**EE 381V: Large Scale Optimization** **Fall 2014**

# Lecture 7 — September 18

*Lecturer: Sanghavi* *Scribe: David Inouye, Jimmy Lin, & Vutha Va*

## 7.1 Recap

In the last lecture we studied coordinate descent and steepest descent method. Toward the end of the lecture we looked at an example illustrating that by change of coordinates ($x = Ay$) we could alter the convergence behavior of gradient descent methods. In this lecture we will study Newton's method, which is affine invariant, i.e. the change of coordinates $x = Ay$ does not affect the convergence property as in gradient descent method.

## 7.2 Introduction to Newton's Method

**Definition 1.** *For $f$ having positive definite Hessian $\nabla^2 f(x) \succ 0$, the Newton updating rule with step size $t$ is defined as,*

$$x^+ = x + t\Delta x_{\mathrm{nt}} = x - t\,\nabla^2 f(x)^{-1}\nabla f(x), \tag{7.1}$$

*where, $\Delta x_{\mathrm{nt}}$ is called the Newton step.*

Note that since $\nabla^2 f(x) \succ 0$ we have,

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

always holds for $\nabla f(x) \neq 0$, so this is a descent direction.

There are various ways to interpret this choice of updating rule.

### Minimizer of Quadratic Approximation

Consider a quadratic approximation of $f$ around $x$,

$$\tilde{f}(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v. \tag{7.2}$$

This quadratic function is minimized at $v^\star = -\nabla^2 f(x)^{-1}\nabla f(x)$. Note that if $f$ is quadratic, this approximation is exact and $x + v^\star$ is the exact minimizer of $f$.

### A Special Case of Steepest Descent

Newton's method can also be viewed as the steepest descent method with the norm

$$\|u\|_{\nabla^2 f(x)} \triangleq \sqrt{u^T \nabla^2 f(x) u}. \tag{7.3}$$

## Linear Approximation of Gradient around $x$

Consider a linear approximation of $\nabla f(x+v)$,

$$\nabla f(x+v) \simeq \nabla f(x) + \nabla^2 f(x) v. \tag{7.4}$$

The Newton updating rule is obtained by setting the right hand side to 0, which is an approximation to the optimality condition $\nabla f(x^\star) = 0$.

## Affine Invariant Property of the Newton's Method

As mentioned earlier, an important feature of the Newton's method is the affine invariant property, i.e. change of coordinates does not alter the convergence behavior as we have seen in the gradient descent methods. We will now state this formally and prove it.

**Lemma 7.1.** *The Newton's method is affine invariant, i.e. if we define $g(y) = f(Ay)$ and*

- *$y^+$ is the Newton update for $g$*
- *$x^+$ is the Newton update for $f$*

*then if $x = Ay$ we have $x^+ = Ay^+$.*

**Proof:** Let $x = Ay$, then

$$\nabla_y^2 g(y) = \nabla_y^2 f(Ay) = A^T \nabla_x^2 f(x) A \tag{7.5}$$

$$\nabla_y g(y) = A^T \nabla_x f(x) \tag{7.6}$$

Substituting these to the Newton updating rule we have,

$$\begin{aligned} y^+ &= y - t \left(A^T \nabla_x^2 f(x) A\right)^{-1} A^T \nabla_x f(x) \\ &= y - t\, A^{-1} \nabla_x^2 f(x)^{-1} A^{-T} A^T \nabla_x f(x) \\ &= y - t\, A^{-1} \nabla_x^2 f(x)^{-1} \nabla_x f(x) \end{aligned}$$

Multiply both sides by $A$,

$$\begin{aligned} Ay^+ &= x - t\, \nabla_x^2 f(x)^{-1} \nabla_x f(x) \\ &= x^+. \end{aligned}$$

□

## 7.3 Convergence of Newton's Method

We make two major assumptions in this analysis:

1. $f$ is strongly convex, such that $mI \preceq \nabla^2 f(x) \preceq MI$.
2. $\nabla^2 f(x)$ is Lipschitz continuous with constant $L > 0$, i.e.

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \le L\|x - y\|_2 \quad \forall x, y. \tag{7.7}$$

   Note that the norm on the left is the spectral norm, defined as the largest singular value. $L$ can be interpreted as a bound on the third derivative of $f$. The smaller $L$ is, the better $f$ can be approximated by a quadratic function. Since each step of Newton's method minimizes a quadratic approximation of $f$, the performance of Newton's method will be best for functions with small $L$.

