Адаптивность градиентных методов

Плетнев Никита Вячеславович

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

Научный руководитель: д. ф.-м. н. Гасников Александр Владимирович

Постановка задачи

Изучается задача безусловной минимизации

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

Предположения

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

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

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

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

Цель работы

Цель

Требуется построить более эффективный, чем существующие, метод безусловной выпуклой оптимизации первого порядка.

Ожидания

- Предложить модификацию быстрого градиентного метода, не требующую знания констант сильной выпуклости и Липшица.
- Оценить скорость сходимости полученного метода.

Исходный алгоритм

В качестве базового метода взят алгоритм первого порядка с фиксированным шагом OGM-G. В Kim, Fessler «Optimizing the Efficiency of First-order Methods for Decreasing the Gradient of Smooth Convex Functions», (2018) [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:

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

Исходный алгоритм

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], следует, что для сокращения нормы градиента вдвое требуется взять

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

Исходный алгоритм

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.

Путь решения

Построение адаптивного по μ и L метода на основе OGM-G.

Ближайшее решение

Статья Lei, Jordan «On the Adaptivity of Stochastic Gradient-Based Optimization», (2019) [2]. Недостаток: требуемое число итераций достигается лишь с точностью до логарифмического множителя.

Адаптивный по μ метод ACGM

Идея принадлежит Нестерову Ю. Е. (пособие [3], 2018).

Алгоритм

Вход: $f \in \mathcal{F}_L(\mathbb{R}^d)$, $\mathbf{x}_0 \in \mathbb{R}^d$, L, $\mu_0, \beta, \varepsilon$.
Пока не выполнен критерий останова $||\nabla f(\mathbf{x}_k)|| \leq \varepsilon$:
k — номер шага.

1. $\mu_k := \beta \mu_{k-1}$;
2. $\mathbf{x}_k = OGMG(f, \mathbf{x}_{k-1}, L, \mu_k)$;
3. Если выполнено условие $||\nabla f(\mathbf{x}_k)|| \leq \frac{1}{2}||\nabla f(\mathbf{x}_{k-1})||$, то перейти к следующему шагу;
4. Иначе $\mu_k := \frac{\mu_k}{\beta}$ и вернуться к пункту 2. Если при этом выполнено условие $||\nabla f(\mathbf{x}_k)|| < ||\nabla f(\mathbf{x}_{k-1})||$, то $\mathbf{x}_{k-1}$ заменяется на x_k.

Теоремы о сходимости ACGM

В работе доказаны следующие теоремы о сходимости ACGM.

Теорема 1

Алгоритм ACGM с оптимальным $\beta = 4$ и $\mu_0 > \max_k \mu_k^{loc}$ достигает точки $\mathbf{x}$, удовлетворяющей критерию останова $||\nabla f(\mathbf{x})|| \leq \varepsilon$, за не более чем $C \sum_{k=0}^{K-1} \sqrt{\frac{L}{\mu_k^{loc}}}$ вычислений градиента, где $K = \log_2 \frac{||\nabla f(\mathbf{x}_0)||}{\varepsilon}$, $C = \frac{\sqrt{8}\beta}{\sqrt{\beta}-1} = 8\sqrt{2}$.

Теорема 2

Алгоритм ACGM с оптимальным $\beta = 4$ и $\mu_0 > \max_k \mu^{loc}$ достигает точки $\mathbf{x}$, удовлетворяющей критерию останова $||\nabla f(\mathbf{x})|| \leq \varepsilon$, за не более чем $CK\sqrt{\frac{L}{\mu}}$ вычислений градиента, где $K = \log_2 \frac{||\nabla f(\mathbf{x}_0)||}{\varepsilon}$, $C = \frac{\sqrt{8}\beta}{\sqrt{\beta}-1} = 8\sqrt{2}$.

Выбор параметров β и μ_0

Оптимальное β

Минимизация коэффициента $C(\beta)$ из теоремы 2, полученного при рассмотрении наихудших оценок, дает значение $\beta = 4$.

