Unconstrained Optimization IV: Newton and Quasi-Newton methods

Note Title 4/12/2022

Recall that the Newton search direction

$$p_k^N = -\nabla^2 f_k^{-1} \nabla f_k$$

is a descent direction if $\nabla^2 f_k$ (and hence also $\nabla^2 f_k^{-1}$) is positive definite.

But it may or may not be a descent direction when $\nabla^2 f_k$ is not positive definite (and it is not well-defined when $\nabla^2 f_k$ is singular.)

There are at least two approaches for obtaining a globally convergent iteration based on the Newton step:

(i) a line search approach, in which $\nabla^2 f_k$ is modified when necessary to make it positive definite and thereby yield descent

(ii) a trust region approach, in which $\nabla^2 f_k$ is used to form a quadratic model that is minimized in a ball around the current iterate x_k.

But let's first study the local ROC properties.

If $\nabla^2 f(x^*) \succ 0$, by continuity $\nabla^2 f(x) \succ 0$ and the Newton direction at x is a well-defined descent direction.

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

Theorem 3.5.

Suppose that f is twice differentiable and that the Hessian $\nabla^2 f(x)$ is Lipschitz continuous (see (A.42)) in a neighborhood of a solution x^* at which the sufficient conditions (Theorem 2.4) are satisfied. Consider the iteration $x_{k+1} = x_k + p_k$, where p_k is given by (3.30). Then

$\nabla f_* = 0$, $\nabla^2 f_* \succ 0$

$p_k = -\nabla^2 f_k^{-1} \nabla f_k$

(i) *if the starting point x_0 is sufficiently close to x^*, the sequence of iterates converges to x^*;*

(ii) *the rate of convergence of $\{x_k\}$ is quadratic; and*

(iii) *the sequence of gradient norms $\{\|\nabla f_k\|\}$ converges quadratically to zero.*

Proof: Note that

$$x_{k+1} - x^* = x_k + p_k^N - x^* = x_k - x^* - \nabla^2 f_k^{-1} \nabla f_k$$

$$= \nabla^2 f_k^{-1} \left[\nabla^2 f_k (x_k - x^*) - (\nabla f_k - \nabla f_*) \right] \qquad (\nabla f_* = 0)$$

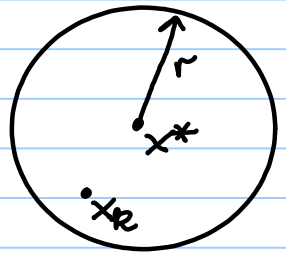
By the fundamental thm of calculus, $\nabla f_k - \nabla f_* = \int_0^1 \nabla^2 f(x_k + t(x^* - x_k))(x_k - x^*)\,dt$, we have

$$\|\nabla^2 f_k(x_k - x^*) - (\nabla f_k - \nabla f_*)\|$$

$$= \left\| \int_0^1 [\nabla^2 f_k - \nabla^2 f(x_k + t(x^* - x_k))](x_k - x^*)\, dt \right\|$$

$$\leq \int_0^1 \underbrace{\|\nabla^2 f(x_k) - \nabla^2 f(x_k + t(x^* - x_k))\|}_{\leq L\,\|t(x^* - x_k)\|} \|x_k - x^*\|\, dt$$

$$\leq L\|x_k - x^*\|^2 \int_0^1 t\, dt = \frac{1}{2} L \|x_k - x_*\|^2 \quad \text{if } x_k \approx x^*.$$



Since $\nabla^2 f(x^*)$ is non-singular, $\|\nabla^2 f(x)^{-1}\| \leq 2\|\nabla^2 f(x^*)^{-1}\|$ for $x \approx x^*$.

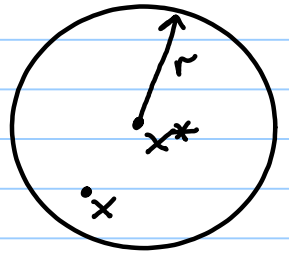
So when x_k is close enough to x^*,

$$\|x_{k+1} - x^*\| \leq \|\nabla^2 f_k^{-1}\| \, \|\nabla^2 f_k(x_k - x^*) - (\nabla f_k - \nabla f_*)\|$$

$$\leq 2\|\nabla^2 f(x^*)^{-1}\| \tfrac{1}{2} L \|x_k - x_*\|^2 = \underbrace{L\|\nabla^2 f(x^*)^{-1}\|}_{=:\tilde{L}} \|x_k - x^*\|^2. \quad \text{—(*)}$$

And this must mean x_{k+1} is even closer to x_* and, in fact, x_k converges to x^* quadratically.

