

# 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 21, 2014

# Newton Step

## Definition

For  $x \in \text{dom } f$ , the vector

$$\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x) \quad (1)$$

is called *Newton step* for function  $f(\cdot)$ , at point  $x$ .

In terms of positive definiteness of  $\nabla^2 f(x)$ ,

$$\nabla f(x)^T \Delta x_{nt} = -\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) < 0 \quad (2)$$

unless  $\nabla f(x) = 0$ . So the Newton step is a descent direction unless  $x$  is already optimal.

# Interpretation I

## Minimizer of second-order Approximation

Second-order Taylor approximation  $\hat{f}$  of  $f$  at  $x$  is

$$\hat{f}(x + v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v \quad (3)$$

where RHS is minimized at direction

$$v = -\nabla^2 f(x)^{-1} \nabla f(x) = \Delta x_{nt} \quad (4)$$

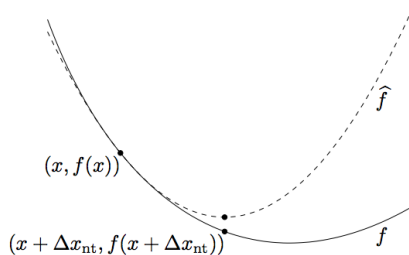
Combined with update rule, we have Newton update

$$x^+ = x - t \nabla^2 f(x)^{-1} \nabla f(x) \quad (5)$$

where  $t$  is fixed step size.

# Interpretation I

## Minimizer of second-order Approximation



**Figure 1 :** The function  $f$  (shown solid) and its second-order approximation  $\hat{f}$  at  $x$  (dashed). The Newton step  $\Delta x_{nt}$  is what must be added to  $x$  to give the minimizer of  $\hat{f}$ .

## Interpretation II

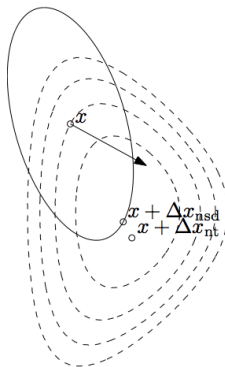
Steepest descent direction in Hessian norm

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

$$\|u\|_{\nabla^2 f(x)} \triangleq \sqrt{u^T \nabla^2 f(x) u} \quad (6)$$

## Interpretation II

Steepest descent direction in Hessian norm



**Figure 2 :** The dashed lines are level curves of a convex function. The ellipsoid shown (with solid line) is  $\{x + v | v^T \nabla^2 f(x) v \leq 1\}$ . The arrow shows  $-\nabla f(x)$ , the gradient descent direction. The Newton step  $\Delta x_{nt}$  is the steepest descent direction in the norm  $\|\cdot\|_{\nabla^2 f(x)}$ .

## Interpretation III

Solution of linearized optimality condition

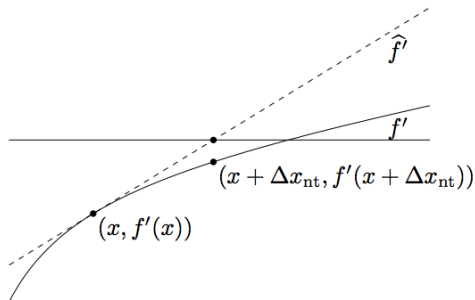
Newton step can also be interpreted as to linear approximation over gradient  $\nabla f(x)$  around  $x$ .

$$\nabla f(x + v) \cong \nabla f(x) + \nabla^2 f(x)v \quad (7)$$

Set RHS to zero gives newton update.

## Interpretation III

### Solution of linearized optimality condition



**Figure 3 :** The solid curve is the derivative  $f'$  of the function  $f$  shown in Figure 2.  $\hat{f}'$  is the linear approximation of  $f'$  at  $x$ . The Newton step  $\Delta x_{nt}$  is the difference between the root of  $\hat{f}'$  and the point  $x$ .



## Affine Invariance of Newton step

### Lemma

*Newton step 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 NOT vulnerable to the selection of coordinate system.*

### Remark

*Gradient Descent Method is not affine invariant. This means that bad coordinate choice may limit the power of Gradient Descent Method.*

