TMA4180 Optimization I

Convergence of descent methods with backtracking (Armijo) linesearch. Bisection algorithm for weak Wolfe conditions.

Anton Evgrafov

Department of Mathematical Sciences, NTNU anton.evgrafov@math.ntnu.no

1 Convergence of descent methods with backtracking (Armijo) linesearch.

Read: Section 3.1 in Nocedal and Wright, "Numerical optimization," in particular Algorithm 3.1, p. 37.

Consider the following iteration:

$$x_{k+1} = x_k + \alpha_k p_k, \qquad k = 0, 1, 2, \dots$$

with

$$B_k p_k = -\nabla f(x_k),$$

where $B_k = B_k^{\mathrm{T}}$, and α_k is selected using the backtracking (Armijo) linesearch with parameters $c, \rho \in (0, 1)$.

Theorem 1. *Assume that*

1. $f : \mathbb{R}^n \to \mathbb{R}$ *is continuously differentiable;*
2. *the set* $S := \{x \in \mathbb{R}^n \mid f(x) \leq f(x_0)\}$ *is bounded;*
3. *the matrices* B_k *are uniformly positive definite and bounded, that is* $\exists m > 0, M > 0 : m \leq \lambda_{\min}(B_k) \leq \lambda_{\max}(B_k) \leq M$*, where* $\lambda_{\min}$ *and* $\lambda_{\max}$ *are the smallest and the largest eigenvalues of* B_k*.*

Then the sequence $\{x_k\}$ *is bounded, and every its limit point* $\hat{x}$ *is a stationary point for* f*.*

Proof. Owing to the sufficient decrease condition in the linesearch procedure the sequence $f(x_k)$, $k = 0, 1, 2, \dots$ is non-increasing; thus $x_k \in S$ for all k; in particular it is bounded and therefore has at least one limit point. The set S is closed because f is continuous, and thus is compact owing to the assumption 2 and Heine–Borel theorem. Therefore, the function f attains its minimum value on S (Weierstrass theorem) and thus is bounded from below on S. As a result, the non-increasing sequence $f(x_k)$ has a finite limit, and furthermore $\lim_{k\to\infty}[f(x_{k+1}) - f(x_k)] = 0$.

Owing to the sufficient decrease condition it holds that

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq c\alpha_k \nabla f(x_k)^T p_k = -c\alpha_k \nabla f(x_k)^T B_k^{-1} \nabla f(x_k) \\ &\leq -c\alpha_k \lambda_{\min}(B_k^{-1}) \|\nabla f(x_k)\|^2 \\ &\leq -cM^{-1}\alpha_k \|\nabla f(x_k)\|^2 \leq 0. \end{aligned} \tag{1}$$

The sequence on the left converges to 0, meaning that the sequence on the right must also converge to zero. We will show that this implies that $\lim_{k\to\infty} \|\nabla f(x_k)\| = 0$.

Suppose that this is not true; then, for some subsequence of indices k' and some $\epsilon > 0$ we must have that $\|\nabla f(x_{k'})\| \geq \epsilon$. From (1) it then follows that $\lim_{k'\to\infty} \alpha_{k'} = 0$. In particular, it means that the step $\alpha_{k'}/\rho$ was not acceptable to the linesearch procedure for all large k', that is

$$f(x_{k'} + \alpha_{k'}\rho^{-1}p_{k'}) > f(x_{k'}) + c\alpha_{k'}\rho^{-1}\nabla f(x_{k'})^T p_{k'}. \tag{2}$$

The sequence of directions $p_k = -B_k^{-1}\nabla f(x_k)$ is bounded. Indeed, by our assumption 3 the norms $\|B_k^{-1}\| = \lambda_{\min}^{-1}(B_k) \leq m^{-1}$. Furthermore, the continuous function $x \mapsto \|\nabla f(x)\|$ attains its maximum over the compact set S, and thus $\|\nabla f(x_k)\|$ is bounded by this maximum value, for all k. As a result, we may assume that for some subsequence of k', say k'', it holds that $\lim_{k''\to\infty} x_{k''} = \hat{x}$ and $\lim_{k''\to\infty} p_{k''} = \hat{p}$. Rearranging the terms in (6) we get

