

На правах рукописи

Грабовой Андрей Валериевич

О СЛОЖНОСТИ МОДЕЛИ 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:
д.ф.-м.н. К. В. Воронцов

Москва — 2025

Оглавление

	Стр.
Введение	5
Глава 1. Сложность моделей и данных	13
1.1. Оценка сложности моделей и данных	14
1.2. Достаточный объем выборки, как мера сложности данных	18
1.3. Сходимость ландшафта оптимизационной задачи, как мера сложности модели	20
1.3.1. Полносвязная нейросетевая модель глубокого обучения	24
1.3.2. Сверточные модели глубокого обучения	27
1.3.3. Трансформер модели глубокого обучения	29
1.4. Результаты вычислительных экспериментов	32
1.4.1. Полносвязная нейросетевая модель глубокого обучения	33
1.4.2. Сверточные модели глубокого обучения	37
1.4.3. Трансформер модели глубокого обучения	38
1.5. Заключение по главе	38
Глава 2. Матрицы Гессе нейросетевых моделей глубокого обучения	45
2.1. Полносвязная нейросетевая модель глубокого обучения	47
2.1.1. Спектральная оценка матрицы Гессе	51
2.2. Матричные модели глубокого обучения	58
2.2.1. Матричная факторизация матрицы Гессе	62
2.2.2. Оценка спектральных норм матрицы Гессе	64
2.3. Матрица Гессе для трансформерной модели глубокого обучения	76
2.3.1. Матрица Гессе для слоя самовнимания	78
2.3.2. Матрица Гессе для LayerNorm слоя	86
2.3.3. Матрица Гессе для нелинейности ReLU	93
2.3.4. Матрица Гессе для трансформера	94

2.3.5. Спектральные оценки матрицы Гессе для трансформера	110
2.4. Результаты вычислительных экспериментов	114
2.5. Заключение по главе	118
 Глава 3. Достаточный объем выборки моделей	120
3.1. Статистические методы определения достаточного размера выборки	122
3.2. Эвристические методы определения достаточного размера выборки .	126
3.3. Байесовские методы определения достаточного размера	128
3.4. Метод определения достаточного размера выборки на основе сэмлирования эмпирической функции ошибки	130
3.5. Метод определения достаточного размера выборки на основе близости апостериорных распределений	134
3.6. Результаты вычислительных экспериментов	144
3.6.1. Определения достаточного размера выборки на основе статистических методов	145
3.6.2. Определение достаточного размера выборки на основе сэмлирования эмпирической функции ошибки	149
3.6.3. Определение достаточного размера выборки на основе близости апостериорных распределений	153
3.7. Заключение по главе	159
 Глава 4. Методы снижения сложности моделей глубокого обучения	161
4.1. Удаления параметров моделей глубокого обучения	161
4.2. Дистилляция моделей глубокого обучения на многодоменных данных	166
4.3. Анти-Дистилляция моделей глубокого обучения	168
4.4. Результаты вычислительных экспериментов	172
4.4.1. Удаления параметров моделей глубокого обучения	172
4.4.2. Дистилляция моделей глубокого обучения на многодоменных данных	177
4.4.3. Анти-Дистилляция моделей глубокого обучения	183
4.5. Заключение по главе	188

Глава 5. Применение теоретических оценок в прикладных задачах	189
Заключение	190
Общие свойства и определения	191
Список иллюстраций	195
Список таблиц	200
Список литературы	202
Приложение А. Дополнительные Леммы и утверждения	218

Введение

Актуальность темы. Ключевым фактором развития нейросетевых моделей с начала XXI века является прогресс в вычислительной технике, характеризующийся экспоненциальным ростом производительности суперкомпьютеров, измеряемой в операциях с плавающей запятой (англ. FLOPS): от значений на уровне терафлопсов в начале столетия до достижения экзафлопсов в настоящий момент. Параллельно наблюдается 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, посвященные анализу сложности таких моделей, в значительной степени опираются на эмпирические корреляции, связывающие их эффективность с количественными метриками — числом параметров и вычислительными затратами на обучение либо применение. Отсутствие же строгой теоретической основы для предсказания поведения моделей при масштабировании делает этот процесс экономически и энергетически нерациональным и зачастую приводит к получению необоснованных и противоречивых результатов.

Развитие больших языковых моделей (англ. LLM) сопряжено с высокими затратами на их обучение, выражющимися в большом потреблении вычислительных, энергетических и финансовых ресурсов. Существенной проблемой в этом процессе является присущая ему непредсказуемость, для нивелирования которой на предварительной стадии обучения (англ. pretrain) эмпирически подбираются оптимальные соотношения между размером модели в параметрах и объемом обучающих данных в токенах при заданной допустимой вычислительном ресурсе в операциях с плавающей запятой. Однако данные, полученные для одной архитектуры, не являются переносимыми на другие модели, что делает подобные оценки зачастую несостоятельными и приводит к непредви-

денным результатам при полномасштабном обучении. В этой связи разработка теоретических оценок сложности моделей представляется критически важной, поскольку она позволила бы получать асимптотические оценки сложности моделей, которые связаны со сложностью выборки еще на этапе проектирования архитектуры, чувствительной к любым модификациям своей структуры.

Оценка сложности моделей машинного обучения является глубоко изученной областью для классических методов и находит ограниченное применение при анализе моделей глубокого обучения. Фундаментальный вклад в теорию оценивания сложности моделей и в теорию машинного обучения в целом внесли работы Владимира Наумовича Вапника и Алексея Яковлевича Червоненкиса, заложившие основы статистической теории обучения [1]. Альтернативный математический аппарат для анализа алгоритмов обучения был разработан Лесли Вэлиантом, предложившим приближенно правильного обучения (англ. PAC-Learning) [2]. Современный подход к определению сложности основан на Радемахеровской сложности, предложенная Владимиром Кольчинским и Дмитрием Панченко [3]. Современные исследования в области теории машинного обучения редко рассматривают современные нейросетевые модели ввиду их сложности и невозможности получения “адекватных” оценок на сложность таких моделей. Значимая часть современных исследований на ведущих конференция посвящена классическим методам машинного анализа и улучшения оценок для известных методов, так как текущие оценки сложности даже для классических моделей машинного обучения являются сильно завышенными.

В работе же проводится теоретический анализ нейросетевых моделей, опирающийся на анализ их матриц Гессе. Матрица Гессе используется как для анализа важности параметров в задачах прореживания нейросетевых моделей, так и для анализа ландшафта функции потерь, который используется для анализа сложности модели.

Степень разработанности темы диссертационного исследования.
Современное состояние анализа сложности моделей машинного обучения опи-

саны в [4]. С другой стороны сложность нейросетевых моделей является слабо изученной темой на текущий момент. В свою очередь рассмотрения сложности выборки сводится только к анализу размера выборки для обучения [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18], что является не совсем корректным, так как сложность выборки также обусловлено и сложностью каждого объекта. Современные методы оценки сложность объектов выборки опираются исследования сложности многообразий, которые аппроксимируют данный элемент выборки. Одним из способов оценки объема выборки для обучения больших языковых моделей являются эмперические оценки вместе с методологией их получения для новых моделей [19, 20] полученные Джаредом Капланом и Джорданом Хоффманом в рамках исследования законов масштабирования нейросетевых моделей [21]. Что касается задачи снижения сложности моделей, на текущий момент существует несколько направлений. Первое направление направлено на квантизацию параметров моделей, второе же, более теоретическое, направлено на дистилляцию и привилегированное обучение [22, 23, 24].

Однако, на текущий момент, в исследованиях не существует теоретического аппарата, для описания сложности моделей и данных, чтобы получить асимптотические оценки сложности моделей и данных. Вместе с тем полученный математический аппарат позволит проводить сравнительный анализ различных нейросетевых архитектур, для выбора лучшего решения для заданной задачи, которая описывается выборкой заданной сложности.

В данной работе проведен комплексный подход к оценке сложности моделей и данных. Предложен альтернативный математический аппарат для анализа сложности на основе анализа ландшафта функции потерь нейросетевых моделей. Получены теоретические оценки для матриц Гессе различных моделей глубокого обучения, которые требуются для оценки сложности модели. Отдельной частью данной работы является анализ связи объема и сложности выборки со сложностью модели. Вводятся частные случаи общей теории сложности, которые позволяют получать практические оценки для прикладных задач.

Объектом исследования в работе являются параметрические семейства функций, которые представимы в виде суперпозиции линейных и нелинейных преобразований; данные, которые используются для настройки параметров параметрических семейств.

Предмет исследования: разработка моделей, методов и теоретического аппарата работы со сложностью моделей глубокого обучения и данными, которые используются для обучения.

Цель и задачи исследования. Целью исследования является получения оценок сложности моделей глубокого обучения, а также получения оценок сложности данных для обучения моделей.

Для достижения цели были поставлены и решены следующие задачи:

- разработка методов оценки сложности моделей глубокого обучения; **Теоретическое введение теории сложности, с понятием условной сложности.**
- вычисление оценок матриц Гессе для различных архитектур моделей глубокого обучения; **Все работы связанные с оценкой матриц Гессе: Киселев/Мешков/Петров.**
- вычисление оценок размера выборки и их связь со сложностью моделей и данных; **Все работы связанные с оценкой размера выборки: Киселев.**
- вычисление оценок важности параметров глубоких нейросетевых моделей при обучении и предсказании; **Работы с прореживанием нейросетевых моделей, дистилляция моделей.** Требуется работа связана с использованием новых оценок матрицы Гессе для вычисления важности параметров.
- оценка влияния регуляризаторов параметров моделей глубокого обучения на оценку сложности моделей глубокого обучения; **Требуются дополнительные, новые исследования: Мешков.**
- практическое применение анализа сложности моделей и данных при обучении моделей глубокого обучения; **Тут все эксперименты, которые есть: Киселев/Мешков.**

- неявная регуляризация сложности модели при мультизадачном обучении. **Все работы связанные с детекцией машингена в мультизадачном подходе: Грицай/Ремизова.**

Методы исследования. Для решения поставленных задач в диссертации используются методы: машинного обучения, статистического анализа данных, статистической теории машинного обучения, линейной алгебры, дискретной математики, теории вероятности, глубокого обучения.

Научная новизна. Научной новизной проведенного исследования являются теоретические оценки сложности моделей глубокого обучения и их связи со сложностью данных для обучения. Диссертация представляет новый подход к рассмотрению сложности моделей глубокого обучения на стыке строго терроризированных оценок и их практической применимостью в реальных задачах. В диссертации предлагается новый подход к оценке сложности модели на основе анализа ландшафта оптимизационной задачи. В рамках диссертационной работы получены оценки матриц Гессе для некоторых моделей глубокого обучения. На основе полученных оценок матриц Гессе в рамках диссертационной работы рассмотрены новые подходы для построения методов оценки важности параметров нейросетевых моделей, которые ранее были слабо исследованы из-за невычислимости данной матрицы для больших нейросетевых моделей.

Теоретическая значимость работы. Диссертационная работа в значительной степени представляет из себя именно теоретический результат. В диссертационной работе рассматриваются фундаментальные вопросы о сложности моделей машинного обучения. Полученные оценки на матрицы Гессе открывают большой спектр задач в теории выбора моделей машинного обучения.

Практическая значимость работы. Предложенные теоретические оценки в работе также рассматривались и из практической стороны. В диссертационной работе предложены адаптации различных полученных оценок, для их применимости в различных прикладных задачах. Отдельно проведены работы с сравнением теоретических оценок и эмпирических оценок закона масштаби-

рования моделей глубокого обучения.

Положения, выносимые на защиту:

1. Оценки сложности моделей глубокого обучения.
2. Оценки сложности данных.
3. Ландшафтная мера сложности моделей глубокого обучения и ее связь со сложностью данных.
4. Оценки матриц Гессе для некоторых классов нейросетевых архитектур и их связь с ландшафтной мерой сложности моделей глубокого обучения.
5. Оценки достаточного размера выборки и их связь со сложностью моделей и данных.
6. Методы снижения размерности пространства параметров моделей глубокого обучения на основе анализа матриц Гессе.

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 ведущих 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 опубликовано 39 работ.

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, соискате-

лю принадлежит определяющая роль в построении теоретических методов и направлении. В работе ...

Структура и объем работы. Диссертация состоит из оглавления, введения, шести разделов, заключения, списка иллюстраций, списка таблиц, перечня основных обозначений и списка литературы из 123 наименований. Основной текст занимает 222 страницы.

Глава 1

Сложность моделей и данных

Современные модели глубокого обучения демонстрируют исключительную эффективность в решении сложных задач, однако их успешное применение требует понимания фундаментального взаимодействия между сложностью модели и характеристиками данных. В главе 2 проводится анализ матриц Гессе для различных архитектур нейронных сетей, что позволяет получить количественные оценки кривизны функции потерь и сложности оптимизационного ландшафта. Эти результаты создают теоретическую основу для формального определения и измерения сложности как моделей, так и данных.

Ключевой идеей настоящей главы является установление 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 получения частных случаев, которые имеют более подробный практический и теоретический анализ.

В рамках данного подхода основным является анализ изменения функции потерь при непрерывном изменении выборки. В разделе 1.3. описывается то как абсолютное изменение функции потерь оценивается через спектральную норму матрицы Гессе:

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

Этот результат, основанный на теоретических выкладках из главы 2, позволяет формализовать понятие *условной сложности выборки* $\mu_D(D_i|f)$ и установить критерии достаточности объема данных для обучения конкретной модели.

Предлагаемый формализм не только углубляет теоретическое понимание процессов обучения глубоких нейронных сетей, но и имеет практическую зна-

чимость для разработки эффективных стратегий обучения, выбора архитектур моделей и планирования экспериментов. Результаты данной 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.

1.1. Оценка сложности моделей и данных

В современной теории глубокого обучения фундаментальной проблемой является установление соответствия между сложностью модели и характеристиками данных. В рамках данного раздела произведем формализацию данного определения.

Определение 1. *Генеральной совокупностью данных Γ назовем произвольное множество объектов, которые исследуются в рамках той либо иной задаче. В общем случае нет никаких ограничений на счетность множества генеральной совокупности.*

Определение 1 позволяет работать как с однородными, так и с многородной генеральными совокупностями.

Определение 2. *Генеральную совокупность Γ назовем однородной, если все объекты генеральной совокупности порождаются из одного распределения. В противном случае выборку назовем k -родной, где k является числом распределений на основе которой была сгенерирована генеральная совокупность;*

В определении 2 примером двуродной генеральной совокупности выступает выборка состоящая из текстов и из изображений в качестве объекта исследования. Например, современные большие языковые модели одновременно работают как с текстами, так и с изображениями и называются многомодальными моделями.

Пусть задана генеральная совокупность данных Γ , где задано множество всех подмножеств объектов образующих кольцо выборок:

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

Определение 3 (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 3 является классическим определением из теории меры, удовлетворяющее свойству конечной-адитивности. Конечной-адитивности нам достаточно, так как в рамках исследований предполагается конечное число объектов при обучении моделей глубокого обучения. Отдельно стоит оговорить, что мы предполагаем, что мы производим сравнение выборок только из одной генеральной совокупности, но заметим, что мы никак не ограничиваем саму генеральную совокупность и в рамках определения возможны мультимодальные генеральные совокупности.

Пусть задано множество параметрических аппроксимирующих моделей

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

где каждое f_i является некоторым множеством параметрических функций. В определении 4 вводиться общее определение характеристики параметрического семейства функций f , которые мы в дальнейшем будем рассматривать в качестве моделей глубокого обучения.

Определение 4. *Мерой сложности модели f назовем отображение $\mu_f(f_i)$:*

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

Заметим, что определение меры сложности модели f не является определением меры в общем случае, так как множество \mathfrak{F} не является кольцом, поэтому данная мера является некоторым отображением, которая является некоторой характеристикой сложности. К примеру число параметров модели удовлетворяет определению 4.

Введя определения меры сложности как для выборки, так и для модели, перейдем к определению обучаемости модели на выборке которое сформулировано в определении 5.

Определение 5. Назовем модель $f \in \mathfrak{F}$ обучаемой на выборке $D \in \mathfrak{D}$, если

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

В рамках определения 5 не вводится никакого ограничения на качество аппроксимации модели после обучения, более подробно это будет определено для частных случаев мер в следующих разделах. Также, определение 5 имеет эмпирическую интерпретацию: сложность модели не должна превышать сложности данных, на которых она обучается, так как в противном случае возникает проблема переобучения, когда модель запоминает шум в данных вместо выявления значимых закономерностей.

В рамках диссертационной работы предполагается исследовать частный случай меры сложности модели на основе оценок ландшафта оптимизационных задач используя матрицы Гессе для различных нейросетевых архитектур, которые описаны в главе 2. Подробнее о частном случае меры сложности описано в разделе 1.3..

Теорема 1. Если для исходной выборки $D \in \mathfrak{D}$ выполняется условие $\mu_f(f) \leq \mu_D(D)$, тогда для новой выборки $D' \in \mathfrak{D}$ модель дообучаема при условии:

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

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

По определению обучаемости модели на выборке D имеем:

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

При добавлении новых данных D' к исходной выборке D сложность объединенной выборки не убывает:

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

Из свойства аддитивности меры сложности выборки, получаем:

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

тогда, объединяя эти три неравенства, получаем цепочку:

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

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

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

□

Это неравенство показывает, что “оставшаяся емкость” модели, а именно, что разность между сложностью модели и сложностью исходных данных не превосходит сложности новых данных D' , что является необходимым условием для успешного дообучения модели на новых данных.

Теорема 1 предполагает, что мера сложности данных обладает свойством монотонности и субаддитивности. В практических приложениях эти свойства должны быть проверены для конкретных выбранных мер сложности.

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

1.2. Достаточный объем выборки, как мера сложности данных

В рамках введенного определения об обучаемости модели на выборке, ключевым понятием становится *условная сложность выборки*:

$$\mu_D(D|f) : \mathfrak{D} \rightarrow \mathbb{R}_+, \quad (1.1)$$

которая характеризует сложность данных $D \in \mathfrak{D}$ относительно заданной параметрической модели f . Эта мера отражает, насколько “трудной” является выборка D для обучения модели f .

Мотивация введения условной сложности выборки проистекает из практического опыта обучения нейронных сетей. Одна и та же выборка данных может представлять различную сложность для разных архитектур моделей.

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

$$\mu_D(D|f) = \inf\{\mu_D(D') : D' \subseteq D, \mu_f(f) \leq \mu_D(D')\}, \quad (1.2)$$

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

Определение 6. Условной сложностью выборки D относительно заданной параметрической модели f назовем отображение (1.1) определяющееся выражением (1.2).

Рассмотрим частный случай меры сложности данных μ_D заданной из определения 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 агрегирована сложность одного объекта выборки. Заметим, что данное предположение является сильным ограничением и может не выполняться на практике, поскольку в реальных задачах различные объекты могут обладать разной сложностью для модели.

Замечание 1. 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 7. Однородную генеральную совокупность Γ_C назовем простой, если она состоит из объектов одинаковой сложности C .

Теорема 2. Для простой генеральной совокупности, мера сложности любой выборки $D \subset \Gamma$ равна ее объему:

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

Доказательство. Докажем, что функция $\mu_D(D) = C \cdot |D|$ удовлетворяет определению меры сложности выборки 3.

Для начала докажем, что функция μ_D является неотрицательной. Поскольку $|D| \geq 0$ для любой выборки $D \subset \Gamma$, и константа $C \in \mathbb{R}_+$, то $\mu_D(D) = C|D| \geq 0$.

Докажем монотонность функции μ_D . Пусть $D_1 \subseteq D_2 \subset \Gamma$. Тогда $|D_1| \leq |D_2|$, следовательно:

$$\mu_D(D_1) = C|D_1| \leq C|D_2| = \mu_D(D_2)$$

Докажем субаддитивность функции μ_D . Для любых непересекающихся выборок $D_1, D_2 \subset \Gamma$ выполняется:

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

Таким образом, функция $\mu_D(D) = C \cdot |D|$ удовлетворяет всем требованиям меры сложности данных в рамках сделанных предположений. \square

Частным случаем *условной сложности* выборки является достаточный объем выборки — минимальный объем данных из выборки D необходимый для обучения модели f .

Определение 8. Размер выборки m^* называется *достаточным согласно критерию T* , если T выполняется для всех $k \geq m^*$.

Таким образом исследования достаточного объема выборки является частным случаем предложенного определения меры сложности данных. Подробный метод оценки достаточного объема выборки рассматривается в главе 3.

1.3. Сходимость ландшафта оптимизационной задачи, как мера сложности модели

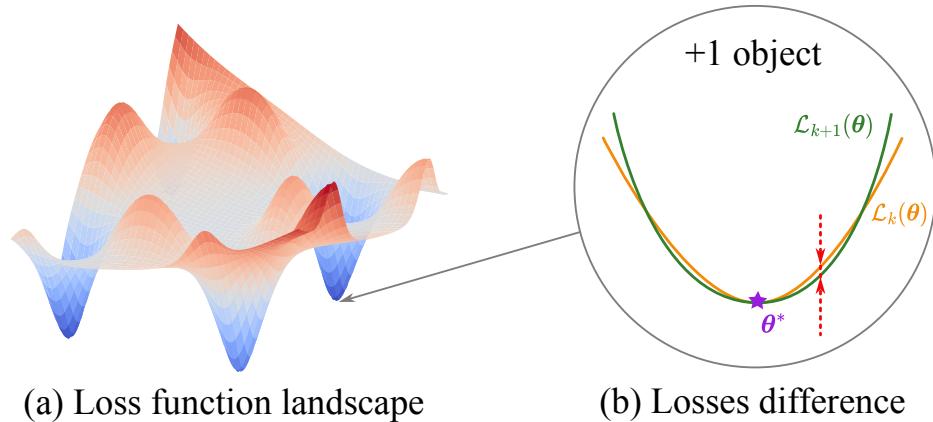


Рис. 1.1: Пример изменения функции потерь при добавлении нового объекта

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 $f_{\boldsymbol{\theta}}$.

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 ошибки либо средняя квадратическая ошибка, в зависимости от рассматриваемой задачи и архитектуры модели.

Заметим, что функция эмпирического риска $\mathcal{L}_m(\boldsymbol{\theta})$ задает некоторую поверхностью в пространстве размерности P . Изменение значения при добавлении одного объекта

$$\begin{aligned}\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta}) &= \frac{1}{k+1} \sum_{i=1}^k \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), \mathbf{y}_i) - \frac{1}{k+1} \sum_{i=1}^k \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), \mathbf{y}_i) = \quad (1.3) \\ &= \frac{1}{k+1} \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_{k+1}), \mathbf{y}_{k+1}) - \sum_{i=1}^k \frac{1}{k(k+1)} \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), \mathbf{y}_i) = \\ &= \frac{1}{k+1} (\ell(f_{\boldsymbol{\theta}}(\mathbf{x}_{k+1}), \mathbf{y}_{k+1}) - \mathcal{L}_k(\boldsymbol{\theta})).\end{aligned}$$

Дальнейшее исследования ландшафта нацелено на исследования данной разницы, причем особенно особенно интересуют предельные свойства при стремлении размера выборки к бесконечности. Для дальнейших оценок данной разности вводится предположение 1, которое в целом подтверждается на практике, но в свою очередь является достаточно сильным, что упрощает дальнейшие выкладки.

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

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

Содержательно, предположение 1 имеет простую эвристическую интерпретацию, что новый объект данных является *репрезентативным* для уже обученной модели — он не приносит “новой информации”, а лишь уточняет ее. В целом при асимптотически большом объеме выборки, данное свойство не противоречит эмпирическим результатам.

Воспользуемся квадратичным приближением Тейлора для упомянутых выше функций потерь в окрестности точки $\boldsymbol{\theta}^*$. Предполагаем, что разложение до

второго порядка будет 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:

$$\mathcal{L}_k(\boldsymbol{\theta}) \approx \mathcal{L}_k(\boldsymbol{\theta}^*) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \mathbf{H}^{(k)}(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*), \quad (1.4)$$

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

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

Следовательно, используя полученное квадратичное приближение (1.4), формула для разности потерь (1.3) принимает вид:

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

причем, используя неравенство треугольника, мы можем вывести следующую оценку:

$$\begin{aligned} |\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})| &\leq \frac{1}{k+1} \left| \ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_{k+1}), \mathbf{y}_{k+1}) - \frac{1}{k} \sum_{i=1}^k \ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_i), \mathbf{y}_i) \right| + \quad (1.5) \\ &\quad + \frac{1}{k+1} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 \left\| \mathbf{H}_{k+1}(\boldsymbol{\theta}^*) - \frac{1}{k} \sum_{i=1}^k \mathbf{H}_i(\boldsymbol{\theta}^*) \right\|_2. \end{aligned}$$

Заметим, что первое слагаемое может быть легко ограничено константой, поскольку сама функция потерь принимает ограниченные значения. Однако выражение с гессианами не так просто оценить. Подробный анализ матриц Гессе для различных типов параметрических моделей глубокого обучения представлен в главе 2. Таким образом, мы анализируем локальную сходимость ландшафта функции, ее матрицу Гессе.

Получаем выражение для анализа, описывающее поведение ландшафта функции потерь:

$$|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})| \leq \frac{M_l}{k+1} + \frac{1}{k+1} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 \left\| \mathbf{H}_{k+1}(\boldsymbol{\theta}^*) - \frac{1}{k} \sum_{i=1}^k \mathbf{H}_i(\boldsymbol{\theta}^*) \right\|_2. \quad (1.6)$$

Таким образом получили, что анализ сходимости ландшафта оптимизационной задачи сводится к анализу нормы матрицы Гессе, которые подробно разобраны в главе 2.

Оценка 1.6 задает некоторое свойство параметрического семейства функций f на заданной выборке D . Определим данное свойство как условную сложность модели f на выборке D :

$$\mu_f(f|D) : \mathfrak{F} \rightarrow \mathbb{R}_+, \quad (1.7)$$

причем, более подробно рассмотрим частный случай условной меры сложности параметрического семейства функций f вида:

$$\mu_f(f|D) = \mathsf{E}_{\mathbf{x}_i \in D} \|\mathbf{H}_i(\boldsymbol{\theta}^*) - \mathsf{E}_{\mathbf{x}_i \in D} \mathbf{H}_i(\boldsymbol{\theta}^*)\|_2. \quad (1.8)$$

Определение 9. Условной сложностью параметрической модели f относительно заданной выборки D назовем выражение (1.7).

Определение 10. Ландшафтная мерой сложности параметрической функции f назовем условную сложность параметрической модели f заданной выражением (1.8).

Определение 9 описывает прикладный способ задания сложности на параметрических семействах функций в контексте оптимизации на заданных выборках. Причем, условная сложность модели $\mu_f(f|D)$ характеризует сложность архитектуры модели f при ее обучении на выборке данных D . Это позволяет количественно оценить, насколько модель “соответствует” данным. Так слишком простая модель может недообучаться, а слишком сложная — переобучаться.

Ландшафтная же мера сложности 10 представляет собой явный вид условной сложности, основанную на анализе оптимизационного ландшафта функции

потерь. Заметим, что выражение (1.8) содержательно указывает на то, насколько сильно добавление нового объекта данных изменяет кривизну функции потерь в окрестности оптимума.

Дальнейшее повествование в главе посвящено к оценкам ландшафтной меры для различных параметрических моделей f . Все результаты основываются на анализе матриц Гессе описанных в главе 2.

Используя выражение 1.6 и определение ландшафтной меры сложности получаем следующую асимптотическую связь между этими оценками:

$$|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})| \leq \frac{M_l}{k+1} + \frac{\mu_f(f|D)}{k+1} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2. \quad (1.9)$$

Лемма 3. Пусть задан некоторый ε , описывающей допустимое изменение ландшафта при добавления одного объекта в некоторой окрестности оптимума $\boldsymbol{\theta}^*$ радиуса R , причем выборка D из 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. Очевидно из определения условных мер сложности моделей и данных и подстановке всей выборки D в выражение (1.9). \square

1.3.1. Полносвязная нейросетевая модель глубокого обучения

Используя результаты, полученные в рамках главы 2, а именно результаты теоремы 9 в которой показана асимптотика нормы матрицы Гессе от гиперпараметров полносвязной нейросетевой модели:

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

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

Теорема 4. 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 8 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, что 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. Используя неравенство треугольника для выражения 1.5, получаем

$$\begin{aligned} & \left| \ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_{k+1}), \mathbf{y}_{k+1}) - \frac{1}{k} \sum_{i=1}^k \ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_i), \mathbf{y}_i) \right| \leq \\ & \leq |\ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_{k+1}), \mathbf{y}_{k+1})| + \left| \frac{1}{k} \sum_{i=1}^k \ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_i), \mathbf{y}_i) \right| \leq \\ & \leq |\ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_{k+1}), \mathbf{y}_{k+1})| + \frac{1}{k} \sum_{i=1}^k |\ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_i), \mathbf{y}_i)| \leq \\ & \leq M_\ell + \frac{1}{k} \sum_{i=1}^k M_\ell = 2M_\ell = \mathcal{O}(1) \text{ при } k \rightarrow \infty. \end{aligned}$$

Аналогично для норм матриц Гессе получаем:

$$\begin{aligned}
& \left\| \mathbf{H}_{k+1}(\boldsymbol{\theta}^*) - \frac{1}{k} \sum_{i=1}^k \mathbf{H}_i(\boldsymbol{\theta}^*) \right\|_2 \leqslant \\
& \leqslant \|\mathbf{H}_{k+1}(\boldsymbol{\theta}^*)\| + \left\| \frac{1}{k} \sum_{i=1}^k \mathbf{H}_i(\boldsymbol{\theta}^*) \right\|_2 \leqslant \\
& \leqslant \|\mathbf{H}_{k+1}(\boldsymbol{\theta}^*)\| + \frac{1}{k} \sum_{i=1}^k \|\mathbf{H}_i(\boldsymbol{\theta}^*)\|_2 \leqslant \\
& \leqslant M_{\mathbf{H}} + \frac{1}{k} \sum_{i=1}^k M_{\mathbf{H}} = 2M_{\mathbf{H}} = \mathcal{O}(1) \text{ при } k \rightarrow \infty,
\end{aligned}$$

где из Теоремы 8 получаем

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

Таким образом, подставляя полученные оценки в выражение для разности, получаем

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

где выбирая окрестность локального минимума $\boldsymbol{\theta}^*$, т.е. $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 \leqslant R^2$, получаем

$$|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})| \leqslant \frac{2}{k+1} (M_{\ell} + M_{\mathbf{H}}R^2) \rightarrow 0 \text{ при } k \rightarrow \infty.$$

□

Используя доказательство теоремы 4 легко получается выражение для ландшафтной меры полносвязной нейросетевой модели глубокого обучения, которое описано в следствии 1.

Следствие 1. *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 зависящая ограничивающая параметры и данные.

В работе [25] показано, что существуют функции, которые могут быть эффективно представлены трехслойными сетями, но требуют экспоненциального числа нейронов в двухслойных сетях. Следствие указывает 1, указывает на схожий результат, описывающий то, что экспоненциальный рост сложность модели увеличивается экспоненциально при увеличении числа слоев, и следовательно, уменьшив число слоем, потребуется экспоненциальный рост параметров модели, чтобы сохранить заданную сложность модели.

1.3.2. Сверточные модели глубокого обучения

В данном подразделе рассматривается оценка ландшафтной меры сложности для сверточных нейронных сетей, которые являются одними из популярных на текущий момент моделей глубокого обучения. В анализе ландшафта используются результаты о матрицах Гессе из главы 2.

Начнем с анализа 1D-сверточных сетей, которые применяются в обработке последовательностей, временных рядов и сигналов. Основные результаты, важные для определения ландшафтной меры сложности, представлены в теореме 5.

Теорема 5. Пусть параметры $\boldsymbol{\theta}$ находятся в R -окрестности оптимума:

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

а функция потерь ограничена некоторой константой:

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

Пусть все объекты в наборе данных ограничены:

$$\exists W_x \forall i \|x_i\| \leq W_x.$$

Тогда в условиях теоремы 12:

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

Доказательство. Доказательство эквивалентно доказательству теореме 4 подставляя оценки нормы матрицы Гессе из теоремы 12. \square

Используя теорему 5 легко получается выражение для ландшафтной меры 1D-сверточной нейросетевой модели глубокого обучения, которое описано в следствии 2.

Следствие 2. *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 зависящая ограничивающая параметры и данные.

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.

Перейдем к анализу 2D-сверточных сетей, которые применяются в задачах обработки изображений. Основные результаты, важные для определения ландшафтной меры сложности, представлены в теореме 6.

Теорема 6. *Пусть параметры θ находятся в R -окрестности оптимума:*

$$\|\theta - \theta^*\| \leq R.$$

Также функция потерь ограничена некоторой константой:

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

Пусть все объекты в наборе данных также ограничены:

$$\exists W_x \forall i \|x_i\| \leq W_x.$$

Тогда при выполнении условий теоремы 13 справедливо:

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

где $q^2 = C^2 k^2 m n$.

Доказательство. Доказательство эквивалентно доказательству теореме 4 подставляя оценки нормы матрицы Гессе из теоремы 13. \square

Используя теорему 6 легко получается выражение для ландшафтной меры 2D-сверточной нейросетевой модели глубокого обучения, которое описано в следствии 3.

Следствие 3. *Ландшафтная мера сложности параметрической функции f 2D-сверточной модели глубокого обучения имеет асимптотику:*

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

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

Для 2D-сверточных сетей наблюдается более быстрый рост сложности по сравнению с 1D-архитектурами, что обусловлено двумерной природой данных.

1.3.3. Трансформер модели глубокого обучения

Архитектура трансформеров представляет собой один из наиболее значительных прорывов в области глубокого обучения последних лет. Эти 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.

Теорема 7. Для одного блока самовнимания и одного блока трансформера 2.10 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 17, а для блока трансформера $M = M_{tr}$ вычисляется в соответствии с Теоремой 26.

Доказательство. Доказательство проводится в несколько этапов. На первом этапе оценивается разность значений функции потерь в оптимальной точке, затем на втором этапе оценивается разность гессианов. После чего используя результаты обоих этапов, проводиться объединение обоих 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 метриц 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:

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

Объединяя полученные оценки, получаем:

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

Теперь оценим вклад второго слагаемого в общую разность функций потерь:

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

Комбинируя оценки для обоих слагаемых, получаем итоговую оценку:

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

□

1.4. Результаты вычислительных экспериментов

В данном разделе описываются результаты вычислительных экспериментов для методов, описанных в данной главе.

На Рис. 1.2 представлены соответствующие результаты, показывающие, что хотя Предположение 1 может быть ослаблено, его выполнимость улучшается с увеличением длины последовательностей.

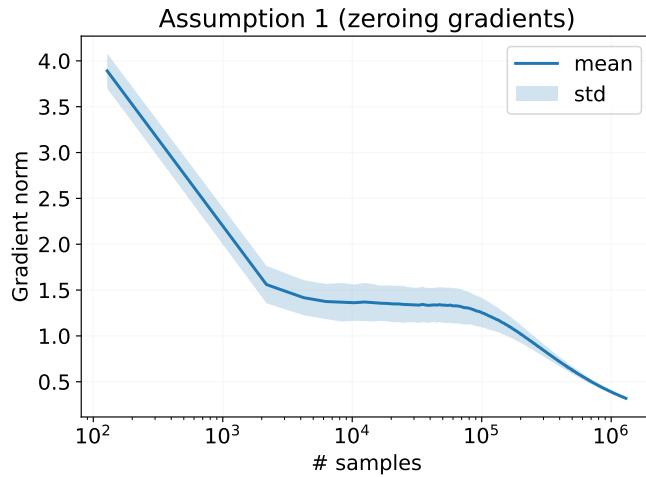


Рис. 1.2: Проверка Предположения 1 о сходимости локальных минимумов при увеличении объема выборки.

1.4.1. Полносвязная нейросетевая модель глубокого обучения

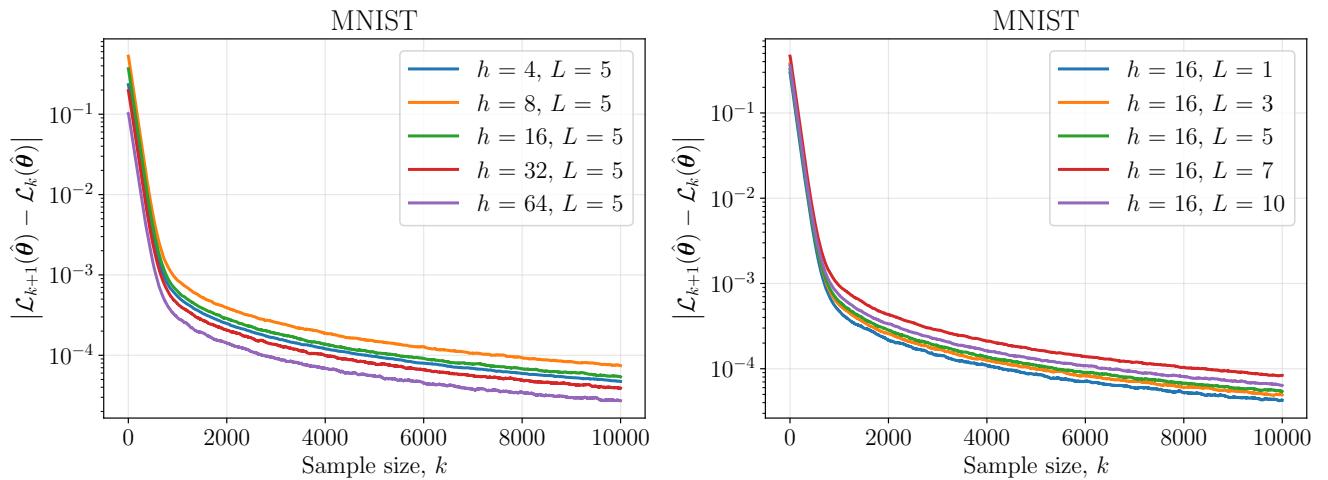


Рис. 1.3: Зависимость абсолютного значения разности функций потерь от доступного размера выборки, прямая классификация изображений. Графики слева показывают уменьшение значений с увеличением размерности скрытого слоя. Графики справа показывают увеличение значений с увеличением количества слоев.

