Основні означення: Optimization Theory

Vladyslav Skrynyk, Olena Mikhailova

19 березня 2025 р.

1 Основні означення

1.1 Опуклість (Convexity)

Функція $f : \mathbb{R}^n \to \mathbb{R}$ називається **опуклою**, якщо для всіх $x, y \in \mathbb{R}^n$ і будь-якого $\lambda \in [0, 1]$ виконується:

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y). \tag{1}$$

1.2 Сильна опуклість (Strong Convexity)

Якщо в (1) виконується строга нерівність.

1.3 L-smoothness

Функція $f : \Omega \to \mathbb{R}$ є L-гладкою, якщо її градієнт є L-Ліпшицевим, тобто виконується нерівність

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y \in \Omega.$$

градієнт не змінюється надто критично

1.4 Лема спуску (Descent Lemma)

Нехай $f : \Omega \to \mathbb{R}$ є L-гладкою на опуклій області Ω. Тоді функцію f можна оцінити зверху наступним чином:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2, \quad \forall x, y \in \Omega. \quad (2)$$

Доведення. Виразимо приріст $f(y) - f(x)$ як інтеграл від градієнта уздовж прямої, що з'єднує x та y, а потім використаємо Ліпшицеву умову на градієнт ∇f для оцінки приросту:

$$f(y) - f(x) = \int_0^1 \langle \nabla f(x + t \cdot (y - x)), y - x \rangle \, dt$$

$$= \left(\int_0^1 \langle \nabla f(x + t \cdot (y - x)) - \nabla f(x), y - x \rangle \, dt \right) + \langle \nabla f(x), y - x \rangle$$

$$\leq \left(\int_0^1 \|\nabla f(x + t \cdot (y - x)) - \nabla f(x)\|_2 \cdot \|y - x\|_2 \, dt \right) + \langle \nabla f(x), y - x \rangle$$

$$\leq \left(\int_0^1 tL\|y - x\|_2^2 \, dt \right) + \langle \nabla f(x), y - x \rangle$$

$$= \frac{L}{2}\|y-x\|_2^2 + \langle \nabla f(x), y-x\rangle.$$

Перегрупувавши вирази, отримуємо твердження теореми.

Теорема (Лема про градієнтний спуск, ДОДАТКОВА). Нехай $f : \mathbb{R}^n \to \mathbb{R}$ є L-гладкою. Тоді для будь-якого $0 < \eta \leq \frac{1}{L}$ кожен крок градієнтного спуску (1) гарантує:

$$f(x_{t+1}) \leq f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|_2^2.$$

1.5 Розмір кроку (Stepsize, Learning rate)

Розмір кроку α_k в ітераційному методі визначає, наскільки далеко ми рухаємося в напрямку градієнта:

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

1.6 Градієнтний спуск (Gradient Descent)

Метод градієнтного спуску визначається рекурентним співвідношенням:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \tag{2}$$

де α_k — це розмір кроку. Це ітераційний алгоритм оптимізації першого порядку, в якому для знаходження локального(глобального) мінімуму функції здійснюються кроки, пропорційні протилежному значенню градієнту функції в поточній точці.

1.7 Пошук вздовж прямої (Linesearch)

Метод визначення оптимального розміру кроку α_k шляхом розв'язання задачі:

$$\alpha_k = \arg\min_{\alpha > 0} f(x_k - \alpha \nabla f(x_k)). \tag{3}$$

1.7.1 Exact Line-Search

Після вибору напрямку (у градієнтному спуску це напрямок негативного градієнта) можна розглянути таку одномірну задачу оптимізації для визначення найкращого розміру кроку:

$$\eta_t = \arg\min_{\eta \geq 0} f(x_t - \eta \nabla f(x_t)).$$

Часто вирішення цієї задачі точно є обчислювально складним, тому на практиці зазвичай використовують наближений підхід.

1.7.2 Backtracking Line-Search

Ідея методу зворотного відстеження (*backtracking line-search*) загалом полягає в тому, щоб спочатку спробувати агресивний (великий) розмір кроку та поступово зменшувати його, якщо він занадто великий.

Алгоритм працює наступним чином: ми вибираємо два параметри $\alpha \in (0, 0.5)$ і $\beta \in (0, 1)$. На ітерації t:

1. Ініціалізуємо $\eta = 1$.

2. **Перевірка умови:** якщо

$$f(x_t - \eta \nabla f(x_t)) > f(x_t) - \alpha \eta \|\nabla f(x_t)\|_2^2,$$

тоді оновлюємо $\eta := \beta \times \eta$ і повторюємо перевірку.

3. В іншому випадку робимо крок:

$$x_{t+1} = x_t - \eta \nabla f(x_t).$$

На практиці часто використовують значення $\alpha = 0.3$ та $\beta = 0.5$, що дає непогані результати.

1.8 Метрична проекція (Metric Projection)

Метрична проекція точки x на опуклу множину C визначається як

$$P_C(x) = \arg\min_{y \in C} \|x - y\|. \tag{4}$$

1.9 Швидкість збіжності (Convergence Rate)

Швидкість збіжності методу оптимізації характеризує, як швидко послідовність $\{x_k\}$ наближається до оптимального розв'язку x^*. Розглядають:

- Лінійну збіжність: $\|x_k - x^*\| \leq Cq^k$, де $q \in (0, 1)$.
- Квадратичну збіжність: $\|x_{k+1} - x^*\| \leq C\|x_k - x^*\|^2$.

1.9.1 Аналіз збіжності

Припустимо, що функція f є опуклою та диференційовною, її область визначення $\mathrm{dom}(f) = \mathbb{R}^n$, а також градієнт ∇f є Ліпшицевим з константою $L > 0$:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y \in \mathbb{R}^n.$$

Або, якщо f двічі диференційовна:

$$\nabla^2 f(x) \preceq LI.$$

Теорема: Градієнтний спуск із фіксованим розміром кроку $t \leq \frac{1}{L}$ задовольняє:

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

Аналогічний результат справедливий для методу зворотного відстеження, якщо замінити t на $\frac{\beta}{L}$.

Таким чином, градієнтний спуск має швидкість збіжності $O(1/k)$

Для того щоб отримати $f(x) - f(x^*) \leq \epsilon$ треба $O(1/\epsilon)$ ітерацій