To rigorize the 'close enough' argument above, let $r>0$ be the radius of a ball around x^* so that for every x st $\|x-x^*\|<r$,



- $\nabla^2 f(x) \succ 0$
- $\|\nabla^2 f(x) - \nabla^2 f(x^*)\| \leq L\|x-x^*\|$
- $\|\nabla^2 f(x)^{-1}\| \leq 2\|\nabla^2 f(x^*)^{-1}\|$

Then if x_0 is close enough to x^* so that (say) $\|x_0-x^*\| \leq \min(r, 1/1.1\tilde{L})$, then we can use the inequality $(*)$ inductively to deduce that all x_k are well-defined and $\|x_k-x^*\| \downarrow 0$ monotonically and quadratically.

For the last claim:

$$\|\nabla f(x_{k+1})\| = \|\overbrace{\nabla f(x_{k+1}) - \nabla f_k - \nabla^2 f(x_k) p_k^N}^{=0}\|$$

$$= \|\int_0^1 \nabla^2 f(x_k + t\overbrace{p_k^N}^{x_{k+1}-x_k})\, dt - \nabla^2 f(x_k) p_k^N\|$$

$$\leq \int_0^1 \|\nabla^2 f(x_k + t p_k^N) - \nabla^2 f(x_k)\|\, \|p_k^N\|\, dt$$

$$\leq \tfrac{1}{2} L \|p_k^N\|^2$$

$$\leq \tfrac{1}{2} L \underbrace{\|\nabla^2 f(x_k)^{-1}\|^2}_{\leq 2^2 \|\nabla^2 f(x^*)^{-1}\|^2} \|\nabla f_k\|^2 \quad \leq 2L \|\nabla^2 f(x^*)^{-1}\|^2 \|\nabla f_k\|^2.$$

So if $\|\nabla f_0\|$ is small enough, the gradient norms converge to zero quadratically.

Q.E.D.

Something red herring (to me) in the way this theorem is stated in N&W:

> Nowhere in the proof did we use the assumption that $\nabla^2 f(x^*) \succ 0$, the proof goes through (with trivial changes) if we merely assume $\nabla^2 f(x^*)$ is non-singular!

Ex: Go through the proof and check it.

And it has two implications:

(I) If x^* is any critical point of f with $\nabla^2 f(x^*)$ non-singular (x^* can be a saddle point, or even a maximizer), when x_0 is close enough to x^*, Newton's

method is guaranteed to produce a sequence $\{x_k\}$ that converges to x^* monotonically (meaning $\|x_k - x^*\| \downarrow$ monotonically, NOT $f(x_k) \downarrow$ monotonically) and quadratically.

Ex. Revisit the function $f(x_1, x_2) = \frac{1}{2}x_1^2 - \frac{1}{2}x_2^2 + \frac{1}{4}x_2^2$, which has a saddle point at $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ with Hessian $\begin{bmatrix} 1 & \\ & -1 \end{bmatrix}$.

According to the theory, if x_0 is close enough to $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$, the Newton iterates x_k converge to $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ (very quickly.) This is NOT the case for gradient descent!

Verify this fact experimentally. Use Beck's Matlab function pure_newton.m.

- for x_0 close enough to $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$, is $f(x_k)$ monotonically $\downarrow$?
- can you observe quadratic convergence empirically ?
- what happens to $\{x_k\}$ if x_0 is not too close to $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$?

Verify also experimentally that for most x_0 — no matter how close to $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ — $\{x_k\}$ generated by (any version of) GD does not converge to $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

(II) The same proof can be easily adapted to prove:

Thm: Let $F: \mathbb{R}^n \to \mathbb{R}^n$ be C^1 in some open neighborhood D of x^*, at which $F(x^*) = 0$ and $DF(x^*)$ non-singular. Assume DF is Lipshitz continuous in a neighborhood of x^* (automatically true if F is C^2.) Consider the iteration generated by Newton's method (for solving $F(x) = 0$):

$$x_{k+1} = x_k + p_k \quad , \quad p_k = -DF(x_k)^{-1} F(x_k).$$

Then:

(i) *if the starting point x_0 is sufficiently close to x^*, the sequence of iterates converges to x^*;*

(ii) *the rate of convergence of $\{x_k\}$ is quadratic; and*

(iii) *the sequence of gradient norms $\{\|\nabla f_k\|\}$ converges quadratically to zero.*

$F(x_k)$

If we apply this (obsentibly more general) result to solve the nonlinear system of equations

$$\nabla f = 0, \quad f: \mathbb{R}^n \to \mathbb{R}$$

we get the same conclusions as the original result. Again, the result has nothing to do with whether x^* is a local min, a local max, or a saddle point.

