УДК 519.85

Методы оптимизации при обобщённой (L_0, L_1, L_2)-гладкости третьего порядка

Е. Р. Обжерин
Физтех-школа прикладной математики и информатики, Московский физико-технический институт

Тезис доклада. Рассматривается задача выпуклой минимизации функции $f : \mathbb{R}^d \to \mathbb{R}$ при ослабленных условиях гладкости. Классическая L-гладкость, выражающаяся в ограничении

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|,$$

недостаточно гибка для описания поведения многих функций потерь в машинном обучении. В ответ на это была предложена модель (L_0, L_1)-гладкости второго порядка [1, 2], в которой

$$\|\nabla^2 f(x)\| \leq L_0 + L_1\|\nabla f(x)\|.$$

Под этим условием были разработаны модификации градиентного метода [3, 4], объяснено поведение clipping-алгоритмов, а также показана применимость адаптивных методов типа Adam [5].

Однако эмпирические наблюдения показывают, что в некоторых архитектурах наблюдается квадратичная зависимость между градиентом и гессианом [6]. Это мотивирует к дальнейшему обобщению: (L_1, L_2)-гладкость третьего порядка:

$$\|\nabla^3 f(x)\| \leq L_1 + L_2\|\nabla^2 f(x)\|,$$

а также (L_0, L_1, L_2)-гладкость [7]:

$$\|\nabla^3 f(x)\| \leq L_0 + L_1\|\nabla f(x)\| + L_2\|\nabla^2 f(x)\|.$$

В настоящей работе предлагается метод оптимизации на основе кубической регуляризации Ньютона, где параметр регуляризации $M(x)$ адаптивно зависит от текущих норм градиента и гессиана. Основной технический вклад заключается в получении оценки остаточного члена разложения Тейлора третьего порядка с использованием обобщённого неравенства Гронуолла. Это позволяет управлять ростом кривизны вдоль траектории и получать корректные аппроксимации без предположения о глобальной ограниченности тензора третьего порядка.

Поставлена задача получения аппроксимации без экспоненциального остатка, что имеет значение для устойчивости метода при больших шагах. Предполагается дальнейшее развитие модели и её применение в высокопроизводительных методах обучения.

Благодарности. Автор благодарит Д. И. Камзолова за постановку задачи и обсуждение теоретических аспектов.

Список литературы

[1] Jingzhao Zhang и др. «Why Gradient Clipping Accelerates Training: A Theoretical Justification for Adaptivity». В: *International Conference on Learning Representations (ICLR)*. 2020. arXiv: 1905.11881 [math.OC].

[2] Bohang Zhang и др. «Improved Analysis of Clipping Algorithms for Non-convex Optimization». В: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020. arXiv: 2006.06200 [cs.LG].

[3] Eduard Gorbunov и др. «Methods for Convex (L0, L1)-Smooth Optimization: Clipping, Acceleration, and Adaptivity». В: *arXiv preprint arXiv:2409.14989* (2024).

[4] Daniil Vankov и др. «Optimizing (L0, L1)-Smooth Functions by Gradient Methods». В: *arXiv preprint arXiv:2410.10800* (2025).

[5] Anonymous. «Provable Benefit of Adaptivity in Adam». Under review at ICLR 2023. 2023.

[6] Anonymous. «An Empirical Study of the (L0, L1)-Smoothness Condition in Deep Learning». Submitted to Mathematics of Modern Machine Learning Workshop, NeurIPS 2024. 2024.

[7] Aleksandr Lobanov и др. «Linear Convergence Rate in Convex Setup is Possible! Gradient Descent Method Variants under (L0, L1)-Smoothness». В: *arXiv preprint arXiv:2412.17050* (2025).