Обобщенная жадная градиентная оптимизация гиперпараметров

К. Д. Яковлев

iakovlev.kd@phystech.edu

Москва, Московский физико-технический институт

Научный руководитель: к.ф.-м.н. Бахтеев Олег Юрьевич

2024

Цель исследования

Цель

Предложить градиентный метод оптимизации гиперпараметров с линейейной по количеству параметров и гиперпараметров сложностью итерации и затратами памяти.

Проблема

Существующие методы не гарантируют выполнения следующих условий одновременно: 1) отсутствие требований на сходимость внутренней процедуры оптимизации к единственному решению, 2) отсутствие смещения из-за короткого горизонта, 3) линейная сложность итерации и затраты памяти.

Метод решения

Предлагаемый метод основан на агрегации жадных гиперградиентов без дополнительных вычислительных затрат.

Агрегация жадных гиперградиентов

Пусть задано $\gamma \in (0, 1)$. Тогда аппроксимация гиперградиента запишется как:

$$\hat{d}_{\boldsymbol{\alpha}} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}; \gamma) = \nabla_{\boldsymbol{\alpha}} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}) + \sum_{t=1}^{T} \gamma^{T-t} \nabla_{\mathbf{w}_t} \mathcal{L}_{\text{val}}(\mathbf{w}_t, \boldsymbol{\alpha}) \mathbf{B}_t.$$

	IFT	RMD	DrMAD	$T1 - T2$	Ours
Онлайн оптимизация	×	✓	×	✓	✓
Длинный горизонт	✓	✓	✓	×	✓
Линейная сложность	✓	×	✓	✓	✓

Постановка задачи оптимизации гиперпараметров

- Пусть задан вектор параметров модели $\mathbf{w} \in \mathbb{R}^p$ и вектор гиперпараметров $\boldsymbol{\alpha} \in \mathbb{R}^h$. Задача оптимизации:

$$\begin{aligned} \boldsymbol{\alpha}^* &= \arg\min_{\boldsymbol{\alpha}} \mathcal{L}_2(\mathbf{w}^*, \boldsymbol{\alpha}), \\ \text{s.t.} \quad \mathbf{w}^* &= \arg\min_{\mathbf{w}} \mathcal{L}_1(\mathbf{w}, \boldsymbol{\alpha}). \end{aligned}$$

- Пусть внутренняя задача решается с помощью оптимизатора $\boldsymbol{\Phi}(.,.)$:

$$\mathbf{w}_{t+1}(\boldsymbol{\alpha}) = \boldsymbol{\Phi}(\mathbf{w}_t, \boldsymbol{\alpha}), \quad t = \overline{1, T}; \quad \boldsymbol{\Phi}(\mathbf{w}_t, \boldsymbol{\alpha}) = \mathbf{w}_t - \eta \nabla_{\mathbf{w}_t} \mathcal{L}_{\text{train}}(\mathbf{w}_t, \boldsymbol{\alpha})$$

- Гиперградиент запишется как:

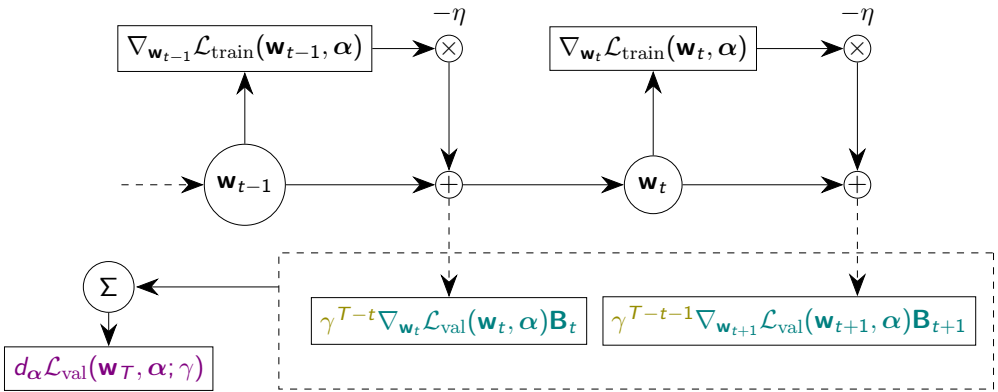
$$d_{\boldsymbol{\alpha}} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}) = \nabla_{\boldsymbol{\alpha}} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}) + \sum_{t=1}^{T} \nabla_{\mathbf{w}_T} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}) \left(\prod_{k=t+1}^{T} \mathbf{A}_k \right) \mathbf{B}_t,$$

$$\mathbf{A}_k = \frac{\partial \boldsymbol{\Phi}(\mathbf{w}_{k-1}, \boldsymbol{\alpha})}{\partial \mathbf{w}_{k-1}}, \quad \mathbf{B}_t = \frac{\partial \boldsymbol{\Phi}(\mathbf{w}_{t-1}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}.$$