Two interesting numerical examples from [Beck]:

Example 5.3. Consider the minimization problem

$$\min_{x,y} 100x^4 + 0.01y^4,$$

whose optimal solution is obviously $(x,y) = (0,0)$. This is a rather poorly scaled problem, and indeed the gradient method with initial vector $\mathbf{x}_0 = (1,1)^T$ and parameters $(s,\alpha,\beta,\varepsilon) = (1,0.5,0.5,10^{-6})$ converges after the huge amount of 14612 iterations:

```
>> f=@(x)100*x(1)^4+0.01*x(2)^4;
>> g=@(x)[400*x(1)^3;0.04*x(2)^3];
>> [x,fun_val]=gradient_method_backtracking(f,g,[1;1],1,0.5,0.5,1e-6)
iter_number =    1 norm_grad = 90.513620 fun_val = 13.799181
iter_number =    2 norm_grad = 32.381098 fun_val = 3.511932
iter_number =    3 norm_grad = 11.472585 fun_val = 0.887929
      :                    :                    :
iter_number = 14611 norm_grad = 0.000001 fun_val = 0.000000
iter_number = 14612 norm_grad = 0.000001 fun_val = 0.000000
```

Invoking pure Newton's method, we obtain convergence after only 17 iterations:

```
>> h=@(x)[1200*x(1)^2,0;0,0.12*x(2)^2];
>> pure_newton(f,g,h,[1;1],1e-6)
iter=  1 f(x)=19.7550617284
iter=  2 f(x)=3.9022344155
iter=  3 f(x)=0.7708117364
       :              :
iter= 15 f(x)=0.0000000027
iter= 16 f(x)=0.0000000005
iter= 17 f(x)=0.0000000001
```

Thomas Apr 20 Reply

Look at the iterates carefully, it is clear that f(x) decreases by about 1/5 in every iteration.

$$\nabla^2 f(x^*) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\Downarrow$$

GD: ROC deterioted from linear to arithmetic
$O(\theta^k)$ $O(1/k^\alpha)$

Newton: ROC deterioted from quadratic to linear

Extra credit: Prove it.

2.

Example 5.4. Consider the minimization problem

$$\min_{x_1,x_2} \sqrt{x_1^2+1}+\sqrt{x_2^2+1},$$

whose optimal solution is $\mathbf{x}=0$. The Hessian of the function is

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} \frac{1}{(x_1^2+1)^{3/2}} & 0 \\ 0 & \frac{1}{(x_2^2+1)^{3/2}} \end{pmatrix} \succ 0.$$

$\Longrightarrow -\nabla^2 f(x)^{-1}\nabla f(x)$ is a descent dir. $\forall x$.

However, when $\|x_0\| \gg 1$, there is no reason to believe that:

the minimizer of f ($= \begin{bmatrix}0\\0\end{bmatrix}$) $\approx$ the minimizer of the quadratic approx. of f at x_0

x_0

$\begin{bmatrix}0\\0\end{bmatrix}$

Note that despite the fact that the Hessian is positive definite, there does not exist an $m > 0$ for which $\nabla^2 f(\mathbf{x}) \succeq m\mathbf{I}$. This violation of the basic assumptions can be seen practically. Indeed, if we employ Newton's method with initial vector $\mathbf{x}_0 = (1,1)^T$ and tolerance parameter $\varepsilon = 10^{-8}$ we obtain convergence after 37 iterations:

```
>> f=@(x)sqrt(1+x(1)^2)+sqrt(1+x(2)^2);
>> g=@(x)[x(1)/sqrt(x(1)^2+1);x(2)/sqrt(x(2)^2+1)];
>> h=@(x)diag([1/(x(1)^2+1)^1.5,1/(x(2)^2+1)^1.5]);
>> pure_newton(f,g,h,[1;1],1e-8)
iter=  1 f(x)=2.8284271247
iter=  2 f(x)=2.8284271247
   :                :
iter= 30 f(x)=2.8105247315
iter= 31 f(x)=2.7757389625
iter= 32 f(x)=2.6791717153
iter= 33 f(x)=2.4507092918
iter= 34 f(x)=2.1223796622
iter= 35 f(x)=2.0020052756
iter= 36 f(x)=2.0000000081
iter= 37 f(x)=2.0000000000
```

slow

quadratic convergence!

