

Line Search Methods Analysis

Saurav Samantaray

Department of Mathematics

Indian Institute of Technology Madras

February 5, 2024



Step Length

- In computing the step length we face a trade-off.
- We want to choose α_k to give a substantial reduction of f , but we don't want to spend too much time making the choice.
- Off-course the ideal choice would be the global minimiser of the univariate function $\phi(\cdot)$ defined by

$$\phi(\alpha) = f(x_k + \alpha p_k), \alpha > 0. \quad (1)$$

- But in general, it is too expensive to identify this value.
- It requires too many evaluations of the objective function and/or the gradient to even find a local minimiser to moderate precision.

Step Length

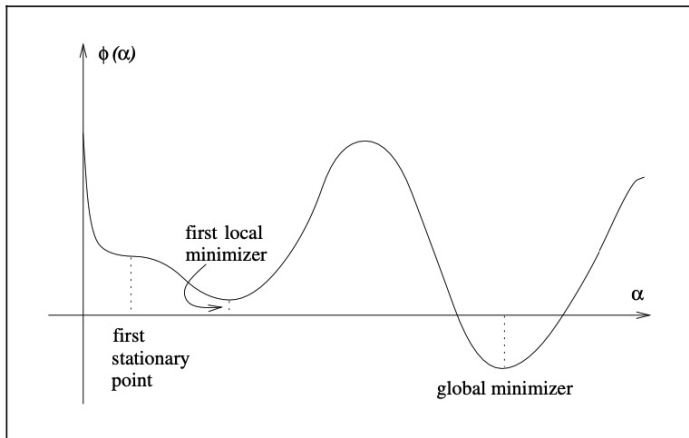


Figure: The ideal step length is the global minimiser

Step Length

- Practically, strategies perform an inexact line search to identify a step length that achieves adequate reductions in f at minimal cost.
- We will discuss these search strategies a little later.
- We will now discuss various termination conditions for line search algorithms and show that effective step lengths need not lie near minimisers of the univariate function $\phi(\alpha)$.
- Is $f(x_k + \alpha_k p_k) < f(x_k)$ good enough to get convergence??
- for example consider the function

$$f(x) = x^2 - 1$$

it has the global minima at $x = 0$, $f = -1$.

Step Length

- Consider a sequence $\{x_k\}$ s.t.

$$f(x_k) = \frac{5}{k}, \quad k = 1, 2, 3, \dots$$

$$\implies f(x_k) > f(x_{k+1})$$

- The reduction in f at each step is not enough to get it to converge to the minimiser.

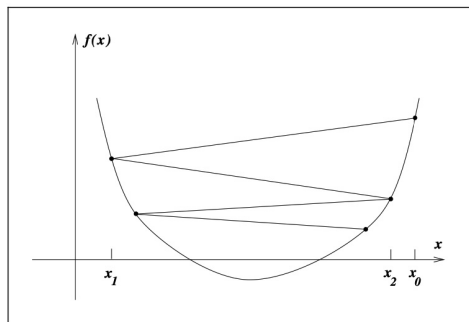


Figure: Insufficient reduction

The Wolfe Condition

Armijo Condition (Sufficient Decrease Condition):

α_k should be chosen such that

$$f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \nabla f_k^T p_k \quad (2)$$

for some constant $c_1 \in (0, 1)$.

- Since p_k is a descent direction and $c_1 > 0$ and $\alpha > 0$ the first thing that the **Armijo condition** asserts that there is a reduction in f from x_k to $x_{k+1} = x_k + \alpha p_k$.
- The reduction in f is at least

$$c_1 \alpha \nabla f_k^T p_k$$

therefore it also says the reduction in f must be proportional to both the step length α_k and the directional derivative $\nabla f_k^T p_k$

The Wolfe Condition

- The right hand side of (2) is a linear function in α (say) $l(\alpha)$.

$$l(\alpha) = f(x_\alpha) + c_1 \alpha \nabla f_k^T p_k$$

- The function $l(\cdot)$ has a negative slope $c_1 \nabla f_k^T p_k$ but $c_1 \in (0, 1)$.
- Therefore, it lies above the graph of ϕ for small positive values of α .
- The sufficient decrease condition states that α is acceptable only if

