Survey of Several Nonlinear Optimization Algorithms

Liang Sun
sun.liang@asu.edu
Arizona State University

September 16th, 2008

Outline

1. Gradient Method
2. Newton Method
3. Gradient Method and Newton Method: What is Different?
4. Conjugate Gradient
5. Constrained Minimization
    - Penalty Function Method
    - Barrier Function Method

Gradient Method

- General scheme of gradient method
  1. Choose $x_0 \in \boldsymbol{R}^n$.
  2. Iterate $x_{k+1} = x_k - h_k f'(x_k), k = 0, 1, \cdots$, where h_k is called the *step size*.
- There are many variants of gradient method, the difference lies in the strategy to determine the step size.

Different Strategies to Determine the Step Size

1. The sequence $\{h_k\}_{k=0}^{\infty}$ is chosen *in advance*, before the gradient method starts its job

   For example,

$$\begin{aligned} h_k &= h > 0; \qquad \text{(constant step)} \\ h_k &= \frac{h}{\sqrt{k+1}}. \end{aligned}$$

2. Full relaxation:

$$h_k = \arg \min_{h \geq 0} f(x_k - hf'(x_k)).$$

3. Goldstein-Armijo condition: Find $x_{k+1} = x_k - hf'(x_k)$ such that

$$\begin{aligned} \alpha \langle f'(x_k), x_k - x_{k+1} \rangle &\leq f(x_k) - f(x_{k+1}) \\ (1-\alpha) \langle f'(x_k), x_k - x_{k+1} \rangle &\geq f(x_k) - f(x_{k+1}), \end{aligned}$$

   where $0 < \alpha < \frac{1}{2}$.

Comments on Different Strategies

- The first strategy is simple, although it is often used only in convex optimization where the behavior of the functions is much more predictable than in the general nonlinear case.
- The second strategy, i.e., the relaxed strategy, is never used in practice since even in one-dimensional case we cannot find an exact minimum of a function in finite time.
- The third strategy is used in the majority of the practical algorithms.
- To investigate the convergence of different strategies, next we prove that for all strategies, the following statement is true:

  $$f(x_k) - f(x_{k+1}) \geq \frac{\omega}{L}\|f'(x_k)\|^2, \tag{1}$$

  where $x_{k+1} = x_k - h_k f'(x_k)$, L is the Lipschitz constant, and ω is some positive constant.

More Comments on Goldstein-Armijo Condition

- The requirement that $f(x_{k+1}) < f(x_k)$ is not sufficient to ensure that the sequence $\{x_k\}$ converges to a minimizer of f.
- An example of insufficient step
  Consider function $f = \frac{1}{2}x^2$ with $h_k = \frac{1}{2^{k+1}}$ starting at the point $x_0 = 2$. We obtain $f'(x) = x$, and $x_{k+1} = x_k - h_k x_k = x_k - \frac{1}{2^{k+1}} x_k$. Thus $f(x_{k+1}) < f(x_k)$ is guarantied. The sequence generated is:

$$2, 1, \frac{3}{4}, \frac{27}{32} \to \bar{x} \approx 0.5776$$

- In this example, although f is decreased at each step, h_k eventually becomes so small that the decrease in f converges to zero before the gradient has time to converge to zero.
- The difficulty can be overcome if at each step the step size h_k is chosen to produce a sufficient decrease in f.

More Comments on Goldstein-Armijo Condition

- In the Armijo condition, we ensure the minimal decrease of f:

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq \alpha\langle f'(x_k), x_k - x_{k+1}\rangle \\ &= \alpha h_k\langle f'(x_k), f'(x_k)\rangle = \alpha h_k\|f'(x_k)\|^2. \end{aligned}$$

- The minimal decrease of f is proportional to the step size h_k and $\|f'(x_k)\|^2$
- Denote $\ell_1(h)$ as follows:

$$\ell_1(h) = f(x_k) - \alpha h\|f'(x_k)\|^2.$$

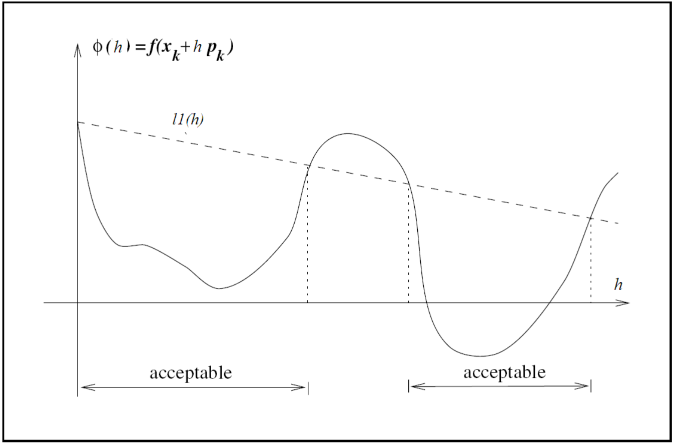
- Denote $\phi(h)$ as:

$$\phi(h) = f(x_k - hf'(x_k)) = f(x_{k+1}).$$

- Thus the Armijo condition can be formulated equivalently as:

$$\phi(h_k) \leq \ell_1(h_k).$$

An Illustration of Armijo Condition



Goldstein Condition

- The sufficient decrease condition is not enough by itself to ensure the algorithm to make reasonable progress.
  In the last example, it is satisfied for all sufficiently small values of h_k.
- The Goldstein condition prevents h_k from being too small:

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\le (1-\alpha)\langle f'(x_k), x_k - x_{k+1}\rangle \\ &= (1-\alpha)h_k\langle f'(x_k), f'(x_k)\rangle = (1-\alpha)h_k\|f'(x_k)\|^2. \end{aligned}$$

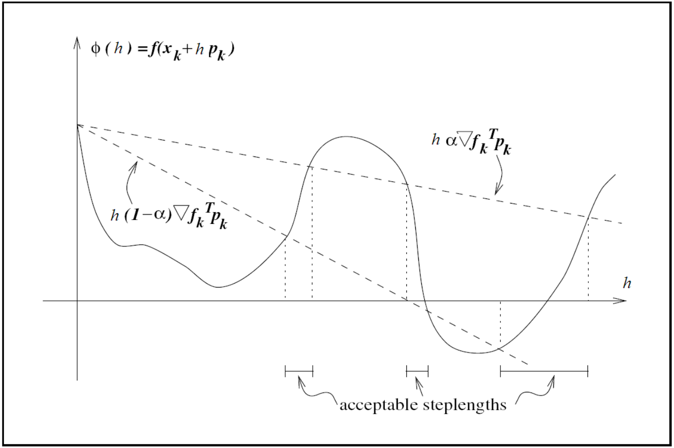
- Denote $\ell_2(h)$ as follows:

$$\ell_2(h) = f(x_k) - (1-\alpha)h\|f'(x_k)\|^2.$$

- Thus the Goldstein-Armijo condition can be stated as:

$$\ell_2(h_k) \le \phi(h_k) \le \ell_1(h_k).$$

An Illustration of Goldstein-Armijo Condition



Goldstein-Armijo Condition

Theorem

Suppose that $f : \boldsymbol{R}^n \to \boldsymbol{R}$ *is continuously differentiable, and assume that* f *is bounded below along the ray* $\{x_k - h_k f'(x_k) | h > 0\}$. *Then if* $0 < \alpha < \frac{1}{2}$, *there exist intervals of step sizes satisfying the Goldstein-Armijo Condition.*

Proof.