Для проверки полученных теоретических оценок мы провели детальное эмпирическое исследование. В данном разделе представлены результаты обучения полносвязной нейронной сети для задачи классификации изображений. Основ-

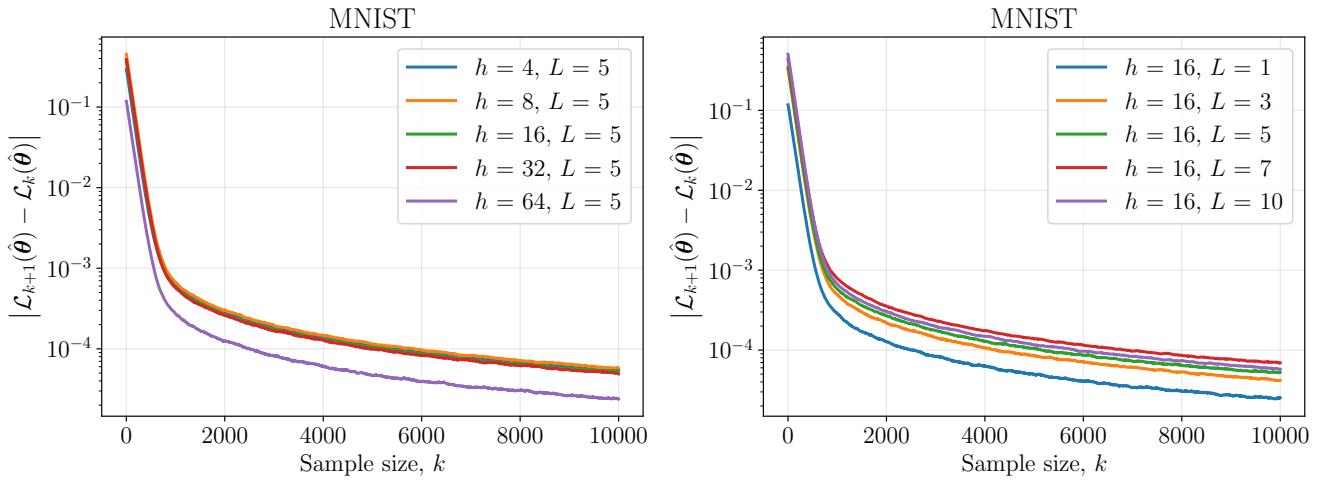


Рис. 1.4: Зависимость абсолютного значения разности функций потерь от доступного размера выборки, **извлечение признаков изображений**. Графики слева показывают уменьшение значений с увеличением размерности скрытого слоя. Графики справа показывают увеличение значений с увеличением количества слоев.

Таблица 1.1: Описание наборов данных для классификации изображений

Название	Описание	Формат	Разрешение
MNIST [26]	Рукописные цифры	Оттенки серого	28×28
FashionMNIST [27]	Элементы одежды	Оттенки серого	28×28
CIFAR10 [28]	Различные объекты	RGB	32×32
CIFAR100 [28]	Различные объекты	RGB	32×32

ной целью этих экспериментов является эмпирическое подтверждение сходимости ландшафта функции потерь с увеличением размера выборки. Для достижения этой цели мы обучили полносвязную нейронную сеть на полном наборе данных и получили соответствующие параметры $\hat{\theta}$ как точку вблизи минимума. Впоследствии мы исследовали зависимость между средней разностью потерь и доступным размером выборки.

Мы использовали библиотеку `pytorch` [29] в качестве Python-фреймворка

для обучения нейронных сетей. Архитектура сети была единообразной и состояла из нескольких линейных слоев с функцией активации ReLU после каждого слоя, за исключением последнего. Размер h был фиксированным для всех скрытых слоев L . Сеть обучалась в течение многочисленных эпох с использованием оптимизатора Adam [30] с постоянной скоростью обучения 10^{-3} . Использовались различные наборы данных для классификации изображений, доступные в библиотеке `torchvision`. Для процесса обучения был выбран размер батча (англ. batch size) 64. Эксперименты проводились на GPU Tesla A100 80GB с 16 CPU ядрами и 243 GB оперативной памяти.

В этом эксперименте мы использовали значения пикселей изображений в качестве входных данных. На Рис. 1.3 представлены результаты, полученные при анализе 10 000 объектов из набора данных MNIST [26], 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. В то же время, график справа иллюстрирует поведение разности потерь при изменении количества скрытых слоев от 1 до 10, при фиксированном размере скрытого слоя $h = 16$. Эта последовательность была повторена 100 раз для усреднения. К полученным результатам было применено экспоненциальное скользящее среднее с коэффициентом сглаживания 0,99. Из наблюдаемых зависимостей видно, 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. Кроме того, на практике константа M , ограничивающая величину весов, оказывается относительно небольшой. Более того, поскольку задача классификации набора данных MNIST считается относительно простой, неглубокая, но широкая нейронная сеть может давать хорошие результаты классификации. Следовательно, наблюдалось, что значения функции потерь ниже для больших значений h , и поэтому их разница также меньше.

В отличие от предыдущего эксперимента, в этой части используется предварительно обученный экстрактор признаков изображений. Полносвязная сеть используется в качестве многоклассового классификатора. Мы выбрали Vision Transformer (ViT) [31] от Google. Мы аналогичным образом выбрали случайным образом 10 000 объектов из набора данных MNIST и варьировали размер скрытого слоя и количество слоев. Результаты согласуются с наблюдениями при прямой классификации изображений. Эта согласованность подтверждает, что представленная сходимость не зависит от природы пространства исходных объектов \mathcal{X} . Для наблюдения сходимости ландшафта функции потерь достаточно ограниченности этого пространства.

Наш эксперимент подтверждает сходимость, доказанную в Теореме 4. Кроме того, верхняя оценка скорости этой сходимости остается верной. Действительно, изменение параметров нейронной сети, таких как количество слоев и размер слоя, приводит к незначительному изменению разности функций потерь.

Приведем расширенную версию проведенных экспериментов. Ниже в Таблице 1.1 представлено описание используемых наборов данных. Мы выбрали четыре набора данных из библиотеки `torchvision`: MNIST [26], FashionMNIST [27], CIFAR10 и CIFAR100 [28]. Единственной предобработкой данных являлась нормализация для приведения значений к диапазону $[-1; 1]$.

На Рис. 1.5 графики слева были получены при фиксированном количестве слоев $L = 5$ в сети. Мы изменяли размер скрытого слоя на всех уровнях от 4 до 64. В то же время, график справа демонстрирует поведение разности потерь при изменении количества скрытых слоев от 1 до 10, при сохранении размера $h = 16$ неизменным. Данная последовательность была повторена 100 раз для усреднения. Для полученных результатов мы применили экспоненциальное скользящее среднее с коэффициентом сглаживания 0.99.

Аналогично Рис. 1.4 на Рис. 1.6 результаты подтверждают сходимость, доказанную в Теореме 4. Также верна верхняя оценка скорости этой сходимости. В частности, изменение параметров нейронной сети: количества слоев и размера

слоя приводит к изменению разности функций потерь.

1.4.2. Сверточные модели глубокого обучения

Для проверки теоретических оценок мы провели комплексное эмпирическое исследование. В данном 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, позиции пулинга, и наблюдение того, как скорость сходимости зависит от этих параметров. Для достижения этой цели мы обучили сверточные сети и получили параметры $\hat{\theta}$ вблизи оптимума. Мы использовали сверточную архитектуру с функцией активации ReLU после каждого слоя. Чтобы проследить влияние конкретного параметра на сходимость, мы фиксировали ключевые параметры нейронной сети, варьировали интересующий гиперпараметр и обучали соответствующий набор моделей. Затем мы исследовали зависимость между средним абсолютным различием между значениями средней функции потерь и доступным размером выборки. Далее, для каждой модели, чтобы получить более надежные результаты, мы усредняли разность потерь по перемешанным выборкам. Дополнительно, для улучшения визуализации, мы использовали экспоненциальное сглаживание с коэффициентом 0.995. Для данного исследования мы использовали числовое представление пикселей изображений в качестве входных данных. Результаты получены на основе анализа выборок из баз данных MNIST [26], FashionMNIST [27] и CIFAR10 [28]. Во всех экспериментах использовались следующие гиперпараметры: постоянная скорость обучения 1e-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.

1.4.3. Трансформер модели глубокого обучения

Для более глубокого изучения зависимости между функцией потерь и ее гессианом мы проводим эксперимент, соответствующий Теореме 7. Здесь мы используем другую конфигурацию модели на наборе данных CIFAR-100 [28]. По сравнению с аналогичной моделью для набора данных MNIST, эта модель имеет в 8 раз больше блоков Трансформера, а также скрытые слои в 8 раз шире. Во время обучения модель также обучалась в течение ряда эпох для достижения точности $>50\%$ на валидационном наборе данных. Результаты представлены на Рис. 1.11. Настройка эксперимента следующая:

1. Обучаем модель до сходимости и сохраняем вектор параметры \mathbf{w}^* .
2. Начинаем с пустого набора данных, добавлять данные батч за батчем (англ. batch) и вычисляем среднее значение потерь по просмотренным батчам;
3. Вычисляем абсолютную разность в соответствии с выражением:

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

1.5. Заключение по главе

В данной главе проведено комплексное исследование сходимости ландшафта функции потерь при увеличении объема выборки для различных архитектур нейронных сетей.

Теоретический анализ и эмпирические результаты для полно связных нейронных сетей демонстрируют, что абсолютная разность между средними значениями функции потерь при добавлении нового объекта в выборку стремится к нулю при увеличении количества доступных объектов до бесконечности. Это достигнуто за счет доказательства теоремы о верхней оценке нормы гессиана. Полученные результаты подтверждают сходимость поверхности функции потерь для задачи классификации изображений как при прямом использовании

исходных представлений, так и при работе с предобученным экстрактором признаков.

Для сверточных нейронных сетей установлено, что абсолютная разность между средними значениями функции потерь демонстрирует монотонную зависимость от размера слоя и позиции пулинга, в то время как зависимость от размера ядра и количества слоев имеет немонотонный характер. Это указывает на преобладающее влияние значения функции потерь в точке оптимума. Предложенный метод оценки нормы гессиана и его использования для анализа сходимости ландшафта потерь предоставляет вклад в теорию анализа локальной геометрии ландшафтов функции потерь.

В случае трансформерных моделей работа восполняет ключевой пробел в анализе путем явного вывода якобианов и гессианов для LayerNorm и FFN. Теоретические результаты показывают гетерогенность гессиана по блокам, причем наибольший вклад вносят компоненты, связанные с Values и Keys. Установленное неравенство сходимости $|\mathcal{L}_{k+1}(\mathbf{w}) - \mathcal{L}_k(\mathbf{w})| \leq 2L/(k+1) + M\|\mathbf{w} - \mathbf{w}^*\|_2^2/(k+1)$ объясняет наблюдаемую стабилизацию ландшафта потерь с ростом объема данных.

Основные ограничения исследования включают детерминистический характер анализа, предположение о существовании единой точки минимума для последовательных размеров выборки, а также возможность улучшения верхних оценок за счет учета специфики разреженных матриц. Полученные результаты открывают перспективы для разработки методов определения достаточного размера выборки и алгоритмов обучения.

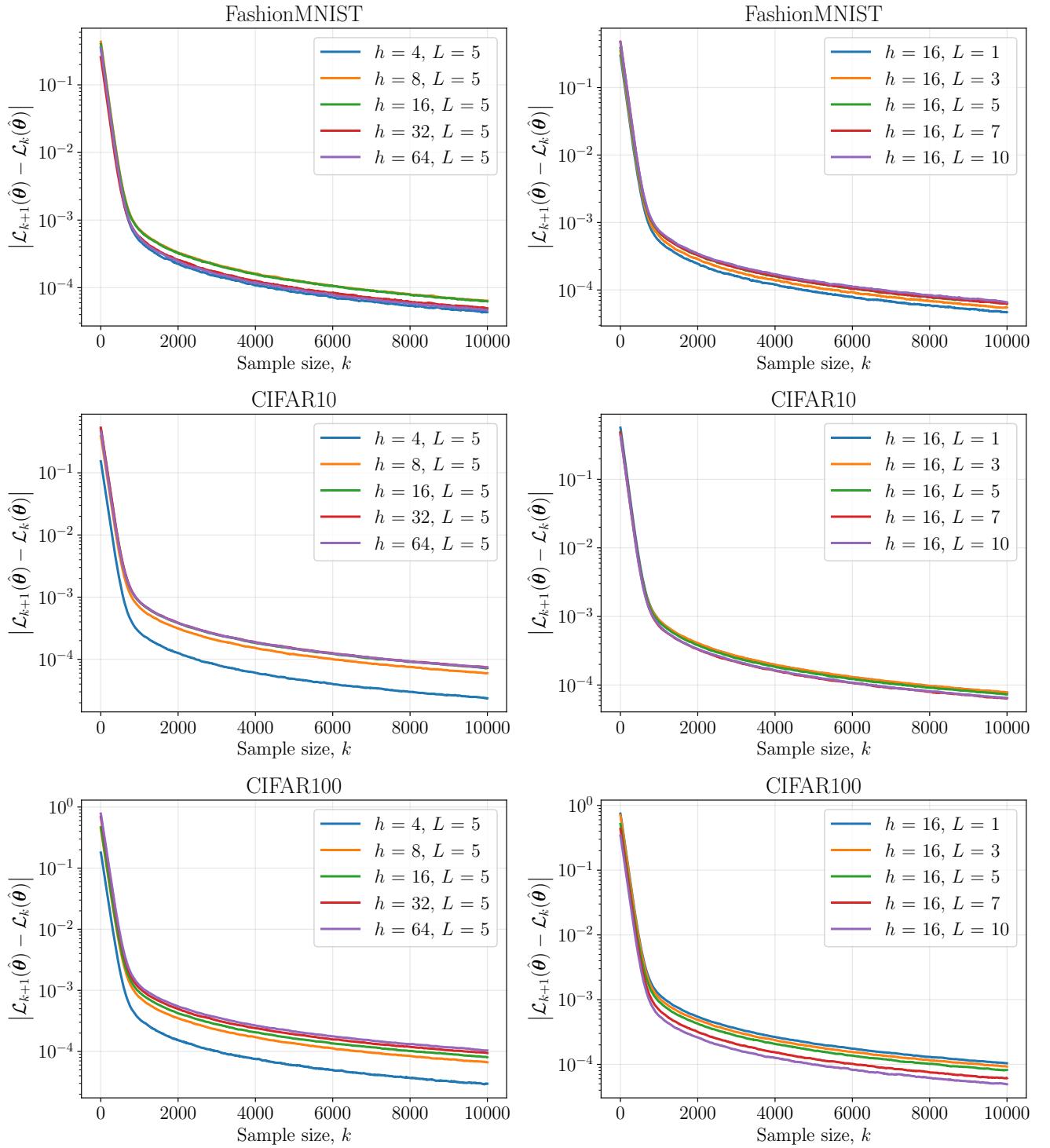


Рис. 1.5: Зависимость абсолютного значения разности функций потерь от доступного размера выборки, прямая классификация изображений. Графики слева показывают уменьшение значений с увеличением размерности скрытого слоя. Графики справа показывают увеличение значений с увеличением количества слоев. Результаты для различных наборов данных: FashionMNIST, CIFAR10 и CIFAR100.

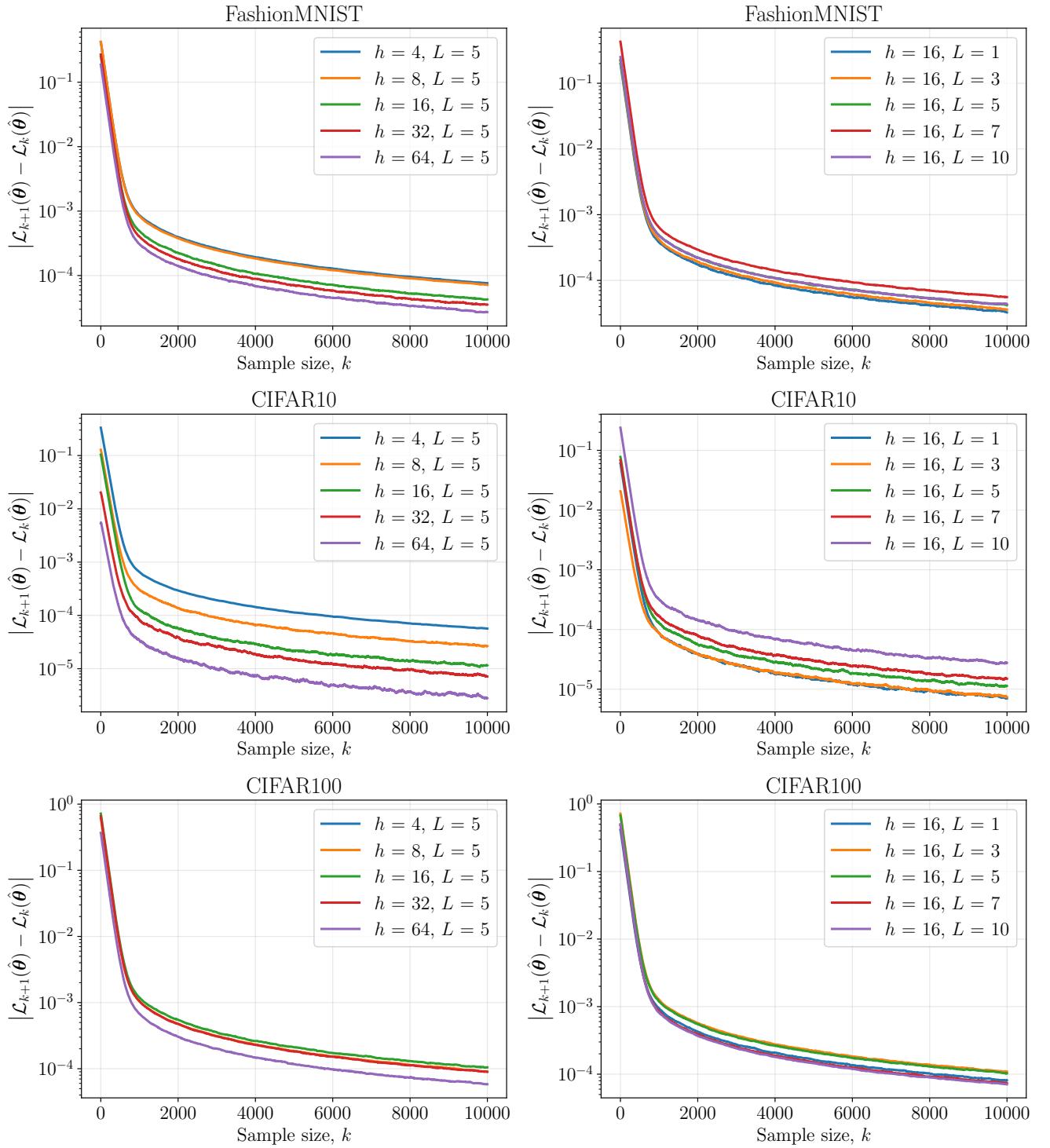


Рис. 1.6: Зависимость абсолютного значения разности функций потерь от доступного размера выборки, извлечение признаков изображений. Графики слева показывают уменьшение значений с увеличением размерности скрытого слоя. Графики справа показывают увеличение значений с увеличением количества слоев. Результаты для различных наборов данных: FashionMNIST, CIFAR10 и CIFAR100.

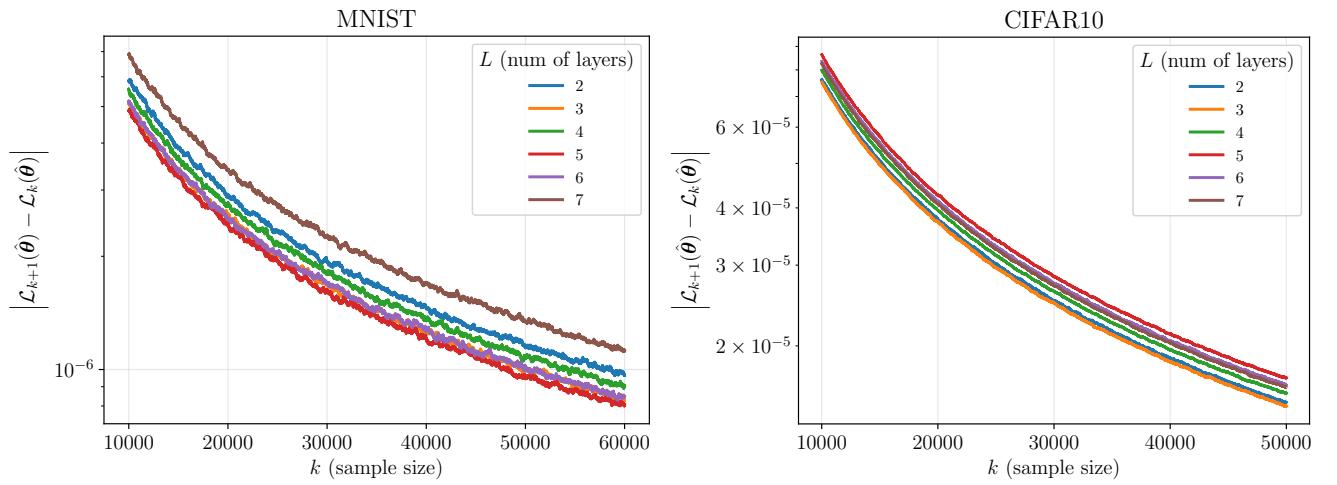


Рис. 1.7: Переменное количество скрытых сверточных слоев L с фиксированным размером ядра $k = 3$ и количеством каналов $C = 6$. Анализ полученных графиков показывает немонотонный характер зависимости выходных значений от количества слоев.

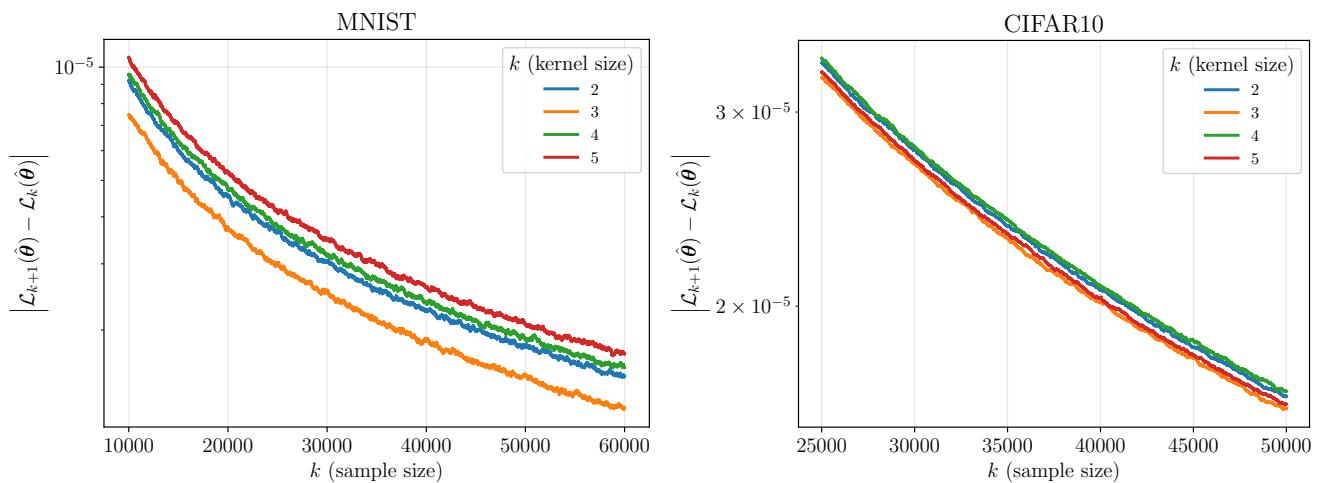


Рис. 1.8: Изменение размера ядра k при фиксированном количестве сверточных слоев L и количестве каналов $C = 6$. Данные демонстрируют немонотонный характер зависимости относительно размера ядра.

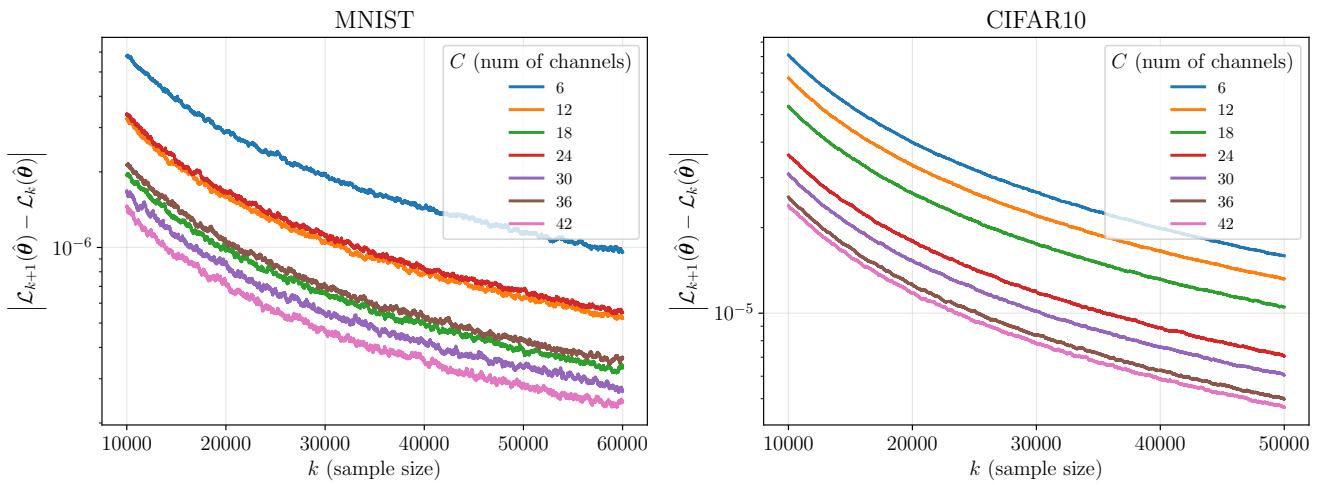


Рис. 1.9: Изменение количества каналов C при фиксированном количестве сверточных слоев L и размере ядра $k = 3$. Зависимость значения от количества каналов имеет монотонный характер.

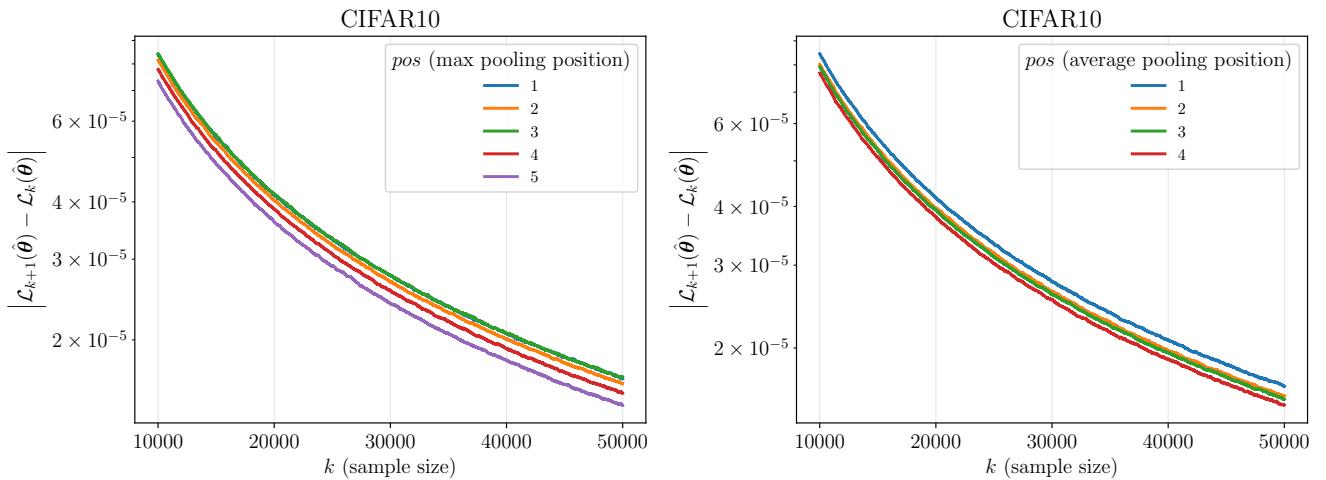


Рис. 1.10: Изменение количества каналов C при фиксированном количестве сверточных слоев L и размере ядра $k = 3$. График показывает монотонную зависимость значения от позиции пулинга в сети.

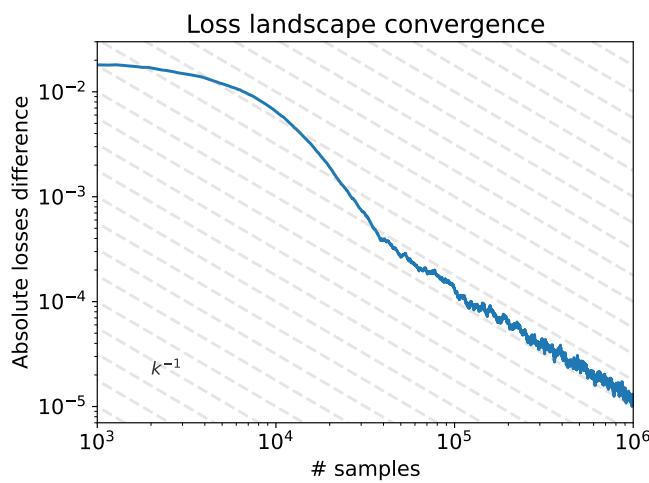


Рис. 1.11: Абсолютная разность потерь в зависимости от количества обучающих примеров в наборе данных, отображенная в двойном логарифмическом масштабе. Синяя линия представляет экспоненциальное скользящее среднее (EMA) желаемой зависимости, а серая линия соответствует линейному тренду.

Глава 2

Матрицы Гессе нейросетевых моделей глубокого обучения

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 не выпуклые 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, описывающая 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.

Применения матрицы Гессе в контексте глубокого обучения проявляется в решении разнообразных теоретических и практических задач. В аспекте оп-

тимизации, на основе матрицы Гессе строятся методы второго порядка, такие как метод Ньютона, который использует обратную матрицу Гессе \mathbf{H}^{-1} для вычисления адаптивных 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. Это описывает “остроту” минимума, которая 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, способность к интерполяции данных и обобщающая способность на тестовых выборках.

Таким образом, матрица Гессе является не только эффективным инструментом для ускорения процесса оптимизации, но и является аналитическим аппаратом для анализа внутренних свойств модели: от устойчивости найденного решения до прогнозирования его способности к обобщению 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 ее ключевых спектральных характеристик, чему и посвящена настоящая глава. 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.

2.1. 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 полно связная 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 удобство для анализа в теоретическом анализе нейросетевых архитектур. Для функции активации ReLU, определяемой как $\sigma(\mathbf{x}) = [\mathbf{x} \geq 0] \mathbf{x}$, 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 [32], применение цепного правила для матриц второго порядка [33] позволяет декомпозировать матрицу Гессе на сумму двух структурно различных компонент:

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

где $\nabla_{\boldsymbol{\theta}} \mathbf{z}_i \in \mathbb{R}^{P \times K}$ 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 [34, 32, 35] демонстрируют, что спектральное распределение матрицы Гессе характеризуется наличием основной массы собственных значений, сосредоточенной вблизи нуля (обусловленной H-компонентой), и выбросов, распределенных в области ненулевых значений (обусловленных G-компонентой). 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:

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

Дополнительное теоретическое обоснование данной аппроксимации представляется в рамках теории ядра нейронного касательного пространства [36, 37], 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 Н-компоненты.

На основе работ [38], предлагающих аналитическую аппроксимацию G-компоненты для полносвязных 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 (для упрощения записи индекс i опущен):

- Матричное представление функции активации ReLU:

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

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 [39]:

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

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 (8).

2.1.1. Спектральная оценка матрицы Гессе

Теорема 8 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 является естественным предположением 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 8. *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:*

$$\|\mathbf{W}^{(p)}\|_2 \leq M_{\mathbf{W}},$$

$$\|\mathbf{x}_i\|_2 \leq M_{\mathbf{x}},$$

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:

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

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

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

а 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:

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

Далее оценим норму $\|\mathbf{F}\|_2$. Для оценки $\|\mathbf{F}\|_2$ проанализируем спектральную норму матриц $\mathbf{G}^{(p)}$, которые определяются как:

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

Используя свойство субмультипликативности спектральной нормы, имеем:

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

Причем, так как матрицы $\mathbf{D}^{(p)}$ — диагональные матрицы с элементами 0 или 1, так как является индикаторной матрицы оператора ReLU, то ее спектральная норма не превосходит 1. Таким образом, получаем:

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

Далее, так как матрица \mathbf{F} представляет собой вертикальную конкатенацию блоков вида $(\mathbf{G}^{(p)})^\top \otimes \mathbf{x}^{(p)}$ и $(\mathbf{G}^{(p)})^\top$, то 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 вектора 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, что для всех p норма матриц $\|\mathbf{W}^{(p)}\|_2 \leq M_{\mathbf{W}}$ и $\|\mathbf{x}_i\|_2 \leq M_{\mathbf{x}}$, то:

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

Причем, заметим, что сумма в правой части представляет собой сумму геометрической прогрессии:

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

а 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 2. В Теореме 8 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 эмперическому представлению о том, что кривизна функции потерь пропорциональна квадрату нормы весов и квадрату нормы входных данных.

Теорема 9 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 9. *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 8 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 $p = 1, \dots, L$, а 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 , и следовательно:

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

В Теореме 8 была получена следующая оценка:

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

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

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

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

$$L\sqrt{2}M_x^2(hM)^{2L} = \propto Lh^{2L},$$

поскольку M и M_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 ассимптотика:

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

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

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

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 . Хотя может показаться, что полученная оценка является завышенной, на самом деле это не так. Дело в том, что если выбрать значение h большим, то ограничивающая константа M , скорее всего, будет очень малой. Благодаря этому значение под знаком степени $2L$, вероятно, окажется меньше единицы. В дальнейшем мы используем эту верхнюю оценку для получения неравенства разности функции потерь.

Замечание 3. Полученная оценка имеет важное значение для теории глубокого обучения, поскольку она раскрывает компромисс между шириной и глубиной нейронной сети с точки зрения сложности оптимизации. Экспоненциальная зависимость от глубины сети L объясняет известные практические трудности при обучении очень глубоких сетей, в то время как степенная зависимость от ширины h указывает на более управляемый рост сложности при увеличении размера слоев. Учет взаимосвязи между h и M позволяет более адекватно интерпретировать полученные оценки и использовать их для практического анализа сходимости методов оптимизации.

2.2. Матричные модели глубокого обучения

В данном 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.

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

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

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

$$f_{\theta}(\mathbf{x}) = \mathbf{T}^{(L+1)} \Lambda^{(L)} \dots \Lambda^{(1)} \mathbf{T}^{(1)} \mathbf{x}. \quad (2.1)$$

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 $\Lambda^{(p)}$, 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:

$$\mathbf{G}^{(p)} := \mathbf{T}^{(L+1)} \mathbf{\Lambda}^{(L)} \dots \mathbf{T}^{(p+1)} \mathbf{\Lambda}^{(p)};$$

$$\mathbf{G}^{(L+1)} := \mathbf{I};$$

и преобразования, которые предшествовали p -му слою:

$$\mathbf{R}^{(p)} := \mathbf{\Lambda}^{(p)} \mathbf{T}^{(p)} \dots \mathbf{\Lambda}^{(1)} \mathbf{T}^{(1)}; \quad p = \overline{1, L};$$

$$\mathbf{R}^{(0)} := \mathbf{I}.$$

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

Используя введенные обозначения, запишем следующие выражения:

$$\mathbf{z} = \mathbf{G}^{(p)} \mathbf{z}^{(p)},$$

$$\mathbf{x}^{(p)} = \mathbf{R}^{(p)} \mathbf{x},$$

$$\mathbf{z} = f_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{G}^{(p)} \mathbf{T}^{(p)} \mathbf{R}^{(p-1)} \mathbf{x}.$$

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.

2.2.1. Матричная факторизация матрицы Гессе

Теорема 10. Пусть функция нейронной сети $f_{\theta}(\mathbf{x})$ представима в виде (2.1), тогда матрица Гессе функции потерю относительно параметров модели может быть представлен в факторизованной форме: $\mathbf{H}_O(\boldsymbol{\theta}) = \mathbf{Q}^T \mathbf{F}^T \mathbf{A} \mathbf{F} \mathbf{Q}$, 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:

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

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 [40], 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 10 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} (зависящего только от логитов и функции потерь) с помощью матриц \mathbf{F} и \mathbf{Q} , которые capture архитектурные свойства сети и параметризацию слоев соответственно.

