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

Lecture 8 — September 25

Lecturer: Caramanis & Sanghavi *Scribe: Srinadh B, Anish Mittal*

8.1 Newton's method

8.1.1 Convergence

In the last class we have seen the two phases of Newton's method and the convergence proof. Newton's method can be also viewed as steepest descent with coordinate change at each step. In this method at each step the direction is determined by the Newton step and the magnitude can be determined by using back tracking line search(BTLS). The Newton step is $\triangle x_{nt} = -\nabla^2 f^{-1}(x)\nabla f(x)$.

Assumptions:

1. f is strongly convex.
$$mI \leq \nabla^2 f(x) \leq MI. \tag{8.1}$$

2. L-Lipschitz hessian.
$$||\nabla^2 f(x) - \nabla^2 f(y)||_{op} \leq L||x - y||_2. \tag{8.2}$$

Under the above two assumptions on the function f it has been shown in the last lecture that Newton's method converges in two distict phases. First phase is a damped phase with slow convergence and the second phase has quadratic convergence and hence very fast.

The transition to second phase happens when $||\nabla f(x)|| < \eta$ where $\eta < m^2/L$. When $||\nabla f(x)|| < \eta$ making $t = 1$ results in quadratic convergence. But in practice we rarely can compute the values m and L. Luckily BTLS chooses $t = 1$ automatically little after the threshold at $\eta < 3(1 - 2\alpha)m^2/L$. Alternately one can show that quadratic convergence is attained when $||x - x^+|| \leq \frac{2m}{3L}$.

8.1.2 Drawbacks of the analysis

While the Newton's method is affine invariant as noticed in previous lectures the convergence analysis is not. This is because of the constants in the assumptions m, M and L are not affine invariant. So we need a new set of assumptions that are affine invariant. Also we cannot compute the values of m and L in general and all the bounds on the number of steps for convergence are practically useless. We seek bounds that don't depend on such unknown constants. Self-Concordant functions as we will see in the coming sections don't

suffer these problems and are of huge practical significance because of their ubiquitous nature.

Consider the Lipschitz continuity of the hessian

$$||\nabla^2 f(x) - \nabla^2 f(y)|| \leq L||x - y||.$$

Note that on the LHS, hessian is a matrix and the norm used is the operator norm.

Definition 1. *The third derivative of f*

$$\nabla^3 f(x)[u] \stackrel{\triangle}{=} \lim_{\alpha \to 0} \frac{1}{\alpha}(\nabla^2 f(x + \alpha u) - \nabla^2 f(x)), \tag{8.3}$$

is a tensor with one more dimension than the hessian and evaluating it at vector u gives $\nabla^3 f(x)[u]$ a matrix. The Lipschitz assumption implies that

$$\nabla^3 f(x)[u] \leq \lim_{\alpha \to 0} \frac{1}{\alpha} L||\alpha u|| = L||u||.$$

Therefore

$$v^T \nabla^3 f(x)[u]v \leq L||u||||v||^2. \tag{8.4}$$

Proposition 1. *In the above equation 8.4 the LHS is affine invariant whereas the RHS is not.*

Proof: Let there be a linear transformation of variables with the invertible matrix A. Let $\phi(x) = f(Ax), x = A\tilde{x}, v = A\tilde{v}$ and $u = A\tilde{u}$. Then

$$\begin{aligned}
\tilde{v}^T \phi'''(\tilde{x})[\tilde{u}]\tilde{v} &= \tilde{v}^T \lim_{\alpha \to 0} \frac{1}{\alpha}(\nabla^2 \phi(\tilde{x} + \alpha\tilde{u}) - \nabla^2 \phi(\tilde{x}))\tilde{v} \\
&= (A^{-1}v)^T \lim_{\alpha \to 0} A^T \frac{1}{\alpha}(\nabla^2 f(A\tilde{x} + \alpha A\tilde{u}) - \nabla^2 f(A\tilde{x}))A(A^{-1}v) \\
&= v^T (A^{-1})^T A^T \lim_{\alpha \to 0} \frac{1}{\alpha}(\nabla^2 f(x + \alpha u) - \nabla^2 f(x))AA^{-1}v \\
&= v^T \lim_{\alpha \to 0} \frac{1}{\alpha}(\nabla^2 f(x + \alpha u) - \nabla^2 f(x))v \\
&= v^T f'''(x)v.
\end{aligned}$$

