Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа прикладной математики и информатики
Кафедра интеллектуальных систем

Мешков Владислав Сергеевич

Анализ сходимости оптимизационной поверхности сверточных нейросетевых моделей на основе Гессиана функции потерь

01.03.02 — Прикладная математика и информатика

Выпускная квалификационная работа бакалавра

Научный руководитель:
Грабовой Андрей Валериевич,
канд. физ.-мат. наук

Москва — 2025

Аннотация

Аннотация — Гессиан нейронной сети является важным аспектом для понимания ландшафта потерь и характеристики сетевой архитектуры. Матрица Гессе фиксирует важную информацию о кривизне, чувствительности и локальном поведении функции потерь. Наша работа предлагает метод, который улучшает понимание локального поведения функции потерь и может использоваться для анализа поведения нейронных сетей, а также для интерпретации параметров в этих сетях. В этой работе мы рассматриваем подход к исследованию свойств глубокой нейронной сети с использованием Гессеана. Мы предлагаем метод оценки нормы матрицы Гессе для определенного типа нейронных сетей, таких как сверточные. Мы получили результаты как для одномерных, так и для двумерных сверток, а также для полностью связанной головы в этих сетях. Наш эмпирический анализ подтверждает эти выводы, демонстрируя сходимость в ландшафте функции потерь. Мы оценили гессианскую норму для нейронных сетей, представленных как произведение матриц, и рассмотрели, как эта оценка влияет на ландшафт функции потерь.

Содержание

Ландшафт функции потерь играет ключевую роль в понимании свойств параметров глубоких нейронных сетей [1, 2, 3]. Его анализ позволяет исследовать различия между архитектурами [4, 5], влияние функций активации [6], а также свойства локальных и глобальных минимумов [7, 8, 9].

Многочисленные исследования посвящены изучению ландшафта функции потерь в современных архитектурах. В частности, работа [10] анализирует самоконтролируемые Vision Transformers (ViT) через призму ландшафта потерь, тогда как [11] исследует ViT и MLP-Mixers с точки зрения геометрии потерь, стремясь улучшить эффективность обучения и обобщающую способность моделей. Исследование [12] предлагает метод визуализации, дающий ценные инсайты о ландшафтах потерь нейронных сетей. Работа [13] сокращает вычислительные затраты на подобный анализ. Такие исследования, как [14, 15], экспериментально исследуют поверхность потерь глубоких нейронных сетей, включая траектории различных алгоритмов оптимизации. Другие работы фокусируются на спектре матрицы Гессе [16, 17, 18] или выводят верхние оценки ее ранга. Например, [19] оценивает ранги блоков Гессе через их декомпозицию.

В данной работе мы выводим теоретические оценки для спектральной нормы матрицы Гессе. Мы показываем, что спектральная норма Гессе дает верхнюю оценку на разность средних значений функции потерь при добавлении нового объекта в выборку. Это открывает возможности для приложений, связанных с определением размера выборки [20], и позволяет глубже понять, как параметры влияют на матрицу Гессе.

Основные цели исследования:

- Теоретический анализ декомпозиции матрицы Гессе на линейные компоненты с конкретными приложениями к сверточным сетям
- Исследование поведения функции потерь в окрестности оптимума и зависимости ландшафта потерь от архитектуры сети
- Анализ влияния нормы параметров, их количества и пространственного распределения в сети на процесс обучения [21, 22, 23]
- Оценка абсолютной разности между средними значениями функции потерь на последовательных шагах обучения, вытекающая из нашей оценки нормы Гессе

Основные результаты работы:

- Предложен метод декомпозиции матрицы Гессе на линейные компоненты и применен для оценки ее нормы
- Продемонстрировано применение результатов к сверточным архитектурам и установлены связи между параметрами и их оценками
- Подтверждена справедливость теоретических результатов экспериментами по классификации изображений с использованием сверточных сетей

1. Обзор литературы

Ландшафт функции потерь в нейронных сетях. В литературе ландшафт функции потерь исследуется с различных точек зрения. В работе [24] установлена связь между количеством классов и направлениями высокой положительной кривизны. Исследование [5] показывает, что ландшафт потерь двухслойной сети с функцией активации ReLU обладает хорошими свойствами при большом количестве скрытых узлов. В работе [25] предложена модель, описывающая необходимые топологические свойства ландшафта потерь для линейной связности мод (LMC). Авторы [26] предлагают классификацию кривых поверхности потерь, построенных вдоль направлений гауссовского шума. Свойства нейронных сетей и спектры их матриц Гессе вблизи порога интерполяции изучены в [27]. Исследования [10, 11] анализируют архитектуру ViT через локальный ландшафт функции потерь. Однако эти работы ограничены конкретными архитектурами.

Анализ матрицы Гессе. Декомпозиция матрицы Гессе на составляющие является ключевым инструментом изучения ее свойств. В работе [28] предложена гипотеза декомпозиции послойных матриц Гессе в виде произведения Кронекера двух матриц меньшей размерности. В исследовании [29] представлено правило дифференцирования для Гессе и полезные тензорные вычисления. Однако эти результаты не были распространены на анализ ландшафта функции потерь через норму матрицы Гессе.

Собственные значения и спектр матрицы Гессе. Спектр матрицы Гессе играет важную роль в понимании структуры ландшафта функции потерь. В работе [16] разработан инструмент для изучения эволюции полного спектра Гессе в процессе оптимизации. Авторы [30] предлагают эффективный метод аппроксимации спектра Гессе нейронных сетей через его декомпозицию на компоненты. Исследование [31] демонстрирует, что распределение собственных значений может состоять из двух частей: сосредоточенной около нуля и удаленной от нуля. В работе [32] выявлена и проанализирована формальная структура "класс/межкласс лежащая в основе многих особенностей спектров глубоких сетей. Проблема выбросов в спектре Гессе рассмотрена в [33], где предпринята попытка их объяснения. Авторы [18] дают характеристику спектров Гессе для широкого класса нелинейных моделей. В исследовании [21] разработан инструмент для отслеживания эволюции спектра Гессе в процессе обучения.

Наиболее полное освещение темы представлено в работе [34], где предложен метод исследования спектра при изменении размера выборки, однако анализ ограничен только полносвязными нейронными сетями.

2. Предварительные сведения

2.1. Общие обозначения

В данном разделе вводятся основные обозначения и предположения, используемые в работе. Аналогично [35], мы рассматриваем матричные производные, используя построчное векторное представление (vec_r). Для заданных матриц $\mathbf{X} \in \mathbb{R}^{m \times n}$ и $\mathbf{Y} \in \mathbb{R}^{p \times q}$ определим:

$$\frac{\partial \mathbf{X}}{\partial \mathbf{Y}} := \frac{\partial vec_r \mathbf{X}}{\partial (vec_r \mathbf{Y})^\intercal}. \tag{1}$$

Для тензоров более высокой размерности определение аналогично.

Рассматривается задача классификации на K классов, где $p(\mathbf{y}|\mathbf{x})$ - вероятность отображения входного вектора $\mathbf{x} \in \mathcal{X}$ в соответствующий выход $\mathbf{y} \in \mathcal{Y} = \mathbb{R}^K$ (one-hot векторы). Нейронная сеть $f_{\boldsymbol{\theta}}$ параметризуется вектором $\boldsymbol{\theta} \in \mathbb{R}^p$.