Using the Taylor expansion, we have

$$\phi(h) = f(x_k - hf'(x_k)) = f(x_k) - h\|f'(x_k)\|^2 + o(|hf'(x_k)|) < f(x_k) - \beta h\|f'(x_k)$$

for $0 < \beta < 1$ and small values of h. Thus for small values of h, we have

$$\phi(h) < \ell_1(h) \text{ and } \phi(h) < \ell_2(h).$$

Note that $\phi(h)$ is bounded below for all $h > 0$ and since $0 < \alpha < \frac{1}{2}$, the line $\ell_1(h)$ is a decreasing linear function for $h > 0$. Thus $\ell_1(h)$ must intersect the graph of $\phi(h)$ at least once.

Goldstein-Armijo Condition

proof continued...

Let h' be the smallest intersection value of h, that is,

$$\phi(h') = f(x_k - h'f'(x_k)) = f(x_k) - \alpha h' \|f'(x_k)\|^2 = \ell_1(h'). \tag{2}$$

Since h' is the smallest value of h satisfying Eq. (2), thus the Armijo condition holds for the interval $(0, h']$. Note that $\ell_1(h) > \ell_2(h)$ for all $h > 0$. Thus, we have

$$\phi(h') = f(x_k - h'f'(x_k)) = \ell_1(h') > \ell_2(h')$$

Also note that $\ell_2(h)$ is also a decreasing linear function for $h > 0$. Therefore, $\ell_2(h)$ must intersect the graph of $\phi(h)$ at least once in the interval $(0, h')$. Let h'' be the last intersection value of h in this interval $(0, h')$. Thus, $\forall h \in [h'', h']$, we have

$$\ell_2(h) \leq \phi(h) \leq \ell_1(h).$$

Convergence Analysis

- Assumption

  We consider the problem

  $$\min_{x \in \boldsymbol{R}^n} f(x),$$

  with $f \in C_L^{1,1}(\boldsymbol{R}^n)$, and let us assume that $f(x)$ is bounded below on $\boldsymbol{R}$.
- According to Lemma 1.2.3, for $y = x - hf'(x)$, we have

  $$\begin{aligned} f(y) &\leq f(x) + \langle f'(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 \\ &= f(x) - h\|f'(x)\|^2 + \frac{h^2}{2}L\|f'(x)\|^2 \\ &= f(x) - h(1 - \frac{h}{2}L)\|f'(x)\|^2 \\ &= f(x) + \Delta(h)\|f'(x)\|^2, \end{aligned}$$

  where $\Delta(h) = -h(1 - \frac{h}{2}L)$

Convergence Analysis

- In order to get the best estimate for the possible decrease of the objective function, we have to minimize $\Delta(h)$.
- It turns out that $h^* = \frac{1}{L}$, and $\Delta(h^*) = -\frac{1}{2L}$.
- Thus, we prove that one step of the gradient method can decrease the objective function as follows

$$f(y) \leq f(x) - \frac{1}{2L}\|f'(x)\|^2.$$

- Next we go back to the three different strategies.

Convergence Analysis – Constant Step Strategy

In constant step strategy, $h_k = h$, thus we have

$$f(x_k) - f(x_{k+1}) \geq h(1 - \frac{h}{2}L)\|f'(x)\|^2.$$

Therefore, if we choose $h_k = \frac{2\alpha}{L}$ with $\alpha \in (0, 1)$, then

$$f(x_k) - f(x_{k+1}) \geq \frac{2}{L}\alpha(1 - \alpha)\|f'(x)\|^2.$$

Let $\omega = 2\alpha(1 - \alpha) > 0$.
It turns out that the optimal choice is $h_k = \frac{1}{L}$.

Convergence Analysis – Full Relaxation Strategy

Note that optimal h_k is computed in each step of full relaxation strategy, we have

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L}\|f'(x)\|^2.$$

since the maximal decrease cannot be less than that with $h_k = \frac{1}{L}$.

Convergence Analysis – Goldstein-Armijo Rule Strategy

It follows from the Goldstein-Armijo rule that ($\beta = 1 - \alpha$)

$\alpha\langle f'(x_k), x_k - x_{k+1}\rangle \leq f(x_k) - f(x_{k+1}) \leq \beta\langle f'(x_k), x_k - x_{k+1}\rangle = \beta h_k\|f'(x_k)\|^2.$

Also we have

$$f(x_k) - f(x_{k+1}) \geq h_k(1 - \frac{h_k}{2}L)\|f'(x)\|^2.$$

Then,

$$\begin{aligned} h_k(1 - \frac{h_k}{2}L)\|f'(x)\|^2 &\leq \beta h_k\|f'(x_k)\|^2 \\ \Leftrightarrow 1 - \frac{h_k}{2}L &\leq \beta \\ \Leftrightarrow h_k &\geq \frac{2}{L}(1-\beta). \end{aligned}$$

$$\begin{aligned} \text{Thus, } f(x_k) - f(x_{k+1}) &\geq \alpha\langle f'(x_k), x_k - x_{k+1}\rangle \\ &= \alpha h_k\|f'(x_k)\|^2 \geq \alpha\frac{2}{L}(1-\beta)\|f'(x_k)\|^2 \end{aligned}$$

Convergence Analysis

- For all strategies, we have proved that

$$f(x_k) - f(x_{k+1}) \geq \frac{\omega}{L}\|f'(x_k)\|^2, \tag{3}$$

where $x_{k+1} = x_k - h_k f'(x_k)$, L is the Lipschitz constant, and ω is some positive constant.

- We sum up the inequalities for $k = 0, \cdots, N$, we obtain

$$\frac{\omega}{L}\sum_{k=0}^{N}\|f'(x)'\|^2 \leq f(x_0) - f(x_N) \leq f(x_0) - f^*.$$

- Thus we have

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

Denote $g_N^* = \min_{0\leq k\leq N} g_k$, where $g_k = \|f'(x_k)\|$, then we have

$$g_N^* \leq \frac{1}{\sqrt{N+1}}\left[\frac{1}{\omega}L(f(x_0) - f^*)\right]^{\frac{1}{2}}. \tag{4}$$

Convergence Analysis

- The inequality (4) describes the rate of convergence of the sequence $\{g_N^*\}$.
- Usually, the rate of convergence can be used to derive upper complexity estimates for the problem classes.
- Example

  Problem class:

  1. Unconstrained minimization.
  2. $f \in C_L^{1,1}(\boldsymbol{R}^n)$.
  3. $f(x)$ in bounded below.

  Oracle: First order black box.

