

Globalizing Newton's method: line searches (II)

Mark S. Gockenbach

I begin by showing that, under mild conditions, the Wolfe conditions can be satisfied.

Theorem 0.1 *Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable, $x^{(k)} \in \mathbb{R}^n$, and $p^{(k)}$ is a descent direction for f at $x^{(k)}$. Suppose further that f is bounded below on the ray $\{x^{(k)} + \alpha p^{(k)} : \alpha \geq 0\}$. If $0 < c_1 < c_2 < 1$, then there exists an interval $[a_1, a_2]$, $0 < a_1 < a_2$, such that*

$$f(x^{(k)} + \alpha p^{(k)}) \leq f(x^{(k)}) + c_1 \alpha \nabla f(x^{(k)}) \cdot p^{(k)}$$

and

$$\nabla f(x^{(k)} + \alpha p^{(k)}) \cdot p^{(k)} \geq c_2 \nabla f(x^{(k)}) \cdot p^{(k)}$$

for all $\alpha \in [a_1, a_2]$.

Proof I use the usual notation

$$\phi(\alpha) = f(x^{(k)} + \alpha p^{(k)}), \quad \alpha \geq 0.$$

Since $\phi'(0) < 0$ and $0 < c_1 < 1$, it is easy to show that

$$\phi(\alpha) < \phi(0) + c_1 \alpha \phi'(0) \tag{1}$$

for all α sufficiently small. Moreover, since ϕ is bounded below by assumption, (1) clearly fails for α sufficiently large. Therefore, if I define

$$\bar{\alpha} = \sup\{\hat{\alpha} > 0 : \phi(\alpha) < \phi(0) + c_1 \alpha \phi'(0) \forall \alpha \in (0, \hat{\alpha})\},$$

then $\bar{\alpha}$ exists and is finite. Moreover, $\phi(\bar{\alpha}) = \phi(0) + c_1 \bar{\alpha} \phi'(0)$. By the Mean Value Theorem, there exists $\beta \in (0, \bar{\alpha})$ such that

$$\begin{aligned} \phi'(\beta) &= \frac{\phi(\bar{\alpha}) - \phi(0)}{\bar{\alpha}} \\ &= \frac{\phi(0) + c_1 \bar{\alpha} \phi'(0) - \phi(0)}{\bar{\alpha}} \\ &= c_1 \phi'(0) \\ &> c_2 \phi'(0) \quad (\text{since } c_2 > c_1 \text{ and } \phi'(0) < 0). \end{aligned}$$

By the continuity of ϕ' , there exists an interval $(a_1, a_2) \subset (0, \bar{\alpha})$ around β such that

$$\phi'(\alpha) \geq c_2 \phi'(0) \tag{2}$$

for all $\alpha \in (a_1, a_2)$. But then, by construction, both (1) and (2) hold for $\alpha \in (a_1, a_2)$. QED

A little thought shows that, in fact, the strong Wolfe conditions are satisfied for α in the interval (a_1, a_2) constructed in the proof of the previous theorem. This follows from the fact that $\phi'(\beta) < 0$, and hence

$$\phi'(\beta) > c_2 \phi'(0)$$

implies that

$$|\phi'(\beta)| < c_2 |\phi'(0)|.$$

I can now show that any descent algorithm employing a line search that satisfies the Wolfe conditions is globally convergent in a certain sense. I begin with a technical result. The reader should recall that the angle θ between two vectors $p, q \in \mathbb{R}^n$ is defined¹ by

$$\cos \theta = \frac{p \cdot q}{\|p\| \|q\|}. \quad (3)$$

Theorem 0.2 (Zoutendijk) *Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable and that ∇f is Lipschitz continuous on an open set N containing the lower level set $\{x \in \mathbb{R}^n : f(x) \leq f(x^{(0)})\}$. Suppose that the sequence $\{x^{(k)}\}$ is defined by*

$$x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}, \quad k = 0, 1, 2, \dots,$$

where, for all k , $p^{(k)}$ is a descent direction for f at $x^{(k)}$ and α_k is chosen to satisfy the Wolfe conditions. Then

$$\sum_{k=1}^{\infty} \cos^2 \theta_k \|\nabla f(x^{(k)})\|^2 < \infty,$$

where θ_k is the angle between $-\nabla f(x^{(k)})$ and $p^{(k)}$.

