CS Space

# Unconstrained Nonlinear Optimization
# Lecture 3: Line Search Methods 2

**Feodor Pisnitchenko**

CMCC/UFABC

**Momoe Sakamori**

IMECC/UNICAMP

An algorithm is said to be:

- **globally convergent** if for any initial point $x_0$, it generates a sequence of points that converges to a point $x^*$ in the solution set;
- **locally convergent** if there exists a $\rho > 0$ such that for any initial point $x_0$ that $\|x^* - x_0\| < \rho$, it generates a sequence os points that converges to $x^*$ in the solution set.

## Convergence of Line Search Methods

- $\theta_k$ is the angle between $p_k$ and the steepest descent direction $-\nabla f_k$, defined by

$$\cos\theta_k = \frac{-\nabla f_k^T p_k}{\|\nabla f_k\| \|p_k\|}$$

- $\{x | f(x) = c\}$ is the level set of $f$ for value $c$, and $\{x | f(x) \leq c\}$ is its corresponding sublevel set.
- Let $f : D \rightarrow \mathbb{R}^m$ where $D \subset \mathbb{R}^n$. The function $f$ is Lipschitz continuous on some set $N \subset D$ if there is a constant $L > 0$ ($L$ is called the Lipschitz constant) such that

$$\|f(x_1) - f(x_0)\| \leq L\|x_1 - x_0\|, \text{for all } x_0, x_1 \in N.$$

## Convergence of Line Search Methods

***Theorem (Zoutendijk's Result)***

Consider any line search iteration where $p_k$ is a descent direction and $\alpha_k$ satisfies the Wolfe conditions. Suppose that $f$ is bounded below in $\mathbb{R}^n$ and that $f$ is continuously differentiable in an open set $N$ containing the level set $L = \{x | f(x) \leq f(x_0)\}$, where $x_0$ is the starting point of the iteration. Assume also that the gradient $\nabla f$ is Lipschitz continuous on $N$. Then

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

## Theorem (Zoutendijk's Result)

Proof. The gradiente $\nabla f$ is Lipschitz continuous in $N$, then there exists a constant $L > 0$ such that
$\|\nabla f(x) - \nabla f(\tilde{x})\| \leq L\|x - \tilde{x}\|, \quad \forall x, \tilde{x} \in N.$
From Cauchy-Schwartz inequality,

$$|(\nabla f_{k+1} - \nabla f_k)^T p_k| \leq \underbrace{\|\nabla f_{k+1} - \nabla f_k\|}_{\leq L\|x_{k+1} - x_k\| = L\|\alpha_k p_k\|} \|p_k\| \leq L\alpha_k \|p_k\|^2$$

$$(\nabla f_{k+1} - \nabla f_k)^T p_k = \underbrace{\nabla f_{k+1}^T p_k}_{\geq c_2 \nabla f_k^T,\ \text{Wolfe (II)}} - \nabla f_k^T p_k \geq \underbrace{(c_2 - 1)}_{< 0} \underbrace{\nabla f_k^T p_k}_{< 0} > 0$$

By combining these two relations,

$$(c_2 - 1)\nabla f_k^T p_k \le (\nabla f_{k+1} - \nabla f_k)^T p_k \le L\alpha_k \|p_k\|^2$$

$$\alpha_k \ge \frac{c_2 - 1}{L} \frac{\nabla f_k^T p_k}{\|p_k\|^2}$$

From Wolfe condition (I),

$$\begin{aligned} f_{k+1} &\le f_k + c_1 \alpha_k \nabla f_k^T p_k \\ &\le f_k - c_1 \frac{(c_2 - 1)}{L} \frac{(\nabla f_k^T p_k)^2}{\|p_k\|^2} \\ &= f_k - c \cos^2 \theta_k \|\nabla f_k p_k\|^2 \end{aligned}$$

where $c = c_1 \frac{(c_2 - 1)}{L}$.

