Lecture 14

Newton Algorithm for Unconstrained Optimization

October 21, 2008
Outline

• Newton Method for System of Nonlinear Equations
• Newton’s Method for Optimization
• Classic Analysis
Newton’s Method for System of Equations

- A numerical method for solving a system of equations
  \[ G(x) = 0, \quad G : \mathbb{R}^n \to \mathbb{R}^n \]

- When \( G \) is continuously differentiable, the classical Newton method is based on a natural (local) approximation of \( G \): linearization
  - Given an iterate \( x_k \), the map \( G \) is approximated at \( x_k \) by the following linear map
    \[
    L(x; x_k) = G(x_k) + JG(x_k)(x - x_k),
    \]
    where \( JG(x) \) is the Jacobian of \( G \) at \( x \)
  - We have \( L(x; x_k) \approx G(x) \)
  - System \( G(x) = 0 \) is “replaced” by the system \( L(x; x_k) = 0 \)
  - The resulting solution is defining a new iterate \( x_{k+1} \),
    \[
    x_{k+1} = x_k - JG(x_k)^{-1}G(x_k)
    \]
Properties of Newton’s Method

\[ x_{k+1} = x_k - JG(x_k)^{-1}G(x_k) \]

- Fast local convergence, but globally the method can fail
  - When started far from a solution
• Numerical instabilities occur when $JG(x^*)$ is singular (or nearly singular)

• Two main properties that make the process work
  • A linear model $L(x; x_k)$ that provides good approximation of $G$ near $x_k$, when $x_k$ is near a solution
  • Solvability of linear equation $L(x; x_k) = 0$ when $JG(x_k)$ is invertible

• These two properties are “guaranteed” when
  • $G$ is continuously differentiable and
  • $JG(x^*)$ is invertible (nonsingular)
Convergence Rate Terminology

**Definition 7.2.1** Let $\{x_k\} \subseteq \mathbb{R}^n$ be a sequence converging to some $x^* \in \mathbb{R}^n$. The convergence rate is said to be

- **$Q$-linear** if

  $$\limsup_{k \to \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} < \infty$$

- **$Q$-superlinear** if

  $$\limsup_{k \to \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0$$

- **$Q$-quadratic** if

  $$\limsup_{k \to \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^2} < \infty$$

- **$R$-linear** if

  $$\limsup_{k \to \infty} (\|x_{k+1} - x^*\|)^{1/k} < 1$$
Unconstrained Minimization

minimize \( f(x) \)

Suppose that:
- The function \( f \) is convex and twice continuously differentiable over \( \text{dom} f \)
- The optimal value is attained: there exists \( x^* \) such that

\[
f(x^*) = \inf_x f(x)
\]

Newton Method can be applied to solve the corresponding optimality condition

\[
\nabla f(x^*) = 0,
\]
resulting in \( x_{k+1} = x_k - \nabla f^2(x_k)^{-1} \nabla f(x_k) \).
This is known as pure Newton method
As discussed, in this form the method may not always converge.
Newton’s direction

\[ d_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k) \]

Interpretations:

• Second-order Taylor’s expansion at \( x_k \) yields

\[ f(x_k + d) = f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T \nabla^2 f(x_k) d + o(\|d\|^2) \]

• The right-hand side without the small order term, provides a (quadratic) approximation of \( f \) in a (small) neighborhood of \( x_k \)

\[ f(x_k + d) \approx f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T \nabla^2 f(x_k) d \]

• Minimizing the quadratic approximation w/r to \( d \) yields:

\[ \nabla f(x_k)^T + \nabla^2 f(x_k) d = 0 \]

• Newton’s \( d_k \) can be also viewed as solving linearized optimality condition

\[ \nabla f(x_k + d) \approx \nabla f(x_k) + \nabla^2 f(x_k) d = 0 \]
Newton Decrement

- Newton’s decrement at \( x_k \) is defined by:
  \[
  \lambda(x_k) = \left( \nabla f(x_k)^T \nabla^2 f(x_k)^{-1} \nabla f(x_k) \right)^{1/2}
  \]

- Provides a measure of the proximity of \( x \) to \( x^* \)

- Obtained by evaluating the difference between \( f(x_k) \) and the quadratic approximation of \( f \) at \( x_k \) evaluated at the optimal \( d \) (Newton’s direction)
  \[
  f(x_k) - \left[ f(x_k) + \nabla f(x_k)^T d_k + \frac{1}{2} d_k^T \nabla^2 f(x_k) d_k \right] = \frac{1}{2} \lambda(x_k)^2
  \]

Properties:
- Equal to the norm of the Newton step in the quadratic Hessian norm
  \[
  \lambda(x_k) = \left[ d_k \nabla^2 f(x_k) d_k \right]^{1/2} = \| d_k \| \nabla^2 f(x_k)
  \]
- Affine invariant (unlike \( \| \nabla f(x) \| \))
Newton’s Method

**Given** a starting point $x \in \text{dom} f$, error tolerance $\epsilon > 0$, and parameters $\sigma \in (0, 1/2)$ and $\beta \in (0, 1)$.