$$\phi(\alpha) \leq l(\alpha).$$

- In practice, c_1 is chosen to be quite small, say

$$c_1 = 10^{-4}$$

The Wolfe Condition

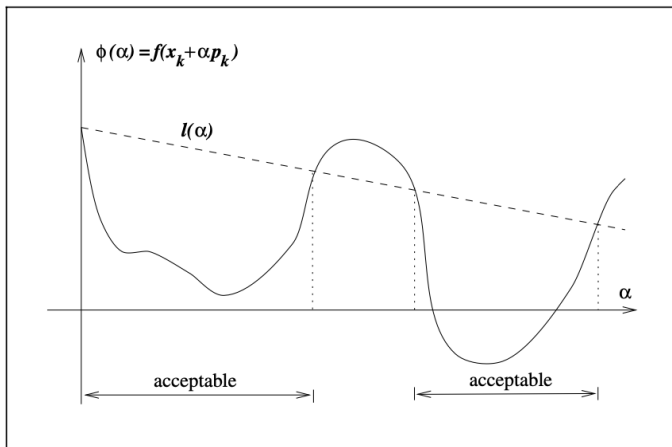


Figure: The intervals on which the Armijo condition is satisfied is shown

The Wolfe Condition

- The sufficient decrease condition is not enough by itself to ensure that the algorithm makes reasonable progress.
- As it is satisfied for all sufficiently small values of α

The Wolfe Condition

- To rule out unacceptable short steps we introduce a second requirement.

Curvature Conditions

α_k should satisfy

$$\nabla f(x_k + \alpha_k p_k)^T p_k \geq c_2 \nabla f_k^T p_k \quad (3)$$

for some constant $c_2 \in (c_1, 1)$.

- The left-hand side is simply the derivative $\phi'(\alpha_k)$.
- So the curvature condition ensures that the slope of ϕ at α_k is greater than c_2 times the initial slope $\phi'(0)$.
- If the **slope $\phi'(\alpha)$ is strongly negative**, we have an indication that **we can reduce f significantly** by moving further along the chosen direction.

The Wolfe Condition

- On the other hand if $\phi'(\alpha)$ is only slightly negative or even positive, it is a sign that we cannot expect much more decrease in f in this direction.
- So it makes sense to terminate the line search. (See Figure 6)
- Typical values of c_2 are 0.9 when the search direction p_k is chosen by a Newton or quasi-Newton method, and 0.1 when p_k is obtained from a non-linear conjugate gradient method.
- The sufficient decrease and curvature conditions are known collectively as the Wolfe conditions.

The Wolfe Condition

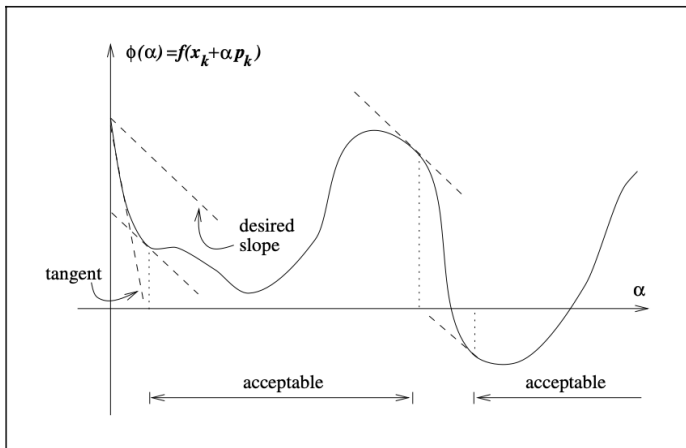


Figure: Insufficient Reduction

Wolfe Conditions

$$\begin{aligned} f(x_k + \alpha_k p_k) &\leq f(x_k) + c_1 \alpha_k \nabla f_k^T p_k \\ \nabla f(x_k + \alpha_k p_k)^T p_k &\geq c_2 \nabla f_k^T p_k. \end{aligned} \tag{4}$$

with $0 < c_1 < c_2 < 1$.

- A step length may satisfy the Wolfe conditions without being particularly close to a minimiser of ϕ . (See previous figure)
- The curvature conditions can be modified to force α_k to lie in at least a broad neighbourhood of a local minimiser or stationary point of ϕ .

α_k is required to satisfy

with $0 < c_1 < c_2 < 1$.

- The only difference with the Wolfe conditions is that we no longer allow the derivative $\phi'(\alpha)$ to be too positive.
- IT excludes points that are far from stationary points of ϕ .
- Is it always possible to find step lengths that satisfy Wolfe conditions ?