Оптимальное μ_0

Если $\mu_0 = \frac{\mu^{loc}}{4^k}$, то на первом шаге ACGM будет выполнено $M = 2\sqrt{2\frac{L \cdot 4^k}{\mu^{loc}}} = 2^k N$ итераций.
Для достижения такого же результата требуется $k+1$ рестартов.
i-ый рестарт требует $\frac{N}{2^i}$ итераций. Суммарное количество не превосходит $2N$.
Применение заниженного значения константы сильной выпуклости значительно увеличивает количество итераций.
Поэтому оптимальным значением μ_0 является L.

В §5 пособия [3] рассматривается метод Нестерова Ю. Е..

Универсальный градиентный спуск

Вход: $f \in \mathcal{F}(\mathbb{R}^d)$, $\mathbf{x}_0 \in \mathbb{R}^d$, L_0, ε.
Пока не выполнен критерий останова $||\nabla f(\mathbf{x}_k)|| \leq \varepsilon$:
k — номер шага.

1. $L_{k+1} := \frac{L_k}{2}$;
2. $\mathbf{x}_{k+1} := \mathbf{x}_k - \frac{1}{L_{k+1}} \nabla f(\mathbf{x}_k)$;
3. Если выполнено условие

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{1}{2L_{k+1}} ||\nabla f(\mathbf{x}_k)||^2 + \frac{\varepsilon}{2},$$

   то перейти к следующему шагу;
4. Иначе $L_{k+1} := 2L_{k+1}$ и вернуться к пункту 2.

OGM-GL

Вход: $f \in \mathcal{F}(\mathbb{R}^d)$, $\mathbf{x}_0 = \mathbf{y}_0 \in \mathbb{R}^d$, L_0, $N \geq 1$, ε.
k — номер попытки выполнения цикла
$L_{k+1} := \dfrac{L_k}{2}$
Для $i = 0 \dots N-1$:
$\mathbf{y}_{i+1} = \mathbf{x}_i - \frac{1}{L_{k+1}} \nabla f(\mathbf{x}_i)$;
Если $f(\mathbf{y}_{i+1}) > f(\mathbf{x}_i) - \frac{1}{2L_{k+1}} ||\nabla f(\mathbf{x}_i)||^2 + \frac{\varepsilon}{2}$, то $L_{k+1} := 2L_{k+1}$ и вернуться к началу цикла.
$\mathbf{x}_{i+1} = \mathbf{y}_{i+1} + \beta_i(\mathbf{y}_{i+1} - \mathbf{y}_i) + \gamma_i(\mathbf{y}_{i+1} - \mathbf{x}_i)$.
Выход: $\mathbf{x}_N$, L_{end}.

Поскольку значение L обновляется при работе OGM-GL, конечное значение добавлено к выходу алгоритма.

Адаптивный по L и μ метод ALGM

Алгоритм

Вход: $f \in \mathcal{F}(\mathbb{R}^d)$, $\mathbf{x}_0 \in \mathbb{R}^d$, L_0, μ_0, β, ε.
Пока не выполнен критерий останова $||\nabla f(\mathbf{x}_k)|| \leq \varepsilon$:
k — номер шага.

1. $\mu_k := \beta \mu_{k-1}$;
2. $\mathbf{x}_k, L_k = OGMGL\left(f, \mathbf{x}_{k-1}, L_{k-1}, \left\lceil \sqrt{\frac{8L_{k-1}}{\mu_k}} \right\rceil, \varepsilon, \right)$;
3. $\mu_k := \mu_k \cdot \frac{L_k}{L_{k-1}}$
4. Если выполнено условие $||\nabla f(x_k)|| \leq \frac{1}{2}||\nabla f(x_{k-1})||$, то перейти к следующему шагу;
5. Иначе $\mu_k := \frac{\mu_k}{\beta}$ и вернуться к пункту 2. Если при этом выполнено условие $||\nabla f(x_k)|| < ||\nabla f(x_{k-1})||$, то x_{k-1} заменяется на x_k.

Выход: $\mathbf{x}_k$.

В тексте работы доказана следующая теорема.