2.2.2. Оценка спектральных норм матрицы Гессе

Теорема 11 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 11. Пусть нейронная сеть $f_{\theta}(\mathbf{x})$ представима в виде (2.1). 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. Using the results of the previous Lemma 10, it is enough for us to evaluate the upper bound of the expression: $\|\mathbf{Q}\|^2 \|\mathbf{F}\|^2 \|\mathbf{A}\|$

In the work [41], the norm of matrix \mathbf{A} was examined, and it was proven that:

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

Norm of block-diagonal matrix is not greater than max of block's norm

$$\|\mathbf{Q}\|^2 \leq \max_{i=1,\dots,L+1} \|\mathbf{Q}^{(i)}\|^2 \leq q^2.$$

Norm of matrix product is less or equal then product of norms:

$$\|\mathbf{G}^{(p)}\|^2 \leq \|\mathbf{T}^{(p+1)}\|^2 \dots \|\mathbf{T}^{(L+1)}\|^2 \leq w_{\mathbf{T}}^{2(L-p+1)}.$$

$$\left\| \mathbf{R}^{(p-1)} \right\|^2 \leqslant \left\| \mathbf{T}^{(1)} \right\|^2 \dots \left\| \mathbf{T}^{(p-1)} \right\|^2 \leqslant w_{\mathbf{T}}^{2(p-1)}.$$

The spectral matrix norm of the Kronecker product is equal to their ordinary product norm. Spectral norm of vertical stacked matrices is less or equal then sum of norms of it's blocks

$$\begin{aligned} \|\mathbf{F}\|^2 &\leqslant \sum_{p=1}^{L+1} \left\| \mathbf{G}^{(p+1)\top} \otimes \mathbf{R}^{(p-1)} \mathbf{x} \right\|^2 = \\ &= \sum_{p=1}^{L+1} \left\| \mathbf{G}^{(p)} \right\|^2 \left\| \mathbf{R}^{(p-1)} \mathbf{x} \right\|^2. \end{aligned}$$

Substituting the obtained estimates into the $\|\mathbf{H}_O\|$ formula we get

$$\begin{aligned} \|\mathbf{F}\|^2 &\leqslant \|\mathbf{x}\|^2 \sum_{p=1}^{L+1} w_{\mathbf{T}}^{2L} \leqslant \|\mathbf{x}\|^2 (L+1) w_{\mathbf{T}}^{2L}. \\ \|\mathbf{H}_O\| &\leqslant \|\mathbf{Q}\|^2 \|\mathbf{F}\|^2 \|\mathbf{A}\| \leqslant \sqrt{2} \|\mathbf{x}\|^2 q^2 (L+1) w_{\mathbf{T}}^{2L}. \end{aligned}$$

□

Доказательство. Согласно Теоремы 10, 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:

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

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 [41], для кросс-энтропийной функции потерь справедлива оценка:

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

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:

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

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

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

причем, 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:

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

□

Рассмотрим частный случай модели глубокого обучения, удовлетворяющего условию матричной факторизации для сверточной нейронной сети (англ. CNN) с одномерной сверткой можно получить оценки на норму матрицы Гессе, как показано далее в Теореме 12. Здесь для простоты мы сохраняем обозначение $\mathbf{T}^{(p)}$, но используем его для одномерных сверток, и поясним, как они представляются в виде линейных операторов. Известно, что сверточные сети часто могут быть представлены в виде линейных сверточных нейронных сетей (LCN). Обычно это относится к представлению сверточных сетей с помощью матриц Тэплица [42, 43]. В данной работе мы используем обозначения для матриц Тэплица из работы [40]. Также в указанной работе авторы определили специальный тип матрицы $\mathbf{Q}^{(p)}$, соответствующий структуре одномерной матрицы Тэплица. Наша 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 $vec(\mathbf{x}^{(p)}) \in \mathbb{R}^{(C_p d_p)}$. Получили:

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

Теорема 12 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 12. Пусть задана сеть вида:

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

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:

$$C_l \leq C,$$

$$k_i \leq k,$$

$$d_i \leq d_1 := d,$$

$$|\mathbf{W}_{i,j,k}^{(p)}|^2 \leq w^2.$$

Тогда норма матрицы Гессе имеет следующую верхнюю оценку:

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

Доказательство. Из Теоремы 11 следует, что требуется доказать следующие неравенства:

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

Согласно работе [40], 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 $C_l \leq C$, 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:

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

Согласно работе [40] 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:

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

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

Собирая все полученные оценки воедино:

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

□

Далее рассматриваются двумерные сверточные сети. Аналогично, для простоты мы сохраняем 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 частности, в случае векторизованного входа, используем ту же структуру Теплица, что и в [44], но в этом случае будет удобнее использовать конкретную матрицу $\mathbf{T}^{(p)}$, 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 13 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 13. 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:

$$C_l \leq C,$$

$$k_i \leq k,$$

$$m_i \leq m_1 := m,$$

$$n_i \leq n_1 := n,$$

$$|\mathbf{W}_{i,j,k}^{(p)}|^2 \leq w^2.$$

Тогда норма матрицы Гессе имеет следующую верхнюю оценку:

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

где $q^2 = C^2 k^2 mn$.

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

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

а следовательно:

$$\left\| \mathbf{T}^{(p)} \right\|_F^2 = \sum_{c_1,i,k,l}^{C_{p-1}, C_p n_p m_p, k_p^1, k_p^2} (\mathbf{W}_{c_1, c_2(i), k, l}^{(p)})^2, \quad (2.2)$$

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

По аналогии с доказательством 12, используя 11 необходимо доказать два неравенства:

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

Сначала оценим норму $\mathbf{T}^{(p)}$, используя (2.2) получем:

$$\|\mathbf{T}^{(p)}\|^2 \leq \|\mathbf{T}^{(p)}\|_F^2 \leq \sum_i Ck^2w^2 \leq C^2k^2w^2mn.$$

Далее оценим норму производной слоя по параметрам:

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

Как упоминалось ранее, строка $\mathbf{T}^{(p)}$ — является $vec_r(\mathbf{W}_{*,i,*,*}^{(p)})$, 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 $\mathbf{T}^{(p)}$, но очевидно, что один 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 $\mathbf{T}_{i,*}^{(p)}$, мы можем зафиксировать i и соответствующий c_2 , а также одновременно мы знаем, что для каждого c_1, k_1, k_2 существует только один столбец j такой, что $\mathbf{T}_{i,j}^{(p)} = \mathbf{W}_{c_1,c_2,k_1,k_2}^{(p)}$, а следовательно получаем:

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

Далее используем свойство нормы Фробениуса как верхней оценки спектральной нормы:

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

□

Замечание 4. Полученные оценки в Теоремах 12 и 13 имеют недостаток связанный с тем, что они не учитывают уменьшение размеров после сверточных операций и зависят только от верхних границ параметров.

Рассмотренные выше теоремы 12 и 13 оперируют с сетями, которые состоят исключительно из сверточных слоев, что редко встречается на практике. Далее в Теоремах 14, 15, 16 рассматриваются случаи добавления других распространенных слоев в сверточных сетях.

Также требует оценки норма слоя 14. Теорема 14 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 14. 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 $q^2 = m n C^2 k^2$.

Доказательство. Введем обозначение $\mathbf{M}^{(l)}$ для слоя 2D-Max-Pool. Аналогично сверточным слоям, мы можем описать каждую строку $\mathbf{M}^{(l)}$. Отметим некоторые свойства \mathbf{M} , которые будут использоваться: во-первых, строка \mathbf{M}_{i*} соответствует определенному окну пулинга и, во-вторых, 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, то в каждой строке имеется только один ненулевой элемент, а 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 13 оцениваем компоненты $\mathbf{G}^{(p)}$ и $\mathbf{R}^{(p-1)}$, однако с учетом нового слоя:

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

а следовательно 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 5. 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 расположена слой пулинга (меньше l), 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 15 устанавливает оценку нормы матрицы Гессе для сверточной сети, содержащей слой усредняющего пулинга.

Теорема 15. *Пусть задана сеть вида:*

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

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

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

где $q^2 = m n C^2 k^2$.

Доказательство. Введем обозначение $\mathbf{A}^{(l)}$ для слоя 2D-Avg-Pool. Заметим, 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, мы 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 $\|\mathbf{A}^{(p)}\| \leq 1$, мы можем полностью повторить вычисления доказательства Теоремы 14 и получить тот же результат. \square

Замечание 6. Полученная оценка в рамках Теоремы 15 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 16 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 16. 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 13 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

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

причем, известно, что

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

а 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:

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

а следовательно получаем:

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

Применяя этот результат к матрице Гессе:

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

□

2.3. Матрица Гессе для трансформерной модели глубокого обучения

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:

$$\mathbf{F}(\mathbf{X}) = \mathbf{A}(\mathbf{X}) \mathbf{X} \mathbf{W}_V, \tag{2.3}$$

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 (2.3), полный блок трансформера выглядит в 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 полносвязной 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. 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. Выполнения данного свойства можно добиться, добавив к знаменателю положительную константу, но это усложнит вычисления.

Для оценки матрицы Гессе рассматривается среднеквадратичная функция ошибки:

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

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

$$\frac{\partial^2(\mathcal{L}_k \circ f_{\mathbf{w}})}{\partial \mathbf{W}_i \partial \mathbf{W}_j} = \frac{\partial f_{\mathbf{w}}}{\partial \mathbf{W}_i}(\cdot)^\top \frac{\partial^2 \mathcal{L}_k}{\partial f_{\mathbf{w}}^2}(f_{\mathbf{w}}(\cdot)) \frac{\partial f_{\mathbf{w}}}{\partial \mathbf{W}_j}(\cdot) + \left(\frac{\partial \mathcal{L}_k}{\partial f_{\mathbf{w}}}(f_{\mathbf{w}}(\cdot)) \otimes \mathbf{I}_{p_i q_i} \right) \frac{\partial^2 f_{\mathbf{w}}}{\partial \mathbf{W}_i \partial \mathbf{W}_j}(\cdot) \quad (2.4)$$

Для начала вычислим обобщенные выражения матрицы Гессе для слоя самовнимания и расширяем их до полного блока модели трансформер. Подход основан на теоретической базе [45], 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} относящиеся к матрицам $\{\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V\}$. Используя разложения Гаусса-Ньютона (2.4):

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

где \mathbf{H}_o является outer-product частью матрицы Гессе, а \mathbf{H}_f является матрицей Гессе функции самовнимания. Результаты этого разложения можно вычислить согласно Теоремам 3.1-3.2 в работе [45].

2.3.1. Матрица Гессе для слоя самовнимания

Проведем оценку нормы матрицы Гессе для одного слоя самовнимания. Результат данной оценки показан в Теореме 17.

Теорема 17. *Пусть $\|\cdot\|_2$ является спектральной нормой, тогда для слоя самовнимания получаем:*

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

где

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

Доказательство. Используя результаты Леммы А.3 из работы [46], 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:

$$\begin{aligned} \|\mathbf{Z}_1\|_2 &\leq \|\mathbf{I}_L \otimes \mathbf{X}^\top\|_2 \left\| \frac{\partial \mathbf{A}}{\partial \mathbf{T}} \right\|_2 \|\mathbf{X} \otimes \mathbf{X}\|_2 \leq \\ &\leq \|\mathbf{X}\|_2 \frac{1}{L} \|\mathbf{X}\|_2^2 = \frac{1}{L} \|\mathbf{X}\|_2^3 \end{aligned} \quad (2.5)$$

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, а следовательно, каждый элемент матрицы $\mathbf{A}_{i,j} \leq 1$. Далее используя свойства 4 получаем $\|\mathbf{A}\|_{\max} \leq \|\mathbf{A}\|_2 \leq \sqrt{LL} \|\mathbf{A}\|_{\max} = L \|\mathbf{A}\|_{\max} \leq L$. 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 различных 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 17.

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 [45]. 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:

$$\begin{aligned}\|\mathbf{R}_m\| &\leq \sqrt{\text{rank}(\mathbf{F}(\mathbf{X}) - \mathbf{Target})} \|\mathbf{F}(\mathbf{X}) - \mathbf{Target}\|_2 \leq \\ &\leq \sqrt{\text{rank}(\mathbf{F}(\mathbf{X}) - \mathbf{Target})} (\|\mathbf{A}\|_2 \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 + \|\mathbf{Target}\|_2) \leq \\ &\leq \sqrt{\text{rank}(\mathbf{F}(\mathbf{X}) - \mathbf{Target})} (L \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 + \|\mathbf{Target}\|_2)\end{aligned}$$

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 С1 с работы [45], где указано, что:

$$\frac{\partial^2 \mathbf{A}_{i,j}}{\partial \mathbf{T}_{i,:} \partial \mathbf{T}_{i,:}} = \mathbf{A}_{i,j} \left(2\mathbf{A}_{i,:} \mathbf{A}_{i,:}^\top + \mathbf{E}_{j,j}^{L,L} - \text{diag}(\mathbf{A}_{i,:}) - \mathbf{e}_j \mathbf{A}_{i,:}^\top - \mathbf{A}_{i,:} \mathbf{e}_j^\top \right) \in \mathbb{R}^{L \times L}, \quad (2.6)$$

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, а следовательно 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 $\mathbf{A}_{i,j} \leq 1$, а следовательно можем оценить матрицу $\|\mathbf{A}_{i,:} \mathbf{A}_{i,:}^\top\|_2$, так как $\mathbf{A}_{i,:}$ является строкой softmax-матрицы, то значение

суммы строки равняются 1. Таким образом, мы можем использовать векторно-матричные неравенства для получения выражение:

$$\|\mathbf{A}_{i,:}\mathbf{A}_{i,:}^\top\|_2 \leq \|\mathbf{A}_{i,:}\|_2^2 \leq \|\mathbf{A}_{i,:}\|_1^2 = 1. \quad (2.7)$$

Аналогично заметим, что

$$\|\mathbf{E}_{j,j}^{m,n}\|_2 = \|\mathbf{e}_j\mathbf{e}_j^\top\|_2 \leq 1. \quad (2.8)$$

Перейдем к оценке нормы диагональной матрицы $\|diag(\mathbf{A}_{i,:})\|_2$. Заметим, что для диагональной матрицы верно следующее выражение:

$$\|diag(\mathbf{A}_{i,:})\|_2 = \max_j \mathbf{A}_{i,j} \leq 1. \quad (2.9)$$

Используя оценки (2.7) и (2.9) и (2.8) оценим нормы $\mathbf{e}_j\mathbf{A}_{i,:}^\top$ и $\mathbf{A}_{i,:}\mathbf{e}_j^\top$. Матрицы $\mathbf{e}_j\mathbf{A}_{i,:}^\top$ and $\mathbf{A}_{i,:}\mathbf{e}_j^\top$ являются матрицами ранга 1, причем только с одной не нулевой строкой и колонкой соответственно с элементами матрицы $\mathbf{A}_{i,:}$. Следовательно их спектральные нормы оценивается сверху нормой матрицы $\|\mathbf{A}_{i,:}\|_2 \leq 1$.

Получаем, что все слагаемые в выражении (2.6) имеют верхнюю оценку 1, а следовательно:

$$\left\| \frac{\partial^2 \mathbf{A}}{\partial \mathbf{T}^2} \right\|_2 \leq 6$$

Возвращаясь к выражению (2.5) 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 :

$$\begin{aligned}
& \|\mathbf{H}_f(\mathbf{W}_Q, \mathbf{W}_K)\| \leq \\
& \leq \frac{2}{Ld_V d_K} \|\mathbf{R}_{d_V d_K} (\mathbf{I}_L \otimes \mathbf{W}_V^\top \otimes \mathbf{I}_{d_V} \otimes \mathbf{W}_K^\top) \mathbf{Z}_2 (\mathbf{W}_Q \otimes \mathbf{I}_{d_V}) \mathbf{K}_{d_K, d_V}\|_2 + \\
& + \frac{2}{Ld_V \sqrt{d_K}} \|\mathbf{R}_{d_V} (\mathbf{I}_L \otimes \mathbf{W}_V^\top \otimes \mathbf{I}_{d_V}) (\mathbf{Z}_1 \otimes \mathbf{I}_{d_V}) \mathbf{S} \otimes \mathbf{I}_{d_K}\|_2 \leq \\
& \leq \frac{2}{Ld_V d_K} \sqrt{\text{rank}(\mathbf{F}(\mathbf{X}) - \mathbf{Target})} \left(L \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 + \right. \\
& \quad \left. + \|\mathbf{Target}\|_2 \right) \|\mathbf{W}_V\|_2 \|\mathbf{W}_K\|_2 \|\mathbf{W}_Q\|_2 6 \|\mathbf{X}\|_2^5 + \\
& + \frac{2}{Ld_V \sqrt{d_K}} \sqrt{\text{rank}(\mathbf{F}(\mathbf{X}) - \mathbf{Target})} \left(L \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 + \right. \\
& \quad \left. + \|\mathbf{Target}\|_2 \right) \|\mathbf{W}_V\|_2 \frac{1}{L} \|\mathbf{X}\|_2^3 \sqrt{d_V} = \\
& = \frac{2 \sqrt{\text{rank}(\mathbf{F}(\mathbf{X}) - \mathbf{Target})} (L \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 + \|\mathbf{Target}\|_2)}{Ld_V \sqrt{d_V d_K}} \|\mathbf{W}_V\|_2 \cdot \\
& \quad \cdot \left(3L \|\mathbf{W}_K\|_2 \|\mathbf{W}_Q\|_2 \|\mathbf{X}\|_2^5 + \frac{d_V}{L} \|\mathbf{X}\|_2^3 \right).
\end{aligned}$$

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:

$$\begin{aligned}
& \|\mathbf{H}(\mathbf{W}_i, \mathbf{W}_j)\|_2 \leq \\
& \leq 3 \max \left(\frac{2L}{d_V} \|\mathbf{X}\|_2^2, \right. \\
& \frac{2}{L^3 d_V d_K} \|\mathbf{W}_K\|_2^2 \|\mathbf{W}_V\|_2^2 \|\mathbf{X}\|_2^6 + \\
& + \frac{12}{d_V d_K} \sqrt{\text{rank}(\mathbf{F}(\mathbf{X}) - \mathbf{Target})} \left(L \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 + \right. \\
& \left. + \|\mathbf{Target}\|_2 \right) \|\mathbf{W}_V\|_2 \|\mathbf{W}_K\|_2^2 \|\mathbf{X}\|_2^5, \\
& \frac{2}{L d_V \sqrt{d_K}} \|\mathbf{W}_V\|_2 \|\mathbf{W}_K\|_2 \|\mathbf{X}\|_2^4 + \\
& + \frac{2\sqrt{\text{rank}(\mathbf{F}(\mathbf{X}) - \mathbf{Target})}}{L^2 \sqrt{d_V d_K}} (L \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 + \|\mathbf{Target}\|_2) \|\mathbf{W}_K\|_2 \|\mathbf{X}\|_2^3, \\
& \frac{2}{L^3 d_V d_K} \|\mathbf{W}_K\|_2 \|\mathbf{W}_Q\|_2 \|\mathbf{W}_V\|_2^2 \|\mathbf{X}\|_2^6 + \\
& + \frac{2\sqrt{\text{rank}(\mathbf{F}(\mathbf{X}) - \mathbf{Target})} \left(L \|\mathbf{X}\|_2 \|\mathbf{W}_V\|_2 + \|\mathbf{Target}\|_2 \right)}{L d_V \sqrt{d_V d_K}} \\
& \cdot \|\mathbf{W}_V\|_2 \left(3L \|\mathbf{W}_K\|_2 \|\mathbf{W}_Q\|_2 \|\mathbf{X}\|_2^5 + \frac{d_V}{L} \|\mathbf{X}\|_2^3 \right) \left. \right).
\end{aligned}$$

Полученное выражение почти полностью соответствует выражению M , где для полного соответствия требуется воспользоваться неравенством $\text{rank}(\mathbf{F}(\mathbf{X}) - \mathbf{Target}) \leq \min(L, d_V)$. \square

Теорема 17 оценивает только один слой самовнивания. Теперь перейдем к оценке полного блока трансформера. Полный трансформер слой содержит слой самовнивания, блок полно связной сети (англ. FFN), и слоя нормализации выходов (англ. LayerNorm). Весь блок описывается следующими выражениями:

$$\mathbf{Y} = \text{LayerNorm}(\mathbf{X} + \mathbf{F}(\mathbf{X})) \quad (2.10)$$

$$\mathbf{Z} = \text{LayerNorm}(\mathbf{Y} + \text{FFN}(\mathbf{Y})),$$

где

$$\text{FFN}(\mathbf{Y}) = \sigma(\mathbf{Y} \mathbf{W}_1 + \mathbf{b}_1) \mathbf{W}_2 + \mathbf{b}_2,$$

с матрицами параметров $\mathbf{W}_1 \in \mathbb{R}^{d_V \times d_{\text{ff}}}$, матрицами $\mathbf{W}_2 \in \mathbb{R}^{d_{\text{ff}} \times d_V}$, векторами $b_1 \in \mathbb{R}^{d_{\text{ff}}}$, $b_2 \in \mathbb{R}^{d_V}$, а также функцией активации σ . Функция 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 предположения, что параметры γ, β являются постоянными и не меняются в процессе оптимизации.

2.3.2. Матрица Гессе для LayerNorm слоя

Для начала вычислим для функции LayerNorm матрицу Якоби относительно параметров модели, для этого докажем Теорему 18.

Теорема 18. 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:

$$\begin{aligned} \frac{\partial \text{LayerNorm}(\mathbf{X})}{\partial \mathbf{X}} &= (\mathbf{P}(\mathbf{X}) \otimes \mathbf{I}_{d_V}) \left(\mathbf{I}_{Ld_V} - \frac{1}{d_V} (\mathbf{I}_L \otimes \mathbf{1}_{d_V \times d_V}) \right) + \\ &\quad + (\mathbf{I}_L \otimes \mathbf{M}(\mathbf{X})^\top) \frac{\partial \mathbf{P}(\mathbf{X})}{\partial \mathbf{X}}, \end{aligned}$$

$\sigma \partial e$

$$\begin{aligned} \frac{\partial \mathbf{P}}{\partial \mathbf{X}} &= \frac{1}{\sqrt{d_V}} \left(-\mathbf{D}^{-1} \otimes \mathbf{D}^{-\top} \right) \cdot \\ &\quad \cdot (\mathbf{e}_1 \otimes \mathbf{e}_1, \dots, \mathbf{e}_L \otimes \mathbf{e}_L) \cdot \\ &\quad \cdot \left(\text{diag}^{-1}(\text{vec}_r^{1/2}(\mathbf{M}^{\circ 2} \mathbf{1}_{d_V})) (\mathbf{I}_L \otimes \mathbf{1}_{d_V}^\top) \text{diag}(\text{vec}_r(\mathbf{M})) \frac{\partial \mathbf{M}}{\partial \mathbf{X}} \right), \end{aligned}$$

$\sigma \partial e \mathbf{D} = \text{diag}(\sigma(\mathbf{X})).$

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 31 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:

$$\frac{\partial \mathbf{M}}{\partial \mathbf{X}} = \frac{\partial (\mathbf{X} - \frac{1}{d_V} \mathbf{X} \mathbf{1}_{d_V \times d_V})}{\partial \mathbf{X}} = (\mathbf{I}_L \otimes \mathbf{I}_{d_V}) - \frac{1}{d_V} (\mathbf{I}_L \otimes \mathbf{1}_{d_V \times d_V})$$

Далее 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 18.

Далее, применив цепное правило получаем:

$$\frac{\partial \mathbf{P}}{\partial \mathbf{X}} = \frac{\partial \mathbf{D}^{-1}}{\partial \mathbf{D}} \frac{\partial \text{diag}(\sigma(\mathbf{X}))}{\partial \sigma(\mathbf{X})} \frac{\partial \sigma(\mathbf{X})}{\partial \mathbf{X}},$$

причем используя свойства 33, 34 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 $\tau = \mathbf{Q} \cdot \mathbf{1}_L$, $\mathbf{Q} = \mathbf{M}^{\circ 2}$, а следовательно 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 35 и 36 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 мы получили все составляющие для вычисления матрицы Якоби для LayerNorm оператора:

$$\begin{aligned} \frac{\partial \text{LayerNorm}(\mathbf{X})}{\partial \mathbf{X}} &= (\mathbf{P}(\mathbf{X}) \otimes \mathbf{I}_{d_V}) \frac{\partial \mathbf{M}}{\partial \mathbf{X}} + (\mathbf{I}_L \otimes \mathbf{M}^\top) \frac{\partial \mathbf{P}}{\partial \mathbf{X}} = \\ &= (\mathbf{P}(\mathbf{X}) \otimes \mathbf{I}_{d_V}) \frac{\partial \mathbf{M}}{\partial \mathbf{X}} + \\ &\quad + (\mathbf{I}_L \otimes \mathbf{M}^\top) \frac{1}{\sqrt{d_V}} (-\mathbf{D}^{-1} \otimes \mathbf{D}^{-\top}) \begin{pmatrix} \mathbf{e}_1 \otimes \mathbf{e}_1 & \dots & \mathbf{e}_L \otimes \mathbf{e}_L \end{pmatrix} \cdot \\ &\quad \cdot \left(\text{diag}^{-1}(\text{vec}_r^{\circ \frac{1}{2}}(\mathbf{M}^{\circ 2} \cdot \mathbf{1}_{d_V})) \cdot (\mathbf{I}_L \otimes \mathbf{1}_{d_V}^T) \cdot \text{diag}(\text{vec}_r(\mathbf{M})) \frac{\partial \mathbf{M}}{\partial \mathbf{X}} \right), \end{aligned}$$

где

$$\begin{aligned}\mathbf{M}(\mathbf{X}) &= (\mathbf{X} - \frac{1}{d_V} \mathbf{X} \mathbf{1}_{d_V \times d_V}) \\ \mathbf{P}(\mathbf{X}) &= \text{diag}^{-1}(\sigma(\mathbf{X})) \\ \frac{\partial \mathbf{M}}{\partial \mathbf{X}} &= (\mathbf{I}_L \otimes \mathbf{I}_{d_V}) - \frac{1}{d_V} (\mathbf{I}_L \otimes \mathbf{1}_{d_V \times d_V}).\end{aligned}$$

□

Теперь же вычислим для функции матрицу Гессе относительно параметров модели, для этого докажем Теорему 19.

Теорема 19. *Пусть задан оператор LayerNorm в виде аналогичном Теореме 18:*

$$\text{LayerNorm}(\mathbf{X}) = \mathbf{P}(\mathbf{X})\mathbf{M}(\mathbf{X}),$$

с матрицей Якоби, полученнюю в Теореме 18:

$$\frac{\partial \text{LayerNorm}}{\partial \mathbf{X}} = (\mathbf{P} \otimes \mathbf{I}_{d_V})\mathbf{G} + (\mathbf{I}_L \otimes \mathbf{M}^\top)\mathbf{H},$$

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 18:

$$\mathbf{H} = \frac{\partial \mathbf{P}}{\partial \mathbf{X}}.$$

Тогда для функции LayerNorm матрица Гессе относительно параметров \mathbf{X} имеет вид:

$$\begin{aligned}\frac{\partial^2 \text{LayerNorm}}{\partial \mathbf{X}^2} &= ((\mathbf{P}(\mathbf{X}) \otimes \mathbf{I}_{d_V}) \otimes \mathbf{I}_{Ld_V}) \frac{\partial^2 \mathbf{M}}{\partial \mathbf{X}^2} + \\ &+ (\mathbf{I}_{Ld_V} \otimes \mathbf{G}^\top) \frac{\partial(\mathbf{P}(\mathbf{X}) \otimes \mathbf{I}_{d_V})}{\partial \mathbf{X}} + \\ &+ ((\mathbf{I}_L \otimes \mathbf{M}^\top) \otimes \mathbf{I}_{Ld_V}) \frac{\partial^2 \mathbf{P}}{\partial \mathbf{X}^2} + (\mathbf{I}_{Ld_V} \otimes \mathbf{H}^\top) \frac{\partial(\mathbf{I}_L \otimes \mathbf{M}^\top)}{\partial \mathbf{X}}.\end{aligned}\tag{2.11}$$

причем все матрицы явно вычислимые и задаются формулами, которые указаны ниже в доказательстве.

Доказательство. Используя свойство матричного произведения 9 получаем следующее выражение матрицы Гессе для оператора LayerNorm:

$$\begin{aligned} \frac{\partial^2 \text{LayerNorm}}{\partial \mathbf{X}^2} &= ((\mathbf{P}(\mathbf{X}) \otimes \mathbf{I}_{d_V}) \otimes \mathbf{I}_{Ld_V}) \frac{\partial^2 \mathbf{M}}{\partial \mathbf{X}^2} + \\ &\quad + \left(\mathbf{I}_{Ld_V} \otimes \left(\frac{\partial \mathbf{M}}{\partial \mathbf{X}} \right)^\top \right) \frac{\partial (\mathbf{P}(\mathbf{X}) \otimes \mathbf{I}_{d_V})}{\partial \mathbf{X}} + \\ &\quad + ((\mathbf{I}_L \otimes \mathbf{M}^\top) \otimes \mathbf{I}_{Ld_V}) \frac{\partial^2 \mathbf{P}}{\partial \mathbf{X}^2} + \\ &\quad + \left(\mathbf{I}_{Ld_V} \otimes \left(\frac{\partial \mathbf{P}}{\partial \mathbf{X}} \right)^\top \right) \frac{\partial (\mathbf{I}_L \otimes \mathbf{M}^\top)}{\partial \mathbf{X}}, \end{aligned}$$

причем заметим, 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 Леммы 37 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 размер вектора $\sigma(\mathbf{X})$ имеет размерность $L \times 1$, тогда матрица $\mathbf{D} \in \mathbb{R}^{L \times L}$, а 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 $(\mathbf{e}_1 \otimes \mathbf{e}_1 \dots \mathbf{e}_L \otimes \mathbf{e}_L) \in \mathbb{R}^{L^2 \times L}$. 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 $(\mathbf{I}_L \otimes \mathbf{1}_{d_V}^T) \in \mathbb{R}^{L \times Ld_V}$ и $\text{diag}(\text{vec}_r(M)) \in \mathbb{R}^{Ld_V \times Ld_V}$. Последнее слагаемое $\frac{\partial \mathbf{M}}{\partial \mathbf{X}}$ уже было посчитано ранее, причем его размерность $Ld_V \times Ld_V$. 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 31 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. Сначала вычислим $\frac{\partial \mathbf{A}_1}{\partial \mathbf{X}}$. Используя Лемму 38 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 37, 35 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 18:

$$\begin{aligned} \frac{\partial \mathbf{D}}{\partial \mathbf{X}} &= \left(\mathbf{e}_1 \otimes \mathbf{e}_1 \dots \mathbf{e}_L \otimes \mathbf{e}_L \right) \cdot \\ &\quad \cdot \left(\text{diag}^{-1}(\text{vec}_r^{\circ \frac{1}{2}}(\mathbf{M}^{\circ 2} \cdot \mathbf{1}_{d_V})) \cdot (\mathbf{I}_L \otimes \mathbf{1}_{d_V}^T) \cdot \text{diag}(\text{vec}_r(\mathbf{M})) \frac{\partial \mathbf{M}}{\partial \mathbf{X}} \right), \end{aligned}$$

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:

$$\begin{aligned}\mathbf{A}_2 &= \text{diag}^{-1}(\text{vec}_r^{\circ\frac{1}{2}}(\mathbf{M}^{\circ 2} \cdot \mathbf{1}_{d_V})) \\ \mathbf{B}_2 &= (\mathbf{I}_L \otimes \mathbf{1}_{d_V}^T) \cdot \text{diag}(\text{vec}_r(\mathbf{M})) \frac{\partial \mathbf{M}}{\partial \mathbf{X}} \\ \mathbf{E} &= \begin{pmatrix} \mathbf{e}_1 \otimes \mathbf{e}_1 & \dots & \mathbf{e}_L \otimes \mathbf{e}_L \end{pmatrix}.\end{aligned}$$

Для начала, заметим, что \mathbf{E} является константной матрицей относительно матрицы \mathbf{X} , а следовательно используя результат Леммы 31 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 31 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

$$\begin{aligned}\frac{\partial \mathbf{A}_3}{\partial \mathbf{X}} &= \frac{\partial \text{diag}(\text{vec}_r(\mathbf{M}))}{\partial \mathbf{X}} = \frac{\partial \text{diag}(\mathbf{v})}{\partial (\mathbf{v})} \frac{\partial \text{vec}_r(\mathbf{M})}{\partial \mathbf{M}} \frac{\partial \mathbf{M}}{\partial \mathbf{X}}, \\ \frac{\partial \mathbf{B}_3}{\partial \mathbf{X}} &= \frac{\partial^2 \mathbf{M}}{\partial \mathbf{X}^2} = 0,\end{aligned}$$

причем, используя Лемму 36 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 $\frac{\partial \text{vec}_r(\mathbf{M})}{\partial \mathbf{M}} = \mathbf{I}_{Ld_V}$. Для вычисления матрицы $\frac{\partial \mathbf{A}_2}{\partial \mathbf{X}}$ воспользуемся Леммами 35, 36, 34, 32 получаем выражение:

$$\frac{\partial \mathbf{A}_2}{\partial \mathbf{X}} = \frac{\partial \text{diag}^{-1}(\text{vec}_r^{\circ\frac{1}{2}}(\mathbf{M}^{\circ 2} \cdot \mathbf{1}_{d_V}))}{\partial \mathbf{X}} =$$

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 (2.11) вычислены, что заканчивает доказательство. \square

2.3.3. Матрица Гессе для нелинейности ReLU

Теорема 20. 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:

$$\mathcal{N} = \bigcup_{i,j} \{\mathbf{X} \in \mathbb{R}^{m \times n} : x_{ij} = 0\}.$$

Заметим, что каждое множество $\{x_{ij} = 0\}$ является гиперплоскостью коразмерности 1 в пространстве $\mathbb{R}^{m \times n}$, а 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 32:

$$\text{vec}_r(d\text{ReLU}(\mathbf{X})) = \text{diag}(\text{vec}_r(\mathbf{1}_{\{\mathbf{X}>0\}}))\text{vec}_r(d\mathbf{X}),$$

причем, используя свойство 10 и Лемму 36 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:

$$\frac{\partial^2 \text{ReLU}(\mathbf{X})}{\partial \mathbf{X}^2} = \mathbf{0}, \quad \mathbf{X} \notin \mathcal{N}.$$

□

2.3.4. Матрица Гессе для трансформера

Лемма 21. *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 17.

Определим блоки обобщенного функционального гессиана, используя результаты [45] 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

$$\mathbf{R}_m := \text{vec}_r(\mathbf{F}(\mathbf{X}) - \mathbf{Target})^\top \otimes \mathbf{I}_m \in \mathbb{R}^{m \times (m \cdot Ld_V)}.$$

Тогда для $i \in \{V, Q, K\}$ с $n_i := p_i q_i$ блоки функционального гессиана могут быть факторизованы как

$$\mathbf{H}_f(\mathbf{W}_i, \mathbf{W}_j) = \mathbf{R}_{n_i} \Phi_{ij}, \quad \Phi_{ij} := \frac{\partial^2 \mathbf{F}}{\partial \mathbf{W}_i \partial \mathbf{W}_j} \in \mathbb{R}^{(Ld_V \cdot n_i) \times n_j}.$$

В частности, блоки кривизны модели Φ_{ij} получаются из соответствующих выражений в [45, 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*

$$\mathbf{Z}_1 := (\mathbf{I}_L \otimes \mathbf{X}^\top) (\partial \mathbf{A} / \partial \mathbf{T}) (\mathbf{X} \otimes \mathbf{X}) \in \mathbb{R}^{Ld_V \times d_V^2},$$

$$\mathbf{Z}_2 := (\mathbf{I}_L \otimes \mathbf{X}^\top \otimes \mathbf{X}^\top \otimes \mathbf{X}^\top) \frac{\partial^2 \mathbf{A}}{\partial \mathbf{T}^2} (\mathbf{X} \otimes \mathbf{X}) \in \mathbb{R}^{Ld_V^3 \times d_V^2},$$

где $\frac{\partial^2 \mathbf{A}}{\partial \mathbf{T}^2}$ обозначает тензор вторых производных *softmax* (построчный), согласованный с vec_r и произведениями Кронекера, как указано выше, а \mathbf{Z}_1 — линейный оператор первой производной *softmax*, используемый в [45].

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 $\mathbf{K}_{\cdot, \cdot}$. Аналогичные симметричные соотношения дают Φ_{QV} и Φ_{KV} из Φ_{VQ} .

Доказательство. По определению обобщенного функционального гессиана в [45],

$$\mathbf{H}_f(\mathbf{W}_i, \mathbf{W}_j) = \left(\frac{\partial \ell}{\partial \mathbf{F}} \otimes \mathbf{I}_{p_i q_i} \right) \frac{\partial^2 \mathbf{F}}{\partial \mathbf{W}_i \partial \mathbf{W}_j}.$$

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 [45, Thm. 3.2]. Используя выражения для \mathbf{H}_f получаем выражения Φ_{ij} просто удаляя ведущей метрицы \mathbf{R}_{n_i} . \square

Лемма 22. 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 следую-

щие оценки спектральных норм:

$$\|\mathbf{Y}\|_2 \leq \|\mathbf{Y}\|_F = \sqrt{Ld_V}, \quad (2.12)$$

$$\|\text{FFN}(\mathbf{Y})\|_2 \leq \sqrt{\min(L, d_{ff})} \|\mathbf{Y}\|_2 \|\mathbf{W}_1\|_2 \|\mathbf{W}_2\|_2, \quad (2.13)$$

$$\|\mathbf{S}\|_2 \leq \|\mathbf{Y}\|_2 + \|\text{FFN}(\mathbf{Y})\|_2 \leq \sqrt{Ld_V} \left(1 + \sqrt{\min(L, d_{ff})} \|\mathbf{W}_1\|_2 \|\mathbf{W}_2\|_2 \right). \quad (2.14)$$

Доказательство. Оценим $\|\mathbf{Y}\|_2$. Согласно определению LayerNorm в рамках Теоремы 18 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 (2.12).

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

$$\|\text{ReLU}(\mathbf{Y}\mathbf{W}_1)\|_2 \leq \|\text{ReLU}(\mathbf{Y}\mathbf{W}_1)\|_F,$$

причем согласно определению 19, 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 (2.13).

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 (2.12) и (2.13), получаем (2.14). \square

Лемма 23. 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 18, а ее гессиан $\mathbf{H}_{\text{LN}}(\mathbf{X}) = \frac{\partial^2 \text{LayerNorm}(\mathbf{X})}{\partial \mathbf{X}^2}$ вычисляется как в Теореме 19.