  $\epsilon-$**solution:** $f(\bar{x}) \leq f(x_0), \|f'(\bar{x})\| \leq \epsilon$.

Convergence Analysis

- We can estimate an upper bound for the number of steps, or the number of calls of the oracle.

$$\begin{aligned} & g_N^* \le \frac{1}{\sqrt{N+1}} \left[\frac{1}{\omega} L(f(x_0) - f^*) \right] \le \epsilon \\ \Rightarrow \quad & N + 1 \ge \frac{L}{\omega \epsilon^2} (f(x_0) - f^*). \end{aligned}$$

- Thus, we can use the value $\frac{L}{\omega\epsilon^2}(f(x_0) - f^*)$ as an upper complexity estimate for our problem class.

Local Convergence Analysis of Gradient Method

We consider the local convergence of gradient method under the following assumptions:

1. $f \in C_M^{2,2}(\boldsymbol{R}^n)$.
2. There exists a local minimum of function f at which the Hessian is *positive definite*.
3. We know some bounds $0 < \ell \leq L < \infty$ for the Hessian at x^*:

$$\ell I_n \leq f''(x^*) \leq L I_n.$$

4. Our starting point x_0 is close enough to x^*.

Local Convergence Analysis of Gradient Method

It follows from the gradient method that

$$x_{k+1} = x_k - h_k f'(x_k).$$

Note that $f'(x^*) = 0$, then we have

$$f'(x_k) = f'(x_k) - f'(x^*) = \int_0^1 f''(x^* + \tau(x_k - x^*))(x_k - x^*) d\tau = G_k(x_k - x^*),$$

where $G_k = \int_0^1 f''(x^* + \tau(x_k - x^*)) d\tau$. Therefore,

$$\begin{aligned} x_{k+1} - x^* &= (x_k - h_k f'(x_k)) - x^* \\ &= x_k - x^* - h_k G_k(x_k - x^*) \\ &= (x_k - x^*)(I - h_k G_k). \end{aligned}$$

Contraction Mapping

Let a sequence $\{a_k\}$ be defined as follows:

$$a_0 \in \boldsymbol{R}^n, a_{k+1} = A_k a_k,$$

where A_k are $(n \times n)$ matrices such that $\|A_k\| \leq 1 - q$ with $q \in (0, 1)$. Then we can estimate the rate of convergence of the sequence $\{a_k\}$ to zero:

$$\|a_{k+1}\| \leq (1 - q)\|a_k\| \leq (1 - q)^{k+1}\|a_0\| \to 0.$$

In Gradient method, we can consider $I - h_k G_k$ as A_k to analyze the local convergence rate.

Local Convergence Analysis of Gradient Method

Denote $r_k = \|x_k - x^*\|$. It follows from Corollary 1.2.1 that

$$
\begin{aligned}
& f''(x^*) - \tau M r_k I_n \leq f''(x^* + \tau(x_k - x^*)) \leq f''(x^*) + \tau M r_k I_n \\
\Rightarrow \quad & \int_0^1 \left(f''(x^*) - \tau M r_k I_n\right) d\tau \leq \int_0^1 f''(x^* + \tau(x_k - x^*)) d\tau \\
& \leq \int_0^1 \left(f''(x^*) + \tau M r_k I_n\right) d\tau \\
\Rightarrow \quad & (\ell - \frac{r_k}{2} M) I_n \leq G_k \leq (L + \frac{r_k}{2} M) I_n \\
\Rightarrow \quad & (1 - h_k(L + \frac{r_k}{2} M)) I_n \leq I_n - h_k G_k \leq (1 - h_k(\ell - \frac{r_k}{2} M)) I_n \\
\Rightarrow \quad & \|I_n - h_k G_k\| \leq \max\{a_k(h_k), b_k(h_k)\},
\end{aligned}
$$

where $a_k(h) = 1 - h(\ell - \frac{r_k}{2} M)$, and $b_k(h) = h(L + \frac{r_k}{2} M) - 1$.

Local Convergence Analysis of Gradient Method

- Note that $a_k(0) = 1$ and $b_k(0) = -1$. We assume $r_k < \bar{r} \equiv \frac{2l}{M}$, then $a_k(h)$ is a strictly decreasing linear function and we can ensure $\|I_n - h_k G_k\| < 1$ for small enough h_k. In this case we have $r_{k+1} < r_k$.
- Next we consider the optimal h_k^* which minimizes the right side of the inequality. Since $a_k(h)$ is a strictly decreasing linear function and $b_k(h)$ is a strictly increasing function, the optimal h_k^* can be found from the equation:

$$a_k(h) = b_k(h) \Leftrightarrow 1 - h(\ell - \frac{r_k}{2}M) = h(L + \frac{r_k}{2}M) - 1 \Leftrightarrow h_k^* = \frac{2}{L+\ell}.$$

Therefore, $a_k(h_k^*) = b_k(h_k^*) = 1 - \frac{2}{L+\ell}(\ell - r_k M/2)$ and

$$\begin{aligned} & x_{k+1} - x^* = (x_k - x^*)(I - h_k G_k) \\ \Rightarrow \quad & r_{k+1} \le \|I_n - h_k G_k\| r_k \\ \Rightarrow \quad & r_{k+1} \le \left(1 - \frac{2}{L+\ell}(\ell - r_k M/2)\right) r_k = \frac{L-\ell}{L+\ell} r_k + \frac{M r_k^2}{L+\ell}. \end{aligned}$$

Local Convergence Analysis of Gradient Method

- Denote $q = \frac{2\ell}{L+\ell}$ and $a_k = \frac{M}{L+\ell} r_k (< q)$, Then we can prove that

$$a_k \leq \frac{q r_0}{\bar{r} - r_0} \left(\frac{1}{1+q} \right)^k$$

Theorem

Let function $f(x)$ satisfy our assumptions and let the starting point x_0 be close enough to a local minimum:

$$r_0 = \|x_0 - x^*\| < \bar{r} = \frac{2\ell}{M}.$$

Then the gradient method with the optimal step size converges with the following rate:

$$\|x_k - x^*\| \leq \frac{\bar{r} r_0}{\bar{r} - r_0} \left(1 - \frac{\ell}{L+\ell} \right)^k$$

This rate of convergence is called linear.