$$\begin{aligned} 0 &\leq \lim_{k''\to\infty} \frac{f(x_{k''} + \alpha_{k''}\rho^{-1}p_{k''}) - f(x_{k''})}{\alpha_{k''}\rho^{-1}} - c\nabla f(x_{k''})^T p_{k''} \\ &= (1-c)\nabla f(\hat{x})^T \hat{p}, \end{aligned} \tag{3}$$

and therefore $\nabla f(\hat{x})^T\hat{p} \geq 0$ as $0 < c < 1$. On the other hand,

$$\begin{aligned} \nabla f(\hat{x})^T \hat{p} &= \lim_{k''\to\infty} \nabla f(x_{k''})^T p_{k''} = -\lim_{k''\to\infty} \nabla f(x_{k''})^T B_{k''}^{-1} \nabla f(x_{k''}) \\ &\leq -M^{-1}\epsilon^2 < 0. \end{aligned} \tag{4}$$

However, equations (7) and (8) contradict each other. This must mean that our assumption that $\|\nabla f(x_{k'})\| \geq \epsilon$ over some subsequence k' is wrong and in fact

$$\lim_{k\to\infty} \|\nabla f(x_k)\| = 0. \tag{5}$$

Finally, let $\hat{x}$ be an arbitrary limit point of $\{x_k\}$, that is, $\hat{x} = \lim_{k'''\to\infty} x_{k'''}$ for some subsequence k'''. Owing to the continuity of the function $x \mapsto \|\nabla f(x)\|$ (assumption 1) and (5) it holds that $\|\nabla f(\hat{x})\| = 0$, as we claimed. □

Note: Assumption 3 of the theorem above implies two things. First, the *length* of the search direction remains “of the same order” as the length of the gradient. Thus the steplengths α_k do not have to go to zero because the ratio $\|p_k\|/\|\nabla f(x_k)\|$ grows indefinitely; or conversely, the unit steps α_k become accepted by the linesearch routine without the algorithm making any significant progress because the ratio $\|p_k\|/\|\nabla f(x_k)\|$ is too small.

Second, the *angle* between the gradient and the search direction remains obtuse (thus p_k is always a descent direction), and the angle is bounded away from $\pi/2$. Thus the steplengths α_k do not have to go to zero because the search directions become almost orthogonal to the gradient, and as a result potentially almost cease to be directions of descent.

The results similar to the theorem above can be formulated in several equivalent ways, but without conditions ensuring the "quality" of the direction of descent as described by the two properties above the convergence of the algorithm (as described by the theorem) cannot be asserted.

2 Bisection algorithm for weak Wolfe conditions

Steplengths satisfying the weak Wolfe conditions can be computed using a very simple algorithm, based on the ideas of bracketing and bisection. Namely: given the constants $0 < c_1 < c_2 < 1$, a point $x \in \mathbb{R}^n$, and a direction $p \in \mathbb{R}^n$ satisfying the condition $\nabla f(x)^{\mathrm{T}} p < 0$, consider the following iteration:

```
 1: $\alpha := 1$, $\alpha_{\min} := 0$, $\alpha_{\max} := +\infty$                      ▷ Initialization
 2: loop
 3:     if $f(x + \alpha p) > f(x) + c_1 \alpha \nabla f(x)^{\mathrm{T}} p$ then        ▷ No sufficient decrease
 4:         $\alpha_{\max} := \alpha$
 5:         $\alpha := (\alpha_{\min} + \alpha_{\max})/2$
 6:     else if $\nabla f(x + \alpha p)^{\mathrm{T}} p < c_2 \nabla f(x)^{\mathrm{T}} p$ then    ▷ No curvature condition
 7:         $\alpha_{\min} := \alpha$
 8:         if $\alpha_{\max} = +\infty$ then
 9:             $\alpha := 2\alpha$
10:         else
11:             $\alpha := (\alpha_{\min} + \alpha_{\max})/2$
12:         end if
13:     else
14:         return $\alpha$                                                 ▷ Success!
15:     end if
16: end loop
```

