1. Line Search Methods

Let $f : \mathbb{R}^n \to \mathbb{R}$ be given and suppose that x_c is our current best estimate of a solution to

$$\mathcal{P} \qquad \min_{x \in \mathbb{R}^n} f(x) \ .$$

A standard method for improving the estimate x_c is to choose a direction of search $d \in \mathbb{R}^n$ and the compute a step length $t^* \in \mathbb{R}$ so that $x_c + t^* d$ approximately optimizes f along the line $\{x + td \mid t \in \mathbb{R}\}$. The new estimate for the solution to $\mathcal{P}$ is then $x_+ = x_c + t^* d$. The procedure for choosing t^* is called a *line search method.* If t^* is taken to be the global solution to the problem

$$\min_{t \in \mathbb{R}} f(x_c + td) \ ,$$

then t^* is called the *Curry* step length. However, except in certain very special cases, the Curry step length is far too costly to compute. For this reason we focus on a few easily computed step lengths. We begin the simplest and the most commonly used line search method called backtracking.

1.1. Basic Backtracking Algorithm.

In the backtracking line search we assume that $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and that we are given a direction d of strict descent at the current point x_c, that is $f'(x_c; d) < 0$.

Initialization: Choose $\gamma \in (0, 1)$ and $c \in (0, 1)$.

Having x_c obtain x_+ as follows:

Step 1: Compute the backtracking stepsize

$$\begin{aligned} t^* \quad := \quad & \max \gamma^\nu \\ & \text{subject to } \nu \in \{0, 1, 2, \ldots\} \text{ and} \\ & f(x_c + \gamma^\nu d) \le f(x_c) + c\gamma^\nu f'(x_c; d). \end{aligned}$$

Step 2: Set $x_+ = x_c + t^* d$.

The backtracking line search method form the basic structure upon which most line search methods are built. Due to the importance of this method, we take a moment to emphasize its key features.

(1) The update to x_c has the form

$$x_+ = x_c + t^* d \ . \tag{1.1}$$

Here d is called the *search direction* while t^* is called the *step length* or *stepsize.*

(2) The search direction d must satisfy

$$f'(x_c; d) < 0.$$

Any direction satisfying this strict inequality is called a *direction of strict descent* for f at x_c. If $\nabla f(x_c) \neq 0$, then a direction of strict descent always exists. Just take $d = -\nabla f'(x_c)$. As we have already seen

$$f'(x_c; -\nabla f'(x_c)) = -\|\nabla f'(x_c)\|^2.$$

It is important to note that if d is a direction of strict descent for f at x_c, then there is a $\bar{t} > 0$ such that

$$f(x_c + td) < f(x_c) \quad \forall\ t \in (0, \bar{t}).$$

In order to see this recall that

$$f'(x_c; d) = \lim_{t \downarrow 0} \frac{f(x_c + td) - f(x_c)}{t}.$$

Hence, if $f'(x_c; d) < 0$, there is a $\bar{t}0$ such that

$$\frac{f(x_c + td) - f(x_c)}{t} < 0 \quad \forall\ t \in (0, \bar{t}),$$

that is

$$f(x_c + td) - f(x_c) \quad \forall\ t \in (0, \bar{t}).$$

(3) In Step 1 of the algorithm, we require that the step length t^* be chosen so that

$$f(x_c + t^* d) \leq f(x_c) + c\gamma^\nu f'(x_c; d). \tag{1.2}$$

This inequality is called the Armijo-Goldstein inequality. It is named after the two researchers to first use it in the design of line search routines (Allen Goldstein is a Professor Emeritus here at the University of Washington). Observe that this inequality guarantees that

$$f(x_c + t^* d) < f(x_c).$$

For this reason, the algorithm described above is called a *descent algorithm*. It was observed in point (2) above that it is always possible to choose t^* so that $f(x_c + t^* d) < f(x_c)$. But the Armijo-Goldstein inequality is a somewhat stronger statement. To see that it too can be satisfied observe that since $f'(x_c; d) < 0$,

$$\lim_{t \downarrow 0} \frac{f(x_c + td) - f(x_c)}{t} = f'(x_c; d) < cf'(x_c; d) < 0.$$

Hence, there is a $\bar{t} > 0$ such that

$$\frac{f(x_c + td) - f(x_c)}{t} \leq cf'(x_c; d) \quad \forall\ t \in (0, \bar{t}),$$

that is

$$f(x_c + td) \leq f(x_c) + tcf'(x_c; d) \quad \forall\ t \in (0, \bar{t}).$$

(4) The Armijo-Goldstein inequality is known as a condition of *sufficient decrease*. However, it is essential that we do not choose t^* too small. This is the reason for setting t^* equal to the first (largest) member of the geometric sequence $\{\gamma^\nu\}$ for which the Armijo-Goldstein inequality is satisfied. In general, we always wish to choose t^* as large as possible since it is often the case that some effort was put into the selection of the search direction d. Indeed, as we will see, for Newton's method we must take $t^* = 1$ in order to achieve rapid local convergence.

