Полное строгое определение локального и глобального Hessian'а второго порядка в произвольной нейронной архитектуре

Автор

May 30, 2025

Abstract

В этом отчёте приводится исчерпывающее математически строгое определение Hessian'а второго порядка для нейронных сетей произвольной архитектуры, заданной направленным ациклическим графом. Учтены все чистые и смешанные вторые производные по входам и параметрам, кросс-блоки между разными параметрами, «шаринг» параметров, а также негладкие активации через Clarke-Гессиан. Для тривиального графа из единственного узла без потомков и предков наши формулы сводятся к стандартному Hessian'у $\nabla_\theta^2 \mathcal{L}(\theta) \in \mathbb{R}^{p \times p}$. Каждый шаг снабжён пояснениями о роли того или иного члена.

Contents

1 Введение

Hessian второго порядка $\nabla^2\mathcal{L}$ играет ключевую роль в анализе кривизны функционала потерь и в разработке методов оптимизации (Newton, trust-region и др.). Типичная Gauss–Newton-аппроксимация учитывает лишь часть всех вторых производных. Здесь мы даём *полный* формализм, закрывающий следующие пробелы:

- чистые и смешанные вторые производные по *входам* каждого узла,
- чистые вторые производные по *параметрам*,
- кросс-блоки $\partial^2/\partial\theta_v\,\partial\theta_w$,
- смешанные вход–параметрические слагаемые,
- учёт «шаринга» одного вектора параметров в нескольких узлах,
- негладкие активации через выбор Clarke-Гессиана.

Особый случай: если граф состоит из одного узла без потомков и предков, все определения сводятся к стандартному Hessian'у $\nabla^2_\theta\mathcal{L}(\theta) \in \mathbb{R}^{p\times p}$.

2 Модель и разделение случаев

Граф: сеть задаётся DAG $G = (V, E)$.

Узел $v \in V$: входы $f_{\mathrm{Pa}(v)} \in \prod_{u\in\mathrm{Pa}(v)} \mathbb{R}^{d_u}$, параметры $\theta_v \in \mathbb{R}^{p_v}$, отображение

$$f_v \;=\; g_v\big(f_{\mathrm{Pa}(v)},\, \theta_v\big) \;\in\; \mathbb{R}^{d_v}.$$

Потеря: $\mathcal{L} : \mathbb{R}^{d_{out}} \to \mathbb{R}$ на выходном узле $out \in V$.

Разделяем два режима:

- **Случай A (гладкий).** Все $g_v \in C^2$ по входам и параметрам.
- **Случай B (негладкий).** Допускаются ReLU, max-pool и пр.; вводится Clarke-Гессиан $\partial_C^2 f_v$.

Вычисление всех блоков Hessian ведётся в *обратном топологическом порядке* по G, начиная с out.

3 Первый порядок

Нам прежде всего нужны:

$$
\begin{aligned}
\delta_v &:= \nabla_{f_v} \mathcal{L} && \in \mathbb{R}^{d_v}, \\
\delta_{v,i} &:= [\delta_v]_i, && i = 1, \dots, d_v, \\
D_{u \leftarrow v} &:= \frac{\partial f_u}{\partial f_v} && \in \mathbb{R}^{d_u \times d_v}, \\
D_v &:= \frac{\partial f_v}{\partial \theta_v} && \in \mathbb{R}^{d_v \times p_v}.
\end{aligned}
$$

Комментарий: градиенты δ_v и якобианы $D_{u \leftarrow v}$, D_v — основа цепного правила первого порядка.

4 Тензоры чистых и смешанных вторых производных

Чтобы учесть *все* вторые производные функций узлов, вводим:

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

Комментарий: индексы i суммируются с весом $\delta_{u,i}$ — это даёт чистые «тензорные» вклады в Hessian.

5 Полный входной Hessian

Вводим блочную матрицу $\{H^f_{v,w}\}_{v,w\in V}$, где каждый блок $H^f_{v,w} \in \mathbb{R}^{d_v\times d_w}$ даётся формулой

$$\boxed{\begin{aligned} H^f_{v,w} = &\sum_{u\in\mathrm{Ch}(v)\cap\mathrm{Ch}(w)} D^\top_{u\leftarrow v}\, H^f_{u,u}\, D_{u\leftarrow w} \quad \text{(Gauss–Newton)} \\ &+ \sum_{u\in\mathrm{Ch}(v)\cap\mathrm{Ch}(w)} \sum_{i=1}^{d_u} [T_{u;\,v,w}]_{i,\bullet,\bullet}\, \delta_{u,i} \quad \text{(смешанные входы)} \\ &+ \delta_{v=w} \sum_{u\in\mathrm{Ch}(v)} \sum_{i=1}^{d_u} [T_{u;v}]_{i,\bullet,\bullet}\, \delta_{u,i} \quad \text{(чистые по одному входу)} \end{aligned}} \tag{1}$$

с условиями-основами

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

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

Симметрия: $H^f_{v,w} = (H^f_{w,v})^\top$.