The Wolfe Condition

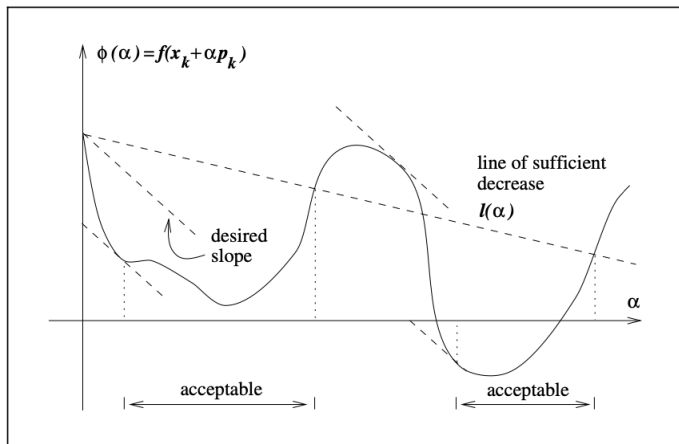


Figure: Step Lengths satisfying the Wolfe conditions.

Existence of α satisfying Wolfe conditions

Lemma

Suppose that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable. Let p_k be a descent direction at x_k , and assume that f is bounded below along the ray

$$\{x_k + \alpha p_k \mid \alpha > 0\}$$

Then if $0 < c_1 < c_2 < 1$, there exist intervals of step lengths satisfying Wolfe conditions and the strong Wolfe conditions.

Sketch of Proof



$$\phi(\alpha) = f(x_k + \alpha p_k)$$

is bounded below for all $\alpha > 0$

- Let $l(\alpha) = f(x_k) + \alpha c_1 \nabla f_k^T p_k$, the line is unbounded below and must intersect the graph of ϕ at least once.

Existence of α satisfying Wolfe conditions

- Note that for very small values if α (we can find such α)

$$\begin{aligned} l(\alpha) &= f(x_k) + \alpha c_1 \nabla f_k^T p_k \\ &> f(x_k) + \alpha \nabla f_k^T p_k \quad \text{as } \nabla f_k^T p_k < 0 \text{ and } c_1 < 1 \\ &\approx f(x_k + \alpha p_k) = \phi(\alpha). \end{aligned}$$

Therefore, to start with, the graph of $l(\alpha)$ stays above $\phi(\alpha)$.

- Now since $\phi(\alpha)$ is bounded below \exists a minimum value and since $l(\alpha)$ is unbounded below it will (for large values of α) attain values lesser than the minimum value of $\phi(\alpha)$.

Therefore, both the graphs will intersect atleast once.

- Let $\alpha' > 0$ be the smallest intersecting value of α that is

$$f(x_k + \alpha' p_k) = f(x_k) + \alpha' c_1 \nabla f_k^T p_k.$$

Existence of α satisfying Wolfe conditions

- α' is the point where the line $l(\alpha)$ meets $\phi(\alpha)$ for the first time . Therefore for all $\alpha < \alpha'$ the sufficient decrease condition holds good.
- Now by applying the mean value theorem on $\phi(\alpha)$ in the interval $[0, \alpha']$ we get

$$\begin{aligned}\frac{\phi(\alpha') - \phi(0)}{\alpha' - 0} &= \phi'(\alpha'') \quad \alpha'' \in (0, \alpha') \\ \implies f(x_k + \alpha' p_k) - f(x_k) &= \alpha' \nabla f(x_k + \alpha'' p_k)^T p_k \\ \implies f(x_k + \alpha' p_k) &= f(x_k) - \alpha' \nabla f(x_k + \alpha'' p_k)^T p_k\end{aligned}$$

$$\nabla f(x_k + \alpha'')^T p_k = c_1 \nabla f_k^T p_k > c_2 \nabla f_k^T p_k \text{ since } c_2 > c_1 \text{ and } \nabla f_k^T p_k <$$

- α'' satisfies the Wolfe conditions and the inequalities hold strictly for both the condition.

Existence of α satisfying Wolfe conditions

- Hence, by our smoothness assumption on f , there is an interval around α'' for which the Wolfe conditions hold.
- Moreover, since the left-hand side term in the curvature condition is negative, the strong Wolfe condition also holds in the same interval.

The Goldstein Conditions

The Goldstein conditions are stated as a pair of inequalities, in the following way:

$$f(x_k) + (1-c)\alpha_k \nabla f_k^T p_k \leq f(x_k + \alpha_k p_k) \leq f(x_k) + c\alpha_k \nabla f_k^T p_k, \quad (6)$$

with $0 < c < \frac{1}{2}$.

- The second inequality is the sufficient decrease condition.
- Whereas the first inequality is introduced to control the step length from below.
- A disadvantage of the Goldstein conditions vis-a-vis the Wolfe conditions is that the first inequality in (6) may exclude all minimizers of ϕ .
- However, the Goldstein and Wolfe conditions have much in common, and their convergence theories are quite similar.

