

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

**Аноним**  
Аффилиация  
anonymous@example.com

June 10, 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 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 Гессиана между узлами $v$ и $w$	$\mathbb{R}^{d_v \times d_w}$
$H_{\theta_v, \theta_w}$	Блок параметрического Гессиана	$\mathbb{R}^{p_v \times p_w}$
$\partial_C^2 f_v$	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

### 3.2 Функциональные пространства и аналитические предпосылки

Прежде чем перейти к определению компонентов и структуры Гессияна, 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  $g_v \in C^2(\mathbb{R}^{\sum_{u \in \text{Pa}(v)} d_u}, \mathbb{R}^{d_v})$ , т.е. дважды непрерывно дифференцируема по всем входам и параметрам.
2. В негладком случае (Случай B): функция узла  $g_v \in PC^2(\mathbb{R}^{\sum_{u \in \text{Pa}(v)} d_u}, \mathbb{R}^{d_v})$  и локально липшицева, т.е. является кусочно дважды дифференцируемой с границами кусков, образующими множество меры нуль.

**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 по теореме Радемахера), 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.

### 3.3 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  $out$ .

### 3.4 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:

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

**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  $t$  относятся к компонентам входов от родительских узлов.
- Индексы  $\alpha$  и  $\beta$  (вместо иногда используемых  $k$ ,  $\ell$ ) относятся к компонентам параметров  $\theta_v$ .

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

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 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 3** (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-Гессiana для ReLU-сетей). Пусть нейронная сеть использует активации  $\text{ReLU}(t) = \max\{0, t\}$  и имеет DAG вычислений  $G = (V, E)$ . Обозначим через  $\mathcal{L} : \mathbb{R}^d \rightarrow \mathbb{R}$  функцию потерь, полученную как композицию сети  $F : \mathbb{R}^d \rightarrow \mathbb{R}^m$  с внешней функцией  $\ell : \mathbb{R}^m \rightarrow \mathbb{R}$ :

$$\mathcal{L}(x) = \ell(F(x)).$$

Предположим, что  $\ell \in C^2(\mathbb{R}^m, \mathbb{R})$  и локально липшицева. Тогда

1. Каждая функция узла  $f_v$  локально липшицева.
2. Для входов  $x$ , где отображение  $x \mapsto [f_{u_1}(x), \dots, f_{u_k}(x)]^\top$  имеет локально полный ранг для всех узлов, функция  $\mathcal{L}$  дважды дифференцируема в  $x$ . Без этого рангового условия множество точек негладкости может иметь положительную меру, как показывает пример 1-D ReLU сети  $F(x) = \max\{0, x\}$  с  $\mathcal{L}(x) = F(x)^2/2$ , где функция дважды недифференцируема на  $(-\infty, 0]$ .
3. Во всех точках  $x$ , где  $\mathcal{L}$  дважды дифференцируема, Clarke-Гессиан  $\partial_C^2 \mathcal{L}(x)$  вырождается в одиночное множество, совпадающее с обычным Гессианом  $\nabla^2 \mathcal{L}(x)$ .
4. На подмногообразии нулевой меры, соответствующем границам линейных регионов ReLU, Clarke-Гессиан  $\partial_C^2 \mathcal{L}(x)$  представляет собой непустое, выпуклое и компактное множество матриц при условии, что градиент  $\nabla \mathcal{L}(x)$  существует.

*Proof.* **Шаг 1 (локальная липшицевость каждого узла).** Рассмотрим топологический порядок вершин  $v_1, \dots, v_{|V|}$ . Для входного узла  $v_1$  функция  $f_{v_1}(x) = x$  очевидно 1-липшицева. Пусть узел  $v$  получает выходы  $f_{u_1}, \dots, f_{u_k}$  предыдущих вершин и применяет линейное преобразование  $W_v(\cdot) + b_v$ , за которым следует ReLU:

$$f_v(x) = \text{ReLU}(W_v[f_{u_1}(x), \dots, f_{u_k}(x)]^\top + b_v).$$

Линейное отображение имеет константу Липшица  $\|W_v\|_2$ , а ReLU — константу

1. Следовательно,  $f_v$   $(\prod_{i=1}^k L_{u_i})\|W_v\|_2$ -липшицева, где  $L_{u_i}$  — константа для  $f_{u_i}$ . Индукцией по порядку вершин получаем локальную липшицевость всех  $f_v$ .

**Шаг 2 (мера множества гладкости).** Заметим, что функция  $\mathcal{L}$  негладка только на подмногообразиях, соответствующих границам линейных регионов ReLU, которые в общем случае могут иметь положительную меру. Для каждого нейрона с ReLU-активацией множество точек, где преактивация равна нулю, есть решение уравнения вида  $W_v[f_{u_1}(x), \dots, f_{u_k}(x)]^\top + b_v = 0$ . При фиксированных параметрах  $W_v$  и  $b_v$  и при условии, что отображение  $x \mapsto [f_{u_1}(x), \dots, f_{u_k}(x)]^\top$  имеет локально полный



ранг, это уравнение задаёт гиперповерхность (подмногообразие коразмерности 1) в пространстве входов, имеющую нулевую меру Лебега.

Однако для узких или свёрточных слоёв это ранговое условие может нарушаться. Рассмотрим простой пример: 1-D ReLU сеть  $F(x) = \max\{0, x\}$  с функцией потерь  $\mathcal{L}(x) = F(x)^2/2$ . Здесь функция  $\mathcal{L}$  дважды недифференцируема на  $(-\infty, 0]$ , который имеет положительную меру Лебега. Таким образом, утверждение о дифференцируемости "почти всюду" справедливо только при дополнительном ранговом предположении.