Тогда выполняются следующие оценки:

$$\|\mathbf{J}_{\text{LN}}(\mathbf{X})\|_2 \leq \frac{1}{\sigma_{\min}} + \frac{\|\mathbf{X}\|_2^2}{\sqrt{n}\sigma_{\min}^3}, \quad (2.15)$$

$$\|\mathbf{H}_{\text{LN}}(\mathbf{X})\|_2 \leq \frac{\|\mathbf{X}\|_2}{\sigma_{\min}^3} \left(1 + \sqrt{\frac{m}{n}}\right) + \frac{\|\mathbf{X}\|_2^2}{\sqrt{n}\sigma_{\min}^3} + \frac{3\|\mathbf{X}\|_2^3}{n\sigma_{\min}^5}, \quad (2.16)$$

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 18:

$$\mathbf{J}_{\text{LN}}(\mathbf{X}) = (\mathbf{P} \otimes \mathbf{I}_n)\mathbf{G} + (\mathbf{I}_m \otimes \mathbf{M}^\top)\mathbf{H},$$

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 $\|\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n \times n}\|_2 \leq 1$, и произведение Кронекера сохраняет оценку спектральной нормы согласно свойствам 1, 2, а также Леммы 39. 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 свойству 1. Для $\|\mathbf{H}\|_2 = \left\| \frac{\partial \mathbf{P}}{\partial \mathbf{X}} \right\|_2$, Теорема 18 вместе с Леммами 35, 36, 33, 34 и свойствами 1, 2 дают оценку:

$$\begin{aligned}\left\| \frac{\partial \mathbf{P}}{\partial \mathbf{X}} \right\|_2 &\leq \frac{1}{\sqrt{n}} \|\mathbf{D}^{-1} \otimes \mathbf{D}^{-\top}\|_2 \left\| \text{diag}^{-1}(\text{vec}_r^{1/2}(\mathbf{M}^{\circ 2} \mathbf{1}_n)) \right\|_2 \cdot \\ &\quad \cdot \|\mathbf{I}_m \otimes \mathbf{1}_n^\top\|_2 \|\text{diag}(\text{vec}_r(\mathbf{M}))\|_2 \left\| \frac{\partial \mathbf{M}}{\partial \mathbf{X}} \right\|_2.\end{aligned}$$

Используя свойство 4 получаем оценки:

$$\begin{aligned}\|\mathbf{D}^{-1} \otimes \mathbf{D}^{-\top}\|_2 &= \|\mathbf{D}^{-1}\|_2^2 = \frac{1}{\sigma_{\min}^2}, \\ \left\| \text{diag}^{-1}(\cdot) \right\|_2 &= \frac{1}{\min_i \sqrt{\sum_v M_{i,v}^2}} = \frac{1}{\sqrt{n} \sigma_{\min}},\end{aligned}$$

$$\|\mathbf{I}_m \otimes \mathbf{1}^\top\|_2 = \sqrt{n},$$

$$\|\text{diag}(\text{vec}_r(\mathbf{M}))\|_2 = \|\mathbf{M}\|_{\max} \leq \|\mathbf{M}\|_2,$$

$$\left\| \frac{\partial \mathbf{M}}{\partial \mathbf{X}} \right\|_2 \leq 1,$$

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 (2.15):

$$\|\mathbf{J}_{\text{LN}}(\mathbf{X})\|_2 \leq \frac{1}{\sigma_{\min}} \cdot 1 + \|\mathbf{X}\|_2 \cdot \frac{\|\mathbf{X}\|_2}{\sqrt{n} \sigma_{\min}^3} = \frac{1}{\sigma_{\min}} + \frac{\|\mathbf{X}\|_2^2}{\sqrt{n} \sigma_{\min}^3}.$$

Из Теоремы 19 (с m, n), используя $\frac{\partial^2 \mathbf{M}}{\partial \mathbf{X}^2} = 0$,

$$\mathbf{H}_{LN}(\mathbf{X}) = (\mathbf{I}_{mn} \otimes \mathbf{G}^\top) \frac{\partial(\mathbf{P} \otimes \mathbf{I}_n)}{\partial \mathbf{X}} + ((\mathbf{I}_m \otimes \mathbf{M}^\top) \otimes \mathbf{I}_{mn}) \frac{\partial^2 \mathbf{P}}{\partial \mathbf{X}^2} + (\mathbf{I}_{mn} \otimes \mathbf{H}^\top) \frac{\partial(\mathbf{I}_m \otimes \mathbf{M}^\top)}{\partial \mathbf{X}}.$$

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:

$$\left\| ((\mathbf{I}_m \otimes \mathbf{M}^\top) \otimes \mathbf{I}_{mn}) \frac{\partial^2 \mathbf{P}}{\partial \mathbf{X}^2} \right\|_2 \leq \|\mathbf{X}\|_2 \left\| \frac{\partial^2 \mathbf{P}}{\partial \mathbf{X}^2} \right\|_2.$$

Далее оценим $\left\| \frac{\partial^2 \mathbf{P}}{\partial \mathbf{X}^2} \right\|_2$, следуя той же цепочке, что и в доказательстве Теоремы 19 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 35, 36, 33, 34 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 37 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 (2.16):

$$\begin{aligned}\|\mathbf{H}_{\text{LN}}(\mathbf{X})\|_2 &\leq \frac{\|\mathbf{X}\|_2}{\sigma_{\min}^3} + \left(\frac{\|\mathbf{X}\|_2^2}{\sqrt{n}\sigma_{\min}^3} + \frac{3\|\mathbf{X}\|_2^3}{n\sigma_{\min}^5} \right) + \frac{\sqrt{m}}{\sqrt{n}} \frac{\|\mathbf{X}\|_2}{\sigma_{\min}^3} = \\ &= \frac{\|\mathbf{X}\|_2}{\sigma_{\min}^3} \left(1 + \sqrt{\frac{m}{n}} \right) + \frac{\|\mathbf{X}\|_2^2}{\sqrt{n}\sigma_{\min}^3} + \frac{3\|\mathbf{X}\|_2^3}{n\sigma_{\min}^5}.\end{aligned}$$

□

Теорема 24. Для модели глубокого обучения архитектуры трансформер 2.10 матрица Якоби $\frac{\partial \mathbf{Z}}{\partial \mathbf{W}_i}$ вычисляется в следующем виде:

$$\frac{\partial \mathbf{Z}}{\partial \mathbf{W}_i} = \frac{\partial \text{LayerNorm}(FFN(\mathbf{Y}) + \mathbf{Y})}{\partial (FFN(\mathbf{Y}) + \mathbf{Y})} \frac{\partial (FFN(\mathbf{Y}) + \mathbf{Y})}{\partial \mathbf{W}_i}, \quad i \in \{1, 2\},$$

где

$$\frac{\partial (FFN(\mathbf{Y}) + \mathbf{Y})}{\partial \mathbf{W}_i} = \begin{cases} (\mathbf{I}_L \otimes \mathbf{W}_2^\top) \text{diag}(\text{vec}_r(\mathbf{1}_{\{\mathbf{X}>0\}})) (\mathbf{Y} \otimes \mathbf{I}_{d_{ff}}), & \text{for } i = 1 \\ \sigma(\mathbf{Y}\mathbf{W}_1) \otimes \mathbf{I}_{d_V}, & \text{for } i = 2 \end{cases},$$

причем $\frac{\partial \text{LayerNorm}(FFN(\mathbf{Y}) + \mathbf{Y})}{\partial (FFN(\mathbf{Y}) + \mathbf{Y})}$ вычисляется согласно Теоремы 18.

$$\frac{\partial \mathbf{Z}}{\partial \mathbf{W}_i} = \frac{\partial \text{LayerNorm}(FFN(\mathbf{Y}) + \mathbf{Y})}{\partial (FFN(\mathbf{Y}) + \mathbf{Y})} \frac{\partial (FFN(\mathbf{Y}) + \mathbf{Y})}{\partial \mathbf{Y}} \frac{\partial \mathbf{Y}}{\partial \mathbf{W}_i}, \quad i \in \{K, Q, V\},$$

где

$$\frac{\partial (FFN(\mathbf{Y}) + \mathbf{Y})}{\partial \mathbf{Y}} = (\mathbf{I}_L \otimes \mathbf{W}_2^\top) \text{diag}(\text{vec}_r(\mathbf{1}_{\{\mathbf{X}>0\}})) (\mathbf{I}_L \otimes \mathbf{W}_1^\top) + (\mathbf{I}_L \otimes \mathbf{I}_{d_V}),$$

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 [46], а матрица $\frac{\partial \text{LayerNorm}(\mathbf{F}(\mathbf{X}) + \mathbf{X})}{\partial (\mathbf{F}(\mathbf{X}) + \mathbf{X})}$ вычисляется согласно Теоремы 18.

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 (2.10), а именно:

$$\mathbf{Y} = \text{LayerNorm}(\mathbf{F}(\mathbf{X}) + \mathbf{X}),$$

$$\mathbf{Z} = \text{LayerNorm}(\text{FFN}(\mathbf{Y}) + \mathbf{Y}).$$

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 20 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 18 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

$$\begin{aligned}
\mathbf{M}(\text{FFN}(\mathbf{Y}) + \mathbf{Y}) &= ((\text{FFN}(\mathbf{Y}) + \mathbf{Y}) - \frac{1}{d_V} (\text{FFN}(\mathbf{Y}) + \mathbf{Y}) \mathbf{1}_{d_V \times d_V}), \\
\mathbf{P}((\text{FFN}(\mathbf{Y}) + \mathbf{Y})) &= \text{diag}^{-1}(\sigma(\text{FFN}(\mathbf{Y}) + \mathbf{Y})), \\
\frac{\partial \mathbf{M}}{\partial (\text{FFN}(\mathbf{Y}) + \mathbf{Y})} &= (\mathbf{I}_L \otimes \mathbf{I}_{d_V}) - \frac{1}{d_V} (\mathbf{I}_L \otimes \mathbf{1}_{d_V \times d_V}),
\end{aligned}$$

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 20 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 [46]:

$$\begin{aligned}
\frac{\partial \mathbf{F}}{\partial \mathbf{W}_V} &= \text{softmax} \left(\frac{\mathbf{X}\mathbf{W}_Q\mathbf{W}_K^\top\mathbf{X}^\top}{\sqrt{d_K}} \right) \mathbf{X} \otimes \mathbf{I}_{d_V} \\
\frac{\partial \mathbf{F}}{\partial \mathbf{W}_Q} &= (\mathbf{I}_L \otimes \mathbf{W}_V^\top\mathbf{X}^\top) \frac{\partial \mathbf{A}}{\partial \mathbf{M}} \left(\frac{\mathbf{X} \otimes \mathbf{X}\mathbf{W}_K}{\sqrt{d_K}} \right),
\end{aligned}$$

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 37 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 Леммы A.2 с работы [46], а матрица $\frac{\partial \text{LayerNorm}(\mathbf{F}(\mathbf{X}) + \mathbf{X})}{\partial (\mathbf{F}(\mathbf{X}) + \mathbf{X})}$ вычисляется согласно Теоремы 18. \square

В Теореме 24 получен вид матрицы Якоби для полного блока трансформера, теперь можно перейти к вычислению матрицы Гессе для, который получен в виде Теоремы 25.

Теорема 25. 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 18 и 19 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 18, 19 и 20 введем следующее:

$$\mathbf{D}_\sigma := \text{diag}(\text{vec}_r(\mathbf{1}_{\{\mathbf{Y}\mathbf{W}_1 > 0\}})) \in \mathbb{R}^{Ld_{ff} \times Ld_{ff}},$$

$$\mathbf{J}_{SY} := \frac{\partial \mathbf{S}}{\partial \mathbf{Y}} = (\mathbf{I}_L \otimes \mathbf{W}_2^\top) \mathbf{D}_\sigma (\mathbf{I}_L \otimes \mathbf{W}_1^\top) + (\mathbf{I}_L \otimes \mathbf{I}_{d_V}) \in \mathbb{R}^{Ld_V \times Ld_V},$$

где для матрицы $\mathbf{Y} = \text{LayerNorm}(\mathbf{F}(\mathbf{X}) + \mathbf{X})$ в условиях Теорем 18, 19 определено:

$$\mathbf{J}_Y := \frac{\partial \text{LayerNorm}(\mathbf{F}(\mathbf{X}) + \mathbf{X})}{\partial (\mathbf{F}(\mathbf{X}) + \mathbf{X})} \in \mathbb{R}^{Ld_V \times Ld_V},$$

$$\mathbf{H}_Y := \frac{\partial^2 \text{LayerNorm}(\mathbf{F}(\mathbf{X}) + \mathbf{X})}{\partial (\mathbf{F}(\mathbf{X}) + \mathbf{X})^2} \in \mathbb{R}^{(Ld_V)^2 \times Ld_V},$$

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 вычислимые в условиях Теоремы 24 в следующем виде:

$$\mathbf{G}_V := \frac{\partial \mathbf{F}}{\partial \mathbf{W}_V} \in \mathbb{R}^{Ld_V \times n_V},$$

$$\mathbf{G}_Q := \frac{\partial \mathbf{F}}{\partial \mathbf{W}_Q} \in \mathbb{R}^{Ld_V \times n_Q},$$

$$\mathbf{G}_K := \frac{\partial \mathbf{F}}{\partial \mathbf{W}_K} \in \mathbb{R}^{Ld_V \times n_K},$$

$$\mathbf{B}_1 := \frac{\partial \mathbf{S}}{\partial \mathbf{W}_1} = (\mathbf{I}_L \otimes \mathbf{W}_2^\top) \mathbf{D}_\sigma (\mathbf{Y} \otimes \mathbf{I}_{d_{ff}}) \in \mathbb{R}^{Ld_V \times n_1},$$

$$\mathbf{B}_2 := \frac{\partial \mathbf{S}}{\partial \mathbf{W}_2} = \sigma(\mathbf{Y}\mathbf{W}_1) \otimes \mathbf{I}_{d_V} \in \mathbb{R}^{Ld_V \times n_2},$$

$$\mathbf{B}_k := \frac{\partial \mathbf{S}}{\partial \mathbf{W}_k} = \mathbf{J}_{SY} \mathbf{J}_Y \mathbf{G}_k \in \mathbb{R}^{Ld_V \times n_k}, \quad k \in \{K, Q, V\}.$$

Тогда матрицы Гессе трансформера \mathbf{Z} по параметрам модели $(\mathbf{W}_i, \mathbf{W}_j)$ задается в виде:

$$\mathbf{H}_{\text{tr}}^{(i,j)} := \frac{\partial^2 \mathbf{Z}}{\partial \mathbf{W}_i \partial \mathbf{W}_j} = (\mathbf{J}_Z \otimes \mathbf{I}_{n_i}) \boldsymbol{\xi}_{ij} + (\mathbf{I}_{Ld_V} \otimes \mathbf{B}_i^\top) \mathbf{H}_Z \mathbf{B}_j, \quad (2.17)$$

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:

$$\boldsymbol{\xi}_{11} = \mathbf{0}_{(Ld_V \cdot n_1) \times n_1},$$

$$\boldsymbol{\xi}_{22} = \mathbf{0}_{(Ld_V \cdot n_2) \times n_2},$$

$$\boldsymbol{\xi}_{12} = (\mathbf{I}_L \otimes \mathbf{K}_{d_V, d_{ff}} \otimes \mathbf{I}_{d_V}) (\mathbf{I}_{Ld_{ff}} \otimes \text{vec}_r(\mathbf{I}_{d_V})) (\mathbf{D}_\sigma(\mathbf{Y} \otimes \mathbf{I}_{d_{ff}})),$$

$$\boldsymbol{\xi}_{21} = (\mathbf{I}_{Ld_V} \otimes ((\mathbf{Y} \otimes \mathbf{I}_{d_{ff}})^\top \mathbf{D}_\sigma^\top)) (\mathbf{I}_L \otimes \mathbf{K}_{d_V, L} \otimes \mathbf{I}_{d_{ff}}) (\text{vec}_r(\mathbf{I}_L) \otimes \mathbf{I}_{d_V d_{ff}}) \mathbf{K}_{d_{ff}, d_V},$$

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 $k \in \{K, Q, V\}$:

$$\boldsymbol{\xi}_{1k} = ((\mathbf{I}_L \otimes \mathbf{W}_2^\top) \mathbf{D}_\sigma \otimes \mathbf{I}_{n_k}) (\mathbf{I}_L \otimes \mathbf{K}_{d_{ff}, d_V} \otimes \mathbf{I}_{d_{ff}}) (\mathbf{I}_{Ld_V} \otimes \text{vec}_r(\mathbf{I}_{d_{ff}})) (\mathbf{J}_Y \mathbf{G}_k),$$

$$\boldsymbol{\xi}_{2k} = (\mathbf{I}_L \otimes \mathbf{K}_{d_V, d_{ff}} \otimes \mathbf{I}_{d_V}) (\mathbf{I}_{Ld_{ff}} \otimes \text{vec}_r(\mathbf{I}_{d_V})) (\mathbf{D}_\sigma(\mathbf{I}_L \otimes \mathbf{W}_1^\top) \mathbf{J}_Y \mathbf{G}_k),$$

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 $k, \ell \in \{K, Q, V\}$:

$$\boldsymbol{\xi}_{k\ell} = (\mathbf{J}_{SY} \otimes \mathbf{I}_{n_k}) [(\mathbf{I}_{Ld_V} \otimes \mathbf{G}_k^\top) (\mathbf{H}_Y \mathbf{G}_\ell) + (\mathbf{J}_Y \otimes \mathbf{I}_{n_k}) \boldsymbol{\Phi}_{k\ell}],$$

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 21. 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 получе-

ны матрицы $\mathbf{H}_Z, \mathbf{H}_Y$ в рамках Теоремы 19 и которые являются симметричными по построению и оператор $ReLU$, для которого матрица Гессе является нулевой согласно Теоремы 20, а во вторых все другие отображения являются линейными, а следовательно согласно Леммы 31 и свойства производной произведения Кронекера их частные производные являются коммутативными почти всюду.

Доказательство. Вычислим производной матрицы Якоби с Теоремы 24 используя Лемму 31 об производной матричного умножения, также свойство производной произведения Кронекера 12, также Лемму 37 об производной транспонированной матрицы, Лемму 32 и Теорему 20. 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 24. Используя Лемму 31 и Теорему 19 получаем выражение для блока матрицы Гессе:

$$\begin{aligned} \frac{\partial^2 \mathbf{Z}}{\partial \mathbf{W}_i \partial \mathbf{W}_j} &= (\mathbf{J}_Z \otimes \mathbf{I}_{n_i}) \boldsymbol{\xi}_{ij} + (\mathbf{I}_{Ld_V} \otimes \mathbf{B}_i^\top) \mathbf{H}_Z \mathbf{B}_j, \\ \boldsymbol{\xi}_{ij} &:= \frac{\partial \mathbf{B}_i}{\partial \mathbf{W}_j} \in \mathbb{R}^{(Ld_V \cdot n_i) \times n_j}. \end{aligned}$$

На 2-м шаге вычисляем размерности и вид матриц \mathbf{B}_i . Используя результаты Теорем 24 и 20 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. Заметим, что матрица \mathbf{B}_1 не зависит от матрицы \mathbf{W}_1 , а следовательно $\boldsymbol{\xi}_{11} = \mathbf{0}$. Аналогично матрица \mathbf{B}_2 не зависит от матрицы \mathbf{W}_2 , а 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 31, 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 37 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 $\boldsymbol{\xi}_{21}$:

$$\boldsymbol{\xi}_{21} = (\mathbf{I}_{Ld_V} \otimes ((\mathbf{Y} \otimes \mathbf{I}_{d_{ff}})^\top \mathbf{D}_\sigma^\top)) (\mathbf{I}_L \otimes \mathbf{K}_{d_V, L} \otimes \mathbf{I}_{d_{ff}}) (\text{vec}_r(\mathbf{I}_L) \otimes \mathbf{I}_{d_V d_{ff}}) \mathbf{K}_{d_{ff}, d_V}.$$

Перейдем к оценке пар FFN с параметрами слоев внимания для всех $k \in \{K, Q, V\}$. Для матрицы $\mathbf{B}_1 = (\mathbf{I}_L \otimes \mathbf{W}_2^\top) \mathbf{D}_\sigma(\mathbf{Y} \otimes \mathbf{I}_{d_{ff}})$, заметим, что почти всюду матрица $\frac{\partial \mathbf{D}_\sigma}{\partial \mathbf{Y}} = \mathbf{0}$ равняется нулю согласно Теоремы 20, а следовательно только последний множитель зависит от матрицы \mathbf{W}_k . Используя Лемму 31, 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 24 матрица $\frac{\partial \mathbf{Y}}{\partial \mathbf{W}_k} = \mathbf{J}_Y \mathbf{G}_k$. Тогда получаем оценку на матрицу $\boldsymbol{\xi}_{1k}$ следующего вида:

$$\boldsymbol{\xi}_{1k} = ((\mathbf{I}_L \otimes \mathbf{W}_2^\top) \mathbf{D}_\sigma \otimes \mathbf{I}_{n_k}) (\mathbf{I}_L \otimes \mathbf{K}_{d_{ff}, d_V} \otimes \mathbf{I}_{d_{ff}}) (\mathbf{I}_{Ld_V} \otimes \text{vec}_r(\mathbf{I}_{d_{ff}})) (\mathbf{J}_Y \mathbf{G}_k).$$

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:

$$\boldsymbol{\xi}_{2k} = (\mathbf{I}_L \otimes \mathbf{K}_{d_V, d_{ff}} \otimes \mathbf{I}_{d_V}) (\mathbf{I}_{Ld_{ff}} \otimes \text{vec}_r(\mathbf{I}_{d_V})) (\mathbf{D}_\sigma(\mathbf{I}_L \otimes \mathbf{W}_1^\top) \mathbf{J}_Y \mathbf{G}_k),$$

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 20 для выражения $\frac{\partial \sigma(\cdot)}{\partial (\cdot)} = \mathbf{D}_\sigma$, а также согласно Теореме 24 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 Теоре-

мы 20, а следовательно используя Лемму 31 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 условия Теоремы 19 легко получить следующее выражение $\frac{\partial \mathbf{J}_Y}{\partial \mathbf{W}_\ell} = \mathbf{H}_Y \mathbf{G}_\ell$, а 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 единственны нелинейности с потенциально ненулевым вторым дифференциалом является оператор LayerNorm, для которого получены матрицы $\mathbf{H}_Z, \mathbf{H}_Y$ в рамках Теоремы 19 и которые являются симметричными по построению и оператор ReLU, для которого матрица Гессе является нулевой согласно Теоремы 20, а во вторых все другие отображения являются линейными, а следовательно согласно Леммы 31 и свойства производной произведения Кронекера их частные производные являются коммутативными почти всюду. \square

2.3.5. Спектральные оценки матрицы Гессе для трансформера

Теорема 26. Пусть матрица Гессе $\mathbf{H}_{\text{tr}}^{(i,j)}$ описывает матрицу Гессе между (i, j) -м блоком трансформер модели (2.17), где $i, j \in \{1, 2, K, Q, V\}$, $n_i = \dim(\mathbf{W}_i)$. Тогда для каждой пары (i, j) получаем оценку нормы:

$$\|\mathbf{H}_{\text{tr}}^{(i,j)}\|_2 \leq \|\mathbf{J}_Z\|_2 \|\boldsymbol{\xi}_{ij}\|_2 + \|\mathbf{B}_i\|_2 \|\mathbf{H}_Z\|_2 \|\mathbf{B}_j\|_2, \quad (2.18)$$

$$e \partial e \boldsymbol{\xi}_{ij} = \frac{\partial}{\partial \mathbf{W}_j} \left(\frac{\partial \mathbf{S}}{\partial \mathbf{W}_i} \right), \mathbf{B}_i = \frac{\partial \mathbf{S}}{\partial \mathbf{W}_i}.$$

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

$$\|\mathbf{H}_{\text{tr}}\|_2 \leq \sqrt{m_b n_b} \max_{i,j} \left(\frac{2}{L d_V} \left\| \frac{\partial \mathbf{Z}}{\partial \mathbf{W}_i} \right\|_2 \left\| \frac{\partial \mathbf{Z}}{\partial \mathbf{W}_j} \right\|_2 + \|\mathbf{R}_m^{\text{tr}}\|_2 \|\mathbf{H}_{\text{tr}}^{(i,j)}\|_2 \right).$$

Доказательство. Рассмотрим блоки матрицы Гессе (2.17):

$$\mathbf{H}_{\text{tr}}^{(i,j)} = (\mathbf{J}_Z \otimes \mathbf{I}_{n_i}) \boldsymbol{\xi}_{ij} + (\mathbf{I}_{L d_V} \otimes \mathbf{B}_i^\top) \mathbf{H}_Z \mathbf{B}_j.$$

Используя свойства норм матриц 6, 1 и 2 получаем оценку

$$\begin{aligned} \|\mathbf{H}_{\text{tr}}^{(i,j)}\|_2 &\leq \|\mathbf{J}_Z \otimes \mathbf{I}_{n_i}\|_2 \|\boldsymbol{\xi}_{ij}\|_2 + \|\mathbf{I}_{L d_V} \otimes \mathbf{B}_i^\top\|_2 \|\mathbf{H}_Z\|_2 \|\mathbf{B}_j\|_2 = \\ &= \|\mathbf{J}_Z\|_2 \|\boldsymbol{\xi}_{ij}\|_2 + \|\mathbf{B}_i\|_2 \|\mathbf{H}_Z\|_2 \|\mathbf{B}_j\|_2, \end{aligned}$$

описанной в условиях Теоремы (2.18).

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 (2.18). Используя свойства матричных норм 1, 2, 6, 4, 3, а также определение 17 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 17 было доказано, что :

$$\begin{aligned} \left\| \frac{\partial \mathbf{A}}{\partial \mathbf{T}} \right\|_2 &\leq \frac{1}{L}, \\ \|\mathbf{Z}_1\|_2 &= \|(\mathbf{I}_L \otimes \mathbf{X}^\top)(\partial \mathbf{A} / \partial \mathbf{T})(\mathbf{X} \otimes \mathbf{X})\|_2 \leq \|\mathbf{X}\|_2 \frac{1}{L} \|\mathbf{X}\|_2^2 = \frac{1}{L} \|\mathbf{X}\|_2^3, \\ \left\| \frac{\partial^2 \mathbf{A}}{\partial \mathbf{T}^2} \right\|_2 &\leq 6, \\ \|\mathbf{Z}_2\|_2 &\leq \|\mathbf{X}\|_2^5 \left\| \frac{\partial^2 \mathbf{A}}{\partial \mathbf{T}^2} \right\|_2 \leq 6 \|\mathbf{X}\|_2^5, \\ \|\mathbf{A}\|_2 &\leq \sqrt{L L} \|\mathbf{A}\|_{\max} = L, \end{aligned}$$

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 21. 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 25, 20. 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:

$$\|\mathbf{B}_1\|_2 \leq \|\mathbf{I}_L \otimes \mathbf{W}_2^\top\|_2 \|\mathbf{D}_\sigma\|_2 \|\mathbf{Y} \otimes \mathbf{I}_{d_{ff}}\|_2 = \|\mathbf{W}_2\|_2 \|\mathbf{Y}\|_2. \quad (2.19)$$

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:

$$\|\mathbf{B}_k\|_2 \leq \|\mathbf{J}_{SY}\|_2 \|\mathbf{J}_Y\|_2 \|\mathbf{G}_k\|_2. \quad (2.20)$$

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:

$$\begin{aligned}
\|\mathbf{J}_{SY}\|_2 &\leq \|\mathbf{I}_L \otimes \mathbf{W}_2^\top\|_2 \|\mathbf{D}_\sigma\|_2 \|\mathbf{I}_L \otimes \mathbf{W}_1^\top\|_2 + \|\mathbf{I}_L \otimes \mathbf{I}_{d_V}\|_2 = \\
&= \|\mathbf{W}_2\|_2 \|\mathbf{W}_1\|_2 + 1.
\end{aligned} \quad (2.21)$$

Для матриц $\|\mathbf{G}_V\|_2, \|\mathbf{G}_Q\|_2, \|\mathbf{G}_K\|_2$, используя свойства 1, 2 получаем оценки на нормы:

$$\begin{aligned}\|\mathbf{G}_V\|_2 &\leq L\|\mathbf{X}\|_2, \\ \|\mathbf{G}_Q\|_2 &\leq \frac{1}{L\sqrt{d_K}}\|\mathbf{W}_V\|_2\|\mathbf{W}_K\|_2\|\mathbf{X}\|_2^3, \\ \|\mathbf{G}_K\|_2 &\leq \frac{1}{L\sqrt{d_K}}\|\mathbf{W}_V\|_2\|\mathbf{W}_Q\|_2\|\mathbf{X}\|_2^3.\end{aligned}\tag{2.22}$$

Для оценки матричных норм $\|\boldsymbol{\xi}_{ij}\|_2$, рассмотрим чему они равны из определения в Теореме 25. 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:

$$\begin{aligned}\|\boldsymbol{\xi}_{11}\|_2 &= 0, \\ \|\boldsymbol{\xi}_{22}\|_2 &= 0, \\ \|\boldsymbol{\xi}_{12}\|_2 &\leq \|\mathbf{I}_L \otimes \mathbf{K}_{d_V, d_{ff}} \otimes \mathbf{I}_{d_V}\|_2 \|\mathbf{I}_{Ld_{ff}} \otimes \text{vec}_r(\mathbf{I}_{d_V})\|_2 \|\mathbf{D}_\sigma\|_2 \|\mathbf{Y} \otimes \mathbf{I}_{d_{ff}}\|_2 \\ &= 1 \cdot \|\text{vec}_r(\mathbf{I}_{d_V})\|_2 \cdot 1 \cdot \|\mathbf{Y}\|_2 = \sqrt{d_V} \|\mathbf{Y}\|_2, \\ \|\boldsymbol{\xi}_{21}\|_2 &\leq \|\mathbf{I}_L \otimes \mathbf{W}_2^\top\|_2 \|\mathbf{D}_\sigma\|_2 \|\mathbf{I}_L \otimes \mathbf{K}_{d_{ff}, d_V} \otimes \mathbf{I}_{d_{ff}}\|_2 \|\mathbf{I}_{Ld_V} \otimes \text{vec}_r(\mathbf{I}_{d_{ff}})\|_2 \\ &= \|\mathbf{W}_2\|_2 \cdot 1 \cdot 1 \cdot \|\text{vec}_r(\mathbf{I}_{d_{ff}})\|_2 = \sqrt{d_{ff}} \|\mathbf{W}_2\|_2.\end{aligned}\tag{2.23}$$

В случае пар FFN с параметрами слоев внимания для всех $k \in \{K, Q, V\}$ получаем оценки:

$$\begin{aligned}\|\boldsymbol{\xi}_{1k}\|_2 &\leq \|(\mathbf{I}_L \otimes \mathbf{W}_2^\top) \mathbf{D}_\sigma \otimes \mathbf{I}_{n_k}\|_2 \|\mathbf{I}_L \otimes \mathbf{K}_{d_{ff}, d_V} \otimes \mathbf{I}_{d_{ff}}\|_2 \cdot \\ &\quad \cdot \|\mathbf{I}_{Ld_V} \otimes \text{vec}_r(\mathbf{I}_{d_{ff}})\|_2 \|\mathbf{J}_Y\|_2 \|\mathbf{G}_k\|_2 \\ &\leq \|\mathbf{W}_2\|_2 \cdot 1 \cdot 1 \cdot \sqrt{d_{ff}} \cdot \|\mathbf{J}_Y\|_2 \|\mathbf{G}_k\|_2 = \\ &= \sqrt{d_{ff}} \|\mathbf{W}_2\|_2 \|\mathbf{J}_Y\|_2 \|\mathbf{G}_k\|_2, \\ \|\boldsymbol{\xi}_{2k}\|_2 &\leq \|\mathbf{I}_L \otimes \mathbf{K}_{d_V, d_{ff}} \otimes \mathbf{I}_{d_V}\|_2 \|\mathbf{I}_{Ld_{ff}} \otimes \text{vec}_r(\mathbf{I}_{d_V})\|_2 \|\mathbf{D}_\sigma\|_2 \|\mathbf{I}_L \otimes \mathbf{W}_1^\top\|_2 \cdot \\ &\quad \cdot \|\mathbf{J}_Y\|_2 \|\mathbf{G}_k\|_2 \\ &\leq 1 \cdot \sqrt{d_V} \cdot 1 \cdot \|\mathbf{W}_1\|_2 \cdot \|\mathbf{J}_Y\|_2 \cdot \|\mathbf{G}_k\|_2 = \\ &= \sqrt{d_V} \|\mathbf{W}_1\|_2 \|\mathbf{J}_Y\|_2 \|\mathbf{G}_k\|_2.\end{aligned}\tag{2.24}$$

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:

$$\begin{aligned} \|\boldsymbol{\xi}_{k\ell}\|_2 &\leq \|\mathbf{J}_{SY}\|_2 \left(\|\mathbf{I}_{Ld_V} \otimes \mathbf{G}_k^\top\|_2 \|\mathbf{H}_Y\|_2 \|\mathbf{G}_\ell\|_2 + \|\mathbf{J}_Y\|_2 \|\boldsymbol{\Phi}_{k\ell}\|_2 \right) = \quad (2.25) \\ &= \|\mathbf{J}_{SY}\|_2 \left(\|\mathbf{G}_k\|_2 \|\mathbf{H}_Y\|_2 \|\mathbf{G}_\ell\|_2 + \|\mathbf{J}_Y\|_2 \|\boldsymbol{\Phi}_{k\ell}\|_2 \right). \end{aligned}$$

Итого собирая все части выражения (2.18), используя для каждой пары (i, j) оценки норм матриц $\|\boldsymbol{\xi}_{ij}\|_2$ с выражений (2.23),(2.24),(2.25), а также оценки норм матриц $\|\mathbf{B}_i\|_2$ с выражений (2.19),(2.20),(2.21),(2.22) и подставляя в выражение (2.18) получаем следующие оценки норм на все блоки матрицы Гесе:

$$\begin{aligned} \|\mathbf{H}_{\text{tr}}^{(1,1)}\|_2 &\leq \|\mathbf{J}_Z\|_2 \cdot 0 + \|\mathbf{B}_1\|_2^2 \|\mathbf{H}_Z\|_2 \leq \\ &\leq \|\mathbf{H}_Z\|_2 (\|\mathbf{W}_2\|_2 \|\mathbf{Y}\|_2)^2, \\ \|\mathbf{H}_{\text{tr}}^{(1,2)}\|_2 &\leq \|\mathbf{J}_Z\|_2 \sqrt{d_V} \|\mathbf{Y}\|_2 + \|\mathbf{H}_Z\|_2 (\|\mathbf{W}_2\|_2 \|\mathbf{Y}\|_2) \|\sigma(\mathbf{Y}\mathbf{W}_1)\|_2, \\ \|\mathbf{H}_{\text{tr}}^{(1,k)}\|_2 &\leq \|\mathbf{J}_Z\|_2 \sqrt{d_{ff}} \|\mathbf{W}_2\|_2 \|\mathbf{J}_Y\|_2 \|\mathbf{G}_k\|_2 + \\ &\quad + \|\mathbf{H}_Z\|_2 (\|\mathbf{W}_2\|_2 \|\mathbf{Y}\|_2) (\|\mathbf{J}_{SY}\|_2 \|\mathbf{J}_Y\|_2 \|\mathbf{G}_k\|_2), \\ \|\mathbf{H}_{\text{tr}}^{(k,\ell)}\|_2 &\leq \|\mathbf{J}_Z\|_2 \|\mathbf{J}_{SY}\|_2 \left(\|\mathbf{G}_k\|_2 \|\mathbf{H}_Y\|_2 \|\mathbf{G}_\ell\|_2 + \|\mathbf{J}_Y\|_2 \|\boldsymbol{\Phi}_{k\ell}\|_2 \right) + \\ &\quad + \|\mathbf{H}_Z\|_2 (\|\mathbf{J}_{SY}\|_2 \|\mathbf{J}_Y\|_2 \|\mathbf{G}_k\|_2) (\|\mathbf{J}_{SY}\|_2 \|\mathbf{J}_Y\|_2 \|\mathbf{G}_\ell\|_2), \end{aligned}$$

В оценке матричной нормы $\|\mathbf{Y}\|_2$ and $\|\mathbf{S}\|_2$ были использованы результаты Леммы 22. Нормы матриц $\mathbf{H}_Z, \mathbf{H}_Y$ оцениваются в рамках Леммы 23. \square

2.4. Результаты вычислительных экспериментов

Данный раздел носит больше теоретический характер по оценке норм матриц Гессе для различных классов моделей глубокого обучения.

В рамках вычислительного эксперимента проанализируем вид матриц Гессе для моделей глубокого обучения на базе архитектуры трансформер.

В этой части мы используем один блок Transformer, который мы обучаем на наборе данных MNIST [26]. Сначала мы подаем всего один батч из обучающего загрузчика данных в инициализированную модель и вычисляем точный гессиан с использованием пакета `curvlinops` для эффективного вычисления линейного оператора гессиана. Визуализируя его в логарифмическом масштабе, на Рис. 2.1 мы подчеркиваем неоднородность в величинах элементов матрицы.

набор данных	размер патча	скрытая размерность	размер FFN	количество блоков
MNIST	4	16	64	1
CIFAR-100	4	128	512	8

Таблица 2.1: Гиперпараметры архитектур Vision Transformer (ViT), используемые в наших экспериментах



Рис. 2.1: Визуализация элементов гессиана для инициализированной модели с одним блоком Трансформера. Мы наблюдаем общую неоднородность величин, при этом блоки, соответствующие **Values**, имеют большие значения.

Мы обучаем модель в течение нескольких эпох, достигая достаточно высокой точности на валидационном наборе данных ($>50\%$), а затем снова визуализируем элементы матрицы Гессе, как показано на Рис. 2.2. На Рис. 2.2 видно, что каждый из блоков гессиана приобретает большую величину, однако блок Values-Values демонстрирует наибольшие значения.



Рис. 2.2: Визуализация элементов гессиана для **модели, обученной в течение нескольких эпох**, с одним блоком Трансформера. Мы наблюдаем общую неоднородность величин, при этом блок, соответствующий Values-Values, имеет наибольшие значения.

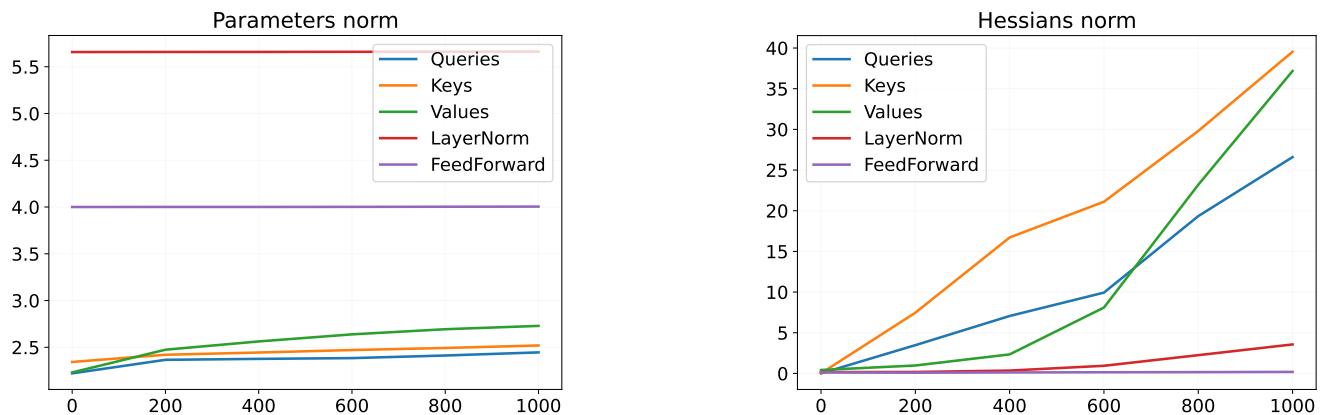


Рис. 2.3: Нормы блоков параметров и нормы их гессианов, рассчитанные точно на одном батче, содержащем 128 примеров из обучающего набора данных MNIST.