By summing this expression over all indices less than or equal to $k$, we obtain

$$f_{k+1} \leq f_0 - c \sum_{j=0}^{k} \cos^2 \theta_j \|\nabla f_j\|^2$$

$$\sum_{j=0}^{k} \cos^2 \theta_j \|\nabla f_j\|^2 \leq \frac{f_0 - f_{k+1}}{c}$$

Since $f$ is bounded below, we have $f_0 - f_{k+1} < b, \forall k$, for some positive constant $b$. Hence, by taking limits, we obtain

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

- Zoutendijk condition implies that $\cos^2 \theta_k \|\nabla f_k\|^2 \to 0$.
- If $p_k$ is chosen such that $\theta_k < \pi/2$, then there is a positive constant $\delta$ such that $\cos \theta_k \geq \delta > 0, \forall k$. Thus,

$$\lim_{k \to \infty} \|\nabla f_k\| = 0.$$

- It means that if the search direction are never too close to orthogonality with the gradient and the step lengths satisfy Wolfe conditions, then the gradient norms $\|\nabla f_k\|$ converge to zero.
- $x_k$ converges to a stationary point of $f$, but there is no guarantee that the stationary point is a minimizer.
- In the steepest descent method, $p_k = -\nabla f_k$, $x_k$ converges to a stationary point if the step lengths satisfy Wolfe conditions.

- Consider a Newton-like method:

$$x_{k+1} = x_k + \alpha_k p_k, p_k = -B_k^{-1} \nabla f_k.$$

- Assume that $B_k > 0$ with a uniformly bounded condition number. It means that $\exists M$ such that

$$\|B_k\| \|B_k^{-1}\| \leq M, \forall k.$$

- It can be shown that $\cos \theta_k \geq 1/M$.
- By combining this bound with the Zoutendijk condition implication, we find $\lim_{k\to\infty} \|\nabla f_k\| = 0$.
- Newton and Quasi-Newton methods are globally convergent if the matrices $B_k$ have a bounded condition number and are positive definite, and if the step lengths satisfy the Wolfe conditions.

## Convergence Rate of Steepest Descent

Suppose that the objective function is quadratic,

$$f(x) = \frac{1}{2}x^T Q x - b^T x + c,$$

where $Q$ is symmetric and positive definite, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$.
The gradient is given by $\nabla f(x) = Qx - b$.
The minimizer $x^*$ is the unique solution of the linear system $Qx = b$.
The step length is given by $\alpha_k = \arg\min\{f(x_k - \alpha \nabla f_k), \alpha > 0\}$.

$$\begin{aligned} f(x_k - \alpha \nabla f_k) &= \frac{1}{2}(x_k - \alpha \nabla f_k)^T Q (x_k - \alpha \nabla f_k) - b^T (x_k - \alpha \nabla f_k) + c \\ &= f_k + \frac{1}{2}\alpha^2 \nabla f_k^T Q \nabla f_k - \alpha (Qx - b)^T \nabla f_k \\ &= f_k + \frac{1}{2}\alpha^2 \nabla f_k^T Q \nabla f_k - \alpha \nabla f_k^T \nabla f_k \end{aligned}$$

Then by differentiating the function with respect to $\alpha$, we have

$$\frac{d}{d\alpha} f(x_k - \alpha \nabla f_k) = \alpha \nabla f_k^T Q \nabla f_k - \nabla f_k^T \nabla f_k.$$

By setting this expression to zero, we obtain the exact step length

$$\alpha_k = \frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k}.$$

The steepest descent iterations become

$$x_{k+1} = x_k - \left( \frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k} \right) \nabla f_k. \tag{1}$$

Since $\nabla f_k = Q x_k - b$, the iterations are closed-form expression for $x_{k+1}$ in terms of $x_k$.