Если ранговое условие выполнено, то согласно результатам Hanin and Rolnick [2019] и Serra et al. [2018], множество точек негладкости ReLU-сети с  $L$  слоями и общим числом нейронов  $N$  может быть покрыто не более чем  $2^N$  аффинными подпространствами коразмерности 1, каждое из которых имеет меру Лебега нуль.

**Шаг 3 (совпадение Гессианов в гладких точках).** Пусть  $x$  — точка, где  $\mathcal{L}$  дважды дифференцируема. Тогда градиент  $\nabla \mathcal{L}$  непрерывен в окрестности  $x$  и дифференцируем в  $x$ , так что по определению обобщённого Гессиана

$$\partial_C^2 \mathcal{L}(x) = \{\nabla^2 \mathcal{L}(x)\}.$$

В общем случае внутри одного линейного региона сети функция  $F$  аффинна, т.е.  $F(x) = Ax + b$  для некоторых  $A$  и  $b$ . Если внешняя функция  $\ell \in C^2$ , то применяя цепное правило, получаем:

$$\nabla^2 \mathcal{L}(x) = A^\top \nabla^2 \ell(F(x)) A.$$

Для типичных функций потерь, таких как квадратичная или кросс-энтропийная,  $\nabla^2 \ell$  хорошо определено и ненулевое.

**Шаг 4 (существование Clarke-Гессиана в негладких точках).** В отличие от стандартного цепного правила для первых производных, для вторых производных композиции функций в негладком случае следует использовать обобщённое цепное правило для Clarke-Гессиана [??].

Для функции  $\mathcal{L}(x) = \ell(F(x))$ , где  $\ell \in C^2$  и  $F$  локально липшицева с существующим градиентом, Clarke-Гессиан  $\partial_C^2 \mathcal{L}(x)$  является непустым, выпуклым и **компактным**, поскольку все множества  $\partial_C^2 F_i(x)$  ограничены (см. [?, Thm 3.46]) и итоговая выпуклая оболочка конечного объединения ограниченных замкнутых множеств остаётся компактной.

Таким образом, все четыре утверждения теоремы доказаны.  $\square$

**Definition 4** (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  $\partial_C^2 f_v$ . Конкретный элемент этого множества выбирается из условия минимизации квадрата нормы Фробениуса:

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

**Remark 3** (О единственности элемента минимальной нормы). Квадрат нормы Фробениуса  $\|M\|_F^2$  является строго выпуклой функцией от  $M$ , а множество  $\partial_C^2 f_v$  выпукло и компактно. Следовательно, задача минимизации  $\|M\|_F^2$  имеет единственное решение, что обеспечивает однозначность выбора элемента из субдифференциала.

### 3.7 Полный входной Гессиан

**Definition 5** (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:

**Definition 6** (Псевдообратная транспонированная матрица). Для матрицы  $A \in \mathbb{R}^{m \times n}$  обозначим через  $A^\dagger$  псевдообратную матрицу Мура-Пенроуза. Тогда  $A^{-\top} := (A^\top)^\dagger$  — псевдообратная транспонированная матрица.

$$\begin{aligned}
H_{v,w}^f = & \sum_{u \in \text{Ch}(v) \cap \text{Ch}(w)} D_{u \leftarrow v}^\top H_{u,u}^f D_{u \leftarrow w} \quad (\text{Гаусс-Ньютон}) \\
& + \sum_{u \in \text{Ch}(v) \cap \text{Ch}(w)} \sum_{i=1}^{d_u} [T_{u;v,w}^{\text{sym}}]_{i,\bullet,\bullet} \delta_{u,i} \quad (\text{смешанные входы}) \\
& + \mathbf{1}_{v=w} \sum_{u \in \text{Ch}(v)} \sum_{i=1}^{d_u} [T_{u;v}]_{i,\bullet,\bullet} \delta_{u,i} \quad (\text{чистые по одному входу}) \\
& + \sum_{u \in \text{Ch}(v)} \sum_{z \in \text{Ch}(u) \cap \text{Pa}(w)} D_{u \leftarrow v}^\top H_{u,z}^f D_{w \leftarrow z}^{-\top} \quad (\text{путь } v \rightarrow^* w) \\
& + \sum_{u \in \text{Ch}(w)} \sum_{z \in \text{Ch}(u) \cap \text{Pa}(v)} D_{v \leftarrow z}^{-\top} H_{z,u}^f D_{u \leftarrow w}^\top \quad (\text{путь } w \rightarrow^* v) \\
& + \frac{\partial^2 \mathcal{L}}{\partial f_v \partial f_w} \quad (\text{прямая зависимость потерь от узлов})
\end{aligned} \tag{1}$$

с базовыми условиями:

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

**Remark 4** (О симметризации тензоров в формуле (1)). Во втором слагаемом формулы (1) используются симметризованные тензоры смешанных вторых производных

$T_{u;v,w}^{sym}$ , определяемые как:

$$[T_{u;v,w}^{sym}]_{i,j,k} = \frac{1}{2}([T_{u;v,w}]_{i,j,k} + [T_{u;w,v}]_{i,k,j})$$

Эта симметризация необходима для обеспечения корректности формулы в негладком случае (Случай B), где равенство смешанных частных производных может нарушаться. В гладком случае (Случай A) справедливо равенство  $T_{u;v,w}^{sym} = T_{u;v,w} = T_{u;w,v}^\top$  согласно теореме Шварца.

В четвертом и пятом слагаемых формулы через  $D_{v \leftarrow z}^{-\top}$  обозначается псевдообратная матрица к  $D_{v \leftarrow z}^\top$ , которая обеспечивает корректную передачу влияния по пути от одного узла к другому.

