

Построение и исследование адаптивного по константе сильной выпуклости градиентного метода

Плетнев Никита Вячеславович

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель: Гасников Александр Владимирович

Задача

Требуется построить эффективный метод безусловной оптимизации первого порядка.

Ожидания

- Предложить модификацию быстрого градиентного метода, избавленную от присущих ему недостатков.
- Сравнить построенный метод с другими подходами к безусловной оптимизации.

Решается задача безусловной минимизации

$$\min_{x \in \mathbb{R}^d} f(x).$$

Предположения

- решение $x^* = \arg \min_{x \in \mathbb{R}^d} f(x)$ существует;
- градиент функции $f(x)$ обладает свойством Липшица с константой $L > 0$:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \forall x, y \in \mathbb{R}^d;$$

- функция $f(x)$ является сильно выпуклой с неизвестной нам константой $\mu > 0$:

$$\frac{\mu}{2}\|x - x^*\|^2 \leq f(x) - f(x^*) \leq \frac{1}{2\mu}\|\nabla f(x)\|^2.$$

В качестве базового метода взят алгоритм первого порядка с фиксированным шагом OGM-G. В [1] показана его оптимальность в классе методов с заданным числом шагов фиксированной длины.

OGM-G

Вход: $f \in \mathcal{F}_L(\mathbb{R}^d)$, $\mathbf{x}_0 = \mathbf{y}_0 \in \mathbb{R}^d$, $N \geq 1$.

Для $i = 0 \dots N-1$:

$$\mathbf{y}_{i+1} = \mathbf{x}_i - \frac{1}{L} \nabla f(\mathbf{x}_i);$$

$$\mathbf{x}_{i+1} = \mathbf{y}_{i+1} + \beta_i(\mathbf{y}_{i+1} - \mathbf{y}_i) + \gamma_i(\mathbf{y}_{i+1} - \mathbf{x}_i).$$

Коэффициенты β_i, γ_i вычисляются по формулам:

$$\beta_i = \frac{(\theta_i - 1)(2\theta_{i+1} - 1)}{\theta_i(2\theta_i - 1)}; \quad \gamma_i = \frac{2\theta_{i+1} - 1}{2\theta_i - 1}.$$

OGM-G

Последовательность $\{\theta_i\}_{i=0}^N$ строится следующим образом:

$$\theta_i = \begin{cases} \frac{1 + \sqrt{1 + 8\theta_1^2}}{2}, & i = 0; \\ \frac{1 + \sqrt{1 + 4\theta_{i+1}^2}}{2}, & 1 \leq i < N; \\ 1, & i = N. \end{cases}$$

Оценка количества шагов

Из оценок, полученных в [1], выводится, что для сокращения нормы градиента вдвое можно взять

$$N = 2\sqrt{2\frac{L}{\mu}}.$$

Проблемы при использовании

- необходимо знать константу сильной выпуклости μ ;
- ее оценивание существенно сложнее, чем исходная задача;
- само ограничение длины шага $\frac{1}{L}$ значительно замедляет сходимость.

Пути решения

- незнание μ можно обойти, построив адаптивный по μ метод на основе OGM-G;
- также можно попытаться получить адаптивность по L ;
- необходимо сравнение эффективности с другими подходами (например, метод сопряженных градиентов).

Идея принадлежит Ю. Е. Нестерову.

Алгоритм

Инициализируем μ произвольным положительным значением, например $\mu_0 = 1$. На каждом шаге (k — номер шага):

- 1 $\mu_k := 2\mu_{k-1}$;
- 2 Выполняем OGM-G с начальным значением — результатом прошлого шага и $N = 2\sqrt{2\frac{L}{\mu}}$ итерациями;
- 3 Если выполнено условие $\|\nabla f(x^N)\| \leq \frac{1}{2}\|\nabla f(x^0)\|$, то переходим к следующему шагу;
- 4 Иначе $\mu_k := \frac{\mu_k}{2}$ и возвращаемся к пункту 2.