Convergence Rate

- Sublinear rate
  This rate is described in terms of a power function of the iteration counter k. For example,

$$r_k \leq \frac{c}{\sqrt{k}} \Leftrightarrow k \leq \frac{c^2}{\epsilon^2}$$

- Linear rate
  This rate is given in terms of an exponential function of the iteration number. For example,

$$r_{k+1} \leq (1-q)r_k, \text{or } r_k \leq c(1-q)^k, \text{where } 0 < q < 1.$$

- Quadratic rate
  This rate has a form of the double exponential function of the iteration number. For example,

$$r_{k+1} \leq c r_k^2 \Leftrightarrow r_k \leq c^{2^k - 1} r_0^{2^k}.$$

Motivations of Newton Method – Zero Finding

- The Newton method was initially proposed for finding a root of a function of one variable $\phi(t), t \in \boldsymbol{R} : \phi(t^*) = 0$.
- We assume we have some t close to t^*, and use the following form to approximate $\phi(t + \Delta t)$:

$$\phi(t + \Delta t) = \phi(t) + \phi'(t)\Delta t = 0 \Leftrightarrow \Delta t = -\frac{\phi(t)}{\phi'(t)}.$$

- Convert it into the algorithmic form, we have the following process

$$t_{k+1} = t_k - -\frac{\phi(t_k)}{\phi'(t_k)}.$$

Motivations of Newton Method – Zero Finding

- The scheme can be extended to the problem of finding solution to the system of nonlinear equations:

  $$F(x) = 0,$$

  where $x \in \boldsymbol{R}^n$ and $F(x) : \boldsymbol{R}^n \rightarrow \boldsymbol{R}^n$
- Similarly, we can obtain

  $$x_{k+1} = x_k - [F'(x_k)]^{-1} F(x_k).$$

Motivations of Newton Method – Zero Finding

- For the unconstrained minimization problem, we can minimize it by solving the following nonlinear system

$$f'(x) = 0.$$

- Note that this replacement is not completely equivalent, but it works in nondegenerate situations.
- Using approximation, the Newton system looks like

$$f'(x) + f''(x)\Delta x = 0.$$

- The algorithmic form is

$$x_{k+1} = x_k - [f''(x_k)]^{-1} f'(x_k).$$

Motivations of Newton Method – Quadratic Approximation

- Suppose the approximation of $f(x)$ near x_k is $\phi(x)$:

$$\phi(x) = f(x_k) + \langle f'(x_k), x - x_k \rangle + \frac{1}{2} \langle f''(x_k)(x - x_k), x - x_k \rangle.$$

- Assume $f''(x_k) > 0$, we can choose x_{k+1} as a point of minimum of the quadratic function $\phi(x)$, thus we have

$$\phi'(x_{k+1}) = f'(x_k) + f''(x_k)(x_{k+1} - x_k) = 0.$$

- Similarly, we have

$$x_{k+1} = x_k - [f''(x_k)]^{-1} f'(x_k).$$

Local Convergence Analysis of Newton Method

- Consider the problem

$$\min_{x \in \boldsymbol{R}^n} f(x)$$

under the following assumptions:

1. $f \in C_M^{2,2}(\boldsymbol{R}^n)$.
2. There exists a local minimum of function f with positive definite Hessian:

$$f''(x^*) \geq \ell I_n, \ell > 0.$$

3. Our starting point x_0 is close enough to x^*.

Local Convergence Analysis of Newton Method

$$
\begin{aligned}
x_{k+1} - x^* &= x_k - x^* - [f''(x_k)]^{-1} f'(x_k) \\
&= x_k - x^* - [f''(x_k)]^{-1} \left(f'(x_k) - f'(x^*) \right) \\
&= x_k - x^* - [f''(x_k)]^{-1} \int_0^1 f''(x^* + \tau(x_k - x^*))(x_k - x^*) d\tau \\
&= [f''(x_k)]^{-1} G_k (x_k - x^*),
\end{aligned}
$$

where $G_k = \int_0^1 [f''(x_k) - f''(x^* + \tau(x_k - x^*))] d\tau$. Denote $r_k = \|x_k - x^*\|$. Thus we can use the technique of contraction mapping to analyze the convergence.

Local Convergence Analysis of Newton Method

$$
\begin{aligned}
\|G_k\| &= \|\int_0^1 [f''(x_k) - f''(x^* + \tau(x_k - x^*))]d\tau\| \\
&\le \int_0^1 \|[f''(x_k) - f''(x^* + \tau(x_k - x^*))]\|d\tau \\
&\le \int_0^1 M(1-\tau)r_k d\tau = \frac{r_k M}{2}.
\end{aligned}
$$

Using Corollary 1.2.1, we have

$$f''(x_k) \ge f''(x^*) - Mr_k I_n \ge (\ell - Mr_k)I_n.$$

We assume $r_k < \frac{\ell}{M}$, thus $f''(x_k)$ is positive definite and $\|[f''(x_k)]^{-1}\| \le (\ell - Mr_k)^{-1}$. Thus we have

$$r_{k+1} \le \frac{Mr_k^2}{2(\ell - Mr_k)}$$

Local Convergence Analysis of Newton Method

Theorem

Let function $f(x)$ satisfy our assumptions. Suppose that the initial starting point x_0 is close enough to x^*:

$$\|x_0 - x^*\| < \bar{r} = \frac{2\ell}{3M}.$$

Then $\|x_k - x^*\| < \bar{r}$ for all k and the Newton method converges quadratically:

$$\|x_{k+1} - x^*\| = \frac{M\|x_k - x^*\|^2}{2(\ell - M\|x_k - x^*\|)}.$$

Comments:
The constraint $r_k < \frac{2\ell}{3M}$ guaranties that $r_{k+1} < r_k$.

What is the Difference between Gradient Method and Newton Method?

- The problem is the unconstrained minimization problem $\min_{x \in \boldsymbol{R}^n} f(x)$ where $f \in C_M^{2,2}$.
- The procedure in gradient method is

$$x_{k+1} = x_k - h_k f'(x_k), h_k > 0.$$

- The procedure in the Newton method is

$$x_{k+1} = x_k - [f''(x_k)]^{-1} f'(x_k).$$

- The gradient method has a linear convergence rate while the Newton method converges quadratically.

Approximation for Gradient Method

Consider the approximation of $f(x)$ near $\bar{x}$:

$$\phi_1(x) = f(\bar{x}) + \langle f'(\bar{x}), x - \bar{x} \rangle + \frac{1}{2h}\|x - \bar{x}\|^2,$$

where the paramter $h > 0$. Thus we have

$$\phi'(x_1^*) = f'(\bar{x}) + \frac{1}{h}(x_1^* - \bar{x}) = 0 \Leftrightarrow x_1^* = \bar{x} - hf'(\bar{x}).$$

Approximation for the Newton Method

Consider the quadratic approximation of $f(x)$ near $\bar{x}$:

$$\phi_2(x) = f(\bar{x}) + \langle f'(\bar{x}), x - \bar{x} \rangle + \frac{1}{2} \langle f''(\bar{x})(x - \bar{x}), x - \bar{x} \rangle.$$

Thus we have

$$\phi_2'(x_2^*) = f'(\bar{x}) + f''(\bar{x})(x_2^* - \bar{x}) = 0 \Leftrightarrow x_2^* = \bar{x} - [f''(\bar{x})]^{-1} f'(\bar{x}).$$

Approximation for the Quasi-Newton Method

- Can we use some other approximations of the function $f(x)$, which is better than $\phi_1(x)$ and less expensive than $\phi_2(x)$?
- Let $G \in \boldsymbol{R}^{n\times n}$ be a positive definite matrix, and consider the following approximation

$$\phi_G(x) = f(\bar{x}) + \langle f'(\bar{x}), x - \bar{x}\rangle + \frac{1}{2}\langle G(x - \bar{x}), x - \bar{x}\rangle.$$

- Computing its minimum from this equation, we obtain

$$\phi'_G(x^*_G) = f'(\bar{x}) + G(x^*_G - \bar{x}) = 0 \Leftrightarrow x^*_G = \bar{x} - G^{-1}f'(\bar{x}).$$

- The first-order methods which forms a sequence $\{G_k\} : G_k \to f''(x^*)$ (or $\{H_k\} : H_k \equiv G_k^{-1} \to [f''(x^*)]^{-1}$) are called *variable metric methods* (or quasi-Newton method).