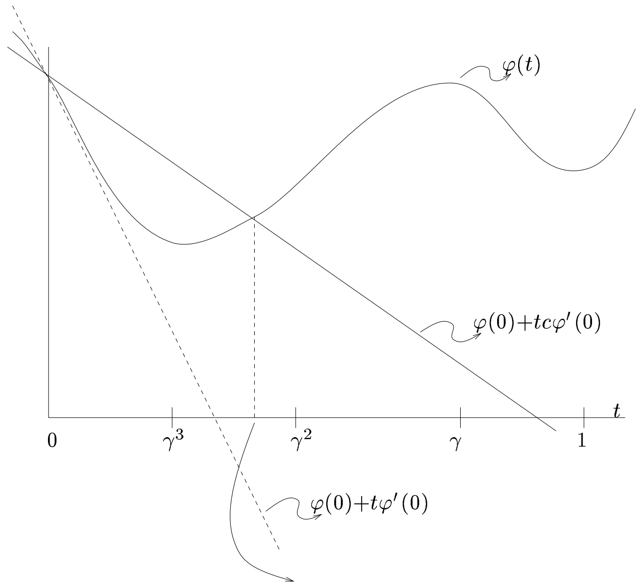
(5) There is a balance that must be struck between taking t^* as large as possible and not having to evaluating the function at many points. Such a balance is obtained with an appropriate selection of the parameters γ and c. Typically one takes $\gamma \in [.5, .8]$ while $c \in [.001, .1]$ with adjustments depending on the cost of function evaluation and degree of nonlinearity.

(6) The backtracking procedure of Step 1 is easy to program. A pseudo-Matlab code follows:

$$
\left[\begin{array}{lrcl}
 & f_c & = & f(x_c) \\
 & \Delta f & = & cf'(x_c; d) \\
 & \text{new}f & = & f(x_c + d) \\
 & t & = & 1 \\
\texttt{while} & \text{new}f & > & f_c + t\Delta f \\
 & t & = & \gamma t \\
 & \text{new}f & = & f(x_c + td) \\
\texttt{endwhile} & & &
\end{array}\right.
$$

Point (3) above guarantees that this procedure is finitely terminating.

(7) The backtracking procedure has a nice graphical illustration. Set $\varphi(t) = f(x_c + td)$ so that $\varphi(0) = f'(x_c; d)$.



$$t^* = \gamma^3, \; x_+ = x_c + \gamma^3 d.$$

Before proceeding to a convergence result for the backtracking algorithm, we consider some possible choices for the search directions d. There are essentially three directions of interest:

(1) Steepest Descent (or Cauchy Direction):

$$d = -\nabla f(x_c)/\|\nabla f(x_c)\| \; .$$

(2) Newton Direction:

$$d = -\nabla^2 f(x_c)^{-1} \nabla f(x_c) \; .$$

(3) Newton-Like Direction:

$$d = -H\nabla f(x_c),$$

where $H \in \mathbb{R}^{n\times n}$ is symmetric and constructed to approximate the inverse of the Hessian.

In order to base a descent method on these directions we must have

$$f'(x_c; d) < 0.$$

For the Cauchy direction $-\nabla f(x_c)/\|\nabla f(x_c)\|$, this inequality always holds when $\nabla f(x_c) \neq 0$;

$$f'(x_c; -\nabla f(x_c)/\|\nabla f(x_c)\|) = -\|\nabla f(x_c)\| < 0.$$

On the other hand the Newton and Newton-like directions do not always satisfy this property:

$$f'(x_c; -H\nabla f(x_c)) = -\nabla f(x_c)^T H\nabla f(x_c).$$

These directions are directions of strict descent if and only if

$$0 < \nabla f(x_c)^T H\nabla f(x_c) \ .$$

This condition is related to second-order sufficiency conditions for optimality when H is an approximation to the inverse of the Hessian.

The advantage of the Cauchy direction is that it always provides a direction of strict descent. However, once the iterates get "close" to a stationary point, the procedure takes a very long time to obtain a moderately accurate estimate of the stationary point. Most often numerical error takes over due to very small stepsizes and the iterates behave chaotically.