Оценки

Каждое очередное уменьшение нормы градиента требует не более

$$\frac{4}{\sqrt{2}-1} \sqrt{\frac{L}{\mu_k}}$$

вычислений градиента функции.

При этом μ_k отличается не более чем вдвое от истинного значения константы сильной выпуклости в окрестности данной части траектории метода. Использование «глобального» значения μ могло бы повысить количество операций.

Достижение условия остановки $\|\nabla f(x)\| \leq \varepsilon$ требует

$O\left(\sqrt{\frac{L}{\mu}} \log_2 \frac{\|\nabla f(x^0)\|}{\varepsilon}\right)$ вычислений $\nabla f(x)$.

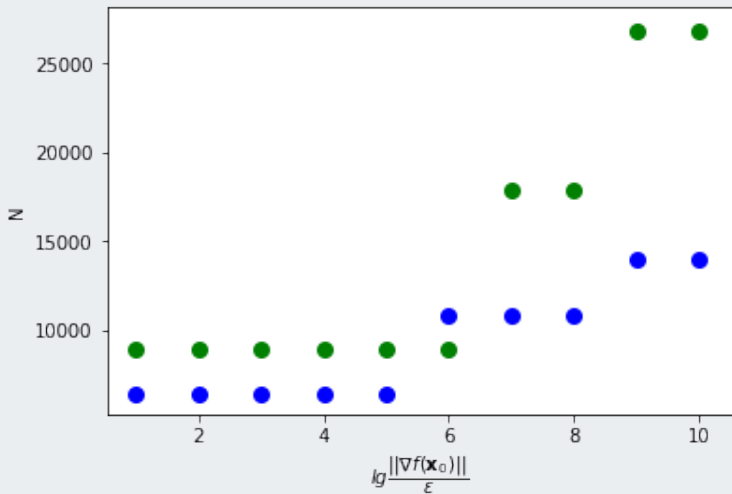
Проверка проводилась на функции $f(\mathbf{x}) = \frac{1}{2} \langle \mathbf{x}, A\mathbf{x} \rangle$, где матрица A — диагональная матрица порядка 10 с элементами $a_{ii} = i^3$. $\mathbf{x}_0 = (1, 2, \dots, 10)$.

Свойства этой функции:

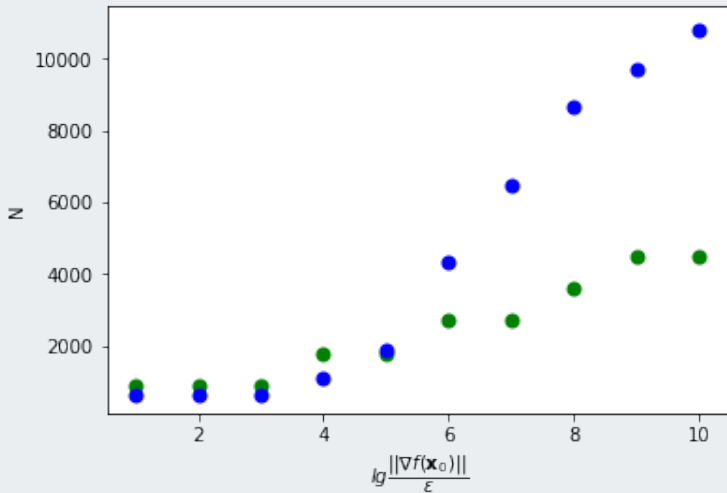
- сильно выпуклая, $\mu = 1$;
- градиент липшицев, $L = 1000$;
- единственный минимум $\mathbf{x} = \mathbf{0}$.

Сравнивались результаты работы OGM-G с количеством итераций, определенным исходя из «предполагаемого» значения константы сильной выпуклости, и результаты работы адаптивного алгоритма с таким же начальным значением μ . Исследовалось количество итераций, необходимых для достижения заданной точности. Мера точности — норма градиента.