**Remark 5** (О псевдообратных матрицах в формуле (1)). В формуле (1) используется псевдообратная матрица Мура-Пенроуза  $D_{v \leftarrow z}^{-\top}$ , которая определяется как псевдообратная к  $D_{v \leftarrow z}^\top$ , т.е.  $D_{v \leftarrow z}^{-\top} = (D_{v \leftarrow z}^\top)^\dagger$ . Для невырожденной квадратной матрицы  $D_{v \leftarrow z}^\top$  псевдообратная совпадает с обычной обратной матрицей  $(D_{v \leftarrow z}^\top)^{-1}$ .

В случае, когда  $D_{v \leftarrow z}^\top$  является прямоугольной или вырожденной матрицей (например, в узких слоях нейронной сети), псевдообратная Мура-Пенроуза обеспечивает решение с минимальной нормой. Для вычисления можно использовать сингулярное разложение (SVD):

$$D_{v \leftarrow z}^\top = U \Sigma V^*, \quad D_{v \leftarrow z}^{-\top} = V \Sigma^\dagger U^*,$$

где  $\Sigma^\dagger$  получается заменой ненулевых сингулярных чисел на их обратные значения, а нулевые сингулярные числа остаются нулями.

Для практических реализаций рекомендуется использовать численно устойчивые алгоритмы вычисления псевдообратной матрицы с регуляризацией при малых сингулярных числах.

**Lemma 1** (О корректности псевдообратного распространения). Пусть  $z \in \text{Pa}(w) \cap \text{Ch}(u)$  и  $v \in \text{Pa}(u)$  — узлы графа, образующие путь  $v \rightarrow u \rightarrow z \rightarrow w$ . Тогда вклад в блок Гессiana  $H_{v,w}^f$  от этого пути корректно выражается через псевдообратную матрицу Мура-Пенроуза:

$$D_{u \leftarrow v}^\top H_{u,z}^f D_{w \leftarrow z}^{-\top},$$

где  $D_{w \leftarrow z}^{-\top} = (D_{w \leftarrow z}^\top)^\dagger$  — псевдообратная к  $D_{w \leftarrow z}^\top$ . Этот результат является следствием обобщенного цепного правила для дифференцирования композиций функций с ранг-дефицитными якобианами.

*Proof.* Рассмотрим функцию потерь  $\mathcal{L}$  как композицию преобразований вдоль пути:

$$\mathcal{L}(f_v) = \mathcal{L}(f_w(f_z(f_u(f_v)))).$$

Для первых производных по цепному правилу имеем:

$$\frac{\partial \mathcal{L}}{\partial f_v} = \frac{\partial f_u}{\partial f_v}^\top \frac{\partial f_z}{\partial f_u}^\top \frac{\partial f_w}{\partial f_z}^\top \frac{\partial \mathcal{L}}{\partial f_w} = D_{u \leftarrow v}^\top D_{z \leftarrow u}^\top D_{w \leftarrow z}^\top \delta_w.$$

При дифференцировании второй раз получаем выражения, включающие вторые производные. В общем случае нелинейных функций  $f_u$ ,  $f_z$  и  $f_w$  эти выражения содержат как первые, так и вторые производные. По обобщенному цепному правилу для композиций функций с ранг-дефицитными якобианами [Ben-Israel and Greville, 2003, Теорема 2.8], вторая производная включает член:

$$D_{u \leftarrow v}^\top \frac{\partial^2 \mathcal{L}}{\partial f_u \partial f_z} (D_{w \leftarrow z}^\top)^\dagger,$$

где используется псевдообратная матрица  $(D_{w \leftarrow z}^\top)^\dagger$  для корректной передачи влияния производных через промежуточные преобразования с неполным рангом.

Выражение  $\frac{\partial^2 \mathcal{L}}{\partial f_u \partial f_z}$  соответствует блоку входного Гессiana  $H_{u,z}^f$ . Таким образом, полное выражение  $D_{u \leftarrow v}^\top H_{u,z}^f D_{w \leftarrow z}^{-\top}$  правильно учитывает влияние изменений  $f_v$  на  $f_w$  через путь  $v \rightarrow u \rightarrow z \rightarrow w$  даже в случаях, когда якобианы имеют неполный ранг.

В специальном случае, когда все якобианы невырождены и квадратные, это выражение сводится к классическому цепному правилу с обычными обратными матрицами. В общем же случае использование псевдообратных матриц обеспечивает математически корректное обобщение, минимизирующее норму решения [Magnus and Neudecker, 2019].  $\square$

**Remark 6** (О неповторяющихся путях в формуле (1)). В четвертом и пятом слагаемых формулы (1) используется двойное суммирование по  $u \in \text{Ch}(v)$  и  $z \in \text{Ch}(u) \cap \text{Pa}(w)$  (для путей  $v \rightarrow^* w$ ) или по  $u \in \text{Ch}(w)$  и  $z \in \text{Ch}(u) \cap \text{Pa}(v)$  (для путей  $w \rightarrow^* v$ ). Может возникнуть вопрос о возможном двойном учёте одних и тех же путей.

Важно отметить, что двойной учёт исключается благодаря структуре направленного ациклического графа (DAG) и рекурсивному характеру вычислений. Для каждой пары узлов  $(v, w)$  каждый возможный путь от  $v$  к  $w$  или от  $w$  к  $v$  учитывается ровно один раз, так как:

1. Узел  $u$  является непосредственным потомком  $v$  (первый уровень пути).
2. Узел  $z$  является потомком  $u$  и родителем  $w$  (промежуточный уровень пути).
3. Мы суммируем по всем таким путям, но каждый путь определяется уникальной комбинацией  $(u, z)$ .