This shows that LHS is affine invariant. Now applying the affine transform to RHS gives

$$L||\tilde{u}||||\tilde{v}||^2 = L||A^{-1}u||||A^{-1}v||^2, \tag{8.5}$$

which is clearly not equal to $L||u||||v||^2$. Hence RHS is not affine invariant.

□

To make this definition affine invariant replace the 2 norm with an affine invariant norm $||.||_{\nabla^2 f(x)}$. Doing that and replacing v with u gives

$$u^T \nabla^3 f(x)[u]u \leq L(u^T \nabla^2 f(x) u)^{\frac{3}{2}}. \tag{8.6}$$

In the 1-dimensional case this gives

$$\nabla^3 f(x) \leq L(\nabla^2 f(x))^{\frac{3}{2}}. \tag{8.7}$$

This motivates the definition of self-concordant functions.

8.2 Self-Concordant functions

Definition 2. *A convex function $f : \mathcal{R} \to \mathcal{R}$ is called self-concordant (S-C) if*

$$|f'''(x)| \leq 2(f''(x))^{\frac{3}{2}}. \tag{8.8}$$

The number 2 can be replaced with any positive constant and is there just for simplicity. To see that let a function f be self-concordant for some positive constant k

$$|f'''(x)| \leq k(f''(x))^{\frac{3}{2}}. \tag{8.9}$$

Now consider $\tilde{f}(x) = \frac{k^2}{4} f(x)$. This function satisfies the S-C property with constant 2.

$$\begin{aligned} |\tilde{f}'''(x)| &= \frac{k^2}{4}|f'''(x)| \\ &\leq \frac{k^3}{4}(f''(x))^{\frac{3}{2}} \\ &= \frac{k^3}{4}(\frac{4}{k^2}\tilde{f}''(x))^{\frac{3}{2}} \\ &= 2(\tilde{f}''(x))^{\frac{3}{2}}. \end{aligned}$$

Thus if a function satisfies 8.9 with some positive constant k it can be scaled to satisfy 8.8 and become S-C with constant 2.

Definition 3. *A function $f : \mathcal{R}^n \to \mathcal{R}$ is S-C if $\tilde{f}(t) = f(x + tv)$ is S-C along all directions $v \in \mathcal{R}^n$.*

8.2.1 Examples

The functions below are S-C on their respective domains.

1. $f(x) = -\log(x)$. **dom** $f = \{x \in \mathcal{R} : x > 0\}$.

Proof:

$$f^{'}(x) = \frac{-1}{x}, f^{''}(x) = \frac{1}{x^2}, f^{'''}(x) = \frac{-2}{x^3}.$$
$$|f^{'''}(x)| = \frac{2}{x^3} = 2(\frac{1}{x^2})^{\frac{3}{2}} = 2f^{''}(x)^{\frac{3}{2}}.$$

□

2. $f(x) = -\log(a^T x - b)$. **dom** $f = \{x \in \mathcal{R}^n : a^T x - b > 0\}$.

   Proof: To show this function is S-C we need to show it is S-C along each line in its domain. Let $\tilde{f}(t) = -\log(a^T x + a^T v t - b)$. **dom** $\tilde{f} = \{t \in \mathcal{R} : a^T x + a^T v t - b > 0\}$.

$$\tilde{f}^{'}(t) = \frac{-a^T v}{(a^T x + a^T v t - b)}, \tilde{f}^{''}(t) = \frac{(a^T v)^2}{(a^T x + a^T v t - b)^2}, \tilde{f}^{'''}(t) = \frac{-2(a^T v)^3}{(a^T x + a^T v t - b)^3}$$

$$|\tilde{f}^{'''}(t)| = \frac{2(a^T v)^3}{(a^T x + a^T v t - b)^3} = 2(\tilde{f}^{''}(t))^{\frac{3}{2}}.$$

   Note that this result is true for all values of v and hence f is S-C. □

3. $f(x) = -\sum_{i=1}^{M} \log(a_i^T x - b_i)$. **dom** $f = \{x \in \mathcal{R} : a_i^T x - b_i > 0 \forall i.\}$

   This function is just sum of S-C functions, so it is S-C. Look at the proposition 2 below.