# Convergence Rate of Steepest Descent

Let $\|x\|_Q^2 = x^T Q x$. By using the relation $Qx^* = b$, we have

$$\frac{1}{2}\|x - x^*\|_Q^2 = f(x) - f(x^*).$$

This norm measures the difference between the current objective value and the optimal value. By using equation (**??**), we can obtain

$$\|x_{k+1} - x^*\|_Q^2 = \left\{1 - \frac{\left(\nabla f_k^T \nabla f_k\right)^2}{\left(\nabla f_k^T Q \nabla f_k\right)\left(\nabla f_k^T Q^{-1} \nabla f_k\right)}\right\} \|x_k - x^*\|_Q^2$$

## Convergence Rate of Steepest Descent

***Theorem***

When the steepest descent method with exact line searches is applied to the strongly convex quadratic function, the error norm satisfies

$$\|x_{k+1} - x^*\|_Q^2 \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right) \|x_k - x^*\|_Q^2,$$

where $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are the eigenvalues of $Q$.

- As the condition $k(Q) = \lambda_n / \lambda_1$ increases the convergence depreciates.
- The rate-of-convergence behavior of the steepest descent method is essentially the same on general nonlinear objective functions.

## Convergence Rate of Steepest Descent

### *Theorem*

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, and that the iterates generated by the steepest descent method with exact line searches converge to a point $x^*$ at which the Hessian matrix $\nabla^2 f(x^*)$ is positive definite. Let $r$ be any scalar satisfying

$$r \in \left( \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}, 1 \right),$$

where $\lambda_1 \leq \cdots \leq \lambda_n$ are the eigenvalues of $\nabla^2 f(x^*)$. Then for all $k$ sufficiently large, we have

$$f(x_{k+1}) - f(x^*) \leq r^2 [f(x_k) - f(x^*)]$$

## Convergence Rate of Steepest Descent

- In general, as the condition number $k(Q) = \lambda_n/\lambda_1$ increases, the contours of the quadratic become more elongated, the zigzagging of the steepest descent method becomes more pronounced.
- In special, if all the eigenvalues are equal, the convergence is achieved in one iteration.

# Convergence Rate of Newton's Method

- Consider the direction of Newton's method iteration

$$p_k^N = \nabla^2 f_k^{-1} \nabla f_k$$

- Since the Hessian is not always positive definite, $p_k$ may not always be a descent direction.
- The following theorem is about local rate-of-convergence properties of Newton's method.

# Convergence Rate of Newton's Method

***Theorem (Local Convergence of Newton's Method)***

Suppose that $f$ is twice differentiable and the Hessian $\nabla^2 f(x)$ is Lipschitz continuous in a neighborhood of a solution $x^*$ at which the second order sufficient conditions are satisfied. Consider the iteration $x_{k+1} = x_k + p_k$, where $p_k^N = \nabla^2 f_k^{-1} \nabla f_k$. Then

(i) if the initial point $x_0$ is sufficiently close to $x^*$, the sequence of iterates converges to $x^*$;

(ii) the rate of convergence of $\{x_k\}$ is quadratic;

(iii) the sequence of gradient norms $\{\|\nabla f_k\|\}$ converges quadratically to zero.

## Convergence Rate of Newton's Method

Proof.

$$\begin{aligned} x_k + p_k^N - x^* &= \left(\nabla^2 f_k^{-1} \nabla^2 f_k\right) x_k - \nabla^2 f_k^{-1} \nabla f_k - \left(\nabla^2 f_k^{-1} \nabla^2 f_k\right) x^* \\ &= \nabla^2 f_k^{-1} \left[\nabla^2 f_k (x_k - x^*) - \nabla f_k\right] \\ &= \nabla^2 f_k^{-1} \left[\nabla^2 f_k (x_k - x^*) - (\nabla f_k - \nabla f_*)\right] \end{aligned}$$

