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 19, 2014

Newton Step

Definition

For $x \in \text{dom}\, f$, 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 $\nabla^2 f(x)$,

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

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 Tylor approximation $\hat{f}$ of f at x is

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

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

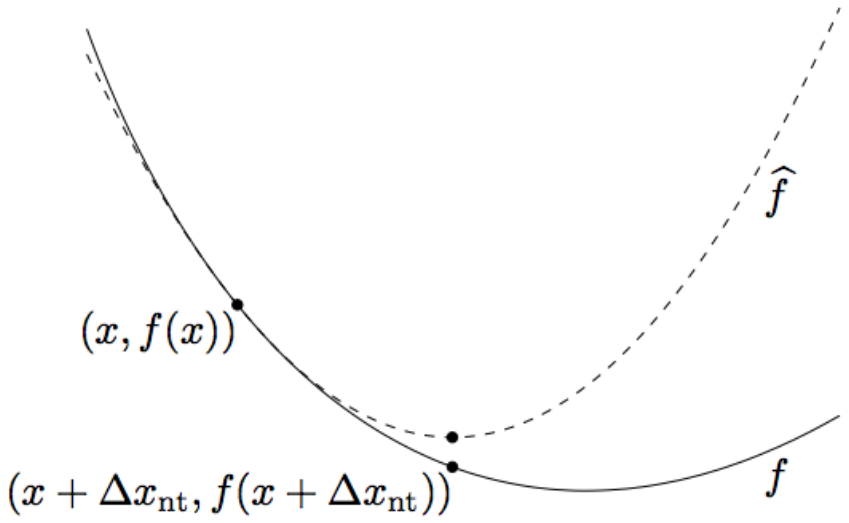

Figure 1 : The function f (shown solid) and its second-order approximation $\hat{f}$ at x (dashed). The Newton step Δx_{nt} is what must be added to x to give the minimizer of $\hat{f}$.

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

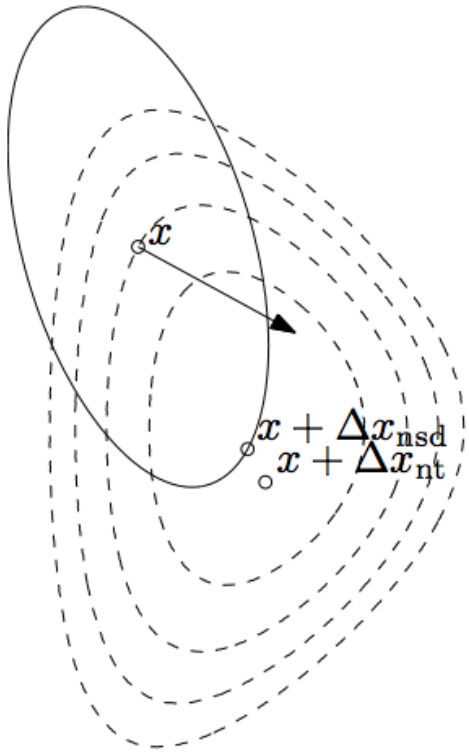

Figure 2 : The dashed lines are level curves of a convex function. The ellipsoid shown (with solid line) is $\{x + v | v^T \nabla^2 f(x) v \leq 1\}$. The arrow shows $-\nabla f(x)$, the gradient descent direction. The Newton step Δx_{nt} is the steepest descent direction in the norm $|| \cdot ||_{\nabla^2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

Interpretation III

Solution of linearized optimality condition

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

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

Set RHS to zero gives newton update.

Interpretation III

Solution of linearized optimality condition

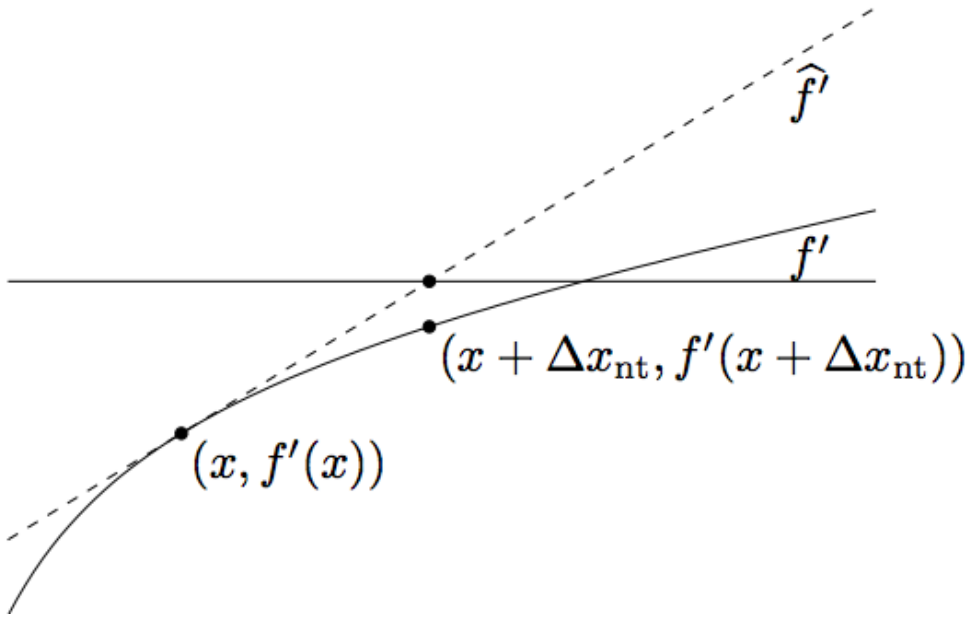

Figure 3 : The solid curve is the derivative f' of the function f shown in Figure 2. $\hat{f}'$ is the linear approximation of f' at x. The Newton step Δx_{nt} is the difference between the root of $\hat{f}'$ and the point x.

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 vulnerable to the selection of coordinate system.

Remark

Gradient Descent Method is not affine invariance. 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

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

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

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

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

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

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

Multiply both sides with affine tranformation A,

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

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

$$= x^+ \tag{15}$$

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*

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

- *$\nabla^2 f(x)$ 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 phrases:*

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

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

Inequality (18) has three 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*

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

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.