Motivation for Quasi-Newton Method

- Note that all inner product are defined with respect to the standard Euclidean inner product on $\boldsymbol{R}^n$:

$$\langle x, y \rangle = \sum_{i=1}^{n} x^{(i)} y^{(i)}, x, y \in \boldsymbol{R}^n, \text{ and } \|x\| = \langle x, x \rangle^{1/2}$$

- Given a symmetric positive definite matrix $A \in \boldsymbol{R}^{n \times n}$, we can define a new inner product based on A:

$$\langle x, y \rangle_A = x^T A y = \langle Ax, y \rangle, \text{ and } \|x\|_A = \langle Ax, x \rangle^{1/2}$$

Motivation for Quasi-Newton Method

- Using the new inner product defined on A, we have

$$\begin{aligned} f(x+h) &= f(x) + \langle f'(x), h \rangle + o(\|h\|) \\ &= f(x) + \langle A^{-1} f'(x), h \rangle_A + o(\|h\|_A) \end{aligned}$$

- We can define the gradient respect to the new metric defined by A:

$$f'_A(x) = A^{-1} f'(x).$$

- If we define $A = f''(x)$, then $f'_A(x) = [f''(x)]^{-1} f'(x)$, thus the direction in the Newton method can be interpreted as the gradient respect to this new metric.

General Scheme of the Variable Metric Methods

1. Choose $x_0 \in \mathbb{R}^n$. Set $H_0 = I_n$. Compute $f(x_0)$ and $f'(x_0)$.
2. kth iteration ($k \geq 0$).
    1. Set $p_k = -H_k f'(x_k)$.
    2. Find $x_{k+1} = x_k + \alpha_k p_k$ using line search.
    3. Compute $f(x_{k+1})$ and $f'(x_{k+1})$.
    4. Update the matrix $H_k : H_k \longrightarrow H_{k+1}$.

Comments:
The variable metric schemes differ one from another only in the implementation of Step 2.4, which updates the matrix H_k. For that, they use the new information, accumulated at Step 2.3, namely the gradient $f'(x_{k+1})$.

Some Examples of Quasi-Newton Method

Denote $\Delta H_k = H_{k+1} - H_k$, $y_k = f'(x_{k+1}) - f'(x_k)$, and $s_k = x_{k+1} - x_k$.

1. Rank-one correction scheme (SR1)

$$\Delta H_k = \frac{(s_k - H_k y_k)(s_k - H_k y_k)^T}{\langle s_k - H_k y_k, y_k \rangle}.$$

2. Davidon-Fletcher-Powell scheme (DFP)

$$\Delta H_k = \frac{s_k s_k^T}{\langle y_k, s_k I \rangle} - \frac{H_k y_k y_k^T H_k}{\langle H_k y_k, y_k \rangle}.$$

3. Broyden-Fletcher-Goldfarb-Shanno scheme (BFGS)

$$\Delta H_k = \frac{H_k y_k s_k^T + s_k y_k^T H_k}{\langle H_k y_k, s_k \rangle} - \beta_k \frac{H_k y_k y_k^T H_k}{\langle H_k y_k, y_k \rangle},$$

where $\beta_k = 1 + \frac{\langle y_k, s_k \rangle}{\langle H_k y_k, y_k \rangle}$.

Secant Equation

- Suppose the quadratic approximation of point $x = x_{k+1} + p$ near x_{k+1} is

$$\phi_{k+1}(x_{k+1} + p) = f(x_{k+1}) + p^T f'(x_{k+1}) + \frac{1}{2} p^T G_{k+1} p.$$

- What requirement should we impose on G_{k+1}?
  - The gradient of ϕ_{k+1} should match the gradient of the objective function f at the latest two iterations x_k and x_{k+1}. That is,

$$\begin{aligned} \phi'_{k+1}(0) &= f'(x_{k+1}) \\ \phi'_{k+1}(-\alpha_k p_k) &= f'(x_{k+1}) - \alpha_k G_{k+1} p_k = f'(x_k). \end{aligned}$$

- By rearranging the secant equation, we obtain

$$\begin{aligned} & \alpha_k G_{k+1} p_k = f'(x_{k+1}) - f'(x_k) \\ \Leftrightarrow \quad & G_{k+1} s_k = y_k \end{aligned} \tag{5}$$

- Eq. (5) is called the **Secant Equation**.

The SR1 (Symmetric-Rank-1) Method

- The symmetric rank-1 update has the general form

$$G_{k+1} = G_k + \sigma v v^T, \text{ where } \sigma \in \{-1, 1\}. \tag{6}$$

- Plugging Eq.(6) into the secant equation, we have

$$y_k = G_k s_k + [\sigma v^T s_k] v \Rightarrow v = \delta(y_k - G_k s_k), \text{ where } \delta \in \boldsymbol{R}.$$

- Plugging $v = \delta(y_k - G_k s_k)$ into $y_k = G_k s_k + [\sigma v^T s_k] v$, we have

$$\begin{aligned} & y_k - G_k s_k = \sigma \delta^2 [s_k^T (y_k - G_k s_k)](y_k - G_k s_k) \\ \Leftrightarrow \quad & \sigma = \operatorname{sign}(s_k^T (y_k - G_k s_k)), \delta = \pm \left| s_k^T (y_k - G_k s_k) \right|^{-\frac{1}{2}} \end{aligned}$$

- In summary, the only symmetric rank-1 updating formula that satisfies the secant equation is

$$G_{k+1} = G_k + \frac{(y_k - G_k s_k)(y_k - G_k s_k)^T}{(y_k - G_k s_k)^T s_k}$$

The SR1 (Symmetric-Rank-1) Method

More analysis on SR1 updating

1. If $(y_k - G_k s_k)^T s_k \neq 0$, there is a unique rank-one updating formula satisfying the secant equation.
2. If $y_k = G_k s_k$, the only updating formula satisfying the secant equation is $G_{k+1} = G_k$.
3. If $(y_k - G_k s_k)^T s_k = 0$ and $y_k \neq G_k s_k$, there is no symmetric rank-one updating formula satisfying the secant equation.
4. In practice, the SR1 update is defined only if $(y_k - G_k s_k)^T s_k \neq 0$.

