

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 линейной 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:

$$\hat{d}_{\alpha} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \alpha; \gamma) = \frac{\partial}{\partial \alpha} \mathcal{L}_2(\mathbf{w}_T(\alpha), \alpha) + \sum_{t=1}^T \mathbf{B}_t \frac{\partial \mathcal{L}_2(\mathbf{w}_t, \alpha)}{\partial \mathbf{w}_t} \gamma^{T-t}.$$

	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 $\alpha \in \mathbb{R}^h$. Задача оптимизации:

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

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

$$\mathbf{w}_{t+1}(\alpha) = \Phi(\mathbf{w}_t, \alpha), \quad t = \overline{1, T}.$$

- ▶ Гиперградиент запишется как:

$$\begin{aligned} d_{\alpha} \mathcal{L}_2(\mathbf{w}_T(\alpha), \alpha) &= \frac{\partial}{\partial \alpha} \mathcal{L}_2(\mathbf{w}_T(\alpha), \alpha) + \sum_{t=1}^T \mathbf{B}_t \mathbf{A}_{t+1} \dots \mathbf{A}_T \frac{\partial \mathcal{L}_2(\mathbf{w}_T(\alpha), \alpha)}{\partial \mathbf{w}}, \\ \mathbf{B}_t &= \frac{\partial \Phi(\mathbf{w}_{t-1}, \alpha)}{\partial \alpha}, \quad \mathbf{A}_t = \frac{\partial \Phi(\mathbf{w}_{t-1}, \alpha)}{\partial \mathbf{w}_{t-1}}. \end{aligned}$$

Аппроксимация гиперградиента

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

$$\hat{d}_{\alpha} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \alpha; \gamma) = \frac{\partial}{\partial \alpha} \mathcal{L}_2(\mathbf{w}_T(\alpha), \alpha) + \sum_{t=1}^T \mathbf{B}_t \frac{\partial \mathcal{L}_2(\mathbf{w}_t, \alpha)}{\partial \mathbf{w}_t} \gamma^{T-t}.$$

Теорема (Яковлев, 2024)

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

$$\lim_{\gamma \rightarrow 0^+} \hat{d}_{\alpha}(\mathbf{w}_T, \alpha; \gamma) = \nabla_{\alpha} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \alpha) + \nabla_{\mathbf{w}_T} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \alpha) \mathbf{B}_T. \quad (1)$$

Таким образом, предложенный подход является обобщением T1 – T2.

Достаточное условие спуска

Предположения

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

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

Постановка вычислительного эксперимента

- ▶ Цель – сравнение качества предложенного подхода с существующими методами подсчета гиперградиента.
- ▶ Эксперимент проводится на задаче мета-обучения. Приводится точность предсказаний на мета-контроле, а также вычислительная сложность итерации подсчета гиперградиента.
- ▶ Сравниваются следующие методы: FO, IFT, $T1 - T2$.

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	46.10 ± 0.82	36.94 ± 1.07	29.79 ± 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.

- ▶ 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.