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

Цель работы

Задача

Требуется построить эффективный метод безусловной оптимизации первого порядка.

Ожидания

- Предложить модификацию быстрого градиентного метода, избавленную от присущих ему недостатков.
- Доказать теоремы о сходимости полученного метода.

Постановка задачи

Решается 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 нам константой $\mu > 0$:

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

Исходный алгоритм

В качестве базового метода взят алгоритм первого порядка с фиксированным шагом OGM-G. В [1] Optimizing the Efficiency of First-order Methods for Decreasing the Gradient of Smooth Convex Functions, Donghwan Kim, Jeffrey A. Fessler 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 $i = 0 \ldots N-1$:
$\mathbf{y}_{i+1} = \mathbf{x}_i - \frac{1}{L}\nabla f(\mathbf{x}_i)$;
$\mathbf{x}_{i+1} = \mathbf{y}_{i+1} + \beta_i(\mathbf{y}_{i+1} - \mathbf{y}_i) + \gamma_i(\mathbf{y}_{i+1} - \mathbf{x}_i)$.

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:

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

Оценка количества шагов

Из оценок, полученных в [1], выводится, 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 метода на основе OGM-G.

Ближайшее решение

Статья [3] On the Adaptivity of Stochastic Gradient-Based Optimization, Lihua Lei, Michael I. Jordan. Недостаток: требуемое число итераций достигается лишь с точностью до логарифмического множителя.

Идея принадлежит Ю. Е. Нестерову (пособие [2]).

Алгоритм

Вход: $f \in \mathcal{F}_L(\mathbb{R}^d)$, $\mathbf{x}_0 \in \mathbb{R}^d$, L, $\mu_0, \beta, \varepsilon$.
Пока не выполнено условие остановки (5):
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 $||\nabla f(x_k)|| \leq \frac{1}{2}||\nabla f(x_{k-1})||$, то перейти к следующему шагу;
4. Иначе $\mu_k := \frac{\mu_k}{\beta}$ и вернуться к пункту 2. Если при этом выполнено условие $||\nabla f(x_k)|| < ||\nabla f(x_{k-1})||$, то x_{k-1} заменяется на x_k.

Оценки

Каждое очередное уменьшение нормы градиента требует не более

$$\frac{\sqrt{8\beta}}{\sqrt{\beta}-1}\sqrt{\frac{L}{\mu_k}}$$

вычислений градиента функции.
При этом μ_k отличается не более чем в β раз от истинного значения константы сильной выпуклости в окрестности данной части траектории метода. Использование значения μ, соответствующего определению сильной выпуклости во всем пространстве повышает 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 $\nabla f(x)$.

Теоремы о сходимости

Теорема 1

Алгоритм ACGM с оптимальным $\beta = 4$ и $\mu_0 > \max_k \mu_k^{loc}$ достигает точки $\mathbf{x}$, удовлетворяющей критерию останова (5), за не более чем $C \sum\limits_{k=0}^{K-1} \sqrt{\frac{L}{\mu_k^{loc}}}$ 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 (5), 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.

Функция для проверки

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

Матрица Гессе

$$\begin{Vmatrix} a(1 + 3x_1^2) & 0 \\ 0 & b(1 + 3x_2^2) \end{Vmatrix}$$

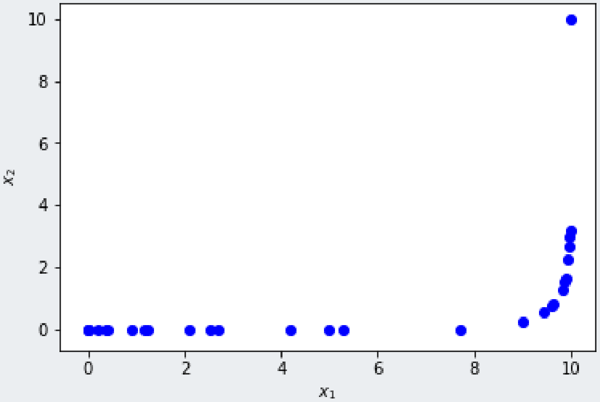
Начальная точка — $\mathbf{x}_0 = (10, 10)$.