The SR1 (Symmetric-Rank-1) Method

- By applying the Sherman-Morrison-Woodbury formula, we obtain the corresponding update formula for the inverse Hessian approximation H_{k+1}

$$H_{k+1} = H_k + \frac{(s_k - H_k y_k)(s_k - H_k y_k)^T}{(s_k - H_k y_k)^T y_k}.$$

- The Sherman-Morrison-Woodbury formula

$$\left(A + \alpha u u^T\right)^{-1} = A^{-1} - \frac{\alpha}{1 + \alpha u^T A^{-1} u} A^{-1} u u^T A^{-1}.$$

- One drawback of SR1 updating is that G_{k+1} may not be positive definite.

The BFGS Method

- Can we preserve both the symmetry and positive definiteness in each iteration?
- We can consider the rank-two updating as follows:

$$G_{k+1} = G_k + \gamma u u^T + \delta v v^T.$$

- Plugging G_{k+1} into the secant equation, we have

$$y_k - G_k s_k = \gamma (u^T s_k) u + \delta (v^T s_k) v,$$

  which implies that the suitable u and v can be found as linear combinations of y_k and $G_k s_k$.
- One of the simplest possible choice is $u = G_k s_k$ and $v = y_k$, thus we have

$$0 = G_{k+1} s_k - y_k = G_k s_k (1 + \gamma s_k^T G_k s_k) - y_k (1 - \delta y_k^T s_k).$$

- An obvious choice of γ and δ that makes the equality hold is:

$$\gamma = -\frac{1}{s_k^T G_k s_k}, \delta = \frac{1}{y_k^T s_k}$$

The BFGS Method

- The resulting updating formula in BFGS is

$$G_{k+1} = G_k - \frac{1}{s_k^T G_k s_k} G_k s_k s_k^T G_k + \frac{1}{y_k^T s_k} y_k y_k^T .$$

- Using the Sherman-Morrison-Woodbury formula, we have

$$H_{k+1} = \left(I - \frac{1}{y_k^T s_k} s_k y_k^T \right) H_k \left(I - \frac{1}{y_k^T s_k} s_k y_k^T \right) + \frac{1}{y_k^T s_k} s_k s_k^T .$$

- Most researchers today agree that the BFGS method is the “best” one in practice, but reasons for its superiority have not yet been fully explained.

The DFP Method

- DFP method is the first quasi-Newton algorithm proposed in 1950s by Davidon. However, Davidon's paper remained as a technical report for more than 30 years until it appeard in the first issue of the SIAM Journal on Optimization in 1991.
- The secant equation can be rewritten as

$$s_k = H_{k+1} y_k, \text{ where } H_{k+1} = G_{k+1}^{-1}.$$

- We focus on the updating of H_{k+1} and use the same technique in BFGS for rank-two updating, we have

$$H_{k+1} = H_k - \frac{1}{y_k^T H_k y_k} H y_k y_k^T H + \frac{1}{s_k^T y_k} s_k s_k^T.$$

The DFP Method

- Using the Sherman-Morrison-Woodbury formula, we have

$$\begin{aligned} G_{k+1} &= G_k - \frac{1}{s_k^T G_k s_k} G_k s_k s_k^T G_k + \frac{1}{y_k^T s_k} y_k y_k^T \\ &+ \left(s_k^T G_k s_k\right) w w^T, \text{ where } w = \frac{1}{y_k^T s_k} y_k - \frac{1}{s_k^T G_k s_k} G_k s_k. \end{aligned}$$

- In other words, we have

$$G_{k+1}^{DFP} = G_{k+1}^{BFGS} + \left(s_k^T G_k s_k\right) w w^T, \text{ if } G_k^{DFP} = G_k^{BFGS}.$$

The Broyden Class of Updates

- It can be verified that

$$w^T s_k = \left(\frac{1}{y_k^T s_k} y_k - \frac{1}{s_k^T G_k s_k} G_k s_k \right)^T s_k = 0.$$

- Hence, the symmetric rank-one matrix ww^T can be added to any matrix G_k without violating the secant equation $G_{k+1} s_k = y_k$.
- The formula used in the Broyden class of updates:

$$G_{k+1} = G_k - \frac{1}{s_k^T G_k s_k} G_k s_k s_k^T G_k + \frac{1}{y_k^T s_k} y_k y_k^T + \phi(s_k^T G_k s_k) ww^T,$$

where ϕ may depend on y_k and G_k.

1. For BFGS, $\phi = 0$.
2. For DFP, $\phi = 1$.
3. For SR1, $\phi = \frac{y_k^T s_k}{y_k^T s_k - s_k^T G_k s_k}$.

Relationships between the Updates

- In 1960s and 1970s, a lot of "new" quasi-Newton updates were proposed, and the researchers claimed that their own updates were the "best".

Theorem (L. Dixon, 1972)

Let $f(x) \in C^2$*, and assume that* x_0 *and* G_0 *are given. Let* $\{x_k\}, \{G_k\}, \{p_k\}$ *and* $\{\alpha_k\}$ *denote the sequence generated by BFGS line search method, with* $\{x_k^\phi\}, \{G_k^\phi\}, \{p_k^\phi\}$ *and* $\{\alpha_k^\phi\}$ *the corresponding values for the line search method in which* G_{k+1} *is defined by any member of the Broyden class. If each of the sequences* $\{G_k\}$ *and* $\{G_k^\phi\}$ *is well-defined, for all* k, α_k *and* α_k^ϕ *are the minimizer of* $f(x_k + \alpha_k p_k)$ *that are nearest to the point* $\alpha = 0$*, then*

$$x_k^\phi = x_k, \alpha_k^\phi p_k^\phi = \alpha_k p_k, \ \ and \ G_k = G_k^\phi + \left(\frac{\phi_{k-1}}{g_{k-1}^T p_{k-1}^\phi} \right) g_k g_k^T$$

where g_k *denotes* $f'(x_k)$.

Relationships between the Updates

- The theorem proved by L. Dixon in 1972 shows that, when an **exact** line search is used, the iterates generated by different quasi-Newton update formulae are identical. Thus, the observed differences in performance were attributable to variations in the step length selection strategy rather than to properties of the updates.

Local Convergence of Variable Metric Method

In the neighborhood of a strict minimum these methods have a superlinear rate of convergence: for any $x_0 \in \boldsymbol{R}^n$, there exists a number N such that for all $k \geq N$ we have

$$\|x_{k+1} - x^*\| \leq \text{const} \cdot \|x_k - x^*\| \cdot \|x_{k-n} - x^*\|.$$

(the proofs are very long and technical.)

