# Large Scale Optimization: Lecture 7

Jimmy Lin, Vutha Va and David Inouye

The University of Texas at Austin

September 23, 2014

# Newton Step

**Definition**

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

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

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

- Note that $f(\cdot)$ is twice differentiable.
- Because $\nabla^2 f(x)$ is positive definite,

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

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

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

where RHS is minimized at direction

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

Combined with update rule, we have Newton update

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

where $t$ is fixed step size.

# Interpretation I

## Minimizer of second-order Approximation

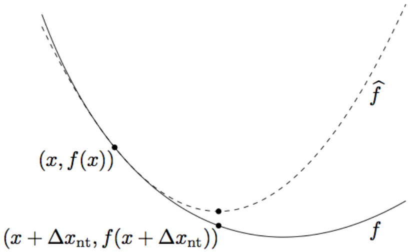


Figure 1 : The function $f$ (shown solid) and its second-order approximation $\hat{f}$ at $x$ (dashed). The Newton step $\Delta x_{nt}$ is what must be added to $x$ to give the minimizer of $\hat{f}$. [Figure and caption from *Convex Optimization* by Boyd and Vandenberghe.]

# Interpretation II

## Steepest descent direction in Hessian norm

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

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

# Interpretation II

## Steepest descent direction in Hessian norm

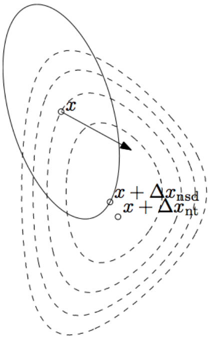


Figure 2 : The dashed lines are level curves of a convex function. The ellipsoid shown (with solid line) is $\{x + v | v^T \nabla^2 f(x) v \leq 1\}$. The arrow shows $-\nabla f(x)$, the gradient descent direction. The Newton step $\Delta x_{nt}$ is the steepest descent direction in the norm $\|\cdot\|_{\nabla^2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

# Interpretation III

## Solution of linearized optimality condition

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

$$\nabla f(x+v) \approx \nabla f(x) + \nabla^2 f(x) v \tag{7}$$

Set RHS to zero gives Newton step $\Delta x_{nt}$

$$v = -\nabla^2 f(x)^{-1} \nabla f(x) = \Delta x_{nt} \tag{8}$$

So the Newton step $\Delta x_{nt}$ is what must be added to $x$ so that the linearized optimality condition holds.
Again, this suggests that when $x$ is near $x^\star$, the update $x + \Delta x_{nt}$ should be a very good approximation of $x^\star$.

# Interpretation III

## Solution of linearized optimality condition

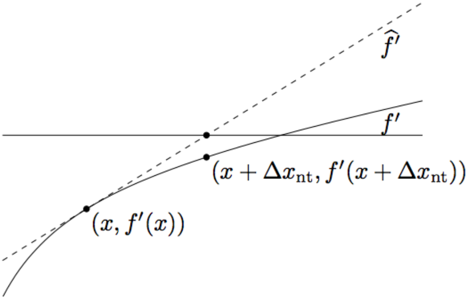


Figure 3 : The solid curve is the derivative $f'$ of the function $f$ shown in Figure 2. $\hat{f}'$ is the linear approximation of $f'$ at $x$. The Newton step $\Delta x_{nt}$ is the difference between the root of $\hat{f}'$ and the point $x$. [Figure and caption from *Convex Optimization* by Boyd and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

# 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 invulnerable 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 of Newton step

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

$$\nabla_y^2 g(y) = \nabla_y^2 f(Ay) = A^T \nabla_x^2 f(x) A \tag{9}$$

$$\nabla_y g(y) = \nabla_y f(Ay) = A^T \nabla_x f(x) \tag{10}$$

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

$$\begin{aligned} y^+ &= y - t(\nabla_y^2 g(y)^{-1} \nabla_y g(y) \\ &= y - t(A^T \nabla_x^2 f(x) A)^{-1} A^T \nabla_x f(x) \\ &= y - t\ A^{-1} \nabla_x^2 f(x)^{-1} \nabla_x f(x) \end{aligned} \tag{11}$$

Multiply both sides by $A$,

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

# Convergence Analysis: Assumption

We assume $f(\cdot)$ satisfies the following:

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

$$mI \preceq \nabla^2 f(x) \preceq MI, \quad \forall x \tag{13}$$

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

$$\|\nabla^2 f(y) - \nabla^2 f(x)\|_2 \leq L\|x - y\|_2, \ \forall x, \ y \tag{14}$$

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

# Convergence Analysis: Theorem

## Theorem (Part I)

*There exist* $\eta$, $\gamma$, *where* $0 < \eta \leq \frac{m^2}{L}$, $\gamma = \frac{\alpha\beta m}{M^2}\eta^2$ *such that Newton method with BTLS has two phases:*

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

$$f(x^+) - f(x) \leq -\gamma \tag{15}$$

Inequality (15) has three implications:

- Every Newton step with BTLS gets closer to global optima by at least $\gamma$.
- Damped phase has at most $\frac{f(x^{(0)}) - f^\star}{\gamma}$ iterations.
- The damped phase is also at least linearly convergent if not better, i.e. $f(x^+) - f^\star \leq c(f(x) - f^\star)$

# 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*

$$\frac{L}{2m^2}\|\nabla f(x^+)\|_2 \leq \left(\frac{L}{2m^2}\|\nabla f(x)\|_2\right)^2 \tag{16}$$

Implication:

- To achieve an accuracy of $\varepsilon$, only $O(\log\log\varepsilon)$ iterations are needed once in the quadratic phase, which is known as quadratic convergence. [detail next slide]

