Федеративное обучение и сверхпараметризация в моделях

Эйдлин Иван

Научный руководитель: А. В. Гасников

МФТИ

22 апреля 2025 г.

Начало работы

Этапы исследования

- Первый этап - исследование работ по обобщенной гладкости [1], [2], [4]
- **Второй этап - изучение сверхпараметризации [3], [5]**

Мотивация текущего исследования

- Исследование началось с изучения свойств функций с обобщенной гладкостью
- Заинтересовала техника рестартов и ее теоретическое обоснование
- Обнаружена связь между обобщенной гладкостью и сверхпараметризацией
- В рамках этого этапа: применение рестартов к сверхпараметризованным моделям

Обобщённая гладкость

Определение (L_0, L_1)-гладкости

Функция $f : \mathbb{R}^d \to \mathbb{R}$ называется (L_0, L_1)-гладкой, если для любых $x, y \in \mathbb{R}^d$ с условием $\|y - x\| \leqslant \frac{1}{L_1}$:

$$\|\nabla f(y) - \nabla f(x)\| \leqslant (L_0 + L_1\|\nabla f(x)\|)\|y - x\| \tag{1}$$

Теоретические гарантии

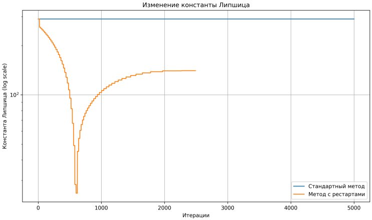
Для функций с (L_0, L_1)-гладкостью градиентный спуск с адаптивным шагом $\eta_k = (L_0 + L_1\|\nabla f(x^k)\|)^{-1}$ гарантирует:

- линейную сходимость, если $\|\nabla f(x^k)\| \geqslant \frac{L_0}{L_1}$
- сублинейную сходимость, если $\|\nabla f(x^{N-1})\| < \frac{L_0}{L_1}$

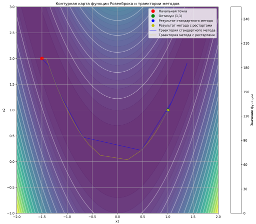
Техника рестартов

Определение

Метод рестартов для ускоренного градиентного спуска - подход, при котором периодически происходит сброс накопленного момента и переоценка параметров метода.



Константа Липшица $L(x)$



Контурная карта

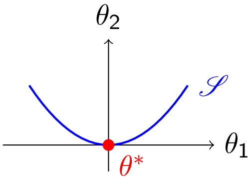
Особенности оптимизации сверхпараметризованных моделей I

Определение сверхпараметризации

Модель считается сверхпараметризованной, когда число параметров d значительно превышает количество обучающих примеров n: $d \gg n$.

Геометрия пространства решений

- Множество глобальных минимумов образует многообразие
- При $d > n$ существует бесконечное множество интерполирующих решений
- Различные решения обладают разной обобщающей способностью



Особенности оптимизации сверхпараметризованных моделей II

Феномен интерполяции и обобщения

- Парадокс: переобучение не наблюдается даже при $d \gg n$
- Существуют "хорошие"и "плохие"минимумы
- Градиентный спуск обладает имплицитной регуляризацией

Проблемы стандартных оптимизаторов

- Плато в ландшафте потерь (медленная сходимость)
- Осцилляции вблизи решения
- Сложность выбора размера шага
- Чувствительность к начальной инициализации

Обобщённая гладкость и сверхпараметризация

Связь с теорией

- Функции потерь сверхпараметризованных моделей проявляют свойства обобщённой гладкости
- Условие (L_0, L_1)-гладкости актуально в окрестности многообразия решений
- Локальные свойства ландшафта требуют адаптивных методов

Теоретические гарантии

Для функций с (L_0, L_1)-гладкостью градиентный спуск с адаптивным шагом имеет линейную сходимость при $\|\nabla f(x^k)\| \geqslant \frac{L_0}{L_1}$.

Выбор архитектуры для первых экспериментов

ResNet как объект исследования

- ResNet-18 с дополнительными полносвязными слоями (увеличение параметров)
- Модификация архитектуры: замена стандартного классификатора на последовательность

$$\begin{aligned} \mathrm{fc} = \mathrm{Linear}(512 \to 1024) \to \mathrm{ReLU} \to \\ \mathrm{Linear}(1024 \to 1024) \to \mathrm{ReLU} \to \mathrm{Linear}(1024 \to 10) \end{aligned} \tag{2}$$

Формализация задачи оптимизации

$$\min_{\theta \in \mathbb{R}^d} \mathscr{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(x_i), y_i) \tag{3}$$

где f_θ - ResNet с параметрами θ, ℓ - функция потерь (кросс-энтропия), $(x_i, y_i)_{i=1}^n$ - обучающие данные.