Note that in the first 30 iterations the method is almost stuck. On the other hand, the gradient method with backtracking and parameters $(s,\alpha,\beta)=(1,0.5,0.5)$ converges after only 7 iterations:

```
>> [x,fun_val]=gradient_method_backtracking(f,g,[1;1],1,0.5,0.5,1e-8);
iter_number =    1 norm_grad = 0.397514 fun_val = 2.084022
iter_number =    2 norm_grad = 0.016699 fun_val = 2.000139
iter_number =    3 norm_grad = 0.000001 fun_val = 2.000000
iter_number =    4 norm_grad = 0.000001 fun_val = 2.000000
iter_number =    5 norm_grad = 0.000000 fun_val = 2.000000
iter_number =    6 norm_grad = 0.000000 fun_val = 2.000000
iter_number =    7 norm_grad = 0.000000 fun_val = 2.000000
```

If we start from the more distant point $(10,10)^T$. The situation is much more severe. The gradient method with backtracking converges after 13 iterations:

```
>> [x,fun_val]=gradient_method_backtracking(f,g,[10;10],1,0.5,0.5,1e-8);
iter_number =    1 norm_grad = 1.405573 fun_val = 18.120635
iter_number =    2 norm_grad = 1.403323 fun_val = 16.146490
     :                 :                  :
iter_number =   12 norm_grad = 0.000049 fun_val = 2.000000
iter_number =   13 norm_grad = 0.000000 fun_val = 2.000000
```

Newton's method, on the other hand, diverges:

```
>> pure_newton(f,g,h,[10;10],1e-8);
iter=   1 f(x)=2000.0009999997   ← ≈ 2·10^3 = 2000
iter=   2 f(x)=1999999999.9999990000
iter=   3 f(x)=19999999999999973000000000.0000000
iter=   4 f(x)=19999999999999923000000000000000000....
iter=   5 f(x)=       Inf
```

Note that it is essentially a 1-D problem.

$$f(x)=\sqrt{1+x^2}\approx\begin{cases}x, & x\gg 1\\ 1+\frac{1}{2}x^2, & x\ll 1\end{cases}$$

$$\begin{aligned}\Phi(\alpha)&=f(x_0-\alpha f'(x_0)^{-1}f'(x_0))\\&=\left[1+(x_0-\alpha x_0(1+x_0^2))^2\right]^{1/2}\end{aligned}$$

$$\operatorname*{argmin}_{\alpha}\Phi(\alpha)=1/(1+x_0^2)\begin{cases}\ll 1 & x_0\gg 1\\ \approx 1 & x_0\ll 1\end{cases}$$

$$\Phi(1)=(1+x_0^6)^{\frac{1}{2}}\gg f(x_0)\quad\text{when } x_0\gg 1.$$

$\approx x_0^3$ $\approx x_0$

The Newton dir. is a descent dir even when $\|x_0\|\gg 1$ (as $\nabla^2 f(x_0)\succ 0$), so it is okay to use it as a search dir. But it is not okay to use the "natural step size" 1.

In this particular example, we just need to replace the "natural" unit step size by a step size chosen by, say, backtracking. When using backtracking in Newton's method, it is wise to set the initial step size $\bar{\alpha}$ to 1. When the iterates approach a local minimizer, $\bar{\alpha}=1$ will always satisfy the c-Armijo condition with $c<\frac{1}{2}$, and no backtracking will be executed.

consequence of a stronger result by Dennis and Moré. It is clear that this statement cannot hold for $c>\frac{1}{2}$ by considering $f(x)=\frac{1}{2}x^2$

Example 5.5. Continuing Example 5.4, invoking Newton's method with the same initial vector $\mathbf{x}_0 = (10,10)^T$ that caused the divergence of pure Newton's method, results in convergence after 17 iterations.

