Maria Cameron

1. Line Search Methods

The line search methods proceed as follows. Each iteration computes a search direction p_k and then decides how far to move along this direction. The iteration is given by

$$x_{k+1} = x_k + \alpha_k p_k,$$

where the scalar α_k is called the step length.

Most line search algorithms require p_k to be a descent direction, i.e.,

$$p_k^T \nabla f(x_k) < 0.$$

Example The steepest descent direction $p_k = -\nabla f_k$ is a descent direction:

$$p_k^T \nabla f(x_k) = -\|\nabla f_k\|^2 < 0.$$

If the Hessian is symmetric positive definite then the Newton direction $p_k = -H_k^{-1}\nabla f_k$ is a descent direction. If the approximate Hessian B_k is symmetric positive definite then the quasi-Newton search direction $p_k = -B_k^{-1}\nabla f_k$is a descent direction:

$$p_k^T \nabla f(x_k) = -\nabla f_k^T H_k^{-1} < 0.$$

Ideally, one would like to find a global minimizer α_k of

$$F(\alpha) := f(x_k + \alpha p_k).$$

However, even finding a local minimizer might take too many iterations and result in a slow routine. Therefore, the goal is just to make a step along the direction p_k that results in a reasonable reduction of f so that the overall algorithm converges to a local minimum of f.

1.1. **Wolfe's conditions.**

A popular inexact line search condition stipulates that α_k should first of all give **sufficient decrease** in the objective function f, as measured by the following inequality

$$f(x_k + \alpha p_k) \le f(x_k) + c_1 \alpha \nabla f_k^T p_k, \quad c_1 \in (0,1) \tag{1}$$

is some fixed constant. Eq. (1) requires that for the picked value of α the graph of $F(\alpha) := f(x_k + \alpha p_k)$ lies below the line $f(x_k) + c_1\alpha\nabla f_k^T p_k$. By Taylor's theorem

$$f(x_k + \alpha p_k) = f(x_k) + \alpha \nabla f_k^T p_k + O(\alpha^2).$$

Since p_k is a descent direction, i.e. $\nabla f_k^T p_k < 0$, such α exists.

The sufficient decrease condition (1) is not enough to ensure convergence since as we have just seen, this condition is satisfied for all small enough α. To rule out unacceptably small steps the second requirement called a **curvature condition** is introduced

$$\nabla f(x_k + \alpha p_k)^T p_k \ge c_2 \nabla f_k^T p_k, \quad c_2 \in (c_1, 1) \tag{2}$$

is a fixed constant. The curvature condition enforces to choose α large enough so that the slope of $F(\alpha)$ is larger than c_2 times the slope of $F(0)$.

Conditions (1) and (2) are the Wolfe conditions. Sometimes the curvature condition can be amplified to out rule α's for which F increases faster than $c_2|\nabla f_k^T p_k$. The resulting conditions are called strong Wolfe's conditions.

$$f(x_k + \alpha p_k) \le f(x_k) + c_1 \alpha \nabla f_k^T p_k, \tag{3}$$

$$|\nabla f(x_k + \alpha p_k)^T p_k| \le c_2 |\nabla f_k^T p_k|, \tag{4}$$

$$0 < c_1 < c_2 < 1.$$

Lemma 1. *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. Let p_k be a descent direction at x_k and assume that f is bounded from below along the ray $\{x_k + \alpha p_k \mid \alpha > 0\}$. then id $0 < c_1 < c_2 < 1$, there exists interval of step lengths α satisfying the Wolfe conditions and the strong Wolfe conditions.*

Proof. Since $F(\alpha) = f(x_k + \alpha p_k)$ is bounded from below for all $\alpha > 0$, the line $l(\alpha) = f(x_k) + \alpha c_1 \nabla f_k^T p_k$ must intersect the graph of ϕ at least once. Let $\alpha' > 0$ be the smallest intersecting value of α, i.e.

$$f(x_k + \alpha' p_k) = f(x_k) + \alpha' c_1 \nabla f_k^T p_k < f(x_k) + c_1 \nabla f_k^T p_k. \tag{5}$$

Hence the sufficient decrease holds for all $0 < \alpha < \alpha'$.

By mean value theorem, there exists $\alpha'' \in (0, \alpha')$ such that

$$f(x_k + \alpha' p_k) - f(x_k) = \alpha' \nabla f(x_k + \alpha'' p_k)^T p_k. \tag{6}$$

Combining Eqs. (5) and (7) we obtain

$$\nabla f(x_k + \alpha'' p_k)^T p_k = c_1 \nabla f_k^T p_k > c_2 \nabla f_k^T p_k, \tag{7}$$

since $c_1 < c_2$ and $\nabla f_k^T p_k < 0$. Therefore, α'' satisfies the Wolfe conditions (1) and (2) and the inequalities are strict. By smoothness assumption on f there is an interval around α'' for which the Wolfe conditions hold. Since $\nabla f(x_k + \alpha'' p_k)^T p_k < 0$, the strong Wolfe conditions (3) and (4) hold in the same interval. $\square$

1.2. **Backtracking.** The sufficient decrease condition alone is not enough to guarantee that the algorithm makes a reasonable progress along the given search direction. However, if the line search algorithm chooses the step length by backtracking, the curvature condition can be dispensed. This means that we first try a large step and the gradually decrease the step length until the sufficient decrease condition is satisfied.