Применение рестартов в ResNet

Критерии определения момента рестарта в эксперименте

1. Фиксированный период (каждые 15 эпох):

$$\text{Restart if: iteration} - \text{last\_restart} \geqslant 15 \tag{4}$$

2. На основе нормы градиента на валидационной выборке:

$$\text{Restart if: } \frac{\|\nabla \mathscr{L}_{\text{val}}(x_k)\|}{\|\nabla \mathscr{L}_{\text{val}}(x_{k-p})\|} > 2.0 \tag{5}$$

3. На основе изменения функции потерь на валидации:

$$\text{Restart if: } \mathscr{L}_{\text{val}}(x_k) > \mathscr{L}_{\text{val}}(x_{k-p}) \tag{6}$$

Сравнение AdamW и AdamW с рестартами

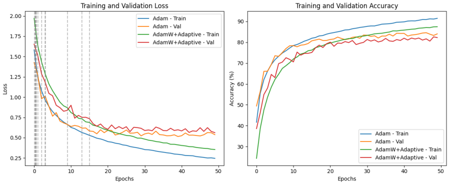
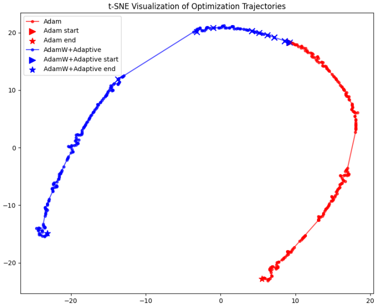


График потерь и точности

- AdamW: лучшая точность на валидации
- Меньший разрыв между обучающей и валидационной точностью



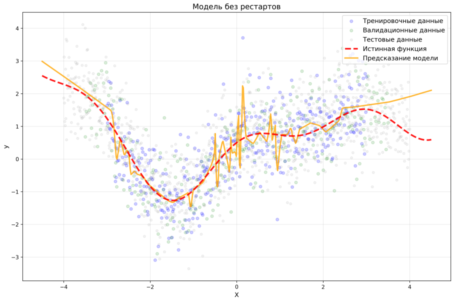
t-SNE визуализация траекторий

- Разные минимумы в пространстве параметров
- Более эффективное исследование пространства

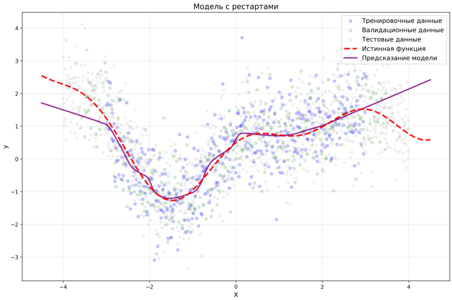
Рестарты против переобучения

Эксперимент по регрессии с большим количеством параметров

- Задача регрессии на зашумленных данных:
  $y = \sin(x) + 0.5\cos(2x) + 0.1x^2 + \varepsilon$
- Сверхпараметризованная нейронная сеть (параметров $\gg$ точек)
- Сравнение оптимизации с рестартами и без них



высокая дисперсия, шумное



низкая дисперсия, гладкое поведение

Эксперименты с YOLO моделями

Сравнение на задаче детекции объектов

- **Архитектура YOLO для задачи детекции объектов**
- **Сравнение стандартного SGD и SGD с рестартами**
- Анализ влияния периода рестартов на качество детекции
- Исследование динамики обучения с помощью t-SNE визуализации

Результаты

- Улучшение mAP (mean Average Precision) на 1-2%
- Ускорение сходимости по числу итераций

Заключение и дальнейшие планы

Основные результаты

- Разработана теория применения рестартов для сверхпараметризованных моделей
- Предложены критерии определения момента рестарта
- Продемонстрировано улучшение обобщающей способности на ResNet

Направления дальнейших исследований

- Расширение на другие архитектуры (трансформеры, диффузионные модели)
- Исследование комбинации рестартов с методами понижения размерности
- Глубокий анализ влияния рестартов на федеративное обучение
- Выявление оптимальных критериев рестарта

Литература

[1] S. S. Ablaev, A. N. Beznosikov, A. V. Gasnikov, D. M. Dvinskikh, A. V. Lobanov, S. M. Puchinin, and F. S. Stonyakin. On some works of boris teodorovich polyak on the convergence of gradient methods and their development. *Computational Mathematics and Mathematical Physics*, 64(4):635–675, Apr. 2024.

[2] A. Lobanov, A. Gasnikov, E. Gorbunov, and M. Takáč. Linear convergence rate in convex setup is possible! gradient descent method variants under (l_0, l_1)-smoothness, 2025.

[3] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever. Deep double descent: Where bigger models and more data hurt, 2019.

[4] Z. Tovmasyan, G. Malinovsky, L. Condat, and P. Richtárik. Revisiting stochastic proximal point methods: Generalized smoothness and similarity, 2025.

[5] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization, 2017.

Спасибо за внимание!