On the other hand, Newton's method (and its approximation, the secant method), may not define directions of strict descent until one is very close to a stationary point satisfying the second-order sufficiency condition. However, once one is near such a stationary point, then Newton's method (and some Newton-Like methods) zoom in on the stationary point very rapidly. This behavior will be made precise when we establish our convergence result from Newton's method.

Let us now consider the basic convergence result for the backtracking algorithm.

Theorem 1.1. (Convergence for Backtracking) *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $x_0 \in \mathbb{R}$ be such that f is differentiable on $\mathbb{R}^n$ with ∇f Lipschitz continuous on an open convex set containing the set $\{x : f(x) \le f(x_0)\}$. Let $\{x^k\}$ be the sequence satisfying $x^{k+1} = x^k$ if $\nabla f(x^k) = 0$; otherwise,*

$$x^{k+1} = x_k + t_k d^k, \quad \text{where } d^k \text{ satisfies } f'(x^k; d^k) < 0,$$

and t_k is chosen by the backtracking stepsize selection method. Then one of the following statements must be true:

(i) *There is a k_0 such that $\nabla f'(x^{k_0}) = 0$.*
(ii) $f(x^k) \searrow -\infty$
(iii) *The sequence $\{\|d^k\|\}$ diverges ($\|d^k\| \to \infty$).*
(iv) *For every subsequence $J \subset \mathbb{N}$ for which $\{d^k : k \in J\}$ is bounded, we have*

$$\lim_{k\in J} f'(x^k; d^k) = 0.$$

Before proving this Theorem, we first consider some important corollaries concerning the Cauchy and Newton search directions. Each corollary assumes the conditions of Theorem 1.1.

Corollary 1.1.1. *If the sequences $\{d^k\}$ and $\{f(x^k)\}$ are bounded, then*
$$\lim_{k\to\infty} f'(x^k; d^k) = 0.$$

Proof. The hypotheses imply that either (i) or (iv) with $J = \mathbb{N}$ occurs in Theorem 1.1. Hence, $\lim_{k\to\infty} f'(x^k; d^k) = 0$. □

Corollary 1.1.2. *If $d^k = -\nabla f'(x^k)/\|\nabla f(x^k)\|$ is the Cauchy direction for all k, then every accumulation point, $\overline{x}$, of the sequence $\{x^k\}$ satisfies $\nabla f(\overline{x}) = 0$.*

Proof. The sequence $\{f(x^k)\}$ is decreasing. If $\overline{x}$ is any accumulation point of the sequence $\{x^k\}$, then we claim that $f(\overline{x})$ is a lower bound for the sequence $\{f(x^k)\}$. Indeed, if this were not the case, then for some k_0 and $\epsilon > 0$
$$f(x^k) + \epsilon < f(\overline{x})$$
for all $k > k_0$ since $\{f(x^k)\}$ is decreasing. But $\overline{x}$ is a cluster point of $\{x^k\}$ and f is continuous. Hence, there is a $\widehat{k} > k_0$ such that
$$|f(\overline{x}) - f(x^{\widehat{k}})| < \epsilon/2.$$
But then
$$f(\overline{x}) < \frac{\epsilon}{2} + f(x^{\widehat{k}}) \quad \text{and} \quad f(x^{\widehat{k}}) + \epsilon < f(\overline{x}).$$
Hence,
$$f(x^{\widehat{k}}) + \epsilon < \frac{\epsilon}{2} + f(x^{\widehat{k}}), \quad \text{or} \quad \frac{\epsilon}{2} < 0.$$
This contradiction implies that $\{f(x^k)\}$ is bounded below by $f(\overline{x})$. But then the sequence $\{f(x^k)\}$ is bounded so that Corollary 1.1.1 applies. That is,
$$0 = \lim_{k\to\infty} f'\left(x^k; \frac{-\nabla f(x^k)}{\|\nabla f(x^k)\|}\right) = \lim_{k\to\infty} -\|\nabla f(x^k)\|.$$
Since ∇f is continuous, $\nabla f(\overline{x}) = 0$. □

Corollary 1.1.3. *Let us further assume that f is twice continuously differentiable and that there is a $\beta > 0$ such that, for all $u \in \mathbb{R}^n$, $\beta\|u\|^2 < u^T\nabla^2 f(x)u$ on $\{x : f(x) \le f(x^0)\}$. If the Basic Backtracking algorithm is implemented using the Newton search directions,*
$$d^k = -\nabla^2 f(x^k)^{-1}\nabla f(x^k),$$
then every accumulation point, $\overline{x}$, of the sequence $\{x^k\}$ satisfies $\nabla f(\overline{x}) = 0$.