4. $f(x) = -\log(\det(X))$. **dom** $f = \{X : X$ is a positive definite matrix $(S_{++}^n)\}$.

   Proof: To show this function is S-C we need to show it is S-C along each line in it's domain. Let $\tilde{f}(t) = -\log(\det(X + tV))$, where $V \in S^n$ is a symmetric $n \times n$ matrix. $\tilde{f}(t)$ can be rewritten as follows.

$$\begin{aligned}
\tilde{f}(t) &= -\log(\det(X + tV)) && (8.10)\\
&= -\log(\det(X^{\frac{1}{2}}(I + tX^{\frac{-1}{2}} V X^{\frac{-1}{2}})X^{\frac{1}{2}})) && (8.11)\\
&= -\log(\det(X)) - \log(\det(I + tX^{\frac{-1}{2}} V X^{\frac{-1}{2}})) && (8.12)\\
&= -\log(\det(X)) - \sum_{i=1}^{n} \log(1 + t\lambda_i), && (8.13)
\end{aligned}$$

   where λ_i are eigenvalues of $X^{\frac{-1}{2}} V X^{\frac{-1}{2}}$. 8.11 follows from the fact that X is positive definite matrix. 8.12 follows from the fact that $\det(AB) = \det(A)\det(B)$. 8.13 follows from the fact that $\det(A) = \prod_{i=1}^{n} \lambda_i(A)$. Thus $\tilde{f}(t)$ is sum of S-C functions hence it is also S-C. This proof doesn't depend on V and is true for all V, hence $f(X)$ is S-C. □

Proposition 2. *Sum of S-C functions is S-C.*

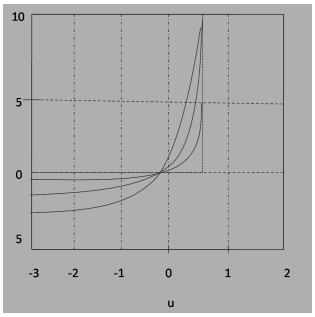

Figure 8.1. An example of barrier function, $-\frac{1}{t}log(-x+.5)$ plotted for different values of t.

Proof: Let $f_1(x)$ and $f_2(x)$ be S-C.

$$\begin{aligned} |(f_1+f_2)'''(x)| &= |f_1'''(x)+f_2'''(x)| \\ &\leq |f_1'''(x)|+|f_2'''(x)| \\ &\leq 2(f_1''(x))^{\frac{3}{2}}+2(f_2''(x))^{\frac{3}{2}} \\ &\leq 2((f_1+f_2)''(x))^{\frac{3}{2}}. \end{aligned}$$

The last inequality is true because $f_1''(x)$ and $f_2''(x)$ are greater than 0 because of convexity and since $(a+b)^{\frac{3}{2}} \geq a^{\frac{3}{2}}+b^{\frac{3}{2}}$, when $a,b \geq 0$. □

Exercise: Read the rules for combining S-C functions to get a S-C function back from the book.

Definition 4. *F is called a barrier function on its domain if $\forall x_k \to x \in \partial($**dom** $F)$, $F(x_k) \to \infty$,*

where ∂(Q) means the boundary of the set Q.

Proposition 3. *Self concordant functions are a barrier of their domain.*

Definition 5. *F is a S-C barrier if F is S-C and $||\nabla f(x)||^2_{\nabla^2 f(x)^{-1}} < \alpha$.*

8.2.2 Convergence

We have seen convergence results for Newton's method in the previous section with various unknown constants m and L. Below is a convergence result in the case of S-C functions and this result does not depend on any unknown constants.