For notational convenience we denote $g = \nabla f(x)$ and $H = \nabla^2 f(x)$ from this point on.

Our main result of this lecture is Theorem 7.2. We will devote the rest of the lecture trying to understand and prove this theorem. Before stating the theorem, let us first recall Backtracking Line Search (BTLS). With BTLS, first $\alpha$ and $\beta$ are chosen such that $0 < \alpha < \frac{1}{2}$ and $0 < \beta < 1$, starting with $t = 1$, repeat

```
while true
    if  $f(x + t\Delta x) \le f(x) + \alpha t g^T \Delta x$
        exit
    else
        $t \leftarrow \beta t$
    end
end
```

We are now ready to state and prove the theorem.

**Theorem 7.2.** *There exist $\eta$ and $\gamma$ with $0 < \eta \le \frac{m^2}{L}$ and $\gamma = \alpha\beta\eta^2\frac{m}{M^2}$ such that Newton's method with BTLS satisfies,*

*(a). Damped Newton Phase: If $\|g\|_2 \ge \eta$ then $f(x^+) - f(x) \le -\gamma$.*

*(b). Quadratic Phase: If $\|g\|_2 < \eta$ then BTLS $t = 1$ and*

$$\frac{L}{2m^2}\|\nabla f(x^+)\|_2 \le \left(\frac{L}{2m^2}\|\nabla f(x)\|_2\right)^2. \tag{7.8}$$

Before proceeding to the proof, let us first interpret the implications of this theorem.

## Implication of (a)

In the damped Newton phase, $f$ decreases by at least $\gamma$ at each iteration, there the total of iterations in this phase cannot exceed,

$$\frac{f(x^{(0)}) - p^\star}{\gamma}$$

since otherwise $f(x)$ would be less than $p^\star$, which contradicts the optimality of $p^\star$. In other words, the quadratic phase will start after $\frac{f(x^{(0)})-p^\star}{\gamma}$ iterations.

## Implication of (b)

Let $k$ be the first iteration in which $\|g\| < \eta$. And let $\ell \geq 0$ be the number of iterations after $k$. For simplicity, let us define:

$$a_\ell = \frac{L}{2m^2}\|\nabla f(x^{(k+\ell-1)})\|_2 \tag{7.9}$$

First, let us establish a bound on $a_1$. In the quadratic phase, since $\|\nabla f(x^{(k)})\|_2 < \eta$ and $\eta < \frac{m^2}{L}$ by assumption, we have

$$\frac{L}{2m^2}\|\nabla f(x^{(k)})\|_2 < \frac{L}{2m^2}\eta < \frac{1}{2}$$

Thus, $a_1 < \frac{1}{2}$. Now from (b) of the theorem, we also have that $a_{\ell+1} \leq a_\ell^2$. Therefore, we have the following sequence:

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

$$\implies a_\ell \leq \left(\frac{1}{2}\right)^{2^{\ell-1}}$$

$$\implies \frac{L}{2m^2}\|\nabla f(x^{(\ell)})\|_2 \leq \left(\frac{1}{2}\right)^{2^{\ell-1}}$$

$$\implies \|\nabla f(x^{(\ell)})\|_2 \leq \frac{2m^2}{L}\left(\frac{1}{2}\right)^{2^{\ell-1}}$$

For strongly convex functions, we also have

$$f(x^{(\ell)}) - p^\star \leq \frac{1}{2m}\|\nabla f(x^{(\ell)})\|_2^2 \tag{7.10}$$

$$\leq \frac{1}{2m}\left(\frac{2m^2}{L}\left(\frac{1}{2}\right)^{2^{\ell-1}}\right)^2 \tag{7.11}$$

$$\leq \frac{2m^3}{L^2}\left(\frac{1}{2}\right)^{2^{\ell-1}} \tag{7.12}$$

thus, $f(x) \to p^\star$ quadratically.

Therefore, if we want $a_\ell \leq \epsilon$, we only need the following number of iterations:

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

### 7.3.1 Convergence Proof

For readability of the proof, we will divide the proof into lemmas to emphasize the flow of the proof and not get lost in the details of the derivation.

**Lemma 7.3.** *With the assumptions in part (a), $t = \frac{m}{M}$ satisfies BTLS exit condition, i.e. $f(x + t\Delta x_{\text{nt}}) \leq f(x) + \alpha t g^T \Delta x_{\text{nt}}$.*

**Lemma 7.4.** *With the assumptions in part (b), $t = 1$ satisfies BTLS exit condition.*

Proofs for both Lemmas in Section 7.3.2.