$\bar{\alpha}$ set to 1 inside the code

```
>> newton_backtracking(f,g,h,[10;10],0.5,0.5,1e-8);
iter=  1 f(x)=4.6688169339
iter=  2 f(x)=2.4101973721
iter=  3 f(x)=2.0336386321
   :                :
iter= 16 f(x)=2.0000000005
iter= 17 f(x)=2.0000000000
```

(0.5 = c, 0.5 = ρ)

see [Beck] sec5.2 for the code.

Ex: Check this last claim carefully. $\left[\phi_{x_0}(\alpha)=\frac{1}{2}(1-\alpha)^2 x_0\right]$

For non-convex objectives, $\nabla^2 f(x)$ is not always positive definite, the Newton dir. is not always a descent direction. In this case, there is a whole development of Newton's method with Hessian modification that goes like this:

Algorithm 3.2 (Line Search Newton with Modification).

Given initial point x_0;

for $k = 0, 1, 2, \ldots$

Factorize the matrix $B_k = \nabla^2 f(x_k) + E_k$, where $E_k = 0$ if $\nabla^2 f(x_k)$ is sufficiently positive definite; otherwise, E_k is chosen to ensure that B_k is sufficiently positive definite;

Solve $B_k p_k = -\nabla f(x_k)$;

Set $x_{k+1} \leftarrow x_k + \alpha_k p_k$, where α_k satisfies the Wolfe, Goldstein, or Armijo backtracking conditions;

end

See N&W, sec 3.4.
Beck, sec 5.3.

Quasi-Newton methods ...

are quite a surprising discovery. They achieve superlinear ROC while computationally more efficient than Newton's method. Converge much faster than GD in practice.

method	memory complexity	time complexity per iteration	ROC (assume strong convexity)
GD	$O(n)$	$O(n)$	Linear
Newton	$O(n^2)$	$O(n^3)$ (without sparsity in Hessian)	quadratic
Quasi-Newton	$O(n^2)$	$O(n^2)$	superlinear

Instead of $p_k^N = -\nabla^2 f_k^{-1} \nabla f_k$, a quasi-Newton search direction is of the form

$$p_k = -B_k^{-1} \nabla f_k \quad \text{with } B_k \text{ satisfying the secant condition}$$

$$B_k(x_k - x_{k-1}) = \nabla f_k - \nabla f_{k-1}$$

E.g. in the BFGS method (to be derived): $s_k = x_{k+1} - x_k$, $y_k = \nabla f_{k+1} - \nabla f_k$

$$B_{k+1}^{-1} = (I - \rho_k s_k y_k^T) B_k^{-1} (I - \rho_k y_k s_k^T) + \rho_k s_k s_k^T \quad (\rho_k = 1/y_k^T s_k)$$

Unlike Newton's method, no matrix inversion needed in each iteration!

Ex: I once hired a bright undergraduate student who implemented the BFGS method by coding up the above formula directly in Matlab:

```
rho = (y'*s)^(-1);
H = (eye(n) - rho*s*y')*H*(eye(n) - rho*y*s') + rho*s*s';
```

This is a $O(n^3)$ implementation.

Can you rewrite his code so that it runs in $O(n^2)$ time?

The fundamental (and somewhat surprising) facts are summarized in the following results:

The proofs are elementary but tricky; they require famalarizing ourselves with the properties of sequences that converge superlinearly.

In this context, assume $x_k \neq x^*$ $\forall$ large enough k.

Lemma: Let $\{x_k\} \subset \mathbb{R}^n$ converge superlinearly to x^* (i.e. $\|x_{k+1} - x^*\| / \|x_k - x^*\| \to 0$).
Then

$$\lim_{k\to\infty} \|x_{k+1} - x_k\| / \|x_k - x^*\| = 1.$$

Proof:

Just note that

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} \geq \left| \frac{\|x_{k+1} - x_k\|}{\|x_k - x^*\|} - \frac{\|x_k - x^*\|}{\|x_k - x^*\|} \right| . \qquad (\|u+v\| \geq |\|u\| - \|v\||)$$

Q.E.D.

(i) The converse is not true. (E.g. $x_{2k-1} = 1/k!$, $x_{2k} = 2x_{2k-1}$.)

(ii) If we know that a method produces $\{x_k\}$ that converge to x^* superlinearly, this result means we can use $\|x_{k+1} - x_k\|$ — computable from the iterates — to approximate the (uncomputable) error $\|x_k - x^*\|$ in a practical setting.

