной нейронной сети при изменении количества каналов C при фиксированных количестве слоев L и размере ядра $k = 3$. Графики демонстрируют монотонный характер зависимости: увеличение числа каналов приводит к увеличению разности потерь, что согласуется с теоретическими оценками.

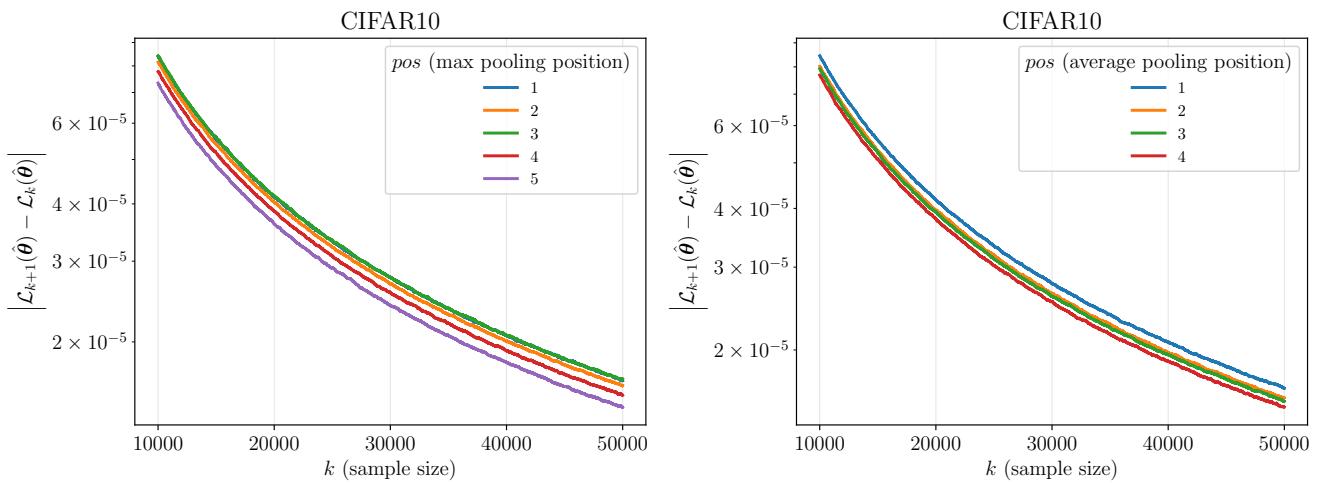


Рис. 2.10: Зависимость абсолютного значения разности функций потерь $|\mathcal{L}_{k+1}(\hat{\theta}) - \mathcal{L}_k(\hat{\theta})|$ от размера выборки k для сверточной нейронной сети при изменении позиции операции пулинга на наборе данных CIFAR10 при фиксированных количестве слоев L , размере ядра $k = 3$ и количестве каналов C . Графики демонстрируют монотонную зависимость: более раннее применение пулинга приводит к меньшей разности потерь.

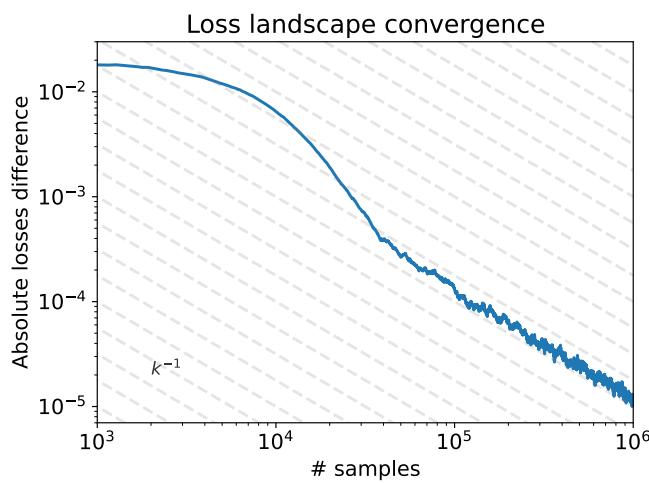


Рис. 2.11: Зависимость абсолютного значения разности функций потерь $|\mathcal{L}_{k+1}(\mathbf{w}) - \mathcal{L}_k(\mathbf{w})|$ от количества обучающих примеров k для трансформерной модели на наборе данных CIFAR-100, отображенная в двойном логарифмическом масштабе. Результаты подтверждают теорему 8 и демонстрируют стабилизацию ландшафта функции потерь с ростом объема данных.

Глава 3

Матрицы Гессе нейросетевых моделей глубокого обучения

Как было установлено в предыдущей главе, ландшафтная мера сложности модели определяется через спектральные свойства матриц Гессе функции потерь. Анализ сходимости ландшафта оптимизационной задачи при увеличении объема выборки сводится к анализу спектральной нормы матрицы Гессе. Для практического применения введенного в главе 2.1. формализма необходимо получить конкретные оценки спектральных норм матриц Гессе для различных архитектур нейронных сетей. Это составляет основную задачу настоящей главы.

В современных задачах оптимизации, к которым редуцируется процесс обучения моделей глубокого обучения, фундаментальную роль играет анализ свойств целевой функции потерь $\mathcal{L}(\boldsymbol{\theta})$, заданной на многомерном пространстве параметров $\boldsymbol{\theta} \in \mathbb{R}^n$.

Глубокие нейронные сети, обладающие способностью к аппроксимации сложных нелинейных зависимостей, порождают высокосложные невыпуклые функции потерь с многочисленными локальными минимумами, седловыми точками и сложным ландшафтом оптимизационной задачи.

Если градиент $\nabla \mathcal{L}(\boldsymbol{\theta})$ характеризует скорость и направление наискорейшего спуска в параметрическом пространстве, то матрица Гессе $\mathbf{H}(\mathcal{L})$ — симметричная матрица вторых частных производных функции потерь — предоставляет информацию о ее локальной кривизне, описывающую геометрические свойства ландшафта. Матрица Гессе содержит информацию о локальном поведении функции в окрестности заданной точки, позволяя не только предсказывать траекторию оптимизации, но и анализировать устойчивость найденных решений.

Формальное определение матрицы Гессе для функции $\mathcal{L}(\boldsymbol{\theta})$ от n параметров задается следующим выражением:

$$\mathbf{H}(\mathcal{L})_{ij} = \frac{\partial^2 \mathcal{L}}{\partial \theta_i \partial \theta_j},$$

где индексы $i, j = 1, \dots, n$ соответствуют компонентам вектора параметров $\boldsymbol{\theta}$. Для функций, обладающих непрерывными вторыми производными, матрица Гессе симметрична в силу Теоремы Шварца-Клеро-Янга о равенстве смешанных производных. Эта симметричность обеспечивает вещественность всех собственных значений и ортогональность соответствующих собственных векторов, что имеет фундаментальное значение для спектрального анализа в дальнейшем.

Применение матрицы Гессе в контексте глубокого обучения проявляется в решении разнообразных теоретических и практических задач. В аспекте оптимизации, на основе матрицы Гессе строятся методы второго порядка, такие как метод Ньютона, который использует обратную матрицу Гессе \mathbf{H}^{-1} для вычисления аддитивных направлений обновления параметров, учитывающих локальную кривизну поверхности потерь. Это свойство обеспечивает существенное ускорение сходимости в окрестности локального минимума по сравнению с методами первого порядка, основанными исключительно на градиентной информации. Однако прямая реализация методов второго порядка сопряжена с существенными вычислительными сложностями, что стимулировало развитие квази-ニュтоновских методов и методов приближенного вычисления обратной матрицы Гессе.

В контексте теоретического анализа моделей машинного обучения матрица Гессе вносит существенный вклад в оценку сложности и обобщающей способности моделей. Собственные значения матрицы Гессе, вычисленные в стационарной точке, содержат информацию о геометрии ландшафта функции потерь.

Спектральный анализ матрицы Гессе позволяет количественно охарактеризовать локальную кривизну функции потерь вдоль различных направлений в параметрическом пространстве, выявить наличие седловых точек и оценить устойчивость найденного решения.

Во-первых, малые собственные значения соответствуют направлениям с незначительной кривизной — так называемым “плоским” регионам, где параметры могут варьироваться без существенного роста ошибки. Эти направления часто ассоциируются с параметрами, оказывающими незначительное влияние на выход модели, или с симметриями в архитектуре сети. В противоположность этому, большие собственные значения указывают на “острые” минимумы с выраженной кривизной. Эмпирические и теоретические исследования подтверждают, что плоские минимумы демонстрируют улучшенную обобщающую способность, обусловленную их пониженней чувствительностью к малым возмущениям в данных и параметрах модели. Понятие “остроты” минимума активно изучается в современной теории глубокого обучения и оптимизации.

Во-вторых, след матрицы Гессе, равный сумме ее собственных значений, является интегральной характеристикой общей кривизны функции потерь. Детальный анализ спектрального состава матрицы Гессе, в частности оценка кратности собственных значений вблизи нуля, позволяет количественно оценить эффективную размерность пространства параметров, влияющих на выход модели,

и идентифицировать структурную избыточность в архитектуре нейронной сети. Кроме того, распределение собственных значений матрицы Гессе тесно связано с устойчивостью модели к шуму, способностью к интерполяции данных и обобщающей способностью на тестовых выборках.

Таким образом, матрица Гессе является не только эффективным инструментом для ускорения процесса оптимизации, но и аналитическим аппаратом для анализа внутренних свойств модели: от устойчивости найденного решения до прогнозирования его способности к общению на новые данные. Однако использование матрицы Гессе в задачах глубокого обучения с миллионами и миллиардами параметров сталкивается с вычислительными ограничениями, поскольку требования к памяти и вычислительным ресурсам для хранения и обращения плотной матрицы размерности $n \times n$ становятся непрактичными для реальных приложений. Это методологическое ограничение обуславливает актуальность разработки эффективных методов аппроксимации матрицы Гессе и получения аналитических оценок ее ключевых спектральных характеристик для различных архитектур нейронных сетей. Современные подходы к решению этой проблемы включают методы случайного проектирования, разложения Кронекера, диагональные и блочно-диагональные аппроксимации, а также методы, основанные на теории случайных матриц.

3.1. Полносвязная нейросетевая модель глубокого обучения

Рассмотрим формальную постановку задачи K -классовой классификации с использованием функции потерь кросс-энтропии. В данной постановке входные данные представляются вектором $\mathbf{x} \in \mathbb{R}^l$, а выходные данные — вектором $\mathbf{y} \in \mathbb{R}^K$, имеющим структуру one-hot кодирования, где все компоненты равны нулю, за исключением позиции $y_k = 1$, соответствующей истинной метке класса для входного образца \mathbf{x} . Такое представление данных является стандартным для задач классификации и позволяет естественным образом использовать категориальное распределение для моделирования неопределенности предсказаний.

Рассматривается L -слойная полносвязная нейронная сеть $f_{\theta}(\cdot)$ с функцией активации ReLU, применяемой после каждого линейного преобразования. Выбор функции активации ReLU обусловлен ее вычислительной эффективностью и свойством устранять проблему затухающих градиентов, а также ее удобством для анализа в теоретическом анализе нейросетевых архитектур. Для функции

активации ReLU, определяемой как $\sigma(\mathbf{x}) = [\mathbf{x} \geq \mathbf{0}] \mathbf{x}$, где $[\cdot]$ обозначает поэлементную индикаторную функцию, выход сети представляет собой вектор логитов $\mathbf{z} \in \mathbb{R}^K$. Вычисление логитов осуществляется посредством последовательного применения следующих рекуррентных соотношений:

$$\begin{aligned}\mathbf{z}^{(p)} &= \mathbf{W}^{(p)} \mathbf{x}^{(p)} + \mathbf{b}^{(p)}, \\ \mathbf{x}^{(p+1)} &= \sigma(\mathbf{z}^{(p)}).\end{aligned}$$

Здесь $\mathbf{x}^{(p)}$ и $\mathbf{z}^{(p)}$ обозначают вход и выход p -го слоя соответственно, при этом полагается $\mathbf{x}^{(1)} = \mathbf{x}$ и $\mathbf{z} = f_{\theta}(\mathbf{x}) = \mathbf{z}^{(L)}$. Совокупность всех параметров модели обозначается как $\boldsymbol{\theta} = \text{col}(\mathbf{w}^{(1)}, \mathbf{b}^{(1)}, \dots, \mathbf{w}^{(L)}, \mathbf{b}^{(L)}) \in \mathbb{R}^n$. Для p -го слоя $\mathbf{w}^{(p)}$ представляет собой векторизованную матрицу весов $\mathbf{W}^{(p)}$, а $\mathbf{b}^{(p)}$ — соответствующий вектор смещений. Общее число параметров n в таких моделях может достигать миллионов и даже миллиардов, что и создает вычислительные трудности для точного вычисления матрицы Гессе.

Выходы модели определяются как $\mathbf{p} = \text{softmax}(\mathbf{z}) \in \mathbb{R}^K$, где каждая компонента вычисляется по формуле:

$$p_i = \text{softmax}(\mathbf{z})_i = \frac{\exp(z_i)}{\sum_{j=1}^K \exp(z_j)} \in (0; 1).$$

Функция потерь представляет собой стандартную кросс-энтропийную функцию ошибки:

$$\ell(\mathbf{z}, \mathbf{y}) = \text{CE}(\mathbf{p}, \mathbf{y}) = - \sum_{k=1}^K y_k \log p_k \in \mathbb{R}^+.$$

Эта функция является выпуклой по логитам \mathbf{z} , но невыпуклой по параметрам сети $\boldsymbol{\theta}$ из-за сложной композиционной структуры нейронной сети.

Согласно установленным результатам в литературе [36], применение цепного правила для матриц второго порядка [37] позволяет декомпозировать матрицу Гессе на сумму двух структурно различных компонент:

$$\mathbf{H}_i(\boldsymbol{\theta}) = \underbrace{\nabla_{\theta} \mathbf{z}_i \frac{\partial^2 \ell(\mathbf{z}_i, \mathbf{y}_i)}{\partial \mathbf{z}_i^2} \nabla_{\theta} \mathbf{z}_i^T}_{\text{G-компоненты}} + \underbrace{\sum_{k=1}^K \frac{\partial \ell(\mathbf{z}_i, \mathbf{y}_i)}{\partial z_{ik}} \nabla_{\theta}^2 z_{ik}}_{\text{H-компоненты}},$$

где $\nabla_{\theta} \mathbf{z}_i \in \mathbb{R}^{P \times K}$ представляет собой матрицу Якоби функции нейронной сети по параметрам, а $\frac{\partial^2 \ell(\mathbf{z}_i, \mathbf{y}_i)}{\partial \mathbf{z}_i^2}$ — матрицу Гессе функции потерь относительно

выходных логитов для i -го наблюдения. Первое слагаемое (G -компоненты) отражает влияние кривизны функции потерь, в то время как второе слагаемое (H -компоненты) — кривизну самой нейронной сети.

Эмпирические исследования [42, 36, 41] демонстрируют, что спектральное распределение матрицы Гессе характеризуется наличием основной массы собственных значений, сосредоточенной вблизи нуля, и выбросов, распределенных в области ненулевых значений. Это бимодальное распределение собственных значений является характерной чертой гессианов глубоких нейронных сетей и отражает фундаментальные свойства параметрического пространства таких моделей. Вследствие данной спектральной структуры, для практического анализа наиболее релевантной является G -компоненты, что обосновывает использование следующей аппроксимации:

$$\mathbf{H}_i(\boldsymbol{\theta}) \approx \nabla_{\boldsymbol{\theta}} \mathbf{z}_i \frac{\partial^2 \ell(\mathbf{z}_i, \mathbf{y}_i)}{\partial \mathbf{z}_i^2} \nabla_{\boldsymbol{\theta}} \mathbf{z}_i^T.$$

Дополнительное теоретическое обоснование данной аппроксимации представляется в рамках теории ядра нейронного касательного пространства [35, 59], где предполагается линейная зависимость логитов \mathbf{z} от параметров $\boldsymbol{\theta}$ в окрестности точки оптимума. Данное предположение имплицирует исчезающую кривизну логитов $\nabla_{\boldsymbol{\theta}}^2 z_{ik}$, что влечет тождественное обращение в ноль H -компоненты.

На основе работ [60], предлагающих аналитическую аппроксимацию G -компоненты для полно связанных нейронных сетей, принимается следующая параметризация: $\mathbf{H}_i(\boldsymbol{\theta}) \approx \mathbf{F}_i^T \mathbf{A}_i \mathbf{F}_i$. Эта факторизация позволяет эффективно вычислять приближения гессиана без явного построения полной матрицы, используя разложения в матрицы меньшей размерности. Введем систему обозначений:

- Матричное представление функции активации ReLU:

$$\mathbf{D}^{(p)} = \text{diag}([\mathbf{z}^{(p)} \geq 0]),$$

Эта диагональная матрица кодирует паттерн активации нейронов на p -м слое и играет основную роль в определении функциональной структуры сети.

- Матрица прямого распространения от p -го слоя к выходу:

$$\mathbf{G}^{(p)} = \frac{\partial \mathbf{z}}{\partial \mathbf{z}^{(p)}} = \mathbf{W}^{(L)} \mathbf{D}^{(L-1)} \mathbf{W}^{(L-1)} \mathbf{D}^{(L-2)} \dots \mathbf{D}^{(p)},$$

Эта матрица описывает, как изменения в активациях на p -м слое через последующие слои к выходу сети.

- Блочная матрица всех производных логитов по параметрам:

$$\mathbf{F}^T = \begin{pmatrix} (\mathbf{G}^{(1)})^T \otimes \mathbf{x}^{(1)} \\ (\mathbf{G}^{(1)})^T \\ \vdots \\ (\mathbf{G}^{(L)})^T \otimes \mathbf{x}^{(L)} \\ (\mathbf{G}^{(L)})^T \end{pmatrix},$$

где \otimes обозначает произведение Кронекера. Эта блочная структура естественным образом отражает слоистую архитектуру сети и позволяет эффективное вычисление.

- Гессиан функции потерь относительно логитов, имеющий структуру ковариационной матрицы [61]:

$$\mathbf{A} = \nabla_{\mathbf{z}}^2 \ell(\mathbf{z}, \mathbf{y}) = \text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T.$$

Эта матрица является положительно полуопределенной и вырожденной, что отражает инвариантность функции softmax к сдвигам в пространстве логитов.

На основе предложенной параметризации получена верхняя оценка спектральной нормы матрицы Гессе в полносвязной нейронной сети, формулируемая в теореме 9.

3.1.1. Спектральная оценка матрицы Гессе

Теорема 9 устанавливает верхнюю оценку спектральной нормы матрицы Гессе для полносвязной нейронной сети с функцией активации ReLU. Условия теоремы включают ограниченность спектральных норм матриц весов всех слоев и норм входных векторов, что представляет собой естественное предположение для большинства практических приложений.

Полученная оценка демонстрирует экспоненциальную зависимость от глубины сети L , что согласуется с известными результатами о возрастании сложности оптимизационного ландшафта с увеличением глубины нейронной сети.

Теорема 9 (Киселев-Грабовой, Оценка нормы матрицы Гессе для полносвязных сетей). Рассмотрим L -слойную полносвязную нейронную сеть с функцией активации ReLU и без членов смещения, применяемую для решения задачи

классификации на K классов. Предположим, что выполнены условия:

$$\begin{aligned}\|\mathbf{W}^{(p)}\|_2 &\leq M_{\mathbf{W}}, \\ \|\mathbf{x}_i\|_2 &\leq M_{\mathbf{x}},\end{aligned}$$

для всех слоев $p = 1, \dots, L$ в сети и для всех объектов $i = 1, \dots, m$. Тогда для любого объекта $i = 1, \dots, m$ выполняется следующее неравенство:

$$\|\mathbf{H}_i(\boldsymbol{\theta})\|_2 \leq L\sqrt{2}M_{\mathbf{x}}^2M_{\mathbf{W}}^{2L} + \sqrt{2}\frac{M_{\mathbf{W}}^2(M_{\mathbf{W}}^{2L}-1)}{M_{\mathbf{W}}^2-1}.$$

Доказательство. Для упрощения, опустим индекс i , соответствующий конкретному объекту в выборке, поскольку оценка проводится для произвольного фиксированного объекта.

Вначале воспользуемся факторизацией гессиана, полученной ранее: $\mathbf{H}(\boldsymbol{\theta}) = \mathbf{F}^\top \mathbf{A} \mathbf{F}$. В силу субмультипликативности спектральной нормы матрицы, получаем следующую оценку:

$$\|\mathbf{H}(\boldsymbol{\theta})\|_2 = \|\mathbf{F}^\top \mathbf{A} \mathbf{F}\|_2 \leq \|\mathbf{F}^\top\|_2 \cdot \|\mathbf{A}\|_2 \cdot \|\mathbf{F}\|_2.$$

Учитывая, что $\|\mathbf{F}^\top\|_2 = \|\mathbf{F}\|_2$ для любой матрицы, приходим к оценке:

$$\|\mathbf{H}(\boldsymbol{\theta})\|_2 \leq \|\mathbf{A}\|_2 \cdot \|\mathbf{F}\|_2^2.$$

Далее детально рассмотрим каждое слагаемое отдельно, начиная с оценки спектральной нормы матрицы \mathbf{A} . В силу эквивалентности матричных норм выполняется неравенство между спектральной нормой и нормой Фробениуса:

$$\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F,$$

где $\|\mathbf{A}\|_F$ — норма Фробениуса, определяемая как квадратный корень из суммы квадратов всех элементов матрицы. Используя свойства этой нормы для оценки указанного слагаемого. Согласно определению нормы Фробениуса получаем:

$$\begin{aligned}\|\mathbf{A}\|_F^2 &= \|\text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top\|_F^2 = \sum_{k=1}^K(p_k - p_k^2)^2 + \sum_{k \neq l}p_k^2p_l^2 = \\ &= \sum_{k=1}^Kp_k^2(1 - p_k)^2 + \sum_{k \neq l}p_k^2p_l^2.\end{aligned}$$

Поскольку все вероятности удовлетворяют следующему неравенству $0 \leq p_k \leq 1$ для всех $k = 1, \dots, K$ получаем:

$$0 \leq p_k^2 \leq p_k \quad \text{и} \quad 0 \leq (1 - p_k)^2 \leq (1 - p_k),$$

Следовательно, для первого слагаемого получаем оценку:

$$\sum_{k=1}^K p_k^2 (1 - p_k)^2 \leq \sum_{k=1}^K p_k (1 - p_k) \leq \sum_{k=1}^K p_k = 1,$$

где последнее неравенство следует из того, что $p_k(1 - p_k) \leq p_k$ и $\sum_{k=1}^K p_k = 1$. Для второго слагаемого получаем:

$$\sum_{k \neq l} p_k^2 p_l^2 \leq \sum_{k \neq l} p_k p_l = \left(\sum_{k=1}^K p_k \right)^2 - \sum_{k=1}^K p_k^2 = 1 - \sum_{k=1}^K p_k^2,$$

поскольку $\left(\sum_{k=1}^K p_k \right)^2 = 1$, а двойная сумма $\sum_{k \neq l} p_k p_l$ равна квадрату суммы за вычетом суммы квадратов. Комбинируя полученные оценки, получаем:

$$\|\mathbf{A}\|_F^2 \leq 1 + \left(1 - \sum_{k=1}^K p_k^2 \right) = 2 - \sum_{k=1}^K p_k^2 \leq 2,$$

где последнее неравенство следует из неотрицательности $\sum_{k=1}^K p_k^2$. Таким образом, норма Фробениуса матрицы \mathbf{A} ограничена сверху значением $\sqrt{2}$, и следовательно:

$$\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F \leq \sqrt{2}.$$

Переходим к оценке нормы $\|\mathbf{F}\|_2$. Для оценки $\|\mathbf{F}\|_2$ анализируем спектральную норму матриц $\mathbf{G}^{(p)}$, которые определяются как:

$$\mathbf{G}^{(p)} = \mathbf{W}^{(L)} \mathbf{D}^{(L-1)} \mathbf{W}^{(L-1)} \mathbf{D}^{(L-2)} \cdots \mathbf{D}^{(p)}.$$

Используя свойство субмультипликативности спектральной нормы, имеем:

$$\|\mathbf{G}^{(p)}\|_2 \leq \|\mathbf{W}^{(L)}\|_2 \cdot \|\mathbf{D}^{(L-1)}\|_2 \cdot \|\mathbf{W}^{(L-1)}\|_2 \cdot \|\mathbf{D}^{(L-2)}\|_2 \cdots \|\mathbf{D}^{(p)}\|_2.$$

Поскольку матрицы $\mathbf{D}^{(p)}$ являются диагональными матрицами с элементами 0 или 1, их спектральная норма не превосходит 1. Таким образом, получаем:

$$\|\mathbf{G}^{(p)}\|_2 \leq \prod_{s=p}^L \|\mathbf{W}^{(s)}\|_2.$$

Поскольку матрица \mathbf{F} представляет собой вертикальную конкатенацию блоков вида $(\mathbf{G}^{(p)})^\top \otimes \mathbf{x}^{(p)}$ и $(\mathbf{G}^{(p)})^\top$, для спектральной нормы такой блочной матрицы справедливо неравенство:

$$\|\mathbf{F}\|_2^2 \leq \sum_{p=1}^L \left(\|(\mathbf{G}^{(p)})^\top \otimes \mathbf{x}^{(p)}\|_2^2 + \|(\mathbf{G}^{(p)})^\top\|_2^2 \right),$$

где используя свойство спектральной нормы произведения Кронекера:

$$\|\mathbf{A} \otimes \mathbf{B}\|_2 = \|\mathbf{A}\|_2 \cdot \|\mathbf{B}\|_2,$$

а также учитывая, что $\|(\mathbf{G}^{(p)})^\top\|_2 = \|\mathbf{G}^{(p)}\|_2$, получаем:

$$\begin{aligned} \|\mathbf{F}\|_2^2 &\leq \sum_{p=1}^L \left(\|\mathbf{G}^{(p)}\|_2^2 \cdot \|\mathbf{x}^{(p)}\|_2^2 + \|\mathbf{G}^{(p)}\|_2^2 \right) = \\ &= \sum_{p=1}^L \|\mathbf{G}^{(p)}\|_2^2 \left(\|\mathbf{x}^{(p)}\|_2^2 + 1 \right). \end{aligned}$$

Подставляя полученную ранее оценку для $\|\mathbf{G}^{(p)}\|_2$, получаем итоговую оценку:

$$\|\mathbf{F}\|_2^2 \leq \sum_{p=1}^L \left(\|\mathbf{x}^{(p)}\|_2^2 + 1 \right) \prod_{s=p}^L \|\mathbf{W}^{(s)}\|_2^2.$$

Собирая полученные оценки норм $\|\mathbf{A}\|, \|\mathbf{F}\|$ получаем итоговую оценку для гессиана:

$$\|\mathbf{H}_i(\boldsymbol{\theta})\|_2 \leq \sqrt{2} \sum_{p=1}^L \left(\|\mathbf{x}_i^{(p)}\|_2^2 + 1 \right) \prod_{s=p}^L \|\mathbf{W}^{(s)}\|_2^2.$$

В случае, когда векторы смещений отсутствуют, т.е. $\mathbf{b}^{(p)} = \mathbf{0}$ для всех $p = 1, \dots, L$. Получаем, что выход каждого слоя оценивается:

$$\|\mathbf{x}_i^{(p)}\|_2 \leq \|\mathbf{x}_i\|_2 \prod_{s=1}^{p-1} \|\mathbf{W}^{(s)}\|_2,$$

поскольку каждый слой осуществляет линейное преобразование с последующей нелинейностью ReLU, которая не увеличивает норму. Тогда используя данную

оценку, получаем:

$$\begin{aligned}
\|\mathbf{H}_i(\boldsymbol{\theta})\|_2 &\leq \sqrt{2} \sum_{p=1}^L \left(\|\mathbf{x}_i\|_2^2 \prod_{s=1}^{p-1} \|\mathbf{W}^{(s)}\|_2^2 + 1 \right) \prod_{s=p}^L \|\mathbf{W}^{(s)}\|_2^2 = \\
&= \sqrt{2} \sum_{p=1}^L \left(\|\mathbf{x}_i\|_2^2 \prod_{s=1}^L \|\mathbf{W}^{(s)}\|_2^2 + \prod_{s=p}^L \|\mathbf{W}^{(s)}\|_2^2 \right) = \\
&= L\sqrt{2}\|\mathbf{x}_i\|_2^2 \prod_{p=1}^L \|\mathbf{W}^{(p)}\|_2^2 + \sqrt{2} \sum_{p=1}^L \prod_{s=p}^L \|\mathbf{W}^{(s)}\|_2^2.
\end{aligned}$$

Воспользовавшись условиями теоремы, согласно которым для всех p норма матриц $\|\mathbf{W}^{(p)}\|_2 \leq M_{\mathbf{W}}$ и $\|\mathbf{x}_i\|_2 \leq M_{\mathbf{x}}$, получаем:

$$\|\mathbf{H}_i(\boldsymbol{\theta})\|_2 \leq L\sqrt{2}M_{\mathbf{x}}^2M_{\mathbf{W}}^{2L} + \sqrt{2} \sum_{p=1}^L M_{\mathbf{W}}^{2(L-p+1)}.$$

Сумма в правой части представляет собой сумму геометрической прогрессии:

$$\sum_{p=1}^L M_{\mathbf{W}}^{2(L-p+1)} = \sum_{k=1}^L M_{\mathbf{W}}^{2k} = \frac{M_{\mathbf{W}}^2(M_{\mathbf{W}}^{2L} - 1)}{M_{\mathbf{W}}^2 - 1},$$

Следовательно, получаем итоговую оценку на матрицу Гессе:

$$\|\mathbf{H}_i(\boldsymbol{\theta})\|_2 \leq L\sqrt{2}M_{\mathbf{x}}^2M_{\mathbf{W}}^{2L} + \sqrt{2} \frac{M_{\mathbf{W}}^2(M_{\mathbf{W}}^{2L} - 1)}{M_{\mathbf{W}}^2 - 1}.$$

Для однослойной сети $L = 1$ оценка упрощается:

$$\|\mathbf{H}_i(\boldsymbol{\theta})\|_2 \leq \sqrt{2}M_{\mathbf{W}}^2(M_{\mathbf{x}}^2 + 1).$$

□

Замечание 4. В теореме 9 получена оценка на матрицу Гессе для однослойной сети $L = 1$ вида:

$$\|\mathbf{H}_i(\boldsymbol{\theta})\|_2 \leq \sqrt{2}M_{\mathbf{W}}^2(M_{\mathbf{x}}^2 + 1).$$

Данная оценка соответствует эмпирическому представлению о том, что кривизна функции потерь пропорциональна квадрату нормы весов и квадрату нормы входных данных.

Теорема 10 указывает на зависимость спектральной нормы матрицы Гессе от размера скрытого слоя h . Полученная оценка демонстрирует, что норма гессиана растет экспоненциально как по глубине сети L , так и по размеру скрытого слоя h , причем степень экспоненты определяется произведением hM . Это подчеркивает влияние ширины и глубины сети на сложность оптимизационного ландшафта.

Экспоненциальный характер зависимости объясняет известную эмпирическую оценку о том, что глубокие и широкие сети обладают значительно более сложной геометрией функции потерь, что создает дополнительные сложности для методов оптимизации. Полученный результат также подчеркивает важность контроля норм параметров на протяжении всего процесса обучения для обеспечения устойчивости алгоритмов оптимизации.

Теорема 10 (*Киселев-Грабовой*, Асимптотика нормы матрицы Гессе для полносвязных сетей). *Пусть все параметры модели ограничены некоторой константой $M > 0$, то есть для всех $i, j = 1, \dots, h$ и для всех слоев $p = 1, \dots, L$ выполняется условие $|w_{ij}^{(p)}| \leq M$, тогда при выполнении условий Теоремы 9 справедливо следующее неравенство:*

$$\|\mathbf{H}_i(\boldsymbol{\theta})\|_2 \leq L\sqrt{2}M_x^2(hM)^{2L} + \sqrt{2}\frac{(hM)^2((hM)^{2L} - 1)}{(hM)^2 - 1}.$$

Таким образом, имеет место следующая пропорциональность для нормы матрицы Гессе:

$$\|\mathbf{H}_i(\boldsymbol{\theta})\|_2 \propto L(hM)^{2L}.$$

Доказательство. Для произвольной матрицы $\mathbf{W}^{(p)} \in \mathbb{R}^{h \times h}$ справедливо неравенство между спектральной и фробениусовой нормами:

$$\|\mathbf{W}^{(p)}\|_2 \leq \|\mathbf{W}^{(p)}\|_F = \sum_{i=1}^h \sum_{j=1}^h \left(w_{ij}^{(p)}\right)^2.$$

Из условия теоремы каждый элемент матрицы ограничен $|w_{ij}^{(p)}| \leq M$ для всех $i, j = 1, \dots, h$ и всех слоев $p = 1, \dots, L$. Следовательно:

$$\left(w_{ij}^{(p)}\right)^2 \leq M^2,$$

Далее, так как матрица имеет размер $h \times h$, общее число элементов равно h^2 . Следовательно:

$$\|\mathbf{W}^{(p)}\|_F^2 \leq h^2 M^2.$$

В теореме 9 была получена следующая оценка:

$$\|\mathbf{H}_i(\boldsymbol{\theta})\|_2 \leq L\sqrt{2}M_{\mathbf{x}}^2M_{\mathbf{W}}^{2L} + \sqrt{2}\frac{M_{\mathbf{W}}^2(M_{\mathbf{W}}^{2L}-1)}{M_{\mathbf{W}}^2-1},$$

где $M_{\mathbf{W}}$ — верхняя оценка спектральных норм матриц весов всех слоев, причем ранее было получено, что $M_{\mathbf{W}} \leq hM$. Следовательно, получаем:

$$\|\mathbf{H}_i(\boldsymbol{\theta})\|_2 \leq L\sqrt{2}M_{\mathbf{x}}^2(hM)^{2L} + \sqrt{2}\frac{(hM)^2((hM)^{2L}-1)}{(hM)^2-1}.$$

Для анализа асимптотического поведения полученной оценки рассмотрим ее поведение при больших значениях h и L . Первое слагаемое имеет асимптотику:

$$L\sqrt{2}M_{\mathbf{x}}^2(hM)^{2L} = \propto Lh^{2L},$$

поскольку M и $M_{\mathbf{x}}$ являются константами. Второе слагаемое:

$$\sqrt{2}\frac{(hM)^2((hM)^{2L}-1)}{(hM)^2-1} \propto h^{2L},$$

так как при $hM > 1$ числитель растет как $(hM)^{2L+2}$, а знаменатель как $(hM)^2$. Таким образом, итоговая асимптотика:

$$\|\mathbf{H}_i(\boldsymbol{\theta})\|_2 \propto Lh^{2L}.$$

Для однослойной сети оценка упрощается, то есть подставляя $L = 1$, получаем:

$$\|\mathbf{H}_i(\boldsymbol{\theta})\|_2 \leq \sqrt{2}M_{\mathbf{W}}^2(M_{\mathbf{x}}^2 + 1) \leq \sqrt{2}(hM)^2(M_{\mathbf{x}}^2 + 1).$$

Следовательно, в случае одного слоя норма гессиана растет квадратично с размером слоя:

$$\|\mathbf{H}_i(\boldsymbol{\theta})\|_2 \propto h^2.$$

□

Полученные результаты указывают на то, что норма матрицы Гессе является степенной функцией от размера скрытого слоя h и экспоненциальной функцией от количества слоев L .

С одной стороны, полученная оценка является завышенной. С другой стороны, если выбрать значение h большим, то ограничивающая константа M будет малой, на что указывают эмпирические исследования.

3.2. Матричные модели глубокого обучения

В настоящем разделе рассматривается общий класс матричных моделей глубокого обучения, которые представляют собой композицию последовательных линейных преобразований и нелинейных функций активации. Частным случаем такого представления являются сверточные нейронные сети (англ. CNN), а также другие архитектуры, где каждый слой может быть представлен в виде линейного оператора.

Данный подход позволяет единообразно анализировать различные архитектуры нейронных сетей через призму матричных операций, что упрощает получение оценок для матриц Гессе.

Пусть $f_{\theta}(\mathbf{x})$ является суперпозицией $L + 1$ слоев с активациями ReLU, что формально записывается как:

$$f_{\theta}(\mathbf{x}) = \mathbf{T}^{(L+1)} \circ \sigma \circ \cdots \circ \sigma \circ \mathbf{T}^{(1)}(\mathbf{x}).$$

В этом представлении каждый $\mathbf{T}^{(p+1)}$ представляет собой линейный оператор или его матричное представление, а σ обозначает функцию активации ReLU, применяемую поэлементно. Такая композиционная структура позволяет описывать глубокие нейронные сети как последовательность преобразований, где каждый слой осуществляет линейное отображение с последующей нелинейной активацией.

Промежуточные результаты вычисления функции сети могут быть представлены в виде системы уравнений:

$$\begin{cases} \mathbf{z}^{(p+1)} &= \mathbf{T}^{(p+1)} \mathbf{x}^{(p)}, \\ \mathbf{x}^{(p+1)} &= \sigma(\mathbf{z}^{(p+1)}) \end{cases}$$

где выход сети определяется как $f_{\theta}(\mathbf{x}) = \mathbf{z} := \mathbf{z}^{(L+1)}$, а входные данные задаются как $\mathbf{x}^{(0)} := \mathbf{x}$. Здесь $\mathbf{z}^{(p+1)}$ представляет собой выход линейного оператора на $(p+1)$ -м слое до применения активации, а $\mathbf{x}^{(p+1)}$ — результат после применения функции активации ReLU, который является входом для следующего слоя.

Рассмотрим матрицу $\Lambda^{(p+1)} := \text{diag}(\mathbf{x}^{(p+1)} > 0)$, зависящую от входных данных, которая кодирует паттерн активации нейронов на $(p+1)$ -м слое. Элементы этой матрицы равны 1 для нейронов с положительной активацией и 0 в противном случае, что отражает свойство функции ReLU "отсекать" отрицательные значения и представляет функцию активации в виде линейного оператора. Ис-

пользуя эти диагональные матрицы, всю функцию нейронной сети можно представить в виде произведения матриц, а именно суперпозиций линейных операторов:

$$f_{\theta}(\mathbf{x}) = \mathbf{T}^{(L+1)} \mathbf{\Lambda}^{(L)} \dots \mathbf{\Lambda}^{(1)} \mathbf{T}^{(1)} \mathbf{x}. \quad (3.1)$$

Данное представление удобно, так как позволяет рассматривать глубокую нейронную сеть с активациями ReLU как кусочно-линейную функцию, где нелинейность возникает исключительно за счет бинарных переключений в матрицах $\mathbf{\Lambda}^{(p)}$, зависящих от входных данных.

Вектор параметров модели объединяет все обучаемые параметры сети: $\boldsymbol{\theta} = \text{col}(\mathbf{W}^{(L+1)}, \dots, \mathbf{W}^{(1)})$, где каждый линейный оператор $\mathbf{T}^{(p)}$ дифференцируемо параметризуется соответствующей частью вектора параметров $\mathbf{W}^{(p)}$. Для анализа модели вводится производная слоя по его параметрам:

$$\mathbf{Q}^{(p)} := \frac{\partial \mathbf{T}^{(p)}}{\partial \mathbf{W}^{(p)}},$$

а затем строится блочно-диагональная матрица, объединяющая эти производные по всем слоям:

$$\mathbf{Q} := \text{diag}(\mathbf{Q}^{(1)}, \dots, \mathbf{Q}^{(L+1)}).$$

Матрица $\mathbf{Q}^{(p)}$ полностью описывает расположение параметров в p -м слое и их влияние на линейное преобразование этого слоя.

Для дальнейшего анализа вводятся дополнительные обозначения, которые описывают предшествующие и последующие преобразования относительно p -го слоя. Преобразования, которые последуют после p -го слоя:

$$\begin{aligned} \mathbf{G}^{(p)} &:= \mathbf{T}^{(L+1)} \mathbf{\Lambda}^{(L)} \dots \mathbf{T}^{(p+1)} \mathbf{\Lambda}^{(p)}; \\ \mathbf{G}^{(L+1)} &:= \mathbf{I}; \end{aligned}$$

и преобразования, которые предшествовали p -му слою:

$$\begin{aligned} \mathbf{R}^{(p)} &:= \mathbf{\Lambda}^{(p)} \mathbf{T}^{(p)} \dots \mathbf{\Lambda}^{(1)} \mathbf{T}^{(1)}; p = \overline{1, L}; \\ \mathbf{R}^{(0)} &:= \mathbf{I}. \end{aligned}$$

Матрица $\mathbf{G}^{(p)}$ описывает линейные преобразования от p -го слоя к выходу сети, в то время как $\mathbf{R}^{(p)}$ описывает линейные преобразования входа до p -го слоя.

Используя введенные обозначения, запишем следующие выражения:

$$\begin{aligned} \mathbf{z} &= \mathbf{G}^{(p)} \mathbf{z}^{(p)}, \\ \mathbf{x}^{(p)} &= \mathbf{R}^{(p)} \mathbf{x}, \\ \mathbf{z} &= f_{\theta}(\mathbf{x}) = \mathbf{G}^{(p)} \mathbf{T}^{(p)} \mathbf{R}^{(p-1)} \mathbf{x}. \end{aligned}$$

Первое уравнение выражает выход сети через промежуточные значения на p -м слое, второе показывает преобразование входного сигнала до p -го слоя, а третье дает полное представление выхода сети через параметры p -го слоя и преобразования до и после него.

Объединяя матрицы $\mathbf{G}^{(p)}$ и $\mathbf{R}^{(p)}$ в единый оператор, получаем расширенную матрицу:

$$\mathbf{F}^T := \begin{pmatrix} \mathbf{G}^{(1)^T} \otimes \mathbf{R}^{(0)} \mathbf{x} \\ \vdots \\ \mathbf{G}^{(k)^T} \otimes \mathbf{R}^{(k-1)} \mathbf{x} \\ \vdots \\ \mathbf{G}^{(L+1)^T} \otimes \mathbf{R}^{(L)} \mathbf{x} \end{pmatrix}.$$

Эта блочная матрица содержит в себе всю информацию о том, как изменения параметров в различных слоях влияют на выход сети. Каждый блок этой матрицы соответствует определенному слою и содержит информацию как о линейных преобразованиях от этого слоя к выходу ($\mathbf{G}^{(k)^T}$), так и о преобразовании входа до этого слоя ($\mathbf{R}^{(k-1)} \mathbf{x}$).

В случае использования функции потерь кросс-энтропии (СЕ) для задач классификации, гессиан функции потерь относительно логитов имеет структуру:

$$\mathbf{A} := \nabla_{\mathbf{z}}^2 \ell = \text{diag}(\mathbf{p}) - \mathbf{p} \mathbf{p}^T,$$

где $\mathbf{p} := \text{softmax}(\mathbf{z})$ представляет вектор вероятностей, предсказанных моделью. Матрица \mathbf{A} является ковариационной матрицей многомерного распределения и обладает свойствами положительной полуопределенности и вырожденности, что отражает инвариантность функции softmax к сдвигам в пространстве логитов.

3.2.1. Матричная факторизация матрицы Гессе

Теорема 11 (Киселев-Грабовой-Мешков, Факторизация матрицы Гессе). *Пусть функция нейронной сети $f_{\theta}(\mathbf{x})$ представима в виде (3.1), тогда матрица Гессе функции потерь относительно параметров модели может быть представлена в факторизованной форме: $\mathbf{H}_O(\theta) = \mathbf{Q}^T \mathbf{F}^T \mathbf{A} \mathbf{F} \mathbf{Q}$, где \mathbf{H}_O описывает G -компоненту матрицы Гессе.*

Доказательство. Доказательство теоремы основано на последовательном применении цепного правила матричного дифференцирования и использовании

свойств произведения Кронекера. Исходное представление выхода матричной нейросетевой модели:

$$\mathbf{z} = f_{\theta}(\mathbf{x}) = \mathbf{T}^{(L+1)} \mathbf{\Lambda}^{(L)} \mathbf{T}^{(L)} \dots \mathbf{\Lambda}^{(1)} \mathbf{T}^{(1)} \mathbf{x}.$$

Производные выхода сети по параметрам модели вычисляются с использованием цепного правила:

$$\frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(p)}} = \frac{\partial \mathbf{z}}{\partial \mathbf{z}^{(p)}} \frac{\partial \mathbf{z}^{(p)}}{\partial \mathbf{T}^{(p)}} \frac{\partial \mathbf{T}^{(p)}}{\partial \mathbf{W}^{(p)}}.$$

Для вычисления $\frac{\partial \mathbf{z}^{(p)}}{\partial \mathbf{T}^{(p)}}$ используется тождество для векторизации матричных произведений: $\text{vec}(\mathbf{B}\mathbf{V}\mathbf{A}^T) = (\mathbf{A} \otimes \mathbf{B})\text{vec}(\mathbf{V})$. Применяя это тождество с $\mathbf{A} = \mathbf{I}$ и векторизуя $\mathbf{z}^{(p)} = \mathbf{T}^{(p)}\mathbf{x}^{(p-1)}$, получаем:

$$\text{vec}(\mathbf{z}^{(p)}) = \text{vec}(\mathbf{T}^{(p)}\mathbf{x}^{(p-1)}) = (\mathbf{I} \otimes \mathbf{x}^{(p-1)^T})\text{vec}(\mathbf{T}^{(p)}).$$

Отсюда следует, что:

$$\frac{\partial \mathbf{z}^{(p)}}{\partial \mathbf{T}^{(p)}} = \mathbf{I} \otimes \mathbf{x}^{(p-1)^T}.$$

Используя выражение $\mathbf{z} = \mathbf{G}^{(p)}\mathbf{z}^{(p)}$, получаем производную выхода сети по промежуточному значению:

$$\frac{\partial \mathbf{z}}{\partial \mathbf{z}^{(p)}} = \mathbf{G}^{(p)}.$$

По определению $\mathbf{Q}^{(p)}$ имеем:

$$\frac{\partial \mathbf{T}^{(p)}}{\partial \mathbf{W}^{(p)}} = \mathbf{Q}^{(p)}.$$

Для объединения этих выражений используется свойство произведения Кронекера: если $\mathbf{A}_i \in \mathbb{R}^{m_i \times n_i}$, то $\mathbf{A}_1 \otimes \mathbf{A}_2 = (\mathbf{A}_1 \otimes \mathbf{I}_{m_2})(\mathbf{I}_{m_1} \otimes \mathbf{A}_2)$. Применяя это свойство с $m_2 = 1$, получаем:

$$\mathbf{G}^{(p)}(\mathbf{I} \otimes \mathbf{x}^{(p-1)^T}) = (\mathbf{G}^{(p)} \otimes \mathbf{I}_1)(\mathbf{I} \otimes \mathbf{x}^{(p-1)^T}) = \mathbf{G}^{(p)} \otimes \mathbf{x}^{(p-1)^T}.$$

Подставляя все компоненты в исходную формулу для производной, получаем окончательное выражение:

$$\frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(p)}} = (\mathbf{G}^{(p)} \otimes \mathbf{I}_1)(\mathbf{I} \otimes \mathbf{x}^{(p-1)^T})\mathbf{Q}^{(p)} = (\mathbf{G}^{(p)} \otimes \mathbf{x}^{(p-1)^T})\mathbf{Q}^{(p)}.$$

Используя результаты работ по анализу гессиана в нейронных сетях [43], получаем выражение для блоков матрицы Гессе:

$$\begin{aligned}\mathbf{H}_O^{(kl)} &= J(\boldsymbol{\theta})^T \mathbf{A} J(\boldsymbol{\theta}) = \\ &= \mathbf{Q}^{(k)^T} (\mathbf{G}^{(k)^T} \otimes \mathbf{R}^{(k-1)} \mathbf{x}) A (\mathbf{G}^{(l)} \otimes \mathbf{x}^T \mathbf{R}^{(l-1)^T}) \mathbf{Q}^{(l)}.\end{aligned}$$

Объединяя все блоки в единую матрицу, получаем итоговое выражение для матрицы Гессе:

$$\mathbf{H}_O = \mathbf{Q}^T \mathbf{F} \mathbf{A} \mathbf{F}^T \mathbf{Q}.$$

□

Теорема 11 устанавливает результат о структуре гессиана в матричных моделях глубокого обучения. Предложенная факторизация позволяет эффективно анализировать и вычислять гессиан без необходимости явного построения полной матрицы вторых производных, что особенно важно для моделей с большим количеством параметров.

Структура $\mathbf{H}_O = \mathbf{Q}^T \mathbf{F} \mathbf{A} \mathbf{F}^T \mathbf{Q}$ подчеркивает, что гессиан может быть представлен как преобразование "внутреннего" гессиана \mathbf{A} с помощью матриц \mathbf{F} и \mathbf{Q} , которые отражают (англ. capture) архитектурные свойства сети и параметризацию слоев соответственно.

3.2.2. Оценка спектральных норм матрицы Гессе

В настоящем подразделе устанавливаются верхние оценки спектральных норм матрицы Гессе для матричных моделей глубокого обучения.

Теорема 12 устанавливает верхнюю оценку для нормы матрицы Гессе в терминах структурных параметров нейронной сети. Полученная оценка демонстрирует экспоненциальную зависимость от глубины сети L и полиномиальную зависимость от норм параметров $\mathbf{Q}^{(p)}$ и $\mathbf{T}^{(p)}$.

Особенностью данной теоремы является учет влияния всех слоев сети через мультипликативные члены $w_{\mathbf{T}}^{2L}$ и аддитивные члены $(L + 1)$, что качественно отражает накопление сложности при увеличении глубины архитектуры. Оценка также подчеркивает важность контроля норм весовых матриц на протяжении всего процесса обучения для обеспечения устойчивости оптимизации.

Теорема 12 (Киселев-Грабовой-Мешков, Оценка нормы матрицы Гессе для матричных сетей). Пусть нейронная сеть $f_{\boldsymbol{\theta}}(\mathbf{x})$ представима в виде (3.1).

Пусть для всех слоев p выполнены условия:

$$\begin{aligned}\|\mathbf{Q}^{(p)}\| &\leq q, \\ \|\mathbf{T}^{(p)}\|^2 &\leq w_{\mathbf{T}}^2.\end{aligned}$$

Тогда справедлива оценка:

$$\|\mathbf{H}_O\| \leq \sqrt{2}q^2 \|\mathbf{x}\|^2 (L+1)w_{\mathbf{T}}^{2L}.$$

Доказательство. Согласно теореме 11, матрица Гессе представима в виде $\mathbf{H}_O = \mathbf{Q}^\top \mathbf{F}^\top \mathbf{A} \mathbf{F} \mathbf{Q}$. Используя субмультипликативное свойство спектральной нормы, получаем:

$$\|\mathbf{H}_O\| \leq \|\mathbf{Q}\|^2 \|\mathbf{F}\|^2 \|\mathbf{A}\|.$$

Таким образом, задача сводится к оценке норм \mathbf{A} , \mathbf{F} и \mathbf{Q} .

Матрица $\mathbf{A} = \text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top$ представляет собой гессиан функции потерь относительно логитов. Согласно результатам работы [62], для кросс-энтропийной функции потерь справедлива оценка:

$$\|\mathbf{A}\| \leq \sqrt{2}.$$

Матрица \mathbf{Q} является блочно-диагональной с блоками $\mathbf{Q}^{(1)}, \dots, \mathbf{Q}^{(L+1)}$. Для блочно-диагональных матриц спектральная норма не превосходит максимальной нормы ее блоков:

$$\|\mathbf{Q}\| \leq \max_{i=1,\dots,L+1} \|\mathbf{Q}^{(i)}\|.$$

Из условия теоремы $\|\mathbf{Q}^{(i)}\| \leq q$ для всех i . Следовательно:

$$\|\mathbf{Q}\|^2 \leq q^2.$$

Матрица $\mathbf{G}^{(p)}$ и матрица $\mathbf{R}^{(p)}$ представляют собой произведения матриц $\mathbf{T}^{(i)}$ и диагональных матриц активаций $\Lambda^{(i)}$. Поскольку диагональные элементы матриц $\Lambda^{(i)}$ равны 0 или 1, их спектральная норма не превосходит 1. Для матрицы $\mathbf{G}^{(p)} = \mathbf{T}^{(L+1)} \Lambda^{(L)} \dots \mathbf{T}^{(p+1)} \Lambda^{(p)}$ применяем субмультипликативное свойство:

$$\|\mathbf{G}^{(p)}\| \leq \|\mathbf{T}^{(L+1)}\| \cdot \|\Lambda^{(L)}\| \dots \|\mathbf{T}^{(p+1)}\| \cdot \|\Lambda^{(p)}\| \leq w_{\mathbf{T}}^{L-p+1}.$$

Аналогично для $\mathbf{R}^{(p-1)} = \Lambda^{(p-1)} \mathbf{T}^{(p-1)} \dots \Lambda^{(1)} \mathbf{T}^{(1)}$ получаем оценку:

$$\|\mathbf{R}^{(p-1)}\| \leq w_{\mathbf{T}}^{p-1}.$$

Матрица \mathbf{F} представляет собой вертикальную конкатенацию блоков вида $\mathbf{G}^{(p)\top} \otimes \mathbf{R}^{(p-1)} \mathbf{x}$. Для вертикально сконкатенированных матриц спектральная норма оценивается как корень из суммы квадратов норм блоков:

$$\|\mathbf{F}\|^2 \leq \sum_{p=1}^{L+1} \|\mathbf{G}^{(p)\top} \otimes \mathbf{R}^{(p-1)} \mathbf{x}\|^2,$$

Используя свойство нормы произведения Кронекера $\|\mathbf{A} \otimes \mathbf{B}\| = \|\mathbf{A}\| \cdot \|\mathbf{B}\|$, получаем:

$$\|\mathbf{F}\|^2 \leq \sum_{p=1}^{L+1} \|\mathbf{G}^{(p)}\|^2 \cdot \|\mathbf{R}^{(p-1)} \mathbf{x}\|^2.$$

Учитывая, что $\|\mathbf{R}^{(p-1)} \mathbf{x}\| \leq \|\mathbf{R}^{(p-1)}\| \cdot \|\mathbf{x}\|$, и подставляя оценки, полученные ранее, получаем:

$$\|\mathbf{F}\|^2 \leq \|\mathbf{x}\|^2 \sum_{p=1}^{L+1} w_{\mathbf{T}}^{2(L-p+1)} \cdot w_{\mathbf{T}}^{2(p-1)} = \|\mathbf{x}\|^2 \sum_{p=1}^{L+1} w_{\mathbf{T}}^{2L} = \|\mathbf{x}\|^2 (L+1) w_{\mathbf{T}}^{2L}.$$

Собирая все полученные оценки, получаем:

$$\|\mathbf{H}_O\| \leq \|\mathbf{Q}\|^2 \cdot \|\mathbf{F}\|^2 \cdot \|\mathbf{A}\| \leq q^2 \cdot \|\mathbf{x}\|^2 (L+1) w_{\mathbf{T}}^{2L} \cdot \sqrt{2}.$$

□

В настоящем подразделе рассматривается частный случай модели глубокого обучения, удовлетворяющей условию матричной факторизации для сверточной нейронной сети (англ. CNN) с одномерной сверткой. Оценки на норму матрицы Гессе устанавливаются в теореме 13.

Для сохранения единообразия обозначений используется символ $\mathbf{T}^{(p)}$, который в данном контексте соответствует одномерным сверткам, представленным в виде линейных операторов. Известно, что сверточные сети часто могут быть представлены в виде линейных сверточных нейронных сетей (англ. LCN). Обычно это относится к представлению сверточных сетей с помощью матриц Теплица [63, 64]. Используются обозначения для матриц Теплица из работы [43]. В указанной работе авторы определили специальный тип матрицы $\mathbf{Q}^{(p)}$, соответствующий структуре одномерной матрицы Теплица.

Одномерная сверточная сеть имеет вид $f_{\theta}(\mathbf{x}) = \mathbf{T}^{(L+1)} * (\sigma(\dots(\sigma(\mathbf{T}^{(1)} * \mathbf{x}))\dots))$, где операция $*$ означает свертку.

Пусть C_p обозначает количество каналов после p -го слоя, а d_p - размер последовательности. Здесь $\mathbf{x}^{(p)} \in \mathbb{R}^{C_p \times d_p}$, $\mathbf{T}^{(p)}$ — одномерный сверточный слой с

ядром $\mathbf{W}^{(p)} \in \mathbb{R}^{C_{p-1} \times C_p \times k_p}$. Для упрощения дальнейших обозначений заменим $\mathbf{x}^{(p)}$ на $\text{vec}(\mathbf{x}^{(p)}) \in \mathbb{R}^{(C_p d_p)}$. Получили:

$$\mathbf{z}^{(p+1)} = \mathbf{T}^{(p+1)} \mathbf{x}^{(p)}.$$

Теорема 13 доказывает верхнюю оценку нормы гессиана для глубокой одномерной сверточной сети. Особенностью полученной оценки является мультипликативная зависимость от глубины сети L и полиномиально-экспоненциальная зависимость от параметров модели — числа каналов C , размера ядра k и длины последовательности d .

Теорема 13 (*Киселев-Грабовой-Мешков*, Оценка нормы матрицы Гессе для 1D-сверточных сетей). *Пусть задана сеть вида:*

$$f_{\theta}x = C_{\mathbf{W}^{(L+1)}} \circ \sigma \circ \cdots \circ \sigma \circ C_{\mathbf{W}^{(1)}},$$

где $C_{\mathbf{W}^{(i)}}$ — одномерная свертка с ядром $\mathbf{W}^{(i)}$, без дополнения (англ. *padding*) и с единичным шагом (англ. *stride*). Пусть заданы следующие верхние оценки на параметры:

$$\begin{aligned} C_l &\leqslant C, \\ k_i &\leqslant k, \\ d_i &\leqslant d_1 := d, \\ |\mathbf{W}_{i,j,k}^{(p)}|^2 &\leqslant w^2. \end{aligned}$$

Тогда норма матрицы Гессе имеет следующую верхнюю оценку:

$$\|\mathbf{H}_O\| \leqslant \sqrt{2} \|x\|^2 d^2 (L + 1) (C^2 w^2 k d)^L.$$

Доказательство. Из теоремы 12 следует, что требуется доказать следующие неравенства:

$$\begin{aligned} \left\| \mathbf{T}^{(p)} \right\|^2 &\leqslant C^2 d k w^2, \\ \left\| \mathbf{Q}^{(p)} \right\|^2 &\leqslant d^2. \end{aligned}$$

Согласно работе [43], матрица $\mathbf{T}^{(p)}$ состоит из блоков размером $C_l \times C_{l-1}$, каждый из которых содержит d_{l-1} строк с расположением элементов ядра в соответствующих позициях. Учитывая ограничения на число каналов $C_l \leqslant C$,

длину последовательности $d_{l-1} \leq d$, размер ядра $k_l \leq k$ и ограниченность элементов ядра $|\mathbf{W}_{i,j,k}^{(p)}|^2 \leq w^2$, получаем оценку:

$$\left\| \mathbf{T}^{(p)} \right\|^2 \leq C^2 dk w^2.$$

Согласно работе [43] рассмотрим матрицы:

$$\frac{\partial \mathbf{T}^{(l)}}{\partial \mathbf{Q}^{(l)}} =: \mathbf{Q}^{(l)} = \mathbf{I}_{C_l} \otimes \begin{pmatrix} \mathbf{I}_{C_{l-1}} \otimes (\pi_R^0 \mathbf{I}_{d_{l-1} \times k_l}) \\ \vdots \\ \mathbf{I}_{C_{l-1}} \otimes (\pi_R^{d_{l-1}-k_l} \mathbf{I}_{d_{l-1} \times k_l}) \end{pmatrix}.$$

Оценим норму этой вертикально сконкатенированной матрицы:

$$\begin{aligned} \left\| \mathbf{Q}^{(l)} \right\| &\leq \sum_{i=0}^{d_{l-1}-k_l} \left\| \pi_R^i \mathbf{I}_{d_{l-1} \times k_l} \right\| \leq \sum_{i=0}^{d_{l-1}-k_l} \left\| \pi_R \right\| = \\ &= \sum_{i=0}^{d_{l-1}-k_l} 1 = d_{l-1} - k_l + 1 = d_l \leq d_1 = d, \end{aligned}$$

Следовательно, получаем, что $\left\| \mathbf{Q}^{(l)} \right\|^2 \leq d^2$.

Собирая все полученные оценки воедино:

$$\left\| \mathbf{H}_O \right\| \leq \left\| \mathbf{Q} \right\|^2 \left\| \mathbf{F} \right\|^2 \left\| \mathbf{A} \right\| \leq \sqrt{2} \|x\|^2 d^2 (L+1) (C^2 w^2 k d)^L.$$

□

Переходим к рассмотрению двумерных сверточных сетей. Аналогично, для простоты сохраним обозначение $\mathbf{T}^{(p)}$ для слоев сверточной сети. Пусть задан $\mathbf{x} \in \mathbb{R}^{m \times n \times C}$ — входное изображение, имеющее размеры (m, n) и C каналов. Обозначим $\mathbf{x}^{(l)} \in \mathbb{R}^{m_i \times n_i \times C_i}$ — вход $(l+1)$ -го слоя, а матрицей $\mathbf{W}^{(l)} \in \mathbb{R}^{C_{l-1} \times C_l \times k_l^1 \times k_l^2}$ — свертку с размерами (k_l^1, k_l^2) , входным и выходным количеством каналов C_{l-1} и C_l соответственно.

Аналогично тому как было сделано для 1D-свертки, используем $\text{vec}(\mathbf{x}) \in \mathbb{R}^{m_i n_i C_i}$ вместо $\mathbf{x} \in \mathbb{R}^{m_i \times n_i \times C_i}$. Исследуется операция свертки над входным тензором. В случае векторизованного входа используется структура Теплица, аналогичная представленной в работе [65], при этом применяется конкретная матрица $\mathbf{T}^{(p)}$, строка которой состоит из элементов $\mathbf{W}_{*,c_2,*,*}^{(p)}$ для c_2 -го канала. То есть каждая строка матрицы $\mathbf{T}^{(p)}$ реализует “применение” ядра к определенной позиции и определенному каналу. Обозначим матрицей $\mathbf{T}_i^{(p)}$ матрицу, которая соответствует $c_2 = c_2(i)$ -му каналу \mathbf{W} .

Далее теорема 14 устанавливает оценку нормы гессиана для глубоких двумерных сверточных сетей. Особенностью полученной оценки является экспоненциальная зависимость от глубины сети L и полиномиальная зависимость от основных параметров архитектуры: числа каналов C , размера ядра k и пространственных размеров $m \times n$.

Теорема 14 (*Киселев-Грабовой-Мешков*, Оценка нормы матрицы Гессе для 2D-сверточных сетей). Пусть задана сеть вида

$$f_{\theta} \mathbf{x} = C_{\mathbf{W}^{(L+1)}} \circ \cdots \circ C_{\mathbf{W}^{(1)}},$$

где $C_{\mathbf{W}^{(l)}}$ — двумерная свертка с ядром $\mathbf{W}^{(i)}$, без дополнения (англ. *padding*) и с единичным шагом (англ. *stride*). Пусть заданы следующие верхние оценки на параметры:

$$\begin{aligned} C_l &\leqslant C, \\ k_i &\leqslant k, \\ m_i &\leqslant m_1 := m, \\ n_i &\leqslant n_1 := n, \\ |\mathbf{W}_{i,j,k}^{(p)}|^2 &\leqslant w^2. \end{aligned}$$

Тогда норма матрицы Гессе имеет следующую верхнюю оценку:

$$\|\mathbf{H}_O\| \leqslant \sqrt{2} \|\mathbf{x}\|^2 q^2 (L+1) (C^2 k^2 w^2 mn)^L,$$

где $q^2 = C^2 k^2 mn$.

Доказательство. Для двумерных сверток в матрице $\mathbf{T}^{(p)}$ выполняется равенство:

$$\left\| \mathbf{T}_{i,*}^{(p)} \right\|^2 = \sum_{c,k,l}^{C_{p-1}, k_p^1, k_p^2} |\mathbf{W}_{c,c_2,k,l}^{(p)}|^2,$$

Следовательно:

$$\left\| \mathbf{T}^{(p)} \right\|_F^2 = \sum_{c_1,i,k,l}^{C_{p-1}, C_p n_p m_p, k_p^1, k_p^2} (\mathbf{W}_{c_1,c_2(i),k,l}^{(p)})^2, \quad (3.2)$$

где предполагается соответствие между выходным каналом c_2 и строкой $\mathbf{T}^{(p)} i$.

По аналогии с доказательством 13, используя 12 необходимо доказать два неравенства:

$$\begin{aligned} \left\| \mathbf{T}^{(p)} \right\| &\leqslant C^2 k^2 w^2 mn, \\ \left\| \mathbf{Q}^{(p)} \right\| &\leqslant C^2 k^2 mn. \end{aligned}$$

Оценим норму $\mathbf{T}^{(p)}$, используя (3.2), получаем:

$$\left\| \mathbf{T}^{(p)} \right\|^2 \leq \left\| \mathbf{T}^{(p)} \right\|_F^2 \leq \sum_i C k^2 w^2 \leq C^2 k^2 w^2 m n.$$

Оценим норму производной слоя по параметрам:

$$\left\| \mathbf{Q}^{(p)} \right\| = \left\| \frac{\partial \mathbf{T}^{(p)}}{\partial \mathbf{W}^{(p)}} \right\|.$$

Как упоминалось ранее, строка $\mathbf{T}^{(p)}$ — является $vec_r(\mathbf{W}_{*,*,*,*}^{(p)})$, расположенная в правильном порядке. Тогда норма строки $\frac{\partial \mathbf{T}_{i,j}^{(p)}}{\partial \mathbf{W}_{c_1,c_2,k_1,k_2}^{(p)}} \neq 0 \iff$ индексы выбраны таким образом, что $T_i^{(p)}$ соответствует c_2 , и в то же время $\mathbf{T}_{i,j}^{(p)}$ соответствует c_1, k_1, k_2 . Это соответствие зависит от конкретной матрицы $\mathbf{T}^{(p)}$. Один индекс i соответствует только одному c_2 , поскольку каждая строка участвует в формировании только одного элемента одного канала. Поскольку только $\mathbf{W}_{*,c_2,*,*}^{(p)}$ участвует в формировании одной строки $\mathbf{T}_{i,*}^{(p)}$, индекс i фиксируется, и соответствующий c_2 также фиксируется. Следовательно, для каждого c_1, k_1, k_2 существует только один столбец j такой, что $\mathbf{T}_{i,j}^{(p)} = \mathbf{W}_{c_1,c_2,k_1,k_2}^{(p)}$. Получаем:

$$\begin{aligned} \sum_{j,c_1,k_1,k_2} \left(\frac{\partial \mathbf{T}_{i,j}^{(p)}}{\partial \mathbf{W}_{c_1,c_2,k_1,k_2}^{(p)}} \right)^2 &= \sum_{c_1,k_1,k_2} \sum_j \left(\frac{\partial \mathbf{T}_{i,j}^{(p)}}{\partial \mathbf{W}_{c_1,c_2,k_1,k_2}^{(p)}} \right)^2 = \\ &= \sum_{c_1,k_1,k_2} 1 = C_{p-1} k_p^1 k_p^2 \leq C k^2. \end{aligned}$$

Используем свойство нормы Фробениуса как верхней оценки спектральной нормы:

$$\begin{aligned} \|\mathbf{Q}\|^2 &\leq \|\mathbf{Q}\|_F^2 = \sum_{i,j,c_1,c_2,k_1,k_2} \left(\frac{\partial \mathbf{T}_{i,j}^{(p)}}{\partial \mathbf{W}_{c_1,c_2,k_1,k_2}^{(p)}} \right)^2 = \\ &= \sum_{i,c_2} \sum_{j,c_1,k_1,k_2} \left(\frac{\partial \mathbf{T}_{i,j}^{(p)}}{\partial \mathbf{W}_{c_1,c_2,k_1,k_2}^{(p)}} \right)^2 \leq \\ &\leq \sum_i C k^2 = C m n C k^2 \leq C^2 k^2 m n. \end{aligned}$$

□

Замечание 5. Полученные оценки в теоремах 13 и 14 имеют недостаток, связанный с тем, что они не учитывают уменьшение размеров после сверточных операций и зависят только от верхних границ параметров.

Рассмотренные выше теоремы 13 и 14 оперируют с сетями, которые состоят исключительно из сверточных слоев, что редко встречается на практике. В теоремах 15, 16, 17 рассматриваются случаи добавления других распространенных слоев в сверточных сетях.

Теорема 15 описывает связь между параметрами сети с операцией макс-пулинга и сложностью оптимизационного ландшафта через норму матрицы Гессе. Полученная оценка показывает, что введение слоя макс-пулинга существенно модифицирует зависимость сложности модели от ее параметров.

Теорема 15 (*Киселев-Грабовой-Мешков*, Оценка нормы матрицы Гессе для MaxPool). *Пусть задана сеть вида:*

$$f_{\theta} \mathbf{x} = \mathbf{T}^{(L+1)} \mathbf{\Lambda}^{(L)} \dots \mathbf{\Lambda}^{(1)} \mathbf{T}^{(1)} \mathbf{x},$$

содержащий слой MaxPool2D в позиции $\mathbf{\Lambda}^{(l)}$ с ядром $k_{\text{pool}} \times k_{\text{pool}}$ вместо активации ReLU. Тогда имеет следующую верхнюю оценку нормы матрицы Гессе:

$$\|\mathbf{H}_O\| \leq \sqrt{2} \|\mathbf{x}\|^2 q^2 \left(\frac{1}{k_{\text{pool}}^2}\right)^{L-l+2} (L+1) (k^2 C^2 w^2 m n)^L,$$

где $q^2 = mnC^2k^2$.

Доказательство. Введем обозначение $\mathbf{M}^{(l)}$ для слоя 2D-Max-Pool. Аналогично сверточным слоям, опишем каждую строку $\mathbf{M}^{(l)}$. Установим некоторые свойства матрицы \mathbf{M} , используемые в дальнейшем: строка \mathbf{M}_{i*} соответствует определенному окну пулинга, а столбец \mathbf{M}_{*j} соответствует элементам, которые умножаются на j -й элемент входа.

Поскольку каждое окно покрывает только один элемент и два различных окна не пересекаются, в каждой строке имеется только один ненулевой элемент. Следовательно,

$$\|\mathbf{M}^{(l)}\| = \sqrt{\lambda_{\max}(\mathbf{M}^{(l)\top} \mathbf{M}^{(l)})} = 1,$$

так как $(\mathbf{M}_{*,i}^{(l)}, \mathbf{M}_{*,j}^{(l)}) \neq 0 \iff (\mathbf{M}_{*,i}^{(l)}, \mathbf{M}_{*,j}^{(l)}) = 1 \iff i = j$, а i -й элемент является максимальным в соответствующем окне.

Для простоты предположим, что $\mathbf{M}^{(l)}$ уменьшает оба размера в k_{pool} раз. Аналогично доказательству теоремы 14 оцениваем компоненты $\mathbf{G}^{(p)}$ и $\mathbf{R}^{(p-1)}$,

однако с учетом нового слоя:

$$\begin{aligned} \|\mathbf{G}^{(p)}\| \|\mathbf{R}^{(p-1)}\| &\leqslant \frac{\prod_{i=1}^{L+1} \|T^{(i)}\|}{\|T^{(p)}\|} \leqslant \\ &\leqslant (C^2 k^2 w^2 m n)^{2L} \left(\frac{1}{k_{\text{pool}}^2} \right)^{L-l+2-I\{p-1 \leqslant l\}} \leqslant \\ &\leqslant (C^2 k^2 w^2 m n)^{2L} \left(\frac{1}{k_{\text{pool}}^2} \right)^{L-l+2}, \end{aligned}$$

Используя данные оценки, получаем

$$\begin{aligned} \|\mathbf{F}\|^2 &\leqslant \|\mathbf{x}\|^2 (L+1) (k^2 C^2 w^2 m n)^{(L)} \left(\frac{1}{k_{\text{pool}}^2} \right)^{L-l+2}, \\ \|\mathbf{H}_O\| &\leqslant \sqrt{2} \|\mathbf{x}\|^2 q^2 \left(\frac{1}{k_{\text{pool}}^2} \right)^{L-l+2} (L+1) (k^2 C^2 w^2 m n)^L, \end{aligned}$$

где $q^2 = m n C^2 k^2$. □

Замечание 6. Как и в предыдущих случаях, норма гессиана растет экспоненциально с глубиной сети L , что согласуется с предыдущими оценками, но появление множителя $\left(\frac{1}{k_{\text{pool}}^2} \right)^{L-l+2}$ указывает на снижающий эффект операции пулинга на сложность оптимизации. Этот фактор отражает уменьшение размерности признакового описания после операции пулинга, что приводит к сокращению эффективной размерности параметрического пространства. Видно, что степень $L-l+2$ показывает, что влияние пулинга распространяется на все последующие слои. Чем раньше расположены слои пулинга, тем сильнее его редуцирующее воздействие на норму гессиана. Следовательно получаем, что операция пулинга частично компенсирует экспоненциальный рост сложности с увеличением глубины сети, однако этот эффект зависит от размера ядра пулинга k_{pool} .

Следующая теорема 16 устанавливает оценку нормы матрицы Гессе для сверточной сети, содержащей слой усредняющего пулинга.

Теорема 16 (Киселев-Грабовой-Мешков, Оценка нормы матрицы Гессе для AvgPool). Пусть задана сеть вида:

$$f_{\theta} \mathbf{x} = \mathbf{T}^{(L+1)} \boldsymbol{\Lambda}^{(L)} \dots \boldsymbol{\Lambda}^{(1)} \mathbf{T}^{(1)} \mathbf{x},$$

содержащая слой AvgPool2D в позиции $\boldsymbol{\Lambda}^{(l)}$ вместо активации ReLU с ядром размера $k_{\text{pool}} \times k_{\text{pool}}$. Тогда оценка нормы матрицы Гессе имеет следующую

верхнюю оценку:

$$\|\mathbf{H}_O\| \leq \sqrt{2} \|\mathbf{x}\|^2 q^2 \left(\frac{1}{k_{\text{pool}}^2}\right)^{L-l+2} (L+1) (k^2 C^2 w^2 m n)^L,$$

$$\text{т.е. } q^2 = m n C^2 k^2.$$

Доказательство. Введем обозначение $\mathbf{A}^{(l)}$ для слоя 2D-Avg-Pool. Установим, что

$$(\mathbf{A}_{*,i}, \mathbf{A}_{*,j}) = \frac{1}{k_{\text{pool}}^4} I\{\text{i, j соответствуют одному и тому же окну}\}.$$

Для обоснования этого рассмотрим формулу:

$$(\mathbf{A}_{*j}, \mathbf{A}_{*i}) = \sum_k \mathbf{A}_{ki} \mathbf{A}_{kj} = \sum_{k: \mathbf{A}_{ki} \neq 0, \mathbf{A}_{kj} \neq 0} \frac{1}{k_{\text{pool}}^4}.$$

После этого, применяя элементарные преобразования над строками и столбцами, приводим матрицу $\mathbf{A}^{(p)\top} \mathbf{A}^{(p)}$ к блочно-диагональной форме, где блоки соответствуют индексам в одном окне усредняющего пулинга. Каждый блок матрицы $\mathbf{A}^{(p)\top} \mathbf{A}^{(p)}$ имеет вид $\mathbf{B}_i = \frac{1}{k_{\text{pool}}^2} \mathbf{1} \mathbf{1}^\top$, где $\mathbf{1} = \mathbf{1}_{k_{\text{pool}}^2} \in \mathbb{R}^{k_{\text{pool}}^2}$ — вектор из единиц, и его норма $\|\mathbf{B}_i\| = \frac{1}{k_{\text{pool}}^2} \|\mathbf{1} \mathbf{1}^\top\| = \frac{1}{k_{\text{pool}}}$. Норма блочно-диагональной матрицы равна максимуму норм блоков:

$$\|\mathbf{A}^{(p)}\| = \max_i \|\mathbf{B}_i\| = \frac{1}{k_{\text{pool}}}.$$

Учитывая, что $\|\mathbf{A}^{(p)}\| \leq 1$, повторим полностью вычисления доказательства Теоремы 15 и получим тот же результат. \square

Замечание 7. Полученная оценка в рамках Теоремы 16 демонстрирует, что операция усредняющего пулинга оказывает аналогичное макс-пулингу влияние на сложность оптимизационного ландшафта, уменьшая норму гессиана за счет множителя $\left(\frac{1}{k_{\text{pool}}^2}\right)^{L-l+2}$. Это подтверждает гипотезу о том, что операции пулинга любого типа способствуют снижению сложности модели и могут рассматриваться как эффективный механизм регуляризации в глубоких сверточных нейронных сетях.

Теорема 17 устанавливает оценку нормы матрицы Гессе для гибридной архитектуры, сочетающей сверточные слои и полносвязный слой. Полученная оценка указывает на мультипликативный вклад обеих частей сети, причем сверточная часть вносит множитель $(k^2 C^2 w^2 m n)^L$, а полносвязная — $(h^2 \tilde{w}^2)^P$.

Теорема 17 (*Киселев-Грабовоий-Мешков*, Оценка нормы матрицы Гессе для полносвязных выходных слоев). Пусть задана сеть с P полносвязными слоями следующего вида:

$$f_{\theta} \mathbf{x} = \mathbf{T}^{(L+P+1)} \mathbf{\Lambda}^{(L+P)} \dots \\ \dots \mathbf{\Lambda}^{(L+1)} \mathbf{T}^{(L+1)} \mathbf{\Lambda}^{(L)} \dots \mathbf{\Lambda}^{(1)} \mathbf{T}^{(1)} \mathbf{x},$$

где $\mathbf{T}^{(L+1+i)}$ — полносвязные слои с h_i параметрами при $i = 1, \dots, P$, а $\mathbf{T}^{(r)}$ — двумерные сверточные слои. Пусть $\left\| \mathbf{T}_{ij}^{(L+1+i)} \right\| \leq \tilde{w}$ и $h_p \leq h$. Тогда в условиях и обозначениях теоремы 14 справедливо неравенство:

$$\|\mathbf{H}_O\| \leq \sqrt{2} \|\mathbf{x}\|^2 q^2 (h^2 \tilde{w}^2)^P (k^2 C^2 w^2 m n)^L \times \\ \times \left(L + 1 + P \frac{h^2 \tilde{w}^2}{k^2 C^2 w^2 m n} \right).$$

Доказательство. Как и в предыдущих доказательствах необходимо оценить

$$\left\| \mathbf{G}^{(p)} \right\|^2 \left\| \mathbf{R}^{(p-1)} \right\|^2,$$

Известно, что

$$\left\| T^{(L+1+p)} \right\|^2 \leq (h^2 \tilde{w}^2) \quad \forall p = 1, \dots, P,$$

Следовательно, получаем:

$$\left\| \mathbf{G}^{(p)} \right\|^2 \left\| \mathbf{R}^{(p-1)} \right\|^2 \leq (h^2 \tilde{w}^2)^P (k^2 C^2 w^2 m n)^L,$$

для $p \leq L + 1$ и

$$\left\| \mathbf{G}^{(p)} \right\|^2 \left\| \mathbf{R}^{(p-1)} \right\|^2 \leq (h^2 \tilde{w}^2)^{P-1} (k^2 C^2 w^2 m n)^{L+1},$$

для $p = L + 2 \dots L + P + 1$. В общей форме:

$$\left\| \mathbf{G}^{(p)} \right\|^2 \left\| \mathbf{R}^{(p-1)} \right\|^2 \leq (h^2 \tilde{w}^2)^{P-I_{\{p>L+1\}}} (k^2 C^2 w^2 m n)^{L+I_{\{p>L+1\}}},$$

Следовательно, получаем:

$$\|F\|^2 \leq \sum_{p=1}^{L+P+1} \left\| \mathbf{G}^{(p)} \right\|^2 \left\| \mathbf{R}^{(p-1)} \right\|^2 \|x\|^2 \leq \\ \leq (h^2 \tilde{w}^2)^P (k^2 C^2 w^2 m n)^L \left(L + 1 + P \frac{h^2 \tilde{w}^2}{k^2 C^2 w^2 m n} \right).$$

Применяя этот результат к матрице Гессе:

$$\begin{aligned}\|\mathbf{H}_O\| &\leq \sqrt{2} \|\mathbf{x}\|^2 q^2 (h^2 \tilde{w}^2)^P (k^2 C^2 w^2 m n)^L \times \\ &\quad \times \left(L + 1 + P \frac{h^2 \tilde{w}^2}{k^2 C^2 w^2 m n} \right).\end{aligned}$$

□

3.3. Матрица Гессе для трансформерной модели глубокого обучения

В настоящем разделе рассматриваются оценки норм матриц Гессе для трансформерных моделей глубокого обучения. Трансформеры представляют собой один из наиболее успешных классов архитектур в современном глубоком обучении, широко применяемых в задачах обработки естественного языка, компьютерного зрения и других областях.