Conjugate Gradient Method for Quadratic Function

- Consider the quadratic function $f(x) = \alpha + \langle a, x\rangle + \frac{1}{2}\langle Ax, x\rangle$, where A is symmetric and positive definite.
- We have the following observations:

$$x^* = -A^{-1}a, f^* = \alpha - \frac{1}{2}\langle Ax^*, x^*\rangle, \text{ and } f'(x) = A(x - x^*).$$

- Suppose we are given by a starting point x_0. Consider the linear subspaces

$$\mathcal{L}_k = \text{Lin}\{A(x_0 - x^*), \cdots, A^k(x_0 - x^*)\}.$$

- The sequence of the conjugate gradient method is defined as follows:

$$x_k = \arg\ \min\{f(x) | x \in x_0 + \mathcal{L}_k\}, k = 1, 2, \cdots .$$

Conjugate Gradient Method for Quadratic Function

Lemma

For any $k \geq 1$, we have $\mathcal{L}_k = Lin\{f'(x_0), \cdots, f'(x_{k-1})\}$.

For $k = 1$ we have $f'(x_0) = A(x_0 - x^*)$. Suppose the statement is true for some $k \geq 1$, note that $x_k = \arg\ \min\{f(x) | x \in x_0 + \mathcal{L}_k\}$, thus we can represent x_k as

$$x_k = x_0 + \sum_{i=1}^{k} \lambda_i A^i (x_0 - x^*).$$

Also we have

$$\begin{aligned}
f'(x_k) &= A(x_k - x^*) = A(x_0 - x^*) + \sum_{i=1}^{k} \lambda_i A^{i+1}(x_0 - x^*) \\
&= y + \lambda_k A^{k+1}(x_0 - x^*), \text{ where } y \in \mathcal{L}_k. \quad \text{Thus,} \\
\mathcal{L}_{k+1} &= \text{Lin}\{\mathcal{L}_k, A^{k+1}(x_0 - x^*)\} = \text{Lin}\{\mathcal{L}_k, f'(x_k)\} \\
&= \text{Lin}\{f'(x_0), \cdots, f'(x_k)\}.
\end{aligned}$$

Conjugate Gradient Method for Quadratic Function

Lemma

For any $k, i \geq 0, k \neq i$, we have $\langle f'(x_k), f'(x_i) \rangle = 0$.

Proof.

We use the statement that $x_k = \arg\ \min\{f(x) | x \in x_0 + \mathcal{L}_k\}$ to prove it. Let $k > i$, consider the new function

$$\phi(\bar{\lambda}) = \phi(\lambda_1, \cdots, \lambda_k) = f\left(x_0 + \sum_{j=1}^{k} \lambda_j f'(x_{j-1})\right).$$

It follows from $x_k = \arg\ \min\{f(x) | x \in x_0 + \mathcal{L}_k\}$ that there exists $\bar{\lambda}^* = (\lambda_1^*, \cdots, \lambda_k^*)$ that $x_k = x_0 + \sum_{j=1}^{k} \lambda_j^* f'(x_{j-1})$. Since x_k is the minimum of $f(x)$ on $x_0 + \mathcal{L}_k$, thus we have:

$$0 = \frac{\partial \phi(\bar{\lambda}^*)}{\partial \lambda_i} = \langle f'(x_k), f'(x_{i-1}) \rangle.$$

Conjugate Gradient Method for Quadratic Function

Lemma

For any $p \in \mathcal{L}_k$ *we have* $\langle f'(x_k), p\rangle = 0$.

Proof.

Since $p \in \mathcal{L}_k = \mathrm{Lin}\{f'(x_0), \cdots, f'(x_{k-1})\}$, we can represent p as

$$p = \sum_{i=0}^{k-1} \lambda_i f'(x_i).$$

Thus we have

$$\langle f'(x_k), p\rangle = \langle f'(x_k), \sum_{i=0}^{k-1} \lambda_i f'(x_i)\rangle = \sum_{i=0}^{k-1} \lambda_i \langle f'(x_k), f'(x_i)\rangle = 0.$$

□

Conjugate Gradient Method for Quadratic Function

Denote $\delta_i = x_{i+1} - x_i$. It is clear that $\mathcal{L}_k = \mathrm{Lin}\{\delta_0, \cdots, \delta_{k-1}\}$. Also we have $A\delta_i = f'(x_{i+1}) - f'(x_i)$.

Lemma

For any $k \neq i$ we have $\langle A\delta_k, \delta_i \rangle = 0$.

Proof.

Assume $k > i$. then

$$\langle A\delta_k, \delta_i \rangle = \langle f'(x_{k+1}) - f'(x_k), \delta_i \rangle = 0.$$

Note that $\delta_i \in \mathcal{L}_{i+1} \subseteq \mathcal{L}_k$, thus $\langle f'(x_j), \delta_i \rangle = 0$ for $j > i$. ☐

It follows from this lemma that all the directions δ_k are conjugate with respect to A.

Conjugate Gradient Method for Quadratic Function

Since $\delta_k \in \mathcal{L}_{k+1} = \mathrm{Lin}\{\delta_0, \cdots, \delta_k\} = \mathrm{Lin}\{f'(x_0), \cdots, f'(x_k)\}$, then we can represent δ_k as

$$\delta_k = -h_k f'(x_k) + \sum_{j=1}^{k-1} \lambda_i \delta_j.$$

Multiplying both sides with A and δ_i $(1 \leq i \leq k-1)$, we have

$$\begin{aligned} 0 &= \langle A\delta_k, \delta_i \rangle = -h_k \langle Af'(x_k), \delta_i \rangle + \sum_{j=0}^{k-1} \lambda_i \langle A\delta_j, \delta_i \rangle \\ &= -h_k \langle f'(x_k), f'(x_{i+1}) - f'(x_i) \rangle + \lambda_i \langle A\delta_i, \delta_i \rangle. \end{aligned}$$

Thus $\lambda_i = 0$ for all $i < k-1$. For $i = k-1$ we have

$$\lambda_{k-1} = \frac{h_k \|f'(x_k)\|^2}{\langle A\delta_{k-1}, \delta_{k-1} \rangle} = \frac{h_k \|f'(x_k)\|^2}{\langle f'(x_k) - f'(x_{k-1}), \delta_{k-1} \rangle}.$$

Conjugate Gradient Method for Quadratic Function

Thus, we have

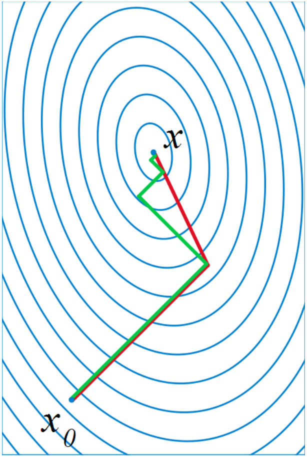
$$x_{k+1} - x_k = \delta_k = -h_k f'(x_k) + \lambda_{k-1}\delta_{k-1}.$$

We can rewrite it as

$$\begin{aligned} x_{k+1} &= x_k - h_k p_k, \\ \text{where } p_k &= f'(x_k) - \frac{\|f'(x_k)\|^2 \delta_{k-1}}{\langle f'(x_k) - f'(x_{k-1}), \delta_{k-1}\rangle} \\ &= f'(x_k) - \frac{\|f'(x_k)\|^2 p_{k-1}}{\langle f'(x_k) - f'(x_{k-1}), p_{k-1}\rangle}, \end{aligned}$$

since $\delta_{k-1} = -h_{k-1}p_{k-1}$.