1.3. **Convergence of line search methods.** To obtain global convergence, we must not only have well chosen step lengths but also well-chosen directions p_k. The angle θ_k between the search direction p_k and the steepest descent direction $-\nabla f_k$ is defined by

$$\cos \theta_k = -\frac{\nabla f_k^T p_k}{\|\nabla f_k\| \|p_k\|}. \tag{8}$$

The following theorem due to Zoutendijk, has far-reaching consequences. It shows that the steepest descent method is globally convergent. For other algorithms it describes how far p_k can deviate from the steepest descent direction and still give rise to a globally convergent iteration.

Theorem 1. *Consider any iteration of the form*

$$x_{k+1} = x_k + \alpha_k p_k,$$

where p_k is a descent direction and α_k satisfies the Wolfe conditions (1), (2). *Suppose f is bounded from below in $\mathbb{R}^n$ and f is continuously differentiable in an open set D containing the sublevel set*

$$L := \{x \in \mathbb{R}^n \mid f(x) < f(x_0)\},$$

where x_0 is the starting point of the iteration. Assume also that ∇f is Lipschitz-continuous in D, i.e.,

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

Then

$$\sum_{k \geq 0} \cos^2 \theta_k \|\nabla f\|^2 < \infty. \tag{9}$$

Proof. Subtracting $\nabla f_k^T p_k$ from Eq. (2) and taking into account that $x_{k+1} = x + k + \alpha_k p_k$ we get

$$(\nabla f_{k+1} - \nabla f_k)^T p_k \geq (c_2 - 1)\nabla f_k^T p_k,$$

while the Lipschitz continuity implies that

$$(\nabla f_{k+1} - \nabla f_k)^T p_k \leq \|\nabla f_{k+1} - \nabla f_k\| \|p_k\| \leq \alpha_k L \|p_k\|^2.$$

Combining these two relations we obtain

$$\alpha_k \geq \frac{c_2 - 1}{L} \frac{\nabla f_k^T p_k}{\|p_k\|^2}.$$

By substituting this inequality in Eq. (1) we get

$$\begin{aligned} f_{k+1} \leq & f_k + \alpha_k c_1 \nabla f_k^T p_k \\ \leq & f_k - c_1 \frac{1 - c_2}{L} \frac{(\nabla f_k^T p_k)^2}{\|p_k\|^2} \\ \leq & f_k - c \cos^2 \theta_k \|\nabla f_k\|^2, \end{aligned}$$

where $c := c_1(1 - c_2)/L$. By summing this expression over all indices we outain

$$f_{k+1} \leq f_0 - c \sum_{j=0}^{k} \cos^2 \theta_j \|\nabla f_j\|^2.$$

Since f is bounded from below, we have that $f_0 - f_{k+1}$ is less than some positive constant. Hence

$$\sum_{j=0}^{\infty} \cos^2 \theta_j \|\nabla f_j\|^2 < \infty.$$

$\square$

We call the inequality (9) the Zoutendijk condition. It implies that

$$\cos^2\theta_k \|\nabla f_k\| \to 0 \quad \text{as} \quad k \to \infty.$$

If an algorithm chooses directions so that $\cos\theta_k$ is bounded away from 0 then

$$\lim_{k\to\infty} \|\nabla f_k\| = 0.$$

For example, for the steepest descent method $\cos\theta_k = 1$ hence $\|\nabla f_k\| \to 0$. Therefore, without any additional requirements on f we only can guarantee convergence to a stationary point rather than to a minimizer.

For Newton-like methods the search direction is of the form

$$p_k = -B_k^{-1}\nabla f_k.$$

If we assume that all matrices B_k are symmetric positive definite with uniformly bounded condition number, i.e.,

$$\|B_k\|\|B_k^{-1}\| \le M \text{ for all } k,$$

then

$$\begin{aligned} \cos\theta_k &= -\frac{\nabla f_k^T p_k}{\|\nabla f_k\|\|p_k\|} \\ &= \frac{\nabla f_k^T B_k^{-1}\nabla f_k}{\|\nabla f_k\|\|B_k^{-1}\nabla f_k\|} \\ &\ge \frac{1}{\|\nabla f_k\|}\frac{\|\nabla f_k\|^2}{\|B_k\|}\frac{1}{\|B_k^{-1}\|\|\nabla f_k\|} \\ &= \frac{1}{\|B_k\|\|B_k^{-1}\|} \ge \frac{1}{M}. \end{aligned}$$

Therefore, $\cos\theta_k$ is bounded away from 0. Hence $\|\nabla f_k\| \to 0$.

Example In the BFGS method (Broyden, Fletcher, Goldfarb, and Shanno)

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k} \tag{10}$$

where

$$s_k = x_{k+1} - x_k, \quad y_k - \nabla f_{k+1} - \nabla y_k.$$

Note that $B_{k+1} - B_k$ is a symmetric rank 2 matrix. One can show that the BFGS update generates positive definite matrices whenever B_0 is positive definite and $s_k^T y_k > 0$.

References

[1] J. Nocedal, S. Wright, Numerical Optimization, Springer, 1999