Пусть $f_{\mathbf{w}}(\cdot)$ обозначает нейронную сеть, в данном случае слой самовнимания (англ. self-attention) или полный блок трансформера (англ. transformer), с параметрами $\mathbf{w} \in \Omega$. При наличии дважды дифференцируемых потерь $l(\cdot, \cdot)$ потери на выборку равны $l_i(\mathbf{w}) := l(f_{\mathbf{w}}(\mathbf{x}_i), \mathbf{y}_i)$. Эмпирические потери для выборок $L = k$ равны $\mathcal{L}_k(\mathbf{w}) = \frac{1}{k} \sum_{i=1}^k l_i(\mathbf{w})$, с гессианом $\mathbf{H}^{(k)}(\mathbf{w}) = \frac{1}{k} \sum_{i=1}^k \nabla_{\mathbf{w}}^2 l_i(\mathbf{w})$.

Пусть заданы входные вектора эмбедингов (англ. embeddings) $\mathbf{X} \in \mathbb{R}^{L \times d_V}$. Выход слоя одной головы (англ. single-head) слоя самовнимания задается в виде:

$$\mathbf{F}(\mathbf{X}) = \mathbf{A}(\mathbf{X}) \mathbf{X} \mathbf{W}_V, \quad (3.3)$$

где $\mathbf{A}(\mathbf{X}) = \text{softmax} \left(\frac{\mathbf{X} \mathbf{W}_Q \mathbf{W}_K^\top \mathbf{X}^\top}{\sqrt{d_K}} \right)$, а $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{d_V \times d_K}$, $\mathbf{W}_V \in \mathbb{R}^{d_V \times d_V}$.

Используя (3.3), полный блок трансформера выглядит следующим образом:

$$\text{LayerNorm} \left(\text{LayerNorm}(\mathbf{X} + \mathbf{F}(\mathbf{X})) + \text{FFN}(\text{LayerNorm}(\mathbf{X} + \mathbf{F}(\mathbf{X}))) \right)$$

где $\text{FFN}(\cdot)$ является блоком полно связной сети с некоторой нелинейностью. Слой LayerNorm для входной матрицы $\mathbf{U} \in \mathbb{R}^{m \times n}$ описывается выражением:

$$\text{LayerNorm}(\mathbf{U})_{i,j} = \gamma_j \frac{\mathbf{U}_{i,j} - \mu_i}{\sqrt{\sigma_i^2}} + \beta_j,$$

где $\mu_i = \frac{1}{m} \sum_{j=1}^m \mathbf{U}_{i,j}$, $\sigma_i^2 = \frac{1}{m} \sum_{j=1}^m (\mathbf{U}_{i,j} - \mu_i)^2$.

Предположение 2 (Условие положительности дисперсии для LayerNorm).

Для входной матрицы слоя LayerNorm: $\mathbf{X} + \mathbf{F}(\mathbf{X})$, $\mathbf{Y} + \text{FFN}(\mathbf{Y})$, построчная дисперсия удовлетворяет условию $\min_i \sigma_i^2 > 0$.

Предположение 2 является техническим и требуется для доказательства ряда теорем. Выполнение данного свойства достигается добавлением к знаменателю положительной константы, что, однако, усложняет вычисления.

Для оценки матрицы Гессе рассматривается среднеквадратичная функция ошибки:

$$l(\cdot, \text{Target}) = \frac{1}{Ld_V} \|\cdot - \text{Target}\|_F^2.$$

В дальнейшем для доказательств используется разложение Гаусса-Ньютона матрицы Гессе $\mathcal{L}_k \circ f_{\mathbf{w}}$:

$$\frac{\partial^2(\mathcal{L}_k \circ f_{\mathbf{w}})}{\partial \mathbf{W}_i \partial \mathbf{W}_j} = \frac{\partial f_{\mathbf{w}}}{\partial \mathbf{W}_i}(\cdot)^\top \frac{\partial^2 \mathcal{L}_k}{\partial f_{\mathbf{w}}^2}(f_{\mathbf{w}}(\cdot)) \frac{\partial f_{\mathbf{w}}}{\partial \mathbf{W}_j}(\cdot) + \left(\frac{\partial \mathcal{L}_k}{\partial f_{\mathbf{w}}}(f_{\mathbf{w}}(\cdot)) \otimes \mathbf{I}_{p_i q_i} \right) \frac{\partial^2 f_{\mathbf{w}}}{\partial \mathbf{W}_i \partial \mathbf{W}_j}(\cdot) \quad (3.4)$$

Вычисляются обобщенные выражения матрицы Гессе для слоя самовнимания, которые затем расширяются до полного блока модели трансформера. Подход основан на теоретической базе работы [44], адаптируя и обобщая ее результаты.

Матрица Гессе функции ошибки \mathcal{L}_k относительно параметров модели \mathbf{w} :

$$\mathbf{H}^{(k)}(\mathbf{w}) = \nabla_{\mathbf{w}}^2 \mathcal{L}_k(\mathbf{w}) = \frac{1}{k} \sum_{i=1}^k \nabla_{\mathbf{w}}^2 l_i(\mathbf{w}) = \frac{1}{k} \sum_{i=1}^k \mathbf{H}_i(\mathbf{w})$$

где $\mathbf{H}_k(\mathbf{w})$ является матрицей Гессе блока самовнимания для параметров \mathbf{w} , относящихся к матрицам $\{\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V\}$. Используя разложения Гаусса-Ньютона (3.4):

$$\mathbf{H}_k(\mathbf{W}_i, \mathbf{W}_j) = \frac{\partial^2 l}{\partial \mathbf{W}_i \partial \mathbf{W}_j} = \mathbf{H}_o(\mathbf{W}_i, \mathbf{W}_j) + \mathbf{H}_f(\mathbf{W}_i, \mathbf{W}_j),$$

где \mathbf{H}_o является внешнепроизведенной (англ. outer-product) частью матрицы Гессе, а \mathbf{H}_f является матрицей Гессе функции самовнимания. Результаты этого разложения вычисляются согласно теоремам 3.1–3.2 в работе [44].

3.3.1. Матрица Гессе для слоя самовнимания

В настоящем подразделе проводится оценка нормы матрицы Гессе для одного слоя самовнимания. Результат данной оценки формулируется в теореме 18.

Теорема 18 (Петров-Грабовой-Киселев, Оценка нормы матрицы Гессе для слоя самовнимания). Пусть $\|\cdot\|_2$ является спектральной нормой, тогда для

слоя самовнимания получаем:

$$\|\mathbf{H}_i(\mathbf{w}^*)\|_2 \leq M,$$

εde

$$\begin{aligned} M = & 3 \cdot \max \left(\frac{2L}{d_V} \|\mathbf{X}\|_2^2, \right. \\ & \frac{8}{L^3 d_V d_K} \|\mathbf{W}_K\|_2^2 \|\mathbf{W}_V\|_2^2 \|\mathbf{X}\|_2^6 + \\ & + \frac{12}{d_V d_K} \sqrt{\min(L, d_V)} (L \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 + \|\mathbf{Target}\|_2) \|\mathbf{W}_V\|_2 \|\mathbf{W}_K\|_2^2 \|\mathbf{X}\|_2^5, \\ & \frac{4}{L d_V \sqrt{d_K}} \|\mathbf{W}_V\|_2 \|\mathbf{W}_K\|_2 \|\mathbf{X}\|_2^4 + \\ & + \frac{4\sqrt{\min(L, d_V)}}{L^2 \sqrt{d_K}} (L \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 + \|\mathbf{Target}\|_2) \|\mathbf{W}_K\|_2 \|\mathbf{X}\|_2^3, \\ & \frac{8}{L^3 d_V d_K} \|\mathbf{W}_K\|_2 \|\mathbf{W}_Q\|_2 \|\mathbf{W}_V\|_2^2 \|\mathbf{X}\|_2^6 + \\ & + \frac{4\sqrt{\min(L, d_V)} (L \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 + \|\mathbf{Target}\|_2)}{L d_V \sqrt{d_K}} \|\mathbf{W}_V\|_2 \cdot \\ & \cdot \left(3L \|\mathbf{W}_K\|_2 \|\mathbf{W}_Q\|_2 \|\mathbf{X}\|_2^5 + \frac{d_V}{L} \|\mathbf{X}\|_2^3 \right) \left. \right). \end{aligned}$$

Доказательство. Используя результаты Леммы А.3 из работы [66], а также свойство 1 и свойство 2 получаем:

$$\left\| \frac{\partial \mathbf{A}}{\partial \mathbf{T}} \right\|_2 = \frac{1}{L} \|\mathbf{I}_L\|_2 \|\mathbf{I}_L - \frac{1}{L} \mathbf{1}_{L \times L}\|_2 \leq \frac{1}{L}$$

Данное неравенство верно в силу того, что $\frac{1}{L} \mathbf{1}_{L \times L}$ является матрицей проекции, поэтому $\mathbf{I}_L - \frac{1}{L} \mathbf{1}_{L \times L}$ также матрица проекции и следовательно норма $\|\mathbf{I}_L - \frac{1}{L} \mathbf{1}_{L \times L}\|_2 \leq 1$.

Далее для аппроксимации нормы матрицы \mathbf{Z}_1 используем те же свойства 1 и 2:

$$\begin{aligned} \|\mathbf{Z}_1\|_2 & \leq \|\mathbf{I}_L \otimes \mathbf{X}^\top\|_2 \left\| \frac{\partial \mathbf{A}}{\partial \mathbf{T}} \right\|_2 \|\mathbf{X} \otimes \mathbf{X}\|_2 \leq \\ & \leq \|\mathbf{X}\|_2 \frac{1}{L} \|\mathbf{X}\|_2^2 = \frac{1}{L} \|\mathbf{X}\|_2^3 \end{aligned} \tag{3.5}$$

где дополнительно было использовано свойство 3 for $\|\mathbf{X}\|_2 = \|\mathbf{X}^\top\|_2$.

Оценим норму матрицы $\|\mathbf{A}\|_2$, которая является матрицей, где каждая строка является результатом применения функции softmax. Следовательно, каждый элемент матрицы $\mathbf{A}_{i,j} \leq 1$. Далее используя свойства 4 получаем $\|\mathbf{A}\|_{\max} \leq \|\mathbf{A}\|_2 \leq \sqrt{Ld_V}\|\mathbf{A}\|_{\max} = L\|\mathbf{A}\|_{\max} \leq L$. Также получаем, что:

$$\|\mathbf{M}_1\|_2 = \|\mathbf{A}\mathbf{X}\|_2 \leq L\|\mathbf{X}\|_2.$$

Итого, легко получаем outer-product матрицы Гессе $\|\mathbf{H}_o(\mathbf{W}_i, \mathbf{W}_j)\|_2$ для разных матриц. В случае матрицы \mathbf{W}_V и матрицы \mathbf{W}_V :

$$\begin{aligned} \|\mathbf{H}_o(\mathbf{W}_V, \mathbf{W}_V)\|_2 &\leq \frac{2}{Ld_V}\|\mathbf{M}_1\|_2^2 \leq \frac{2}{Ld_V}\|\mathbf{A}\|_2^2\|\mathbf{X}\|_2^2 \leq \\ &\leq \frac{2}{Ld_V}L^2\|\mathbf{X}\|_2^2 = \frac{2L}{d_V}\|\mathbf{X}\|_2^2. \end{aligned}$$

Для матриц \mathbf{W}_Q и \mathbf{W}_Q получаем:

$$\begin{aligned} \|\mathbf{H}_o(\mathbf{W}_Q, \mathbf{W}_Q)\|_2 &\leq \left\| \frac{2}{Ld_V d_K} (\mathbf{I}_{d_V} \otimes \mathbf{W}_K^\top) \mathbf{Z}_1^\top (\mathbf{I}_L \otimes \mathbf{W}_V \mathbf{W}_V^\top) \mathbf{Z}_1 (\mathbf{I}_{d_V} \otimes \mathbf{W}_K) \right\|_2 \leq \\ &\leq \frac{2}{Ld_V d_K} \|\mathbf{W}_K\|_2^2 \|\mathbf{Z}_1\|_2^2 \|\mathbf{W}_V\|_2^2 \leq \\ &\leq \frac{2}{Ld_V d_K} \|\mathbf{W}_K\|_2^2 \|\mathbf{W}_V\|_2^2 \frac{1}{L^2} \|\mathbf{X}\|_2^6 = \\ &= \frac{2}{L^3 d_V d_K} \|\mathbf{W}_K\|_2^2 \|\mathbf{W}_V\|_2^2 \|\mathbf{X}\|_2^6. \end{aligned}$$

Между матрицей \mathbf{W}_V и матрицей \mathbf{W}_Q :

$$\begin{aligned} \|\mathbf{H}_o(\mathbf{W}_V, \mathbf{W}_Q)\|_2 &\leq \frac{2}{Ld_V \sqrt{d_K}} \|\mathbf{M}_1^\top \otimes \mathbf{W}_V^\top\|_2 \|\mathbf{Z}_1\|_2 \|\mathbf{I}_{d_V} \otimes \mathbf{W}_K\|_2 \leq \\ &\leq \frac{2}{Ld_V \sqrt{d_K}} L \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 \frac{1}{L} \|\mathbf{X}\|_2^3 \|\mathbf{W}_K\|_2 = \\ &= \frac{2}{Ld_V \sqrt{d_K}} \|\mathbf{W}_V\|_2 \|\mathbf{W}_K\|_2 \|\mathbf{X}\|_2^4 \end{aligned}$$

Между матрицей \mathbf{W}_Q и матрицей \mathbf{W}_K :

$$\begin{aligned} \|\mathbf{H}_o(\mathbf{W}_Q, \mathbf{W}_K)\|_2 &\leq \\ &\leq \frac{2}{Ld_V d_K} \|(\mathbf{I}_{d_V} \otimes \mathbf{W}_K^\top) \mathbf{Z}_1^\top (\mathbf{I}_L \otimes \mathbf{W}_V \mathbf{W}_V^\top) \mathbf{Z}_1 (\mathbf{W}_Q \otimes \mathbf{I}_{d_V}) \mathbf{K} d_K, d_V\|_2 \leq \\ &\leq \frac{2}{L^3 d_V d_K} \|\mathbf{W}_K\|_2 \|\mathbf{W}_Q\|_2 \|\mathbf{W}_V\|_2^2 \|\mathbf{X}\|_2^6. \end{aligned}$$

Для всех оценок использовались свойства 1, 2, а также свойства $\|\mathbf{K}_{d_V d_K}\|_2 = 1$, потому что $\mathbf{K}_{m,n}$ является коммутативной матрицей, описанной в определении 29.

Далее проведем анализ матрицы \mathbf{H}_f . Для этого начнем анализ с матрицы $\mathbf{R}_m = \text{vec}_r(\mathbf{F}(\mathbf{X}) - \text{Target})^T \otimes \mathbf{I}_m$, который описан в рамках Теоремы 3.2 в работе [44]. Так, как $\text{vec}_r(\cdot)$ является функцией векторизации:

$$\begin{aligned} \|\text{vec}_r(\mathbf{F}(\mathbf{X}) - \text{Target})\|_2 &= \|\mathbf{F}(\mathbf{X}) - \text{Target}\|_F \leq \\ &\leq \sqrt{\text{rank}(\mathbf{F}(\mathbf{X}) - \text{Target})} \|\mathbf{F}(\mathbf{X}) - \text{Target}\|_2, \end{aligned}$$

тогда согласно свойству 4 получаем:

$$\begin{aligned} \|\mathbf{R}_m\| &\leq \sqrt{\text{rank}(\mathbf{F}(\mathbf{X}) - \text{Target})} \|\mathbf{F}(\mathbf{X}) - \text{Target}\|_2 \leq \\ &\leq \sqrt{\text{rank}(\mathbf{F}(\mathbf{X}) - \text{Target})} (\|\mathbf{A}\|_2 \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 + \|\text{Target}\|_2) \leq \\ &\leq \sqrt{\text{rank}(\mathbf{F}(\mathbf{X}) - \text{Target})} (L \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 + \|\text{Target}\|_2) \end{aligned}$$

где для получения оценок были использованы свойства матриц 1 и 6. В свою очередь норма матрицы перемешивания (англ. shuffling matrix) оценивается следующим образом:

$$\begin{aligned} \|\mathbf{S}\|_2 &= \|(\mathbf{I}_{d_V} \otimes \mathbf{K}_{d_V, d_V})(\text{vec}_r(\mathbf{I}_{d_V}) \otimes \mathbf{I}_{d_V})\|_2 \leq \\ &\leq \|\text{vec}_r(\mathbf{I}_{d_V})\|_2 = \|\mathbf{I}_{d_V}\|_F = \sqrt{d_V}. \end{aligned}$$

Для верхней оценки нормы матрицы $\left\| \frac{\partial^2 \mathbf{A}}{\partial \mathbf{T}^2} \right\|_2$ воспользуемся Леммой C1 с работы [44], где указано, что:

$$\frac{\partial^2 \mathbf{A}_{i,j}}{\partial \mathbf{T}_{i,:} \partial \mathbf{T}_{i,:}} = \mathbf{A}_{i,j} \left(2\mathbf{A}_{i,:} \mathbf{A}_{i,:}^\top + \mathbf{E}_{j,j}^{L,L} - \text{diag}(\mathbf{A}_{i,:}) - \mathbf{e}_j \mathbf{A}_{i,:}^\top - \mathbf{A}_{i,:} \mathbf{e}_j^\top \right) \in \mathbb{R}^{L \times L}, \quad (3.6)$$

где

$$\mathbf{E}_{j,j}^{L,L} = \mathbf{e}_j \mathbf{e}_j^\top \in \mathbb{R}^{L \times L},$$

поэтому он содержит только один ненулевой элемент, который равен 1 в позиции (j, j) . Кроме того, вторая производная softmax имеет блочно-диагональную структуру. Используя свойство 7 нормы блочно-диагональной матрицы, получаем:

$$\left\| \frac{\partial^2 \mathbf{A}}{\partial \mathbf{T}^2} \right\|_2 = \max_{i,j} \left\| \frac{\partial^2 \mathbf{A}_{i,j}}{\partial \mathbf{T}_{i,:} \partial \mathbf{T}_{i,:}} \right\|_2.$$

Приходим к тому, что требуется оценить следующую норму:

$$\left\| \frac{\partial^2 \mathbf{A}_{i,j}}{\partial \mathbf{T}_{i,:} \partial \mathbf{T}_{i,:}} \right\|_2.$$

Как было указано ранее, $\mathbf{A}_{i,j} \leq 1$. Следовательно, можем оценить матрицу $\|\mathbf{A}_{i,:}\mathbf{A}_{i,:}^\top\|_2$. Поскольку $\mathbf{A}_{i,:}$ является строкой softmax-матрицы, сумма элементов строки равна 1. Таким образом, используя векторно-матричные неравенства, получаем выражение:

$$\|\mathbf{A}_{i,:}\mathbf{A}_{i,:}^\top\|_2 \leq \|\mathbf{A}_{i,:}\|_2^2 \leq \|\mathbf{A}_{i,:}\|_1^2 = 1. \quad (3.7)$$

Аналогично установим, что

$$\|\mathbf{E}_{j,j}^{m,n}\|_2 = \|\mathbf{e}_j\mathbf{e}_j^\top\|_2 \leq 1. \quad (3.8)$$

Перейдем к оценке нормы диагональной матрицы $\|diag(\mathbf{A}_{i,:})\|_2$. Для диагональной матрицы справедливо следующее выражение:

$$\|diag(\mathbf{A}_{i,:})\|_2 = \max_j \mathbf{A}_{i,j} \leq 1. \quad (3.9)$$

Используя оценки (3.7) и (3.9) и (3.8) оценим нормы $\mathbf{e}_j\mathbf{A}_{i,:}^\top$ и $\mathbf{A}_{i,:}\mathbf{e}_j^\top$. Матрицы $\mathbf{e}_j\mathbf{A}_{i,:}^\top$ и $\mathbf{A}_{i,:}\mathbf{e}_j^\top$ являются матрицами ранга 1, при этом каждая из них содержит только одну ненулевую строку или столбец соответственно с элементами матрицы $\mathbf{A}_{i,:}$. Следовательно их спектральные нормы оцениваются сверху нормой матрицы $\|\mathbf{A}_{i,:}\|_2 \leq 1$.

Все слагаемые в выражении (3.6) имеют верхнюю оценку 1. Следовательно:

$$\left\| \frac{\partial^2 \mathbf{A}}{\partial \mathbf{T}^2} \right\|_2 \leq 6$$

Возвращаясь к выражению (3.5) получаем оценку матрицы $\|\mathbf{Z}_2\|_2$:

$$\begin{aligned} \|\mathbf{Z}_2\|_2 &= \| (\mathbf{I}_L \otimes \mathbf{X}^\top \otimes \mathbf{X}^\top \otimes \mathbf{X}^\top) (\partial^2 \mathbf{A} / \partial \mathbf{T}^2) (\mathbf{X} \otimes \mathbf{X}) \|_2 \leq \\ &\leq \|\mathbf{X}\|_2^5 \left\| \frac{\partial^2 \mathbf{A}}{\partial \mathbf{T}^2} \right\|_2 \leq 6 \|\mathbf{X}\|_2^5 \end{aligned}$$

Оцениваем часть \mathbf{H}_f . Для нормы между матрицами \mathbf{W}_V и \mathbf{W}_V :

$$\|\mathbf{H}_f(\mathbf{W}_V, \mathbf{W}_V)\|_2 = 0$$

Для нормы между матрицами \mathbf{W}_Q и \mathbf{W}_Q :

$$\begin{aligned}
\|\mathbf{H}_f(\mathbf{W}_Q, \mathbf{W}_Q)\|_2 &= \frac{2}{Ld_V d_K} \|\mathbf{R}_{d_V d_K} (\mathbf{I}_L \otimes \mathbf{W}_V^\top \otimes \mathbf{I}_{d_V} \otimes \mathbf{W}_K^\top) \mathbf{Z}_2 (\mathbf{I}_{d_V} \otimes \mathbf{W}_K) \|_2, \\
&\leq \frac{2}{Ld_V d_K} \|\mathbf{R}_{d_V d_K}\|_2 \|\mathbf{W}_V\|_2 \|\mathbf{W}_K\|_2 \|\mathbf{Z}_2\|_2 \|\mathbf{W}_K\|_2 \\
&\leq 6 \frac{2}{Ld_V d_K} \sqrt{\text{rank}(\mathbf{F}(\mathbf{X}) - \mathbf{Target})} \left(L \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 + \right. \\
&\quad \left. + \|\mathbf{Target}\|_2 \right) \|\mathbf{W}_V\|_2 \|\mathbf{W}_K\|_2^2 \|\mathbf{X}\|_2^5 = \\
&= \frac{12}{d_V d_K} \sqrt{\text{rank}(\mathbf{F}(\mathbf{X}) - \mathbf{Target})} \left(L \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 + \right. \\
&\quad \left. + \|\mathbf{Target}\|_2 \right) \|\mathbf{W}_V\|_2 \|\mathbf{W}_K\|_2^2 \|\mathbf{X}\|_2^5
\end{aligned}$$

Для нормы между матрицами \mathbf{W}_V и \mathbf{W}_Q :

$$\begin{aligned}
\|\mathbf{H}_f(\mathbf{W}_V, \mathbf{W}_Q)\|_2 &= \frac{2}{Ld_V \sqrt{d_K}} \|\mathbf{R}_{d_V^2} (\mathbf{I}_L \otimes \mathbf{S}) \mathbf{Z}_1 (\mathbf{I}_{d_V} \otimes \mathbf{W}_K) \|_2 \leq \\
&\leq \frac{2}{Ld_V \sqrt{d_K}} \|\mathbf{R}_{d_V^2}\|_2 \|\mathbf{S}\|_2 \|\mathbf{Z}_1\|_2 \|\mathbf{W}_K\|_2 \leq \\
&\leq \frac{2}{Ld_V \sqrt{d_K}} \sqrt{\text{rank}(\mathbf{F}(\mathbf{X}) - \mathbf{Target})} \left(L \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 + \right. \\
&\quad \left. + \|\mathbf{Target}\|_2 \right) \sqrt{d_V} \frac{1}{L} \|\mathbf{X}\|_2^3 \|\mathbf{W}_K\|_2 = \\
&= \frac{2 \sqrt{\text{rank}(\mathbf{F}(\mathbf{X}) - \mathbf{Target})}}{L^2 \sqrt{d_V d_K}} \left(L \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 + \right. \\
&\quad \left. + \|\mathbf{Target}\|_2 \right) \|\mathbf{W}_K\|_2 \|\mathbf{X}\|_2^3
\end{aligned}$$

Для нормы между матрицами \mathbf{W}_Q и \mathbf{W}_K :

$$\begin{aligned}
& \|\mathbf{H}_f(\mathbf{W}_Q, \mathbf{W}_K)\| \leq \\
& \leq \frac{2}{Ld_V d_K} \|\mathbf{R}_{d_V d_K} (\mathbf{I}_L \otimes \mathbf{W}_V^\top \otimes \mathbf{I}_{d_V} \otimes \mathbf{W}_K^\top) \mathbf{Z}_2 (\mathbf{W}_Q \otimes \mathbf{I}_{d_V}) \mathbf{K}_{d_K, d_V}\|_2 + \\
& + \frac{2}{Ld_V \sqrt{d_K}} \|\mathbf{R}_{d_V} (\mathbf{I}_L \otimes \mathbf{W}_V^\top \otimes \mathbf{I}_{d_V}) (\mathbf{Z}_1 \otimes \mathbf{I}_{d_V}) \mathbf{S} \otimes \mathbf{I}_{d_K}\|_2 \leq \\
& \leq \frac{2}{Ld_V d_K} \sqrt{\text{rank}(\mathbf{F}(\mathbf{X}) - \text{Target})} \left(L \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 + \right. \\
& \left. + \|\text{Target}\|_2 \right) \|\mathbf{W}_V\|_2 \|\mathbf{W}_K\|_2 \|\mathbf{W}_Q\|_2 6 \|\mathbf{X}\|_2^5 + \\
& + \frac{2}{Ld_V \sqrt{d_K}} \sqrt{\text{rank}(\mathbf{F}(\mathbf{X}) - \text{Target})} \left(L \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 + \right. \\
& \left. + \|\text{Target}\|_2 \right) \|\mathbf{W}_V\|_2 \frac{1}{L} \|\mathbf{X}\|_2^3 \sqrt{d_V} = \\
& = \frac{2 \sqrt{\text{rank}(\mathbf{F}(\mathbf{X}) - \text{Target})} (L \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 + \|\text{Target}\|_2)}{Ld_V \sqrt{d_V d_K}} \|\mathbf{W}_V\|_2 \cdot \\
& \cdot \left(3L \|\mathbf{W}_K\|_2 \|\mathbf{W}_Q\|_2 \|\mathbf{X}\|_2^5 + \frac{d_V}{L} \|\mathbf{X}\|_2^3 \right).
\end{aligned}$$

Собирая все оценки вместе, используя матричное свойство 4 для всех блоков $\{K, Q, V\}$:

$$\|\mathbf{H}(\mathbf{W}_i, \mathbf{W}_j)\|_2 \leq 3 \max_{i,j \in \{Q, K, V\}} \left(\|\mathbf{H}_o(\mathbf{W}_i, \mathbf{W}_j)\|_2 + \|\mathbf{H}_f(\mathbf{W}_i, \mathbf{W}_j)\|_2 \right)$$

Подставляя оценки получаем следующую оценку на матрицу Гессе:

$$\begin{aligned}
& \|\mathbf{H}(\mathbf{W}_i, \mathbf{W}_j)\|_2 \leq \\
& \leq 3 \max \left(\frac{2L}{d_V} \|\mathbf{X}\|_2^2, \right. \\
& \quad \frac{2}{L^3 d_V d_K} \|\mathbf{W}_K\|_2^2 \|\mathbf{W}_V\|_2^2 \|\mathbf{X}\|_2^6 + \\
& \quad + \frac{12}{d_V d_K} \sqrt{\text{rank}(\mathbf{F}(\mathbf{X}) - \mathbf{\text{Target}})} \left(L \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 + \right. \\
& \quad \left. + \|\mathbf{\text{Target}}\|_2 \right) \|\mathbf{W}_V\|_2 \|\mathbf{W}_K\|_2^2 \|\mathbf{X}\|_2^5, \\
& \quad \frac{2}{L d_V \sqrt{d_K}} \|\mathbf{W}_V\|_2 \|\mathbf{W}_K\|_2 \|\mathbf{X}\|_2^4 + \\
& \quad + \frac{2\sqrt{\text{rank}(\mathbf{F}(\mathbf{X}) - \mathbf{\text{Target}})}}{L^2 \sqrt{d_V d_K}} (L \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 + \|\mathbf{\text{Target}}\|_2) \|\mathbf{W}_K\|_2 \|\mathbf{X}\|_2^3, \\
& \quad \frac{2}{L^3 d_V d_K} \|\mathbf{W}_K\|_2 \|\mathbf{W}_Q\|_2 \|\mathbf{W}_V\|_2^2 \|\mathbf{X}\|_2^6 + \\
& \quad + \frac{2\sqrt{\text{rank}(\mathbf{F}(\mathbf{X}) - \mathbf{\text{Target}})} \left(L \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 + \|\mathbf{\text{Target}}\|_2 \right)}{L d_V \sqrt{d_V d_K}} \cdot \\
& \quad \cdot \|\mathbf{W}_V\|_2 \left(3L \|\mathbf{W}_K\|_2 \|\mathbf{W}_Q\|_2 \|\mathbf{X}\|_2^5 + \frac{d_V}{L} \|\mathbf{X}\|_2^3 \right) \left. \right).
\end{aligned}$$

Полученное выражение почти полностью соответствует выражению M , где для полного соответствия требуется воспользоваться неравенством $\text{rank}(\mathbf{F}(\mathbf{X}) - \mathbf{\text{Target}}) \leq \min(L, d_V)$. \square

Теорема 18 устанавливает верхнюю оценку нормы матрицы Гессе для одного слоя самовнимания. Полученная оценка демонстрирует сложную зависимость от размерностей модели, норм весовых матриц и нормы входных данных $\|\mathbf{X}\|_2$. Особенностью оценки является наличие слагаемых, пропорциональных различным степеням $\|\mathbf{X}\|_2$, что отражает нелинейный характер преобразований в слое самовнимания.

Теорема 18 оценивает только один слой самовнимания. Переходим к оценке полного блока трансформера. Полный трансформерный слой содержит слой самовнимания, блок полно связной сети (англ. FFN) и слои нормализации выходов (англ. LayerNorm). Весь блок описывается следующими выражениями:

$$\mathbf{Y} = \text{LayerNorm}(\mathbf{X} + \mathbf{F}(\mathbf{X})) \tag{3.10}$$

$$\mathbf{Z} = \text{LayerNorm}(\mathbf{Y} + \text{FFN}(\mathbf{Y})),$$

где

$$\text{FFN}(\mathbf{Y}) = \sigma(\mathbf{Y}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2,$$

с матрицами параметров $\mathbf{W}_1 \in \mathbb{R}^{d_V \times d_{\text{ff}}}$, $\mathbf{W}_2 \in \mathbb{R}^{d_{\text{ff}} \times d_V}$, векторами $\mathbf{b}_1 \in \mathbb{R}^{d_{\text{ff}}}$, $\mathbf{b}_2 \in \mathbb{R}^{d_V}$, а также функцией активации σ . Функция $\text{LayerNorm}(\mathbf{X})$ определяется для входной матрицы $\mathbf{X} \in \mathbb{R}^{L \times d_V}$ следующим образом:

$$\text{LayerNorm}(\mathbf{X})_{i,j} = \gamma_j \cdot \frac{\mathbf{X}_{i,j} - \mu_i}{\sqrt{\sigma_i^2}} + \beta_j,$$

где параметры μ_i, σ_i определяются следующим образом:

$$\mu_i = \frac{1}{d_V} \sum_{j=1}^{d_V} \mathbf{X}_{i,j}, \quad \sigma_i^2 = \frac{1}{d_V} \sum_{j=1}^{d_V} (\mathbf{X}_{i,j} - \mu_i)^2,$$

а параметры γ_j, β_j являются настраиваемым в процессе оптимизации.

Итого, получаем полный список параметров полного слоя трансформера:

$$\mathbf{w} = \{\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V, \mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2, \gamma, \beta\}$$

, где γ, β являются параметрами LayerNorm. Для простоты вычисления в некоторых случаях введем предположение, что параметры γ, β являются постоянными и не меняются в процессе оптимизации.

3.3.2. Матрица Гессе для LayerNorm слоя

Для начала вычислим для функции LayerNorm матрицу Якоби относительно параметров модели, для этого докажем теорему 19.

Теорема 19 (Петров-Грабовой-Киселев, Производная LayerNorm). *Пусть задана матрица $\mathbf{X} \in \mathbb{R}^{L \times d_V}$. Определим функцию $\mathbf{M}(\mathbf{X})$ следующим образом:*

$$\begin{aligned} \mathbf{M}(\mathbf{X}) &= \mathbf{X} - \frac{1}{d_V} \mathbf{X} \mathbf{1}_{d_V} \mathbf{1}_{d_V}^\top, \\ \sigma(\mathbf{X}) &= \frac{1}{\sqrt{d_V}} (\mathbf{M}(\mathbf{X})^{\circ 2} \mathbf{1}_{d_V})^{\circ 1/2}, \\ \mathbf{P}(\mathbf{X}) &= \text{diag}^{-1}(\sigma(\mathbf{X})). \end{aligned}$$

Тогда для функции LayerNorm :

$$\text{LayerNorm}(\mathbf{X}) = \mathbf{P}(\mathbf{X}) \mathbf{M}(\mathbf{X}),$$

матрица Якоби относительно переменной \mathbf{X} определяется следующим образом:

$$\begin{aligned} \frac{\partial \text{LayerNorm}(\mathbf{X})}{\partial \mathbf{X}} &= (\mathbf{P}(\mathbf{X}) \otimes \mathbf{I}_{d_V}) \left(\mathbf{I}_{Ld_V} - \frac{1}{d_V} (\mathbf{I}_L \otimes \mathbf{1}_{d_V \times d_V}) \right) + \\ &\quad + (\mathbf{I}_L \otimes \mathbf{M}(\mathbf{X})^\top) \frac{\partial \mathbf{P}(\mathbf{X})}{\partial \mathbf{X}}, \end{aligned}$$

$\varepsilon \partial e$

$$\begin{aligned} \frac{\partial \mathbf{P}}{\partial \mathbf{X}} &= \frac{1}{\sqrt{d_V}} \left(-\mathbf{D}^{-1} \otimes \mathbf{D}^{-\top} \right) \cdot \\ &\quad \cdot (\mathbf{e}_1 \otimes \mathbf{e}_1, \dots, \mathbf{e}_L \otimes \mathbf{e}_L) \cdot \\ &\quad \cdot \left(\text{diag}^{-1}(\text{vec}_r^{1/2}(\mathbf{M}^{\circ 2} \mathbf{1}_{d_V})) (\mathbf{I}_L \otimes \mathbf{1}_{d_V}^\top) \text{diag}(\text{vec}_r(\mathbf{M})) \frac{\partial \mathbf{M}}{\partial \mathbf{X}} \right), \end{aligned}$$

$\varepsilon \partial e \mathbf{D} = \text{diag}(\sigma(\mathbf{X})).$

Доказательство. Представим функцию LayerNorm в матричном виде:

$$\text{LayerNorm}(\mathbf{X}) = \mathbf{P}(\mathbf{X}) \mathbf{M}(\mathbf{X}),$$

где матрица $\mathbf{P}(\mathbf{X}) = \mathbf{D}^{-1}$, а матрица $\mathbf{D} = \text{diag}(\sigma(\mathbf{X}))$, в свою очередь согласно свойству 8 матрица $\mathbf{M}(\mathbf{X}) = (\mathbf{X} - \mu(\mathbf{X}) \mathbf{1}_{d_V}^\top)$.

Используя Лемму 35 получаем выражение для произведения матричнозначных функций:

$$\frac{\partial \text{LayerNorm}(\mathbf{X})}{\partial \mathbf{X}} = (\mathbf{P}(\mathbf{X}) \otimes \mathbf{I}_{d_V}) \frac{\partial \mathbf{M}}{\partial \mathbf{X}} + (\mathbf{I}_L \otimes \mathbf{M}^\top) \frac{\partial \mathbf{P}}{\partial \mathbf{X}}$$

Вычислим значение производной $\frac{\partial \mathbf{M}}{\partial \mathbf{X}}$, используя матричные вычисления $\mathbf{M}(\mathbf{X}) = (\mathbf{X} - \mu(\mathbf{X}) \mathbf{1}_{d_V}^\top) = (\mathbf{X} - \frac{1}{d_V} \mathbf{X} \mathbf{1}_{d_V} \mathbf{1}_{d_V}^\top) = (\mathbf{X} - \frac{1}{d_V} \mathbf{X} \mathbf{1}_{d_V \times d_V})$. Получаем:

$$\frac{\partial \mathbf{M}}{\partial \mathbf{X}} = \frac{\partial (\mathbf{X} - \frac{1}{d_V} \mathbf{X} \mathbf{1}_{d_V \times d_V})}{\partial \mathbf{X}} = (\mathbf{I}_L \otimes \mathbf{I}_{d_V}) - \frac{1}{d_V} (\mathbf{I}_L \otimes \mathbf{1}_{d_V \times d_V})$$

Вычислим значение производной $\frac{\partial \mathbf{P}}{\partial \mathbf{X}}$. Для начала получим выражение для нелинейного преобразования $\sigma(\mathbf{X})$. Данное выражение в матричном виде принимает вид:

$$\sigma(\mathbf{X}) = \left(\frac{1}{d_V} (\mathbf{X} - \mu(\mathbf{X}) \mathbf{1}_{d_V}^\top)^{\circ 2} \mathbf{1}_{d_V} \right)^{\circ \frac{1}{2}} = \frac{1}{\sqrt{d_V}} (\mathbf{M}(\mathbf{X})^{\circ 2} \mathbf{1}_{d_V})^{\circ \frac{1}{2}},$$

где $\circ\alpha$ операция поэлементного взятия степени α описанного в определении 30.
Далее, применив цепное правило получаем:

$$\frac{\partial \mathbf{P}}{\partial \mathbf{X}} = \frac{\partial \mathbf{D}^{-1}}{\partial \mathbf{D}} \frac{\partial \text{diag}(\sigma(\mathbf{X}))}{\partial \sigma(\mathbf{X})} \frac{\partial \sigma(\mathbf{X})}{\partial \mathbf{X}},$$

Используя свойства 37, 38 и 9, получаем выражение для $\frac{\partial \sigma(\mathbf{X})}{\partial \mathbf{X}}$:

$$\frac{\partial \sigma(\mathbf{X})}{\partial \mathbf{X}} = \frac{1}{\sqrt{d_V}} \frac{\partial \tau^{\circ \frac{1}{2}}}{\partial \tau} \frac{\partial \tau}{\partial \mathbf{Q}} \frac{\partial \mathbf{Q}}{\partial \mathbf{X}},$$

где $\tau = \mathbf{Q} \cdot \mathbf{1}_L$, $\mathbf{Q} = \mathbf{M}^{\circ 2}$. Подставляя, получаем:

$$\begin{aligned} \frac{\partial \sigma(\mathbf{X})}{\partial \mathbf{X}} &= \frac{1}{\sqrt{d_V}} \frac{\partial \tau^{\circ \frac{1}{2}}}{\partial \tau} \frac{\partial \mathbf{Q} \cdot \mathbf{1}_{d_V}}{\partial \mathbf{Q}} \frac{\partial \mathbf{M}^{\circ 2}}{\partial \mathbf{M}} \frac{\partial \mathbf{M}}{\partial \mathbf{X}} = \\ &= \frac{1}{\sqrt{d_V}} \frac{1}{2} \text{diag}^{-1}(\text{vec}_r^{\circ \frac{1}{2}}(\tau)) (\mathbf{I}_L \otimes \mathbf{1}_{d_V}^T) 2 \cdot \text{diag}(\text{vec}_r(\mathbf{M})) \frac{\partial \mathbf{M}}{\partial \mathbf{X}} = \\ &= \frac{1}{\sqrt{d_V}} \text{diag}^{-1}(\text{vec}_r^{\circ \frac{1}{2}}(\mathbf{M}^{\circ 2} \cdot \mathbf{1}_{d_V})) \cdot (\mathbf{I}_L \otimes \mathbf{1}_{d_V}^T) \cdot \text{diag}(\text{vec}_r(\mathbf{M})) \frac{\partial \mathbf{M}}{\partial \mathbf{X}}. \end{aligned}$$

Используя Леммы 39 и 40 получаем:

$$\begin{aligned} \frac{\partial \mathbf{P}}{\partial \mathbf{X}} &= \frac{1}{\sqrt{d_V}} (-\mathbf{D}^{-1} \otimes \mathbf{D}^{-\top}) \begin{pmatrix} \mathbf{e}_1 \otimes \mathbf{e}_1 & \dots & \mathbf{e}_L \otimes \mathbf{e}_L \end{pmatrix} \cdot \\ &\quad \cdot \left(\text{diag}^{-1}(\text{vec}_r^{\circ \frac{1}{2}}(\mathbf{M}^{\circ 2} \cdot \mathbf{1}_{d_V})) \cdot (\mathbf{I}_L \otimes \mathbf{1}_{d_V}^T) \cdot \text{diag}(\text{vec}_r(\mathbf{M})) \frac{\partial \mathbf{M}}{\partial \mathbf{X}} \right). \end{aligned}$$

Итого получили все составляющие для вычисления матрицы Якоби для оператора LayerNorm:

$$\begin{aligned} \frac{\partial \text{LayerNorm}(\mathbf{X})}{\partial \mathbf{X}} &= (\mathbf{P}(\mathbf{X}) \otimes \mathbf{I}_{d_V}) \frac{\partial \mathbf{M}}{\partial \mathbf{X}} + (\mathbf{I}_L \otimes \mathbf{M}^\top) \frac{\partial \mathbf{P}}{\partial \mathbf{X}} = \\ &= (\mathbf{P}(\mathbf{X}) \otimes \mathbf{I}_{d_V}) \frac{\partial \mathbf{M}}{\partial \mathbf{X}} + \\ &\quad + (\mathbf{I}_L \otimes \mathbf{M}^\top) \frac{1}{\sqrt{d_V}} (-\mathbf{D}^{-1} \otimes \mathbf{D}^{-\top}) \begin{pmatrix} \mathbf{e}_1 \otimes \mathbf{e}_1 & \dots & \mathbf{e}_L \otimes \mathbf{e}_L \end{pmatrix} \cdot \\ &\quad \cdot \left(\text{diag}^{-1}(\text{vec}_r^{\circ \frac{1}{2}}(\mathbf{M}^{\circ 2} \cdot \mathbf{1}_{d_V})) \cdot (\mathbf{I}_L \otimes \mathbf{1}_{d_V}^T) \cdot \text{diag}(\text{vec}_r(\mathbf{M})) \frac{\partial \mathbf{M}}{\partial \mathbf{X}} \right), \end{aligned}$$

где

$$\mathbf{M}(\mathbf{X}) = (\mathbf{X} - \frac{1}{d_V} \mathbf{X} \mathbf{1}_{d_V \times d_V})$$

$$\mathbf{P}(\mathbf{X}) = \text{diag}^{-1}(\sigma(\mathbf{X}))$$

$$\frac{\partial \mathbf{M}}{\partial \mathbf{X}} = (\mathbf{I}_L \otimes \mathbf{I}_{d_V}) - \frac{1}{d_V} (\mathbf{I}_L \otimes \mathbf{1}_{d_V \times d_V}).$$

□

Теперь вычислим для функции матрицу Гессе относительно параметров модели, для этого докажем теорему 20.

Теорема 20 (Вторая производная LayerNorm). *Пусть задан оператор LayerNorm в виде, аналогичном теореме 19:*

$$\text{LayerNorm}(\mathbf{X}) = \mathbf{P}(\mathbf{X})\mathbf{M}(\mathbf{X}),$$

с матрицей Якоби, полученной в теореме 19:

$$\frac{\partial \text{LayerNorm}}{\partial \mathbf{X}} = (\mathbf{P} \otimes \mathbf{I}_{d_V})\mathbf{G} + (\mathbf{I}_L \otimes \mathbf{M}^\top)\mathbf{H},$$

где дополнительно введены обозначения константы:

$$\mathbf{G} = \left(\mathbf{I}_{Ld_V} - \frac{1}{d_V} (\mathbf{I}_L \otimes \mathbf{1}_{d_V \times d_V}) \right),$$

а также оператор, аналогичный оператору в теореме 19:

$$\mathbf{H} = \frac{\partial \mathbf{P}}{\partial \mathbf{X}}.$$

Тогда для функции LayerNorm матрица Гессе относительно параметров \mathbf{X} имеет вид:

$$\begin{aligned} \frac{\partial^2 \text{LayerNorm}}{\partial \mathbf{X}^2} &= ((\mathbf{P}(\mathbf{X}) \otimes \mathbf{I}_{d_V}) \otimes \mathbf{I}_{Ld_V}) \frac{\partial^2 \mathbf{M}}{\partial \mathbf{X}^2} + \\ &+ (\mathbf{I}_{Ld_V} \otimes \mathbf{G}^\top) \frac{\partial(\mathbf{P}(\mathbf{X}) \otimes \mathbf{I}_{d_V})}{\partial \mathbf{X}} + \\ &+ ((\mathbf{I}_L \otimes \mathbf{M}^\top) \otimes \mathbf{I}_{Ld_V}) \frac{\partial^2 \mathbf{P}}{\partial \mathbf{X}^2} + (\mathbf{I}_{Ld_V} \otimes \mathbf{H}^\top) \frac{\partial(\mathbf{I}_L \otimes \mathbf{M}^\top)}{\partial \mathbf{X}}. \end{aligned} \quad (3.11)$$

Все матрицы в данном выражении явно вычислимы и задаются формулами, которые указаны ниже в доказательстве.

Доказательство. Используя свойство матричного произведения 9, получаем следующее выражение матрицы Гессе для оператора LayerNorm:

$$\begin{aligned} \frac{\partial^2 \text{LayerNorm}}{\partial \mathbf{X}^2} &= ((\mathbf{P}(\mathbf{X}) \otimes \mathbf{I}_{d_V}) \otimes \mathbf{I}_{Ld_V}) \frac{\partial^2 \mathbf{M}}{\partial \mathbf{X}^2} + \\ &+ \left(\mathbf{I}_{Ld_V} \otimes \left(\frac{\partial \mathbf{M}}{\partial \mathbf{X}} \right)^\top \right) \frac{\partial(\mathbf{P}(\mathbf{X}) \otimes \mathbf{I}_{d_V})}{\partial \mathbf{X}} + \\ &+ ((\mathbf{I}_L \otimes \mathbf{M}^\top) \otimes \mathbf{I}_{Ld_V}) \frac{\partial^2 \mathbf{P}}{\partial \mathbf{X}^2} + \\ &+ \left(\mathbf{I}_{Ld_V} \otimes \left(\frac{\partial \mathbf{P}}{\partial \mathbf{X}} \right)^\top \right) \frac{\partial(\mathbf{I}_L \otimes \mathbf{M}^\top)}{\partial \mathbf{X}}, \end{aligned}$$

Отметим, что $\mathbf{P} \in \mathbb{R}^{L \times L}$, $\mathbf{M} \in \mathbb{R}^{L \times d_V}$, $\frac{\partial \mathbf{M}}{\partial \mathbf{X}} \in \mathbb{R}^{Ld_V \times Ld_V}$, $\frac{\partial \mathbf{P}}{\partial \mathbf{X}} \in \mathbb{R}^{L^2 \times Ld_V}$. Используя свойства 12 и лемму 41, получаем следующие выражения первых и вторых производных:

$$\begin{aligned} \frac{\partial^2 \mathbf{M}}{\partial \mathbf{X}^2} &= 0, \\ \frac{\partial(\mathbf{P}(\mathbf{X}) \otimes \mathbf{I}_{d_V})}{\partial \mathbf{X}} &= \frac{\partial(\mathbf{P} \otimes \mathbf{I}_L)}{\partial \mathbf{P}} \frac{\partial \mathbf{P}}{\partial \mathbf{X}} = (\mathbf{I}_L \otimes \mathbf{K}_{L,L} \otimes \mathbf{I}_L) (\mathbf{I}_{L^2} \otimes \text{vec}_r(\mathbf{I}_L)) \frac{\partial \mathbf{P}}{\partial \mathbf{X}}, \\ \frac{\partial(\mathbf{I}_L \otimes \mathbf{M}^\top)}{\partial \mathbf{X}} &= \frac{\partial(\mathbf{I}_L \otimes \mathbf{M}^\top)}{\partial \mathbf{M}^\top} \frac{\partial \mathbf{M}^\top}{\partial \mathbf{M}} \frac{\partial \mathbf{M}}{\partial \mathbf{X}} = \\ &= (\mathbf{I}_L \otimes \mathbf{K}_{d_V,L} \otimes \mathbf{I}_L) (\text{vec}_r(\mathbf{I}_L) \otimes \mathbf{I}_{Ld_V}) \mathbf{K}_{d_V,L} \frac{\partial \mathbf{M}}{\partial \mathbf{X}}. \end{aligned}$$

Перейдем к оценке вторых производных матрицы \mathbf{P} . Рассмотрим каждое слагаемое в матрице подробнее. Матрица \mathbf{D} является диагональной матрицей $\text{diag}(\sigma(\mathbf{X}))$, причем вектор $\sigma(\mathbf{X})$ имеет размерность $L \times 1$, тогда матрица $\mathbf{D} \in \mathbb{R}^{L \times L}$. Следовательно, слагаемое $(-\mathbf{D}^{-1} \otimes \mathbf{D}^{-\top}) \in \mathbb{R}^{L^2 \times L^2}$. Каждый базисный вектор \mathbf{e}_i имеет размерность $L \times 1$, а следовательно $\mathbf{e}_i \otimes \mathbf{e}_i \in \mathbb{R}^{L^2 \times 1}$, и тогда $(\mathbf{e}_1 \otimes \mathbf{e}_1 \quad \dots \quad \mathbf{e}_L \otimes \mathbf{e}_L) \in \mathbb{R}^{L^2 \times L}$. Ранее было доказано, что $\mathbf{M}(\mathbf{X}) \in \mathbb{R}^{L \times d_V}$, тогда $M \cdot \mathbf{1}_{d_V} \in \mathbb{R}^{L \times 1}$, и следовательно слагаемое $\text{diag}^{-1}(\text{vec}_r^{\circ \frac{1}{2}}(\mathbf{M}^{\circ 2} \cdot \mathbf{1}_{d_V}))$ имеет размерность $L \times L$. Следующие слагаемые $(\mathbf{I}_L \otimes \mathbf{1}_{d_V}^\top) \in \mathbb{R}^{L \times Ld_V}$ и $\text{diag}(\text{vec}_r(M)) \in \mathbb{R}^{Ld_V \times Ld_V}$. Последнее слагаемое $\frac{\partial \mathbf{M}}{\partial \mathbf{X}}$ было вычислено ранее, его размерность составляет $Ld_V \times Ld_V$. Итого матрица $\frac{\partial \mathbf{P}}{\partial \mathbf{X}}$ принадлежит пространству $\mathbb{R}^{L^2 \times Ld_V}$.

Для удобства введем обозначение:

$$\frac{\partial \mathbf{P}}{\partial \mathbf{X}} = \frac{1}{\sqrt{d_V}} \mathbf{A}_1(\mathbf{X}) \cdot \mathbf{B}_1(\mathbf{X}),$$

где $\mathbf{A}_1 = (-\mathbf{D}^{-1} \otimes \mathbf{D}^{-\top})$, а \mathbf{B}_1 все остальное, обе матрицы были вычислены ранее. Используя свойство произведения матричнозначных функций 35 вторая производная принимает вид:

$$\frac{\partial^2 \mathbf{P}}{\partial \mathbf{X}^2} = \frac{1}{\sqrt{d_V}} \frac{\partial \mathbf{A}_1(\mathbf{X}) \cdot \mathbf{B}_1(\mathbf{X})}{\partial \mathbf{X}} = \frac{1}{\sqrt{d_V}} (\mathbf{A}_1 \otimes \mathbf{I}_{Ld_V}) \frac{\partial \mathbf{B}_1}{\partial \mathbf{X}} + (\mathbf{I}_{L^2} \otimes \mathbf{B}_1^\top) \frac{\partial \mathbf{A}_1}{\partial \mathbf{X}}.$$

Разберем вторую производную по частям. Вычислим $\frac{\partial \mathbf{A}_1}{\partial \mathbf{X}}$. Используя лемму 42,

получаем следующее выражение:

$$\begin{aligned}\frac{\partial \mathbf{A}_1}{\partial \mathbf{X}} &= \frac{\partial (-\mathbf{D}^{-1} \otimes \mathbf{D}^{-\top})}{\partial \mathbf{X}} = \\ &= (\mathbf{I}_L \otimes \mathbf{K}_{L,L} \otimes \mathbf{I}_L) \left((\mathbf{I}_{L^2} \otimes \text{vec}_r(\mathbf{D}^{-\top})) \cdot \frac{\partial -\mathbf{D}^{-1}}{\partial \mathbf{X}} + \right. \\ &\quad \left. + (\text{vec}_r(-\mathbf{D}^{-1}) \otimes \mathbf{I}_{L^2}) \cdot \frac{\partial \mathbf{D}^{-\top}}{\partial \mathbf{X}} \right).\end{aligned}$$

Далее используя Леммы 41, 39 получаем выражение на:

$$\begin{aligned}\frac{\partial -\mathbf{D}^{-1}}{\partial \mathbf{X}} &= \frac{\partial -\mathbf{D}^{-1}}{\partial \mathbf{D}} \frac{\partial \mathbf{D}}{\partial \mathbf{X}} = (\mathbf{D}^{-1} \otimes \mathbf{D}^{-\top}) \frac{\partial \mathbf{D}}{\partial \mathbf{X}}, \\ \frac{\partial \mathbf{D}^{-\top}}{\partial \mathbf{X}} &= \frac{\partial \mathbf{D}^{-\top}}{\partial \mathbf{D}^{-1}} \frac{\partial \mathbf{D}^{-1}}{\partial \mathbf{D}} \frac{\partial \mathbf{D}}{\partial \mathbf{X}} = \mathbf{K}_{L,L} (-\mathbf{D}^{-1} \otimes \mathbf{D}^{-\top}) \frac{\partial \mathbf{D}}{\partial \mathbf{X}},\end{aligned}$$

где $\frac{\partial \mathbf{D}}{\partial \mathbf{X}}$ вычисляется аналогично тому, как в Теореме 19:

$$\begin{aligned}\frac{\partial \mathbf{D}}{\partial \mathbf{X}} &= \left(\mathbf{e}_1 \otimes \mathbf{e}_1 \dots \mathbf{e}_L \otimes \mathbf{e}_L \right) \cdot \\ &\quad \cdot \left(\text{diag}^{-1}(\text{vec}_r^{\circ \frac{1}{2}}(\mathbf{M}^{\circ 2} \cdot \mathbf{1}_{d_V})) \cdot (\mathbf{I}_L \otimes \mathbf{1}_{d_V}^T) \cdot \text{diag}(\text{vec}_r(\mathbf{M})) \frac{\partial \mathbf{M}}{\partial \mathbf{X}} \right),\end{aligned}$$

заканчивая вывод оценки $\frac{\partial \mathbf{A}_1}{\partial \mathbf{X}}$.

Перейдем к оценке $\frac{\partial \mathbf{B}_1}{\partial \mathbf{X}}$. Для начала снова представим матрицу \mathbf{B}_1 в виде произведения матриц:

$$\mathbf{B}_1 = \mathbf{E} \mathbf{A}_2 \mathbf{B}_2,$$

где введены следующие обозначения матриц:

$$\begin{aligned}\mathbf{A}_2 &= \text{diag}^{-1}(\text{vec}_r^{\circ \frac{1}{2}}(\mathbf{M}^{\circ 2} \cdot \mathbf{1}_{d_V})) \\ \mathbf{B}_2 &= (\mathbf{I}_L \otimes \mathbf{1}_{d_V}^T) \cdot \text{diag}(\text{vec}_r(\mathbf{M})) \frac{\partial \mathbf{M}}{\partial \mathbf{X}} \\ \mathbf{E} &= \left(\mathbf{e}_1 \otimes \mathbf{e}_1 \quad \dots \quad \mathbf{e}_L \otimes \mathbf{e}_L \right).\end{aligned}$$

Матрица \mathbf{E} является константной относительно матрицы \mathbf{X} . Используя результат леммы 35, получаем

$$\begin{aligned}\frac{\partial \mathbf{B}_1}{\partial \mathbf{X}} &= \frac{\partial \mathbf{E} \mathbf{A}_2 \mathbf{B}_2}{\partial (\mathbf{A}_2 \mathbf{B}_2)} \frac{\partial \mathbf{A}_2 \mathbf{B}_2}{\partial \mathbf{X}} = (\mathbf{E} \otimes \mathbf{I}_{Ld_V}) \frac{\partial \mathbf{A}_2 \mathbf{B}_2}{\partial \mathbf{X}} \\ &= (\mathbf{E} \otimes \mathbf{I}_{Ld_V}) \left((\mathbf{A}_2 \otimes \mathbf{I}_{Ld_V}) \frac{\partial \mathbf{B}_2}{\partial \mathbf{X}} + (\mathbf{I}_L \otimes \mathbf{B}_2^\top) \frac{\partial \mathbf{A}_2}{\partial \mathbf{X}} \right).\end{aligned}$$

Далее осталось оценить матрицы $\frac{\partial \mathbf{A}_2}{\partial \mathbf{X}}$ и $\frac{\partial \mathbf{B}_2}{\partial \mathbf{X}}$. Для оценки $\frac{\partial \mathbf{B}_2}{\partial \mathbf{X}}$ разобьем на части:

$$\mathbf{B}_2 = \mathbf{J} \mathbf{A}_3 \mathbf{B}_3,$$

где $\mathbf{J} = (\mathbf{I}_L \otimes \mathbf{1}_{d_V}^T)$, $\mathbf{A}_3 = \text{diag}(\text{vec}_r(\mathbf{M}))$, $\mathbf{B}_3 = \frac{\partial \mathbf{M}}{\partial \mathbf{X}}$. Аналогично, используя Лемму 35 получаем:

$$\begin{aligned} \frac{\partial \mathbf{B}_2}{\partial \mathbf{X}} &= \frac{\partial \mathbf{J} \mathbf{A}_3 \mathbf{B}_3}{\partial (\mathbf{A}_3 \mathbf{B}_3)} \frac{\partial \mathbf{A}_3 \mathbf{B}_3}{\partial \mathbf{X}} = (\mathbf{J} \otimes \mathbf{I}_{Ld_V}) \frac{\partial \mathbf{A}_3 \mathbf{B}_3}{\partial \mathbf{X}} \\ &= (\mathbf{J} \otimes \mathbf{I}_{Ld_V}) \left((\mathbf{A}_3 \otimes \mathbf{I}_{Ld_V}) \frac{\partial \mathbf{B}_3}{\partial \mathbf{X}} + (\mathbf{I}_{Ld_V} \otimes \mathbf{B}_3^\top) \frac{\partial \mathbf{A}_3}{\partial \mathbf{X}} \right), \end{aligned}$$

где

$$\begin{aligned} \frac{\partial \mathbf{A}_3}{\partial \mathbf{X}} &= \frac{\partial \text{diag}(\text{vec}_r(\mathbf{M}))}{\partial \mathbf{X}} = \frac{\partial \text{diag}(\mathbf{v})}{\partial (\mathbf{v})} \frac{\partial \text{vec}_r(\mathbf{M})}{\partial \mathbf{M}} \frac{\partial \mathbf{M}}{\partial \mathbf{X}}, \\ \frac{\partial \mathbf{B}_3}{\partial \mathbf{X}} &= \frac{\partial^2 \mathbf{M}}{\partial \mathbf{X}^2} = 0, \end{aligned}$$

Используя лемму 40, получаем, что $\frac{\partial \text{diag}(\mathbf{v})}{\partial (\mathbf{v})} = \begin{pmatrix} \mathbf{e}_1 \otimes \mathbf{e}_1 & \dots & \mathbf{e}_L \otimes \mathbf{e}_L \end{pmatrix}$, где $\mathbf{e}_i \in \mathbb{R}^{Ld_V \times 1}$, а также $\frac{\partial \text{vec}_r(\mathbf{M})}{\partial \mathbf{M}} = \mathbf{I}_{Ld_V}$.

Собирая все полученные выражения воедино, получаем выражение для $\frac{\partial^2 \mathbf{P}}{\partial \mathbf{X}^2}$:

$$\frac{\partial^2 \mathbf{P}}{\partial \mathbf{X}^2} = \frac{1}{\sqrt{d_V}} (\mathbf{A}_1 \otimes \mathbf{I}_{Ld_V}) \frac{\partial \mathbf{B}_1}{\partial \mathbf{X}} + (\mathbf{I}_{L^2} \otimes \mathbf{B}_1^\top) \frac{\partial \mathbf{A}_1}{\partial \mathbf{X}},$$

где $\frac{\partial \mathbf{B}_1}{\partial \mathbf{X}}$, $\frac{\partial \mathbf{A}_1}{\partial \mathbf{X}}$, \mathbf{B}_1 , \mathbf{A}_1 определены и получены выше.

Итого, все матрицы выражения (3.11) вычислены, что заканчивает доказательство. \square

3.3.3. Матрица Гессе для нелинейности ReLU

Теорема 21 (Петров-Грабовой-Киселев, Производные ReLU). Пусть задана матрица $\mathbf{X} \in \mathbb{R}^{m \times n}$, тогда для оператора ReLU почти всюду верно следующее выражение:

$$\begin{aligned} \frac{\partial \text{ReLU}(\mathbf{X})}{\partial \mathbf{X}} &= \text{diag}(\text{vec}_r(\mathbf{1}_{\{\mathbf{X} > 0\}})), \\ \frac{\partial^2 \text{ReLU}(\mathbf{X})}{\partial \mathbf{X}^2} &= \mathbf{0}. \end{aligned}$$

Доказательство. Оператор ReLU принимает следующий вид:

$$\text{ReLU}(x) = \max(0, x),$$

то есть, для каждого элемента x_{ij} в матрице $\mathbf{X} \in \mathbb{R}^{m \times n}$ получаем:

$$\frac{\partial \text{ReLU}(x_{ij})}{\partial x_{ij}} = \begin{cases} 1 & \text{если } x_{ij} > 0, \\ 0 & \text{если } x_{ij} < 0, \\ \text{неопределено (субградиент } \in [0, 1]) & \text{если } x_{ij} = 0. \end{cases}$$

В случае скалярной величины $x \in \mathbb{R}$, множеством с неопределенным градиентом является множество $\{0\}$, которое является множеством меры 0. Рассматривая же матрицу $\mathbf{X} \in \mathbb{R}^{m \times n}$ как точку в $\mathbb{R}^{m \times n}$, дифференцируемым множеством является множество:

$$\mathcal{N} = \bigcup_{i,j} \{\mathbf{X} \in \mathbb{R}^{m \times n} : x_{ij} = 0\}.$$

Каждое множество $\{x_{ij} = 0\}$ является гиперплоскостью коразмерности 1 в пространстве $\mathbb{R}^{m \times n}$, следовательно, является множеством меры 0. Так как, множество \mathcal{N} является конечным объединением множеств меры 0, то и множество \mathcal{N} также имеет меру 0. Получили, что оператор ReLU является почти всюду дифференцируем в пространстве $\mathbb{R}^{m \times n}$.

Для каждой дифференцируемой точки $\mathbf{X} \notin \mathcal{N}$, применим построчную векторизацию и Лемму 36:

$$\text{vec}_r(d\text{ReLU}(\mathbf{X})) = \text{diag}(\text{vec}_r(\mathbf{1}_{\{\mathbf{X} > 0\}}))\text{vec}_r(d\mathbf{X}),$$

Используя свойство 10 и лемму 40 для диагональной матрицы, получаем:

$$\frac{\partial \text{ReLU}(\mathbf{X})}{\partial \mathbf{X}} = \text{diag}(\text{vec}_r(\mathbf{1}_{\{\mathbf{X} > 0\}})).$$

В силу того, что матрица Якоби является кусочно-постоянной, то ее дифференциал равен нулю почти всюду:

$$d\left(\frac{\partial \text{ReLU}(\mathbf{X})}{\partial \mathbf{X}}\right) = \mathbf{0}, \quad \mathbf{X} \notin \mathcal{N},$$

а следовательно и матрица Гессе почти всюду равна нулевой матрице:

$$\frac{\partial^2 \text{ReLU}(\mathbf{X})}{\partial \mathbf{X}^2} = \mathbf{0}, \quad \mathbf{X} \notin \mathcal{N}.$$

□

3.3.4. Матрица Гессе для трансформера

Лемма 22 (Петров-Грабовой-Киселев, Оценка нормы для слоя внимания). Рассмотрим слой внимания следующего вида:

$$\mathbf{F}(\mathbf{X}) = \mathbf{A}(\mathbf{T})\mathbf{X}\mathbf{W}_V, \quad \mathbf{T} = \frac{1}{\sqrt{d_K}}\mathbf{X}\mathbf{W}_Q\mathbf{W}_K^\top\mathbf{X}^\top,$$

где $\mathbf{X} \in \mathbb{R}^{L \times d_V}$, $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{d_V \times d_K}$, $\mathbf{W}_V \in \mathbb{R}^{d_V \times d_V}$. Матрица внимания $\mathbf{A}(\cdot)$ применяет построчный softmax. Используем построчную векторизацию $\text{vec}_r(\cdot)$ и матрицы перестановки $\mathbf{K}_{m,n}$ из определения 29.

Определим блоки обобщенного функционального гессиана, используя результаты [44] в наших обозначениях vec_r как

$$\mathbf{H}_f(\mathbf{W}_i, \mathbf{W}_j) = \left(\frac{\partial \ell}{\partial \mathbf{F}} \otimes \mathbf{I}_{p_i q_i} \right) \frac{\partial^2 \mathbf{F}}{\partial \mathbf{W}_i \partial \mathbf{W}_j},$$

где $p_i q_i$ — размер матрицы \mathbf{W}_i , а $\frac{\partial \ell}{\partial \mathbf{F}} \in \mathbb{R}^{L \times d_V}$ — градиент функции потерь.

Для квадратичной функции потерь $\ell(\mathbf{F}) = \frac{1}{2}\|\mathbf{F} - \mathbf{Target}\|_F^2$ имеем $\frac{\partial \ell}{\partial \mathbf{F}} = \mathbf{F} - \mathbf{Target}$ и матрицу построчного свертывания

$$\mathbf{R}_m := \text{vec}_r(\mathbf{F}(\mathbf{X}) - \mathbf{Target})^\top \otimes \mathbf{I}_m \in \mathbb{R}^{m \times (m \cdot L d_V)}.$$

Тогда для $i \in \{V, Q, K\}$ с $n_i := p_i q_i$ блоки функционального гессиана могут быть факторизованы как

$$\mathbf{H}_f(\mathbf{W}_i, \mathbf{W}_j) = \mathbf{R}_{n_i} \Phi_{ij}, \quad \Phi_{ij} := \frac{\partial^2 \mathbf{F}}{\partial \mathbf{W}_i \partial \mathbf{W}_j} \in \mathbb{R}^{(L d_V \cdot n_i) \times n_j}.$$

В частности, блоки кривизны модели Φ_{ij} получаются из соответствующих выражений в [44, Теорема. 3.2] удалением левого свертывания \mathbf{R}_{n_i} .

Теперь перечислим явные блоки, необходимые для вывода. Определим фиксированный оператор изменения формы

$$\mathbf{S} := (\mathbf{I}_{d_V} \otimes \mathbf{K}_{d_V, d_V}) (\text{vec}_r \mathbf{I}_{d_V} \otimes \mathbf{I}_{d_V}) \in \mathbb{R}^{d_V^2 \times d_V},$$

и операторы производных softmax

$$\mathbf{Z}_1 := (\mathbf{I}_L \otimes \mathbf{X}^\top) (\partial \mathbf{A} / \partial \mathbf{T}) (\mathbf{X} \otimes \mathbf{X}) \in \mathbb{R}^{L d_V \times d_V^2},$$

$$\mathbf{Z}_2 := (\mathbf{I}_L \otimes \mathbf{X}^\top \otimes \mathbf{X}^\top \otimes \mathbf{X}^\top) \frac{\partial^2 \mathbf{A}}{\partial \mathbf{T}^2} (\mathbf{X} \otimes \mathbf{X}) \in \mathbb{R}^{L d_V^3 \times d_V^2},$$

где $\frac{\partial^2 \mathbf{A}}{\partial \mathbf{T}^2}$ обозначает тензор вторых производных softmax, согласованный с вектором произведениями Кронекера, как указано выше, а \mathbf{Z}_1 — линейный оператор первой производной softmax, используемый в [44].

Тогда вторые производные внимания имеют вид:

$$\begin{aligned}\Phi_{VV} &= \mathbf{0}_{(Ld_V \cdot d_V^2) \times d_V^2}, \\ \Phi_{QQ} &= \frac{2}{Ld_V d_K} (\mathbf{I}_L \otimes \mathbf{W}_V^\top \otimes \mathbf{I}_{d_V} \otimes \mathbf{W}_K^\top) \mathbf{Z}_2 (\mathbf{I}_{d_V} \otimes \mathbf{W}_K) \in \mathbb{R}^{(Ld_V \cdot d_V d_K) \times d_V d_K}, \\ \Phi_{VQ} &= \frac{2}{Ld_V \sqrt{d_K}} (\mathbf{I}_L \otimes \mathbf{S}) \mathbf{Z}_1 (\mathbf{I}_{d_V} \otimes \mathbf{W}_K) \in \mathbb{R}^{(Ld_V \cdot d_V^2) \times d_V d_K}, \\ \Phi_{QK} &= \frac{2}{Ld_V d_K} (\mathbf{I}_L \otimes \mathbf{W}_V^\top \otimes \mathbf{I}_{d_V} \otimes \mathbf{W}_K^\top) \mathbf{Z}_2 (\mathbf{W}_Q \otimes \mathbf{I}_{d_V}) \mathbf{K}_{d_K, d_V} \\ &\quad + \frac{2}{Ld_V \sqrt{d_K}} (\mathbf{I}_{d_V} \otimes \mathbf{W}_V^\top \otimes \mathbf{I}_{d_V}) (\mathbf{Z}_1 \otimes \mathbf{I}_{d_V}) \mathbf{S} \otimes \mathbf{I}_{d_K} \in \mathbb{R}^{(Ld_V \cdot d_V d_K) \times d_V d_K}.\end{aligned}$$

Более того, в силу симметрии вторых производных, Φ_{KQ} равен Φ_{QK} с переставленными $\mathbf{W}_Q, \mathbf{W}_K$ и корректировкой перестановки с помощью \mathbf{K}_{\dots} . Аналогичные симметричные соотношения дают Φ_{QV} и Φ_{KV} из Φ_{VQ} .

Доказательство. По определению обобщенного функционального гессиана в [44],

$$\mathbf{H}_f(\mathbf{W}_i, \mathbf{W}_j) = \left(\frac{\partial \ell}{\partial \mathbf{F}} \otimes \mathbf{I}_{p_i q_i} \right) \frac{\partial^2 \mathbf{F}}{\partial \mathbf{W}_i \partial \mathbf{W}_j}.$$

Для квадратичной функции потерь $\frac{\partial \ell}{\partial \mathbf{F}} = \mathbf{R}_{p_i q_i}$, определенную выше, а следовательно

$$\mathbf{H}_f(\mathbf{W}_i, \mathbf{W}_j) = \mathbf{R}_{n_i} \Phi_{ij},$$

где $\Phi_{ij} = \frac{\partial^2 \mathbf{F}}{\partial \mathbf{W}_i \partial \mathbf{W}_j}$.

Явные формы для \mathbf{H}_f описаны в [44, Thm. 3.2]. Используя выражения для \mathbf{H}_f получаем выражения Φ_{ij} просто удаляя ведущей метрицы \mathbf{R}_{n_i} . \square

Лемма 23 (Петров-Грабовой-Киселев, Оценка норм для FFN и LayerNorm). Пусть заданы матрицы $\mathbf{X} \in \mathbb{R}^{L \times d_V}$, $\mathbf{Y} = \text{LayerNorm}(\mathbf{F}(\mathbf{X}) + \mathbf{X}) \in \mathbb{R}^{L \times d_V}$ и задана сеть

$$\text{FFN}(\mathbf{Y}) = \sigma(\mathbf{Y} \mathbf{W}_1) \mathbf{W}_2, \quad \mathbf{W}_1 \in \mathbb{R}^{d_V \times d_{ff}}, \quad \mathbf{W}_2 \in \mathbb{R}^{d_{ff} \times d_V},$$

пусть также задана $\mathbf{S} = \mathbf{Y} + \text{FFN}(\mathbf{Y}) \in \mathbb{R}^{L \times d_V}$. Тогда выполняются следующие

щие оценки спектральных норм:

$$\|\mathbf{Y}\|_2 \leq \|\mathbf{Y}\|_F = \sqrt{Ld_V}, \quad (3.12)$$

$$\|\text{FFN}(\mathbf{Y})\|_2 \leq \sqrt{\min(L, d_{ff})} \|\mathbf{Y}\|_2 \|\mathbf{W}_1\|_2 \|\mathbf{W}_2\|_2, \quad (3.13)$$

$$\|\mathbf{S}\|_2 \leq \|\mathbf{Y}\|_2 + \|\text{FFN}(\mathbf{Y})\|_2 \leq \sqrt{Ld_V} \left(1 + \sqrt{\min(L, d_{ff})} \|\mathbf{W}_1\|_2 \|\mathbf{W}_2\|_2 \right). \quad (3.14)$$

Доказательство. Оценим $\|\mathbf{Y}\|_2$. Согласно определению LayerNorm в рамках Теоремы 19 получаем:

$$\mathbf{Y} = \mathbf{P}(\mathbf{S}_0)\mathbf{M}(\mathbf{S}_0), \quad \mathbf{S}_0 := \mathbf{F}(\mathbf{X}) + \mathbf{X},$$

где $\mathbf{M}(\mathbf{S}_0) = \mathbf{S}_0 - \frac{1}{d_V} \mathbf{S}_0 \mathbf{1}_{d_V} \mathbf{1}_{d_V}^\top$ и $\mathbf{P} = \text{diag}^{-1}(\sigma)$ с $\sigma = \frac{1}{\sqrt{d_V}} (\mathbf{M}^{\circ 2} \mathbf{1})^{\circ 1/2}$, применяемым построчно. Для любой строки i обозначим \mathbf{m}_i как i -ю строку \mathbf{M} и $\sigma_i = \frac{1}{\sqrt{d_V}} \|\mathbf{m}_i\|_2$. Тогда i -я строка \mathbf{Y} имеет вид $\mathbf{y}_i = \mathbf{m}_i / \sigma_i$, а следовательно

$$\|\mathbf{y}_i\|_2^2 = \frac{\|\mathbf{m}_i\|_2^2}{\sigma_i^2} = \frac{\|\mathbf{m}_i\|_2^2}{(1/d_V) \|\mathbf{m}_i\|_2^2} = d_V.$$

Таким образом, каждая строка \mathbf{Y} имеет евклидову норму $\sqrt{d_V}$. Получаем:

$$\|\mathbf{Y}\|_F^2 = \sum_{i=1}^L \|\mathbf{y}_i\|_2^2 = Ld_V, \quad \text{откуда} \quad \|\mathbf{Y}\|_F = \sqrt{Ld_V}.$$

Используя из свойства 4 неравенство для норм $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F$, получаем (3.12).

Оценим $\|\text{FFN}(\mathbf{Y})\|_2$. Используя свойство 1 получаем следующую оценку:

$$\|\text{FFN}(\mathbf{Y})\|_2 = \|\text{ReLU}(\mathbf{Y}\mathbf{W}_1)\mathbf{W}_2\|_2 \leq \|\text{ReLU}(\mathbf{Y}\mathbf{W}_1)\|_2 \|\mathbf{W}_2\|_2,$$

далее, используя свойство 4, получаем

$$\|\text{ReLU}(\mathbf{Y}\mathbf{W}_1)\|_2 \leq \|\text{ReLU}(\mathbf{Y}\mathbf{W}_1)\|_F,$$

Согласно определению 31, норма $\|\cdot\|_F^2$ представляет собой сумму квадратов. Поэлементно $\sigma(\cdot)$ удовлетворяет условию $0 \leq \sigma(a) \leq |a|$, а следовательно $\sigma(a)^2 \leq a^2$ для каждого элемента $a \in \mathbb{R}$. Поэтому получаем:

$$\|\sigma(\mathbf{Y}\mathbf{W}_1)\|_F \leq \|\mathbf{Y}\mathbf{W}_1\|_F.$$

Используя неравенство $\|\cdot\|_F \leq \sqrt{d} \|\cdot\|_2$ с $d = \text{rank}(\cdot)$ из свойства 4 получаем:

$$\|\mathbf{Y}\mathbf{W}_1\|_F \leq \sqrt{\text{rank}(\mathbf{Y}\mathbf{W}_1)} \|\mathbf{Y}\mathbf{W}_1\|_2,$$

так как $\mathbf{Y}\mathbf{W}_1 \in \mathbb{R}^{L \times d_{ff}}$, $\text{rank}(\mathbf{Y}\mathbf{W}_1) \leq \min(L, d_{ff})$. Таким образом получаем используя свойство 1:

$$\|\mathbf{Y}\mathbf{W}_1\|_F \leq \sqrt{\min(L, d_{ff})} \|\mathbf{Y}\mathbf{W}_1\|_2 \leq \sqrt{\min(L, d_{ff})} \|\mathbf{Y}\|_2 \|\mathbf{W}_1\|_2$$

Собирая все вместе получаем:

$$\|\text{FFN}(\mathbf{Y})\|_2 \leq \|\sigma(\mathbf{Y}\mathbf{W}_1)\|_F \|\mathbf{W}_2\|_2 \leq \sqrt{\min(L, d_{ff})} \|\mathbf{Y}\|_2 \|\mathbf{W}_1\|_2 \|\mathbf{W}_2\|_2,$$

что заканчивает оценку (3.13).

