Lecture-7

Step Length Selection

Homework (Due 2/13/03)

- 3.1
- 3.2
- 3.5
- 3.6
- 3.7
- 3.9
- 3.10
- Show equation 3.44
- The last step in the proof of Theorem 3.6. (see slides)
- Show that if \( c_1 > 0.5 \), the line search would exclude the minimizer of a quadratic, and unit step length may not be admissible. (Theorem 3.5)
Sufficient condition

\[ f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \nabla f_k^T p_k, \quad c_1 \in (0, 1) \]

\[ c_1 = 10^{-4} \]

\[ f(x_k + \alpha p_k) - f(x_k) \leq c_1 \alpha \nabla f_k^T p_k, \quad c_1 \in (0, 1) \]

The reduction should be proportional to both the step length, and directional derivative.

\[ f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \nabla f_k^T p_k, \quad c_1 \in (0, 1) \]

\[ f(x_k + \alpha p_k) \leq l(\alpha) \]

St line

Problem:

The sufficient decrease condition is satisfied for all small values of step length.
Curvature condition

\[ \nabla f(x_k + \alpha p_k)^T p_k \geq c_2 \nabla f_k^T(x_k) p_k, \quad c_2 \in (c_1, 1) \]

The slope of \( \phi(\alpha_k) \) is greater than \( c_2 \) times the gradient \( \phi'(0) \).

Derivative

\[ c_2 = .9 \text{ for Newton and Quasi - Newton} \]
\[ c_2 = .1 \text{ for conjugate gradient} \]

Curvature condition

If the slope is strongly negative, that means we can reduce \( f \) further along the chosen direction.

If the slope is positive, it indicates we can not decrease \( f \) further in this direction.
Wolfe conditions

\[ f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \nabla f_k^T p_k, \quad c_1 \in (0, 1) \]  
Sufficient decrease

\[ \nabla f(x_k + \alpha p_k)^T p_k \geq c_2 \nabla f_k^T (x_k) p_k, \quad c_2 \in (c_1, 1) \]  
Curvature

Backtracking Line Search

If line search method chooses its step length appropriately, we can dispense with the second condition

Choose \( \overline{\alpha} > 0, \rho, c \in (0, 1) \); set \( \alpha \leftarrow \overline{\alpha} \);

repeat until \( f(x_k + \alpha p_k) \leq f(x_k) + c\alpha \nabla f_k^T p_k \)

\( \alpha \leftarrow \rho \alpha \);

end(repeat)

Terminate with \( \alpha_k = \alpha \)

This ensures that the step length is short enough to satisfy the sufficient decrease condition, but not too short.
Searching Step Length Using Interpolation

\[ f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \nabla f_k^T p_k, \quad c_1 \in (0, 1) \quad \text{Sufficient decrease} \]
\[ \phi(\alpha_k) \leq \phi(0) + c_1 \alpha \phi'(0) \]
\[ \phi(\alpha) = f(x_k + \alpha p_k) \]

1. Assume \( \alpha_0 \) is the initial guess. Then if we have:
\[ \phi(\alpha_0) \leq \phi(0) + c_1 \alpha_0 \phi'(0) \]
Then this step length satisfies the condition, we terminate the search.

2. Otherwise, we know \([0, \alpha_0]\) contains the acceptable step lengths. We fit quadratic polynomial to three pieces of information:
\[ \phi_i(0) = \phi(0), \phi_i'(0) = \phi'(0), \phi_i(\alpha_0) = \phi(\alpha_0) \]
If necessary we can repeat this process with \( \alpha_0 \) and two most recent values of \( \phi \).

3. If not we fit cubic polynomial to interpolate four pieces of information, and analytically minimize this polynomial to find:
\[ \phi_i(0) = \phi(0), \phi_i'(0) = \phi'(0), \phi_i(\alpha_0) = \phi(\alpha_0), \phi_i(\alpha_i) = \phi(\alpha_i) \]
If necessary we can repeat this process with \( \phi(0), \phi'(0) \) and two most recent values of \( \phi \).
Quadratic Interpolation

\[ \phi_2(\alpha) = a \alpha^2 + b \alpha + c \]
\[ \phi_2(0) = \phi(0), \phi'_2(0) = \phi'(0), \phi_2(\alpha_o) = \phi(\alpha_o) \]
\[ \phi_2(\alpha) = \left( \frac{\phi(\alpha_o) - \phi(0) - \alpha_o \phi'(0)}{\alpha_o^2} \right) \alpha^2 + \phi'(0) \alpha + \phi(0) \]
\[ \frac{d}{d\alpha} \phi_2(\alpha) = 2 \left( \frac{\phi(\alpha_o) - \phi(0) - \alpha_o \phi'(0)}{\alpha_o^2} \right) \alpha + \phi'(0) = 0 \]
\[ \alpha_i = -\left( \frac{\phi'(0) \alpha_o^2}{2(\phi(\alpha_o) - \phi(0) - \alpha_o \phi'(0))} \right) \]

