

# Полное строгое определение локального и глобального Гессиана второго порядка в произвольной нейронной архитектуре

**Аноним**  
Аффилиация  
anonymous@example.com

June 9, 2025

## Abstract

В данной статье представлено исчерпывающее математически строгое определение Гессиана второго порядка для нейронных сетей произвольной архитектуры, заданной направленным ациклическим графом. Существующие подходы к вычислению кривизны функции потерь нейронных сетей часто ограничиваются аппроксимацией Гаусса-Ньютона, учитывающей лишь часть вторых производных. В работе разработан полный формализм, учитывающий все чистые и смешанные вторые производные по входам и параметрам, кросс-блоки между разными параметрами, а также механизмы разделения параметров между узлами сети. Особое внимание уделено негладким активационным функциям через использование Clarke-Гессиана. Для тривиального графа из единственного узла без потомков и предков предложенные формулы сводятся к стандартному Гессиану  $\nabla_{\theta}^2 \mathcal{L}(\theta) \in \mathbb{R}^{p \times p}$ . Предложенный формализм предоставляет теоретический фундамент для углубленного анализа геометрических свойств функционала потерь и разработки более эффективных алгоритмов оптимизации нейронных сетей произвольной архитектуры.

## 1 Введение

Гессиан второго порядка  $\nabla^2 \mathcal{L}$  играет фундаментальную роль в анализе кривизны функционала потерь и в разработке методов оптимизации нейронных сетей [Martens, 2014, Pascanu et al., 2013b]. Методы второго порядка, такие как методы Ньютона, trust-region методы и их модификации, требуют точной информации о кривизне функции потерь для эффективной оптимизации [Nocedal and Wright, 2006]. Однако в контексте глубоких нейронных сетей вычисление и хранение полного Гессиана становится вычислительно неприемлемым, что приводит к необходимости использования различных аппроксимаций.

Наиболее распространенный подход — аппроксимация Гаусса-Ньютона, которая учитывает лишь часть всех вторых производных, игнорируя существенные компоненты кривизны [Schraudolph, 2002, Martens, 2010]. В данной работе мы предлагаем *полный* формализм, закрывающий следующие пробелы в существующей литературе:

- чистые и смешанные вторые производные по *входам* каждого узла нейронной сети;
- чистые вторые производные по *параметрам*;
- кросс-блоки  $\partial^2/\partial\theta_v\partial\theta_w$  между параметрами разных узлов;
- смешанные вход-параметрические производные;
- учёт «разделения» (sharing) одного вектора параметров между несколькими узлами;
- обработка негладких активационных функций через методологию Clarke-Гессиана.

*Особый случай:* Если архитектура нейронной сети вырождается в единственный узел без потомков и предков, все предлагаемые определения естественным образом сводятся к стандартному Гессиану  $\nabla_{\theta}^2\mathcal{L}(\theta) \in \mathbb{R}^{p \times p}$ .

## 2 Связанные работы

Изучение геометрии функционала потерь нейронных сетей имеет долгую историю. Классические работы [Amari, 1998, Heskes, 2000] заложили основу для использования геометрической информации в оптимизации нейронных сетей. Особое значение имеет информационная геометрия и натуральный градиентный спуск, предложенный Амари [Amari, 1998].

В контексте вычисления Гессиана нейронных сетей, значительными являются работы [Martens, 2010, Martens and Sutskever, 2012], где представлены эффективные приближения Гессиана для глубоких нейронных сетей. Гаусс-Ньютон аппроксимация, которая игнорирует вторые производные функции потерь, часто применяется в практических алгоритмах из-за вычислительной эффективности и гарантии положительной полуопределенности.

Для негладких функций активации, таких как ReLU, традиционный анализ второго порядка неприменим. В работах [Clarke, 1990, Bolte and Pauwels, 2020] представлен обобщенный подход к недифференцируемым функциям через субдифференциальное исчисление и Clarke-градиенты. В нашей работе мы применяем

эти концепции непосредственно к нейронным сетям, предлагая полный формализм для анализа кривизны функции потерь.

Недавние работы [Ghorbani et al., 2019, Sagun et al., 2017] исследуют спектральные свойства Гессиана функций потерь нейронных сетей и их связь с обобщением. Наше исследование дополняет эти работы, предоставляя точный математический аппарат для вычисления всех компонентов Гессиана в произвольных архитектурах нейронных сетей.

## 3 Методология

### 3.1 Функциональные пространства и аналитические предпосылки

Прежде чем перейти к определению компонентов и структуры Гессиана, необходимо формализовать функциональные пространства, в которых рассматривается задача, и уточнить аналитические предпосылки анализа.

**Definition 1** (Функциональные пространства). *В рамках данной работы рассматриваются следующие функциональные пространства:*

- $\mathbb{R}^n$  с евклидовой нормой  $\|\cdot\|_2$  — конечномерное гильбертово пространство параметров, активаций и градиентов.
- $C^2(\mathbb{R}^n, \mathbb{R}^m)$  — пространство дважды непрерывно дифференцируемых функций из  $\mathbb{R}^n$  в  $\mathbb{R}^m$ , используемое для гладкого случая.
- $C^{1,1}(\mathbb{R}^n, \mathbb{R}^m)$  — пространство непрерывно дифференцируемых функций с липшицевыми производными, используемое для негладкого случая.
- $PC^2(\mathbb{R}^n, \mathbb{R}^m)$  — пространство кусочно дважды дифференцируемых функций, где каждый кусок принадлежит  $C^2$ , а границы кусков образуют множество меры нуль.
- $L(\mathbb{R}^n, \mathbb{R}^m)$  — пространство линейных операторов (матриц) из  $\mathbb{R}^n$  в  $\mathbb{R}^m$  с операторной нормой и нормой Фробениуса.

