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

Lecture 6 — September 18

Lecturer: Caramanis & Sanghavi *Scribe: Yuhuan Du, Zheng Lu*

6.1 Topics Covered

- Convergence Analysis for Steepest Descent
- Newton's Method

In the last lecture, we talked about coordinate descent method and steepest descent method. We also started the discussion of the convergence analysis for steepest descent and we will finish this part in this lecture. After some comments on the steepest descent convergence theorem, we will introduce a new method: Newton's Method.

6.2 Steepest Descent

Given a norm $||\cdot||$, the normalized steepest descent direction is defined as:

$$\Delta x_{\text{nsd}} = \arg\min_v \{\langle \nabla f(x), v\rangle, \text{ s.t. } ||v|| = 1\}.$$

It is also convenient to consider a steepest descent step Δx_{sd} that is *unnormalized*:

$$\Delta x_{\text{sd}} = \Delta x_{\text{nsd}} ||\nabla f(x)||_*,$$

where $||z||_* = \sup_v \{\langle z, v\rangle, \text{ s.t. } ||v|| = 1\}$.

Then the algorithm can be defined as:

$$x_{t+1} = x_t + \eta \Delta x_{\text{sd}},$$

where η represents the step size.

6.2.1 Convergence for BTLS (Backtracking Line Search)

Fact: For any norm $||\cdot||$and its dual $||\cdot||_*$, there exists finite, positive constants γ and $\tilde{\gamma}$, such that for all x

$$||x|| \geq \gamma ||x||_2, \quad ||x||_* \geq \tilde{\gamma} ||x||_2$$

(see A.1.4 in textbook).

Theorem 6.1. *Suppose f is a strongly convex function with $mI \preceq \nabla^2 f(x) \preceq MI$. Then by using steepest descent method with BTLS, we can get*

$$f(x^{(k)}) - f^* \leq c^k (f(x^{(0)}) - f^*)$$

where $c = 1 - 2m\alpha\tilde{\gamma}^2 \min\{1, \frac{\beta\gamma^2}{M}\}$.

Proof: First we want to show $\eta = \frac{\gamma^2}{M}$ always satisfies the BTLS exit condition. To show this, we need to show

$$f(x + \frac{\gamma^2}{M}\Delta x_{\text{sd}}) \leq f(x) - \frac{1}{2}\frac{\gamma^2}{M}||\nabla f(x)||_*^2.$$

By the property of strong convexity, we have

$$\begin{aligned} f(x_+) &= f(x + \eta\Delta x_{\text{sd}}) \\ &\leq f(x) + \eta\langle\nabla f(x), \Delta x_{\text{sd}}\rangle + \frac{M}{2}||\eta\Delta x_{\text{sd}}||_2^2 \\ &= f(x) - \eta||\nabla f(x)||_*^2 + \frac{M}{2}\eta^2||\nabla f(x)||_*^2||\Delta x_{\text{nsd}}||_2^2. \end{aligned}$$

Since $||\Delta x_{\text{nsd}}||_2^2 \leq \frac{1}{\gamma^2}||\Delta x_{\text{nsd}}||^2 = \frac{1}{\gamma^2}$, we have

$$f(x_+) \leq f(x) - \eta||\nabla f(x)||_*^2 + \frac{M\eta^2}{2\gamma^2}||\nabla f(x)||_*^2.$$

Letting $\eta = \frac{\gamma^2}{M}$, we get

$$f(x + \frac{\gamma^2}{M}\Delta x_{\text{sd}}) \leq f(x) - \frac{1}{2}\frac{\gamma^2}{M}||\nabla f(x)||_*^2.$$

Knowing $\eta = \frac{\gamma^2}{M}$ always satisfies the exit condition, we can say

$$\eta \geq \min\left\{1, \frac{\beta\gamma^2}{M}\right\}.$$

By the exit condition of BTLS, we can get

$$\begin{aligned} f(x_+) &\leq f(x) - \alpha\eta||\nabla f(x)||_*^2 \\ &= f(x) - \alpha\min\left\{1, \frac{\beta\gamma^2}{M}\right\}||\nabla f(x)||_*^2 \\ &\leq f(x) - \alpha\min\left\{1, \frac{\beta\gamma^2}{M}\right\}\tilde{\gamma}^2||\nabla f(x)||_2^2 \\ &\leq f(x) - \alpha\min\left\{1, \frac{\beta\gamma^2}{M}\right\}\tilde{\gamma}^2 2m(f(x) - f^*). \end{aligned}$$

Equivalently,

$$f(x_+) - f^* \leq f(x) - f^* - 2m\alpha\tilde{\gamma}^2 \min\left\{1, \frac{\beta\gamma^2}{M}\right\}(f(x) - f^*).$$

So

$$c = 1 - 2m\alpha\tilde{\gamma}^2 \min\left\{1, \frac{\beta\gamma^2}{M}\right\}.$$

□

Comments on the Theorem
The good thing is obvious: this theorem proves linear convergence for any norm $||\cdot||$ and sufficiently well-conditioned strongly convex function f.

