Градиентный спуск. Скорости сходимости

Семинар

Оптимизация для всех! ЦУ

Градиентный спуск

Предположим, у нас есть задача минимизации гладкой функции $f(x) : \mathbb{R}^n \to \mathbb{R}$:

$$f(x) \to \min_{x \in \mathbb{R}}$$

Градиентный спуск

Предположим, у нас есть задача минимизации гладкой функции $f(x) : \mathbb{R}^n \to \mathbb{R}$:

$$f(x) \to \min_{x \in \mathbb{R}}$$

Один из методов решения этой задачи — **градиентный спуск**:

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

Градиентный спуск

Предположим, у нас есть задача минимизации гладкой функции $f(x) : \mathbb{R}^n \to \mathbb{R}$:

$$f(x) \to \min_{x \in \mathbb{R}}$$

Один из методов решения этой задачи — **градиентный спуск**:

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

Бутылочным горлышком (почти для всех градиентных методов) является выбор шага, который может привести к драматическим различиямв поведении метода.

Как выбрать шаг

- Одно из теоретических предложений: выбирать шаг, обратно пропорциональный константе Липшица градиента

$$\eta_k = \frac{1}{L}$$

Как выбрать шаг

- Одно из теоретических предложений: выбирать шаг, обратно пропорциональный константе Липшица градиента

$$\eta_k = \frac{1}{L}$$

- **Линейный поиск с возвратом**. Зафиксируем два параметра: $0 < \beta < 1$ и $0 < \alpha \leq 0.5$. На каждой итерации начинаем с $\mathrm{t} = 1$, и пока

$$f(x_k - t\nabla f(x_k)) > f(x_k) - \alpha t \|\nabla f(x_k)\|_2^2,$$

уменьшаем $t = \beta t$. Иначе выполняем обновление градиентного спуска $x_{k+1} = x_k - t\nabla f(x_k)$.

Как выбрать шаг

- Одно из теоретических предложений: выбирать шаг, обратно пропорциональный константе Липшица градиента

$$\eta_k = \frac{1}{L}$$

- **Линейный поиск с возвратом**. Зафиксируем два параметра: $0 < \beta < 1$ и $0 < \alpha \leq 0.5$. На каждой итерации начинаем с $\mathrm{t} = 1$, и пока

$$f(x_k - t\nabla f(x_k)) > f(x_k) - \alpha t \|\nabla f(x_k)\|_2^2,$$

уменьшаем $t = \beta t$. Иначе выполняем обновление градиентного спуска $x_{k+1} = x_k - t\nabla f(x_k)$.

- **Точный линейный поиск**.

$$\eta_k = \arg\min_{\eta \geq 0} f(x_k - \eta \nabla f(x_k))$$

Направление локального наискорейшего спуска

Рассмотрим линейное приближение дифференцируемой функции f вдоль некоторого направления $h, \|h\|_2 = 1$:

Направление локального наискорейшего спуска

Рассмотрим линейное приближение дифференцируемой функции f вдоль некоторого направления $h, \|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle f'(x), h \rangle + o(\alpha)$$

Направление локального наискорейшего спуска

Рассмотрим линейное приближение дифференцируемой функции f вдоль некоторого направления $h, \|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle f'(x), h \rangle + o(\alpha)$$

Мы хотим, чтобы h было убывающим направлением:

$$f(x + \alpha h) < f(x)$$

$$f(x) + \alpha \langle f'(x), h \rangle + o(\alpha) < f(x)$$

Направление локального наискорейшего спуска

Рассмотрим линейное приближение дифференцируемой функции f вдоль некоторого направления $h, \|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle f'(x), h \rangle + o(\alpha)$$

Мы хотим, чтобы h было убывающим направлением:

$$f(x + \alpha h) < f(x)$$

$$f(x) + \alpha \langle f'(x), h \rangle + o(\alpha) < f(x)$$

и переходя к пределу при $\alpha \to 0$:

$$\langle f'(x), h \rangle \leq 0$$

Направление локального наискорейшего спуска

Рассмотрим линейное приближение дифференцируемой функции f вдоль некоторого направления $h, \|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle f'(x), h \rangle + o(\alpha)$$