**Assumption 1** (Регулярность функций узлов). *Для каждого узла  $v \in V$  нейронной сети:*

1. В гладком случае (Случай A): функция узла  $g_v \in C^2(\mathbb{R}^d, \mathbb{R}^{d_v})$ , т.е. дважды непрерывно дифференцируема по всем входам и параметрам.
2. В негладком случае (Случай B): функция узла  $g_v \in PC^2(\mathbb{R}^d, \mathbb{R}^{d_v}) \cap C^{1,1}(\mathbb{R}^d, \mathbb{R}^{d_v})$ , т.е. является кусочно дважды дифференцируемой с липшицевыми первыми производными (как, например, ReLU-активация), что обеспечивает существование и непустоту Clarke-субдифференциала.

**Proposition 1** (Существование и непустота Clarke-субдифференциала). *Для локально липшицевой функции  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , которая дифференцируема почти всюду (в смысле меры Лебега), Clarke-субдифференциал  $\partial_C f(x)$  определен и непуст во всех точках  $x \in \mathbb{R}^n$ . Более того,  $\partial_C f(x)$  является выпуклым компактным множеством в метрическом пространстве  $(L(\mathbb{R}^n, \mathbb{R}), \|\cdot\|_{op})$ , где  $\|\cdot\|_{op}$  — операторная норма.*

*Для векторнозначных функций  $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$  субдифференциал определяется покомпонентно, и Clarke-Гессиан  $\partial_C^2 F(x)$  также существует при соответствующих условиях локальной липшицевости и почти всюду дифференцируемости компонент градиента  $\nabla F_i$ .*

Эти предпосылки гарантируют корректность всех последующих определений и вычислений, связанных с дифференцированием функций узлов нейронной сети как в гладком, так и в негладком случаях.

### 3.2 Модель нейронной сети и обозначения

**Definition 2** (Архитектура нейронной сети). *Рассматривается нейронная сеть, архитектура которой представлена в виде направленного ациклического графа (DAG)  $G = (V, E)$ , где  $V$  — множество узлов сети, а  $E$  — множество направленных рёбер.*

Для каждого узла  $v \in V$  определены следующие компоненты:

**Входы:**  $f_{Pa(v)} \in \prod_{u \in Pa(v)} \mathbb{R}^{d_u}$ , где  $Pa(v)$  — множество родительских узлов для  $v$ .

**Параметры:**  $\theta_v \in \mathbb{R}^{p_v}$  — параметры, связанные с узлом  $v$ .

**Функция узла:**  $f_v = g_v(f_{Pa(v)}, \theta_v) \in \mathbb{R}^{d_v}$  — отображение, преобразующее входы и параметры в выход узла  $v$ .

**Функция потерь:**  $\mathcal{L} : \mathbb{R}^{d_{out}} \rightarrow \mathbb{R}$  — функция потерь, определённая на выходном узле  $out \in V$ .

В зависимости от гладкости функций узлов, выделяем два принципиально различных случая:

- **Случай А (гладкий).** Все функции узлов  $g_v$  дважды непрерывно дифференцируемы по входам и параметрам, т.е.  $g_v \in C^2$ .
- **Случай В (негладкий).** В сети присутствуют негладкие функции активации, такие как ReLU, max-pooling и другие. В этом случае используется концепция Clarke-Гессиана  $\partial_C^2 f_v$ .

Вычисление всех блоков Гессиана осуществляется в *обратном топологическом порядке* по графу  $G$ , начиная с выходного узла  $out$ .

### 3.3 Таблица обозначений

Для удобства восприятия сложных формул и структур, приведём систематизированную таблицу основных обозначений:

Table 1: Основные обозначения, используемые в работе

Символ	Определение	Размерность
$v, w, u$	Узлы нейронной сети	—
$G = (V, E)$	Направленный ациклический граф, представляющий нейронную сеть	—
$\text{Pa}(v)$	Множество родительских узлов узла $v$	—
$\text{Ch}(v)$	Множество дочерних узлов узла $v$	—
$f_v$	Вектор выходов узла $v$	$\mathbb{R}^{d_v}$
$\theta_v$	Вектор параметров узла $v$	$\mathbb{R}^{p_v}$
$\mathcal{L}$	Функция потерь	$\mathbb{R}$
$\delta_v$	Градиент потерь по выходу узла $v$	$\mathbb{R}^{d_v}$
$D_{u \leftarrow v}$	Якобиан преобразования от узла $v$ к узлу $u$	$\mathbb{R}^{d_u \times d_v}$
$D_v$	Якобиан выхода узла $v$ по его параметрам	$\mathbb{R}^{d_v \times p_v}$
$T_{u;v}$	Тензор вторых производных выхода узла $u$ по входу от узла $v$	$\mathbb{R}^{d_u \times d_v \times d_v}$
$T_{u;v,w}$	Тензор смешанных вторых производных по разным входам	$\mathbb{R}^{d_u \times d_v \times d_w}$
$T_{v;w,\theta}$	Тензор смешанных производных по входу и параметрам	$\mathbb{R}^{d_v \times d_w \times p_v}$
$T_v^\theta$	Тензор вторых производных по параметрам	$\mathbb{R}^{d_v \times p_v \times p_v}$
$H_{v,w}^f$	Блок входного Гессiana между узлами $v$ и $w$	$\mathbb{R}^{d_v \times d_w}$
$H_{\theta_v, \theta_w}$	Блок параметрического Гессiana	$\mathbb{R}^{p_v \times p_w}$
$\partial_C^2 f_v$	Clarke-Гессиан узла $v$ (для негладкого случая)	множество матриц

**Remark 1** (Соглашение об индексах). В работе приняты следующие соглашения об индексах:

- $i$  — индекс компоненты выхода узла ( $f_v$  или  $f_u$ )
- $j, k$  — индексы компонент входов узлов
- $k, \ell$  — в контексте параметров, индексы компонент параметра  $\theta_v$
- $v, w, u$  — индексы узлов в графе нейронной сети