**Repeat**

1. *Compute the Newton’s direction and decrement:*

   \[ d := -\nabla^2 f(x)^{-1} \nabla f(x), \quad \lambda^2 := \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \]

2. *Stopping criterion: quit if* $\lambda^2 / 2 \leq \epsilon$

3. *Line search: Choose stepsize $\alpha$ by backtracking line search, i.e., starting with* $\alpha = 1$ do

   (a) If $f(x + \alpha d) < f(x) + \sigma \alpha \nabla f(x)^T d$ go to Step 4

   (b) Else $\alpha = \beta \alpha$ and go to (b).

4. *Update* $x := x + \alpha d$. 
Classical Convergence Analysis - Main Results

Assumption 1:

- The level set \( L_0 = \{ x \mid f(x) \leq f(x_0) \} \) is closed
- \( f \) strongly convex on \( L_0 \) with a constant \( m \)
- \( \nabla^2 f \) is Lipschitz continuous on \( L_0 \), with a constant \( L > 0 \):
  \[
  \| \nabla^2 f(x) - \nabla^2 f(y) \| \leq L \| x - y \|_2
  \]
  \( L \) measures how well \( f \) can be approximated by a quadratic function

Analysis outline: there exists a constant \( \eta \in (0, 2m^2/L) \) such that

- When \( \| \nabla f(x_k) \| \geq \eta \), then \( f(x_{k+1}) - f(x_k) \leq -\gamma \) for \( \gamma = \sigma \beta \eta^2 \frac{m}{M^2} \)
- When \( \| \nabla f(x_k) \| < \eta \), then
  \[
  \frac{L}{2m^2} \| \nabla f(x_{k+1}) \|_2 \leq \left[ \frac{L}{2m^2} \| \nabla f(x_k) \|_2 \right]^2
  \]
Two Phases of Newton’s Method

Damped Newton phase \( (\|\nabla f(x)\| \geq \eta; \text{ far from } x^*) \)

- Most iterations require backtracking steps
- Function value decreases by at least \( \gamma \) at each iteration
- This phase ends after at most \( (f(x_0) - f^*)/\gamma \) iterations

Quadratically convergent phase \( (\|\nabla f(x)\| < \eta; \text{ locally near } x^*) \)

- All iterations in this phase use stepsize \( \alpha = 1 \)
- The gradient \( \|\nabla f(x)\| \) converges to zero quadratically:

\[
\frac{L}{2m^2} \|\nabla f(x_l)\| \leq \left[ \frac{L}{2m^2} \|\nabla f(x_k)\| \right]^{2^{l-k}} \leq \left( \frac{1}{2} \right)^{2^{l-k}} \quad \text{for } l \geq k
\]
Analysis of Newton Method

**Theorem** Let Assumption 1 hold. Then, there exists a constant \( \eta \in (0, 2m^2/L) \) such that

- When \( \|\nabla f(x_k)\| \geq \eta \), then \( f(x_{k+1}) - f(x_k) \leq -\gamma \) for \( \gamma = \sigma \beta \eta^2 \frac{m}{M^2} \)
- When \( \|\nabla f(x_k)\| < \eta \), then

\[
\frac{L}{2m^2} \|\nabla f(x_{k+1})\|_2 \leq \left[ \frac{L}{2m^2} \|\nabla f(x_k)\|_2 \right]^2
\]

**Proof** Let \( \|\nabla f(x_k)\| \geq \eta \). Let estimate the stepsize \( \alpha_k \) at the end of the backtracking line search. First note that \( \{x_k\} \subset L_0 \). Since \( f \) is strongly convex, the level set \( L_0 \) is bounded, hence \( \nabla f \) is Lipschitz continuous over \( L_0 \) with some constant \( M \).