Мы хотим, чтобы h было убывающим направлением:

$$f(x + \alpha h) < f(x)$$

$$f(x) + \alpha \langle f'(x), h \rangle + o(\alpha) < f(x)$$

и переходя к пределу при $\alpha \to 0$:

$$\langle f'(x), h \rangle \leq 0$$

Также из неравенства Коши — Буняковского — Шварца:

$$|\langle f'(x), h \rangle| \leq \|f'(x)\|_2 \|h\|_2$$
$$\langle f'(x), h \rangle \geq -\|f'(x)\|_2 \|h\|_2 = -\|f'(x)\|_2$$

Направление локального наискорейшего спуска

Рассмотрим линейное приближение дифференцируемой функции f вдоль некоторого направления $h, \|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle f'(x), h \rangle + o(\alpha)$$

Мы хотим, чтобы h было убывающим направлением:

$$f(x + \alpha h) < f(x)$$

$$f(x) + \alpha \langle f'(x), h \rangle + o(\alpha) < f(x)$$

и переходя к пределу при $\alpha \to 0$:

$$\langle f'(x), h \rangle \leq 0$$

Также из неравенства Коши — Буняковского — Шварца:

$$|\langle f'(x), h \rangle| \leq \|f'(x)\|_2 \|h\|_2$$
$$\langle f'(x), h \rangle \geq -\|f'(x)\|_2 \|h\|_2 = -\|f'(x)\|_2$$

Таким образом, направление антиградиента

$$h = -\frac{f'(x)}{\|f'(x)\|_2}$$

дает направление **локального наискорейшего** убывания функции f.

Направление локального наискорейшего спуска

Рассмотрим линейное приближение дифференцируемой функции f вдоль некоторого направления $h, \|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle f'(x), h \rangle + o(\alpha)$$

Мы хотим, чтобы h было убывающим направлением:

$$f(x + \alpha h) < f(x)$$

$$f(x) + \alpha \langle f'(x), h \rangle + o(\alpha) < f(x)$$

и переходя к пределу при $\alpha \to 0$:

$$\langle f'(x), h \rangle \leq 0$$

Также из неравенства Коши — Буняковского — Шварца:

$$|\langle f'(x), h \rangle| \leq \|f'(x)\|_2 \|h\|_2$$
$$\langle f'(x), h \rangle \geq -\|f'(x)\|_2 \|h\|_2 = -\|f'(x)\|_2$$

Таким образом, направление антиградиента

$$h = -\frac{f'(x)}{\|f'(x)\|_2}$$

дает направление **локального наискорейшего** убывания функции f.
Результатом этого метода является

$$x_{k+1} = x_k - \alpha f'(x_k)$$

Точка минимума липшицевой параболы

Если функция $f : \mathbb{R}^n \to \mathbb{R}$ непрерывно дифференцируема и ее градиент удовлетворяет условиям Липшица с константой L, то $\forall x, y \in \mathbb{R}^n$:

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2} \|y - x\|^2,$$

Точка минимума липшицевой параболы

Если функция $f : \mathbb{R}^n \to \mathbb{R}$ непрерывно дифференцируема и ее градиент удовлетворяет условиям Липшица с константой L, то $\forall x, y \in \mathbb{R}^n$:

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2}\|y - x\|^2,$$

что геометрически означает, что если мы зафиксируем некоторую точку $x_0 \in \mathbb{R}^n$ и определим две параболы:

$$\phi_1(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle - \frac{L}{2}\|x - x_0\|^2,$$
$$\phi_2(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{L}{2}\|x - x_0\|^2.$$

Точка минимума липшицевой параболы

Если функция $f : \mathbb{R}^n \to \mathbb{R}$ непрерывно дифференцируема и ее градиент удовлетворяет условиям Липшица с константой L, то $\forall x, y \in \mathbb{R}^n$:

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2}\|y - x\|^2,$$

что геометрически означает, что если мы зафиксируем некоторую точку $x_0 \in \mathbb{R}^n$ и определим две параболы:

$$\phi_1(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle - \frac{L}{2}\|x - x_0\|^2,$$
$$\phi_2(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{L}{2}\|x - x_0\|^2.$$