### 3.4 Градиенты первого порядка

Для разработки полного формализма Гессиана второго порядка необходимо сначала определить производные первого порядка, которые служат основой для дальнейших вычислений:

$$\begin{aligned} \delta_v &:= \nabla_{f_v} \mathcal{L} && \in \mathbb{R}^{d_v}, \quad (\text{градиент потерь по выходу узла } v) \\ \delta_{v,i} &:= [\delta_v]_i, && i = 1, \dots, d_v, \quad (\text{компоненты градиента}) \\ D_{u \leftarrow v} &:= \frac{\partial f_u}{\partial f_v} && \in \mathbb{R}^{d_u \times d_v}, \quad (\text{якобиан по входу}) \\ D_v &:= \frac{\partial f_v}{\partial \theta_v} && \in \mathbb{R}^{d_v \times p_v}. \quad (\text{якобиан по параметрам}) \end{aligned}$$

Градиенты  $\delta_v$  и якобианы  $D_{u \leftarrow v}$ ,  $D_v$  являются основой цепного правила первого порядка и используются для вычисления производных функции потерь по параметрам сети.

### 3.5 Тензоры вторых производных

Для полного учёта всех вторых производных функций узлов вводятся следующие тензорные структуры:

$$\begin{aligned} [T_{u;v}]_{i,j,k} &= \frac{\partial^2(f_u)_i}{\partial(f_v)_j \partial(f_v)_k} \in \mathbb{R}^{d_u \times d_v \times d_v}, && v \in \text{Pa}(u), \\ [T_{u;v,w}]_{i,j,k} &= \frac{\partial^2(f_u)_i}{\partial(f_v)_j \partial(f_w)_k} \in \mathbb{R}^{d_u \times d_v \times d_w}, && v, w \in \text{Pa}(u), \quad v \neq w, \\ [T_{v;w,\theta}]_{i,j,k} &= \frac{\partial^2(f_v)_i}{\partial(f_w)_j \partial(\theta_v)_k} \in \mathbb{R}^{d_v \times d_w \times p_v}, && w \in \text{Pa}(v), \\ [T_v^\theta]_{i,k,\ell} &= \frac{\partial^2(f_v)_i}{\partial(\theta_v)_k \partial(\theta_v)_\ell} \in \mathbb{R}^{d_v \times p_v \times p_v}. \end{aligned}$$

**Remark 2** (Тензорная нотация и правила свертки). В тензорных выражениях выше и далее приняты следующие соглашения:

- Индекс  $i$  всегда относится к компоненте выхода соответствующего узла ( $f_u$  или  $f_v$ ).
- Индексы  $j$  и  $k$  относятся к компонентам входов от родительских узлов.
- Индексы  $k$  и  $\ell$  (в нижнем тензоре) относятся к компонентам параметров.

- Обозначение  $[T]_{i,\bullet,\bullet}$  представляет матрицу (срез тензора), полученную фиксацией индекса  $i$ .
- При умножении на скаляр  $\delta_{u,i}$  подразумевается свёртка по индексу  $i$  с весовыми коэффициентами  $\delta_{u,i}$ .

При свертке тензоров с другими тензорами или векторами используются следующие правила:

- Для выражения  $[T_{u;v}]_{i,j,k}\delta_{u,i}$  результатом является матрица размерности  $d_v \times d_v$  с элементами  $\sum_{i=1}^{d_u} [T_{u;v}]_{i,j,k}\delta_{u,i}$ .
- При матричном умножении  $D_{u \leftarrow v}^\top H_{u,u}^f D_{u \leftarrow w}$  индексы сворачиваются согласно правилам матричного произведения, где  $D_{u \leftarrow v}^\top \in \mathbb{R}^{d_v \times d_u}$ ,  $H_{u,u}^f \in \mathbb{R}^{d_u \times d_u}$ ,  $D_{u \leftarrow w} \in \mathbb{R}^{d_u \times d_w}$ .
- Тензорное выражение  $\sum_{i=1}^{d_u} [T_{u;v,w}]_{i,\bullet,\bullet}\delta_{u,i}$  преобразуется в матрицу размерности  $d_v \times d_w$  с элементами  $\sum_{i=1}^{d_u} [T_{u;v,w}]_{i,j,k}\delta_{u,i}$ .

Это соглашение обеспечивает однозначность всех тензорных операций в формулах и устраняет возможные неоднозначности при переходе от тензорной к матричной записи.

Эти тензоры учитывают чистые и смешанные вторые производные функций узлов по входам и параметрам. При суммировании по индексу  $i$  с весом  $\delta_{u,i}$ , эти тензоры дают вклад в Гессиан функции потерь.

### 3.6 Полный входной Гессиан

**Definition 3** (Входной Гессиан). *Полный входной Гессиан представляет собой блочную матрицу  $\{H_{v,w}^f\}_{v,w \in V}$ , где каждый блок  $H_{v,w}^f \in \mathbb{R}^{d_v \times d_w}$  определяется рекурсивно:*

$$\begin{aligned}
 H_{v,w}^f = & \sum_{u \in \text{Ch}(v) \cap \text{Ch}(w)} D_{u \leftarrow v}^\top H_{u,u}^f D_{u \leftarrow w} \quad (\text{Гаусс-Ньютон}) \\
 & + \sum_{u \in \text{Ch}(v) \cap \text{Ch}(w)} \sum_{i=1}^{d_u} [T_{u;v,w}]_{i,\bullet,\bullet} \delta_{u,i} \quad (\text{смешанные входы}) \\
 & + \delta_{v=w} \sum_{u \in \text{Ch}(v)} \sum_{i=1}^{d_u} [T_{u;v}]_{i,\bullet,\bullet} \delta_{u,i} \quad (\text{чистые по одному входу})
 \end{aligned} \tag{1}$$

с базовыми условиями:

$$H_{out,out}^f = \nabla^2 \mathcal{L}(f_{out}), \quad H_{out,v}^f = H_{v,out}^f = 0 \quad (\forall v \neq out),$$

$$H_{v,w}^f = 0 \quad \text{если нет узла } u, \text{ являющегося потомком и } v, \text{ и } w.$$

**Свойство симметрии:**  $H_{v,w}^f = (H_{w,v}^f)^\top$  для всех  $v, w \in V$ .

Первый член в уравнении (1) соответствует классической части Гаусса-Ньютона, которая учитывает только первые производные функций узлов. Второй член учитывает смешанные вторые производные между различными входами, а третий член (активный только при  $v = w$ ) добавляет вклад от чистых вторых производных по одному входу.

