Федеративное обучение и сверхпараметризация в моделях

Эйдлин Иван

Научный руководитель: А. В. Гасников

МФТИ

17 мая 2025 г.

Введение и мотивация исследования

Актуальность проблемы

- Современные модели машинного обучения содержат миллиарды параметров
- Оптимизационные методы требуют адаптации для сверхпараметризованных моделей

Состояние исследований в этой области

- Исследования сверхпараметризации: [3], [5]
- Ранние работы по обобщенной гладкости: [1], [2] [4]

Постановка задачи

Цель исследования

- Разработать эффективный метод оптимизации для обучения сверхпараметризованных моделей
- Адаптировать технику рестартов для улучшения сходимости и обобщения
- Предложить критерии определения моментов рестарта

Формальная постановка задачи

$$\min_{\theta \in \mathbb{R}^d} \mathscr{L}(\theta) = \sum_{k=1}^{K} w_k \mathscr{L}_k(\theta) \tag{1}$$

где $\mathscr{L}_k(\theta) = \frac{1}{n_k} \sum_{i=1}^{n_k} \ell(f_\theta(x_i^k), y_i^k)$ - функция потерь k-го клиента, w_k - вес клиента, f_θ - модель с параметрами θ.

Методы исследования

Методологический подход

- Теоретический анализ связи сверхпараметризации и обобщенной гладкости
- Экспериментальное исследование сверхпараметризованных моделей (ResNet, YOLO)

Стек технологий

- Фреймворк: PyTorch
- Модели: модифицированные ResNet-18, YOLOv5
- Данные: CIFAR
- Методы визуализации: t-SNE для анализа траекторий оптимизации

Техника рестартов

Ускоренный метод Нестерова

Итерации ускоренного метода градиентного спуска:

$$y_k = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}) \tag{2}$$

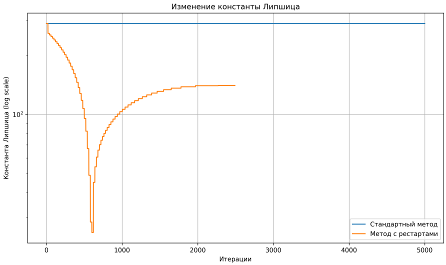
$$x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k) \tag{3}$$

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2} \tag{4}$$

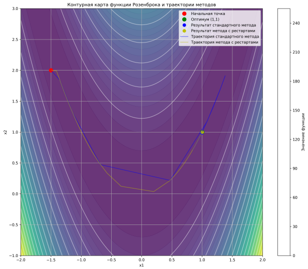
Определение

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

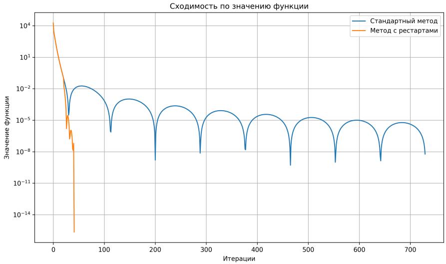
Техника рестартов



Константа Липшица $L(x)$



Контурная карта



Сравнение методов

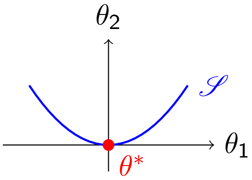
Особенности оптимизации сверхпараметризованных моделей I

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 $\|\nabla f(x^k)\| \geqslant \frac{L_0}{L_1}$.

Применение рестартов в ResNet

Критерии определения момента рестарта в эксперименте

1. Фиксированный период (каждые 15 эпох):

$$\text{Restart if: iteration} - \text{last\_restart} \geqslant 15 \tag{5}$$

2. На основе нормы градиента на валидационной выборке:

$$\text{Restart if: } \frac{\|\nabla \mathscr{L}_{\text{val}}(x_k)\|}{\|\nabla \mathscr{L}_{\text{val}}(x_{k-p})\|} > 2.0 \tag{6}$$

3. На основе изменения функции потерь на валидации:

$$\text{Restart if: } \mathscr{L}_{\text{val}}(x_k) > \mathscr{L}_{\text{val}}(x_{k-p}) \tag{7}$$

Сравнение AdamW и AdamW с рестартами

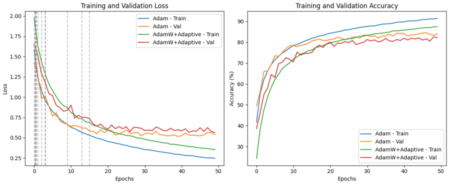
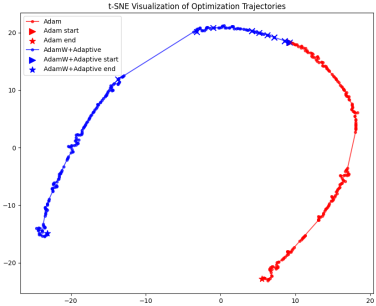


График потерь и точности

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

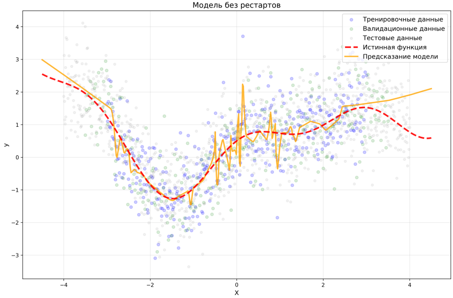


t-SNE визуализация траекторий

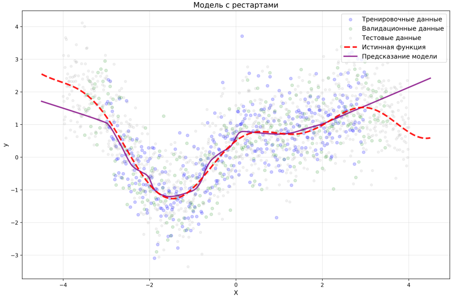
- Разные минимумы в пространстве параметров
- Более эффективное исследование пространства

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

Рестарты и методы понижения размерности

Проблема высокой размерности

- Большинство обновлений в сверхпараметризованных моделях происходит в низкоразмерных подпространствах
- Методы понижения размерности способны значительно сократить объем передаваемых данных

Предложенный подход

- Применение SVD-разложения для выделения ключевых направлений обновлений:

$$\Delta_k \approx U_k \Sigma_k V_k^T, \quad \text{где } \operatorname{rank}(\Sigma_k) \ll d \tag{8}$$

- Рестарты при смене доминирующих направлений в пространстве параметров

FedRestarts

Концепция адаптивных рестартов

1. **Локальные рестарты**: на уровне клиентов
   - При обнаружении плато или увеличения локальных потерь
   - При сильном отклонении от глобальной модели
2. **Глобальные рестарты**: на уровне сервера
   - При сильной несогласованности обновлений от клиентов
   - При деградации глобальной модели на валидации

Основные результаты

Теоретические результаты

- Выявлена связь между обобщенной (L_0, L_1)-гладкостью и поведением сверхпараметризованных моделей
- Разработаны адаптивные критерии рестартов для обучения
- Разработана и проанализирована техника комбинирования рестартов с SVD-разложением
- Предложен метод FedRestart

Заключение

Ключевые выводы исследования

- Рестарты существенно потенциально улучшают сходимость и обобщающую способность сверхпараметризованных моделей
- Теоретические гарантии сходимости подтверждаются экспериментально
- Комбинация SVD-разложения с адаптивными рестартами открывает новые возможности для масштабирования

Благодарности

Автор выражает благодарность научному руководителю А.В. Гасникову за ценные рекомендации и постановку задачи.

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!