Transition to local convergence (p 46)

If: i) $$\lim_{k\to\infty} \frac{\|\nabla^2 f(x_k) p_k + \nabla f(x_k)\|}{\|p_k\|} = 0 \qquad (*)$$

ii) $x_k \to x^*$ at which 2nd order suff optim. cdt° hold

iii) t_k chosen according to wolfe, Goldstein

$\Rightarrow$ $\exists \bar{k} \in \mathbb{N}$ s.t. $\forall k > \bar{k}$ $t_k = 1$.

what it means: line search method behaves like Newton method asymptotically, regardless of starting point.

$\Rightarrow$ Globally convergent method.

$(*)$ is satisfied if p_k is computed as the solution:

$$(\nabla^2 f(x_k) + \mu_k I)\, p_k = -\nabla f(x_k) \qquad \begin{matrix} \mu_k \geq 0 \\ \mu_k \to 0 \end{matrix}$$

μ_k Guarantees Hessian is pos def but this can be done in various ways by playing with Cholesky facto of ~~Hessian~~.

Cholesky: $A = LL^T$ (LU facto for symm matrices, very robust facto and reveals whether a matrix is pos def or not).

Gradient method.

$p_k = -\nabla f_k$: $(*)$ is not satisfied, will $t_k = 1$ be acceptable eventually?

(SDC):

$$f(x_k - t\nabla f_k) \leq f_k - c_1 t \|\nabla f_k\|_2^2$$

Taylor $$f(x_k - t\nabla f_k) = f_k - t\|\nabla f_k\|_2^2 + \frac{1}{2} t^2 \nabla f_k^T \nabla^2 f(x_k - z\nabla f_k)$$

for $z \in [0,t]$

Combining:

$$\frac{1}{2}t^2 \nabla f_k^T \nabla^2 f(x_k - \tau \nabla f_k)\nabla f_k \le t(1-c_1)\|\nabla f_k\|_2^2$$

$$\Rightarrow t \le 2(1-c_1)\frac{\|\nabla f_k\|_2^2}{\underbrace{\nabla f_k^T \nabla^2 f(x_k - \tau \nabla f_k)\nabla f_k}_{\sim \text{Rayleigh quot}}}$$

$$\lesssim \frac{1}{\lambda_{min}(\nabla^2 f_k)}$$

may or may not be <1 depending on scaling of f!

Scaling: $f(x)$

$$\tilde f(\tilde x) = f(\underset{\uparrow}{D}\tilde x)$$

diag scaling matrix (units km, m etc...)

$$\nabla \tilde f(\tilde x) = D\nabla f(D\tilde x)$$

$$\nabla^2 \tilde f(\tilde x) = D\nabla^2 f(D\tilde x)D$$

Grad method in new var.

$$D\tilde x_{k+1} = D\tilde x_k - D\tilde t\,\nabla \tilde f(\tilde x)$$

$$= D\tilde x_k - \tilde t\,\underline{D^2}\nabla f(D\tilde x)$$

does ~~not~~ scale like variable → not scale invariant.

Newton's method is scale invariant: (does not care about formulation (~ units))

$$x_{k+1} = D\tilde x_{k+1} = D\tilde x_k - \cancel{D D^{-1}}[\nabla^2 f(D\tilde x_k)]^{-1}\cancel{D^{-1} D}\nabla f(D\tilde x_k)$$

Quasi-Newton methods (Chap 6)

Model $f(x_k+p)$ by $m_k(p) = f_k + \nabla f_k^T p + \frac{1}{2} p^T B_k p$

- B_k = symm matrix approximating $\nabla^2 f(x_k)$ in some sense.

e.g. $B_k = \nabla^2 f(x)$ Newton's method

$B_k = I$ Steepest descent

Idea: replace $\nabla^2 f(x)$ by cheaper to compute B_k (or B_k^{-1})

If B_k is symm pos def $p_k = -B_k^{-1} \nabla f_k$ is descent direction.