Пусть задана независимая выборка размера m:

$$\mathfrak{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1,\dots,m}.$$

Функция потерь на одном объекте $\mathbf{x}_i$:

$$\ell_i(\boldsymbol{\theta}) := \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), \mathbf{y}_i).$$

Эмпирическая функция потерь для первых k элементов:

$$\mathcal{L}_k(\boldsymbol{\theta}) := \frac{1}{k} \sum_{i=1}^{k} \ell_i(\boldsymbol{\theta}), \quad \mathcal{L}(\boldsymbol{\theta}) := \mathcal{L}_m(\boldsymbol{\theta}).$$

Основная цель - оценка эмпирической функции потерь на всей выборке:

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

Особый интерес представляет оценка разности:

$$\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta}) = \frac{1}{k+1} \big(\ell_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})\big).$$

Введем ключевые определения производных:

$$J(\boldsymbol{\theta}) := J_{f_{\boldsymbol{\theta}}(\mathbf{x})} = (\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\mathbf{x}))^{\intercal}.$$

Матрица Якоби нейронной сети:

$$J(\boldsymbol{\theta}) := J_{f_{\boldsymbol{\theta}}(\mathbf{x})} = (\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\mathbf{x}))^{\intercal}.$$

Полная матрица Гессе:

$$\mathbf{H}^{(k)}(\boldsymbol{\theta}) := \nabla^2_{\boldsymbol{\theta}} \mathcal{L}_k(\boldsymbol{\theta}) = \frac{1}{k} \sum_{i=1}^{k} \nabla^2_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}).$$

2.2. Основное Предположение

Для сравнения функций потерь в одной точке введем следующее предположение:

Предположение 1. *Пусть $\boldsymbol{\theta}^*$ - локальный минимум как для $\mathcal{L}_k(\boldsymbol{\theta})$, так и для $\mathcal{L}_{k+1}(\boldsymbol{\theta})$. В частности, это означает, что $\nabla_{\boldsymbol{\theta}} \mathcal{L}_k(\boldsymbol{\theta}^*) = \nabla_{\boldsymbol{\theta}} \mathcal{L}_{k+1}(\boldsymbol{\theta}^*) = 0$.*

Данное предположение позволяет изучать поведение ландшафта, используя лишь одну точку.

2.3. Аппроксимация и декомпозиция

Используя это предположение и квадратичную аппроксимацию, в работе [34] показано, что для изучения локального поведения можно использовать разложение Тейлора второго порядка:

$$\begin{aligned}\left|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})\right| \approx \frac{1}{k+1}\left|\ell_{k+1}(\boldsymbol{\theta}^{)} - \mathcal{L}_k(\boldsymbol{\theta}^{)}\right| + \\ \frac{1}{k+1} \left\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\right\|^2 \left\| \mathbf{H}_{k+1}(\boldsymbol{\theta}^*) - \frac{1}{k} \sum_{i=1}^{k} \mathbf{H}_i(\boldsymbol{\theta}^*) \right\|. \quad (2)\end{aligned}$$

Согласно [36], применяя правило дифференцирования сложной функции,

можно декомпозировать матрицу Гессе:

$$\mathbf{H} = \mathbf{H}_O + \mathbf{H}_F = J(\boldsymbol{\theta})^\intercal \left[\nabla^2_{f_{\boldsymbol{\theta}}} \ell(\boldsymbol{\theta})\right] J(\boldsymbol{\theta}) + \sum_{c=1}^{K} \left[\nabla_{f_{\boldsymbol{\theta}}} \ell(\boldsymbol{\theta})\right] c \nabla^2 \boldsymbol{\theta} f^c_{\boldsymbol{\theta}}(\mathbf{x}).$$

Как показано в [24, 37], вблизи локального минимума можно учитывать только член $\mathbf{H}_O$, поскольку средний градиент близок к нулю, и членом $\mathbf{H}_F$ можно пренебречь. На основе этой аппроксимации мы будем рассматривать норму матрицы $\mathbf{H}_O$:

$$\|\mathbf{H}\| \approx \left\| J(\boldsymbol{\theta})^\intercal \left[\nabla^2_{f_{\boldsymbol{\theta}}} \ell(\boldsymbol{\theta})\right] J(\boldsymbol{\theta}) \right\|. \tag{3}$$

2.4. Разложение матрицы Гессе

Мы используем термин "внешне-произведенная"матрица Гессе для члена $\mathbf{H}_O$, следуя [38]. Именно в такой форме матрица Гессе наиболее удобна для анализа. Отметим, что $\nabla^2_{f_{\boldsymbol{\theta}}} \ell(\boldsymbol{\theta})$ зависит только от функции потерь:

- Для MSE: $\nabla^2_{f_{\boldsymbol{\theta}}} \ell(\boldsymbol{\theta}) = \mathbf{I}$
- Для кросс-энтропии: $\nabla^2_{f_{\boldsymbol{\theta}}} \ell(\boldsymbol{\theta}) = diag(\mathbf{p}) - \mathbf{p}\mathbf{p}^\intercal$, где $\mathbf{p} := \text{SoftMax}(\mathbf{z})$

Выбор функции потерь влияет лишь на мультипликативную константу, не меняя общего анализа. Оценка нормы произведения матриц сводится к произведению их норм, что приводит к квадратичной зависимости нормы Гессе от Якобиана. Как отмечено в [39], Якобиан содержит важную структурную информацию о сети, что будет исследовано далее.

3. Представление сети в виде произведения матриц

Пусть $f_{\boldsymbol{\theta}}(\mathbf{x})$ представляет собой композицию $L+1$ слоев с активациями ReLU:

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

Здесь $\mathbf{T}^{(p+1)}$ - линейный оператор (или его матрица), σ - функция активации ReLU. Промежуточные результаты можно представить как:

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

где выходные логиты $f_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{z} := \mathbf{z}^{(L+1)}$, и вход $\mathbf{x}^{(0)} := \mathbf{x}$.

Пусть $\mathbf{\Lambda}^{(p+1)} := diag(\mathbf{x}^{(p+1)} > 0)$ - входозависимая матрица. Тогда $f_{\boldsymbol{\theta}}(\mathbf{x})$ можно представить в виде произведения (возможно, входозависимых) матриц:

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{T}^{(L+1)}\mathbf{\Lambda}^{(L)} \dots \mathbf{\Lambda}^{(1)}\mathbf{T}^{(1)}\mathbf{x}. \tag{4}$$

Вектор весов рассматривается как $\boldsymbol{\theta} = col(\mathbf{W}^{(L+1)}, \dots, \mathbf{W}^{(1)})$, где $\mathbf{T}^{(p)}$ дифференцируем и параметризуется частью $\mathbf{W}^{(p)}$. Тогда производная слоя по его параметрам может быть определена как:

где матрица $\mathbf{Q}^{(p)}$ полностью описывает расположение параметров в p-м слое.

Для упрощения дальнейших формул определим:

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

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

Используя эти обозначения, можно переписать:

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

Объединенные матрицы $\mathbf{G}^{(p)}$ и $\mathbf{R}^{(p)}$ дают:

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

Матрица Гессе нейронной сети по логитам в случае функции потерь кросс-энтропии:

$$\mathbf{A} := \nabla^2_{\mathbf{z}}\ell = diag(\mathbf{p}) - \mathbf{p}\mathbf{p}^\intercal,$$

where $\mathbf{p} := \text{softmax}(\mathbf{z})$.

3.1. Структура матрицы Гессе

Согласно работам [19, 28, 34], мы можем декомпозировать outer-product ($\mathbf{H}_O$) матрицу Гессе на более простые компоненты, в частности, нам нужно декомпозировать только матрицу Якоби.

