Transition to local convergence \((p \ 46)\)

\[
\text{If:} \quad \lim_{{k \to \infty}} \frac{\|D^2 f(x_k) p_k + \nabla f(x_k)\|}{\|p_k\|} = 0 \quad (*)
\]

\(\Rightarrow \) \(x_k \to x^*\) at which 2nd order optimality conditions hold

\(\Rightarrow\) the chosen according to we get, Goldstein

\(\Rightarrow\) \(\exists\) \( u_k \geq 0 \) \( u_k = 1 \)

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

\(\Rightarrow\) globally convergent method.

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

\[
(D^2 f(x_k) + \mu_k I) p_k = -\nabla f(x_k) \quad \mu_k \geq 0
\]

\(\mu_k\) guarantees Hessian is positive definite, but this can be done in various ways by playing with Cholesky factor of Hessian.

Cholesky: \( A = L L^T \) (LU factor for symm matrices, very robust factor and reveals whether a matrix is posi def or not)

**Gradient method**

\[
p_k = -\nabla f_k \quad (\ast) \text{ is not satisfied, recall } t_k \to 0 \text{ be acceptable eventually?}
\]

**SDC**:

\[
f(x_k - t\nabla f_k) \leq f_k - c_1 t \|\nabla f_k\|^2
\]

Taylor:

\[
f(x_k - t\nabla f_k) = f_k - t\|\nabla f_k\|^2 + \frac{t^2}{2} \nabla^2 f_k^T (x_k - 2t\nabla f_k)
\]

for \(c_1 \geq 0\)
Combining:

$$
\frac{1}{2} t^2 \mathbf{Df_k}^T \mathbf{Df} (x_k - 2 \mathbf{Df_k}) \mathbf{Df_k} \leq t (1 - c_1) \| \mathbf{Df_k} \|^2
$$

$$
\Rightarrow t \leq 2 (1 - c_1) \frac{\| \mathbf{Df_k} \|^2}{\mathbf{Df_k}^T \mathbf{Df} (x_k - 2 \mathbf{Df_k}) \mathbf{Df_k}}
$$

Rayleigh quot

$$
\leq \frac{1}{\text{Amn} (\mathbf{Df_k})}
$$

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

**Scaling**

\( f(x) \)

\( \hat{f}(\tilde{x}) = f(D\tilde{x}) \)

def. scaling matrix

(units km, m etc.)

\( D \hat{f}(\tilde{x}) = D \mathbf{Df} (D\tilde{x}) \)

\( D^2 \hat{f}(\tilde{x}) = D \mathbf{Df} (D\tilde{x}) D \)

**Grad method**

in new var.

\( D \tilde{x}_{k+1} = D \tilde{x}_k - D \tilde{f} (D\tilde{x}) \)

\( = D \tilde{x}_k - D \tilde{E} D^2 \mathbf{Df} (D\tilde{x}) \)

---

does not scale like var

---

Newton's method so scale invariant (does not care about formula)

\( x_{k+1} = D \tilde{x}_{k+1} = D \tilde{x}_k - D^{-1} [D^2 \mathbf{Df} (D\tilde{x})]^{-1} D^{-T} D \mathbf{Df} (D\tilde{x}) \)
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 \( \nabla^2 f(x) \) 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 posdef \( \nabla f_k \) 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:
\[ \begin{aligned}
\quad m_{k+1}(p) &\approx f(x_{k+1} + p) \\
\nabla m_{k+1}(-\alpha_k p_k) &= \nabla f(x_k) \quad \text{(*)}
\end{aligned} \]

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

\( (*) \Rightarrow \)
\[ B_{k+1} \alpha_k p_k = \frac{\nabla f_{k+1} - \nabla f_k}{\alpha_k} \]

Secant condition (in equations)
What else do we need $B_{k+1}$ to be?

2. $B_{k+1}$ symmetric ($\sim \frac{m^2}{2}$ equations)
3. $B_{k+1}$ positive definite ($\sim m$ eq)

1 & 3 $\implies$ \[ \forall k \quad B_{k+1} y_k = \begin{pmatrix} 2k^T y_k \end{pmatrix} > 0 \]

"Curvature condition"

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

=) Find $B_{k+1}$ as:

\[ \begin{array}{l}
\text{(A)} \\
\min_{B, B_k} \| B - B_k \|_* \\
\text{st.} \\
B = B_k \\
B_{k+1} = y_k
\end{array} \]

this does not guarantee $B_{k+1}$ positive definite, but so it turns out to be positive definite.

