Методы оптимизации, ВМК, осень 2018

Лекция 12: Быстрый градиентный метод

1 Введение

Рассмотрим простейшую гладкую задачу безусловной минимизации

$$\min_{x\in\mathbb{R}^n} f(x),$$

где $f : \mathbb{R}^n \to \mathbb{R}$ — дифференцируемая функция. Возможно, самым известным численным методом решения этой задачи является *градиентный метод*. Он начинает с некоторой точки $x_0 \in \mathbb{R}^n$ и рекуррентно строит последовательность $(x_k)_{k=0}^{\infty}$ точек в $\mathbb{R}^n$ по правилу

$$x_{k+1} := x_k - \alpha_k \nabla f(x_k)$$

для $k \geq 0$, где $\alpha_k \geq 0$. Известно, что если функция f выпуклая, а ее градиент ∇f липшицев, то градиентный метод сходится со скоростью $O(1/k)$.

Одним из первых прорывов в теории численных методов оптимизации было изобретение *быстрого градиентного метода* (БГМ, также называемого *ускоренным градиентным методом*) Ю.Е. Нестеровым в 1983 году. Основной особенностью БГМ является то, что он сходится со скоростью $O(1/k^2)$, в то время как сложность его итерации практически такая же, как и обычного градиентного метода. Более того, согласно теории сложности итерационных методов оптимизации, разработанной Немировским и Юдиным в конце 70-х годов, БГМ обладает *оптимальной* скоростью сходимости среди черноящичных методов первого порядка, т. е. никакой другой итерационный метод не может иметь более быструю скорость сходимости, при условии, что на каждой итерации методу разрешается только вычислять значение функции и ее градиента в любой точке.

2 Минимизация выпуклых функций

Будем рассматривать задачу минимизации

$$\min_{x\in\mathbb{R}^n} f(x),$$

где $f : \mathbb{R}^n \to \mathbb{R}$ — дифференцируемая функция.

В отличие от классического градиентного метода, который обычно мотивируется геометрическими соображениями, быстрый градиентный метод является результатом чисто алгебраических построений. Вместо того, чтобы выписать схему некоторого метода, и затем приложить большие усилия для того, чтобы доказать, что он сходится с определенной скоростью, начнем с оценки скорости сходимости, которую хочется получить, и затем придумаем метод, удовлетворяющий этой оценке.

Определение 2.1 (Оценочные последовательности)**.** Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — функция, и пусть $(x_k)_{k=0}^{\infty}$ — последовательность точек в $\mathbb{R}^n$. *Оценочными последовательностями* для $(x_k)_{k=0}^{\infty}$ относительно функции f называются пара последовательностей $(A_k)_{k=0}^{\infty}$ и $(\psi_k)_{k=0}^{\infty}$, где $(A_k)_{k=0}^{\infty}$ — последовательность неотрицательных вещественных чисел (называемых *шкалирующими коэффициентами*), удовлетворяющая $A_k > 0$ для всех $k \geq 1$, и $(\psi_k)_{k=0}^{\infty}$ — последовательность функций $\mathbb{R}^n$ to $\mathbb{R}$ (называемых *оценочными функциями*), таких, что

$$A_k f(x_k) \leq \psi_k(x) \leq A_k f(x) + \frac{1}{2}\|x - x_0\|^2. \tag{2.1}$$

для всех $x \in \mathbb{R}^n$ и всех $k \geq 0$.

Смысл оценочных последовательностей состоит в следующем. Предположим, что некоторый итеративный метод генерирует последовательность точек $(x_k)_{k=0}^{\infty}$, удовлетворяющих (2.1), причем шкалирующие коэффициенты неограниченно возрастают: $A_k \to \infty$ при $k \to \infty$. Тогда отсюда сразу же следует, что $(x_k)_{k=0}^{\infty}$ является минимизирующей последовательностью для функции f, и ее скорость сходимости определяется скоростью роста шкалирующих коэффициентов:

Лемма 2.2. *Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — функция, $(x_k)_{k=0}^{\infty}$ — последовательность точек в $\mathbb{R}^n$, $(A_k)_{k=0}^{\infty}$ и $(\psi_k)_{k=0}^{\infty}$ — оценочные последовательности для $(x_k)_{k=0}^{\infty}$ относительно f. Если f имеет минимум, тогда*

$$f(x_k) - \min f \le \frac{R^2}{2A_k}$$

для всех $k \ge 1$, где $R := \inf_{x \in \mathbb{R}^n : f(x) = \min f} \|x_0 - x\|$ — расстояние от начальной точки x_0 до множества минимумов f.