Proof. Let $\overline{x}$ be an accumulation point of the sequence $\{x^k\}$ and let $J \subset \mathbb{N}$ be such that $x^k \xrightarrow{J} \overline{x}$. Clearly, $\{x^k : k \in J\}$ is bounded. Hence, the continuity of ∇f and $\nabla^2 f$, along with the Weierstrass Compactness Theorem, imply that the sets $\{\nabla f(x^k) : k \in J\}$ and $\{\nabla^2 f(x^k) : k \in J\}$ are also bounded. Let M_1 be a bound on the values $\{\|\nabla f(x^k)\| : k \in J\}$ and let M_2 be an upper bound on the values $\{\|\nabla^2 f(x^k)\| : k \in J\}$. Recall that by hypotheses

$\beta\|u\|^2$ is a uniform lower bound on the values $\{u^T\nabla^2 f(x^k)u\}$ for every $u \in \mathbb{R}^n$. Take $u = d^k$ to obtain the bound

$$\beta\|d^k\|^2 \le \nabla f(x^k)^T\nabla^2 f(x^k)^{-1}\nabla f(x^k) \le \|d^k\|\|\nabla f(x^k)\|,$$

and so

$$\|d^k\| \le \beta^{-1}M_1 \ \forall\, k \in J.$$

Therefore, the sequence $\{d^k : k \in J\}$ is bounded. Moreover, as in the proof of Corollary 1.1.2, the sequence $\{f(t_k)\}$ is also bounded. On the other hand,

$$\|\nabla f(x^k)\| = \|\nabla^2 f(x^k)d^k\| \le M_2\|d^k\| \ \forall\, k \in J.$$

Therefore,

$$M_2^{-1}\|\nabla f(x^k)\| \le \|d^k\| \ \forall\, k \in J.$$

Consequently, Theorem 1.1 Part (iv) implies that

$$\begin{aligned} 0 &= \lim_{k\in J} |f'(x^k; d^k)| \\ &= \lim_{k\in J} |\nabla f(x^k)^T\nabla^2 f(x^k)^{-1}\nabla f(x^k)| \\ &\ge \lim_{k\in J} \beta\|d^k\|^2 \\ &\ge \lim_{k\in J} \beta M_2^{-2}\|\nabla f(x^k)\|^2 \\ &= \beta M_2^{-2}\|\nabla f(\overline{x})\|^2. \end{aligned}$$

Therefore, $\nabla f(\overline{x}) = 0$. □

Proof of Theorem 1.1: We assume that none of (i), (ii), (iii), and (iv) hold and establish a contradiction.

Since (i) does not occur, $\nabla f(x^k) \ne 0$ for all $k = 1, 2, \ldots$. Since (ii) does not occur, the sequence $\{f(x^k)\}$ is bounded below. Since $\{f(x^k)\}$ is a bounded decreasing sequence in $\mathbb{R}$, we have $f(x^k) \searrow \overline{\varphi}$ for some $\overline{f}$. In particular, $(f(x^{k+1}) - f(x^k)) \to 0$. Next, since (iii) and (iv) do not occur, there is a subsequence $J \subset \mathbb{N}$ and a vector $\overline{d}$ such that $d^k \xrightarrow{J} \overline{d}$ and

$$\sup_{k\in J} f'(x^k; d^k) =: \beta < 0.$$

The Armijo-Goldstein inequality combined with the fact that $(f(x^{k+1}) - f(x^k)) \to 0$, imply that

$$t_k f'(x^k; d^k) \to 0.$$

Since $f'(x^k; d^k) \le \beta < 0$ for $k \in J$, we must have $t_k \xrightarrow{J} 0$. With no loss in generality, we assume that $t_k < 1$ for all $k \in J$. Hence,

$$c\gamma^{-1}t_k f'(x^k; d^k) < f(x^k + t_k\gamma^{-1}d^k) - f(x^k) \tag{1.3}$$

for all $k \in J$ due to Step 1 of the line search and the fact that $\tau_k < 1$. By the Mean Value Theorem, there exists for each $k \in J$ a $\theta_k \in (0, 1)$ such that

$$f(x^k + t_k\gamma^{-1}d^k) - f(x^k) = t_k\gamma^{-1}f'(\widehat{x}^k; d^k)$$

where

$$\begin{aligned}\widehat{x}^n &:= (1-\theta_k)x^k + \theta_k(x^k + t_k\gamma^{-1}d^k)\\ &= x^k + \theta_k t_k \gamma^{-1} d^k.\end{aligned}$$

