Быстрая оптимизация мультизадачных моделей

Филатов Андрей

Московский физико-технический институт
Кафедра интеллектуальных систем

Апрель, 2021

Постановка задачи

Задача оптимизации

Рассмотрим T функции (задач) $\mathscr{L}^t$. Цель заключается в их одновременной оптимизации, то есть:

$$\min_{\substack{\boldsymbol{\theta}^{sh},\\ \boldsymbol{\theta}^1,\ldots,\boldsymbol{\theta}^T}} \mathscr{L}\left(\boldsymbol{\theta}^{sh},\boldsymbol{\theta}^1,\ldots,\boldsymbol{\theta}^T\right) = \min_{\substack{\boldsymbol{\theta}^{sh},\\ \boldsymbol{\theta}^1,\ldots,\boldsymbol{\theta}^T}} \left(\mathscr{L}^1\left(\boldsymbol{\theta}^{sh},\boldsymbol{\theta}^1\right),\ldots,\mathscr{L}^T\left(\boldsymbol{\theta}^{sh},\boldsymbol{\theta}^T\right)\right)^\top \quad (1)$$

где $\boldsymbol{\theta}^{sh}$ - параметры общие параметры, а $\boldsymbol{\theta}^t$ отдельные параметры для каждой задачи

Обозначения

Парето доминирование

Считаем, что точка $\boldsymbol{\theta}_1$ доминируется точкой $\boldsymbol{\theta}_2$ если:

$$\forall i = 1 \dots T \ \mathscr{L}^i(\boldsymbol{\theta}_2) \leq \mathscr{L}^i(\boldsymbol{\theta}_1)$$

$$\exists j : \mathscr{L}^j(\theta_2) < \mathscr{L}^j(\theta_1)$$

Парето оптимальность

Точка $\boldsymbol{\theta} = [\boldsymbol{\theta}^{sh}, \boldsymbol{\theta}^t]$ для задачи (1) Парето оптимальная тогда и только тогда, когда она допустимая и не существует другой точки, которая Парето доминирует $\boldsymbol{\theta}$:

- Для любого допустимой $\hat{\boldsymbol{\theta}}$: $\mathscr{L}^i(\boldsymbol{\theta}) \leq \mathscr{L}^i(\hat{\boldsymbol{\theta}})$ для всех $i \in \{1, \dots, T\}$
- $\mathscr{L}\left(\boldsymbol{\theta}^{sh}, \boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^T\right) \neq \mathscr{L}\left(\hat{\boldsymbol{\theta}}^{sh}, \hat{\boldsymbol{\theta}}^1, \dots, \hat{\boldsymbol{\theta}}^T\right)$

Обозначения

Парето станционарость

Точка $\boldsymbol{\theta}^* = [\boldsymbol{\theta}^{sh}, \boldsymbol{\theta}^t]$ Парето станционарная, если существует набор коэффициентов $\alpha^1, \ldots, \alpha^T \geq 0$, таких что $\sum_{t=1}^T \alpha^t = 1$ и $\sum_{t=1}^T \alpha^t \nabla_{\boldsymbol{\theta}^{*h}} \hat{\mathscr{L}}^t \left(\boldsymbol{\theta}^{sh}, \boldsymbol{\theta}^t \right) = 0$ и для всех задач $t, \nabla_{\boldsymbol{\theta}^t} \hat{\mathscr{L}}^t \left(\boldsymbol{\theta}^{sh}, \boldsymbol{\theta}^t \right) = 0$

Методы решения

- Взвешивание: сводим мультикритериальную оптимизацию к одномерной оптимизации следующим образом:

$$L(x) = \sum_{t=1}^{T} w_i \mathscr{L}^t(x)$$

- Методы нулевого порядка: эволюционные алгоритмы и мультикритериальная байесовская оптимизация
- Методы более старшего порядка: градиентный спуск, метод Ньютона

Методы первого порядка

- MGDA

$$\min_{\alpha^1,\dots,\alpha^T} \left\{ \left\| \sum_{t=1}^{T} \alpha^t \nabla_{\boldsymbol{x}} \hat{\mathscr{L}}^t(\boldsymbol{x}) \right\|_2^2 \mid \sum_{t=1}^{T} \alpha^t = 1, \alpha^t \geq 0 \quad \forall t \right\} \tag{2}$$

$$d^* = \sum_{t=1}^{T} \tilde{\alpha}^t \nabla_x \hat{\mathscr{L}}^t$$

- Минмакс подход

$$\min_u \max_t \left(\frac{\partial \mathscr{L}}{\partial \boldsymbol{x}} d\right)_t + g^t(d) \tag{3}$$

Рассматривая $g^t(d) = \|d\|^2$ получаем двойственную к MGDA.
Рассматривая $g^t(d) = \frac{1}{2} d^T H^t d$ получаем метод Ньютона (H^t гессиан $\mathscr{L}^t$).