The Goldstein Conditions

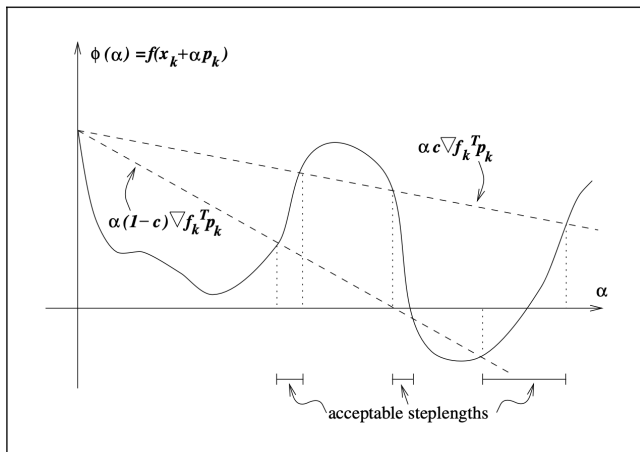


Figure: The Goldstein conditions.

Sufficient Decrease and Backtracking

- The sufficient decrease condition alone is not sufficient to ensure that the algorithm makes reasonable progress along the given search direction.
- However, the extra curvature condition can be dispensed off by using a so-called backtracking approach to choose candidate step length.

Backtracking Line Search

- 1 Choose $\bar{\alpha} > 0$, $\rho \in (0, 1)$, $c \in (0, 1)$;
- 2 Set $\alpha = \bar{\alpha}$
- 3 While $f(x_k + \alpha p_k) > f(x_k) + c\alpha \nabla f_K^T p_k$
- 4 $\alpha = \rho\alpha$;
- 5 end.

Terminate with $\alpha_k = \alpha$.

Sufficient Decrease and Backtracking

- The initial step length $\bar{\alpha}$ is chosen to be 1 in Newton and quasi-Newton methods, but can have different values in other algorithms, such as steepest descent or conjugate gradient.
- An acceptable step length will be found in a finite number of steps as α_k will eventually become small enough to satisfy the sufficient decrease condition.
- In practice the contraction factor " ρ " is allowed to vary at each iteration of the line search.
- One may need to ensure that $\rho \in [\rho_{lo}, \rho_{hi}]$ for some fixed constants $0 < \rho_{lo} < \rho_{hi} < 1$.

Sufficient Decrease and Backtracking

- The backtracking approach either choose $\alpha_k = \bar{\alpha}$ the initial choice or else α_k is short enough to satisfy the sufficient decrease condition.
- Still α_k is not very small as, $\frac{\alpha_k}{\rho}$ doesn't satisfy the sufficient decrease condition.
- It only by a factor of $\frac{1}{\rho}$ its shorter from the previous choice of α_k which doesn't work.
- It is a very simple and quite a popular strategy to terminate line search algorithms.
- Well suited for Newton methods but less appropriate for quasi-Newton and conjugate gradient methods.

Convergence of Line Search Methods

Global Convergence

$$||\nabla f_k|| \rightarrow 0 \quad \text{as } k \rightarrow \infty$$

i.e. convergence to a stationary point for any starting point x_0 .

To obtain global convergence:

- ① Need to choose step lengths well;
 - ② Choose search directions p_k appropriately as well.
- Let p_k be a chosen direction at the k th iteration of the line search method.
 - We define θ_k to be the angle between p_k and the steepest descent direction $-\nabla f_k$ given by

$$\cos \theta = \frac{-\nabla f_k^T p_k}{||\nabla f_k|| ||p_k||} \quad (7)$$

Global Convergence

Theorem

Zoutendijk 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. Suppose that f is bounded below in \mathbb{R}^n and that f is continuously differentiable in an open set \mathcal{N} containing the level set

$$\mathcal{L} \stackrel{\text{def}}{=} \{x : f(x) \leq f(x_0)\}$$

where x_0 is the starting point of the iteration. Assume also that the gradient " ∇f " is Lipschitz continuous on \mathcal{N} , i.e. there exists a constant $L > 0$ s.t.

$$\|\nabla f(x) - \nabla f(\tilde{x})\| \leq L\|x - \tilde{x}\|, \quad \text{for all } x, \tilde{x} \in \mathcal{N}$$

Then

$$\sum_{k \geq 0} \cos^2 \theta_k \|\nabla f_k\|^2 < \infty$$