BACKTRACKING LINE SEARCH

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

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

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 $\{x + td \,|\, t \in \mathbb{R}\}$. The new estimate for the solution to $\mathcal{P}$ is then $x_n = x_c + t^* d$. 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.

2. The 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_n as follows:

Step 1: Compute the backtracking stepsize

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

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

The backtracking line search method forms 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

$$x_n = x_c + t^* d \ . \tag{1}$$

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

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

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

$$f(x_c + t^* d) \le f(x_c) + c\gamma^\nu f'(x_c;d). \tag{2}$$

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

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

that is

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

(4) The Armijo-Goldstein inequality is known as a condition of *sufficient decrease*. 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.

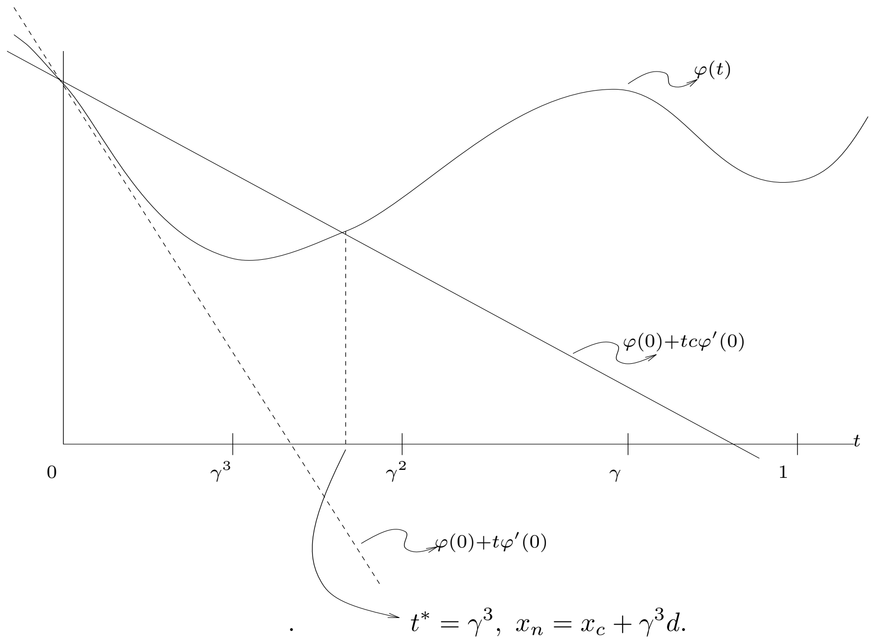
(5) There is a balance that must be struck between taking t^* as large as possible and not having to evaluating the function at many points. Such a balance is obtained with an appropriate selection of the parameters γ and c. Typically one takes $\gamma \in [.5, .8]$ while $c \in [.001, .1]$ with adjustments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

```
      $f_c = f(x_c)$
      $\Delta f = cf'(x_c;d)$
      $\text{new}f = f(x_c + d)$
      $t = 1$
while $\text{new}f > f_c + t\Delta f$
      $t = \gamma t$
      $\text{new}f = f(x_c + td)$
endwhile
```

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 $\varphi'(0) = f'(x_c;d)$.



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

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):

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

(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 $\nabla^2 f(x_c)$.

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

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

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

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

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 2.1. (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

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

Remark 2.1. *It is important to note that this theorem says nothing about the convergence of the sequence $\{x^k\}$. Indeed, this sequence may diverge. The theorem only concerns the function values and the first-order necessary condition for optimality.*

Before proving this Theorem, we first consider some important corollaries concerning the Cauchy and Newton search directions. Each corollary assumes that the hypotheses of Theorem 2.1 hold.

Corollary 2.1.1. *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 2.1. Hence, $\lim_{k \to \infty} f'(x^k; d^k) = 0$. □

Corollary 2.1.2. *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 2.1.1 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 2.1.3. *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

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

and so

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

Therefore, the sequence $\{d^k : k \in J\}$ is bounded. Moreover, as in the proof of Corollary 2.1.2, the sequence $\{f(t_k)\}$ is also bounded. On the other hand,

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

Therefore,

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

Consequently, Theorem 2.1 Part (iv) implies that

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

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

Proof of Theorem 2.1: We assume that none of (i), (ii), (iii), and (iv) hold and establish a contradiction.

Since (i) does not occur, $\nabla f(x^k) \neq 0$ 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 $f(x^k) \searrow \overline{f}$ 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,

$$c\gamma^{-1} t_k f'(x^k; d^k) < f(x^k + t_k \gamma^{-1} d^k) - f(x^k) \tag{3}$$

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

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

Now, since ∇f is Lipschitz continuous, we have

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

Combining this inequality with inequality (3) yields the inequality

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

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

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

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

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

□