What matrix norm?

\[ \| B - B_k \|_* = \| G^{-\frac{1}{2}} (B - B_k) G^{-\frac{1}{2}} \|_F \]

where

\[ G = \int_0^1 D^2 f(x_k + \tau k p_k) \, d\tau \]

\[ \text{average Hessian} \]

Note: If 2nd order suffcond are satisfied at $x^*$

\[ D^2 f(x^*) > 0 \]

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

Thus if $x_k + \tau k p_k$ is close to $x^*$ $D f (-) > 0$.

\[ \implies B > 0 \implies B \text{ is invertible and } B^{-1} \text{ is pos def} \]

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

\[ G^{-\frac{1}{2}} = G^{-\frac{1}{2}} G^{-\frac{1}{2}} \]

("matrix square roots")
Recall if $A = \text{symm} \ p.d \ \forall i \ \ A = \sum_{i=1}^{n} \lambda_i \ e_i e_i^T$

$A^{-1} = \sum_{i=1}^{n} \lambda_i^{-1} e_i e_i^T$

$G$ is not fortuitous choice as it makes norm $\|.\|^*$ scale invariant, and $G_A = y_k$ (M.V.T.)

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

$B_{k+1} = (I - \gamma_k y_k y_k^T) B_k (I - \gamma_k y_k y_k^T) + \gamma_k y_k y_k^T$

where $\gamma_k = \frac{1}{\lambda_k y_k}$

Recall outer prod $\mu \nu^T = \begin{pmatrix} \mu \nu_1 & \mu \nu_2 & \cdots & \mu \nu_n \\ \nu_1 \nu^T & \nu_2 \nu^T & \cdots & \nu_n \nu^T \end{pmatrix} = \text{rank one matrix}$

It can be shown that:

- $B_{k+1}$ is symm if $B_k$ is symm
- $B_{k+1} y_k = y_k$ (Secant condition)
- $B_{k+1}$ p.d. provided $B_k$ is p.d. and $y_k^T y_k > 0$ ("Curvature condition")

(see p. 141)

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

If $A \in \mathbb{R}^{n \times n}$ invertible, $UV \in \mathbb{R}^{n \times k}$ with $rk(UV^T) = k$
\[(A + UV^T)^{-1} = A^{-1} - A^{-1} U (I + V^T A^{-1} U)^{-1} V^T A^{-1}\]

\[\text{provided } I + V^T A^{-1} U \in \mathbb{R}^{k \times k} \text{ is invertible.} \]

\[\text{proof (sketch using Neumann series) } (I - A)^{-1} = I + A + A^2 + \ldots\]

\[(A + UV^T)^{-1} = (I + A^{-1} UV^T)^{-1} A^{-1}\]
\[= (I - A^{-1} UV^T + (A^{-1} UV^T)^2 - \ldots)^{-1} A^{-1}\]
\[= A^{-1} - A^{-1} U [I - V^T A^{-1} U + (V^T A^{-1} U)^2 - \ldots]^{-1} V^T A^{-1}\]

If we let \(H_k = B_k^{-1}\), \(H_{k+1} = B_k^{-1}\) then

\[H_{k+1} = H_k - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k} + \frac{y_k^T s_k}{y_k^T s_k}\]

\[\text{DFP update (\(=nk2\) update)}\]

\[\text{David \& Fletcher Powell}\]

\[\text{Note: at each update we only need to store two vectors: } y_k \text{ and } s_k\]

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.

\[\Rightarrow \text{only keep last } p \text{ pairs } y_k, s_k \Rightarrow \text{forget some info to get cheap method} \Rightarrow \text{Limited Memory ON}\]
Instead of finding $B_{k+1}$ close to $B_k$ and then taking its inverse why not find $H_{k+1}$ close to $H_k$?

Find $H_{k+1}$ as:

$$ \min \| H - H_k \| $$

$$ H = H_k $$

$$ H y_k = y_k $$

$$ B \Delta x = y_k $$

$$ \Rightarrow \Delta x = H y_k $$

Solution is:

$$ H_{k+1} = (I - S_k \Delta x y_k^T) H_k (I - S_k y_k y_k^T) + S_k \Delta x \Delta x^T $$

where 

$$ S_k = \frac{I}{\Delta x^T y_k} $$

as before.

BFGS: Broyden-Fletcher-Goldfarb-Shanno.

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

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

This guarantees that:

$$ \Delta x_k = c_2 \nabla f_k^T \Delta x_k $$

$$ y_k^T \Delta x = (\nabla f_{k+1} - \nabla f_k)^T \Delta x_k $$

$$ \geq \frac{(c_2 - 1)}{c_0} \nabla f_k^T \Delta x_k $$

$$ \Rightarrow \frac{(c_2 - 1)}{c_0} \nabla f_k^T \Delta x_k > 0 $$

2) If $y_k^T \Delta x \leq 0$ let $B_{k+1} = B_k$ at close to $x^*$.

$y_k^T \Delta x > 0$ should be satisfied.
Convergence result

Let $f: \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable and Lipschitz continuous. Let $x^*$ be a local solution to $f$. If $\nabla^2 f(x^*)$ is positive definite, then there exists $\varepsilon > 0$, $\delta > 0$ such that

$$
\|x_0 - x^*\| < \varepsilon \quad \Rightarrow \quad \|x_k - x^*\| \to 0 \quad \text{for} \quad \|x_k - x^*\| < \delta
$$

$q$-superlinear convergence

Other similar methods: SYMM Rank one update SR1 and also Broyden class method.

Most popular and useful: BFGS & L-BFGS.