Теорема 3

Алгоритм ALGM с оптимальным $\beta = 4$ достигает точки $\mathbf{x}$, удовлетворяющей критерию останова $||\nabla f(\mathbf{x}_k)|| \leq \varepsilon$, за не более чем $C\sqrt{\frac{L}{\mu}}\left(3K + \log_2 \frac{L}{L_0}\right)$ вычислений градиента и $2C\sqrt{\frac{L}{\mu}}\left(3K + \log_2 \frac{L}{L_0}\right)$ вычислений функции, где $K = \log_2 \frac{||\nabla f(\mathbf{x}^0)||}{\varepsilon}$, $C = \frac{\sqrt{8}\beta}{\sqrt{\beta}-1} = 8\sqrt{2}$.

Численные эксперименты

Функция для проверки

$$f(x_1, x_2) = a(\frac{x_1^2}{2} + \frac{x_1^4}{4}) + b(\frac{x_2^2}{2} + \frac{x_2^4}{4}),\text{ где } a = 0.1, b = 100.$$

Матрица Гессе

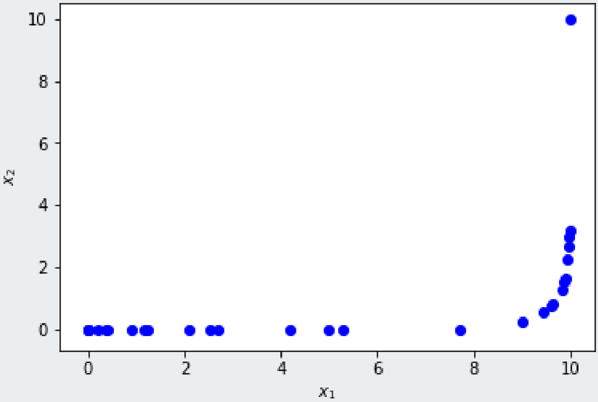
$$\left\| \begin{matrix} a(1+3x_1^2) & 0 \\ 0 & b(1+3x_2^2) \end{matrix} \right\|$$

Начальная точка — $\mathbf{x}_0 = (10, 10)$.

Свойства функции

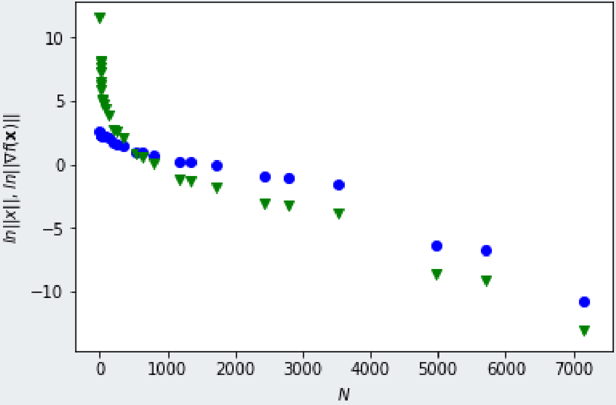
- сильно выпуклая, $\mu = 0.1$;
- градиент липшицев, $L = 30100$;
- единственный минимум $\mathbf{x} = \mathbf{0}$, в его окрестности $\frac{L}{\mu} = \frac{b}{a} = 1000$.

Траектория метода ACGM



Траектория ACGM при $\mu_0 = L$, $\beta = 4$, $\varepsilon = \frac{||\nabla f(x_0)||}{10^{10}}$.

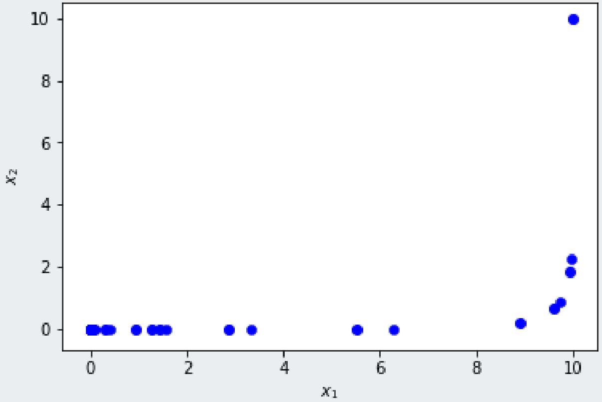
Убывание логарифма нормы при работе ACGM



Синими кружками отмечен натуральный логарифм $||\mathbf{x}||$, зелеными треугольниками — $||\nabla f(\mathbf{x})||$.