Модификации

- PCGrad (Tianhe Yu et al. 2020)
  Предложил преобразование искомых градиентов, которое позволяет получить направление минимизирующие всех функции.
- EDM (Katrutsa et al. 2020)
  Добавил нормирование градиентов, чтобы сделать метод более устойчивым в случае, когда градиенты разных задач сильно различаются по норме.

Линейный поиск

Пусть получено d — направление убывания всех функций.
Необходимо найти шаг η, чтобы:

$$\forall t = 1 \dots T : \ \mathscr{L}^t(x - \eta d) < \mathscr{L}^t(x) \tag{4}$$

Теорема

Если это условие (4) будет выполнено на каждой итерации, то каждый частичный предел последовательности, созданной градиентным спуском, — Парето стационарная точка.

Линейный поиск

Armijo rule

На каждом шаге градиентного спуска нам найти шаг η, чтобы выполнялось следующее Armijo rule:

$$\mathscr{L}^t(\boldsymbol{\theta}^{sh} - \eta d_{sh}, \boldsymbol{\theta}^t - \eta \nabla_{\boldsymbol{\theta}^t} \mathscr{L}^t) \leq \mathscr{L}^t - \eta\beta \left\| \frac{\partial \mathscr{L}^t}{\partial \boldsymbol{\theta}^t} \right\|^2 - \eta\beta \left(\frac{\partial \mathscr{L}^t}{\partial \boldsymbol{\theta}^{sh}} \right)^\top d_{sh} \qquad (5)$$

Мотивация

- Получив направление для градиентного спуска нужно выбрать шаг;
- Теоретически обосновано лишь использование линейного поиска для нахождение шага;
- Линейный поиск неэффективен на практике из высокой вычислительной стоимость из-за дополнительных forward и backward проходов;

Цель

Предложить вычислительно эффективный метод, сохранив теоретические результаты.

Линейный поиск

Новое правило Армихо

На каждом шаге градиентного спуска нам найти шаг η, чтобы выполнялось следующее правило Армихо:

$$\mathscr{L}^t(\boldsymbol{Z} - \eta d_z, \boldsymbol{\theta}^t - \eta \nabla_{\boldsymbol{\theta}^t} \mathscr{L}^t) \leq \mathscr{L}^t - \eta\beta \left\| \frac{\partial \mathscr{L}^t}{\partial \boldsymbol{\theta}^t} \right\|^2 - \eta\beta \left(\frac{\partial \mathscr{L}^t}{\partial \boldsymbol{Z}} \right)^\top d_z \qquad (6)$$

Теорема (Филатов 2021)

Каждый частичный предел последовательности, сгенерированной в результате оптимизации с правилом Армихо (6), является Парето станционарной точкой.

Классический и быстрый алгоритмы

Algorithm 1: Backtracking(γ, lr_{ub})

Ensure: Learning rate $\eta = lr_{ub}/\gamma$

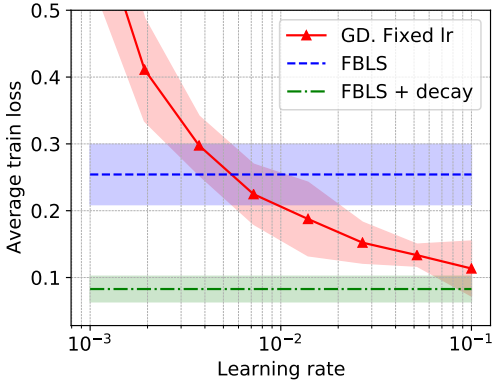
repeat
  $\eta \leftarrow \gamma \cdot \eta$
  $\tilde{\boldsymbol{\theta}}^{sh} \leftarrow \boldsymbol{\theta}^{sh} - \eta \cdot d_{sh}$
  for $t \leftarrow 1$ to T **do**
    $\tilde{\boldsymbol{\theta}}^{t} \leftarrow \boldsymbol{\theta}^{t} - \eta \cdot \nabla_{\boldsymbol{\theta}^t} L^t$
  end for
until Armijo condition 5
for $t \leftarrow 1$ to T **do**
  $\boldsymbol{\theta}^{t}_{new} \leftarrow \tilde{\boldsymbol{\theta}}^{t}$
end for
$\boldsymbol{\theta}^{sh}_{new} \leftarrow \tilde{\boldsymbol{\theta}}^{sh}$

