Федеративное обучение и сверхпараметризация в моделях

Эйдлин Иван

Научный руководитель: А. В. Гасников

МФТИ

25 марта 2025 г.

Начало работы

Первый этап - исследование [1], [4], [2]
Второй этап - становление исследования неактуальным [5]
Третий этап - исследование [3]

Общие определения

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 $\|y - x\| \leqslant \frac{1}{L_1}$:

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

Определение сильной выпуклости

Функция $f : \mathbb{R}^d \to \mathbb{R}$ называется $\mu \geqslant 0$ сильно выпуклой, если для любых $x, y \in \mathbb{R}^d$:

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2 \tag{2}$$

Градиентный спуск с адаптивным шагом

Для функций с обобщенной гладкостью (L_0, L_1):

$$x_{k+1} = x_k - \eta_k \nabla f(x_k), \quad \text{где} \quad \eta_k = \frac{1}{L_0 + L_1\|\nabla f(x_k)\|} \tag{3}$$

Обобщённая гладкость

Теорема

Пусть функция f удовлетворяет условию (L_0, L_1)-гладкости и условию выпуклости. Тогда градиентный спуск с размером шага $\eta_k = (L_0 + L_1\|\nabla f(x^k)\|)^{-1}$ гарантирует:

- линейную сходимость, если $\|\nabla f(x^k)\| \geqslant \frac{L_0}{L_1}$ для $k \in [N-1]$:

$$\boxed{f(x^N) - f^* \leqslant \left(1 - \frac{1}{4L_1R}\right)^N F_0;}$$

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

$$\boxed{f(x^N) - f^* < \frac{4L_0R^2}{N}.}$$

Техника рестартов

Ускоренный метод Нестерова

Итерации ускоренного метода градиентного спуска:

$$y_k = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}) \tag{4}$$

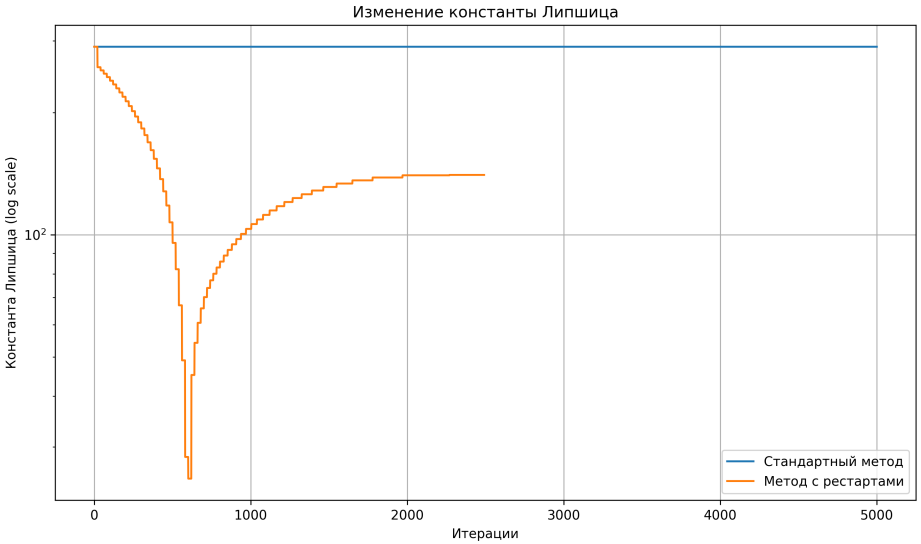
$$x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k) \tag{5}$$