Тогда

$$\phi_1(x) \leq f(x) \leq \phi_2(x) \quad \forall x \in \mathbb{R}^n.$$

Точка минимума липшицевой параболы

Если функция $f : \mathbb{R}^n \to \mathbb{R}$ непрерывно дифференцируема и ее градиент удовлетворяет условиям Липшица с константой L, то $\forall x, y \in \mathbb{R}^n$:

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2}\|y - x\|^2,$$

что геометрически означает, что если мы зафиксируем некоторую точку $x_0 \in \mathbb{R}^n$ и определим две параболы:

$$\phi_1(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle - \frac{L}{2}\|x - x_0\|^2,$$
$$\phi_2(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{L}{2}\|x - x_0\|^2.$$

Тогда

$$\phi_1(x) \leq f(x) \leq \phi_2(x) \quad \forall x \in \mathbb{R}^n.$$

Теперь, если у нас есть глобальная верхняя граница функции в виде параболы, мы можем попробовать сразу попасть в ее минимум.

Точка минимума липшицевой параболы

Если функция $f : \mathbb{R}^n \to \mathbb{R}$ непрерывно дифференцируема и ее градиент удовлетворяет условиям Липшица с константой L, то $\forall x, y \in \mathbb{R}^n$:

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2}\|y - x\|^2,$$

что геометрически означает, что если мы зафиксируем некоторую точку $x_0 \in \mathbb{R}^n$ и определим две параболы:

$$\phi_1(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle - \frac{L}{2}\|x - x_0\|^2,$$
$$\phi_2(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{L}{2}\|x - x_0\|^2.$$

Тогда

$$\phi_1(x) \leq f(x) \leq \phi_2(x) \quad \forall x \in \mathbb{R}^n.$$

Теперь, если у нас есть глобальная верхняя граница функции в виде параболы, мы можем попробовать сразу попасть в ее минимум.

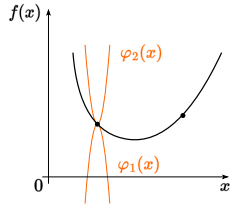

Рис. 1: Иллюстрация

Точка минимума липшицевой параболы

Если функция $f : \mathbb{R}^n \to \mathbb{R}$ непрерывно дифференцируема и ее градиент удовлетворяет условиям Липшица с константой L, то $\forall x, y \in \mathbb{R}^n$:

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2}\|y - x\|^2,$$

что геометрически означает, что если мы зафиксируем некоторую точку $x_0 \in \mathbb{R}^n$ и определим две параболы:

$$\phi_1(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle - \frac{L}{2}\|x - x_0\|^2,$$

$$\phi_2(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{L}{2}\|x - x_0\|^2.$$

Тогда

$$\phi_1(x) \leq f(x) \leq \phi_2(x) \quad \forall x \in \mathbb{R}^n.$$

Теперь, если у нас есть глобальная верхняя граница функции в виде параболы, мы можем попробовать сразу попасть в ее минимум.

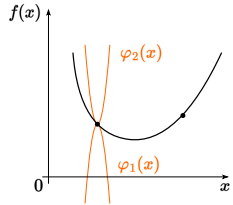

Рис. 1: Иллюстрация

$$\nabla \phi_2(x) = 0$$

$$\nabla f(x_0) + L(x^* - x_0) = 0$$

$$x^* = x_0 - \frac{1}{L}\nabla f(x_0)$$

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$$

Таким образом, мы получаем шаг $\frac{1}{L}$. Однако часто константа L неизвестна.

Любая μ-строго выпуклая дифференцируемая функция является PL-функцией

Theorem

Если функция $f(x)$ дифференцируема и μ-строго выпукла, то она является PL-функцией.

Доказательство

По критерию строгой выпуклости первого порядка:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|y - x\|_2^2$$

Положим $y = x^*$:

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\mu}{2}\|x^* - x\|_2^2$$

Любая μ-строго выпуклая дифференцируемая функция является PL-функцией

Theorem

Если функция $f(x)$ дифференцируема и μ-строго выпукла, то она является PL-функцией.