Свойства функции

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

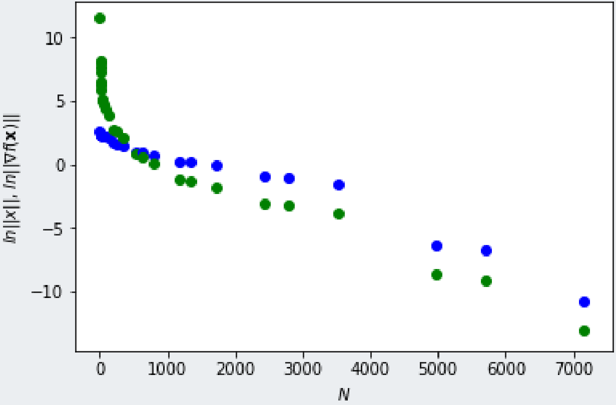
Эксперименты

Траектория метода



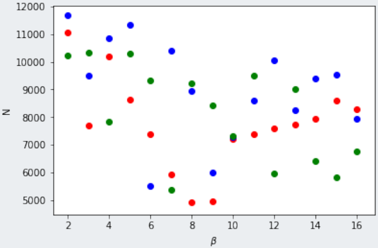
Траектория ACGM при $\mu_0 = L$, $\beta = 4$, $\varepsilon = \frac{||\nabla f(x_0)||}{10^{10}}$.

Убывание логарифма нормы



Синим цветом отмечен натуральный логарифм $||\mathbf{x}||$, зеленым — $||\nabla f(\mathbf{x})||$.

Зависимость числа итераций от параметра β



$$\varepsilon = \frac{||\nabla f(\mathbf{x}_0)||}{10^{10}};$$

$\mu_0 = 1$ (красные точки),

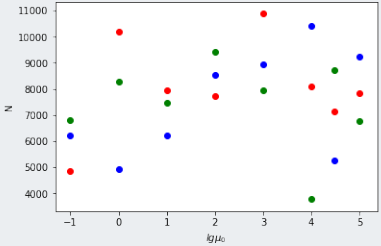
$\mu_0 = 1000$ (синие точки),

$\mu_0 = 100000$ (зеленые точки).

Интерпретация

Минимизация зависящего от β коэффициента выполнялась в предположении о достижении всех верхних, наихудших оценок. Данное предположение не выполняется, поэтому количество итераций оказывается минимальным при других значениях β. Однако именно $\beta = 4$ дает наименьшее значение $C(\beta)$.

Зависимость числа итераций от параметра μ_0



$\varepsilon = \frac{||\nabla f(\mathbf{x}_0)||}{10^{10}}$;

$\beta = 4$ (красные точки),

$\beta = 8$ (синие точки),

$\beta = 16$ (зеленые точки).

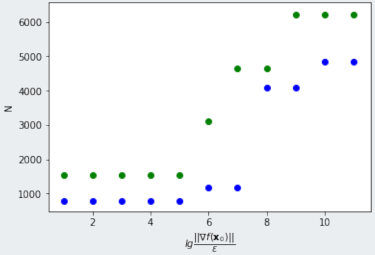
Интерпретация

Закономерность в расположении точек не обнаруживается, для каждого из рассмотренных значений β выделяется два минимума: в окрестности L и в окрестности истинного значения μ, что частично подтверждает полученные выводы об оптимальном значении μ_0.

Эксперименты

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

$\mu = 0.1$

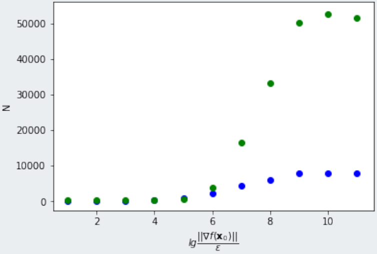


Интерпретация

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

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

$\mu = 10$



Интерпретация

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

Результаты

Интерпретация графиков

Адаптивный метод превосходит по скорости сходимости базовый алгоритм; теорема 2 экспериментально подтверждена.

Выводы

- Построенный адаптивный метод оказался эффективнее, чем исходный метод OGM-G.
- Метод не требует знания константы сильной выпуклости и предопределенного числа итераций.
- Доказаны теоремы о скорости сходимости.

Результаты, выносимые на защиту

- Алгоритм безусловной выпуклой оптимизации ACGM;
- Теорема 2 о скорости сходимости ACGM.

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

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