## Proof of Affine Invariance

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

$$\nabla_y^2 g(y) = \nabla_y^2 f(Ay) = A^T \nabla_x^2 f(x) A \quad (8)$$

$$\nabla_y g(y) = \nabla_y f(Ay) = A^T \nabla_x f(x) \quad (9)$$

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

$$y^+ = y - t(\nabla_y^2 g(y))^{-1} \nabla_y g(y) \quad (10)$$

$$= y - t(A^T \nabla_x^2 f(x) A)^{-1} A^T \nabla_x f(x) \quad (11)$$

$$= y - t A^{-1} \nabla_x^2 f(x)^{-1} \nabla_x f(x) \quad (12)$$

Multiply both sides with affine tranformation  $A$ ,

$$Ay^+ = Ay - A \cdot t A^{-1} \nabla_x^2 f(x)^{-1} \nabla_x f(x) \quad (13)$$

$$= x - t \nabla_x^2 f(x)^{-1} \nabla_x f(x) \quad (14)$$

$$= x^+ \quad (15)$$

# 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*

$$mI \preceq \nabla^2 f(x) \preceq MI \quad (16)$$

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

$$\|\nabla^2 f(y) - \nabla^2 f(x)\|_2 \leq L\|x - y\|_2, \forall x, y \quad (17)$$

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

# Convergence Analysis: Theorem

## Theorem (Part I)

*There exists  $f$ ,  $\eta$ ,  $\gamma$ , 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 phases:*

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

$$f(x^+) - f(x) \leq -\gamma, \text{ also } f(x^+) - f^* \leq c(f(x) - f^*) \quad (18)$$

Inequality (18) has three implications:

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

# Convergence Analysis: Theorem

## Theorem (Part II)

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

$$\frac{L}{2m^2} \|\nabla f(x^+)\|_2 \leq \left( \frac{L}{2m^2} \|\nabla f(x)\|_2 \right)^2 \quad (19)$$

Implications:

- To achieve an accuracy of  $\epsilon$ , only  $O(\log \log \epsilon)$  iterations are needed once in the quadratic phase.
- Also, for strongly convex functions,  $f(x) \rightarrow p^*$  quadratically.

## Convergence Analysis: Proof Part (a)

### Lemma (BTLS Damped Lemma)

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

Proof omitted because of space. See p. 489-490 in Boyd.

### Lemma (Damped Phase)

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: Lemma Damped Phase

$$t \leq \beta \frac{m}{M} \quad (\text{BTLS Damped Lemma})$$

$$f(x^+) \leq f(x) - \alpha \left( \beta \frac{m}{M} \right) g^T H^{-1} g \quad (\text{BTLS condition})$$

$$\leq f(x) - \alpha \left( \beta \frac{m}{M} \right) \left( \frac{1}{M} \|g\|_2^2 \right) \quad (H^{-1} \preceq I/m)$$

$$= f(x) - \alpha \beta \frac{m}{M^2} \|g\|_2^2$$

$$= f(x) - \underbrace{\alpha \beta \frac{m}{M^2} \eta^2}_{\gamma} \quad (\text{Assumptions})$$

$$\implies f(x^+) - f(x) = -\gamma \quad (20)$$

## Convergence Analysis: Proof Part (b)

### Lemma (BTLS Quad. Lemma)

*With the assumptions in (b),  $t = 1$  satisfies the exit condition of BTLS.*

Proof omitted because of space. See p. 490-491 in Boyd.

### Lemma (Quad. Phase)

*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: Quad. Phase

$$x^+ = x - H^{-1}g \quad (\text{BTLS Quad. Lemma})$$

$$\|\nabla f(x^+)\|_2 = \|\nabla f(x - H^{-1}g) - g + HH^{-1}g\|_2 \quad (\text{Add zero})$$

$$= \left\| \int_0^1 \nabla^2 f(x - tH^{-1}g)(-H^{-1}g) + HH^{-1}g dt \right\|_2$$

(Fund. Theorem of Calculus)

$$= \left\| \int_0^1 (\nabla^2 f(x - tH^{-1}g) - H)(-H^{-1}g) dt \right\|_2$$

(Rearrange)

$$\leq \int_0^1 \left\| (\nabla^2 f(x - tH^{-1}g) - H) \right\|_2 \left\| (-H^{-1}g) \right\|_2 dt$$

(Triangle inequality of norms)

# Convergence Analysis: Quad. Phase

$$\begin{aligned}
 \|\nabla f(x^+)\|_2 &\leq \int_0^1 \left\| (\nabla^2 f(x - tH^{-1}g) - H) \right\|_2 \|H^{-1}g\|_2 dt \\
 &\leq \int_0^1 L \| -tH^{-1}g \|_2 \|H^{-1}g\|_2 dt \\
 &\quad \text{(Lipschitz Continuity of Hessian)} \\
 &= L \|H^{-1}g\|_2^2 \int_0^1 t dt = \frac{L}{2} \|H^{-1}g\|_2^2 \\
 &\leq \frac{L}{2m^2} \|g\|_2^2 \quad \text{(Strong convexity } (H^{-1} \preceq I/m)) \\
 \\
 \implies \frac{L}{2m^2} \|\nabla f(x^+)\|_2 &\leq \left( \frac{L}{2m^2} \|g\|_2 \right)^2 \quad (21)
 \end{aligned}$$