Доказательство

По критерию строгой выпуклости первого порядка:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|y - x\|_2^2$$

Положим $y = x^*$:

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\mu}{2}\|x^* - x\|_2^2$$

$$f(x) - f(x^*) \leq \nabla f(x)^T (x - x^*) - \frac{\mu}{2}\|x^* - x\|_2^2 =$$

Любая μ-строго выпуклая дифференцируемая функция является PL-функцией

Theorem

Если функция $f(x)$ дифференцируема и μ-строго выпукла, то она является PL-функцией.

Доказательство

По критерию строгой выпуклости первого порядка:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|y - x\|_2^2$$

Положим $y = x^*$:

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\mu}{2}\|x^* - x\|_2^2$$

$$f(x) - f(x^*) \leq \nabla f(x)^T (x - x^*) - \frac{\mu}{2}\|x^* - x\|_2^2 =$$

$$= \left(\nabla f(x)^T - \frac{\mu}{2}(x^* - x)\right)^T (x - x^*) =$$

Любая μ-строго выпуклая дифференцируемая функция является PL-функцией

Theorem

Если функция $f(x)$ дифференцируема и μ-строго выпукла, то она является PL-функцией.

Доказательство

По критерию строгой выпуклости первого порядка:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|y - x\|_2^2$$

Положим $y = x^*$:

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\mu}{2}\|x^* - x\|_2^2$$

$$f(x) - f(x^*) \leq \nabla f(x)^T (x - x^*) - \frac{\mu}{2}\|x^* - x\|_2^2 =$$

$$= \left(\nabla f(x)^T - \frac{\mu}{2}(x^* - x)\right)^T (x - x^*) =$$

$$= \frac{1}{2}\left(\frac{2}{\sqrt{\mu}}\nabla f(x)^T - \sqrt{\mu}(x^* - x)\right)^T \sqrt{\mu}(x - x^*)$$

Любая μ-строго выпуклая дифференцируемая функция является PL-функцией

Theorem

Если функция $f(x)$ дифференцируема и μ-строго выпукла, то она является PL-функцией.

Доказательство

По критерию строгой выпуклости первого порядка:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|y - x\|_2^2$$

Положим $y = x^*$:

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\mu}{2}\|x^* - x\|_2^2$$

$$f(x) - f(x^*) \leq \nabla f(x)^T (x - x^*) - \frac{\mu}{2}\|x^* - x\|_2^2 =$$

$$= \left(\nabla f(x)^T - \frac{\mu}{2}(x^* - x)\right)^T (x - x^*) =$$

$$= \frac{1}{2}\left(\frac{2}{\sqrt{\mu}}\nabla f(x)^T - \sqrt{\mu}(x^* - x)\right)^T \sqrt{\mu}(x - x^*)$$

Любая μ-строго выпуклая дифференцируемая функция является PL-функцией

Theorem

Если функция $f(x)$ дифференцируема и μ-строго выпукла, то она является PL-функцией.

Доказательство

По критерию строгой выпуклости первого порядка:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|y - x\|_2^2$$

Положим $y = x^*$:

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\mu}{2}\|x^* - x\|_2^2$$

$$f(x) - f(x^*) \leq \nabla f(x)^T (x - x^*) - \frac{\mu}{2}\|x^* - x\|_2^2 =$$

$$= \left(\nabla f(x)^T - \frac{\mu}{2}(x^* - x)\right)^T (x - x^*) =$$

$$= \frac{1}{2}\left(\frac{2}{\sqrt{\mu}}\nabla f(x)^T - \sqrt{\mu}(x^* - x)\right)^T \sqrt{\mu}(x - x^*)$$

Пусть $a = \frac{1}{\sqrt{\mu}}\nabla f(x) \setminus$ и
$b = \sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}}\nabla f(x)$

Любая μ-строго выпуклая дифференцируемая функция является PL-функцией

Theorem

Если функция $f(x)$ дифференцируема и μ-строго выпукла, то она является PL-функцией.