Proof Let L be the Lipschitz constant for ∇f on N , so that

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \text{ for all } x, y \in N.$$

Then, using the Cauchy-Schwarz inequality and the Lipschitz continuity of ∇f ,

$$\left(\nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) \right) \cdot p^{(k)} \leq \|\nabla f(x^{(k+1)}) - \nabla f(x^{(k)})\| \|p^{(k)}\| \leq L\alpha_k \|p^{(k)}\|^2.$$

On the other hand, the curvature condition implies that

$$\left(\nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) \right) \cdot p^{(k)} \geq (c_2 - 1) \nabla f(x^{(k)}) \cdot p^{(k)}.$$

Putting together the previous two inequalities yields a lower bound on α_k :

$$\alpha_k \geq \frac{(c_2 - 1) \nabla f(x^{(k)}) \cdot p^{(k)}}{L \|p^{(k)}\|^2}. \quad (4)$$

(This result shows explicitly how the curvature condition prevents short steps.) Substituting (4) into the sufficient decrease condition, I can derive a lower bound on the decrease in f :

$$\begin{aligned} f(x^{(k+1)}) - f(x^{(k)}) &\leq c_1 \alpha_k \nabla f(x^{(k)}) \cdot p^{(k)} \\ &\leq \frac{c_1 (c_2 - 1) (\nabla f(x^{(k)}) \cdot p^{(k)})^2}{L \|p^{(k)}\|^2} \\ &= \frac{c_1 (c_2 - 1)}{L} \frac{(\nabla f(x^{(k)}) \cdot p^{(k)})^2}{\|\nabla f(x^{(k)})\|^2 \|p^{(k)}\|^2} \|\nabla f(x^{(k)})\|^2 \\ &= \frac{c_1 (c_2 - 1)}{L} \cos^2 \theta_k \|\nabla f(x^{(k)})\|^2, \end{aligned}$$

which implies that

$$\frac{c_1 (1 - c_2)}{L} \cos^2 \theta_k \|\nabla f(x^{(k)})\|^2 \leq f(x^{(k)}) - f(x^{(k+1)}).$$

¹In \mathbb{R}^2 or \mathbb{R}^3 , (3) follows from trigonometry, while for $n > 3$, (3) is a definition.

Since

$$\sum_{k=0}^N \left(f(x^{(k)}) - f(x^{(k+1)}) \right) = f(x^{(0)}) - f(x^{(N+1)})$$

(telescoping sum) and f is bounded below, it follows that

$$\sum_{k=0}^{\infty} \left(f(x^{(k)}) - f(x^{(k+1)}) \right) < \infty$$

and hence that

$$\sum_{k=0}^{\infty} \cos^2 \theta_k \|\nabla f(x^{(k)})\|^2 < \infty.$$

QED

The consequence of Zoutendijk's result is the following: If f is bounded below and the angle between $p^{(k)}$ and $-\nabla f(x^{(k)})$ is bounded away from 90° , say $0 \leq \theta < \hat{\theta} < 90$, where $\hat{\theta}$ is a constant, then

$$\cos^2 \theta_k \geq \cos^2 \hat{\theta} > 0 \text{ for all } k.$$

It then follows that

$$\|\nabla f(x^{(k)})\| \rightarrow 0 \text{ as } k \rightarrow \infty. \quad (5)$$

This is more or less the best global convergence result that can be proved for line search methods. If, for example, the lower level set $S = \{x \in \mathbb{R}^n : f(x) \leq f(x^{(0)})\}$ is compact, then the sequence $\{x^{(k)}\}$ is bounded and hence at least a subsequence converges to some x^* , and (5) shows that the limit point x^* is a stationary point. Since $\{x^{(k)}\}$ is generated by a descent algorithm, it is easy to show that the stationary point cannot be a local maximizer; however, it is not possible to rule out convergence to a saddle point.

To guarantee that the angle between $-\nabla f(x^{(k)})$ and $p^{(k)}$ is bounded away from 90° , it suffices to ensure that the condition numbers of the Hessian approximations H_k are uniformly bounded.

Definition 0.3 *The condition number of $A \in \mathbb{R}^{n \times n}$ is defined by*

$$\text{cond}(A) = \|A\| \|A^{-1}\|,$$

where the norm is taken to be the operator norm.²

For a symmetric positive definite matrix A ,

$$\|A\| = \lambda_{\max}(A), \quad \|A^{-1}\| = \lambda_{\min}(A)^{-1},$$

and hence the condition number reduces to ratio of the largest eigenvalue of A to the smallest:

$$\text{cond}(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

Theorem 0.4 *Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable and $p^{(k)} = -H_k^{-1} \nabla f(x^{(k)})$, where $\{H_k\}$ is a sequence of symmetric positive definite matrices. If there exists $M > 0$ such that*

$$\text{cond}(H_k) \leq M \text{ for all } k = 0, 1, 2, \dots,$$

then

$$\cos \theta_k \geq \frac{1}{M},$$

where $\cos \theta_k$ is the angle between $-\nabla f(x^{(k)})$ and $p^{(k)}$.