If we use line search:

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

Need replacement B_{k+1} of Hessian $\nabla^2 f(x_{k+1})$ s.t.

$$m_{k+1}(p) = f_{k+1} + \nabla f_{k+1}^T p + \frac{1}{2} p^T B_{k+1} p$$

with:

$$\begin{cases} m_{k+1}(p) \approx f(x_{k+1}+p) \\ \nabla m_{k+1}(\underbrace{-\alpha_k p_k}_{\text{goes back to } x_k}) = \nabla f(x_k) \end{cases} \quad (*)$$

since $\nabla m_{k+1}(p) = \nabla f_{k+1} + B_{k+1} p$

$$(*) \Rightarrow \boxed{B_{k+1} \underbrace{\alpha_k p_k}_{s_k} = \underbrace{\nabla f_{k+1} - \nabla f_k}_{y_k}}$$

SECANT CONDITION (1)

(n equations)

What else do we need B_{k+1} to be?

② B_{k+1} symm $\left(\sim \frac{n^2}{2} \text{ equations}\right)$

③ B_{k+1} pos def $(\sim n \text{ eq})$

① & ③ $\Rightarrow$ $s_k^T B_{k+1} s_k = \boxed{s_k^T y_k > 0}$

"Curvature condition"

However ① & ② & ③ do not determine B_{k+1} uniquely.

$\Rightarrow$ Find B_{k+1} as:

(☆) $\min \|B - B_k\|_*$
s.t.
$B^T = B$
$B s_k = y_k$

this does not guarantee B_{k+1} pos def, but sol turns out to be pos def.

What matrix norm?

$$\|B - B_k\|_* = \|\bar{G}^{-1/2}(B - B_k)\bar{G}^{-1/2}\|_F .$$

weighted Frobenius norm

where

$$\bar{G} = \int_0^1 D^2 f(x_k + t \overbrace{\alpha_k p_k}^{s_k})\, dt$$

averaged Hessian

Recall: $\|A\|_F^2 = \sum_{i,j=1}^n |a_{ij}|^2$

Note: If 2nd order suff cond are satisfied at x^*

$D^2 f(x^*) > 0$

and also $D^2 f(x) > 0$ for x sufficiently close to x^*.

thus if $x_k + t\alpha_k p_k$ is close to x^* $Df(-) > 0$

$\Rightarrow \bar{G} > 0 \Rightarrow \bar{G}$ is inversible and $\bar{G}^{-1}$ is pos def

So there is matrix $\bar{G}^{-1/2}$ s.t. $\bar{G}^{-1} = \bar{G}^{-1/2}\bar{G}^{-1/2}$

("matrix square root")

Recall if A = symm pos def w/ $A = \sum_{i=1}^{n} \lambda_i u_i u_i^T$

$$A^{\frac{1}{2}} = \sum_{i=1}^{n} \lambda_i^{\frac{1}{2}} u_i u_i^T$$

$\bar{G}$ is not fortuitous choice as it makes norm $\|\cdot\|_*$ scale invariant and $\bar{G} s_k = y_k$ (M.V.T.)

Solution to ($\bigstar$) can be found explicitly:

$$\boxed{B_{k+1} = \left(I - \rho_k y_k s_k^T\right) B_k \left(I - \rho_k s_k y_k^T\right) + \rho_k y_k y_k^T}$$

where $\rho_k = \dfrac{1}{s_k^T y_k}$

recall outer prod $u v^T = \left(u v_1 \,|\, u v_2 \,|\, \dots \,|\, u v_n\right)$

$$= \begin{pmatrix} u_1 v^T \\ u_2 v^T \\ \vdots \\ u_n v^T \end{pmatrix} = \text{rank one matrix}$$

It can be shown that:

- B_{k+1} is symm if B_k is symm
- $B_{k+1} s_k = y_k$ (Secant condition)
- B_{k+1} pos def provided B_k is pos def and $y_k^T s_k > 0$ ("curvature condition")

(see p 141)

But we need B_{k+1}^{-1}, to compute it efficiently we use Sherman-Morrison-Woodbury (SMW) formula:

If $A \in \mathbb{R}^{n \times n}$ invertible, $U, V \in \mathbb{R}^{n \times k}$ with $\operatorname{rk}(UV^T) = k$

□ □,□

then:

$$(A+\underbrace{UV^T}_{\text{rk k update}})^{-1} = A^{-1} - A^{-1}U(I+V^TA^{-1}U)^{-1}V^TA^{-1}$$

provided $I + V^TA^{-1}U \in \mathbb{R}^{k\times k}$ is invertible

□

()

proof (sketch using Neumann series $(I-A)^{-1} = I+A+A^2+\dots$

$$(A+UV^T)^{-1} = (I + A^{-1}UV^T)^{-1}A^{-1}$$
$$= (I - A^{-1}UV^T + (A^{-1}UV^T)^2 - (\dots)^3 + \dots)A^{-1}$$
$$= A^{-1} - A^{-1}U\left[I - V^TA^{-1}U + (V^TA^{-1}U)^2 - (\ )^3 \dots\right]V$$
$$= A^{-1} - A^{-1}U(I+V^TA^{-1}U)^{-1}V^TA^{-1}.$$

If we let $H_k = B_k^{-1}$, $H_{k+1} = B_{k+1}^{-1}$ then

$$\boxed{H_{k+1} = H_k - \overset{\text{rk 1}}{\frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k}} + \overset{\text{rk 1}}{\frac{s_k s_k^T}{y_k^T s_k}}}$$

DFP update (= rk 2 update)
Davidon Fletcher Powell

note: at each update we only need to store two vectors: y_k and s_k

If $H_0 = \alpha I$ easy to find H_{k+1} from history of $\{y_k, s_k\}$, with cheap operations.

Problem: each iteration becomes more expensive

⇝ only keep last p pairs $\{y_k, s_k\}$ (forget some info to get cheap method ⇝ Limited Memory QN

Instead of finding B_{k+1} close to B_k and then taking inverse why not find H_{k+1} close to H_k ?

Find H_{k+1} as: $\min \|H - H_k\|$

$H = H^T$

$H y_k = s_k$

$B s_k = y_k$

$\Rightarrow s_k = H y_k$

Solution is:

$$\boxed{H_{k+1} = (I - \rho_k s_k y_k^T) H_k (I - \rho_k y_k s_k^T) + \rho_k s_k s_k^T}$$

where $\rho_k = \frac{1}{s_k^T y_k}$ as before   BFGS

$$\left[\text{SMW} \Rightarrow B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{s_k^T y_k} \right]$$

BFGS: Broydon Fletcher Goldfarb & Shanno.

⚠ Note we have assumed all along that $y_k^T s_k > 0$.

1) Use step size with Wolfe (or SW) cdtn:

this guarantees that: $\nabla f_{k+1}^T s_k \geq c_2 \nabla f_k^T s_k$ $0 < c_2 <$

$$\Rightarrow \quad y_k^T s_k = (\nabla f_{k+1} - \nabla f_k)^T s_k$$

$$\geq \underbrace{(c_2 - 1)}_{<0} \underbrace{\nabla f_k^T s_k}_{<0} > 0$$

2) If $y_k^T s_k \leq 0$ set $B_{k+1} = B_k$ as close to x^* $y_k^T s_k > 0$ should be satisfied.

Convergence result

Let $f: \mathbb{R}^n \to \mathbb{R}$ be twice cont. diff'ble w/ Lipschitz cont. 2nd der. Let x_* be a local sol s.t. 2nd order suff optim. cdt° are satisfied (i.e. $\nabla^2 f(x^*)$ pos def), then there exists $\varepsilon > 0$, $\delta > 0$ s.t.

if $\|x_0 - x_*\| < \varepsilon$ and $\|B_0 - \nabla^2 f(x_*)\| < \delta$ $\Rightarrow$ $x_k \to x^*$ with $\lim_{k\to\infty} \frac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|} = 0$

q-superlinear convergenc

other similar methods: symm rank one update SR1 and also Broyden class method.

Most popular and useful: BFGS & L-BFGS.