Рассмотрим ключевые леммы данной работы, которые описывают декомпозицию матрицы Гессе в произведение 5 матриц и использование этого представления для оценки нормы. Доказательства приведены в приложениях 8.1. и 8.2..

Лемма 1. *Если наша сеть* $f_{\boldsymbol{\theta}}(\mathbf{x})$ *может быть представлена как* (4)*, то* $\mathbf{H}_O(\boldsymbol{\theta}) = \mathbf{Q}^T\mathbf{F}^T\mathbf{A}\mathbf{F}\mathbf{Q}$.

Лемма 2. *Пусть нейронная сеть* $f_{\boldsymbol{\theta}}(\mathbf{x})$ *представляется в виде* (4)
Пусть $\forall p: \ \left\|\mathbf{Q}^{(p)}\right\| \leqslant q, \ \left\|\mathbf{T}^{(p)}\right\|^2 \leqslant w^2_{\mathbf{T}}$.
Тогда имеет место:
$\|\mathbf{H}_O\| \leqslant \sqrt{2}q^2 \|\mathbf{x}\|^2 (L+1) w^{2L}_{\mathbf{T}}$.

Эти леммы используются для оценки норм матрицы Гессе.

4. Сверточные сети

4.1. Одномерные свертки

В данном разделе для простоты сохраняем обозначение $\mathbf{T}^{(p)}$ для одномерных сверток и поясняем их представление в виде линейных операторов. Как известно, сверточные сети часто могут быть представлены линейными сверточными нейронными сетями (LCN), что обычно относится к представлению CNN через матрицы Тёплица [40, 41].

В работе используется обозначение для матриц Тёплица из [19], где авторы нашли специфический вид матрицы $\mathbf{Q}^{(p)}$ согласно структуре одномерной матрицы Тёплица.

Наша одномерная сверточная сеть: $f_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{T}^{(L+1)} * (\sigma(\dots(\sigma(\mathbf{T}^{(1)} * \mathbf{x}))\dots)$ где операция $*$ означает свертку.

Пусть C_p - количество каналов после p-го слоя, d_p - размер последовательности. Тогда $\mathbf{x}^{(p)} \in \mathbb{R}^{C_p \times d_p}$, $\mathbf{T}^{(p)}$ - слой одномерной свертки с ядром $\mathbf{W}^{(p)} \in \mathbb{R}^{C_{p-1} \times C_p \times k_p}$.

Для упрощения обозначений заменяем $\mathbf{x}^{(p)}$ на $vec(\mathbf{x}^{(p)}) \in \mathbb{R}^{(C_p d_p)}$. Теперь имеем:

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

Основные результаты для одномерных сверток, использующие матрицы Тёплица для вычисления $\mathbf{T}^{(p)}$ и $\mathbf{Q}^{(p)}$ (наши обозначения для сверток и матриц Тёплица упрощены), а также Леммы 1 и 2. Подробности доказательства - в приложении 8.3..

Теорема 1. *Рассмотрим сеть $f_{\mathbf{x}} = C_{\mathbf{W}^{(L+1)}} \circ \sigma \circ \cdots \circ \sigma \circ C_{\mathbf{W}^{(1)}}$, где $C_{\mathbf{W}^{(i)}}$ - одномерная свертка с ядром $\mathbf{W}^{(i)}$, без дополнения и шагом 1. Пусть заданы верхние границы: $C_l \leqslant C$, $k_i \leqslant k$, $d_i \leqslant d_1 := d$, $|\mathbf{W}^{(p)}_{i,j,k}|^2 \leqslant w^2$. Тогда можем оценить норму outer-product матрицы Гессе:*

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

Применяя эту теорему к разности потерь, аналогично [34], получаем:

Следствие. *Пусть $\boldsymbol{\theta}$ находится в R-окрестности оптимума: $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leqslant R$. Функция потерь ограничена константой: $\exists W_l > 0 : \forall i |\ell_i| \leqslant W_l$. Все объекты в*

выборке ограничены: $\exists W_x \forall i |x_i| \leqslant W_x$. *Тогда в условиях теоремы 1 и наших предположений имеем:*

$$\left|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})\right| \leqslant \frac{2}{k+1} W_\ell + \\ + \frac{2}{k+1} R^2 \sqrt{2} d^2 W_x^2 (L+1)(C^2 w^2 k d)^L.$$

Как видно, эта оценка чрезмерно завышена по сравнению с реальной нормой. Однако можно предположить, что зависимость от числа каналов, норм весов и размеров действительно соответствует представленной выше. Если параметры свертки уже достаточны для обучения, то увеличение, например, размера ядра свертки k неоправданно увеличит оценку примерно в $(1 + \frac{\Delta k}{k})^L$ раз, что может привести к замедлению скорости сходимости без существенного улучшения качества.

4.2. Двумерные свертки

Рассматриваем двумерные сверточные сети, сохраняя обозначение $\mathbf{T}^{(p)}$ для слоев сверточной сети. $\mathbf{x} \in \mathbb{R}^{m \times n \times C}$ - входное изображение размером (m, n) с C каналами. $\mathbf{x}^{(l)} \in \mathbb{R}^{m_i \times n_i \times C_i}$ - вход $(l+1)$-го слоя. $\mathbf{W}^{(l)} \in \mathbb{R}^{C_{l-1} \times C_l \times k_l^1 \times k_l^2}$ - свертка с размерами ядра (k_l^1, k_l^2), числом входных и выходных каналов C_{l-1}, C_l соответственно.

Аналогично разделу A, используем $vec(\mathbf{x}) \in \mathbb{R}^{m_i n_i C_i}$ вместо $\mathbf{x} \in \mathbb{R}^{m_i \times n_i \times C_i}$. Операция свертки рассматривается для векторизованного входа. Можно использовать тот же подход с матрицами Тёплица, что и в [42], но проще использовать специфическую матрицу $\mathbf{T}^{(p)}$, строка которой состоит из элементов $\mathbf{W}^{(p)}_{*,c_2,*,*}$ для c_2-го канала.

Теорема 2. *Пусть сеть* $f_{\mathbf{x}} = C_{\mathbf{W}^{(L+1)}} \circ \cdots \circ C_{\mathbf{W}^{(1)}}$, *где* $C_{\mathbf{W}^{(l)}}$ *- двумерная свертка с ядром* $\mathbf{W}^{(i)}$, *без дополнения и шагом 1. Пусть заданы верхние границы:* $C_l \leqslant C$, $k_i \leqslant k$, $m_i \leqslant m_1 := m$, $n_i \leqslant n_1 := n$, $|\mathbf{W}^{(p)}_{i,j,k}|^2 \leqslant w^2$. *Тогда норма матрицы Гессе:*

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

где $q^2 = C^2 k^2 mn$.

Следствие. *Пусть* $\boldsymbol{\theta}$ *в* R*-окрестности оптимума:* $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leqslant R$. *Функция потерь ограничена:* $\exists W_l > 0 : \forall i |\ell_i| \leqslant W_l$. *Объекты ограничены:* $\exists W_x \forall i |x_i| \leqslant$

W_x. *Тогда в условиях теоремы 2:*

$$\left|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})\right| \leqslant \frac{2}{k+1} W_\ell + \\ + \frac{2}{k+1} R^2 \sqrt{2} q^2 W_x^2 (L+1)(C^2 k^2 w^2 mn)^L,$$

где $q^2 = C^2 k^2 mn$.