Рекурсивная природа вычислений гарантирует, что длинные пути декомпозируются на последовательность коротких, и что вклад каждого пути учитывается ровно один раз в окончательной сумме.

**Theorem 2** (О ненулевых блоках входного Гессиана). Блок  $H_{v,w}^f$  может быть ненулевым в любом из следующих случаев:

1. Существует путь от  $v$  к некоторому узлу  $u$  и путь от  $w$  к тому же узлу  $u$ , формально:  $\exists u \in V : v \rightarrow^* u$  и  $w \rightarrow^* u$ , где  $\rightarrow^*$  обозначает наличие пути в графе  $G$ .
2. Существует путь от  $v$  к  $w$  или от  $w$  к  $v$ , т.е.  $v \rightarrow^* w$  или  $w \rightarrow^* v$ .
3. Существует функциональная зависимость  $\mathcal{L}$  от обоих узлов  $f_v$  и  $f_w$  напрямую, т.е.  $\frac{\partial^2 \mathcal{L}}{\partial f_v \partial f_w} \neq 0$ .

*Proof.* Рассмотрим формулу (1) для блока  $H_{v,w}^f$ :

В случае 1, если существуют пути  $v \rightarrow^* u$  и  $w \rightarrow^* u$ , то возможны два случая: (а) существует общий потомок  $c \in \text{Ch}(v) \cap \text{Ch}(w)$ , что даёт ненулевой вклад через первые два слагаемых; (б) такого общего потомка нет, но через последовательность других узлов существует путь к общему узлу  $u$ .

В случае 2, если  $v \rightarrow^* w$ , то четвертое слагаемое формулы становится ненулевым, учитывая передачу влияния по пути от  $v$  к  $w$ . Аналогично, если  $w \rightarrow^* v$ , то ненулевым становится пятое слагаемое.

В случае 3, последнее слагаемое  $\frac{\partial^2 \mathcal{L}}{\partial f_v \partial f_w}$  явно ненулевое.

Если ни одно из этих условий не выполняется, то изменения выходов  $f_v$  и  $f_w$  влияют на непересекающиеся подмножества переменных, от которых зависит  $\mathcal{L}$ , и следовательно  $H_{v,w}^f = 0$ .  $\square$

**Proposition 2** (О вычислении ненулевых блоков). Для эффективного вычисления ненулевых блоков  $H_{v,w}^f$  в случае одностороннего пути (например,  $v \rightarrow^* w$ ), можно использовать рекуррентное соотношение:

$$H_{v,w}^f = \sum_{u \in \text{Ch}(v)} D_{u \leftarrow v}^\top H_{u,w}^f$$

для узлов  $v$ , которые не имеют общих потомков с  $w$ . Это позволяет избежать явного вычисления псевдообратных матриц в формуле (1).

**Свойство симметрии:** В гладком случае (Случай А)  $H_{v,w}^f = (H_{w,v}^f)^\top$  для всех  $v, w \in V$ , что следует из равенства смешанных частных производных для дважды непрерывно дифференцируемых функций.

В негладком случае (Случай В) для элементов Clarke-Гессиана минимальной нормы симметрия может не выполняться. В этом случае можно произвести симметризацию:  $\hat{H}_{v,w}^f = \frac{1}{2}(H_{v,w}^f + (H_{w,v}^f)^\top)$ .

**Remark 7** (Об использовании симметризации в негладком случае). Следует отметить, что симметризация Clarke-Гессiana изменяет его спектральные свойства. Если исходные матрицы  $H_{v,w}^f$  и  $(H_{w,v}^f)^\top$  имеют разные собственные значения, то симметризованная версия  $\hat{H}_{v,w}^f$  будет иметь другой спектр. Это может влиять на методы оптимизации, использующие обратный Гессиян  $H^{-1}$ , такие как метод Ньютона.

Симметризация рекомендуется в следующих случаях:

- Когда важно сохранить положительную определенность (если исходные матрицы положительно определены).
- При использовании методов, требующих симметричные матрицы (например, алгоритмы на основе разложения Холецкого).

Симметризацию не рекомендуется применять, когда асимметрия Гессiana несет важную информацию о кривизне функции потерь в негладких точках или когда требуется точное вычисление направления Ньютона.

### 3.8 Полный параметрический Гессиян

**Definition 7** (Параметрический Гессиян). Полный Гессиян 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:

$$\begin{aligned}
 H_{\theta_v, \theta_w} = & D_v^\top H_{v,w}^f D_w && \text{(блок Гаусса-Ньютона)} \\
 & + \mathbf{1}_{v=w} \sum_{i=1}^{d_v} \delta_{v,i} [T_v^\theta]_{i,\bullet,\bullet} && \text{(чистые по } \theta_v) \\
 & + \sum_{u \in \text{Pa}(v) \cap \text{Ch}(w)} \sum_{i=1}^{d_v} \delta_{v,i} [T_{v;u,\theta}]_{i,:,:} (D_{w \leftarrow u} D_w)^\top && \text{(вход-парам. } v \rightarrow w) \\
 & + \sum_{u \in \text{Pa}(w) \cap \text{Ch}(v)} \sum_{i=1}^{d_w} \delta_{w,i} [T_{w;u,\theta}]_{i,:,:} (D_{v \leftarrow u} D_v)^\top && \text{(вход-парам. } w \rightarrow v)
 \end{aligned} \tag{3}$$

**Remark 8** (О согласованности размерностей в формуле (3)). При вычислении третьего слагаемого в формуле (3), тензор  $[T_{v;u,\theta}]_{i,:,:}$  имеет размерность  $d_u \times p_v$ . Для согласованности размерностей необходимо использовать  $(D_{w \leftarrow u} D_w)^\top$  размерности  $p_w \times d_u$ , а не  $D_{w \leftarrow u} D_w$  (размерности  $d_u \times p_w$ ). Это обеспечивает получение матрицы  $p_v \times p_w$ , соответствующей требуемой размерности блока  $H_{\theta_v, \theta_w}$ . Аналогичное замечание справедливо для четвертого слагаемого.