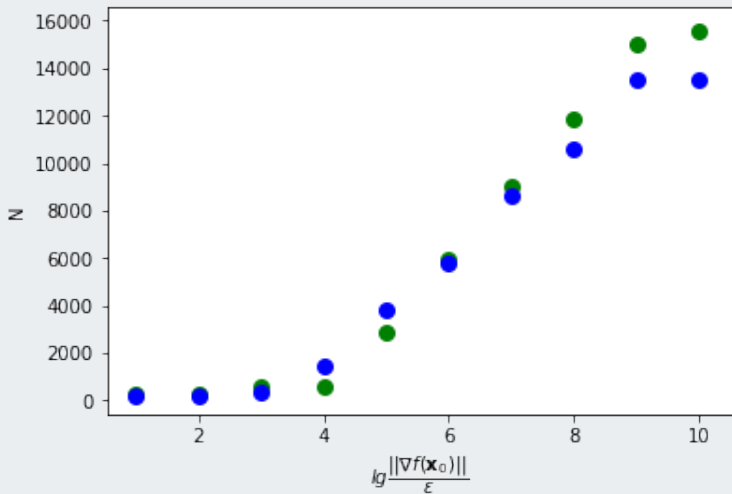
$\mu = 0.01$



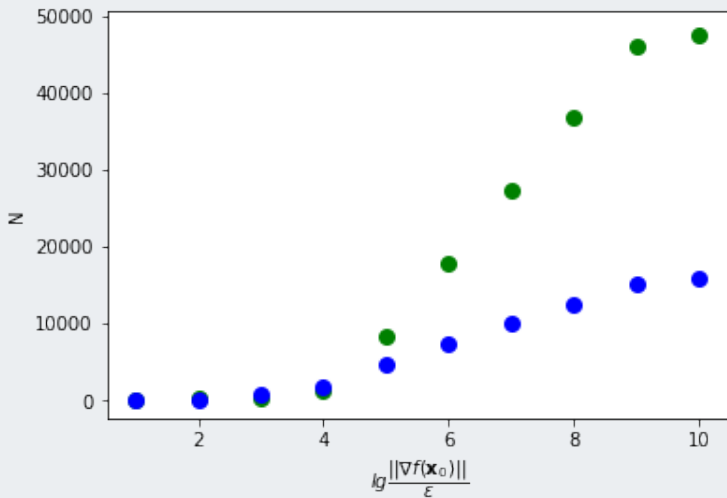
$\mu = 1$



$\mu = 10$



$\mu = 100$



Интерпретация графиков

При $\mu = 1$ — то есть, истинном значении константы сильной выпуклости — адаптивный метод работает хуже обычного. Это ожидаемо, потому что подбор константы и перескакивания вокруг истинного значения тратят много итераций зря. Во всех остальных случаях адаптивный метод требует в 2-4 раза меньше итераций для достижения заданной точности.

Выводы

- Построенный адаптивный метод оказался эффективнее, чем исходный метод OGM-G.
- Метод не требует знания константы сильной выпуклости и предопределенного числа итераций.
- Однако проблема неизвестности константы Липшица для градиента по-прежнему не решена.

- Исследовать вопрос об адаптивности по константе Липшица.
- Исследовать сходимость с использованием других критериев точности.
- Сравнить эффективность полученных методов с другими подходами.
- Сформулировать и доказать теоремы о скорости сходимости.

- [1] Optimizing the Efficiency of First-order Methods for Decreasing the Gradient of Smooth Convex Functions, Donghwan Kim, Jeffrey A. Fessler, 2018, 14 с., arXiv:1803.06600v2;
- [2] Современные численные методы оптимизации. Метод универсального градиентного спуска. Учебное пособие, Гасников А. В., 2018, 220 с., ISBN 978-5-7417-0667-1
- [3] Введение в выпуклую оптимизацию, Нестеров Ю. Е., 2013, 279 с., ISBN 978-5-9405-7623-5