By \( Q \)-approximation Lemma, we have for any \( \alpha > 0 \)

\[
f(x_k + \alpha d) \leq f(x_k) + \alpha \nabla f(x_k)^T d_k + \frac{\alpha^2 M}{2} \|d_k\|^2.
\]
Using the Newton decrement $\lambda_k = (\nabla f(x_k)^T \nabla^2 f(x_k)^{-1} \nabla f(x_k)^T)^{1/2}$, and the fact $d_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$ we can write

$$f(x_k + \alpha d_k) \leq f(x_k) - \alpha \lambda_k^2 + \frac{\alpha^2 M}{2} \|d_k\|^2.$$ 

Note that

$$\lambda_k^2 = \nabla f(x_k)^T \nabla^2 f(x_k)^{-1} \nabla f(x_k) = d_k^T \nabla^2 f(x_k) d_k \geq m \|d_k\|^2,$$

by strong convexity of $f$. Hence $\|d_k\|^2 \leq \lambda_k^2 / m$ implying that

$$f(x_k + \alpha d_k) \leq f(x_k) - \alpha \lambda_k^2 + \frac{\alpha^2 M}{2m} \lambda_k^2$$

$$\leq f(x_k) - \alpha \left(1 - \frac{\alpha M}{2m}\right) \lambda_k^2.$$
Thus the stepsize $\alpha = \frac{m}{M}$ satisfies the exit condition in the backtracking line search (with $\sigma \leq 1/2$), since for $\alpha = \frac{m}{M}$ we have

$$f(x_k + \alpha d_k) \leq f(x_k) - \frac{m}{2M} \lambda_k^2 < f(x_k) - \sigma \frac{m}{M} \lambda_k^2.$$  

Therefore, the backtracking line search stops with some $\alpha_k \geq \frac{m}{M} \geq \beta \frac{m}{M}$. Thus, when the line search is exited with step $\alpha_k$, we have

$$f(x_k + \alpha_k d_k) < f(x_k) - \sigma \alpha_k \lambda_k^2 \leq f(x_k) - \sigma \beta \frac{m}{M} \lambda_k^2.$$  

Since $\nabla^2 f(x) \leq MI$, we have

$$\lambda_k^2 = \nabla f(x_k)^T \nabla^2 f(x_k)^{-1} \nabla f(x_k)^T \geq \frac{1}{M} \|\nabla f(x_k)\|.$$
Hence, when $\|\nabla f(x_k)\| \leq \eta$, we have

$$f(x_k + \alpha_k d_k) < f(x_k) - \sigma \beta \frac{m}{M^2} \eta^2.$$ 

Suppose now $\|\nabla f(x_k)\| \leq \eta$.

**Strong Q-lemma**: For a continuously differentiable function $g$ over $\mathbb{R}^n$ with Lipschitz gradients with constant $L$, we have

$$\|g(x + y) - g(x) - \nabla g(x)^T y\| \leq \frac{L}{2} ||y||^2.$$ 

Applying this lemma to $\nabla f(x)$, we have

$$\|\nabla f(x_k + d) - \nabla f(x) - \nabla^2 f(x) d\| \leq \frac{L}{2} ||d||^2.$$
letting $d$ be the Newton direction, we obtain

$$\|\nabla f(x_{k+1})\| \leq \frac{L}{2}\|\nabla^2 f(x_k)^{-1}\nabla f(x_k)\|^2 \leq \frac{L}{2}\|\nabla^2 f(x_k)^{-1}\|^2 \|\nabla f(x_k)\|^2.$$ 

Since $mI \leq \nabla^2 f(x)$, it follows

$$\|\nabla f(x_{k+1})\| \leq \frac{L}{2m^2}\|\nabla f(x_k)\|^2.$$ 

Hence, $\|\nabla f(x_{k+1})\| \leq \eta$. Furthermore, for any $K \geq 0$,

$$\|\nabla f(x_{K+1})\| \leq \frac{2m^2}{L}\left(\frac{L}{2m^2}\|\nabla f(x_K)\|\right)^2 \leq \cdots \leq \left(\frac{L}{2m^2}\|\nabla f(x_k)\|\right)^{2K},$$

showing the quadratic convergence rate.
Conclusions

The number of iterations until \( f(x) - f^* \leq \epsilon \) is bounded above by

\[
\frac{f(x_0) - f^*}{\gamma} + \log_2 \log_2 \left( \frac{\epsilon_0}{\epsilon} \right)
\]

- \( \gamma = \sigma \beta \eta^2 \frac{m}{M^2} \), \( \epsilon_0 = 2m^3/L^2 \)
- The second term is small (of the order of 6) and almost constant for practical purposes:
  six iterations of the quadratically convergent phase results in accuracy

\[
\approx 5 \cdot 10^{-20} \epsilon_0
\]

- In practice, the constants \( m, L \) (hence \( \gamma, \epsilon_0 \)) are usually unknown
- The analysis provides qualitative insight in the convergence properties (i.e., explains two algorithm phases)