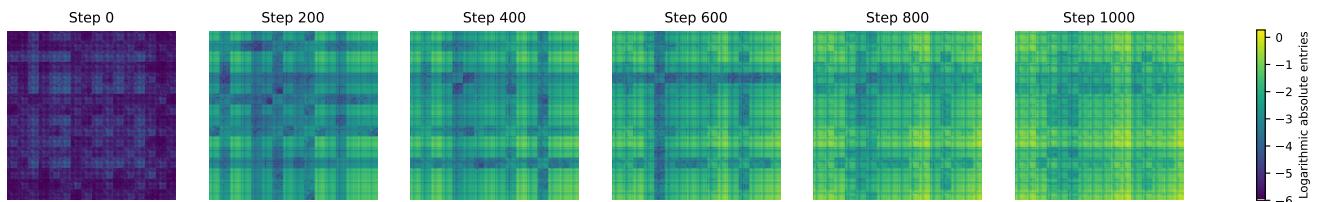


Рис. 2.4: Визуализация элементов (Queries).

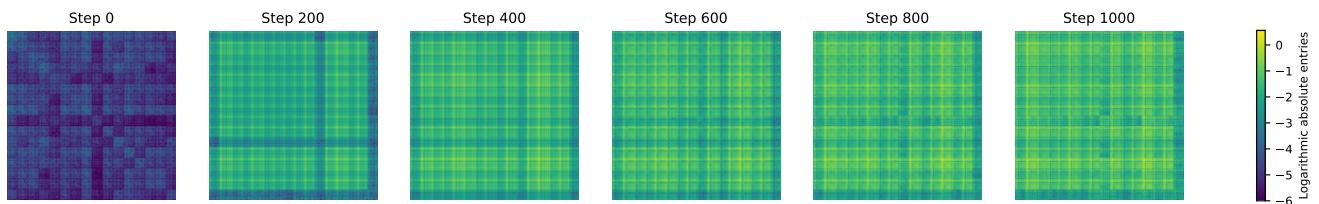


Рис. 2.5: Визуализация элементов (Keys).

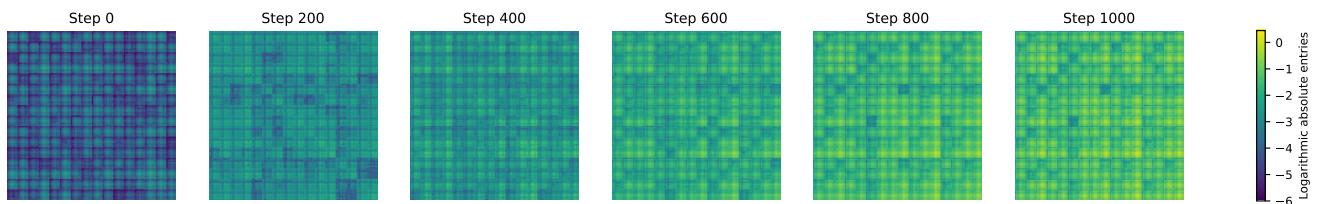


Рис. 2.6: Визуализация элементов (Values).

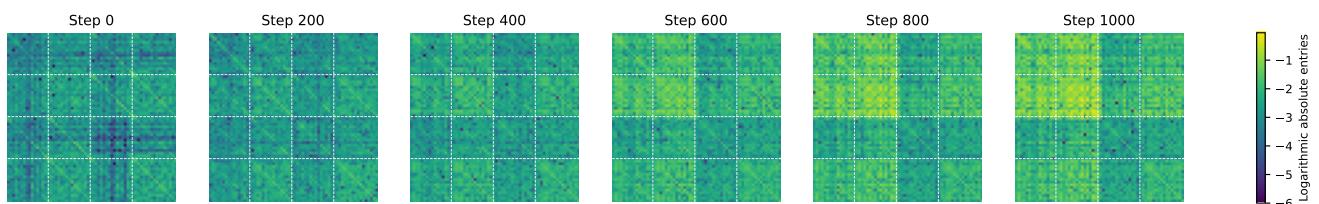


Рис. 2.7: Визуализация элементов (LayerNorm).

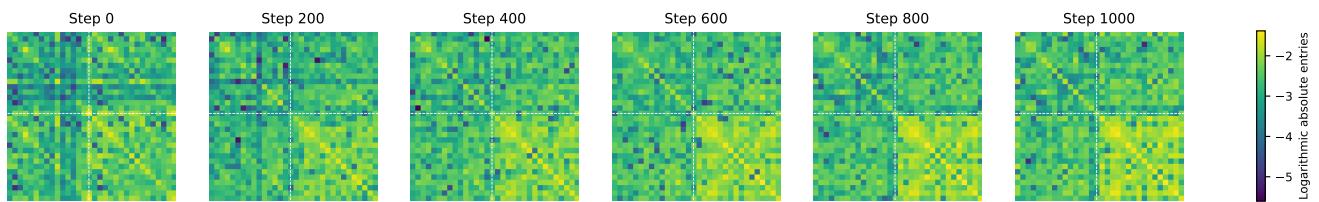


Рис. 2.8: Визуализация элементов (FeedForward).

Данный эксперимент наглядно показывает, как организована матрица Гессе Трансформера в целом, что позволяет нам исследовать каждую его часть отдельно. Далее продолжается предыдущий эксперимент, детализируя изменения элементов для отдельных блоков параметров, что показано на Рис. 2.4, 2.5, 2.6, 2.7, 2.8.

Далее мы вычисляем нормы матриц и нормы их гессианов и отображаем их

на Рис. 2.3. Результаты показывают, что наибольшая величина соответствует блокам Keys и Values, в то время как остальные блоки демонстрируют значительно меньшие абсолютные значения элементов.

2.5. Заключение по главе

Проведенное во второй 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.

Для полносвязных нейронных сетей установлены точные верхние оценки спектральной нормы матрицы Гессе, демонстрирующие экспоненциальную зависимость от глубины сети и полиномиальную зависимость от ширины слоев. Разработанный метод матричной факторизации гессиана позволил унифицировать анализ для широкого класса моделей, включая сверточные архитектуры. Для одномерных и двумерных сверточных сетей получены явные оценки нормы гессиана через структурные параметры архитектуры, а также исследовано регуляризующее влияние операций пулинга на сложность оптимизационного ландшафта.

Наиболее значительным результатом главы является комплексный анализ матриц Гессе для трансформерных архитектур. Впервые получены явные выражения для матриц Якоби и Гессе ключевых компонентов трансформера: механизма самовнимания, слоя LayerNorm и блока FFN. Теоретически установлена гетерогенность вклада различных параметров в общую кривизну функции потерь.

Теоретическая значимость полученных результатов заключается в установлении прямой связи между спектральными свойствами матриц Гессе и введен-

ной в первой главе ландшафтной мерой сложности моделей. Практическая ценность работы состоит в создании аппарата для сравнительного анализа архитектур и прогнозирования их поведения при масштабировании.

Перспективными направлениями дальнейших исследований являются уточнение оценок за счет учета стохастичности и спектральных распределений, а также разработка методов оптимизации и регуляризации на основе полученных теоретических результатов.

Глава 3

Достаточный объем выборки моделей

В главе 1 было показано, что частным случаем условной меры сложности данных является достаточный размер выборки. В рамках текущей главы рассмотрим различные методы определения достаточного размера выборки для различных моделей, от линейных до моделей глубокого обучения.

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 объема выборки, такие как статистические, байесовские и эвристические методы.

Статистические методы предполагают, что выборка соответствует некоторым предварительным условиям, сформулированным ранее. Эти условия сформулированы как статистический критерий [14, 15, 16, 6]. 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 [9]. Эти модели выбираются в соответствии с постановкой задачи регрессии или классификации. В этой статье представлены обобщенные линейные модели. В статье [15] предложен подход к оценке мощности и размера выборки, связанной с ней, на основе теста отношения максимального правдоподобия. Этот подход оказался более точным для ряда независимых переменных. Кроме того, в статье [17] предложен метод оценки мощности для статистики Вальда. В статье [11] в случае логистической регрессии предлагается использовать метод, использующий кривую ROC-AUC

и концепцию сдвига. Классические методы [14, 15, 16, 17, 6] имеют ряд ограничений, связанных с практическим применением этих методов. Чтобы оценить размер выборки, необходимо знать дисперсию оценки параметра или, в более общем случае, иметь оценку параметра нецентральности в распределении статистики, используемой, когда альтернативная гипотеза верна. Эти методы не показывают как получить эти значения. Кроме того, дисперсия оценки и параметр нецентральности не будут получены с определенной дисперсией, влияние которой на результат оценки размера выборки не имеет значения.

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, можно использовать методы, основанные 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 [11], определяется достаточный размер выборки, который соответствует началу переобучения. В статье [12] для оценки достаточного размера выборки используется процедуру бутстррап. Превышение текущего размера выборки проверяется на основе анализа доверительных интервалов оцениваемого параметра. Ширина доверительного интервала с различными 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 [10, 13, 18], где оценка размера выборки определяется на основе максимизации ожидаемого значение некоторой функции качества [10]. Функция качества может включать в себя

явные функции распределения параметров и штрафы за увеличение размера выборки. Альтернативой подходам [18], основанным на функции качества, является выборка размера выборки путем установления ограничений на определенный критерий качества оценки параметров модели. Примеры критериев: критерий средней апостериорной дисперсии (AVPC), критерий средней длины (ALC), критерий 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 статье [11] 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.

3.1. Статистические методы определения достаточного размера выборки

Задана выборка размера m :

$$\mathfrak{D}_m = \{\mathbf{x}_i, y_i\}_{i=1}^m,$$

где $\mathbf{x}_i \in \mathbb{R}^n$, $y_i \in \mathbb{Y}$. Вектор признаков $\mathbf{x} = [\mathbf{u}, \mathbf{v}]$ соединяет $\mathbf{u}_i \in \mathbb{R}^k$ and $\mathbf{v}_i \in \mathbb{R}^{n-k}$. Выборка \mathfrak{D}_m случайным образом делится на обучающую и тестовую части:

$$\mathfrak{D}_{\mathcal{T}_m} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i \in \mathcal{T}_m}, \quad \mathfrak{D}_{\mathcal{L}_m} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i \in \mathcal{L}_m}, \quad \mathcal{T}_m \sqcup \mathcal{L}_m = \{1, \dots, m\}.$$

Введем 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 правдоподобия выборки $\mathfrak{D}_{\mathcal{L}}$ с заданным вектором 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 имеет вид:

$$\mathbf{I}(\mathfrak{D}, \mathbf{w}) = -\nabla \nabla^T l(\mathfrak{D}, \mathbf{w}), \quad \mathbf{V} = \mathbf{I}^{-1}(\mathfrak{D}, \mathbf{m}), \quad (3.1)$$

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 выборки. Они позволяют прогнозировать необходимое число образцов на ранней стадии эксперимента.

Плотность распределения целевой переменной

$$p(y|\mathbf{u}, \mathbf{v}, \mathbf{w}_u, \mathbf{w}_v) = \exp(y\theta - b(\theta) + c(y)),$$

где θ является параметром распределения, полученный с помощью функции связи $\theta = \theta(\mathbf{u}, \mathbf{v}, \mathbf{w}_u, \mathbf{w}_v)$.

Тестируемая гипотеза

$$H_0 : \mathbf{m}_u = \mathbf{m}_u^0, \quad H_1 : \mathbf{m}_u \neq \mathbf{m}_u^0.$$

Пусть статистики $S_{m,u}(\mathbf{w}_u, \mathbf{w}_v)$ и $S_{m,v}(\mathbf{w}_u, \mathbf{w}_v)$ являются производными логарифма правдоподобия выборки \mathfrak{D}_m в точках \mathbf{w}_u и \mathbf{w}_v . 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 Лагранджа равняется

$$LM = \mathbf{s}_m^\top \mathbf{Q}_m^{-1} \mathbf{s}_m.$$

где \mathbf{Q}_m ковариационная матрица вектора \mathbf{s}_m .

В случае истинности гипотезы H_0 статистика LM асимптотически имеет распределения $\chi^2(k)$. В [14] показано, что при альтернативной гипотезе H_1 статистика LM асимптотически имеет распределения $\chi^2(k, \gamma)$, где γ является параметром нецентральности

$$\gamma = \boldsymbol{\xi}_m^\top \boldsymbol{\Sigma}_m^{-1} \boldsymbol{\xi}_m = m \boldsymbol{\xi}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\xi} = m\gamma^0, \quad (3.2)$$

где $\boldsymbol{\xi}_m$ и $\boldsymbol{\Sigma}_m$ матрицы математического ожидания и 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 :

$$\gamma^* : \chi_{k,1-\alpha}^2 = \chi_{k,\beta}^2(\gamma). \quad (3.3)$$

Используя (3.2) и (3.3) получаем

$$m^* = \frac{\gamma^*}{\gamma^0}.$$

Это достаточный минимальный размер выборки, чтобы различить вектор \mathbf{m}_u от \mathbf{m}_u^0 .

Пусть правдоподобие выборки задается выражением

$$p(y|\mathbf{u}, \mathbf{v}, \mathbf{w}_u, \mathbf{w}_v) = \exp \left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right), \quad (3.4)$$

где θ является параметром распределения, который вычисляются с помощью функции связи $\theta = \theta(\mathbf{u}, \mathbf{v}, \mathbf{w}_u, \mathbf{w}_v)$.

Проверяемая гипотеза

$$H_0 : \mathbf{m}_u = \mathbf{m}_u^0, \quad H_1 : \mathbf{m}_u \neq \mathbf{m}_u^0.$$

Введем логарифм статистики отношения правдоподобий:

$$LR = 2 \left(l(\mathfrak{D}, \hat{\mathbf{w}}) - l(\mathfrak{D}, \hat{\mathbf{w}}^0) \right),$$

где $\hat{\mathbf{w}} = [\hat{\mathbf{w}}_u, \hat{\mathbf{w}}_v]$ является вектором, который максимизирует правдоподобие (3.4), $\hat{\mathbf{w}}^0 = [\mathbf{m}_u^0, \hat{\mathbf{w}}_v^0]$ является вектором, который максимизирует правдоподобие (3.4) с фиксируемым подвектором параметров \mathbf{m}_u^0 .

В случае истинности гипотезы H_0 статистика LR асимптотически имеет распределения $\chi^2(k)$. В [16] показано, что при альтернативной гипотезе H_1 статистика LR асимптотически имеет распределения $\chi^2(k, \gamma)$, 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 $\chi_{k,1-\alpha}^2$, $\chi_{k,\beta}^2(\gamma^*)$ квантили распределений χ_k^2 and $\chi_k^2(\gamma^*)$. Правдоподобие выборки:

$$p(y|\mathbf{u}, \mathbf{v}, \mathbf{w}_u, \mathbf{w}_v) = \exp \left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right), \quad (3.5)$$

где θ является параметром распределения, который вычисляются 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:

$$W = (\hat{\mathbf{w}}_u - \mathbf{m}_u^0)^\top \hat{\mathbf{V}}_u^{-1} (\hat{\mathbf{w}}_u - \mathbf{m}_u^0),$$

где $\hat{\mathbf{w}} = [\hat{\mathbf{w}}_u, \hat{\mathbf{w}}_v]$ вектор параметров, который максимизирует правдоподобие выборки (3.5), где матрица $\hat{\mathbf{V}}_u$ задается в выражении (3.1).

В случае истинности гипотезы H_0 статистика Вальда W асимптотически имеет распределение χ^2 . В [17] показано, что в случае истинности альтернативной гипотезы H_1 статистика Вальда W асимптотически имеет распределение $\chi^2(k, \gamma)$ с параметром нецентральности γ :

$$\gamma = m\delta, \quad \delta = (\hat{\mathbf{w}}_u - \mathbf{m}_u^0)^\top \Sigma_u^{-1} (\hat{\mathbf{w}}_u - \mathbf{m}_u^0), \quad \Sigma_u = m\hat{\mathbf{V}}_u.$$

Использую 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 $\chi_{k,1-\alpha^*}^2$, $\chi_{k,\beta}^2(\gamma^*)$ квантили распределения, а параметр α^* это поправка 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 $\mathbf{w}^* = [\mathbf{m}_u^0, \mathbf{w}_v^*]$ является решением уравнения

$$\lim_{m \rightarrow \infty} m^{-1} \mathbb{E} \left(\frac{\partial l(\mathfrak{D}, [\mathbf{m}_u^0, \mathbf{w}_v])}{\partial \mathbf{w}_v} \right) = 0.$$

3.2. Эвристические методы определения достаточного размера выборки

В методе, основанном на эвристике, используются популярные статистические эвристики, такие как бутстррап, перекрестная проверка и задание функции по-

лезности. Введем 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 $Z_{1-\beta}$ являются квантилями распределения $\mathcal{N}(0, 1)$.

Данный метод не рассматривается далее, поскольку его можно использовать только в задаче логистической регрессии.

Рассмотрим метод на основе кроссвалидации. Определим критерий переобучения как

$$RS(m) = \ln \frac{L(\mathfrak{D}_{\mathcal{L}(m)}, \hat{\mathbf{w}})}{L(\mathfrak{D}_{\mathcal{T}(m)}, \hat{\mathbf{w}})}, \quad \frac{|\mathcal{T}(m)|}{|\mathcal{L}(m)|} = \text{const} \leq 0.5.$$

Заметим, что

$$\lim_{m \rightarrow \infty} RS(m) \rightarrow 0.$$

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:

$$m^* : \forall m \geq m^* \max_i (b_i^m - a_i^m) < l.$$

Важно, что этот метод является покоординатным и следовательно для повышения точности прогноза требуется значительное увеличение размера выборки.

3.3. Байесовские методы определения достаточного размера

Байесовские методы оценки размера выборки основаны 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:

$$\forall m \geq m^* E_{\mathfrak{D}_m} D [\hat{\mathbf{w}} | \mathfrak{D}_m] \leq l.$$

где l некоторый заданный экспертизой 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^* определя-

ется критерием среднего покрытия:

$$\forall m \geq m^* \mathbb{E}_{\mathfrak{D}_m} P \{ \mathbf{w} \in A(\mathfrak{D}_m) \} \geq 1 - \alpha, \quad (3.6)$$

где α некоторый параметр заданный эксперто.

Определим функцию $A(\mathfrak{D})$:

$$P(A(\mathfrak{D})) = 1 - \alpha.$$

Оценки критерия средней длины m^* заданный в (3.6):

$$\forall m \geq m^* \mathbb{E}_{\mathfrak{D}_m} r_m \leq l,$$

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 функция штрафов 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 [11] показано, что если размер набора выборок $\mathfrak{D}_{\mathcal{B}_1}$ достаточно велик, чем 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 and p_2 апостериорные вероятности вектора параметров \mathbf{w} рассчитаны 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:

$$\forall \mathfrak{D}_{\mathcal{B}_1} : |\mathfrak{D}_{\mathcal{B}_1}| \geq m^* \mathbb{E}_{\mathfrak{D}_{\mathcal{B}_2}} D_{KL}(p_1, p_2) \leq \varepsilon.$$

3.4. Метод определения достаточного размера выборки на основе сэмлирования эмпирической функции ошибки

В данном разделе предполагаем, что выполняется условие $m^* \leq m$. Это означает, что требуется определить, какой объем выборки можно считать достаточным, и что для этого есть достаточно объектов в самой выборке D . Для определения достаточности мы будем использовать функцию правдоподобия. Когда доступно достаточное количество объектов, естественно ожидать, что полученная оценка параметров не будет существенно изменяться от одной реализации выборки к другой [7, 8]. Аналогичное утверждение справедливо и для функции правдоподобия. Таким образом, мы формализуем критерии, позволяющие определить достаточный объем выборки. Критерий определяется в определении 11.

Определение 11. Пусть задано некоторое $\varepsilon > 0$. Размер выборки m^* назовем D -достаточным, если для всех $k \geq m^*$ выполняется условие:

$$D(k) = \mathbb{D}_{\hat{\mathbf{w}}_k} L(\mathfrak{D}_m, \hat{\mathbf{w}}_k) \leq \varepsilon.$$

С другой стороны, при наличии достаточного количества объектов вполне естественно, что при добавлении еще одного объекта к рассмотрению результи-

рующая оценка параметра изменится незначительно, на основе данного свойства получаем определение 12.

Определение 12. Пусть задано некоторое $\varepsilon > 0$. Размер выборки t^* назовем *M-достаточным*, если для всех $k \geq t^*$ выполняется условие:

$$M(k) = |\mathbb{E}_{\hat{\mathbf{w}}_{k+1}} L(\mathfrak{D}_m, \hat{\mathbf{w}}_{k+1}) - \mathbb{E}_{\hat{\mathbf{w}}_k} L(\mathfrak{D}_m, \hat{\mathbf{w}}_k)| \leq \varepsilon.$$

В приведенных выше определениях вместо функции правдоподобия $L(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$ мы можем рассмотреть ее логарифм $l(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$. Предположим, что $\mathbb{W} = \mathbb{R}^n$, информацией Фишера задана матрицей:

$$[\mathcal{I}(\mathbf{w})]_{ij} = -\mathbb{E} \left[\frac{\partial^2 \log p(\mathbf{y}|\mathbf{x}, \mathbf{w})}{\partial w_i \partial w_j} \right],$$

заметим, что известным результатом является асимптотическая нормальность оценки максимального правдоподобия, то есть

$$\sqrt{k} (\hat{\mathbf{w}}_k - \mathbf{w}) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}^{-1}(\mathbf{w})).$$

Сходимость по распределению, вообще говоря, не влечет сходимости моментов случайного вектора. Тем не менее, если предположить последнее, то в некоторых моделях можно доказать корректность нашего предложенного определения M-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 27, которая доказывает сходимость параметров.

Теорема 27. 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 M-достаточного размера выборки корректно. А именно, для любого $\varepsilon > 0$ существует такой t^* , что для всех $k \geq t^*$ выполняется $M(k) \leq \varepsilon$.

Доказательство. Рассмотрим определение M-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

мости:

$$\begin{aligned}
|(\mathbf{m}_k - \mathbf{m}_{k+1})^\top \mathbf{X}^\top \mathbf{X} (\mathbf{m}_k + \mathbf{m}_{k+1})| &\leq \|\mathbf{X}(\mathbf{m}_k - \mathbf{m}_{k+1})\|_2 \|\mathbf{X}(\mathbf{m}_k + \mathbf{m}_{k+1})\|_2 \leq \\
&\leq \|\mathbf{X}\|_2^2 \|\mathbf{m}_k - \mathbf{m}_{k+1}\|_2 \|\mathbf{m}_k + \mathbf{m}_{k+1}\|_2 \leq \\
&\leq C \|\mathbf{X}\|_2^2 \|\mathbf{m}_k - \mathbf{m}_{k+1}\|_2.
\end{aligned}$$

Последнее слагаемое оцениваем, используя неравенство Гельдера для нормы Фробениуса:

$$|\text{tr}(\mathbf{X}^\top \mathbf{X} (\Sigma_k - \Sigma_{k+1}))| \leq \|\mathbf{X}^\top \mathbf{X}\|_F \|\Sigma_k - \Sigma_{k+1}\|_F.$$

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

Следствие 4. 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 M -достаточного размера выборки корректно.

По условию, задана только одна выборка, а следовательно, в эксперименте невозможно вычислить математическое ожидание и дисперсию, указанные в определениях. Поэтому для их оценки используется метод бутстрэпирования, то есть сгенерируем из заданной \mathfrak{D}_m некоторое число B подвыборок размера k с возвращением. Для каждой из них получим оценку параметров $\hat{\mathbf{w}}_k$ и вычислим значение $L(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$. Для оценки используется выборочное среднее и несмещенную выборочную дисперсию.

Предложенные выше определения также могут быть применены в тех задачах, где минимизируется произвольная функция потерь, а не максимизируется функция правдоподобия. Мы не приводим какого-либо теоретического обоснования для этого, но на практике такая эвристика оказывается вполне успешной.

3.5. Метод определения достаточного размера выборки на основе близости апостериорных распределений

В работе [11] предлагается использовать расхождение Кульбака-Лейблера для оценки достаточного размера выборки в задаче бинарной классификации. Идея 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.

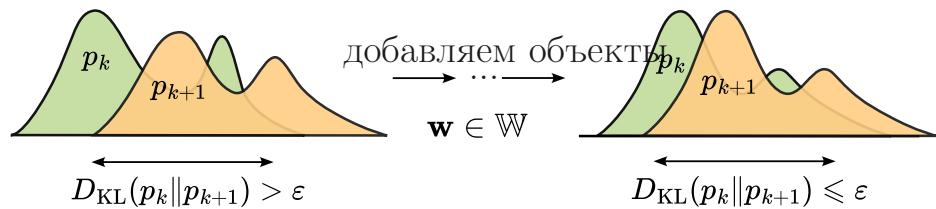


Рис. 3.1: Пример сдвига распределений при добавления объектов

In this paper, the question of the correctness of this approach is considered. The method is studied in an arbitrary probabilistic model. As a measure of proximity, it is proposed to use not only the Kullback-Leibler divergence, but also the s-score similarity function from [47].

В рамках данного раздела предлагается использовать не только расхождение Кульбака-Лейблера, но и функцию схожести s-score из работы [47], 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 13. 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, то есть

$$|\mathcal{I}_1 \Delta \mathcal{I}_2| = |(\mathcal{I}_1 \setminus \mathcal{I}_2) \cup (\mathcal{I}_2 \setminus \mathcal{I}_1)| = 1.$$

Consider two similar subsamples $\mathfrak{D}_k = (\mathbf{X}_k, \mathbf{y}_k)$ and $\mathfrak{D}_{k+1} = (\mathbf{X}_{k+1}, \mathbf{y}_{k+1})$ of sizes k and $k+1$, respectively. This means that the larger one is obtained by adding one element to the smaller one. Let's find the posterior distribution of the model

parameters over these subsamples:

$$p_j(\mathbf{w}) = p(\mathbf{w}|\mathfrak{D}_j) = \frac{p(\mathfrak{D}_j|\mathbf{w})p(\mathbf{w})}{p(\mathfrak{D}_j)} \propto p(\mathfrak{D}_j|\mathbf{w})p(\mathbf{w}), \quad j = k, k+1.$$

Рассмотрим две похожие 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, то есть выборка \mathfrak{D}_{k+1} получена путём добавления одного элемента к выборке \mathfrak{D}_k . Найдем апостериорное распределение параметров модели по этим подвыборкам:

$$p_j(\mathbf{w}) = p(\mathbf{w}|\mathfrak{D}_j) = \frac{p(\mathfrak{D}_j|\mathbf{w})p(\mathbf{w})}{p(\mathfrak{D}_j)} \propto p(\mathfrak{D}_j|\mathbf{w})p(\mathbf{w}), \quad j = k, k+1.$$

Определение 14. Пусть задано некоторое $\varepsilon > 0$. Размер выборки m^* называется **KL-достаточным**, если для всех $k \geq m^*$

$$KL(k) = D_{KL}(p_k \| p_{k+1}) = \int p_k(\mathbf{w}) \log \frac{p_k(\mathbf{w})}{p_{k+1}(\mathbf{w})} d\mathbf{w} \leq \varepsilon.$$

For a pair of normal distributions, the Kullback-Leibler divergence has a fairly simple form. Assume that the posterior distribution is normal, that is, $p_k(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_k, \Sigma_k)$. Guided by the heuristic that the convergence of the moments of such a distribution should entail the proximity of posterior distributions on similar subsamples, the following statement can be formulated.

Для пары нормальных распределений расхождение Кульбака-Лейблера имеет достаточно простой вид, а следовательно предположив, что апостериорное распределение является нормальным, то есть $p_k(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_k, \Sigma_k)$ получаем формулировку теоремы 28

Теорема 28. 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: $\Sigma_{k+1} = \Sigma_k + \Delta\Sigma$, 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

ма $\|\Sigma_{k+1}^{-1}\|_2$ 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: 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 $D_{\text{KL}}(p_k \| p_{k+1}) \rightarrow 0$ 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 28 implies что расстояние между двумя гауссовскими распределениями стремится к нулю по мере сходимости их векторов математических ожиданий и ковариационных матриц. Это позволяет нам глубже изучить вопрос сходимости расхождения Кульбака-Лейблера, анализируя аналитические выражения для математических ожиданий и дисперсий.

Рассмотрим функцию схожести s-score из работы [47] 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 15. Пусть задано некоторое $\varepsilon > 0$. Размер выборки m^* называется **S-достаточным**, если для всех $k \geq m^*$

$$S(k) = s\text{-score}(p_k, p_{k+1}) \geq 1 - \varepsilon.$$

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 29

Теорема 29. 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. Let's use the s-score expression for a pair of normal posterior distributions from [47]:

$$s\text{-score}(p_k, p_{k+1}) = \exp \left(-\frac{1}{2} (\mathbf{m}_{k+1} - \mathbf{m}_k)^\top (\boldsymbol{\Sigma}_k + \boldsymbol{\Sigma}_{k+1})^{-1} (\mathbf{m}_{k+1} - \mathbf{m}_k) \right).$$

Because

$$\left| (\mathbf{m}_{k+1} - \mathbf{m}_k) (\boldsymbol{\Sigma}_k + \boldsymbol{\Sigma}_{k+1})^{-1} (\mathbf{m}_{k+1} - \mathbf{m}_k) \right| \leq \|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2^2 \|(\boldsymbol{\Sigma}_k + \boldsymbol{\Sigma}_{k+1})^{-1}\|_2 \rightarrow 0$$

if $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \rightarrow 0$, then the value of the quadratic form inside the exponent tends to zero. Therefore, $s\text{-score}(p_k, p_{k+1}) \rightarrow 1$ as $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \rightarrow 0$. \square

Значимость теоремы 29 аналогична значимости теоремы 28. По существу, близость нормальных распределений в терминах функции схожести s-score сводится к сходимости их математических ожиданий. Примечательно, что в отличие от теоремы 28, сходимость ковариационных матриц здесь не требуется.

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, мы приходим к одному из простейших примеров модели, для которой справедливы приведенные выше теоремы. Фактически, для линейной регрессии могут быть сформулированы более простые утверждения.

Теорема 30. *Пусть множество значений признаков и целевой переменной ограничены, то есть существует $M \in \mathbb{R}$ такое, что $\|\mathbf{x}\|_2 \leq M$ и $|y| \leq M$. 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:

$$\|\Sigma_{k+1} - \Sigma_k\|_2 \leq \left\| (\alpha\mathbf{I} + \mathbf{A}_{k+1})^{-1} \right\|_2 \left\| (\alpha\mathbf{I} + \mathbf{A}_k)^{-1} \right\|_2 \|\mathbf{A}_{k+1} - \mathbf{A}_k\|_2.$$

Теперь проанализируем каждый из множителей. Спектральная норма обратной матрицы выражается через минимальное собственное значение:

$$\left\| (\alpha\mathbf{I} + \mathbf{A})^{-1} \right\|_2 = \frac{1}{\lambda_{\min}(\alpha\mathbf{I} + \mathbf{A})}.$$

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 мы используем более слабую оценку:

$$\left\| (\alpha\mathbf{I} + \mathbf{A})^{-1} \right\|_2 \leq \frac{1}{\lambda_{\min}(\mathbf{A})}.$$

Таким образом, 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 $\mathbf{x}_{k+1} \mathbf{x}_{k+1}^\top$ имеет единичный ранг, 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:

$$\mathbf{X}_{k+1}^\top \mathbf{X}_{k+1} = \mathbf{X}_k^\top \mathbf{X}_k + \mathbf{x}_{k+1} \mathbf{x}_{k+1}^\top,$$

$$\mathbf{X}_{k+1}^\top \mathbf{y}_{k+1} = \mathbf{X}_k^\top \mathbf{y}_k + \mathbf{x}_{k+1} y_{k+1}.$$

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:

$$\begin{aligned}\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 &\leqslant \frac{M^2}{\lambda_{\min}(\mathbf{X}_k^\top \mathbf{X}_k)} \cdot \frac{kM^2}{\lambda_{\min}(\mathbf{X}_k^\top \mathbf{X}_k)} + \frac{M^2}{\lambda_{\min}(\mathbf{X}_{k+1}^\top \mathbf{X}_{k+1})} \\ &= \frac{kM^4}{\lambda_{\min}^2(\mathbf{X}_k^\top \mathbf{X}_k)} + \frac{M^2}{\lambda_{\min}(\mathbf{X}_{k+1}^\top \mathbf{X}_{k+1})}.\end{aligned}$$

Из условия $\lambda_{\min}(\mathbf{X}_k^\top \mathbf{X}_k) = \omega(\sqrt{k})$ следует:

$$\begin{aligned}\frac{1}{\lambda_{\min}(\mathbf{X}_k^\top \mathbf{X}_k)} &= o\left(\frac{1}{\sqrt{k}}\right), \\ \frac{1}{\lambda_{\min}^2(\mathbf{X}_k^\top \mathbf{X}_k)} &= o\left(\frac{1}{k}\right).\end{aligned}$$

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 30 является ключевой в данном разделе, так как при слабых и понятных предположениях из нее следует сходимость моментов апостериорного распределения параметров. Первое предположение в теореме 30 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 30 представляет больший интерес, поскольку оно углубляется в поведение минимального собственного значения выборочной ковариационной матрицы признаков. К сожалению, в данной работе не приводятся теоретические гарантии для этой сходимости, хотя сходимость проверяется экспериментально.

3.6. Результаты вычислительных экспериментов

В данном разделе описываются результаты вычислительных экспериментов для методов, описанных в данной главе.

3.6.1. Определения достаточного размера выборки на основе статистических методов

Таблица 3.1: Описание выборок для анализа качества определения оптимального размера выборки

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

Проводится эксперимент для анализа свойств методов оценки достаточного размера выборки. Эксперимент состоит из трех частей. В первой части рассматриваются оценки достаточного размера выборки для разных наборов данных с фиксированным набором гиперпараметров различных методов. 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 качестве данных использовались выборки, описанные в таблице 3.1. Методы в строках таблицы 3.2 показывают оценки размера выборки для соответствующих выборок.

В этой части вычислительного эксперимента анализируется сходимость различных методов на разных выборках. В эксперименте используются выборки: Boston Housing [48], Diabetes, Forest Fires, Servo [49], NBA. Результат анализа представлен в таблице 3.2. Символ “–” обозначает, что исходный размер выборки недостаточный для прогноза.

Гиперпараметры каждого метода для всех выборок описаны в таблице 3.3. Поскольку критерии Лагранжа, отношения правдоподобия и Вальда асимпто-

Таблица 3.2: Эксперимент по оценке размера выборки для различных наборов выборок

Методы и наборы данных	Boston Housing	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

тически 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 бутстррап семплируется множество подвыборок размером t . Для разных значений t вычисляется t^* .

Рис. 3.2 демонстрирует зависимость статистических значений каждого метода для разных выборок с фиксированным размером выборки t . Пороговые значения для каждого метода устанавливаются экспериментально, что позволяет контролировать различные статистические характеристики выборки. Рис. 3.2 показывает адекватность различных методов определения достаточного размера выборки. Представленные функции монотонны и асимптотически стремятся к константе. На рис. 3.3 показаны результаты методов на выборках разных размера. Показано различие методов в дисперсии вычисленного t^* . Анализируются различные методы в случае небольшого размера выборки. Все представленные

Таблица 3.3: Экспертные оценки гиперпараметров для разных методов оценки объема выборки

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	—	—

методы сходятся, причем результат предсказания в асимптотике не зависит от доступного размера выборки m .

Небольшое значение дисперсии интерпретируется как вычислительная устойчивость рассмотренных методов. Показано, что некоторые методы не дают оценку достаточного размера выборки, если у них нет соответствующего размера выборки. Это значит, что они не эффективны с точки зрения прогноза, но могут быть использованы для ретроспективы и анализа уже проведенного эксперимента.

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. 3.2 показана зависимость скалярных функций от размера подвыборки. На рис. 3.2 показано, что эти функции монотонны. Тип поведения функции зависит от метода. Изменяя ограничения, установленные

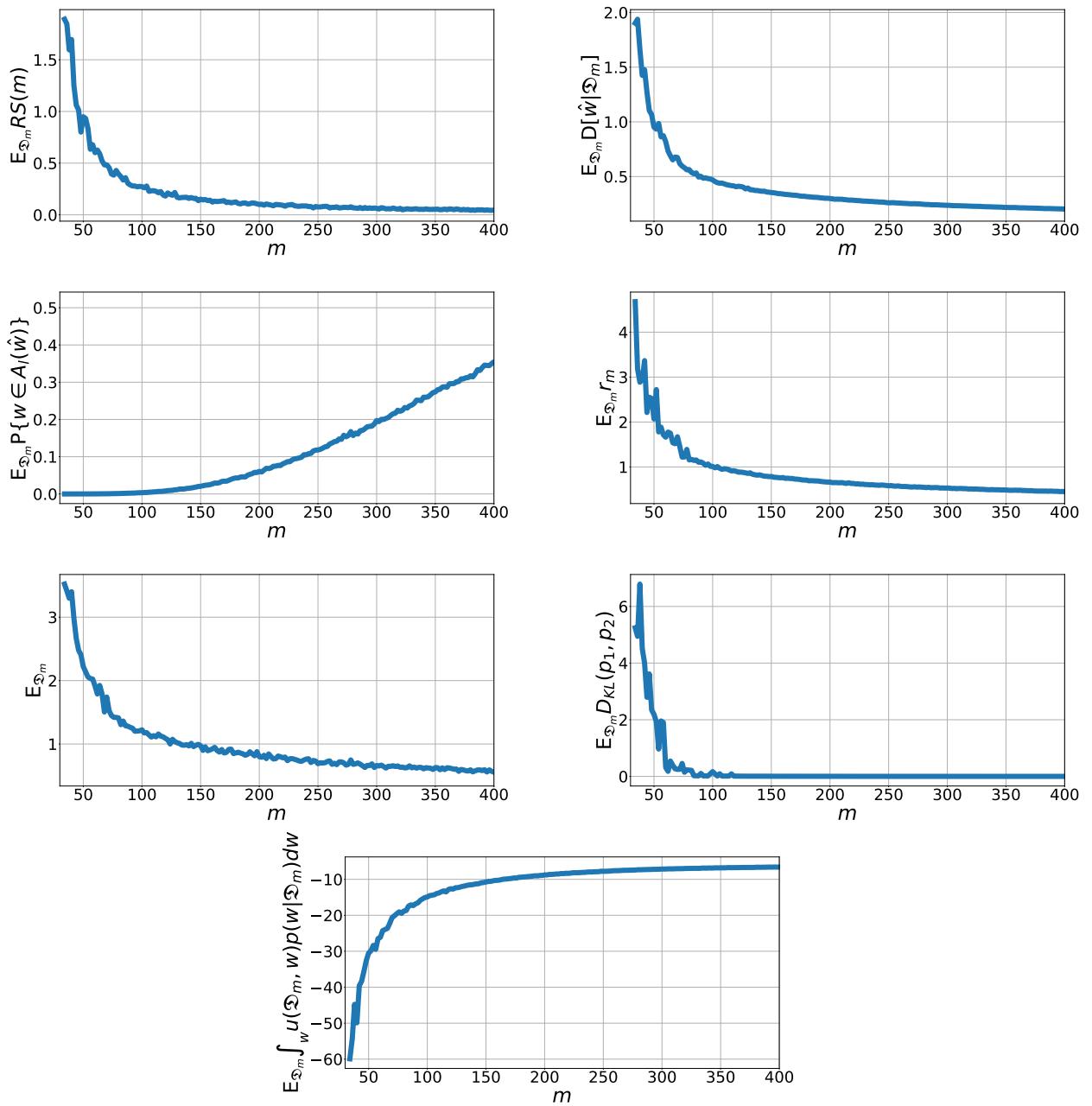


Рис. 3.2: Зависимость статистических значений различных методов

экспертно, можно изменить размер выборки, который будет соответствовать этим ограничениям.