An Example of Conjugate Gradient Method



Conjugate Gradient Method for Quadratic Function

- We can write down the conjugate gradient scheme in terms of the gradients of the objective function $f(x)$.
- All discussions are based on the quadratic function.
- Can we extend the scheme to the general nonlinear functions?

General Scheme of Conjugate Gradient Method

1. Choose $x_0 \in \mathbb{R}^n$. Compute $f(x_0)$ and $f'(x_0)$. Set $p_0 = f'(x_0)$.
2. kth iteration ($k \geq 0$).
    1. Find $x_{k+1} = x_k - h_k p_k$ (using an "exact" line search procedure).
    2. Compute $f(x_{k+1})$ and $f'(x_{k+1})$
    3. Compute the coefficienet β.
    4. Set $p_{k+1} = f'(x_{k+1}) - \beta_k p_k$.

Some popular methods to update β:

- $\beta_k = \frac{\|f'(x_{k+1}\|^2}{\langle f'(x_{k+1}) - f'(x_k), p_k \rangle}$.
- Fletcher-Rieves: $\beta_k = -\frac{\|f'(x_{k+1})\|^2}{\|f'(x_k)\|^2}$.
- Polak-Ribbiere: $\beta_k = -\frac{\langle f'(x_{k+1}), f'(x_{k+1}) - f'(x_k) \rangle}{\|f'(x_k)\|^2}$.

Conjugate Gradient Method

- In the first step, we usually decrease the function along the direction of anti-gradient, and in later steps we try to find directions conjugate with respect to A, thus this method is called *Conjugate Gradient Method*.
- For quadratic function, the conjugate gradient method terminates in n iterations.
- In nonlinear case this is not true. In practical schemes, we apply the **restart** strategy, which at some point sets $\beta_k = 0$ and follow the direction of anti-gradient(usually after every n iterations). Thus the global convergence of the scheme is ensured.

Constrained Minimization

- Problem Formulation

$$\begin{aligned} f_0(x) &\rightarrow \min, \\ f_i(x) &\leqslant 0, i = 1, \cdots, m. \end{aligned}$$

  where $f_i(x)$ are smooth functions.

- The ideology of *Sequential Unconstrained Minimization*
  1. We have several efficient methods for unconstrained minimization.
  2. An unconstrained minimization problem is simpler than a constrained one.
  3. Therefore, let us try to approximate the solution of the constrained problem by a sequence of solutions of some auxiliary unconstrained problems.
- Two main groups of Sequential Unconstrained Minimization
  1. Penalty function methods
  2. Barrier function methods

Penalty Function Method

Definition (Penalty Function)

A continuous function $\Phi(x)$ is called a penalty function for a closed set Q if

- $\Phi(x) = 0$ for any $x \in Q$,
- $\Phi(x) > 0$ for any $x \notin Q$

The penalty function sometimes is called just *penalty*.

Property

If $\Phi_1(x)$ is a penalty function for Q_1 and $\Phi_2(x)$ is a penalty function for Q_2 then $\Phi_1(x) + \Phi_2(x)$ is a penalty function for the intersection $Q_1 \cap Q_2$.

Penalty Function Method – Some Examples of Penalty Function

- Denote $(a)_+ = \max\{a, 0\}$. Let $Q = \{x \in R^n | f_i(x) \leq 0, i = 1, ..., m\}$.
- Quadratic penalty:

$$F(x) = \sum_{i=1}^{m} (f_i(x))_+^2.$$

- Nonsmooth penalty:

$$F(x) = \sum_{i=1}^{m} (f_i(x))_+.$$

Penalty Function Method

- General scheme of penalty function method
    1. Choose $x_0 \in R^n$. Choose a sequence of penalty coefficients: $0 < t_k < t_{k+1}, t_k \to \infty$.
    2. kth iteration ($k \leq 0$). Find a point $x_{k+1} = \arg\ \min_{x \in R^n}\{f_0(x) + t_k \Phi(x)\}$ using x_k as a starting point.
- It is easy to prove the convergence of this scheme assuming that x_{k+1} is a global minimum of the auxiliary function $\Psi_k(x) = f_0(x) + t_k \Phi(x)$.

Penalty Function Method – Proof of Convergence

Assumption

There exists $\bar{t} > 0$ *such that the set* $S = \{x \in R^n | f_0(x) + \bar{t}\Phi(x) \leq f^*\}$ *is bounded, where* f^* *is the optimal value of the constrained problem.*

Theorem

If the constrained optimization problem satisfies this assumption, then $\lim_{k\to\infty} f(x_k) = f^*, \lim_{k\to\infty} \Phi(x_k) = 0$.

Barrier Function Method

Definition (Barrier Function)

A continuous function $F(x)$ is called a barrier function for a closed set Q with nonempty interior if $F(x) \to \infty$ when x approaches the boundary of the set Q. The barrier function sometimes is called just *barrier*.

Property

If $F_1(x)$ *is a barrier function for* Q_1 *and* $F_2(x)$ *is a barrier function for* Q_2 *then* $F_1(x) + F_2(x)$ *is a barrier function for the intersection* $Q_1 \cap Q_2$.

Barrier Function Method – Some Examples of Barrier Function

- Let $Q = \{x \in R^n | f_i(x) \leq 0, i = 1, ..., m\}$.
- Power-function barrier:

$$F(x) = \sum_{i=1}^{m} \frac{1}{(-f_i(x))^p}, p \geq 1.$$

- Logarithmic barrier:

$$F(x) = -\sum_{i=1}^{m} \ln(-f_i(x)).$$

- Exponential barrier:

$$F(x) = \sum_{i=1}^{m} \exp\left(\frac{1}{-f_i(x)}\right).$$

Barrier Function Method

- General scheme of penalty function method
  1. Choose $x_0 \in \mathrm{int} Q$. Choose a sequence of penalty coefficients: $0 < t_k < t_{k+1}, t_k \to \infty$.
  2. kth iteration ($k \leq 0$). Find a point $x_{k+1} = \arg\ \min_{x \in Q} \{f_0(x) + \frac{1}{t_k} F(x)\}$ using x_k as a starting point.
- Similarly, we can prove the convergence of this scheme assuming that x_{k+1} is a global minimum of the auxiliary function $\Psi_k(x) = f_0(x) + \frac{1}{t_k} F(x)$.

Barrier Function Method – Proof of Convergence

Assumption

The barrier $F(x)$ is below bounded: $F(x) \geq F^*$ for all $\in Q$.

Theorem

Let the constrained optimization problem satisfies this assumption. Then

$$\lim_{k \to \infty} \Psi_k^* = f^*.$$