**Proof (Theorem 7.2 part (a)):** Using Lemma 7.3, with $t \geq \beta\frac{m}{M}$ and substituting $\Delta x_{\text{nt}} = -H^{-1}g$, we have

$$f(x^+) \leq f(x) - \alpha\beta\frac{m}{M}g^T H^{-1}g \tag{7.13}$$

By strong convexity $H \preceq MI$, so $H^{-1} \succeq \frac{1}{M}I$, and we have

$$g^T H^{-1} g \geq \frac{1}{M}\|g\|_2^2, \tag{7.14}$$

therefore,

$$f(x^+) \leq f(x) - \alpha\beta\frac{m}{M}\frac{1}{M}\|g\|_2^2 \tag{7.15}$$

$$f(x^+) - f(x) \leq -\underbrace{\alpha\beta\frac{m}{M^2}\eta^2}_{\gamma} \tag{7.16}$$

where the last inequality follows because $\|g\|_2 \geq \eta$. □

**Proof (Theorem 7.2 part (b)):** Using Lemma 7.4, we can set $t = 1$. Thus, we have $x^+ = x - H^{-1}g$ and

$$\nabla f(x^+) = \nabla f(x - H^{-1}g) - g + HH^{-1}g \tag{7.17}$$

Applying the fundamental theorem of calculus to $\nabla f(x - H^{-1}g) - g$ we have[1]

$$\nabla f(x - H^{-1}g) - g = \int_0^1 \nabla^2 f(x - tH^{-1}g)(-H^{-1}g)\mathrm{d}t \tag{7.18}$$

thus,

$$\nabla f(x^+) = \int_0^1 \left(\nabla^2 f(x - tH^{-1}g) - H\right)(-H^{-1}g)\mathrm{d}t \tag{7.19}$$

Taking the norm of both sides,

$$\begin{aligned}\|\nabla f(x^+)\|_2 &= \left\|\int_0^1 \left(\nabla^2 f(x - tH^{-1}g) - H\right)(-H^{-1}g)\mathrm{d}t\right\|_2 && (7.20)\\ &\le \int_0^1 \|\nabla^2 f(x - tH^{-1}g) - H\|_2 \|H^{-1}g\|_2 \mathrm{d}t && (7.21)\end{aligned}$$

By Lipschitz continuity of $\nabla^2 f$ we have

$$\|\nabla^2 f(x - tH^{-1}g) - H\|_2 \le L\|tH^{-1}g\|_2 = Lt\|H^{-1}g\|_2, \tag{7.22}$$

thus,

$$\|\nabla f(x^+)\|_2 \le \int_0^1 Lt\|H^{-1}g\|_2^2 \mathrm{d}t = \frac{L}{2}\|H^{-1}g\|_2^2 \tag{7.23}$$

Also, $H \succeq mI$ so $H^{-1} \preceq \frac{1}{m}I$ and,

$$\|H^{-1}g\|_2^2 \le \frac{1}{m^2}\|g\|_2^2. \tag{7.24}$$

Substituting this,

$$\|\nabla f(x^+)\|_2 \le \frac{L}{2m^2}\|g\|_2^2, \tag{7.25}$$

and multiplying both sides by $\frac{L}{2m^2}$ we obtain the result stated in the theorem. □

### 7.3.2 Proof of Lemmas

**Proof (Lemma 7.3):**

$$\begin{aligned}f(x^+) &= f(x - tH^{-1}g) && (7.26)\\ &\le f(x) - tg^T H^{-1}g + \frac{M}{2}t^2 g^T H^{-1}H^{-1}g && (7.27)\end{aligned}$$

[1] This point can be easily seen for scalar case. Let $a(t) = \nabla f(x - tH^{-1}g)$, then $a(0) = \nabla f(x)$, $a(1) = \nabla f(x - H^{-1}g)$, and $\int_0^1 \frac{d}{dt}a(t)\mathrm{d}t = a(1) - a(0)$.