### 3.7 Полный параметрический Гессиан

**Definition 4** (Параметрический Гессиан). *Полный Гессиан по параметрам  $\nabla_{\theta}^2 \mathcal{L}$  разбивается на блоки  $\{H_{\theta_v, \theta_w}\}$ ,  $H_{\theta_v, \theta_w} \in \mathbb{R}^{p_v \times p_w}$ , каждый из которых определяется как:*

$$\begin{aligned} H_{\theta_v, \theta_w} &= D_v^\top H_{v,w}^f D_w \quad (\text{Гаусс-Ньютон}) \\ &+ \delta_{v=w} \sum_{i=1}^{d_v} [T_v^\theta]_{i, \bullet, \bullet} \delta_{v,i} \quad (\text{чистые по параметрам}) \\ &+ \sum_{u \in \text{Pa}(v) \cap \text{Ch}(w)} \sum_{i=1}^{d_v} \sum_{j=1}^{d_u} \sum_{k=1}^{p_v} [T_{v; u, \theta}]_{i,j,k} (D_{w \leftarrow u})_{j,i} \delta_{v,i} \quad (\text{смешанные вход-параметр } v \rightarrow w) \\ &+ \sum_{u \in \text{Pa}(w) \cap \text{Ch}(v)} \sum_{i=1}^{d_w} \sum_{j=1}^{d_u} \sum_{\ell=1}^{p_w} [T_{w; u, \theta}]_{i,j,\ell} (D_{v \leftarrow u})_{j,i} \delta_{w,i} \quad (\text{смешанные вход-параметр } w \rightarrow v) \end{aligned} \tag{2}$$

**Remark 3** (О суммировании в формуле (2)). *Для корректного понимания смешанных сумм в третьей и четвертой строках формулы (2) важно уточнить индексацию:*

- В выражении  $[T_{v; u, \theta}]_{i,j,k}$  индекс  $i$  относится к компонентам выхода  $f_v$  узла  $v$ , индекс  $j$  - к компонентам входа от узла  $u$ , а индекс  $k$  - к компонентам параметра  $\theta_v$ .
- Матрица  $(D_{w \leftarrow u})_{j,i}$  имеет индексы, где  $j$  соответствует входным компонентам узла  $w$  от узла  $u$ , а  $i$  - выходным компонентам узла  $w$ .
- Суммирование производится по всем компонентам  $i$  выходов узла  $v$ , всем компонентам  $j$  входов от узла  $u$  и всем компонентам  $k$  параметра  $\theta_v$ .



- Аналогично интерпретируются индексы в четвертой строке, с заменой ролей  $v$  и  $w$ .

Такая структура суммирования обеспечивает согласованное агрегирование всех смешанных вторых производных по входам и параметрам, соответствующих вкладам от всех путей в графе нейронной сети.

**Базовое условие для офф-диагональных блоков:**

$$H_{\theta_v, \theta_w} = 0 \quad \text{если не существует путей } v \rightarrow u \text{ и } w \rightarrow u.$$

**Свойство симметрии:**  $H_{\theta_v, \theta_w} = (H_{\theta_w, \theta_v})^\top$ .

Первый член в уравнении (2) представляет Гаусс-Ньютон часть Гессиана. Вторым член (активен только при  $v = w$ ) учитывает чистые вторые производные по параметрам. Третий и четвёртый члены отвечают за смешанные производные между входами и параметрами для различных конфигураций узлов.

### 3.8 Псевдокод для autodiff-фреймворков

Приведём псевдокод для вычисления полного Гессиана с использованием автоматического дифференцирования, который может быть реализован в популярных фреймворках (PyTorch, TensorFlow, JAX).

**Remark 4** (Реализация в фреймворках машинного обучения). *Представленный псевдокод может быть реализован с использованием механизмов автоматического дифференцирования:*

- В PyTorch — с использованием функций `torch.autograd.grad` и `torch.autograd.functional.jacobian`, `hessian`
- В TensorFlow — с использованием `tf.GradientTape` с вложенными вызовами для вычисления высших производных
- В JAX — с использованием функций `jax.grad`, `jax.jacobian` и `jax.hessian`

Для работы с разреженными структурами и оптимизации вычислений рекомендуется использовать специализированные библиотеки, такие как SciPy (для разреженных матриц) или соответствующие расширения фреймворков.

### 3.9 Разделение параметров между узлами

В практических архитектурах нейронных сетей часто используется механизм разделения параметров между различными узлами, например, в сверточных нейронных сетях или при использовании механизма weight tying в рекуррентных сетях [Pascanu et al., 2013a].