Now, since ∇f is Lipschitz continuous, we have

$$\begin{aligned}f(x^k + t_k\gamma^{-1}d^k) - f(x^k) &= t_k\gamma^{-1}f'(\widehat{x}^k; d^k)\\ &= t_k\gamma^{-1}f'(x^k; d^k) + t_k\gamma^{-1}[f'(\widehat{x}^k; d^k) - f'(x^k; d^k)]\\ &= t_k\gamma^{-1}f'(x^k; d^k) + t_k\gamma^{-1}[\nabla f(\widehat{x}^k) - \nabla f(x^k)]^T d^k\\ &\leq t_k\gamma^{-1}f'(x^k; d^k) + t_k\gamma^{-1}L\|\widehat{x}^k - x^k\|\|d^k\|\\ &= t_k\gamma^{-1}f'(x^k; d^k) + L(t_k\gamma^{-1})^2\theta_k\|d^k\|^2.\end{aligned}$$

Combining this inequality with inequality (1.3) yields the inequality

$$ct_k\gamma^{-1}f'(x^k; d^k) < t_k\gamma^{-1}f'(x^k; d^k) + L(t_k\gamma^{-1})^2\theta_k\|d^k\|^2.$$

By rearranging and then substituting β for $f'(x^k; d^k)$ we obtain

$$0 < (1-c)\beta + (t_k\gamma^{-1})L\|\delta_k\|^2 \quad \forall\ k \in J.$$

Now taking the limit over $k \in J$, we obtain the contradiction

$$0 \leq (1-c)\beta < 0.$$

■

1.2. **The Wolfe Conditions.** We now consider a couple of modifications to the basic backtracking line search that attempt to better approximate an exact line-search (Curry line search), i.e. the stepsize t_k is chosen to satisfy

$$f(x^k + t_k d^k) = \min_{t\in\mathbb{R}} f(x^k + td^k).$$

In this case, the first-order optimality conditions tell us that $0 = \nabla f(x^k + t_k d^k)^T d^k$. The Wolfe conditions try to combine the Armijo-Goldstein sufficient decrease condition with a condition that tries to push $\nabla f(x^k + t_k d^k)^T d^k$ either toward zero, or at least to a point where the search direction d^k is less of a direction of descent. To describe these line search conditions, we take parameters $0 < c_1 < c_2 < 1$.

Weak Wolfe Conditions

$$f(x^k + t_k d^k) \leq f(x^k) + c_1 t_k f'(x^k; d^k) \tag{1.4}$$

$$c_2 f'(x^k; d^k) \leq f'(x^k + t_k d^k; d^k)\ . \tag{1.5}$$

Strong Wolfe Conditions

$$f(x^k + t_k d^k) \leq f(x^k) + c_1 t_k f'(x^k; d^k) \tag{1.6}$$

$$|f'(x^k + t_k d^k; d^k)| \leq c_2|f'(x^k; d^k)|\ . \tag{1.7}$$

The weak Wolfe condition (1.5) tries to make d^k less of a direction of descent (and possibly a direction of ascent) at the new point, while the strong Wolfe condition tries to push the directional derivative in the direction d^K closer to zero at the new point. Imposing one or the other of the Wolfe conditions on a line search procedure has become standard practice for optimization software based on line search methods.

We now give a result showing that there exists stepsizes satisfying the weak Wolfe conditions. A similar result (with a similar proof) holds for the strong Wolfe conditions.

Lemma 1.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and suppose that $x, d \in \mathbb{R}^n$ are such that the set $\{f(x+td) : t \geq 0\}$ is bounded below and $f'(x;d) < 0$, then for each $0 < c_1 < c_2 < 1$ the set*

$$\left\{ t \;\middle|\; \begin{array}{c} t > 0, f'(x+td;d) \geq c_2 f'(x;d), \text{ and} \\ f(x+td) \leq f(x) + c_1 t f'(x;d) \end{array} \right\}$$

has non–empty interior.

Proof. Set $\phi(t) = f(x+td) - (f(x) + c_1 t f'(x;d))$. Then $\phi(0) = 0$ and $\phi'(0) = (1-c_1)f'(x;d) < 0$. So there is a $\bar{t} > 0$ such that $\phi(t) < 0$ for $t \in (0, \bar{t})$. Moreover, since $f'(x;d) < 0$ and $\{f(x+td) : t \geq 0\}$ is bounded below, we have $\phi(t) \to +\infty$ as $t \uparrow \infty$. Hence, by the continuity of f, there exists $\hat{t} > 0$ such that $\phi(\hat{t}) = 0$. Let $t^* = \inf\left\{\hat{t} \;\middle|\, 0 \leq t,\ \phi(\hat{t}) = 0\right\}$. Since $\phi(t) < 0$ for $t \in (0, \bar{t})$, $t^* > 0$ and by continuity $\phi(t^*) = 0$. By Rolle's theorem (or the mean value theorem) there must exist $\tilde{t} \in (0, t^*)$ with $\phi'(\tilde{t}) = 0$. That is,