Оценим $\|\mathbf{S}\|_2$. Используя из свойства 6 неравенство для нормы суммы, получаем:

$$\|\mathbf{S}\|_2 = \|\mathbf{Y} + \text{FFN}(\mathbf{Y})\|_2 \leq \|\mathbf{Y}\|_2 + \|\text{FFN}(\mathbf{Y})\|_2,$$

откуда подставляя (3.12) и (3.13), получаем (3.14). \square

Лемма 24 (*Петров-Грабовой-Киселев*, Оценка норм производных LayerNorm). Пусть заданы матрицы $\mathbf{X} \in \mathbb{R}^{m \times n}$. Производная LayerNorm $\mathbf{J}_{\text{LN}}(\mathbf{X}) = \frac{\partial \text{LayerNorm}(\mathbf{X})}{\partial \mathbf{X}}$ вычисляется в соответствии с Теоремой 19, а ее гессиан $\mathbf{H}_{\text{LN}}(\mathbf{X}) = \frac{\partial^2 \text{LayerNorm}(\mathbf{X})}{\partial \mathbf{X}^2}$ вычисляется как в Теореме 20.

Тогда выполняются следующие оценки:

$$\|\mathbf{J}_{\text{LN}}(\mathbf{X})\|_2 \leq \frac{1}{\sigma_{\min}} + \frac{\|\mathbf{X}\|_2^2}{\sqrt{n}\sigma_{\min}^3}, \quad (3.15)$$

$$\|\mathbf{H}_{\text{LN}}(\mathbf{X})\|_2 \leq \frac{\|\mathbf{X}\|_2}{\sigma_{\min}^3} \left(1 + \sqrt{\frac{m}{n}}\right) + \frac{\|\mathbf{X}\|_2^2}{\sqrt{n}\sigma_{\min}^3} + \frac{3\|\mathbf{X}\|_2^3}{n\sigma_{\min}^5}, \quad (3.16)$$

где σ_{\min} обозначает $\min_i \|\mathbf{M}_i\|_2$, где $\mathbf{M}(\mathbf{X}) = \mathbf{X}(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top)$

Доказательство. Согласно Теореме 19:

$$\mathbf{J}_{\text{LN}}(\mathbf{X}) = (\mathbf{P} \otimes \mathbf{I}_n)\mathbf{G} + (\mathbf{I}_m \otimes \mathbf{M}^\top)\mathbf{H},$$

где $\mathbf{G} = \mathbf{I}_{mn} - \frac{1}{n}(\mathbf{I}_m \otimes \mathbf{1}_{n \times n})$, $\mathbf{H} = \frac{\partial \mathbf{P}}{\partial \mathbf{X}}$, и $\mathbf{P} = \text{diag}^{-1}(\boldsymbol{\sigma})$. Используя свойства 2, 1, 6 получаем:

$$\begin{aligned} \|\mathbf{J}_{\text{LN}}(\mathbf{X})\|_2 &\leq \|\mathbf{P} \otimes \mathbf{I}_n\|_2 \|\mathbf{G}\|_2 + \|\mathbf{I}_m \otimes \mathbf{M}^\top\|_2 \|\mathbf{H}\|_2 = \\ &= \|\mathbf{P}\|_2 \|\mathbf{G}\|_2 + \|\mathbf{M}\|_2 \|\mathbf{H}\|_2. \end{aligned}$$

Оценим каждый множитель по отдельности. Множитель $\|\mathbf{G}\|_2 \leq 1$, поскольку $\frac{1}{n}\mathbf{1}_{n \times n}$ является проекцией, следовательно $\|\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\|_2 \leq 1$, и произведение

Кронекера сохраняет оценку спектральной нормы согласно свойств 1, 2, а также Леммы 43. Множитель $\|\mathbf{P}\|_2 = \|\mathbf{D}^{-1}\|_2 = 1/\sigma_{\min}$, где $\mathbf{D} = \text{diag}(\boldsymbol{\sigma})$. Множитель $\|\mathbf{M}\|_2 \leq \|\mathbf{X}\|_2$, потому что $\mathbf{M}(\mathbf{X}) = \mathbf{X}(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top)$, и правый множитель является проектором с нормой ≤ 1 согласно свойства 1. Для $\|\mathbf{H}\|_2 = \left\| \frac{\partial \mathbf{P}}{\partial \mathbf{X}} \right\|_2$, Теорема 19 вместе с Леммами 39, 40, 37, 38 и свойствами 1, 2 дают оценку:

$$\begin{aligned} \left\| \frac{\partial \mathbf{P}}{\partial \mathbf{X}} \right\|_2 &\leq \frac{1}{\sqrt{n}} \left\| \mathbf{D}^{-1} \otimes \mathbf{D}^{-\top} \right\|_2 \left\| \text{diag}^{-1}(\text{vec}_r^{1/2}(\mathbf{M}^{\circ 2}\mathbf{1}_n)) \right\|_2 \\ &\cdot \left\| \mathbf{I}_m \otimes \mathbf{1}_n^\top \right\|_2 \left\| \text{diag}(\text{vec}_r(\mathbf{M})) \right\|_2 \left\| \frac{\partial \mathbf{M}}{\partial \mathbf{X}} \right\|_2. \end{aligned}$$

Используя свойство 4 получаем оценки:

$$\begin{aligned} \left\| \mathbf{D}^{-1} \otimes \mathbf{D}^{-\top} \right\|_2 &= \left\| \mathbf{D}^{-1} \right\|_2^2 = \frac{1}{\sigma_{\min}^2}, \\ \left\| \text{diag}^{-1}(\cdot) \right\|_2 &= \frac{1}{\min_i \sqrt{\sum_v M_{i,v}^2}} = \frac{1}{\sqrt{n}\sigma_{\min}}, \\ \left\| \mathbf{I}_m \otimes \mathbf{1}^\top \right\|_2 &= \sqrt{n}, \\ \left\| \text{diag}(\text{vec}_r(\mathbf{M})) \right\|_2 &= \|\mathbf{M}\|_{\max} \leq \|\mathbf{M}\|_2, \\ \left\| \frac{\partial \mathbf{M}}{\partial \mathbf{X}} \right\|_2 &\leq 1, \end{aligned}$$

откуда получаем оценку:

$$\|\mathbf{H}\|_2 \leq \frac{1}{\sqrt{n}\sigma_{\min}^2} \cdot \frac{1}{\sqrt{n}\sigma_{\min}} \cdot \sqrt{n} \cdot \|\mathbf{M}\|_2 \cdot 1 \leq \frac{\|\mathbf{X}\|_2}{\sqrt{n}\sigma_{\min}^3}.$$

Собирая все полученные оценки, получаем (3.15):

$$\|\mathbf{J}_{\text{LN}}(\mathbf{X})\|_2 \leq \frac{1}{\sigma_{\min}} \cdot 1 + \|\mathbf{X}\|_2 \cdot \frac{\|\mathbf{X}\|_2}{\sqrt{n}\sigma_{\min}^3} = \frac{1}{\sigma_{\min}} + \frac{\|\mathbf{X}\|_2^2}{\sqrt{n}\sigma_{\min}^3}.$$

Из Теоремы 20, используя $\frac{\partial^2 \mathbf{M}}{\partial \mathbf{X}^2} = 0$,

$$\mathbf{H}_{\text{LN}}(\mathbf{X}) = (\mathbf{I}_{mn} \otimes \mathbf{G}^\top) \frac{\partial(\mathbf{P} \otimes \mathbf{I}_n)}{\partial \mathbf{X}} + ((\mathbf{I}_m \otimes \mathbf{M}^\top) \otimes \mathbf{I}_{mn}) \frac{\partial^2 \mathbf{P}}{\partial \mathbf{X}^2} + (\mathbf{I}_{mn} \otimes \mathbf{H}^\top) \frac{\partial(\mathbf{I}_m \otimes \mathbf{M}^\top)}{\partial \mathbf{X}}.$$

Оценим три слагаемых отдельно с помощью свойств 1, 2. Первое слагаемое. По Предложению 12 получаем:

$$\frac{\partial(\mathbf{P} \otimes \mathbf{I}_n)}{\partial \mathbf{X}} = (\mathbf{I}_m \otimes \mathbf{K}_{n,m} \otimes \mathbf{I}_n)(\mathbf{I}_{m^2} \otimes \text{vec}_r(\mathbf{I}_n)) \frac{\partial \mathbf{P}}{\partial \mathbf{X}},$$

а следовательно

$$\begin{aligned} \left\| (\mathbf{I}_{mn} \otimes \mathbf{G}^\top) \frac{\partial(\mathbf{P} \otimes \mathbf{I}_n)}{\partial \mathbf{X}} \right\|_2 &\leq \|\mathbf{G}\|_2 \|\mathbf{I}_{m^2} \otimes \text{vec}_r(\mathbf{I}_n)\|_2 \left\| \frac{\partial \mathbf{P}}{\partial \mathbf{X}} \right\|_2 = \\ &= 1 \cdot \sqrt{n} \cdot \frac{\|\mathbf{X}\|_2}{\sqrt{n}\sigma_{\min}^3} = \frac{\|\mathbf{X}\|_2}{\sigma_{\min}^3}. \end{aligned}$$

Второе слагаемое. Используя $\|\mathbf{I}_m \otimes \mathbf{M}^\top\|_2 = \|\mathbf{M}\|_2 \leq \|\mathbf{X}\|_2$ и оценку для $\left\| \frac{\partial^2 \mathbf{P}}{\partial \mathbf{X}^2} \right\|_2$ получаем:

$$\left\| ((\mathbf{I}_m \otimes \mathbf{M}^\top) \otimes \mathbf{I}_{mn}) \frac{\partial^2 \mathbf{P}}{\partial \mathbf{X}^2} \right\|_2 \leq \|\mathbf{X}\|_2 \left\| \frac{\partial^2 \mathbf{P}}{\partial \mathbf{X}^2} \right\|_2.$$

Оценим $\left\| \frac{\partial^2 \mathbf{P}}{\partial \mathbf{X}^2} \right\|_2$, следуя той же цепочке, что и в доказательстве теоремы 20 запишем $\frac{\partial \mathbf{P}}{\partial \mathbf{X}} = \frac{1}{\sqrt{n}} \mathbf{A}_1(\mathbf{X}) \mathbf{E} \mathbf{B}_1(\mathbf{X})$ и продифференцируем, используя свойство 1 с Леммами 39, 40, 37, 38 и свойствами 1, 2, 4. Получаем:

$$\left\| \frac{\partial^2 \mathbf{P}}{\partial \mathbf{X}^2} \right\|_2 \leq \frac{1}{\sqrt{n}\sigma_{\min}^3} \|\mathbf{X}\|_2 + \frac{3}{n\sigma_{\min}^5} \|\mathbf{X}\|_2^2,$$

а следовательно,

$$\left\| ((\mathbf{I}_m \otimes \mathbf{M}^\top) \otimes \mathbf{I}_{mn}) \frac{\partial^2 \mathbf{P}}{\partial \mathbf{X}^2} \right\|_2 \leq \frac{\|\mathbf{X}\|_2^2}{\sqrt{n}\sigma_{\min}^3} + \frac{3\|\mathbf{X}\|_2^3}{n\sigma_{\min}^5}.$$

Третье слагаемое. По свойству 12 и Лемме 41 получаем:

$$\frac{\partial(\mathbf{I}_m \otimes \mathbf{M}^\top)}{\partial \mathbf{X}} = (\mathbf{I}_m \otimes \mathbf{K}_{n,m} \otimes \mathbf{I}_m)(\text{vec}_r(\mathbf{I}_m) \otimes \mathbf{I}_{mn}) \frac{\partial \mathbf{M}}{\partial \mathbf{X}},$$

откуда получаем:

$$\begin{aligned} \left\| (\mathbf{I}_{mn} \otimes \mathbf{H}^\top) \frac{\partial(\mathbf{I}_m \otimes \mathbf{M}^\top)}{\partial \mathbf{X}} \right\|_2 &\leq \|\mathbf{H}\|_2 \|\text{vec}_r(\mathbf{I}_m) \otimes \mathbf{I}_{mn}\|_2 \left\| \frac{\partial \mathbf{M}}{\partial \mathbf{X}} \right\|_2 = \\ &= \frac{\|\mathbf{X}\|_2}{\sqrt{n}\sigma_{\min}^3} \cdot \sqrt{m} \cdot 1 = \frac{\sqrt{m}}{\sqrt{n}} \frac{\|\mathbf{X}\|_2}{\sigma_{\min}^3}. \end{aligned}$$

Суммируя все слагаемые с помощью свойства 6 получаем (3.16):

$$\begin{aligned} \|\mathbf{H}_{LN}(\mathbf{X})\|_2 &\leq \frac{\|\mathbf{X}\|_2}{\sigma_{\min}^3} + \left(\frac{\|\mathbf{X}\|_2^2}{\sqrt{n}\sigma_{\min}^3} + \frac{3\|\mathbf{X}\|_2^3}{n\sigma_{\min}^5} \right) + \frac{\sqrt{m}}{\sqrt{n}} \frac{\|\mathbf{X}\|_2}{\sigma_{\min}^3} = \\ &= \frac{\|\mathbf{X}\|_2}{\sigma_{\min}^3} \left(1 + \sqrt{\frac{m}{n}} \right) + \frac{\|\mathbf{X}\|_2^2}{\sqrt{n}\sigma_{\min}^3} + \frac{3\|\mathbf{X}\|_2^3}{n\sigma_{\min}^5}. \end{aligned}$$

□

Теорема 25 (Петров-Грабовой-Киселев, Матрица Якоби трансформера). Для модели глубокого обучения архитектуры трансформер 3.10 матрица Якоби $\frac{\partial \mathbf{Z}}{\partial \mathbf{W}_i}$ вычисляется в следующем виде:

$$\frac{\partial \mathbf{Z}}{\partial \mathbf{W}_i} = \frac{\partial \text{LayerNorm}(\text{FFN}(\mathbf{Y}) + \mathbf{Y})}{\partial (\text{FFN}(\mathbf{Y}) + \mathbf{Y})} \frac{\partial (\text{FFN}(\mathbf{Y}) + \mathbf{Y})}{\partial \mathbf{W}_i}, \quad i \in \{1, 2\},$$

где

$$\frac{\partial (\text{FFN}(\mathbf{Y}) + \mathbf{Y})}{\partial \mathbf{W}_i} = \begin{cases} (\mathbf{I}_L \otimes \mathbf{W}_2^\top) \text{diag}(\text{vec}_r(\mathbf{1}_{\{\mathbf{X} > 0\}})) (\mathbf{Y} \otimes \mathbf{I}_{d_{ff}}), & \text{for } i = 1 \\ \sigma(\mathbf{Y}\mathbf{W}_1) \otimes \mathbf{I}_{d_V}, & \text{for } i = 2 \end{cases},$$

причем $\frac{\partial \text{LayerNorm}(\text{FFN}(\mathbf{Y}) + \mathbf{Y})}{\partial (\text{FFN}(\mathbf{Y}) + \mathbf{Y})}$ вычисляется согласно Теоремы 19.

$$\frac{\partial \mathbf{Z}}{\partial \mathbf{W}_i} = \frac{\partial \text{LayerNorm}(\text{FFN}(\mathbf{Y}) + \mathbf{Y})}{\partial (\text{FFN}(\mathbf{Y}) + \mathbf{Y})} \frac{\partial (\text{FFN}(\mathbf{Y}) + \mathbf{Y})}{\partial \mathbf{Y}} \frac{\partial \mathbf{Y}}{\partial \mathbf{W}_i}, \quad i \in \{K, Q, V\},$$

где

$$\frac{\partial (\text{FFN}(\mathbf{Y}) + \mathbf{Y})}{\partial \mathbf{Y}} = (\mathbf{I}_L \otimes \mathbf{W}_2^\top) \text{diag}(\text{vec}_r(\mathbf{1}_{\{\mathbf{X} > 0\}})) (\mathbf{I}_L \otimes \mathbf{W}_1^\top) + (\mathbf{I}_L \otimes \mathbf{I}_{d_V}),$$

причем $\frac{\partial \mathbf{Y}}{\partial \mathbf{W}_i} = \frac{\partial \text{LayerNorm}(\mathbf{F}(\mathbf{X}) + \mathbf{X})}{\partial (\mathbf{F}(\mathbf{X}) + \mathbf{X})} \frac{\partial \mathbf{F}(\mathbf{X})}{\partial \mathbf{W}_i}$, где $\frac{\partial \mathbf{F}(\mathbf{X})}{\partial \mathbf{W}_i}$ вычисляется при помощи Леммы A.2 в работе [66], а матрица $\frac{\partial \text{LayerNorm}(\mathbf{F}(\mathbf{X}) + \mathbf{X})}{\partial (\mathbf{F}(\mathbf{X}) + \mathbf{X})}$ вычисляется согласно Теоремы 19.

Доказательство. В дальнейшем при доказательстве вводим следующие обозначения и предположения $\mathbf{X} \in R^{L \times d_V}, \mathbf{Y} \in R^{L \times d_V}, \mathbf{W}_1 \in R^{d_V \times d_{ff}}, \text{ReLU}(\mathbf{Y}\mathbf{W}_1) \in R^{L \times d_{ff}}, \mathbf{W}_2 \in R^{d_{ff} \times d_V}$. Трансформер блок определен в выражении (3.10), а именно:

$$\begin{aligned} \mathbf{Y} &= \text{LayerNorm}(\mathbf{F}(\mathbf{X}) + \mathbf{X}), \\ \mathbf{Z} &= \text{LayerNorm}(\text{FFN}(\mathbf{Y}) + \mathbf{Y}). \end{aligned}$$

Начнем вычисления матрицы Гессе для полного трансформера с матрицей $\frac{\partial \mathbf{Z}}{\partial \mathbf{W}_i}$, тогда для $i \in \{1, 2\}$ получаем:

$$\frac{\partial \mathbf{Z}}{\partial \mathbf{W}_i} = \frac{\partial \text{LayerNorm}(\text{FFN}(\mathbf{Y}) + \mathbf{Y})}{\partial (\text{FFN}(\mathbf{Y}) + \mathbf{Y})} \frac{\partial (\text{FFN}(\mathbf{Y}) + \mathbf{Y})}{\partial \mathbf{W}_i},$$

где

$$\frac{\partial (\text{FFN}(\mathbf{Y}) + \mathbf{Y})}{\partial \mathbf{W}_i} = \frac{\partial (\text{FFN}(\mathbf{Y}))}{\partial \mathbf{W}_i} = \frac{\partial \mathbf{I}_L \sigma(\mathbf{Y}\mathbf{W}_1) \mathbf{W}_2 \mathbf{I}_{d_V}}{\partial \mathbf{W}_i},$$

причем используя свойство 9 об производной произведения матриц получаем:

$$\begin{aligned}\frac{\partial \mathbf{I}_L \sigma(\mathbf{Y} \mathbf{W}_1) \mathbf{W}_2 \mathbf{I}_{d_V}}{\partial \mathbf{W}_2} &= \sigma(\mathbf{Y} \mathbf{W}_1) \otimes \mathbf{I}_{d_V} \\ \frac{\partial \mathbf{I}_L \sigma(\mathbf{Y} \mathbf{W}_1) \mathbf{W}_2 \mathbf{I}_{d_V}}{\partial \mathbf{W}_1} &= \frac{\partial \sigma(\mathbf{Y} \mathbf{W}_1) \mathbf{W}_2}{\partial \sigma(\mathbf{Y} \mathbf{W}_1)} \frac{\partial \sigma(\mathbf{Y} \mathbf{W}_1)}{\partial \mathbf{Y} \mathbf{W}_1} \frac{\partial \mathbf{Y} \mathbf{W}_1}{\partial \mathbf{W}_1} = \\ &= (\mathbf{I}_L \otimes \mathbf{W}_2^\top) \frac{\partial \sigma(\mathbf{Y} \mathbf{W}_1)}{\partial \mathbf{Y} \mathbf{W}_1} (\mathbf{I}_L \otimes \mathbf{W}_1^\top).\end{aligned}$$

Используя результаты Теоремы 21 для производной оператора ReLU для матрицы $\frac{\partial \sigma(\mathbf{Y} \mathbf{W}_1)}{\partial \mathbf{Y} \mathbf{W}_1}$ получаем следующее выражение:

$$\frac{\partial \mathbf{I}_L \sigma(\mathbf{Y} \mathbf{W}_1) \mathbf{W}_2 \mathbf{I}_{d_V}}{\partial \mathbf{W}_i} = (\mathbf{I}_L \otimes \mathbf{W}_2^\top) \text{diag}(\text{vec}_r(\mathbf{1}_{\{\mathbf{X} > 0\}})) (\mathbf{Y} \otimes \mathbf{I}_{d_{ff}}).$$

Тогда в общем виде для $i \in \{1, 2\}$ получаем следующее выражение:

$$\frac{\partial (\text{FFN}(\mathbf{Y}) + \mathbf{Y})}{\partial \mathbf{W}_i} = \begin{cases} (\mathbf{I}_L \otimes \mathbf{W}_2^\top) \text{diag}(\text{vec}_r(\mathbf{1}_{\{\mathbf{X} > 0\}})) (\mathbf{Y} \otimes \mathbf{I}_{d_{ff}}), & \text{если } i = 1 \\ \sigma(\mathbf{Y} \mathbf{W}_1) \otimes \mathbf{I}_{d_V}, & \text{если } i = 2 \end{cases}.$$

Тогда весь блок трансформера имеет следующую производную:

$$\frac{\partial \mathbf{Z}}{\partial \mathbf{W}_i} = \begin{cases} \frac{\partial \text{LayerNorm}(\text{FFN}(\mathbf{Y}) + \mathbf{Y})}{\partial (\text{FFN}(\mathbf{Y}) + \mathbf{Y})} (\mathbf{I}_L \otimes \mathbf{W}_2^\top) \text{diag}(\text{vec}_r(\mathbf{1}_{\{\mathbf{X} > 0\}})) (\mathbf{Y} \otimes \mathbf{I}_{d_{ff}}), & i = 1 \\ \frac{\partial \text{LayerNorm}(\text{FFN}(\mathbf{Y}) + \mathbf{Y})}{\partial (\text{FFN}(\mathbf{Y}) + \mathbf{Y})} \sigma(\mathbf{Y} \mathbf{W}_1) \otimes \mathbf{I}_{d_V}, & i = 2 \end{cases}$$

причем, согласно Теореме 19 об производной LayerNorm получаем следующее выражение в нашем случае:

$$\begin{aligned}\frac{\partial \text{LayerNorm}(\text{FFN}(\mathbf{Y}) + \mathbf{Y})}{\partial (\text{FFN}(\mathbf{Y}) + \mathbf{Y})} &= \\ &= (\mathbf{P}(\text{FFN}(\mathbf{Y}) + \mathbf{Y}) \otimes \mathbf{I}_{d_V}) \frac{\partial \mathbf{M}}{\partial (\text{FFN}(\mathbf{Y}) + \mathbf{Y})} + \\ &+ (\mathbf{I}_L \otimes \mathbf{M}^\top) \frac{1}{\sqrt{d_V}} (-\mathbf{D}^{-1} \otimes \mathbf{D}^{-\top}) \begin{pmatrix} \mathbf{e}_1 \otimes \mathbf{e}_1 & \dots & \mathbf{e}_L \otimes \mathbf{e}_L \end{pmatrix} \cdot \\ &\cdot \left(\text{diag}^{-1}(\text{vec}_r^{\circ \frac{1}{2}}(\mathbf{M}^{\circ 2} \cdot \mathbf{1}_{d_V})) \cdot (\mathbf{I}_L \otimes \mathbf{1}_{d_V}^T) \cdot \text{diag}(\text{vec}_r(\mathbf{M})) \frac{\partial \mathbf{M}}{\partial (\text{FFN}(\mathbf{Y}) + \mathbf{Y})} \right),\end{aligned}$$

где

$$\mathbf{M}(\text{FFN}(\mathbf{Y}) + \mathbf{Y}) = ((\text{FFN}(\mathbf{Y}) + \mathbf{Y}) - \frac{1}{d_V} (\text{FFN}(\mathbf{Y}) + \mathbf{Y}) \mathbf{1}_{d_V \times d_V}),$$

$$\mathbf{P}((\text{FFN}(\mathbf{Y}) + \mathbf{Y})) = \text{diag}^{-1}(\sigma(\text{FFN}(\mathbf{Y}) + \mathbf{Y})),$$

$$\frac{\partial \mathbf{M}}{\partial (\text{FFN}(\mathbf{Y}) + \mathbf{Y})} = (\mathbf{I}_L \otimes \mathbf{I}_{d_V}) - \frac{1}{d_V} (\mathbf{I}_L \otimes \mathbf{1}_{d_V \times d_V}),$$

где σ вычисляется согласно определению оператору LayerNorm.

Далее, перейдем к вычислению матриц Гессе $\frac{\partial \mathbf{Z}}{\partial \mathbf{W}_i}$ для $i \in \{K, Q, V\}$, где получаем:

$$\frac{\partial \mathbf{Z}}{\partial \mathbf{W}_i} = \frac{\partial \text{LayerNorm}(\text{FFN}(\mathbf{Y}) + \mathbf{Y})}{\partial (\text{FFN}(\mathbf{Y}) + \mathbf{Y})} \frac{\partial (\text{FFN}(\mathbf{Y}) + \mathbf{Y})}{\partial \mathbf{Y}} \frac{\partial \mathbf{Y}}{\partial \mathbf{W}_i},$$

где используя свойство 9 и результат Теоремы 21 получаем:

$$\begin{aligned} \frac{\partial (\text{FFN}(\mathbf{Y}) + \mathbf{Y})}{\partial \mathbf{Y}} &= \frac{\partial \text{FFN}(\mathbf{Y})}{\partial \mathbf{Y}} + \frac{\partial \mathbf{Y}}{\partial \mathbf{Y}} = \\ &= \frac{\partial \text{FFN}(\mathbf{Y})}{\partial \mathbf{Y}} + (\mathbf{I}_L \otimes \mathbf{I}_{d_V}) = \\ &= \frac{\partial \sigma(\mathbf{Y}\mathbf{W}_1)\mathbf{W}_2}{\partial \mathbf{Y}} + (\mathbf{I}_L \otimes \mathbf{I}_{d_V}) = \\ &= (\mathbf{I}_L \otimes \mathbf{W}_2^\top) \frac{\partial \sigma(\mathbf{Y}\mathbf{W}_1)}{\partial \mathbf{Y}\mathbf{W}_1} \frac{\partial \mathbf{Y}\mathbf{W}_1}{\partial \mathbf{Y}} + (\mathbf{I}_L \otimes \mathbf{I}_{d_V}) = \\ &= (\mathbf{I}_L \otimes \mathbf{W}_2^\top) \text{diag}(\text{vec}_r(\mathbf{1}_{\{\mathbf{X}>0\}})) (\mathbf{I}_L \otimes \mathbf{W}_1^\top) + (\mathbf{I}_L \otimes \mathbf{I}_{d_V}). \end{aligned}$$

Для вычисления матрицы $\frac{\partial \mathbf{Y}}{\partial \mathbf{W}_i}$ используем результат Леммы A.2 с работы [66]:

$$\begin{aligned} \frac{\partial \mathbf{F}}{\partial \mathbf{W}_V} &= \text{softmax} \left(\frac{\mathbf{X}\mathbf{W}_Q\mathbf{W}_K^\top\mathbf{X}^\top}{\sqrt{d_K}} \right) \mathbf{X} \otimes \mathbf{I}_{d_V} \\ \frac{\partial \mathbf{F}}{\partial \mathbf{W}_Q} &= (\mathbf{I}_L \otimes \mathbf{W}_V^\top\mathbf{X}^\top) \frac{\partial \mathbf{A}}{\partial \mathbf{M}} \left(\frac{\mathbf{X} \otimes \mathbf{X}\mathbf{W}_K}{\sqrt{d_K}} \right), \end{aligned}$$

где

$$\frac{\partial \mathbf{A}}{\partial \mathbf{M}} = \text{blockdiag} \left(\frac{\partial \mathbf{A}_i}{\partial \mathbf{M}_i^\top} \right),$$

причем данное выражение сильно упрощается используя свойства матрицы \mathbf{A} :

$$\frac{\partial \mathbf{A}_i}{\partial \mathbf{M}_i^\top} = \text{diag}(\mathbf{A}_i) - \mathbf{A}_i \mathbf{A}_i^\top,$$

где \mathbf{A}_i является i -й строкой матрицы \mathbf{A} в формате вектора. Итого в условия равномерного внимания (англ. uniform-attention) данное выражение упрощается до:

$$\frac{\partial \mathbf{A}}{\partial \mathbf{M}} = \frac{1}{n} \mathbf{I}_L \otimes \left(\mathbf{I}_L - \frac{1}{L} \mathbf{1}_{L \times L} \right)$$

Аналогично, используя Лемму 41 вычисляем производную относительно матрицы \mathbf{W}_K :

$$\frac{\partial \mathbf{F}}{\partial \mathbf{W}_K} = (\mathbf{I}_L \otimes \mathbf{W}_V^\top\mathbf{X}^\top) \frac{\partial \mathbf{A}}{\partial \mathbf{M}} \left(\frac{(\mathbf{X}\mathbf{W}_Q \otimes \mathbf{X})\mathbf{K}_{d_V d_K}}{\sqrt{d_k}} \right).$$

Получаем, что матрица $\frac{\partial \mathbf{Y}}{\partial \mathbf{W}_i}$ для $i \in \{K, Q, V\}$ вычисляется следующим образом:

$$\begin{aligned}\frac{\partial \mathbf{Y}}{\partial \mathbf{W}_i} &= \frac{\partial \text{LayerNorm}(\mathbf{F}(\mathbf{X}) + \mathbf{X})}{\partial \mathbf{W}_i} = \\ &= \frac{\partial \text{LayerNorm}(\mathbf{F}(\mathbf{X}) + \mathbf{X})}{\partial (\mathbf{F}(\mathbf{X}) + \mathbf{X})} \frac{\partial \mathbf{F}(\mathbf{X})}{\partial \mathbf{W}_i},\end{aligned}$$

где $\frac{\partial \mathbf{F}(\mathbf{X})}{\partial \mathbf{W}_i}$ вычисляется согласно лемме A.2 из работы [66], а матрица $\frac{\partial \text{LayerNorm}(\mathbf{F}(\mathbf{X}) + \mathbf{X})}{\partial (\mathbf{F}(\mathbf{X}) + \mathbf{X})}$ вычисляется согласно теореме 19. \square

В теореме 25 получен вид матрицы Якоби для полного блока трансформера. Переходим к вычислению матрицы Гессе, который формулируется в теореме 26.

Теорема 26 (Петров-Грабовой-Киселев, Матрица Гессе трансформера). *Пусть заданы матрицы параметров модели трансформера $\mathbf{X} \in \mathbb{R}^{L \times d_V}$, $\mathbf{Y} \in \mathbb{R}^{L \times d_V}$, $\mathbf{W}_1 \in \mathbb{R}^{d_V \times d_{ff}}$, $\mathbf{W}_2 \in \mathbb{R}^{d_{ff} \times d_V}$, $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{d_V \times d_K}$, $\mathbf{W}_V \in \mathbb{R}^{d_V \times d_V}$, где блок трансформатор описан в виде следующих матричнозначных функций:*

$$\mathbf{S}(\mathbf{Y}, \mathbf{W}_1, \mathbf{W}_2) = \sigma(\mathbf{Y}\mathbf{W}_1)\mathbf{W}_2 + \mathbf{Y} \in \mathbb{R}^{L \times d_V}, \quad \mathbf{Z} = \text{LayerNorm}(\mathbf{S}) \in \mathbb{R}^{L \times d_V},$$

для которых в условиях теорем 19 и 20 вычислимые матрицы Якоби и Гессе вида:

$$\mathbf{J}_Z := \frac{\partial \text{LayerNorm}(\mathbf{S})}{\partial \mathbf{S}} \in \mathbb{R}^{Ld_V \times Ld_V}, \quad \mathbf{H}_Z := \frac{\partial^2 \text{LayerNorm}(\mathbf{S})}{\partial \mathbf{S}^2} \in \mathbb{R}^{(Ld_V)^2 \times Ld_V}.$$

Также в условиях теорем 19, 20 и 21 введем следующее:

$$\begin{aligned}\mathbf{D}_\sigma &:= \text{diag}(\text{vec}_r(\mathbf{1}_{\{\mathbf{Y}\mathbf{W}_1 > 0\}})) \in \mathbb{R}^{Ld_{ff} \times Ld_{ff}}, \\ \mathbf{J}_{SY} &:= \frac{\partial \mathbf{S}}{\partial \mathbf{Y}} = (\mathbf{I}_L \otimes \mathbf{W}_2^\top) \mathbf{D}_\sigma (\mathbf{I}_L \otimes \mathbf{W}_1^\top) + (\mathbf{I}_L \otimes \mathbf{I}_{d_V}) \in \mathbb{R}^{Ld_V \times Ld_V},\end{aligned}$$

где для матрицы $\mathbf{Y} = \text{LayerNorm}(\mathbf{F}(\mathbf{X}) + \mathbf{X})$ в условиях теорем 19, 20 определено:

$$\begin{aligned}\mathbf{J}_Y &:= \frac{\partial \text{LayerNorm}(\mathbf{F}(\mathbf{X}) + \mathbf{X})}{\partial (\mathbf{F}(\mathbf{X}) + \mathbf{X})} \in \mathbb{R}^{Ld_V \times Ld_V}, \\ \mathbf{H}_Y &:= \frac{\partial^2 \text{LayerNorm}(\mathbf{F}(\mathbf{X}) + \mathbf{X})}{\partial (\mathbf{F}(\mathbf{X}) + \mathbf{X})^2} \in \mathbb{R}^{(Ld_V)^2 \times Ld_V},\end{aligned}$$

где для удобства введем следующие обозначения: $n_1 = d_V d_{ff}$, $n_2 = d_{ff} d_V$, $n_Q = n_K = d_V d_K$, $n_V = d_V^2$. Пусть матрицы Якоби вычислимы в условиях теоре-

мы 25 в следующем виде:

$$\begin{aligned}
\mathbf{G}_V &:= \frac{\partial \mathbf{F}}{\partial \mathbf{W}_V} \in \mathbb{R}^{Ld_V \times n_V}, \\
\mathbf{G}_Q &:= \frac{\partial \mathbf{F}}{\partial \mathbf{W}_Q} \in \mathbb{R}^{Ld_V \times n_Q}, \\
\mathbf{G}_K &:= \frac{\partial \mathbf{F}}{\partial \mathbf{W}_K} \in \mathbb{R}^{Ld_V \times n_K}, \\
\mathbf{B}_1 &:= \frac{\partial \mathbf{S}}{\partial \mathbf{W}_1} = (\mathbf{I}_L \otimes \mathbf{W}_2^\top) \mathbf{D}_\sigma(\mathbf{Y} \otimes \mathbf{I}_{d_{ff}}) \in \mathbb{R}^{Ld_V \times n_1}, \\
\mathbf{B}_2 &:= \frac{\partial \mathbf{S}}{\partial \mathbf{W}_2} = \sigma(\mathbf{Y} \mathbf{W}_1) \otimes \mathbf{I}_{d_V} \in \mathbb{R}^{Ld_V \times n_2}, \\
\mathbf{B}_k &:= \frac{\partial \mathbf{S}}{\partial \mathbf{W}_k} = \mathbf{J}_{SY} \mathbf{J}_Y \mathbf{G}_k \in \mathbb{R}^{Ld_V \times n_k}, \quad k \in \{K, Q, V\}.
\end{aligned}$$

Тогда матрицы Гессе трансформера \mathbf{Z} по параметрам модели $(\mathbf{W}_i, \mathbf{W}_j)$ задается в виде:

$$\mathbf{H}_{\text{tr}}^{(i,j)} := \frac{\partial^2 \mathbf{Z}}{\partial \mathbf{W}_i \partial \mathbf{W}_j} = (\mathbf{J}_Z \otimes \mathbf{I}_{n_i}) \boldsymbol{\xi}_{ij} + (\mathbf{I}_{Ld_V} \otimes \mathbf{B}_i^\top) \mathbf{H}_Z \mathbf{B}_j, \quad (3.17)$$

где размерность матрицы Гессе $\mathbf{H}_{\text{tr}}^{(i,j)} \in \mathbb{R}^{(Ld_V \cdot n_i) \times n_j}$, также введены дополнительные матрицы для удобства:

$$\boldsymbol{\xi}_{ij} := \frac{\partial}{\partial \mathbf{W}_j} \left(\frac{\partial \mathbf{S}}{\partial \mathbf{W}_i} \right) \in \mathbb{R}^{(Ld_V \cdot n_i) \times n_j}.$$

Матрицы $\boldsymbol{\xi}_{ij}$ вычисляются для всех пар (i, j) почти всюду.

Для пар FFN:

$$\begin{aligned}
\boldsymbol{\xi}_{11} &= \mathbf{0}_{(Ld_V \cdot n_1) \times n_1}, \\
\boldsymbol{\xi}_{22} &= \mathbf{0}_{(Ld_V \cdot n_2) \times n_2}, \\
\boldsymbol{\xi}_{12} &= (\mathbf{I}_L \otimes \mathbf{K}_{d_V, d_{ff}} \otimes \mathbf{I}_{d_V}) (\mathbf{I}_{Ld_{ff}} \otimes \text{vec}_r(\mathbf{I}_{d_V})) (\mathbf{D}_\sigma(\mathbf{Y} \otimes \mathbf{I}_{d_{ff}})), \\
\boldsymbol{\xi}_{21} &= (\mathbf{I}_{Ld_V} \otimes ((\mathbf{Y} \otimes \mathbf{I}_{d_{ff}})^\top \mathbf{D}_\sigma^\top)) (\mathbf{I}_L \otimes \mathbf{K}_{d_V, L} \otimes \mathbf{I}_{d_{ff}}) (\text{vec}_r(\mathbf{I}_L) \otimes \mathbf{I}_{d_V d_{ff}}) \mathbf{K}_{d_{ff}, d_V},
\end{aligned}$$

где матрицы $\boldsymbol{\xi}_{12}, \boldsymbol{\xi}_{21}$ имеют размерности $(Ld_V \cdot n_1) \times n_2$ и $(Ld_V \cdot n_2) \times n_1$ соответственно.

Для пар FFN с параметрами слоев внимания для всех $k \in \{K, Q, V\}$:

$$\begin{aligned}
\boldsymbol{\xi}_{1k} &= ((\mathbf{I}_L \otimes \mathbf{W}_2^\top) \mathbf{D}_\sigma \otimes \mathbf{I}_{n_k}) (\mathbf{I}_L \otimes \mathbf{K}_{d_{ff}, d_V} \otimes \mathbf{I}_{d_{ff}}) (\mathbf{I}_{Ld_V} \otimes \text{vec}_r(\mathbf{I}_{d_{ff}})) (\mathbf{J}_Y \mathbf{G}_k), \\
\boldsymbol{\xi}_{2k} &= (\mathbf{I}_L \otimes \mathbf{K}_{d_V, d_{ff}} \otimes \mathbf{I}_{d_V}) (\mathbf{I}_{Ld_{ff}} \otimes \text{vec}_r(\mathbf{I}_{d_V})) (\mathbf{D}_\sigma(\mathbf{I}_L \otimes \mathbf{W}_1^\top) \mathbf{J}_Y \mathbf{G}_k),
\end{aligned}$$

где размерности матрицы $\boldsymbol{\xi}_{1k} \in \mathbb{R}^{(Ld_V \cdot n_1) \times n_k}$ и матрицы $\boldsymbol{\xi}_{2k} \in \mathbb{R}^{(Ld_V \cdot n_2) \times n_k}$.

Для пар слоев внимания $k, \ell \in \{K, Q, V\}$:

$$\boldsymbol{\xi}_{k\ell} = (\mathbf{J}_{SY} \otimes \mathbf{I}_{n_k}) \left[(\mathbf{I}_{Ld_V} \otimes \mathbf{G}_k^\top) (\mathbf{H}_Y \mathbf{G}_\ell) + (\mathbf{J}_Y \otimes \mathbf{I}_{n_k}) \boldsymbol{\Phi}_{k\ell} \right],$$

где $\boldsymbol{\Phi}_{k\ell} := \frac{\partial \mathbf{G}_k}{\partial \mathbf{W}_\ell} \in \mathbb{R}^{(Ld_V \cdot n_k) \times n_\ell}$ является второй производной слоя внимания \mathbf{F} по ее параметрам, которые вычислены в рамках Леммы 22. Все матрицы имеют следующие размерности $\boldsymbol{\xi}_{k\ell} \in \mathbb{R}^{(Ld_V \cdot n_k) \times n_\ell}$.

Также матрица Гессе удовлетворяет следующим свойствам почти везде:

$$\mathbf{H}_{\text{tr}}^{(i,j)} = \mathbf{H}_{\text{tr}}^{(j,i)},$$

так как, во первых единственны нелинейности с потенциалью ненулевым вторым дифференциалом является оператор *LayerNorm*, для которого получены матрицы $\mathbf{H}_Z, \mathbf{H}_Y$ в рамках Теоремы 20 и которые являются симметричными по построению и оператор *ReLU*, для которого матрица Гессе является нулевой согласно Теоремы 21, а во вторых все другие отображения являются линейными, а следовательно согласно Леммы 35 и свойства производной произведения Кронекера их частные производные являются коммутативными почти всюду.

Доказательство. Вычислим производной матрицы Якоби с Теоремы 25 используя Лемму 35 об производной матричного умножения, также свойство производной произведения Кронекера 12, также Лемму 41 об производной транспонированной матрицы, Лемму 36 и Теорему 21. Для удобства доказательства разделим его на 4 шага.

На 1-м шаге для всех $i \in \{1, 2, K, Q, V\}$ получаем:

$$\frac{\partial \mathbf{Z}}{\partial \mathbf{W}_i} = \mathbf{J}_Z \mathbf{B}_i, \quad \mathbf{J}_Z \in \mathbb{R}^{Ld_V \times Ld_V},$$

где $\mathbf{B}_i := \frac{\partial \mathbf{S}}{\partial \mathbf{W}_i}$ задается следующим образом:

$$\begin{aligned} \mathbf{B}_1 &= (\mathbf{I}_L \otimes \mathbf{W}_2^\top) \mathbf{D}_\sigma (\mathbf{Y} \otimes \mathbf{I}_{d_{ff}}) \in \mathbb{R}^{Ld_V \times n_1}, \\ \mathbf{B}_2 &= \sigma(\mathbf{Y} \mathbf{W}_1) \otimes \mathbf{I}_{d_V} \in \mathbb{R}^{Ld_V \times n_2}, \\ \mathbf{B}_k &= \mathbf{J}_{SY} \mathbf{J}_Y \mathbf{G}_k \in \mathbb{R}^{Ld_V \times n_k}, \quad k \in \{K, Q, V\}, \end{aligned}$$

где матрица \mathbf{J}_{SY} вычисляется следующим образом:

$$\mathbf{J}_{SY} = \frac{\partial \mathbf{S}}{\partial \mathbf{Y}} = (\mathbf{I}_L \otimes \mathbf{W}_2^\top) \mathbf{D}_\sigma (\mathbf{I}_L \otimes \mathbf{W}_1^\top) + (\mathbf{I}_L \otimes \mathbf{I}_{d_V}) \in \mathbb{R}^{Ld_V \times Ld_V},$$

матрица имеет следующую размерность $\mathbf{J}_Y \in \mathbb{R}^{Ld_V \times Ld_V}$, а матрица \mathbf{G}_k описана в Теореме 25. Используя Лемму 35 и Теорему 20 получаем выражение для блока матрицы Гессе:

$$\frac{\partial^2 \mathbf{Z}}{\partial \mathbf{W}_i \partial \mathbf{W}_j} = (\mathbf{J}_Z \otimes \mathbf{I}_{n_i}) \boldsymbol{\xi}_{ij} + (\mathbf{I}_{Ld_V} \otimes \mathbf{B}_i^\top) \mathbf{H}_Z \mathbf{B}_j,$$

$$\boldsymbol{\xi}_{ij} := \frac{\partial \mathbf{B}_i}{\partial \mathbf{W}_j} \in \mathbb{R}^{(Ld_V \cdot n_i) \times n_j}.$$

На 2-м шаге вычисляем размерности и вид матриц \mathbf{B}_i . Используя результаты Теорем 25 и 21 получаем следующие выражения:

$$\mathbf{B}_1 = (\mathbf{I}_L \otimes \mathbf{W}_2^\top) \mathbf{D}_\sigma (\mathbf{Y} \otimes \mathbf{I}_{d_{ff}}) \in \mathbb{R}^{Ld_V \times n_1},$$

$$\mathbf{B}_2 = \sigma(\mathbf{Y} \mathbf{W}_1) \otimes \mathbf{I}_{d_V} \in \mathbb{R}^{Ld_V \times n_2},$$

где матрица $\mathbf{D}_\sigma \in \mathbb{R}^{Ld_{ff} \times Ld_{ff}}$, матрица $(\mathbf{Y} \otimes \mathbf{I}_{d_{ff}}) \in \mathbb{R}^{Ld_{ff} \times d_V d_{ff}}$. Тогда для всех матриц \mathbf{B}_k , где $k \in \{K, Q, V\}$ получаем:

$$\mathbf{B}_k = \mathbf{J}_{SY} \mathbf{J}_Y \mathbf{G}_k \in \mathbb{R}^{Ld_V \times n_k}.$$

На 3-м шаге вычисляем матрицы $\boldsymbol{\xi}_{ij}$ для всех пар (i, j) .

Начнем вычисления с пар FFN. Матрица \mathbf{B}_1 не зависит от матрицы \mathbf{W}_1 , следовательно, $\boldsymbol{\xi}_{11} = \mathbf{0}$. Аналогично, матрица \mathbf{B}_2 не зависит от матрицы \mathbf{W}_2 , следовательно, $\boldsymbol{\xi}_{22} = \mathbf{0}$. Вычислим $\frac{\partial \mathbf{B}_2}{\partial \mathbf{W}_1}$ используя свойство 12 производной произведения Кронекера для $\frac{\partial(\mathbf{X} \otimes \mathbf{Y})}{\partial \mathbf{X}}$, где $\mathbf{X} = \sigma(\mathbf{Y} \mathbf{W}_1)$ и $\mathbf{Y} = \mathbf{I}_{d_V}$:

$$\frac{\partial \mathbf{B}_2}{\partial \mathbf{W}_1} = (\mathbf{I}_L \otimes \mathbf{K}_{d_V, d_{ff}} \otimes \mathbf{I}_{d_V}) (\mathbf{I}_{Ld_{ff}} \otimes \text{vec}_r(\mathbf{I}_{d_V})) \frac{\partial \text{vec}_r(\sigma(\mathbf{Y} \mathbf{W}_1))}{\partial \mathbf{W}_1},$$

далее используя то, что $\frac{\partial \text{vec}_r(\sigma(\mathbf{Y} \mathbf{W}_1))}{\partial \mathbf{W}_1} = \mathbf{D}_\sigma (\mathbf{Y} \otimes \mathbf{I}_{d_{ff}})$ получаем оценку на $\boldsymbol{\xi}_{12}$:

$$\boldsymbol{\xi}_{12} = (\mathbf{I}_L \otimes \mathbf{K}_{d_V, d_{ff}} \otimes \mathbf{I}_{d_V}) (\mathbf{I}_{Ld_{ff}} \otimes \text{vec}_r(\mathbf{I}_{d_V})) (\mathbf{D}_\sigma (\mathbf{Y} \otimes \mathbf{I}_{d_{ff}})).$$

Вычислим $\frac{\partial \mathbf{B}_1}{\partial \mathbf{W}_2}$ используя Лемму 35, где в качестве левого множителя выступает $(\mathbf{I}_L \otimes \mathbf{W}_2^\top)$:

$$\frac{\partial \mathbf{B}_1}{\partial \mathbf{W}_2} = (\mathbf{I}_{Ld_V} \otimes ((\mathbf{Y} \otimes \mathbf{I}_{d_{ff}})^\top \mathbf{D}_\sigma^\top)) \frac{\partial (\mathbf{I}_L \otimes \mathbf{W}_2^\top)}{\partial \mathbf{W}_2},$$

далее используя свойство 12 об производной произведения Кронекера и Лемму 41 об производной транспонированной матрицы получим:

$$\frac{\partial (\mathbf{I}_L \otimes \mathbf{W}_2^\top)}{\partial \mathbf{W}_2} = (\mathbf{I}_L \otimes \mathbf{K}_{d_V, L} \otimes \mathbf{I}_{d_{ff}}) (\text{vec}_r(\mathbf{I}_L) \otimes \mathbf{I}_{d_V d_{ff}}) \mathbf{K}_{d_{ff}, d_V}.$$

Далее собирая все полученные матрицы, получаем оценку на ξ_{21} :

$$\xi_{21} = (\mathbf{I}_{Ld_V} \otimes ((\mathbf{Y} \otimes \mathbf{I}_{d_{ff}})^\top \mathbf{D}_\sigma^\top)) (\mathbf{I}_L \otimes \mathbf{K}_{d_V, L} \otimes \mathbf{I}_{d_{ff}}) (\text{vec}_r(\mathbf{I}_L) \otimes \mathbf{I}_{d_V d_{ff}}) \mathbf{K}_{d_{ff}, d_V}.$$

Переходим к оценке пар FFN с параметрами слоев внимания для всех $k \in \{K, Q, V\}$. Для матрицы $\mathbf{B}_1 = (\mathbf{I}_L \otimes \mathbf{W}_2^\top) \mathbf{D}_\sigma (\mathbf{Y} \otimes \mathbf{I}_{d_{ff}})$ установим, что почти всюду матрица $\frac{\partial \mathbf{D}_\sigma}{\partial \mathbf{Y}} = \mathbf{0}$ равна нулю согласно теореме 21. Следовательно, только последний множитель зависит от матрицы \mathbf{W}_k . Используя лемму 35, где первый множитель является константой, а также цепное правило относительно переменной \mathbf{Y} , получаем:

$$\frac{\partial (\mathbf{Y} \otimes \mathbf{I}_{d_{ff}})}{\partial \mathbf{W}_k} = \left(\frac{\partial (\mathbf{Y} \otimes \mathbf{I}_{d_{ff}})}{\partial \mathbf{Y}} \right) \frac{\partial \mathbf{Y}}{\partial \mathbf{W}_k},$$

причем согласно свойства 12 об производной произведения Кронекера с матрицей $\mathbf{X} = \mathbf{Y}$ и матрицей $\mathbf{Y} = \mathbf{I}_{d_{ff}}$ получаем:

$$\frac{\partial (\mathbf{Y} \otimes \mathbf{I}_{d_{ff}})}{\partial \mathbf{Y}} = (\mathbf{I}_L \otimes \mathbf{K}_{d_{ff}, d_V} \otimes \mathbf{I}_{d_{ff}}) (\mathbf{I}_{Ld_V} \otimes \text{vec}_r(\mathbf{I}_{d_{ff}})),$$

а в свою очередь согласно Теоремы 25 матрица $\frac{\partial \mathbf{Y}}{\partial \mathbf{W}_k} = \mathbf{J}_Y \mathbf{G}_k$. Тогда получаем оценку на матрицу ξ_{1k} следующего вида:

$$\xi_{1k} = ((\mathbf{I}_L \otimes \mathbf{W}_2^\top) \mathbf{D}_\sigma \otimes \mathbf{I}_{n_k}) (\mathbf{I}_L \otimes \mathbf{K}_{d_{ff}, d_V} \otimes \mathbf{I}_{d_{ff}}) (\mathbf{I}_{Ld_V} \otimes \text{vec}_r(\mathbf{I}_{d_{ff}})) (\mathbf{J}_Y \mathbf{G}_k).$$

Перейдем к матрице $\mathbf{B}_2 = \sigma(\mathbf{Y} \mathbf{W}_1) \otimes \mathbf{I}_{d_V}$, заметим, что только первый множитель произведения Кронекера зависит от матриц \mathbf{W}_k , а следовательно используя свойство 12 производной произведения Кронекера и цепное правило получим следующее выражение:

$$\xi_{2k} = (\mathbf{I}_L \otimes \mathbf{K}_{d_V, d_{ff}} \otimes \mathbf{I}_{d_V}) (\mathbf{I}_{Ld_{ff}} \otimes \text{vec}_r(\mathbf{I}_{d_V})) (\mathbf{D}_\sigma (\mathbf{I}_L \otimes \mathbf{W}_1^\top) \mathbf{J}_Y \mathbf{G}_k),$$

где использовано свойство 9 для преобразования $\frac{\partial (\mathbf{Y} \mathbf{W}_1)}{\partial \mathbf{Y}} = \mathbf{I}_L \otimes \mathbf{W}_1^\top$, а также Теорема 21 для выражения $\frac{\partial \sigma(\cdot)}{\partial (\cdot)} = \mathbf{D}_\sigma$, а также согласно Теореме 25 матрица $\frac{\partial \mathbf{Y}}{\partial \mathbf{W}_k} = \mathbf{J}_Y \mathbf{G}_k$.

Перейдем к оценке пар слоев внимания (k, ℓ) with $k, \ell \in \{K, Q, V\}$. Рассмотрим матрицу $\mathbf{B}_k = \mathbf{J}_{SY} \mathbf{J}_Y \mathbf{G}_k$, заметим, что почти всюду $\frac{\partial \mathbf{J}_{SY}}{\partial \mathbf{Y}} = \mathbf{0}$, так как матричнозначная функция \mathbf{D}_σ является кусочно-постоянной, согласно Теоремы 21, а следовательно используя Лемму 35 с матрицей $\mathbf{A}(\cdot) = \mathbf{J}_Y$, и с матрицей $\mathbf{B}(\cdot) = \mathbf{G}_k$ получаем:

$$\frac{\partial \mathbf{B}_k}{\partial \mathbf{W}_\ell} = (\mathbf{J}_{SY} \otimes \mathbf{I}_{n_k}) \frac{\partial (\mathbf{J}_Y \mathbf{G}_k)}{\partial \mathbf{W}_\ell},$$

далее вычислим матрицу $\frac{\partial(\mathbf{J}_Y \mathbf{G}_k)}{\partial \mathbf{W}_\ell}$ используя свойство производной матричного произведения:

$$\frac{\partial(\mathbf{J}_Y \mathbf{G}_k)}{\partial \mathbf{W}_\ell} = (\mathbf{J}_Y \otimes \mathbf{I}_{n_k}) \Phi_{k\ell} + (\mathbf{I}_{Ld_V} \otimes \mathbf{G}_k^\top) \frac{\partial \mathbf{J}_Y}{\partial \mathbf{W}_\ell}.$$

В условиях теоремы 20 легко получить следующее выражение $\frac{\partial \mathbf{J}_Y}{\partial \mathbf{W}_\ell} = \mathbf{H}_Y \mathbf{G}_\ell$. Следовательно, получаем оценки:

$$\boldsymbol{\xi}_{k\ell} = (\mathbf{J}_{SY} \otimes \mathbf{I}_{n_k}) [(\mathbf{I}_{Ld_V} \otimes \mathbf{G}_k^\top) (\mathbf{H}_Y \mathbf{G}_\ell) + (\mathbf{J}_Y \otimes \mathbf{I}_{n_k}) \Phi_{k\ell}].$$

На 4-м шаге проверим симметричность полученных выражений. Во-первых, единственные нелинейности с потенциально ненулевым вторым дифференциалом являются оператор LayerNorm, для которого получены матрицы $\mathbf{H}_Z, \mathbf{H}_Y$ в рамках теоремы 20 и которые являются симметричными по построению, и оператор ReLU, для которого матрица Гессе является нулевой согласно теореме 21. Во-вторых, все другие отображения являются линейными, следовательно, согласно лемме 35 и свойству производной произведения Кронекера их частные производные являются коммутативными почти всюду. \square

3.3.5. Спектральные оценки матрицы Гессе для трансформера

Теорема 27 (Петров-Грабовой-Киселев, Оценка нормы матрицы Гессе трансформера). Пусть матрица Гессе $\mathbf{H}_{\text{tr}}^{(i,j)}$ описывает матрицу Гессе между (i, j) -м блоком трансформер модели (3.17), где $i, j \in \{1, 2, K, Q, V\}$, $n_i = \dim(\mathbf{W}_i)$. Тогда для каждой пары (i, j) получаем оценку нормы:

$$\|\mathbf{H}_{\text{tr}}^{(i,j)}\|_2 \leq \|\mathbf{J}_Z\|_2 \|\boldsymbol{\xi}_{ij}\|_2 + \|\mathbf{B}_i\|_2 \|\mathbf{H}_Z\|_2 \|\mathbf{B}_j\|_2, \quad (3.18)$$

$$\text{где } \boldsymbol{\xi}_{ij} = \frac{\partial}{\partial \mathbf{W}_j} \left(\frac{\partial \mathbf{S}}{\partial \mathbf{W}_i} \right), \mathbf{B}_i = \frac{\partial \mathbf{S}}{\partial \mathbf{W}_i}.$$

Пусть матрица \mathbf{H}_{tr} является полной матрицей Гессе размера $m_b \times n_b$ состоящей из блоков матрицы $\mathbf{H}_{\text{tr}}^{(i,j)}$, где $m_b = n_b = 5, i \in \{1, 2, K, Q, V\}, j \in \{1, 2, K, Q, V\}$). Тогда

$$\|\mathbf{H}_{\text{tr}}\|_2 \leq \sqrt{m_b n_b} \max_{i,j} \left(\frac{2}{Ld_V} \left\| \frac{\partial \mathbf{Z}}{\partial \mathbf{W}_i} \right\|_2 \left\| \frac{\partial \mathbf{Z}}{\partial \mathbf{W}_j} \right\|_2 + \|\mathbf{R}_m^{tr}\|_2 \|\mathbf{H}_{\text{tr}}^{(i,j)}\|_2 \right).$$

Доказательство. Рассмотрим блоки матрицы Гессе (3.17):

$$\mathbf{H}_{\text{tr}}^{(i,j)} = (\mathbf{J}_Z \otimes \mathbf{I}_{n_i}) \boldsymbol{\xi}_{ij} + (\mathbf{I}_{Ld_V} \otimes \mathbf{B}_i^\top) \mathbf{H}_Z \mathbf{B}_j.$$

Используя свойства норм матриц 6, 1 и 2 получаем оценку

$$\begin{aligned}\|\mathbf{H}_{\text{tr}}^{(i,j)}\|_2 &\leq \|\mathbf{J}_Z \otimes \mathbf{I}_{n_i}\|_2 \|\boldsymbol{\xi}_{ij}\|_2 + \|\mathbf{I}_{Ld_V} \otimes \mathbf{B}_i^\top\|_2 \|\mathbf{H}_Z\|_2 \|\mathbf{B}_j\|_2 = \\ &= \|\mathbf{J}_Z\|_2 \|\boldsymbol{\xi}_{ij}\|_2 + \|\mathbf{B}_i\|_2 \|\mathbf{H}_Z\|_2 \|\mathbf{B}_j\|_2,\end{aligned}$$

описанной в условиях теоремы (3.18).

Оценим все слагаемые операторных норм в полученной оценке, а именно норму $\|\mathbf{B}_i\|_2$, и норму $\|\boldsymbol{\xi}_{ij}\|_2$, которые используются в формуле (3.18). Используя свойства матричных норм 1, 2, 6, 4, 3, а также определение 29 коммутативных матрицы, получаем, что $\|\mathbf{K}_{m,n}\|_2 = 1$, а также согласно свойству 4 получаем нормы $\|\text{vec}_r(\mathbf{I}_d)\|_2 = \|\mathbf{I}_d\|_F = \sqrt{d}$ и $\|\mathbf{I}_p\|_2 = 1$. В доказательстве Теоремы 18 было доказано, что :

$$\begin{aligned}\left\|\frac{\partial \mathbf{A}}{\partial \mathbf{T}}\right\|_2 &\leq \frac{1}{L}, \\ \|\mathbf{Z}_1\|_2 &= \|(\mathbf{I}_L \otimes \mathbf{X}^\top)(\partial \mathbf{A}/\partial \mathbf{T})(\mathbf{X} \otimes \mathbf{X})\|_2 \leq \|\mathbf{X}\|_2 \frac{1}{L} \|\mathbf{X}\|_2^2 = \frac{1}{L} \|\mathbf{X}\|_2^3, \\ \left\|\frac{\partial^2 \mathbf{A}}{\partial \mathbf{T}^2}\right\|_2 &\leq 6, \\ \|\mathbf{Z}_2\|_2 &\leq \|\mathbf{X}\|_2^5 \left\|\frac{\partial^2 \mathbf{A}}{\partial \mathbf{T}^2}\right\|_2 \leq 6 \|\mathbf{X}\|_2^5, \\ \|\mathbf{A}\|_2 &\leq \sqrt{LL} \|\mathbf{A}\|_{\max} = L,\end{aligned}$$

а следовательно согласно свойству 1 получаем оценку $\|\mathbf{A}\mathbf{X}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{X}\|_2 \leq L \|\mathbf{X}\|_2$. Оценим матрицы $\Phi_{k\ell}$ полученные в рамках Леммы 22. Используя свойства матричных норм 1, 2, а также верхние оценки на матрицы $\|\mathbf{Z}_1\|_2$, $\|\mathbf{Z}_2\|_2$ получаем:

$$\begin{aligned}\|\Phi_{VV}\|_2 &= 0, \\ \|\Phi_{QQ}\|_2 &\leq \frac{2}{Ld_V d_K} \|\mathbf{W}_V\|_2 \|\mathbf{W}_K\|_2 \|\mathbf{Z}_2\|_2 \|\mathbf{W}_K\|_2 \leq \\ &\leq \frac{12}{Ld_V d_K} \|\mathbf{W}_V\|_2 \|\mathbf{W}_K\|_2^2 \|\mathbf{X}\|_2^5, \\ \|\Phi_{VQ}\|_2 &\leq \frac{2}{Ld_V \sqrt{d_K}} \|\mathbf{I}_L \otimes \mathbf{S}\|_2 \|\mathbf{Z}_1\|_2 \|\mathbf{I}_{d_V} \otimes \mathbf{W}_K\|_2 \leq \\ &\leq \frac{2}{L^2 \sqrt{d_V d_K}} \|\mathbf{W}_K\|_2 \|\mathbf{X}\|_2^3, \\ \|\Phi_{QK}\|_2 &\leq \frac{2}{Ld_V d_K} \|\mathbf{W}_V\|_2 \|\mathbf{W}_K\|_2 \|\mathbf{Z}_2\|_2 \|\mathbf{W}_Q\|_2 + \frac{2}{Ld_V \sqrt{d_K}} \|\mathbf{W}_V\|_2 \|\mathbf{Z}_1\|_2 \|\mathbf{S}\|_2 \leq \\ &\leq \frac{12}{Ld_V d_K} \|\mathbf{W}_V\|_2 \|\mathbf{W}_K\|_2 \|\mathbf{W}_Q\|_2 \|\mathbf{X}\|_2^5 + \frac{2}{L^2 \sqrt{d_V d_K}} \|\mathbf{W}_V\|_2 \|\mathbf{X}\|_2^3.\end{aligned}$$

Для оценки матричных норм $\|\mathbf{B}_i\|_2$ рассмотрим чему они равны при разных i из определения в Теоремах 26, 21. Для матрицы $\mathbf{B}_1 = (\mathbf{I}_L \otimes \mathbf{W}_2^\top) \mathbf{D}_\sigma (\mathbf{Y} \otimes \mathbf{I}_{d_{ff}})$, тогда используя свойства матричных норм 2, 1, 3, а также оценку $\|\mathbf{D}_\sigma\|_2 \leq 1$ получаем:

$$\|\mathbf{B}_1\|_2 \leq \|\mathbf{I}_L \otimes \mathbf{W}_2^\top\|_2 \|\mathbf{D}_\sigma\|_2 \|\mathbf{Y} \otimes \mathbf{I}_{d_{ff}}\|_2 = \|\mathbf{W}_2\|_2 \|\mathbf{Y}\|_2. \quad (3.19)$$

Для матрицы $\mathbf{B}_2 = \sigma(\mathbf{Y}\mathbf{W}_1) \otimes \mathbf{I}_{d_V}$ воспользовавшись свойством 2 получаем:

$$\|\mathbf{B}_2\|_2 = \|\sigma(\mathbf{Y}\mathbf{W}_1)\|_2.$$

Для матриц $\mathbf{B}_k = \mathbf{J}_{SY} \mathbf{J}_Y \mathbf{G}_k$, где $k \in \{K, Q, V\}$, используя свойство 1:

$$\|\mathbf{B}_k\|_2 \leq \|\mathbf{J}_{SY}\|_2 \|\mathbf{J}_Y\|_2 \|\mathbf{G}_k\|_2. \quad (3.20)$$

Для матрицы $\mathbf{J}_{SY} = (\mathbf{I}_L \otimes \mathbf{W}_2^\top) \mathbf{D}_\sigma (\mathbf{I}_L \otimes \mathbf{W}_1^\top) + (\mathbf{I}_L \otimes \mathbf{I}_{d_V})$ используя свойства 6, 1, 2, 3, а также оценку $\|\mathbf{D}_\sigma\|_2 \leq 1$ получаем оценку матричной нормы:

$$\begin{aligned} \|\mathbf{J}_{SY}\|_2 &\leq \|\mathbf{I}_L \otimes \mathbf{W}_2^\top\|_2 \|\mathbf{D}_\sigma\|_2 \|\mathbf{I}_L \otimes \mathbf{W}_1^\top\|_2 + \|\mathbf{I}_L \otimes \mathbf{I}_{d_V}\|_2 = \\ &= \|\mathbf{W}_2\|_2 \|\mathbf{W}_1\|_2 + 1. \end{aligned} \quad (3.21)$$

Для матриц $\|\mathbf{G}_V\|_2, \|\mathbf{G}_Q\|_2, \|\mathbf{G}_K\|_2$, используя свойства 1, 2 получаем оценки на нормы:

$$\begin{aligned} \|\mathbf{G}_V\|_2 &\leq L \|\mathbf{X}\|_2, \\ \|\mathbf{G}_Q\|_2 &\leq \frac{1}{L\sqrt{d_K}} \|\mathbf{W}_V\|_2 \|\mathbf{W}_K\|_2 \|\mathbf{X}\|_2^3, \\ \|\mathbf{G}_K\|_2 &\leq \frac{1}{L\sqrt{d_K}} \|\mathbf{W}_V\|_2 \|\mathbf{W}_Q\|_2 \|\mathbf{X}\|_2^3. \end{aligned} \quad (3.22)$$

Для оценки матричных норм $\|\boldsymbol{\xi}_{ij}\|_2$, рассмотрим чему они равны из определения в Теореме 26. В случае пар FFN для матриц $\|\boldsymbol{\xi}_{11}\|_2, \|\boldsymbol{\xi}_{12}\|_2, \|\boldsymbol{\xi}_{21}\|_2, \|\boldsymbol{\xi}_{22}\|_2$ используя свойства 2, 1, 4 матричных норм, а также свойство коммутативных матриц $\|\mathbf{K}_{m,n}\|_2 = 1$ получаем оценки:

$$\begin{aligned} \|\boldsymbol{\xi}_{11}\|_2 &= 0, \\ \|\boldsymbol{\xi}_{22}\|_2 &= 0, \\ \|\boldsymbol{\xi}_{12}\|_2 &\leq \|\mathbf{I}_L \otimes \mathbf{K}_{d_V, d_{ff}} \otimes \mathbf{I}_{d_V}\|_2 \|\mathbf{I}_{d_{ff}} \otimes \text{vec}_r(\mathbf{I}_{d_V})\|_2 \|\mathbf{D}_\sigma\|_2 \|\mathbf{Y} \otimes \mathbf{I}_{d_{ff}}\|_2 \\ &= 1 \cdot \|\text{vec}_r(\mathbf{I}_{d_V})\|_2 \cdot 1 \cdot \|\mathbf{Y}\|_2 = \sqrt{d_V} \|\mathbf{Y}\|_2, \\ \|\boldsymbol{\xi}_{21}\|_2 &\leq \|\mathbf{I}_L \otimes \mathbf{W}_2^\top\|_2 \|\mathbf{D}_\sigma\|_2 \|\mathbf{I}_L \otimes \mathbf{K}_{d_{ff}, d_V} \otimes \mathbf{I}_{d_{ff}}\|_2 \|\mathbf{I}_{d_V} \otimes \text{vec}_r(\mathbf{I}_{d_{ff}})\|_2 \\ &= \|\mathbf{W}_2\|_2 \cdot 1 \cdot 1 \cdot \|\text{vec}_r(\mathbf{I}_{d_{ff}})\|_2 = \sqrt{d_{ff}} \|\mathbf{W}_2\|_2. \end{aligned} \quad (3.23)$$

В случае пар FFN с параметрами слоев внимания для всех $k \in \{K, Q, V\}$ получаем оценки:

$$\begin{aligned}
\|\boldsymbol{\xi}_{1k}\|_2 &\leq \|(\mathbf{I}_L \otimes \mathbf{W}_2^\top) \mathbf{D}_\sigma \otimes \mathbf{I}_{n_k}\|_2 \|\mathbf{I}_L \otimes \mathbf{K}_{d_{ff}, d_V} \otimes \mathbf{I}_{d_{ff}}\|_2 \cdot \\
&\quad \cdot \|\mathbf{I}_{Ld_V} \otimes \text{vec}_r(\mathbf{I}_{d_{ff}})\|_2 \|\mathbf{J}_Y\|_2 \|\mathbf{G}_k\|_2 \\
&\leq \|\mathbf{W}_2\|_2 \cdot 1 \cdot 1 \cdot \sqrt{d_{ff}} \cdot \|\mathbf{J}_Y\|_2 \|\mathbf{G}_k\|_2 = \\
&= \sqrt{d_{ff}} \|\mathbf{W}_2\|_2 \|\mathbf{J}_Y\|_2 \|\mathbf{G}_k\|_2, \\
\|\boldsymbol{\xi}_{2k}\|_2 &\leq \|\mathbf{I}_L \otimes \mathbf{K}_{d_V, d_{ff}} \otimes \mathbf{I}_{d_V}\|_2 \|\mathbf{I}_{Ld_{ff}} \otimes \text{vec}_r(\mathbf{I}_{d_V})\|_2 \|\mathbf{D}_\sigma\|_2 \|\mathbf{I}_L \otimes \mathbf{W}_1^\top\|_2 \cdot \\
&\quad \cdot \|\mathbf{J}_Y\|_2 \|\mathbf{G}_k\|_2 \\
&\leq 1 \cdot \sqrt{d_V} \cdot 1 \cdot \|\mathbf{W}_1\|_2 \cdot \|\mathbf{J}_Y\|_2 \cdot \|\mathbf{G}_k\|_2 = \\
&= \sqrt{d_V} \|\mathbf{W}_1\|_2 \|\mathbf{J}_Y\|_2 \|\mathbf{G}_k\|_2.
\end{aligned} \tag{3.24}$$

Для пар слоев внимания $k, \ell \in \{K, Q, V\}$

$$\boldsymbol{\xi}_{k\ell} = (\mathbf{J}_{SY} \otimes \mathbf{I}_{n_k}) \left[(\mathbf{I}_{Ld_V} \otimes \mathbf{G}_k^\top) (\mathbf{H}_Y \mathbf{G}_\ell) + (\mathbf{J}_Y \otimes \mathbf{I}_{n_k}) \boldsymbol{\Phi}_{k\ell} \right],$$

используя свойства 1, 2 получаем следующие оценки:

$$\begin{aligned}
\|\boldsymbol{\xi}_{k\ell}\|_2 &\leq \|\mathbf{J}_{SY}\|_2 \left(\|\mathbf{I}_{Ld_V} \otimes \mathbf{G}_k^\top\|_2 \|\mathbf{H}_Y\|_2 \|\mathbf{G}_\ell\|_2 + \|\mathbf{J}_Y\|_2 \|\boldsymbol{\Phi}_{k\ell}\|_2 \right) = \tag{3.25} \\
&= \|\mathbf{J}_{SY}\|_2 \left(\|\mathbf{G}_k\|_2 \|\mathbf{H}_Y\|_2 \|\mathbf{G}_\ell\|_2 + \|\mathbf{J}_Y\|_2 \|\boldsymbol{\Phi}_{k\ell}\|_2 \right).
\end{aligned}$$

Итого собирая все части выражения (3.18), используя для каждой пары (i, j) оценки норм матриц $\|\boldsymbol{\xi}_{ij}\|_2$ с выражений (3.23),(3.24),(3.25), а также оценки норм матриц $\|\mathbf{B}_i\|_2$ с выражений (3.19),(3.20),(3.21),(3.22) и подставляя в выражение (3.18) получаем следующие оценки норм на все блоки матрицы Гесее:

$$\begin{aligned}
\|\mathbf{H}_{\text{tr}}^{(1,1)}\|_2 &\leq \|\mathbf{J}_Z\|_2 \cdot 0 + \|\mathbf{B}_1\|_2^2 \|\mathbf{H}_Z\|_2 \leq \\
&\leq \|\mathbf{H}_Z\|_2 (\|\mathbf{W}_2\|_2 \|\mathbf{Y}\|_2)^2, \\
\|\mathbf{H}_{\text{tr}}^{(1,2)}\|_2 &\leq \|\mathbf{J}_Z\|_2 \sqrt{d_V} \|\mathbf{Y}\|_2 + \|\mathbf{H}_Z\|_2 (\|\mathbf{W}_2\|_2 \|\mathbf{Y}\|_2) \|\sigma(\mathbf{Y}\mathbf{W}_1)\|_2, \\
\|\mathbf{H}_{\text{tr}}^{(1,k)}\|_2 &\leq \|\mathbf{J}_Z\|_2 \sqrt{d_{ff}} \|\mathbf{W}_2\|_2 \|\mathbf{J}_Y\|_2 \|\mathbf{G}_k\|_2 + \\
&\quad + \|\mathbf{H}_Z\|_2 (\|\mathbf{W}_2\|_2 \|\mathbf{Y}\|_2) (\|\mathbf{J}_{SY}\|_2 \|\mathbf{J}_Y\|_2 \|\mathbf{G}_k\|_2), \\
\|\mathbf{H}_{\text{tr}}^{(k,\ell)}\|_2 &\leq \|\mathbf{J}_Z\|_2 \|\mathbf{J}_{SY}\|_2 \left(\|\mathbf{G}_k\|_2 \|\mathbf{H}_Y\|_2 \|\mathbf{G}_\ell\|_2 + \|\mathbf{J}_Y\|_2 \|\boldsymbol{\Phi}_{k\ell}\|_2 \right) + \\
&\quad + \|\mathbf{H}_Z\|_2 (\|\mathbf{J}_{SY}\|_2 \|\mathbf{J}_Y\|_2 \|\mathbf{G}_k\|_2) (\|\mathbf{J}_{SY}\|_2 \|\mathbf{J}_Y\|_2 \|\mathbf{G}_\ell\|_2),
\end{aligned}$$

В оценке матричных норм $\|\mathbf{Y}\|_2$ и $\|\mathbf{S}\|_2$ были использованы результаты леммы 23. Нормы матриц $\mathbf{H}_Z, \mathbf{H}_Y$ оцениваются в рамках леммы 24. \square

3.4. Результаты вычислительных экспериментов

Настоящий раздел посвящен эмпирической валидации теоретических оценок норм матриц Гессе для различных классов моделей глубокого обучения. Основной целью экспериментов является проверка соответствия теоретических предсказаний, полученных в предыдущих разделах, эмпирическим наблюдениям, а также качественный и количественный анализ структуры матриц Гессе для трансформерных архитектур.

В рамках вычислительного эксперимента анализируется вид матриц Гессе для моделей глубокого обучения на базе архитектуры трансформера. Выбор трансформерной архитектуры обусловлен ее практической значимостью и сложностью структуры, требующей детального теоретического анализа, представленного в разделе 3.3.. Теоретические результаты для трансформеров получены в разделе 3.3., где установлены оценки норм матриц Гессе для различных компонентов трансформера.

Эксперименты проводятся на архитектуре Vision Transformer (ViT) с параметрами, указанными в таблице 3.1. В экспериментах используется один блок трансформера, который обучается на наборе данных MNIST [53]. Выбор MNIST обусловлен простотой задачи, что позволяет сосредоточиться на анализе структуры гессиана без усложнения, связанного с особенностями сложных данных.

Вычисление матриц Гессе осуществляется с использованием пакета `curvlinops`, который обеспечивает эффективное вычисление линейного оператора гессиана без явного построения полной матрицы вторых производных. Данный подход позволяет анализировать гессианы для моделей с большим количеством параметров, что было бы непрактично при прямом вычислении.

Экспериментальная процедура включает следующие этапы:

1. Инициализация модели с заданными гиперпараметрами.
2. Вычисление матрицы Гессе на одном батче из обучающего загрузчика данных (128 примеров).
3. Визуализация структуры гессиана для инициализированной модели.
4. Обучение модели в течение нескольких эпох до достижения точности на валидационном наборе данных более 50%.
5. Вычисление и визуализация матрицы Гессе для обученной модели.
6. Детальный анализ отдельных блоков гессиана, соответствующих различным компонентам трансформера.

7. Вычисление спектральных норм блоков параметров и соответствующих блоков матрицы Гессе.

Визуализация гессиана в логарифмическом масштабе (рис. 3.1) демонстрирует неоднородность в величинах элементов матрицы для инициализированной модели. Наблюдается блочная структура гессиана, соответствующая различным компонентам трансформера: механизму самовнимания, блоку полно связной сети (FFN) и слоям нормализации (LayerNorm).

набор данных	размер патча	скрытая размерность	размер FFN	количество блоков
MNIST	4	16	64	1
CIFAR-100	4	128	512	8

Таблица 3.1: Гиперпараметры архитектур Vision Transformer (ViT) для экспериментов по анализу матриц Гессе. Параметры включают размер патча, скрытую размерность, размер блока FFN и количество трансформерных блоков для наборов данных MNIST и CIFAR-100.



Рис. 3.1: Визуализация элементов матрицы Гессе для инициализированной модели трансформера (один блок, набор данных MNIST, батч из 128 примеров) в логарифмическом масштабе. Наблюдаются неоднородность величин элементов, при этом блоки, соответствующие параметрам Values, демонстрируют наибольшие значения.