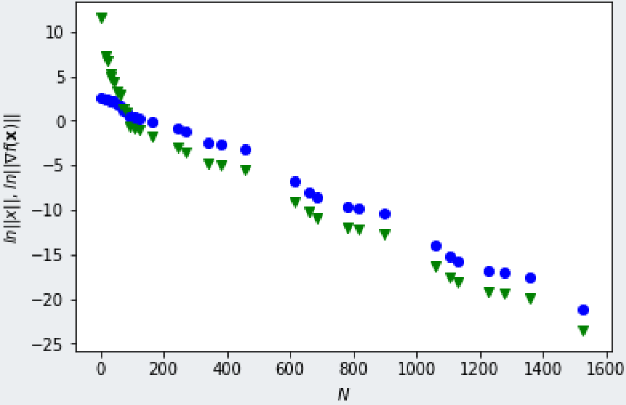
Численные эксперименты

Траектория метода ALGM



Траектория ALGM при $\mu_0 = L_0 = 1$, $\beta = 4$, $\varepsilon = 10^{-10}$.

Убывание логарифма нормы при работе ALGM

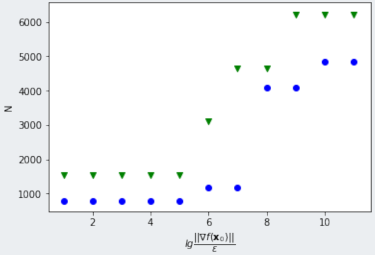


Синими кружками отмечен натуральный логарифм $||\mathbf{x}||$, зелеными треугольниками — $||\nabla f(\mathbf{x})||$.

Численные эксперименты

Зависимость количества итераций от ε в критерии останова для ACGM (синие кружки) и OGM-G (зеленые треугольники).

$\mu = 0.1$



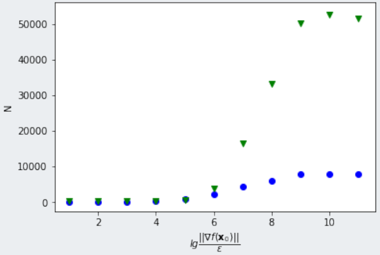
Интерпретация

В данном случае, когда начальное значение μ_0 совпадает с истинным значением константы сильной выпуклости, ACGM требует меньше итераций, чем OGM-G.

Численные эксперименты

Зависимость количества итераций от ε в критерии останова для ACGM (синие кружки) и OGM-G (зеленые треугольники).

$\mu = 10$

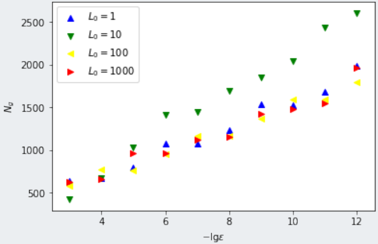


Интерпретация

Когда начальное значение μ_0 превышает на порядок истинное значение константы сильной выпуклости, ACGM требует значительно меньше итераций, чем OGM-G.

Зависимость количества итераций от ε в критерии останова для ALGM.

$L_0 \in \{1, 10, 100, 1000\}$



Интерпретация

График подтверждает линейный характер зависимости числа итераций от логарифма требуемой погрешности.

Результаты

Интерпретация графиков

- ACGM превосходит по скорости сходимости базовый алгоритм;
- ALGM сходится и превосходит ACGM по скорости сходимости;
- теоремы 2 и 3 экспериментально подтверждены.

Выводы

- Построенный метод ACGM, не требующий знания константы сильной выпуклости, оказался эффективнее, чем исходный метод OGM-G.
- ALGM не требует знания констант Липшица и сильной выпуклости, а также предопределенного числа итераций.
- Доказаны теоремы о скорости сходимости ACGM и ALGM.

Результаты, выносимые на защиту

- Алгоритмы безусловной выпуклой оптимизации ACGM и ALGM;
- Теорема 2 о скорости сходимости ACGM;
- Теорема 3 о скорости сходимости ALGM.

Использованные источники

[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] On the Adaptivity of Stochastic Gradient-Based Optimization, Lihua Lei, Michael I. Jordan, 2019, 46 с., arXiv:1904.04480v2;

[3] Современные численные методы оптимизации. Метод универсального градиентного спуска. Учебное пособие, Гасников А. В., 2018, 220 с., ISBN 978-5-7417-0667-1.