Lecture-6

Convergence and order of convergence

Line Search Methods

$$x_{k+1} \leftarrow x_k + a_k p_k$$

$$p_k \leftarrow -B_k^{-1} \nabla f_k$$

Steepest descent is and identity matrix
Newton is a Hessian matrix
Quasi-Newton is approximation to the Hessian matrix

Line Search Methods

$$x_{k+1} \leftarrow x_k + a_k p_k$$

$$p_i^T A p_j = 0 \qquad \forall \mathrm{i} \neq \mathrm{j}$$

Conjugate gradient

Important Questions

- What are the conditions under which, the method converges?
- What is the rate of convergence?

Conditions of convergence

- Steepest Descent: Wolf's conditions
- Newton and Quasi-Newton: In addition to Wolfe's conditions, PD Hessian, and bounded condition number
- Conjugate Gradient: subsequence of direction cosines $\cos\theta_k$ is bounded away from zero.

Convergence Rate

- Steepest descent: Linear
- Quasi-Newton: Super-linear
- Newton: Quadratic
- Conjugate Gradient: *n* steps

Convergence of Line Search Methods

- The steepest descent method is globally convergent
- For other algorithms how far p_k can deviate from the steepest descent direction and still gives rise to globally convergent iteration.

Convergence of Line Search Methods (Theorem 3.2)

The angle between p_k and steepest descent direction $-\nabla f_k^T$

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

We will show (Theorem 3.2):

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

Convergence of Line Search Methods

$x_{k+1} = x_k + a_k p_k$ Iteration scheme

$\nabla f(x_k + a p_k)^T p_k \geq c_2 \nabla f_k^T (x_k) p_k, \quad c_2 \in (c_1, 1)$ Curvature condition

Therefore

$$(\nabla f_{k+1} - \nabla f_k)^T p_k \geq (c_2 - 1)\nabla f_k^T p_k \qquad (1)$$

$\| \nabla f(x) - \nabla f(\tilde{x}) \| \leq L \| x - \tilde{x} \|$ Lipschitz continuous

$$(\nabla f_{k+1} - \nabla f_k)^T p_k \leq \| (\nabla f_{k+1} - \nabla f_k)^T \| \| p_k \|$$

$$\leq a_k L \| p_k \| \| p_k \|$$

$$(\nabla f_{k+1} - \nabla f_k)^T p_k \leq a_k L \| p_k \|^2 \qquad (2)$$

Convergence of Line Search Methods

$$(\nabla f_{k+1} - \nabla f_k)^T p_k \leq a_k L \| p_k \|^2 \qquad (2)$$

$$\frac{(\nabla f_{k+1} - \nabla f_k)^T p_k}{L \| p_k \|^2} \leq a_k$$

$$a_k \geq \frac{(\nabla f_{k+1} - \nabla f_k)^T p_k}{L \| p_k \|^2}$$

$$(\nabla f_{k+1} - \nabla f_k)^T p_k \geq (c_2 - 1)\nabla f_k^T p_k \qquad (1)$$

Combining (1) and (2)

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

Convergence of Line Search Methods

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

$$f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha_k \nabla f_k^T p_k \qquad \text{Sufficient decrease}$$

Therefore

$$f_{k+1} \leq f_k - c_1 \frac{1 - c_2}{L} \frac{(\nabla f_k^T p_k)^2}{\| p_k \|^2}$$

$$f_{k+1} \leq f_k - c \cos^2 \theta_k \| \nabla f_k \|^2, \quad c = c_1(1 - c_2)/L$$

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

Convergence of Line Search Methods

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

Since f is bounded below, we have f_0-f_{k+1} is less than some positive constant for all k

Taking the limits:

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

Convergence of Line Search Methods

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

$$\cos\theta_k \geq \delta > 0$$

If angle is bounded away
From 90^0

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

We can be sure that gradient norms converges to zero, provided that
the search directions are never too close to orthogonality with the gradient

Therefore, the steepest descent produces a gradient sequence that converges to zero,
provided that it uses a line search satisfying Wolf's conditions.

We can not guarantee that the method converges to a minimizer,
but only that it is attracted by stationary points.

