Адаптивная по константе сильной выпуклости модификация быстрого градиентного метода

Плетнев Н. В., Гасников А. В., Нестеров Ю. Е..

Аннотация Эта статья посвящена построению варианта быстрого градиентного метода, адаптивного по константе сильной выпуклости. Применяются полученные в [1] метод OGM-G и связанные с ним оценки для нормы градиента. Для достижения адаптивности по константе сильной выпуклости используется предложенный в [2] способ, связанный с последовательным делением предполагаемого значения константы пополам.

Ключевые слова Методы первого порядка, градиентные методы, гладкая выпуклая оптимизация, сильная выпуклость, адаптивность, конструкция рестартов.

1 Введение

Задачи оптимизации функций высокой размерности имеют многообразные приложения, например, в машинном обучении и управлении. Методы первого порядка пользуются большой популярностью, потому что их реализация требует вычисления только значения функции, ее градиента и простейших векторных операций.

Однако эти методы требуют выполнения большого числа итераций для достижения заданной точности. В статье [1] рассматривается метод OGM-G повышенной эффективности. Но у этого метода, как и в целом у быстрых градиентных методов, есть существенная проблема: требуется заранее, до начала вычислений, определить количество итераций.

В пособии [2] предлагается способ оценки количества требуемых итераций, но он требует знания параметра сильной выпуклости μ. Также там указана идея эффективного для применения быстрого градиентного метода оценивания данного параметра. Реализации данной идеи и посвящена эта работа.

2 Обозначения и предположения

Решается задача минимизации:

$$\min_{x \in \mathrm{R}^d} f(x). \tag{1}$$

Предполагается, что решение

$$x^* = \arg\min_{x \in \mathrm{R}^d} f(x) \tag{2}$$

существует, а градиент функции $f(x)$ обладает свойством Липшица с константой $L > 0$:

$$||\nabla f(x) - \nabla f(y)|| \leq L||x - y|| \ \forall x, y \in \mathrm{R}^d. \tag{3}$$

Также считается, что функция $f(x)$ является сильно выпуклой с неизвестной нам константой $\mu > 0$:

$$\frac{\mu}{2}||x - x^*||^2 \leq f(x) - f(x^*) \leq \frac{1}{2\mu}||\nabla f(x)||^2. \tag{4}$$

Второе неравенство легко доказывается от противного с использованием определения сильной выпуклости.

3 Необходимые оценки

По теореме 2 из [1], при применении OGM-G

$$||\nabla f(x^N)||^2 \leq \frac{4L(f(x^0) - f(x^*))}{N^2}. \tag{5}$$

Оттуда же,

$$f(x^N) - f(x^*) \leq \frac{L||x^0 - x^*||^2}{N^2}. \tag{6}$$

Из (4) и (5):

$$||\nabla f(x^N)||^2 \leq \frac{4L}{N^2}\frac{1}{2\mu}||\nabla f(x^0)||^2, \tag{7}$$

или

$$||\nabla f(x^N)|| \leq \sqrt{\frac{2L}{\mu N^2}}||\nabla f(x^0)||. \tag{8}$$

Таким образом, если выполнить $N = 2\sqrt{2\frac{L}{\mu}}$ итераций, то норма градиента $f(x)$ уменьшится хотя бы вдвое.

4 Адаптивность по константе сильной выпуклости

Итак, в случае известной константы сильной выпуклости получена оценка числа итераций, гарантирующего уменьшение вдвое нормы градиента. В этом случае эффективное применение метода не представляет затруднений.

Однако предположение об известности константы μ практически никогда не выполняется в реальности. Оценить её, не затратив на это больше вычислительных ресурсов, чем на саму задачу оптимизации, не представляется возможным. Поэтому Ю. Е. Нестеровым предложен способ адаптивного по μ применения алгоритма, основанный на рестартах метода OGM-G.

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.

В результате очередное уменьшение вдвое нормы градиента будет выполнено за

$$2\sqrt{2\frac{L}{\mu_k^{init}}}+2\sqrt{2\frac{L}{\mu_k^{init}/2}}+\ldots+2\sqrt{2\frac{L}{\mu_k^{init}/2^m}} = \sqrt{8\frac{L}{\mu_k}}\sum_{i=0}^{m}\frac{1}{\sqrt{2^i}} \lesssim \frac{4}{\sqrt{2}-1}\sqrt{\frac{L}{\mu_k}}$$

итераций метода OGM-G, где m — количество повторений цикла на шаге k, а индекс $init$ указывает на то, что в формуле используется не конечное значение переменной, а то, которым она была инициализирована.

Оценим суммарное количество итераций. Если критерий останова выглядит как $||\nabla f(x)|| \leq \varepsilon$, то требуется выполнить $K = \log_2 \frac{||\nabla f(x^0)||}{\varepsilon}$ шагов. Каждый шаг содержит $O\left(\sqrt{\frac{L}{\mu}}\right)$ итераций, поэтому алгоритм завершит

работу, выполнив $O\left(\sqrt{\frac{L}{\mu}}\log_2\frac{||\nabla f(x^0)||}{\varepsilon}\right)$ итераций — то есть, вычислений $f(x)$ и $\nabla f(x)$.

5 Результаты

Предложен алгоритм, основанный на использовании быстрого градиентного метода и адаптивный по константе сильной выпуклости. Получена оценка количества обращений к вычислению функции и ее градиента при использовании построенного алгоритма.

Ссылки

[1] Optimizing the Efficiency of First-order Methods for Decreasing the Gradient of Smooth Convex Functions, Donghwan Kim, Jeffrey A. Fessler, 2018, 14 c., 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