После обучения модели в течение нескольких эпох до достижения точности на валидационном наборе данных более 50% элементы матрицы Гессе снова визуализируются (рис. 3.2). Наблюдаются существенное изменение структуры гессиана: каждый из блоков приобретает большую величину по сравнению с инициализированной моделью, что отражает накопление кривизны функции потерь



Рис. 3.2: Визуализация элементов матрицы Гессе для обученной модели трансформера (один блок, набор данных MNIST, точность на валидации $>50\%$) в логарифмическом масштабе. После обучения все блоки гессиана приобретают большие значения, при этом блок Values-Values демонстрирует максимальные величины, что подтверждает гетерогенность вклада различных параметров в кривизну функции потерь.

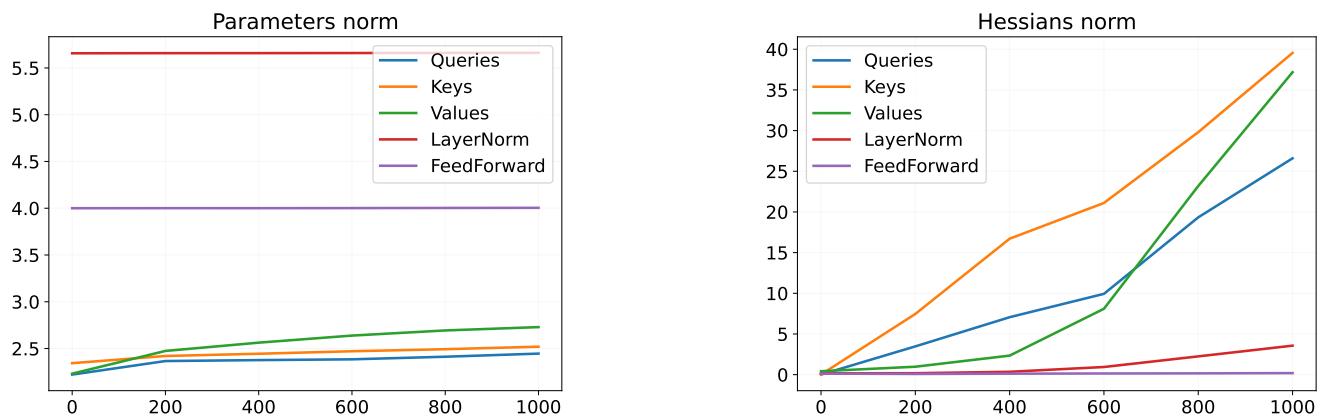


Рис. 3.3: Спектральные нормы блоков параметров трансформера (слева) и спектральные нормы соответствующих блоков матрицы Гессе (справа), вычисленные на одном батче из 128 примеров набора данных MNIST. Наибольшие значения норм соответствуют блокам Keys и Values, что согласуется с теоретическими оценками теоремы 27.

в процессе обучения. Блок Values-Values демонстрирует наибольшие значения среди всех блоков гессиана, что согласуется с теоретическими предсказаниями теоремы 27.

Для детального анализа структуры гессиана рассматриваются отдельные блоки, соответствующие различным компонентам трансформера

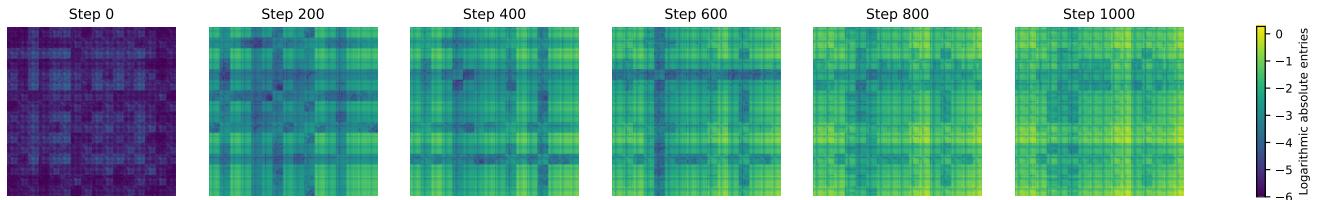


Рис. 3.4: Визуализация элементов блока матрицы Гессе, соответствующего параметрам Queries трансформера (один блок, набор данных MNIST). Блок демонстрирует структурированное распределение элементов, отражающее взаимосвязи между параметрами механизма самовнимания.

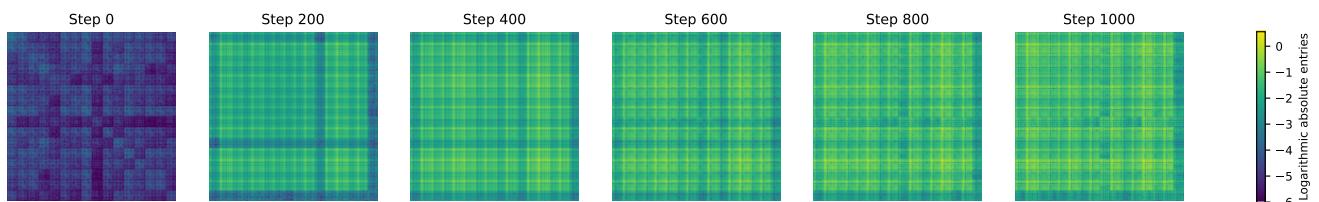


Рис. 3.5: Визуализация элементов блока матрицы Гессе, соответствующего параметрам Keys трансформера (один блок, набор данных MNIST). Блок демонстрирует значительные значения элементов, что подтверждает важную роль параметров Keys в формировании кривизны функции потерь.

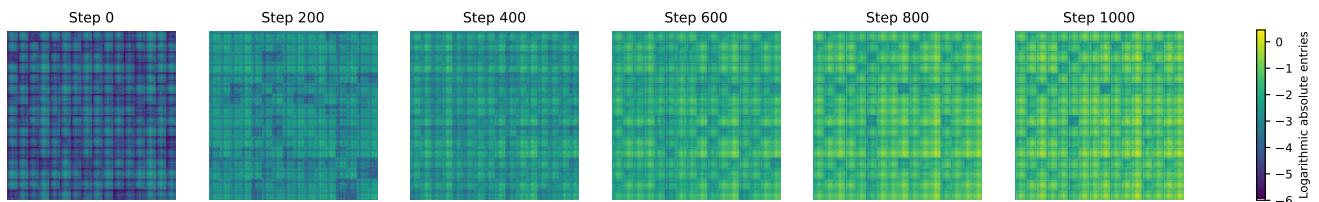


Рис. 3.6: Визуализация элементов блока матрицы Гессе, соответствующего параметрам Values трансформера (один блок, набор данных MNIST). Блок демонстрирует наибольшие значения элементов среди всех блоков гессиана, что согласуется с теоретическими оценками и подтверждает доминирующий вклад параметров Values в общую кривизну функции потерь.

(рис. 3.4, 3.5, 3.6, 3.7, 3.8).

Блок, соответствующий параметрам Queries (рис. 3.4), демонстрирует структурированное распределение элементов, отражающее взаимосвязи между параметрами механизма самовнимания. Блок Keys (рис. 3.5) показывает значительные значения элементов, подтверждая важную роль параметров Keys в форми-

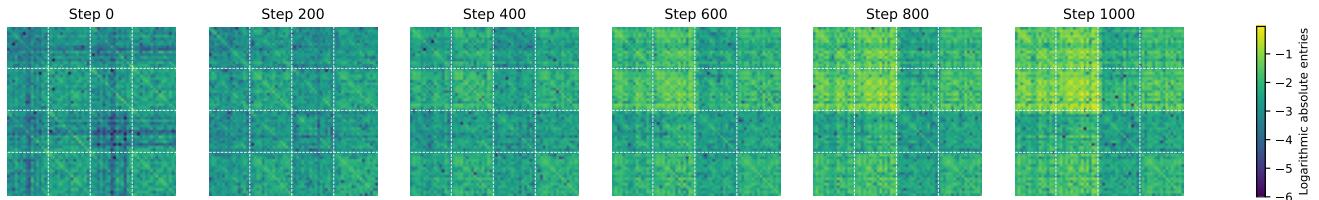


Рис. 3.7: Визуализация элементов блока матрицы Гессе, соответствующего параметрам LayerNorm трансформера (один блок, набор данных MNIST). Блок демонстрирует структурированное распределение элементов, отражающее влияние нормализации слоев на локальную геометрию функции потерь.

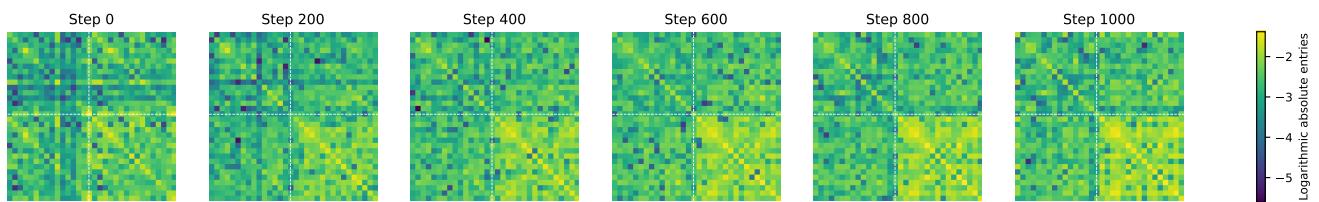


Рис. 3.8: Визуализация элементов блока матрицы Гессе, соответствующего параметрам блока FFN (Feed-Forward Network) трансформера (один блок, набор данных MNIST). Блок демонстрирует умеренные значения элементов по сравнению с блоками Keys и Values, что отражает специфику вклада полно связанных слоев в кривизну функции потерь.

ровании кривизны функции потерь.

Наиболее выраженные значения наблюдаются в блоке Values (рис. 3.6), что согласуется с теоретическими оценками и подтверждает доминирующий вклад параметров Values в общую кривизну функции потерь. Данное наблюдение согласуется с результатами теоремы 18, где оценки для блока Values содержат наибольшие слагаемые.

Блок LayerNorm (рис. 3.7) демонстрирует структурированное распределение элементов, отражающее влияние нормализации слоев на локальную геометрию функции потерь. Блок FFN (рис. 3.8) показывает умеренные значения элементов по сравнению с блоками Keys и Values, что отражает специфику вклада полно связанных слоев в кривизну функции потерь.

Для количественной оценки вклада различных компонентов трансформера в кривизну функции потерь вычисляются спектральные нормы блоков параметров и соответствующих блоков матрицы Гессе (рис. 3.3).

Результаты демонстрируют четкую иерархию вклада различных блоков:

наибольшие значения спектральных норм соответствуют блокам Keys и Values, в то время как остальные блоки (Queries, FFN, LayerNorm) демонстрируют значительно меньшие абсолютные значения элементов. Данное наблюдение качественно согласуется с теоретическими оценками, полученными в теореме 27, где оценки для блоков Keys и Values содержат наибольшие слагаемые, пропорциональные высоким степеням норм весовых матриц и входных данных.

Гетерогенность вклада различных блоков параметров в общую кривизну функции потерь, установленная теоретически, подтверждается эмпирически. Это указывает на то, что при практическом применении методов регуляризации или оптимизации следует учитывать различную важность различных компонентов трансформера, что может быть использовано для разработки более эффективных алгоритмов обучения.

Следует отметить ограничения проведенных экспериментов. Во-первых, эксперименты проводятся на относительно простой задаче (MNIST) с небольшой моделью (один блок трансформера), что может не отражать полную картину для более сложных задач и крупных моделей. Во-вторых, вычисления выполняются на одном батче данных, что может не полностью характеризовать поведение гессиана на всей выборке. В-третьих, количественное сравнение с теоретическими оценками ограничено отсутствием точных численных значений констант в теоретических границах.

Тем не менее, качественное соответствие между теоретическими предсказаниями и эмпирическими наблюдениями подтверждает корректность теоретического анализа и демонстрирует практическую применимость полученных оценок. Наблюдаемая гетерогенность вклада различных блоков параметров указывает на важность учета архитектурных особенностей при разработке методов оптимизации и регуляризации для трансформерных моделей.

3.5. Заключение по главе

Проведенное в настоящей главе исследование матриц Гессе для нейросетевых моделей глубокого обучения представляет собой систематический теоретический анализ локальных свойств оптимизационного ландшафта параметрических семейств функций. Полученные результаты формируют строгий математический аппарат для количественной оценки сложности современных архитектур глубокого обучения и обеспечивают теоретическую основу для применения ландшафтной меры сложности, введенной в главе 2.1..

Основным методологическим вкладом главы является разработка унифицированного подхода к анализу матриц Гессе на основе матричной факторизации. Теорема 11 устанавливает, что матрица Гессе для широкого класса матричных моделей может быть представлена в факторизованной форме $\mathbf{H}_O(\boldsymbol{\theta}) = \mathbf{Q}^T \mathbf{F}^T \mathbf{A} \mathbf{F} \mathbf{Q}$, что позволяет эффективно анализировать и вычислять гессиан без необходимости явного построения полной матрицы вторых производных. Данный подход унифицирует анализ для различных архитектур и создает основу для получения спектральных оценок, сформулированных в теореме 12.

Для полносвязных нейронных сетей установлены точные верхние оценки спектральной нормы матрицы Гессе. Теорема 9 демонстрирует, что норма гессиана ограничена выражением $\|\mathbf{H}_i(\boldsymbol{\theta})\|_2 \leq L\sqrt{2}M_x^2 M_{\mathbf{W}}^{2L} + \sqrt{2}\frac{M_{\mathbf{W}}^2(M_{\mathbf{W}}^{2L}-1)}{M_{\mathbf{W}}^{2}-1}$, что устанавливает экспоненциальную зависимость от глубины сети L и полиномиальную зависимость от ширины слоев. Теорема 10 уточняет данную оценку, показывая асимптотическую пропорциональность $\|\mathbf{H}_i(\boldsymbol{\theta})\|_2 \propto L(hM)^{2L}$, где h — размер скрытого слоя, а M — константа, ограничивающая параметры.

Для сверточных нейронных сетей получены явные оценки нормы гессиана через структурные параметры архитектуры. Теорема 13 устанавливает оценку для одномерных сверточных сетей: $\|\mathbf{H}_O\| \leq \sqrt{2}\|\mathbf{x}\|^2 d^2(L+1)(C^2 w^2 k d)^L$, демонстрируя мультипликативную зависимость от глубины и полиномиально-экспоненциальную зависимость от числа каналов C , размера ядра k и длины последовательности d . Теорема 14 определяет аналогичную оценку для двумерных сверточных сетей: $\|\mathbf{H}_O\| \leq \sqrt{2}\|\mathbf{x}\|^2 q^2(L+1)(C^2 k^2 w^2 m n)^L$, где зависимость от пространственных размеров $m \times n$ является квадратичной. Теоремы 15 и 16 исследуют регуляризующее влияние операций пулинга на сложность оптимизационного ландшафта, демонстрируя, что применение пулинга снижает норму гессиана и способствует стабилизации процесса обучения.

Наиболее значительным результатом главы является комплексный анализ матриц Гессе для трансформерных архитектур, восполняющий существенный пробел в теоретическом понимании этих моделей. Впервые получены явные выражения для матриц Якоби и Гессе ключевых компонентов трансформера. Теорема 19 устанавливает выражение для матрицы Якоби слоя LayerNorm, а теорема 20 — для матрицы Гессе, что представляет собой нетривиальный результат ввиду сложной нелинейной структуры операции нормализации. Теорема 18 определяет оценки нормы гессиана для механизма самовнимания, демонстрируя зависимость от структурных параметров трансформера. Теорема 27

устанавливает верхние оценки для полной матрицы Гессе трансформера, показывая гетерогенность вклада различных блоков параметров (Queries, Keys, Values, FFN, LayerNorm) в общую кривизну функции потерь. Эмпирические результаты, представленные в разделе 3.3, подтверждают теоретические оценки и демонстрируют, что наибольшие значения норм соответствуют блокам Keys и Values, что согласуется с полученными теоретическими предсказаниями.

Теоретическая значимость полученных результатов заключается в установлении прямой связи между спектральными свойствами матриц Гессе и введенной в главе 2.1. ландшафтной мерой сложности моделей. Полученные оценки спектральных норм матриц Гессе обеспечивают вычислимые методы оценки сложности для различных архитектур, что решает проблему непрактичности прямого вычисления матриц Гессе для крупных моделей, выявленную в обзоре литературы. Практическая ценность работы состоит в создании аппарата для сравнительного анализа архитектур, прогнозирования их поведения при масштабировании и разработки методов регуляризации на основе теоретических оценок.

Основные ограничения исследования связаны с детерминистическим характером анализа, предположениями о ограниченности параметров и данных, а также возможностью улучшения верхних оценок за счет учета специфики разреженных матриц и стохастических эффектов в процессе обучения. Перспективными направлениями дальнейших исследований являются уточнение оценок за счет учета стохастичности и спектральных распределений, разработка методов оптимизации и регуляризации на основе полученных теоретических результатов, а также расширение анализа на другие классы архитектур, такие как рекуррентные нейронные сети и графовые нейронные сети. Полученные результаты вносят существенный вклад в теорию анализа локальной геометрии ландшафтов функции потерь и создают основу для дальнейших исследований в области формализации сложности моделей глубокого обучения.

Глава 4

Достаточный объем выборки моделей

В главе 2 был разработан единый теоретический аппарат для формализации соотношения между сложностью модели и сложностью данных. Ключевым результатом этой главы является введение формальных определений меры сложности выборки $\mu_D(D)$ и меры сложности модели $\mu_f(f)$, а также установление критерия обучаемости модели на выборке: $\mu_f(f) \leq \mu_D(D)$.

В разделе 2.2. главы 2 было показано, что частным случаем условной меры сложности данных $\mu_D(D|f)$ является достаточный размер выборки m^* — минимальный объем данных из выборки D , необходимый для обучения модели f . Для простой генеральной совокупности Γ_C , состоящей из объектов одинаковой сложности C , мера сложности выборки принимает вид $\mu_D(D) = C \cdot |D|$, что устанавливает прямую связь между размером выборки и ее сложностью.

В разделе 2.3. главы 2 была введена ландшафтная мера сложности модели $\mu_f(f|D)$, определяемая через спектральные свойства матриц Гессе функции потерь. Установлено, что анализ сходимости ландшафта оптимизационной задачи при увеличении объема выборки сводится к анализу спектральной нормы матрицы Гессе, что позволяет количественно оценить влияние добавления новых объектов данных на локальную геометрию функции потерь в окрестности оптимума.

В главе 3 были получены конкретные оценки спектральных норм матриц Гессе для различных архитектур нейронных сетей. Теорема 9 устанавливает оценку для полносвязных сетей: $\|\mathbf{H}_i(\boldsymbol{\theta})\|_2 \propto L(hM)^{2L}$, теоремы 13 и 14 — для сверточных сетей, а теорема 27 — для трансформерных моделей. Эти оценки обеспечивают вычислимые методы оценки ландшафтной меры сложности модели $\mu_f(f|D)$, что создает теоретическую основу для практического применения формализма, разработанного в главе 2.

Однако для практического использования введенного теоретического аппарата необходимо решить следующую задачу: как на основе имеющихся данных и выбранной модели определить конкретное численное значение достаточного размера выборки m^* ? Теоретические результаты предыдущих глав устанавливают формальные связи между сложностью модели и сложностью данных, но не предоставляют алгоритмических процедур для вычисления m^* в конкретных прикладных задачах. Более того, для моделей глубокого обучения прямое вычисление матриц Гессе и оценка ландшафтной меры сложности остаются вы-

числительно дорогостоящими, что ограничивает практическую применимость теоретических результатов.

В рамках настоящей главы рассматриваются различные методы определения достаточного размера выборки для различных моделей, от линейных до моделей глубокого обучения. Предлагаемые методы опираются на теоретический аппарат, разработанный в предыдущих главах, и обеспечивают практические инструменты для оценки необходимого объема данных при планировании экспериментов. В отличие от теоретических оценок, основанных на анализе матриц Гессе, предлагаемые методы используют наблюдаемые характеристики процесса обучения для определения момента, когда добавление новых объектов данных перестает существенно влиять на свойства модели.

Планирование эксперимента требует оценки минимального размера выборки: числа выполненных измерений набора характеристик, необходимых для построения сформулированных условий. Выбор метода оценки размера выборки зависит от решаемой задачи, которая определяет формулировку статистической гипотезы и статистики для ее проверки.

В целом существуют различные подходы к определению достаточного размера выборки, такие как статистические, байесовские и эвристические методы. Каждый из этих подходов имеет свои преимущества и ограничения, которые определяют область их применимости.

Статистические методы предполагают, что выборка соответствует некоторым предварительным условиям, сформулированным ранее. Эти условия сформулированы как статистический критерий [22, 23, 24, 14]. Метод оценки размера выборки, связанный с этим критерием, гарантирует достижение фиксированной статистической мощности $1 - \beta$ со степенью ошибки первого рода, не превышающей установленное значение α . Такой размер выборки называется достаточным.

Однако практическое применение методов оценки размера выборки предполагает, что модель соответствует измеренным данным [17]. Эти модели выбираются в соответствии с постановкой задачи регрессии или классификации. В настоящей главе рассматриваются обобщенные линейные модели.

В работе [23] предложен подход к оценке мощности и размера выборки на основе теста отношения максимального правдоподобия. Этот подход оказался более точным для ряда независимых переменных. В работе [25] предложен метод оценки мощности для статистики Вальда. В работе [19] в случае логистической регрессии предлагается использовать метод, использующий кривую

ROC-AUC и концепцию сдвига.

Классические методы [22, 23, 24, 25, 14] имеют ряд ограничений, связанных с практическим применением. Чтобы оценить размер выборки, необходимо знать дисперсию оценки параметра или, в более общем случае, иметь оценку параметра нецентральности в распределении статистики, используемой при альтернативной гипотезе. Указанные методы не предоставляют алгоритмических процедур для получения этих значений. Кроме того, дисперсия оценки и параметр нецентральности оцениваются с неопределенностью, влияние которой на результат оценки размера выборки не учитывается.

Статистические методы позволяют оценить размер выборки на основе предположений о распределении данных и информации о соответствии между наблюдаемыми значениями и предположениями нулевой гипотезы.

Когда размер исследуемой выборки является достаточным или чрезмерным, возможно применение методов, основанных на наблюдении изменения определенной характеристики процедуры построения модели при увеличении размера выборки. В частности, наблюдая за соотношением качества прогнозирования с контрольной выборкой и обучающей выборкой [19], определяется достаточный размер выборки, который соответствует началу переобучения.

В работе [20] для оценки достаточного размера выборки используется процедура бутстрата. Превышение текущего размера выборки проверяется на основе анализа доверительных интервалов оцениваемого параметра. Ширина доверительного интервала с разными значениями объема выборки оценивается с помощью метода бутстрата. Для этого выборки меньшего размера отбираются заданное число раз и вычисляется доверительный интервал ошибки при оценке параметра модели. Размер выборки считается достаточным, если ширина доверительного интервала не превышает заранее установленного значения.

Перечисленные выше ограничения статистических методов оценки размера выборки подробно исследуются в байесовской процедуре [18, 21, 26]. В рамках данного подхода оценка размера выборки определяется на основе максимизации ожидаемого значения некоторой функции качества [18]. Функция качества может включать в себя явные функции распределения параметров и штрафы за увеличение размера выборки.

Альтернативой подходам [26], основанным на функции качества, является выбор размера выборки путем установления ограничений на определенный критерий качества оценки параметров модели. Примеры критериев: критерий средней апостериорной дисперсии (AVPC), критерий средней длины (ALC), крите-

рий среднего покрытия (ACC). Для каждого перечисленного критерия оценка размера выборки определяется как минимальное значение размера выборки, для которого ожидаемое значение выбранного критерия не превышает какого-либо фиксированного порога.

В работе [19] предлагается считать размер выборки достаточным, если расстояние Кульбака-Лейблера между распределениями, оцененными на основе подвыборок такого размера, достаточно мало. Такой подход не требует дальнейшего обобщения в случае нескольких переменных. Кроме того, оценка может производиться как при наличии предположений о распределении данных, так и при их отсутствии. Недостаток этого подхода заключается в том, что количественная оценка может быть получена только при чрезмерно большом размере выборки.

4.1. Статистические методы определения достаточного размера выборки

В настоящем разделе рассматриваются статистические методы определения достаточного размера выборки для обобщенных линейных моделей. Основой данных методов является использование информационной матрицы Фишера и статистических критериев для проверки гипотез о параметрах модели.

Рассмотрим выборку размера m :

$$\mathfrak{D}_m = \{\mathbf{x}_i, y_i\}_{i=1}^m,$$

где $\mathbf{x}_i \in \mathbb{R}^n$ — вектор признаков, $y_i \in \mathbb{Y}$ — целевая переменная. Вектор признаков $\mathbf{x} = [\mathbf{u}, \mathbf{v}]$ объединяет $\mathbf{u}_i \in \mathbb{R}^k$ и $\mathbf{v}_i \in \mathbb{R}^{n-k}$. Выборка \mathfrak{D}_m случайным образом разделяется на обучающую и тестовую части:

$$\mathfrak{D}_{\mathcal{T}_m} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i \in \mathcal{T}_m}, \quad \mathfrak{D}_{\mathcal{L}_m} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i \in \mathcal{L}_m}, \quad \mathcal{T}_m \sqcup \mathcal{L}_m = \{1, \dots, m\}.$$

Определим параметрическое семейство функций для аппроксимации неизвестного распределения $p(y|\mathbf{x}, \mathfrak{D}_{\mathcal{L}_m})$:

$$\mathfrak{F} = \left\{ f(y, \mathbf{x}, \mathbf{w}) \mid \mathbf{w} \in \mathbb{W}, \int_{y \in \mathbb{Y}, \mathbf{x} \in \mathbb{R}^n} f(y, \mathbf{x}, \mathbf{w}) dy d\mathbf{x} = 1 \right\}.$$

Для модели f с вектором параметров \mathbf{w} определим функцию правдоподобия и логарифм функции правдоподобия выборки \mathfrak{D} :

$$L(\mathfrak{D}, \mathbf{w}) = \prod f(y, \mathbf{x}, \mathbf{w}), \quad l(\mathfrak{D}, \mathbf{w}) = \sum \log f(y, \mathbf{x}, \mathbf{w}),$$

где $f(y, \mathbf{x}, \mathbf{w})$ является оценкой плотности вероятности для объекта (y, \mathbf{x}) при заданном векторе параметров \mathbf{w} .

Используя принцип максимального правдоподобия для оценки параметров \mathbf{w} :

$$\hat{\mathbf{w}} = \arg \max_{\mathbf{w} \in \mathbb{W}} L(\mathfrak{D}_{\mathcal{L}}, \mathbf{w}).$$

Информационная матрица Фишера (англ. Fisher Information Matrix) имеет вид:

$$\mathbf{I}(\mathfrak{D}, \mathbf{w}) = -\nabla \nabla^T l(\mathfrak{D}, \mathbf{w}), \quad \mathbf{V} = \mathbf{I}^{-1}(\mathfrak{D}, \mathbf{m}), \quad (4.1)$$

где \mathbf{V} — ковариационная матрица оценок параметров. Статистические методы и байесовские методы используют информационную матрицу Фишера для оценки размера выборки.

Основным преимуществом методов, основанных на статистике, является их способность оценивать достаточный размер выборки при недостаточном наборе данных. Они позволяют прогнозировать необходимое число объектов на ранней стадии эксперимента.

Рассмотрим обобщенную линейную модель, в которой плотность распределения целевой переменной задается выражением

$$p(y|\mathbf{u}, \mathbf{v}, \mathbf{w}_u, \mathbf{w}_v) = \exp(y\theta - b(\theta) + c(y)),$$

где θ является каноническим параметром распределения, получаемым с помощью функции связи $\theta = \theta(\mathbf{u}, \mathbf{v}, \mathbf{w}_u, \mathbf{w}_v)$, а функции $b(\theta)$ и $c(y)$ определяют конкретный тип распределения.

Тестируемая гипотеза

$$H_0 : \mathbf{m}_u = \mathbf{m}_u^0, \quad H_1 : \mathbf{m}_u \neq \mathbf{m}_u^0.$$

Пусть статистики $S_{m,u}(\mathbf{w}_u, \mathbf{w}_v)$ и $S_{m,v}(\mathbf{w}_u, \mathbf{w}_v)$ представляют собой производные логарифма правдоподобия выборки \mathfrak{D}_m по параметрам \mathbf{w}_u и \mathbf{w}_v соответственно. Рассмотрим $\mathbf{s}_m = S_{m,u}(\mathbf{m}_u^0, \hat{\mathbf{w}}_v^0)$, где $\hat{\mathbf{w}}_v^0$ получается из уравнения

$$S_{m,v}(\mathbf{m}_u^0, \mathbf{w}_v) = 0.$$

Статистика множителей Лагранжа (англ. Lagrange Multiplier) определяется как

$$LM = \mathbf{s}_m^T \mathbf{Q}_m^{-1} \mathbf{s}_m,$$

где \mathbf{Q}_m — ковариационная матрица вектора \mathbf{s}_m .

В случае истинности гипотезы H_0 статистика LM асимптотически имеет центральное распределение $\chi^2(k)$. В [22] показано, что при альтернативной гипотезе H_1 статистика LM асимптотически имеет нецентральное распределение $\chi^2(k, \gamma)$, где γ является параметром нецентральности

$$\gamma = \boldsymbol{\xi}_m^\top \boldsymbol{\Sigma}_m^{-1} \boldsymbol{\xi}_m = m \boldsymbol{\xi}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\xi} = m\gamma^0, \quad (4.2)$$

где $\boldsymbol{\xi}_m$ и $\boldsymbol{\Sigma}_m$ — соответственно вектор математического ожидания и матрица ковариации \mathbf{s}_m . Обозначим $\boldsymbol{\xi}_1 = \boldsymbol{\xi}$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$.

Альтернативный метод получения γ включает условия на уровне значимости α и вероятность ошибки II рода β :

$$\gamma^* : \chi_{k,1-\alpha}^2 = \chi_{k,\beta}^2(\gamma). \quad (4.3)$$

Используя соотношения (4.2) и (4.3), получаем

$$m^* = \frac{\gamma^*}{\gamma^0}.$$

Полученное значение m^* представляет собой достаточный минимальный размер выборки, необходимый для различия вектора \mathbf{m}_u от \mathbf{m}_u^0 с заданными уровнями значимости α и мощности $1 - \beta$.

Рассмотрим случай, когда правдоподобие выборки задается выражением

$$p(y|\mathbf{u}, \mathbf{v}, \mathbf{w}_u, \mathbf{w}_v) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right), \quad (4.4)$$

где θ является параметром распределения, который вычисляется с помощью функции связи $\theta = \theta(\mathbf{u}, \mathbf{v}, \mathbf{w}_u, \mathbf{w}_v)$.

Тестируемая гипотеза

$$H_0 : \mathbf{m}_u = \mathbf{m}_u^0, \quad H_1 : \mathbf{m}_u \neq \mathbf{m}_u^0.$$

Определим логарифм статистики отношения правдоподобий:

$$LR = 2 \left(l(\mathfrak{D}, \hat{\mathbf{w}}) - l(\mathfrak{D}, \hat{\mathbf{w}}^0) \right),$$

где $\hat{\mathbf{w}} = [\hat{\mathbf{w}}_u, \hat{\mathbf{w}}_v]$ — вектор параметров, максимизирующий правдоподобие (4.4), а $\hat{\mathbf{w}}^0 = [\mathbf{m}_u^0, \hat{\mathbf{w}}_v^0]$ — вектор параметров, максимизирующий правдоподобие (4.4) при фиксированном подвекторе параметров \mathbf{m}_u^0 .

В случае истинности гипотезы H_0 статистика LR асимптотически имеет центральное распределение $\chi^2(k)$. В [24] показано, что при альтернативной

гипотезе H_1 статистика LR асимптотически имеет нецентральное распределение $\chi^2(k, \gamma)$, где γ является параметром нецентральности

$$\gamma = m\Delta^*, \quad \Delta^* = \mathbb{E} [2a^{-1}(\phi) \{(\theta - \theta^*) \nabla b(\theta) - b(\theta) + b(\theta^*)\}],$$

где параметры θ и θ^* рассчитываются с использованием параметров $\mathbf{w} = [\mathbf{w}_u, \mathbf{w}_v]$ и $\mathbf{w}^* = [\mathbf{w}_u^0, \mathbf{w}_v^*]$. Параметры \mathbf{w}_v^* вычисляются на основе решения уравнения

$$\lim_{m \rightarrow \infty} m^{-1} \mathbb{E} \left(\frac{\partial l (\mathfrak{D}, [\mathbf{m}_u^0, \mathbf{w}_v])}{\partial \mathbf{w}_v} \right) = 0.$$

Тогда с учетом α и β достаточный размер выборки m^* вычисляется

$$m^* = \frac{\gamma^*}{\Delta^*}, \quad \gamma^* : \chi_{k,1-\alpha}^2 = \chi_{k,\beta}^2(\gamma),$$

где $\chi_{k,1-\alpha}^2$ и $\chi_{k,\beta}^2(\gamma^*)$ — квантили распределений χ_k^2 и $\chi_k^2(\gamma)$ соответственно. Правдоподобие выборки:

$$p(y|\mathbf{u}, \mathbf{v}, \mathbf{w}_u, \mathbf{w}_v) = \exp \left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right), \quad (4.5)$$

где θ является параметром распределения, который вычисляется с помощью функции связи $\theta = \theta(\mathbf{u}, \mathbf{v}, \mathbf{w}_u, \mathbf{w}_v)$.

Тестируемая гипотеза:

$$H_0 : \mathbf{m}_u = \mathbf{m}_u^0, \quad H_1 : \mathbf{m}_u \neq \mathbf{m}_u^0.$$

Тест Вальда для гипотезы:

$$W = (\hat{\mathbf{w}}_u - \mathbf{m}_u^0)^T \hat{\mathbf{V}}_u^{-1} (\hat{\mathbf{w}}_u - \mathbf{m}_u^0),$$

где $\hat{\mathbf{w}} = [\hat{\mathbf{w}}_u, \hat{\mathbf{w}}_v]$ вектор параметров, который максимизирует правдоподобие выборки (4.5), где матрица $\hat{\mathbf{V}}_u$ задается в выражении (4.1).

В случае истинности гипотезы H_0 статистика Вальда W асимптотически имеет центральное распределение $\chi^2(k)$. В [25] показано, что в случае истинности альтернативной гипотезы H_1 статистика Вальда W асимптотически имеет нецентральное распределение $\chi^2(k, \gamma)$ с параметром нецентральности γ :

$$\gamma = m\delta, \quad \delta = (\hat{\mathbf{w}}_u - \mathbf{m}_u^0)^T \Sigma_u^{-1} (\hat{\mathbf{w}}_u - \mathbf{m}_u^0), \quad \Sigma_u = m\hat{\mathbf{V}}_u.$$

Используя заданный уровень значимости α и заданную ошибку второго рода β , определим оптимальный размер выборки:

$$m^* = \frac{\gamma^*}{\delta}, \quad \gamma^* : \chi_{k,1-\alpha^*}^2 = \chi_{k,\beta}^2(\gamma),$$

где $\chi_{k,1-\alpha^*}^2$ и $\chi_{k,\beta}^2(\gamma^*)$ — квантили соответствующих распределений, а параметр α^* представляет собой поправку на уровень значимости:

$$\alpha^* = P(\boldsymbol{\xi}^\top \Sigma^{*-1} \boldsymbol{\xi} > \chi_{k,1-\alpha}^2), \quad \Sigma^* = \mathbf{I}^{-1}(\mathfrak{D}, \mathbf{w}^*),$$

где $\mathbf{w}^* = [\mathbf{m}_u^0, \mathbf{w}_v^*]$ представляет собой решение уравнения

$$\lim_{m \rightarrow \infty} m^{-1} \mathsf{E} \left(\frac{\partial l(\mathfrak{D}, [\mathbf{m}_u^0, \mathbf{w}_v])}{\partial \mathbf{w}_v} \right) = 0.$$

Статистические методы, рассмотренные выше, требуют знания дисперсии оценки параметра или параметра нецентральности, что ограничивает их практическое применение. В следующем разделе рассматриваются эвристические методы, которые не требуют таких предположений и могут применяться в более широком классе задач.

4.2. Эвристические методы определения достаточного размера выборки

В настоящем разделе рассматриваются эвристические методы определения достаточного размера выборки, основанные на популярных статистических эвристиках, таких как бутстррап, перекрестная проверка и задание функции полезности. В отличие от статистических методов, эвристические подходы не требуют строгих предположений о распределении данных и могут применяться в ситуациях, когда теоретические гарантии недоступны.

Определим набор индексов \mathcal{A} для параметров логистической регрессии \mathbf{w} . Тестируется гипотеза

$$H_0 : j \notin \mathcal{A} (\mathbf{w}_j = 0), \quad H_1 : j \in \mathcal{A}^* (\mathbf{w}_j \neq 0),$$

где \mathbf{w}_j является j -м элементом вектора \mathbf{w} . Установим параметр отступа c_0 для задачи логистической регрессии:

$$H_0 : 1 - c_0 = p_0, \quad H_1 : 1 - c_0 = p_1,$$

где c_0 оптимальное решение, когда исключен j -й элемент вектора. Используя статистику

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}} \sqrt{m}, \quad \hat{p} = \frac{1}{m} \sum_{i=1}^m y_i.$$

В случае истинности нулевой гипотезы H_0 статистика Z асимптотически имеет распределение $\mathcal{N}(0, 1)$. В случае истинности альтернативной гипотезы H_1 статистика Z асимптотически имеет распределение $\mathcal{N}\left(p_1 - p_0, \sqrt{\frac{p_1(1-p_1)}{p_0(1-p_0)}}\right)$.

Достаточный объем выборки задается выражением

$$m^* = \frac{p_0(1 - p_0) \left(Z_{1-\alpha/2} + Z_{1-\beta} \sqrt{\frac{p_1(1-p_1)}{p_0(1-p_0)}} \right)^2}{(p_1 - p_0)^2},$$

где $Z_{1-\alpha/2}$ и $Z_{1-\beta}$ — квантили стандартного нормального распределения $\mathcal{N}(0, 1)$.

Данный метод не рассматривается далее, поскольку его можно использовать только в задаче логистической регрессии.

Рассмотрим метод на основе кросс-валидации (англ. cross-validation). Определим критерий переобучения как

$$RS(m) = \ln \frac{L(\mathfrak{D}_{\mathcal{L}(m)}, \hat{\mathbf{w}})}{L(\mathfrak{D}_{\mathcal{T}(m)}, \hat{\mathbf{w}})}, \quad \frac{|\mathcal{T}(m)|}{|\mathcal{L}(m)|} = \text{const} \leq 0.5.$$

Справедливо следующее предельное соотношение:

$$\lim_{m \rightarrow \infty} RS(m) = 0.$$

Достаточный размер выборки m^* определяется согласно условию:

$$m^* : \forall m \geq m^* \mathbb{E}_{\mathfrak{D}_m} RS(m) \leq \varepsilon,$$

где ε некоторый параметр, который задается экспертизой.

Этот метод предполагает, что длины доверительных интервалов квантиля не превышают некоторого фиксированного значения l . Для некоторого размера выборки m вычисляются квантильные доверительные интервалы $(a_1^m, b_1^m), (a_2^m, b_2^m), \dots, (a_n^m, b_n^m)$ с уровнем значимости α с использованием начальной загрузки для каждого параметра модели. Достаточный размер выборки задается выражением:

$$m^* : \forall m \geq m^* \max_i (b_i^m - a_i^m) < l.$$

Данный метод является покоординатным, и следовательно для повышения точности прогноза требуется значительное увеличение размера выборки.

Эвристические методы, рассмотренные выше, основаны на наблюдении за поведением модели при изменении размера выборки, но не предоставляют строгих теоретических гарантий. В следующем разделе рассматриваются байесовские методы, которые позволяют формализовать задачу определения достаточного размера выборки в рамках вероятностного подхода.

4.3. Байесовские методы определения достаточного размера выборки

В настоящем разделе рассматриваются байесовские методы оценки размера выборки, основанные на ограничении некоторых характеристик модели. Для анализа эффективности определяется функция размера выборки. Увеличение этой функции интерпретируется как снижение эффективности модели. Размер выборки m^* выбирается таким, чтобы исследуемая функция принимала значения меньше некоторого порогового значения ε .

Размер выборки m^* определяется условием:

$$\forall m \geq m^* \mathbb{E}_{\mathfrak{D}_m} D [\hat{\mathbf{w}} | \mathfrak{D}_m] \leq l.$$

где l некоторый заданный экспертино параметр, который количественно определяет неопределенность оценки параметра.

Обозначим через $A(\mathfrak{D}) \subset \mathbb{R}^n$ некоторый набор параметров модели \mathbf{w} :

$$A(\mathfrak{D}) = \{\mathbf{w} : \|\mathbf{w} - \hat{\mathbf{w}}\| \leq l\},$$

где l — некоторый фиксированный радиус шара. Размер выборки m^* определяется критерием среднего покрытия:

$$\forall m \geq m^* \mathbb{E}_{\mathfrak{D}_m} P \{ \mathbf{w} \in A(\mathfrak{D}_m) \} \geq 1 - \alpha, \quad (4.6)$$

где α некоторый параметр заданный экспертино.

Определим функцию $A(\mathfrak{D})$:

$$P(A(\mathfrak{D})) = 1 - \alpha.$$

Оценка критерия средней длины m^* , заданная в (4.6):

$$\forall m \geq m^* \mathbb{E}_{\mathfrak{D}_m} r_m \leq l,$$

где r_m является радиусом шара $A(\mathfrak{D}_m)$.

Следующие методы максимизируют ожидание некоторой функции полезности $u(\mathfrak{D}, \mathbf{w})$ по размеру выборки:

$$m^* = \arg \max_m \mathbb{E}_{\mathfrak{D}_m} \int_{\mathbf{w}} u(\mathfrak{D}_m, \mathbf{w}) p(\mathbf{w} | \mathfrak{D}_m) d\mathbf{w},$$

где функция полезности $u(\mathfrak{D}, \mathbf{w})$ задается в виде:

$$u(\mathfrak{D}_m, \mathbf{w}) = l(\mathfrak{D}_m, \mathbf{w}) - cm,$$

где c — коэффициент штрафа для каждого элемента в наборе выборки.

Назовем индексы $\mathcal{B}_1, \mathcal{B}_2 \subset \{1, \dots, m\}$ по соседству, если

$$|\mathcal{B}_1 \Delta \mathcal{B}_2| = 1.$$

Таким образом, \mathcal{B}_2 можно преобразовать в \mathcal{B}_1 путем удаления, замены или добавления одного элемента. В [19] показано, что если размер набора выборок $\mathfrak{D}_{\mathcal{B}_1}$ достаточно велик, то параметры модели $\hat{\mathbf{w}}_1$, оптимизированные с помощью $\mathfrak{D}_{\mathcal{B}_1}$, должны находиться в окрестности параметров модели $\hat{\mathbf{w}}_2$, которые оптимизированы с помощью $\mathfrak{D}_{\mathcal{B}_2}$.

Используя дивергенцию Кульбака-Лейблера в качестве функции близости между распределениями параметров модели, оптимизированных с помощью $\mathfrak{D}_{\mathcal{B}_1}$ и $\mathfrak{D}_{\mathcal{B}_2}$:

$$D_{KL}(p_1, p_2) = \int_{\mathbf{w} \in \mathbb{W}} p_1(\mathbf{w}) \log \frac{p_1(\mathbf{w})}{p_2(\mathbf{w})} d\mathbf{w},$$

где p_1 и p_2 — апостериорные распределения вектора параметров \mathbf{w} , рассчитанные на подвыборках $\mathfrak{D}_{\mathcal{B}_1}$ и $\mathfrak{D}_{\mathcal{B}_2}$ соответственно. Также предполагается, что $\mathfrak{D}_{\mathcal{B}_1}$ и $\mathfrak{D}_{\mathcal{B}_2}$ находятся по соседству. Достаточный размер выборки m^* оценивается:

$$\forall \mathfrak{D}_{\mathcal{B}_1} : |\mathfrak{D}_{\mathcal{B}_1}| \geq m^* \mathsf{E}_{\mathfrak{D}_{\mathcal{B}_2}} D_{KL}(p_1, p_2) \leq \varepsilon.$$

Рассмотренные выше классические методы определения достаточного размера выборки имеют существенные ограничения: статистические методы требуют знания дисперсии оценок параметров, байесовские методы — вычислительно сложны для моделей с большим числом параметров, а эвристические методы не предоставляют строгих теоретических гарантий. В следующих разделах предлагаются новые методы, основанные на анализе стабильности функции правдоподобия и близости апостериорных распределений, которые преодолевают указанные ограничения.

4.4. Метод определения достаточного размера выборки на основе сэмплирования эмпирической функции ошибки

В настоящем разделе рассматривается метод определения достаточного размера выборки, основанный на анализе стабильности функции правдоподобия

при изменении объема данных. Предполагается, что выполняется условие $m^* \leq m$, где m — размер доступной выборки D , а m^* — искомый достаточный размер. Таким образом, требуется определить минимальный объем выборки, который следует считать достаточным для обучения модели, при условии наличия достаточного количества объектов в самой выборке D .

Для определения достаточности используется функция правдоподобия. Когда доступно достаточное количество объектов, естественно ожидать, что полученная оценка параметров не будет существенно изменяться от одной реализации выборки к другой [15, 16]. Аналогичное утверждение справедливо и для функции правдоподобия. Таким образом, формализуем критерии, позволяющие определить достаточный объем выборки.

Критерий определяется в определении 22.

Определение 22 (D-достаточный размер выборки). *Пусть задано некоторое $\varepsilon > 0$. Размер выборки m^* назовем D-достаточным, если для всех $k \geq m^*$ выполняется условие:*

$$D(k) = \mathbb{D}_{\hat{\mathbf{w}}_k} L(\mathfrak{D}_m, \hat{\mathbf{w}}_k) \leq \varepsilon.$$

Определение 22 формализует идею о том, что при достаточном размере выборки дисперсия функции правдоподобия по различным реализациям подвыборок должна быть мала, что указывает на стабильность оценки параметров.

С другой стороны, при наличии достаточного количества объектов естественно ожидать, что при добавлении еще одного объекта к рассмотрению результирующая оценка параметра изменится незначительно; на основе данного свойства получаем определение 23.

Определение 23 (M-достаточный размер выборки). *Пусть задано некоторое $\varepsilon > 0$. Размер выборки m^* назовем M-достаточным, если для всех $k \geq m^*$ выполняется условие:*

$$M(k) = |\mathbb{E}_{\hat{\mathbf{w}}_{k+1}} L(\mathfrak{D}_m, \hat{\mathbf{w}}_{k+1}) - \mathbb{E}_{\hat{\mathbf{w}}_k} L(\mathfrak{D}_m, \hat{\mathbf{w}}_k)| \leq \varepsilon.$$

Определение 23 формализует условие, при котором добавление нового объекта к выборке не приводит к существенному изменению математического ожидания функции правдоподобия, что указывает на достижение достаточного объема данных для стабильной оценки параметров модели.

В приведенных выше определениях вместо функции правдоподобия $L(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$ рассматривается ее логарифм $l(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$, что упрощает математический анализ за счет перехода от произведения к сумме.

Предположим, что $\mathbb{W} = \mathbb{R}^n$, информационная матрица Фишера задана матрицей:

$$[\mathcal{I}(\mathbf{w})]_{ij} = -\mathbb{E} \left[\frac{\partial^2 \log p(\mathbf{y}|\mathbf{x}, \mathbf{w})}{\partial w_i \partial w_j} \right],$$

Известным результатом является асимптотическая нормальность оценки максимального правдоподобия:

$$\sqrt{k} (\hat{\mathbf{w}}_k - \mathbf{w}) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}^{-1}(\mathbf{w})),$$

где \xrightarrow{d} обозначает сходимость по распределению. Следует отметить, что сходимость по распределению, вообще говоря, не влечет сходимости моментов случайного вектора. Однако если предположить сходимость моментов, то в некоторых моделях можно доказать корректность предложенного определения М-достаточного размера выборки.

Для удобства обозначим параметры распределения $\hat{\mathbf{w}}_k$ следующим образом: математическое ожидание $\mathbb{E}\hat{\mathbf{w}}_k = \mathbf{m}_k$ и матрица ковариаций $\mathbb{D}\hat{\mathbf{w}}_k = \Sigma_k$. Тогда справедлива теорема 28, которая доказывает сходимость параметров.

Теорема 28 (*Киселев-Грабовой*, Корректность М-достаточного размера выборки). *Пусть $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \rightarrow 0$ и $\|\Sigma_{k+1} - \Sigma_k\|_F \rightarrow 0$ при $k \rightarrow \infty$. Тогда в модели линейной регрессии определение М-достаточного размера выборки корректно. А именно, для любого $\varepsilon > 0$ существует такой m^* , что для всех $k \geq m^*$ выполняется $M(k) \leq \varepsilon$.*

Доказательство. Рассмотрим определение М-достаточного размера выборки в терминах логарифма функции правдоподобия. В модели линейной регрессии

$$\begin{aligned} L(\mathfrak{D}_m, \hat{\mathbf{w}}_k) &= p(\mathbf{y}|\mathbf{X}, \hat{\mathbf{w}}_k) = \prod_{i=1}^m p(y_i|\mathbf{x}_i, \hat{\mathbf{w}}_k) = \prod_{i=1}^m \mathcal{N}(y_i|\hat{\mathbf{w}}_k^\top \mathbf{x}_i, \sigma^2) = \\ &= (2\pi\sigma^2)^{-m/2} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}_k\|_2^2\right). \end{aligned}$$

Возьмем логарифм:

$$l(\mathfrak{D}_m, \hat{\mathbf{w}}_k) = \log p(\mathbf{y}|\mathbf{X}, \hat{\mathbf{w}}_k) = -\frac{m}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}_k\|_2^2.$$

Возьмем математическое ожидание по \mathfrak{D}_k , учитывая что $\mathbb{E}_{\mathfrak{D}_k} \hat{\mathbf{w}}_k = \mathbf{m}_k$ и $\text{cov}(\hat{\mathbf{w}}_k) = \Sigma_k$:

$$\mathbb{E}_{\mathfrak{D}_k} l(\mathfrak{D}_m, \hat{\mathbf{w}}_k) = -\frac{m}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left(\|\mathbf{y} - \mathbf{X}\mathbf{m}_k\|_2^2 + \text{tr}(\mathbf{X}^\top \mathbf{X} \Sigma_k) \right).$$

Запишем выражение для разности математических ожиданий:

$$\begin{aligned}
& \mathbb{E}_{\mathfrak{D}_{k+1}} l(\mathfrak{D}_m, \hat{\mathbf{w}}_{k+1}) - \mathbb{E}_{\mathfrak{D}_k} l(\mathfrak{D}_m, \hat{\mathbf{w}}_k) = \\
&= \frac{1}{2\sigma^2} \left(\|\mathbf{y} - \mathbf{X}\mathbf{m}_k\|_2^2 - \|\mathbf{y} - \mathbf{X}\mathbf{m}_{k+1}\|_2^2 \right) + \frac{1}{2\sigma^2} \text{tr} \left(\mathbf{X}^\top \mathbf{X} (\Sigma_k - \Sigma_{k+1}) \right) = \\
&= \frac{1}{2\sigma^2} \left(2\mathbf{y}^\top \mathbf{X} (\mathbf{m}_{k+1} - \mathbf{m}_k) + (\mathbf{m}_k - \mathbf{m}_{k+1})^\top \mathbf{X}^\top \mathbf{X} (\mathbf{m}_k + \mathbf{m}_{k+1}) \right) + \\
&\quad + \frac{1}{2\sigma^2} \text{tr} \left(\mathbf{X}^\top \mathbf{X} (\Sigma_k - \Sigma_{k+1}) \right).
\end{aligned}$$

Значение функции $M(k)$ представляет собой модуль от приведенного выше выражения. Применим неравенство треугольника для модуля, а затем оценим каждое слагаемое.

Оценим первое слагаемое, используя неравенство Коши-Буняковского:

$$|\mathbf{y}^\top \mathbf{X} (\mathbf{m}_{k+1} - \mathbf{m}_k)| \leq \|\mathbf{X}^\top \mathbf{y}\|_2 \|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2.$$

Второе слагаемое оцениваем с помощью неравенства Коши-Буняковского, свойства согласованности спектральной нормы матрицы, а также ограниченности последовательности векторов \mathbf{m}_k , что следует из приведенного условия сходимости:

$$\begin{aligned}
|(\mathbf{m}_k - \mathbf{m}_{k+1})^\top \mathbf{X}^\top \mathbf{X} (\mathbf{m}_k + \mathbf{m}_{k+1})| &\leq \|\mathbf{X}(\mathbf{m}_k - \mathbf{m}_{k+1})\|_2 \|\mathbf{X}(\mathbf{m}_k + \mathbf{m}_{k+1})\|_2 \leq \\
&\leq \|\mathbf{X}\|_2^2 \|\mathbf{m}_k - \mathbf{m}_{k+1}\|_2 \|\mathbf{m}_k + \mathbf{m}_{k+1}\|_2 \leq \\
&\leq C \|\mathbf{X}\|_2^2 \|\mathbf{m}_k - \mathbf{m}_{k+1}\|_2.
\end{aligned}$$

Последнее слагаемое оцениваем, используя неравенство Гельдера для нормы Фробениуса:

$$|\text{tr} \left(\mathbf{X}^\top \mathbf{X} (\Sigma_k - \Sigma_{k+1}) \right)| \leq \|\mathbf{X}^\top \mathbf{X}\|_F \|\Sigma_k - \Sigma_{k+1}\|_F.$$

Наконец, поскольку $\|\mathbf{m}_k - \mathbf{m}_{k+1}\|_2 \rightarrow 0$ и $\|\Sigma_k - \Sigma_{k+1}\|_F \rightarrow 0$ при $k \rightarrow \infty$, то $M(k) \rightarrow 0$ при $k \rightarrow \infty$, что и доказывает теорему. \square

Теорема 28 устанавливает корректность определения М-достаточного размера выборки для модели линейной регрессии при выполнении условий сходимости моментов оценок параметров. Указанные условия являются естественными для асимптотически нормальных оценок максимального правдоподобия и выполняются при стандартных предположениях о регулярности модели.

Следствие 4 (Корректность М-достаточного размера выборки при сходимости к истинным параметрам). *Пусть $\|\mathbf{m}_k - \mathbf{w}\|_2 \rightarrow 0$ и $\|\Sigma_k - [k\mathcal{I}(\mathbf{w})]^{-1}\|_F \rightarrow 0$ при $k \rightarrow \infty$.*

Тогда в модели линейной регрессии определение М-достаточного размера выборки корректно.

По условию, задана только одна выборка, а следовательно, в эксперименте невозможно вычислить математическое ожидание и дисперсию, указанные в определениях 22 и 23. Поэтому для их оценки используется метод бутстрэпирования (англ. bootstrap): из заданной выборки \mathfrak{D}_m генерируется некоторое число B подвыборок размера k с возвращением. Для каждой подвыборки получается оценка параметров $\hat{\mathbf{w}}_k$ и вычисляется значение $L(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$. Для оценки математического ожидания и дисперсии используются выборочное среднее и несмещенная выборочная дисперсия соответственно. Количество подвыборок B выбирается достаточно большим (обычно $B \geq 1000$) для обеспечения точности оценок.

Предложенные выше определения также могут быть применены в тех задачах, где минимизируется произвольная функция потерь, а не максимизируется функция правдоподобия. В этом случае вместо функции правдоподобия $L(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$ используется функция потерь $\mathcal{L}(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$, а критерии достаточности формулируются аналогичным образом. Строгое теоретическое обоснование данного обобщения отсутствует, однако эмпирические результаты демонстрируют применимость таких методов на практике.

Метод, рассмотренный выше, основан на анализе стабильности функции правдоподобия при изменении объема данных. Альтернативный подход заключается в анализе близости апостериорных распределений параметров модели на близких подвыборках, что составляет содержание следующего раздела.

4.5. Метод определения достаточного размера выборки на основе близости апостериорных распределений

В настоящем разделе рассматривается метод определения достаточного размера выборки, основанный на анализе близости апостериорных распределений параметров модели. В работе [19] предлагается использовать расхождение Кульбака-Лейблера для оценки достаточного размера выборки в задаче бинарной классификации.

Идея метода основана на том, что если две подвыборки отличаются друг

от друга одним объектом, то полученные по ним апостериорные распределения должны быть близки. Эта близость определяется расхождением Кульбака-Лейблера.



Рис. 4.1: Визуализация сдвига апостериорных распределений параметров модели при последовательном добавлении объектов в выборку. Иллюстрация демонстрирует концепцию близости распределений, используемую в методах KL- и S-достаточности для определения достаточного размера выборки.

В рамках данного раздела предлагается использовать не только расхождение Кульбака-Лейблера, но и функцию схожести s-score из работы [67], для этого рассмотрим две подвыборки $\mathfrak{D}^1 \subseteq \mathfrak{D}_m$ и $\mathfrak{D}^2 \subseteq \mathfrak{D}_m$. Пусть $\mathcal{I}_1 \subseteq \mathcal{I} = \{1, \dots, m\}$ и $\mathcal{I}_2 \subseteq \mathcal{I} = \{1, \dots, m\}$ — соответствующие им подмножества индексов.

Определение 24 (Близкие подвыборки). *Подвыборки \mathfrak{D}^1 и \mathfrak{D}^2 назовем близкими, если \mathcal{I}_2 может быть получено из \mathcal{I}_1 путем удаления, замены или добавления одного элемента. Формально это условие записывается как*

$$|\mathcal{I}_1 \Delta \mathcal{I}_2| = |(\mathcal{I}_1 \setminus \mathcal{I}_2) \cup (\mathcal{I}_2 \setminus \mathcal{I}_1)| = 1,$$

где Δ обозначает симметрическую разность множеств.

Рассмотрим две близкие подвыборки $\mathfrak{D}_k = (\mathbf{X}_k, \mathbf{y}_k)$ и $\mathfrak{D}_{k+1} = (\mathbf{X}_{k+1}, \mathbf{y}_{k+1})$ размеров k и $k+1$ соответственно, где выборка \mathfrak{D}_{k+1} получена путем добавления одного элемента к выборке \mathfrak{D}_k .

Вычислим апостериорное распределение параметров модели по каждой из этих подвыборок:

$$p_j(\mathbf{w}) = p(\mathbf{w} | \mathfrak{D}_j) = \frac{p(\mathfrak{D}_j | \mathbf{w}) p(\mathbf{w})}{p(\mathfrak{D}_j)} \propto p(\mathfrak{D}_j | \mathbf{w}) p(\mathbf{w}), \quad j = k, k + 1.$$

Определение 25 (KL-достаточный размер выборки). *Пусть задано некоторое $\varepsilon > 0$. Размер выборки m^* называется KL-достаточным, если для всех $k \geq m^*$*

$$KL(k) = D_{KL}(p_k \| p_{k+1}) = \int p_k(\mathbf{w}) \log \frac{p_k(\mathbf{w})}{p_{k+1}(\mathbf{w})} d\mathbf{w} \leq \varepsilon.$$

Для пары нормальных распределений расхождение Кульбака-Лейблера имеет аналитическое выражение. Предположив, что апостериорное распределение является нормальным, $p_k(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_k, \Sigma_k)$, где \mathbf{m}_k — вектор математического ожидания, а Σ_k — ковариационная матрица, получаем формулировку теоремы 29

Теорема 29 (*Киселев-Грабовой*, Корректность KL-достаточного размера выборки). Пусть $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \rightarrow 0$ и $\|\Sigma_{k+1} - \Sigma_k\|_F \rightarrow 0$ при $k \rightarrow \infty$.

Тогда в модели с нормальным апостериорным распределением параметров определение KL-достаточного размера выборки корректно. А именно, для любого $\varepsilon > 0$ существует такой m^* , что для всех $k \geq m^*$ выполняется $KL(k) \leq \varepsilon$.

Доказательство. Рассмотрим выражение для расхождения Кульбака-Лейблера между двумя нормальными апостериорными распределениями $p_k = \mathcal{N}(\mathbf{m}_k, \Sigma_k)$ и $p_{k+1} = \mathcal{N}(\mathbf{m}_{k+1}, \Sigma_{k+1})$. Для двух многомерных нормальных распределений данная метрика имеет аналитическое выражение:

$$D_{\text{KL}}(p_k \| p_{k+1}) = \\ = \frac{1}{2} \left(\text{tr}(\Sigma_{k+1}^{-1} \Sigma_k) + (\mathbf{m}_{k+1} - \mathbf{m}_k)^\top \Sigma_{k+1}^{-1} (\mathbf{m}_{k+1} - \mathbf{m}_k) - n + \log \left(\frac{\det \Sigma_{k+1}}{\det \Sigma_k} \right) \right).$$

Для анализа поведения каждого слагаемого при $k \rightarrow \infty$ введем обозначение для разности ковариационных матриц: $\Delta\Sigma = \Sigma_{k+1} - \Sigma_k$, где по условию теоремы $\|\Delta\Sigma\|_F = \|\Sigma_{k+1} - \Sigma_k\|_F \rightarrow 0$. Первое слагаемое представляет собой след произведения матриц:

$$\text{tr}(\Sigma_{k+1}^{-1} \Sigma_k) = \text{tr}\left((\Sigma_k + \Delta\Sigma)^{-1} \Sigma_k\right).$$

Используя разложение в ряд для обратной матрицы при малых $\Delta\Sigma$, получаем:

$$(\Sigma_k + \Delta\Sigma)^{-1} = \Sigma_k^{-1} - \Sigma_k^{-1} \Delta\Sigma \Sigma_k^{-1} + O(\|\Delta\Sigma\|_F^2).$$

Тогда:

$$(\Sigma_k + \Delta\Sigma)^{-1} \Sigma_k = \mathbf{I}_n - \Sigma_k^{-1} \Delta\Sigma + O(\|\Delta\Sigma\|_F^2).$$

Взяв след от этого выражения и учитывая, что $\text{tr}(\mathbf{I}_n) = n$, а $\|\Sigma_k^{-1} \Delta\Sigma\|_F \rightarrow 0$ при $\|\Delta\Sigma\|_F \rightarrow 0$, получаем:

$$\text{tr}(\Sigma_{k+1}^{-1} \Sigma_k) \rightarrow n \quad \text{при} \quad \|\Delta\Sigma\|_F \rightarrow 0.$$

Второе слагаемое представляет собой квадратичную форму:

$$(\mathbf{m}_{k+1} - \mathbf{m}_k)^\top \Sigma_{k+1}^{-1} (\mathbf{m}_{k+1} - \mathbf{m}_k).$$

По условию теоремы $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \rightarrow 0$. Квадратичная форма оценивается сверху следующим образом:

$$|(\mathbf{m}_{k+1} - \mathbf{m}_k)^\top \Sigma_{k+1}^{-1} (\mathbf{m}_{k+1} - \mathbf{m}_k)| \leq \|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2^2 \cdot \|\Sigma_{k+1}^{-1}\|_2.$$

Поскольку ковариационная матрица Σ_{k+1} является положительно определенной и сходится к некоторой предельной матрице, ее спектральная норма $\|\Sigma_{k+1}^{-1}\|_2$ ограничена. Следовательно, при $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \rightarrow 0$ данное слагаемое стремится к нулю. Третье и четвертое слагаемые составляют:

$$-n + \log \left(\frac{\det \Sigma_{k+1}}{\det \Sigma_k} \right) = \log \left(\frac{\det \Sigma_{k+1}}{\det \Sigma_k} \right) - n.$$

Преобразуем отношение определителей:

$$\frac{\det \Sigma_{k+1}}{\det \Sigma_k} = \frac{\det(\Sigma_k + \Delta \Sigma)}{\det \Sigma_k} = \det(\mathbf{I}_n + \Sigma_k^{-1} \Delta \Sigma).$$

Для малых $\Delta \Sigma$ используем приближение:

$$\det(\mathbf{I}_n + \Sigma_k^{-1} \Delta \Sigma) = 1 + \text{tr}(\Sigma_k^{-1} \Delta \Sigma) + O(\|\Delta \Sigma\|_F^2).$$

Тогда:

$$\log \left(\frac{\det \Sigma_{k+1}}{\det \Sigma_k} \right) = \log \det(\mathbf{I}_n + \Sigma_k^{-1} \Delta \Sigma) = \text{tr}(\Sigma_k^{-1} \Delta \Sigma) + O(\|\Delta \Sigma\|_F^2).$$

Поскольку $\|\Delta \Sigma\|_F \rightarrow 0$, то $\text{tr}(\Sigma_k^{-1} \Delta \Sigma) \rightarrow 0$, и следовательно:

$$\log \left(\frac{\det \Sigma_{k+1}}{\det \Sigma_k} \right) \rightarrow 0.$$

Таким образом, все четыре слагаемых в выражении для D_{KL} сходятся к своим пределам при $k \rightarrow \infty$: первое слагаемое стремится к n , второе — к 0 , третье равно $-n$, четвертое — к 0 . Сумма этих пределов равна $n + 0 - n + 0 = 0$, что доказывает, что $D_{KL}(p_k \| p_{k+1}) \rightarrow 0$ при $k \rightarrow \infty$. Следовательно, для любого $\varepsilon > 0$ существует такой m^* , что для всех $k \geq m^*$ выполняется $KL(k) \leq \varepsilon$, что и требовалось доказать. \square

Теорема 29 устанавливает, что расхождение Кульбака-Лейблера между двумя нормальными апостериорными распределениями стремится к нулю по мере сходимости их векторов математических ожиданий и ковариационных матриц. Это позволяет использовать KL-дивергенцию в качестве критерия достаточности размера выборки, анализируя аналитические выражения для моментов апостериорных распределений.

Рассмотрим функцию схожести s-score из работы [67] в качестве меры близости распределений по аналогии, как это было с KL-дивергенцией:

$$\text{s-score}(g_1, g_2) = \frac{\int_{\mathbf{w}} g_1(\mathbf{w})g_2(\mathbf{w})d\mathbf{w}}{\max_{\mathbf{b}} \int_{\mathbf{w}} g_1(\mathbf{w} - \mathbf{b})g_2(\mathbf{w})d\mathbf{w}}.$$

Определение 26 (S-достаточный размер выборки). *Пусть задано некоторое $\varepsilon > 0$. Размер выборки m^* называется S-достаточным, если для всех $k \geq m^*$*

$$S(k) = \text{s-score}(p_k, p_{k+1}) \geq 1 - \varepsilon.$$

Как и в случае с KL-достаточным размером выборки, в модели с нормальным апостериорным распределением можно записать выражение для используемого критерия, который записан в виде теоремы 30

Теорема 30 (Киселев-Грабовой). Корректность S-достаточного размера выборки). *Пусть $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \rightarrow 0$ при $k \rightarrow \infty$.*

Тогда в модели с нормальным апостериорным распределением параметров определение S-достаточного размера выборки корректно. А именно, для любого $\varepsilon > 0$ существует такой m^ , что для всех $k \geq m^*$ выполняется $S(k) \geq 1 - \varepsilon$.*

Доказательство. Используем выражение для s-score пары нормальных апостериорных распределений из работы [67]:

$$\text{s-score}(p_k, p_{k+1}) = \exp \left(-\frac{1}{2} (\mathbf{m}_{k+1} - \mathbf{m}_k)^{\top} (\boldsymbol{\Sigma}_k + \boldsymbol{\Sigma}_{k+1})^{-1} (\mathbf{m}_{k+1} - \mathbf{m}_k) \right).$$

Оценим квадратичную форму в показателе экспоненты:

$$\left| (\mathbf{m}_{k+1} - \mathbf{m}_k)^{\top} (\boldsymbol{\Sigma}_k + \boldsymbol{\Sigma}_{k+1})^{-1} (\mathbf{m}_{k+1} - \mathbf{m}_k) \right| \leq \|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2^2 \|(\boldsymbol{\Sigma}_k + \boldsymbol{\Sigma}_{k+1})^{-1}\|_2.$$

Поскольку ковариационные матрицы $\boldsymbol{\Sigma}_k$ и $\boldsymbol{\Sigma}_{k+1}$ являются положительно определенными и сходятся к некоторой предельной матрице, спектральная норма $\|(\boldsymbol{\Sigma}_k + \boldsymbol{\Sigma}_{k+1})^{-1}\|_2$ ограничена. При условии $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \rightarrow 0$ значение квадратичной формы в показателе экспоненты стремится к нулю. Следовательно, $\text{s-score}(p_k, p_{k+1}) \rightarrow 1$ при $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \rightarrow 0$, что и требовалось доказать. \square

Теорема 30 устанавливает корректность определения S-достаточного размера выборки при более слабых условиях, чем теорема 29. В отличие от KL-дивергенции, для сходимости s-score к единице требуется только сходимость математических ожиданий апостериорных распределений, что делает данный критерий менее консервативным и более применимым на практике.

Пусть в модели линейной регрессии задано нормальное априорное распределение параметров. В силу свойства сопряженности априорного распределения и правдоподобия, апостериорное распределение также будет нормальным. Таким образом, приходим к одному из простейших примеров модели, для которой справедливы приведенные выше теоремы. Фактически, для линейной регрессии могут быть сформулированы более простые утверждения.

Теорема 31 (*Киселев-Грабовой*, Сходимость апостериорных распределений в линейной регрессии). *Пусть множества значений признаков и целевой переменной ограничены, то есть существует константа $M \in \mathbb{R}$ такая, что $\|\mathbf{x}\|_2 \leq M$ и $|y| \leq M$ для всех объектов выборки. Если $\lambda_{\min}(\mathbf{X}_k^\top \mathbf{X}_k) = \omega(\sqrt{k})$ при $k \rightarrow \infty$, то в модели линейной регрессии с нормальным априорным распределением параметров $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \rightarrow 0$ и $\|\Sigma_{k+1} - \Sigma_k\|_F \rightarrow 0$ при $k \rightarrow \infty$.*

Доказательство. Рассмотрим линейную регрессионную модель с нормальным априорным распределением параметров: $p(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} \mathbf{I})$. Данное априорное распределение является сопряженным для нормального правдоподобия, что существенно упрощает анализ. Нормальное правдоподобие задается в виде:

$$p(\mathbf{y} | \mathbf{X}, \mathbf{w}) = \mathcal{N}(\mathbf{y} | \mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I}) = (2\pi\sigma^2)^{-m/2} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2\right).$$

Благодаря свойству сопряженности нормального априорного распределения и нормального правдоподобия, апостериорное распределение также является нормальным:

$$p(\mathbf{w} | \mathbf{X}, \mathbf{y}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \Sigma),$$

где параметры распределения имеют аналитическое выражение:

$$\begin{aligned} \Sigma &= \left(\alpha \mathbf{I} + \frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X} \right)^{-1}, \\ \mathbf{m} &= \frac{1}{\sigma^2} \Sigma \mathbf{X}^\top \mathbf{y} = (\mathbf{X}^\top \mathbf{X} + \alpha \sigma^2 \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}. \end{aligned}$$

Перейдем к анализу сходимости ковариационных матриц. Рассмотрим подвыборки размера k и $k+1$, полученные из исходных данных. Нас интересует поведение разности $\|\Sigma_{k+1} - \Sigma_k\|_2$ при $k \rightarrow \infty$. Введем обозначение $\mathbf{A}_k = \frac{1}{\sigma^2} \mathbf{X}_k^\top \mathbf{X}_k$

для нормированной матрицы ковариации признаков. Тогда разность обратных матриц можно преобразовать, используя матричное тождество:

$$\|\Sigma_{k+1} - \Sigma_k\|_2 = \left\| (\alpha\mathbf{I} + \mathbf{A}_{k+1})^{-1} - (\alpha\mathbf{I} + \mathbf{A}_k)^{-1} \right\|_2.$$

Применяя матричное тождество для разности обратных матриц, получаем:

$$(\alpha\mathbf{I} + \mathbf{A}_{k+1})^{-1} - (\alpha\mathbf{I} + \mathbf{A}_k)^{-1} = (\alpha\mathbf{I} + \mathbf{A}_{k+1})^{-1} (\mathbf{A}_k - \mathbf{A}_{k+1}) (\alpha\mathbf{I} + \mathbf{A}_k)^{-1}.$$

Используя субмультипликативное свойство спектральной нормы, оцениваем:

$$\|\Sigma_{k+1} - \Sigma_k\|_2 \leq \left\| (\alpha\mathbf{I} + \mathbf{A}_{k+1})^{-1} \right\|_2 \left\| (\alpha\mathbf{I} + \mathbf{A}_k)^{-1} \right\|_2 \|\mathbf{A}_{k+1} - \mathbf{A}_k\|_2.$$

Проанализируем каждый из множителей в полученной оценке. Спектральная норма обратной матрицы выражается через минимальное собственное значение исходной матрицы:

$$\left\| (\alpha\mathbf{I} + \mathbf{A})^{-1} \right\|_2 = \frac{1}{\lambda_{\min}(\alpha\mathbf{I} + \mathbf{A})}.$$

Поскольку $\alpha > 0$ и матрица \mathbf{A} положительно полуопределенна, имеем $\lambda_{\min}(\alpha\mathbf{I} + \mathbf{A}) \geq \alpha + \lambda_{\min}(\mathbf{A})$. Однако для получения точной асимптотики используем более слабую оценку:

$$\left\| (\alpha\mathbf{I} + \mathbf{A})^{-1} \right\|_2 \leq \frac{1}{\lambda_{\min}(\mathbf{A})}.$$

Объединяя полученные оценки, получаем цепочку неравенств:

$$\begin{aligned} \|\Sigma_{k+1} - \Sigma_k\|_2 &\leq \frac{1}{\lambda_{\min}(\mathbf{A}_{k+1})} \frac{1}{\lambda_{\min}(\mathbf{A}_k)} \|\mathbf{A}_{k+1} - \mathbf{A}_k\|_2 = \\ &= \sigma^2 \frac{1}{\lambda_{\min}(\mathbf{X}_{k+1}^\top \mathbf{X}_{k+1})} \frac{1}{\lambda_{\min}(\mathbf{X}_k^\top \mathbf{X}_k)} \|\mathbf{X}_{k+1}^\top \mathbf{X}_{k+1} - \mathbf{X}_k^\top \mathbf{X}_k\|_2. \end{aligned}$$

Поскольку выборка \mathfrak{D}_{k+1} получается из \mathfrak{D}_k добавлением одного наблюдения, имеем:

$$\|\mathbf{X}_{k+1}^\top \mathbf{X}_{k+1} - \mathbf{X}_k^\top \mathbf{X}_k\|_2 = \left\| \sum_{i=1}^{k+1} \mathbf{x}_i \mathbf{x}_i^\top - \sum_{i=1}^k \mathbf{x}_i \mathbf{x}_i^\top \right\|_2 = \|\mathbf{x}_{k+1} \mathbf{x}_{k+1}^\top\|_2.$$

Матрица $\mathbf{x}_{k+1} \mathbf{x}_{k+1}^\top$ является матрицей ранга 1, и ее спектральная норма равна квадрату евклидовой нормы вектора \mathbf{x}_{k+1} :

$$\|\mathbf{x}_{k+1} \mathbf{x}_{k+1}^\top\|_2 = \lambda_{\max}(\mathbf{x}_{k+1} \mathbf{x}_{k+1}^\top) = \|\mathbf{x}_{k+1}\|_2^2.$$

Из условия ограниченности признаков следует $\|\mathbf{x}_{k+1}\|_2^2 \leq M^2$. Таким образом:

$$\|\mathbf{X}_{k+1}^\top \mathbf{X}_{k+1} - \mathbf{X}_k^\top \mathbf{X}_k\|_2 \leq M^2.$$

Теперь рассмотрим условие на минимальное собственное значение. Из предположения $\lambda_{\min}(\mathbf{X}_k^\top \mathbf{X}_k) = \omega(\sqrt{k})$ следует, что:

$$\frac{1}{\lambda_{\min}(\mathbf{X}_k^\top \mathbf{X}_k)} = o\left(\frac{1}{\sqrt{k}}\right).$$

Комбинируя полученные оценки, приходим к:

$$\|\Sigma_{k+1} - \Sigma_k\|_2 \leq \sigma^2 M^2 \cdot o\left(\frac{1}{\sqrt{k}}\right) \cdot o\left(\frac{1}{\sqrt{k}}\right) = o\left(\frac{1}{k}\right).$$

Для перехода к норме Фробениуса воспользуемся неравенством $\|\mathbf{A}\|_F \leq \sqrt{n}\|\mathbf{A}\|_2$, где n — размерность пространства параметров:

$$\|\Sigma_{k+1} - \Sigma_k\|_F \leq \sqrt{n}\|\Sigma_{k+1} - \Sigma_k\|_2 = \sqrt{n} \cdot o\left(\frac{1}{k}\right) = o\left(\frac{1}{k}\right).$$

Теперь перейдем к анализу сходимости математических ожиданий. Требуется оценить:

$$\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 = \left\| (\mathbf{X}_{k+1}^\top \mathbf{X}_{k+1} + \alpha\sigma^2 \mathbf{I})^{-1} \mathbf{X}_{k+1}^\top \mathbf{y}_{k+1} - (\mathbf{X}_k^\top \mathbf{X}_k + \alpha\sigma^2 \mathbf{I})^{-1} \mathbf{X}_k^\top \mathbf{y}_k \right\|_2.$$

Представим расширенную матрицу признаков и вектор ответов через предыдущие значения:

$$\mathbf{X}_{k+1} = \begin{bmatrix} \mathbf{X}_k \\ \mathbf{x}_{k+1}^\top \end{bmatrix}, \quad \mathbf{y}_{k+1} = \begin{bmatrix} \mathbf{y}_k \\ y_{k+1} \end{bmatrix}.$$

Тогда матричные произведения принимают вид:

$$\begin{aligned} \mathbf{X}_{k+1}^\top \mathbf{X}_{k+1} &= \mathbf{X}_k^\top \mathbf{X}_k + \mathbf{x}_{k+1} \mathbf{x}_{k+1}^\top, \\ \mathbf{X}_{k+1}^\top \mathbf{y}_{k+1} &= \mathbf{X}_k^\top \mathbf{y}_k + \mathbf{x}_{k+1} y_{k+1}. \end{aligned}$$

Подставляя эти выражения, получаем:

$$\begin{aligned} \|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 &= \left\| (\mathbf{X}_k^\top \mathbf{X}_k + \alpha\sigma^2 \mathbf{I} + \mathbf{x}_{k+1} \mathbf{x}_{k+1}^\top)^{-1} (\mathbf{X}_k^\top \mathbf{y}_k + \mathbf{x}_{k+1} y_{k+1}) \right. \\ &\quad \left. - (\mathbf{X}_k^\top \mathbf{X}_k + \alpha\sigma^2 \mathbf{I})^{-1} \mathbf{X}_k^\top \mathbf{y}_k \right\|_2. \end{aligned}$$

Для упрощения первого слагаемого применим лемму о матричном обращении:

$$\begin{aligned} & (\mathbf{X}_k^\top \mathbf{X}_k + \alpha\sigma^2 \mathbf{I} + \mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top)^{-1} = \\ &= (\mathbf{X}_k^\top \mathbf{X}_k + \alpha\sigma^2 \mathbf{I})^{-1} - \frac{(\mathbf{X}_k^\top \mathbf{X}_k + \alpha\sigma^2 \mathbf{I})^{-1} \mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top (\mathbf{X}_k^\top \mathbf{X}_k + \alpha\sigma^2 \mathbf{I})^{-1}}{1 + \mathbf{x}_{k+1}^\top (\mathbf{X}_k^\top \mathbf{X}_k + \alpha\sigma^2 \mathbf{I})^{-1} \mathbf{x}_{k+1}}. \end{aligned}$$

После алгебраических преобразований получаем:

$$\begin{aligned} & \|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 = \\ &= \left\| \left[\left(\mathbf{I} + (\mathbf{X}_k^\top \mathbf{X}_k + \alpha\sigma^2 \mathbf{I})^{-1} \mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top \right)^{-1} - \mathbf{I} \right] (\mathbf{X}_k^\top \mathbf{X}_k + \alpha\sigma^2 \mathbf{I})^{-1} \mathbf{X}_k^\top \mathbf{y}_k \right. \\ & \quad \left. + (\mathbf{X}_{k+1}^\top \mathbf{X}_{k+1} + \alpha\sigma^2 \mathbf{I})^{-1} \mathbf{x}_{k+1} y_{k+1} \right\|_2. \end{aligned}$$

Применяя неравенство треугольника и свойства норм, оцениваем каждый член отдельно:

$$\begin{aligned} & \|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \leqslant \\ & \leqslant \left\| \left(\mathbf{I} + (\mathbf{X}_k^\top \mathbf{X}_k + \alpha\sigma^2 \mathbf{I})^{-1} \mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top \right)^{-1} - \mathbf{I} \right\|_2 \|(\mathbf{X}_k^\top \mathbf{X}_k + \alpha\sigma^2 \mathbf{I})^{-1}\|_2 \|\mathbf{X}_k^\top \mathbf{y}_k\|_2 \\ & \quad + \left\| (\mathbf{X}_{k+1}^\top \mathbf{X}_{k+1} + \alpha\sigma^2 \mathbf{I})^{-1} \right\|_2 \|\mathbf{x}_{k+1} y_{k+1}\|_2. \end{aligned}$$

Проанализируем первый множитель в первом слагаемом. Используя матричное тождество, получаем:

$$\begin{aligned} & \left\| \left(\mathbf{I} + (\mathbf{X}_k^\top \mathbf{X}_k + \alpha\sigma^2 \mathbf{I})^{-1} \mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top \right)^{-1} - \mathbf{I} \right\|_2 = \\ &= \left\| - \left(\mathbf{I} + (\mathbf{X}_k^\top \mathbf{X}_k + \alpha\sigma^2 \mathbf{I})^{-1} \mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top \right)^{-1} (\mathbf{X}_k^\top \mathbf{X}_k + \alpha\sigma^2 \mathbf{I})^{-1} \mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top \right\|_2. \end{aligned}$$

Применяя субмультипликативность нормы и учитывая, что спектральная норма произведения матриц не превышает произведения их норм, получаем оценку:

$$\begin{aligned} & \left\| \left(\mathbf{I} + (\mathbf{X}_k^\top \mathbf{X}_k + \alpha\sigma^2 \mathbf{I})^{-1} \mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top \right)^{-1} - \mathbf{I} \right\|_2 \leqslant \\ & \leqslant \left\| \left(\mathbf{I} + (\mathbf{X}_k^\top \mathbf{X}_k + \alpha\sigma^2 \mathbf{I})^{-1} \mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top \right)^{-1} \right\|_2 \|(\mathbf{X}_k^\top \mathbf{X}_k + \alpha\sigma^2 \mathbf{I})^{-1}\|_2 \|\mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top\|_2. \end{aligned}$$

Оценим каждый из этих множителей. Для первого множителя используем тот факт, что матрица $(\mathbf{X}_k^\top \mathbf{X}_k + \alpha\sigma^2 \mathbf{I})^{-1} \mathbf{x}_{k+1}\mathbf{x}_{k+1}^\top$ имеет единичный ранг, и ее

максимальное собственное значение равно:

$$\lambda_{\max} \left((\mathbf{X}_k^\top \mathbf{X}_k + \alpha \sigma^2 \mathbf{I})^{-1} \mathbf{x}_{k+1} \mathbf{x}_{k+1}^\top \right) = \mathbf{x}_{k+1}^\top (\mathbf{X}_k^\top \mathbf{X}_k + \alpha \sigma^2 \mathbf{I})^{-1} \mathbf{x}_{k+1}.$$

Тогда спектральная норма обратной матрицы оценивается как:

$$\left\| \left(\mathbf{I} + (\mathbf{X}_k^\top \mathbf{X}_k + \alpha \sigma^2 \mathbf{I})^{-1} \mathbf{x}_{k+1} \mathbf{x}_{k+1}^\top \right)^{-1} \right\|_2 \leq 1.$$

Второй множитель оценивается через минимальное собственное значение:

$$\left\| (\mathbf{X}_k^\top \mathbf{X}_k + \alpha \sigma^2 \mathbf{I})^{-1} \right\|_2 \leq \frac{1}{\lambda_{\min}(\mathbf{X}_k^\top \mathbf{X}_k)}.$$

Третий множитель, как уже было установлено, равен $\|\mathbf{x}_{k+1}\|_2^2 \leq M^2$. Таким образом, получаем оценку для первого множителя:

$$\left\| \left(\mathbf{I} + (\mathbf{X}_k^\top \mathbf{X}_k + \alpha \sigma^2 \mathbf{I})^{-1} \mathbf{x}_{k+1} \mathbf{x}_{k+1}^\top \right)^{-1} - \mathbf{I} \right\|_2 \leq \frac{M^2}{\lambda_{\min}(\mathbf{X}_k^\top \mathbf{X}_k)}.$$

Теперь оценим второй множитель первого слагаемого вместе с третьим множителем:

$$\left\| (\mathbf{X}_k^\top \mathbf{X}_k + \alpha \sigma^2 \mathbf{I})^{-1} \right\|_2 \|\mathbf{X}_k^\top \mathbf{y}_k\|_2 \leq \frac{\|\mathbf{X}_k^\top \mathbf{y}_k\|_2}{\lambda_{\min}(\mathbf{X}_k^\top \mathbf{X}_k)}.$$

Норма $\|\mathbf{X}_k^\top \mathbf{y}_k\|_2$ оценивается с использованием условия ограниченности:

$$\|\mathbf{X}_k^\top \mathbf{y}_k\|_2 = \left\| \sum_{i=1}^k \mathbf{x}_i y_i \right\|_2 \leq \sum_{i=1}^k \|\mathbf{x}_i y_i\|_2 \leq kM^2.$$

Следовательно:

$$\left\| (\mathbf{X}_k^\top \mathbf{X}_k + \alpha \sigma^2 \mathbf{I})^{-1} \right\|_2 \|\mathbf{X}_k^\top \mathbf{y}_k\|_2 \leq \frac{kM^2}{\lambda_{\min}(\mathbf{X}_k^\top \mathbf{X}_k)}.$$

Теперь рассмотрим второе слагаемое:

$$\left\| (\mathbf{X}_{k+1}^\top \mathbf{X}_{k+1} + \alpha \sigma^2 \mathbf{I})^{-1} \right\|_2 \|\mathbf{x}_{k+1} y_{k+1}\|_2 \leq \frac{M^2}{\lambda_{\min}(\mathbf{X}_{k+1}^\top \mathbf{X}_{k+1})}.$$

Комбинируя все полученные оценки, приходим к итоговой оценке:

$$\begin{aligned} \|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 &\leq \frac{M^2}{\lambda_{\min}(\mathbf{X}_k^\top \mathbf{X}_k)} \cdot \frac{kM^2}{\lambda_{\min}(\mathbf{X}_k^\top \mathbf{X}_k)} + \frac{M^2}{\lambda_{\min}(\mathbf{X}_{k+1}^\top \mathbf{X}_{k+1})} \\ &= \frac{kM^4}{\lambda_{\min}^2(\mathbf{X}_k^\top \mathbf{X}_k)} + \frac{M^2}{\lambda_{\min}(\mathbf{X}_{k+1}^\top \mathbf{X}_{k+1})}. \end{aligned}$$

Из условия $\lambda_{\min}(\mathbf{X}_k^\top \mathbf{X}_k) = \omega(\sqrt{k})$ следует:

$$\frac{1}{\lambda_{\min}(\mathbf{X}_k^\top \mathbf{X}_k)} = o\left(\frac{1}{\sqrt{k}}\right),$$

$$\frac{1}{\lambda_{\min}^2(\mathbf{X}_k^\top \mathbf{X}_k)} = o\left(\frac{1}{k}\right).$$

Поэтому первое слагаемое оценивается как $k \cdot o\left(\frac{1}{k}\right) = o(1)$, а второе слагаемое как $o\left(\frac{1}{\sqrt{k}}\right) = o(1)$. Таким образом:

$$\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 = o(1) \quad \text{при } k \rightarrow \infty.$$

Это завершает доказательство сходимости как ковариационных матриц, так и математических ожиданий апостериорного распределения параметров. \square

Теорема 31 является ключевой в настоящем разделе, так как при слабых и понятных предположениях из нее следует сходимость моментов апостериорного распределения параметров. Первое предположение в теореме 31 касается ограничения на область значений признаков и целевой переменной. Это условие обычно выполняется в практических приложениях, поэтому оно служит в первую очередь для целей теоретического анализа. Второе условие теоремы 31 представляет больший интерес, поскольку оно углубляется в поведение минимального собственного значения выборочной ковариационной матрицы признаков. Следует отметить, что в рамках настоящей работы не приводятся строгие теоретические гарантии для данной сходимости, однако эмпирические результаты подтверждают выполнение указанного условия.

