EE 381V Large Scale Optimization: Lecture 07

Prof. Sujay Sanghavi

The University of Texas at Austin
Scribes: Jimmy Lin, Vutha Va and David Inouye

September 18, 2014

Newton Method

Definition (Basic Idea and Update Rule)

Newton Method aims at minimizing a quadratic approximation of function f.

$$f(x+v) = f(x) + \nabla f(x)^T v + v^T \nabla^2 f(x) v \tag{1}$$

where RHS is minimized at direction

$$v = -\nabla^2 f(x)^{-1} \nabla f(x) \tag{2}$$

And the update rule for Newton Method is

$$x^+ = x - t\nabla^2 f(x)^{-1} \nabla f(x) \tag{3}$$

where t is fixed step size for optimization.

Other Interpretations

Remark (Relation to Steepest Descent Method)

Newton Method can be interpreted as steepest descent method when the norm is defined as

$$||u||_{\nabla^2 f(x)} \triangleq \sqrt{u^T \nabla^2 f(x) u} \tag{4}$$

Remark (First-Order Approximation)

Newton Method can also be interpreted as to linear approximation over gradient $\nabla f(x)$ around x.

$$\nabla f(x+v) \approxeq \nabla f(x) + \nabla^2 f(x) \tag{5}$$

Set RHS to zero gives newton update.

Affine Invariance of Newton Method

Lemma

Newton Method is affine invariant.

For example, let $g(y) = f(Ay)$, y^+ be newton update on function $g(\cdot)$, and x^+ be newton update on function $f(\cdot)$. Then if $x = Ay$, we have $x^+ = Ay^+$.

Remark

Affine Invariance indicates that Newton Method is vulnerable to the selection of coordinate system. Note thtat Gradient Descent Method is not affine invariance. This means that bad coordinate choice may disable Gradient Descent Method.

Proof of Affine Invariance

Let $x = Ay$ and $g(y) = f(Ay)$, then we have

$$\begin{aligned} \nabla^2_y g(y) &= \nabla^2_y f(Ay) &= A^T \nabla^2_x f(x) A \quad (6) \\ \nabla_y g(y) &= \nabla_y f(Ay) &= A^T \nabla_x f(x) \quad (7) \end{aligned}$$

Newton update y^+ for $g(\cdot)$ can be extended as

$$\begin{aligned} y^+ &= y - t(\nabla^2_y g(y))^{-1} \nabla_y g(y) & (8) \\ &= y - t(A^T \nabla^2_x f(x) A)^{-1} A^T \nabla_x f(x) & (9) \\ &= y - t\, A^{-1} \nabla^2_x f(x)^{-1} \nabla_x f(x) & (10) \end{aligned}$$

Multiply both sides with affine tranformation A,

$$\begin{aligned} Ay^+ &= Ay - A \cdot t\, A^{-1} \nabla^2_x f(x)^{-1} \nabla_x f(x) & (11) \\ &= x - t\, \nabla^2_x f(x)^{-1} \nabla_x f(x) & (12) \\ &= x^+ & (13) \end{aligned}$$

Convergence Analysis: Assumption

Assumption

Let $f(\cdot)$ be the function discussed for Convergence of Newton Method. Both of following assumptions are what convergence analysis is based on.

- *Function $f(\cdot)$ is strongly convex, such that*

$$mI \leq \nabla^2 f(x) \leq MI \tag{14}$$

- *$\nabla^2 f(x)$ is L-Lipschitz with constant $L > 0$, such that*

$$||\nabla^2 f(y) - \nabla^2 f(x)||_2 \leq L||x - y||_2, \ \forall x, \ y \tag{15}$$

  Note that induced matrix norm $||\cdot||_2$ equals to the largest singluar value of inside matrix.

Convergence Analysis: Theorem

Theorem (Part I)

There exists f, η, γ, *where* $0 \leq \eta \leq \frac{m^2}{L}$, $\gamma = \frac{\alpha\beta m}{M^2}\eta^2$ *such that Newton Method with BTLS has two phrases:*

(a) *Global or Damped Phrase: If* $||\nabla f(x)||_2 \geq \eta$, *then*

$$f(x^+) - f(x) \leq -\gamma, \text{ also } f(x^+) - f^* \leq c(f(x) - f^*) \quad (16)$$

Inequality (16) has two implications:

- Every newton step with BTLS gets closer to global optima by at least γ.
- Damped phrase has at most $\frac{f(x^{(0)}) - f^*}{\gamma}$ iterations.
- The damped phrase essentially conforms to property of linear convergence.

Convergence Analysis: Theorem

Theorem (Part II)

(b) *Local or Quadratic Phrase: If* $||\nabla f(x)||_2 < \eta$*, then BTLS will give* $t = 1$ *and we have*

$$\frac{L}{2m^2}||\nabla f(x^+)||_2 \leq \left(\frac{L}{2m^2}||\nabla f(x)||_2\right)^2 \tag{17}$$

Convergence Analysis: Proof

Lemma

$t = \frac{m}{M}$ satisfies the exit condition of BTLS.

Convergence Analysis: Proof

Lemma

If $||\nabla f(x)||_2 \geq \eta$, *then* $f(x^+) - f(x) \leq -\gamma$, *where* $\gamma = \frac{\alpha\beta m}{M^2}\eta^2$

Convergence Analysis: Proof

Lemma

If $||\nabla f(x)||_2 < \eta$, *then* $\frac{L}{2m^2}||\nabla f(x^+)||_2 \leq \left(\frac{L}{2m^2}||\nabla f(x)||_2\right)^2$

Convergence Analysis: Proof

Lemma

$t = 1$ *satisfies the exit condition of BTLS.*