$$\nabla f(x + \tilde{t}d)^T d = c_1 \nabla f(x)^T d > c_2 \nabla f(x)^T d.$$

From the definition of t^* and the fact that $\tilde{t} \in (0, t^*)$, we also have

$$f(x + td) - (f(x) + c_1 \tilde{t} \nabla f(x)^T d) < 0\ .$$

The result now follows from the continuity of f and ∇f. $\square$

We now describe a bisection method that either computes a stepsize satisfying the weak Wolfe conditions or sends the function values to $-\infty$. Let x and d in $\mathbb{R}^n$ be such that $f'(x;d) < 0$.

A Bisection Method for the Weak Wolfe Conditions

Initialization: Choose $0 < c_1 < c_2 < 1$, and set $\alpha = 0$, $t = 1$, and $\beta = +\infty$.

Repeat
- If $f(x+td) > f(x) + c_1 t f'(x;d)$,
  - set $\beta = t$ and reset $t = \frac{1}{2}(\alpha + \beta)$.
- Elseif $f'(x+td;d) < c_2 f'(x;d)$,
  - set $\alpha = t$ and reset

$$t = \begin{cases} 2\alpha, & \text{if } \beta = +\infty \\ \frac{1}{2}(\alpha+\beta), & \text{otherwise.} \end{cases}$$

- Else, STOP.

End Repeat

In the above procedure, a stepsize satisfying the weak Wolfe conditions is always bounded below by α and bounded above by β. If the Armijo-Goldstein condition (1.4) fails, then the current stepsize t is too large so we set the new upper bound to t ($\beta = t$), and compute a new guess at the stepsize as the midpoint of the upper and lower bounds. If the Armijo-Goldstein condition is satisfied but the second Wolfe condition (1.5) is not, then the current stepsize t is too small so we raise the lower bound to t ($\alpha = t$), and compute a new guess at the stepsize as the midpoint of the upper and lower bounds when possible; otherwise, we set it to twice the current stepsize.

Observe that if we never set β to a finite value, then the Armijo-Goldstein condition is always satisfied and either we double the stepsize at every iteration, or we terminate at a point satisfying both Wolfe conditions (1.4) and (1.5). If, in this case, we do not terminate, then $t \uparrow +\infty$ and the function values $f(x+td)$ are driven to $-\infty$. On the other hand, if β is set to a finite value at some iteration, then from that iteration forward the length of the interval $[\alpha, \beta]$ is halved at each iteration. Since this interval necessarily contains an open subinterval containing points satisfying both of the Wolfe conditions (1.4) and (1.5), the procedure must terminate finitely.

A global convergence result for a line search routine based on the Weak Wolfe conditions now follows.

Theorem 1.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$, $x^0 \in \mathbb{R}^n$, and $0 < c_1 < c_2 < 1$. Assume that $\nabla f(x)$ exists and is Lipschitz continuous on an open set containing the set $\{x \mid f(x) \le f(x^0)\}$. Let $\{x^\nu\}$ be a sequence initiated at x^0 and generated by the following algorithm:*

Step 0: *Set $k = 0$.*

Step 1: *Choose $d^k \in \mathbb{R}^n$ such that $f'(x^k; d^k) < 0$.*
If no such d^k exists, then STOP.
First-order necessary conditions for optimality are satisfied at x^k.

Step 2: *Let t^k be a stepsize satisfying the Weak Wolfe conditions (1.4) and (1.5).*
If no such t^k exists, then STOP.
The function f is unbounded below.

Step 3: *Set $x^{k+1} = x^k + t_k d^k$, reset $k = k+1$, and return to Step 1.*

One of the following must occur:

(i) *The algorithm terminates finitely at a first-order stationary point for f.*
(ii) *For some k the stepsize selection procedure generates a sequence of trial stepsizes $t_{k\nu} \uparrow +\infty$ such that $f(x^k + t_{k\nu} d^k) \to -\infty$.*
(iii) $f(x^k) \downarrow -\infty$.
(iv) $\displaystyle\sum_{k=0}^{\infty} \|\nabla f(x^k)\|^2 \cos^2\theta_k < +\infty$, *where* $\cos\theta_k = \dfrac{\nabla f(x^k)^T d^k}{\|\nabla f(x^k)\| \|d^k\|}$ *for all* $k = 1, 2, \ldots$.