However, does this theorem give a better rate for poorly-conditioned functions and some good norm? The answer is NO.

Let us see some examples.

Example 1. Steepest Descent for $||\cdot||_1$
Suppose we use $||\cdot||_1$ to find the steepest descent direction. We know the dual for $||\cdot||_1$ is $||\cdot||_\infty$, and it is easy to show that for $x \in \mathbb{R}^n$

$$||x||_1 \geq ||x||_2, \quad \text{(We get equality when only one } x_i \text{ is non-zero.)}$$
$$||x||_\infty \geq \frac{1}{\sqrt{n}}||x||_2. \quad \text{(We get equality when } x_1 = x_2 = ... = x_n. \text{)}$$

Therefore, $\gamma = 1, \tilde{\gamma} = \frac{1}{\sqrt{n}}$. Compared with the convergence rate in gradient descent with BTLS, where $c = 1 - 2m\alpha \min\left\{1, \frac{\beta}{M}\right\}$, the introduction of γ and $\tilde{\gamma}$ makes the convergence rate c for steepest descent even bigger, which means a slower convergence rate.

Thus, according to the theorem, we cannot get a better rate for any functions in this case.

Example 2. Change of Coordinates
For some poorly-conditioned functions, for example $f(x_1, x_2) = x_1^2 + 10x_2^2$, we hope to convert it into a well-conditioned problem by changing coordinates and then using the gradient descent method to solve it.

Specifically, let $x = Ay$ and $g(y) = f(Ay)$. Then $\nabla g(y) = A^T \nabla f(Ay)$, $\nabla^2 g(y) = A^T \nabla^2 f(Ay) A$.

We want to make sure $g(y)$ is a well-conditioned function so that gradient descent works well for $g(y)$.

We know $mI \preceq \nabla^2 f(x) \preceq MI$, so if $\nabla^2 f(x) = P \succ 0$ is a constant matrix, which means $f(x)$ is a multi-variable quadratic function like the function mentioned above, by letting $A = P^{-\frac{1}{2}}$ we can get $\nabla^2 g(y) = I$. So if we use gradient descent for $g(y)$, we can get the best descent direction and best convergence rate. Specifically,

$$\begin{aligned} y_+ &= y - \eta \nabla g(y) \\ &= y - \eta A^T \nabla f(Ay), \\ Ay_+ &= Ay - \eta A A^T \nabla f(Ay) \\ x_+ &= x - \eta A A^T \nabla f(x). \end{aligned}$$

This is the same as using steepest descent method for $f(x)$ with the definition of norm shown below:

$$||x||_Q = (x^T Q x)^{\frac{1}{2}},$$

where $Q = (AA^T)^{-1} = P = \nabla^2 f(x)$.

This means using steepest descent method with this norm definition can get the best convergence rate.

However, if we analyse the convergence rate for this method using the theorem above, we have

$$\begin{aligned} mI \preceq & Q \preceq MI, \\ x^T Q x &\geq m||x||_2^2, \\ x^T Q^{-1} x &\geq \frac{1}{M}||x||_2^2, \end{aligned}$$

i.e.

$$\begin{aligned} \gamma &= \sqrt{m}, \\ \tilde{\gamma} &= \frac{1}{\sqrt{M}}. \end{aligned}$$

Applying the theorem directly, the convergence rate for the steepest descent with BTLS is

$$c = 1 - 2m\alpha \min\left\{\frac{1}{M}, \frac{\beta m}{M^2}\right\}.$$

This is even worse than gradient descent with BTLS whose convergence rate is

$$c = 1 - 2m\alpha \min\left\{1, \frac{\beta}{M}\right\}.$$

Now we can conclude that the theorem above is not useful for better rate.

6.3 Newton's Method

In the last section we have seen that if the objective function f is quadratic and strongly convex, then its Hessian $\nabla^2 f$ is a constant positive definite matrix, and we can use the *change of coordinate* method to bring the condition number to 1 before doing gradient descent. We also showed that the change of coordinate method is actually equivalent to the steepest descent method with norm $||\cdot||_{\nabla^2 f}$.

Now what if the objective function f is not quadratic but still strongly convex? One consequence is that the Hessian $\nabla^2 f(x)$ will vary at different x, which means if we still want to use the change of coordinate method, the norm $||\cdot||_{\nabla^2 f(x)}$ is going to be different at each step. This gives us an intuition on the Newton step.