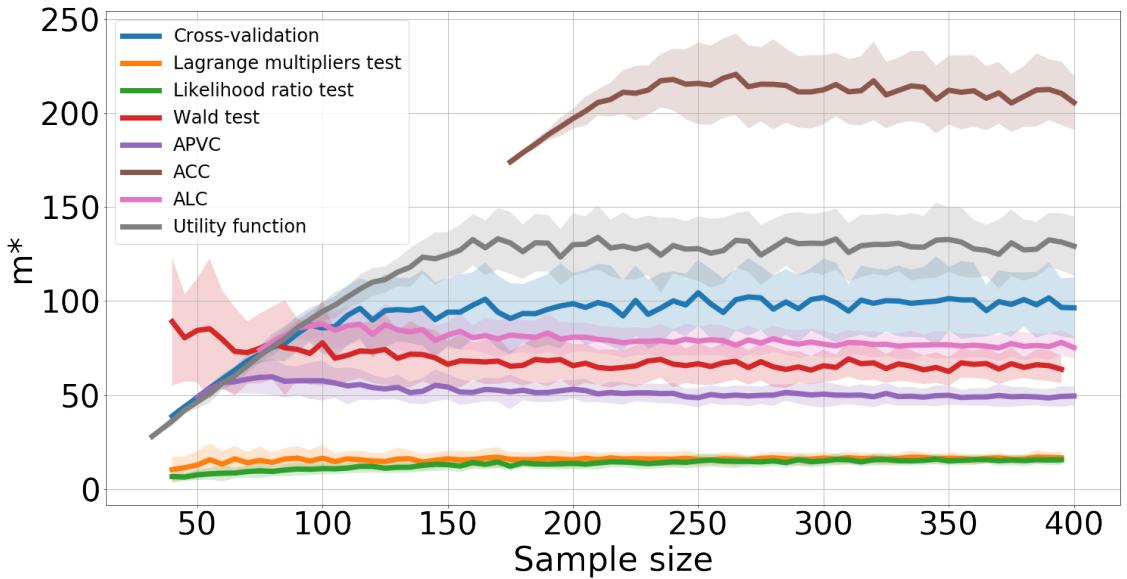


Рис. 3.3: Анализ методов в зависимости от доступного размера выборки

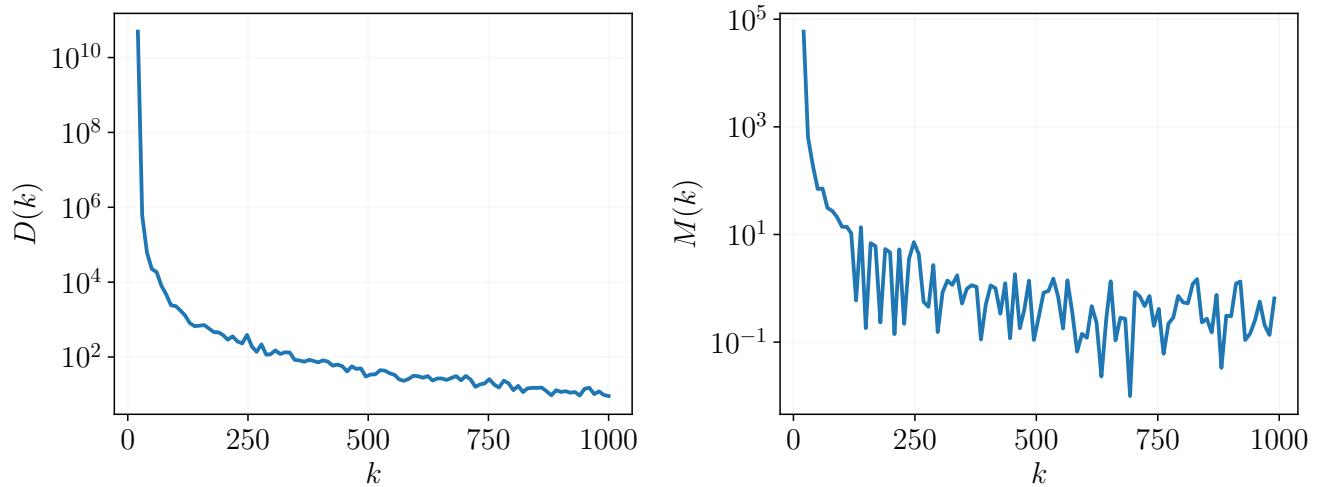


Рис. 3.4: Сходимость предложенных функций $D(k)$ и $M(k)$ для синтетического набора данных регрессии, то есть модели линейной регрессии. Обе функции стремятся к нулю с увеличением размера выборки.

3.6.2. Определение достаточного размера выборки на основе сэмплирования эмпирической функции ошибки

В данном разделе представлено эмпирическое исследование предложенных методов. Эксперименты проводились на синтетических данных и наборе данных Liver Disorders из [50].

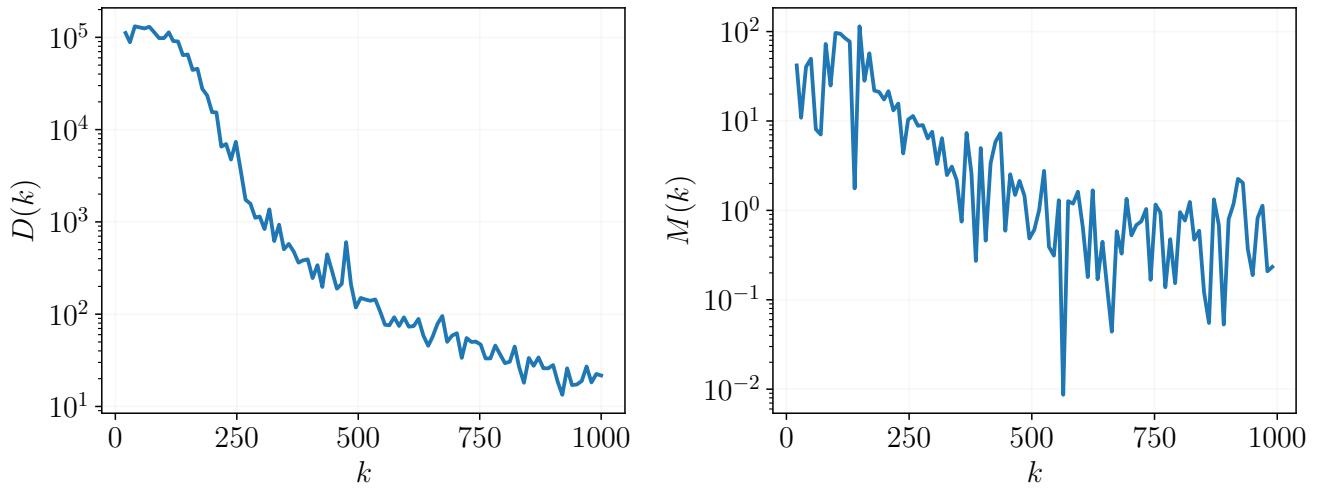


Рис. 3.5: Сходимость предложенных функций $D(k)$ и $M(k)$ для синтетического набора данных классификации, то есть модели логистической регрессии. Обе функции стремятся к нулю с увеличением размера выборки.

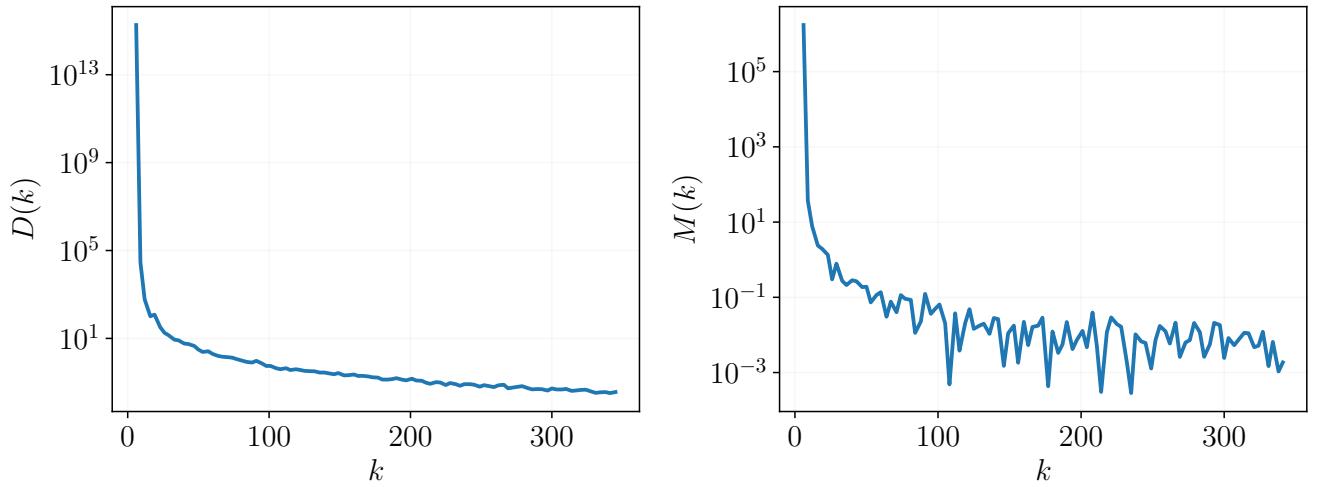


Рис. 3.6: Сходимость предложенных функций $D(k)$ и $M(k)$ для набора данных Liver Disorders. Обе функции стремятся к нулю с увеличением размера выборки.

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 объектов и 5 признаков. Мы также использовали $B = 1000$ подвыборок, полученных методом бутстрэпа, для оценки математического ожидания и дисперсии

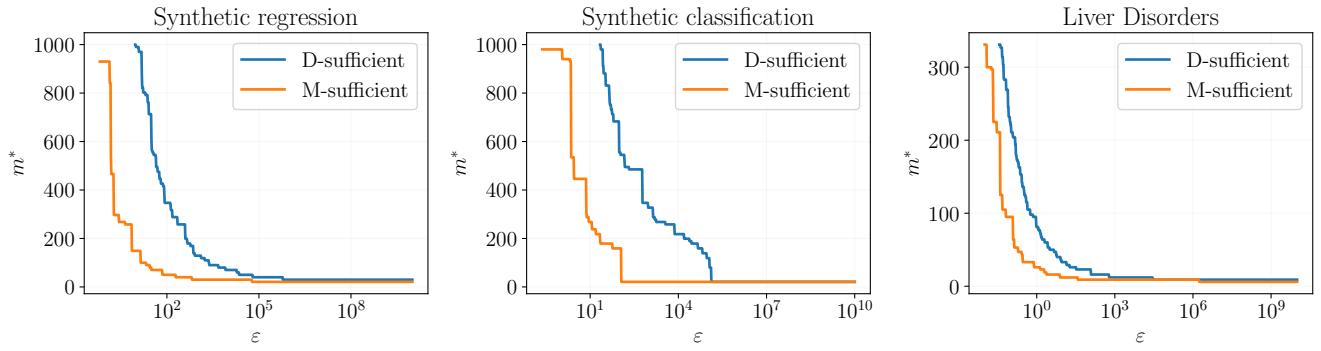


Рис. 3.7: Зависимость достаточного размера выборки от порога для трех наборов данных: синтетическая регрессия, синтетическая классификация и Liver Disorders. С увеличением значения порога ε достаточный размер выборки уменьшается. Это означает, что можно выбрать меньше объектов для достижения желаемых значений предложенных функций $D(k)$ и $M(k)$.

функции потерь.

Рис. 3.4 показывает 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. 3.5. В то же время, на Рис. 3.6 видны графики для набора данных Liver Disorders. Можно заметить, 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^* . Чтобы изучить зависимость между ними, построена зависимость на Рис. 3.7, который показывает, какие размеры выборки можно выбрать для обеспечения определенного уровня достоверности.

Для сравнения производительности предложенных методов на различных наборах данных были выбраны выборки из открытого репозитория [50]. Подробная информация о каждом наборе данных, количестве наблюдений и ко-

Таблица 3.4: Сравнение предложенных методов определения размера выборки: на основе функций $D(k)$ и $M(k)$. Для каждой из предложенных функций пороговое значение ε подбиралось таким образом, чтобы начальное значение функции уменьшалось вдвое. Результаты получены для различных наборов данных с задачей регрессии. Прочерки в таблице означают, что исходный размер выборки недостаточен.

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	—	—

личестве признаков представлена в Таблице 3.4. 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 3.4. Пропуски означают, что исходный размер выборки недостаточен.

3.6.3. Определение достаточного размера выборки на основе близости апостериорных распределений

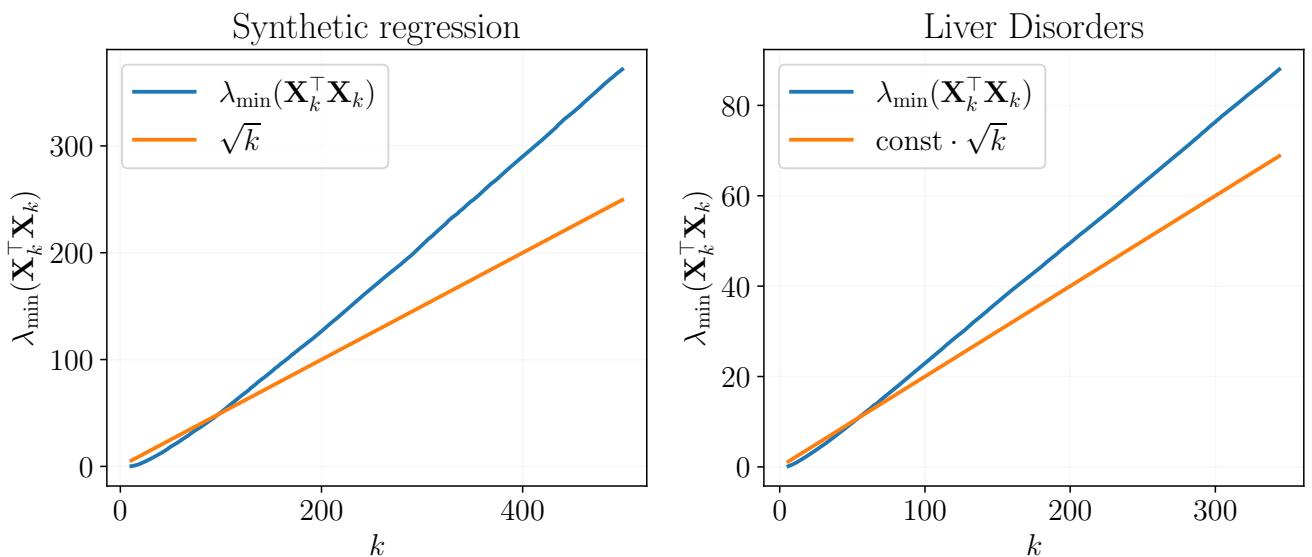


Рис. 3.8: Зависимость минимального собственного значения от размера доступной выборки. Оценочный график, изображенный синим цветом, для больших размеров выборки лежит выше оранжевого графика степенной функции. Такое асимптотическое поведение минимального собственного значения соответствует полученным теоретическим результатам.

Таблица 3.5: Описание выборок, используемых в эксперименте.

Выборка	Количество признаков, n	Количество объектов, m
Boston Housing	14	506
Diabetes	10	576
Forest Fires	13	517
Servo	4	167

В данном разделе представлено расширенное эмпирическое исследование предложенных методов. Эксперименты состоят из нескольких частей. В первой мы проверяем сходимости, полученные в ходе теоретического анализа. Далее

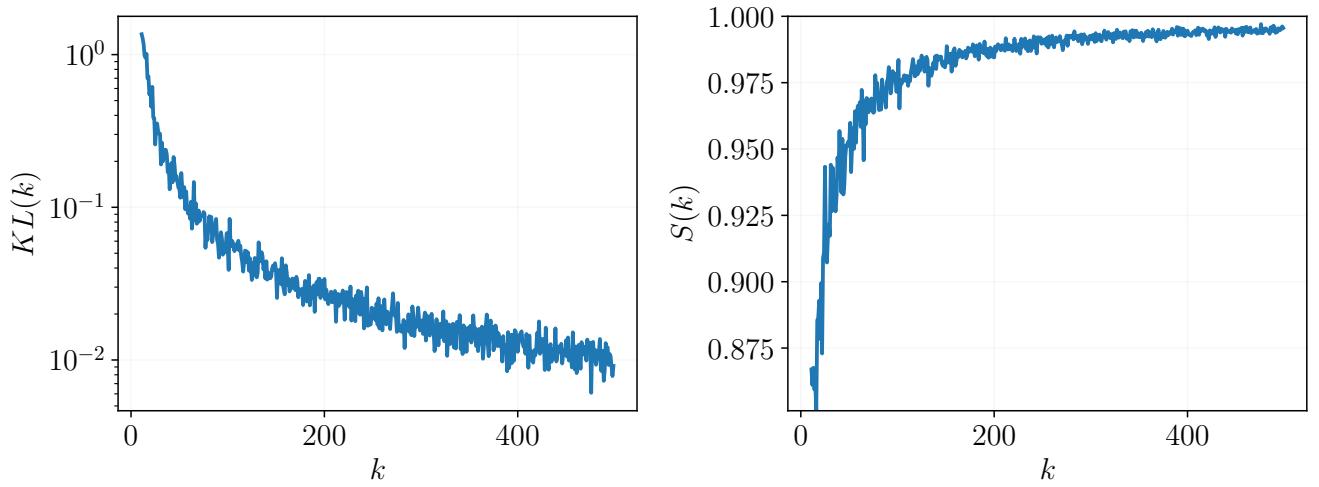


Рис. 3.9: Синтетический набор данных регрессии демонстрирует результаты сходимости предложенных функций оценки размера выборки. График слева соответствует расхождению Кульбака-Лейблера и стремится к нулю, в то время как график справа стремится к единице, отражая поведение функции схожести s-score.

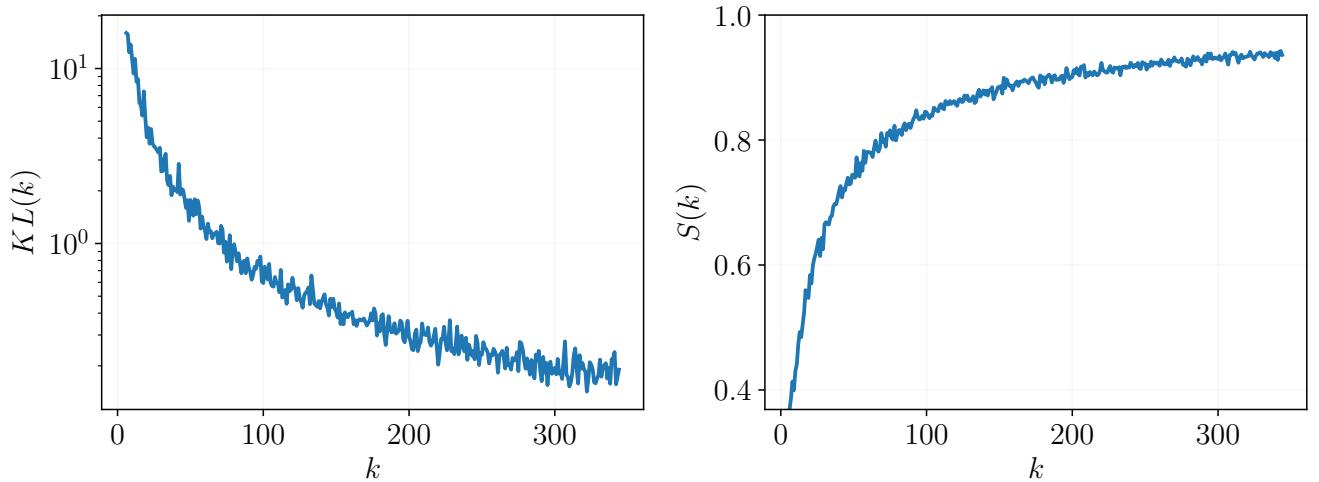


Рис. 3.10: Набор данных Liver Disorders демонстрирует результаты сходимости предложенных функций оценки размера выборки. Слева представлено расхождение Кульбака-Лейблера, которое стремится к нулю с увеличением размера выборки. Справа представлена функция схожести s-score, которая стремится к единице при приближении размера выборки к бесконечности.

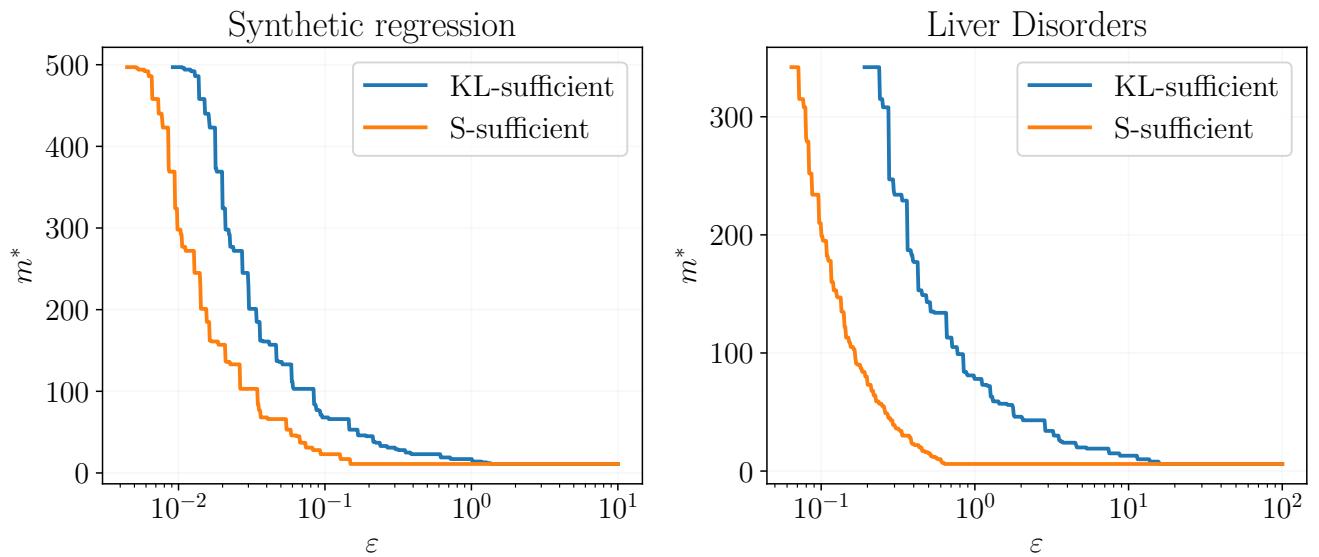


Рис. 3.11: Зависимость достаточного размера выборки от порогового параметра. Для S-достаточного размера выборки требуются более низкие значения порога. Таким образом, он оказывается более требовательным к этому значению.

мы оцениваем размеры выборок для различных наборов данных, используя разные подходы. Наконец, мы изучаем зависимость достаточного размера выборки от объема доступных данных.

Здесь мы исследуем, наблюдаются ли на практике полученные теоретические сходимости. А именно, сначала мы рассматриваем поведение минимального собственного значения матрицы $\mathbf{X}_k^\top \mathbf{X}_k$ при увеличении размера выборки. Затем мы исследуем сходимость предложенных функций $KL(k)$ и $S(k)$. Наконец, изучается зависимость достаточного размера выборки от пороговых параметров. Эксперимент проводится на двух наборах данных: синтетическая регрессия и Liver Disorders.

Синтетические данные генерируются из модели линейной регрессии. Количество объектов составляет 500, количество признаков - 10. Для генерации синтетического набора данных регрессии мы выполнили выборку исходных признаков, параметров модели и шумовых остатков из стандартного нормального распределения. Априорное распределение параметров также было задано как стандартное нормальное, как для синтетической регрессии, так и для набора

Таблица 3.6: Экспериментальная оценка достаточного размера выборки, согласно предложенным методам для различных наборов данных.

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

данных Liver Disorders, который содержит 345 объектов и 5 признаков. Мы предобработали входные признаки, используя стандартный масштабатор (Standard Scaler).

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 мы вычисляли минимальное собственное значение матрицы $\mathbf{X}_k^\top \mathbf{X}_k$. Также вычислялись значения $KL(k)$ и $S(k)$. Этот процесс повторялся $B = 100$ раз.

На Рис. 3.8 показано асимптотическое поведение минимального собственного значения матрицы $\mathbf{X}_k^\top \mathbf{X}_k$. Мы видим, что когда размер выборки стремится к бесконечности, минимальное собственное значение также стремится к бесконечности. При этом, как и требуется для Теоремы 30, график лежит выше \sqrt{k} .

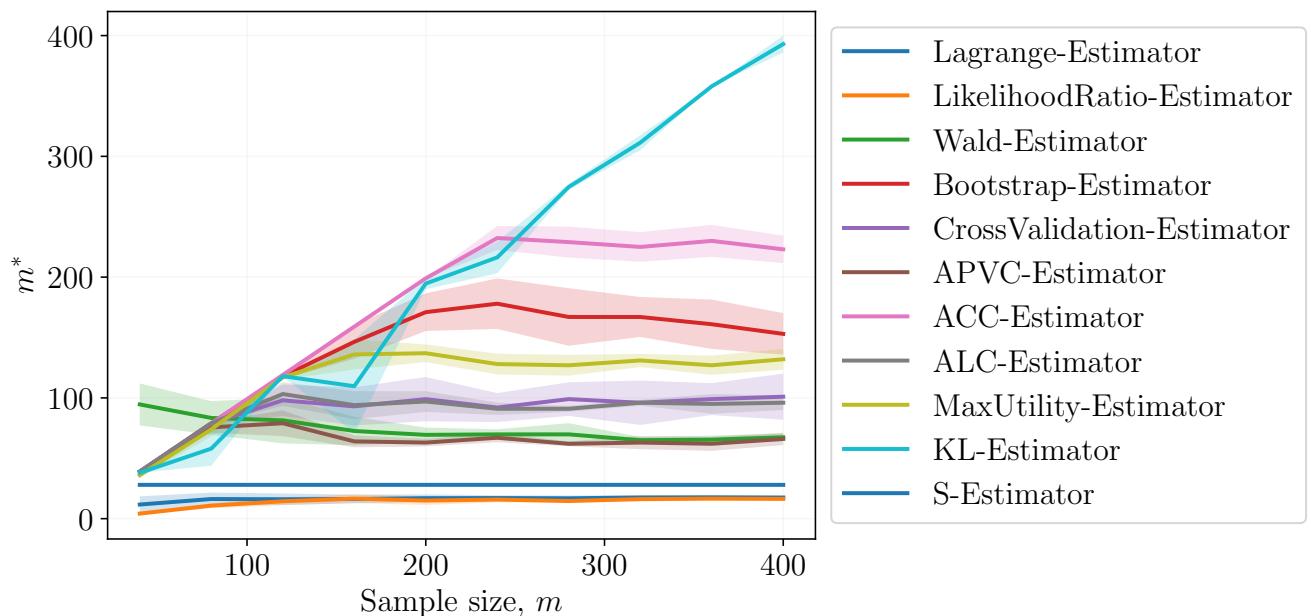


Рис. 3.12: Зависимость оцененного достаточного размера выборки m^* от доступного размера выборки m для каждого метода. Критерий на основе расхождения Кульбака-Лейблера является более консервативным и требует большего размера выборки. Критерий S-достаточности, напротив, предполагает, что может быть достаточен минимальный размер выборки.

На Рис. 3.9 мы можем наблюдать полученные зависимости между доступным размером выборки k и предложенными функциями $KL(k)$ и $S(k)$ для синтетического набора данных регрессии. В то же время, на Рис. 3.10 мы видим аналогичные графики для набора данных Liver Disorders. Можно заметить, 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^* . Чтобы изучить зависимость между ними, мы представляем Рис. 3.11, который показывает, какие размеры выборки можно выбрать для обеспечения определенного уровня достоверности.

Для сравнения наших предложенных методов с базовыми мы использовали

следующую схему эксперимента. Модель машинного обучения - линейная регрессия. Мы выбрали 4 набора данных с задачей регрессии из открытых источников: Boston, Diabetes, Forestfires и Servo. Их описательная статистика представлена в Таблице 3.5. Мы применили к ним 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. Они были подробно проанализированы в [5]. Использовались значения параметров этих методов по умолчанию.

Что касается наших методов, мы немного изменили определения достаточности, переведя их в термины относительного изменения. А именно, мы считаем размер выборки достаточным, если функция $KL(k)$ имеет относительное отклонение от своего значения на всей выборке не более чем ε . Аналогично с функцией $S(k)$. Мы зафиксировали $\varepsilon = 0.05$ и получили результатирующие размеры выборок. Это значение было выбрано потому, что другие методы, особенно статистические, используют 0.05 как значение ошибки I рода.

Результаты в Таблице 3.6 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. Мы полагаем, 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.

Данная часть включает наиболее комплексный анализ различных методов определения размера выборки. Мы анализируем, как достаточный размер выборки зависит от объема доступного набора данных. В частности, мы увеличиваем объем доступной выборки и вычисляем достаточный размер на основе

различных методов. Таким образом, мы получаем Рис. 3.12, который позволяет нам сравнить вышеупомянутые методы с точки зрения их консервативности.

Можно видеть, что S-достаточный размер выборки часто является минимальным. Мы уже обсуждали причину этого в предыдущем подразделе. Также, KL-достаточный размер выборки, как правило, требует почти полной выборки. По нашему мнению, это связано с тем, что расхождение Кульбака-Лейблера чрезвычайно чувствительно к изменениям математического ожидания и дисперсии сравниваемых распределений. Таким образом, стабилизация расстояния между ними происходит довольно поздно.

3.7. Заключение по главе

В данной главе была рассмотрена задача оценки достаточного объема выборки для задач машинного обучения. Были предложены различные методы решения данной задачи.

Предложен метод, основанный на анализе функции правдоподобия, включает методы D-достаточности и M-достаточности. Метод D-достаточности использует дисперсию функции правдоподобия на бутстрэп-подвыборках, в то время как метод M-достаточности анализирует разность математических ожиданий функции правдоподобия при последовательном добавлении объектов в выборку. Для линейной регрессионной модели была строго доказана корректность определения M-достаточного объема выборки при выполнении определенных условий на параметры модели.

Также, предложен метод, основанный на анализе близости апостериорных распределений параметров модели на похожих подвыборках. В рамках этого подхода были предложены критерии KL-достаточности и S-достаточности, использующие расхождение Кульбака-Лейблера и функцию схожести s-score соответственно. Для нормального апостериорного распределения удалось получить аналитические выражения для этих мер близости, что значительно упростило

теоретический анализ.

Экспериментальные исследования на синтетических и реальных данных подтвердили эффективность предложенных методов. Показано, что функции $D(k)$, $M(k)$, $KL(k)$ стремятся к нулю, а функция $S(k)$ — к единице с ростом объема выборки. Сравнительный анализ выявил особенности различных критериев: KL-дивергенция дает более консервативные оценки, требующие больших объемов данных, в то время как s-score часто указывает на достаточность минимального размера выборки.

Практические рекомендации включают использование относительных отклонений от значений на полной выборке с порогами 0.05 – 0.1. Основное ограничение методов связано с вычислительной сложностью обращения ковариационных матриц для моделей с большим количеством параметров.

Перспективы дальнейших исследований включают расширение теоретического обоснования методов на более сложные модели, в том числе нейронные сети, а также преодоление ограничения о нормальности апостериорного распределения. Полученные результаты создают основу для разработки практических инструментов оценки достаточного объема данных в прикладных задачах машинного обучения.

Глава 4

Методы снижения сложности моделей глубокого обучения

В данной главе рассмотрим 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.

4.1. Удаления параметров моделей глубокого обучения

Задана выборка:

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m, \quad \mathbf{x}_i \in \mathbb{X} = \mathbb{R}^n, \quad y_i \in \mathbb{Y},$$

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.

Задано 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 регрессии структура модели имеет вид f_{reg} , а в случае классификации имеет вид

f_{cl} . Задана функция 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 $\hat{\mathbf{w}}$ получим 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:

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \Delta \mathbf{w}(\mathbf{g}_{S,t}, \mathbf{w}_{t-1}, \mathbf{w}_{t-2}, \dots), \quad \mathbf{g}_{S,t} = \frac{\partial \mathcal{L}(\mathbf{w}_t, \mathbf{X}_S, \mathbf{Y}_S)}{\partial \mathbf{w}}, \quad (4.1)$$

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 [51], которая вычисляется на каждой итерации (4.1) 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 [51], 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 $\mathcal{I}_{\mathbf{w}_{t_0}} = [3, 1, 2]$.

Фиксация параметров модели в процессе обучения. Для фиксации параметров \mathbf{w}_{t_0} при помощи вектора индексов $\mathcal{I}_{\mathbf{w}_{t_0}}$ используется бинарный вектор $\boldsymbol{\alpha}(\zeta)$:

$$\alpha_i(\zeta) = \begin{cases} 1, & \text{если } \mathcal{I}_{\mathbf{w}_{t_0}}[j] \leq \zeta; \\ 0 & \text{иначе,} \end{cases} \quad (4.2)$$

где ζ — число фиксирующих параметров.

Учитывая (4.2), уравнение (4.1) 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

$$\mathbf{A}_{ps} = \mathbf{U} \Lambda \mathbf{V}^T.$$

Индекс обусловленности η_j определим как отношение максимального элемента к j -му элементу матрицы Λ . Для нахождения мультиколлинеарных признаков требуется найти индекс ξ вида:

$$\xi = \arg \max_{j \in \mathcal{A}} \eta_j.$$

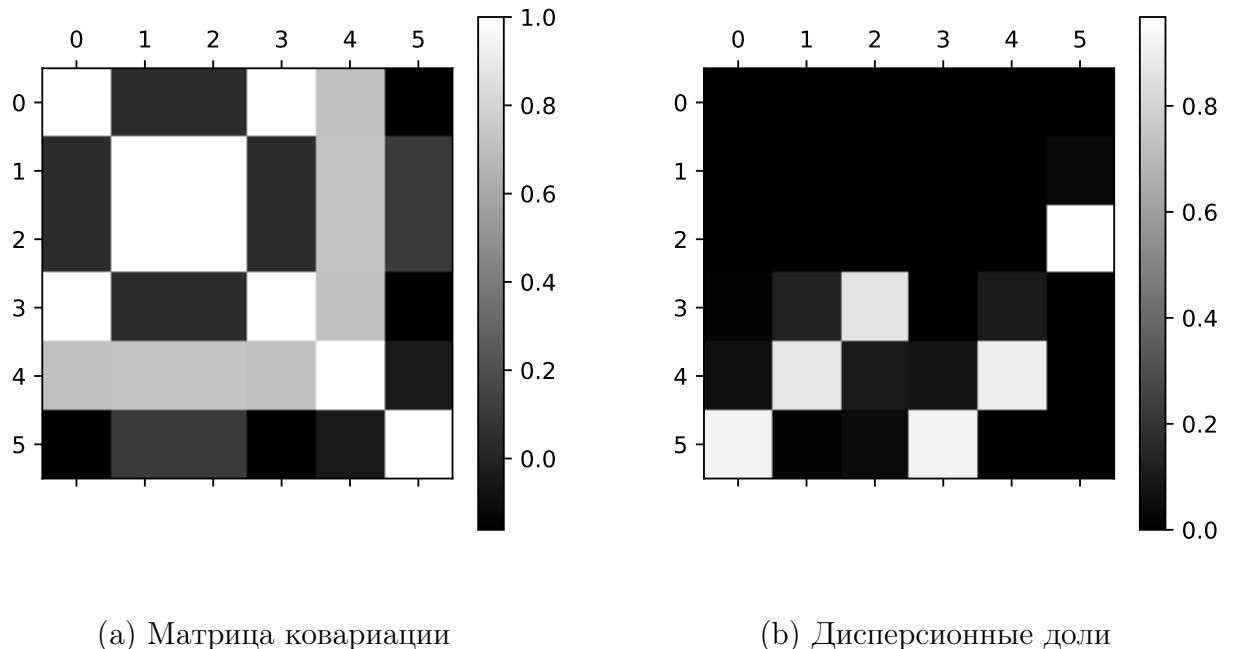


Рис. 4.1: Иллюстрация метода Белсли для анализа мультиколлинеарности параметров

Дисперсионный долевой коэффициент q_{ij} определим как вклад j -го признака в дисперсию i -го элемента вектора параметра \mathbf{w} :

$$q_{ij} = \frac{u_{ij}^2 / \lambda_{jj}}{\sum_{j=1}^n u_{ij}^2 / \lambda_{jj}}.$$

Большие значение дисперсионных долей указывают на наличие зависимости между параметрами. Находим долевые коэффициенты, которые вносят максимальный вклад в дисперсию параметра w_ξ :

$$\zeta = \arg \max_{j \in \mathcal{A}} q_{\xi j}.$$

Таблица 4.1: Иллюстрация метода Белсли для анализа мультиколлинеарности параметров

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	$9 \cdot 10^{-18}$	$1 \cdot 10^{-17}$	$2 \cdot 10^{-17}$	$9 \cdot 10^{-18}$	$2 \cdot 10^{-19}$	$9 \cdot 10^{-1}$
$2 \cdot 10^{15}$	$1 \cdot 10^{-2}$	$1 \cdot 10^{-1}$	$8 \cdot 10^{-1}$	$2 \cdot 10^{-3}$	$9 \cdot 10^{-2}$	$1 \cdot 10^{17}$
$8 \cdot 10^{15}$	$6 \cdot 10^{-2}$	$8 \cdot 10^{-1}$	$9 \cdot 10^{-2}$	$8 \cdot 10^{-2}$	$9 \cdot 10^{-1}$	$2 \cdot 10^{17}$
$1 \cdot 10^{16}$	$9 \cdot 10^{-1}$	$1 \cdot 10^{-2}$	$4 \cdot 10^{-2}$	$9 \cdot 10^{-1}$	$1 \cdot 10^{-3}$	$5 \cdot 10^{-21}$

Параметр с индексом ζ определим 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. 4.1.a, где $x \in [0.0, 0.02, \dots, 20.0]$.