4.6. Результаты вычислительных экспериментов

В настоящем разделе представлены результаты вычислительных экспериментов для методов определения достаточного размера выборки, описанных в предыдущих разделах главы. Эксперименты направлены на валидацию теоретических результатов и сравнение эффективности различных подходов.

4.6.1. Определения достаточного размера выборки на основе статистических методов

Проводится эксперимент для анализа свойств методов оценки достаточного размера выборки. Эксперимент состоит из трех частей.

Таблица 4.1: Характеристики выборок, используемых для анализа качества методов определения достаточного размера выборки. Таблица содержит информацию о типе задачи, количестве признаков и общем размере выборки для каждого набора данных.

Выборка	Задача	Число признаков	Размер выборки
Boston Housing	regression	14	506
Diabets	regression	20	576
Forest Fires	regression	13	517
Servo	regression	4	167
NBA	classification	12	2235

В первой части рассматриваются оценки достаточного размера выборки для различных наборов данных с фиксированным набором гиперпараметров различных методов. В качестве данных использовались выборки, описанные в таблице 4.1. Результаты представлены в таблице 4.2, где показаны оценки размера выборки для соответствующих выборок.

Во второй части исследуется зависимость достаточного размера выборки от имеющегося размера выборки. В третьей части исследуется поведение методов в зависимости от изменения гиперпараметров методов.

В данной части вычислительного эксперимента анализируется сходимость различных методов на различных выборках. В эксперименте используются выборки: Boston Housing [68], Diabetes, Forest Fires, Servo [69], NBA. Результат анализа представлен в таблице 4.2. Символ “–” обозначает, что исходный размер выборки недостаточный для прогноза.

Гиперпараметры каждого метода для всех выборок описаны в таблице 4.3. Поскольку критерии Лагранжа, отношения правдоподобия и Вальда асимптотически эквивалентны, то параметры этих методов задавались одинаково. Параметры методов «Average Coverage» и «Average Length» также задаются одинаково.

Вычислительный эксперимент проводился для анализа описанных методов. Выбирается некоторый размер выборки t и методом бутстрепа семплируется множество подвыборок размером t . Для разных значений t вычисляется t^* .

На рис. 4.2 демонстрируется зависимость статистических показателей каждого метода для разных выборок с фиксированным размером выборки t . Поро-

Таблица 4.2: Сравнение оценок достаточного размера выборки, полученных различными статистическими и байесовскими методами для пяти наборов данных. Результаты демонстрируют значительный разброс оценок между методами, что указывает на различную консервативность подходов.

Методы	Boston	Diabetes	Forest Fires	Servo	NBA
Lagrange Multipliers Test	18	25	44	38	218
Likelihood Ratio Test	17	25	43	18	110
Wald Test	66	51	46	76	200
Cross Validation	178	441	172	120	—
Bootstrap	113	117	86	60	405
APVC	98	167	351	20	—
ACC	228	441	346	65	—
ALC	98	267	516	25	—
Utility Function	148	172	206	105	925

Таблица 4.3: Гиперпараметры методов оценки достаточного размера выборки, установленные экспертизно для экспериментов. Параметры включают уровни значимости α , вероятности ошибки второго рода β , пороговые значения ε и l , а также параметры обобщенных линейных моделей.

Method	GLM parameters	l	ε	α	β
Lagrange Multipliers Test	\mathbf{w}_u^0	—	0.2	0.05	0.2
Likelihood Ratio Test	\mathbf{w}_u^0	—	0.2	0.05	0.2
Wald Test	\mathbf{w}_u^0	—	0.2	0.05	0.2
Cross Validation	—	—	0.05	—	—
Bootstrap	—	0.5	—	0.05	—
APVC	—	0.5	—	—	—
ACC	—	0.25	—	0.05	—
ALC	—	0.5	—	0.05	—
Utility function	—	—	0.005	—	—

говые значения для каждого метода устанавливаются экспертизно, что позволяет контролировать различные статистические характеристики выборки.

Представленные функции являются монотонными и асимптотически стре-

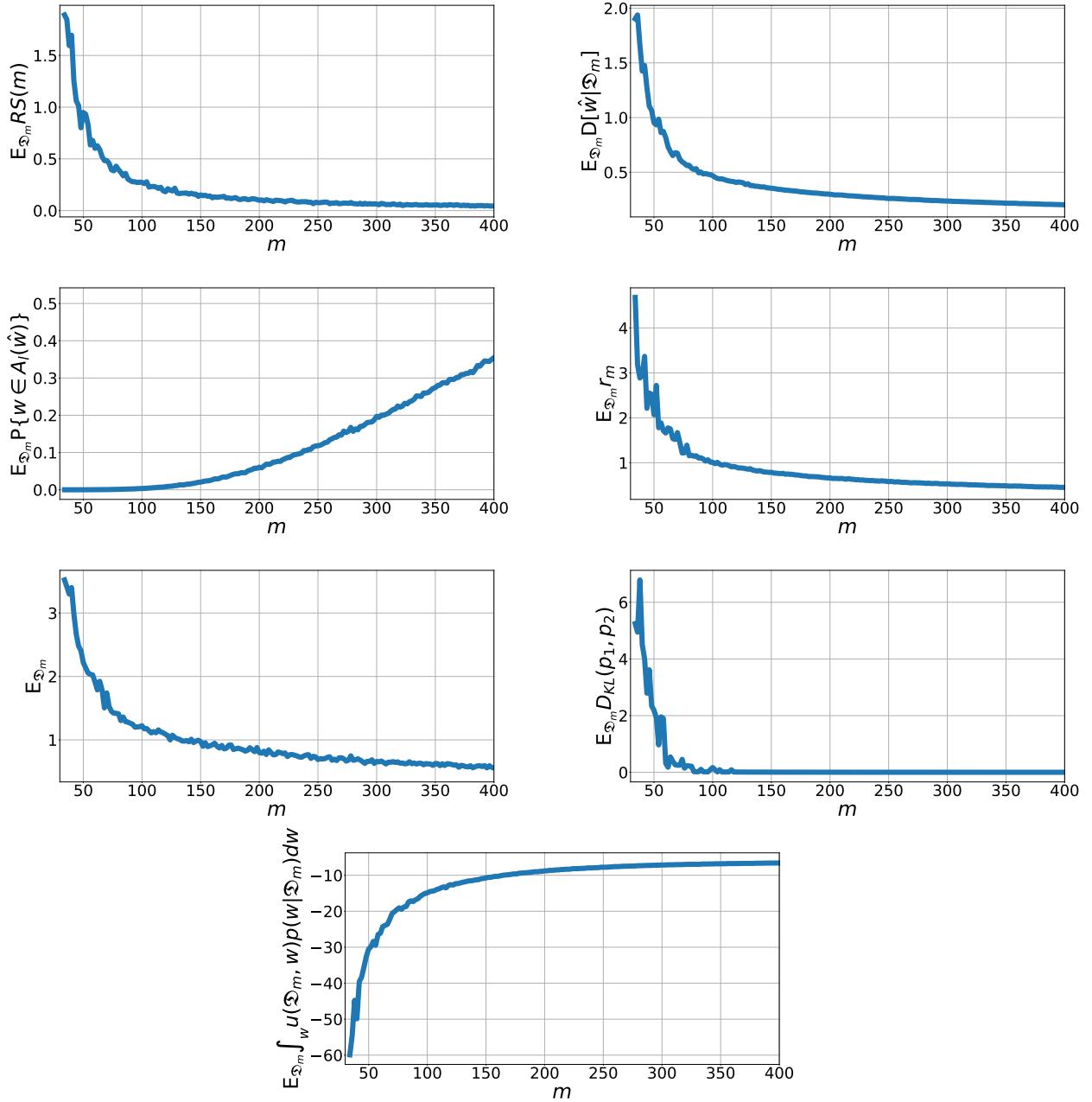


Рис. 4.2: Зависимость статистических значений различных методов определения достаточного размера выборки от размера подвыборки для наборов данных Boston Housing, Diabetes, Forest Fires, Servo и NBA. Все представленные функции монотонны и асимптотически стремятся к константе, что подтверждает корректность методов.

мятся к константе, что подтверждает корректность различных методов определения достаточного размера выборки.

На рис. 4.3 показаны результаты методов на выборках различного размера. Наблюдается различие методов в дисперсии вычисленного t^* . Все представленные методы демонстрируют сходимость, причем результат предсказания в

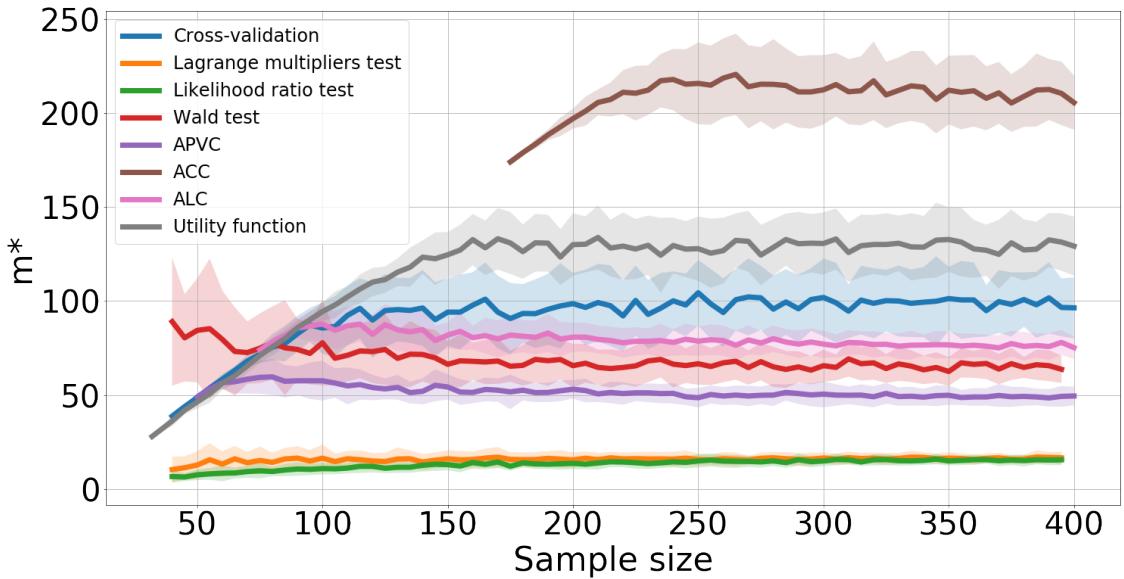


Рис. 4.3: Зависимость оцененного достаточного размера выборки t^* от доступного размера выборки t для различных методов на наборе данных Boston Housing. Результаты демонстрируют сходимость методов и низкую дисперсию оценок, что указывает на вычислительную устойчивость рассмотренных подходов.

асимптотике не зависит от доступного размера выборки t .

Небольшое значение дисперсии интерпретируется как вычислительная устойчивость рассмотренных методов.

Показано, что некоторые методы не дают оценку достаточного размера выборки, если доступный размер выборки недостаточен для применения метода. Это означает, что указанные методы не эффективны с точки зрения прогнозирования необходимого объема данных на ранних этапах эксперимента, однако могут быть использованы для ретроспективного анализа уже проведенных экспериментов.

Анализируется оценка достаточного размера выборки в зависимости от гиперпараметров для байесовских методов, а также эвристических методов. Для анализа рассмотрена выборка Boston Housing.

Байесовские методы используют решающее правило над скалярной функцией для определения достаточного размера выборки. На рис. 4.2 показана зависимость скалярных функций от размера подвыборки. Наблюдается, что указанные функции являются монотонными. Характер поведения функции определяется выбранным методом. Изменение ограничений, установленных экспертом, позволяет варьировать размер выборки, соответствующий заданным ограничениям.

ниям.

4.6.2. Определение достаточного размера выборки на основе сэмплирования эмпирической функции ошибки

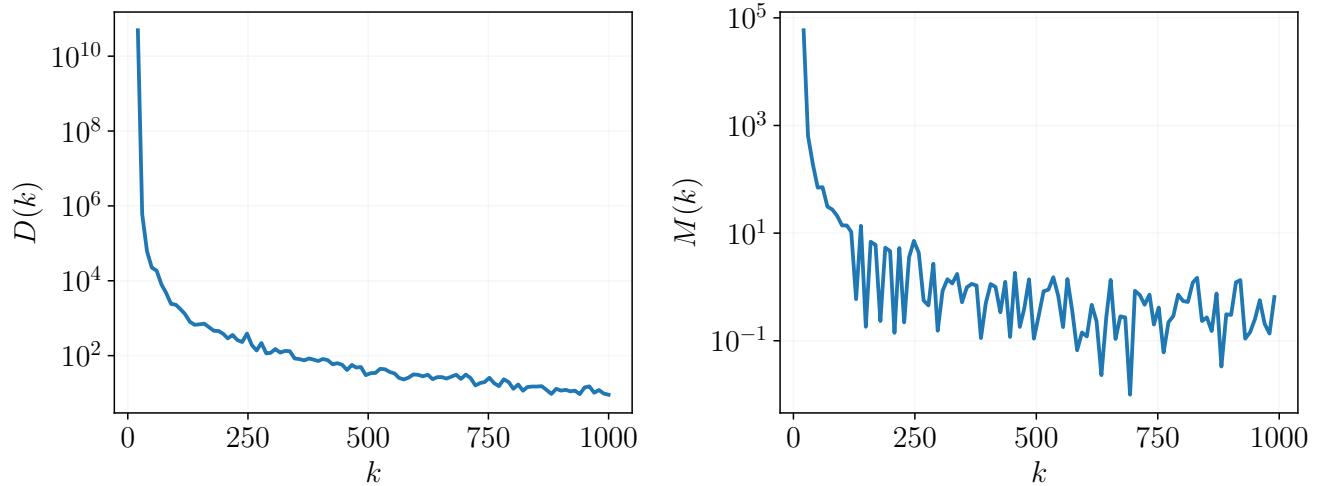


Рис. 4.4: Сходимость функций $D(k)$ и $M(k)$ для синтетического набора данных регрессии. Обе функции стремятся к нулю с увеличением размера выборки, что подтверждает теорему 28.

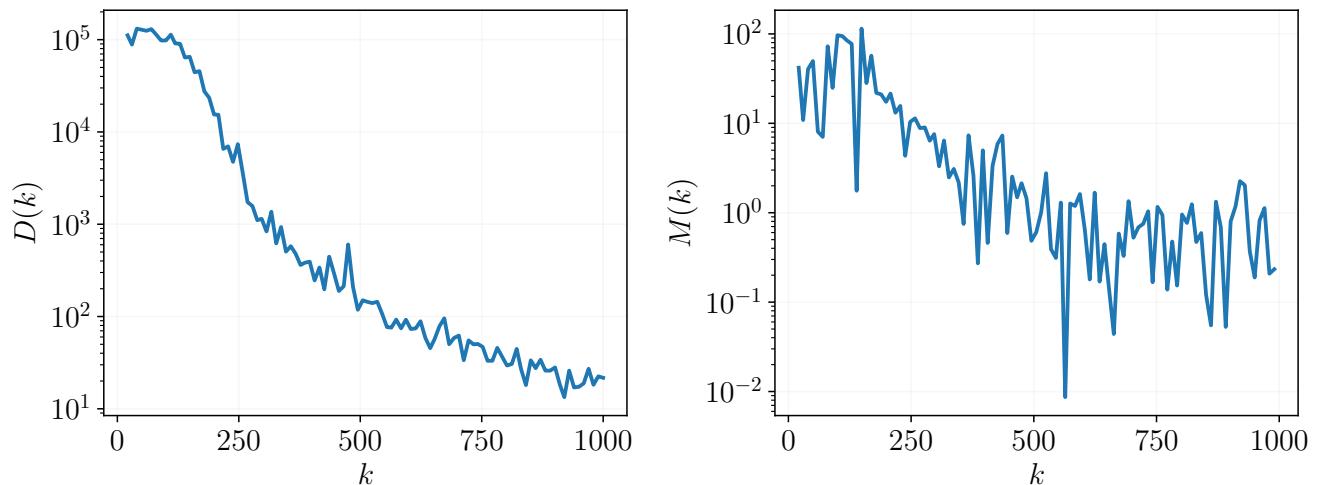


Рис. 4.5: Сходимость функций $D(k)$ и $M(k)$ для синтетического набора данных классификации. Обе функции демонстрируют монотонное убывание к нулю с увеличением размера выборки, подтверждая применимость методов D- и M-достаточности для задач классификации.

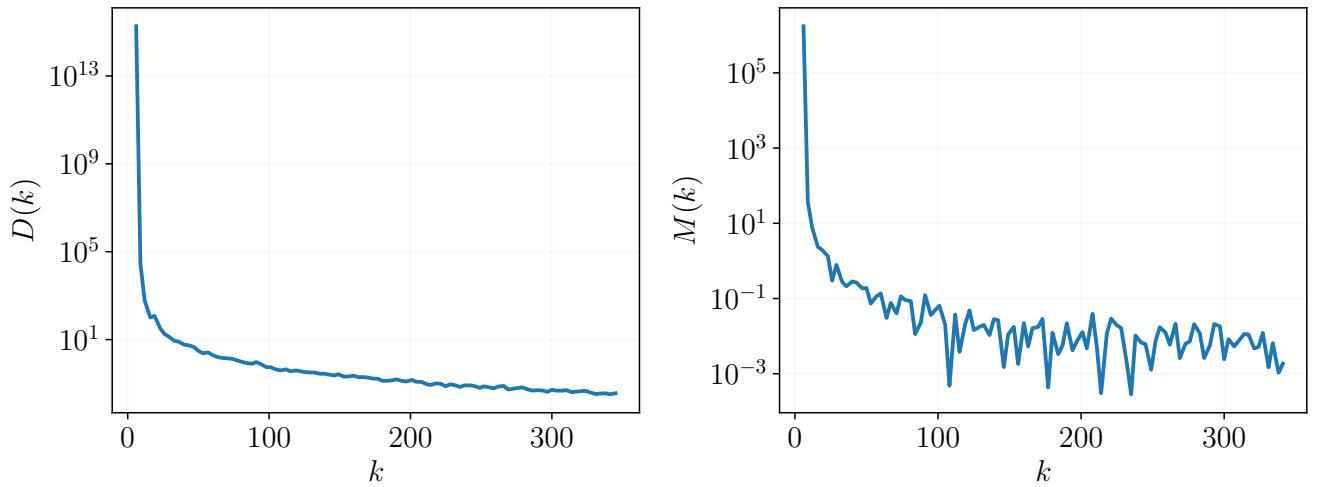


Рис. 4.6: Сходимость функций $D(k)$ и $M(k)$ для набора данных Liver Disorders (345 объектов, 5 признаков, $B = 1000$ бутстрэп-подвыборок). Обе функции демонстрируют сходимость к нулю, что подтверждает теоретические результаты и демонстрирует применимость методов на реальных данных.

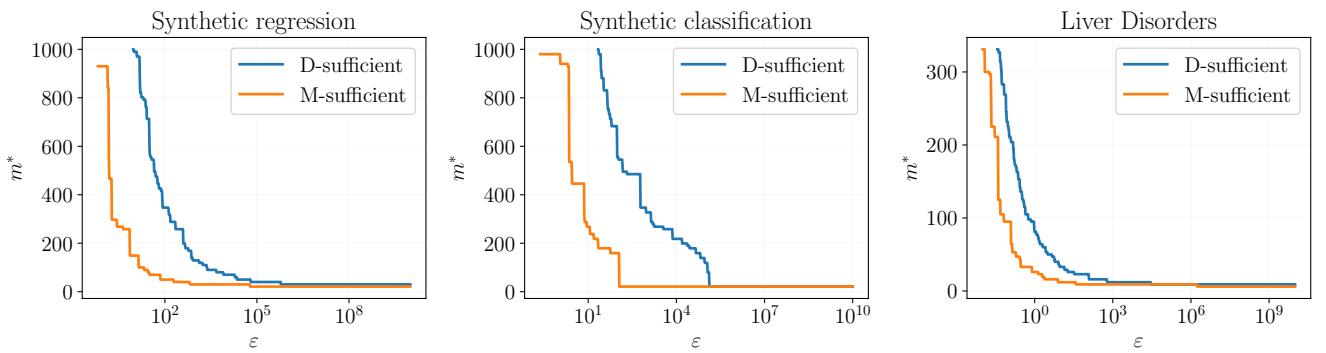


Рис. 4.7: Зависимость достаточного размера выборки m^* от порогового параметра ε для методов D- и M-достаточности на трех наборах данных. С увеличением значения порога ε достаточный размер выборки монотонно уменьшается, что позволяет выбирать меньше объектов для достижения заданного уровня стабильности функций $D(k)$ и $M(k)$.

В настоящем разделе представлено эмпирическое исследование предложенных методов. Эксперименты проводились на синтетических данных и наборе данных Liver Disorders из [70].

Синтетические данные были сгенерированы из моделей линейной регрессии и логистической регрессии. Количество объектов составляет 1000, количество признаков — 20. Использовалось $B = 1000$ бутстрэп-подвыборок. Вычислялись значения $D(k)$ и $M(k)$. Набор данных регрессии Liver Disorders содержит 345

Таблица 4.4: Сравнение оценок достаточного размера выборки методами D- и M-достаточности для 13 наборов данных с задачей регрессии. Методы демонстрируют сопоставимые результаты для большинства наборов данных, при этом M-достаточность иногда требует большего размера выборки.

Dataset name	Objects m	Features n	D	M
Abalone	4177	8	96	96
Auto MPG	392	8	15	15
Automobile	159	25	70	156
Liver Disorders	345	6	12	19
Servo	167	4	41	—
Forest fires	517	12	208	—
Wine Quality	6497	12	144	144
Energy Efficiency	768	9	24	442
Student Performance	649	32	129	177
Facebook Metrics	495	18	31	388
Real Estate Valuation	414	7	15	23
Heart Failure Clinical Records	299	12	63	224
Bone marrow transplant: children	142	36	—	—

объектов и 5 признаков. Также использовалось $B = 1000$ подвыборок, полученных методом бутстрэпа, для оценки математического ожидания и дисперсии функции потерь.

На рис. 4.4 показаны полученные зависимости между доступным размером выборки k и предложенными функциями $D(k)$ и $M(k)$ для синтетического набора данных регрессии. Результаты для синтетического набора данных классификации представлены на рис. 4.5. На рис. 4.6 представлены графики для набора данных Liver Disorders. Наблюдается, что во всех случаях значения $D(k)$ и $M(k)$ приближаются к нулю с увеличением размера выборки. Эти эмпирические результаты подтверждают полученные ранее теоретические выводы.

В определениях D-достаточности и M-достаточности присутствует гиперпараметр ε , который соответствует порогу для достаточного размера выборки m^* . Для изучения зависимости между указанными величинами построена зависимость на рис. 4.7, которая демонстрирует возможные размеры выборки для обеспечения заданного уровня достоверности.

Для сравнения производительности предложенных методов на различных

наборах данных были выбраны выборки из открытого репозитория [70]. Подробная информация о каждом наборе данных, количестве наблюдений и количестве признаков представлена в Таблице 4.4. В демонстрационных целях было выбрано значение гиперпараметра ε , при котором значение целевой функции, $D(k)$ или $M(k)$, уменьшается вдвое. Соответствующие результаты представлены в Таблице 4.4. Пропуски означают, что исходный размер выборки недостаточен.

4.6.3. Определение достаточного размера выборки на основе близости апостериорных распределений

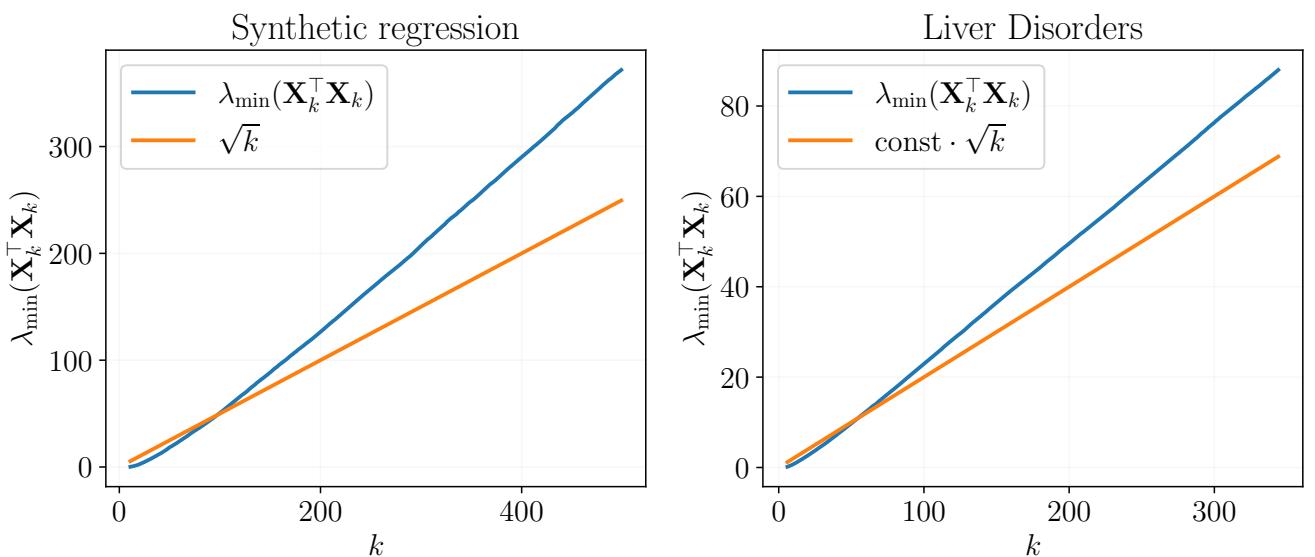


Рис. 4.8: Зависимость минимального собственного значения $\lambda_{\min}(\mathbf{X}_k^\top \mathbf{X}_k)$ от размера выборки k для синтетических данных регрессии (500 объектов, 10 признаков) и набора данных Liver Disorders (345 объектов, 5 признаков). Асимптотическое поведение соответствует условию теоремы 31: $\lambda_{\min} = \omega(\sqrt{k})$ при $k \rightarrow \infty$.

В настоящем разделе представлено расширенное эмпирическое исследование предложенных методов определения достаточного размера выборки на основе близости апостериорных распределений. Эксперименты состоят из трех частей.

В первой части проверяются сходимости, полученные в ходе теоретического анализа. А именно, сначала рассматривается поведение минимального собственного значения матрицы $\mathbf{X}_k^\top \mathbf{X}_k$ при увеличении размера выборки, что необходимо для выполнения условий теоремы 31. Затем исследуется сходимость предложенных функций $KL(k)$ и $S(k)$ к их предельным значениям. Наконец, изу-

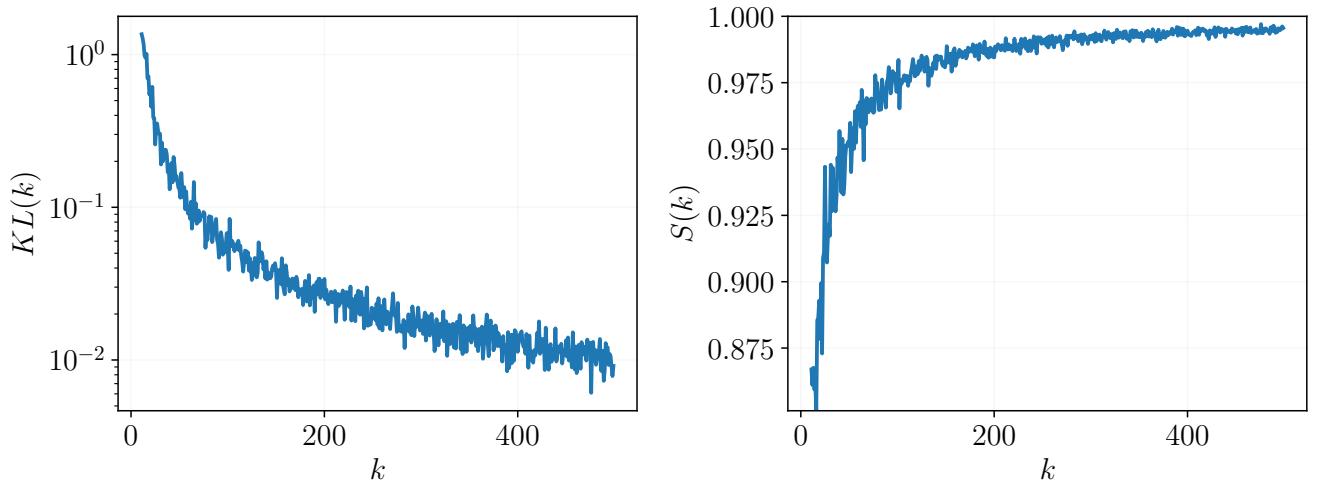


Рис. 4.9: Сходимость функций $KL(k)$ и $S(k)$ для синтетического набора данных регрессии. Функция $KL(k)$ стремится к нулю, а $S(k)$ стремится к единице, что подтверждает теоремы 29 и 30.

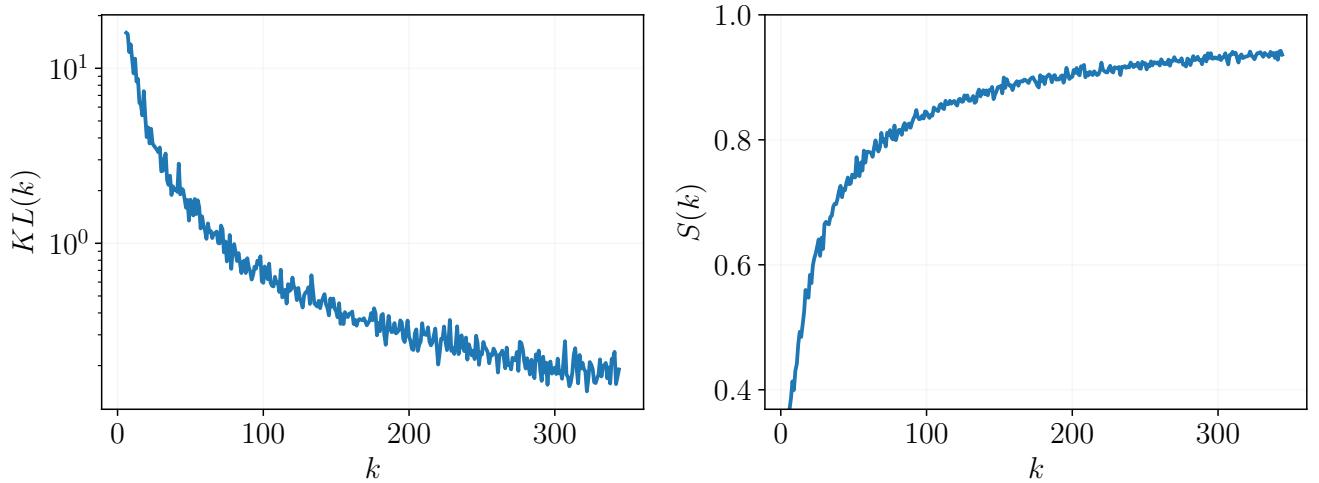


Рис. 4.10: Сходимость функций $KL(k)$ и $S(k)$ для набора данных Liver Disorders (345 объектов, 5 признаков, нормальное априорное распределение, $B = 100$ повторений). Результаты демонстрируют сходимость $KL(k)$ к нулю и $S(k)$ к единице, подтверждая теоретические предсказания и применимость методов KL- и S-достаточности на реальных данных.

чается зависимость достаточного размера выборки от пороговых параметров ε . Эксперимент проводится на двух наборах данных: синтетическая регрессия и Liver Disorders.

Во второй части оцениваются размеры выборок для различных наборов данных, используя разные подходы (KL- и S-достаточность, а также классические

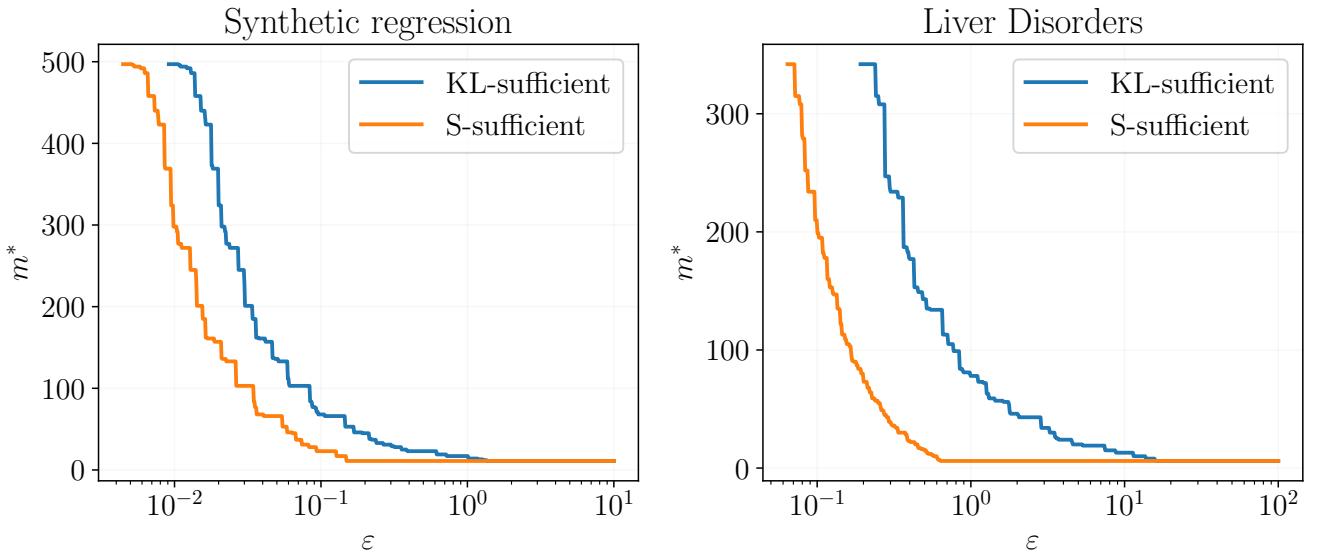


Рис. 4.11: Зависимость достаточного размера выборки m^* от порогового параметра ε для методов KL- и S-достаточности на синтетических данных регрессии и наборе данных Liver Disorders. Метод S-достаточности требует более низких значений порога для достижения заданного уровня близости распределений, что указывает на его более строгие требования к качеству оценки.

Таблица 4.5: Характеристики выборок, используемых для сравнения методов определения достаточного размера выборки на основе близости апостериорных распределений. Все наборы данных соответствуют задаче регрессии и используются для оценки методов KL- и S-достаточности.

Выборка	Количество признаков, n	Количество объектов, m
Boston Housing	14	506
Diabetes	10	576
Forest Fires	13	517
Servo	4	167

методы). В третьей части изучается зависимость достаточного размера выборки от объема доступных данных, что позволяет оценить стабильность методов при различных объемах выборок.

Синтетические данные генерируются из модели линейной регрессии. Количество объектов составляет 500, количество признаков — 10. Для генерации синтетического набора данных регрессии исходные признаки, параметры модели и шумовые остатки генерируются из стандартного нормального распределения.

Таблица 4.6: Сравнение оценок достаточного размера выборки, полученных классическими методами и предложенными методами KL- и S-достаточности для четырех наборов данных регрессии. Метод KL-достаточности дает более консервативные оценки, требующие почти полной выборки, в то время как S-достаточность указывает на минимальные размеры выборки.

Methods and sample sets	Boston	Diabetes	Forest Fires	Servo
Lagrange Multipliers Test	18	25	44	38
Likelihood Ratio Test	17	25	43	18
Wald Test	66	51	46	76
Cross Validation	178	441	171	120
Bootstrap	113	117	86	60
APVC	98	167	351	20
ACC	228	441	346	65
ALC	98	267	516	25
Utility function	148	172	206	105
KL (ours)	493	437	86	165
S (ours)	28	22	26	10

Априорное распределение параметров также задано как стандартное нормальное, как для синтетической регрессии, так и для набора данных Liver Disorders, который содержит 345 объектов и 5 признаков. Входные признаки предобращаются с использованием стандартного метода масштабирования данных (англ. Standard Scaler).

Процедура эксперимента организована следующим образом. Один объект последовательно удалялся из заданной выборки до тех пор, пока количество объектов в подвыборке не становилось равным количеству признаков. Для каждого размера выборки k вычисляется минимальное собственное значение матрицы $\mathbf{X}_k^\top \mathbf{X}_k$, а также значения $KL(k)$ и $S(k)$. Этот процесс повторялся $B = 100$ раз для обеспечения статистической надежности результатов.

На рис. 4.8 показано асимптотическое поведение минимального собственного значения матрицы $\mathbf{X}_k^\top \mathbf{X}_k$ при увеличении размера выборки. Наблюдается, что при стремлении размера выборки к бесконечности минимальное собственное значение также стремится к бесконечности. При этом, как и требуется для теоремы 31, график лежит выше \sqrt{k} .

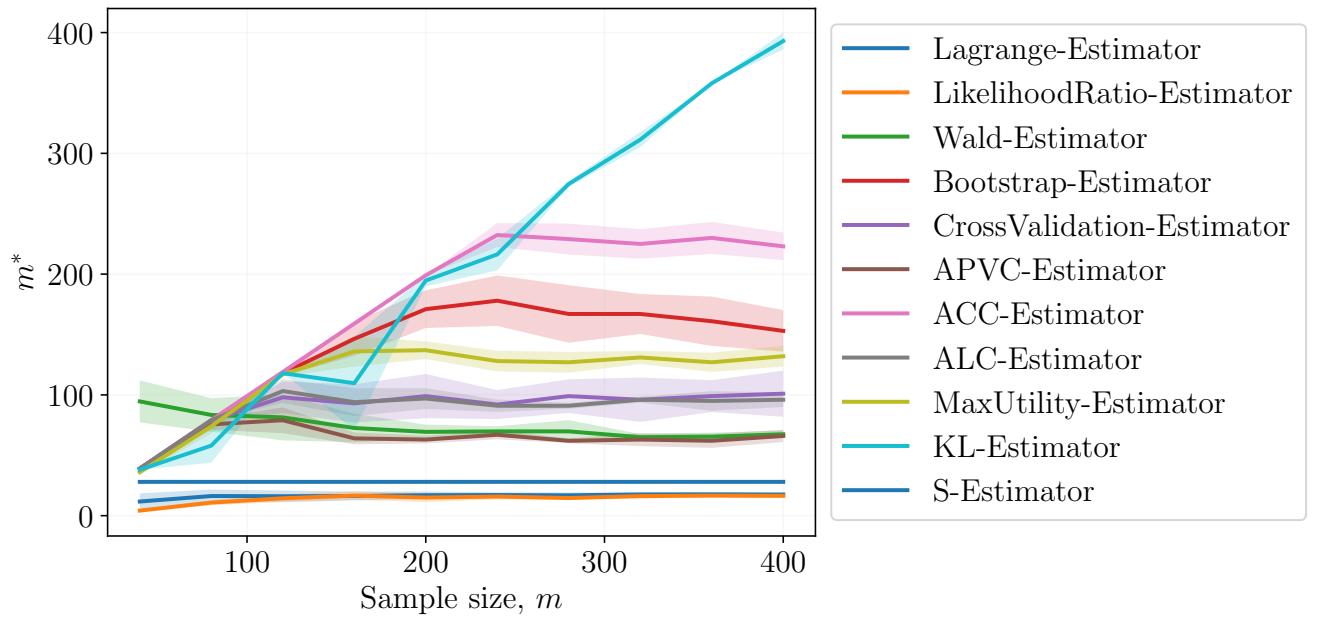


Рис. 4.12: Зависимость оцененного достаточного размера выборки m^* от доступного размера выборки m для классических методов и предложенных методов KL- и S-достаточности на наборе данных Boston Housing. Критерий KL-достаточности является наиболее консервативным и требует почти полной выборки, в то время как S-достаточность указывает на минимальные размеры, что связано с высокой чувствительностью расхождения Кульбака-Лейблера к изменениям распределений.

На рис. 4.9 представлены зависимости между доступным размером выборки k и предложенными функциями $KL(k)$ и $S(k)$ для синтетического набора данных регрессии. На рис. 4.10 представлены аналогичные графики для набора данных Liver Disorders. Наблюдается, что в обоих случаях значение $KL(k)$ приближается к нулю с увеличением размера выборки, а $S(k)$ стремится к единице. Эти эмпирические результаты подтверждают полученные ранее теоретические выводы.

В определениях KL-достаточности и S-достаточности присутствует гиперпараметр ε , который соответствует порогу для достаточного размера выборки m^* . На рис. 4.11 показана зависимость между уровнем достоверности и размером выборки. Для сравнения предложенных методов с базовыми использовалась следующая схема эксперимента. Модель машинного обучения — линейная регрессия. Выбрано 4 набора данных с задачей регрессии из открытых источников: Boston, Diabetes, Forestfires и Servo. Их описательная статистика представлена в Таблице 4.5. К данным применяются 9 различных базовых методов

оценки размера выборки: Тест множителей Лагранжа, Тест отношения правдоподобий, Тест Вальда, Кросс-валидация, Бутстрэп, Критерий средней апостериорной дисперсии (APVC), Критерий среднего покрытия (ACC), Критерий средней длины (ALC) и Функция полезности.

Далее в экспериментах достаточность определяется в терминах относительного изменения, а именно, размер выборки достаточным, если функция $KL(k)$ имеет относительное отклонение от своего значения на всей выборке не более чем ε . Аналогично с функцией $S(k)$, зафиксировав $\varepsilon = 0.05$ получаем результатирующие размеры выборок.

Результаты в Таблице 4.6 указывают на то, что критерий на основе расхождения Кульбака-Лейблера является более консервативным и требует большего размера выборки, в то время как критерий S-достаточности предполагает, что минимального размера выборки может быть достаточно. Предполагается, что это типичный результат для функции схожести s-score, которая была разработана для сравнения различных моделей машинного обучения, особенно в случаях с неинформативными распределениями. Если распределения имеют высокую дисперсию, функция близости приближается к единице, что приводит к тому, что критерий считает достаточным даже небольшой размер выборки.

Далее проводится комплексный анализ различных методов определения размера выборки. Анализируется зависимость достаточного размера выборки от объема доступного набора данных. В частности, при увеличении объема доступной выборки вычисляется достаточный размер на основе различных методов. Результаты представлены на рис. 4.12, который сравнивает вышеупомянутые методы с точки зрения их консервативности.

Наблюдается, что S-достаточный размер выборки часто является минимальным. KL-достаточный размер выборки, как правило, требует почти полной выборки. Предполагается, что это связано с тем, что расхождение Кульбака-Лейблера чрезвычайно чувствительно к изменениям математического ожидания и дисперсии сравниемых распределений. Таким образом, стабилизация расстояния между ними происходит довольно поздно.

4.7. Заключение по главе

В настоящей главе решена задача разработки практических методов определения достаточного размера выборки для задач машинного обучения, что восполняет пробел между теоретическим аппаратом, разработанным в главах 2

и 3, и практическими потребностями планирования экспериментов. В отличие от теоретических оценок, основанных на анализе матриц Гессе, предложенные методы используют наблюдаемые характеристики процесса обучения, что делает их применимыми в реальных задачах.

В главе проведен систематический обзор существующих подходов к определению достаточного размера выборки, включая статистические методы, байесовские методы и эвристические методы. Выявлены ключевые ограничения классических подходов: необходимость знания дисперсии оценки параметра или параметра нецентральности, отсутствие алгоритмических процедур для их получения, а также вычислительная сложность для моделей глубокого обучения.

Основным теоретическим вкладом главы является разработка двух новых методов определения достаточного размера выборки, основанных на анализе стабильности процесса обучения.

Первый метод основан на анализе функции правдоподобия при изменении объема данных. Введены два критерия достаточности: D-достаточность, использующая дисперсию функции правдоподобия на бутстрэп-подвыборках, и M-достаточность, анализирующая разность математических ожиданий функции правдоподобия при последовательном добавлении объектов в выборку. Теорема 28 строго доказывает корректность определения M-достаточного размера выборки для модели линейной регрессии при выполнении условий сходимости математических ожиданий и ковариационных матриц оценок параметров. Следствие 4 устанавливает достаточные условия сходимости к истинным значениям параметров и информационной матрице Фишера.

Второй метод основан на анализе близости апостериорных распределений параметров модели на близких подвыборках, отличающихся одним объектом. Введены критерии KL-достаточности и S-достаточности, использующие расхождение Кульбака-Лейблера и функцию схожести s-score соответственно. Для нормального апостериорного распределения получены аналитические выражения для этих мер близости, что позволило провести строгий теоретический анализ. Теорема 29 доказывает корректность определения KL-достаточного размера выборки при сходимости математических ожиданий и ковариационных матриц апостериорных распределений. Теорема 30 устанавливает корректность S-достаточности при более слабых условиях — требуется только сходимость математических ожиданий. Теорема 31 для модели линейной регрессии с нормальным априорным распределением устанавливает достаточные условия сходимости моментов апостериорного распределения при условии роста минимального

собственного значения матрицы $\mathbf{X}_k^\top \mathbf{X}_k$ как $\omega(\sqrt{k})$ при $k \rightarrow \infty$.

Проведены обширные вычислительные эксперименты на синтетических и реальных данных, подтвердившие эффективность предложенных методов. Эксперименты на синтетических данных регрессии и классификации, а также на наборах данных Liver Disorders, Boston Housing, Diabetes, Forest Fires, Servo и других продемонстрировали сходимость функций $D(k)$, $M(k)$ и $KL(k)$ к нулю, а функции $S(k)$ — к единице с ростом объема выборки, что согласуется с теоретическими предсказаниями. Сравнительный анализ с классическими методами выявил особенности различных критериев: KL-дивергенция дает более консервативные оценки, требующие почти полной выборки, в то время как S-достаточность часто указывает на достаточность минимального размера выборки, что связано с высокой чувствительностью расхождения Кульбака-Лейблера к изменениям распределений.

Практические рекомендации включают использование относительных отклонений от значений на полной выборке с порогами 0.05 – 0.1 для методов на основе функции правдоподобия, а также применение бутстрэпирования с $B = 1000$ подвыборок для оценки математических ожиданий и дисперсий. Для методов на основе близости апостериорных распределений рекомендуется использовать порог $\varepsilon = 0.05$ для S-достаточности.

Основные ограничения методов связаны с вычислительной сложностью обращения ковариационных матриц для моделей с большим количеством параметров, а также с предположением о нормальности апостериорного распределения для методов KL- и S-достаточности. Кроме того, теоретическое обоснование методов в настоящее время ограничено моделями линейной регрессии и логистической регрессии.

Перспективными направлениями дальнейших исследований являются расширение теоретического обоснования методов на более сложные модели, в том числе нейронные сети, преодоление ограничения о нормальности апостериорного распределения, разработка приближенных методов для оценки близости распределений в высокомерных пространствах параметров, а также интеграция предложенных методов с теоретическими оценками ландшафтной меры сложности из главы 2.

Полученные результаты создают основу для разработки практических инструментов оценки достаточного объема данных в прикладных задачах машинного обучения и обеспечивают мост между теоретическим формализмом сложности моделей и данных, разработанным в предыдущих главах, и практическими

ми потребностями планирования экспериментов.

Глава 5

Методы снижения сложности моделей глубокого обучения

В настоящей главе рассматриваются методы снижения сложности параметрических моделей глубокого обучения. Предполагается, что число параметров нейросети можно существенно снизить без значимой потери качества и значимого повышения дисперсии функции ошибки.

Предлагаются методы снижения сложности моделей на основе ковариационной матрицы градиентов функции ошибки по параметрам модели. Разработанные методы опираются на теоретический аппарат, введенный в главах 2 и 3, и обеспечивают практические инструменты для уменьшения сложности моделей при сохранении их качества.

5.1. Удаление параметров моделей глубокого обучения

В настоящем разделе рассматривается метод удаления параметров моделей глубокого обучения на основе анализа ковариационной матрицы градиентов функции ошибки. Предложенный подход позволяет упорядочить параметры по их важности и последовательно удалять наименее значимые параметры без существенной потери качества модели.

Рассмотрим выборку:

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m, \quad \mathbf{x}_i \in \mathbb{X} = \mathbb{R}^n, \quad y_i \in \mathbb{Y},$$

где n — размерность признакового пространства, m — число объектов в выборке. Пространство ответов $\mathbb{Y} = \mathbb{R}$ в случае задачи регрессии и $\mathbb{Y} = \{1, \dots, R\}$ в случае задачи классификации, где R — число классов.

Определим семейство моделей параметрических функций с наперед заданной структурой:

$$\begin{aligned}\mathfrak{F} &= \{f(\mathbf{w}) : \mathbb{X} \rightarrow \mathbb{Y} | \mathbf{w} \in \mathbb{R}^p\}, \\ \mathbf{h}(\mathbf{w}, \mathbf{x}) &= \mathbf{W}_1 \boldsymbol{\sigma}(\mathbf{W}_2 \boldsymbol{\sigma}(\dots \boldsymbol{\sigma}(\mathbf{W}_r \mathbf{x}) \dots)), \\ f_{\text{cl}}(\mathbf{w}, \mathbf{x}) &= \arg \max_{j \in \{1, \dots, R\}} \text{softmax}(\mathbf{h}(\mathbf{w}, \mathbf{x}))_j, \\ f_{\text{reg}}(\mathbf{w}, \mathbf{x}) &= \mathbf{h}(\mathbf{w}, \mathbf{x}),\end{aligned}$$

где p — размерность пространства параметров, r — число слоев нейросети, $\mathbf{w} = \text{vec}[\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_r]$, а $\boldsymbol{\sigma}$ — функция активации. В случае задачи ре-

грессии структура модели имеет вид f_{reg} , а в случае классификации имеет вид f_{cl} . Определим функцию потерь:

$$\begin{aligned}\mathcal{L}(\mathbf{w}, \mathfrak{D}) &= \frac{1}{m} \sum_{i=1}^m l(\mathbf{x}_i, y_i, \mathbf{w}), \\ l_{\text{reg}}(\mathbf{x}, y, \mathbf{w}) &= (y - f(\mathbf{w}, \mathbf{x}))^2, \\ l_{\text{cl}}(\mathbf{x}, y, \mathbf{w}) &= - \sum_{j=1}^R ([y = j] \ln \text{softmax}_j(\mathbf{h}(\mathbf{w}, \mathbf{x}))),\end{aligned}$$

где l_{reg} — это функция ошибки на одном элементе для задачи регрессии, l_{cl} — для задачи классификации. Оптимальный вектор параметров $\hat{\mathbf{w}}$ получается минимизацией функции потерь:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^p} \mathcal{L}(\mathbf{w}, \mathfrak{D}).$$

Для поиска оптимальных параметров модели используется градиентный метод оптимизации:

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \Delta \mathbf{w}(\mathbf{g}_{S,t}, \mathbf{w}_{t-1}, \mathbf{w}_{t-2}, \dots), \quad \mathbf{g}_{S,t} = \frac{\partial \mathcal{L}(\mathbf{w}_t, \mathbf{X}_S, \mathbf{Y}_S)}{\partial \mathbf{w}}, \quad (5.1)$$

где t — номер итерации, $\mathbf{g}_{S,t}$ — значение градиента на подвыборке размера S , $\Delta \mathbf{w}$ — приращение вектора параметров.

Порядок на множестве параметров модели задается при помощи ковариационной матрицы \mathbf{C} градиентов функции ошибки \mathcal{L} по параметрам модели \mathbf{w} . Для вычисления ковариационной матрицы \mathbf{C} используется итерационная формула [71], вычисляемая на каждой итерации (5.1) градиентного метода оптимизации параметров:

$$\mathbf{C}_t = (1 - \kappa_t) \mathbf{C}_{t-1} + \kappa_t (\mathbf{g}_{1,t} - \mathbf{g}_{S,t}) (\mathbf{g}_{1,t} - \mathbf{g}_{S,t})^\top,$$

где t — номер итерации, $\mathbf{g}_{S,t}$ — значение градиента на подвыборке размера S , $\mathbf{g}_{1,t}$ — значение градиента на первом элементе подвыборки, $\kappa_t = \frac{1}{t}$ — параметр сглаживания, \mathbf{C}_0 инициализируются из равномерного распределения.

Пусть известно t_0 — число итераций, после которого все параметры находятся в некоторой локальной окрестности минимума, тогда, как показано в работе [71], матрица \mathbf{C}_{t_0} аппроксимирует истинную ковариационную матрицу \mathbf{C} . Ковариационная матрица \mathbf{C}_{t_0} используется для упорядочения параметров модели \mathbf{w}_{t_0} .

Пусть \mathcal{I} — упорядоченный вектор индексов $[1, 2, \dots, p]$. Обозначим $\mathcal{I}_{\mathbf{w}_{t_0}}$ вектор индексов, порядок которого задан при помощи ковариационной матрицы \mathbf{C}_{t_0} .

Например, если ковариационная матрица \mathbf{C}_{t_0} имеет вид

$$\begin{bmatrix} 0,3 & 0 & 0 \\ 0 & 0,2 & 0 \\ 0 & 0 & 0,25 \end{bmatrix},$$

то вектор индексов $\mathcal{I}_{\mathbf{w}_{t_0}} = [3, 1, 2]$.

Для фиксации параметров \mathbf{w}_{t_0} при помощи вектора индексов $\mathcal{I}_{\mathbf{w}_{t_0}}$ используется бинарный вектор $\boldsymbol{\alpha}(\zeta)$:

$$\alpha_i(\zeta) = \begin{cases} 1, & \text{если } \mathcal{I}_{\mathbf{w}_{t_0}}[j] \leq \zeta; \\ 0 & \text{иначе,} \end{cases} \quad (5.2)$$

где ζ — число фиксирующих параметров.

Учитывая (5.2), уравнение (5.1) приводится к виду

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \boldsymbol{\alpha}(\zeta) \cdot \Delta \mathbf{w}(\mathbf{g}_{S,t}, \mathbf{w}_{t-1}, \mathbf{w}_{t-2}, \dots),$$

где t — номер итерации, $\mathbf{g}_{S,t}$ — значение градиента на подвыборке размера S , $\Delta \mathbf{w}$ — приращение вектора параметров. После умножения на бинарный вектор $\boldsymbol{\alpha}$ часть параметров не оптимизируется, что приводит к фиксации параметров.

Предлагается метод, основанный на модификации метода Белсли. Пусть \mathbf{w} — вектор параметров, доставляющий минимум функционалу потерь \mathcal{L} на множестве $\mathbb{W}_{\mathcal{A}}$, а \mathbf{A}_{ps} — соответствующая ему ковариационная матрица.

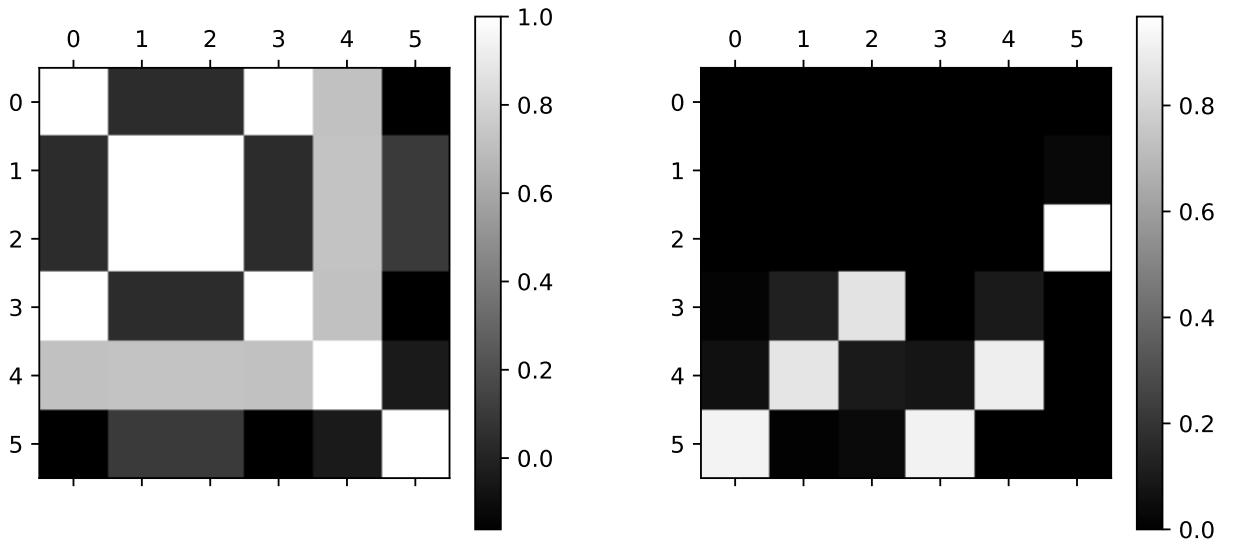
Выполним сингулярное разложение матрицы

$$\mathbf{A}_{ps} = \mathbf{U} \boldsymbol{\Lambda} \mathbf{V}^T.$$

Индекс обусловленности η_j определим как отношение максимального элемента к j -му элементу матрицы $\boldsymbol{\Lambda}$. Для нахождения мультиколлинеарных признаков требуется найти индекс ξ вида:

$$\xi = \arg \max_{j \in \mathcal{A}} \eta_j.$$

Дисперсионный долевой коэффициент q_{ij} определим как вклад j -го признака в дисперсию i -го элемента вектора параметра \mathbf{w} :



(a) Матрица ковариации

(b) Дисперсионные доли

Рис. 5.1: Иллюстрация метода Белсли для анализа мультиколлинеарности параметров на синтетических данных. Слева показана матрица ковариации параметров, справа — дисперсионные доли, демонстрирующие вклад каждого признака в дисперсию параметров модели.

Таблица 5.1: Индексы обусловленности η и дисперсионные доли q_j для синтетических данных, демонстрирующие работу метода Белсли. Максимальный индекс обусловленности $\eta_6 = 1.2 \cdot 10^{16}$ соответствует максимальным дисперсионным долям признаков с индексами 1 и 4, которые являются линейно зависимыми.

η	q_1	q_2	q_3	q_4	q_5	q_6
1.0	$2 \cdot 10^{-17}$	$4 \cdot 10^{-17}$	$1 \cdot 10^{-16}$	$2 \cdot 10^{-17}$	$6 \cdot 10^{-17}$	$3 \cdot 10^{-4}$
1.5	$5 \cdot 10^{-17}$	$9 \cdot 10^{-17}$	$2 \cdot 10^{-16}$	$5 \cdot 10^{-17}$	$3 \cdot 10^{-20}$	$3 \cdot 10^{-2}$
3.3	$9 \cdot 10^{-18}$	$1 \cdot 10^{-17}$	$2 \cdot 10^{-17}$	$9 \cdot 10^{-18}$	$2 \cdot 10^{-19}$	$9 \cdot 10^{-1}$
$2 \cdot 10^{15}$	$1 \cdot 10^{-2}$	$1 \cdot 10^{-1}$	$8 \cdot 10^{-1}$	$2 \cdot 10^{-3}$	$9 \cdot 10^{-2}$	$1 \cdot 10^{17}$
$8 \cdot 10^{15}$	$6 \cdot 10^{-2}$	$8 \cdot 10^{-1}$	$9 \cdot 10^{-2}$	$8 \cdot 10^{-2}$	$9 \cdot 10^{-1}$	$2 \cdot 10^{17}$
$1 \cdot 10^{16}$	$9 \cdot 10^{-1}$	$1 \cdot 10^{-2}$	$4 \cdot 10^{-2}$	$9 \cdot 10^{-1}$	$1 \cdot 10^{-3}$	$5 \cdot 10^{-21}$

$$q_{ij} = \frac{u_{ij}^2 / \lambda_{jj}}{\sum_{j=1}^n u_{ij}^2 / \lambda_{jj}}.$$

Большие значения дисперсионных долей указывают на наличие зависимости между параметрами. Находим долевые коэффициенты, вносящие максимальный вклад в дисперсию параметра w_ξ :

$$\zeta = \arg \max_{j \in \mathcal{A}} q_{\xi j}.$$

Параметр с индексом ζ определяется как наименее релевантный параметр нейросети.

Проиллюстрируем принцип работы метода Белсли на примере. Гипотеза порождения данных:

$$\mathbf{w} = \begin{bmatrix} \sin(x) \\ \cos(x) \\ 2+\cos(x) \\ 2+\sin(x) \\ \cos(x) + \sin(x) \\ x \end{bmatrix}$$

с матрицей ковариации на рис. 5.1.a, где $x \in [0.0, 0.02, \dots, 20.0]$.

В таблице 5.1 приведены индексы обусловленности и соответствующие им дисперсионные доли, изображенные на рис. 5.1.b. Согласно этим данным, максимальный индекс обусловленности $\eta_6 = 1.2 \cdot 10^{16}$. Ему соответствуют максимальные дисперсионные доли признаков с индексами 1 и 4, которые, согласно построению выборки, являются линейно зависимыми.

5.2. Дистилляция моделей глубокого обучения на многодоменных данных

В настоящем разделе рассматривается метод дистилляции моделей глубокого обучения на многодоменных данных. В отличие от классической дистилляции, где модель учителя и модель ученика обучаются на данных из одного домена, предлагаемый метод позволяет передавать знания между моделями, обученными на данных из различных доменов, связанных инъективным отображением.

Определение 27 (Близкие генеральные совокупности). *Генеральная совокупность объектов B называется близкой к генеральной совокупности A , если существует инъективное отображение $\varphi : A \rightarrow B$.*

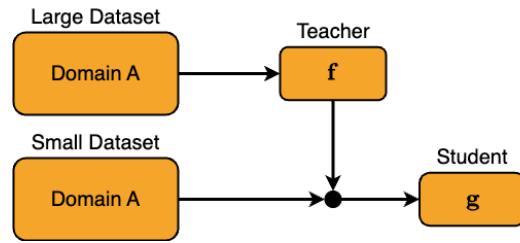


Рис. 5.2: Схема базовой дистилляции моделей глубокого обучения, где модель учителя обучается на большом наборе данных из генеральной совокупности A , а затем ее выходы используются для обучения модели ученика на меньшем наборе данных из того же домена.

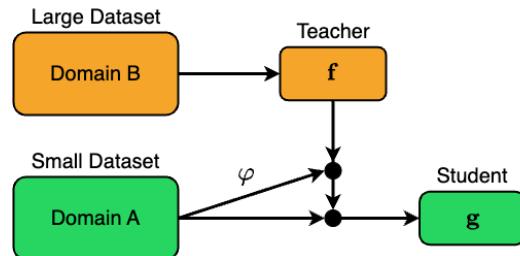


Рис. 5.3: Схема дистилляции моделей глубокого обучения с доменной адаптацией, где модель учителя обучается на данных из домена B , а модель ученика — на данных из домена A , связанных инъективным отображением φ .

Предлагается использовать, помимо меток учителя на одном из доменов, связь между доменами при обучении модели ученика. В этом случае в качестве доменов должны выступать близкие генеральные совокупности.

На рис. 5.2 показан процесс обучения модели ученика в базовой постановке задачи дистилляции. Модель учителя обучается на большом наборе данных из генеральной совокупности A , затем ее выходы используются для обучения модели ученика на меньшем наборе данных из того же домена. На рис. 5.3 представлен предложенный метод, задействующий выходы модели учителя, обученной на другом домене, и связь между доменами.

Рассмотрим базовую постановку задачи дистилляции. Задан набор данных

$$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n, \quad \mathbf{x}_i \in \mathbb{X}, \quad \mathbf{y}_i \in \{1, \dots, R\},$$

где R — количество классов в задаче классификации.

Предполагается, что задана обученная модель с большим количеством параметров — модель учителя. Требуется обучить модель ученика с меньшим количеством параметров, учитывая ответы учителя. Модель учителя f и модель

ученика \mathbf{g} принадлежат параметрическому семейству функций:

$$\mathfrak{F} = \{\mathbf{f} | \mathbf{f} = \text{softmax}(\mathbf{v}(\mathbf{x})/T), \mathbf{v} : \mathbb{R}^n \rightarrow \mathbb{R}^R\},$$

где \mathbf{v} — дифференцируемая параметрическая функция заданной структуры, T — параметр температуры, используемый для смягчения распределения вероятностей.

Функция потерь \mathcal{L} , учитывающая модель учителя \mathbf{f} при выборе модели ученика \mathbf{g} , имеет вид:

$$\begin{aligned}\mathcal{L}(\mathbf{w}, \mathbf{X}, \mathbf{Y}, \mathbf{f}) = & - \sum_{i=1}^m \sum_{r=1}^R y_i^r \log g^r(x_i) \Big|_{T=1} \\ & - \sum_{i=1}^m \sum_{r=1}^R f^r(x_i) \Big|_{T=T_0} \log g^r(x_i) \Big|_{T=T_0},\end{aligned}$$

где $\cdot \Big|_{T=t}$ означает, что параметр температуры T в предыдущей функции равен t . Первое слагаемое в функции потерь соответствует стандартной задаче классификации с температурой $T = 1$, второе слагаемое — дистилляции с температурой $T = T_0$, где $T_0 > 1$ позволяет получить более мягкое распределение вероятностей от модели учителя.

Задача оптимизации формулируется следующим образом:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{W}} \mathcal{L}(\mathbf{w}, \mathbf{X}, \mathbf{Y}, \mathbf{f}).$$

Перейдем к постановке задачи дистилляции для многодоменной выборки. Даны две выборки:

$$\mathfrak{D}_s = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n, \quad \mathbf{x}_i \in \mathbb{X}_s, \quad \mathbf{y}_i \in \mathbb{Y}$$

$$\mathfrak{D}_t = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^m, \quad \mathbf{x}_i \in \mathbb{X}_t, \quad \mathbf{y}_i \in \mathbb{Y},$$

где $\mathfrak{D}_s, \mathfrak{D}_t$ — исходный и целевой наборы данных. В базовой постановке задачи дистилляции предполагается, что $\mathfrak{D}_t \subset \mathfrak{D}_s, \mathbb{X}_t = \mathbb{X}_s$. Предполагается, что количество объектов в наборах данных не совпадает:

$$n \gg m$$

Пусть задана модель учителя на выборке большей мощности:

$$\mathbf{f} : \mathbb{X}_s \rightarrow \mathbb{Y}',$$

где \mathbf{f} — модель учителя, \mathbb{Y}' — пространство вероятностей классов.

Связь между исходной и целевой выборками задается:

$$\varphi : \mathbb{X}_t \rightarrow \mathbb{X}_s,$$

где φ — инъективное отображение. Требуется получить модель ученика для малоресурсной выборки:

$$\mathbf{g} : \mathbb{X}_t \rightarrow \mathbb{Y}',$$

где \mathbf{g} — модель ученика.

В работе рассматривается функция потерь, учитывающая метки учителя и связь между доменами:

$$\begin{aligned} \mathcal{L}(\mathbf{w}, \mathbf{X}, \mathbf{Y}, \mathbf{f}, \varphi) = & \\ & -\lambda \sum_{i=1}^m \sum_{r=1}^R \mathbb{I}[y_i = r] \log g^r(\mathbf{x}_i, \mathbf{w}) \\ & -(1-\lambda) \sum_{i=1}^m \sum_{r=1}^R (f \circ \varphi)^r(\mathbf{x}_i) \log g^r(\mathbf{x}_i, \mathbf{w}), \end{aligned}$$

где λ — метапараметр, задающий вес дистилляции, \mathbb{I} — индикаторная функция.

Задача оптимизации для мультидоменной дистилляции формулируется следующим образом:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{W}} \mathcal{L}(\mathbf{w}, \mathbf{X}, \mathbf{Y}, \mathbf{f}, \varphi).$$

Функция потерь объединяет информацию от истинных меток и от модели учителя, примененной к преобразованным данным другого домена.

5.3. Антидистилляция моделей глубокого обучения

В настоящем разделе описывается постановка задачи анти-дистилляции для задачи классификации. В отличие от классической дистилляции, где знания передаются от более сложной модели к более простой, анти-дистилляция решает обратную задачу: передачу знаний от простой модели к более сложной для работы с более сложными наборами данных. Аналогичный подход может быть применен для произвольных задач машинного обучения.

Даны два набора данных

$$\mathfrak{D}_1 = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m_1}, \quad \mathbf{x}_i \in \mathbb{R}^n, \quad y_i \in C_1 = \{1, \dots, c_1\},$$

$$\mathfrak{D}_2 = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m_2}, \quad \mathbf{x}_i \in \mathbb{R}^n, \quad y_i \in C_2 = \{1, \dots, c_2\},$$

где m_1 и m_2 — количество объектов в \mathfrak{D}_1 и \mathfrak{D}_2 соответственно, n — размерность входного пространства. C_1 и C_2 — множества меток классов $1, \dots, c_1, \dots, c_2$.

Предполагается, что объекты \mathbf{x}_i порождены из генеральной совокупности, общей для обоих наборов данных $\mathfrak{D}_1, \mathfrak{D}_2$, и имеют схожие свойства для этих наборов. Предполагается, что набор данных \mathfrak{D}_2 является более сложным для классификации и требует более сложной модели классификации.