# Implication of Part (b)

If we have $a_1(< 1), a_2, \ldots$ such that $a_{i+1} \leq a_i^2$, then

$$a_\ell \leq (a_{\ell-1})^2 \leq (a_{\ell-2})^{2^2} \leq (a_{\ell-3})^{2^3} \leq \cdots \leq (a_1)^{2^{\ell-1}}$$

for $a_\ell \leq \varepsilon$ need,

$$\begin{aligned} (a_1)^{2^{\ell-1}} &\leq \varepsilon \\ 2^{\ell-1} \log a_1 &\leq \log \varepsilon \\ \ell &\geq \log \log \frac{1}{\varepsilon} + \text{constant} \end{aligned}$$

- In this case to make it more precise, set $a_\ell = \frac{L}{2m^2} \|\nabla f(x^{(k+\ell-1)})\|_2$ and note that $a_1 < \frac{L}{2m^2} \eta < \frac{1}{2}$

# Convergence Analysis: Part (a)

For readability of the proof, we will divide the proof into lemmas. Use the following lemma for part (a).

**Lemma (Global BTLS)**

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

We will first prove the lemma and then the main result, which is

**Theorem (Part (a))**

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

- Notation: For convenience, $g \triangleq \nabla f(x)$ and $H \triangleq \nabla^2 f(x)$

# Convergence Analysis: Part (a)

## Proof of Global BTLS Lemma

$$
\begin{aligned}
f(x^+) &= f(x - tH^{-1}g) \\
&\leq f(x) - tg^T H^{-1} g + \frac{M}{2} t^2 g^T H^{-1} H^{-1} g && (17) \\
&\leq f(x) - tg^T H^{-1} g + \frac{M}{2m} t^2 g^T H^{-1} g && (18) \\
&= f(x) - \frac{m}{2M} g^T H^{-1} g && \text{by setting } t = \frac{m}{M} \\
&\leq f(x) - \alpha \frac{m}{M} g^T H^{-1} g && \text{since } \alpha < \frac{1}{2}
\end{aligned}
$$

Hence, $t = \frac{m}{M}$ satisfies the BTLS exit condition.
Note (17) $\Rightarrow$ (18) follows because,

$$
g^T H^{-1} H^{-1} g = g^T H^{-1/2} H^{-1} H^{-1/2} g \leq \frac{1}{m} g^T H^{-1} g
$$

# Convergence Analysis: Main Proof of Part (a)

$$
\begin{aligned}
t &\leq \beta \frac{m}{M} && \text{(Global BTLS Lemma)} \\
f(x^+) &\leq f(x) - \alpha \left(\beta \frac{m}{M}\right) g^T H^{-1} g && \text{(BTLS condition)} \\
&\leq f(x) - \alpha \left(\beta \frac{m}{M}\right) \left(\frac{1}{M} \|g\|_2^2\right) && (H^{-1} \preceq I/m) \\
&= f(x) - \alpha\beta \frac{m}{M^2} \|g\|_2^2 \\
&\leq f(x) - \underbrace{\alpha\beta \frac{m}{M^2} \eta^2}_{\gamma} && (19) \\
\implies f(x^+) - f(x) &= -\gamma && (20)
\end{aligned}
$$

# Convergence Analysis: Part (b)

We will use the following lemma for this part.

### Lemma (Local BTLS)

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

Proof: See lecture note and text book.

### Theorem (Part (b))

*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: Part (b)

$$
\begin{aligned}
\text{Let } x^+ &= x - H^{-1}g && \text{(BTLS Quad. Lemma)} \\
\|\nabla f(x^+)\|_2 &= \|\nabla f(x - H^{-1}g) - g + HH^{-1}g\|_2 && \text{(Add zero)} \\
&= \left\| \int_0^1 \nabla^2 f(x - tH^{-1}g)(-H^{-1}g) + HH^{-1}g \mathrm{d}t \right\|_2 \\
& && \text{(Fund. Theorem of Calculus)} \\
&= \left\| \int_0^1 (\nabla^2 f(x - tH^{-1}g) - H)(-H^{-1}g) \mathrm{d}t \right\|_2 \\
& && \text{(Rearrange)} \\
&\le \int_0^1 \left\| (\nabla^2 f(x - tH^{-1}g) - H) \right\|_2 \left\| (-H^{-1}g) \right\|_2 \mathrm{d}t \\
& && \text{(Triangle inequality of norms)}
\end{aligned}
$$

# Convergence Analysis: Part (b) Cont.

$$\begin{aligned}
\|\nabla f(x^+)\|_2 &\le \int_0^1 \left\|(\nabla^2 f(x - tH^{-1}g) - H)\right\|_2 \|H^{-1}g\|_2 \mathrm{d}t \\
&\le \int_0^1 L\|-tH^{-1}g\|_2 \|H^{-1}g\|_2 \mathrm{d}t \\
&\qquad\qquad\qquad \text{(Liptschitz Continuity of Hessian)} \\
&= L\|H^{-1}g\|_2^2 \int_0^1 t\mathrm{d}t = \frac{L}{2}\|H^{-1}g\|_2^2 \\
&\le \frac{L}{2m^2}\|g\|_2^2 \qquad \text{(Strong convexity } (H^{-1} \preceq I/m))
\end{aligned}$$

$$\implies \frac{L}{2m^2}\|\nabla f(x^+)\|_2 \le \left(\frac{L}{2m^2}\|g\|_2\right)^2 \tag{21}$$

# Summary

- Mainly cover convergence analysis of Newton method
- Newton method is affine invariant
- There are two phases
  - Damped phase: linear convergence
  - Quadratic phase: quadratic convergence