(I) **Thm** (Dennis - Moré 1974)

Let $F: \mathbb{R}^n \to \mathbb{R}^n$ be C^1 in some open neighborhood D of x^*, at which $F(x^*) = 0$ and $DF(x^*)$ non-singular.

Let $\{B_k\}$ be a sequence of non-singular matrices and suppose that $\exists x_0 \in D$ s.t. the sequence $\{x_k\}$ defined by $x_{k+1} = x_k \underbrace{- B_k^{-1} F(x_k)}_{+ p_k}$ remains in D and converges to x^*.

we saw how to justify it rigorously for Newton's method for x_0 close enough to x^*

Then $\{x_k\}$ converges to x^* superlinearly

$$\Leftrightarrow$$

$$\lim_{k \to \infty} \frac{\| [B_k - DF(x^*)] p_k \|}{\| p_k \|} = 0 \qquad (*)$$

[Assume we are not so lucky that the method solves the system $F(x) = 0$ in a finite # of iterations. This means $F(x_k) \neq 0 \ \forall k$, which also means $p_k \neq 0 \ \forall k$, and there is never a division by 0 in $(*)$.]

Note: $B_k - DF(x_k) \to 0 \quad \not\Rightarrow \quad (*)$.

Proof : ($\Leftarrow$) Assume (*) holds.

$$[B_k - DF(x^*)]\, p_k = -F(x_k) - DF(x^*)(x_{k+1} - x_k)$$

$$= \underbrace{[F(x_{k+1}) - F(x_k)}_{= DF(n_k)(x_{k+1} - x_k)} - DF(x^*)(x_{k+1} - x_k)] - F(x_{k+1}) \quad - (1)$$

$$= [DF(n_k) - DF(x^*)](x_{k+1} - x_k)$$

So

$$\frac{\|[B_k - DF(x^*)]\, p_k\|}{\|p_k\|} \geq -\frac{\|[DF(n_k) - DF(x^*)]\, p_k\|}{\|p_k\|} + \frac{\|F(x_{k+1})\|}{\|p_k\|}$$

$\|u - v\| \geq \|u\| - \|v\|$

$\downarrow$ by (*) $\to 0$

$\downarrow$ by continuity of $DF(\cdot)$ $\to 0$

So $\|F(x_{k+1})\| / \|p_k\| \to 0$ as $k \to \infty$. $\quad - (2)$

Since $F(x^*) = 0$ and $DF(x^*)$ is non-singular,

$$\|F(x_{k+1})\| = \|F(x_{k+1}) - F(x^*)\| = \|DF(n'_k)(x_{k+1} - x^*)\| \geq \beta \|x_{k+1} - x^*\|$$

for some $\beta > 0$.

Therefore,

$$\frac{\|F(x_{k+1})\|}{\|p_k\|} \geq \frac{\beta \|x_{k+1} - x^*\|}{\|x_{k+1} - x^*\| + \|x^* - x_k\|} = \frac{\beta \rho_k}{1 + \rho_k}$$

where $\rho_k = \dfrac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|}$.

$\|p_k\| = x_{k+1} - x_k = x_{k+1} - x^* + x^* - x_k$

This and ② $\Rightarrow$ $\rho_k/(1+\rho_k) \to 0$ $\Rightarrow$ $\rho_k \to 0$, as desired.

($\Leftarrow$) Assume $x_k \to x^*$ superlinearly, $F(x^*) = 0$.

Since

$$\frac{\|F(x_{k+1})\|}{\|x_{k+1} - x_k\|} = \frac{\|F(x_{k+1}) - F(x^*)\|}{\|x_k - x^*\|} \cdot \frac{\|x_k - x^*\|}{\|x_{k+1} - x_k\|}$$

$$= \frac{\|DF(\eta_k)(x_{k+1} - x^*)\|}{\|x_k - x^*\|}$$

$\to 0$ by superlinear convergence

$\to 1$ by lemma

This means ② holds. Then by ①, (*) holds. Q.E.D.

Note: Apply this result to $f: \mathbb{R}^n \to \mathbb{R}$ with $F = \nabla f$, we have Theorem 3.7 in N&W (the proof is incomplete there.)

Just like the damped Newton's method, in practical quasi-Newton methods, the new iterate is updated as

$$x_{k+1} = x_k + \alpha_k p_k$$

with the step size selected to guarantee a sufficient decrease in the objective value. We expect $\alpha_k = 1$ for large k so as to obtain superlinear convergence.

