EE 381V: Large Scale Optimization **Fall 2012**

Lecture 7 — September 20

Lecturer: Caramanis & Sanghavi *Scribe: Ye Wang & Ryan Buckley*

7.1 Introduction and Newton Decrement

In the previous lecture we covered the steepest descent method and Newton's method, and we noted that Newton's method is affine invariant. In this lecture, we discuss the Newton decrement and the convergence of Newton's method.

Recall that the typical stopping criterion for the steepest descent method, $\|\nabla f(x)\|_2 \leq \epsilon$, is not affine invariant. Since the Newton step is affine invariant, we would like the Newton stopping condition to be affine invariant as well. To that end, we introduce the Newton decrement.

Definition 1. *The Newton decrement at x, denoted by $\lambda(x)$, is defined as*

$$\lambda(x) = \|\nabla f(x)\|_{\nabla^2 f(x)^{-1}} = \left(\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)\right)^{1/2}. \tag{7.1}$$

Note that we can express the Newton decrement in terms of the Newton step. Recall that Δx_{nt} is defined to be

$$\Delta x_{\text{nt}} = -\nabla^2 f(x)^{-1} \nabla f(x). \tag{7.2}$$

Therefore,

$$\lambda^2(x) = -\nabla f(x)^T \Delta x_{\text{nt}}. \tag{7.3}$$

We can apply the definition of the Newton step once more to see that

$$\lambda^2(x) = \Delta x_{\text{nt}}^T \nabla^2 f(x) \Delta x_{\text{nt}}. \tag{7.4}$$

A typical stopping condition for Newton's method is

$$\frac{\lambda^2(x)}{2} \leq \epsilon. \tag{7.5}$$

We now prove the affine invariance of this stopping criterion.

Proposition 1. $\lambda^2(x)$ *is affine invariant.*

Proof: Consider an affine change of coordinates with a nonsingular $T \in \mathbb{R}^{n\times n}$ and some $z \in \mathbb{R}^n$. Define $\hat{f}(y) = f(Ty + z)$. If $x = Ty + z$, then

$$\nabla \hat{f}(y) = T^T \nabla f(x), \quad \nabla^2 \hat{f}(y) = T^T \nabla^2 f(x) T \text{ , and } \nabla^2 \hat{f}(y)^{-1} = T^{-1} \nabla^2 f(x)^{-1} T^{-T}.$$

Hence,

$$\begin{aligned}\lambda^2(y) &= \nabla\hat{f}(y)^T\nabla^2\hat{f}(y)^{-1}\nabla\hat{f}(y)\\ &= \nabla f(x)^T T T^{-1}\nabla^2 f(x)^{-1}T^{-T}T^T\nabla f(x)\\ &= \nabla f(x)^T\nabla^2 f(x)^{-1}\nabla f(x)\\ &= \lambda^2(x).\end{aligned}$$

So the value of λ^2 is unaffected by an affine change of variables. □

7.2 Outline of Convergence Analysis

We make two major assumptions in this section:

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 $L > 0$:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq L\|x-y\|_2 \quad \forall x, y. \tag{7.6}$$

Note that the norm on the left is the spectral norm, defined in Boyd §A.1.5. 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. 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.

The iterations of Newton's method fall into two phases:

- Phase I: the "global" or "damped" phase
- Phase II: the "local" or "quadratic" phase

In the convergence proof, we show that there exist η and γ with $0 < \eta \leq m^2/L$ and $\gamma > 0$ such that:

- If $\|\nabla f(x^{(k)}\|_2 \geq \eta$ then we say we are in the damped convergence phase, and

$$f(x^{(k)}) - f(x^{(k+1)}) \geq \gamma, \tag{7.7}$$

i.e. each step brings us closer to the optimal value by at least γ.

- If $\|\nabla f(x^{(k)})\|_2 < \eta$ then we say we are in the quadratic convergence phase. The step size in backtracking line search will be $t = 1$, and

$$\frac{L}{2m^2}\|\nabla f(x^{(k+1)})\|_2 \leq \left(\frac{L}{2m^2}\|\nabla f(x^{(k)})\|_2\right)^2. \tag{7.8}$$

By applying this inequality over l steps and using $\|\nabla f(x)\|_2 < \eta \leq m^2/L$, we have

$$\frac{L}{2m^2}\|\nabla f(x^{(k+l)})\|_2 \leq \left(\left(\frac{L}{2m^2}\|\nabla f(x^{(k)})\|_2\right)^2\right)^l \leq \left(\frac{1}{2}\right)^{2l}. \tag{7.9}$$