Cubic Interpolation

\[ \phi_i(\alpha) = a \alpha^3 + b \alpha^2 + c \alpha + d \]
\[ \phi_i(0) = \phi(0), \phi'_i(0) = \phi'(0), \phi_i(\alpha_o) = \phi(\alpha_o), \phi_i(\alpha_i) = \phi(\alpha_i) \]
\[ \phi_i(\alpha) = a \alpha^3 + b \alpha^2 + \phi'(0) \alpha + \phi(0) \]

\[
\begin{bmatrix}
    a \\
    b
\end{bmatrix} = 
\begin{bmatrix}
    \alpha_o^2 & -\alpha_i^2 & \alpha_o \\
    \alpha_o \alpha_i (\alpha_i - \alpha_o) & -\alpha_o^2 & \alpha_o \\
    \alpha_i^2 & \alpha_o \alpha_i (\alpha_i - \alpha_o) & -\alpha_o^2
\end{bmatrix} 
\begin{bmatrix}
    \phi(\alpha_i) - \phi(0) - \phi'(0) \alpha_i \\
    \phi(\alpha_o) - \phi(0) - \phi'(0) \alpha_o \\
    \phi(\alpha_o) - \phi(0) - \phi'(0) \alpha_o
\end{bmatrix}
\]
\[ \alpha_z = \left( \frac{-b + \sqrt{b^2 - 3a\phi'(0)}}{3a} \right) \]
Algorithm 3.2 (Line Search Algorithm)

Set $\alpha_0 \leftarrow 0$, choose $\alpha_i > 0$, and $\alpha_{\text{max}}$:

$i \leftarrow 1$

repeat

Evaluate $\phi(\alpha_i)$;

if $\phi(\alpha_i) > \phi(0) + c_i \alpha_i \phi'(0)$ or $[\phi(\alpha_i) > \phi(\alpha_{i-1}), i > 1]$

$\alpha_i \leftarrow \text{zoom}(\alpha_{i-1}, \alpha_i)$, and stop;

Evaluate $\phi'(\alpha_i)$;

if $|\phi'(\alpha_i)| \leq -c_i \phi'(0)$

set $\alpha_i \leftarrow \alpha_i$, and stop;

if $\phi'(\alpha_i) \geq 0$

set $\alpha_i \leftarrow \text{zoom}(\alpha_{i}, \alpha_{i-1})$, and stop;

choose $\alpha_{i+1} \in (\alpha_i, \alpha_{\text{max}})$

$i \leftarrow i + 1$

end(repeat)

Algorithm 3.3 (Zoom)

repeat

Interpolate to find a trial step length

$\alpha_j$ between $\alpha_{lo}, \alpha_{hi}$;

Evaluate $\phi(\alpha_j)$;

if $\phi(\alpha_j) > \phi(0) + c_i \alpha_j \phi'(0)$ or $[\phi(\alpha_j) > \phi(\alpha_{lo})]$

$\alpha_{hi} \leftarrow \alpha_j$;

esle

Evaluate $\phi'(\alpha_j)$;

if $|\phi'(\alpha_j)| \leq -c_i \phi'(0)$

set $\alpha_{hi} \leftarrow \alpha_j$, and stop;

if $\phi'(\alpha_j)(\alpha_{hi} - \alpha_{lo}) \geq 0$

set $\alpha_{hi} \leftarrow \alpha_{lo}$

set $\alpha_{lo} \leftarrow \alpha_j$

end(repeat)
Figure 8.1: The ideal step-length is the global minimizer.
Theorem 3.5 (Any Descent Direction)

Suppose \( f \) is three times continuously differentiable. Consider iteration \( x_{k+1} = x_k + \alpha_k p_k \), where \( p_k \) is a descent direction, \( \alpha_k \) Satisfies Wolfe’s conditions, with \( c_1 \sim \frac{1}{2} \). If the \( \{x_k\} \) converges to a point \( x^* \) such that \( \nabla f(x^*) = 0 \) and \( \nabla^2 f(x^*) \) 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) \( \alpha_k = 1 \) is admissible for all \( k > k_0 \) and

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

Theorem 3.6 (Quasi-Newton)

Suppose \( f \) is three times continuously differentiable. Consider iteration \( x_{k+1} = x_k + p_k \), where \( p_k \) is given by Quasi-Newton direction. Assume the sequence \( \{x_k\} \) converges to a point \( x^* \) such that \( \nabla f(x^*) = 0 \) and \( \nabla^2 f(x^*) \) is pd, the \( \{x_k\} \) converges superlinearly if

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

Given two non-negative infinite sequences

\[ \eta_k = O(v_k) \]

if \[ |\eta_k| \leq C |v_k| \], for \( C > 0, \forall k \)

\[ \eta_k = o(v_k) \]

if \( \lim_{k \to \infty} \frac{\eta_k}{v_k} = 0 \)

Sketch of a Proof

\[ p_k - p_k^N = \nabla^2 f_k^{-1} (\nabla^2 f_k p_k + \nabla f_k) \]
\[ = \nabla^2 f_k^{-1} (\nabla^2 f_k - B_k) p_k \]
\[ = O(\| (\nabla^2 f_k - B_k) p_k \|) \]
\[ = o(\| p_k \|) \]

\[ \| (B_k - \nabla^2 f(x^*) p_k \| = 0 \]

Norm of Hessian is bounded.
Sketch of a Proof

\[ \| x_k + p_k - x^* \| = \| x_k + p_k^N - p_k^N + p_k - x^* \| \leq \| x_k + p_k^N - x^* \| + \| p_k - p_k^N \| \]

= \text{O}(\| x_k - x^* \|^2) + o(\| p_k \|)

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

\[ \eta_k = o(\nu_k) \]

\[ \text{if } \lim_{k \to \infty} \frac{\eta_k}{\nu_k} = 0 \]

Theorem 3.7

Super-linear

Show this in Homework

Theorem 3.7 (Newton)

Suppose that \( f \) is twice differentiable and that Hessian is Lipschitze continuous. Consider the iteration \( x_{k+1} = x_k + p_k \) where \( p_k \) 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 \( \| \nabla f_k \| \) 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 $\cos \theta$ approach to zero, but not the gradient.