Аппроксимация гиперградиента

Пусть задано $\gamma \in (0, 1)$. Тогда аппроксимация гиперградиента запишется как:

$$\hat{d}_{\boldsymbol{\alpha}}\mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}; \gamma) = \nabla_{\boldsymbol{\alpha}}\mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}) + \sum_{t=1}^{T} \gamma^{T-t} \nabla_{\mathbf{w}_t}\mathcal{L}_{\text{val}}(\mathbf{w}_t, \boldsymbol{\alpha})\mathbf{B}_t.$$



Обобщение метода $T1 - T2$

Определение

Аппроксимация гиперградиента, определяемая методом $T1 - T2$ запишется как:

$$\hat{d}_{\boldsymbol{\alpha}}^{T1-T2}\mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}) = \nabla_{\boldsymbol{\alpha}}\mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}) + \nabla_{\mathbf{w}_T}\mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha})\mathbf{B}_T.$$

Теорема (Яковлев, 2024)

Пусть $\hat{d}_{\boldsymbol{\alpha}}(\mathbf{w}_T, \boldsymbol{\alpha}; \gamma)$ – предложенная аппроксимация гиперградиента. Тогда имеет место следующий предел:

$$\lim_{\gamma \to 0^+} \hat{d}_{\boldsymbol{\alpha}}(\mathbf{w}_T, \boldsymbol{\alpha}; \gamma) = \nabla_{\boldsymbol{\alpha}}\mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}) + \nabla_{\mathbf{w}_T}\mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha})\mathbf{B}_T.$$

Таким образом, предложенный подход является обобщением $T1 - T2$.

Достаточное условие спуска

Предположения

1. $\mathcal{L}_{\text{val}}(., \boldsymbol{\alpha})$ является L-гладкой and μ-сильно выпуклой для любого $\boldsymbol{\alpha}$.
2. $\frac{\partial \boldsymbol{\Phi}(., \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}$ является C_B-Липшицевой для любого $\boldsymbol{\alpha}$.
3. $\| \frac{\partial \boldsymbol{\Phi}(\mathbf{w}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \| \leq B$ для любой пары $(\mathbf{w}, \boldsymbol{\alpha})$ для некоторого $B \geq 0$.
4. $\mathbf{w}$ принадлежит некоторому выпуклому множеству с диаметром $D < \infty$.
5. $\boldsymbol{\Phi}(\mathbf{w}, \boldsymbol{\alpha}) = \mathbf{w} - \eta \nabla_{\mathbf{w}} \mathcal{L}_{\text{train}}(\mathbf{w}, \boldsymbol{\alpha})$ для некоторого $\eta \geq 0$.
6. $\nabla^2_{\mathbf{w}} \mathcal{L}_{\text{train}}(., \boldsymbol{\alpha}) = \mathbf{I}$ для любого $\boldsymbol{\alpha}$, а также $\nabla_{\boldsymbol{\alpha}} \mathcal{L}_{\text{val}}(\mathbf{w}, \boldsymbol{\alpha}) = \mathbf{0}$ для любого $\mathbf{w}$.
7. $\mathbf{B}_t \mathbf{B}_t^\top \succeq \kappa \mathbf{I}$ для некоторого $\kappa > 0$.
8. Определим $\mathbf{w}_\infty := \arg\min_{\mathbf{w}} \mathcal{L}_{\text{train}}(\mathbf{w}, \boldsymbol{\alpha})$, $\mathbf{w}_2^* := \arg\min_{\mathbf{w}} \mathcal{L}_{\text{val}}(\mathbf{w}, \boldsymbol{\alpha})$. Пусть $\|\mathbf{w}_\infty - \mathbf{w}_2^*\| \geq 2De^{-\mu\eta T} + \delta$, для некоторого $\delta > 0$.

Теорема (Яковлев, 2024)

Пусть $\gamma = 1 - \eta \in (0, 1)$. Пусть также выполнены предположения (1-8), тогда найдется достаточно большое T и универсальная константа $c > 0$ такая, что:

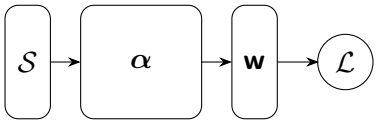
$$d_{\boldsymbol{\alpha}} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}) \hat{d}_{\boldsymbol{\alpha}} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}; \gamma)^\top \geq c \| d_{\boldsymbol{\alpha}} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}) \|_2^2.$$

Постановка вычислительного эксперимента

- Цель – сравнение качества предложенного подхода с существующими методами подсчета гиперградиента.
- Эксперимент проводится на задаче мета-обучения.