So $\|\nabla f(x^{(k+l)})\|_2$ converges to 0 very quickly. Since f is strongly convex, we can write

$$\begin{aligned} f(y) &\geq f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|_2^2 && (7.10)\\ &\geq f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2, && (7.11)\end{aligned}$$

for all y. Finally,

$$f(x^{(k+l)}) - f^* \leq \frac{1}{2m}\|\nabla f(x^{(k+l)})\|_2^2 \leq \frac{2m^3}{L^2}\left(\frac{1}{2}\right)^{2(l+1)}. \tag{7.12}$$

7.3 Proof for damped convergence phase (7.7)

Proposition 2. *There exist constants η, γ such that if $\|\nabla f(x)\|_2 \geq \eta$ then $f(x) - f(x^+) \geq \gamma$.*

Proof: Suppose $\|\nabla f(x)\|_2 \geq \eta$. By strong convexity,

$$f(x^+) = f(x + t\Delta x_{\text{nt}}) \leq f(x) + t\nabla f(x)^T\Delta x_{\text{nt}} + \frac{M\|\Delta x_{\text{nt}}\|_2^2}{2}t^2.$$

Applying strong convexity to (7.4) gives

$$\lambda^2(x) = \Delta x_{\text{nt}}^T \nabla^2 f(x)\Delta x_{\text{nt}} \geq m\|\Delta x_{\text{nt}}\|_2^2.$$

By applying this and (7.3), we have

$$f(x + t\Delta x_{\text{nt}}) \leq f(x) - t\lambda^2(x) + \frac{M}{2m}t^2\lambda^2(x).$$

If we set the step size to $\hat{t} = \frac{m}{M}$, we see that

$$f(x + \hat{t}\Delta x_{\text{nt}}) \leq f(x) - \frac{2m}{M}\lambda^2(x) \leq f(x) - \alpha\hat{t}\lambda^2(x) = f(x) + \alpha\hat{t}\nabla f(x)^T\Delta x_{\text{nt}},$$

showing that $\hat{t}$ satisfies the backtracking line search exit condition. At worst, therefore, we exit backtracking line search with $t = \beta\frac{m}{M}$. Then

$$\begin{aligned} f(x^+) - f(x) &\leq -\alpha t\lambda^2(x)\\ &\leq -\alpha\beta\frac{m}{M}\lambda^2(x)\\ &\leq -\alpha\beta\frac{m}{M}\frac{\|\nabla f(x)\|_2^2}{M},\end{aligned}$$

where the last inequality holds because

$$\lambda^2(x) = \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \geq \frac{\|\nabla f(x)\|_2^2}{M}.$$

Recall our supposition of phase I convergence:

$$\|\nabla f(x)\|_2 \geq \eta.$$

Thus

$$f(x^+) - f(x) \leq -\alpha\beta \frac{m}{M^2}\eta^2 = -\gamma,$$

or finally, $f(x) - f(x^+) \geq \gamma$. □

7.4 Proof for Quadratic Convergence Phase (7.8)

Proposition 3. *If $\|\nabla f(x)\|_2 < \eta$ then backtracing line search selects unit step sizes ($t = 1$) and $\frac{L}{2m^2}\|\nabla f(x^+)\|_2 \leq \left(\frac{L}{2m^2}\|\nabla f(x)\|_2\right)^2$.*

Proof: We first assume that backtracking line search will accept unit steps and prove the inequality, then we prove that unit step sizes hold.

Suppose $\|\nabla f(x)\|_2 < \eta$ and $t = 1$. Recall from (7.2) that $\nabla f(x) = -\nabla^2 f(x)\Delta x_{\text{nt}}$, so

$$\|\nabla f(x^+)\|_2 = \|\nabla f(x^+) - \nabla f(x) - \nabla^2 f(x)\Delta x_{\text{nt}}\|_2.$$

By applying the fundamental theorem of calculus, we have

$$\begin{aligned} &\leq \left\| \int_0^1 (\nabla^2 f(x + t(x^+ - x)) - \nabla^2 f(x))\Delta x_{\text{nt}}\, dt \right\|_2 \\ &\leq \int_0^1 \|\nabla^2 f(x + t\Delta x_{\text{nt}}) - \nabla^2 f(x)\|_2 \; \|\Delta x_{\text{nt}}\|_2 \, dt. \end{aligned}$$

