Operations Research Techniques and Algorithms (620-361)

Dr Yao-ban Chan

y.chan@ms.unimelb.edu.au
Telephone: 8344 9073
Office: Room 198, Richard Berry Building

Christina Burt
c.burt@ms.unimelb.edu.au
Telephone: 8344 1797
Office: 139 Barry St

Monday 31st March, 2008

Today's Lecture

- Unimodal n-D unconstrained optimisation
  - Rates of convergence

Rates of convergence

Definition: Let $\{x^k\}$ converge to x^* in $\Re^n$.

1. If, for some constant $c \in (0, 1)$ and all large enough k, we have
$$\|x^{k+1} - x^*\| \le c\|x^k - x^*\|,$$
then we say that the *rate of convergence of* $\{x^k\}$ *is linear*, or $x^k \to x^*$ *linearly* (as $k \to \infty$).

2. If, for some sequence $\{c_k\}$ of positive scalars such that $c_k \downarrow 0$, we have
$$\|x^{k+1} - x^*\| \leq c_k \|x^k - x^*\|,$$
then the rate of convergence is *superlinear*, and $x^k \to x^*$ superlinearly.
3. If, for some $c > 0$ and all large enough k,
$$\|x^{k+1} - x^*\| \ \leq \ c\|x^k - x^*\|^2,$$
the rate of convergence is *quadratic*, and $x^k \to x^*$ *quadratically*.

Linear convergence is considered rather slow, especially compared with quadratic convergence.

For instance consider the sequence $x^k = 10^{-k}$; then $x^k \to x^* = 0$ linearly. Actually we get one extra decimal place of accuracy per iteration, in the sense that x^{k+1} is one tenth the size of x^k (or ten times closer to x^*). So it takes 10 (respectively 20) iterations for the iterates to be less than 10^{-10} (resp. 10^{-20}).

Compare this with the sequence $x^k = 10^{-k^2} \to 0$; note $x^{k+1} = 10^{-k^2-2k-1}$ has $2k+1$ more decimal places of accuracy than x^k. In particular, it only takes 4 (resp. 5) iterations for the iterates to be less than 10^{-10} (10^{-20}).

It is easy to show that superlinear convergence is also linear (although the converse is not true!): as $c_k \to 0$, there must be a k' such that for all $k \geq k'$, $c_k < 1$. Then take $c = \max_{k \geq k'} c_k$.
It is also easy to show that quadratic convergence is also superlinear (again the converse is not true!): take $c_k = c\|x^k - x^*\|$.
If the sequence converges then this must converge to 0!

Examples:

$$x^k = 10^{-k^2} \to 0$$

$$\|x^{k+1} - x^*\| = 10^{-(k+1)^2} = 10^{-k^2-2k-1} = 10^{-2k-1}10^{-k^2}$$
$$= 10^{-2k-1}\|x^k - x^*\|$$

If we take $c_k = 10^{-2k-1}$, this shows that the sequence converges superlinearly.

$$x^k = 10^{-2^k} \to 0$$

$$\|x^{k+1} - x^*\| = 10^{-2^{k+1}} = 10^{-2^k \cdot 2} = \left[10^{-2^k}\right]^2$$

so we can take $c = 1$ to satisfy the quadratic criterion.

$$x^k = (\frac{k-1}{k+2}, \frac{1}{2}) \to (1, \frac{1}{2})$$

$$\|x^k - x^*\| = \sqrt{\left(\frac{k-1}{k+2} - 1\right)^2} = \frac{1}{k+2}\sqrt{(k-1-k-2)^2} = \frac{3}{k+2}$$

$$\frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = \frac{3}{k+3}\frac{k+2}{3} = \frac{k+2}{k+3} \to 1$$

so there is no c which satisfies the linearity criterion. Hence this sequence converges slower that linearly.

In practice we don't know the value of the minimum x^*, but we do know that $\nabla f(x^*) = 0$.
So instead of looking at $\|x^k - x^*\|$ we look at $\|\nabla f(x^k) - \nabla f(x^*)\| = \|\nabla f(x^k)\|$, the norm of the "residual" $\nabla f(x^k)$. There is a close relationship between $\|x^k - x^*\|$ and $\|\nabla f(x^k)\|$.

Lemma:

Let f be C^2, $\nabla f(x^*) = 0$, and the Hessian matrix $\nabla^2 f(x^*)$ be invertible. If $x^k \rightarrow x^*$ then $\nabla f(x^k) \rightarrow \nabla f(x^*)$, and the rate of convergence of these two sequences is identical.

This justifies using a test of the form $\|\nabla f(x^k)\| < \epsilon$ when we write computer code to whether x^k "close enough" to a stationary point.

Newton's method

To increase the rate of convergence of an iterative descent method, we must use or approximate second-order information, that is knowledge of the Hessian function $\nabla^2 f$.
Since it does not use knowledge of the Hessian, the steepest descent algorithm is a first-order method.
The classical second-order method is *Newton's Method*.

The idea behind Newton's Method is to minimise at each iteration the *quadratic approximation* of f around the current iterate x^k. This generates at each iteration a sub-problem that is "simpler", in the sense that we know the mathematical properties of the quadratic approximation.