Доказательство. Пусть $k \ge 1$, и пусть $x^* \in \{x \in \mathbb{R}^n : f(x) = \min f\}$ — произвольное. Тогда, по определению оценочных последовательностей,

$$A_k f(x_k) \le \psi_k(x^*) \le A_k f(x^*) + \frac{1}{2}\|x_0 - x^*\|^2.$$

Учитывая, что $f(x^*) = \min f$ и переупорядочивая, получаем

$$f(x_k) - \min f \le \frac{1}{2A_k}\|x_0 - x^*\|^2.$$

Остается взять инфимум от обеих частей. □

Таким образом, чтобы построить метод со скоростью сходимости $O(1/k^2)$, нужно указать правила вычисления точек x_k и оценочные последовательности для $(x_k)_{k=0}^{\infty}$, такие, что $A_k = O(k^2)$.

Следующая лемма предоставляет систематический способ построения функций ψ_k и коэффициентов A_k, удовлетворяющих второму неравенству в (2.1):

Лемма 2.3. *Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — дифференцируемая выпуклая функция, и пусть $x_0 \in \mathbb{R}^n$. Пусть $(a_k)_{k=0}^{\infty}$ — последовательность положительных чисел, $(y_k)_{k=0}^{\infty}$ — последовательность точек в $\mathbb{R}^n$, и пусть последовательность $(A_k)_{k=0}^{\infty}$ неотрицательных чисел и последовательность $(\psi_k)_{k=0}^{\infty}$ функций из $\mathbb{R}^n$ в $\mathbb{R}$ определены рекурсивно по правилам*

$$A_0 := 0, \qquad \psi_0(x) := \frac{1}{2}\|x - x_0\|^2,$$

и

$$A_{k+1} := A_k + a_k, \qquad \psi_{k+1}(x) := \psi_k(x) + a_k[f(y_k) + \langle \nabla f(y_k), x - y_k\rangle]$$

для $k \ge 0$. Тогда $A_k > 0$ для всех $k \ge 1$ и

$$\psi_k(x) \le A_k f(x) + \frac{1}{2}\|x - x_0\|^2 \tag{2.2}$$

для всех $x \in \mathbb{R}^n$ и всех $k \ge 0$.

Доказательство. Неравенство $A_k > 0$ для всех $k \ge 1$ очевидно.

Пусть $x \in \mathbb{R}^n$. Докажем (2.2) индукцией по k. База индукции $k = 0$ тривиальна.

Теперь предположим индуктивно, что (2.2) справедливо для некоторого $k \ge 0$ и докажем

$$\psi_{k+1}(x) \le A_{k+1} f(x) + \frac{1}{2}\|x - x_0\|^2.$$

По определению,

$$\psi_{k+1}(x) = \psi_k(x) + a_k[f(y_k) + \nabla f(y_k), x - y_k\rangle].$$

Применяя индуктивное предположение, получаем

$$\psi_{k+1}(x) \le A_k f(x) + a_k[f(y_k) + \langle \nabla f(y_k), x - y_k\rangle] + \frac{1}{2}\|x - x_0\|^2.$$

Из дифференциального неравенства выпуклости, имеем

$$f(y_k) + \langle \nabla f(y_k), x - y_k \rangle \le f(x).$$

Поскольку a_k положительное, отсюда следует, что

$$a_k[f(y_k) + \langle \nabla f(y_k), x - y_k \rangle] \le a_k f(x).$$

Таким образом,

$$\psi_{k+1}(x) \le A_k f(x) + a_k f(x) + \frac{1}{2}\|x - x_0\|^2 = A_{k+1} f(x) + \frac{1}{2}\|x - x_0\|^2,$$

где последнее равенство следует из определения A_{k+1}. □

Предыдущая лемма является источником богатого параметрического семейства оценочных последовательностей: выбирая *произвольную* последовательность $(y_k)_{k=0}^{\infty}$ точек в $\mathbb{R}^n$ и *произвольную* последовательность $(a_k)_{k=0}^{\infty}$ положительных чисел, мы сразу же получаем последовательности $(A_k)_{k=0}^{\infty}$ и $(\psi_k)_{k=0}^{\infty}$, удовлетворяющие второму неравенству в (2.1). Используя эту свободу, становится возможным построить последовательность $(x_k)_{k=0}^{\infty}$, удовлетворяющую также и первому неравенству в (2.1).