For a reason we shall see later, the construction of quasi-Newton methods works well in tandem with the Wolfe conditions. (This is yet another surprise of the subject, as the two seem unrelated at first glance.)

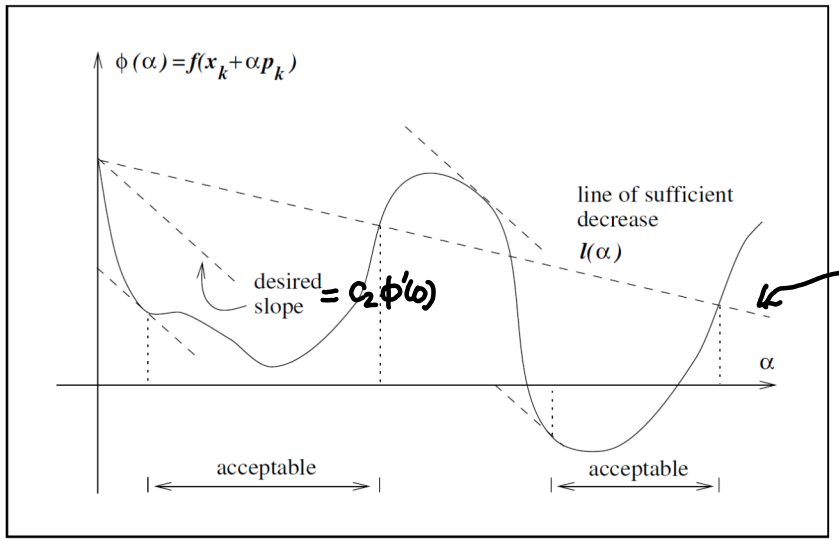


Figure 3.5 Step lengths satisfying the Wolfe conditions.

$l(\alpha) = \phi(0) + c_1\alpha\phi'(0)$
[slope $= c_1\phi'(0)$]

$0 < c_1 < c_2 < 1$

Wolfe conditions ask for a step size α that satisfies

(i) $\phi(\alpha) \leq l(\alpha)$ and

(ii) $\phi'(\alpha) \geq c_2\phi'(0)$

See N&W section 3.5 for algorithm.

(II) Thm (Dennis - Moré 1977)

Let $f:\mathbb{R}^n \to \mathbb{R}$ be C^2 in an open set D. Consider iteration $x_{k+1} = x_k + \alpha_k p_k$, where p_k is a descent direction and α_k satisfies the Wolfe conditions with $c_1 < \frac{1}{2}$. If $\{x_k\} \to x^* \in D$ at which $\nabla^2 f(x^*) > 0$ and

$$\lim_{k\to\infty} \|\nabla f_k + \nabla^2 f_k p_k\| / \|p_k\| = 0, \quad -(**)$$

$$\left[\Leftrightarrow p_k - p_k^N = o(\|p_k\|)\right]$$

then

(i) $\exists\, k_0 \geq 0$ s.t. $\lambda_k = 1$ is admissible for the Wolfe's conditions $\forall k > k_0$, and

(ii) if $\alpha_k = 1$ for all $k > k_0$, $\nabla f(x^*) = 0$ and $\{x_k\} \to x^*$ superlinearly.

Proof of (ii): if $\alpha_k = 1$ for large k, $p_k = x_{k+1} - x_k \to 0$ as $k \to \infty$, $(**)$ shows $\|\nabla f_k\| \to 0$. By continuity, $\nabla f(x^*) = \lim_{k\to\infty} \nabla f(x_k) = 0$.

We can always write $p_k = -B_k^{-1} \nabla f_k$ for some invertible matrix B_k. Then $(**)$ is equivalent to

$$\lim_{k\to\infty} \|[B_k - \nabla^2 f_k] p_k\| / \|p_k\|.$$

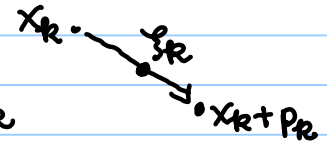
Superlinear convergence then follows from the previous theorem.