Рассмотрим модель учителя \mathbf{g}_{tr} , обученную на первом наборе данных \mathfrak{D}_1 :

$$\mathbf{g}_{\text{tr}} : \mathbb{R}^n \rightarrow \Delta^{c_1}, \quad \mathbf{g}_{\text{tr}}(\mathbf{x}) = \mathbf{g}(\mathbf{x}, \hat{\mathbf{u}}),$$

где Δ^c — множество c -мерных вероятностных векторов,

Параметры модели учителя \mathbf{g}_{tr} определяются следующим образом:

$$\hat{\mathbf{u}} = \arg \min_{\mathbf{u}} \mathcal{L}_{\text{ce}}(\mathbf{u}, \mathfrak{D}_1) = \arg \min_{\mathbf{u}} \sum_{i=1}^{m_1} l(y_i, g(\mathbf{x}_i, \mathbf{u})),$$

здесь l — перекрестная энтропия:

$$l(y, \hat{y}) = - \sum_{k=1}^c [y = k] \log \hat{y}_k, \quad y \in C, \quad \hat{y} \in \Delta^c.$$

Задача состоит в построении модели ученика

$$\mathbf{f}_{\text{st}} : \mathbb{R}^n \rightarrow \Delta^{c_2}, \quad \mathbf{f}_{\text{st}}(\mathbf{x}) = \mathbf{f}(\mathbf{x}, \hat{\mathbf{w}}),$$

, минимизирующей перекрестную энтропию на валидационной части второго набора данных \mathfrak{D}_2

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \mathcal{L}_{\text{ce}}(\mathbf{w}, \mathfrak{D}_2^{\text{val}}),$$

, где $\mathfrak{D}_2 = \mathfrak{D}_2^{\text{train}} \sqcup \mathfrak{D}_2^{\text{val}}$ и $\hat{\mathbf{w}}$ — оптимальные параметры модели.

Поскольку валидационная функция потерь не оптимизируется напрямую, используются градиентные методы оптимизации на обучающей части $\mathfrak{D}_2^{\text{train}}$ набора данных \mathfrak{D}_2 . Чтобы уменьшить переобучение и использовать больше информации о данных, используется информация от модели учителя \mathbf{g}_{tr} . Предполагается, что наборы данных \mathfrak{D}_1 и \mathfrak{D}_2 имеют общие свойства.

Функция

$$\varphi : \mathbb{R}^{N_{\text{tr}}} \rightarrow \mathbb{R}^{N_{\text{st}}}$$

отображает параметры модели учителя в начальные параметры модели ученика $\mathbf{w} = \varphi(\hat{\mathbf{u}})$.

Гипотеза 1. *Модели ученика, инициализированные результатом применения функции φ к параметрам предварительно обученной модели учителя, являются более устойчивыми и достигают более высокой точности, чем модели с параметрами по умолчанию.*

Основная проблема предложенного метода заключается в том, что модель учителя \mathbf{g}_{tr} , обученная на простом наборе данных \mathfrak{D}_1 , может быть намного проще, чем модель ученика \mathbf{f}_{st} . Для использования большего объема информации из параметров модели учителя $\hat{\mathbf{u}}$ требуется расширить размерность пространства параметров модели учителя N_{tr} до размерности N_{st} пространства параметров модели ученика.

Для решения данной проблемы оптимизируется следующая составная функция потерь:

$$\varphi(\mathbf{u}) = \arg \min_{\mathbf{w} \in \mathbb{R}^{N_{\text{st}}}} \mathcal{L}(\mathbf{w}), \quad (5.3)$$

где

$$\mathcal{L}(\mathbf{w}) = \lambda_1 \mathcal{L}_{\text{ce}}(\mathbf{w}, \mathfrak{D}_1) + \lambda_2 \mathcal{L}_2(\mathbf{w}, \mathbf{u}) + \lambda_3 \mathcal{L}_3^\delta(\mathbf{w}, \mathfrak{D}_1) + \lambda_4 \mathcal{L}_4(\mathbf{w}),$$

$$\forall i \in \overline{1, 4} \lambda_i \geq 0$$

Первое слагаемое $\mathcal{L}_{\text{ce}}(\mathbf{w}, \mathfrak{D}_1)$ представляет собой перекрестную энтропию, отвечающую за качество модели ученика на простом наборе данных \mathfrak{D}_1 . Это слагаемое обеспечивает сохранение способности модели ученика решать исходную задачу классификации.

Второе слагаемое

$$\mathcal{L}_2(\mathbf{w}, \mathbf{u}) = \|\mathbf{u} - \mathbf{Pr}[\mathbf{w}]\|_2^2$$

обеспечивает близость параметров модели учителя и модели ученика в соответствующих местах, где \mathbf{Pr} — оператор проекции, выбирающий только первые параметры, общие для обеих моделей. В случае моделей многослойного перцептрона оператор \mathbf{Pr} выбирает параметры тех же нейронов для каждого слоя модели. Данное слагаемое способствует передаче знаний от модели учителя к модели ученика через близость параметров.

Третье слагаемое

$$\mathcal{L}_3^\delta(\mathbf{w}, \mathfrak{D}_1) = \sum_{(\mathbf{x}, y) \in \mathfrak{D}_1} \mathbb{E}_{\mathbf{x}' \in U_\delta(\mathbf{x})} \mathcal{L}_{\text{ce}}(\mathbf{w}, \mathbf{x}', y)$$

отвечает за устойчивость решения к шуму во входных данных, где $U_\delta(\mathbf{x})$ представляет равномерное распределение в окрестности $[\delta - \mathbf{x}, \delta + \mathbf{x}]$. Данное слагаемое обеспечивает сглаживание функции потерь в окрестности обучающих

примеров, что повышает устойчивость модели к небольшим искажениям входных данных.

Четвертое слагаемое

$$\mathcal{L}_4(\mathbf{w}) = \text{tr} \left(\frac{\partial^2 \mathcal{L}_{\text{ce}}}{\partial \mathbf{w}^2} \right)$$

выполняет регуляризацию гессиана функции потерь, что способствует нахождению решений в более плоских областях ландшафта функции потерь и повышает устойчивость модели к возмущениям параметров.

Последнее слагаемое \mathcal{L}_4 включает вычисление гессиана, наивное вычисление которого может быть ресурсоемким. Для эффективного вычисления следа гессиана используется метод стохастической аппроксимации [72] с быстрым умножением гессиан-вектор [45]. Сложность такой процедуры линейна от количества параметров модели ученика \mathbf{f}_{st} .

В интересующем нас случае анти-дистилляции подразумевается $\lambda_2 > 0$, то есть оптимизация, делающая параметры модели учителя и ученика достаточно близкими. Это обеспечивает передачу знаний от модели учителя к модели ученика через близость параметров.

Важным свойством модели является устойчивость к искажению входных данных. Для обеспечения этого свойства используются слагаемые \mathcal{L}_3 и \mathcal{L}_4 . Оба этих слагаемых регулируют гессиан функции перекрестной энтропии [46, 73], способствуя нахождению решений в более плоских областях ландшафта функции потерь.

5.4. Результаты вычислительных экспериментов

В настоящем разделе представлены результаты вычислительных экспериментов для методов снижения сложности моделей, описанных в предыдущих разделах главы. Эксперименты направлены на валидацию предложенных методов и сравнение их эффективности с существующими подходами.

5.4.1. Удаление параметров моделей глубокого обучения

Для анализа свойств предложенного алгоритма и сравнения его с существующими подходами проведен вычислительный эксперимент, в котором параметры нейросети удалялись методами, описанными в разделах 3.1–3.3, и методом Белсли.

В качестве данных использовались три выборки. Выборки Wine [74] и Boston Housing [68] представляют собой реальные данные, широко используемые в задачах машинного обучения. Синтетические данные сгенерированы таким образом, чтобы параметры сети были мультиколлинеарными, что позволяет проверить эффективность метода Белсли в условиях, для которых он был разработан. Генерация синтетических данных состояла из двух этапов. На первом этапе генерировался вектор параметров $\mathbf{w}_{\text{synthetic}}$:

$$\mathbf{w}_{\text{synthetic}} \sim \mathcal{N}(\mathbf{m}_{\text{synthetic}}, \mathbf{A}_{\text{synthetic}}),$$

$$\text{где } \mathbf{m}_{\text{synthetic}} = \begin{bmatrix} 1.0 \\ 0.0025 \\ \dots \\ 0.0025 \end{bmatrix}, \quad \mathbf{A}_{\text{synthetic}} = \begin{bmatrix} 1.0 & 10^{-3} & \dots & 10^{-3} & 10^{-3} \\ 10^{-3} & 1.0 & \dots & 0.95 & 0.95 \\ \dots & \dots & \dots & \dots & \dots \\ 10^{-3} & 0.95 & \dots & 0.95 & 1.0 \end{bmatrix}.$$

На втором этапе генерировалась выборка $\mathfrak{D}_{\text{synthetic}}$:

$$\mathfrak{D}_{\text{synthetic}} = \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \sim \mathcal{N}(\mathbf{1}, \mathbf{I}), y_i = x_{i0}, i = 1 \dots 10000\}.$$

В приведенном выше векторе параметров $\mathbf{w}_{\text{synthetic}}$ для выборки $\mathfrak{D}_{\text{synthetic}}$ наиболее релевантным является первый параметр, а все остальные параметры нерелевантны. Матрица ковариации выбрана таким образом, чтобы все нерелевантные параметры были зависимыми величинами, что приводит к максимальной эффективности метода Белсли.

Таблица 5.2: Характеристики выборок, использованных для анализа метода задания порядка параметров методом Белсли. Включает реальные данные (Wine для классификации, Boston Housing для регрессии) и синтетические данные с мультиколлинеарными параметрами.

Выборка	Тип задачи	Размер выборки	Число признаков
Wine	классификация	178	13
Boston Housing	регрессия	506	13
Synthetic data	регрессия	10000	100

Для всех алгоритмов тренировочная и тестовая выборки составили 80% и 20% соответственно. Критерием качества прореживания служит процент параметров нейросети, удаление которого не влечет значимой потери качества прогноза. Дополнительным критерием качества служит устойчивость нейросети к зашумленности данных.

Качеством прогноза R_{cl} модели для задачи классификации является точность прогноза модели:

$$R_{\text{cl}} = \frac{\sum_{(\mathbf{x}, y) \in \mathfrak{D}} [f(\mathbf{x}, \mathbf{w}) = y]}{|\mathfrak{D}|},$$

где $[\cdot]$ — индикаторная функция.

Качеством прогноза R_{rg} модели для задачи регрессии является среднеквадратическое отклонение результата модели от точного:

$$R_{\text{rg}} = \frac{\sum_{(\mathbf{x}, y) \in \mathfrak{D}} (f(\mathbf{x}, \mathbf{w}) - y)^2}{|\mathfrak{D}|}.$$

Для эксперимента на выборке Wine [74] рассматривается нейронная сеть с 13 нейронами на входе, 13 нейронами в скрытом слое и 3 нейронами на выходе.

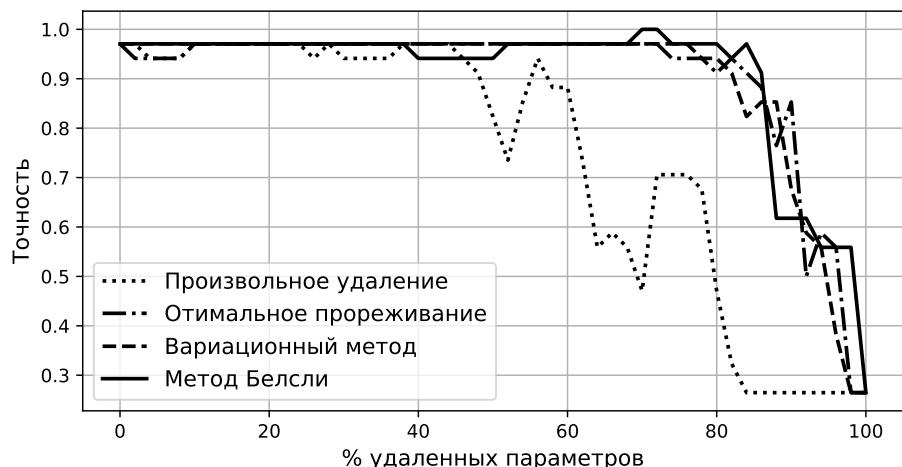


Рис. 5.4: Зависимость точности классификации R_{cl} от процента удаленных параметров для различных методов прореживания на выборке Wine. Метод оптимального прореживания, вариационный метод и метод Белсли позволяют удалить $\approx 80\%$ параметров без существенной потери качества.

На рис. 5.4 показана зависимость точности прогноза R_{cl} от процента удаленных параметров для различных методов прореживания. Результаты демонстрируют, что метод оптимального прореживания, вариационный метод и метод Белсли позволяют удалить до $\approx 80\%$ параметров без существенной потери качества. При удалении $\approx 90\%$ параметров качество всех методов начинает снижаться, что указывает на достижение предела эффективности прореживания для данной архитектуры.

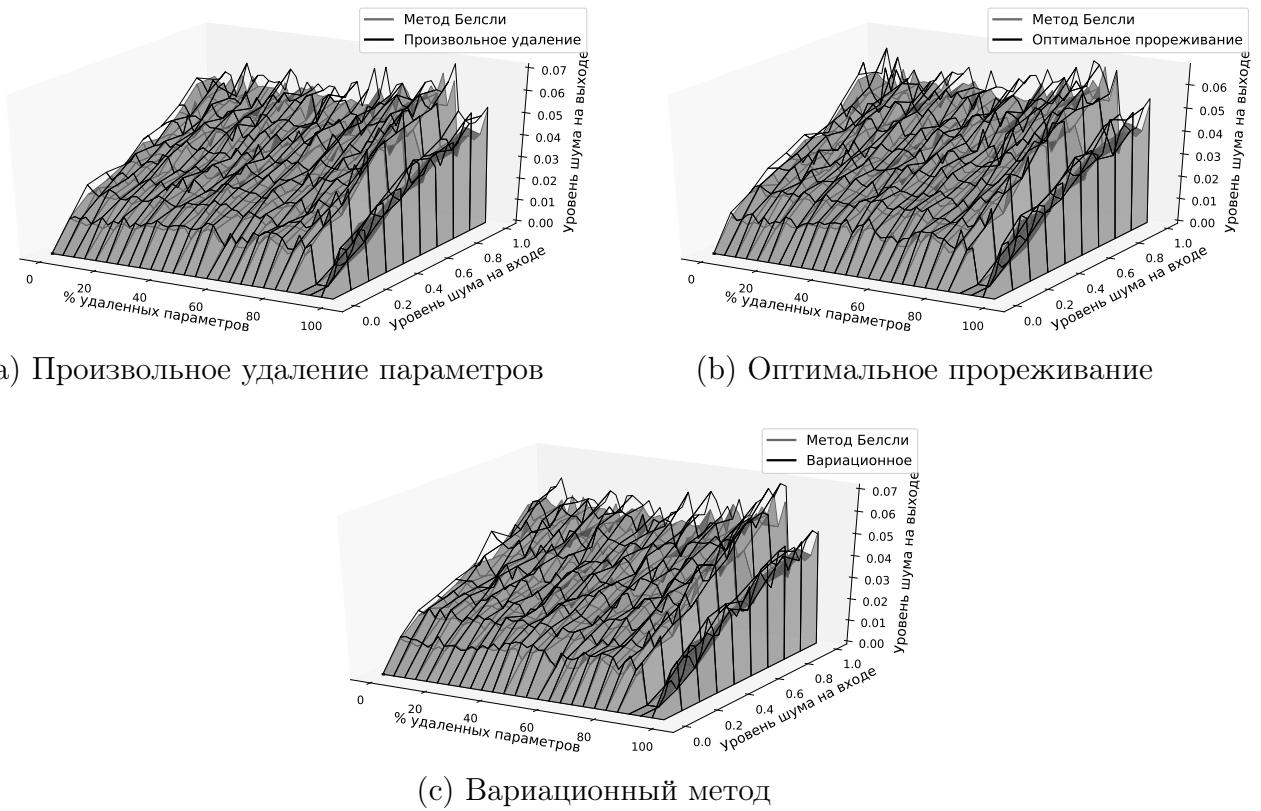


Рис. 5.5: Поверхности изменения уровня шума ответов нейросети при изменении процента удаленных параметров и уровня шума входных данных для различных методов прореживания на выборке Wine. Метод Белсли демонстрирует наименьший уровень шума, что видно по более низкому положению соответствующей поверхности.

На рис. 5.5 показаны поверхности изменения уровня шума ответов нейросети при изменении процента удаленных параметров и уровня шума входных данных для разных методов прореживания. Анализ поверхностей показывает, что метод Белсли демонстрирует наименьший уровень шума в выходных данных по сравнению с другими методами, на что указывает более низкое положение соответствующей поверхности. Это свидетельствует о лучшей устойчивости модели, полученной методом Белсли, к зашумленности входных данных.

Для эксперимента на выборке Boston Housing [68] рассматривается нейронная сеть с 13 нейронами на входе, 39 нейронами в скрытом слое и одним нейроном на выходе.

На рис. 5.6 показана зависимость среднеквадратического отклонения прогноза R_{rg} от точного ответа от процента удаленных параметров для различных методов. График демонстрирует, что метод Белсли является более эффектив-

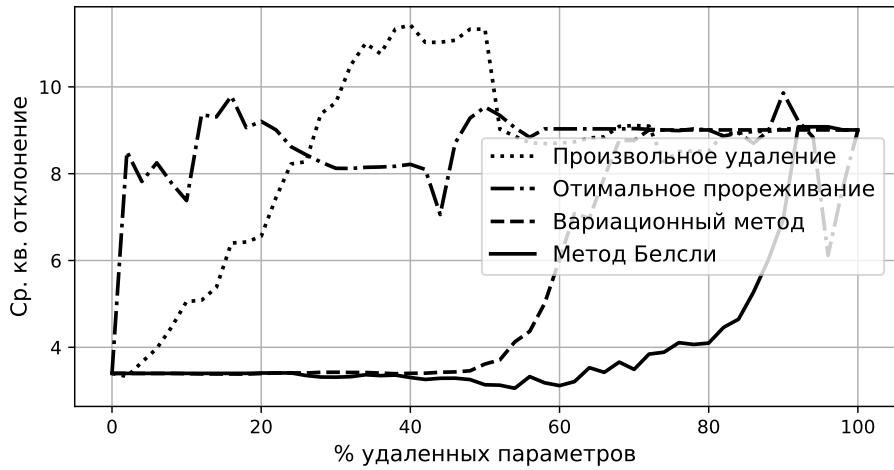


Рис. 5.6: Зависимость среднеквадратического отклонения прогноза R_{rg} от процента удаленных параметров для различных методов прореживания на выборке Boston Housing. Метод Белсли является наиболее эффективным, позволяя удалить больше параметров без потери качества.

ным, чем другие методы, поскольку позволяет удалить больше параметров нейросети без потери качества.

На рис. 5.7 показаны поверхности изменения уровня шума ответов нейросети при изменении процента удаленных параметров и уровня шума входных данных для разных методов прореживания. Анализ показывает, что уровень шума всех методов примерно одинаковый, поскольку соответствующие поверхности находятся на близком уровне. Это указывает на то, что для данной задачи регрессии устойчивость к шуму во входных данных не зависит существенно от выбранного метода прореживания.

Для эксперимента на синтетических данных рассматривается нейронная сеть с 100 нейронами на входе и одним нейроном на выходе.

На рис. 5.8 показана зависимость среднеквадратического отклонения прогноза от точного ответа от процента удаленных параметров для различных методов. График демонстрирует, что удаление параметров методом Белсли является более эффективным, чем другие методы прореживания, поскольку качество прогноза нейросети повышается при удалении шумовых параметров.

На рис. 5.9 показаны поверхности изменения уровня шума ответов нейросети при изменении процента удаленных параметров и уровня шума входных данных для разных методов прореживания. Анализ показывает, что метод Белсли демонстрирует наименьший уровень шума, поскольку соответствующая поверх-

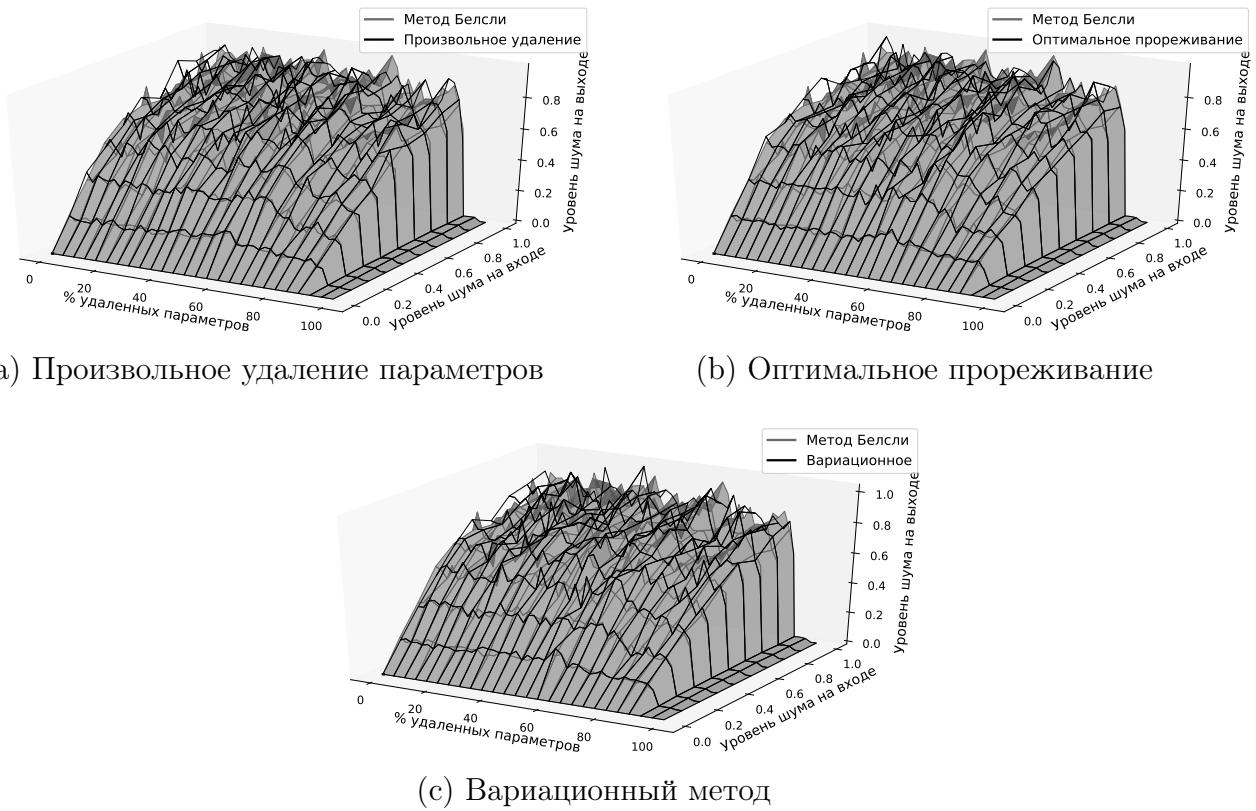


Рис. 5.7: Поверхности изменения уровня шума ответов нейросети при изменении процента удаленных параметров и уровня шума входных данных для различных методов прореживания на выборке Boston Housing. Все методы демонстрируют одинаковый уровень шума, так как соответствующие поверхности находятся на одном уровне.

ность находится ниже других поверхностей. Это подтверждает эффективность метода Белсли для работы с мультиколлинеарными параметрами, что соответствует его теоретическому обоснованию.

5.4.2. Дистилляция моделей глубокого обучения на многодоменных данных

В настоящем подразделе представлены результаты вычислительных экспериментов для метода мультидоменной дистилляции, описанного в разделе 5. Цель вычислительного эксперимента — сравнить производительность моделей учителя и ученика на реальных наборах данных с использованием отображения φ и без него для задач компьютерного зрения и обработки естественного языка. Для анализа качества дистилляции используется интегральный критерий качества [75].

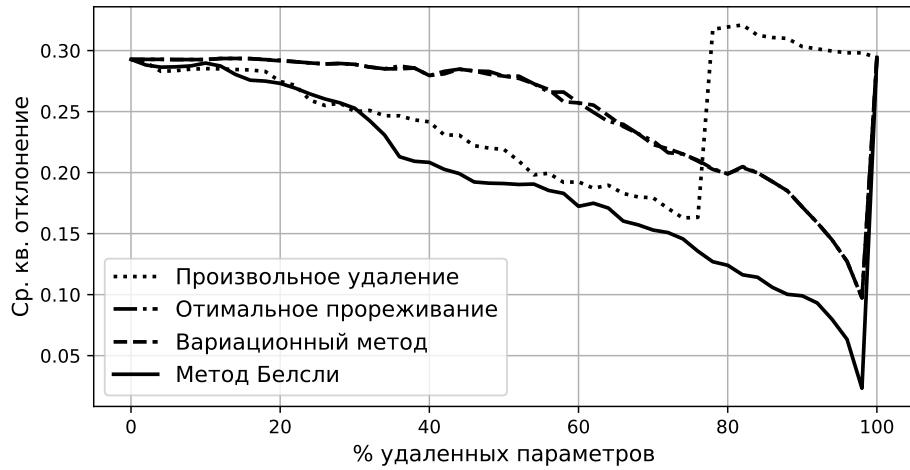


Рис. 5.8: Зависимость среднеквадратичного отклонения прогноза от процента удаленных параметров для различных методов прореживания на синтетической выборке с мультиколлинеарными параметрами. Метод Белсли является наиболее эффективным, поскольку качество прогноза повышается при удалении шумовых параметров.

Эксперименты проводятся на двух типах данных:

- Подмножество ImageNet [76] — набор изображений для задачи классификации на 10 классов. Набор данных состоит из обучающей и тестовой частей, причем обучающая часть разделена на мультиресурсную и мало-ресурсную части, что позволяет моделировать сценарий передачи знаний от модели, обученной на большом наборе данных, к модели, обучаемой на малом наборе данных.
- OPUS-100 [77] — англоцентричный набор данных для машинного перевода, в котором все обучающие пары включают английский язык на стороне источника или цели. Используются языковые пары fr-en и de-en, что позволяет проверить эффективность мультидоменной дистилляции при работе с различными языковыми парами.

Конфигурация алгоритма многодоменной дистилляции для задачи компьютерного зрения.

Структуры модели учителя \mathbf{f} и модели ученика \mathbf{g} описаны в таблице 5.3 и таблице 5.4. Функция активации после каждого скрытого слоя — ReLu. Для решения задачи оптимизации используется метод градиентной оптимизации Adam [57].

В таблице 5.5 описаны наборы данных для вычислительного эксперимента

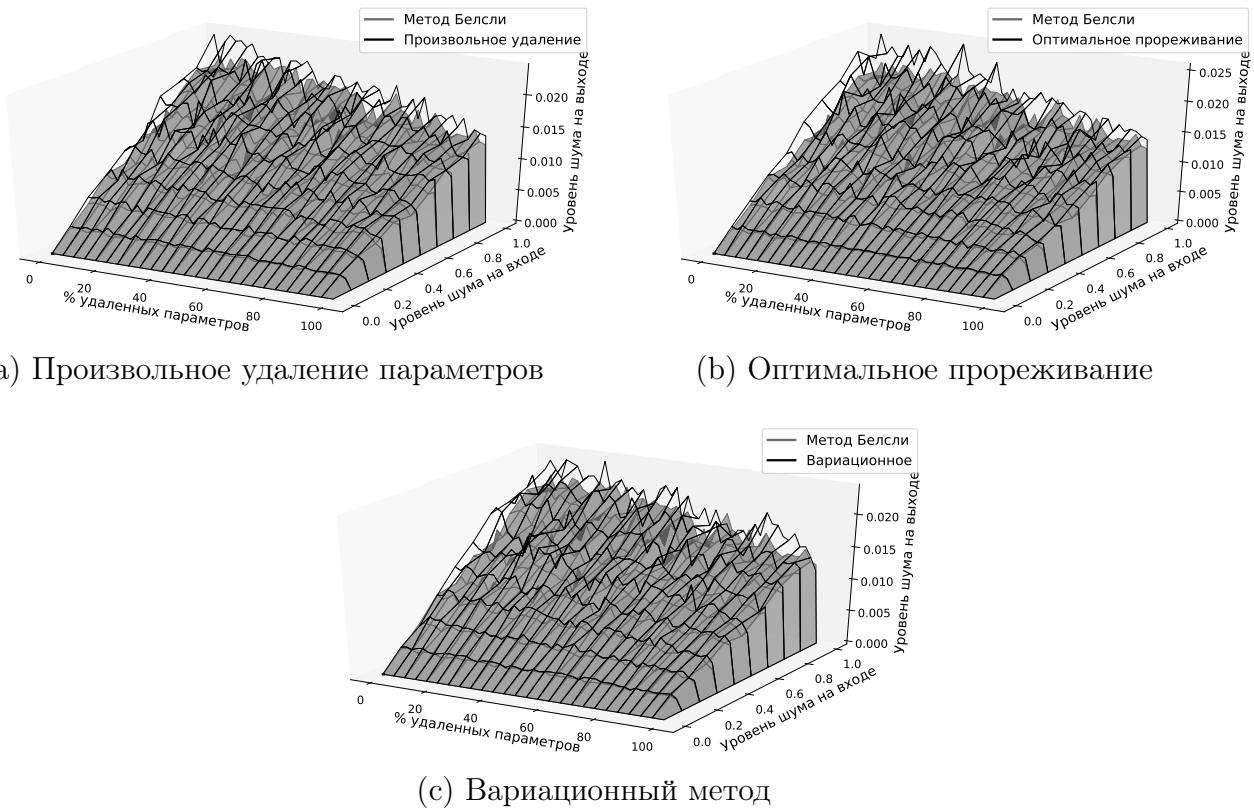


Рис. 5.9: Поверхности изменения уровня шума ответов нейросети при изменении процента удаленных параметров и уровня шума входных данных для различных методов прореживания на синтетической выборке. Метод Белсли демонстрирует наименьший уровень шума, что видно по более низкому положению соответствующей поверхности.

по компьютерному зрению. Каждый из наборов данных состоит из обучающей и тестовой части, при этом обучающая часть разделена на мультиресурсную и малоресурсную части. Обучающая часть содержит 9 469 объектов, мультиресурсная часть содержит 8 469 объектов, малоресурсная часть содержит 1 000 объектов, а тестовая часть содержит 3 925 объектов.

Цель эксперимента — сравнить производительность модели ученика, обучающейся без учителя, с учителем и с учителем на другом домене с использованием адаптации домена. Экспериментальная процедура включает три этапа. На первом этапе обучается модель ученика на малоресурсной части и происходит ее тестирование на тестовой части для получения базовой оценки качества. На втором этапе используются метки модели учителя, обученной на мультиресурсной части, для обучения модели ученика. На третьем этапе происходит обучение модели учителя на изображениях из мультиресурсной части и используется

Таблица 5.3: Архитектура модели учителя для эксперимента по мультидоменной дистилляции в задаче компьютерного зрения. Модель представляет собой сверточную нейронную сеть с пятью сверточными слоями и четырьмя полносвязными слоями, общее число параметров составляет 4 455 984.

Слой	Размер входного вектора	Количество параметров
Входной слой	(3, 200, 200)	0
CONV1 (ядро=5)	(24, 196, 196)	1800
POOL1	(24, 98, 98)	0
CONV2 (ядро = 5)	(48, 94, 94)	28800
POOL2	(48, 47, 47)	0
CONV3 (ядро = 8)	(96, 40, 40)	294912
POOL3	(96, 20, 20)	0
CONV4 (ядро = 5)	(192, 16, 16)	460800
POOL4	(192, 8, 8)	0
CONV5 (ядро = 7)	(384, 2, 2)	3612672
POOL5	(384, 1, 1)	0
Полносвязный слой	(384)	0
Полносвязный слой	(120)	46080
Полносвязный слой	(84)	10080
Полносвязный слой	(10)	840
		$\sum = 4455984$

модель учителя и отображение для обучения модели ученика. Предобученная модель CycleGAN [78] используется в качестве отображения φ для доменной адаптации.

На рис. 5.10 показано одно из изображений в наборе данных ImageNet [76] и то же изображение после преобразования с помощью модели CycleGAN [78]. Результаты усредняются по 5 запускам для вычисления среднего значения и дисперсии метрик.

Конфигурация алгоритма многодоменной дистилляции для задачи обработки естественного языка.

Набор данных OPUS-100 был разделен следующим образом: обучающая часть для учителя состояла из 5 000 немецко-английских предложений, обучающая часть для модели ученика — из 2 000 французско-английских предло-

Таблица 5.4: Архитектура модели ученика для эксперимента по мультидоменной дистилляции в задаче компьютерного зрения. Модель имеет упрощенную структуру по сравнению с моделью учителя, с двумя сверточными слоями и тремя полно связанными слоями, общее число параметров составляет 12 755 640.

Слой	Размер входного вектора	Количество параметров
Входной слой	(3, 200, 200)	0
CONV1 (ядро=5)	(24, 196, 196)	1800
POOL1	(24, 98, 98)	0
CONV2 (ядро = 5)	(48, 94, 94)	28800
POOL2	(48, 47, 47)	0
Полносвязный слой	(106032)	0
Полносвязный слой	(120)	12723840
Полносвязный слой	(10)	1200
		$\sum = 12755640$

Таблица 5.5: Характеристики подмножеств набора данных ImageNet, использованных в эксперименте по мультидоменной дистилляции. ImageNet-Big содержит 1 281 167 изображений для обучения и 50 000 для валидации, ImageNet-Small — 64 058 и 2 500 соответственно, оба набора содержат 200 классов.

Набор данных	Описание	Размер набора
ImageNet-Train	Обучающая часть	9469
ImageNet-Big	Мультиресурсная часть	8469
ImageNet-Small	Малоресурсная часть	1000
ImageNet-Test	Тестовая часть	3925

жений, тестовая часть — из 500 французско-английских предложений. Такое разделение позволяет проверить эффективность мультидоменной дистилляции при работе с различными языковыми парами.

В таблице 5.6 описаны наборы данных для вычислительного эксперимента по обработке естественного языка.

Использовались модель ученика **g** и модель учителя **f** в качестве трансформерной модели на основе статьи [79] и метод градиентной оптимизации Adam [57] для решения задачи оптимизации. Модель NLLB [80] использовалась



Рис. 5.10: Пример применения инъективного отображения φ для доменной адаптации между ImageNet-Small и ImageNet-Big. Демонстрирует визуальное преобразование объекта из исходного домена в целевой домен для использования в мультидоменной дистилляции.

Таблица 5.6: Характеристики подмножеств набора данных OPUS-100, использованных в эксперименте по мультидоменной дистилляции для задачи машинного перевода. Показаны размеры обучающих и валидационных выборок для языковых пар fr-en и de-en.

Набор данных	Описание	Язык	Размер набора
Teacher-Train	Обучающая часть модели учителя	de-en	5000
Student-Train	Обучающая часть модели ученика	fr-en	2000
Student-Test	Тестовая часть	fr-en	500

в качестве отображения φ , переводящего французские предложения в немецкие.

Аналогично эксперименту по компьютерному зрению, сравнивается производительность модели ученика без учителя, с учителем и с учителем и адаптацией домена. Результаты усреднялись по 5 запускам для вычисления среднего значения и дисперсии метрик, что обеспечивает статистическую надежность полученных оценок.

На рис. 5.11 и рис. 5.12 представлены результаты обучения моделей на наборе данных ImageNet. Анализ графиков показывает, что модели, обученные с использованием учителя, достигают лучшего качества и точности по сравнению с моделью, обученной без учителя. Модель ученика, обученная с использованием меток учителя на том же домене, достигает наивысшей точности и наименьших потерь, что ожидаемо, поскольку отсутствует необходимость в адаптации домена. Модель ученика, обученная с использованием меток учителя и адапта-

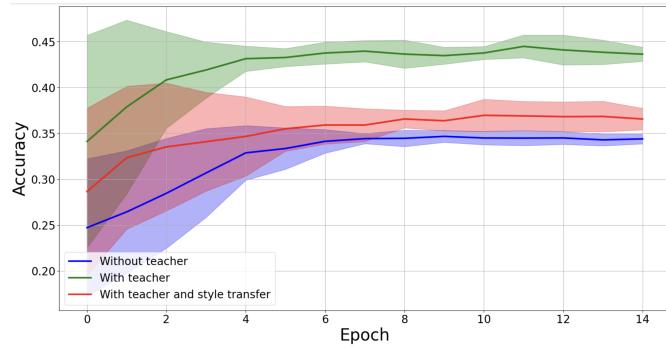


Рис. 5.11: Зависимость точности аппроксимации от числа эпох обучения на тестовой выборке ImageNet для различных методов дистилляции. Результаты усреднены по 5 запускам и демонстрируют преимущество методов с использованием учителя и доменной адаптации.

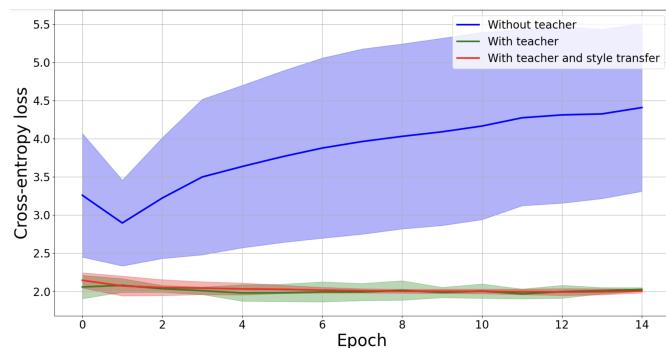


Рис. 5.12: Зависимость ошибки перекрестной энтропии между истинными и предсказанными метками от числа эпох обучения на тестовой выборке ImageNet для различных методов дистилляции. Результаты усреднены по 5 запускам и показывают снижение ошибки при использовании методов дистилляции.

ции домена, демонстрирует существенное улучшение качества аппроксимации по сравнению с моделью без использования учителя, что подтверждает эффективность предложенного метода мультидоменной дистилляции.

Таким образом, экспериментально показано, что дистилляция с использованием адаптации домена приводит к более эффективным нейронным сетям с меньшим количеством параметров, сохраняя при этом высокое качество аппроксимации.

Таблица 5.7: Сравнение качества моделей для задачи компьютерного зрения на наборе данных ImageNet. Показаны валидационная точность, потери и интегральный критерий для различных комбинаций учителя и ученика с использованием доменной адаптации и без нее, результаты усреднены по нескольким запускам.

Ученика	Учитель	Отображение φ	Точность	Потери перекрестной энтропии	Интегральный критерий
ImageNet-Small	—	—	$0,34 \pm 0,01$	$4,41 \pm 1,10$	$53,89 \pm 14,99$
ImageNet-Small	ImageNet-Big	StyleTransfer	$0,37 \pm 0,01$	$2,01 \pm 0,03$	$28,30 \pm 0,79$
ImageNet-Small	ImageNet-Big	—	$0,44 \pm 0,01$	$2,03 \pm 0,02$	$28,08 \pm 1,22$

В таблице 5.7 представлены количественные результаты эксперимента по компьютерному зрению: валидационная точность, потери и интегральный критерий для моделей, обученных с дистилляцией и адаптацией домена и без них.

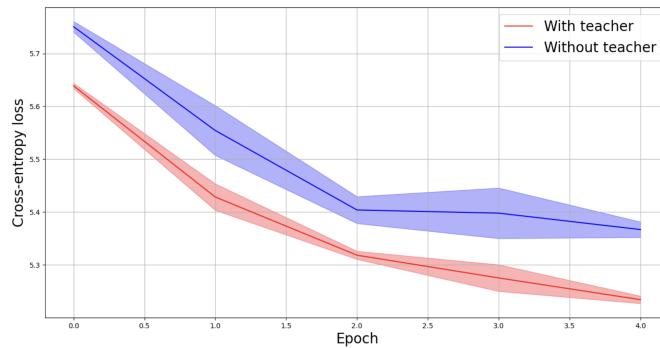


Рис. 5.13: Зависимость ошибки перекрестной энтропии от числа эпох обучения на тестовом наборе данных OPUS-100 для задачи машинного перевода. Результаты усреднены по 3 запускам и демонстрируют преимущество моделей, обученных с использованием учителя.

На рис. 5.13 представлена зависимость ошибки перекрестной энтропии от числа эпох обучения для задачи машинного перевода. Анализ показывает, что

модели, обученные с использованием учителя, достигают лучшего качества, демонстрируя более быстрое снижение ошибки и более низкие значения потерь по сравнению с моделью, обученной без учителя.

В таблице 5.8 представлены количественные результаты сравнения моделей ученика, полученных с использованием дистилляции и без нее. Результаты подтверждают эффективность мультидоменной дистилляции для задачи машинного перевода.

Таблица 5.8: Сравнение качества моделей для задачи машинного перевода на наборе данных OPUS-100. Показаны потери перекрестной энтропии и метрика BLEU для модели ученика, обученной с использованием учителя и доменной адаптации (NLLB) и без них.

Ученик	Учитель	Отображение φ	Потери перекрестной энтропии	BLEU
Student-Train	—	—	$5,367 \pm 0,015$	0,0282
Student-Train	Teacher-Train	NLLB	$5,233 \pm 0,007$	0,0572

5.4.3. Антидистилляция моделей глубокого обучения

В настоящем подразделе представлены результаты вычислительных экспериментов для метода анти-дистилляции, описанного в разделе 5. Цель вычислительного эксперимента — сравнить производительность моделей в зависимости от инициализации параметров.

Производится сравнение различных подходов к инициализации:

1. Xavier — заполнение всех параметров модели $U[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}]$, где n — количество нейронов входного слоя [81], т.е. инициализация параметров модели по умолчанию.
2. Zero pad — заполнение расширенных параметров нулями.
3. Uniform pad — заполнение расширенных параметров равномерно распределенными случайными величинами $U[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}]$, где n — количество нейронов входного слоя.
4. Transfer learning — взятие предобученной модели и изменение только классификационного слоя для новой задачи классификации. Сначала модель обучалась с замороженными параметрами на всех слоях, кроме классификационного. После 3 эпох обучения все параметры размораживались.

Начиная с четвертой эпохи, оптимизировались все параметры нейронной сети.

5. Net2Net — инкрементальный алгоритм расширения пространства параметров модели[82].
6. With Data Noise — получение инициализации модели ученика путем решения задачи оптимизации 5.3 с $\lambda_1, \lambda_3 = 1$ и $\lambda_2, \lambda_4 = 0$.
7. Anti-Distillation, $\lambda_4 = 0$ — инициализация методом Анти-Дистилляции с оптимизацией гиперпараметров $\lambda_1, \lambda_2, \lambda_3$ с помощью байесовской оптимизации ($\lambda_4 = 0$) [83].
8. Anti-Distillation — оптимизация всех λ_i .

Критериями качества являются: точность на валидационном наборе, иска-
женном атакой FSGM [84], и точность на валидационном наборе при условии,
что параметры модели искажены шумом: $\mathbf{w}_\varepsilon = \mathbf{w} + \varepsilon \boldsymbol{\xi}$, где $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

В качестве набора данных используется Fashion-MNIST [85] — набор данных изображений статей Zalando, состоящий из обучающего набора из 60 000 примеров и тестового набора из 10 000 примеров. Каждый пример представляет собой полутоновое изображение размером 28×28 пикселей, связанное с меткой из 10 классов.

Экспериментальная процедура включает следующие этапы: обучение модели учителя, увеличение ее сложности и сравнение различных способов инициализации параметров модели.

Рассматривается модель полносвязной сети. Модель учителя имеет следую-
щие размеры скрытых слоев: [128, 64, 32]. Модель ученика имеет [256, 128, 64]
нейронов в скрытых слоях.

Модель учителя обучалась в течение 30 эпох с начальной скоростью обуче-
ния 10^{-2} , уменьшавшейся до 10^{-3} после 10 эпох. Модели ученика сравнивались
при обучении в течение 10 эпох со скоростью обучения 10^{-3} . Оптимизация про-
водится с использованием алгоритма оптимизации Adam [57]. Методы иници-
ализации сравниваются по следующим критериям: точность предсказаний на
валидационной выборке, значение функции потерь перекрестной энтропии и
дисперсия предсказаний. Дополнительно исследуется устойчивость методов к
зашумленным входным данным, для чего указанные критерии качества анали-
зируются в зависимости от процента искаженных изображений.

Набор данных \mathfrak{D}_2 состоит из Fashion-MNIST, а $\mathfrak{D}_1 = \{(\mathbf{x}, y) | (\mathbf{x}, y) \in \mathfrak{D}_2, y \in C_1\}$, где $C_1 \subset C_2$, $C_1 = \{0, \dots, 4\}$, $C_2 = \{0, \dots, 9\}$. Таким образом, набор дан-
ных \mathfrak{D}_1 представляет собой подмножество \mathfrak{D}_2 , содержащее только первые пять

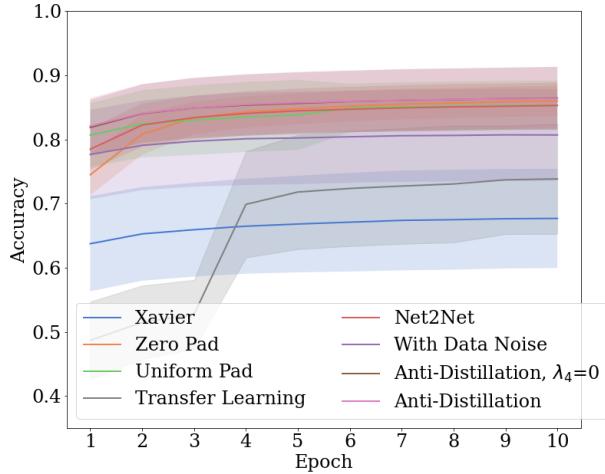


Рис. 5.14: Сравнение валидационной точности для различных методов инициализации параметров модели на наборе данных Fashion-MNIST. Модели, использующие Анти-Дистилляцию, демонстрируют меньшую дисперсию и более высокую точность по сравнению с моделями с различной инициализацией.

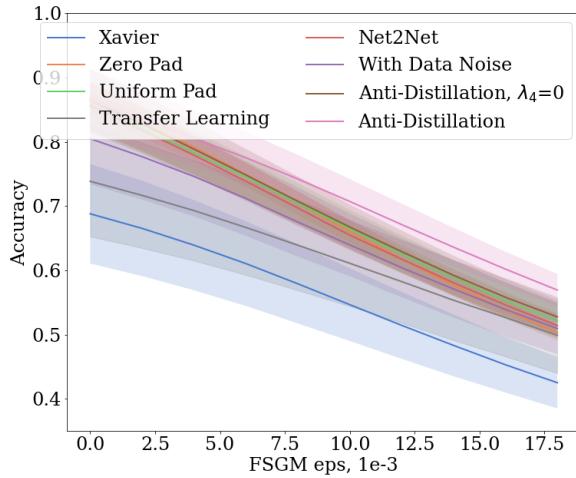


Рис. 5.15: Зависимость валидационной точности от уровня адверсарного шума в данных для различных методов инициализации на наборе данных Fashion-MNIST. Анти-Дистилляция является наиболее устойчивым методом, демонстрируя наивысшую точность при высоких уровнях шума.

классов, что позволяет моделировать сценарий перехода от простой задачи к более сложной.

На рис. 5.14 представлено сравнение валидационной точности для различных методов инициализации параметров модели. Анализ результатов показы-

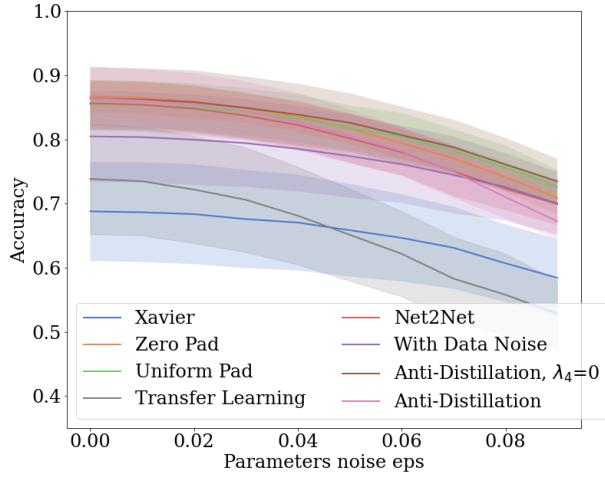


Рис. 5.16: Зависимость валидационной точности от параметра интенсивности нормального шума ϵ в параметрах модели для различных методов инициализации на наборе данных Fashion-MNIST. Метод Анти-Дистилляции без регуляризации гессиана ($\lambda_4 = 0$) является наиболее устойчивым, сохраняя наивысшую точность при максимальном уровне шума.

вает, что модели, использующие анти-дистилляцию, в среднем имеют меньшую дисперсию и более высокую точность, чем модели с различной инициализацией параметров. Обучение модели с нуля оказалось наименее эффективным решением. Предложенный метод анти-дистилляции обеспечивает лучшие результаты с меньшим количеством итераций для сходимости, что указывает на эффективность использования информации от предобученной модели учителя.

В представленных экспериментах не учитывалось количество итераций, необходимых для расширения модели учителя, которое также требует процедуры оптимизации. Предполагается, что во многих реальных случаях этим временем можно пренебречь, поскольку предложенный метод позволяет расширить модель учителя один раз, используя только базовый набор данных \mathcal{D}_1 , для последующего использования в множественных задачах обучения модели ученика [86].

На рис. 5.15 представлена зависимость валидационной точности от уровня адвокарного шума в данных. Анализ показывает, что Анти-Дистилляция является наиболее устойчивым к адвокарным атакам методом инициализации параметров модели, поскольку демонстрирует наивысшую валидационную точность с большим отрывом при высоких уровнях шума.

На рис. 5.16 представлена зависимость валидационной точности от уровня

нормального шума в параметрах модели. Анализ показывает, что метод Анти-Дистилляции без регуляризации гессиана ($\lambda_4 = 0$) является наиболее устойчивым к нормальному шуму в параметрах модели, поскольку сохраняет наивысшую точность при максимальном рассматриваемом уровне шума.

Таблица 5.9: Сравнение точности на валидационном наборе Fashion-MNIST для различных методов инициализации параметров модели. Показаны базовая точность, точность при адвверсарной атаке FSGM и точность при нормальном шуме в параметрах, результаты усреднены по нескольким запускам.

Метод инициализации	Точность	Атака FSGM	Шум в параметрах
Xavier	0.68 ± 0.08	0.42 ± 0.04	0.58 ± 0.06
Zero Pad	0.86 ± 0.02	0.50 ± 0.01	0.71 ± 0.03
Uniform Pad	0.85 ± 0.04	0.52 ± 0.03	0.73 ± 0.03
Transfer Learning	0.74 ± 0.09	0.50 ± 0.06	0.53 ± 0.05
Net2Net	0.85 ± 0.04	0.51 ± 0.02	0.70 ± 0.03
With Data Noise	0.81 ± 0.07	0.51 ± 0.03	0.70 ± 0.05
Anti-Distillation, $\lambda_4=0$	0.86 ± 0.05	0.53 ± 0.03	0.73 ± 0.04
Anti-Distillation	0.86 ± 0.05	0.57 ± 0.03	0.67 ± 0.03

В таблице 5.9 представлены количественные результаты сравнения методов инициализации: валидационная точность после последней эпохи обучения, точность при наивысшем уровне шума изображения от адвверсарной атаки FSGM и точность при наивысшем уровне шума в параметрах модели.

5.5. Заключение по главе

В настоящей главе рассмотрены методы снижения сложности параметрических моделей глубокого обучения, опирающиеся на теоретический аппарат, введенный в главах 2 и 3. Предложены три класса методов: удаление параметров на основе анализа ковариационной матрицы градиентов, мультидоменная дистилляция и анти-дистилляция. Все методы направлены на практическое применение теории сложности моделей для уменьшения числа параметров при сохранении качества и повышении устойчивости моделей.

В разделе об удалении параметров предложен метод, основанный на анализе ковариационной матрицы \mathbf{C} градиентов функции ошибки по параметрам модели, вычисляемой итеративно в процессе оптимизации. Метод позволяет

упорядочить параметры по их важности на основе диагональных элементов ковариационной матрицы и последовательно удалять наименее значимые параметры без существенной потери качества модели. Дополнительно разработан метод, основанный на модификации метода Белсли, использующий сингулярное разложение ковариационной матрицы и анализ дисперсионных долей для выявления мультиколлинеарных параметров. Экспериментальные результаты на наборах данных Wine, Boston Housing и синтетических данных с мультиколлинеарными параметрами продемонстрировали эффективность предложенных подходов. Метод оптимального прореживания, вариационный метод и метод Белсли позволяют удалить до $\approx 80\%$ параметров без существенной потери качества, при этом метод Белсли показал наилучшие результаты по устойчивости к шуму во входных данных.

Мультидоменная дистилляция решает задачу передачи знаний между моделями, обученными на данных из различных доменов, связанных инъективным отображением φ . Предложена функция потерь, учитывающая как метки учителя, так и связь между доменами через композицию $(f \circ \varphi)$, что позволяет эффективно использовать информацию от модели учителя, обученной на другом домене. Эксперименты в области компьютерного зрения на наборе данных ImageNet подтвердили улучшение качества аппроксимации модели ученика при использовании доменной адаптации, при этом обучение в рамках одного домена показало лучшие результаты, однако адаптация домена обеспечила значительное улучшение по сравнению с базовыми подходами. Эксперименты в области обработки естественного языка на наборе данных OPUS-100 для задачи машинного перевода продемонстрировали снижение ошибки перекрестной энтропии и увеличение метрики BLEU при использовании мультидоменной дистилляции.

Анти-дистилляция решает обратную задачу по сравнению с классической дистилляцией: передачу знаний от простой модели к более сложной для работы с более сложными наборами данных. Предложена составная функция потерь, включающая четыре компонента: перекрестную энтропию на простых данных, регуляризацию близости параметров учителя и ученика, регуляризацию устойчивости к шуму во входных данных и регуляризацию гессиана функции потерь. Для эффективного вычисления следа гессиана использован метод стохастической аппроксимации с быстрым умножением гессиан-вектор, обеспечивающий линейную сложность от числа параметров. Эксперименты на наборе данных Fashion-MNIST продемонстрировали, что модели, инициализированные методом анти-дистилляции, достигают валидационной точности $0,86 \pm 0,05$, что пре-

восходит результаты методов Xavier ($0,68 \pm 0,08$), Transfer Learning ($0,74 \pm 0,09$) и других базовых подходов. Кроме того, анти-дистилляция показала наивысшую устойчивость к адвверсарным атакам FSGM (точность $0,57 \pm 0,03$ при высоких уровнях шума) и к нормальному шуму в параметрах модели (точность $0,73 \pm 0,04$ при максимальном уровне шума для варианта без регуляризации гессиана).

Рассмотренные методы открывают перспективы для практического применения в условиях ограниченных данных и необходимости адаптации моделей к новым доменам. Разработанные подходы обеспечивают эффективное снижение сложности моделей при сохранении их качества и повышении устойчивости, что особенно важно для развертывания моделей в ресурсно-ограниченных средах. Дальнейшие исследования будут направлены на интеграцию байесовских методов дистилляции, учитывающих распределения параметров, на применение разработанных подходов к другим архитектурам нейронных сетей и наборам данных, а также на теоретический анализ связи между предложенными методами и ландшафтной мерой сложности, введенной в главе 2.

Глава 6

Применение теоретических оценок в прикладных задачах

В предыдущих главах был разработан единый теоретический аппарат для формализации соотношения между сложностью модели и сложностью данных. В главе 2 введены формальные определения меры сложности выборки $\mu_D(D)$ и меры сложности модели $\mu_f(f)$, а также установлен критерий обучаемости $\mu_f(f) \leq \mu_D(D)$. Кроме того, разработана ландшафтная мера сложности модели $\mu_f(f|D)$, определяемая через спектральные свойства матриц Гессе функции потерь. В главе 3 получены конкретные оценки спектральных норм матриц Гессе для различных архитектур нейронных сетей, что обеспечивает вычислимо осуществимые методы оценки ландшафтной меры сложности. В главе 4 разработаны практические методы определения достаточного размера выборки на основе анализа стабильности процесса обучения. В главе 5 предложены методы снижения сложности моделей через удаление параметров, дистилляцию и анти-дистилляцию.

Однако для полной валидации разработанного теоретического аппарата и демонстрации его практической значимости необходимо применить полученные результаты к реальным прикладным задачам машинного обучения. В отличие от предыдущих глав, где основной акцент делался на разработке строгих математических оценок и алгоритмических процедур, в настоящей главе демонстрируется адаптация теоретических подходов к конкретным практическим проблемам. Это позволяет оценить эффективность предложенного формализма в условиях реальных данных и ограниченных вычислительных ресурсов.

Настоящая глава охватывает три ключевых направления практического применения теоретического аппарата сложности моделей и данных. Во-первых, рассматривается применение меры сложности Радемахера для анализа многозадачного обучения с адаптерами LoRA, что позволяет количественно оценить преимущества совместного использования параметров энкодера и эффективность низкоранговой параметризации. Во-вторых, демонстрируется снижение сложности данных в задаче декодирования фМРТ-изображений из видеопоследовательностей, где предварительное сжатие данных обеспечивает существенное сокращение времени обучения без потери качества реконструкции. В-третьих, разрабатываются методы оценки качества данных в задаче детекции машино-генерированного текста, основанные на топологической статистике и анализе устойчивости к адверсарным возмущениям, что позволяет выявить сме-

щения в обучающих наборах данных и улучшить надежность детекторов.

Полученные результаты создают основу для практического применения теоретического аппарата сложности моделей и данных в реальных задачах машинного обучения, демонстрируя эффективность предложенных подходов и открывая перспективы для дальнейшего развития методов управления сложностью в прикладных задачах.

6.1. Сложность моделей в многозадачном обучении

6.1.1. Радемахеровская сложность моделей глубокого обучения

Пусть \mathcal{X} обозначает входное пространство, а $\mathcal{Y} = \{1, \dots, C\}$ — общее пространство меток для всех доменов. Домен определяется как пара $\mathcal{D} = (\mathcal{X}, P(X))$, где $P(X)$ — распределение над \mathcal{X} .

Предполагается наличие K размеченных исходных доменов

$$\mathcal{D}_{S_k} = (\mathcal{X}, P_{S_k}(X)), \quad k = 1, \dots, K,$$

с размеченными выборками

$$\mathcal{S}_k = \{(x_i^{S_k}, y_i^{S_k})\}_{i=1}^{n_{S_k}}, \quad x_i^{S_k} \sim P_{S_k}(X), y_i^{S_k} \in \mathcal{Y}.$$

Также рассматривается целевой домен

$$\mathcal{D}_T = (\mathcal{X}, P_T(X)),$$

с выборками

$$\mathcal{T} = \{(x_j^T, y_j^T)\}_{j=1}^{n_T^{\text{lab}}} \cup \{x_j^T\}_{j=1}^{n_T^{\text{unlab}}}.$$

Цель кросс-доменной классификации — обучить классификатор

$$f_\theta : \mathcal{X} \rightarrow \mathcal{Y}, \quad f_\theta(x) = \arg \max_{y \in \mathcal{Y}} p_\theta(y \mid x),$$

такой, чтобы ожидаемая целевая ошибка

$$\epsilon_T(f_\theta) = \mathbb{E}_{(x,y) \sim P_T(X,Y)} [\mathbf{1}\{f_\theta(x) \neq y\}]$$

минимизировалась за счет использования знаний из всех источников $\{\mathcal{S}_k\}_{k=1}^K$. Цель состоит в обучении классификатора, который способен достигать высоких метрик качества на примерах из доменов, не встречавшихся во время обучения или представленных лишь ограниченными данными.

Таблица 6.1: Сравнение формул обновления параметров для полной тонкой настройки и метода LoRA. В полной тонкой настройке обновление ΔW изучается напрямую, тогда как в LoRA обновление параметризуется в виде низкоранговой факторизации AB , где $r \ll \min(d, k)$, что значительно сокращает количество обучаемых параметров.

Полная тонкая настройка	Тонкая настройка с LoRA
$W_{\text{upd}} = W + \Delta W$	$W_{\text{upd}} = W + AB$
$\hat{y} = x(W + \Delta W)$	$\hat{y} = x(W + AB)$

Современные задачи классификации часто решаются путем адаптации больших предобученных моделей к конкретному домену или набору данных [87]. Распространенная стратегия — тонкая настройка (fine-tuning), когда параметры предобученной модели обновляются с использованием размеченных данных из целевой задачи. Хотя этот подход обычно дает высокую производительность, он может быть вычислительно дорогим, поскольку количество обучаемых параметров очень велико.

Основная идея LoRA заключается в том, что для адаптации к задаче не обязательно обновлять всю матрицу параметров. Вместо этого обновления могут быть эффективно представлены в низкоразмерном подпространстве, которое захватывает существенные вариации, необходимые для адаптации. Это значительно сокращает количество обучаемых параметров, в значительной степени сохраняя выразительную способность модели.

Формально, пусть $\Delta W \in \mathbb{R}^{d \times k}$ обозначает обновление параметров для данного слоя. При стандартной тонкой настройке ΔW изучается напрямую. LoRA ограничивает ΔW условием низкого ранга путем факторизации:

$$\Delta W \approx AB,$$

где $A \in \mathbb{R}^{d \times r}$ и $B \in \mathbb{R}^{r \times k}$ с $r \ll \min(d, k)$. Обновленная матрица параметров тогда имеет вид:

$$W_{\text{upd}} = W + \Delta W = W + AB,$$

где W обозначает замороженные предобученные веса, а AB — изучаемое низкоранговое обновление. Ранг r выступает в качестве гиперпараметра, контролирующего размер изучаемого подпространства.

В задачах классификации предсказание обычно имеет вид

$$p(c \mid \mathbf{x}) = \text{softmax}(W^\top \mathbf{x}),$$

где \mathbf{x} — это вектор признаков входного объекта, а W — обучаемая матрица весов. LoRA может быть применена здесь путем замены W на ее низкоранговую адаптированную форму $W + AB$.

Многозадачное обучение — это подход, в котором несколько связанных задач изучаются совместно с целью достижения лучшей обобщающей способности по сравнению с обучением каждой задачи независимо. Ключевое предположение заключается в том, что задачи разделяют некоторую базовую структуру, так что обмен информацией между ними уменьшает переобучение и улучшает прогнозную производительность.

Формально, пусть $\{(x_{t,i}, y_{t,i})\}_{i=1}^{n_t}$ обозначает обучающие данные для задачи $t \in \{1, \dots, T\}$, взятые из распределения P_t над $\mathcal{X} \times \mathcal{Y}$. Цель — обучить дискриминативные функции $f_t : \mathcal{X} \rightarrow \mathcal{Y}$ для всех задач. Распространенный подход формулирует это как задачу регуляризованной минимизации эмпирического риска:

$$\min_{f=(f_1, \dots, f_T)} \frac{1}{T} \sum_{t=1}^T \frac{1}{n_t} \sum_{i=1}^{n_t} \ell(f_t(x_{t,i}), y_{t,i}) + \lambda \Omega(f),$$

где первое слагаемое — это средняя эмпирическая ошибка по задачам, а $\Omega(f)$ — регуляризующее слагаемое, поощряющее обмен информацией.

С теоретической точки зрения, преимущество MTL может быть интерпретировано как эффективное сокращение пространства гипотез за счет использования связности задач, что приводит к более узким обобщающим границам.

Традиционные методы машинного обучения часто обучаются в однозадачной постановке. В отличие от этого, многозадачное обучение (MTL) совместно обучается на нескольких задачах с целью улучшения общей производительности за счет использования общих представлений [88].

Такая постановка особенно эффективна, когда задачи связаны, так как передача информации может уменьшить переобучение и улучшить обобщение. Кроме того, даже для одной целевой задачи вспомогательные задачи могут служить формой индуктивного смещения, помогая модели изучать более устойчивые представления.

Центральный вопрос в многозадачном обучении, как и в стандартном обучении с учителем, заключается в том, насколько хорошо изученные функции

обобщаются на новые данные. Границы обобщающей ошибки предоставляют теоретические гарантии, связывая истинный риск с его эмпирическим аналогом. В постановке MTL границы часто выводятся в рамках упомянутого выше подхода регуляризованной минимизации эмпирического риска.

Пусть даны обучающие примеры $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$, взятые i.i.d. из неизвестного распределения P . Цель — обучить функцию $f \in \mathcal{F}$ с малой ожидаемой потерей $\mathbb{E}_{(X,Y) \sim P}[\ell(f(X), Y)]$. Поскольку P неизвестно, обычно связывают этот риск со средним эмпирическим значением функции потерь плюс слагаемое сложности, которое зависит от богатства гипотезного класса \mathcal{F} :

$$\begin{aligned}\mathbb{E}_{(X,Y) \sim P}[\ell(f(X), Y)] &\leq \frac{1}{n} \sum_{i=1}^n \ell(f(X_i), Y_i) \\ &+ h(\text{complexity of } \mathcal{F}, n).\end{aligned}$$

Среди различных мер сложности широко используется сложность Радемахера. Эта мера количественно определяет, насколько хорошо гипотезный класс может аппроксимировать случайный шум, что напрямую связано с обобщающей способностью модели. Формально:

Определение 28 (Сложность Радемахера). Пусть $\mathcal{G} := \{g : \mathcal{Z} \rightarrow \mathbb{R}\}$ — гипотезный класс, а $S := \{z_1, \dots, z_n\}$ — i.i.d. выборка из P . Эмпирическая сложность Радемахера для \mathcal{G} определяется как

$$\widehat{R}(\mathcal{G}) := \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i g(z_i) \right],$$

где σ_i — независимые случайные величины Радемахера, равномерно распределенные в $\{\pm 1\}$. (Ожидаемая) сложность Радемахера определяется как

$$R(\mathcal{G}) := \mathbb{E}_{S \sim P^n} [\widehat{R}(\mathcal{G})].$$

Интуитивно, меньшая сложность Радемахера соответствует менее выразительному предсказательному классу, что, в свою очередь, приводит к более узким границам обобщения. В настоящем исследовании выполняется сопоставление сложности модели в условиях многозадачного и однозадачного обучения. В отличие от других работ, мы не сравниваем различные типы границ обобщения, а фокусируемся на сравнении сложности моделей.

Для эмпирической проверки нашего теоретического анализа сфокусируемся на задаче обнаружения машинно-генерированного текста. Эта задача требует

высокой точности классификации с низкой частотой ложных срабатываний, чтобы избежать некорректного помечания контента, написанного человеком. Стремительное улучшение больших языковых моделей (LLMs) сделало тексты, сгенерированные ИИ, все более неотличимыми от написанных человеком [89, 90], создавая насущную потребность в устойчивых методах обнаружения для сохранения академической честности и борьбы с дезинформацией.

Большинство подходов рассматривают это как задачу бинарной классификации текста [91, 92]. Тонкая настройка архитектур на основе Transformer стала доминирующим решением [93], демонстрируя высокую производительность, когда тестовые данные поступают из того же или близкородственного домена. Однако производительность часто резко падает при появлении новых доменов, делая модели менее надежными [94]. Чтобы решить эту проблему, можно либо использовать более мощные модели, либо интегрировать дополнительную информацию о тексте, такую как его источник, домен или стилистические особенности. Эти сигналы могут выступать в качестве вспомогательных регуляризаторов, улучшая обобщение. Многозадачное обучение предоставляет естественную основу для объединения нескольких задач, включая интеграцию такой вспомогательной информации, при этом сохраняя основную цель классификации.

6.1.2. Радемахеровская сложность многозадачного обучения в LoRA-адаптерах

Предполагается, что решается задача классификации с использованием моделей с архитектурой Transformer [79]. Тонкая настройка таких моделей в режиме обучения с учителем является современным методом для задач классификации.

Анализ LoRA показывает, что минимизация эмпирического риска при параметризации LoRA остается согласованной с минимизацией истинного риска, гарантируя, что низкоранговая адаптация сохраняет асимптотические статистические гарантии полной тонкой настройки.

Теорема 32 (*Грабовой-Грицай-Ремизова*, Состоятельность оценки). *Пусть $\mathcal{D} = \{(X_i, c_i)\}_{i=1}^n$ — независимые одинаково распределенные выборки из истинного распределения P_{true} , где $c_i \in [N_c]$ обозначает метки классов. Предположим следующее:*

1. *Существует параметр Θ^* такой, что модель распределения $P_{\text{model}}(\cdot | \Theta)$ аппроксимирует истинное распределение с минимальным расхождением*

Кульбака-Лейблера:

$$\Theta^* \in \arg \min_{\Theta} D_{\text{KL}}(P_{\text{true}} \| P_{\text{model}}(\cdot | \Theta)).$$

2. При $n \rightarrow \infty$ эмпирическое распределение сходится по вероятности к P_{true} .
3. Функция потерь задается как отрицательное логарифмическое правдоподобие

$$\mathcal{L}_n(\Theta) = -\frac{1}{n} \sum_{i=1}^n \log \left(P_{\Phi_0 + \Theta}(c_i | X_i) \right),$$

где Φ_0 — замороженные предобученные веса, а Θ соответствует обучаемым низкоранговым параметрам в LoRA. Предполагается, что $\mathcal{L}_n(\Theta)$ непрерывна и дифференцируема.

Тогда минимизация эмпирического риска является состоятельной:

$$\lim_{n \rightarrow \infty} \arg \min_{\Theta} \mathcal{L}_n(\Theta) = \Theta^*.$$

Доказательство. Пусть истинный риск и его эмпирический аналог определены как

$$L(\Theta) = \mathbb{E}_{(X, c) \sim P_{\text{true}}} [\mathcal{L}(X, c; \Theta)], \quad \widehat{L}_n(\Theta) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(X_i, c_i; \Theta),$$

где $\mathcal{L}(X_i, c_i; \Theta) = -\log P_{\Phi_0 + \Theta}(c_i | X_i)$ — отрицательное логарифмическое правдоподобие для отдельного примера.

В соответствии с равномерным законом больших чисел, для непрерывной и ограниченной функции \mathcal{L} имеем

$$\sup_{\Theta} | L(\Theta) - \widehat{L}_n(\Theta) | \xrightarrow[n \rightarrow \infty]{p} 0.$$

Следовательно, последовательность эмпирических рисков $\widehat{L}_n(\Theta)$ сходится равномерно к истинному риску $L(\Theta)$. Равномерная сходимость влечет состоятельность минимизации эмпирического риска, т.е.

$$\arg \min_{\Theta} \widehat{L}_n(\Theta) \xrightarrow{n \rightarrow \infty} \arg \min_{\Theta} L(\Theta).$$

Из определения $L(\Theta)$ и с использованием тождества

$$D_{\text{KL}}(P_{\text{true}} \| P_{\text{model}}(\cdot | \Theta)) = \mathbb{E}_{(X, c) \sim P_{\text{true}}} \left[\log \frac{P_{\text{true}}(c | X)}{P_{\text{model}}(c | X; \Theta)} \right],$$

минимизация ожидаемых потерь $L(\Theta)$ эквивалентна минимизации расхождения Кульбака-Лейблера между P_{true} и $P_{\text{model}}(\cdot \mid \Theta)$. Согласно Предположению (1), минимизатором этого расхождения является Θ^* . Следовательно,

$$\lim_{n \rightarrow \infty} \arg \min_{\Theta} \widehat{L}_n(\Theta) = \arg \min_{\Theta} L(\Theta) = \Theta^*,$$

что завершает доказательство. \square

Теорема 33 (*Грабовой-Грицай-Ремизова*, Корректность при низкоранговых обновлениях). *Предположим следующее:*

1. *Выходной слой задается в виде*

$$\hat{\mathbf{y}} = \text{softmax}(W_{\text{upd}}^\top \mathbf{x}),$$

где $\mathbf{x} \in \mathbb{R}^d$ — это вектор признаков BERT, $W \in \mathbb{R}^{d \times k}$ — замороженные предобученные веса, а

$$W_{\text{upd}} = W + \Delta W.$$

2. *Вместо непосредственного обучения $\Delta W \in \mathbb{R}^{d \times k}$, обновление параметризуется в низкоранговой форме:*

$$\Delta W = AB, \quad A \in \mathbb{R}^{d \times r}, B \in \mathbb{R}^{r \times k}, r \ll \min(d, k). \quad (6.1)$$

3. *Выполнены условия Теоремы 32, т.е. модель остается статистически состоятельной при минимизации эмпирического риска.*

Тогда при параметризации (6.1) выход модели $\hat{\mathbf{y}}$ сохраняется в том смысле, что низкоранговое обновление не искажает корректность классификационного слоя.

Доказательство. Пусть выходной слой определен как

$$\hat{\mathbf{y}} = \text{softmax}(W_{\text{upd}}^\top \mathbf{x}), \quad W_{\text{upd}} = W + \Delta W,$$

где $W \in \mathbb{R}^{d \times k}$ — замороженные предобученные веса, а $\mathbf{x} \in \mathbb{R}^d$ — вектор признаков энкодера. В силу дистрибутивности матричного сложения,

$$W_{\text{upd}}^\top \mathbf{x} = W^\top \mathbf{x} + \Delta W^\top \mathbf{x}.$$

Следовательно, предсказанные вероятности могут быть записаны как

$$\hat{\mathbf{y}} = \frac{\exp(W^\top \mathbf{x} + \Delta W^\top \mathbf{x})}{\sum_{i=1}^k \exp((W^\top \mathbf{x} + \Delta W^\top \mathbf{x})_i)}. \quad (6.2)$$

При параметризации LoRA $\Delta W = AB$ с $A \in \mathbb{R}^{d \times r}$ и $B \in \mathbb{R}^{r \times k}$, член обновления принимает вид

$$\Delta W^\top \mathbf{x} = (AB)^\top \mathbf{x} = B^\top (A^\top \mathbf{x}) \in \mathbb{R}^k.$$

Это показывает, что низкоранговая форма лишь ограничивает ΔW рангом r в подпространстве $\mathbb{R}^{d \times k}$, но не изменяет выходную размерность или отображение $\mathbf{x} \mapsto \hat{\mathbf{y}}$. Логиты в (6.2) остаются хорошо определенными и дифференцируемыми для всех \mathbf{x} .

Более того, поскольку обновление ΔW изучается через минимизацию эмпирического риска, согласованную с Теоремой 1, результирующие параметры (A, B) дают тот же минимизатор истинного ожидаемого риска, что и неограниченный ΔW . Следовательно, низкоранговая параметризация сохраняет корректность выходного слоя в том смысле, что $\hat{\mathbf{y}}_{\text{LoRA}} = \text{softmax}((W + AB)^\top \mathbf{x})$ порождает то же решающее правило, что и $\text{softmax}((W + \Delta W)^\top \mathbf{x})$. Таким образом, введение модулей LoRA не искажает выходное распределение или классификационные границы модели. \square

Следствие 5 (LoRA с теоретическими гарантиями). *Объединяя Теорему 32 и Теорему 33, мы получаем, что тонкая настройка на основе LoRA одновременно сохраняет:*

1. *Статистическую состоятельность: минимизация эмпирического риска сходится к минимизатору истинного риска с ростом объема выборки;*
2. *Корректность выходного слоя: низкоранговые обновления AB не искажают выход классификации.*

Таким образом, адаптация с помощью LoRA наследует те же асимптотические гарантии, что и полная тонкая настройка, при этом требуя значительно меньшего количества обучаемых параметров.

Приведенные выше результаты показывают, что LoRA не ослабляет теоретические основы тонкой настройки. Со статистической точки зрения, метод сохраняет состоятельность: с увеличением количества обучающих примеров адаптированная модель сходится к тому же оптимальному решению, что и при полной тонкой настройке. С функциональной точки зрения, ограничение обновлений низкоранговым подпространством не искажает выход классификационного слоя. В совокупности эти свойства объясняют, почему LoRA достигает сопоставимой производительности с полной тонкой настройкой на практике, будучи при этом существенно более эффективной.

В настоящем разделе анализируется влияние многозадачного обучения на сложность обобщения для каждой задачи. Для этого выполняется сравнение однозадачного обучения модели на основе Transformer с MTL-подходом, использующим общий энкодер. Для обеспечения справедливого сравнения предполагается, что архитектуры энкодера и головы идентичны и состоят из линейных слоев. Это позволяет изолировать эффект многозадачного обучения от влияния различий в архитектуре. В случае STL присутствует единственная голова для целевой задачи, тогда как в настройке MTL вводятся дополнительные головы для связанных задач, обеспечивая неявную регуляризацию для целевой задачи.

Формально, в случаях STL и MTL гипотезные классы имеют вид:

$$\begin{aligned}\mathcal{F}_{\text{STL}} &= \left\{ x \mapsto w_{\text{head}}^\top \phi(x; w_{\text{enc}}) \right. \\ &\quad \left. \mid w_{\text{enc}} \in \mathcal{W}_{\text{enc}}, w_{\text{head}} \in \mathcal{W}_{\text{head}} \right\}, \\ \mathcal{F}_{\text{MTL}} &= \left\{ (x \mapsto w_t^\top \phi(x; w_{\text{shared}}))_{t=1}^T \right. \\ &\quad \left. \mid w_{\text{shared}} \in \mathcal{W}_{\text{shared}}, w_t \in \mathcal{W}_{\text{head}} \right\}.\end{aligned}$$

Теорема 34 (Сложность Радемахера на задачу при MTL). *Пусть $S_t = \{x_i\}_{i=1}^n$ — выборка фиксированной целевой задачи t с $\|x_i\|_2 \leq R$. Пусть $\phi(\cdot; w)$ — энкодер, и рассмотрим линейные головы $f_{w,h}(x) = h^\top \phi(x; w)$ с $\|h\|_2 \leq B_{\text{head}}$. Предположим:*

1. *Ограничение на признаки: для всех x, w $\|\phi(x; w)\|_2 \leq L \|w\| \|x\|_2$.*
2. *Энкодер STL удовлетворяет $\|w_{\text{enc}}\| \leq B_{\text{enc}}$, общий энкодер MTL удовлетворяет $\|w_{\text{shared}}\| \leq B_{\text{shared}}$.*
3. *Многозадачное масштабирование: $B_{\text{shared}} \leq B_{\text{enc}}/\sqrt{T}$.*

Обозначим через $\widehat{\mathfrak{R}}_n(\cdot; S_t)$ эмпирическую сложность Радемахера на S_t . Тогда

$$\widehat{\mathfrak{R}}_n(\mathcal{F}_{\text{MTL}}^{(t)}; S_t) \leq \frac{1}{\sqrt{T}} \widehat{\mathfrak{R}}_n(\mathcal{F}_{\text{STL}}^{(t)}; S_t).$$

Доказательство. Пусть $S_t = \{x_i\}_{i=1}^n$ — выборка для целевой задачи t с $\|x_i\|_2 \leq R$. Для $B > 0$ определим гипотезный класс

$$\mathcal{F}(B) = \left\{ x \mapsto h^\top \phi(x; w) \mid \|h\|_2 \leq B_{\text{head}}, \|w\|_2 \leq B \right\}.$$

Эмпирическая сложность Радемахера на S_t равна

$$\widehat{\mathfrak{R}}_n(\mathcal{F}(B); S_t) = \frac{1}{n} \mathbb{E}_\sigma \sup_{\|h\| \leq B_{\text{head}}, \|w\| \leq B} \sum_{i=1}^n \sigma_i h^\top \phi(x_i; w),$$

где $\sigma_i \in \{\pm 1\}$ — независимые одинаково распределенные величины Радемахера.