Лемма 2.4. *Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — дифференцируемая выпуклая функция, $(a_k)_{k=0}^{\infty}$ — последовательность положительных чисел. Пусть $x_0 \in \mathbb{R}^n$ — произвольное, и пусть последовательности $(x_k)_{k=1}^{\infty}$, $(y_k)_{k=0}^{\infty}$, $(v_k)_{k=0}^{\infty}$ точек в $\mathbb{R}^n$, последовательность $(A_k)_{k=0}^{\infty}$ неотрицательных вещественных чисел, и последовательность $(\psi_k)_{k=0}^{\infty}$ функций из $\mathbb{R}^n$ в $\mathbb{R}$ определены рекурсивно по правилам*

$$\psi_0(x) := \frac{1}{2}\|x - x_0\|^2, \qquad v_0 := x_0, \qquad A_0 := 0,$$

и

$$A_{k+1} := A_k + a_k, \qquad y_k := \frac{A_k x_k + a_k v_k}{A_{k+1}},$$
$$\psi_{k+1}(x) := \psi_k(x) + a_k[f(y_k) + \langle \nabla f(y_k), x - y_k \rangle],$$
$$v_{k+1} := \operatorname{argmin} \psi_{k+1}, \qquad x_{k+1} := \frac{A_k x_k + a_k v_{k+1}}{A_{k+1}}$$

для $k \ge 0$. Тогда

$$\psi_{k+1}(x) \ge A_{k+1}\left[f(y_k) + \langle \nabla f(y_k), x_{k+1} - y_k \rangle + \frac{A_{k+1}}{2a_k^2}\|x_{k+1} - y_k\|^2\right] \tag{2.3}$$

для всех $x \in \mathbb{R}^n$ и всех $k \ge 0$.

Заметим, что для каждого $k \ge 0$, функция ψ_{k+1} строго выпукла как сумма строго выпуклой функции ψ_k и аффинной функции $x \mapsto a_k[f(y_k) + \langle \nabla f(y_k), x - y_k \rangle]$; поэтому точка v_{k+1} определена корректно (сильно выпуклая функция имеет, и при том единственный, минимум).

Доказательство. Заметим, что, по определению, $v_k = \operatorname{argmin} \psi_k$ для всех $k \ge 0$. Поэтому (2.3) эквивалентно

$$A_k f(x_k) \le \psi_k(v_k) \tag{2.4}$$

для всех $k \ge 0$. Действительно, импликация (2.3) $\Rightarrow$ (2.4) следует из подстановки $x = v_k$; обратная импликация (2.4) $\Rightarrow$ (2.3) следует из неравенства $\psi_k(v_k) \le \psi_k(x)$, которое является определением минимума.

Докажем (2.4) индукцией по k. База индукции $k = 0$ тривиальна. Теперь предположим индуктивно, что (2.4) выполнено для некоторого $k \ge 0$ и докажем, что

$$A_{k+1} f(x_{k+1}) \le \psi_{k+1}(v_{k+1}).$$

По определению ψ_{k+1}, имеем

$$\psi_{k+1}(v_{k+1}) = \psi_k(v_{k+1}) + a_k[f(y_k) + \langle \nabla f(y_k), v_{k+1} - y_k \rangle].$$

Заметим, что функция ψ_k квадратичная. Используя разложение Тейлора в точке v_k и учитывая, что $\nabla \psi_k(v_k) = 0$ (напомним, что v_k является минимумом ψ_k), получаем

$$\psi_k(v_{k+1}) = \psi_k(v_k) + \frac{1}{2}\|v_{k+1} - v_k\|^2.$$

Отсюда, по предположению индукции, имеем

$$\psi_k(v_{k+1}) \geq A_k f(x_k) + \frac{1}{2}\|v_{k+1} - v_k\|^2,$$

откуда

$$\psi_{k+1}(v_{k+1}) \geq A_k f(x_k) + a_k[f(y_k) + \langle \nabla f(y_k), v_{k+1} - y_k \rangle] + \frac{1}{2}\|v_{k+1} - v_k\|^2.$$

Используя дифференциальное неравенство выпуклости, получаем

$$f(x_k) \geq f(y_k) + \langle \nabla f(y_k), x_k - y_k \rangle.$$

Отсюда

$$\begin{aligned}\psi_{k+1}(v_{k+1}) &\geq A_k f(y_k) + A_k \langle \nabla f(y_k), x_k - y_k \rangle + a_k[f(y_k) + \langle \nabla f(y_k), v_{k+1} - y_k \rangle] + \frac{1}{2}\|v_{k+1} - v_k\|^2 \\ &= A_{k+1}\left[f(y_k) + \left\langle \nabla f(y_k), \frac{A_k x_k + a_k v_{k+1}}{A_{k+1}} - y_k \right\rangle + \frac{1}{2A_{k+1}}\|v_{k+1} - v_k\|^2\right].\end{aligned}$$