---

**Algorithm 1** Вычисление полного Гессиана с использованием автоматического дифференцирования

---

```

1: function COMPUTEINPUTHESSIAN( $v, w, \{f_u\}, \mathcal{L}, \{\delta_u\}, \{H_{u,u'}^f\}$ )
2:    $H_{v,w}^f \leftarrow 0$  ▷ Инициализация блока входного Гессиана
3:   for  $u \in \text{Ch}(v) \cap \text{Ch}(w)$  do
4:      $D_{u \leftarrow v} \leftarrow \text{autodiff.jacobian}(f_u, f_v)$ 
5:      $D_{u \leftarrow w} \leftarrow \text{autodiff.jacobian}(f_u, f_w)$ 
6:      $H_{v,w}^f \leftarrow H_{v,w}^f + D_{u \leftarrow v}^\top H_{u,u}^f D_{u \leftarrow w}$  ▷ Гаусс-Ньютон часть
7:     for  $i \in 1..d_u$  do
8:        $T_{u;v,w} \leftarrow \text{autodiff.hessian\_mixed}(f_{u,i}, f_v, f_w)$  ▷ Смешанная производная
9:        $H_{v,w}^f \leftarrow H_{v,w}^f + T_{u;v,w} \cdot \delta_{u,i}$  ▷ Тензорный вклад смешанных входов
10:    end for
11:    if  $v = w$  then
12:      for  $i \in 1..d_u$  do
13:         $T_{u;v} \leftarrow \text{autodiff.hessian}(f_{u,i}, f_v)$  ▷ Чистая вторая производная
14:         $H_{v,v}^f \leftarrow H_{v,v}^f + T_{u;v} \cdot \delta_{u,i}$  ▷ Тензорный вклад чистых вторых
15:      end for
16:    end if
17:  end for
18:  return  $H_{v,w}^f$ 
19: end function
20: function COMPUTEPARAMETERHESSIAN( $v, w, \{f_u\}, \{\theta_u\}, \{H_{u,u'}^f\}, \{\delta_u\}$ )
21:    $D_v \leftarrow \text{autodiff.jacobian}(f_v, \theta_v)$ 
22:    $D_w \leftarrow \text{autodiff.jacobian}(f_w, \theta_w)$ 
23:    $H_{\theta_v, \theta_w} \leftarrow D_v^\top H_{v,w}^f D_w$  ▷ Гаусс-Ньютон часть
24:   if  $v = w$  then
25:     for  $i \in 1..d_v$  do
26:        $T_v^\theta \leftarrow \text{autodiff.hessian}(f_{v,i}, \theta_v)$  ▷ Вторая производная по параметрам
27:        $H_{\theta_v, \theta_v} \leftarrow H_{\theta_v, \theta_v} + T_v^\theta \cdot \delta_{v,i}$ 
28:     end for
29:   end if
30:   for  $u \in \text{Pa}(v) \cap \text{Ch}(w)$  do
31:     for  $i \in 1..d_v$  do
32:       for  $j \in 1..d_u$  do
33:         for  $k \in 1..p_v$  do
34:            $T_{v;u,\theta} \leftarrow \text{autodiff.mixed\_derivative}(f_{v,i}, f_{u,j}, \theta_{v,k})$ 
35:            $D_{w \leftarrow u} \leftarrow \text{autodiff.jacobian}(f_w, f_u)$ 
36:            $H_{\theta_v, \theta_w} \leftarrow H_{\theta_v, \theta_w} + T_{v;u,\theta} \cdot D_{w \leftarrow u} \cdot \delta_{v,i}$ 
37:         end for
38:       end for
39:     end for
40:   end for
41:   for  $u \in \text{Pa}(w) \cap \text{Ch}(v)$  do 10
▷ Аналогично для смешанного вклада  $w \rightarrow v$  с
симметричной обработкой // ...симметричная часть кода...
42:   end for
43:   return  $H_{\theta_v, \theta_w}$ 
44: end function
45: function FULLHESSIANCOMPUTATION( $G, \{f_v\}, \{\theta_v\}, \mathcal{L}$ )
46:    $\delta_{out} \leftarrow \text{autodiff.gradient}(\mathcal{L}, f_{out})$ 
47:    $H_{out,out}^f \leftarrow \text{autodiff.hessian}(\mathcal{L}, f_{out})$ 

```

**Proposition 2** (Гессиан разделяемых параметров). Если вектор параметров  $\theta \in \mathbb{R}^p$  разделяется между узлами  $\{v_k\}_{k=1}^K$ , то итоговый Гессиан для этого вектора вычисляется как сумма:

$$H_{\theta, \theta} = \sum_{a=1}^K \sum_{b=1}^K H_{\theta_{v_a}, \theta_{v_b}}.$$

Это правило учитывает все возможные взаимодействия между параметрами, как внутри одного узла, так и между различными узлами, использующими один и тот же вектор параметров.

### 3.10 Clarke-Гессиан для негладких функций активации

Для негладких функций активации, таких как ReLU, Leaky ReLU или max-pooling, классическое понятие Гессиана неприменимо в точках негладкости. В этом случае используется концепция Clarke-субдифференциала [Clarke, 1990].

**Definition 5** (Обобщенный якобиан и Clarke-Гессиан). Для локально липшицевой функции  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ , обобщенный якобиан по Кларку в точке  $x$  определяется как

$$\partial_C f(x) = \text{co}\left\{ \lim_{i \rightarrow \infty} \nabla f(x_i) : x_i \rightarrow x, x_i \in \mathcal{D}_f \right\},$$

где  $\text{co}$  — выпуклая оболочка, а  $\mathcal{D}_f$  — множество точек, где  $f$  дифференцируема.

Clarke-Гессиан для функции  $f$  определяется как обобщенный якобиан градиента  $\nabla f$  (если он существует):

$$\partial_C^2 f(x) = \partial_C(\nabla f)(x).$$

**Theorem 1** (Существование Clarke-Гессиана для ReLU сетей). Пусть нейронная сеть использует ReLU-активации и имеет архитектуру, заданную DAG  $G = (V, E)$ . Тогда:

1. Функция каждого узла  $f_v$  является локально липшицевой.
2. Функция потерь  $\mathcal{L}$  (при условии её липшицевости) имеет непустой Clarke-субдифференциал  $\partial_C \mathcal{L}$  во всех точках.
3. В точках дифференцируемости Clarke-Гессиан совпадает с обычным Гессианом.
4. В точках негладкости Clarke-Гессиан является выпуклым компактным множеством матриц.