Suppose f is C^2. Consider an iterate x^k such that $\nabla f(x^k) \neq 0$ and the Hessian $\nabla^2 f(x^k)$ is invertible. Then the *Newton direction* is given by

$$d^k := -\nabla^2 f(x^k)^{-1} \nabla f(x^k). \tag{1}$$

By choosing the direction d^k according as the Newton direction, we are choosing the direction that would get to the minimum in one step with a step length of $t_k = 1$ if the second order approximation held exactly, that is if f were a quadratic function (of course, in general, f is not quadratic, so we have an approximation to the minimum of f - hopefully, an iterative method based on this will converge to an actual minimum of f).

For an arbitrary function in C^2, there is no guarantee that $\nabla^2 f(x^k)$ is invertible, nor is it always true that the Newton direction is a descent direction.

Therefore, at each iteration we must check both of these conditions. Specifically we must check

- that $\nabla^2 f(x^k)$ is invertible and,
- that the Newton direction $d^k = -\nabla^2 f(x^k)^{-1} \nabla f(x^k)$ (if it is well-defined) is a descent direction.

If both of these conditions are satisfied, we can then use a line search procedure to find a suitable step size $t_k > 0$, and finally set $x^{k+1} = x^k + t_k d^k$.

Lemma:

If $\nabla f(x^k) \neq 0$ and $\nabla^2 f(x^k)$ is positive definite, then $\nabla^2 f(x^k)$ is invertible and the Newton direction is a descent direction for f at x^k.

Proof: Because $\nabla^2 f(x^k)$ is positive definite and symmetric, it has all positive eigenvalues. Therefore the determinant, which is the product of the eigenvalues, is nonzero, so it is invertible. Furthermore, the eigenvalues of the inverse are the reciprocals of the eigenvalues of $\nabla^2 f(x^k)$, so the inverse is also positive definite.

$$\langle d^k, \nabla f(x^k) \rangle = \langle \nabla f(x^k), -\nabla^2 f(x^k)^{-1} \nabla f(x^k) \rangle < 0$$

from positive definiteness. So the Newton direction is a descent direction.

Algorithm for Newton's Method.

This follows the framework of the general descent method. At iteration k, the descent direction is $d^k = -\nabla^2 f(x^k)^{-1} \nabla f(x^k)$ if $\nabla^2 f(x^k)$ is positive definite, otherwise d^k is any descent direction for f at x^k, for example $-\nabla f(x^k)$; and t_k is determined by the line search procedure.

Newton's Method

To minimise a unimodal function $f : \Re^N \rightarrow \Re$ to within tolerance ϵ.

1. Select $x^0 \in \Re^N$.
   Set $k = 0$.
2. If $||\nabla f(x^k)|| < \epsilon$ then stop.
   If $\nabla^2 f(x^k)$ is positive definite, then
   Set $d^k = -\nabla^2 f(x^k)^{-1} \nabla f(x^k)$.
   Else, set $d^k = -\nabla f(x^k)$

3. Select step length t_k either
   - by solving the single-variable minimisation problem: $\min q(t) = f(x^k + td^k)$.
   - by using our procedure for finding a step length that satisfies the Armijo-Goldstein and Wolff conditions.
4. Set $k = k + 1$.
   Set $x^{k+1} = x^k + t_k d^k$.
   Return to step 2.

Theorem: Suppose f is C^2, $x^0 \in \Re^n$, $0 < \sigma < 1/2 \leq \mu < 1$, and we implement Newton's Method as above with the step length chosen to satisfy the Armijo-Goldstein and Wolff conditions. If $\{x^k\}$ has a cluster point x^* such that $\nabla^2 f(x^*)$ is positive definite then

1. x^* is a local minimum of f.
2. For sufficiently large k, d^k is the Newton direction, $t_k = 1$ and $x^{k+1} = x^k + d^k$.
3. $x^k \to x^*$ superlinearly, indeed the rate of convergence is quadratic if f is C^3.

So, for example, if we applied Newton's method to the function $f(x) = \sin(x)$ to find the minimum $3\pi/2$, the Theorem shows that if the step length satisfies the Armijo-Goldstein and Wolff conditions, then since

$$\nabla^2 f(3\pi/2) = -\sin(3\pi/2) = 1 > 0,$$

we will find converge to the minimum at a quadratic rate.
On the other hand, if we applied Newton's method to the function $f(x) = x^4$, then as we got close to the minimum of 0, we would not be able to tell if $\nabla^2 f(x^k)$ was positive definite (read: positive) or not. So this would depend on the thresholds we use for ∇f and $\nabla^2 f$. If we set them badly, then we may revert to the steepest descent method.

Example. Consider the problem of minimizing $f(x) = x_1^2 + x_2^2 - x_1x_2 - 3x_1 + 3x_2 + 3$. Apply one step of the steepest descent method starting with the point $x^0 = (0,0)^T$. Apply one step of Newton's Method under the same conditions. By finding all stationary points and checking second order conditions, find an exact solution.