Остается заметить, что

$$v_{k+1} - v_k = \frac{A_{k+1}}{a_k}(x_{k+1} - y_k),$$

и подставить эту в предыдущую формулу. □

Используя квадратичную оценку сверху на функции, следующую из липшицевости градиента, получаем

Следствие 2.5. *Если* $La_k^2 \leq A_{k+1}$, *то*

$$\psi_{k+1}(v_{k+1}) \geq A_{k+1} f(x_{k+1}).$$

Предыдущая лемма дает некоторый итеративный метод, строящий последовательность $(x_k)_{k=0}^{\infty}$ точек в $\mathbb{R}^n$. Если условие (**??**) выполнено, тогда $(A_k)_{k=0}^{\infty}$ и $(\psi_k)_{k=0}^{\infty}$ являются оценочными последовательностями для $(x_k)_{k=0}^{\infty}$, и, по лемме 2.2, шкалирующие коэффициенты A_k определяют скорость сходимости этого метода. Таким образом, чтобы завершить построение, остается обеспечить (**??**). К счастью, коэффициенты a_k до сих пор не выбраны. Воспользуемся этой последней степенью свободы, чтобы обеспечить (**??**).

Пусть $k \geq 0$. Чтобы обеспечить (**??**), выберем $a_k > 0$ из квадратного уравнения

$$La_k^2 = A_k + a_k,$$

т. е.

$$a_k = \frac{1 + \sqrt{1 + 4LA_k}}{2L}.$$

Мы практически закончили. Последнее, что осталось сделать — оценить, насколько быстро растут шкалирующие коэффициенты A_k.

Лемма 2.6. *Пусть $L > 0$, и пусть $(a_k)_{k=0}^{\infty}$ и $(A_k)_{k=0}^{\infty}$ — последовательности неотрицательных чисел, определенные рекуррентно по правилу $A_0 := 0$ и*

$$a_k := \frac{1 + \sqrt{1 + 4LA_k}}{2L}, \qquad A_{k+1} := A_k + a_k,$$

для $k \geq 0$. Тогда $La_k^2 = A_{k+1}$ и

$$A_k \geq \frac{k^2}{4L}$$

для всех $k \geq 0$.

Доказательство. Утверждение $La_k^2 = A_{k+1}$ для всех $k \geq 0$ легко проверяется непосредственно.

Для доказательства неравенства подставим в это тождество $a_k = A_{k+1} - A_k$ и получим

$$A_{k+1} = L(A_{k+1} - A_k)^2 = L\left(\sqrt{A_{k+1}} - \sqrt{A_k}\right)^2 \left(\sqrt{A}_{k+1} + \sqrt{A_k}\right)^2$$

для всех $k \geq 0$. Поскольку $A_k \leq A_{k+1}$ для всех $k \geq 0$, тогда

$$A_{k+1} \leq 4LA_{k+1}\left(\sqrt{A_{k+1}} - \sqrt{A_k}\right)^2$$

для всех $k \geq 0$. Выполняя алгебраические преобразования, получаем

$$\sqrt{A_{k+1}} \geq \sqrt{A_k} + \frac{1}{2\sqrt{L}}$$

для всех $k \geq 0$, откуда, по индукции,

$$\sqrt{A}_k \geq \frac{k}{2\sqrt{L}}$$

для всех $k \geq 0$. Остается возвести обе части в квадрат. □

Резюмируем полученный метод вместе с оценкой его скорости сходимости.

Быстрый градиентный метод (БГМ)

Вход: дифференцируемая функция $f : \mathbb{R}^n \to \mathbb{R}$, начальная точка $x_0 \in \mathbb{R}^n$, константа Липшица $L > 0$ для ∇f.

1 Положить $A_0 := a_0$ и $v_0 := x_0$.

2 Итерация $k \geq 0$:

(a) Положить $a_k := \frac{1+\sqrt{1+4LA_k}}{2L}$ и $A_{k+1} := A_k + a_k$.

(b) Положить $y_k := \frac{A_k x_k + a_k v_k}{A_{k+1}}$.

(c) Вычислить $v_{k+1} := \operatorname{argmin}_{x \in \mathbb{R}^n}\{\frac{1}{2}\|x - x_0\|^2 + \sum_{i=0}^{k} a_i[f(y_i) + \langle \nabla f(y_i), x - y_i\rangle]\}$.