Newton-Like

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

Assume Hessian is a PD with a uniformly bounded condition number

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

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

Show that (Homework)

$$\cos\theta_k \geq \frac{1}{M}$$

Newton-Like

$$\cos q_k \geq \frac{1}{M}$$

Using $\sum_{k \geq 0} \cos^2 q_k \, \| \nabla f_k \|^2 < \infty$ Theorem 3.2

$$\cos^2 q_k \, \| \nabla f_k \|^2 \to 0$$

Therefore

$$\cos^2 q_k \geq 0$$

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

Therefore:
We have shown that :
Newton and Quasi Newton
are globally convergent
if Hessians have bounded condition
numbers and are PD, and if the step
lengths satisfy Wolf's conditions

Conjugate Gradient

Only subsequence of the gradient norms converges to zero, rather than the whole sequence.

$$\lim_{k \to \infty} \inf \| \nabla f_k \|^2 = 0$$

Sketch of proof by contradiction:

$$\| \nabla f_k \| \geq g$$

Then $\cos^2 q_k \, \| \nabla f_k \|^2 \to 0$

Implies

$$\cos q_k \to 0$$

Therefore it is enough to show that a subsequence $\{\cos q_k\}$ is bounded away from zero.

General Class of Algorithms

- Algorithm
  - Every iteration produces a decrease in the objective function
  - Every *m the*-th iteration is a steepest descent step, with the step length chosen to satisfy the Wolf's conditions.
- Then
  - Since $\cos q_k = 1$ for steepest descent, then following holds

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

Convergence Rate of Steepest Descent: Quadratic Function

$$\| x_{k+1} - x^* \|_Q^2 \le \left(\frac{l_n - l_1}{l_n + l_1} \right)^2 \| x_k - x^* \|_Q^2$$

Theorem 3.3

As the condition number increases the contours of the quadratic become more elongated, the zigzags of line search becomes more pronounced.

Theorem 3.4: Steepest Descent

$$f(x_{k+1}) - f(x^*) \le \left(\frac{l_n - l_1}{l_n + l_1} \right)^2 (f(x_k) - f(x^*))$$

where $0 \le l_1 \le l_2 \le \dots l_n$ are eigenvalues of Hessian

If the condition number is 800, and $f(x_1)=1$ and $f(x^*)=0$,
After 1000 iterations the value of function will be .08.

Theorem 3.5

Suppose *f* is three times continuously differentiable. Consider iteration , where is a descent direction, Satisfies Wolfe's conditions, with . If the converges to a point such that and is pd, and if the search direction satisfies

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

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

Then

(i) if is admissible for all $k>k_0$

(ii) If for all $k>k_0$, then $\{x_k\}$ converges to x^* superlinearly.

Theorem 3.6

Suppose *f* is three times continuously differentiable. Consider iteration , where is given by Quasi-Newton direction. Assume the sequence converges to a point such that and is pd, the converges superlinearly ifif the following condition holds.

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

Order Notations

Given two non-negative infinite sequences

$$\boldsymbol{h}_k = O(\boldsymbol{n}_k)$$
$$if \quad |\boldsymbol{h}_k| \le C |\boldsymbol{n}_k|, \text{ for } C > 0, \forall k$$

$$\boldsymbol{h}_k = o(\boldsymbol{n}_k)$$
$$if \quad \lim_{\mathrm{k} \to \infty} \frac{\boldsymbol{h}_k}{\boldsymbol{n}_k} = 0$$

Sketch of a Proof

$$\boldsymbol{h}_k = O(\boldsymbol{n}_k)$$

$$if \quad |\boldsymbol{h}_k| \le C|\boldsymbol{n}_k|, \text{for } C > 0$$

$$\boldsymbol{h}_k = o(\boldsymbol{n}_k)$$

$$if \quad \lim_{k \to \infty} \frac{\boldsymbol{h}_k}{\boldsymbol{n}_k} = 0$$

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

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

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

$$\begin{aligned} \| x_k + p_k - x^* \| &\le \| x_k + p_k^N - x^* \| + \| p_k - p_k^N \| \\ &= O(\| x_k - x^* \|^2) + o(\| p_k \|) \end{aligned}$$

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

Theorem 3.7

Suppose that f is twice differentiable and that Hessian is Lipschitze continuous. Consider the iteration where is given by

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

Then:

1. If the starting point x_0 is sufficiently close to x^*, the sequence converges to x^*.

2. The rate of convergence is quadratic
3. The sequence of gradient norms converges quadratically to zero.

Coordinate Descent Method

Cycle through *n* coordinate directions $e_1, e_2, \ldots e_n$ using each in turn as a search direction.

Fix all other variables except one, and minimize the function.

It is an inefficient method, it can iterate infinitely without Ever approaching a point, where the gradient vanishes.

The gradient may become more and more perpendicular to search Directions, making $\cos q$ approach to zero, but not the gradient.