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 $x \in \mathrm{dom}\, f$, 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.

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

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

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

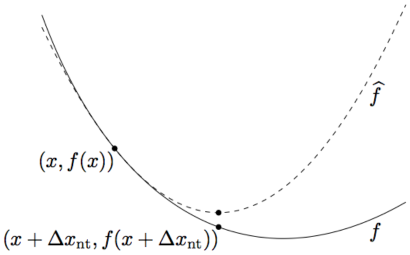

Figure 1 : The function f (shown solid) and its second-order approximation $\hat{f}$ at x (dashed). The Newton step Δx_{nt} is what must be added to x to give the minimizer of $\hat{f}$.

Interpretation II

Steepest descent direction in Hessian norm

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

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

Interpretation II

Steepest descent direction in Hessian norm

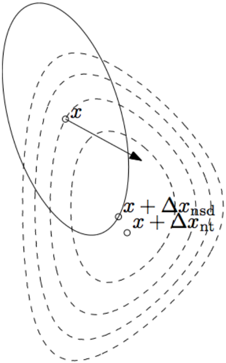

Figure 2 : The dashed lines are level curves of a convex function. The ellipsoid shown (with solid line) is $\{x + v | v^T \nabla^2 f(x) v \leq 1\}$. The arrow shows $-\nabla f(x)$, the gradient descent direction. The Newton step Δx_{nt} is the steepest descent direction in the norm $\| \cdot \|_{\nabla^2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

Interpretation III
Solution of linearized optimality condition

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

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

Set RHS to zero gives newton update.

Interpretation III
Solution of linearized optimality condition

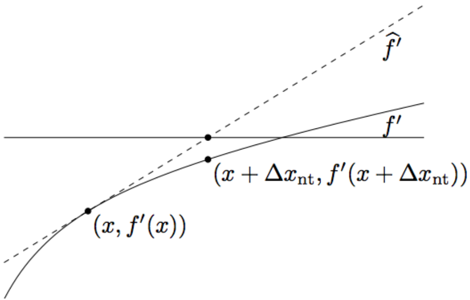

Figure 3 : The solid curve is the derivative f' of the function f shown in Figure 2. $\hat{f'}$ is the linear approximation of f' at x. The Newton step Δx_{nt} is the difference between the root of $\hat{f'}$ and the point x.

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

$$\begin{aligned}
\nabla_y^2 g(y) &= \nabla_y^2 f(Ay) &= A^T \nabla_x^2 f(x) A && (8)\\
\nabla_y g(y) &= \nabla_y f(Ay) &= A^T \nabla_x f(x) && (9)
\end{aligned}$$

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

$$\begin{aligned}
y^+ &= y - t(\nabla_y^2 g(y))^{-1} \nabla_y g(y) && (10)\\
&= y - t(A^T \nabla_x^2 f(x) A)^{-1} A^T \nabla_x f(x) && (11)\\
&= y - t\, A^{-1} \nabla_x^2 f(x)^{-1} \nabla_x f(x) && (12)
\end{aligned}$$

Multiply both sides with affine tranformation A,

$$\begin{aligned}
Ay^+ &= Ay - A \cdot t\, A^{-1} \nabla_x^2 f(x)^{-1} \nabla_x f(x) && (13)\\
&= x - t\, \nabla_x^2 f(x)^{-1} \nabla_x f(x) && (14)\\
&= x^+ && (15)
\end{aligned}$$

Convergence Analysis: Assumption

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

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

$$mI \preceq \nabla^2 f(x) \preceq MI \tag{16}$$

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

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

  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 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 γ.
- 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*

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

Implications:

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

Convergence Analysis: Proof of Part (a)

Lemma (BTLS Damped Lemma)

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

Lemma (Damped Phase 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: BTLS Damped Lemma

$$
\begin{aligned}
f(x^+) &= f(x - tH^{-1}g) \\
&\leq f(x) - tg^T H^{-1} g + \frac{M}{2} t^2 g^T H^{-1} H^{-1} g && (20) \\
&\leq f(x) - tg^T H^{-1} g + \frac{M}{2m} t^2 g^T H^{-1} g && (21) \\
&= f(x) - \frac{m}{2M} g^T H^{-1} g && \text{by setting } t = \frac{m}{M} \\
&\leq f(x) - \alpha \frac{m}{M} g^T H^{-1} g && (\alpha < \frac{1}{2}) \quad (22)
\end{aligned}
$$

Hence, $t = \frac{m}{M}$ satisfies the BTLS exit condition.
Note that derivation from (20) to (21) is given by

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

Convergence Analysis: Damped Phase Lemma

$$
\begin{aligned}
t &\leq \beta \frac{m}{M} && \text{(BTLS Damped Lemma)} \\
f(x^+) &\leq f(x) - \alpha \left(\beta \frac{m}{M}\right) g^T H^{-1} g && \text{(BTLS condition)} \\
&\leq f(x) - \alpha \left(\beta \frac{m}{M}\right) \left(\frac{1}{M} \|g\|_2^2\right) && (H^{-1} \preceq I/m) \\
&= f(x) - \alpha\beta \frac{m}{M^2} \|g\|_2^2 \\
&= f(x) - \underbrace{\alpha\beta \frac{m}{M^2} \eta^2}_{\gamma} && (23) \\
\implies & f(x^+) - f(x) = -\gamma && (24)
\end{aligned}
$$

Convergence Analysis: Proof Part (b)

Lemma (BTLS Quad. Lemma)

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

Proof of BTLS Quad. Lemma will come after the proof of the following Quad. Phase Lemma.

Lemma (Quad. Phase 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: Quad. Phase Lemma

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

Convergence Analysis: Quad. Phase Lemma

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

$$
\implies \frac{L}{2m^2}\|\nabla f(x^+)\|_2 \leq \left(\frac{L}{2m^2}\|g\|_2\right)^2 \tag{25}
$$

Convergence Analysis: BTLS Quad. Lemma

Now we show that $t = 1$ satisfies the exit condition of BTLS under the assumption of (b).
Setting $t = 1$ we have,

$$
\begin{aligned}
f(x + \Delta x_{\text{nt}}) &\leq f(x) - \frac{1}{2}\lambda^2(x) + \frac{L}{6m^{3/2}}\lambda^3(x) && (26) \\
&= f(x) - \lambda^2(x)\left(\frac{1}{2} - \frac{L\lambda(x)}{6m^{3/2}}\right) && (27) \\
&= f(x) + g^T \Delta x_{\text{nt}}\left(\frac{1}{2} - \frac{L\lambda(x)}{6m^{3/2}}\right) && (28)
\end{aligned}
$$

Again using strong convexity, we have

$$\lambda(x) = (g^T H^{-1} g)^{1/2} \leq \frac{1}{m^{1/2}} \|g\|_2 < \frac{1}{m^{1/2}} \eta. \tag{29}$$

where the last inequality follows from the assumption $\|g\|_2 < \eta$. Hence if we choose α such that,

$$\begin{aligned} \alpha &< \frac{1}{2} - \frac{L\lambda(x)}{6m^{3/2}} && (30) \\ &< \frac{1}{2} - \frac{L}{6m^2}\eta && (31) \end{aligned}$$

then $t = 1$ satisfies BTLS exit condition.