Theorem 2. *Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, $0 < c_1 < c_2 < 1$, $x, p \in \mathbb{R}^n$,are such that $\nabla f(x)^{\mathrm{T}} p < 0$. Then, either (i) the algorithm successfully stops at line 14 with a steplength $\alpha > 0$ satisfying the weak Wolfe conditions after finitely many loop iterations, or (ii) $\alpha_{\max}$ remains infinite, the steplength α is doubled every time and eventually grows to infinity, and $f(x + \alpha p) \to -\infty$.*

Proof. Assume that the algorithm does not terminate after finitely many steps (thus line 14 is never reached). If $\alpha_{\max}$ is never set to a finite value, it means that the lines 4–5 are never visited, and as a result we keep repeating lines 7 and 9. Thus indeed, α is doubled at each iteration, and since the condition on line 3 is evaluated to false, we have that $f(x + \alpha p) \leq f(x) + c_1 \alpha \nabla f(x)^{\mathrm{T}} p$ at each iteration. Since $\nabla f(x)^{\mathrm{T}} p < 0$ by assumption, the quantity $f(x + \alpha p)$ must go to minus infinity. As a result, the alternative (ii) assessed in the theorem holds in this case.

Now assume that at some point $\alpha_{\max}$ becomes finite. Suppose that this happens at iteration K of the loop. From now on we will write α^k, $\alpha^k_{\min}$, and $\alpha^k_{\max}$ to denote the algorithmic quantities at iteration k of the loop.

Note that from iteration K, the sequence $\alpha^k_{\max}$ is finite and non-increasing, the sequence $\alpha^k_{\min}$ is non-negative and non-decreasing, $\alpha^k_{\min} < \alpha^k < \alpha^k_{\max}$, and finally $\alpha^{k+1}_{\max} - \alpha^{k+1}_{\min} = (\alpha^k_{\max} - \alpha^k_{\min})/2$. Therefore, there is $\hat{\alpha} \geq 0$, such that $\alpha^k \to \hat{\alpha}$, $\alpha^k_{\min} \uparrow \hat{\alpha}$, and $\alpha^k_{\max} \downarrow \hat{\alpha}$.

Finally, by algorithmic construction we have (for iterations $k > K$)

$$f(x + \alpha^k_{\max} p) > f(x) + c_1 \alpha^k_{\max} \nabla f(x)^{\mathrm{T}} p, \tag{6}$$

see lines 3–5;

$$\nabla f(x + \alpha^k_{\min} p) p < c_2 \nabla f(x) p, \tag{7}$$

see lines 6–11, and finally

$$f(x + \alpha^k_{\min} p) \leq f(x) + c_1 \alpha^k_{\min} \nabla f(x)^{\mathrm{T}} p, \tag{8}$$

because the algorithm arrives at lines 6–11 only when the condition on line 3 fails. Note that (7) and (8) hold even if lines 7–11 have never been visited and $\alpha^k_{\min} = 0$.

Taking the limit in (7) we arrive at $\nabla f(x + \hat{\alpha} p) p \leq c_2 \nabla f(x) p$. We can also add (6) and (8) to obtain

$$c_1(\alpha^k_{\max} - \alpha^k_{\min}) \nabla f(x)^{\mathrm{T}} p < f(x + \alpha^k_{\max} p) - f(x + \alpha^k_{\min} p) = (\alpha^k_{\max} - \alpha^k_{\min}) \nabla f(x + \tilde{\alpha}^k p)^{\mathrm{T}} p,$$

where the last equality is owing to the mean value theorem for some $\alpha^k_{\min} \leq \tilde{\alpha}^k \leq \alpha^k_{\max}$. We divide the last inequality by $\alpha^k_{\max} - \alpha^k_{\min} > 0$ and take a limit to obtain the inequality $\nabla f(x + \hat{\alpha} p)^{\mathrm{T}} p \geq c_1 \nabla f(x)^{\mathrm{T}} p$, which is a contradiction with the previously obtained inequality $\nabla f(x + \hat{\alpha} p) p \leq c_2 \nabla f(x) p$ as $c_2 \nabla f(x) p < c_1 \nabla f(x)^{\mathrm{T}} p$.

Therefore, the algorithm must terminate after finitely many iterations, or alternative (i) must hold in this case. □