Optimization Theory

Lecture 08

Fudan University

luoluo@fudan.edu.cn

Outline

Outline

Line Search Methods

A line search method computes a search direction $\mathbf{p}_k$ and then decides how far to move along that direction.

The iteration is given by

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha_t \mathbf{p}_t,$$

where the positive scalar α_t is called step size, step length or learning rate.

We typically require $\mathbf{p}_t$ to be a descent direction that satisfies

$$\langle \mathbf{p}_t, \nabla f(\mathbf{x}_k) \rangle < 0.$$

For example

1. $\mathbf{p}_t = -\nabla f(\mathbf{x}_t)$
2. $\mathbf{p}_t = -\mathbf{G}_t^{-1} \nabla f(\mathbf{x}_t)$ with some positive definite $\mathbf{G}_t \in \mathbb{R}^{d \times d}$

The ideal choice for α is based on

$$\min_{\alpha > 0} \phi(\alpha) \triangleq f(\mathbf{x}_t + \alpha \mathbf{p}_t),$$

but it is not practical.

We want to efficiently select α_t that leads to sufficient reduction in f.

The simple decrease condition

$$f(\mathbf{x}_t + \alpha_t \mathbf{p}_t) < f(\mathbf{x}_t)$$

is not enough.

Wolfe Conditions

We require

$$\begin{aligned} f(\mathbf{x}_t + \alpha_t \mathbf{p}_t) &\leq f(\mathbf{x}_t) + c_1 \alpha_t \langle \nabla f(\mathbf{x}_t), \mathbf{p}_t \rangle, \\ \langle \nabla f(\mathbf{x}_t + \alpha_t \mathbf{p}_t), \mathbf{p}_t \rangle &\geq c_2 \langle \nabla f(\mathbf{x}_t), \mathbf{p}_t \rangle \end{aligned} \tag{1}$$

for some $c_1 \in (0, 1)$ and $c_2 \in (c_1, 1)$, that is Wolfe conditions.

Theorem

Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ *is continuously differentiable and lower bounded. Let* $\mathbf{p}_t$ *be a descent direction at* $\mathbf{x}_t$*, then there exist intervals of step lengths satisfying the conditions (1) with* $0 < c_1 < c_2 < 1$.

Wolfe Conditions

We still consider Wolfe conditions

$$\begin{aligned} f(\mathbf{x}_t + \alpha_t \mathbf{p}_t) \leq & f(\mathbf{x}_t) + c_1 \alpha_t \langle \nabla f(\mathbf{x}_t), \mathbf{p}_t \rangle, \\ \langle \nabla f(\mathbf{x}_t + \alpha_t \mathbf{p}_t), \mathbf{p}_t \rangle \geq & c_2 \langle \nabla f(\mathbf{x}_t), \mathbf{p}_t \rangle \end{aligned} \tag{2}$$

for some $c_1 \in (0, 1)$ and $c_2 \in (c_1, 1)$, that is Wolfe condition.

Theorem

Let $\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha_t \mathbf{p}_t$, where $\mathbf{p}_t$ is a descent direction and α_k satisfies the Wolfe conditions. Suppose that continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and lower bounded on $\mathbb{R}^d$ and continuously differentiable. Then

$$\sum_{t=0}^{+\infty} (\cos \theta_t)^2 \|\nabla f(\mathbf{x}_t)\|_2^2 < +\infty, \quad \textit{where } \cos \theta_t = \frac{-\langle \nabla f(\mathbf{x}_t), \mathbf{p}_t \rangle}{\|\nabla f(\mathbf{x}_t)\|_2 \|\mathbf{p}_t\|_2}.$$

Backtracking Line Search

If the algorithm chooses candidate step lengths appropriately, we can use just the sufficient decrease condition.

Algorithm 1 Backtracking Line Search Method

Input: $\mathbf{x}_t, \mathbf{p}_t \in \mathbb{R}^d, \ \ \hat{\alpha} > 0, \ \ \tau, c_1 \in (0, 1)$

$\alpha = \hat{\alpha}$

while $f(\mathbf{x}_t + \alpha \mathbf{p}_t) > f(\mathbf{x}_t) + c_1 \alpha \langle \nabla f(\mathbf{x}_t), \mathbf{p}_t \rangle$ **do**

    $\alpha \leftarrow \tau \alpha$

Output: $\alpha_t = \alpha$

Outline

Barzilai-Borwein Step Size

Gradient descent methods with Barzilai-Borwein step size has the forms of

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha_t \nabla f(\mathbf{x}_t)$$

where

$$\alpha_t = \frac{\|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2^2}{\langle \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}), \mathbf{x}_t - \mathbf{x}_{t-1} \rangle}$$

or

$$\alpha_t = \frac{\langle \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}), \mathbf{x}_t - \mathbf{x}_{t-1} \rangle}{\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|_2^2}.$$

Outline

Parameter-Free Methods

Algorithm 2 Adaptive Gradient Descent

Input: $\mathbf{x}_0 \in \mathbb{R}^d, \;\; \lambda_0 > 0, \;\; \theta_0 = +\infty$

$\mathbf{x}_1 = \mathbf{x}_0 - \lambda_0 \nabla f(\mathbf{x}_0)$

for $t = 1, 2, \dots$ **do**

$$\lambda_t = \min\left\{ \sqrt{1 + \theta_{t-1}}\,\lambda_{t-1}, \; \frac{\|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2}{2\,\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|_2} \right\}$$

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \lambda_t \nabla f(\mathbf{x}_t)$$

$$\theta_t = \frac{\lambda_t}{\lambda_{t-1}}$$

end for