Доказательство

По критерию строгой выпуклости первого порядка:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|y - x\|_2^2$$

Положим $y = x^*$:

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\mu}{2}\|x^* - x\|_2^2$$

$$f(x) - f(x^*) \leq \nabla f(x)^T (x - x^*) - \frac{\mu}{2}\|x^* - x\|_2^2 =$$

$$= \left(\nabla f(x)^T - \frac{\mu}{2}(x^* - x)\right)^T (x - x^*) =$$

$$= \frac{1}{2}\left(\frac{2}{\sqrt{\mu}}\nabla f(x)^T - \sqrt{\mu}(x^* - x)\right)^T \sqrt{\mu}(x - x^*)$$

Пусть $a = \frac{1}{\sqrt{\mu}}\nabla f(x)$ \ и
$b = \sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}}\nabla f(x)$
Тогда $a + b = \sqrt{\mu}(x - x^*)$ и
$a - b = \frac{2}{\sqrt{\mu}}\nabla f(x) - \sqrt{\mu}(x - x^*)$

Любая μ-строго выпуклая дифференцируемая функция является PL-функцией

$$f(x) - f(x^*) \leq \frac{1}{2}\left(\frac{1}{\mu}\|\nabla f(x)\|_2^2 - \left\|\sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}}\nabla f(x)\right\|_2^2\right)$$

Любая μ-строго выпуклая дифференцируемая функция является PL-функцией

$$f(x) - f(x^*) \leq \frac{1}{2}\left(\frac{1}{\mu}\|\nabla f(x)\|_2^2 - \left\|\sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}}\nabla f(x)\right\|_2^2\right)$$

$$f(x) - f(x^*) \leq \frac{1}{2\mu}\|\nabla f(x)\|_2^2,$$

Любая μ-строго выпуклая дифференцируемая функция является PL-функцией

$$f(x) - f(x^*) \leq \frac{1}{2}\left(\frac{1}{\mu}\|\nabla f(x)\|_2^2 - \left\|\sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}}\nabla f(x)\right\|_2^2\right)$$

$$f(x) - f(x^*) \leq \frac{1}{2\mu}\|\nabla f(x)\|_2^2,$$

Любая μ-строго выпуклая дифференцируемая функция является PL-функцией

$$f(x) - f(x^*) \leq \frac{1}{2}\left(\frac{1}{\mu}\|\nabla f(x)\|_2^2 - \left\|\sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}}\nabla f(x)\right\|_2^2\right)$$

$$f(x) - f(x^*) \leq \frac{1}{2\mu}\|\nabla f(x)\|_2^2,$$

что точно соответствует условию PL. Это означает, что мы уже имеем доказательство линейной сходимости для любой строго выпуклой функции.

Точный линейный поиск, или наискорейший спуск

$$\alpha_k = \arg\min_{\alpha \in \mathbb{R}^+} f(x_{k+1}) = \arg\min_{\alpha \in \mathbb{R}^+} f(x_k - \alpha \nabla f(x_k))$$

Более теоретический, чем практический подход. Он также позволяет анализировать сходимость, но часто точный линейный поиск может быть сложным, если вычисление функции занимает слишком много времени или стоит слишком дорого.
Интересное теоретическое свойство этого метода заключается в том, что каждая следующая итерация ортогональна предыдущей:

$$\alpha_k = \arg\min_{\alpha \in \mathbb{R}^+} f(x_k - \alpha \nabla f(x_k))$$

Условия оптимальности:

$$\nabla f(x_{k+1})^\top \nabla f(x_k) = 0$$

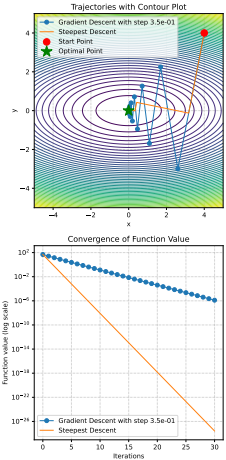

Рис. 2: Наискорейший спуск

Открыть в Colab ♣

Анализ сходимости. Линейный поиск с возвратом

Предположим, что f выпукла, дифференцируема и имеет константу Липшица $L > 0$.