(d) Положить $x_{k+1} := \frac{A_k x_k + a_k v_{k+1}}{A_{k+1}}$.

Замечание 2.7. Дифференцируя, можно получить следующую явную формулу для v_{k+1}:

$$v_{k+1} = x_0 - \sum_{i=0}^{k} a_i \nabla f(y_i).$$

Другими словами, шаг 3 приведенного метода может быть явно записан следующим образом:

$$v_{k+1} = v_k - a_k \nabla f(y_k).$$

Теорема 2.8. *Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — дифференцируемая выпуклая функция, обладающая липшицевым градиентом с параметром $L > 0$, и пусть $x_0 \in \mathbb{R}^n$. Если f имеет минимум, то БГМ строит последовательность точек $(x_k)_{k=1}^{\infty}$ в $\mathbb{R}^n$, такую, что*

$$f(x_k) - \min f \leq \frac{2LR^2}{k^2}$$

для всех $k \geq 1$, где $R := \inf_{x \in \mathbb{R}^n : f(x) = \min f} \|x_0 - x\|$ — расстояние от начальной точки x_0 до множества минимумов f.

Доказательство. Согласно лемме 2.4, лемме 2.3 и лемме 2.6, для $(x_k)_{k=0}^{\infty}$ существуют оценочные последовательности $(A_k)_{k=0}^{\infty}$ и $(\psi_k)_{k=0}^{\infty}$) относительно f, такие, что $A_k \geq \frac{k^2}{4L}$ для всех $k \geq 1$. Остается применить лемму 2.2. □

3 Адаптивная оценка константы Липшица

Адаптивный быстрый градиентный метод (БГМ II)

Вход: дифференцируемая функция $f : \mathbb{R}^n \to \mathbb{R}$, начальная точка $x_0 \in \mathbb{R}^n$, начальная оценка константа Липшица $L_0 > 0$ для ∇f.

1 Положить $A_0 := a_0$ и $v_0 := x_0$.

2 Итерация $k \geq 0$:

(a) Положить $\bar{L}_k := L_k$.

(b) Положить $a_k := \frac{1+\sqrt{1+4\bar{L}_k A_k}}{2\bar{L}_k}$ и $A_{k+1} := A_k + a_k$.

(c) Положить $y_k := \frac{A_k x_k + a_k v_k}{A_{k+1}}$.

(d) Вычислить $v_{k+1} := \operatorname{argmin}_{x \in \mathbb{R}^n} \{\frac{1}{2}\|x - x_0\|^2 + \sum_{i=0}^{k} a_i [f(y_i) + \langle \nabla f(y_i), x - y_i \rangle]\}$.

(e) Положить $x_{k+1} := \frac{A_k x_k + a_k v_{k+1}}{A_{k+1}}$.

(f) Если $f(x_{k+1}) > f(y_k) + \langle \nabla f(y_k), x_{k+1} - y_k \rangle + \frac{\bar{L}_k}{2}\|x_{k+1} - y_k\|^2$, положить $\bar{L}_k := 2\bar{L}_k$ и вернуться к шагу 2(b).

(g) Положить $L_{k+1} := \bar{L}_k / 2$.

Теорема 3.1. *Пусть градиент ∇f липшицев с параметром $L > 0$. Тогда суммарное число выполнений внутреннего цикла не превосходит*

$$N(k) \leq 2(k+1) + 2\log_2 \frac{L}{L_0}.$$

Доказательство. Пусть $n_k \geq 1$ — число выполнений внутреннего цикла подбора L_k на итерации $k \geq 0$. По построению

$$L_{k+1} = \frac{1}{2} 2^{n_k - 1} L_k = 2^{n_k - 2} L_k.$$

Отсюда

$$n_k = 2 + \log_2 L_{k+1} - \log_2 L_k.$$

Таким образом, суммарное число выполнений внутреннего цикла

$$N(k) = \sum_{i=0}^{k} n_i = 2(k+1) + \log_2 \frac{L_{k+1}}{L_0}.$$

Остается заметить, что по построению $L_k \leq L$ для всех $k \geq 0$. □

Поскольку БГМ на каждой итерации внутреннего цикла делает два вызова оракула (в y_k и x_{k+1}), то в среднем приходится 4 вызова оракула на итерацию.

Скорость сходимости: оцениваем

$$A_{k+1} = \bar{L}_k a_k^2 \leq 2La_k^2.$$

Дальше применяем прежние рассуждения. В итоге в предыдущей оценке константа L в два раза увеличивается.

TODO: Добавить про сильно выпуклые функции, условную и композитную оптимизацию.