Как можно видеть, сеть в этом примере состоит исключительно из сверточных слоев, что является редким явлением на практике. В 5 мы обсуждали случай добавления полностью связанной головы к сверточной нейронной сети. Эти результаты позволяют нам построить гипотезу о том, что гессианова норма может быть экспоненциальной функцией числа слоев, а также зависеть от размера ядра, размеров изображения и каналов как описано выше. Главным недостатком этих результатов является то, что на них не влияет уменьшение размеров после сверток и они зависят только от верхних границ параметров.

4.3. Пуллинги

Мы также предоставляем результаты, связанные с добавлением пула в сеть. Сначала речь идет о максимальном пуле, доказательство можно найти в приложении 8.5..

Лемма 3. *Пусть сверточная нейронная сеть* $f_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{T}^{(L+1)} \boldsymbol{\Lambda}^{(L)} \dots \boldsymbol{\Lambda}^{(1)} \mathbf{T}^{(1)} \mathbf{x}$ *стодержит MaxPool2D в слое* $\boldsymbol{\Lambda}^{(l)}$ *с ядром* $k_{\text{pool}} \times k_{\text{pool}}$ *вместо ReLU активации. Тогда* $\|\mathbf{H}_O\| \leqslant \sqrt{2}\,\|\mathbf{x}\|^2 q^2 \left(\frac{1}{k_{\text{pool}}^2}\right)^{L-l+2} (L+1)(k^2 C^2 w^2 mn)^L$,
где $q^2 = mnC^2k^2$.

И результат заключается в добавлении среднего пула, доказательство можно найти в 8.6.

Лемма 4. *Пусть Сверточная сеть*
$f_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{T}^{(L+1)} \boldsymbol{\Lambda}^{(L)} \dots \boldsymbol{\Lambda}^{(1)} \mathbf{T}^{(1)} \mathbf{x}$
содержит AvgPool2D в качестве слоя $\boldsymbol{\Lambda}^{(l)}$ *вместо ReLU активации с ядром размера* $k_{\text{pool}} \times k_{\text{pool}}$. *Тогда*

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

где $q^2 = mnC^2k^2$.

4.4. Полносвязная голова

Видно, что наша сеть состояла исключительно из сверточных слоев, что почти никогда не случается, рассмотрим сеть, которая последними P слоями является полносвязной. Докказательство можно найти в 8.7..

Лемма 5. *Пусть сверточная сеть с полносвязной классификационной головой размера p:*

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

, где $\mathbf{T}^{(L+1+i)}$ *- Линейный слой с* h_i *скрытыми параметрами, где* $i = 1, \dots, P$, $\mathbf{T}^{(r)}$*-2D-сверточный слой как в 4.2.. Предполагается, что* $\left\|\mathbf{T}_{ij}^{(L+1+i)}\right\| \leqslant \tilde{w}$ *и* $h_p \leqslant h$*. То, в рамках ограничений и обозначений Теоремы 2, имеем*

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

5. Эксперименты

Для проверки теоретических оценок было проведено всестороннее эмпирическое исследование. В данном разделе представлены результаты обучения сверточных сетей с различными параметрами.

Основная цель экспериментов - продемонстрировать зависимость ландшафта функции потерь от таких параметров как количество слоев, размер ядра, число каналов, позиции пулинга, а также наблюдение зависимости скорости сходимости от этих параметров. Для этого мы обучали сверточные сети и получали параметры $\hat{\boldsymbol{\theta}}$ вблизи оптимума.

Использовалась сверточная архитектура с активацией ReLU после каждого слоя. Для отслеживания влияния конкретного параметра на сходимость мы фиксировали ключевые параметры нейронной сети, варьировали интересующий гиперпараметр и обучали соответствующий набор моделей.

Затем исследовалась зависимость между средней абсолютной разностью значений средней функции потерь и доступным размером выборки. Для каждой модели, чтобы получить более надежные результаты, усреднялась разность потерь по перемешанным выборкам. Для улучшенной визуализации применялось экспоненциальное сглаживание с коэффициентом 0.995.

В исследовании использовалось числовое представление пикселей изображений в качестве входных данных. Результаты получены на выборках из баз данных MNIST[43], FashionMNIST[44] и CIFAR10[45].

Во всех экспериментах использовались следующие гиперпараметры:

- постоянная скорость обучения 1e-3
- оптимизатор Adam
- размер мини-батча 64
- 10 эпох обучения на MNIST и Fashion-MNIST
- 15 эпох на CIFAR-10

Если параметр не варьировался, он сохранялся одинаковым для всех слоев.

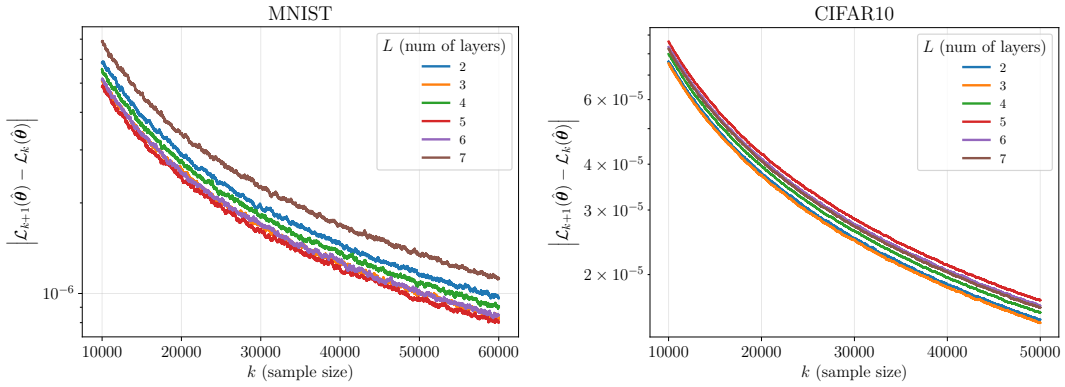

Рис. 1: Изменение количества скрытых сверточных слоев L при фиксированном размере ядра $k = 3$ и числе каналов $C = 6$. Анализ полученных графиков выявляет немонотонную зависимость между выходными значениями и количеством слоев.

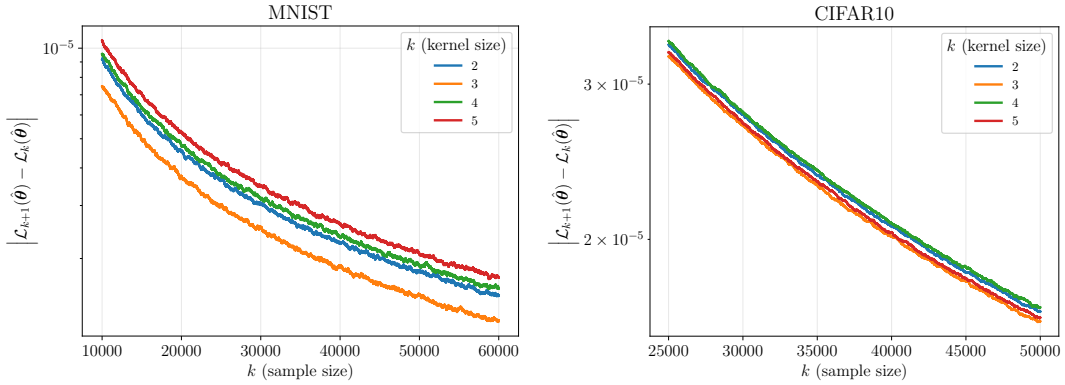

Рис. 2: Изменение размера ядра свертки k при фиксированном количестве сверточных слоев L и числе каналов $C = 6$. Данные демонстрируют немонотонную зависимость от размера ядра.