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2} \tag{6}$$

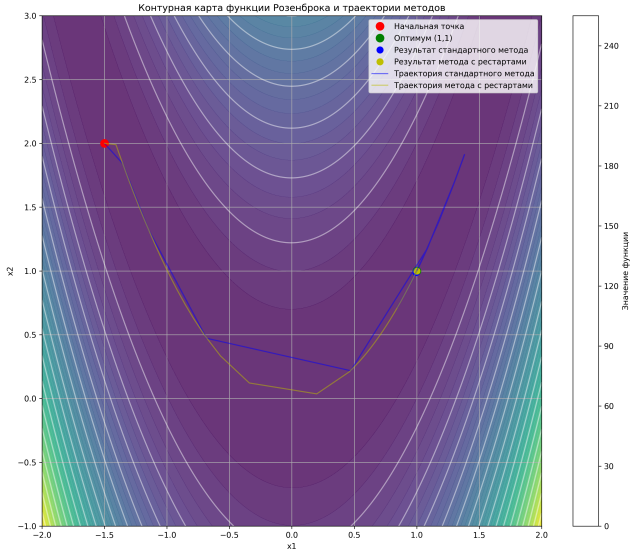
Определение

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

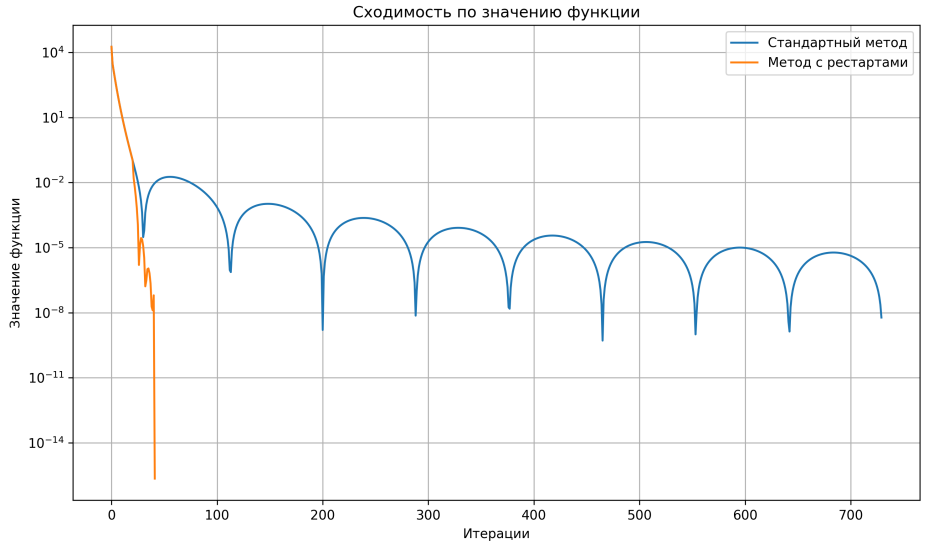
Техника рестартов



Константа Липшица $L(x)$



Контурная карта



Сравнение методов

Сверхпараметризованные модели

Определение и свойства

Сверхпараметризованная модель: $|\theta| \gg n$ (параметров больше, чем примеров)

- Множество глобальных минимумов: $\mathscr{S} = \{\theta : \mathscr{L}(\theta) = 0\}$
- Интерполяционный режим: идеальное соответствие обучающим данным
- Свойство неединственности решения создаёт проблему выбора модели

Примеры

- Современные нейронные сети (ResNet, трансформеры)
- Языковые модели (GPT, BERT): миллиарды параметров
- Модели компьютерного зрения с избыточным числом фильтров

Федеративное обучение и сверхпараметризация

Применение рестартов

- Периодические рестарты между циклами агрегации помогают избежать застревания в локальных минимумах
- Адаптивная частота рестартов на основе неоднородности данных на разных устройствах
- Рестарты с координацией глобальной и локальных моделей: $\theta_k^{t+1} = \theta_k^t - \eta \nabla \mathscr{L}_k(\theta_k^t) + \beta(\theta^t - \theta_k^t)$
- Выход из плато сверхпараметризованных ландшафтов за счет сброса накопленного момента

Открытые проблемы и перспективные направления

Открытые проблемы

- Теоретическое обоснование влияния сверхпараметризации на сходимость градиентных методов с рестартами
- Оптимальные критерии для определения момента рестарта в гетерогенных федеративных системах
- Влияние неоднородности данных на поведение сверхпараметризованных моделей

Перспективные направления

- Адаптивные схемы рестартов на основе локальной информации о гладкости функций потерь
- Комбинация техник рестартов с методами редукции размерности для эффективного обучения

План на семестр

Шаг 1

Применение рестартов в ускоренных алгоритмах, в условиях обобщённой гладкости.

Шаг 2

Переход к изучению сверхпараметризованных моделей и федеративного обучения

Шаг ?

Теоретический анализ и экспериментальная проверка адаптивных стратегий рестартов для федеративного обучения

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. Kulunchakov and J. Mairal. A generic acceleration framework for stochastic composite optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[3] 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.

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

[5] D. Vankov, A. Rodomanov, A. Nedich, L. Sankar, and S. U. Stich. Optimizing (l_0, l_1)-smooth functions by gradient methods, 2025.

Спасибо за внимание!