Для фиксированных w и σ , по дуальности Коши-Буняковского,

$$\sup_{\|h\| \leq B_{\text{head}}} \sum_{i=1}^n \sigma_i h^\top \phi(x_i; w) = B_{\text{head}} \left\| \sum_{i=1}^n \sigma_i \phi(x_i; w) \right\|_2.$$

Следовательно,

$$\widehat{\mathfrak{R}}_n(\mathcal{F}(B); S_t) \leq \frac{B_{\text{head}}}{n} \mathbb{E}_\sigma \sup_{\|w\| \leq B} \left\| \sum_{i=1}^n \sigma_i \phi(x_i; w) \right\|_2.$$

По неравенствам Йенсена и Хинчина,

$$\mathbb{E}_\sigma \left\| \sum_{i=1}^n \sigma_i a_i \right\|_2 \leq \left(\sum_{i=1}^n \|a_i\|_2^2 \right)^{1/2},$$

что дает

$$\widehat{\mathfrak{R}}_n(\mathcal{F}(B); S_t) \leq \frac{B_{\text{head}}}{n} \sup_{\|w\| \leq B} \left(\sum_{i=1}^n \|\phi(x_i; w)\|_2^2 \right)^{1/2}.$$

Используя ограничение на признаки $\|\phi(x; w)\|_2 \leq L \|w\|_2 \|x\|_2$ и $\|x_i\|_2 \leq R$, получаем

$$\widehat{\mathfrak{R}}_n(\mathcal{F}(B); S_t) \leq \frac{B_{\text{head}} L R B}{\sqrt{n}}.$$

Применяя это с $B = B_{\text{enc}}$ для STL и $B = B_{\text{shared}}$ для MTL, находим

$$\widehat{\mathfrak{R}}_n(\mathcal{F}(B_{\text{shared}}); S_t) \leq \frac{B_{\text{head}} L R B_{\text{shared}}}{\sqrt{n}} \leq \frac{1}{\sqrt{T}} \frac{B_{\text{head}} L R B_{\text{enc}}}{\sqrt{n}} = \frac{1}{\sqrt{T}} \widehat{\mathfrak{R}}_n(\mathcal{F}(B_{\text{enc}}); S_t),$$

где последнее неравенство следует из предположения многозадачного масштабирования $B_{\text{shared}} \leq B_{\text{enc}}/\sqrt{T}$. \square

Полученный результат указывает на снижение эмпирической сложности Радемахера на задачу в $1/\sqrt{T}$ раз при справедливом масштабировании. Это согласуется с современными анализами многозадачного обучения, основанными на средних Радемахера, как показано в [95].

Данный результат имеет важное практическое значение: при совместном использовании параметров энкодера между T задачами эффективная сложность модели на каждую задачу снижается пропорционально $1/\sqrt{T}$, что количественно объясняет преимущества многозадачного обучения в терминах меры сложности Радемахера. Однако более ранняя работа [96] установила более сильное улучшение типа $1/T$ в другой постановке. В постановке Бакстера, количество

задач T само по себе способствует оценке общего индуктивного смещения: с ростом T сложность данных на задачу уменьшается пропорционально $1/T$. В отличие от этого, наш анализ сохраняет размер выборки целевой задачи n фиксированным и сравнивает STL и MTL на одном и том же n , поэтому улучшение проявляется как множитель $1/\sqrt{T}$ в сложности Радемахера на задачу.

Обе сложности на задачу оцениваются на одной и той же выборке S_t размера n , поэтому общий множитель $1/\sqrt{n}$ сокращается; разница обусловлена только бюджетом энкодера.

Более того, в анализе обеспечивается, чтобы обе модели STL и MTL обучались на одинаковом общем количестве примеров (nT). Для STL энкодер обучается на nT примерах только из целевой задачи. Для MTL каждая из T голов получает n примеров из своей соответствующей задачи, так что общий энкодер видит те же самые nT примеров, причем одна из голов соответствует целевой задаче. Это гарантирует, что наблюдаемое снижение сложности на задачу обусловлено совместным использованием параметров, а не неравными бюджетами данных.

Наш теоретический анализ установил, что LoRA сохраняет статистическую состоятельность полной тонкой настройки и поддерживает корректность выходного слоя, в то время как многозадачное обучение при соответствующем масштабировании снижает сложность Радемахера на задачу в $1/\sqrt{T}$ раз. Вместе эти результаты показывают, что оба подхода сохраняют фундаментальные гарантии классической тонкой настройки, но достигают большей эффективности с точки зрения параметров (LoRA) или обобщения (MTL).

Полученные теоретические результаты создают основу для практического применения LoRA и многозадачного обучения в реальных задачах, обеспечивая теоретическое обоснование их эффективности и количественную оценку преимуществ по сравнению с полной тонкой настройкой.

6.2. Снижение сложности данных в задаче декодирования фМРТ-снимков

В предыдущем разделе рассматривалось снижение сложности моделей через низкоранговую параметризацию и многозадачное обучение. В настоящем разделе демонстрируется альтернативный подход — снижение сложности данных без изменения архитектуры модели. Такой подход особенно актуален в задачах, где данные имеют высокую размерность, но могут быть эффективно сжаты без су-

щественной потери информации. В задаче декодирования фМРТ-изображений из видеопоследовательностей предварительное сжатие томографических данных позволяет существенно сократить время обучения при сохранении качества реконструкции, что иллюстрирует практическую значимость управления сложностью данных в прикладных задачах нейровизуализации.

Частота кадров $\nu \in \mathbb{R}$ и продолжительность $t \in \mathbb{R}$ видеопоследовательности задаются. Видеопоследовательность задается как

$$\mathbf{P} = [\mathbf{p}_1, \dots, \mathbf{p}_{\nu t}], \quad \mathbf{p}_\ell \in \mathbb{R}^{W \times H \times C},$$

где W , H и C — ширина, высота и количество каналов изображения соответственно.

Обозначим частоту фМРТ-изображений через $\mu \in \mathbb{R}$. Зададим последовательность изображений

$$\mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_{\mu t}], \quad \mathbf{s}_\ell \in \mathbb{R}^{X \times Y \times Z},$$

где X , Y и Z — размеры воксельного изображения.

Задача состоит в построении отображения, учитывающего задержку Δt между фМРТ-изображением и видеопоследовательностью, а также предыдущие томографические данные. Формально необходимо найти такое отображение \mathbf{g} , что

$$\mathbf{g}(\mathbf{p}_1, \dots, \mathbf{p}_{k_\ell - \nu \Delta t}; \mathbf{s}_1, \dots, \mathbf{s}_{\ell-1}) = \mathbf{s}_\ell, \quad \ell = 1, \dots, \mu t,$$

где для ℓ -го фМРТ-изображения номер соответствующего кадра k_ℓ определяется по формуле

$$k_\ell = \frac{\ell \cdot \nu}{\mu}.$$

Схема предложенного метода реконструкции фМРТ-изображений показана на рис. 6.1.

Обозначим фМРТ-изображение как $\mathbf{s}_\ell = [v_{ijk}^\ell] \in \mathbb{R}^{X \times Y \times Z}$, где $v_{ijk}^\ell \in \mathbb{R}_+$ — значение соответствующего вокселя. Для сокращения времени работы метода предлагается использовать сжатие фМРТ-изображений путем снижения размерности. Сжатие в 2 раза представляется в виде отображения

$$\chi : \mathbb{R}^{X \times Y \times Z} \rightarrow \mathbb{R}^{X/2 \times Y/2 \times Z/2}.$$

Сжатие в 2^k раз получается последовательным применением χ k раз. В дальнейшем для простоты сохраним обозначения размерностей изображения $X \times Y \times Z$.

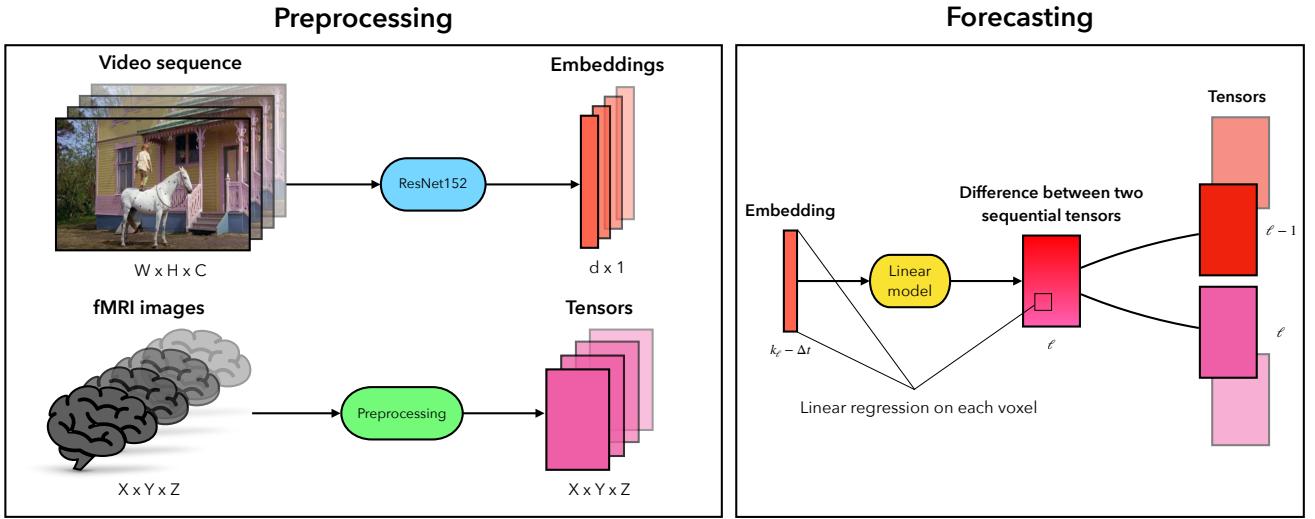


Рис. 6.1: Схема метода декодирования фМРТ-изображений из видеопоследовательностей. Метод использует архитектуру ResNet152 для векторизации видеофрагментов и линейную регрессию с L_2 -регуляризацией для предсказания разностей между последовательными фМРТ-снимками с учетом временной задержки Δt .

Предположим, что для последовательности снимков выполняется свойство Маркова, т.е. каждый снимок зависит только от одного изображения и предыдущего снимка. Тогда соответствующее отображение записывается в виде

$$\mathbf{g}(\mathbf{p}_{k_\ell - \nu \Delta t}) = \mathbf{s}_\ell - \mathbf{s}_{\ell-1} = \boldsymbol{\delta}_\ell, \quad \ell = 2, \dots, \nu t.$$

где $\boldsymbol{\delta}_\ell = [v_{ijk}^\ell - v_{ijk}^{\ell-1}] = [\delta_{ijk}^\ell] \in \mathbb{R}^{X \times Y \times Z}$ — разность двух последовательных снимков.

Отображение $\mathbf{g} : \mathbf{P} \rightarrow \mathbf{S}$ представляется как композиция двух других:

$$\mathbf{g} = \varphi \circ \psi,$$

где

$$\begin{aligned} \psi : \mathbf{P} \rightarrow \mathbb{R}^d &— векторизация изображения, \\ \varphi : \mathbb{R}^d \rightarrow \mathbf{S} &— целевое отображение. \end{aligned}$$

Для каждого изображения из видеопоследовательности имеем вектор внедрения размерности d :

$$\mathbf{x}_\ell = [x_1^\ell, \dots, x_d^\ell]^\top \in \mathbb{R}^d, \quad \ell = 1, \dots, \nu t.$$

В качестве архитектуры используется нейронная сеть ResNet152 без последнего линейного слоя.

При заданном $k_\ell = \ell \cdot \nu / \mu$ общее количество пар составляет $N = \mu(t - \Delta t)$. Таким образом, для каждого вокселя задана выборка

$$\mathfrak{D}_{ijk} = \{(\mathbf{x}_\ell, \delta_{ijk}^\ell) \mid \ell = 2, \dots, N\}.$$

Задача формулируется как регрессия

$$y_{ijk} : \mathbb{R}^d \rightarrow \mathbb{R}.$$

В качестве модели используется линейная регрессия с вектором параметров

$$\mathbf{w}_{ijk} = [w_1^{ijk}, \dots, w_d^{ijk}]^\top \in \mathbb{R}^d :$$

$$f_{ijk}(\mathbf{x}, \mathbf{w}_{ijk}) = \langle \mathbf{x}, \mathbf{w}_{ijk} \rangle.$$

Для модели f_{ijk} с соответствующим вектором параметров $\mathbf{w}_{ijk} \in \mathbb{R}^d$ определим квадратичную функцию потерь с L_2 -регуляризацией:

$$\mathcal{L}_{ijk}(\mathbf{w}_{ijk}) = \sum_{\ell=2}^N (f_{ijk}(\mathbf{x}_\ell, \mathbf{w}_{ijk}) - \delta_{ijk}^\ell)^2 + \alpha \|\mathbf{w}_{ijk}\|_2^2,$$

где $\alpha \in \mathbb{R}$ — коэффициент регуляризации.

Требуется найти параметры, доставляющие минимум функционалу потерь $\mathcal{L}_{ijk}(\mathbf{w}_{ijk})$ при заданных гиперпараметрах Δt и α . Задача оптимизации формулируется следующим образом:

$$\hat{\mathbf{w}}_{ijk} = \arg \min_{\mathbf{w}_{ijk}} \mathcal{L}_{ijk}(\mathbf{w}_{ijk}).$$

Минимум функции потерь находится методом наименьших квадратов. Введем матрицу объектов-признаков

$$\mathbf{X} = [\mathbf{x}_2, \dots, \mathbf{x}_N]^\top = [x_j^i] \in \mathbb{R}^{(N-1) \times d}$$

и вектор, компонентами которого являются разности значений одного и того же вокселя на разных изображениях:

$$\Delta_{ijk} = [\delta_{ijk}^2, \dots, \delta_{ijk}^N]^\top \in \mathbb{R}^{N-1}.$$

Решение задачи оптимизации записывается в виде

$$\hat{\mathbf{w}}_{ijk} = (\mathbf{X}^\top \mathbf{X} + \alpha \mathbf{I})^{-1} \mathbf{X}^\top \Delta_{ijk}.$$

Выведем формулу для восстановленных фМРТ-изображений. Введем матрицу весов

$$\hat{\mathbf{W}} = [\hat{\mathbf{w}}_1, \dots, \hat{\mathbf{w}}_{XYZ}]^\top = [\hat{w}_j^i] \in \mathbb{R}^{XYZ \times d}.$$

Введем для тензоров $\mathbf{s}_\ell, \boldsymbol{\delta}_\ell \in \mathbb{R}^{X \times Y \times Z}$ векторы

$$\mathbf{s}_\ell^R = [v_1^\ell, \dots, v_{XYZ}^\ell]^\top, \boldsymbol{\delta}_\ell^R = [\delta_1^\ell, \dots, \delta_{XYZ}^\ell]^\top \in \mathbb{R}^{XYZ}.$$

Тогда вектор прогнозируемого изображения находится по формуле

$$\hat{\mathbf{s}}_\ell^R = \mathbf{s}_{\ell-1}^R + \hat{\boldsymbol{\delta}}_\ell^R = \mathbf{s}_{\ell-1}^R + \hat{\mathbf{W}} \mathbf{x}_\ell.$$

6.3. Качество данных в задаче детекции машинно-генерированного контента

В предыдущих разделах рассматривались методы управления сложностью моделей и данных в контексте обучения. В настоящем разделе акцент смещается на оценку качества самих данных, что является критически важным аспектом для обеспечения надежности моделей машинного обучения. Низкое качество обучающих данных может приводить к завышенным оценкам производительности моделей и их неспособности к обобщению на реальные данные. Это особенно актуально для задачи детекции машинно-генерированного контента, где качество обучающих наборов напрямую влияет на надежность детекторов.

В последние годы появилось значительное количество AI-детекторов и коллекций с AI-фрагментами. Несколько методов детекции продемонстрировали качество распознавания до 99,9% согласно целевым метрикам в таких коллекциях. Однако качество таких детекторов часто резко падает в реальных условиях, что ставит вопрос о надежности детекторов: действительно ли они высоконадежны, или их высокие бенчмарк-показатели связаны с низким качеством оценочных наборов данных? Подчеркнем необходимость создания надежных и качественных методов оценки сгенерированных данных, которые были бы устойчивы к смещениям и низкой способности к обобщению будущих моделей. В настоящем разделе рассматриваются работы, посвященные детекции AI-генерированного контента, и предлагаются методы оценки качества наборов данных, содержащих AI-фрагменты.

Предлагается оценить различные наборы данных с едиными настройками, чтобы увидеть, насколько хорошо стандартные подходы работают на них. Целью исследования является не достичь максимальных показателей, а сравнить производительность одного и того же метода на разных наборах данных.

Базовые методы. Для методов, основанных на возмущениях, использовался фреймворк DetectGPT с GPT-2 [97] в качестве базовой модели и T5-Large [98] в качестве генератора возмущений. Однако из-за высоких вычислительных затрат DetectGPT использовался Fast-DetectGPT [99], который заменяет этап возмущения в DetectGPT более эффективным этапом сэмплирования. Для методов zero-shot использовался Binoculars [100] с улучшенной оценкой перплексии. Эти два базовых метода не требуют тонкой настройки, что является важным аспектом для задачи детекции, поскольку обучать детектор для каждого домена и генератора непрактично. Наконец, в качестве метода на основе энкодера использовалась модель mDeBERTa [101], которая является современной моделью для детекции машинно-генерированного текста в мультиязычном контексте [102]. Эти три детектора, охватывают все основные категории детекторов.

Топологическая статистика. В работе [103] было показано, что если рассмотреть внутреннюю размерность многообразия на множестве эмбеддингов, то можно отделить тексты, написанные человеком, от машинно-сгенерированных. Авторы использовали размерность персистентной гомологии (англ. PHD) и показали, что статистически тексты, сгенерированные человеком, имеют более высокую PHD, чем машинно-сгенерированные тексты, предложив таким образом новый детектор. Дополнительно, в [104] было предложено вычислять PHD внутри скользящего окна. Эти внутренние размерности текста в пределах скользящего окна могут быть использованы в качестве признака для детекторов. Авторы демонстрируют, что метрика устойчива к изменению домена и генераторов. Для возможности сравнения наборов данных между собой разработана симметричная оценка, использующая KL-дивергенцию. Пусть h_d , m_d — распределения внутренних размерностей для двух типов текстов из одного набора данных, человеческого и машинного происхождения, тогда наша оценка KL_{TTS} выглядит следующим образом:

$$\text{KL}_{\text{TTS}}(h_d, m_d) = |D_{\text{KL}}(h_d || m_d) - D_{\text{KL}}(m_d || h_d)|$$

Чем ниже эта оценка, тем ближе друг к другу h_d и m_d , что означает почти неразличимые тексты, и наоборот.

Возмущения и перемешивание. Основываясь на результатах исследований модификации текста [105, 106], которые показывают, как небольшие возмущения влияют на системы машинного понимания текста, рассматривается этот способ как возможный метод оценки качества набора данных. Ключевая идея

заключается в том, что ИИ-модели чувствительны к таким адверсарным изменениям, в отличие от людей. Рассматриваются две модификации: Адверсарное возмущение токенов и перемешивание предложений.

В данном подходе текст разбивается на токены, и каждый токен случайным образом заменяется на синоним из коллекции WordNet [107] с вероятностью 70%. Далее данная техника применяется к каждому представленному классу, и с использованием модели-энкодера получаются эмбеддинги для каждого из текстов в текущем наборе данных. Наконец, измеряются средние сдвиги эмбеддингов для классов человеческих и сгенерированных текстов, используя косинусное расстояние между эмбеддингами исходных текстов и модифицированных. В итоге, после модификаций получаем Δ_{shift} — логарифм разности средних сдвигов эмбеддингов.

$$\Delta_{\text{shift}} = \log \frac{\frac{1}{n} \sum_{i=1}^n \cos_d(h_{h_i}^o, h_{h_i}^p)}{\frac{1}{m} \sum_{j=1}^m \cos_d(h_{m_j}^o, h_{m_j}^p)},$$

где n и m — количество примеров в человеческой и сгенерированной частях набора данных соответственно, $h_{h_i}^o$ — эмбеддинг i -го фрагмента человеческой части данных, $h_{h_i}^p$ — тот же эмбеддинг после пертурбации. Аналогично, $h_{m_i}^o$ и $h_{m_i}^p$ — эмбеддинги для машинно-сгенерированных текстов. Функция \cos_d измеряет косинусное расстояние между двумя векторами.

В данном подходе предложения случайным образом меняются местами, что влияет на связность текста. Фрагмент разделяется на предложения, и случайным образом изменяется порядок 70% выбранных предложений. Данная техника применяется к каждому представленному классу. Затем, используя модель кодирования текста, получаются эмбеддинги для каждого из текстов текущего набора данных. Наконец, выполняется измерение сдвига эмбеддингов для класса человеческих и сгенерированных текстов, после чего сдвиги преобразуются в распределения, подобные вероятностным. Это в итоге приводит к $\text{KL}_{\text{shuffle}}(H, M)$ — дивергенции Кульбака-Лейблера между сдвигами человеческих и сгенерированных текстов.

$$\text{KL}_{\text{shuffle}}(H, M) = \sum_i H(i) \log \frac{H(i)}{M(i)},$$

$$H(i) = \frac{\cos_d(h_{h_i}^o, h_{h_i}^p) + \epsilon}{\sum_j (\cos_d(h_{h_j}^o, h_{h_j}^p) + \epsilon)},$$

где $M(i)$ имеет ту же структуру, что и $H(i)$, за исключением того, что вместо текстов человеческого класса используются тексты сгенерированного класса, ϵ — малая константа, добавляемая для избежания деления на ноль.

6.4. Результаты вычислительных экспериментов

6.4.1. Сложность моделей в многозадачном обучении

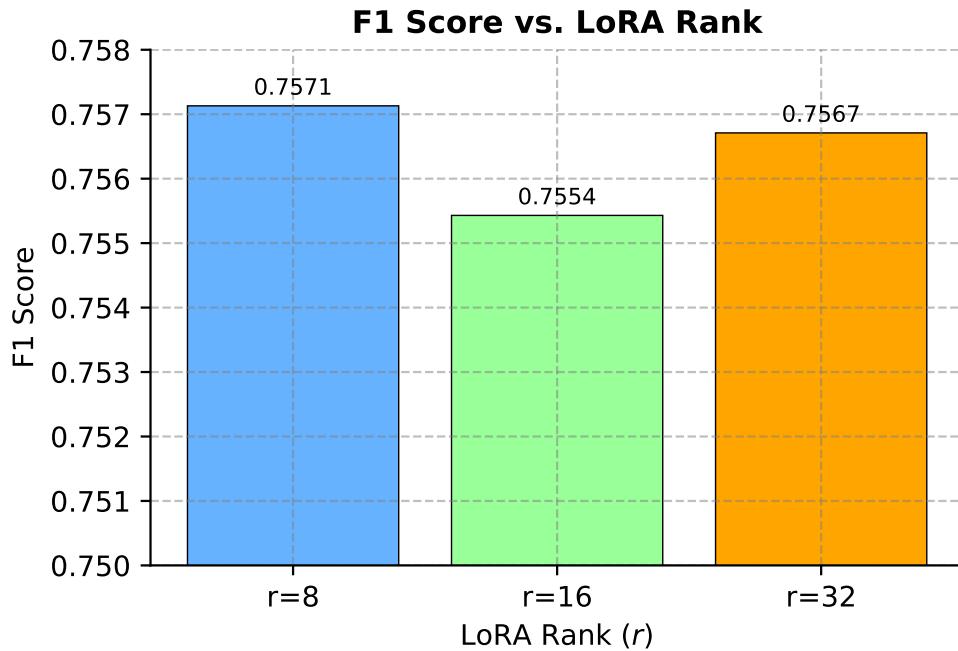


Рис. 6.2: Зависимость метрики F_1 от ранга r адаптеров LoRA для модели DeBERTa-v3-base на задаче детекции машинно-генерированного текста. Наилучшая производительность достигается при $r = 8$, что указывает на снижение отдачи при увеличении ранга и подтверждает эффективность низкоранговой параметризации.

Для эмпирической проверки теоретических свойств, установленных выше, рассматривается задача обнаружения машинно-генерированного текста. Данная задача формулируется как проблема бинарной классификации, что служит компактным, но информативным эталоном для оценки как прогнозной производительности, так и эффективности обучения.

Выбор данной задачи мотивирован ее практической значимостью, разнообразием доменов в наборе данных и необходимостью эффективных стратегий адаптации для достижения сильного обобщения.

В качестве набора данных используется GenAI Detection Challenge (COLING 2025, Task 1), представленный в [108]. Из исходных 600 000 примеров выбирается 60 000 для обучения и 5 000 для тестирования, сохраняя распределение меток.

Набор данных охватывает широкий спектр доменов. Внутрираспределительные категории включают финансы, право, психологию, новости и медицину. Внера распределительные категории включают эссе IELTS, научные статьи, биомедицинские аннотации и юридические статьи.

Результаты предыдущих соревнований указывают на то, что стандартные методы тонкой настройки испытывают трудности при достижении сильного обобщения на этом гетерогенном тестовом наборе, что подчеркивает необходимость более эффективных стратегий адаптации.

В экспериментах фиксируется энкодер как DeBERTa-base [109], который продемонстрировал высокую производительность в недавних задачах классификации текста. Для анализа качества рассматриваются следующие метрики: точность (accuracy), точность (precision), полнота (recall), F1-мера, потеря на валидации и общее время обучения.

Для эмпирической проверки гарантии корректности для моделей Transformer с добавлением LoRA (теорема 33) выполняется сравнение DeBERTa-v3-base в двух режимах тонкой настройки: полная тонкая настройка всех параметров и адаптация с помощью LoRA. Цели эксперимента состоят в следующем: (i) подтвердить, что LoRA сохраняет прогнозную корректность, существенно сокращая сложность параметров, и (ii) проверить, что теоретические выгоды в эффективности транслируются в сокращение времени обучения. Все условия обучения сохраняются идентичными, а полные детали гиперпараметров приведены в дополнительных материалах.

При полной тонкой настройке обновляются все параметры DeBERTa-v3-base. Для LoRA исходные веса замораживаются, а низкоранговые адаптеры вставляются в каждый блок Transformer с рангом $r = 8$, коэффициентом масштабирования $\alpha = 32$ и долей отсева (dropout) для адаптеров 0.1. Эта конфигурация была выбрана после дополнительного анализа чувствительности к рангу, представленного на рис. 6.2.

Результаты суммированы в таблице 6.2. По сравнению с полной тонкой настройкой, LoRA демонстрирует лишь незначительное снижение прогнозных метрик (1.4–3.2%), при этом обеспечивая существенно более низкие потери на валидации (снижение на 0.657, или 36.3%).

Дополнительная диагностика, представленная в дополнительных материалах, показывает более гладкие нормы градиентов и траектории потерь при обучении для LoRA. Это свидетельствует о более стабильной сходимости и большей эмпирической устойчивости LoRA по сравнению с полной тонкой настройкой. Полученные результаты согласуются с теоретической гарантией о том, что низкоранговые обновления сохраняют корректность выходного слоя (теорема 33). Для оценки вычислительной эффективности также рассматривается общее время обучения. LoRA обеспечивает ускорение в 12.6%, что указывает на то, что низкоранговая адаптация снижает эффективную сложность и ускоряет оптимизацию при незначительной потере в метриках, основанных на точности.

Таблица 6.2: Сравнение производительности модели DeBERTa-v3-base при полной тонкой настройке и адаптации с LoRA на задаче детекции машинно-генерированного текста. LoRA демонстрирует незначительное снижение метрик точности (1.4–3.2%), но обеспечивает существенное снижение потерь на валидации (36.3%) и ускорение обучения (12.6%), что подтверждает теоретические гарантии корректности выходного слоя при низкоранговых обновлениях.

Metric	DeBERTa	DeBERTa & LoRA	Change (%)
Accuracy \uparrow	0.7104	0.6876	−3.2
Precision \uparrow	0.6573	0.6413	−2.4
Recall \uparrow	0.9608	0.9470	−1.4
F_1 -score \uparrow	0.7806	0.7648	−2.0
Validation Loss \downarrow	1.8094	1.1522	+36.3
Training Time (s) \downarrow	5570	4867	+12.6

Задача GenAI Detection предоставляет, помимо бинарной метки, вспомогательные метаданные, такие как поддомен и источник benchmark. Эта информация используется для создания многозадачной постановки с тремя выходными головками: исходная задача бинарной классификации и две вспомогательные многоклассовые задачи. Такая конфигурация побуждает модель изучать более богатые текстовые представления и отражает теоретическое преимущество MTL в снижении сложности на задачу.

На основе аналитического результата об эффективности Радемахера (теорема 34) выполняется адаптация DeBERTa-base для выполнения необходимых

Таблица 6.3: Сравнение производительности и эмпирической сложности Радемахера (ERC) модели DeBERTa-v3-base в условиях однозадачного (STL) и многозадачного (MTL) обучения на задаче бинарной классификации GenAI Detection. MTL демонстрирует улучшение всех метрик качества (F1: $0.781 \rightarrow 0.826$, ROC-AUC: $0.788 \rightarrow 0.834$) и снижение ERC с 0.0159 ± 0.0009 до 0.0111 ± 0.0010 , что согласуется с теоретическим предсказанием снижения сложности на задачу в $1/\sqrt{T}$ раз.

Model	Mode	F1 \uparrow	ROC–AUC \uparrow	Acc. \uparrow	ERC \downarrow
DeBERTa	STL	0.781	0.788	0.710	0.0159 ± 0.0009
DeBERTa	MTL	0.826	0.834	0.781	0.0111 ± 0.0010

предположений. Обеспечивается: (i) ограничения нормы весов через проецирование параметров на L_2 -шары после каждого обновления; (ii) липшицева непрерывность с помощью спектральной нормализации всех линейных слоев; и (iii) ограниченность входов через нормализацию токенных эмбеддингов.

Эти модификации гарантируют, что эмпирическая установка соответствует теоретическим условиям, позволяя проводить валидную оценку преимуществ MTL в рамках установленной схемы. Подробный процесс подготовки модели описан в дополнительных материалах.

Сначала оценивается прогнозная производительность в условиях STL и MTL. Выполняется обучение двух вариантов: (i) базовой STL-модели с частичной тонкой настройкой энкодера и классификационной головой, и (ii) MTL-модели с общим энкодером, обновляемым совместно по всем трем задачам.

На этапе тестирования для MTL-модели используется только общий энкодер и бинарная голова, отбрасывая вспомогательные головы. Результаты в таблице 6.3 показывают, что включение вспомогательных задач в процесс обучения улучшает точность на исходной бинарной задаче, подтверждая, что многозадачные сигналы действуют как полезное индуктивное смещение.

Далее, для оценки не только прогнозных улучшений, но и сложности модели на целевой задаче, выполняется прямое сравнение эмпирической сложности Радемахера (ERC). В соответствии с определением, ERC измеряется на тех же данных, что использовались для обучения, после замены истинных меток случайным шумом.

Это гарантирует, что оценка отражает только емкость гипотезного класса,

а не информацию о реальных метках. После обучения ERC оценивается исключительно для целевой задачи как для STL, так и для MTL. Результаты представлены в колонке ERC в таблице 6.3.

Более подробное описание процесса проведения обоих экспериментов приведено в дополнительных материалах.

Анализ показывает, что MTL не только улучшает прогнозную производительность, но и достигает более низкой эмпирической сложности Радемахера по сравнению с STL. Это демонстрирует, что вспомогательные задачи обеспечивают полезное индуктивное смещение, которое как улучшает обобщающую способность на целевой задаче, так и эффективно снижает сложность на задачу, что согласуется с предсказанным снижением на $1/\sqrt{T}$ из нашего теоретического анализа.

Таблица 6.4: Сравнение моделей DeBERTa-v3-base при полной тонкой настройке и адаптации с LoRA ($r = 8$) после 5 эпох обучения. LoRA обеспечивает более низкие потери (0.2972 vs 0.3058), более гладкие нормы градиентов (17.41 vs 24.65) и требует только 0.16% параметров от полной модели (296,450 vs 184M), демонстрируя эффективность низкоранговой адаптации.

Model	Loss	Grad. Norm	Runtime (s)	Params	Epochs
Full Tuning	0.3058	24.65	679,660	184M (100%)	5
LoRA ($r = 8$)	0.2972	17.41	700,284	296,450 ($\approx 0.16\%$)	5

Таблица 6.5: Сравнение конфигураций LoRA с различными рангами $r \in \{8, 16, 32\}$ для модели DeBERTa-v3-base. Анализ показывает, что более высокие ранги ($r = 32$) обеспечивают более низкие потери при обучении, но $r = 8$ обеспечивает лучшее обобщение, что подтверждает оптимальность низкоранговой параметризации для данной задачи.

Rank (r)	Loss	Grad. Norm	Runtime (s)
8	0.3413	2.95	1681
16	0.3358	5.18	1684
32	0.3194	2.55	1688

Все исходные параметры DeBERTa были заморожены, а низкоранговые

адаптерные модули были вставлены в каждый блок трансформера. Ранг адаптера выбран $r = 8$, коэффициент масштабирования $\alpha = 32$ и использовался dropout 0.1 в слоях адаптера. Данная конфигурация выбрана как наиболее эффективная после проведения поиска по сетке для $r = \{8, 16, 32\}$. Это оставляет обучаемыми только 296 450 параметров из 184 млн общих ($\approx 0.16\%$), что согласуется с теоретической мотивацией снижения размерности. Все конфигурации экспериментов указаны в таблице 6.4.

Кроме того, на рис. 6.4 представлена производительность модели: модели DeBERTa и DeBERTa & LoRA обучались в течение 5 эпох с размороженными 6 слоями из 12. Анализ показывает, что норма градиента является значительно более гладкой для модели, использующей LoRA, по сравнению с моделью без нее. Тот же эффект наблюдается на графике функции потерь обучения: потери при обучении модели с LoRA являются более гладкими и сходятся быстрее, чем у модели без нее.

Для определения оптимального ранга адаптера сравниваются $r = 8, 16, 32$ с точки зрения потерь при обучении, представленных на рис. 6.3. Хотя более высокие ранги, в частности $r = 32$, демонстрировали более быструю сходимость и более низкие финальные потери при обучении, $r = 8$ обеспечивает лучшее обобщение, несмотря на более медленную сходимость. В целом, $r = 8$ был выбран в качестве наиболее эффективной конфигурации в текущей настройке. Конфигурации экспериментов показаны в таблице 6.5.

На основе аналитических результатов об эффективности Радемахера выполняется адаптация DeBERTa-base для выполнения необходимых предположений. Введенные модификации гарантируют, что эмпирическая установка соответствует теоретическим условиям, позволяя провести валидную оценку преимуществ MTL в рамках установленной схемы. Подробный процесс подготовки описан в Алгоритме 1.

Для оценки ERC обучается как STL, так и MTL модели в контролируемых и сопоставимых условиях. В обоих случаях использовалось одинаковое общее количество обучающих примеров (nT), что гарантирует, что любые наблюдаемые различия в ERC вызваны совместным использованием параметров, а не дисбалансом данных.

В эксперименте STL использовался энкодер DeBERTa-v3-base с одной бинарной классификационной головой, соответствующей целевой задаче. Модель обучалась на $nT = 90,000$ примерах ($n = 30,000, T = 3$), выбранных исключительно из набора данных целевой бинарной классификации. Чтобы гарантиро-

Algorithm 1: Алгоритм обеспечения условий теоремы 34 для оценки эмпирической сложности Радемахера (ERC) с использованием модели DeBERTa. Алгоритм применяет спектральную нормализацию к линейным слоям энкодера, нормализует входные последовательности и проецирует параметры на L_2 -шары для выполнения ограничений на нормы весов, необходимых для теоретического анализа.

Input : Модель $f_t(x) = w_t^\top \phi(x; w_{\text{shared}})$, с ограничениями $B_{\text{shared}}, B_{\text{head}}, R$

Output: Модель удовлетворяющая условиям теоремы 34.

for each *linear layer* W *in encoder* $\phi(\cdot; w_{\text{shared}})$ **do**

- $W \leftarrow \text{SpectralNorm}(W)$; $// \|\phi(x; w)\|_2 \leq L \|w\| \|x\|_2$

for each *input sequence* $x = [x_1 \dots x_m]$ **do**

- $\tilde{x}_i \leftarrow x_i / \max(1, \|x_i\|_2)$ $\forall i$ $x \leftarrow [\tilde{x}_1 \dots \tilde{x}_m] \cdot \min\left(1, \frac{R}{\|\tilde{x}\|_F}\right)$; $// \|x\| \leq R$

for each *training step* **do**

- for each** *parameter group* (w, B) *in*

 - $\{(w_{\text{shared}}, B_{\text{shared}})\} \cup \{(w_t, B_{\text{head}}) \mid t \in [T]\}$ **do**
 - $w \leftarrow w \cdot \min\left(1, \frac{B}{\|w\|_2}\right)$; $// \|w\|_2 \leq B$

вать, что ERC отражает репрезентационную способность модели, а не ее соответствие истинному распределению меток, все целевые метки были заменены случайным шумом из $\{-1, 1\}$. Энкодер и классификационная головка подвергались тонкой настройке на этих зашумленных метках в течение 3 эпох. После обучения ERC вычислялась на том же наборе данных со случайными метками.

В настройке MTL использовался тот же энкодер DeBERTa-v3-base, разделяемый между $T = 3$ классификационными головками. Одна головка соответствовала целевой бинарной задаче, в то время как две другие обучались на вспомогательных многоклассовых задачах: предсказание поддомена с 5 классами и предсказание поддомена с 6 классами. Каждая задача предоставляла $n = 30,000$ примеров, так что общий энкодер обрабатывал в сумме $nT = 90,000$ обучающих примеров. Только вспомогательные задачи использовали свои исходные метки; целевая задача использовала зашумленные метки для обеспечения независимости от емкости модели. После обучения ERC оценивалась только на подмножестве целевой задачи.

Данная процедура гарантирует, что как STL, так и MTL модели обучались



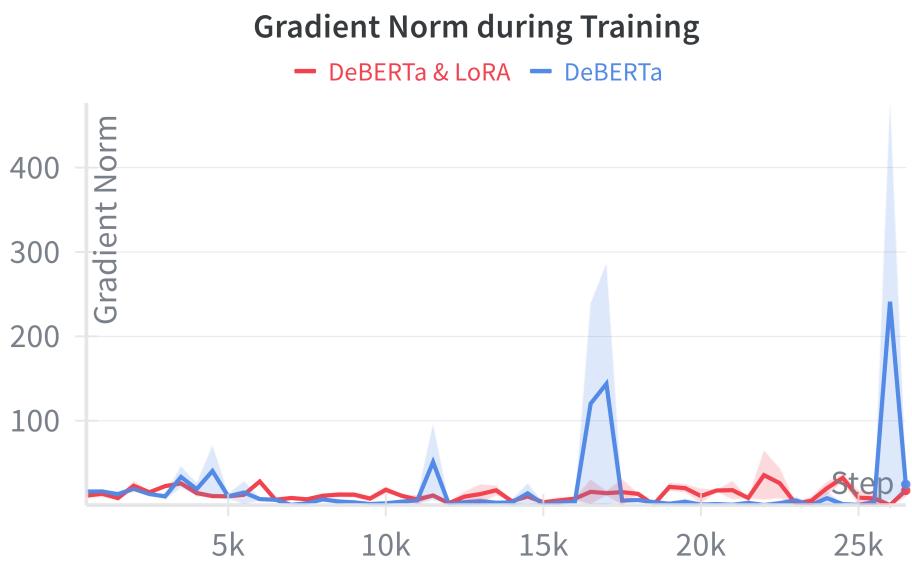
Рис. 6.3: Динамика потерь при обучении модели DeBERTa-v3-base с адаптерами LoRA различных рангов ($r \in \{8, 16, 32\}$) на задаче детекции машинно-генерированного текста. Анализ показывает, что более высокие ранги обеспечивают более быструю сходимость и более низкие финальные потери, но $r = 8$ обеспечивает лучшее обобщение, что подтверждает оптимальность низкоранговой параметризации.

в идентичных условиях по объему данных и вычислительным ресурсам. Следовательно, наблюдаемое снижение ERC типа $1/\sqrt{T}$ непосредственно количественно оценивает эффект совместного использования параметров энкодера на эффективную емкость модели.

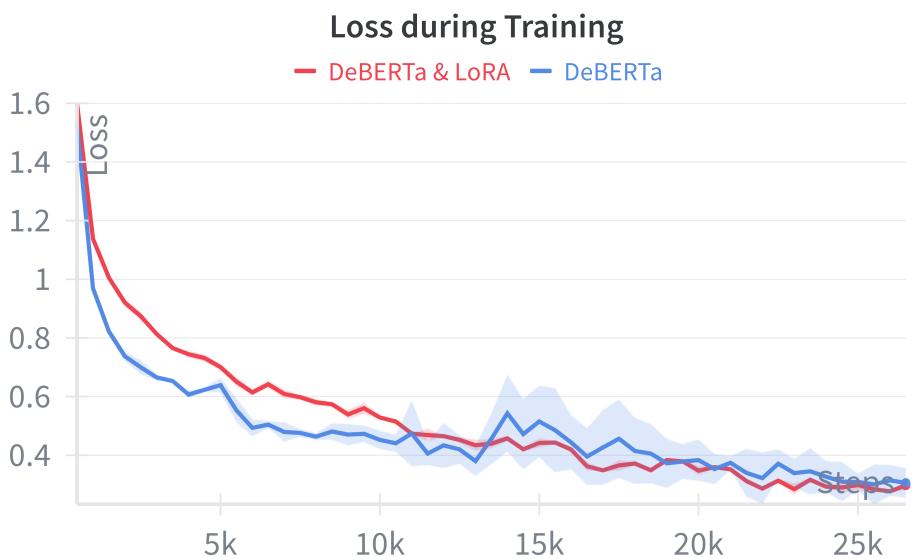
6.4.2. Снижение сложности данных в задаче декодирования фМРТ-снимков

Для анализа производительности предложенного метода и проверки гипотез был проведен вычислительный эксперимент. В качестве данных использовалась выборка, представленная в [110].

Набор данных содержит результаты обследования 63 испытуемых. Для тридцати из них известны данные фМРТ. В выборке 16 мужчин и 14 женщин в возрасте от 7 до 47 лет. Средний возраст испытуемых составляет 22 года. Характеристики выборки: продолжительность обследования, частоты кадров видеопоследовательностей фМРТ и изображений, а также их размеры суммированы в таблице 6.6.



(a) Норма градиента



(b) Ошибка на обучающей выборке

Рис. 6.4: Динамика нормы градиента и потерь при обучении моделей DeBERTa-v3-base с полной тонкой настройкой и адаптацией LoRA ($r = 8$) в течение 5 эпох на задаче детекции машинно-генерированного текста. LoRA демонстрирует более гладкие траектории нормы градиента и потерь, что указывает на более стабильную сходимость и большую эмпирическую устойчивость по сравнению с полной тонкой настройкой.

Таблица 6.6: Характеристики выборки для эксперимента по декодированию фМРТ-изображений из видеопоследовательностей. Выборка содержит данные 30 испытуемых с продолжительностью обследования 390 с, частотой кадров видео 25 Гц и частотой фМРТ-изображений 1.64 Гц, что обеспечивает временное соответствие между видеокадрами и томографическими данными.

Name	Notation	Value
Duration of examination	t	390 s
Video frame rate	ν	25 Hz
fMRI frame rate	μ	1.64 Hz
Video dimensions	W, H, C	640, 480, 3
fMRI dimensions	X, Y, Z	40, 64, 64

Выборка разделена на обучающую и тестовую части в соотношении 70% и 30% соответственно. Критерием качества реконструкции фМРТ-изображений является MSE — сумма квадратов отклонений между истинными и реконструированными изображениями, усредненная по всем voxelам каждого изображения из тестовой выборки.

Для сокращения времени работы алгоритма фМРТ-изображение предварительно сжимается с использованием слоя MaxPool3D. Рассматриваются коэффициенты сжатия 1, 2, 4 и 8. Значения voxelей нормализованы к диапазону [0; 1] с помощью процедуры MinMaxScale.

Была проанализирована зависимость MSE от параметра регуляризации α . Рассматривались коэффициенты сжатия 1, 2, 4 и 8. Соответствующие графики показаны на рис. 6.5. Для построения графика было выполнено усреднение по испытуемым. Показаны границы стандартного отклонения. Графики демонстрируют, что оптимальное значение коэффициента $\alpha \approx 1000$. Форма кривой сохраняется независимо от коэффициента сжатия фМРТ-изображений.

Выполняется сравнение времени обучения модели при использовании различных коэффициентов сжатия фМРТ-изображений. Рассматриваются коэффициенты 1, 2, 4 и 8. Для каждого значения коэффициента сжатия вычисляется среднее значение времени обучения модели для всех испытуемых.

Результаты экспериментов представлены в таблице 6.7. Время работы метода существенно сокращается при использовании предварительного сжатия фМРТ-изображений. Эксперимент с подбором оптимального коэффициента ре-

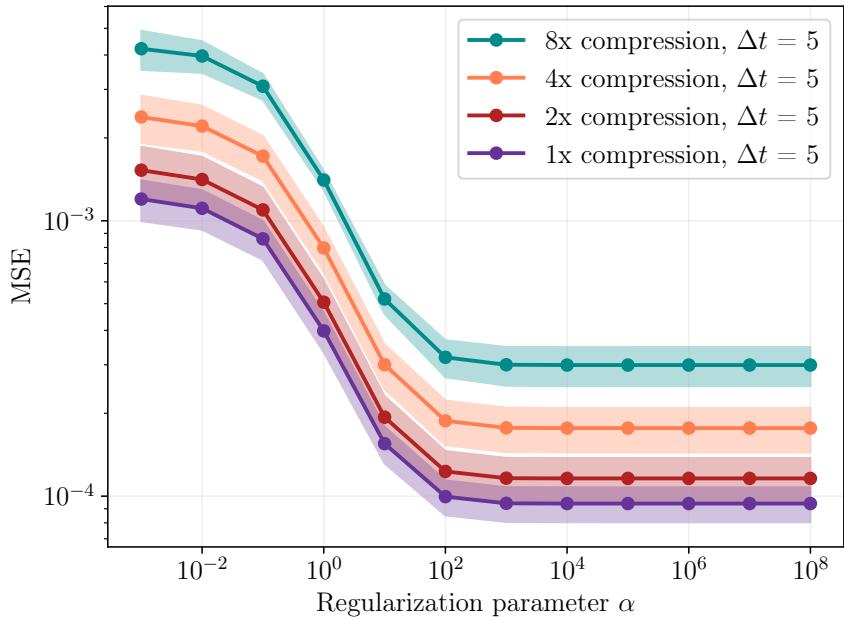


Рис. 6.5: Зависимость метрики MSE от параметра регуляризации α для метода декодирования фМРТ-изображений на тестовой выборке при различных коэффициентах сжатия (1, 2, 4, 8). Оптимальное значение коэффициента $\alpha \approx 1000$ сохраняется независимо от коэффициента сжатия, что подтверждает стабильность метода при снижении размерности данных.

гуляризации подтверждает, что сжатие изображений не изменяет характер зависимостей между параметрами и качеством реконструкции.

Таблица 6.7: Зависимость среднего времени обучения модели декодирования фМРТ-изображений от коэффициента сжатия данных. Использование предварительного сжатия фМРТ-изображений с коэффициентами 2, 4 и 8 обеспечивает существенное сокращение времени обучения (с 36.3 с до 6.7 с, 1.6 с и 1.4 с соответственно), что демонстрирует эффективность снижения сложности данных без потери качества реконструкции.

Compression coefficient	Mean time, s	Std, s
1	36.3	6.1
2	6.7	0.5
4	1.6	0.1
8	1.4	0.3

6.4.3. Качество данных в задаче детекции машинно-генерированного контента

Dataset	DeBERTa	Binoculars	DetectGPT
GPT-2	0.972	0.495	0.412
HC3	0.998	0.931	0.972
GhostBuster	0.910	0.683	0.711
MGTBench	0.961	0.364	0.447
MAGE	0.835	0.632	0.654
M4	0.987	0.871	0.881
OutFox	0.901	0.692	0.707
TweepFake	0.941	0.845	0.864
SemEval24 Mono	0.991	0.913	0.924
SemEval24 Multi	0.994	—	—
RuATD	0.765	—	—
DAGPap22	0.968	0.333	0.562
PAN24	0.826	0.411	0.890
AuTex23en	0.941	0.783	0.911
AuTex23es	0.933	—	—
IberAuTex	0.964	—	—
MGT-1 Mono	0.904	0.665	0.683
MGT-1 Multi	0.934	—	—

Таблица 6.8: Результаты классификации различных детекторов машинно-генерированного текста (DeBERTa, Binoculars, DetectGPT) на множественных наборах данных, оцененные с помощью метрики F_1 -score. Детектор на основе DeBERTa демонстрирует наиболее стабильную производительность на различных наборах данных, тогда как Binoculars и DetectGPT показывают значительные вариации, что указывает на проблемы с устойчивостью этих методов к различным доменам.

Из каждого набора данных выбрано 1000 документов из тестовой выборки, сбалансированных между двумя классами. Для базовых методов выполняется тонкая настройка mdeberta-v3-base для каждого набора данных, после чего модель оценивается. Для оценки качества Binoculars и Fast-DetectGPT использо-

Dataset	$\text{KL}_{\text{TTS}} \downarrow$	$\text{PHD}_{\text{human}}$	$\text{PHD}_{\text{machine}}$	$\Delta_{\text{shift}} \downarrow$	$\text{KL}_{\text{shuffle}} \downarrow$
GPT-2	0.014	9.23 ± 1.98	10.27 ± 1.84	0.084	1.255
HC3	0.053	8.76 ± 1.83	7.38 ± 1.05	0.264	1.167
GhostBuster	0.053	9.84 ± 1.18	9.76 ± 1.15	0.024	0.359
MGTBench	0.043	8.77 ± 1.31	9.97 ± 1.02	0.031	0.421
MAGE	0.011	9.8 ± 2.14	9.38 ± 3.04	0.094	0.310
M4	0.036	7.26 ± 1.99	8.59 ± 1.4	0.107	0.483
OutFox	0.025	8.96 ± 1.21	11.48 ± 1.13	0.095	0.237
TweepFake	-	9.02 ± 3.19	8.12 ± 4.02	0.116	1.001
SemEval24 Mono	0.012	9.11 ± 1.19	9.41 ± 1.2	0.191	2.576
SemEval24 Multi	0.001	9.65 ± 1.81	9.42 ± 1.44	0.059	2.046
RuATD	<u>0.007</u>	7.33 ± 1.4	7.46 ± 1.41	0.315	14.028
DAGPap22	0.083	8.35 ± 1.33	7.48 ± 2.01	0.039	0.472
PAN24	0.053	9.4 ± 1.05	8.52 ± 1.59	0.050	0.331
AuTex23 Eng	<u>0.021</u>	8.07 ± 2.26	8.1 ± 2.68	0.110	4.331
AuTex23 Esp	<u>0.001</u>	9.16 ± 3.49	9.25 ± 3.26	0.105	1.306
IberAuTex	0.012	9.33 ± 2.45	8.47 ± 2.73	0.223	5.516
MGT-1 Mono	0.019	9.19 ± 1.75	8.96 ± 2.24	0.031	0.587
MGT-1 Multi	0.006	8.76 ± 1.85	8.6 ± 2.29	0.027	0.522

Таблица 6.9: Статистика качества данных для выбранных наборов данных детекции машинно-генерированного текста, включающая метрики KL_{TTS} , PHD , Δ_{shift} и $\text{KL}_{\text{shuffle}}$. Высокие значения метрик указывают на различимость текстов разного происхождения, тогда как низкие значения отражают схожую устойчивость к модификациям, что является признаком качественных данных.

вался falcon-rw-1b [111] и gpt-neo-2.7B [112] соответственно. Следует отметить, что для последних двух методов качество измерялось только на англоязычных выборках. В эксперименте с топологическими признаками использовалась модель roberta-base, как и авторы оригинальной работы. В эксперименте с пертурбациями и перемешиванием энкодер multilingual-e5-large использовался для построения эмбеддингов текстов, который показывает высокие метрики для кодирования высокоресурсных языков [113]. Результаты сравнения разработанных признаков в выбранных наборах данных представлены в таблице 6.9.

Относительно PHD и оценки TTS , в предыдущих работах было показано,

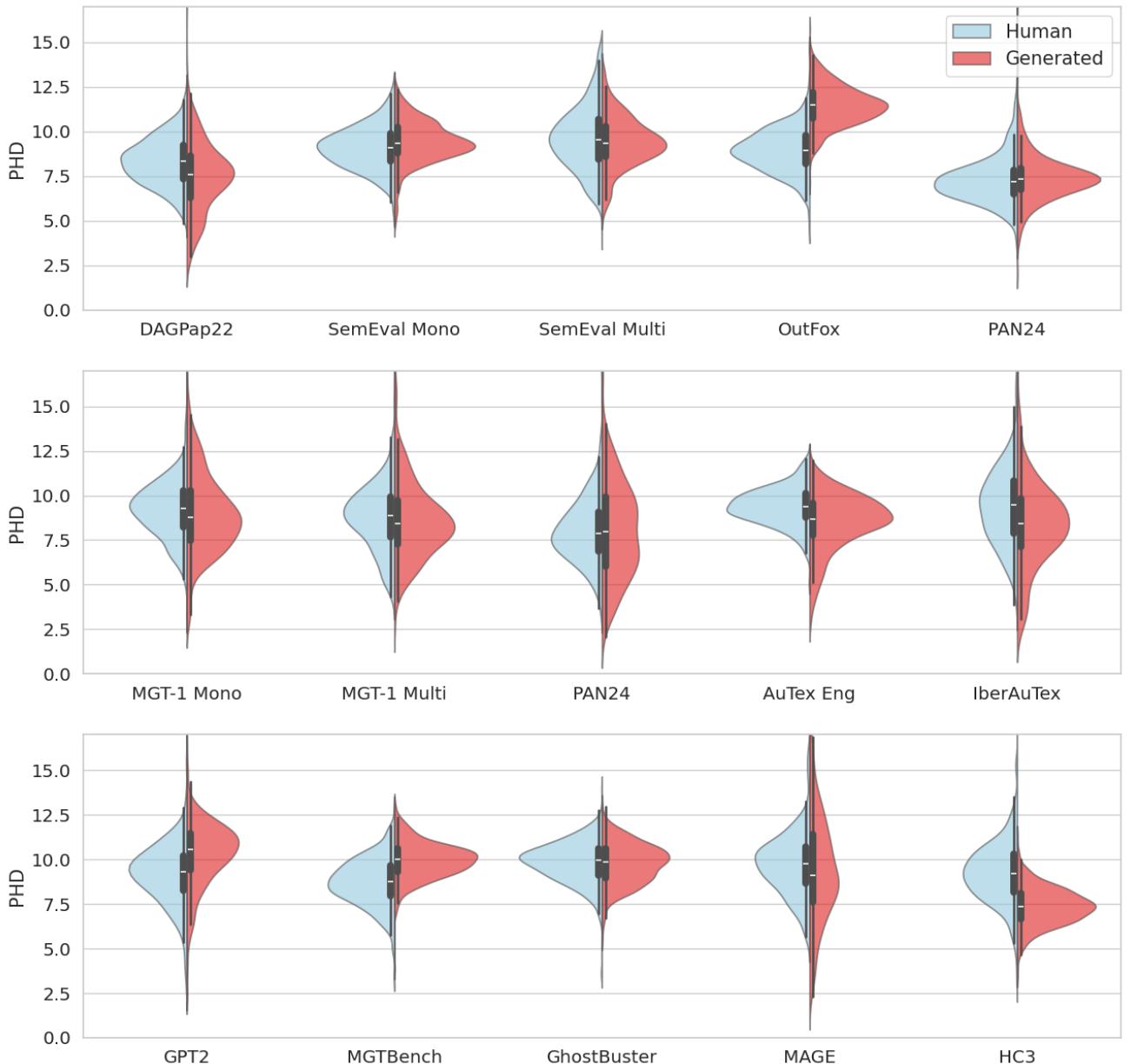


Рис. 6.6: Распределения размерности персистентной гомологии (PHD) для человеческих и машинно-сгенерированных текстов в различных наборах данных детекции. Качественные наборы данных (SemEval, PAN24, MGT-1) демонстрируют схожие распределения PHD для обоих типов текстов, что указывает на высокое качество сгенерированных данных и их близость к человеческим текстам по топологическим характеристикам.

что тексты языковых моделей имеют меньшие значения PHD, чем написанные человеком. Однако данный результат был получен для моделей GPT-2, GPT-3.5 и OPT. Для более современных языковых моделей, которые генерируют более похожие на человеческие тексты, данная тенденция могла измениться.

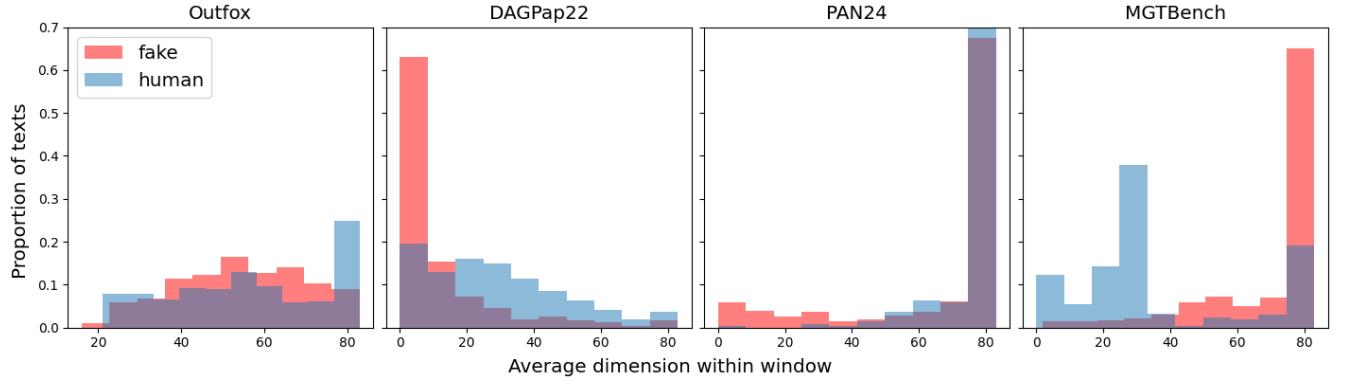


Рис. 6.7: Топологические временные ряды (TTS) для четырех наборов данных с высокими значениями метрики KL_{TTS}: GhostBuster, PAN24, MGTBench и DAGPap22. Высокие значения KL_{TTS} для GhostBuster и PAN24 обусловлены расхождением в текстах с более высокими размерностями, тогда как для MGTBench и DAGPap22 — разницей в самих распределениях PHD, что указывает на различные типы различий между человеческими и машинно-сгенерированными текстами.

Высокое значение KL_{TTS} для текстов разного происхождения означает, что детектору легче разделить такие тексты. Следует отметить, что KL_{TTS} также ограничена для более коротких текстов. Относительно PHD предполагается, что сгенерированные тексты хорошего качества должны иметь PHD, аналогичный написанным человеком.

Дополнительно сравниваются распределения PHD для всех наборов данных на рис. 6.6. Распределения для текстов обоих типов происхождения должны быть схожими, что в основном соблюдается для текстов из SemEval, PAN24 и MGT-1.

В следующих столбцах приводится статистика, наблюдаемая на модифицированных текстах. Для обеих метрик (Δ_{shift} и KL_{shuffle}) чем ниже значение, тем лучше, так как это отражает схожую степень устойчивости сгенерированных и человеческих текстов к адвокарным атакам. Качественно сгенерированные данные без смещений должны принимать значения, близкие к человеческим.

В таблице 6.8 показаны результаты применения современных детекторов к выбранным тестовым наборам данных. На наборах данных с низкими значениями метрик качества в таблице 6.9 может быть достигнуто качество, близкое к 1, что указывает на явное наличие смещения детектора в их сторону или структурной особенности, которая слишком очевидна для модели детекции.

Невозможно судить о качестве данных только по достижению значений F_1 ,

близких к 1. Однако комбинирование значений двух таблиц позволяет оценить, какой набор имеет данные лучшего качества, а какой — худшего.

Относительно KL_{TTS}, на рис. 6.7 показаны 4 набора данных с высоким значением этого показателя. GhostBuster и PAN24 получили такой высокий балл из-за расхождения в текстах с более высокими размерностями, тогда как MGBTbench и DAGPap22 — из-за разницы в самих распределениях.

Следует отметить, что KL_{TTS} может плохо работать с очень короткими текстами, поскольку внутренний метод вычисления PHD требует достаточно длинных текстов для стабильного вычисления. По этой причине KL_{TTS} для RuATD, AuTex23-es и Tweepfake отбрасываются, так как они не соответствуют критериям. Кроме того, показано, что тексты должны быть достаточной длины [114] для построения надежных детекторов.

Анализ значений в таблице 6.9 показывает наличие данных достаточно высокого качества в выбранных наборах данных. Разработанные атрибуты в совокупности способны отражать качество сгенерированного набора данных с различных точек зрения.

Предлагается использовать эти атрибуты в сочетании с другими статистическими инструментами для оценки качества данных, например, с законом Ципфа [115].

Представленная статистика может быть использована для оценки качества коллекций и их улучшения. Кроме того, наборы данных, собирающие машинно-генерированный контент, могут быть полезны для двух более общих целей. Во-первых, высококачественные сгенерированные данные могут быть использованы для оценки качества каузальной модели во время обучения, что служит одной из целей обучения для улучшения ответов модели и приближения их к человеческим. Во-вторых, хорошие детекторы могут помочь очистить обучающие наборы, поскольку большая доля низкокачественных сгенерированных текстов в этих наборах может привести к возникновению смещений в сторону некорректной структуры и артефактов в выходных данных модели в будущем.

Вопрос о том, означает ли низкая производительность детекторов низкое качество набора данных, не имеет однозначного ответа. Например, в [100] метод Binoculars достигает оценки F_1 , близкой к 1.0, в то время как в наших экспериментах был получен широкий диапазон оценок: от 0.33 на DAGPap22 до 0.93 на НСЗ. Для НСЗ все три детектора показали схожие результаты, что позволяет предположить, что тексты НСЗ относительно легко обнаружить. Однако данная согласованность не распространяется на DAGPap22. Детектор на осно-

ве DeBERTa достиг оценки F_1 0.96, в то время как DetectGPT показал только 0.562. Данная закономерность, когда детектор на основе DeBERTa демонстрирует заметно более высокие результаты, чем два других метода, наблюдалась на значительной части проанализированных наборов данных.

Низкие оценки для Binoculars заслуживают дополнительного изучения. Даже при фокусировке на доменах, специально протестированных его авторами, таких как PAN24 и Outfox, оценки оказываются значительно ниже почти идеальных результатов, представленных в [100]. Данное расхождение позволяет предположить, что детектор Binoculars может быть нерепрезентативным. Аналогично, в наших экспериментах оценки DetectGPT сопоставимы с оценками Binoculars, что указывает на схожие базовые проблемы с устойчивостью этих детекторов.

6.5. Заключение по главе

В настоящей главе продемонстрировано практическое применение теоретического аппарата оценки сложности моделей и данных, разработанного в главах 2 и 3, к решению прикладных задач машинного обучения.

В отличие от предыдущих глав, где основной акцент делался на разработке строгих математических оценок и методов анализа, здесь представлена адаптация теоретических подходов к реальным проблемам. Глава охватывает три ключевых направления: (1) управление сложностью моделей в многозадачном обучении, (2) снижение сложности данных в задачах нейровизуализации и (3) оценка качества данных в задачах детекции машино-генерированного контента.

В разделе о сложности моделей в многозадачном обучении получены фундаментальные теоретические результаты для адаптеров LoRA и многозадачного обучения. Теорема 32 строго доказывает статистическую состоятельность минимизации эмпирического риска при параметризации LoRA, устанавливая, что низкоранговая адаптация сохраняет асимптотические статистические гарантии полной тонкой настройки. Теорема 33 доказывает корректность выходного слоя при низкоранговых обновлениях, показывая, что параметризация $\Delta W = AB$ с $r \ll \min(d, k)$ не искажает выходное распределение или классификационные границы модели. Теорема 34 устанавливает снижение эмпирической сложности Радемахера на задачу в $1/\sqrt{T}$ раз при многозадачном обучении с соответствующим масштабированием, что количественно характеризует преимущества

совместного использования параметров энкодера.

Эмпирические эксперименты на задаче детекции машинно-генерированного текста (GenAI Detection Challenge) подтвердили теоретические предсказания. LoRA-адаптеры с рангом $r = 8$ демонстрируют незначительное снижение метрик точности (1.4–3.2%) по сравнению с полной тонкой настройкой, но обеспечивают существенное снижение потерь на валидации (36.3%) и ускорение обучения (12.6%), требуя при этом только 0.16% параметров от полной модели. Многозадачное обучение с тремя выходными головками улучшает метрики качества ($F1: 0.781 \rightarrow 0.826$, ROC-AUC: $0.788 \rightarrow 0.834$) и снижает эмпирическую сложность Радемахера с 0.0159 ± 0.0009 до 0.0111 ± 0.0010 , что количественно подтверждает теоретическое предсказание снижения сложности на задачу в $1/\sqrt{T}$ раз.

В разделе о снижении сложности данных в задаче декодирования фМРТ-изображений разработан метод реконструкции томографических данных из видеопоследовательностей, использующий архитектуру ResNet152 для векторизации видеокадров и линейную регрессию с L_2 -регуляризацией для предсказания разностей между последовательными фМРТ-снимками. Предложенный метод предварительного сжатия фМРТ-изображений с коэффициентами 2, 4 и 8 обеспечивает существенное сокращение времени обучения (с 36.3 с до 6.7 с, 1.6 с и 1.4 с соответственно) без потери качества реконструкции. Экспериментальные результаты на выборке из 30 испытуемых показали, что оптимальное значение коэффициента регуляризации $\alpha \approx 1000$ сохраняется независимо от коэффициента сжатия, что подтверждает стабильность метода при снижении размерности данных и демонстрирует практическую эффективность управления сложностью данных в задачах нейровизуализации.

В разделе о качестве данных в задаче детекции машинной генерации разработаны методы оценки качества наборов данных, содержащих AI-фрагменты, основанные на топологической статистике, анализе устойчивости к адверсарным возмущениям и перемешиванию предложений. Предложены четыре метрики качества: KL_{TTS}, PHD, Δ_{shift} и KL_{shuffle}. Экспериментальный анализ множественных наборов данных с использованием трех детекторов выявил значительные вариации в качестве данных: детектор на основе DeBERTa демонстрирует наиболее стабильную производительность на различных наборах данных, тогда как другие детекторы показывают значительные вариации, что указывает на проблемы с устойчивостью этих методов к различным доменам. Качественные наборы данных (SemEval, PAN24, MGT-1) демонстрируют схожие распределения

ния PHD для человеческих и машинно-сгенерированных текстов, что указывает на высокое качество сгенерированных данных и их близость к человеческим текстам по топологическим характеристикам.

Полученные результаты создают основу для практического применения теоретического аппарата сложности моделей и данных в реальных задачах машинного обучения. Теоретические гарантии для LoRA и многозадачного обучения обеспечивают обоснованный выбор архитектурных решений при разработке эффективных и компактных моделей. Методы снижения сложности данных в задачах нейровизуализации демонстрируют практическую значимость управления размерностью данных без потери качества. Метрики оценки качества данных открывают возможности для создания надежных бенчмарков и улучшения качества обучающих наборов данных.

Основные ограничения представленных подходов связаны с вычислительной сложностью анализа топологических характеристик для очень коротких текстов, а также с необходимостью адаптации методов оценки качества данных к различным доменам и языкам. Теоретическое обоснование методов снижения сложности данных в настоящее время ограничено задачами регрессии с линейными моделями.

Перспективными направлениями дальнейших исследований являются разработка упрощенных практических метрик сложности, которые могли бы служить мостом между строгим теоретическим аппаратом предыдущих глав и потребностями прикладных задач, расширение методов оценки качества данных на другие модальности, интеграция предложенных метрик в процесс создания и валидации обучающих наборов данных, а также разработка адаптивных методов выбора ранга LoRA-адаптеров на основе теоретических оценок сложности Радемахера.

Заключение

В диссертации рассмотрена крупная научная проблема, имеющая важное теоретическое и прикладное значение в области математических методов моделей глубокого обучения, связанная с отсутствием системного подхода к оценке и управлению сложностью моделей и данных. В рамках диссертационной работы предложено решение фундаментальной проблемы разработки единого математического аппарата для оценки и управления сложностью как моделей глубокого обучения, так и данных, что позволяет перейти к системному проектированию архитектур нейронных сетей и оптимизации процессов их обучения. Предложенный подход основан на введении формальных мер сложности в рамках теории мер, анализе матриц Гессе и ландшафта оптимизационной задачи, что обеспечивает теоретическую основу для предсказания поведения моделей при масштабировании и количественной оценки соответствия между сложностью модели и сложностью данных.

Основные результаты диссертационной работы перечислены в следующих пунктах:

1. Разработан единый формальный аппарат для оценки сложности моделей и данных на основе теории мер и анализа ландшафта оптимизационной задачи. Введены формальные определения меры сложности выборки и меры сложности модели, установлен критерий обучаемости модели на выборке, определяющий необходимое условие предотвращения переобучения.
2. Введена ландшафтная мера сложности модели, определяемая через спектральные свойства матриц Гессе функции потерь, и установлена ее связь с условной сложностью выборки.
3. Получены строгие теоретические оценки ландшафтных мер на основе спектральных норм матриц Гессе для основных архитектур глубокого обучения: полносвязных, сверточных и трансформерных сетей. Установлены зависимости ландшафтных мер от глубины сети, размеров слоев и других структурных параметров архитектур.
4. Разработаны практические методы оценки достаточного объема выборки, восполняющие пробел между теоретическим аппаратом оценки сложности и практическими потребностями планирования экспериментов. Проведен систематический сравнительный анализ методов на основе сэмплирования эмпирической функции ошибки и анализа близости апостериорных

распределений параметров с классическими статистическими и байесовскими методами, подтверждена эффективность предложенных подходов.

5. Предложены методы мультидоменной дистилляции и анти-дистилляции для передачи знаний между моделями различной сложности и между различными доменами данных. Для многозадачного обучения получены фундаментальные теоретические результаты о статистической состоятельности низкоранговых адаптеров и снижении эмпирической сложности при совместном использовании параметров энкодера.

В целом совокупность полученных в диссертации теоретических и практических результатов позволяет сделать вывод о том, что цель исследований достигнута, сформулированная научная проблема решена. Разработан единый математический аппарат, связывающий сложность моделей и данных в рамках строгой теоретической основы, применимой к широкому классу архитектур нейронных сетей. Получены вычислимые методы оценки сложности, не требующие прямого вычисления матриц Гессе для крупных моделей. Созданы практические инструменты для оценки достаточного объема выборки и снижения сложности моделей, подтвержденные экспериментально на реальных задачах. Перечисленные результаты получили высокую оценку научного сообщества при апробации на ведущих международных конференциях и в научных публикациях.

Общие свойства и определения

Свойство 1 (Норма матричного произведения). Пусть матрицы $\mathbf{A} \in \mathbb{R}^{m \times n}$ и $\mathbf{B} \in \mathbb{R}^{n \times q}$, тогда справедливо неравенство

$$\|\mathbf{AB}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{B}\|_2$$

Свойство 2 (Норма матричного произведения Кронекера). Пусть матрицы $\mathbf{A} \in \mathbb{R}^{m \times n}$ и $\mathbf{B} \in \mathbb{R}^{p \times q}$, тогда справедливо равенство

$$\|\mathbf{A} \otimes \mathbf{B}\|_2 = \|\mathbf{A}\|_2 \|\mathbf{B}\|_2$$

Свойство 3 (Норма матричного транспонирования). Пусть матрица $\mathbf{A} \in \mathbb{R}^{m \times n}$, тогда справедливо равенство

$$\|\mathbf{A}\|_2 = \|\mathbf{A}^\top\|_2$$

Свойство 4 (Соотношения между матричными нормами). Пусть матрица $\mathbf{A} \in \mathbb{R}^{m \times n}$, тогда следующие неравенства между матричными нормами справедливы:

$X \setminus Y$	$\ \mathbf{A}\ _{\max}$	$\ \mathbf{A}\ _1$	$\ \mathbf{A}\ _\infty$	$\ \mathbf{A}\ _2$	$\ \mathbf{A}\ _F$
$\ \mathbf{A}\ _{\max}$		1	1	1	1
$\ \mathbf{A}\ _1$	m		m	\sqrt{m}	\sqrt{m}
$\ \mathbf{A}\ _\infty$	n	n		\sqrt{n}	\sqrt{n}
$\ \mathbf{A}\ _2$	\sqrt{mn}	\sqrt{n}	\sqrt{m}		1
$\ \mathbf{A}\ _F$	\sqrt{mn}	\sqrt{n}	\sqrt{m}	\sqrt{d}	1

где $d = \text{rank}(\mathbf{A})$. Таблица читается следующим образом: для каждой пары норм $\|\cdot\|_X$ и $\|\cdot\|_Y$,

$$\|\mathbf{A}\|_X \leq c \cdot \|\mathbf{A}\|_Y$$

где $c \geq 0$ – неотрицательная константа на пересечении строки X и столбца Y .

Свойство 5 (Связь между vec и vec_r). Пусть задана матрица $\mathbf{A} \in \mathbb{R}^{m \times n}$. Тогда оператор построчной векторизации vec_r и стандартный оператор покоординатной векторизации vec связаны через операцию транспонирования:

$$\text{vec}_r(\mathbf{A}) = \text{vec}(\mathbf{A}^\top)$$

Свойство 6 (Норма матричной суммы). Пусть матрицы \mathbf{A} и \mathbf{B} принадлежат пространству матриц $\mathbb{R}^{m \times n}$, тогда

$$\|\mathbf{A} + \mathbf{B}\|_2 \leq \|\mathbf{A}\|_2 + \|\mathbf{B}\|_2$$

Свойство 7 (Неравенство для нормы блочной матрицы). Пусть матрица $\mathbf{A} \in \mathbb{R}^{m \times n}$ представляет собой блочную матрицу, разбитую на блоки $\mathbf{B}_{i,j}$ размеров $m_i \times n_j$, где $\sum_i m_i = m$ и $\sum_j n_j = n$. Тогда справедливо неравенство:

$$\|\mathbf{A}\|_2 \leq \sqrt{mn} \max_{i,j} \|\mathbf{B}_{i,j}\|_2$$

Отметим, что если матрица \mathbf{A} является блочно-диагональной, то имеет место строгое равенство $\|\mathbf{A}\|_2 = \max_i \|\mathbf{B}_{i,i}\|_2$, что следует из того, что спектральная норма блочно-диагональной матрицы равна максимальной спектральной норме диагональных блоков.

Свойство 8 (Поэлементное деление). Пусть задана матрица $\mathbf{A} \in \mathbb{R}^{m \times n}$ и вектор $\mathbf{b} \in \mathbb{R}^{m \times 1}$ такой, что $b_i \neq 0$ для всех $i = 1, \dots, m$. Тогда для матрицы $\mathbf{C} \in \mathbb{R}^{m \times n}$, где $c_{i,j} = \frac{a_{i,j}}{b_i}$, справедливо равенство

$$\mathbf{C} = \text{diag}^{-1}(\mathbf{b})\mathbf{A}$$

Свойство 9 (Производная матричного произведения). Пусть заданы матрицы $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{X} \in \mathbb{R}^{n \times q}$ и $\mathbf{B} \in \mathbb{R}^{q \times p}$, тогда

$$\frac{\partial \mathbf{AXB}}{\partial \mathbf{X}} = \mathbf{A} \otimes \mathbf{B}^\top$$

где \mathbf{A} и \mathbf{B} не зависят от \mathbf{X} .

Свойство 10 (Построчная векторизация произведения Адамара). Пусть заданы матрицы $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$. Тогда

$$\text{vec}_r(\mathbf{A} \circ \mathbf{B}) = \text{diag}(\text{vec}_r(\mathbf{A}))\text{vec}_r(\mathbf{B})$$

где \circ обозначает произведение Адамара.

Утверждение непосредственно следует из [116], где был получен аналогичный результат для покоординатной векторизации.

Свойство 11 (Построчная векторизация матричного произведения). Пусть заданы матрицы $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{X} \in \mathbb{R}^{n \times q}$ и $\mathbf{B} \in \mathbb{R}^{q \times p}$, тогда

$$\text{vec}_r(\mathbf{AXB}) = (\mathbf{A} \otimes \mathbf{B}^\top)\text{vec}_r(\mathbf{X})$$

Свойство 12 (Производная произведения Кронекера). Пусть заданы матрица $\mathbf{X} \in \mathbb{R}^{n \times q}$ и матрица $\mathbf{Y} \in \mathbb{R}^{p \times r}$. Тогда

$$\frac{\partial(\mathbf{X} \otimes \mathbf{Y})}{\partial \mathbf{X}} = (\mathbf{I}_n \otimes \mathbf{K}_{p,q} \otimes \mathbf{I}_r) (\mathbf{I}_{nq} \otimes \text{vec}_r(\mathbf{Y})),$$

и аналогично

$$\frac{\partial(\mathbf{X} \otimes \mathbf{Y})}{\partial \mathbf{Y}} = (\mathbf{I}_n \otimes \mathbf{K}_{p,q} \otimes \mathbf{I}_r) (\text{vec}_r(\mathbf{X}) \otimes \mathbf{I}_{pr}).$$

Определение 29 (Матрица перестановки). Матрица перестановки $\mathbf{K}_{m,n} \in \mathbb{R}^{mn \times mn}$ — это единственная матрица такая, что для любой матрицы $\mathbf{A} \in \mathbb{R}^{m \times n}$ выполняется:

$$\mathbf{K}_{m,n} \text{vec}(\mathbf{A}) = \text{vec}(\mathbf{A}^\top)$$

В силу свойства 5 справедливо соотношение:

$$\text{vec}_r(\mathbf{A}) = \mathbf{K}_{m,n} \text{vec}(\mathbf{A}) \quad \text{и} \quad \text{vec}(\mathbf{A}) = \mathbf{K}_{n,m} \text{vec}_r(\mathbf{A})$$

поскольку $\mathbf{K}_{n,m} \mathbf{K}_{m,n} = \mathbf{I}_{mn}$.

Определение 30 (Векторизация и поэлементные операции). Пусть задана матрица $\mathbf{A} \in \mathbb{R}^{m \times n}$ и вектор $\mathbf{v} \in \mathbb{R}^n$. Тогда

- $\text{vec}_r(\mathbf{A})$ обозначает построчную векторизацию матрицы \mathbf{A} , отображающую матрицу \mathbf{A} в вектор размерности $m n$ путем последовательной записи строк матрицы.
- $\mathbf{A}^{\circ \alpha}$ обозначает поэлементное возведение матрицы \mathbf{A} в степень α , то есть $(\mathbf{A}^{\circ \alpha})_{ij} = (\mathbf{A}_{ij})^\alpha$ для всех $i = 1, \dots, m$ и $j = 1, \dots, n$.
- $\text{diag}(\mathbf{v})$ создает диагональную матрицу размерности $n \times n$ с вектором \mathbf{v} на главной диагонали и нулями вне главной диагонали.