**Theorem 3** (Сборка локальных блоков в глобальный Гессиан). Пусть параметры всей сети

$$\theta = \begin{pmatrix} \theta_{v_1} \\ \theta_{v_2} \\ \vdots \\ \theta_{v_n} \end{pmatrix} \in \mathbb{R}^P, \quad P = \sum_{k=1}^n p_{v_k},$$

и функция потерь  $\mathcal{L} = \mathcal{L}(\theta) \in C^2(\mathbb{R}^P)$ . Обозначим

$$H_{\theta_{v_i}, \theta_{v_j}} = \frac{\partial^2 \mathcal{L}}{\partial \theta_{v_i} \partial \theta_{v_j}} \in \mathbb{R}^{p_{v_i} \times p_{v_j}}, \quad i, j = 1, \dots, n.$$

Тогда полный Гессиан  $\nabla_{\theta}^2 \mathcal{L} \in \mathbb{R}^{P \times P}$  разбивается на блоки

$$\nabla_{\theta}^2 \mathcal{L} = \begin{pmatrix} H_{\theta_{v_1}, \theta_{v_1}} & H_{\theta_{v_1}, \theta_{v_2}} & \cdots & H_{\theta_{v_1}, \theta_{v_n}} \\ H_{\theta_{v_2}, \theta_{v_1}} & H_{\theta_{v_2}, \theta_{v_2}} & \cdots & H_{\theta_{v_2}, \theta_{v_n}} \\ \vdots & \vdots & \ddots & \vdots \\ H_{\theta_{v_n}, \theta_{v_1}} & H_{\theta_{v_n}, \theta_{v_2}} & \cdots & H_{\theta_{v_n}, \theta_{v_n}} \end{pmatrix}.$$

*Proof.* По определению Гессиана

$$\nabla_{\theta}^2 \mathcal{L} = \frac{\partial}{\partial \theta} (\nabla_{\theta} \mathcal{L}) \in \mathbb{R}^{P \times P},$$

где  $\nabla_{\theta} \mathcal{L} \in \mathbb{R}^P$  записывается в виде  $(\partial \mathcal{L} / \partial \theta_{v_1}, \dots, \partial \mathcal{L} / \partial \theta_{v_n})^{\top}$ . Разбиение вектора  $\theta$  на блоки по  $\theta_{v_i}$  естественным образом даёт блочную структуру у матрицы вторых производных:

$$[\nabla_{\theta}^2 \mathcal{L}]_{(v_i), (v_j)} = \frac{\partial}{\partial \theta_{v_j}} \left( \frac{\partial \mathcal{L}}{\partial \theta_{v_i}} \right) = \frac{\partial^2 \mathcal{L}}{\partial \theta_{v_i} \partial \theta_{v_j}} = H_{\theta_{v_i}, \theta_{v_j}}.$$

Поскольку  $\mathcal{L} \in C^2$ , блоки симметричны:

$$H_{\theta_{v_i}, \theta_{v_j}} = \left( H_{\theta_{v_j}, \theta_{v_i}} \right)^{\top}.$$

Собирая все  $n^2$  блоков, получаем заявленную матрицу. □

### 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].

**Proposition 3** (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.



## 4 Алгоритмы вычисления

### 4.1 Общий алгоритм вычисления полного Гессиана

---

**Algorithm 1** 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 (3)
  - 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**  $\nabla_{\theta}^2 \mathcal{L}$
- 

## 5 Теоретические результаты

### 5.1 Функционально-аналитические свойства Гессиана

**Theorem 4** (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 (3) обеспечивается согласованность размерностей всех тензорных операций.

## 5.2 Интеграция специализированных архитектурных компонентов

**Theorem 5** (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 (3).  $\square$

## 5.3 Стохастические узлы и вариационные подходы

**Definition 8** (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 6** (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 (3) 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 4** (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.

## 6 Анализ вычислительной сложности

**Theorem 7** (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,  $s = \max_{v \in V} |\text{Pa}(v) \cup \text{Ch}(v)|$  — максимальная степень узла. Тогда:

1. Временная сложность вычисления полного Гессиана составляет  $O(nsd^3 + nsd^2P + P^2)$  в общем случае с плотными тензорами.
2. Для сетей с поэлементными функциями активации (например, ReLU, sigmoid), где тензоры  $T_{u,v}$  и  $T_{u,v,w}$  диагональны или разреженные со сложностью  $O(d)$ , общая временная сложность снижается до  $O(nsd^2 + nsdP + P^2)$ , поскольку операция  $D_{u \leftarrow v}^\top H_{u,u}^f D_{u \leftarrow w}$  всё равно требует  $O(d^2)$  операций даже при диагональных тензорах.
3. Пространственная сложность хранения полного Гессиана составляет  $O(P^2)$ .
4. Для полносвязного DAG ( $s = O(n)$ ) временная сложность составляет  $O(n^2d^3 + n^2d^2P + P^2)$  в общем случае и  $O(n^2d^2 + n^2dP + P^2)$  для диагональных тензоров.