Algorithm 2: Fast backtracking(γ, lr_{ub})

Ensure: Learning rate $\eta = lr_{ub}/\gamma$

repeat
  $\eta \leftarrow \gamma \cdot \eta$
  $z \leftarrow z - \eta \cdot d_z$
  for $t \leftarrow 1$ to T **do**
    $\tilde{\boldsymbol{\theta}}^{t} \leftarrow \boldsymbol{\theta}^{t} - \eta \cdot \nabla_{\boldsymbol{\theta}^t} L^t$
  end for
until Armijo condition 6
for $t \leftarrow 1$ to T **do**
  $\boldsymbol{\theta}^{t}_{new} \leftarrow \tilde{\theta}^{t}$
end for
$\boldsymbol{\theta}^{sh}_{new} \leftarrow \boldsymbol{\theta}^{sh} - \eta \cdot \frac{\partial \boldsymbol{\theta}^{sh}}{\partial z} d_z$

Эксперименты



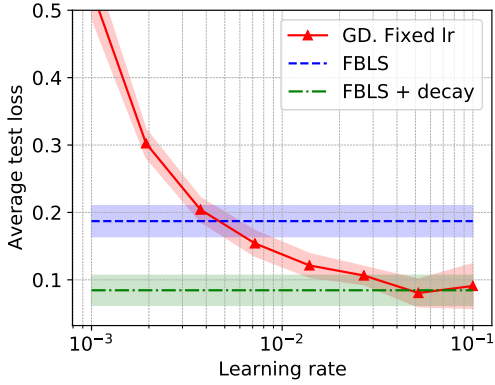
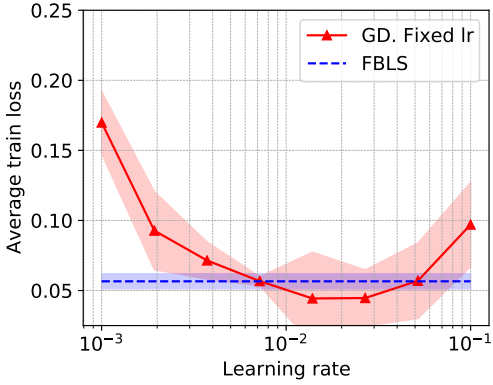


Рис.: Сравнение работы предложенного метода против градиентного спуска с фиксированным шагом после 50 эпох на датасете MultiMNIST

Эксперименты



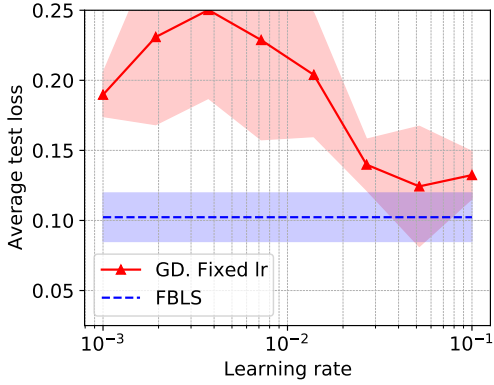


Рис.: Сравнение работы предложенного метода против градиентного спуска с фиксированным шагом после 50 эпох на датасете CIFAR-10

Эксперименты

	MNIST ↓	CIFAR-10 ↓	Cityscapes ↓
Fast backtracking	1.05 (143)	0.15 (85)	1.28 (76800)
Backtracking	1.37 (195)	1.18 (650)	-
Classical SGD	1.0 (143)	1.0 (550)	1.0 (60000)
Latent space SGD	0.95 (136)	0.14 (80)	-

Таблица: Сравнение времени работы (секунд на эпоху), меньше лучше

Результаты, выносимые на защиту

- Предложен метод оптимизации мультизадачных модели;
- Подтверждена теоретическая сходимость предложенного метода;
- Проверена эффективность метода на прикладных задачах при различных условиях.

Дискуссия

- В рамках работы не были исследованы другие методы линейного поиска. Однако, предложенные выводы не ограничены backtracking line search.
- В рамках работы не были исследованы современные методы оптимизации. Рассмотрение методов более высокого порядка и методов редукции дисперсии будут рассмотрены в дальнейшем.