Определение 31 (Матричные нормы). Для матрицы $\mathbf{A} \in \mathbb{R}^{m \times n}$, где $r = \text{rank}(\mathbf{A})$ и $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$ — упорядоченные по убыванию сингулярные

числа матрицы \mathbf{A} (с учетом кратности), определены следующие нормы:

$$\|\mathbf{A}\|_2 = \sigma_1,$$

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2} = \sqrt{\sum_{i=1}^r \sigma_i^2},$$

$$\|\mathbf{A}\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}|,$$

$$\|\mathbf{A}\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}|,$$

$$\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|.$$

Дополнительные Леммы и утверждения

Лемма 35 (Производная умножения матричнозначных функций). *Пусть заданы $\mathbf{A}(\mathbf{X}) \in \mathbb{R}^{p \times r}$ и $\mathbf{B}(\mathbf{X}) \in \mathbb{R}^{r \times q}$ матричнозначные функции переменной \mathbf{X} , тогда справедливо равенство*

$$\frac{\partial \mathbf{A}(\mathbf{X})\mathbf{B}(\mathbf{X})}{\partial \mathbf{X}} = (\mathbf{A} \otimes \mathbf{I}_q) \frac{\partial \mathbf{B}}{\partial \mathbf{X}} + (\mathbf{I}_p \otimes \mathbf{B}^\top) \frac{\partial \mathbf{A}}{\partial \mathbf{X}}$$

Доказательство. Применим цепное правило для вычисления производной сложной функции, а затем объединим его со свойством 9

$$\begin{aligned} \frac{\partial \mathbf{A}(\mathbf{X})\mathbf{B}(\mathbf{X})}{\partial \mathbf{X}} &= \frac{\partial \mathbf{AB}}{\partial \mathbf{B}} \frac{\partial \mathbf{B}}{\partial \mathbf{X}} + \frac{\partial \mathbf{AB}}{\partial \mathbf{A}} \frac{\partial \mathbf{A}}{\partial \mathbf{X}} = \\ &= \frac{\partial \mathbf{ABI}_q}{\partial \mathbf{B}} \frac{\partial \mathbf{B}}{\partial \mathbf{X}} + \frac{\partial \mathbf{I}_p \mathbf{AB}}{\partial \mathbf{A}} \frac{\partial \mathbf{A}}{\partial \mathbf{X}} = \\ &= (\mathbf{A} \otimes \mathbf{I}_q) \frac{\partial \mathbf{B}}{\partial \mathbf{X}} + (\mathbf{I}_p \otimes \mathbf{B}^\top) \frac{\partial \mathbf{A}}{\partial \mathbf{X}} \end{aligned}$$

□

Лемма 36 (Теорема идентификации для построчной векторизации). *Пусть отображение $\mathbf{F} : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{p \times q}$ является дифференцируемой матричнозначной функцией от $\mathbf{X} \in \mathbb{R}^{m \times n}$. Если дифференциал функции \mathbf{F} может быть записан в виде:*

$$d\text{vec}_r(\mathbf{F}(\mathbf{X})) = \mathbf{J} \cdot d\text{vec}_r(\mathbf{X}),$$

где матрица $\mathbf{J} \in \mathbb{R}^{pq \times mn}$ является некоторой константной матрицей относительно переменной $d\mathbf{X}$, тогда матрица \mathbf{J} является матрицей Якоби преобразования $\mathbf{F}(\mathbf{X})$ относительно построчной векторизации. Обозначим это в следующем виде:

$$\frac{\partial \mathbf{F}(\mathbf{X})}{\partial \mathbf{X}} := \frac{\partial \text{vec}_r(\mathbf{F}(\mathbf{X}))}{\partial (\text{vec}_r(\mathbf{X}))^\top} = \mathbf{J}$$

Доказательство. Это построчный vec_r -аналог первой теоремы об идентификации (англ. first identification theorem) в главе 12 [116] для векторизации по столбцам. □

Лемма 37 (Производная квадрата Адамара). *Для матрицы $\mathbf{A} \in \mathbb{R}^{m \times n}$ справедливо равенство*

$$\frac{\partial \mathbf{A}^{\circ 2}}{\partial \mathbf{A}} = 2 \cdot \text{diag}(\text{vec}_r(\mathbf{A})).$$

Доказательство. Используем определение 30, а именно $(\mathbf{A}^{\circ 2})_{ij} = (\mathbf{A}_{ij})^2$. Выполняя поэлементное взятие дифференциала, получим:

$$d(\mathbf{A}^{\circ 2}) = 2\mathbf{A} \circ d\mathbf{A}.$$

Далее, применяя оператор vec_r и используя свойство 10, получим выражение:

$$\text{vec}_r(d(\mathbf{A}^{\circ 2})) = 2\text{diag}(\text{vec}_r(\mathbf{A}))\text{vec}_r(d\mathbf{A}),$$

причем, используя построчный аналог первой теоремы об идентификации 36, получим следующий вид:

$$\frac{\partial \mathbf{A}^{\circ 2}}{\partial \mathbf{A}} = \frac{\partial \text{vec}_r(\mathbf{A}^{\circ 2})}{\partial \text{vec}_r(\mathbf{A})} = 2 \cdot \text{diag}(\text{vec}_r(\mathbf{A})).$$

□

Лемма 38 (Производная корня Адамара). Для матрицы $\mathbf{A} \in \mathbb{R}^{m \times n}$ такой, что $a_{ij} > 0$ для всех $i = 1, \dots, m$ и $j = 1, \dots, n$, справедливо равенство

$$\frac{\partial \mathbf{A}^{\circ \frac{1}{2}}}{\partial \mathbf{A}} = \frac{1}{2}\text{diag}^{-1}(\text{vec}_r^{\circ \frac{1}{2}}(\mathbf{A})).$$

Доказательство. Аналогично доказательству леммы 37 получим $d(\mathbf{A}^{\circ 1/2}) = \frac{1}{2}\mathbf{A}^{\circ -1/2} \circ d\mathbf{A}$, откуда в векторном виде получим следующее выражение:

$$\frac{\partial \mathbf{A}^{\circ \frac{1}{2}}}{\partial \mathbf{A}} = \frac{\partial \text{vec}_r(\mathbf{A}^{\circ \frac{1}{2}})}{\partial \text{vec}_r(\mathbf{A})} = \frac{1}{2}\text{diag}^{-1}(\text{vec}_r^{\circ \frac{1}{2}}(\mathbf{A})).$$

□

Лемма 39 (Производная обратной матрицы). Пусть задана обратимая квадратная матрица $\mathbf{D} \in \mathbb{R}^{n \times n}$ такая, что $\det(\mathbf{D}) \neq 0$, тогда справедливо равенство:

$$\frac{\partial \mathbf{D}^{-1}}{\partial \mathbf{D}} = -\mathbf{D}^{-1} \otimes \mathbf{D}^{-\top}.$$

Доказательство. По определению из [117] и [116]:

$$d(\mathbf{D}^{-1}) = -\mathbf{D}^{-1}(d\mathbf{D})\mathbf{D}^{-1}.$$

Используя оператор vec_r и свойство 11, получим

$$\text{vec}_r(-\mathbf{D}^{-1}(d\mathbf{D})\mathbf{D}^{-1}) = (-\mathbf{D}^{-1} \otimes \mathbf{D}^{-\top})\text{vec}_r(d\mathbf{D}),$$

причем, используя лемму 36, получим:

$$\text{vec}_r(d\mathbf{D}^{-1}) = \frac{\partial \text{vec}_r \mathbf{D}^{-1}}{\partial \text{vec}_r \mathbf{D}} \text{vec}_r(d\mathbf{D}).$$

Следовательно, получим выражение, которое завершает доказательство леммы:

$$\frac{\partial \text{vec}_r \mathbf{D}^{-1}}{\partial \text{vec}_r \mathbf{D}} = (-\mathbf{D}^{-1} \otimes \mathbf{D}^{-\top}).$$

□

Лемма 40 (Производная $\text{diag}(\cdot)$). Для вектора $\mathbf{v} \in \mathbb{R}^{L \times 1}$ справедливо равенство:

$$\frac{\partial \text{diag}(\mathbf{v})}{\partial \mathbf{v}} = (\mathbf{e}_1 \otimes \mathbf{e}_1 \quad \dots \quad \mathbf{e}_L \otimes \mathbf{e}_L),$$

где векторы \mathbf{e}_i являются базисными в пространстве \mathbb{R}^L .

Доказательство. По определению 30 оператор $\text{diag}(\mathbf{v})$ отображает элемент v_i в позицию (i, i) результирующей диагональной матрицы. Тогда производная оператора $\text{diag}(\mathbf{v})$ является матрицей $\mathbf{E}_{ii} = \mathbf{e}_i \mathbf{e}_i^\top$, в которой 1 в позиции (i, i) и 0 иначе. Причем, используя свойство 11 и применяя оператор построчной векторизации, получим:

$$\text{vec}_r(\mathbf{E}_{i,i}) = \mathbf{e}_i \otimes \mathbf{e}_i.$$

Таким образом, применяя для всех $i = 1, \dots, L$, матрица Якоби принимает вид:

$$\frac{\partial \text{diag}(\mathbf{v})}{\partial \mathbf{v}} = (\mathbf{e}_1 \otimes \mathbf{e}_1 \quad \dots \quad \mathbf{e}_L \otimes \mathbf{e}_L).$$

□

Лемма 41 (Производная транспонированной матрицы). Пусть задана матрица $\mathbf{A} \in \mathbb{R}^{m \times n}$, тогда справедливо следующее равенство:

$$\frac{\partial \mathbf{A}^\top}{\partial \mathbf{A}} = \mathbf{K}_{n,m},$$

где матрица $\mathbf{K}_{n,m}$ является коммутационной матрицей (англ. *commutation matrix*) описанной в определении 29.

Доказательство. Объединяя аналогичное утверждение из [116] для постолбцовой векторизации с правилом соединения столбцов и строк 5 и 29, получим утверждение леммы. □

Лемма 42 (Производная произведения Кронекера матричнозначных функций). Пусть заданы матричнозначные функции $\mathbf{A}(\mathbf{X}) \in \mathbb{R}^{n \times q}$ и $\mathbf{B}(\mathbf{X}) \in \mathbb{R}^{p \times r}$ переменной $\mathbf{X} \in \mathbb{R}^{m \times s}$, тогда справедливо равенство

$$\frac{\partial \mathbf{A}(\mathbf{X}) \otimes \mathbf{B}(\mathbf{X})}{\partial \mathbf{X}} = (\mathbf{I}_n \otimes \mathbf{K}_{p,q} \otimes \mathbf{I}_r) \left((\text{vec}_r \mathbf{A} \otimes \mathbf{I}_{pr}) \frac{\partial \mathbf{B}}{\partial \mathbf{X}} + (\mathbf{I}_{nq} \otimes \text{vec}_r \mathbf{B}) \frac{\partial \mathbf{A}}{\partial \mathbf{X}} \right).$$

Доказательство. Применим цепное правило для вычисления производной сложной функции, а затем объединим его со свойством 12

$$\begin{aligned} \frac{\partial \mathbf{A}(\mathbf{X}) \otimes \mathbf{B}(\mathbf{X})}{\partial \mathbf{X}} &= \frac{\partial \mathbf{A} \otimes \mathbf{B}}{\partial \mathbf{B}} \frac{\partial \mathbf{B}}{\partial \mathbf{X}} + \frac{\partial \mathbf{A} \otimes \mathbf{B}}{\partial \mathbf{A}} \frac{\partial \mathbf{A}}{\partial \mathbf{X}} = \\ &= (\mathbf{I}_n \otimes \mathbf{K}_{p,q} \otimes \mathbf{I}_r) (\text{vec}_r \mathbf{A} \otimes \mathbf{I}_{pr}) \frac{\partial \mathbf{B}}{\partial \mathbf{X}} + \\ &\quad + (\mathbf{I}_n \otimes \mathbf{K}_{p,q} \otimes \mathbf{I}_r) (\mathbf{I}_{nq} \otimes \text{vec}_r \mathbf{B}) \frac{\partial \mathbf{A}}{\partial \mathbf{X}} = \\ &= (\mathbf{I}_n \otimes \mathbf{K}_{p,q} \otimes \mathbf{I}_r) \left((\text{vec}_r \mathbf{A} \otimes \mathbf{I}_{pr}) \frac{\partial \mathbf{B}}{\partial \mathbf{X}} + (\mathbf{I}_{nq} \otimes \text{vec}_r \mathbf{B}) \frac{\partial \mathbf{A}}{\partial \mathbf{X}} \right). \end{aligned}$$

□

Лемма 43 (Спектральная норма матрицы из единиц). Пусть задана матрица $\mathbf{A} = \mathbf{1}_{L \times L}$, все элементы которой равны единице. Тогда ее спектральная норма принимает следующее значение:

$$\|\mathbf{A}\|_2 = L$$

Доказательство. Используя основные свойства линейной алгебры, получим $\text{tr}(\mathbf{A}) = L$ и $\text{rank}(\mathbf{A}) = 1 = \dim(\text{Im}(\mathbf{A}))$. Следовательно, используя $\dim(\text{Im}(\mathbf{A})) + \dim(\text{Ker}(\mathbf{A})) = L$, получим $\dim(\text{Ker}(\mathbf{A})) = L - 1$. Таким образом, для $i \in \{2, \dots, L\}$ имеем $\lambda_i = 0$, а $\lambda_1 = L$. Тогда единственное ненулевое сингулярное число матрицы \mathbf{A} равно $\sqrt{L^2} = L$. Следовательно, получим, что $\|\mathbf{A}\|_2 = L$, в соответствии с определением 31. □

Список иллюстраций

2.1	Изменение функции потерь $\mathcal{L}_k(\boldsymbol{\theta})$ при добавлении нового объекта в выборку. Иллюстрация демонстрирует зависимость абсолютной разности $ \mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta}) $ от размера выборки k , что является основой для анализа сходимости ландшафта и определения ландшафтной меры сложности модели.	43
2.2	Проверка предположения 1 о сходимости локальных минимумов при увеличении объема выборки. График демонстрирует выполнимость предположения о том, что локальный минимум $\boldsymbol{\theta}^*$ функции потерь $\mathcal{L}_k(\boldsymbol{\theta})$ остается локальным минимумом функции $\mathcal{L}_{k+1}(\boldsymbol{\theta})$ при добавлении нового объекта, причем выполнимость улучшается с увеличением длины последовательностей.	58
2.3	Зависимость абсолютного значения разности функций потерь $ \mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta}) $ от размера выборки k для полносвязной нейронной сети на наборе данных MNIST. Левый график: при $L = 5$ уменьшение разности с увеличением размера скрытого слоя h от 4 до 64; правый график: при $h = 16$ увеличение разности с увеличением количества слоев L от 1 до 10. Результаты подтверждают теорему 5.	58
2.4	Зависимость абсолютного значения разности функций потерь $ \mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta}) $ от размера выборки k для полносвязной нейронной сети с предобученным экстрактором признаков (Vision Transformer, ViT) на наборе данных MNIST. Левый график: при $L = 5$ уменьшение разности с увеличением h от 4 до 64; правый график: при $h = 16$ увеличение разности с увеличением L от 1 до 10. Результаты подтверждают независимость сходимости от природы пространства исходных объектов и согласуются с теоремой 5.	59
2.5	Зависимость абсолютного значения разности функций потерь $ \mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta}) $ от размера выборки k для полносвязной нейронной сети при прямой классификации на наборах данных FashionMNIST, CIFAR10 и CIFAR100. Верхний ряд: при $L = 5$ уменьшение разности с увеличением h от 4 до 64; нижний ряд: при $h = 16$ увеличение разности с увеличением L от 1 до 10. Результаты подтверждают теорему 5 для различных наборов данных.	60

2.6	Зависимость абсолютного значения разности функций потерь $ \mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta}) $ от размера выборки k для полносвязной нейронной сети с предобученным экстрактором признаков (англ. Vision Transformer, ViT) на наборах данных FashionMNIST, CIFAR10 и CIFAR100. Верхний ряд: при $L = 5$ уменьшение разности с увеличением h от 4 до 64; нижний ряд: при $h = 16$ увеличение разности с увеличением L от 1 до 10. Результаты подтверждают независимость сходимости от природы пространства исходных объектов и согласуются с теоремой 5.	66
2.7	Зависимость абсолютного значения разности функций потерь $ \mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta}) $ от размера выборки k для сверточной нейронной сети при изменении количества сверточных слоев L при фиксированных размере ядра $k = 3$ и количестве каналов $C = 6$. Графики демонстрируют немонотонный характер зависимости разности потерь от количества слоев.	67
2.8	Зависимость абсолютного значения разности функций потерь $ \mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta}) $ от размера выборки k для сверточной нейронной сети при изменении размера ядра свертки k при фиксированных количестве слоев L и количестве каналов $C = 6$. Графики демонстрируют немонотонный характер зависимости разности потерь от размера ядра.	67
2.9	Зависимость абсолютного значения разности функций потерь $ \mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta}) $ от размера выборки k для сверточной нейронной сети при изменении количества каналов C при фиксированных количестве слоев L и размере ядра $k = 3$. Графики демонстрируют монотонный характер зависимости: увеличение числа каналов приводит к увеличению разности потерь, что согласуется с теоретическими оценками.	68
2.10	Зависимость абсолютного значения разности функций потерь $ \mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta}) $ от размера выборки k для сверточной нейронной сети при изменении позиции операции пулинга на наборе данных CIFAR10 при фиксированных количестве слоев L , размере ядра $k = 3$ и количестве каналов C . Графики демонстрируют монотонную зависимость: более раннее применение пулинга приводит к меньшей разности потерь.	68

2.11 Зависимость абсолютного значения разности функций потерь $ \mathcal{L}_{k+1}(\mathbf{w}) - \mathcal{L}_k(\mathbf{w}) $ от количества обучающих примеров k для трансформерной модели на наборе данных CIFAR-100, отображенная в двойном логарифмическом масштабе. Результаты подтверждают теорему 8 и демонстрируют стабилизацию ландшафта функции потерь с ростом объема данных.	69
3.1 Визуализация элементов матрицы Гессе для инициализированной модели трансформера (один блок, набор данных MNIST, батч из 128 примеров) в логарифмическом масштабе. Наблюдается неоднородность величин элементов, при этом блоки, соответствующие параметрам Values, демонстрируют наибольшие значения.	133
3.2 Визуализация элементов матрицы Гессе для обученной модели трансформера (один блок, набор данных MNIST, точность на валидации $>50\%$) в логарифмическом масштабе. После обучения все блоки гессиана приобретают большие значения, при этом блок Values-Values демонстрирует максимальные величины, что подтверждает гетерогенность вклада различных параметров в кривизну функции потерь.	134
3.3 Спектральные нормы блоков параметров трансформера (слева) и спектральные нормы соответствующих блоков матрицы Гессе (справа), вычисленные на одном батче из 128 примеров набора данных MNIST. Наибольшие значения норм соответствуют блокам Keys и Values, что согласуется с теоретическими оценками теоремы 27.	134
3.4 Визуализация элементов блока матрицы Гессе, соответствующего параметрам Queries трансформера (один блок, набор данных MNIST). Блок демонстрирует структурированное распределение элементов, отражающее взаимосвязи между параметрами механизма самовнимания.	135
3.5 Визуализация элементов блока матрицы Гессе, соответствующего параметрам Keys трансформера (один блок, набор данных MNIST). Блок демонстрирует значительные значения элементов, что подтверждает важную роль параметров Keys в формировании кривизны функции потерь.	135

3.6	Визуализация элементов блока матрицы Гессе, соответствующего параметрам Values трансформера (один блок, набор данных MNIST). Блок демонстрирует наибольшие значения элементов среди всех блоков гессиана, что согласуется с теоретическими оценками и подтверждает доминирующий вклад параметров Values в общую кривизну функции потерь.	135
3.7	Визуализация элементов блока матрицы Гессе, соответствующего параметрам LayerNorm трансформера (один блок, набор данных MNIST). Блок демонстрирует структурированное распределение элементов, отражающее влияние нормализации слоев на локальную геометрию функции потерь.	136
3.8	Визуализация элементов блока матрицы Гессе, соответствующего параметрам блока FFN (Feed-Forward Network) трансформера (один блок, набор данных MNIST). Блок демонстрирует умеренные значения элементов по сравнению с блоками Keys и Values, что отражает специфику вклада полно связанных слоев в кривизну функции потерь.	136
4.1	Визуализация сдвига апостериорных распределений параметров модели при последовательном добавлении объектов в выборку. Иллюстрация демонстрирует концепцию близости распределений, используемую в методах KL- и S-достаточности для определения достаточного размера выборки.	155
4.2	Зависимость статистических значений различных методов определения достаточного размера выборки от размера подвыборки для наборов данных Boston Housing, Diabetes, Forest Fires, Servo и NBA. Все представленные функции монотонны и асимптотически стремятся к константе, что подтверждает корректность методов.	167
4.3	Зависимость оцененного достаточного размера выборки t^* от доступного размера выборки t для различных методов на наборе данных Boston Housing. Результаты демонстрируют сходимость методов и низкую дисперсию оценок, что указывает на вычислительную устойчивость рассмотренных подходов.	168
4.4	Сходимость функций $D(k)$ и $M(k)$ для синтетического набора данных регрессии. Обе функции стремятся к нулю с увеличением размера выборки, что подтверждает теорему 28.	169

- 4.5 Сходимость функций $D(k)$ и $M(k)$ для синтетического набора данных классификации. Обе функции демонстрируют монотонное убывание к нулю с увеличением размера выборки, подтверждая применимость методов D- и M-достаточности для задач классификации. 169
- 4.6 Сходимость функций $D(k)$ и $M(k)$ для набора данных Liver Disorders (345 объектов, 5 признаков, $B = 1000$ бутстрэп-подвыборок). Обе функции демонстрируют сходимость к нулю, что подтверждает теоретические результаты и демонстрирует применимость методов на реальных данных. 170
- 4.7 Зависимость достаточного размера выборки m^* от порогового параметра ε для методов D- и M-достаточности на трех наборах данных. С увеличением значения порога ε достаточный размер выборки монотонно уменьшается, что позволяет выбирать меньше объектов для достижения заданного уровня стабильности функций $D(k)$ и $M(k)$ 170
- 4.8 Зависимость минимального собственного значения $\lambda_{\min}(\mathbf{X}_k^\top \mathbf{X}_k)$ от размера выборки k для синтетических данных регрессии (500 объектов, 10 признаков) и набора данных Liver Disorders (345 объектов, 5 признаков). Асимптотическое поведение соответствует условию теоремы 31: $\lambda_{\min} = \omega(\sqrt{k})$ при $k \rightarrow \infty$ 172
- 4.9 Сходимость функций $KL(k)$ и $S(k)$ для синтетического набора данных регрессии. Функция $KL(k)$ стремится к нулю, а $S(k)$ стремится к единице, что подтверждает теоремы 29 и 30. 173
- 4.10 Сходимость функций $KL(k)$ и $S(k)$ для набора данных Liver Disorders (345 объектов, 5 признаков, нормальное априорное распределение, $B = 100$ повторений). Результаты демонстрируют сходимость $KL(k)$ к нулю и $S(k)$ к единице, подтверждая теоретические предсказания и применимость методов KL- и S-достаточности на реальных данных. 173
- 4.11 Зависимость достаточного размера выборки m^* от порогового параметра ε для методов KL- и S-достаточности на синтетических данных регрессии и наборе данных Liver Disorders. Метод S-достаточности требует более низких значений порога для достижения заданного уровня близости распределений, что указывает на его более строгие требования к качеству оценки. 174

4.12 Зависимость оцененного достаточного размера выборки m^* от до- ступного размера выборки m для классических методов и предло- женных методов KL- и S-достаточности на наборе данных Boston Housing. Критерий KL-достаточности является наиболее консер- вативным и требует почти полной выборки, в то время как S- достаточность указывает на минимальные размеры, что связано с высокой чувствительностью расхождения Кульбака-Лейблера к изменениям распределений.	176
5.1 Иллюстрация метода Белсли для анализа мультиколлинеарности параметров на синтетических данных. Слева показана матрица ковариации параметров, справа — дисперсионные доли, демон- стрирующие вклад каждого признака в дисперсию параметров модели.	184
5.2 Схема базовой дистилляции моделей глубокого обучения, где мо- дель учителя обучается на большом наборе данных из генераль- ной совокупности A , а затем ее выходы используются для обуче- ния модели ученика на меньшем наборе данных из того же домена.	186
5.3 Схема дистилляции моделей глубокого обучения с доменной адап- тацией, где модель учителя обучается на данных из домена A , а модель ученика — на данных из домена B , связанных инъективи- вным отображением φ	186
5.4 Зависимость точности классификации R_{cl} от процента удаленных параметров для различных методов прореживания на выборке Wine. Метод оптимального прореживания, вариационный метод и метод Белсли позволяют удалить $\approx 80\%$ параметров без суще- ственной потери качества.	193
5.5 Поверхности изменения уровня шума ответов нейросети при из- менении процента удаленных параметров и уровня шума вход- ных данных для различных методов прореживания на выборке Wine. Метод Белсли демонстрирует наименьший уровень шума, что видно по более низкому положению соответствующей поверх- ности.	194

5.6	Зависимость среднеквадратического отклонения прогноза R_{rg} от процента удаленных параметров для различных методов прореживания на выборке Boston Housing. Метод Белсли является наиболее эффективным, позволяя удалить больше параметров без потери качества.	195
5.7	Поверхности изменения уровня шума ответов нейросети при изменении процента удаленных параметров и уровня шума входных данных для различных методов прореживания на выборке Boston Housing. Все методы демонстрируют одинаковый уровень шума, так как соответствующие поверхности находятся на одном уровне.	196
5.8	Зависимость среднеквадратического отклонения прогноза от процента удаленных параметров для различных методов прореживания на синтетической выборке с мультиколлинеарными параметрами. Метод Белсли является наиболее эффективным, поскольку качество прогноза повышается при удалении шумовых параметров.	197
5.9	Поверхности изменения уровня шума ответов нейросети при изменении процента удаленных параметров и уровня шума входных данных для различных методов прореживания на синтетической выборке. Метод Белсли демонстрирует наименьший уровень шума, что видно по более низкому положению соответствующей поверхности.	198
5.10	Пример применения инъективного отображения φ для доменной адаптации между ImageNet-Small и ImageNet-Big. Демонстрирует визуальное преобразование объекта из исходного домена в целевой домен для использования в мультидоменной дистилляции. . .	201
5.11	Зависимость точности аппроксимации от числа эпох обучения на тестовой выборке ImageNet для различных методов дистилляции. Результаты усреднены по 5 запускам и демонстрируют преимущество методов с использованием учителя и доменной адаптации. .	202
5.12	Зависимость ошибки перекрестной энтропии между истинными и предсказанными метками от числа эпох обучения на тестовой выборке ImageNet для различных методов дистилляции. Результаты усреднены по 5 запускам и показывают снижение ошибки при использовании методов дистилляции.	202

5.13	Зависимость ошибки перекрестной энтропии от числа эпох обучения на тестовом наборе данных OPUS-100 для задачи машинного перевода. Результаты усреднены по 3 запускам и демонстрируют преимущество моделей, обученных с использованием учителя.	203
5.14	Сравнение валидационной точности для различных методов инициализации параметров модели на наборе данных Fashion-MNIST. Модели, использующие Анти-Дистилляцию, демонстрируют меньшую дисперсию и более высокую точность по сравнению с моделями с различной инициализацией.	206
5.15	Зависимость валидационной точности от уровня адверсарного шума в данных для различных методов инициализации на наборе данных Fashion-MNIST. Анти-Дистилляция является наиболее устойчивым методом, демонстрируя наивысшую точность при высоких уровнях шума.	206
5.16	Зависимость валидационной точности от параметра интенсивности нормального шума ε в параметрах модели для различных методов инициализации на наборе данных Fashion-MNIST. Метод Анти-Дистилляции без регуляризации гессиана ($\lambda_4 = 0$) является наиболее устойчивым, сохраняя наивысшую точность при максимальном уровне шума.	207
6.1	Схема метода декодирования фМРТ-изображений из видеопоследовательностей. Метод использует архитектуру ResNet152 для векторизации видеокадров и линейную регрессию с L_2 -регуляризацией для предсказания разностей между последовательными фМРТ-снимками с учетом временной задержки Δt	224
6.2	Зависимость метрики F_1 от ранга r адаптеров LoRA для модели DeBERTa-v3-base на задаче детекции машинно-генерированного текста. Наилучшая производительность достигается при $r = 8$, что указывает на снижение отдачи при увеличении ранга и подтверждает эффективность низкоранговой параметризации.	229

6.3	Динамика потерь при обучении модели DeBERTa-v3-base с адаптерами LoRA различных рангов ($r \in \{8, 16, 32\}$) на задаче детекции машинно-генерированного текста. Анализ показывает, что более высокие ранги обеспечивают более быструю сходимость и более низкие финальные потери, но $r = 8$ обеспечивает лучшее обобщение, что подтверждает оптимальность низкоранговой параметризации.	236
6.4	Динамика нормы градиента и потерь при обучении моделей DeBERTa-v3-base с полной тонкой настройкой и адаптацией LoRA ($r = 8$) в течение 5 эпох на задаче детекции машинно-генерированного текста. LoRA демонстрирует более гладкие траектории нормы градиента и потерь, что указывает на более стабильную сходимость и большую эмпирическую устойчивость по сравнению с полной тонкой настройкой.	237
6.5	Зависимость метрики MSE от параметра регуляризации α для метода декодирования фМРТ-изображений на тестовой выборке при различных коэффициентах сжатия (1, 2, 4, 8). Оптимальное значение коэффициента $\alpha \approx 1000$ сохраняется независимо от коэффициента сжатия, что подтверждает стабильность метода при снижении размерности данных.	239
6.6	Распределения размерности персистентной гомологии (PHD) для человеческих и машинно-сгенерированных текстов в различных наборах данных детекции. Качественные наборы данных (SemEval, PAN24, MGT-1) демонстрируют схожие распределения PHD для обоих типов текстов, что указывает на высокое качество сгенерированных данных и их близость к человеческим текстам по топологическим характеристикам.	242
6.7	Топологические временные ряды (TTS) для четырех наборов данных с высокими значениями метрики KL _{TTS} : GhostBuster, PAN24, MGBTBench и DAGPap22. Высокие значения KL _{TTS} для GhostBuster и PAN24 обусловлены расхождением в текстах с более высокими размерностями, тогда как для MGBTBench и DAGPap22 — разницей в самих распределениях PHD, что указывает на различные типы различий между человеческими и машинно-сгенерированными текстами.	243

Список таблиц

2.1	Описание наборов данных для классификации изображений, использованных в экспериментах по валидации теоретических оценок сходимости ландшафта функции потерь. Все наборы данных из библиотеки <code>torchvision</code> с нормализацией значений пикселей к диапазону $[-1, 1]$	59
3.1	Гиперпараметры архитектур Vision Transformer (ViT) для экспериментов по анализу матриц Гессе. Параметры включают размер патча, скрытую размерность, размер блока FFN и количество трансформерных блоков для наборов данных MNIST и CIFAR-100.133	
4.1	Характеристики выборок, используемых для анализа качества методов определения достаточного размера выборки. Таблица содержит информацию о типе задачи, количестве признаков и общем размере выборки для каждого набора данных.	165
4.2	Сравнение оценок достаточного размера выборки, полученных различными статистическими и байесовскими методами для пяти наборов данных. Результаты демонстрируют значительный разброс оценок между методами, что указывает на различную консервативность подходов.	166
4.3	Гиперпараметры методов оценки достаточного размера выборки, установленные экспертизно для экспериментов. Параметры включают уровни значимости α , вероятности ошибки второго рода β , пороговые значения ε и l , а также параметры обобщенных линейных моделей.	166
4.4	Сравнение оценок достаточного размера выборки методами D- и M-достаточности для 13 наборов данных с задачей регрессии. Методы демонстрируют сопоставимые результаты для большинства наборов данных, при этом M-достаточность иногда требует большего размера выборки.	171
4.5	Характеристики выборок, используемых для сравнения методов определения достаточного размера выборки на основе близости апостериорных распределений. Все наборы данных соответствуют задаче регрессии и используются для оценки методов KL- и S-достаточности.	174

4.6 Сравнение оценок достаточного размера выборки, полученных классическими методами и предложенными методами KL- и S-достаточности для четырех наборов данных регрессии. Метод KL-достаточности дает более консервативные оценки, требующие почти полной выборки, в то время как S-достаточность указывает на минимальные размеры выборки.	175
5.1 Индексы обусловленности η и дисперсионные доли q_j для синтетических данных, демонстрирующие работу метода Белсли. Максимальный индекс обусловленности $\eta_6 = 1.2 \cdot 10^{16}$ соответствует максимальным дисперсионным долям признаков с индексами 1 и 4, которые являются линейно зависимыми.	184
5.2 Характеристики выборок, использованных для анализа метода задания порядка параметров методом Белсли. Включает реальные данные (Wine для классификации, Boston Housing для регрессии) и синтетические данные с мультиколлинеарными параметрами.	192
5.3 Архитектура модели учителя для эксперимента по мультидоменной дистилляции в задаче компьютерного зрения. Модель представляет собой сверточную нейронную сеть с пятью сверточными слоями и четырьмя полно связанными слоями, общее число параметров составляет 4 455 984.	199
5.4 Архитектура модели ученика для эксперимента по мультидоменной дистилляции в задаче компьютерного зрения. Модель имеет упрощенную структуру по сравнению с моделью учителя, с двумя сверточными слоями и тремя полно связанными слоями, общее число параметров составляет 12 755 640.	200
5.5 Характеристики подмножеств набора данных ImageNet, использованных в эксперименте по мультидоменной дистилляции. ImageNet-Big содержит 1 281 167 изображений для обучения и 50 000 для валидации, ImageNet-Small — 64 058 и 2 500 соответственно, оба набора содержат 200 классов.	200
5.6 Характеристики подмножеств набора данных OPUS-100, использованных в эксперименте по мультидоменной дистилляции для задачи машинного перевода. Показаны размеры обучающих и валидационных выборок для языковых пар fr-en и de-en.	201

5.7	Сравнение качества моделей для задачи компьютерного зрения на наборе данных ImageNet. Показаны валидационная точность, потери и интегральный критерий для различных комбинаций учителя и ученика с использованием доменной адаптации и без нее, результаты усреднены по нескольким запускам.	203
5.8	Сравнение качества моделей для задачи машинного перевода на наборе данных OPUS-100. Показаны потери перекрестной энтропии и метрика BLEU для модели ученика, обученной с использованием учителя и доменной адаптации (NLLB) и без них.	204
5.9	Сравнение точности на валидационном наборе Fashion-MNIST для различных методов инициализации параметров модели. Показаны базовая точность, точность при адверсарной атаке FSGM и точность при нормальном шуме в параметрах, результаты усреднены по нескольким запускам.	208
6.1	Сравнение формул обновления параметров для полной тонкой настройки и метода LoRA. В полной тонкой настройке обновление ΔW изучается напрямую, тогда как в LoRA обновление параметризуется в виде низкоранговой факторизации AB , где $r \ll \min(d, k)$, что значительно сокращает количество обучаемых параметров.	213
6.2	Сравнение производительности модели DeBERTa-v3-base при полной тонкой настройке и адаптации с LoRA на задаче детекции машинно-генерированного текста. LoRA демонстрирует незначительное снижение метрик точности (1.4–3.2%), но обеспечивает существенное снижение потерь на валидации (36.3%) и ускорение обучения (12.6%), что подтверждает теоретические гарантии корректности выходного слоя при низкоранговых обновлениях.	231
6.3	Сравнение производительности и эмпирической сложности Радемахера (ERC) модели DeBERTa-v3-base в условиях однозадачного (STL) и многозадачного (MTL) обучения на задаче бинарной классификации GenAI Detection. MTL демонстрирует улучшение всех метрик качества (F1: $0.781 \rightarrow 0.826$, ROC-AUC: $0.788 \rightarrow 0.834$) и снижение ERC с 0.0159 ± 0.0009 до 0.0111 ± 0.0010 , что согласуется с теоретическим предсказанием снижения сложности на задачу в $1/\sqrt{T}$ раз.	232

6.4 Сравнение моделей DeBERTa-v3-base при полной тонкой настройке и адаптации с LoRA ($r = 8$) после 5 эпох обучения. LoRA обеспечивает более низкие потери (0.2972 vs 0.3058), более гладкие нормы градиентов (17.41 vs 24.65) и требует только 0.16% параметров от полной модели (296,450 vs 184M), демонстрируя эффективность низкоранговой адаптации.	233
6.5 Сравнение конфигураций LoRA с различными рангами $r \in \{8, 16, 32\}$ для модели DeBERTa-v3-base. Анализ показывает, что более высокие ранги ($r = 32$) обеспечивают более низкие потери при обучении, но $r = 8$ обеспечивает лучшее обобщение, что подтверждает оптимальность низкоранговой параметризации для данной задачи.	233
6.6 Характеристики выборки для эксперимента по декодированию фМРТ-изображений из видеопоследовательностей. Выборка содержит данные 30 испытуемых с продолжительностью обследования 390 с, частотой кадров видео 25 Гц и частотой фМРТ-изображений 1.64 Гц, что обеспечивает временное соответствие между видеокадрами и томографическими данными.	238
6.7 Зависимость среднего времени обучения модели декодирования фМРТ-изображений от коэффициента сжатия данных. Использование предварительного сжатия фМРТ-изображений с коэффициентами 2, 4 и 8 обеспечивает существенное сокращение времени обучения (с 36.3 с до 6.7 с, 1.6 с и 1.4 с соответственно), что демонстрирует эффективность снижения сложности данных без потери качества реконструкции.	239
6.8 Результаты классификации различных детекторов машинно-генерированного текста (DeBERTa, Binoculars, DetectGPT) на множественных наборах данных, оцененные с помощью метрики F_1 -score. Детектор на основе DeBERTa демонстрирует наиболее стабильную производительность на различных наборах данных, тогда как Binoculars и DetectGPT показывают значительные вариации, что указывает на проблемы с устойчивостью этих методов к различным доменам.	240

- 6.9 Статистика качества данных для выбранных наборов данных детекции машинно-генерированного текста, включающая метрики KL_{TTS} , PHD , Δ_{shift} и $KL_{shuffle}$. Высокие значения метрик указывают на различимость текстов разного происхождения, тогда как низкие значения отражают схожую устойчивость к модификациям, что является признаком качественных данных. 241

Список литературы

1. Стрижов Вадим Викторович. Порождение и выбор моделей в задачах прогнозирования. — ВЦ РАН, Москва, Россия: Рукопись, 2014.
2. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов: статистические проблемы обучения. — Nauka, 1974.
3. Valiant L. G. A theory of the learnable // *Commun. ACM*. — 1984. — nov. — Vol. 27, no. 11. — P. 1134–1142.
4. Koltchinskii Vladimir, Panchenko Dmitriy. Rademacher Processes and Bounding the Risk of Function Learning // High Dimensional Probability II / Ed. by Evarist Giné, David M. Mason, Jon A. Wellner. — Boston, MA: Birkhäuser Boston, 2000. — Pp. 443–457.
5. Vorontsov K. V. Combinatorial Bounds for Learning Performance // *Doklady Mathematics*. — 2004. — Vol. 69, no. 1. — Pp. 145—148.
6. Cybenko George V. Approximation by superpositions of a sigmoidal function // *Mathematics of Control, Signals and Systems*. — 1989. — Vol. 2. — Pp. 303–314. <https://api.semanticscholar.org/CorpusID:3958369>.
7. Håstad Johan. Computational limitations of small-depth circuits. — Cambridge, MA, USA: MIT Press, 1987.
8. Bengio Yoshua, Lecun Yann. Scaling Learning Algorithms towards AI // Large-Scale Kernel Machines / Ed. by L. Bottou, O. Chapelle, D. Decoste, J. Weston. — MIT Press, 2007.
9. Delalleau Olivier, Bengio Yoshua. Shallow vs. Deep Sum-Product Networks // Advances in Neural Information Processing Systems / Ed. by J. Shawe-Taylor, R. Zemel, P. Bartlett et al. — Vol. 24. — Curran Associates, Inc., 2011.
10. Cohen Nadav, Sharir Or, Shashua Amnon. On the Expressive Power of Deep Learning: A Tensor Analysis // Annual Conference Computational Learning Theory. — 2015. <https://api.semanticscholar.org/CorpusID:215826733>.
11. Eldan Ronen, Shamir Ohad. The Power of Depth for Feedforward Neural Networks // *Conference on Learning Theory*. — 2015. — 12.
12. MacKay David J. C. Information Theory, Inference, and Learning Algorithms. — Copyright Cambridge University Press, 2003.

13. Numerical Methods of Sufficient Sample Size Estimation for Generalised Linear Models / A. V. Grabovoy, T. T. Gadaev, A. P. Motrenko, V. V. Strijov // *Lobachevskii Journal of Mathematics*. — 2022. — Vol. 43, no. 9. — Pp. 2453–2462.
14. Demidenko E. Sample size determination for logistic regression revisited // *Statistics in medicine*. — 2006. — Vol. 26. — Pp. 3385–97.
15. Joseph L., Berger R., Bélisle P. Bayesian and mixed Bayesian/likelihood criteria for sample size determination // *Statistician*. — 1997. — Vol. 16, no. 7. — Pp. 769–781.
16. Lawrence J., Wolfson D., Berger R. Sample Size Calculations for Binomial Proportions Via Highest Posterior Density Intervals // *Statistician*. — 1995. — Vol. 44. — Pp. 143–154.
17. Kloek T. Note on a large-sample result in specification analysis // *Econometrica*. — 1975. — Vol. 43. — Pp. 933–936.
18. Lindley D. The choice of sample size // *Statistician*. — 1997. — Vol. 46. — Pp. 129–138.
19. Motrenko A., Strijov V., Weber G. Sample Size Determination for Logistic Regression // *J. Comput. Appl. Math.* — 2014. — Vol. 255, no. C. — Pp. 743–752.
20. Qumsiyeh M. Using the bootstrap for estimation the sample size in statistical experiments // *Journal of modern applied statistical methods*. — 2013. — Vol. 8. — Pp. 305–321.
21. Rubin D., Stern H. Sample size determination using posterior predictive distributions // *Sankhya: The Indian Journal of Statistics Special Issue on Bayesian Analysis*. — 1998. — Vol. 60. — Pp. 161–175.
22. Self S., Mauritsen R. Power sample size calculations for generalized linear models // *Biometrics*. — 1988. — Vol. 44. — Pp. 79–86.
23. Self S., Mauritsen R., Ohara J. Power calculations for likelihood ratio tests in generalized linear models // *Biometrics*. — 1992. — Vol. 48. — Pp. 31–39.
24. Shieh G. On power and sample size calculations for likelihood ratio tests in generalized linear models // *Biometrics*. — 2000. — Vol. 56. — Pp. 1192–1196.
25. Shieh G. On power and sample size calculations for Wald tests in generalized linear models // *Journal of Statistical Planning and Inference*. — 2005. — Vol. 128. — Pp. 43–59.

26. Wang F., Gelfand A. A Simulation-based Approach to Bayesian Sample Size Determination for Performance under a Given Model and for Separating Models // *Statistical Science*. — 2002. — Vol. 17. — Pp. 193–208.
27. Training compute-optimal large language models / Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch et al. // Proceedings of the 36th International Conference on Neural Information Processing Systems. — NIPS '22. — Red Hook, NY, USA: Curran Associates Inc., 2022. — 15 pp.
28. Scaling Laws for Neural Language Models / Jared Kaplan, Sam McCandlish, Tom Henighan et al. // *CoRR*. — 2020. — Vol. abs/2001.08361. <https://arxiv.org/abs/2001.08361>.
29. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. — 2022. <https://arxiv.org/abs/2112.11446>.
30. Hinton G., Vinyals O., Dean J. Distilling the Knowledge in a Neural Network // NIPS Deep Learning and Representation Learning Workshop. — 2015.
31. Unifying distillation and privileged information / D. Lopez-Paz, L. Bottou, B. Schölkopf, V. Vapnik // ICLR. — 2016.
32. Grabovoy A. V., Strijov V. V. Bayesian Distillation of Deep Learning Models // *Automation and Remote Control*. — 2021. — Vol. 82, no. 11. — Pp. 1846–1856.
33. Воронцов Константин Вячеславович. Комбинаторная теория надежности обучения по прецедентам: Ph.D. thesis / МГУ. — 2010. <http://www.machinelearning.ru/wiki/images/b/b6/Voron10doct.pdf>.
34. McAllester David A. A PAC-Bayesian Tutorial with A Dropout Bound // *ArXiv*. — 2013. — Vol. abs/1307.2118. <https://api.semanticscholar.org/CorpusID:1936248>.
35. Jacot Arthur, Gabriel Franck, Hongler Clement. Neural Tangent Kernel: Convergence and Generalization in Neural Networks // Advances in Neural Information Processing Systems. — Vol. 31. — 2018.
36. Sagun Levent, Evci Utku, Guney V. Ugur et al. Empirical Analysis of the Hessian of Over-Parametrized Neural Networks. — 2018. <https://arxiv.org/abs/1706.04454>.
37. Skorski Maciej. Chain Rules for Hessian and Higher Derivatives Made Easy by Tensor Calculus. — 2019. <https://arxiv.org/abs/1911.13292>.

38. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima / Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal et al. // *ArXiv*. — 2016. — Vol. abs/1609.04836. <https://api.semanticscholar.org/CorpusID:5834589>.
39. Sharp Minima Can Generalize For Deep Nets / Laurent Dinh, Razvan Pascanu, Samy Bengio, Yoshua Bengio // International Conference on Machine Learning. — 2017. <https://api.semanticscholar.org/CorpusID:7636159>.
40. Visualizing the Loss Landscape of Neural Nets / Hao Li, Zheng Xu, Gavin Taylor et al. // Advances in Neural Information Processing Systems / Ed. by S. Bengio, H. Wallach, H. Larochelle et al. — Vol. 31. — Curran Associates, Inc., 2018. https://proceedings.neurips.cc/paper_files/paper/2018/file/a41b3bb3e6b050b6c9067c67f663b915-Paper.pdf.
41. *Papyan Vardan*. The Full Spectrum of Deepnet Hessians at Scale: Dynamics with SGD Training and Sample Size. — 2019. <https://arxiv.org/abs/1811.07062>.
42. *Ghorbani Behrooz, Krishnan Shankar, Xiao Ying*. An Investigation into Neural Net Optimization via Hessian Eigenvalue Density // Proceedings of the 36th International Conference on Machine Learning / Ed. by Kamalika Chaudhuri, Ruslan Salakhutdinov. — Vol. 97 of *Proceedings of Machine Learning Research*. — PMLR, 2019. — 09–15 Jun. — Pp. 2232–2241. <https://proceedings.mlr.press/v97/ghorbani19b.html>.
43. *Singh Sidak Pal, Hofmann Thomas, Schölkopf Bernhard*. The Hessian perspective into the nature of convolutional neural networks // Proceedings of the 40th International Conference on Machine Learning. — ICML’23. — JMLR.org, 2023. — 39 pp.
44. *Ormaniec Weronika, Dangel Felix, Singh Sidak Pal*. What Does It Mean to Be a Transformer? Insights from a Theoretical Hessian Analysis // *arXiv preprint arXiv:2410.10986*. — 2024. — Self-Attention Block decomposition.
45. *Pearlmutter Barak A*. Fast exact multiplication by the Hessian // *Neural computation*. — 1994. — Vol. 6, no. 1. — Pp. 147–160.
46. Pyhessian: Neural networks through the lens of the hessian / Zhewei Yao, Amir Gholami, Kurt Keutzer, Michael W Mahoney // 2020 IEEE international conference on big data (Big data) / IEEE. — 2020. — Pp. 581–590.

47. *Pennington Jeffrey, Worah Pratik*. The Spectrum of the Fisher Information Matrix of a Single-Hidden-Layer Neural Network // Advances in Neural Information Processing Systems / Ed. by S. Bengio, H. Wallach, H. Larochelle et al. — Vol. 31. — Curran Associates, Inc., 2018. https://proceedings.neurips.cc/paper_files/paper/2018/file/18bb68e2b38e4a8ce7cf4f6b2625768c-Paper.pdf.
48. *Pennington Jeffrey, Schoenholz Samuel S., Ganguli Surya*. The Emergence of Spectral Universality in Deep Networks // ArXiv. — 2018. — Vol. abs/1802.09979. <https://api.semanticscholar.org/CorpusID:3569532>.
49. Gradient Descent Finds Global Minima of Deep Neural Networks / Simon Du, Jason Lee, Haochuan Li et al. // Proceedings of the 36th International Conference on Machine Learning / Ed. by Kamalika Chaudhuri, Ruslan Salakhutdinov. — Vol. 97 of *Proceedings of Machine Learning Research*. — PMLR, 2019. — 09–15 Jun. — Pp. 1675–1685. <https://proceedings.mlr.press/v97/du19c.html>.
50. Entropy-SGD: biasing gradient descent into wide valleys / Pratik Chaudhari, Anna Choromańska, Stefano Soatto et al. // *Journal of Statistical Mechanics: Theory and Experiment*. — 2016. — Vol. 2019. <https://api.semanticscholar.org/CorpusID:13807351>.
51. *Fort Stanislav, Jastrzebski Stanislaw*. Large Scale Structure of Neural Network Loss Landscapes // Advances in Neural Information Processing Systems / Ed. by H. Wallach, H. Larochelle, A. Beygelzimer et al. — Vol. 32. — Curran Associates, Inc., 2019. https://proceedings.neurips.cc/paper_files/paper/2019/file/48042b1dae4950fef2bd2aafa0b971a1-Paper.pdf.
52. *Fort Stanislav, Scherlis Adam*. The Goldilocks zone: Towards better understanding of neural network loss landscapes // ArXiv. — 2018. — Vol. abs/1807.02581. <https://api.semanticscholar.org/CorpusID:49655834>.
53. *Deng Li*. The mnist database of handwritten digit images for machine learning research // *IEEE Signal Processing Magazine*. — 2012. — Vol. 29, no. 6. — Pp. 141–142.
54. *Xiao Han, Rasul Kashif, Vollgraf Roland*. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. — 2017. <https://arxiv.org/abs/1708.07747>.
55. *Krizhevsky Alex*. Learning Multiple Layers of Features from Tiny Images // *International Journal of Intelligence Science*. — 2009. — Vol. 13. <https://api.semanticscholar.org/CorpusID:18268744>.

56. PyTorch: An Imperative Style, High-Performance Deep Learning Library / Adam Paszke, Sam Gross, Francisco Massa et al. // Advances in Neural Information Processing Systems 32. — Curran Associates, Inc., 2019. — Pp. 8024–8035. <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-lib.pdf>.
57. *Kingma Diederik, Ba Jimmy.* Adam: A Method for Stochastic Optimization // International Conference on Learning Representations (ICLR). — San Diego, CA, USA: 2015.
58. *Wu Bichen, Xu Chenfeng, Dai Xiaoliang et al.* Visual Transformers: Token-based Image Representation and Processing for Computer Vision. — 2020.
59. Wide neural networks of any depth evolve as linear models under gradient descent* / Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz et al. // *Journal of Statistical Mechanics: Theory and Experiment.* — 2020. — dec. — Vol. 2020, no. 12. — P. 124002. <https://dx.doi.org/10.1088/1742-5468/abc62b>.
60. *Wu Yikai, Zhu Xingyu, Wu Chenwei et al.* Dissecting Hessian: Understanding Common Structure of Hessian in Neural Networks. — 2022.
61. *Singla Sahil, Wallace Eric, Feng Shi, Feizi Soheil.* Understanding Impacts of High-Order Loss Approximations and Features in Deep Learning Interpretation. — 2019.
62. *Kiselev N. S., Grabovoy A. V.* Unraveling the Hessian: A Key to Smooth Convergence in Loss Function Landscapes // *Doklady Mathematics.* — 2024. — Vol. 110, no. S1. — Pp. S49–S61.
63. Geometry of Linear Convolutional Networks / Kathlén Kohn, Thomas Merkh, Guido Montúfar, Matthew Trager // *SIAM Journal on Applied Algebra and Geometry.* — 2022. — Vol. 6, no. 3. — Pp. 368–406. <https://doi.org/10.1137/21M1441183>.
64. *Qin Zhen, Han Xiaodong, Sun Weixuan et al.* Toeplitz Neural Network for Sequence Modeling. — 2023. <https://arxiv.org/abs/2305.04749>.
65. *Gnacik Michal, Lapa Krystian.* Using Toeplitz Matrices to obtain 2D convolution. — 2022. — 10.
66. *Noci Lorenzo, Anagnostidis Sotiris, Biggio Luca et al.* Signal Propagation in Transformers: Theoretical Perspectives and the Role of Rank Collapse. — 2022. <https://arxiv.org/abs/2206.03126>.

67. *Aduenko Alexander*. Selection of multimodels in classification tasks: Ph.D. thesis / MIPT. — 2017. https://www.frccsc.ru/diss-council/00207305/diss/list/aduenko_aa.
68. *Harrison D., Rubinfeld D.* Hedonic housing prices and the demand for clean air // *Journal of environmental economics and management*. — 1978. — Vol. 5, no. 1. — Pp. 81–102.
69. *Quinlan J.* Learning With Continuous Classes // Proceedings of Australian Joint Conference on Artificial Intelligence. — World Scientific, 1992. — Pp. 343–348.
70. *Markelle Kelly, Rachel Longjohn, Kolby Nottingham*. The UCI Machine Learning Repository. <https://archive.ics.uci.edu>.
71. Preconditioned stochastic gradient Langevin dynamics for deep neural networks / C. Li, C. Chen, D. Carlson, L. Carin // AAAI. — 2016. — Pp. 1788–1794.
72. *Bai Zhaojun, Fahey Gark, Golub Gene*. Some large-scale matrix computation problems // *Journal of Computational and Applied Mathematics*. — 1996. — Vol. 74, no. 1-2. — Pp. 71–89.
73. *Chen Xiangning, Hsieh Cho-Jui*. Stabilizing differentiable architecture search via perturbation-based regularization // International conference on machine learning / PMLR. — 2020. — Pp. 1554–1565.
74. Aeberhard S. Wine Dataset. — <https://archive.ics.uci.edu/ml/datasets/Wine>.
75. *Grabovoy A. V., Strijov V. V.* Probabilistic Interpretation of the Distillation Problem // *Automation and Remote Control*. — 2022. — Vol. 83, no. 1. — Pp. 123–137.
76. ImageNet: A large-scale hierarchical image database. / J. Deng, W. Dong, R. Socher et al. // CVPR. — IEEE Computer Society, 2009. — Pp. 248–255.
77. Improving massively multilingual neural machine translation and zero-shot translation / Biao Zhang, Philip Williams, Ivan Titov, Rico Sennrich // *arXiv preprint arXiv:2004.11867*. — 2020.
78. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks / Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A Efros // Computer Vision (ICCV), 2017 IEEE International Conference on. — 2017.
79. Attention is All you Need / A. Vaswani, N. Shazeer, N. Parmar et al. // Advances in Neural Information Processing Systems. — Vol. 30. — Curran Associates, Inc., 2017.

80. No language left behind: Scaling human-centered machine translation / Marta R Costa-jussà, James Cross, Onur Çelebi et al. // *arXiv preprint arXiv:2207.04672*. — 2022.
81. *Glorot Xavier, Bengio Yoshua*. Understanding the difficulty of training deep feedforward neural networks // Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics / Ed. by Yee Whye Teh, Mike Titterington. — Vol. 9 of *Proceedings of Machine Learning Research*. — Chia Laguna Resort, Sardinia, Italy: PMLR, 2010. — 13–15 May. — Pp. 249–256. <https://proceedings.mlr.press/v9/glorot10a.html>.
82. *Chen Tianqi, Goodfellow Ian, Shlens Jonathon*. Net2Net: Accelerating Learning via Knowledge Transfer. — 2015. <https://arxiv.org/abs/1511.05641>.
83. Optuna: A next-generation hyperparameter optimization framework / Takuya Akiba, Shotaro Sano, Toshihiko Yanase et al. // Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. — 2019. — Pp. 2623–2631.
84. *Goodfellow Ian J, Shlens Jonathon, Szegedy Christian*. Explaining and harnessing adversarial examples // *arXiv preprint arXiv:1412.6572*. — 2014.
85. *Xiao Han, Rasul Kashif, Vollgraf Roland*. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. — 2017.
86. Meta-transfer learning for few-shot learning / Qianru Sun, Yaoyao Liu, Tat-Seng Chua, Bernt Schiele // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2019. — Pp. 403–412.
87. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT). — 2019. — Pp. 4171–4186.
88. *Anonymous*. Title Redacted for Double-Blind Review // *Journal Name Redacted*. — 2024.
89. Defending Against Neural Fake News / Rowan Zellers, Ari Holtzman, Hannah Rashkin et al. // Advances in Neural Information Processing Systems (NeurIPS). — 2019. — Pp. 9051–9062.
90. Release strategies and the social impacts of language models / Irene Solaiman, Miles Brundage, Jack Clark et al. // *arXiv preprint arXiv:1908.09203*. — 2019.

91. *Jawahar Ganesh, Abdul-Mageed Muhammad, Lyu Chin-Yew*. Automatic detection of machine generated text: A critical survey // Proceedings of the 28th International Conference on Computational Linguistics (COLING). — 2020. — Pp. 2296–2309.
92. Authorship Attribution for Neural Text Generation / Adaku Uchendu, Thai Le, Kai Shu, Dongwon Lee // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). — 2021. — Pp. 4275–4291.
93. *Valiaiev Dmytro*. Detection of Machine-Generated Text: Literature Survey. — 2024. <https://arxiv.org/abs/2402.01642>.
94. *Anonymous*. Title Redacted for Double-Blind Review // *Journal Name Redacted*. — 2025.
95. *Maurer Andreas*. Bounds for Linear Multi-Task Learning // *J. Mach. Learn. Res.* — 2006. — dec. — Vol. 7. — P. 117–139.
96. *Baxter J.* A Model of Inductive Bias Learning // *Journal of Artificial Intelligence Research*. — 2000. — mar. — Vol. 12. — P. 149–198. <http://dx.doi.org/10.1613/jair.731>.
97. Language Models are Unsupervised Multitask Learners / Alec Radford, Jeffrey Wu, Rewon Child et al. // *OpenAI*. — 2019.
98. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer / Colin Raffel, Noam M. Shazeer, Adam Roberts et al. // *J. Mach. Learn. Res.* — 2019. — Vol. 21. — Pp. 140:1–140:67. <https://api.semanticscholar.org/CorpusID:204838007>.
99. *Bao Guangsheng, Zhao Yanbin, Teng Zhiyang et al.* Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. — 2024. <https://arxiv.org/abs/2310.05130>.
100. *Hans Abhimanyu, Schwarzschild Avi, Cherepanova Valeria et al.* Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text. — 2024.
101. *He Pengcheng, Liu Xiaodong, Gao Jianfeng, Chen Weizhu*. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. — 2021.
102. MULTITuDE: Large-Scale Multilingual Machine-Generated Text Detection Benchmark / Dominik Macko, Robert Moro, Adaku Uchendu et al. // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing / Ed. by Houda Bouamor, Juan Pino, Kalika Bali.

- Singapore: Association for Computational Linguistics, 2023. — dec. — Pp. 9960–9987. <https://aclanthology.org/2023.emnlp-main.616>.
103. Intrinsic dimension estimation for robust detection of AI-generated texts / Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva et al. // Proceedings of the 37th International Conference on Neural Information Processing Systems. — NIPS '23. — Red Hook, NY, USA: Curran Associates Inc., 2023. — 20 pp.
 104. Boundary detection in mixed AI-human texts / Laida Kushnareva, Tatiana Gaintseva, Dmitry Abulkhanov et al. // First Conference on Language Modeling. — 2024. <https://openreview.net/forum?id=kzzwTrt04Z>.
 105. *Sadasivan Vinu Sankar, Kumar Aounon, Balasubramanian Sriram et al.* Can AI-Generated Text be Reliably Detected? — 2024. <https://openreview.net/forum?id=NvSwR4IvL0>.
 106. DetectGPT: zero-shot machine-generated text detection using probability curvature / Eric Mitchell, Yoonho Lee, Alexander Khazatsky et al. // Proceedings of the 40th International Conference on Machine Learning. — ICML'23. — 2023. — 13 pp.
 107. *Miller George A.* WordNet: A Lexical Database for English // Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994. — 1994. <https://aclanthology.org/H94-1111>.
 108. *Wang Yuxia, Shelmanov Artem, Mansurov Jonibek et al.* GenAI Content Detection Task 1: English and Multilingual Machine-Generated Text Detection: AI vs. Human. — 2025. <https://arxiv.org/abs/2501.11012>.
 109. *He Pengcheng, Gao Jianfeng, Chen Weizhu.* DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. — 2023. <https://arxiv.org/abs/2111.09543>.
 110. Open multimodal iEEG-fMRI dataset from naturalistic stimulation with a short audiovisual film / Julia Berezutskaya, Mariska J. Vansteensel, Erik J. Aarnoutse et al. // *Scientific Data*. — 2022. — mar. — Vol. 9, no. 1. <https://doi.org/10.1038/s41597-022-01173-0>.
 111. *Almazrouei Ebtesam, Alobeidli Hamza, Alshamsi Abdulaziz et al.* The Falcon Series of Open Language Models. — 2023. <https://arxiv.org/abs/2311.16867>.
 112. *Black Sid, Gao Leo, Wang Phil et al.* GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. — 2021. — mar. — If you use this

software, please cite it using these metadata. <https://doi.org/10.5281/zenodo.5297715>.

113. Wang Liang, Yang Nan, Huang Xiaolong et al. Multilingual E5 Text Embeddings: A Technical Report. — 2024. <https://arxiv.org/abs/2402.05672>.
114. Gritsay German, Grabovoy Andrey, Chekhovich Yury. Automatic Detection of Machine Generated Texts: Need More Tokens // Ivannikov Memorial Workshop Proceedings 2022. — 2022.
115. Powers David M. W. Applications and explanations of Zipf's law // Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning. — NeMLaP3/CoNLL '98. — USA: Association for Computational Linguistics, 1998. — P. 151–160.
116. Magnus Jan R., Neudecker Heinz. Matrix Differential Calculus with Applications in Statistics and Econometrics. — Chichester: Wiley, 1988.
117. Petersen Kaare Brandt, Pedersen Michael Syskind. The Matrix Cookbook. — <https://www2.imm.dtu.dk/pubdb/doc/imm3274.pdf>. — 2012. — Version November 15, 2012.
118. Варламова К. Д., Дорин Д. Д., Грабовой А. В. За чертой знакомых доменов: исследование обобщающей способности детекторов машинно сгенерированных изображений // Доклады Российской академии наук. Математика, информатика, процессы управления. — 2025. — Т. 527. — С. 103–116.
119. Дорин Д. Д., Грабовой А. В., Стрижов В. В. Улучшение декодирования данных ФМРТ с использованием пространственно-временных характеристик в условиях ограниченного набора данных // Доклады Российской академии наук. Математика, информатика, процессы управления. — 2025. — Т. 527. — С. 11–30.
120. Зверева А. К., Грабовой А. В., Каприелова М. С. Динамическое разделение труда в гибридном ии: стратегии кодировщиков и их воздействие на модуляторы на основе сетей долгой краткосрочной памяти // Доклады Российской академии наук. Математика, информатика, процессы управления. — 2025. — Т. 527. — С. 117–133.
121. Грицай Г. М., Грабовой А. В. Интерпретация классификаторов на основе архитектуры трансформер с помощью кластеризации // Доклады Российской академии наук. Математика, информатика, процессы управления. — 2025. — Т. 527. — С. 432–448.

122. Дорин Д. Д., Варламова К. Д., Грабовой А. В. Попарное сравнение изображений для обнаружения плагиата // Доклады Российской академии наук. Математика, информатика, процессы управления. — 2025. — Т. 527. — С. 68–83.
123. Kiselev N. S., Grabovoy A. V. Sample Size Determination: Likelihood Bootstrapping // Computational Mathematics and Mathematical Physics. — 2025. — Vol. 65, no. 2. — Pp. 416–423.
124. Zvereva Anna K., Kaprielova Mariam, Grabovoy Andrey. AnomLite: Efficient binary and multiclass video anomaly detection // Results in Engineering. — 2025. — Vol. 25. — P. 104162.
125. Kiselev Nikita, Grabovoy Andrey. Sample size determination: posterior distributions proximity // Computational Management Science. — 2025. — Vol. 22, no. 1. — P. 1.
126. Gritsay G. M., Khabutdinov I. A., Grabovoy A. V. Stack More LLM's: Efficient Detection of Machine-Generated Texts via Perplexity Approximation // Doklady Mathematics. — 2024. — Vol. 110, no. S1. — Pp. S203–S211.
127. Forecasting fMRI images from video sequences: linear model analysis / Daniil Dorin, Nikita Kiselev, Andrey Grabovoy, Vadim Strijov // HEALTH INFORMATION SCIENCE AND SYSTEMS. — 2024. — Vol. 12, no. 1. — P. 55.
128. RuGECToR: Rule-Based Neural Network Model for Russian Language Grammatical Error Correction / I. A. Khabutdinov, A. V. Chashchin, A. V. Grabovoy et al. // Programming and Computer Software. — 2024. — Vol. 50, no. 4. — Pp. 315–321.
129. Text Reuse Detection in Handwritten Documents / A. V. Grabovoy, M. S. Kaprielova, A. S. Kildyakov et al. // Doklady Mathematics. — 2024.
130. Artificially Generated Text Fragments Search in Academic Documents / G. M. Gritsay, A. V. Grabovoy, A. S. Kildyakov, Yu V. Chekhovich // Doklady Mathematics. — 2024.
131. Bazarova A. I., Grabovoy A. V., Strijov V. V. Analysis of the Properties of Probabilistic Models in Expert-Augmented Learning Problems // Automation and Remote Control. — 2022. — Vol. 83, no. 10. — Pp. 1527–1537.
132. Grabovoy A. V., Strijov V. V. Prior Distribution Selection for a Mixture of Experts // Computational Mathematics and Mathematical Physics. — 2021. — Vol. 61, no. 7. — Pp. 1140–1152.

133. *Grabovoy A. V., Strijov V. V.* Quasi-Periodic Time Series Clustering for Human Activity Recognition // *Lobachevskii Journal of Mathematics*. — 2020. — Vol. 41, no. 3. — Pp. 333–339.
134. *Грабовой А. Б., Бахтееев О. Ю., Стрижов В. В.* Ordering the set of neural network parameters // *Информатика и ее применение*. — 2020. — Vol. 14, no. 2.
135. *Грабовой А. Б., Бахтееев О. Ю., Стрижов В. В.* Estimation of the relevance of the neural network parameters // *Информатика и ее применение*. — 2019.
136. *Voznyuk A., Gritsai G., Grabovoy A.* Team advacheck at PAN: multitasking does all the magic // Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum. — Vol. 4038 of *CEUR Workshop Proceedings (CEUR-WS.org)*. — CEUR-WS.org: 2025. — Pp. 4007–4014.
137. *Voznyuk Anastasia, Gritsai German, Grabovoy Andrey.* Advacheck at SemEval-2025 Task 3: Combining NER and RAG to Spot Hallucinations in LLM Answers // Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025). — Association for Computational Linguistics Vienna, Austria: 2025. — Pp. 1204–1210.
138. Structure Extractor: Multilingual Extraction of Sections from Scientific Document / Ilia Kopanichuk, Artem Chashchin, Inga Ochneva et al. // 2025 37th Conference of Open Innovations Association (FRUCT). — IEEE: 2025. — Pp. 122–128.
139. Advacheck at GenAI Detection Task 1: AI Detection Powered by Domain-Aware Multi-Tasking / German Gritsai, Anastasia Voznyuk, Ildar Khabutdinov, Andrey Grabovoy // Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect). — International Conference on Computational Linguistics Abu Dhabi, UAE: 2025. — Pp. 236–243.
140. *Asvarov Alidar, Grabovoy Andrey.* Neural Machine Translation System for Lezgian, Russian and Azerbaijani Languages // 2024 Ivannikov Ispras Open Conference (ISPRAS). — IEEE: 2024. — Pp. 1–7.
141. *Poimanov Dmitrii, Mestetsky Leonid, Grabovoy Andrey.* N-Gram Perplexity-Based AI-Generated Text Detection // 2024 Ivannikov Ispras Open Conference (ISPRAS). — IEEE: 2024. — Pp. 1–8.
142. *Meshkov Vladislav, Kiselev Nikita, Grabovoy Andrey.* ConvNets Landscape Convergence: Hessian-Based Analysis of Matricized Networks // 2024 Ivannikov Ispras Open Conference (ISPRAS). — IEEE: 2024. — Pp. 1–10.

143. *Boeva G., Gritsay G., Grabovoy A.* Team ap-team at PAN: LLM Adapters for Various Datasets // Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). — Vol. 3740 of *CEUR Workshop Proceedings*. — CEUR-WS.org: 2024. — Pp. 2527–2535.
144. *Gritsay G., Grabovoy A.* Automated Text Identification on Languages of the Iberian Peninsula: LLM and BERT-based Models Aggregation // Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2024). — Vol. 3756 of *CEUR Workshop Proceedings*. — CEUR-WS.org: 2024.
145. *Gritsai German, Khabutdinov Ildar, Grabovoy Andrey.* Multi-head Span-based Detector for AI-generated Fragments in Scientific Papers // Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024). — Association for Computational Linguistics Bangkok, Thailand: 2024. — Pp. 220–225.
146. *Asvarov Alidar, Grabovoy Andrey.* The Impact of Multilinguality and Tokenization on Statistical Machine Translation // 2024 35th Conference of Open Innovations Association (FRUCT). — IEEE: 2024. — Pp. 149–157.
147. Automated Text Identification: Multilingual Transformer-based Models Approach / G. Gritsay, Andrey Grabovoy, A. Kildyakov, Yury Chekhovich // Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2023). — Vol. 3496 of *CEUR Workshop Proceedings*. — CEUR-WS.org: 2023.
148. Anti-Distillation: Knowledge Transfer from a Simple Model to the Complex One / Kseniia Petrushina, Oleg Bakhteev, Andrey Grabovoy, Vadim Strijov // 2022 Ivannikov Ispras Open Conference (ISPRAS). — IEEE: 2022.
149. Зверева Анна Константиновна, Грабовой Андрей Валерьевич, Каприлова Мариам Семеновна. Структурно-ориентированная синтетическая аугментация как регуляризатор в задаче распознавания пространственно-временных паттернов // Тезисы докладов 22-й всероссийской конференции с международным участием «Математические методы распознавания образов (ММРО-2025)», Муром, 22–26 сентября, 2025 г. — Математические методы распознавания образов. — Москва: ООО "МАКС Пресс", 2025. — С. 18–21.

150. Грабовой Андрей Валерьевич. О сложности моделей и данных в задачах обучения нейросетевых моделей // Тезисы докладов 22-й всероссийской конференции с международным участием «Математические методы распознавания образов (ММРО-2025)», Муром, 22–26 сентября, 2025 г. — Математические методы распознавания образов. — Москва: ООО "МАКС Пресс", 2025. — С. 72–73.
151. Киселев Никита Сергеевич, Мешков Владислав Сергеевич, Грабовой Андрей Валерьевич. Достаточный размер обучающей выборки и его связь со сходимостью поверхности функции потерь // Тезисы докладов 22-й всероссийской конференции с международным участием «Математические методы распознавания образов (ММРО-2025)», Муром, 22–26 сентября, 2025 г. — Математические методы распознавания образов. — Москва: ООО "МАКС Пресс", 2025. — С. 73–76.
152. Грицай Герман Михайлович, Грабовой Андрей Валерьевич. Неявная регуляризация векторного представления текстов в подходе многозадачного обучения // Тезисы докладов 22-й всероссийской конференции с международным участием «Математические методы распознавания образов (ММРО-2025)», Муром, 22–26 сентября, 2025 г. — Математические методы распознавания образов. — Москва: ООО "МАКС Пресс", 2025. — С. 78–79.
153. Дорин Даниил Дмитриевич, Киселев Никита Сергеевич, Грабовой Андрей Валерьевич. Декодирование визуальных стимулов из мультимодальных сигналов мозга // Тезисы докладов 22-й всероссийской конференции с международным участием «Математические методы распознавания образов (ММРО-2025)», Муром, 22–26 сентября, 2025 г. — Математические методы распознавания образов. — Москва: ООО "МАКС Пресс", 2025. — С. 124–126.
154. Применение синтетических данных, полученных с помощью генеративной нейросети, для повышения качества моделей детекции / И. Д. Степанов, А. В. Филатов, Д. Д. Дорин и др. // Труды 67-й Всероссийской научной конференции МФТИ. — Прикладная математика и информатика. — Физматкнига: 2025.
155. Мешков В. С., Киселев Н. С., Грабовой А. В. Сходимость ландшафта сверточных нейронных сетей: анализ матричных сетей на основе гесиана // Труды 67-й Всероссийской научной конференции МФТИ. — Прикладная математика и информатика. — Физматкнига: 2025.

156. Киселев Н. С., Грабовой А. В. Сходимость поверхности функции потерь как признак достаточного размера выборки // Труды 67-й Всероссийской научной конференции МФТИ. — Прикладная математика и информатика. — Физматкнига: 2025.
157. Дорин Д. Д., Грабовой А. В. Улучшение декодирования фМРТ в условиях ограниченной выборки // Труды 67-й Всероссийской научной конференции МФТИ. — Прикладная математика и информатика. — Физматкнига: 2025.
158. Грицай Г. М., Грабовой А. В. Выравнивание представлений в подходе многозадачного обучения для детектирования машинно-сгенерированных текстов // Труды 67-й Всероссийской научной конференции МФТИ. — Прикладная математика и информатика. — Физматкнига: 2025.
159. Киселев Никита, Грабовой Андрей. Определение достаточного размера выборки по апостериорному распределению параметров модели // Труды 66-й Всероссийской научной конференции МФТИ, 1–6 апреля 2024 г. — Прикладная математика и информатика. — Физматкнига: 2024. — С. 141.
160. Дорин Даниил, Киселев Никита, Грабовой Андрей. Пространственно-временные методы анализа временных рядов // Труды 66-й Всероссийской научной конференции МФТИ, 1–6 апреля 2024 г. — Прикладная математика и информатика. — Физматкнига: 2024. — С. 127.
161. Грицай Г., Грабовой А. Многозадачное обучение для распознавания машинно-сгенерированных текстов // Труды 65-й Всероссийской научной конференции МФТИ в честь 115-летия Л.Д.Ландау, 3–8 апреля 2023 г. Прикладная математика и информатика. — Физматкнига Москва: 2023. — С. 117–119.
162. Грабовой Андрей Валерьевич, Хабутдинов Ильдар. Анализ работы BERT-подобных моделей в задачах классификации грамматических ошибок на русском языке // Труды 65-й Всероссийской научной конференции МФТИ в честь 115-летия Л.Д.Ландау, 3–8 апреля 2023 г. Прикладная математика и информатика. — Физматкнига Москва: 2023. — С. 117.
163. Баязитов К. М., Грабовой А. В., Стрижов В. В. Дистилляция моделей глубокого обучения на многодоменных выборках // Интеллектуализация обработки информации: Тезисы докладов 14-й Международной конференции. — Москва: Российская академия наук, 2022. — С. 98–99.

164. Грабовой Андрей Валерьевич, Стрижов Vadim Viktorovich. Байесовское выравнивание структур нейросетевых моделей // Труды 64-й Всероссийской научной конференции МФТИ 22-25 ноября 2021. Прикладная математика и информатика. — МФТИ Москва: 2021. — С. 148–149.
165. Грабовой A. B., Стрижов B. B. Априорное распределение параметров в задачах выбора моделей глубокого обучения // «Математические методы распознавания образов» (ММРО-2021): Тезисы докладов 20-й Всероссийской конференции с международным участием. — Москва: Российская академия наук, 2021. — С. 142–143.
166. Грабовой Андрей Валерьевич, Стрижов Vadim Viktorovich. Вероятностный подход к задаче привилегированного обучения и дистилляции // Труды 63-й Всероссийской научной конференции МФТИ. 23-29 ноября 2020 года. Прикладные математика и информатика. — М.: МФТИ, 2020. — С. 197–198.
167. Грабовой A. B., Стрижов B. B. Задача обучения с экспертом для построение интерпретируемых моделей машинного обучения // Тезисы докладов 13-й Международной конференции "Интеллектуализация обработки информации". — РАН Москва: 2020. — С. 16–17.
168. Выбор моделей и ансамблей / В. В. Стрижов, А. А. Адуенко, О. Ю. Бахтеев и др. // Тезисы докладов 13-й Международной конференции "Интеллектуализация обработки информации". — РАН Москва: 2020. — С. 16–17.
169. Грабовой Андрей Валерьевич, Стрижов Vadim Viktorovich. Анализ априорных распределений в задаче смеси экспертов // Труды 62-й Всероссийской научной конференции МФТИ. — Прикладные математика и информатика. — МФТИ М: 2019.
170. Численные методы оценки оптимального объема выборки для логистической и линейной регрессии / Тамаз Тезикоевич Гадаев, Андрей Валерьевич Грабовой, Анастасия Петровна Мотренко, Вадим Викторович Стрижов // Тезисы докладов 19-й Всероссийской конференции с международным участием. — Математические методы распознавания образов. — Москва: Российская академия наук, 2019. — С. 40–41.
171. Грабовой A. B., Бахтеев O. Ю., Стрижов B. B. Введение отношения порядка на множестве параметров нейронной сети // Тезисы докладов 19-й Всероссийской конференции с международным участием. — Математические методы распознавания образов. — Москва: Российская академия наук, 2019. — С. 38–39.

172. Грабовой Андрей Валерьевич, Бахтеев Олег Юрьевич, Стрижов Вадим Викторович. Прореживание нейросетевых моделей методом Белсли // Труды 61-й всероссийской научной конференции МФТИ. — Прикладная математика и информатика. — МФТИ Москва: 2018. — С. 114–115.
173. Грабовой Андрей Валерьевич, Бахтеев Олег Юрьевич, Стрижов Вадим Викторович. Определение релевантности параметров нейросети методом белсли // Тезисы докладов 12-й Международной конференции «Интеллектуализация обработки информации (ИОИ-2018)», Москва, Россия – Гаэта, Италия. — Интеллектуализация обработки информации. — TORUS PRESS: 2018. — С. 36–37.
174. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization / Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre et al. // Advances in Neural Information Processing Systems / Ed. by Z. Ghahramani, M. Welling, C. Cortes et al. — Vol. 27. — Curran Associates, Inc., 2014. https://proceedings.neurips.cc/paper_files/paper/2014/file/04192426585542c54b96ba14445be996-Paper.pdf.
175. Krizhevsky Alex, Nair Vinod, Hinton Geoffrey. CIFAR-10 (Canadian Institute for Advanced Research). <http://www.cs.toronto.edu/~kriz/cifar.html>.
176. Cortical network responses map onto data-driven features that capture visual semantics of movie fragments / Julia Bereznitskaya, Zachary V. Freudenburg, Luca Ambrogioni et al. // *Scientific Reports*. — 2020. — jul. — Vol. 10, no. 1. <https://doi.org/10.1038/s41598-020-68853-y>.