*Proof.* 1. **Вычисление входного Гессиана  $H_{v,w}^f$ :**

По формуле (1), для каждой пары узлов  $(v, w)$  необходимо:

- Вычислить якобианы  $D_{u \leftarrow v}$  и  $D_{u \leftarrow w}$  для всех  $u \in \text{Ch}(v) \cap \text{Ch}(w)$ , что требует  $O(|\text{Ch}(v) \cap \text{Ch}(w)| \cdot d^2)$  операций.
- Умножить матрицы  $D_{u \leftarrow v}^\top H_{u,u}^f D_{u \leftarrow w}$  для всех  $u \in \text{Ch}(v) \cap \text{Ch}(w)$ , что требует  $O(|\text{Ch}(v) \cap \text{Ch}(w)| \cdot d^3)$  операций.
- Вычислить свертки тензоров смешанных производных  $[T_{u;v,w}]_{i,\bullet,\bullet} \delta_{u,i}$ , что требует  $O(|\text{Ch}(v) \cap \text{Ch}(w)| \cdot d^3)$  операций с учетом разреженности тензора.
- Для случая  $v = w$ , вычислить свертки тензоров чистых производных  $[T_{u;v}]_{i,\bullet,\bullet} \delta_{u,i}$ , что требует  $O(|\text{Ch}(v)| \cdot d^3)$  операций.

Для прореженного DAG с максимальной степенью узла  $s$ , число узлов  $u \in \text{Ch}(v) \cap \text{Ch}(w)$  не превышает  $\min(|\text{Ch}(v)|, |\text{Ch}(w)|) \leq s$ . Поэтому для всех пар узлов  $(v, w)$  общая сложность составляет  $O(n^2 \cdot s \cdot d^3)$ .

С учетом разреженности графа, число пар  $(v, w)$  с непустым пересечением  $\text{Ch}(v) \cap \text{Ch}(w)$  не превышает  $O(n \cdot s)$ , что дает сложность  $O(nsd^3)$ .

## 2. Вычисление параметрического Гессиана $H_{\theta_v, \theta_w}$ :

По формуле (3), для каждой пары узлов  $(v, w)$  необходимо:

- Вычислить якобианы  $D_v$  и  $D_w$ , что требует  $O(d_v \cdot p_v + d_w \cdot p_w)$  операций.
- Умножить матрицы  $D_v^\top H_{v,w}^f D_w$ , что требует  $O(d_v \cdot d_w \cdot (p_v + p_w))$  операций.
- Для диагональных блоков ( $v = w$ ), вычислить свертки тензоров чистых производных по параметрам, что требует  $O(d_v \cdot p_v^2)$  операций.
- Вычислить смешанные производные, что требует  $O(|\text{Pa}(v) \cap \text{Ch}(w)| \cdot d_v \cdot d_u \cdot p_v)$  операций.

Общая сложность для всех пар  $(v, w)$  составляет  $O(n^2 \cdot d^2 \cdot P)$ . Учитывая разреженность графа, число пар с ненулевыми блоками снижается до  $O(n \cdot s)$ , что дает сложность  $O(nsdP)$ . Если  $T_{u;v}$  и  $T_{u;v,w}$  диагональны (поэлементные активации), временная сложность понижается до  $O(nsdP + P^2)$  вместо прежнего  $O(nsd^2P)$ .

## 3. Сборка полного Гессиана:

Сборка требует  $O(P^2)$  операций для размещения всех блоков в общей матрице размера  $P \times P$ .

Суммируя все составляющие, получаем общую временную сложность  $O(nsd^3 + nsd^2P + P^2)$ .

Для полносвязного DAG, где  $s = O(n)$ , сложность возрастает до  $O(n^2d^3 + n^2d^2P + P^2)$ .

Пространственная сложность определяется размером полной матрицы Гессиана  $P \times P$ , т.е.  $O(P^2)$ .  $\square$

**Theorem 8** (Методы снижения вычислительных затрат). Для снижения вычислительной сложности вычисления полного Гессiana 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  $O(n^2d^3 + n^2d^2r + Pr)$ , где  $r \ll P$  — ранг аппроксимации.
3. **Гаусс-Ньютон аппроксимация:** использование только первого члена в формулах (1) и (3) снижает сложность и гарантирует положительную полуопределенность.
4. **Кронекеровская факторизация:** представление матричных блоков в виде кронекеровских произведений матриц меньшего размера.

## 7 Анализ сходимости методов оптимизации

**Theorem 9** (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.

**Corollary 1** (Сходимость в негладком случае с минимальной нормой Clarke-Гессiana). В негладком случае (Случай B), когда вместо обычного Гессiana используется элемент Clarke-Гессiana минимальной нормы  $H_t$ , метод Ньютона не гарантирует положительную определенность матрицы  $H_t$ . Для обеспечения сходимости следует использовать регуляризованную версию:

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

где  $\lambda_t > 0$  — параметр регуляризации, выбираемый так, чтобы матрица  $H_t + \lambda_t I$  была положительно определена.

При правильном выборе последовательности  $\{\lambda_t\}$ , сходящейся к нулю достаточно медленно, и шага  $\alpha_t$ , удовлетворяющего условиям Армихо-Гольдштейна, метод сохраняет супер-линейную скорость сходимости в окрестности регулярных точек минимума.