After applying the Lipschitz condition, we obtain

$$\begin{aligned} &\leq \int_0^1 L\|t\Delta x_{\text{nt}}\|_2 \; \|\Delta x_{\text{nt}}\|_2 \, dt \\ &= \frac{L}{2}\|\Delta x_{\text{nt}}\|_2^2 \\ &= \frac{L}{2}\|\nabla^2 f(x)^{-1}\nabla f(x)\|_2^2 \\ &\leq \frac{L}{2m^2}\|\nabla f(x)\|_2^2. \end{aligned}$$

This completes the proof of the inequality. Now we prove that in the quadratic phase of convergence, backtracking line search will select the step size $t = 1$. Recall that for $t = 1$ to satisfy the backtracking line search exit condition, we need that

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

Let us define a function $\tilde{f}(t)$:

$$\begin{aligned}\tilde{f}(t) &= f(x + t\Delta x_{\text{nt}}) \\ \tilde{f}''(t) &= \Delta x_{\text{nt}}^T \nabla^2 f(x + t\Delta x_{\text{nt}}) \Delta x_{\text{nt}}.\end{aligned}$$

By the Lipschitz condition,

$$|\tilde{f}''(t) - \tilde{f}''(0)| \leq Lt\|\Delta x_{\text{nt}}\|_2 \|\Delta x_{\text{nt}}\|_2^2 = Lt\|\Delta x_{\text{nt}}\|_2^3.$$

Since each factor on the right is nonnegative, we can drop the absolute value and move the $\tilde{f}''(0)$ term:

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

By applying strong convexity to (7.4), we see that $\lambda^2(x) \geq m\|\Delta x_{\text{nt}}\|_2^2$. We raise this identity to the 3/2 power and substitute to obtain

$$\tilde{f}''(t) \leq \tilde{f}''(0) + Lt\frac{\lambda^3(x)}{m^{3/2}}.$$

Now note that $\tilde{f}''(0) = \lambda^2(x)$ and substitute:

$$\tilde{f}''(t) \leq \lambda^2(x) + Lt\frac{\lambda^3(x)}{m^{3/2}}.$$

Integrate both sides with respect to t:

$$\tilde{f}'(t) \leq \tilde{f}'(0) + t\lambda^2(x) + Lt^2\frac{\lambda^3(x)}{2m^{3/2}}.$$

Now substitute $\tilde{f}'(0) = \nabla f(x)^T \Delta x_{\text{nt}} = -\lambda^2(x)$ and integrate again:

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

Recall our supposition that $\|\nabla f(x)\|_2 < \eta$ and that we have specified η to be such that $\eta \leq m^2/L$. Remember also that for backtracking line search, the $0 < \alpha < 0.5$. Therefore

$$\|\nabla f(x)\|_2 < \eta \leq m^2/L \leq 3(1 - 2\alpha)m^2/L.$$

Now suppose that $t = 1$ to obtain

$$\begin{aligned} \tilde{f}(1) = f(x + \Delta x_{\mathrm{nt}}) &\leq f(x) - \frac{1}{2}\lambda^2(x) + \frac{L}{6m^{3/2}}\lambda^3(x) && (7.13) \\ &= f(x) - \lambda^2(x)\left(\frac{1}{2} - \frac{L\lambda(x)}{6m^{3/2}}\right). && (7.14) \end{aligned}$$

Since $mI \preceq \nabla^2 f(x) \preceq MI$, the inverse of the Hessian must also be strongly convex, with $I/M \preceq \nabla^2 f(x)^{-1} \preceq I/m$. This combined with (7.1) gives

$$\lambda(x) = \left(\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)\right)^{1/2} \leq \frac{\|\nabla f(x)\|_2}{m^{1/2}}.$$

Hence,

$$\lambda(x) < \frac{3(1-2\alpha)m^2/L}{m^{1/2}} = 3(1-2\alpha)m^{3/2}/L,$$

so

$$\alpha < \frac{1}{2} - \frac{L\lambda(x)}{6m^{3/2}}.$$

Substituting this expression for α into (7.14) gives

$$\begin{aligned} f(x + \Delta x_{\mathrm{nt}}) &< f(x) - \lambda^2(x)\alpha \\ &= f(x) + \alpha \nabla f(x)^T \Delta x_{\mathrm{nt}}. \end{aligned}$$

This proves that a step size $t = 1$ satisfies the exit condition for backtracking line search. □