6. Обсуждение результатов

Как демонстрируют графики, абсолютная разность между средними значениями функции потерь не имеет прямой зависимости от размера ядра или количества слоев сети, но показывает монотонную зависимость от размера слоя и позиции пулинга. Мы предполагаем, что это в первую очередь свидетельствует о более значительном влиянии первой части уравнения (2) на эту величину.

Стоит отметить, что в экспериментах использовались относительно небольшие сети, поэтому увеличение количества параметров улучшало качество модели, существенно влияя на результаты, в частности на значение функции потерь в точке оптимума.

Мы выделили несколько потенциальных решений этой проблемы. Во-первых, предлагается исследовать более сложные структуры сетей, где уве-

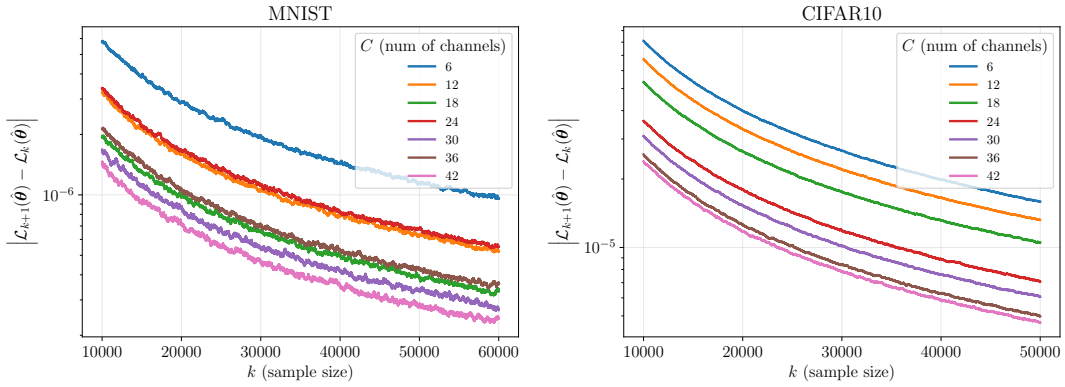

Рис. 3: Изменение количества каналов C при фиксированном числе сверточных слоёв L и размере ядра $k = 3$. Наблюдается монотонная зависимость величины от числа каналов.

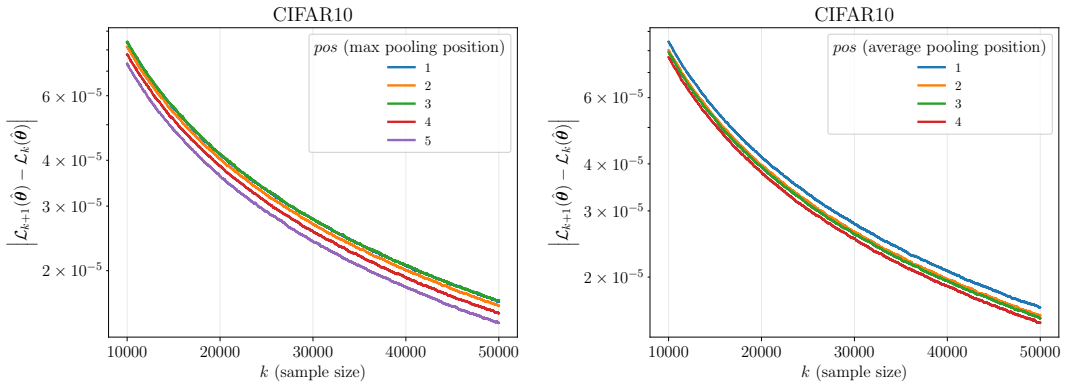

Рис. 4: Изменение количества каналов C при фиксированном числе сверточных слоев L и размере ядра $k = 3$. График демонстрирует монотонную зависимость величины от позиции операции пулинга в сети.

личение количества параметров не будет так значительно влиять на наши оценки.

Кроме того, очевидно, что наша оценка существенно превышает реалистичные значения и служит в первую очередь теоретической конструкцией, а не практической мерой. Основная причина завышения оценки связана с тем, что норма матрицы $\mathbf{T}^{(p)}$ оценивалась через произведение норм (см. доказательства 8.3. или 8.4.). Такой подход для нашего случая с разреженными матрицами неизбежно приводит к значительному завышению результатов.

Мы полагаем, что наше исследование имеет потенциальные приложения в нескольких областях, включая анализ ландшафта функции через гессиан, разработку методов определения оптимального размера выборки и исследование структурных свойств гессиана нейронных сетей.

7. Заключение

В данной работе мы предложили метод оценки нормы гессиана и способ использования этой нормы для оценки сходимости ландшафта потерь. Используя квадратичную аппроксимацию функции потерь, наш теоретический анализ показал, как сходимость ландшафта функции потерь может зависеть от нормы гессиана, а также как норма гессиана зависит от параметров сети.

Основные теоретические результаты включают:

- Аналитическую оценку нормы гессиана через параметры сети (количество слоев, размер ядра, число каналов)
- Декомпозицию гессиана на структурные компоненты
- Исследование влияния различных архитектурных элементов (пулингов, полносвязных головок)
- Экспериментальные результаты на наборах данных MNIST, Fashion-MNIST и CIFAR-10 показали, что:
- Зависимость абсолютной разности между средними значениями функции потерь имеет сложный характер

Мы считаем, что наши результаты дают ценную информацию о:

- Локальной геометрии ландшафтов потерь
- Структурных свойствах гессиана сверточных сетей
- Взаимосвязи между архитектурными параметрами и сходимостью

Перспективные направления дальнейших исследований включают:

- Уточнение оценок с учетом разреженности матриц
- Изучение более сложных архитектур
- Применение результатов для задач подбора размера выборки
- Анализ связи с обобщающей способностью моделей

- Полученные результаты вносят вклад в теоретическое понимание свойств нейронных сетей и могут найти применение при проектировании эффективных архитектур.

Список литературы

1. *Hoffmann Jordan, Borgeaud Sebastian, Mensch Arthur et al.* Training Compute-Optimal Large Language Models. — 2022. — URL: https://arxiv.org/abs/2203.15556.

2. *Grabovoy Andrey, Bakhteev Oleg, Strijov V.* ESTIMATION OF THE RELEVANCE OF THE NEURAL NETWORK PARAMETERs // *Informatics and Applications*. — 2019. — . — URL: http://dx.doi.org/10.14357/19922264190209.

3. *Grabovoy Andrey, Bakhteev O, Strijov V.* ORDERING THE SET OF NEURAL NETWORK PARAMETERS // *Informatics and Applications*. — 2020. — . — URL: http://dx.doi.org/10.14357/19922264200208.

4. Visualizing the Loss Landscape of Neural Nets / Hao Li, Zheng Xu, Gavin Taylor et al. // Advances in Neural Information Processing Systems / Ed. by S. Bengio, H. Wallach, H. Larochelle et al. — Vol. 31. — Curran Associates, Inc., 2018. — URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/a41b3bb3e6b050b6c9067c67f663b915-Paper.pdf.

5. Is the skip connection provable to reform the neural network loss landscape? / Lifu Wang, Bo Shen, Ning Zhao, Zhiyuan Zhang // Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. — IJCAI'20. — 2021. — 7 pp.