**Proposition 5** (Требования к Гессиану для методов разложения). *В методах оптимизации, использующих разложения матриц (например, разложение Холецкого для систем линейных уравнений в методе Ньютона), матрица Гессиана должна быть симметричной. В негладком случае (Случай B) следует:*

1. *Перед применением методов разложения Холецкого симметризовать блоки Гессиана:*

$$H_{\theta_v, \theta_w}^{sym} = \frac{1}{2}(H_{\theta_v, \theta_w} + H_{\theta_w, \theta_v}^\top) \quad (4)$$

$$\nabla_\theta^2 \mathcal{L}^{sym} = \frac{1}{2}(\nabla_\theta^2 \mathcal{L} + (\nabla_\theta^2 \mathcal{L})^\top) \quad (5)$$

2. *Для методов сопряженных градиентов и квази-Ньютоновских методов, которые не требуют явного разложения, можно использовать асимметричные блоки при условии обеспечения сходимости.*

Симметризация обеспечивает совместимость с широким спектром методов оптимизации второго порядка, хотя может терять некоторую информацию о кривизне в негладких точках.

**Proposition 6** (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 10** (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, выбираемый так, чтобы матрица  $H_t + \lambda I$  была положительно определена.

## 8 Результаты и обсуждение

### 8.1 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.

### 8.2 Сравнение с существующими подходами

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

1. **Полнота:** В отличие от Гаусс-Ньютона 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.

## 9 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.
- Исправлен базовый случай для листовых узлов.

- Добавлена ссылка на теорему Бьярнасона и уточнены условия.
- Уточнён вопрос симметрии в негладком случае.
- Дополнен анализ сложности с учетом разреженности тензоров.
- Согласована теорема с алгоритмической реализацией.
- Исправлены минорные недочёты.

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.
- Ben-Israel, A. and Greville, T. N. (2003). *Generalized inverses: theory and applications*, volume 15. Springer Science & Business Media.
- 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.
- Magnus, J. R. and Neudecker, H. (2019). *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons.
- 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.
- Federer, H. (2014). *Geometric measure theory*. Springer.
- Hanin, B. and Rolnick, D. (2019). Complexity of linear regions in deep networks. In *International Conference on Machine Learning*, pp. 2596–2604.
- Serra, T., Tjandraatmadja, C., and Ramalingam, S. (2018). Bounding and counting linear regions of deep neural networks. In *International Conference on Machine Learning*, pp. 4558–4566.



## А Реализация с использованием autodiff-фреймворков

**Алгоритм 1:** Вычисление блока входного Гессиана

```

1: function COMPUTEINPUTHESSIAN( $v, w, \{f_u\}, \mathcal{L}, \{\delta_u\}, \{H_{u,u'}^f\}, \text{computed\_blocks}$ )
2:   if  $(v, w)$  in  $\text{computed\_blocks}$  then
3:     return  $H_{v,w}^f$  ▷ Блок уже вычислен
4:   end if
5:    $H_{v,w}^f \leftarrow 0$  ▷ Инициализация блока входного Гессиана
6:   if  $v$  и  $w$  напрямую влияют на  $\mathcal{L}$  then
7:      $H_{v,w}^f \leftarrow \frac{\partial^2 \mathcal{L}}{\partial f_v \partial f_w}$  ▷ Прямая зависимость от обоих узлов
8:   end if
9:   for  $u \in \text{Ch}(v) \cap \text{Ch}(w)$  do
10:     $D_{u \leftarrow v} \leftarrow \text{autodiff.jacobian}(f_u, f_v)$ 
11:     $D_{u \leftarrow w} \leftarrow \text{autodiff.jacobian}(f_u, f_w)$ 
12:    if  $(u, u)$  not in  $\text{computed\_blocks}$  then
13:       $H_{u,u}^f \leftarrow \text{ComputeInputHessian}(u, u, \{f_u\}, \mathcal{L}, \{\delta_u\}, \{H_{u,u'}^f\}, \text{computed\_blocks})$ 
14:      ▷ Вычислить  $H_{u,u}^f$  ровно один раз
15:      end if
16:       $H_{v,w}^f \leftarrow H_{v,w}^f + D_{u \leftarrow v}^\top H_{u,u}^f D_{u \leftarrow w}$ 
17:      for  $i \in 1..d_u$  do
18:         $T_{u;v,w} \leftarrow \text{ComputeMixedHessian}(f_u, f_v, f_w)$ 
19:         $T_{u;v,w}^{sym} \leftarrow \text{SymmetrizeHessian}(T_{u;v,w}, T_{u;w,v})$  ▷ Симметризация для
20:        негладкого случая
21:         $H_{v,w}^f \leftarrow H_{v,w}^f + T_{u;v,w}^{sym} \cdot \delta_{u,i}$ 
22:      end for
23:      if  $v = w$  then
24:        for  $i \in 1..d_u$  do
25:           $T_{u;v} \leftarrow \text{autodiff.hessian}(f_u, f_v)$ 
26:           $H_{v,v}^f \leftarrow H_{v,v}^f + T_{u;v} \cdot \delta_{u,i}$ 
27:        end for
28:      end if
29:    end for
30:    ▷ Обработка одностороннего пути от  $v$  к  $w$ 
31:    for  $u \in \text{Ch}(v) \setminus \text{Ch}(w)$  do
32:      if  $(u, w)$  not in  $\text{computed\_blocks}$  then
33:         $H_{u,w}^f \leftarrow \text{ComputeInputHessian}(u, w, \{f_u\}, \mathcal{L}, \{\delta_u\}, \{H_{u,u'}^f\}, \text{computed\_blocks})$ 
34:      end if
35:       $D_{u \leftarrow v} \leftarrow \text{autodiff.jacobian}(f_u, f_v)$ 
36:       $H_{v,w}^f \leftarrow H_{v,w}^f + D_{u \leftarrow v}^\top H_{u,w}^f$ 
37:    end for
38:    ▷ Обработка одностороннего пути от  $w$  к  $v$ 
39:    for  $u \in \text{Ch}(w) \setminus \text{Ch}(v)$  do
40:      if  $(v, u)$  not in  $\text{computed\_blocks}$  then
41:         $H_{v,u}^f \leftarrow \text{ComputeInputHessian}(v, u, \{f_u\}, \mathcal{L}, \{\delta_u\}, \{H_{u,u'}^f\}, \text{computed\_blocks})$ 
42:      end if
43:       $D_{u \leftarrow w} \leftarrow \text{autodiff.jacobian}(f_u, f_w)$ 
44:       $H_{v,w}^f \leftarrow H_{v,w}^f + H_{v,u}^f D_{u \leftarrow w}$ 
45:    end for
46:     $\text{computed\_blocks} \leftarrow \text{computed\_blocks} \cup \{(v, w)\}$  ▷ Отметить как вычисленный блок
47:  return  $H_{v,w}^f$ 
48: end function