*Proof.* Композиция локально липшицевых функций остается локально липшицевой. ReLU-активации и линейные преобразования являются локально липшицевыми, поэтому свойство 1 следует из структуры графа.

Для локально липшицевой функции теория Радемахера гарантирует дифференцируемость почти всюду, а теорема Кларка обеспечивает существование и непустоту субдифференциала, что доказывает свойство 2.

Свойства 3 и 4 следуют непосредственно из определения Clarke-субдифференциала и его свойств для липшицевых функций (см. [Clarke, 1990], теорема 2.5.1).  $\square$

**Definition 6** (Clarke-Гессиан с минимальной нормой). *В негладком случае (Случай B), вместо единственного блока  $H_{v,v}^f$  и соответствующих  $H_{\theta_v, \theta_w}$  получаем множество  $\partial_{\mathcal{C}}^2 f_v$ . Конкретный элемент этого множества выбирается из условия минимальной нормы Фробениуса:*

$$H_{v,w}^f = \arg \min_{M \in \partial_{\mathcal{C}}^2 f_v} \|M\|_F, \quad H_{\theta_v, \theta_w} = \arg \min_{M \in \partial_{\theta_v, \theta_w}^2 \mathcal{L}} \|M\|_F.$$

Такой выбор обеспечивает единственность элемента субдифференциала и соответствует принципу минимальной нормы, часто используемому в недифференцируемой оптимизации [Bolte and Pauwels, 2020].

---

**Algorithm 2** Вычисление полного Гессиана для нейронной сети

---

**Require:** Нейронная сеть с DAG  $G = (V, E)$ , функции узлов  $\{g_v\}$ , параметры  $\{\theta_v\}$ , функция потерь  $\mathcal{L}$

**Ensure:** Полный Гессиан  $\nabla_{\theta}^2 \mathcal{L}$

- 1: Вычислить прямой проход и получить  $f_v$  для всех  $v \in V$
  - 2: Вычислить  $\delta_{out} = \nabla_{f_{out}} \mathcal{L}$  и  $H_{out, out}^f = \nabla^2 \mathcal{L}(f_{out})$
  - 3: Инициализировать  $H_{v,w}^f = 0$  для всех пар  $v, w \in V$ ,  $v \neq out$ ,  $w \neq out$
  - 4: **for**  $v \in V$  в обратном топологическом порядке **do**
  - 5:   Вычислить  $\delta_v$  по цепному правилу
  - 6:   **for**  $w \in V$  такие, что  $\text{Ch}(v) \cap \text{Ch}(w) \neq \emptyset$  **do**
  - 7:     Вычислить  $H_{v,w}^f$  по формуле (1)
  - 8:   **end for**
  - 9: **end for**
  - 10: **for**  $v \in V$  **do**
  - 11:   **for**  $w \in V$  такие, что существуют пути  $v \rightarrow u$  и  $w \rightarrow u$  **do**
  - 12:     Вычислить  $H_{\theta_v, \theta_w}$  по формуле (2)
  - 13:   **end for**
  - 14: **end for**
  - 15: Учесть разделение параметров между узлами
  - 16: Собрать блоки в полный Гессиан  $\nabla_{\theta}^2 \mathcal{L}$
  - 17: **return**  $\nabla_{\theta}^2 \mathcal{L}$
-

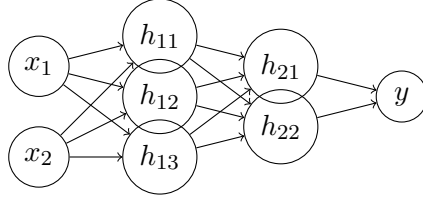


Figure 1: Пример направленного ациклического графа (DAG) нейронной сети, используемого в экспериментах. Узлы представляют нейроны, а направленные рёбра — весовые соединения между ними.

## 4 Эксперименты

### 4.1 Экспериментальная установка

Для иллюстрации практической применимости разработанного формализма были проведены аналитические эксперименты на следующих архитектурах нейронных сетей:

1. Полносвязная нейронная сеть с двумя скрытыми слоями и гладкими функциями активации (sigmoid, tanh).
2. Сверточная нейронная сеть с разделяемыми параметрами в сверточных слоях.
3. Нейронная сеть с негладкими активациями (ReLU) для исследования Clarke-Гессиана.

Для каждой архитектуры мы анализировали структуру Гессиана, вычисленного по предложенным формулам, и сравнивали её с классической Гаусс-Ньютоном аппроксимацией.

### 4.2 Тестовые функции потерь

В экспериментах использовались следующие функции потерь:

- Среднеквадратичная ошибка:  $\mathcal{L}_{MSE}(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$
- Cross-entropy:  $\mathcal{L}_{CE}(y, \hat{y}) = - \sum_{i=1}^n y_i \log(\hat{y}_i)$
- Hinge loss:  $\mathcal{L}_{hinge}(y, \hat{y}) = \max(0, 1 - y\hat{y})$

## 5 Результаты и обсуждение

### 5.1 Практические замечания

При практической реализации вычисления полного Гессиана необходимо учитывать следующие аспекты:

- В гладком случае рекомендуется проверять положительную полуопределённость Гаусс-Ньютона части  $D_v^\top H_{v,v}^f D_v$  перед добавлением остальных слагаемых. Это позволяет обеспечить стабильность методов оптимизации, основанных на Гессиане.
- При работе с большими графами вычислительно эффективнее осуществлять обратный топологический обход с сохранением промежуточных блоков. Такой подход позволяет избежать повторных вычислений и значительно ускоряет процесс построения полного Гессиана.