6. *Bosman Anna Sergeevna, Engelbrecht Andries, Helbig Marde.* Empirical Loss Landscape Analysis of Neural Network Activation Functions // Proceedings of the Companion Conference on Genetic and Evolutionary Computation. — Vol. 33 of *GECCO '23 Companion*. — ACM, 2023. — URL: http://dx.doi.org/10.1145/3583133.3596321.

7. *Anna Sergeevna Bosman Andries Engelbrecht Mardé Helbig.* Visualising basins of attraction for the cross-entropy and the squared error neural network loss functions // *Neurocomputing*. — 2020. — Vol. 400. — Pp. 113–136. — URL: https://www.sciencedirect.com/science/article/pii/S0925231220303593.

8. The role of over-parametrization in generalization of neural networks / Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli et al. // International

Conference on Learning Representations. — 2019. — URL: https://openreview.net/forum?id=BygfghAcYX.

9. *Zou Difan, Gu Quanquan.* An Improved Analysis of Training Over-parameterized Deep Neural Networks // Advances in Neural Information Processing Systems / Ed. by H. Wallach, H. Larochelle, A. Beygelzimer et al. — Vol. 32. — Curran Associates, Inc., 2019. — URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/6a61d423d02a1c56250dc23ae7ff12f3-Paper.pdf.

10. *Lee Youngwan, Willette Jeffrey Ryan, Kim Jonghee, Hwang Sung Ju.* Visualizing the loss landscape of Self-supervised Vision Transformer. — 2024. — URL: https://arxiv.org/abs/2405.18042.

11. *Chen Xiangning, Hsieh Cho-Jui, Gong Boqing.* When Vision Transformers Outperform ResNets without Pre-training or Strong Data Augmentations // International Conference on Learning Representations. — 2022. — URL: https://openreview.net/forum?id=LtKcMgGOeLt.

12. *Elhamod Mohannad, Karpatne Anuj.* Neuro-Visualizer: An Auto-encoder-based Loss Landscape Visualization Method. — 2023. — URL: https://arxiv.org/abs/2309.14601.

13. *Bain Robert.* Visualizing the Loss Landscape of Winning Lottery Tickets. — 2021. — URL: https://arxiv.org/abs/2112.08538.

14. *Yuan Qunyong, Xiao Nanfeng.* Experimental exploration on loss surface of deep neural network // *International Journal of Imaging Systems and Technology.* — 2020. — Vol. 30, no. 4. — Pp. 860–873. — URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ima.22434.

15. *Im Daniel Jiwoong, Tao Michael, Branson Kristin.* An empirical analysis of the optimization of deep network loss surfaces. — 2017. — URL: https://arxiv.org/abs/1612.04010.

16. *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. — URL: https://proceedings.mlr.press/v97/ghorbani19b.html.

17. *Papyan Vardan.* The Full Spectrum of Deep Net Hessians At Scale: Dynamics with Sample Size // *CoRR.* — 2018. — Vol. abs/1811.07062. — URL: http://arxiv.org/abs/1811.07062.

18. *Liao Zhenyu, Mahoney Michael W.* Hessian eigenspectra of more realistic nonlinear models // Proceedings of the 35th International Conference on Neural Information Processing Systems. — NIPS '21. — Red Hook, NY, USA: Curran Associates Inc., 2024. — 14 pp.

19. *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.

20. Numerical Methods of Sufficient Sample Size Estimation for Generalised Linear Models / Andrey Grabovoy, Tamaz Gadaev, A. Motrenko, Vadim Strijov // *Lobachevskii Journal of Mathematics.* — 2022. — 12. — Vol. 43. — Pp. 2453–2462.

21. *Azadbakht Alireza, Kheradpisheh Saeed Reza, Khalfaoui-Hassani Ismail, Masquelier Timothée.* Drastically Reducing the Number of Trainable Parameters in Deep CNNs by Inter-layer Kernel-sharing. — 2022. — URL: https://arxiv.org/abs/2210.14151.

22. *Kroshchanka A. A., Golovko V. A., Chodyka M.* Method for Reducing Neural-Network Models of Computer Vision // *Pattern Recognit. Image Anal.* — 2022. — . — Vol. 32, no. 2. — P. 294–300. — URL: https://doi.org/10.1134/S1054661822020146.

23. *Kahatapitiya Kumara, Rodrigo Ranga.* Exploiting the Redundancy in Convolutional Filters for Parameter Reduction // 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). — 2021. — Pp. 1409–1419.

24. *Fort Stanislav, Ganguli Surya.* Emergent properties of the local geometry of neural loss landscapes // *CoRR.* — 2019. — Vol. abs/1910.05929. — URL: http://arxiv.org/abs/1910.05929.

25. *Singh Sidak Pal, Adilova Linara, Kamp Michael et al.* Landscaping Linear Mode Connectivity. — 2024. — URL: https://arxiv.org/abs/2406.16300.

26. *Li Xin-Chun, Li Lan, Zhan De-Chuan.* Visualizing, Rethinking, and Mining the Loss Landscape of Deep Neural Networks. — 2024. — URL: https://arxiv.org/abs/2405.12493.

27. Phenomenology of Double Descent in Finite-Width Neural Networks / Sidak Pal Singh, Aurelien Lucchi, Thomas Hofmann, Bernhard Schölkopf // International Conference on Learning Representations. — 2022. — URL: https://openreview.net/forum?id=lTqGXfn9Tv.

28. *Wu Yikai, Zhu Xingyu, Wu Chenwei et al.* Dissecting Hessian: Understanding Common Structure of Hessian in Neural Networks. — 2021. — URL: https://openreview.net/forum?id=0rNLjXgchOC.

29. *Skorski Maciej.* Chain Rules for Hessian and Higher Derivatives Made Easy by Tensor Calculus. — 2019. — URL: https://arxiv.org/abs/1911.13292.

30. *Papyan Vardan.* The Full Spectrum of Deepnet Hessians at Scale: Dynamics with SGD Training and Sample Size. — 2019. — URL: https://arxiv.org/abs/1811.07062.

31. *Sagun Levent, Bottou Leon, LeCun Yann.* Eigenvalues of the Hessian in Deep Learning: Singularity and Beyond. — 2017. — URL: https://arxiv.org/abs/1611.07476.

32. *Papyan Vardan.* Traces of Class/Cross-Class Structure Pervade Deep Learning Spectra // *Journal of Machine Learning Research.* — 2020. — Vol. 21, no. 252. — Pp. 1–64. — URL: http://jmlr.org/papers/v21/20-933.html.

33. *Papyan Vardan.* Measurements of Three-Level Hierarchical Structure in the Outliers in the Spectrum of Deepnet Hessians // 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. 5012–5021. — URL: https://proceedings.mlr.press/v97/papyan19a.html.

34. *Kiselev Nikita, Grabovoy Andrey.* Unraveling the Hessian: A Key to Smooth Convergence in Loss Function Landscapes. — 2024. — URL: https://arxiv.org/abs/2409.11995.

35. *Magnus Jan R., Neudecker Heinz.* Matrix Differential Calculus with Applications in Statistics and Econometrics. — Second edition. — John Wiley, 1999.

36. *Schraudolph Nicol N.* Fast curvature matrix-vector products for second-order gradient descent // *Neural Comput.* — 2002. — . — Vol. 14, no. 7. — P. 1723–1738. — URL: https://doi.org/10.1162/08997660260028683.