From Taylor's theorem,

$$\nabla f_k - \nabla f_* = \int_0^1 \nabla^2 f(x_k + t(x^* - x_k))(x_k - x^*) dt$$

and also using Lipschitz condition ($L$ is the Lipschitz constant for $\nabla^2 f(x)$ near $x^*$)

$$\begin{aligned} \|\nabla^2 f(x_k) - \nabla^2 f(x_k + t(x^* - x_k))\| &\leq L\|x_k - [x_k + t(x^* - x_k)]\| \\ &= Lt\|x^* - x_k\|, t \in (0,1) \end{aligned}$$

## Convergence Rate of Newton's Method

$$\|\nabla^2 f_k(x_k - x^*) - (\nabla f_k - \nabla f_*)\| =$$

$$\begin{aligned}
&= \left\| \int_0^1 \left[\nabla^2 f_k(x_k - x^*) - \nabla^2 f(x_k + t(x^* - x_k))(x_k - x^*)\right] dt \right\| \\
&= \left\| \int_0^1 \left[\nabla^2 f_k - \nabla^2 f(x_k + t(x^* - x_k))\right] (x_k - x^*)] dt \right\| \\
&\leq \int_0^1 \|\nabla^2 f_k - \nabla^2 f(x_k + t(x^* - x_k))\| \|x_k - x^*\| dt \\
&\leq \|x_k - x^*\| \int_0^1 Lt \|x^* - x_k\| dt \\
&= \|x_k - x^*\|^2 L \int_0^1 t dt = \frac{1}{2} L \|x_k - x^*\|^2,
\end{aligned}$$

where $L$ is the Lipschitz constant for $\nabla^2 f(x)$ for $x$ near $x^*$.

## Convergence Rate of Newton's Method

Since $\nabla^2 f(x^*)$ is nonsingular, $\exists r > 0$ such that $\|\nabla^2 f_k^{-1}\| \leq 2\|\nabla^2 f(x^*)^{-1}\|$ for all $x_k$ with $\|x_k - x^*\| \leq r$.
Then

$$\begin{aligned}
\|x_k + p_k^N - x^*\| &= \|\nabla^2 f_k^{-1}\left[\nabla^2 f_k(x_k - x^*) - (\nabla f_k - \nabla f_*)\right]\| \\
&\leq \|\nabla^2 f_k^{-1}\| \|\nabla^2 f_k(x_k - x^*) - (\nabla f_k - \nabla f_*)\| \\
&\leq 2\|\nabla^2 f(x^*)^{-1}\| \frac{1}{2} L \|x_k - x^*\|^2 \\
&= \tilde{L}\|x_k - x^*\|^2,
\end{aligned}$$

where $\tilde{L} = L\|\nabla^2 f(x^*)^{-1}\|$.
Choosing $x_0$ such that $\|x_0 - x^*\| \leq \min(r, 1/(2\tilde{L}))$, we obtain that the sequence converges to $x^*$, and the rate of convergence is quadratic.

## Convergence Rate of Newton's Method

By using Taylor's theorem, $x_{k+1} - x_k = p_k$ and $\nabla f_k + \nabla^2 f_k p_k = 0$, we have

$$\begin{aligned}
\|\nabla f(x_{k+1})\| &= \|\left[\nabla f_k + \int_0^1 \nabla^2 f(x_k + tp_k)p_k dt\right] - (\nabla f_k + \nabla^2 f_k p_k)\| \\
&= \|\int_0^1 \left[\nabla^2 f(x_k + tp_k) + \nabla^2 f_k\right] p_k dt\| \\
&\leq \|p_k\| \int_0^1 \|\nabla^2 f(x_k + tp_k) + \nabla^2 f_k\| dt \\
&\leq \|p_k\| \int_0^1 L\|(x_k + tp_k) - x_k\| dt = L\|p_k\| \int_0^1 \|tp_k\| dt \\
&\leq \frac{L}{2}\|p_k\|^2 = \frac{L}{2}\| - \nabla^2 f_k^{-1} \nabla f_k\|^2 \leq \frac{L}{2}\|\nabla^2 f_k^{-1}\|^2 \|\nabla f_k\|^2 \\
&\leq 2L\|\nabla^2 f(x^*)^{-1}\|^2 \|\nabla f_k\|^2 = \hat{L}\|\nabla f_k\|^2,
\end{aligned}$$

where $\hat{L} = 2L\|\nabla^2 f(x^*)^{-1}\|^2$. Then the gradient norms converge to zero quadratically.