Theorem

Градиентный спуск с фиксированным шагом $t \leq 1/L$ удовлетворяет

$$f\left(x^{(k)}\right) - f^* \leq \frac{\left\|x^{(0)} - x^*\right\|_2^2}{2tk}$$

Анализ сходимости. Линейный поиск с возвратом

Предположим, что f выпукла, дифференцируема и имеет константу Липшица $L > 0$.

Theorem

Градиентный спуск с фиксированным шагом $t \leq 1/L$ удовлетворяет

$$f\left(x^{(k)}\right) - f^* \leq \frac{\left\|x^{(0)} - x^*\right\|_2^2}{2tk}$$

Покажем, что скорость сходимости для линейного поиска с возвратом не хуже, чем $O(1/k)$

Анализ сходимости. Линейный поиск с возвратом

Предположим, что f выпукла, дифференцируема и имеет константу Липшица $L > 0$.

Theorem

Градиентный спуск с фиксированным шагом $t \leq 1/L$ удовлетворяет

$$f\left(x^{(k)}\right) - f^* \leq \frac{\left\|x^{(0)} - x^*\right\|_2^2}{2tk}$$

Покажем, что скорость сходимости для линейного поиска с возвратом не хуже, чем $O(1/k)$

Поскольку ∇f непрерывно дифференцируема с константой Липшица $L > 0$, мы имеем

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\|y - x\|_2^2, \forall x, y$$

Анализ сходимости. Линейный поиск с возвратом

Предположим, что f выпукла, дифференцируема и имеет константу Липшица $L > 0$.

Theorem

Градиентный спуск с фиксированным шагом $t \leq 1/L$ удовлетворяет

$$f\left(x^{(k)}\right) - f^* \leq \frac{\left\|x^{(0)} - x^*\right\|_2^2}{2tk}$$

Покажем, что скорость сходимости для линейного поиска с возвратом не хуже, чем $O(1/k)$

Поскольку ∇f непрерывно дифференцируема с константой Липшица $L > 0$, мы имеем

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\|y - x\|_2^2, \forall x, y$$

Пусть $y = x^+ = x - t\nabla f(x)$, тогда:

$$f\left(x^+\right) \leq f(x) - \left(1 - \frac{Lt}{2}\right) t\|\nabla f(x)\|_2^2 \leq f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2$$

Это соответствует условию остановки линейного поиска с возвратом при $\alpha = 0.5, t = \frac{1}{L}$. Следовательно, при липшицевом градиенте такой поиск гарантирует скорость сходимости $O(1/k)$.

Задача

Рассмотрим задачу

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

где $f(x)$ выпукла и L-гладкая. Найдите скорость сходимости градиентного спуска с оптимальным теоретическим шагом $\eta_k = \frac{1}{L}$ для *усредненной точки* и для *лучшей точки*. Другими словами, получите верхние границы на

- $f(\bar{x}_N) - f^*$, where $\bar{x}_N = \frac{1}{N} \sum_{i=0}^{N-1} x_i$,
- $\min_{0 \le i \le N-1} f(x_i) - f^*$.

Шаг градиентного спуска

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ \Psi_k(x) \equiv f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2} \|x - x_k\|_2^2 \right\}$$

Задача

Рассмотрим задачу

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

где $f(x)$ выпукла и L-гладкая. Найдите скорость сходимости градиентного спуска с оптимальным теоретическим шагом $\eta_k = \frac{1}{L}$ для *усредненной точки* и для *лучшей точки*. Другими словами, получите верхние границы на

- $f(\bar{x}_N) - f^*$, where $\bar{x}_N = \frac{1}{N} \sum_{i=0}^{N-1} x_i$,
- $\min_{0 \le i \le N-1} f(x_i) - f^*$.

Шаг градиентного спуска

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ \Psi_k(x) \equiv f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2} \|x - x_k\|_2^2 \right\}$$

Совет

Используйте факт, что $\Psi_k(x)$ является L-строго выпуклой из-за квадратичного регуляризатора.

Код

Примеры: code snippet.