Proof. We assume that (i), (ii), and (iii) do not occur and show that (iv) occurs. Since (i) and (ii) do not occur the sequence $\{x^\nu\}$ is infinite and $f'(x^k; d^k) < 0$ for all $k = 1, 2, \ldots$. Since (ii) does not occur, the weak Wolfe conditions are satisfied at every iteration. The condition (1.4) implies that the sequence $\{f(x^k)\}$ is strictly decreasing. In particular, this implies that $\{x^\nu\} \subset \{x \mid f(x) \le f(x^0)\}$. The condition (1.5) implies that

$$(c_2 - 1)\nabla f(x^k)^T d^k \le (\nabla f(x^{k+1}) - \nabla f(x^k))^T d^k$$

for all k. Combining this with the Lipschitz continuity of ∇f on an open neighborhood of $\{x \mid f(x) \le f(x^0)\}$, gives

$$(c_2 - 1)\nabla f(x^k)^T d^k \le (\nabla f(x^{k+1}) - \nabla f(x^k))^T d^k \le L t_k \|d^k\|^2 .$$

Hence

$$t_k \geq \frac{c_2 - 1}{L} \frac{\nabla f(x^k)^T d^k}{\|d^k\|^2} > 0.$$

Plugging this into (1.4) give the inequality

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

Setting $c = c_1 \frac{1-c_2}{L}$ and summing over k gives

$$f(x^{k+1}) \leq f(x^0) - c \sum_{\nu=0}^{k} \|\nabla f(x^\nu)\|^2 \cos^2 \theta_\nu \ .$$

Since (iii) does not occur, we can take the limit in k and obtain

$$\sum_{\nu=0}^{\infty} \|\nabla f(x^\nu)\|^2 \cos^2 \theta_\nu < +\infty \, .$$

□

If the function f is bounded below and the algorithm does not terminate finitely, then Part (iv) of this theorem states that

$$\|\nabla f(x^k)\| \cos^2 \theta_k \to 0 \ .$$

Hence, if the search directions d^k are chosen so that there is a $\delta > 0$, independent of the iteration k, such that $\cos \theta_k < -\delta$ for all k, then it must be the case that $\|\nabla f(x^k)\| \to 0$ so that every cluster point of the sequence $\{x^k\}$ is a first-order stationary point for f. For example, we have the following corollary to the theorem.

Corollary 1.2.1. *Let f and $\{x^k\}$ be as in the theorem, and let $\{B_k\}$ be a sequence of symmetric positive definite matrices for which there exists $\overline{\lambda} > \underline{\lambda} > 0$ such that*

$$\underline{\lambda} \|u\|^2 \leq u^T B_k u \leq \overline{\lambda} \|u\|^2 \ \forall \, u \in \mathbb{R}^n \ \text{and} \ k = 1, 2, \ldots \ . \tag{1.8}$$

Let us further assume that f is bounded below. If the search directions d^k are given by

$$d^k = -B_k \nabla f(x^k) \ \forall \, k = 1, 2, \ldots \ ,$$

then $\nabla f(x^k) \to 0$.

Proof. It is easily shown (see exercises) that the condition (1.8) implies that the eigenvalues of the sequence $\{B_k\}$ are uniformly lower bounded by $\underline{\lambda}$ and uniformly upper bounded by $\overline{\lambda}$. In particular, this implies that

$$\underline{\lambda} \|u\| \leq \|B_k u\| \leq \overline{\lambda} \|u\| \ \forall \, u \in \mathbb{R}^n \ \text{and} \ k = 1, 2, \ldots$$

(see exercises). Hence for all k

$$
\begin{aligned}
\cos\theta_k &= \frac{\nabla f(x^k)^T d^k}{\|\nabla f(x^k)\|\|d^k\|} \\
&= -\frac{\nabla f(x^k)^T B_k \nabla f(x^k)}{\|\nabla f(x^k)\|\|B_k \nabla f(x^k)\|} \\
&\leq -\frac{\underline{\lambda}\|\nabla f(x^k)\|^2}{\|\nabla f(x^k)\|\|B_k \nabla f(x^k)\|} \\
&\leq -\frac{\underline{\lambda}\|\nabla f(x^k)\|^2}{\|\nabla f(x^k)\|\overline{\lambda}\|\nabla f(x^k)\|} \\
&= -\underline{\lambda}/\overline{\lambda} \\
&< 0\,.
\end{aligned}
$$

Therefore $\nabla f(x^k) \to 0$. □

A possible choice for the matrices B_k in the above result is $B_k = I$ for all k. This essentially gives the method of steepest descent.