Комментарий:

- Первая строка — классический Gauss–Newton.
- Вторая — учёт смешанных $\partial^2 f_u/(\partial f_v\, \partial f_w)$.
- Третья (только при $v = w$) — чистые $\partial^2 f_u/\partial f_v^2$.
- Off-diag-base-case сразу обнуляет блоки без связи.

6 Полный параметрический Hessian

Разбиваем $\nabla^2_\theta\mathcal{L}$ на блочные элементы $\{H_{\theta_v,\theta_w}\}$, $H_{\theta_v,\theta_w} \in \mathbb{R}^{p_v\times p_w}$:

$$\boxed{\begin{aligned} H_{\theta_v,\theta_w} = &\, D^\top_v\, H^f_{v,w}\, D_w \quad \text{(Gauss–Newton)} \\ &+ \delta_{v=w} \sum_{i=1}^{d_v} [T^\theta_v]_{i,\bullet,\bullet}\, \delta_{v,i} \quad \text{(чистые по параметрам)} \\ &+ \sum_{u\in\mathrm{Pa}(v)\cap\mathrm{Ch}(w)} \sum_{i=1}^{d_v}\sum_{j=1}^{d_w}\sum_{k=1}^{p_v} [T_{v;\,u,\theta}]_{i,j,k}\, (D_{w\leftarrow v})_{j,i}\, \delta_{v,i} \quad \text{(смешанные вход–параметр } v\to w) \\ &+ \sum_{u\in\mathrm{Pa}(w)\cap\mathrm{Ch}(v)} \sum_{i=1}^{d_w}\sum_{j=1}^{d_v}\sum_{\ell=1}^{p_w} [T_{w;\,u,\theta}]_{i,j,\ell}\, (D_{v\leftarrow w})_{j,i}\, \delta_{w,i} \quad \text{(смешанные вход–параметр } w\to v) \end{aligned}} \tag{2}$$

Офф-диагональный base-case:

$$H_{\theta_v,\theta_w} = 0 \quad \text{если нет путей } v \to u \text{ и } w \to u.$$

Симметрия: $H_{\theta_v,\theta_w} = (H_{\theta_w,\theta_v})^\top$.
Комментарий:

- Первая строка — Gauss–Newton-часть.
- Вторая — чистые вторые по θ_v, только при $v = w$.
- Третья и четвёртая — смешанные вход–параметр для диагональных и офф-диагональных блоков.

7 Sharing параметров

Если один вектор $\theta \in \mathbb{R}^p$ разделяют узлы $\{v_k\}_{k=1}^K$, то итоговый Hessian

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

8 Clarke-Гессиан (негладкий случай)

В случае B вместо каждого блока $H^f_{v,v}$ и соответствующих H_{θ_v,θ_w} получается множество $\partial^2_C f_v$. Фиксируем *measurable selection*:

$$H^f_{v,w} = \arg\min_{M \in \partial^2_C f_v} \|M\|_F, \quad H_{\theta_v,\theta_w} = \arg\min_{M \in \partial^2_{\theta_v,\theta_w} \mathcal{L}} \|M\|_F.$$

9 Практические замечания

- В **гладком** случае имеет смысл проверять положительную полуопределённость Gauss–Newton-части $D_v^\top H^f_{v,v} D_v$ перед добавлением остальных слагаемых.
- При большом графе эффективнее осуществлять обратный топологический обход с запоминанием промежуточных блоков.

10 Заключение

В отчёте представлен исчерпывающий формализм Hessian'а второго порядка:

- Полная блочная структура по $\{f_v\}$ и $\{\theta_v\}$.
- Учет всех чистых и смешанных вторых производных.
- Явные base-case для off-diag блоков.
- Отдельные правила для гладкого и негладкого случаев.
- Упоминание тривиального предельного случая одного узла.
- Практические рекомендации по проверке PSD Gauss–Newton части.

Теперь этот формализм готов к любым теоретическим выкладкам и практической реализации.