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} \mathbb{E}_{\mathcal{T}} \mathbb{E}_{\mathcal{S}|\mathcal{T}} \mathcal{L}_{\text{val}}(\mathbf{w}^*, \boldsymbol{\alpha}; \mathcal{S}),$$

$$\text{s.t.} \quad \mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{L}_{\text{train}}(\mathbf{w}, \boldsymbol{\alpha}; \mathcal{S}).$$



- В сравнении участвуют следующие базовые методы:

(FO) : $\hat{d}^{\text{FO}}_{\boldsymbol{\alpha}} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}) = \nabla_{\boldsymbol{\alpha}} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}),$

(IFT) : $\hat{d}^{\text{IFT}}_{\boldsymbol{\alpha}} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}) = \nabla_{\boldsymbol{\alpha}} \mathcal{L}_{\text{val}} - \nabla_{\mathbf{w}} \mathcal{L}_{\text{val}} \left(\sum_{j \le i} \left[\mathbf{I} - \nabla^2_{\mathbf{w},\mathbf{w}} \mathcal{L}_{\text{train}} \right]^j \right) \nabla^2_{\mathbf{w},\boldsymbol{\alpha}} \mathcal{L}_{\text{train}} \Big|_{(\mathbf{w}_T, \boldsymbol{\alpha})},$

(T1-T2) : $\hat{d}^{T1-T2}_{\boldsymbol{\alpha}} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}) = \nabla_{\boldsymbol{\alpha}} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}) + \nabla_{\mathbf{w}_T} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}) \mathbf{B}_T.$

Результаты вычислительного эксперимента

- Рассматриваются задачи классификации на n классов с k примерами на каждый класс (n-way, k-shot).
- Приводится точность предсказаний на мета-контроле, а также вычислительная сложность итерации подсчета гиперградиента.

Method	#JVPs	3-way, 10-shot	4-way, 10-shot	5-way, 10-shot
FO	0	43.48 ± 0.69	34.15 ± 0.53	28.59 ± 0.47
$T1 - T2$	1	42.96 ± 0.79	33.95 ± 0.64	27.59 ± 0.46
IFT	11	40.14 ± 0.73	33.23 ± 0.41	27.20 ± 0.52
Ours ($\gamma = 0.99$)	10	$\mathbf{46.10 \pm 0.82}$	$\mathbf{36.94 \pm 1.07}$	$\mathbf{29.79 \pm 0.62}$

Из таблицы видно, что предложенный метод превосходит существующие методы градиентной оптимизации гиперпараметров в терминах точности предсказаний на мета-контроле, имея сопоставимые вычислительные затраты.

Выносится на защиту

- Рассмотрена задача оптимизации гиперпараметров.
- Предложен метод оптимизации гиперпараметров, удовлетворяющий одновременно трем условиям:
  - онлайн оптимизация
  - отсутствие смещения из-за короткого горизонта
  - линейная сложность итерации и затраты памяти.
- Продемонстрирована работоспособность предлагаемого решения.
- Проведен теоретический анализ предложенного метода.

Список публикаций

- **(core-A*)** Yakovlev K. et al. GEC-DePenD: Non-Autoregressive Grammatical Error Correction with Decoupled Permutation and Decoding //Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2023. – С. 1546-1558.
- **(core-A*)** Yakovlev K. et al. Sinkhorn Transformations for Single-Query Postprocessing in Text-Video Retrieval //Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. – 2023. – С. 2394-2398.
- Yakovlev K. D. et al. Neural Architecture Search with Structure Complexity Control //International Conference on Analysis of Images, Social Networks and Texts. – Cham : Springer International Publishing, 2021. – С. 207-219.

Выступления на конференциях

- Яковлев К.Д. Обобщенная жадная градиентная оптимизация гиперпараметров. //Труды 66-й Всероссийской научной конференции МФТИ. 2024.
- Яковлев К.Д. Поиск согласованных нейросетевых моделей в задаче мультидоменного обучения. //Труды 65-й Всероссийской научной конференции МФТИ в честь 115-летия Л.Д. Ландау. - 2023.
- Яковлев К.Д., Гребенькова О.С., Бахтеев О.Ю., Стрижов В.В. Выбор архитектуры модели с контролем сложности // Труды 64-й Всероссийской научной конференции МФТИ. - 2021.