Exercises for Chapter on Line Search Methods

(1) Let Q be an $n \times n$ symmetric positive definite matrix.
  (a) Show that the eigenvalues of Q^2 are the square of the eigenvalues of Q.
  (b) If $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ are the eigen values of Q, show that
$$\lambda_n \|u\|_2^2 \leq u^T Q u \leq \lambda_1 \|u\|_2^2 \ \forall \ u \in \mathbb{R}^n .$$
  (c) If $0 < \underline{\lambda} < \overline{\lambda}$ are such that
$$\underline{\lambda} \|u\|_2^2 \leq u^T Q u \leq \overline{\lambda} \|u\|_2^2 \ \forall \ u \in \mathbb{R}^n ,$$
  then all of the eigenvalues of Q must lie in the interval $[\underline{\lambda}, \overline{\lambda}]$.
  (d) Let $\underline{\lambda}$ and $\overline{\lambda}$ be as in Part (c) above. Show that
$$\underline{\lambda} \|u\|_2 \leq \|Bu\|_2 \leq \overline{\lambda} \|u\|_2 \ \forall \ u \in \mathbb{R}^n \ .$$
  Hint: $\|Qu\|_2^2 = u^T Q^2 u$.

(2) Let Q be an $n \times n$ symmetric positive definite matrix, $g \in \mathbb{R}^n$, and $\alpha \in \mathbb{R}$. Consider the quadratic function
$$f(x) = \frac{1}{2} x^T Q x + g^T x + \alpha \ .$$
  (a) Show that there exists a $\underline{\lambda} > 0$ such that
$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\underline{\lambda}}{2} \|y - x\|_2^2$$
  for all $x, y \in \mathbb{R}^n$.
  (b) Show that f is a strictly convex function.
  (c) Given x and d in $\mathbb{R}^n$ compute the solution t^* to the one dimensional optimization problem
$$\min_{t \in \mathbb{R}^n} f(x + td)$$
  (you must also verify that it is the unique global solution). This is the Curry stepsize for this function.

(3) Let $f : \mathbb{R}^m \to \bar{\mathbb{R}}$ be convex, $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^m$. Show that the function $h : \mathbb{R}^n \to \bar{\mathbb{R}}$ defined by
$$h(y) = f(Ay + b)$$
is also a convex function.

(4) Let $M \in \mathbb{R}^{m \times n}$.
  (a) Show that the matrices $M^T M$ and $M M^T$ are always symmetric and positive semi-definite.
  (b) Provide a necessary and sufficient condition on the matrix M for the matrix $M^T M$ to be positive definite.

(5) Let $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ and consider the function
$$f(x) = \frac{1}{2} \|Ax + b\|_2^2 \ .$$
  (a) Show that f is a convex function.

(b) Provide a necessary and sufficient condition on the matrix A for f to be strictly convex.

(c) Assume that A satisfies the condition that you have identified in Part (b). Given x and d in $\mathbb{R}^n$ compute the solution t^* to the one dimensional optimization problem

$$\min_{t\in\mathbb{R}} f(x+td) \ .$$

(d)* Show that a solution to the problem $\min_{x\in\mathbb{R}^n} f(x)$ must always exist for every $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$.

(6) Provide an example to show that the set

$$\left\{ t \ \middle| \ \begin{array}{c} t > 0, f'(x+td;d) \geq c_2 f'(x;d), \text{ and} \\ f(x+td) \leq f(x) + c_1 t f'(x;d) \end{array} \right\}$$

may be empty if $0 < c_2 < c_1 < 1$.

(7)* Let $f : \mathbb{R}^n \to \mathbb{R}$ and $x_0 \in \mathbb{R}$ be such that f is differentiable on $\mathbb{R}^n$ with ∇f Lipschitz continuous on an open convex set containing the set $\{x : f(x) \leq f(x_0)\}$. Let $\{H_k\}$ be a sequence of symmetric positive definite matrices for which there exists $\overline{\lambda} > \underline{\lambda} > 0$ such that

$$\underline{\lambda}\|u\|^2 \leq u^T H_k u \leq \overline{\lambda}\|u\|^2 \ \forall\, u \in \mathbb{R}^n \text{ and } k = 1, 2, \ldots \ .$$

Finally, let $\{x^k\}$ be the sequence satisfying $x^{k+1} = x^k$ if $\nabla f(x^k) = 0$; otherwise,

$$x^{k+1} = x_k + t_k d^k, \quad \text{where } d^k = -H_k \nabla f(x^k),$$

and t_k is chosen by the backtracking stepsize selection method. Show that every accumulation point, $\overline{x}$, of the sequence $\{x^k\}$ satisfies $\nabla f(\overline{x}) = 0$.