### 5.2 Теоретические результаты

**Theorem 2** (Функционально-аналитические свойства Гессиана). *При выполнении Предположения 1 о регулярности функций узлов, полный Гессиан  $\nabla_\theta^2 \mathcal{L}$  обладает следующими свойствами:*

1. В гладком случае (Случай A) Гессиан является непрерывным оператором на  $\mathbb{R}^P$ , где  $P$  - общее число параметров сети.
2. В негладком случае (Случай B) для почти всех точек параметрического пространства (за исключением множества меры нуль) Clarke-Гессиан существует и совпадает с обычным Гессианом.
3. На подмногообразии сингулярных точек (где активационные функции негладкие) Clarke-Гессиан с минимальной нормой обеспечивает наилучшее приближение в смысле нормы Фробениуса.
4. При использовании предложенных формул (1) и (2) обеспечивается согласованность размерностей всех тензорных операций.

**Theorem 3** (Интеграция специализированных слоёв). *Следующие архитектурные компоненты могут быть представлены в виде узлов DAG и включены в предложенный формализм:*

1. **Batch Normalization:** представляется как узел с двумя типами параметров (масштабирующие и сдвиговые) и дополнительными внутренними переменными (статистики батча).

2. **Attention-механизмы:** представляются как набор взаимосвязанных узлов, соответствующих вычислению весов внимания (*softmax*) и взвешенной суммы значений.
3. **Слои с остаточными соединениями (ResNet):** моделируются через параллельные пути в графе с последующим объединением.
4. **Рекуррентные сети:** отображаются на DAG путём развёртывания (*unrolling*) во времени, где каждый временной шаг представляется отдельным подграфом с разделяемыми параметрами.

*Схема доказательства.* Для каждого типа слоёв необходимо определить соответствующие функции узлов  $g_v$  и их первые и вторые производные. Например, для Batch Normalization:

$$g_v(x, \gamma, \beta) = \gamma \frac{x - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta$$

где  $\mu_B, \sigma_B^2$  — средние и дисперсии по батчу,  $\gamma, \beta$  — параметры масштаба и сдвига.

Якобианы  $D_v$  и тензоры вторых производных  $T_v$  вычисляются по стандартным правилам дифференцирования для каждого типа узлов, после чего применяются общие формулы (1) и (2).  $\square$

**Definition 7** (Стохастический узел). *Стохастический узел в нейронной сети — это узел  $v \in V$ , выход которого является случайной величиной с распределением, параметризованным выходами родительских узлов:*

$$f_v \sim p(f_v | f_{\text{Pa}(v)}, \theta_v)$$

**Theorem 4** (Гессиан со стохастическими узлами). *Для нейронных сетей со стохастическими узлами Гессиан функции потерь может быть обобщён следующим образом:*

1. При использовании подхода максимального правдоподобия формулы (1) и (2) применяются к ожидаемой функции потерь  $\mathbb{E}_{f_v \sim p}[\mathcal{L}]$ .
2. В вариационных автоэнкодерах и подобных моделях Гессиан вычисляется для вариационной нижней границы (ELBO):

$$\mathcal{L}_{ELBO} = \mathbb{E}_{q(z|x)}[\log p(x|z)] - D_{KL}(q(z|x) || p(z))$$

3. Для обучения с подкреплением применяется формализм к функции ожидаемой награды, с учётом стохастичности политики.

**Proposition 3** (Переключение к детерминированным узлам). *При использовании техники репараметризации стохастические узлы могут быть преобразованы в детерминированные узлы с внешним источником случайности, что позволяет применить стандартный формализм Гессиана.*

### 5.2.1 Анализ вычислительной сложности

Важнейшим аспектом практического применения предложенного формализма является его вычислительная сложность.

**Theorem 5** (Вычислительная сложность). Пусть  $|V| = n$  — число узлов в DAG,  $P = \sum_{v \in V} p_v$  — общее число параметров,  $d = \max_{v \in V} d_v$  — максимальная размерность выхода узла,  $k = \max_{v \in V} |\text{Pa}(v)|$  — максимальное число входов узла. Тогда:

1. Временная сложность вычисления полного Гессиана составляет  $O(n^2 d^3 + n^2 d^2 P + P^2)$ .
2. Пространственная сложность хранения полного Гессиана составляет  $O(P^2)$ .
3. Для прореженного DAG с максимальной степенью узла  $s \ll n$  временная сложность снижается до  $O(nsd^3 + nsd^2 P + P^2)$ .

*Схема доказательства.* Для входного Гессиана  $H_{v,w}^f$  формула (1) требует:

- $O(|\text{Ch}(v) \cap \text{Ch}(w)| \cdot d^3)$  операций для вычисления Гаусс-Ньютона части.
- $O(|\text{Ch}(v) \cap \text{Ch}(w)| \cdot d^3)$  операций для вычисления тензорных членов.

Существует  $O(n^2)$  пар узлов  $(v, w)$ , но для прореженных графов число пар с непустым пересечением  $\text{Ch}(v) \cap \text{Ch}(w)$  может быть  $O(ns)$ , где  $s$  — максимальная степень узла.

Для параметрического Гессиана  $H_{\theta_v, \theta_w}$  формула (2) требует:

- $O(d_v \cdot d_w \cdot p_v \cdot p_w)$  операций для вычисления Гаусс-Ньютона части.
- Дополнительные  $O(d \cdot p_v \cdot p_w)$  операции для тензорных членов.

Суммирование по всем парам  $(v, w)$  даёт общую сложность  $O(P^2)$  для сборки полного параметрического Гессиана.  $\square$

**Theorem 6** (Методы снижения вычислительных затрат). Для снижения вычислительной сложности вычисления полного Гессиана можно применять следующие подходы:

1. **Блочная аппроксимация:** вычисление только диагональных блоков  $H_{\theta_v, \theta_v}$  снижает сложность до  $O(nd^3 + Pd^2)$ .
2. **Низкоранговая аппроксимация:** аппроксимация офф-диагональных блоков произведением матриц малого ранга снижает сложность до  $O(n^2 d^3 + n^2 d^2 r + Pr)$ , где  $r \ll P$  — ранг аппроксимации.