```

**Алгоритм 2:** Вычисление блока параметрического Гессиана

```

1: function COMPUTEPARAMETERHESSIAN( $v, w, \{f_u\}, \{\theta_u\}, \{H_{u,w}^f\}, \{\delta_u\}$ )
2:    $D_v \leftarrow \text{autodiff.jacobian}(f_v, \theta_v)$ 
3:    $D_w \leftarrow \text{autodiff.jacobian}(f_w, \theta_w)$ 
4:    $H_{\theta_v, \theta_w} \leftarrow D_v^\top H_{v,w}^f D_w$ 
5:   if  $v = w$  then
6:     for  $i \in 1..d_v$  do
7:        $T_v^\theta \leftarrow \text{autodiff.hessian}(f_{v,i}, \theta_v)$ 
8:        $H_{\theta_v, \theta_v} \leftarrow H_{\theta_v, \theta_v} + T_v^\theta \cdot \delta_{v,i}$ 
9:     end for
10:  end if
11:  for  $u \in \text{Pa}(v) \cap \text{Ch}(w)$  do
12:    for  $i \in 1..d_v$  do
13:      for  $j \in 1..d_u$  do
14:        for  $\alpha \in 1..p_v$  do
15:           $T_{v;u,\theta} \leftarrow \text{ComputeMixedDerivative}(f_{v,i}, f_{u,j}, \theta_{v,\alpha})$ 
16:           $D_{w \leftarrow u} \leftarrow \text{autodiff.jacobian}(f_w, f_u)$ 
17:           $H_{\theta_v, \theta_w} \leftarrow H_{\theta_v, \theta_w} + T_{v;u,\theta} \cdot D_{w \leftarrow u} \cdot \delta_{v,i}$ 
18:        end for
19:      end for
20:    end for
21:  end for
22:  return  $H_{\theta_v, \theta_w}$ 
23: end function

```

**Алгоритм 3:** Полное вычисление Гессиана

```

1: function FULLHESSIANCOMPUTATION( $G, \{f_v\}, \{\theta_v\}, \mathcal{L}$ )
2:    $\delta_{out} \leftarrow \text{autodiff.gradient}(\mathcal{L}, f_{out})$ 
3:    $H_{out,out}^f \leftarrow \text{autodiff.hessian}(\mathcal{L}, f_{out})$ 
4:    $\text{topo\_order} \leftarrow \text{TopologicalSort}(G).\text{reverse}()$ 
5:   Initialize  $\{\delta_v\}, \{H_{v,w}^f\}$  as zero matrices
6:    $\text{computed\_blocks} \leftarrow \emptyset$  ▷ Отслеживание вычисленных блоков
7:    $\text{input\_dep\_nodes} \leftarrow \text{FindNodesDirectlyInfluencingLoss}(\mathcal{L})$ 
8:   for  $v \in \text{input\_dep\_nodes}$  do
9:      $H_{v,v}^f \leftarrow \text{autodiff.hessian}(\mathcal{L}, f_v)$ 
10:     $\text{computed\_blocks} \leftarrow \text{computed\_blocks} \cup \{(v, v)\}$  ▷ Отметить как
11:    end for
12:    for  $v, w \in \text{input\_dep\_nodes}, v \neq w$  do
13:       $H_{v,w}^f \leftarrow \text{autodiff.mixed\_hessian}(\mathcal{L}, f_v, f_w)$ 
14:       $\text{computed\_blocks} \leftarrow \text{computed\_blocks} \cup \{(v, w)\}$  ▷ Отметить как
15:    end for
16:    for  $v \in \text{topo\_order}$  do
17:       $\text{BackpropagateGradients}(v)$ 
18:      for  $w \in V$  do
19:        if  $\text{Ch}(v) \cap \text{Ch}(w) \neq \emptyset$  OR  $(v, w)$  напрямую влияют на  $\mathcal{L}$  then
20:           $H_{v,w}^f \leftarrow \text{ComputeInputHessian}(v, w, \{f_u\}, \mathcal{L}, \{\delta_u\}, \{H_{u,u'}^f\}, \text{computed\_blocks})$ 
21:        end if
22:      end for
23:    end for
24:    Initialize full Hessian matrix  $H$  of size  $P \times P$ 
25:    for  $v, w \in V$  do
26:      if  $\exists u : v \rightarrow^* u$  и  $w \rightarrow^* u$  OR  $(v, w)$  напрямую влияют на  $\mathcal{L}$  then
27:         $H_{\theta_v, \theta_w} \leftarrow \text{ComputeParameterHessian}(v, w, \{f_u\}, \{\theta_u\}, \{H_{u,u'}^f\}, \{\delta_u\})$ 
28:        Update corresponding blocks in  $H$ 
29:      end if
30:    end for
31:    return  $H$ 
32: end function

```