6.3.1 The Newton Step

Definition 1. *For a strongly convex objective function f, the **Newton step** at x is defined as the steepest descent direction using norm $||\cdot||_{\nabla^2 f(x)}$:*

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

The Newton step can be interpreted in the following three ways.

Interpretation 1. As we have already discussed, the Newton step is the steepest descent direction using the norm corresponding to the best change of coordinate method locally at every single step.

Interpretation 2. The Newton step minimizes the best (locally) quadratic approximation. Fig. 6.1 exhibits this idea. Suppose we are minimizing the function f and currently we are at x. We use a quadratic function $\tilde{f}$ to approximate f locally at x. Then we end up with

$$\tilde{f}(x + \Delta x) \approx f(x) + \nabla f(x)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 f(x) \Delta x.$$

Minimizing the right hand side of the above equation with respect to Δx yields

$$\Delta x = -\nabla^2 f(x)^{-1} \nabla f(x),$$

which is nothing but the Newton step defined earlier.

Interpretation 3. The Newton step at x is the first order approximation solution to the equation $\nabla f(x + \Delta x) = 0$. Recall that for any non-linear function ϕ, the first order approximation at x is

$$\phi(x + \Delta x) \approx \phi(x) + \phi'(x) \Delta x.$$

So $\phi(x + \Delta x) = 0$ yields

$$\Delta x \approx -\phi'(x)^{-1} \phi(x).$$

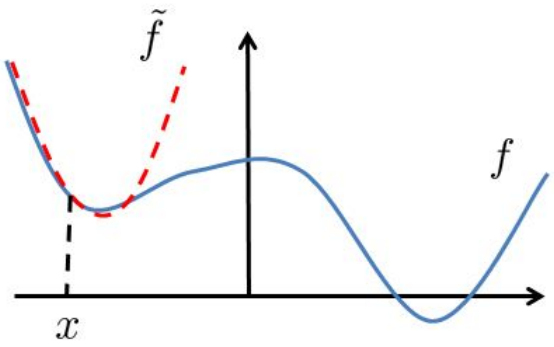

Figure 6.1. The quadratic fit problem. We use a quadratic function $\tilde{f}$ to approximate f locally at x.

Replacing ϕ with ∇f, we obtain

$$\Delta x \approx -\nabla^2 f(x)^{-1} \nabla f(x),$$

which is the Newton step.

6.3.2 Algorithm for Newton's Method

We give an outline of the algorithm for Newton's method as follows,

Repeat the following three steps.

1. Compute the Newton step $\Delta x_{\text{nt}}(x) = -\nabla^2 f(x)^{-1} \nabla f(x)$;
2. Choose step size η by backtracking line search (BTLS) or other line search methods;
3. Update $x^+ = x + \eta \Delta x_{\text{nt}}(x)$.

Note in the above algorithm, we did not give a *stopping rule*. This will be discussed in the next lecture.

6.3.3 Basic Properties of Newton's Method

In Subsection 6.3.1, Interpretation 1 of the Newton step implies that under Newton's method, one cannot do any better using change of coordinate method at any step, since Newton's method is already doing the best. This point can be further verified by the idea of *affine invariance*.

Definition 2. *Consider a descent algorithm which starts at $x^{(0)}$, and updates as $x^{(k)}$, $k = 1, 2, ...$ Then we apply an arbitrary affine transformation A to get $x^{(0)} = Ay^{(0)}$, and use the same descent algorithm on $y^{(0)}$ to get updates $y^{(k)}$, $k = 1, 2, ...$ If we have*

$$\{x^{(0)} = Ay^{(0)}\} \Rightarrow \{x^{(k)} = Ay^{(k)}, \forall k\}, \forall A,$$

then the descent algorithm is said to be ***affine invariant****.*

One consequence of affine invariance is that the sequences $\{x^{(k)}\}_k$ and $\{y^{(k)}\}_k$ are equivalent (under constant linear transformation) and thus have the same convergence behavior. As a result, for an affine invariant descent algorithm, any change of coordinates only changes the original updating sequence to an equivalent one and thus cannot improve the performance of the descent algorithm. Hence we obtain the following proposition.

Proposition 1. *Any affine invariant descent algorithm cannot be further improved by any change of coordinate method.*

One basic property of Newton's method is that it is affine invariant.