37. Empirical Analysis of the Hessian of Over-Parametrized Neural Networks / Levent Sagun, Utku Evci, V. Ugur Güney et al. // *CoRR.* — 2017. — Vol. abs/1706.04454. — URL: http://arxiv.org/abs/1706.04454.

38. *Latrémolière Frédéric, Narayanappa Sadananda, Vojtěchovský Petr.* Estimating the Jacobian matrix of an unknown multivariate function from sample values by means of a neural network. — 2022. — URL: https://arxiv.org/abs/2204.00523.

39. *Hayou Soufiane, Dadoun Benjamin, Youssef Pierre et al.* A Theoretical Study of the Jacobian Matrix in Deep Neural Networks. — 2024. — URL: https://openreview.net/forum?id=pvhyBB86Bt.

40. 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. — URL: https://doi.org/10.1137/21M1441183.

41. *Qin Zhen, Han Xiaodong, Sun Weixuan et al.* Toeplitz Neural Network for Sequence Modeling. — 2023. — URL: https://arxiv.org/abs/2305.04749.

42. *Gnacik Michal, Łapa Krystian.* Using Toeplitz Matrices to obtain 2D convolution. — 2022. — 10.

43. *Deng Li.* The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web] // *IEEE Signal Processing Magazine.* — 2012. — Vol. 29. — Pp. 141–142. — URL: https://api.semanticscholar.org/CorpusID:5280072.

44. *Xiao Han, Rasul Kashif, Vollgraf Roland.* Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. — 2017. — cite arxiv:1708.07747Comment: Dataset is freely available at https://github.com/zalandoresearch/fashion-mnist Benchmark is available at http://fashion-mnist.s3-website.eu-central-1.amazonaws.com/. URL: http://arxiv.org/abs/1708.07747.

45. *Krizhevsky Alex.* Learning Multiple Layers of Features from Tiny Images. — 2009. — URL: https://api.semanticscholar.org/CorpusID:18268744.

8. Дополнение

8.1. Доказательство Леммы 1

Доказательство. Выходы из сверточной нейронной сети - логиты:

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

Рассматривается производная логитов по параметрам нейронной сети

$$\frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(p)}} = \frac{\partial \mathbf{z}}{\partial \mathbf{z}^{(p)}}\frac{\partial \mathbf{z}^{(p)}}{\partial \mathbf{T}^{(p)}}\frac{\partial \mathbf{T}^{(p)}}{\partial \mathbf{W}^{(p)}} \text{ как Якобиан композиции}$$

Используя свойство $vec(\mathbf{BVA}^\intercal) = (\mathbf{A} \otimes \mathbf{B})vec(\mathbf{V})$ где $\mathbf{A} = \mathbf{I}$ и получим векторизованный $\mathbf{z}^{(p)}$

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

Видно, что,

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

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

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

Используя определение $\mathbf{Q}^{(p)}$:

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

Также используем

$$\mathbf{A}_i \in \mathbb{R}^{m_i \times n_i}, \mathbf{A}_1 \otimes \mathbf{A}_2 = (\mathbf{A}_1 \otimes \mathbf{I}_{m_2})(\mathbf{I}_{m_1} \otimes \mathbf{A}_2).$$

с $m_2 = 1$ получаем

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

подставляем приведенные выше утверждения в одну формулу и получаем

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

Как в [19] рассматривая блок $\mathbf{H}_O^{(kl)}$:

$$\mathbf{H}_O^{(kl)} = J(\boldsymbol{\theta})^\intercal \mathbf{A} J(\boldsymbol{\theta}) =$$
$$= \mathbf{Q}^{(k)^\intercal}(\mathbf{G}^{(k)^\intercal} \otimes \mathbf{R}^{(k-1)}\mathbf{x}) A (\mathbf{G}^{(l)} \otimes \mathbf{x}^\intercal \mathbf{R}^{(l-1)^\intercal}) \mathbf{Q}^{(l)}$$

А значит $\mathbf{H}_O = \mathbf{Q}^\intercal \mathbf{F}^\intercal \mathbf{A} \mathbf{F} \mathbf{Q}$.

□

8.2. Доказательство Леммы 2

Доказательство. Используя результаты предыдущей леммы 1, нам достаточно оценить верхнюю границу выражения: $\|\mathbf{Q}\|^2 \|\mathbf{F}\|^2 \|\mathbf{A}\|$
In the work [34], норма матрицы $\mathbf{A}$ было проверено, и было доказано, что:

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

Норма блочно-диагональной матрицы не больше максимальной нормы блока

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

Норма произведения матриц меньше или равна произведению норм :

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

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

Спектральная норма матрицы произведения Кронекера равна их обычной норме произведения. Спектральная норма вертикально сложенных матриц меньше или равна сумме норм ее блоков.

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

Подставляя полученные оценки в $\|\mathbf{H}_O\|$ получим

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

□

8.3. Доказательство Теоремы 1

Доказательство. Ясно, что на основании Леммы 2, нам нужно доказать только 2 утверждения:

$$\left\|\mathbf{T}^{(p)}\right\|^2 \leqslant C^2 dk w^2,$$
$$\left\|\mathbf{Q}^{(p)}\right\|^2 \leqslant d^2.$$

В [19], легко видеть, что в $\mathbf{T}^{(p)}$ каждый блок $C_l C_{l-1}$ содержит d_{l-1} строки с ядром в правильном положении, которые приводят нас к

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

Для доказательства второго неравенства снова обратимся к [19] и оцените норму вертикально сложенных матриц:

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

$$\left\|\mathbf{Q}^{(l)}\right\| \leqslant \sum_{i=0}^{d_{l-1}-k_l} \left\|\pi_R^i \mathbf{I}_{d_{l-1} \times k_l}\right\| \leqslant \sum_{i=1}^{d_{l-1}-k_l} \|\pi_R\| =$$
$$= \sum_{i=0}^{d_{l-1}-k_l} 1 = d_{l-1} - k_l + 1 = d_l \leqslant d_1 = d.$$

□

8.4. Доказательство Теоремы 2

Доказательство. из описания 4.2. мтарицы $\mathbf{T}^{(p)}$ можно увидеть, что

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

И как очевидное следствие

$$\left\|\mathbf{T}^{(p)}\right\|_F^2 = \sum_{c_1,i,k,l}^{C_{p-1},C_p n_p m_p,k_p^1,k_p^2} \left(\mathbf{W}^{(p)}_{c_1,c_2(i),k,l}\right)^2. \tag{5}$$

Здесь мы предполагаем простое соответствие между выходным каналом c_2 и i-й строкой $\mathbf{T}^{(p)}$.

По аналогии с доказательством 8.3., используя 2 нам нужно доказать 2 утверждения:

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

Первоначально норма $\mathbf{T}^{(p)}$ оценена:

$$\left\|\mathbf{T}^{(p)}\right\|^2 \leqslant \left\|\mathbf{T}^{(p)}\right\|_F^2 \leqslant |(5)| \leqslant \sum_i Ck^2w^2 \leqslant C^2k^2w^2mn.$$

Далее оценим норму производной слоя по параметрам.