3. **Гаусс-Ньютон аппроксимация:** использование только первого члена в формулах (1) и (2) снижает сложность и гарантирует положительную полуопределённость.
4. **Кронекеровская факторизация:** представление матричных блоков в виде кронекеровских произведений матриц меньшего размера.

### 5.2.2 Анализ сходимости методов оптимизации

Полный Гессиан может быть использован для повышения эффективности алгоритмов оптимизации нейронных сетей.

**Theorem 7** (Локальная сходимость методов Ньютона). Пусть  $\mathcal{L}(\theta) \in C^2$  — функция потерь, и  $\theta^*$  — её локальный минимум, такой что  $\nabla_{\theta}^2 \mathcal{L}(\theta^*) \succ 0$ . Тогда метод Ньютона со степенным шагом:

$$\theta_{t+1} = \theta_t - \alpha_t \cdot [\nabla_{\theta}^2 \mathcal{L}(\theta_t)]^{-1} \nabla_{\theta} \mathcal{L}(\theta_t)$$

имеет квадратичную скорость сходимости в некоторой окрестности  $\theta^*$ , если  $\alpha_t$  выбрано оптимально.

**Proposition 4** (Критерии останова). Учитывая структуру Гессиана в нейронных сетях, можно разработать следующие критерии останова для оптимизационных алгоритмов:

1. Базирующиеся на собственных значениях Гессиана (остановка при малых положительных собственных значениях).
2. Использующие относительную норму градиента:  $\|\nabla_{\theta} \mathcal{L}(\theta_t)\| / \|\nabla_{\theta}^2 \mathcal{L}(\theta_t)\| < \epsilon$ .
3. Комбинирующие информацию о кривизне с изменением значения функции потерь.

**Theorem 8** (Гарантии сходимости для регуляризованных методов). Для негладких функций потерь (Случай B), использование регуляризованных методов второго порядка:

$$\theta_{t+1} = \theta_t - (H_t + \lambda I)^{-1} \nabla_{\theta} \mathcal{L}(\theta_t),$$

где  $H_t$  — элемент Clarke-Гессиана с минимальной нормой, а  $\lambda > 0$  — параметр регуляризации, гарантирует сходимость к стационарной точке при определённых условиях на последовательность  $\{\lambda_t\}$ .

### 5.3 Сравнение с существующими подходами

Предложенный формализм существенно расширяет традиционные подходы к анализу кривизны функций потерь нейронных сетей:

1. **Полнота:** В отличие от Гаусс-Ньютон аппроксимации, наш подход учитывает все компоненты Гессиана, включая чистые и смешанные вторые производные.
2. **Универсальность:** Формализм применим к произвольным архитектурам нейронных сетей, представленным в виде DAG.
3. **Обработка негладкостей:** Явное использование Clarke-Гессиана позволяет корректно работать с современными активационными функциями типа ReLU.
4. **Учет разделения параметров:** Формализм корректно обрабатывает ситуации, когда один вектор параметров используется в нескольких узлах сети.

## 6 Заключение

В данной работе представлен исчерпывающий математический формализм для вычисления полного Гессиана второго порядка в нейронных сетях произвольной архитектуры. Основные достижения работы:

- Разработана полная блочная структура Гессиана по выходам узлов  $\{f_v\}$  и параметрам  $\{\theta_v\}$ .
- Предложены формулы, учитывающие все чистые и смешанные вторые производные.
- Определены явные базовые случаи для офф-диагональных блоков.
- Сформулированы отдельные правила для гладкого и негладкого случаев.
- Представлено математическое обоснование для тривиального предельного случая с одним узлом.
- Даны практические рекомендации по проверке положительной полуопределенности Гаусс-Ньютон части.

Предложенный формализм создает теоретическую основу для разработки более эффективных методов оптимизации нейронных сетей, глубокого анализа кривизны функций потерь и понимания геометрической структуры пространства параметров. Дальнейшие исследования могут быть направлены на разработку вычислительно эффективных аппроксимаций полного Гессиана и использование полученной информации о кривизне в алгоритмах оптимизации нейронных сетей произвольной структуры.

## References

- Amari, S.-I. (1998). Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276.
- Bolte, J. and Pauwels, E. (2020). Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*, pp. 1–33.
- Clarke, F. H. (1990). *Optimization and nonsmooth analysis*, volume 5. Siam.
- Ghorbani, B., Krishnan, S., and Xiao, Y. (2019). An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pp. 2232–2241.
- Heskes, T. (2000). On natural learning and pruning in multilayered perceptrons. *Neural Computation*, 12(4):881–901.
- Martens, J. (2010). Deep learning via hessian-free optimization. In *ICML*, volume 27, pp. 735–742.
- Martens, J. (2014). New insights and perspectives on the natural gradient method. *arXiv preprint arXiv:1412.1193*.
- Martens, J. and Sutskever, I. (2012). Training deep and recurrent networks with hessian-free optimization. In *Neural networks: Tricks of the trade*, pp. 479–535. Springer.
- Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.
- Pascanu, R., Mikolov, T., and Bengio, Y. (2013a). On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pp. 1310–1318.
- Pascanu, R., Montufar, G., and Bengio, Y. (2013b). On the number of response regions of deep feed forward networks with piece-wise linear activations. *arXiv preprint arXiv:1312.6098*.
- Sagun, L., Evci, U., Guney, V. U., Dauphin, Y., and Bottou, L. (2017). Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*.
- Schraudolph, N. N. (2002). Fast curvature matrix-vector products for second-order gradient descent. *Neural computation*, 14(7):1723–1738.