Theorem 8.1. *If f is self concordant with constant $L=2$, we are guaranteed to be in the region of quadratic convergence as long as*

$$\lambda_f(x) \stackrel{\triangle}{=} ||\nabla f(x)||_{\nabla^2 f^{-1}(x)} < \frac{3-\sqrt{(}5)}{2} = \overline{\lambda}. \tag{8.14}$$

8.2.3 Summary

1. Linear functions are S-C.
2. S-C functions are barriers for their domain.
3. Sums of S-C functions is S-C.
4. S-C functions are easy to optimize by the Newton's method.

8.3 Interior point method

So far we have seen how to solve an unconstrained optimization problem using various descent methods. In this section we will get a brief idea of how Newton's method with S-C function's barrier property can be used to solve constrained optimization problems. In particular we will see an example of the interior point methods. Consider the linear program(LP), a constrained convex optimization problem

$$\begin{aligned} &\min \langle c, x \rangle \\ \text{s.t. } &Ax = b, \\ &x \geq 0. \end{aligned}$$

On a side note the generalization of this LP to matrix scenarios gives rise to semi-definite program(SDP).

$$\begin{aligned} &\min \langle C, X \rangle \\ \text{s.t. } &\langle A_i, X \rangle &&= b_i, \forall i, \\ &X &&\succeq 0, \end{aligned}$$

where $\langle A, B \rangle$ is the trace innerproduct between two matrices defined as $= \sum_{i,j} A_{i,j} B_{i,j}$ and $X \succeq 0$ denotes that X has to be a positive semi-definite matrix.

Let the feasible set determined by the constraints be Q. Let

$$x^* = \arg\min_x : \langle c, x \rangle \text{ s.t. } x \in Q. \tag{8.15}$$

Now define a S-C function F such that the domain of F is Q. Converting it to an unconstrained problem let

$$x_t = \arg\min_x : t \langle c, x \rangle + F(x), \tag{8.16}$$

where $t \geq 0$. Since a S-C function is also a barrier function x will always lie in Q. It is also easy to see that as $t \to \infty$, $x_t \to x^*$. This is an interior point method. So to solve 8.15 we solve 8.16 by increasing t at each iteration. Let $f(x, t) = t \langle c, x \rangle + F(x)$.

1. x_k is the newton step in k_{th} iteration.

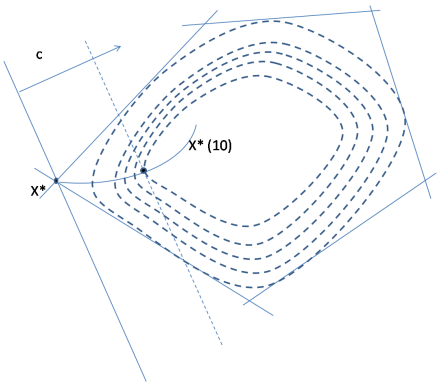

Figure 8.2. An example of central path followed by interior path method for solving an LP with the logarithmic barrier function. The solid lines represent the halfspace constraints. The dotted lines are contours of the objective. x^* is the optimal point and $x^*(10)$ is the value at 10th iteration. The central path converges to x^* as $t \to \infty$.

2. update $t^+ = t + \triangle$.

3. x_{k+1} is obtained via Newton's method initialized at x_k for $f(x, t^+)$.

The path followed by x_k is called central path.

One important question is whether Newton's method converges for this new function $f(x,t)$. It doesn't satisfy the hessian Lipschitz assumption required for convergence because the S-C function $F(x)$ is not Lipschitz. But since $f(x,t)$ is sum of two S-C functions it is also S-C and by theorem 8.1 Newton's method converges.

For faster convergence we want $\triangle$ as big as possible while still in the quadratically convergent phase/region.

Lemma 8.2. *If $\lambda_f^2(x) = ||\nabla f(x)||^2_{\nabla^2 f(x)^{-1}}$ is uniformly bounded on the domain of f, t increases linearly i.e. $t^+ = t + \triangle$ where $\triangle \sim O(t)$, hence doubling t each time.*