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

Как было сказано ранее, строка $\mathbf{T}^{(p)}$ - это в точности $vec_r(\mathbf{W}^{(p)}_{*,i,*,*})$ расположены в правильном порядке. Тогда норма строки равна:
$\frac{\partial \mathbf{T}^{(p)}_{(i,j)}}{\partial \mathbf{W}^{(p)}_{c_1,c_2,k_1,k_2}} \neq 0 \iff$ индексы подобраны таким образом, что $T_i^{(p)}$ соответствует c_2 и в то же время $\mathbf{T}^{(p)}_{i,j}$ соответствует c_1, k_1, k_2 и это соответствие зависит от конкретной матрицы $\mathbf{T}^{(p)}$, но очевидно, что один i соответствует только одному c_2, потому что каждая строка участвует в формировании только одного элемента

одного канала. Так как только $\mathbf{W}^{(p)}_{*,c_2,*,*}$ участвует в формировании одной строки $\mathbf{T}^{(p)}_{i,*}$, можно исправить i и соответственноc_2, и в то же время мы знаем, что для каждого c_1, k_1, k_2, есть только один столбец j: $\mathbf{T}^{(p)}_{i,j} = \mathbf{W}^{(p)}_{c_1,c_2,k_1,k_2}$:

$$\sum_{j,c_1,k_1,k_2} \left(\frac{\partial \mathbf{T}^{(p)}_{i,j}}{\partial \mathbf{W}^{(p)}_{c_1,c_2,k_1,k_2}} \right)^2 =$$

$$\sum_{c_1,k_1,k_2} \sum_{j} \left(\frac{\partial \mathbf{T}^{(p)}_{i,j}}{\partial \mathbf{W}^{(p)}_{c_1,c_2,k_1,k_2}} \right)^2 =$$

$$= \left|\text{во внутренней сумме есть только один ненулевой член}\right| =$$

$$= \sum_{c_1,k_1,k_2} 1 = C_{p-1} k_p^1 k_p^2 \leqslant Ck^2.$$

Рассмотрим норму Фробениуса как верхнюю границу спектральной нормы:

$$\|\mathbf{Q}\|^2 \leqslant \|\mathbf{Q}\|_F^2 = \sum_{i,j,c_1,c_2,k_1,k_2} \left(\frac{\partial \mathbf{T}^{(p)}_{i,j}}{\partial \mathbf{W}^{(p)}_{c_1,c_2,k_1,k_2}} \right)^2 =$$

$$= \sum_{i,c_2} \sum_{j,c_1,k_1,k_2} \left(\frac{\partial \mathbf{T}^{(p)}_{i,j}}{\partial \mathbf{W}^{(p)}_{c_1,c_2,k_1,k_2}} \right)^2 =$$

$$= \left|\text{оценили внутреннюю сумму ранее и только при соответствующих } i \text{ и } c_2\right| \leqslant$$

$$\leqslant \sum_{i} Ck^2 = CmnCk^2 \leqslant C^2 k^2 mn.$$

□

8.5. Доказательство Леммы 3

Доказательство. Используя обозначение $\mathbf{M}^{(l)}$ для 2D-Max-Pool слоя
Как и в случае со свертками, мы можем описать каждую строку $\mathbf{M}^{(l)}$:
Прежде чем начать, рассмотрим некоторые свойства $\mathbf{M}$, то, что будет использоваться: Во-первых, что строка $\mathbf{M}_{i*}$ соответствует определенному окну пула (элементам, охватываемым окном), и, как и второй, является тем столбцом $\mathbf{M}_{*j}$ соответствует элементам, умноженным на j'й элемент входа.
Так как каждое окно покрывает только один элемент и два разных окна не

пересекаются, то в каждой строке находится только один элемент, таким образом

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

так как $(\mathbf{M}^{(l)}_{*,i}, \mathbf{M}^{(l)}_{*,j}) \neq 0 \iff (\mathbf{M}^{(l)}_{*,i}, \mathbf{M}^{(l)}_{*,j}) = 1 \iff i = j$ и i-й элемент — это максимум в соответствующем окне. Для простоты предположим, что $\mathbf{M}^{(l)}$ уменьшает обе размерности в k_{pool} раз, аналогично 8.4. оценивается $\mathbf{G}^{(p)}$ и $\mathbf{R}^{(p-1)}$ компонента, однако, в соответствии с новым слоем.

$$\left\|\mathbf{G}^{(p)}\right\| \left\|\mathbf{R}^{(p-1)}\right\| \leqslant \frac{\prod_{i=1}^{L+1} \left\|T^{(i)}\right\|}{\left\|T^{(p)}\right\|} \leqslant$$
$$(C^2k^2w^2mn)^{2L}\left(\frac{1}{k^2_{\text{pool}}}\right)^{L-l+2-I\{p-1\leqslant l\}} \leqslant$$
$$(C^2k^2w^2mn)^{2L}\left(\frac{1}{k^2_{\text{pool}}}\right)^{L-l+2}.$$

Далее, получим

$$\|\mathbf{F}\|^2 \leqslant \|\mathbf{x}\|^2 (L+1)(k^2C^2w^2mn)^{(L)}\left(\frac{1}{k^2_{\text{pool}}}\right)^{L-l+2},$$
$$\|\mathbf{H}_O\| \leqslant \sqrt{2}\left\|\mathbf{x}^2\right\| q^2\left(\frac{1}{k^2_{\text{pool}}}\right)^{L-l+2}(L+1)(k^2C^2w^2mn)^L,$$

где $q^2 = mnC^2k^2$. □

8.6. Доказательство Леммы 4

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

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

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

После этого, применяя элементарные преобразования над строками и столбцами, приводим матрицу $\mathbf{A}^{(p)^\intercal}\mathbf{A}^{(p)}$ в блочно-диагональную форму, где блоки соответствуют индексам в том же окне avg-pool. Каждый блок $\mathbf{A}^{(p)^\intercal}\mathbf{A}^{(p)}$ - это $\mathbf{B}_i = \frac{1}{k_{\text{pool}}^2}\mathbf{1}\mathbf{1}^\intercal$, где $\mathbf{1} = \mathbf{1}_{k_{\text{pool}}^2} \in \mathbb{R}^{k_{\mathbf{A}}^2}$ - вектор из едениц, и его норма $\|\mathbf{B}_i\| = \frac{1}{k_{\text{pool}}^2}\|\mathbf{1}\mathbf{1}^\intercal\| = \frac{1}{k_{\text{pool}}}$
Норма блочно-диагональной матрицы (и норма матрицы, которая может быть приведена к этому виду) равна максимуму норм:
$\left\|\mathbf{A}^{(p)}\right\| = \max\limits_i \|\mathbf{B}_i\| = \frac{1}{k_{\text{pool}}}$,
потому, что $\left\|\mathbf{A}^{(p)}\right\| \leqslant 1$, мы можем полностью повторно использовать вычисления предыдущего доказательства и получить тот же результат. □

8.7. Доказательство Леммы 5

Доказательство. Как и в предыдущих доказательствах, нам нужно оценить $\left\|\mathbf{G}^{(p)}\right\|^2 \left\|\mathbf{R}^{(p-1)}\right\|^2$.
Также известно, что $\left\|T^{(L+1+p)}\right\|^2 \leqslant (h^2\tilde{w}^2)\ \forall p = 1, \ldots, P$ Тогда мы можем оценить

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

Для $p \leqslant L+1$ и

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

для $p = L+2, \ldots, L+P+1$.

Или в одной записи:

$$\left\|\mathbf{G}^{(p)}\right\|^2 \left\|\mathbf{R}^{(p-1)}\right\|^2 \leqslant$$
$$(h^2\tilde{w}^2)^{P-I_{\{p>L+1\}}}(k^2C^2w^2mn)^{L+I_{\{p>L+1\}}}.$$

А значит имеем:

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

И применяя этот результат к матрицы Гессе

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

□