Note that[2],

$$g^T H^{-1} H^{-1} g = g^T H^{-1/2} H^{-1} H^{-1/2} g \leq \frac{1}{m} g^T H^{-1} g \tag{7.28}$$

Thus,

$$f(x^+) \leq f(x) - t g^T H^{-1} g + \frac{M}{2m} t^2 g^T H^{-1} g \tag{7.29}$$

Setting $t = \frac{m}{M}$,

$$f(x^+) \leq f(x) - \frac{m}{2M} g^T H^{-1} g \tag{7.30}$$

This satisfies the exit condition for $t = \frac{m}{M}$, $f(x^+) \leq f(x) - \alpha \frac{m}{M} g^T H^{-1} g$ since $\alpha < 1/2$. □

**Proof (Lemma 7.4):** In this proof we will find $\alpha < \frac{1}{2}$ such that $t = 1$ satisfies BTLS exit condition. Our goal is to find $\alpha$ such that,

$$f(x + \Delta x_{\text{nt}}) \leq f(x) + \alpha g^T \Delta x_{\text{nt}}. \tag{7.31}$$

For notational convenience we denote

$$\lambda(x) = (\Delta x_{\text{nt}}^T H \Delta x_{\text{nt}})^{1/2} = (g^T H^{-1} g)^{1/2} \tag{7.32}$$

which is known as the Newton decrement at $x$. The second equality follows because $\Delta x_{\text{nt}} = -H^{-1} g$.

By Lipschitz condition for $t \geq 0$,

$$\|\nabla^2 f(x + t\Delta x_{\text{nt}}) - H\|_2 \leq tL \|\Delta x_{\text{nt}}\|_2, \tag{7.33}$$

and we have,

$$|\Delta x_{\text{nt}}^T (\nabla^2 f(x + t\Delta x_{\text{nt}}) - H) \Delta x_{\text{nt}}| \leq tL \|\Delta x_{\text{nt}}\|_2^3, \tag{7.34}$$

Now define $\tilde{f}(t) = f(x + t\Delta x_{\text{nt}})$, then the second derivative with respect to $t$ becomes $\tilde{f}''(t) = \Delta x_{\text{nt}}^T \nabla^2 f(x + t\Delta x_{\text{nt}}) \Delta x_{\text{nt}}$. Substituting $\tilde{f}''$ into the above inequality we have,

$$|\tilde{f}''(t) - \tilde{f}''(0)| \leq tL \|\Delta x_{\text{nt}}\|_2^3, \tag{7.35}$$

We will use this inequality to find an upper bound on $\tilde{f}(t)$. Starting with [3]

$$\tilde{f}''(t) \leq \tilde{f}''(0) + tL \|\Delta x_{\text{nt}}\|_2^3 \tag{7.36}$$

$$\leq \lambda^2(x) + t \frac{L}{m^{3/2}} \lambda^3(x) \tag{7.37}$$

[2] Recall the definition of square root of a matrix. If A is positive definite then we can write $A = U \Lambda U^T$, where $U$ is unitary and $\Lambda$ is diagonal, and $A^{1/2} = U \Lambda^{1/2} U^T$.

[3] Note that if $a, b, c > 0$ and $|a - b| \leq c$, then $a \leq b + c$. Consider if $a \geq b$, then we can remove the absolute value sign. Suppose $a \leq b$, then we have $a \leq b \leq b + c$ since $c$ is positive.

since, $\tilde{f}''(0) = \Delta x_{\rm nt}^T H \Delta x_{\rm nt} = \lambda^2(x)$ and from strong convexity, $H \succeq mI$,

$$\begin{aligned} \lambda^2(x) &= \Delta x_{\rm nt}^T H \Delta x_{\rm nt} \geq m\|\Delta x_{\rm nt}\|_2^2 \\ \Rightarrow \|\Delta x_{\rm nt}\|_2 &\leq m^{-1/2}\lambda(x). \end{aligned}$$

Now integrate both sides of (7.37) with respect to $t$

$$\begin{aligned} \tilde{f}'(t) &\leq \tilde{f}'(0) + t\lambda^2(x) + t^2\frac{L}{2m^{3/2}}\lambda^3(x) \\ &= -\lambda^2(x) + t\lambda^2(x) + t^2\frac{L}{2m^{3/2}}\lambda^3(x) \end{aligned}$$

since $\tilde{f}'(0) = \Delta x_{\rm nt}^T g = -g^T H^{-1} g = -\lambda^2(x)$. Integrating once more,

$$\tilde{f}(t) \leq \tilde{f}(0) - t\lambda^2(x) + \frac{t^2}{2}\lambda^2(x) + t^3\frac{L}{6m^{3/2}}\lambda^3(x)$$

Setting $t = 1$ we have,

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

Again using strong convexity, we have

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

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

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

then $t = 1$ satisfies BTLS exit condition.

□