²Actually, there are many condition numbers, one for each choice of vector norm and induced operator norm. I am using the Euclidean norm for vectors, which induces a particular operator norm for matrices and hence a particular condition number.

Proof By definition,

$$\cos \theta_k = \frac{-\nabla f(x^{(k)}) \cdot p^{(k)}}{\|\nabla f(x^{(k)})\| \|p^{(k)}\|} = \frac{\nabla f(x^{(k)}) \cdot H_k^{-1} \nabla f(x^{(k)})}{\|\nabla f(x^{(k)})\| \|H_k^{-1} \nabla f(x^{(k)})\|}.$$

Any symmetric matrix $A \in \mathbb{R}^{n \times n}$ satisfies

$$\lambda_{\min}(A) \|x\|^2 \leq x \cdot Ax \leq \lambda_{\max}(A) \|x\|^2 \text{ for all } x \in \mathbb{R}^n,$$

and hence

$$\nabla f(x^{(k)}) \cdot H_k \nabla f(x^{(k)}) \geq \lambda_{\max}(H_k) \|\nabla f(x^{(k)})\|^2,$$

since the eigenvalues of H_k^{-1} are the reciprocals of the eigenvalues of H_k . Also,

$$\|H_k^{-1} \nabla f(x^{(k)})\| \leq \|H_k^{-1}\| \|\nabla f(x^{(k)})\| = \lambda_{\min}(H_k)^{-1} \|\nabla f(x^{(k)})\|.$$

Therefore, for all k ,

$$\cos \theta_k \geq \frac{\lambda_{\max}(H_k)^{-1} \|\nabla f(x^{(k)})\|^2}{\lambda_{\min}(H_k)^{-1} \|\nabla f(x^{(k)})\|^2} = \frac{\lambda_{\min}(H_k)}{\lambda_{\max}(H_k)} = \frac{1}{\text{cond}(H_k)} \geq \frac{1}{M}.$$

QED

If the Hessian approximations H_k are generated by direct modification of $\nabla^2 f(x^{(k)})$, then the condition numbers of the matrices H_k can be bounded by design, thus ensuring convergence. In the case of the BFGS method, it is apparently not possible to bound the condition numbers, and therefore it is more difficult to prove global convergence (global convergence *can* be proved for BFGS, albeit under more restrictive hypotheses).

There are a couple of theoretical points left to address. The reader will recall that the idea of a line search and the Wolfe conditions is to create a globally convergent algorithm out of Newton's method or a quasi-Newton method (such as the BFGS method), which are only locally convergent. However, the line search should not interfere with the fast local convergence of quasi-Newton framework. Therefore, it is essential to know that $\alpha_k = 1$ satisfies the Wolfe conditions for all k sufficiently large, so that, for instance, Newton's method with line search reduces simply to Newton's method when a neighborhood of the solution is reached.

I will state the result only for Newton's method itself, although a similar result can be stated and proved for secant methods such as BFGS. The proof of the following theorem is left as an exercise.

Theorem 0.5 *Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is three times continuously differentiable and x^* satisfies the sufficient conditions for a local minimizer of f ($\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ positive definite). If Newton's method generates a sequence $\{x^{(k)}\}$ converging to x^* , then $\alpha_k = 1$ satisfies the Wolfe conditions (in fact, the strong Wolfe conditions) for all k sufficiently large, provided $c_1 \leq 1/2$ in the sufficient decrease condition. In other words, if*

$$p^{(k)} = -\nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)}),$$

then for all k sufficiently large,

$$f(x^{(k)} + p^{(k)}) \leq f(x^{(k)}) + c_1 \nabla f(x^{(k)}) \cdot p^{(k)}$$

and

$$\nabla f(x^{(k)} + p^{(k)}) \cdot p^{(k)} \geq c_2 \nabla f(x^{(k)}) \cdot p^{(k)}.$$

Finally, I asserted earlier that, if the line search is of the backtracking type, then it is not necessary to enforce the curvature condition (unless it is needed for BFGS updating) in order to obtain global convergence.