В табл. 4.1 приведены индексы обусловленности и соответствующие им дисперсионные доли, которые также изображены на рис. 4.1.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, как видно из построения выборки, являются линейно зависимые.

4.2. Дистилляция моделей глубокого обучения на многодоменных данных

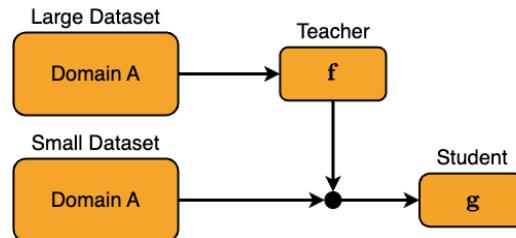


Рис. 4.2: Basic distillation

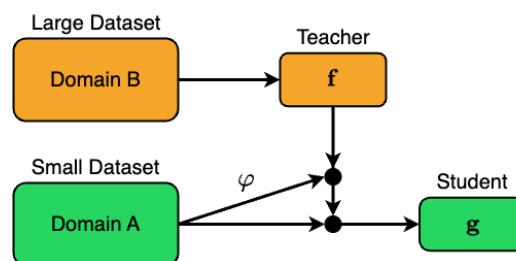


Рис. 4.3: Distillation with domain adaptation

Определение 16. Генеральная совокупность объектов B называется близкой к генеральной совокупности A , если существует инъективное отображение $\varphi : A \rightarrow B$.

Мы предлагаем использовать, помимо меток учителя на одном из доменов, связь между доменами при обучении студенческой модели. В этом случае в качестве доменов должны выступать близкие генеральные совокупности.

На Рис. 4.2 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 студенческой модели на меньшем наборе данных из того же домена. На Рис. 4.3 представлен предложенный метод, который задействует выходы модели учителя, обученной на другом домене, и связь между доменами.

Базовая постановка 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. И требуется обучить студенческую модель с меньшим количеством параметров, учитывая ответы учителя. Модель учителя \mathbf{f} и студенческая модель \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.

Функция потерь \mathcal{L} , учитывающая модель учителя \mathbf{f} при выборе студенческой модели \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 .

Получаем задачу оптимизации:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{W}} \mathcal{L}(\mathbf{w}, \mathbf{X}, \mathbf{Y}, \mathbf{f}).$$

Постановка 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 студенческую модель для малоресурсной выборки:

$$\mathbf{g} : \mathbb{X}_t \rightarrow \mathbb{Y},$$

где \mathbf{g} - студенческая модель.

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.

Получаем задачу оптимизации:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{W}} \mathcal{L}(\mathbf{w}, \mathbf{X}, \mathbf{Y}, \mathbf{f}, \varphi).$$

4.3. Анти-Дистилляция моделей глубокого обучения

В этом разделе мы описываем постановку задачи анти-дистилляции для задачи классификации. Отметим, что аналогичный подход может быть применен для произвольных задач.

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 $1, \dots, c_1, \dots, c_2$.

Мы предполагаем, 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. Мы также предполагаем, что набор данных \mathfrak{D}_2 является более сложным для классификации и требует более сложной модели классификации.

Задана модель учителя \mathbf{g}_{tr} , обученная 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:

$$l(y, \hat{y}) = - \sum_{k=1}^c [y = k] \log \hat{y}_k, \quad y \in C, \quad \hat{y} \in \Delta^c.$$

Наша задача - построить студенческую модель

$$\mathbf{f}_{\text{st}} : \mathbb{R}^n \rightarrow \Delta^{c_2}, \quad \mathbf{f}_{\text{st}}(\mathbf{x}) = \mathbf{f}(\mathbf{x}, \hat{\mathbf{w}}),$$

, которая минимизирует 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 мы не можем напрямую оптимизировать валидационную функцию потерь, общепринятой практикой является использование градиентных методов оптимизации на обучающей части $\mathfrak{D}_2^{\text{train}}$ набора данных \mathfrak{D}_2 . Чтобы уменьшить переобучение и использовать больше информации о данных, мы получаем информацию от модели учителя \mathbf{g}_{tr} . Здесь мы используем наше предположение, 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 студента $\mathbf{w} = \varphi(\hat{\mathbf{u}})$.

Гипотеза 1. *Студенческие модели, 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 студенческая модель \mathbf{f}_{st} . Чтобы использовать больше информации из параметров модели учителя $\hat{\mathbf{u}}$, нам нужно расширить размерность пространства параметров модели учителя N_{tr} до размерности N_{st} пространства параметров студенческой модели.

Чтобы справиться с этим, мы оптимизируем следующую составную функцию потерь:

$$\varphi(\mathbf{u}) = \arg \min_{\mathbf{w} \in \mathbb{R}^{N_{\text{st}}}} \mathcal{L}(\mathbf{w}), \quad (4.3)$$

где

$$\mathcal{L}(\mathbf{w}) = \lambda_1 \mathcal{L}_{\text{ce}}(\mathbf{w}, \mathfrak{D}_1) + \lambda_2 \mathcal{L}_2(\mathbf{w}, \mathbf{u}) + \lambda_3 \mathcal{L}_3^\delta(\mathbf{w}, \mathfrak{D}_1) + \lambda_4 \mathcal{L}_4(\mathbf{w}),$$

$$\forall i \in \overline{1, 4} \quad \lambda_i \geq 0$$

Здесь $\mathcal{L}_{\text{ce}}(\mathbf{w}, \mathfrak{D}_1)$ - перекрестная энтропия, отвечающая за качество студенческой модели на \mathfrak{D}_1 .

Второе слагаемое

$$\mathcal{L}_2(\mathbf{w}, \mathbf{u}) = \|\mathbf{u} - \mathbf{Pr}[\mathbf{w}]\|_2^2$$

обеспечивает малую разницу между параметрами модели учителя и студенческой модели в соответствующих местах, где \mathbf{Pr} берет только первые параметры, общие для обеих моделей (в случае моделей многослойного перцептрона, \mathbf{Pr} берет параметры тех же нейронов для каждого слоя модели).

Компонента

$$\mathcal{L}_3^\delta(\mathbf{w}, \mathfrak{D}_1) = \sum_{\substack{\mathbf{x}, y \in \mathfrak{D}_1}} \mathbb{E}_{\mathbf{x}' \in U_\delta(\mathbf{x})} \mathcal{L}_{\text{ce}}(\mathbf{w}, \mathbf{x}', y)$$

отвечает за устойчивость решения к шуму во входных данных, где $U_\delta(\mathbf{x})$ представляет равномерное распределение в $[\delta - \mathbf{x}; \delta + \mathbf{x}]$.

Последнее слагаемое

$$\mathcal{L}_4(\mathbf{w}) = \text{tr} \left(\frac{\partial^2 \mathcal{L}_{\text{ce}}}{\partial \mathbf{w}^2} \right)$$

выполняет регуляризацию гессиана, что также повышает устойчивость модели.

Заметим, что последнее слагаемое \mathcal{L}_4 включает вычисление гессиана, наивное вычисление которого может быть ресурсоемким. В этой статье мы используем метод стохастической аппроксимации [52] следа гессиана с быстрым умножением гессиан-вектор [53]. Сложность такой процедуры линейна от количества параметров модели \mathbf{f}_{st} .

В интересующем нас случае, Анти-Дистилляции, подразумевается $\lambda_2 > 0$, т.е. оптимизация, которая делает параметры модели учителя и студента достаточно близкими. Мы также заинтересованы в получении модели, устойчивой к искажению входных данных. Для этого свойства мы используем слагаемые \mathcal{L}_3 и \mathcal{L}_4 . Оба этих слагаемых регулируют гессиан функции перекрестной энтропии [54, 55].

4.4. Результаты вычислительных экспериментов

4.4.1. Удаления параметров моделей глубокого обучения

Для анализа свойств предложенного алгоритма и сравнения его с существующими проведен вычислительный эксперимент в котором параметры нейросети удалялись методами, которые описаны в разделах 3.1—3.3 и методом Белсли.

В качестве данных использовались три выборки. Выборки Wine [56] и Boston Housing [48] — это реальные данные. Синтетические данные сгенерированы таким образом чтобы параметры сети мультиколлинеарными. Генерация 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 являются нерелевантными. Матрица ковариации выбрана таким образом, чтобы все нерелевантные параметры являлись зависимыми величинами, что приводит к максимальной эффективности метода Белсли.

Для 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. Также критерием качества служит устойчивость нейросети к зашумленности данных.

Таблица 4.2: Описание выборок для анализа метода задания порядка методом Белсли

Выборка	Тип задачи	Размер выборки	Число признаков
Wine	классификация	178	13
Boston Housing	регрессия	506	13
Synthetic data	регрессия	10000	100

Качеством прогноза R_{cl} модели для задачи классификации является точность прогноза модели:

$$R_{\text{cl}} = \frac{\sum_{(\mathbf{x}, y) \in \mathcal{D}} [f(\mathbf{x}, \mathbf{w}) = y]}{|\mathcal{D}|},$$

Качеством прогноза R_{rg} модели для задачи регрессии является среднеквадратическое отклонение результата модели от точного:

$$R_{\text{rg}} = \frac{\sum_{(\mathbf{x}, y) \in \mathcal{D}} (f(\mathbf{x}, \mathbf{w}) - y)^2}{|\mathcal{D}|},$$

Wine. Рассмотрим нейронную сеть с 13 нейронами на входе, 13 нейронами в скрытом слое и 3 нейронами на выходе.

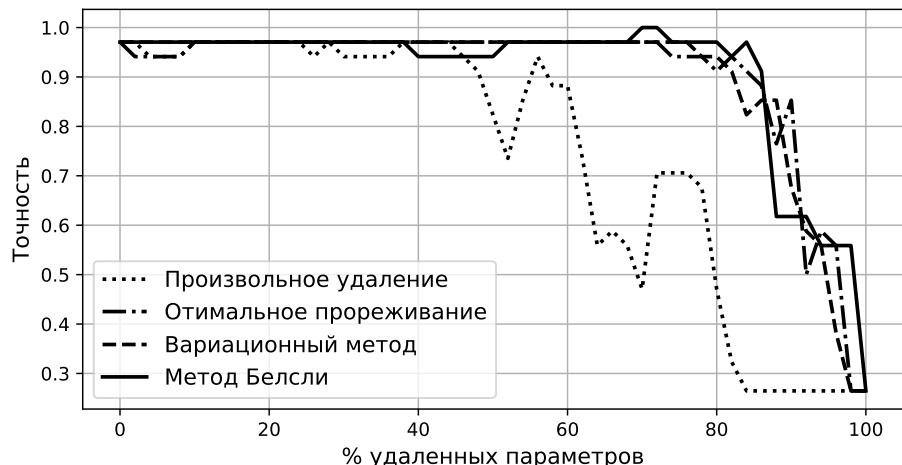


Рис. 4.4: Качество прогноза при удалении параметров на выборке Wine

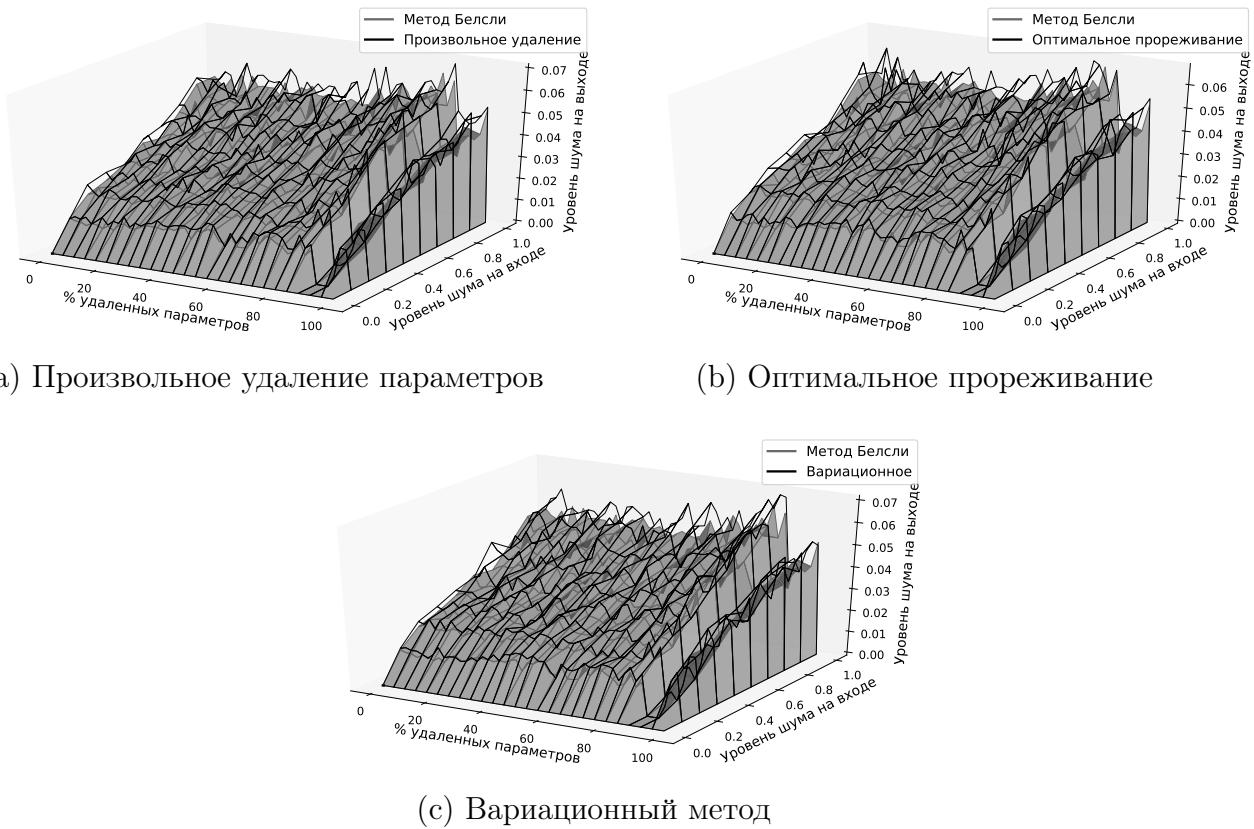


Рис. 4.5: Влияние шума в начальных данных на шум выхода нейросети на выборке Wine

На рис. 4.4 показано как меняется точность прогноза R_{cl} при удалении параметров указанными методами. Из графика видно, что метод оптимального прореживания, вариационный метод и метод Белсли позволяют удалить $\approx 80\%$ параметров и качество всех этих методов падает при удалении $\approx 90\%$ параметров нейросети.

На рис. 4.5 показаны поверхности изменения уровня шума ответов нейросети при изменении процента удаленных параметров и уровня шума входных данных для разных методов прореживания. На графиках показано, что при удалении параметров нейросети методом Белсли шум меньше, чем при удалении параметров другими методами, на это указывает то что поверхность которая соответствует методу Белсли ниже других поверхностей.

Boston Housing. Рассмотрим нейронную сеть с 13 нейронами на входе, 39

нейронами в скрытом слое и одним нейроном на выходе.

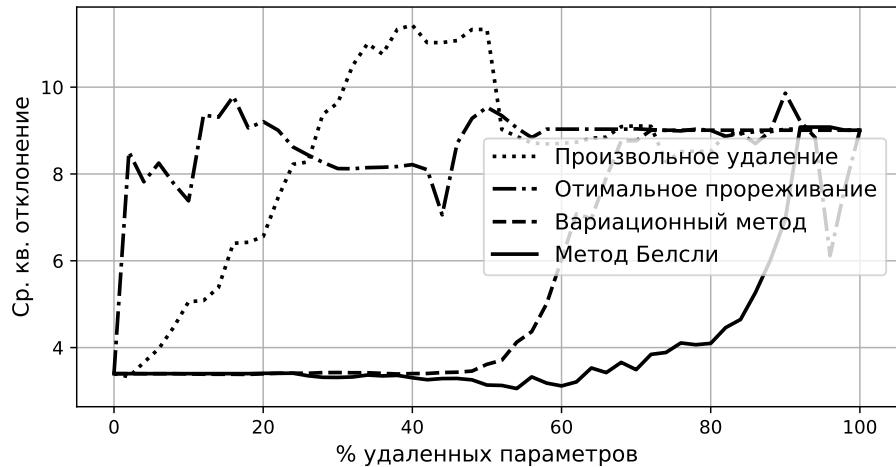


Рис. 4.6: Качество прогноза при удаление параметров на выборке Boston

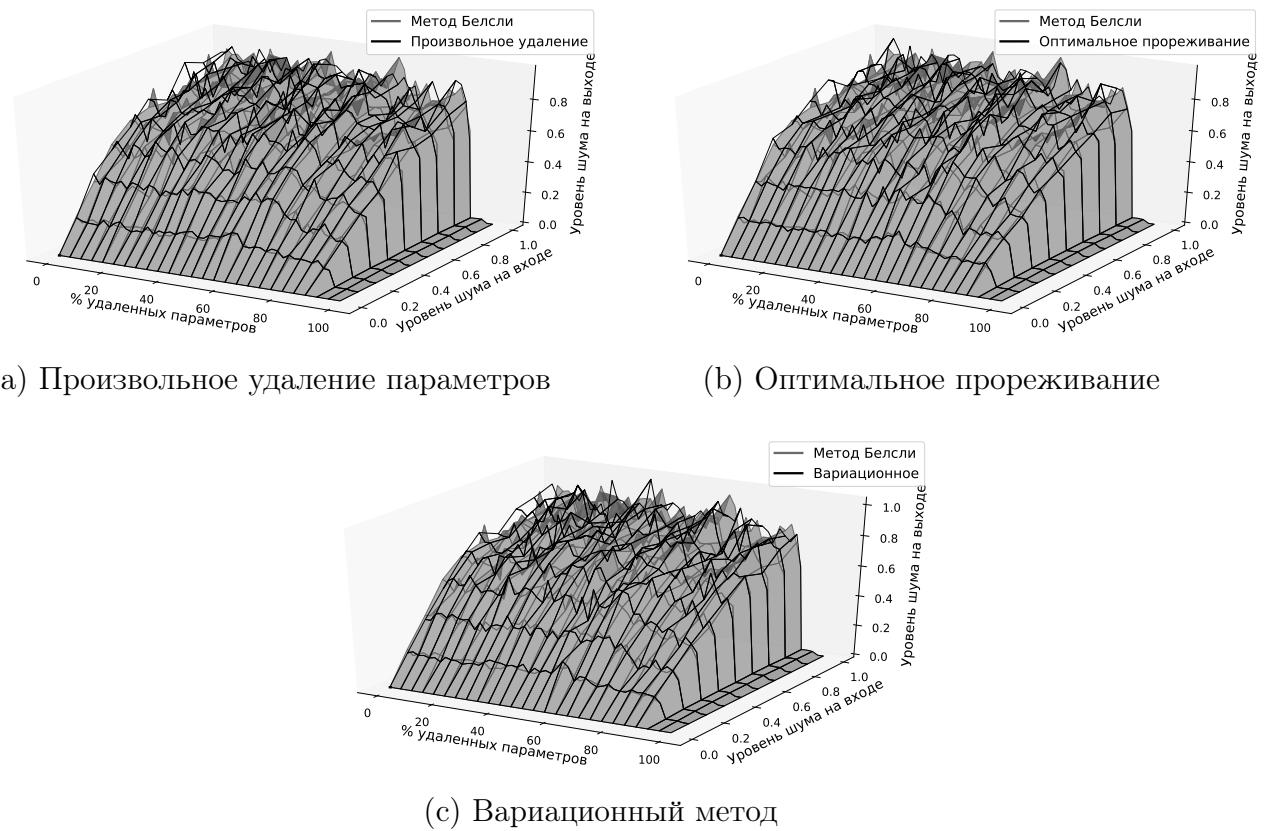


Рис. 4.7: Влияние шума в начальных данных на шум выхода нейросети на выборке Boston

На рис. 4.6 показано как меняется среднеквадратическое отклонение про-

гноза R_{rg} от точного ответа при удалении параметров указанными методами. График показывает, что метод Белсли является более эффективным, чем другие методы, так как позволяет удалить больше параметров нейросети без потери качества.

На рис. 4.7 показаны поверхности изменения уровня шума ответов нейросети при изменении процента удаленных параметров и уровня шума входных данных для разных методов прореживания. График показывает, что уровень шума всех методов одинаковый, так как поверхности всех методов находятся на одном уровне.

Синтетические данные. Рассмотрим нейронную сеть с 100 нейронами на входе и одним нейроном на выходе.

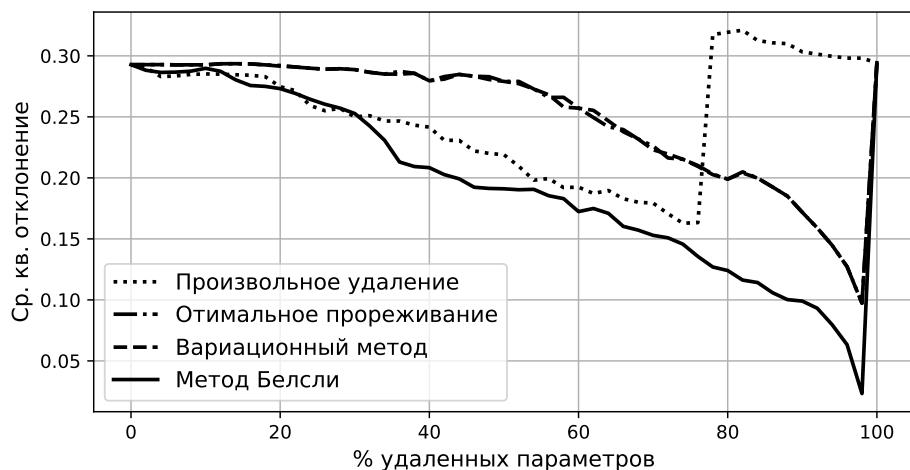


Рис. 4.8: Качество прогноза при удаление параметров на синтетической выборке

На рис. 4.8 показано как меняется среднеквадратическое отклонение прогноза от R_{rg} точного ответа при удалении параметров указанными методами. График показывает, что удаление параметров методом Белсли является более эффективным чем другие методы прореживания, так как качество прогноза нейросети повышается при удалении шумовых параметров.

На рис. 4.9 показаны поверхности изменения уровня шума ответов нейросети при изменении процента удаленных параметров и уровня шума входных

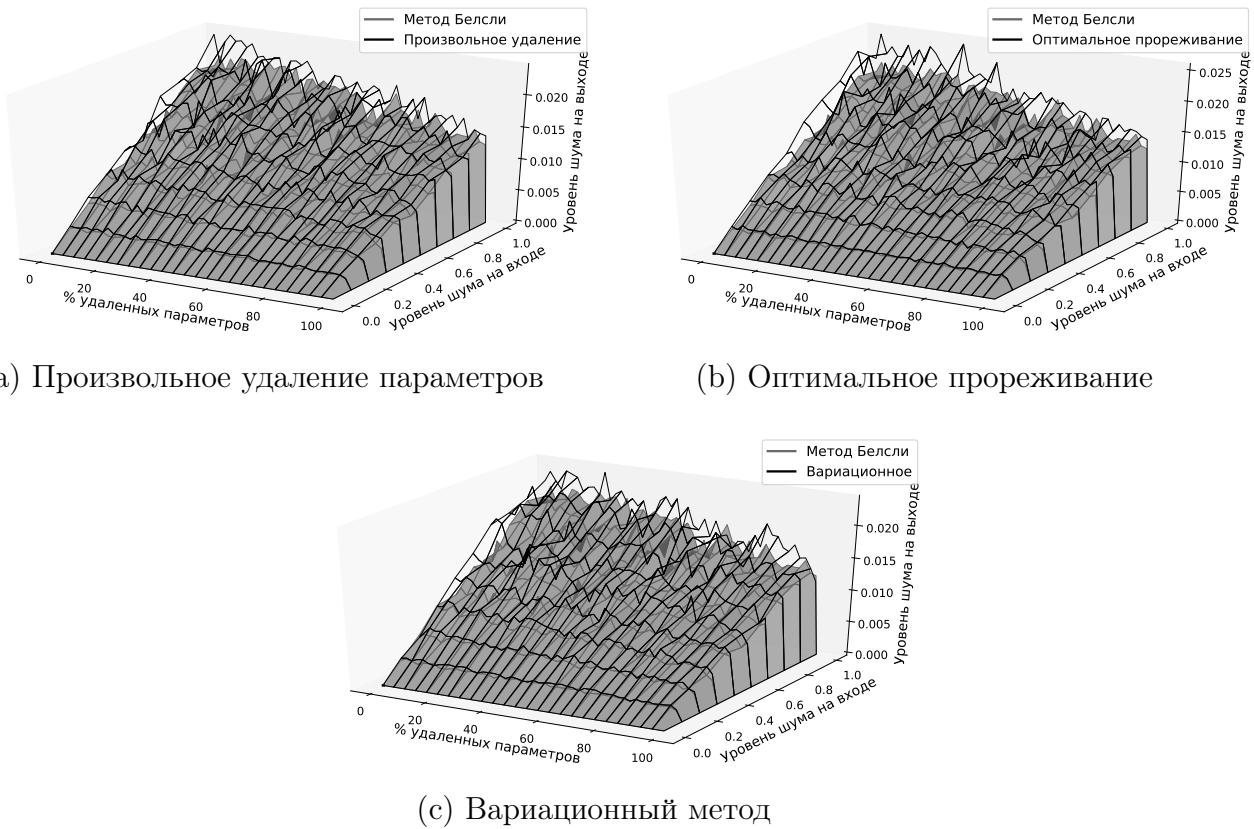


Рис. 4.9: Влияние шума в начальных данных на шум выхода нейросети на синтетической выборке

данных для разных методов прореживания. На графиках показано, что при удалении параметров нейросети методом Белсли шум меньше, чем при удалении параметров другими методами, так как поверхность которая соответствует методу Белсли ниже других поверхностей.

4.4.2. Дистилляция моделей глубокого обучения на многодоменных данных

Цель вычислительного эксперимента — сравнить производительность моделей учителя и студента на реальных наборах данных с использованием отображения φ и без него для задач компьютерного зрения и обработки естественного языка. Для анализа качества дистилляции предложен интегральный критерий качества [57].

- Мы используем подмножество ImageNet — набор изображений, для которого необходимо решить задачу классификации на 10 классов [58]. Набор данных состоит из обучающей и тестовой частей, причём обучающая часть разделена на мультиресурсную и малоресурсную части.
- OPUS-100 является англоцентричным, то есть все обучающие пары включают английский язык на стороне источника или цели [59]. Мы используем наборы данных fr-en и de-en.

Конфигурация алгоритма многодоменной дистилляции для задачи компьютерного зрения

Таблица 4.3: Структура учителя

Слой	Размер входного вектора	Количество параметров
Входной слой	(3, 200, 200)	0
CONV1 (размер ядра=5)	(24, 196, 196)	1800
POOL1	(24, 98, 98)	0
CONV2 (размер ядра = 5)	(48, 94, 94)	28800
POOL2	(48, 47, 47)	0
CONV3 (размер ядра = 8)	(96, 40, 40)	294912
POOL3	(96, 20, 20)	0
CONV4 (размер ядра = 5)	(192, 16, 16)	460800
POOL4	(192, 8, 8)	0
CONV5 (размер ядра = 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
		$\sum = 4455984$

Таблица 4.4: Структура студента

Слой	Размер входного вектора	Количество параметров
Входной слой	(3, 200, 200)	0
CONV1 (размер ядра=5)	(24, 196, 196)	1800
POOL1	(24, 98, 98)	0
CONV2 (размер ядра = 5)	(48, 94, 94)	28800
POOL2	(48, 47, 47)	0
Полносвязный слой	(106032)	0
Полносвязный слой	(120)	12723840
Полносвязный слой	(10)	1200
		$\Sigma = 12755640$

Структуры модели учителя **f** и модели студента **g** описаны в Таблице 4.3 и Таблице 4.4. Функция активации после каждого скрытого слоя — ReLu. Мы используем метод градиентной оптимизации Adam [30] для решения задачи оптимизации.

Таблица 4.5: Набор данных ImageNet

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

В Таблице 4.5 описаны наборы данных для вычислительного эксперимента 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, мультире-

сурсная часть содержит 8 469 объектов, малоресурсная часть содержит 1 000 объектов, а тестовая часть содержит 3 925 объектов.

Цель эксперимента — сравнить производительность студенческой модели, обучающейся без учителя, с учителем и с учителем на другом домене с использованием адаптации домена. Сначала мы обучили студенческую модель на малоресурсной части и протестировали её на тестовой части, затем использовали метки модели учителя, обученной на мультиресурсной части, для обучения студенческой модели. Наконец, мы обучили модель учителя на стилизованных изображениях из мультиресурсной части и использовали модель учителя и отображение для обучения студенческой модели. Предобученная модель CycleGAN [60] использовалась в качестве отображения φ .

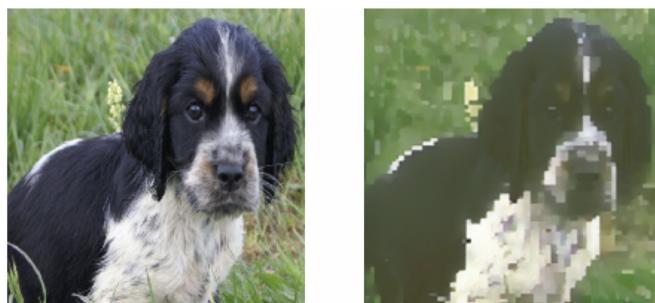


Рис. 4.10: Сравнение примера объекта до и после преобразования.

Рис. 4.10 показывает одно из изображений в наборе данных ImageNet [58] и то же изображение после преобразования с помощью модели CycleGAN [60]. Мы усредняем результаты по 5 запускам и изучаем среднее значение и дисперсию метрик.

Конфигурация алгоритма многодоменной дистилляции для задачи обработки естественного языка. Набор данных OPUS100 был разделен на обучающую часть для учителя, состоящую из немецко-английских предложений, и обучающую часть и тестовую часть для студента, состоящую из французско-английских предложений. Обучающая часть учителя содержала 5 000 предложений, обучающая часть студента содержала 2 000 предложений, а тестовая часть содержала 500 предложений.

Таблица 4.6: Набор данных OPUS100

Набор данных	Описание	Язык	Размер набора
Teacher-Train	Обучающая часть учителя	de-en	5000
Student-Train	Обучающая часть студента	fr-en	2000
Student-Test	Тестовая часть	fr-en	500

Таблица 4.6 описывает наборы данных для вычислительного эксперимента по обработке естественного языка.

Мы использовали студенческую модель \mathbf{g} и модель учителя \mathbf{f} в качестве трансформерной модели на основе статьи "Attention Is All You Need"[61] и метод градиентной оптимизации Adam [30] для решения задачи оптимизации. Модель NLLB[62] использовалась в качестве отображения φ . Эта модель переводила французские предложения в немецкие.

Аналогично эксперименту по компьютерному зрению, мы сравниваем производительности студенческой модели без учителя, с учителем и с учителем и адаптацией домена. Мы усредняем результаты по 5 запускам и изучаем среднее значение и дисперсию метрик.

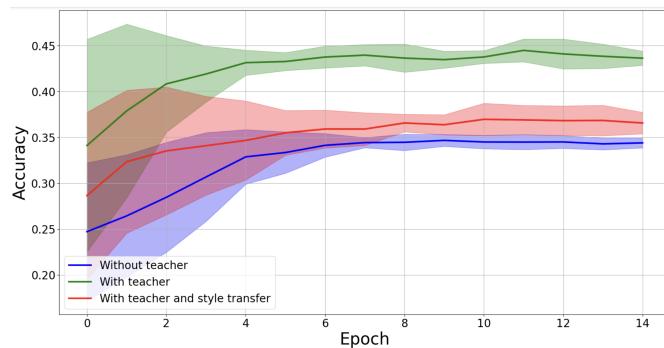


Рис. 4.11: Точность аппроксимации на тестовой выборке. Все результаты усреднены по 5 запускам.

Как видно из Рис. 4.11 и Рис. 4.12, модели, обученные с использованием учителя, достигают лучшего качества и точности. Можно заметить, что студен-

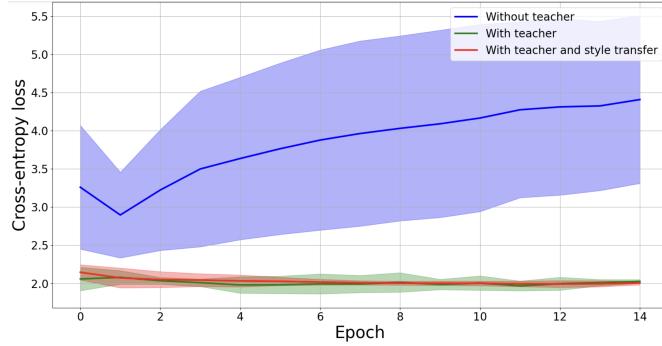


Рис. 4.12: Ошибка перекрестной энтропии между истинными и предсказанными студенческими метками на тестовой выборке. Все результаты усреднены по 5 запускам.

ческая модель, обученная с использованием меток учителя на том же домене (зеленые линии), достигает наивысшей точности и наименьших потерь. Студенческая модель, обученная с использованием меток учителя и адаптации домена (красные линии), показывает лучшее качество аппроксимации, чем модель без использования учителя.

Таким образом, мы экспериментально показали, что дистилляция с использованием адаптации домена может приводить к более эффективным нейронным сетям с меньшим количеством параметров.

Таблица 4.7: Качество моделей для компьютерного зрения

Студент	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	—	$0,44 \pm 0,01$	$2,03 \pm 0,02$	$28,08 \pm 1,22$

Результаты также представлены в табличной форме. Таблица 4.7 содержит данные о валидационной точности, потерях и интегральном критерии моделей, обученных с дистилляцией и адаптацией домена и без них, в эксперименте по компьютерному зрению.

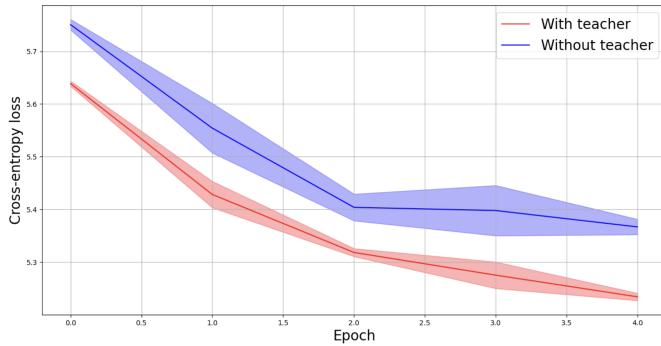


Рис. 4.13: Ошибка перекрестной энтропии на тестовом наборе данных. Все результаты усреднены по 3 запускам.

Как видно из Рис. 4.13, модели, обученные с использованием учителя, достигают лучшего качества.

Таблица 4.8 показывает результаты сравнения студенческих моделей, полученных с использованием дистилляции и без.

Таблица 4.8: Качество моделей для NLP

Студент	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

4.4.3. Анти-Дистилляция моделей глубокого обучения

Цель вычислительного эксперимента — сравнить производительность моделей в зависимости от инициализации параметров.

Мы сравниваем различные подходы к инициализации:

1. Xavier — заполнение всех параметров модели $U[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}]$, где n — количество нейронов входного слоя [63], 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[64].
6. With Data Noise — получение инициализации студенческой модели путем решения задачи оптимизации 4.3 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 ($\lambda_4 = 0$) [65].
8. Anti-Distillation — оптимизация всех λ_i .

Критериями качества являются: точность на валидационном наборе, точность на валидационном наборе, искаженном атакой FSGM [66], точность на валидационном наборе при условии, что параметры модели искажены шумом: $\mathbf{w}_\varepsilon = \mathbf{w} + \varepsilon \boldsymbol{\xi}$, где $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Fashion-MNIST — это 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 28x28, связанное с меткой из 10 классов [67].

Проведение экспериментов осуществляется следующим образом: обучаем модель учителя, увеличиваем её сложность и сравниваем различные способы инициализации параметров модели. Мы рассматриваем полно связные сети.

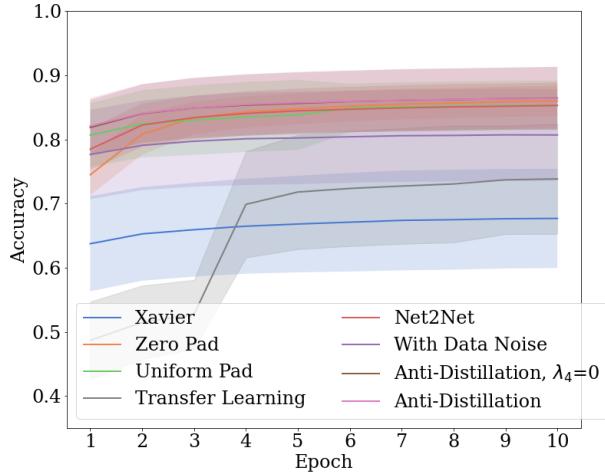


Рис. 4.14: Сравнение валидационной точности для различных методов инициализации

Учителя имеют следующие размеры скрытых слоев: [128, 64, 32]. Студенческие модели имеют [256, 128, 64] нейронов в скрытых слоях.

Учителя обучались в течение 30 эпох с начальной скоростью обучения 1e-2, которая затем уменьшалась до 1e-3 после 10 эпох. Студенты сравнивались при обучении в течение 10 эпох со скоростью обучения 1e-3. Оптимизация проводится с использованием алгоритма оптимизации Adam [30]. Мы сравниваем методы инициализации, измеряя точность предсказаний, значение функции потерь перекрестной энтропии на валидационной выборке и дисперсию предсказаний. Также мы исследуем случай зашумленных входных данных, рассматривая указанные критерии качества в зависимости от процента искаженных изображений.