# Convergence Rate of Quasi-Newton Methods

***Theorem***

Suppose that $f$ is twice continuously differentiable. Consider the iteration $x_{k+1} = x_k + \alpha_k p_k$, where $p_k$ is a descent direction and $\alpha_k$ satisfies the Wolfe conditions with $c_1 \le 1/2$. If the sequence $\{x_k\}$ converges to a point $x^*$ such that $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite, and if the search direction satisfies

$$\lim_{k\to\infty} \frac{\|\nabla f_k + \nabla^2 f_k p_k\|}{\|p_k\|} = 0$$

then

(i) the step length $\alpha_k = 1$ is admissible for all $k$ greater than a certain index $k_0$;

(ii) if $\alpha_k = 1$ for all $k > k_0$, $\{x_k\}$ converges to $x^*$ superlinearly.

# Convergence Rate of Quasi-Newton Methods

If $p_k$ is a quasi-Newton search direction of the form

$$p_k = -B_k^{-1} \nabla f_k,$$

then

$$\lim_{k \to \infty} \frac{\|\nabla f_k + \nabla^2 f_k p_k\|}{\|p_k\|} = 0$$

is equivalent to

$$\lim_{k \to \infty} \frac{\|\left(B_k - \nabla^2 f(x^*)\right) p_k\|}{\|p_k\|} = 0.$$

**Theorem (Convergence of Quasi-Newton Methods)**

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable. Consider the iteration $x_{k+1} = x_k + p_k$ and that $p_k = B_k^{-1} \nabla f_k$, where $B_k$ is symetric positive definite and updated at every iteration. Let us assume also that $\{x_k\}$ converges to a point $x^*$ such that $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) > 0$. Then $\{x_k\}$ converges superlinearly if and only if

$$\lim_{k \to \infty} \frac{\| \left( B_k - \nabla^2 f(x^*) \right) p_k \|}{\|p_k\|} = 0. \quad (2)$$

## Convergence Rate of Quasi-Newton Method

Proof. First, show that condition (2) is equivalent to

$$p_k - p_k^N = o(\|p_k\|).$$

Let $p_k^N = -\nabla^2 f_k^{-1} \nabla f_k$ be the Newton step.
Since $p_k = -B_k^{-1} \nabla f_k$, then we can write $\nabla f_k = -B_k p_k$. Assume that (**??**) holds.

$$\begin{aligned}
p_k - p_k^N &= (\nabla^2 f_k^{-1} \nabla^2 f_k) p_k + \nabla^2 f_k^{-1} \nabla f_k \\
&= \nabla^2 f_k^{-1} (\nabla^2 f_k - B_k) p_k \\
&= \mathcal{O}(\|(\nabla^2 f_k - B_k) p_k\|) \\
&= o(\|p_k\|)
\end{aligned}$$

where $\|\nabla^2 f_k^{-1}\|$ is assumed to be bounded above for $x_k$ sufficiently close to $x^*$, since $\nabla^2 f(x^*)$ is positive definite.

# Convergence Rate of Quasi-Newton Method

$$\begin{aligned}
\|x_k + p_k - x^*\| &\leq \underbrace{\|x_k + p_k^N - x^*\|} + \|p_k - p_k^N\| \\
&\leq \quad \tilde{L}\|x_k - x^*\|^2 \\
&= \mathcal{O}(\|x_k - x^*\|^2) + o(\|p_k\|)
\end{aligned}$$

Manipulating this inequality, we have that $\|p_k\| = \mathcal{O}(\|x_k - x^*\|)$.

$$\|x_k + p_k - x^*\| \leq o(\|x_k - x^*\|)$$

Then the convergence is superlinear.