45010: Optimization I **2019W**

Lecture 5: Wolfe Condition

Lecturer: Yimin Zhong *Scribes: None*

Note: *In all notes, bold face letters denote vectors.*

5.1 Wolfe Condition

In the previous class, we have learned that α_k is the step length, and does not have to be the exact line search optmization problem's minimizer. In this class, we learn one condition for the choices of step length, which is called Wolfe condition, it tells us what kind of step length should be taken.

5.1.1 Sufficient decrease (Armijo's condition)

The first condition is called "sufficient decrease" condition. The idea is, if we move from the previous $\mathbf{x}_k$ to the next iterate $\mathbf{x}_{k+1} := \mathbf{x}_k + \alpha_k \mathbf{p}_k$, the decrease in f must be sufficiently large. We know that from Taylor expansion,

$$f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) = f(\mathbf{x}_k) + \alpha_k \mathbf{p}_k^T \cdot \nabla f(\mathbf{x}_k + t\alpha_k \mathbf{p}_k) \quad \text{for some } t \in (0,1) \tag{5.1}$$

and the decrease is the second term on the right hand side, we want that term to be sufficiently negative, observe that this term is already proportional to both α_k and $\mathbf{p}_k^T \cdot \nabla f(\mathbf{x}_k + t\alpha_k \mathbf{p}_k)$, hopefully, $\mathbf{p}_k^T \cdot \nabla f(\mathbf{x}_k + t\alpha_k \mathbf{p}_k)$ is not too far from $\mathbf{p}_k^T \cdot \nabla f(\mathbf{x}_k)$, but we cannot be sure about it, so we (strongly) relax such assumption by a parameter $0 < c_1 < 1$ really small (usually $c_1 = 10^{-4}$), that

$$\mathbf{p}_k^T \cdot \nabla f(\mathbf{x}_k + t\alpha_k \mathbf{p}_k) \leq c_1 \mathbf{p}_k^T \cdot \nabla f(\mathbf{x}_k) \tag{5.2}$$

this means, the decrease has to be at least a portion of $\mathbf{p}_k^T \cdot \nabla f(\mathbf{x}_k)$, this is negative by descent direction assumption. Combine this and the previous formula,

$$f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) \leq f(\mathbf{x}_k) + c_1 \alpha_k \mathbf{p}_k^T \cdot \nabla f(\mathbf{x}_k) \tag{5.3}$$

Geometrically, if we put the above inequality in terms of α_k only, it means the right hand side's straight line has to be over the curve defined by $\phi(\alpha) := f(\mathbf{x}_k + \alpha \mathbf{p}_k)$ near α_k.

Example 5.1 *In a two dimensional example, $f(x,y) = (x-1)^2 + (2y-1)^2$, let $\mathbf{x}_k = (x_k, y_k)$, the gradient $\nabla f(\mathbf{x}_k) := [2(x_k - 1), 2(2y_k - 1)]$, let the descent direction be the steepest direction $\mathbf{p}_k = -\nabla f(\mathbf{x}_k) = -[2(x_k-1), 2(2y_k-1)]$, then put everything into (5.3), the sufficient decrease condition is*

$$(1-2\alpha_k)^2(x_k-1)^2 + (1-4\alpha_k)^2(2y_k-1)^2 \leq (x_k-1)^2 + (2y_k-1)^2 - c_1\alpha_k\left(4(x_k-1)^2 + 4(2y_k-1)^2\right) \tag{5.4}$$

The left hand side is a parabola, the right hand side is a straight line, so we must have two roots if $c_1 < 1$ (why is that? try to simplify the above inequality.)

5.1.2 Curvature condition

The second condition is called curvature condition. Intuitively, we only need the Armijo condition to get decrease in f, however, only the first condition of Wolfe does not guarantee the optimizer can be found due to very small step lengths are permitted, if each time, I only move very small length, then it is not possible to get to the optimal solution.

Therefore, the curvature condition is to prevent small steps. The decrease between $f(\mathbf{x}_k + \alpha_k \mathbf{p}_k)$ and $f(\mathbf{x}_k)$ is

$$f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) - f(\mathbf{x}_k) = \alpha_k \mathbf{p}_k^T \cdot \nabla f(\mathbf{x}_k + t\alpha_k \mathbf{p}_k) \tag{5.5}$$

If α_k is small, then $\mathbf{p}_k^T \cdot \nabla f(\mathbf{x}_k + \alpha_k \mathbf{p}_k)$ will be close to $\mathbf{p}_k^T \cdot \nabla f(\mathbf{x}_k)$, so is $\mathbf{p}_k^T \cdot \nabla f(\mathbf{x}_k + t\alpha_k \mathbf{p}_k)$, so we do not want that! We set another $1 > c_2 > c_1$ (usually $c_2 = 0.9$), which satisfies

$$\mathbf{p}_k^T \cdot \nabla f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) \leq c_2 \mathbf{p}_k^T \cdot \nabla f(\mathbf{x}_k) \tag{5.6}$$

which means, there is a fixed portion of difference between $\mathbf{p}_k^T \cdot \nabla f(\mathbf{x}_k + \alpha_k \mathbf{p}_k)$ and $\mathbf{p}_k^T \cdot \nabla f(\mathbf{x}_k)$.

Example 5.2 *Under the same objective function as before, if we use the steepest descent, this condition means*

$$-4(x_k - 1)^2(1 - 2\alpha_k) - 4(2y_k - 1)^2(1 - 4\alpha_k) \leq -c_2 \left(4(x_k - 1)^2 + 4(2y_k - 1)^2\right) \tag{5.7}$$

Try to simplify this and conclude what α_k should be.

Lemma 5.3 *If f is differentiable and is bounded from below, then we can always find an interval of α_k that satisfies Wolfe condition.*

Proof: The idea for the proof is in two aspects: (1). The straight line and the curve will intersect at some point(s), we locate the first one as α_0, for any $\alpha < \alpha_0$, the straight line is over the curve, which means all the $\alpha \in (0, \alpha_0)$ satisfy the sufficient decrease condition. (2). There are two ways to present $f(\mathbf{x}_k + \alpha_0 \mathbf{p}_k) - f(\mathbf{x}_k)$, either using the fact of intersection or Taylor expansion, we connect them using $c_1 < c_2$. ■

This theorem confirms that Wolfe condition is a practical condition for step length. The proof of this Lemma is not required. Interested students can refer the full proof in the textbook's Lemma 3.1.