Proposition 2. *Newton's method is affine invariant.*

Proof: Suppose the objective function is f, which is strongly convex. Newton's method starts at $x^{(0)}$ and updates as $x^{(k)}$, $k = 1, 2, ...$ For an arbitrary affine transformation A, take $x^{(0)} = Ay^{(0)}$. Applying Newton's method on $y^{(0)}$ yields $y^{(k)}$, $k = 1, 2, ...$ Define $g(y) = f(Ax)$, then we have

$$\nabla g(y) = A^T \nabla f(Ay),$$
$$\nabla^2 g(y) = A^T \nabla^2 f(Ay) A.$$

We use induction to prove $x^{(k)} = Ay^{(k)}, \forall k$. Note we already have $x^{(0)} = Ay^{(0)}$. As induction hypothesis, suppose $x^{(n)} = Ay^{(n)}$, then we have

$$\begin{aligned} Ay^{(n+1)} &= Ay^{(n)} - \eta A \nabla^2 g(y^{(n)})^{-1} \nabla g(y^{(n)}) \\ &= Ay^{(n)} - \eta A (A^T \nabla^2 f(Ay^{(n)}) A)^{-1} A^T \nabla f(Ay^{(n)}) \\ &= Ay^{(n)} - \eta A A^{-1} (\nabla^2 f(Ay^{(n)}))^{-1} \nabla f(Ay^{(n)}) \\ &= x^{(n)} - \eta (\nabla^2 f(x^{(n)}))^{-1} \nabla f(x^{(n)}) \\ &= x^{(n+1)}. \end{aligned}$$

Thus by induction we have proved $x^{(k)} = Ay^{(k)}, \forall k$, which implies Newton's method is affine invariant. □

Propositions 1 and 2 imply that Newton's method cannot be further improved by a change of coordinates, which is consistent with Interpretation 1 in Subsection 6.3.1.

In fact, we can also check that the gradient descent method is NOT affine invariant by a similar approach as the proof to Proposition 2.

Proposition 3. *Gradient descent method is not affine invariant.*

Proof: Suppose the objective function is f, which is strongly convex. Gradient descent starts at $x^{(0)}$ and updates as $x^{(k)}$, $k = 1, 2, ...$ For an arbitrary affine transformation A, take $x^{(0)} = Ay^{(0)}$. Applying gradient descent on $y^{(0)}$ yields $y^{(k)}$, $k = 1, 2, ...$ Define $g(y) = f(Ax)$, then we have

$$\nabla g(y) = A^T \nabla f(Ay),$$

Note that

$$x^{(1)} = x^{(0)} - \eta \nabla f(x^{(0)}),$$

and that

$$\begin{aligned} Ay^{(1)} &= Ay^{(0)} - \eta A \nabla g(y^{(0)}) \\ &= Ay^{(0)} - \eta A(A^T \nabla f(Ay^{(0)})) \\ &= x^{(0)} - \eta A A^T \nabla f(x^{(0)}). \end{aligned}$$

Thus in general we have

$$x^{(1)} \neq Ay^{(1)}.$$

As a result, gradient descent is not affine invariant. □

Proposition 3 implies that it is possible to improve gradient descent using change of coordinate method, which is consistent with our earlier results.

At the end of this lecture, we state the outline of some basic results regarding the convergence behavior of Newton's method. More details will be discussed next time.

Fact. If objective function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies:

1. smooth and strongly convex, and
2. L-Lipschitz Hessian, i.e., $||\nabla^2 f(x) - \nabla^2 f(y)||_{\text{op}} \leq L||x - y||_2$, where $|| \cdot ||_{\text{op}}$ is the operator norm defined by $||A||_{\text{op}} = \sup_{||v||_2=1} ||Av||_2$,

then Newton's method converges with different convergence behaviors in two phases:

1. if start far from x^*, then convergence is linear on linear scale, i.e. $f(x^k) - f^*$ is linear w.r.t. k;
2. if start close to x^*, then convergence is so-called quadratic convergence, i.e., $\log\log(f(x^k) - f^*)$ is linear w.r.t. k.

Fig. 6.2 shows a typical convergence plot for Newton's method.

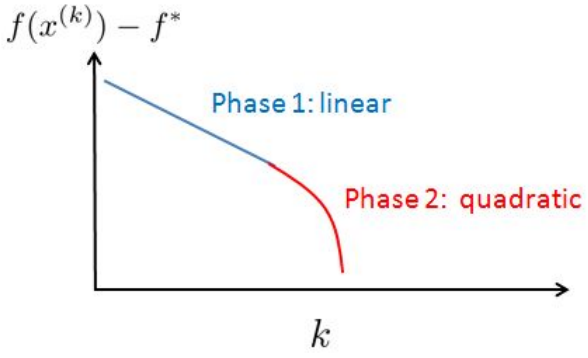

Figure 6.2. A typical convergence plot for Newton's method. There are two phases of different convergence behaviors: linear phase and quadratic phase.