Proof of (i):

For the c_1-Armijo condition, our goal is to show that

$$f(x_k + p_k) \leq f_k + c_1 \langle \nabla f_k, p_k \rangle \quad \text{for large enough } k.$$

$$f(x_k + p_k) = f_k + \langle \nabla f_k, p_k \rangle + \tfrac{1}{2} p_k^T \nabla^2 f(z_k) p_k$$



so $f(x_k + p_k) - f_k - \tfrac{1}{2}\langle \nabla f_k, p_k \rangle = \tfrac{1}{2}\langle \nabla f_k, p_k \rangle + \tfrac{1}{2} p_k^T \nabla^2 f(z_k) p_k$

$$= \tfrac{1}{2}\langle \nabla f_k + \nabla^2 f(z_k) p_k, p_k \rangle$$

$$f(x_k + p_k) - f_k - c_1 \langle \nabla f_k, p_k \rangle = \tfrac{1}{2}\langle \nabla f_k + \nabla^2 f(z_k) p_k, p_k \rangle + (\tfrac{1}{2} - c_1)\langle \nabla f_k, p_k \rangle$$

$$= \tfrac{1}{2}\langle \nabla f_k + \nabla^2 f(x_k) p_k, p_k \rangle + (\tfrac{1}{2} - c_1)\langle \nabla f_k, p_k \rangle = T_3$$

$$+ \langle [\nabla^2 f(z_k) - \nabla^2 f(x_k)] p_k, p_k \rangle = T_2$$

$T_1 = 0$

We show below that when k is large T_3 is the dominant term and it is negative. Specifically, $T_3 \leq$ (some neg. const.) $\|p_k\|^2$, $T_1, T_2 = o(\|p_k\|^2)$, which means $T_1 + T_2 + T_3 < 0$ for large k.

First, note that $-\langle \nabla f_k, p_k \rangle \geq \eta \|p_k\|^2$ for large k, (***)
because

$$-\langle \nabla f_k, p_k \rangle = \langle \nabla^2 f_k p_k, p_k \rangle - \langle \nabla^2 f_k p_k + \nabla f_k, p_k \rangle .$$

$\geq 0.99\,\lambda_{\min}(\nabla^2 f(x^*))\|p_k\|^2$ (say) by continuity $\qquad o(\|p_k\|^2)$ by (**) $\qquad k \to \infty$.

So, $T_3 \leq \underbrace{-(\tfrac{1}{2} - c_1)\eta}_{<0 \text{ as } c_1 \in (0,\frac{1}{2})} \|p_k\|^2$.

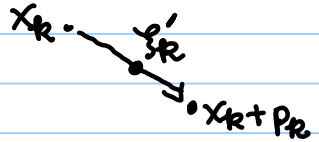
$|T_2| \leq \|p_k\| \underbrace{\|\nabla^2 f(z_k) - \nabla^2 f(x_k)\|}_{o(1)} \|p_k\| = o(\|p_k\|^2)$,

$|T_1| \leq \tfrac{1}{2}\|\nabla f_k + \nabla^2 f_k p_k\| \|p_k\| = o(\|p_k\|^2)$ by (**).

For the second Wolfe condition, our goal is to show

$$\langle \nabla f(x_k + p_k), p_k \rangle \geq c_2 \langle \nabla f(x_k), p_k \rangle \quad \text{for large } k.$$

$$\langle \nabla f(x_k + p_k), p_k \rangle = \langle \nabla f_k + \nabla^2 f(z_k') p_k, p_k \rangle$$
$$= \underbrace{\langle \nabla f_k + \nabla^2 f_k p_k, p_k \rangle}_{=T_1'} + \underbrace{\langle [\nabla^2 f(z_k') - \nabla^2 f(x_k)] p_k, p_k \rangle}_{=T_2'}$$



$|\langle \nabla f(x_k + p_k), p_k \rangle| \leq |T_1'| + |T_2'|$
$= o(\|p_k\|^2) + o(\|p_k\|^2)$ (as before)
$\leq$ (any positive const.) $\|p_k\|^2$ k large

so $|\langle \nabla f(x_k + p_k), p_k \rangle| \leq \tilde{c}_2 \eta \|p_k\|^2 \overset{(***)}{\leq} c_2 |\langle f_k, p_k \rangle|$.

We have shown more than we claimed: $\alpha_k = 1$ satisfies the strong Wolfe conditions for large k. Q.E.D.

Note: This is basically Theorem 3.6 in N&W, except that N&W claim without proof that this 1977 result by Dennis and Moré holds true also for $c_1 = 1/2$.

We know for sure that the result does not hold true for $c_1 > 1/2$, and we proved above that it is true for $c_1 < 1/2$.

In practice, quasi-Newton solvers choose c_1 to be way smaller than $1/2$.