Набор данных \mathfrak{D}_2 состоит из Fasion-MNIST, а $\mathfrak{D}_1 = \{(\mathbf{x}, y) \mid (\mathbf{x}, y) \in \mathfrak{D}_2, y \in C_1\}$, $C_1 \subset C_2$, $C_1 = \{0, \dots, 4\}$, $C_2 = \{0, \dots, 9\}$.

Как видно на Рисунке 4.14, модели, использующие Анти-Дистилляцию, в среднем имеют меньшую дисперсию и более высокую точность, чем модели с различной инициализацией параметров. Обучение модели с нуля оказалось не лучшим решением. Предложенный метод дает нам лучшие результаты с мень-

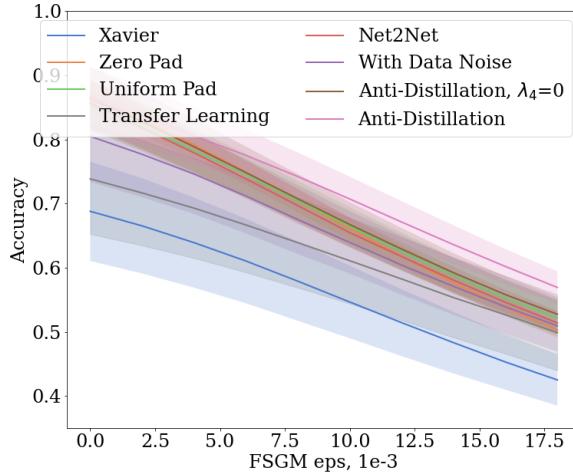


Рис. 4.15: Зависимость валидационной точности от адвверсарного шума в данных

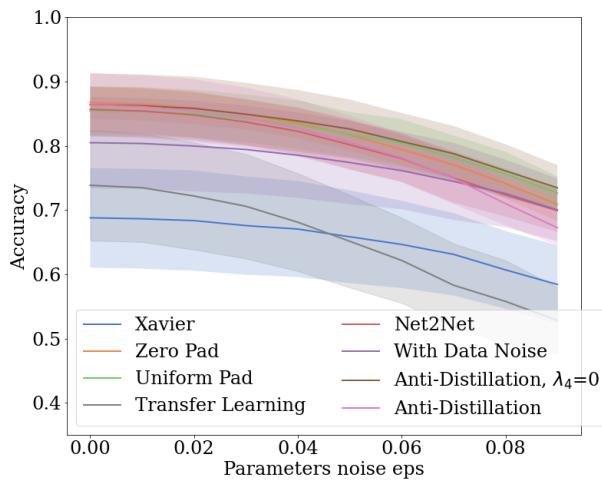


Рис. 4.16: Зависимость валидационной точности от параметра интенсивности шума ε

шим количеством итераций для сходимости. Отметим, что мы не учитывали количество итераций, необходимых для расширения модели учителя, которое также требует процедуры оптимизации. Мы полагаем, что во многих реальных случаях этим временем можно пренебречь, поскольку предложенный метод позволяет нам расширить модель учителя один раз, используя только базовый набор данных \mathfrak{D}_1 , для последующего использования в множественных задачах

обучения студента [68].

Рисунок 4.15 показывает, что Анти-Дистилляция является наиболее устойчивым к адверсарным атакам методом инициализации параметров модели, поскольку она имеет наивысшую валидационную точность с большим отрывом при высоких уровнях шума.

На Рисунке 4.16 мы видим, 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 4.9: Точность на валидационном наборе.

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

Результаты также представлены в табличной форме. Таблица 4.9 содержит данные о валидационной точности для различных методов инициализации после последней эпохи обучения, значения валидационной точности при наивысшем уровне шума изображения от адверсарной атаки и информацию о значениях точности для наивысшего уровня шума в параметрах модели.

4.5. Заключение по главе

В главе были рассмотрены методы передачи знаний между нейронными сетями в условиях различных доменов и сложности данных.

Мультидоменная дистилляция продемонстрировала эффективность передачи знаний от более сложной модели, обученной на большом наборе данных, к менее сложной модели, обучаемой на малом наборе данных того же или другого домена. Эксперименты в областях компьютерного зрения и обработки естественного языка подтвердили улучшение качества аппроксимации студенческой модели. Ожидаемо, обучение в рамках одного домена показало лучшие результаты по сравнению с использованием адаптации домена, однако последняя также обеспечила значительное улучшение по сравнению с базовыми подходами.

Антидистилляция решила задачу расширения модели для работы с более сложными наборами данных. Предложенный метод передачи знаний от простой модели к более сложной не только повысил точность на сложных данных, но и увеличил устойчивость модели к шуму во входных данных и нормальному шуму в параметрах модели. Эксперименты на наборе данных Fashion-MNIST подтвердила эффективность подхода.

Оба метода открывают перспективы для практического применения в условиях ограниченных данных и необходимости адаптации моделей к новым доменам. Дальнейшие исследования будут направлены на интеграцию байесовских методов дистилляции, учитывающих распределения параметров, а также на применение разработанных подходов к другим архитектурам нейронных сетей и наборам данных.

Глава 5

Применение теоретических оценок в прикладных задачах

Заключение

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 выполняется следующее 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 выполняется следующее 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*

$$\|\mathbf{A}\|_2 = \|\mathbf{A}^\top\|_2$$

Свойство 4 (Соотношения между матричными нормами). *Пусть матрица $\mathbf{A} \in \mathbb{R}^{m \times n}$, тогда следующие неравенства между матричными нормами являются верными:*

$X \setminus Y$	$\ \mathbf{A}\ _{\max}$	$\ \mathbf{A}\ _1$	$\ \mathbf{A}\ _\infty$	$\ \mathbf{A}\ _2$	$\ \mathbf{A}\ _F$
$\ \mathbf{A}\ _{\max}$	1	1	1	1	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}	

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 это константа на пересечения строки X и колонки 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 матрица \mathbf{A} и матрица \mathbf{B} являются матрицами из пространства 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 задана матрица $\mathbf{A} \in \mathbb{R}^{m \times n}$ представляет собой блочную матрицу, каждый блок которой является матрицей $\mathbf{B}_{i,j}$. Тогда выполняется следующее неравенство:

$$\|\mathbf{A}\|_2 \leq \sqrt{mn} \max_{i,j} \|\mathbf{B}_{i,j}\|_2$$

Заметим, что если матрица \mathbf{A} является блочно-диагональной, то имеет место строгое равенство $\|\mathbf{A}\|_2 = \max_i \|\mathbf{B}_{i,i}\|_2$.

Свойство 8 (Поэлементное деление). Пусть задана матрица $\mathbf{A} \in \mathbb{R}^{m \times n}$ и вектор $\mathbf{b} \in \mathbb{R}^{m \times 1}$. Тогда для матрицы $\mathbf{C} \in \mathbb{R}^{m \times n}$, где $c_{i,j} = \frac{a_{i,j}}{b_i}$, выполняется

$$\mathbf{C} = \text{diag}^{-1}(\mathbf{b})\mathbf{A}$$

Свойство 9 (Производная матричного произведения). Пусть заданы матрицы $\mathbf{X}, \mathbf{A}, \mathbf{B}$ соответствующих размерностей, 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.

Данный результат непосредственно следует из [69], где был получен аналогичный результат для покоординатной векторизации.

Свойство 11 (Построчная векторизация матричного произведения). Пусть заданы матрицы $\mathbf{X}, \mathbf{A}, \mathbf{B}$ соответствующих размерностей, 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 17 (Матрица перестановки). *Матрица перестановки $\mathbf{K}_{m,n} \in \mathbb{R}^{mn \times mn}$ это такая единственная матрица, что для любой матрицы $\mathbf{A} \in \mathbb{R}^{m \times n}$ выполняется:*

$$\mathbf{K}_{m,n} \text{vec}(\mathbf{A}) = \text{vec}(\mathbf{A}^\top)$$

Используя свойство 5, получаем 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 18 (Векторизация и поэлементные операции). *Пусть задана матрица \mathbf{A} и вектор \mathbf{v} . Тогда*

- $\text{vec}_r(\mathbf{A})$ обозначает построчную векторизацию матрицы \mathbf{A} .
 - $\mathbf{A}^{\circ\alpha}$ обозначает поэлементное возведение матрицы \mathbf{A} в степень α , т.е.
- $$(\mathbf{A}^{\circ\alpha})_{ij} = (\mathbf{A}_{ij})^\alpha.$$
- $\text{diag}(\mathbf{v})$ создаёт диагональную матрицу с вектором \mathbf{v} на главной диагонали.

Определение 19. Для матрицы $\mathbf{A} \in \mathbb{R}^{m \times n}$:

$$\|\mathbf{A}\|_2 = \sigma_1,$$

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2} = \sqrt{\sum_{i=1}^r \sigma_i^2},$$

$$\|\mathbf{A}\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}|,$$

$$\|\mathbf{A}\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}|,$$

$$\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|.$$

Список иллюстраций

1.1	Пример изменения функции потерь при добавлении нового объекта	20
1.2	Проверка Предположения 1 о сходимости локальных минимумов при увеличении объема выборки.	33
1.3	Зависимость абсолютного значения разности функций потерь от доступного размера выборки, прямая классификация изображений. Графики слева показывают уменьшение значений с увеличением размерности скрытого слоя. Графики справа показывают увеличение значений с увеличением количества слоев.	33
1.4	Зависимость абсолютного значения разности функций потерь от доступного размера выборки, извлечение признаков изображений . Графики слева показывают уменьшение значений с увеличением размерности скрытого слоя. Графики справа показывают увеличение значений с увеличением количества слоев.	34
1.5	Зависимость абсолютного значения разности функций потерь от доступного размера выборки, прямая классификация изображений. Графики слева показывают уменьшение значений с увеличением размерности скрытого слоя. Графики справа показывают увеличение значений с увеличением количества слоев. Результаты для различных наборов данных: FashionMNIST, CIFAR10 и CIFAR100.	40
1.6	Зависимость абсолютного значения разности функций потерь от доступного размера выборки, извлечение признаков изображений. Графики слева показывают уменьшение значений с увеличением размерности скрытого слоя. Графики справа показывают увеличение значений с увеличением количества слоев. Результаты для различных наборов данных: FashionMNIST, CIFAR10 и CIFAR100.	41

1.7	Переменное количество скрытых сверточных слоев L с фиксированным размером ядра $k = 3$ и количеством каналов $C = 6$. Анализ полученных графиков показывает немонотонный характер зависимости выходных значений от количества слоев.	42
1.8	Изменение размера ядра k при фиксированном количестве сверточных слоев L и количестве каналов $C = 6$. Данные демонстрируют немонотонный характер зависимости относительно размера ядра.	42
1.9	Изменение количества каналов C при фиксированном количестве сверточных слоев L и размере ядра $k = 3$. Зависимость значения от количества каналов имеет монотонный характер.	43
1.10	Изменение количества каналов C при фиксированном количестве сверточных слоев L и размере ядра $k = 3$. График показывает монотонную зависимость значения от позиции пулинга в сети. . .	43
1.11	Абсолютная разность потерь в зависимости от количества обучающих примеров в наборе данных, отображенная в двойном логарифмическом масштабе. Синяя линия представляет экспоненциальное скользящее среднее (EMA) желаемой зависимости, а серая линия соответствует линейному тренду.	44
2.1	Визуализация элементов гессиана для инициализированной модели с одним блоком Трансформера. Мы наблюдаем общую неоднородность величин, при этом блоки, соответствующие Values, имеют большие значения.	115
2.2	Визуализация элементов гессиана для модели, обученной в течение нескольких эпох , с одним блоком Трансформера. Мы наблюдаем общую неоднородность величин, при этом блок, соответствующий Values-Values, имеет наибольшие значения.	116

2.3	Нормы блоков параметров и нормы их гессианов, рассчитанные точно на одном батче, содержащем 128 примеров из обучающего набора данных MNIST.	116
2.4	Визуализация элементов (Queries).	116
2.5	Визуализация элементов (Keys).	117
2.6	Визуализация элементов (Values).	117
2.7	Визуализация элементов (LayerNorm).	117
2.8	Визуализация элементов (FeedForward).	117
3.1	Пример сдвига распределений при добавления объектов	134
3.2	Зависимость статистических значений различных методов	148
3.3	Анализ методов в зависимости от доступного размера выборки .	149
3.4	Сходимость предложенных функций $D(k)$ и $M(k)$ для синтетического набора данных регрессии, то есть модели линейной регрессии. Обе функции стремятся к нулю с увеличением размера выборки.	149
3.5	Сходимость предложенных функций $D(k)$ и $M(k)$ для синтетического набора данных классификации, то есть модели логистической регрессии. Обе функции стремятся к нулю с увеличением размера выборки.	150
3.6	Сходимость предложенных функций $D(k)$ и $M(k)$ для набора данных Liver Disorders. Обе функции стремятся к нулю с увеличением размера выборки.	150
3.7	Зависимость достаточного размера выборки от порога для трех наборов данных: синтетическая регрессия, синтетическая классификация и Liver Disorders. С увеличением значения порога ε достаточный размер выборки уменьшается. Это означает, что можно выбрать меньше объектов для достижения желаемых значений предложенных функций $D(k)$ и $M(k)$	151

3.8	Зависимость минимального собственного значения от размера доступной выборки. Оценочный график, изображенный синим цветом, для больших размеров выборки лежит выше оранжевого графика степенной функции. Такое асимптотическое поведение минимального собственного значения соответствует полученным теоретическим результатам.	153
3.9	Синтетический набор данных регрессии демонстрирует результаты сходимости предложенных функций оценки размера выборки. График слева соответствует расхождению Кульбака-Лейблера и стремится к нулю, в то время как график справа стремится к единице, отражая поведение функции схожести s-score.	154
3.10	Набор данных Liver Disorders демонстрирует результаты сходимости предложенных функций оценки размера выборки. Слева представлено расхождение Кульбака-Лейблера, которое стремится к нулю с увеличением размера выборки. Справа представлена функция схожести s-score, которая стремится к единице при приближении размера выборки к бесконечности.	154
3.11	Зависимость достаточного размера выборки от порогового параметра. Для S-достаточного размера выборки требуются более низкие значения порога. Таким образом, он оказывается более требовательным к этому значению.	155
3.12	Зависимость оцененного достаточного размера выборки t^* от доступного размера выборки t для каждого метода. Критерий на основе расхождения Кульбака-Лейблера является более консервативным и требует большего размера выборки. Критерий S-достаточности, напротив, предполагает, что может быть достаточен минимальный размер выборки.	157

4.1	Иллюстрация метода Белсли для анализа мультиколлинеарности параметров	164
4.2	Basic distillation	166
4.3	Distillation with domain adaptation	166
4.4	Качество прогноза при удаление параметров на выборке Wine	173
4.5	Влияние шума в начальных данных на шум выхода нейросети на выборке Wine	174
4.6	Качество прогноза при удаление параметров на выборке Boston	175
4.7	Влияние шума в начальных данных на шум выхода нейросети на выборке Boston	175
4.8	Качество прогноза при удаление параметров на синтетической выборке	176
4.9	Влияние шума в начальных данных на шум выхода нейросети на синтетической выборке	177
4.10	Сравнение примера объекта до и после преобразования.	180
4.11	Точность аппроксимации на тестовой выборке. Все результаты усреднены по 5 запускам.	181
4.12	Ошибка перекрестной энтропии между истинными и предсказанными студенческими метками на тестовой выборке. Все результаты усреднены по 5 запускам.	182
4.13	Ошибка перекрестной энтропии на тестовом наборе данных. Все результаты усреднены по 3 запускам.	183
4.14	Сравнение валидационной точности для различных методов инициализации	185
4.15	Зависимость валидационной точности от адвверсарного шума в данных	186
4.16	Зависимость валидационной точности от параметра интенсивности шума ε	186

Список таблиц

1.1	Описание наборов данных для классификации изображений	34
2.1	Гиперпараметры архитектур Vision Transformer (ViT), используемые в наших экспериментах	115
3.1	Описание выборок для анализа качества определения оптимального размера выборки	145
3.2	Эксперимент по оценке размера выборки для различных наборов выборок	146
3.3	Экспертные оценки гиперпараметров для разных методов оценки объема выборки	147
3.4	Сравнение предложенных методов определения размера выборки: на основе функций $D(k)$ и $M(k)$. Для каждой из предложенных функций пороговое значение ε подбиралось таким образом, чтобы начальное значение функции уменьшалось вдвое. Результаты получены для различных наборов данных с задачей регрессии. Прочерки в таблице означают, что исходный размер выборки недостаточен.	152
3.5	Описание выборок, используемых в эксперименте.	153
3.6	Экспериментальная оценка достаточного размера выборки, согласно предложенным методам для различных наборов данных.	156
4.1	Иллюстрация метода Белсли для анализа мультиколлинеарности параметров	165
4.2	Описание выборок для анализа метода задания порядка методом Белсли	173
4.3	Структура учителя	178
4.4	Структура студента	179
4.5	Набор данных ImageNet	179

4.6	Набор данных OPUS100	181
4.7	Качество моделей для компьютерного зрения	182
4.8	Качество моделей для NLP	183
4.9	Точность на валидационном наборе.	187

Список литературы

1. *Вапник В.Н., Червоненкис А.Я.* Теория распознавания образов: статистические проблемы обучения. — Nauka, 1974.
2. *Valiant L. G.* A theory of the learnable // *Commun. ACM.* — 1984. — nov. — Vol. 27, no. 11. — P. 1134–1142.
3. *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.
4. *MacKay David J. C.* Information Theory, Inference, and Learning Algorithms. — Copyright Cambridge University Press, 2003.
5. 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.
6. *Demidenko E.* Sample size determination for logistic regression revisited // *Statistics in medicine.* — 2006. — Vol. 26. — Pp. 3385–97.
7. *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.
8. *Lawrence J., Wolfson D., Berger R.* Sample Size Calculations for Binomial Proportions Via Highest Posterior Density Intervals // *Statistician.* — 1995. — Vol. 44. — Pp. 143–154.
9. *Kloek T.* Note on a large-sample result in specification analysis // *Econometrica.* — 1975. — Vol. 43. — Pp. 933–936.
10. *Lindley D.* The choice of sample size // *Statistician.* — 1997. — Vol. 46. — Pp. 129–138.

11. Motrenko A., Strijov V., Weber G. Sample Size Determination for Logistic Regression // *J. Comput. Appl. Math.* — 2014. — Vol. 255, no. C. — Pp. 743–752.
12. 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.
13. 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.
14. Self S., Mauritsen R. Power sample size calculations for generalized linear models // *Biometrics*. — 1988. — Vol. 44. — Pp. 79–86.
15. Self S., Mauritsen R., Ohara J. Power calculations for likelihood ratio tests in generalized linear models // *Biometrics*. — 1992. — Vol. 48. — Pp. 31–39.
16. Shieh G. On power and sample size calculations for likelihood ratio tests in generalized linear models // *Biometrics*. — 2000. — Vol. 56. — Pp. 1192–1196.
17. 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.
18. 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.
19. 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.
20. 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>.

21. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. — 2022. <https://arxiv.org/abs/2112.11446>.
22. *Hinton G., Vinyals O., Dean J.* Distilling the Knowledge in a Neural Network // NIPS Deep Learning and Representation Learning Workshop. — 2015.
23. Unifying distillation and privileged information / D. Lopez-Paz, L. Bottou, B. Schölkopf, V. Vapnik // ICLR. — 2016.
24. *Grabovoy A. V., Strijov V. V.* Bayesian Distillation of Deep Learning Models // Automation and Remote Control. — 2021. — Vol. 82, no. 11. — Pp. 1846–1856.
25. *Eldan Ronen, Shamir Ohad.* The Power of Depth for Feedforward Neural Networks // Conference on Learning Theory. — 2015. — 12.
26. *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.
27. *Xiao Han, Rasul Kashif, Vollgraf Roland.* Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. — 2017. <https://arxiv.org/abs/1708.07747>.
28. *Krizhevsky Alex.* Learning Multiple Layers of Features from Tiny Images // International Journal of Intelligence Science. — 2009. — Vol. 13. <https://api.semanticscholar.org/CorpusID:18268744>.
29. 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>.

30. *Kingma Diederik, Ba Jimmy.* Adam: A Method for Stochastic Optimization // International Conference on Learning Representations (ICLR). — San Diega, CA, USA: 2015.
31. *Wu Bichen, Xu Chenfeng, Dai Xiaoliang et al.* Visual Transformers: Token-based Image Representation and Processing for Computer Vision. — 2020.
32. *Sagun Levent, Evcı Utku, Guney V. Ugur et al.* Empirical Analysis of the Hessian of Over-Parametrized Neural Networks. — 2018. <https://arxiv.org/abs/1706.04454>.
33. *Skorski Maciej.* Chain Rules for Hessian and Higher Derivatives Made Easy by Tensor Calculus. — 2019. <https://arxiv.org/abs/1911.13292>.
34. *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>.
35. *Papyan Vardan.* The Full Spectrum of Deepnet Hessians at Scale: Dynamics with SGD Training and Sample Size. — 2019. <https://arxiv.org/abs/1811.07062>.
36. *Jacot Arthur, Gabriel Franck, Hongler Clement.* Neural Tangent Kernel: Convergence and Generalization in Neural Networks // Advances in Neural Information Processing Systems. — Vol. 31. — 2018.
37. 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>.
38. *Wu Yikai, Zhu Xingyu, Wu Chenwei et al.* Dissecting Hessian: Understanding Common Structure of Hessian in Neural Networks. — 2022.

39. *Singla Sahil, Wallace Eric, Feng Shi, Feizi Soheil.* Understanding Impacts of High-Order Loss Approximations and Features in Deep Learning Interpretation. — 2019.
40. *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.
41. *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.
42. 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>.
43. *Qin Zhen, Han Xiaodong, Sun Weixuan et al.* Toeplitz Neural Network for Sequence Modeling. — 2023. <https://arxiv.org/abs/2305.04749>.
44. *Gnacik Michal, Lapa Krystian.* Using Toeplitz Matrices to obtain 2D convolution. — 2022. — 10.
45. *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.
46. *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>.
47. *Aduenko Alexander.* Selection of multimodels in classification tasks: Ph.D. thesis / MIPT. — 2017. https://www.frccsc.ru/diss-council/00207305/diss/list/aduenko_aa.

48. *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.
49. *Quinlan J.* Learning With Continuous Classes // Proceedings of Australian Joint Conference on Artificial Intelligence. — World Scientific, 1992. — Pp. 343–348.
50. *Markelle Kelly, Rachel Longjohn, Kolby Nottingham.* The UCI Machine Learning Repository. <https://archive.ics.uci.edu>.
51. Preconditioned stochastic gradient Langevin dynamics for deep neural networks / C. Li, C. Chen, D. Carlson, L. Carin // AAAI. — 2016. — Pp. 1788–1794.
52. *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.
53. *Pearlmutter Barak A.* Fast exact multiplication by the Hessian // *Neural computation*. — 1994. — Vol. 6, no. 1. — Pp. 147–160.
54. 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.
55. *Chen Xiangning, Hsieh Cho-Jui.* Stabilizing differentiable architecture search via perturbation-based regularization // International conference on machine learning / PMLR. — 2020. — Pp. 1554–1565.
56. *Aeberhard S.* Wine Dataset. — <https://archive.ics.uci.edu/ml/datasets/Wine>.
57. *Grabovoy A. V., Strijov V. V.* Probabilistic Interpretation of the Distillation Problem // *Automation and Remote Control*. — 2022. — Vol. 83, no. 1. — Pp. 123–137.
58. ImageNet: A large-scale hierarchical image database. / J. Deng, W. Dong, R. Socher et al. // CVPR. — IEEE Computer Society, 2009. — Pp. 248–255.

59. Improving massively multilingual neural machine translation and zero-shot translation / Biao Zhang, Philip Williams, Ivan Titov, Rico Sennrich // *arXiv preprint arXiv:2004.11867*. — 2020.
60. 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.
61. 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.
62. 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.
63. *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>.
64. *Chen Tianqi, Goodfellow Ian, Shlens Jonathon.* Net2Net: Accelerating Learning via Knowledge Transfer. — 2015. <https://arxiv.org/abs/1511.05641>.
65. 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.
66. *Goodfellow Ian J, Shlens Jonathon, Szegedy Christian.* Explaining and harnessing adversarial examples // *arXiv preprint arXiv:1412.6572*. — 2014.
67. *Xiao Han, Rasul Kashif, Vollgraf Roland.* Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. — 2017.

68. 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.
69. *Magnus Jan R., Neudecker Heinz.* Matrix Differential Calculus with Applications in Statistics and Econometrics. — Chichester: Wiley, 1988.
70. *Kiselev N. S., Grabovoy A. V.* Sample Size Determination: Likelihood Bootstrapping // *Computational Mathematics and Mathematical Physics*. — 2025. — Vol. 65, no. 2. — Pp. 416–423.
71. *Zvereva Anna K., Kaprielova Mariam, Grabovoy Andrey.* AnomLite: Efficient binary and multiclass video anomaly detection // *Results in Engineering*. — 2025. — Vol. 25. — P. 104162.
72. *Kiselev Nikita, Grabovoy Andrey.* Sample size determination: posterior distributions proximity // *Computational Management Science*. — 2025. — Vol. 22, no. 1. — P. 1.
73. *Gritsai 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.
74. 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.
75. 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.
76. Text Reuse Detection in Handwritten Documents / A. V. Grabovoy, M. S. Kaprielova, A. S. Kildyakov et al. // *Doklady Mathematics*. — 2024.

77. Artificially Generated Text Fragments Search in Academic Documents / G. M. Gritsay, A. V. Grabovoy, A. S. Kildyakov, Yu V. Chekhovich // *Doklady Mathematics*. — 2024.
78. *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.
79. *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.
80. *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.
81. Грабовой А. В., Бахтееев О. Ю., Стрижков В. В. Ordering the set of neural network parameters // *Информатика и ее применение*. — 2020. — Vol. 14, no. 2.
82. Грабовой А. В., Бахтееев О. Ю., Стрижков В. В. Estimation of the relevance of the neural network parameters // *Информатика и ее применение*. — 2019.
83. 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.
84. 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.
85. Structure Extractor: Multilingual Extraction of Sections from Scientific Document / Ilia Kopanichuk, Artem Chashchin, Inga Ochnova et al. // 2025

37th Conference of Open Innovations Association (FRUCT). — IEEE: 2025. — Pp. 122–128.

86. Advacheck at GenAI Detection Task 1: AI Detection Powered by Domain-Aware Multi-Tasking / German Gritsai, Anastasia Voznuyk, 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.
87. 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.
88. 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.
89. 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.
90. 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.
91. 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.
92. 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.
93. *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.
 94. 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.
 95. 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.
 96. *Gritsay German, Grabovoy Andrey, Chekhovich Yury*. Automatic Detection of Machine Generated Texts: Need More Tokens // Ivannikov Memorial Workshop Proceedings 2022. — 2022.
 97. Зверева Анна Константиновна, Грабовой Андрей Валерьевич, Каприелова 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.
 98. 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.

99. 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.
100. 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.
101. 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.
102. 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 науч-

ной конференции МФТИ. — Прикладная математика и информатика. — Физматкнига: 2025.

103. *Мешков В. С., Киселев Н. С., Грабовой А. В.* Сходимость ландшафта свёрточных нейронных сетей: анализ матричных сетей на основе гесиана // Труды 67-й Всероссийской научной конференции МФТИ. — Прикладная математика и информатика. — Физматкнига: 2025.
104. *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.
105. *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.
106. *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.
107. *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.
108. *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.
109. *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.

110. *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.
111. *Баязитов К. М., Грабовой А. В., Стрижсов 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.
112. *Грабовой Андрей Валерьевич, Стрижсов Вадим Викторович*. 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.
113. *Грабовой А. В., Стрижсов В. В.* 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.
114. *Грабовой Андрей Валерьевич, Стрижсов Вадим Викторович*. 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.
115. *Грабовой А. В., Стрижсов В. В.* Задача обучения с экспертом для построение интерпретируемых моделей машинного обучения // Тезисы докла-

- дов 13-й Международной конференции "Интеллектуализация обработки информации". — РАН Москва: 2020. — С. 16–17.
116. 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.
117. *Грабовой Андрей Валерьевич, Стрижов Вадим Викторович*. Анализ априорных распределений в задаче смеси экспертов // Труды 62-й Всероссийской научной конференции МФТИ. — Прикладные математика и информатика. — МФТИ М: 2019.
118. Численные методы оценки оптимального объёма 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.
119. *Грабовой А. В., Бахтеев О. Ю., Стрижов В. В.* 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.
120. *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.
121. *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.

122. *Petersen Kaare Brandt, Pedersen Michael Syskind*. The Matrix Cookbook. — <https://www2.imm.dtu.dk/pubdb/doc/imm3274.pdf>. — 2012. — Version November 15, 2012.
123. *Krizhevsky Alex, Nair Vinod, Hinton Geoffrey*. CIFAR-10 (Canadian Institute for Advanced Research). <http://www.cs.toronto.edu/~kriz/cifar.html>.

Приложение А

Дополнительные Леммы и утверждения

Лемма 31 (Производная умножения матричнозначных функций). *Пусть заданы $\mathbf{A}(\mathbf{X}) \in \mathbb{R}^{p \times r}$ и $\mathbf{B}(\mathbf{X}) \in \mathbb{R}^{r \times q}$ матрицезначные функции переменной \mathbf{X} , тогда*

$$\frac{\partial \mathbf{A}(\mathbf{X})\mathbf{B}(\mathbf{X})}{\partial \mathbf{X}} = (\mathbf{A} \otimes \mathbf{I}_q) \frac{\partial \mathbf{B}}{\partial \mathbf{X}} + (\mathbf{I}_p \otimes \mathbf{B}^\top) \frac{\partial \mathbf{A}}{\partial \mathbf{X}}$$

Доказательство. Применить цепное правило для вычисления производной сложной функции, а затем объединим его со свойством 9

$$\begin{aligned} \frac{\partial \mathbf{A}(\mathbf{X})\mathbf{B}(\mathbf{X})}{\partial \mathbf{X}} &= \frac{\partial \mathbf{AB}}{\partial \mathbf{B}} \frac{\partial \mathbf{B}}{\partial \mathbf{X}} + \frac{\partial \mathbf{AB}}{\partial \mathbf{A}} \frac{\partial \mathbf{A}}{\partial \mathbf{X}} = \\ &= \frac{\partial \mathbf{AB}\mathbf{I}_q}{\partial \mathbf{B}} \frac{\partial \mathbf{B}}{\partial \mathbf{X}} + \frac{\partial \mathbf{I}_p\mathbf{AB}}{\partial \mathbf{A}} \frac{\partial \mathbf{A}}{\partial \mathbf{X}} = \\ &= (\mathbf{A} \otimes \mathbf{I}_q) \frac{\partial \mathbf{B}}{\partial \mathbf{X}} + (\mathbf{I}_p \otimes \mathbf{B}^\top) \frac{\partial \mathbf{A}}{\partial \mathbf{X}} \end{aligned}$$

□

Лемма 32 (Теорема идентификации для построчной векторизации). *Пусть отображение $\mathbf{F} : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{p,q}$ является дифференциальной 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 [69] для векторизации по столбцам.

Лемма 33 (Производная квадрата Адамара). Для матрицы $\mathbf{A} \in \mathbb{R}^{m \times n}$, производная поэлементного квадрата равна

$$\frac{\partial \mathbf{A}^{\circ 2}}{\partial \mathbf{A}} = 2 \cdot \text{diag}(\text{vec}_r(\mathbf{A})).$$

Доказательство. Используя определение 18, а именно $(\mathbf{A}^{\circ 2})_{ij} = (\mathbf{A}_{ij})^2$. Выполняя поэлементное взятие дифференциала, получаем, что:

$$d(\mathbf{A}^{\circ 2}) = 2\mathbf{A} \circ d\mathbf{A}.$$

Далее, применяя оператор vec_r и используя свойство 10 получаем 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 32 получаем 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 34 (Производная корня Адамара). Для матричнозначной функции $\mathbf{A} \in \mathbb{R}_+^{m \times n}$ с положительными элементами, производная поэлементного корня равна

$$\frac{\partial \mathbf{A}^{\circ \frac{1}{2}}}{\partial \mathbf{A}} = \frac{1}{2}\text{diag}^{-1}(\text{vec}_r^{\circ \frac{1}{2}}(\mathbf{A})).$$

Доказательство. Аналогично, доказательству леммы 33 получаем $d(\mathbf{A}^{\circ 1/2}) = \frac{1}{2}\mathbf{A}^{\circ -1/2} \circ d\mathbf{A}$, откуда в векторном виде получаем, 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 35 (Производная обратной матрицы). Пусть задана обратимая квадратная матрица $\mathbf{D} \in \mathbb{R}^{n \times n}$, тогда производная операции обращения равно:

$$\frac{\partial \mathbf{D}^{-1}}{\partial \mathbf{D}} = -\mathbf{D}^{-1} \otimes \mathbf{D}^{-\top}.$$

Доказательство. По определению из [122] и [69]:

$$d(\mathbf{D}^{-1}) = -\mathbf{D}^{-1}(d\mathbf{D})\mathbf{D}^{-1}.$$

Используя оператор vec_r и свойство 11, получаем

$$\text{vec}_r(-\mathbf{D}^{-1}(d\mathbf{D})\mathbf{D}^{-1}) = (-\mathbf{D}^{-1} \otimes \mathbf{D}^{-\top})\text{vec}_r(d\mathbf{D}),$$

причем, используя лемму 32 получаем:

$$\text{vec}_r(d\mathbf{D}^{-1}) = \frac{\partial \text{vec}_r \mathbf{D}^{-1}}{\partial \text{vec}_r \mathbf{D}} \text{vec}_r(d\mathbf{D}).$$

Следовательно получаем выражение, которое заканчивает доказательство леммы $\frac{\partial \text{vec}_r \mathbf{D}^{-1}}{\partial \text{vec}_r \mathbf{D}} = (-\mathbf{D}^{-1} \otimes \mathbf{D}^{-\top})$ \square

Лемма 36 (Производная $\text{diag}(\cdot)$). Для вектора $\mathbf{v} \in \mathbb{R}^{L \times 1}$, производная оператора диагонализации является:

$$\frac{\partial \text{diag}(\mathbf{v})}{\partial \mathbf{v}} = (\mathbf{e}_1 \otimes \mathbf{e}_1 \quad \dots \quad \mathbf{e}_L \otimes \mathbf{e}_L),$$

где вектора \mathbf{e}_i являются базисными в пространстве \mathbb{R}^L .

Доказательство. По определению 18 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, применяя оператор построчной векторизации, получаем:

$$\text{vec}_r(\mathbf{E}_{i,i}) = \mathbf{e}_i \otimes \mathbf{e}_i.$$

Итого, применяя для всех $i = 1, \dots, L$, матрица Якоби принимает вид:

$$\frac{\partial \text{diag}(\mathbf{v})}{\partial \mathbf{v}} = (\mathbf{e}_1 \otimes \mathbf{e}_1 \quad \dots \quad \mathbf{e}_L \otimes \mathbf{e}_L).$$

\square

Лемма 37 (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 17.

Доказательство. Объединяя аналогичное свойство из [69] для постолбцовой векторизации с правилом соединения столбцов и строк 5 и 17, получаем утверждение теоремы. \square

Лемма 38 (Производная произведения Кронекера матричнозначных функций). *Пусть заданы матричнозначные функции $\mathbf{A}(\mathbf{X}) \in \mathbb{R}^{n \times q}$ и $\mathbf{B}(\mathbf{X}) \in \mathbb{R}^{p \times r}$ матрицы \mathbf{X} , тогда*

$$\frac{\partial \mathbf{A}(\mathbf{X}) \otimes \mathbf{B}(\mathbf{X})}{\partial \mathbf{X}} = (\mathbf{I}_n \otimes \mathbf{K}_{p,q} \otimes \mathbf{I}_r) \left((\text{vec}_r \mathbf{A} \otimes \mathbf{I}_{pr}) \frac{\partial \mathbf{B}}{\partial \mathbf{X}} + (\mathbf{I}_{nq} \otimes \text{vec}_r \mathbf{B}) \frac{\partial \mathbf{A}}{\partial \mathbf{X}} \right).$$

Доказательство. Применить цепное правило для вычисления производной сложной функции, а затем объединим его со свойством 12

$$\begin{aligned} \frac{\partial \mathbf{A}(\mathbf{X}) \otimes \mathbf{B}(\mathbf{X})}{\partial \mathbf{X}} &= \frac{\partial \mathbf{A} \otimes \mathbf{B}}{\partial \mathbf{B}} \frac{\partial \mathbf{B}}{\partial \mathbf{X}} + \frac{\partial \mathbf{A} \otimes \mathbf{B}}{\partial \mathbf{A}} \frac{\partial \mathbf{A}}{\partial \mathbf{X}} = \\ &= (\mathbf{I}_n \otimes \mathbf{K}_{p,q} \otimes \mathbf{I}_r) (\text{vec}_r \mathbf{A} \otimes \mathbf{I}_{pr}) \frac{\partial \mathbf{B}}{\partial \mathbf{X}} + \\ &\quad + (\mathbf{I}_n \otimes \mathbf{K}_{p,q} \otimes \mathbf{I}_r) (\mathbf{I}_{nq} \otimes \text{vec}_r \mathbf{B}) \frac{\partial \mathbf{A}}{\partial \mathbf{X}} = \\ &= (\mathbf{I}_n \otimes \mathbf{K}_{p,q} \otimes \mathbf{I}_r) \left((\text{vec}_r \mathbf{A} \otimes \mathbf{I}_{pr}) \frac{\partial \mathbf{B}}{\partial \mathbf{X}} + (\mathbf{I}_{nq} \otimes \text{vec}_r \mathbf{B}) \frac{\partial \mathbf{A}}{\partial \mathbf{X}} \right). \end{aligned}$$

\square

Лемма 39. *Пусть задана единичная матрица $\mathbf{A} = \mathbf{1}_{L \times L}$. Тогда ее спектральная норма принимает следующее значение:*

$$\|\mathbf{A}\|_2 = L$$

Доказательство. Используя основные свойства линейной алгебры, получаем $\text{tr}(\mathbf{A}) = L$ и $\text{rank}(\mathbf{A}) = 1 = \dim(\text{Im}(\mathbf{X}))$. Следовательно, используя $\dim(\text{Im}(\mathbf{X})) + \dim(\text{Ker}(\mathbf{X})) = L$, получаем $\dim(\text{Ker}(\mathbf{X})) = L - 1$. 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, получаем, что $\|\mathbf{A}\|_2 = L$, в соответствии с определением 19. \square