Problems with Newton's method:

1) Calculating & storing Hessian is expensive for large problems.
2) Solving the Newton system can be expensive.
3) We only get local convergence, initial point needs to be sufficiently close to true solution.

4.2: Quasi-Newton methods
3: Globalization strategies (line search and trust region).

Line search methods

Definition (descent direction): Let $f$ be continuously differentiable, $p \in \mathbb{R}^n$ is a descent direction if $D_f(x)^T p < 0$.

For descent directions, $\exists \varepsilon > 0$ s.t. $x + tp \in \mathbb{R}^n$, $f(x + tp) < f(x)$, i.e. we can improve objective function if we look in dir $p$.

Example:
- $p = -D_f(x)$
- $p = -H^{-1}D_f(x)$ where $H$ is positive definite ($H = D^2 f(x)$ = Newton).

Plan:
- Compute descent direction $p_k$ given $x_k$.
- Compute step size $\alpha_k > 0$.
- Update $x_{k+1} = x_k + \alpha_k p_k$.

How to choose $\alpha_k$?

1) Want sufficient decrease in obj. fun.
2) Avoid unnecessary small step sizes.

4) Simple decrease: $f(x_k + \alpha_k p_k) < f(x_k)$ is not enough.

Example: could imagine method that enforces:
- $f(x_k + \alpha_k p_k) \geq f(x_k) - \frac{1}{2\alpha_k}$
- $f(x_{k+1}) \geq f(x_k) - \frac{1}{\alpha_k}$
- $f(x_0) \geq f(x_0) - \frac{1}{\alpha_0 \alpha_k}$
Sufficient decrease condition (SDC)

\[ f(x_k + t_p k) \leq f(x_k) + c_1 t_k \nabla f(x_k)^T p_k \]

where \( 0 < c_1 < 1, \quad c_1 \approx 10^{-4} \) typical.

Motivation:
\[
\begin{align*}
\Phi(t) &= f(x_k + t_p k) \\
\Phi'(t) &= \nabla f(x_k+tpk)^T p_k
\end{align*}
\]

\[ SDC \Rightarrow \Phi(t_k) \leq \Phi(0) + c_1 t_k \Phi'(0) \]

\( c_1 \) decrease is a function of decrease predicted by Taylor's theorem.

(SDC) is not enough to guarantee convergence need \( c_1 t_k \) which ensure \( tk \) is sufficient large.

A type of \( c_1 t_k \): Wolfe, Strong Wolfe, Goldstein & Backtracking
1) Wolfe conditions:
\[ f(x_k + tk\, p_k) \leq f(x_k) + C_1 tk \nabla f(x_k)^T p_k \]  
\[ \nabla f(x_k + tk\, p_k)^T p_k \geq C_2 \nabla f(x_k)^T p_k \]  
where \( 0 < C_1 < C_2 < 1 \)  
\[ (W1) = (SDC) \]  
\[ (W2) \]

2) Strong Wolfe conditions:
Wolfe conditions might through step sizes that are far away from minimizers.

\[ \lim_{t \to 0} \frac{f(x_k + tk)}{f(x_k)} = 0 \]  
\[ \leftarrow \text{minimize } f(x_k) \text{ by above} \]  
\[ \left\{ \begin{array}{l}
  f(x_k + tk\, p_k) \leq f(x_k) + C_1 tk \nabla f(x_k)^T p_k \\
  \left| \nabla f(x_k + tk\, p_k)^T p_k \right| \leq C_2 \left| \nabla f(x_k)^T p_k \right| - \phi''(0)
\end{array} \right. \]  
\[ (SW1) \]  
\[ (SW2) \]

meaning: avoid too positive slopes  
- respect to things that look like extremum.

\[ \frac{C_2 (\phi'(0))}{0} \leq \frac{\phi'(tk)}{\phi''(0)} < C_2 \phi''(0) \]  
\[ (SW2) \]
problem: requires derivative expert at initial pt \( x_k + tk\, p_k \) 
can be expensive
3) Goldstein cd^0

\[
\begin{align*}
    f(x_k + tkp_k) &\leq f(x_k) + c_1 tk \, D_k^T p_k \\
    f(x_k + tkp_k) &\geq f(x_k) + (1-c_1) tk \, D_k^T p_k
\end{align*}
\]

where \( c_1 \in (0, \frac{1}{2}) \) for two \( cd^0 \) to be simultaneous.

4) Backtracking (simple)

Let \( 0 < \beta_1 < \beta_2 < 1 \)

Set \( t^{(k)} \)

for \( i = 0 \ldots \)

if \( t^{(i)} \) satisfies (SDC) STOP

else select \( t^{(i+1)} \in [\beta_1 t^{(i)}, \beta_2 t^{(i)}] \)

\( \beta_1 = \beta_2 = \frac{1}{2} \) Armijo rule (step size gets halved each time)

Then if the step size \( tk \) satisfies the Wolfe conditions in Wolfe or Goldstein or Backtracking Then:

\[ c > 0 \text{ (independent of } x) \text{ s.t.} \]

\[ tk \geq -c \frac{D_k^T p_k}{\|p_k\|^2} \]
Then (Convergence of line search methods)
If $f$ is bounded below then:

$$
- \sum_{k=0}^{\infty} \frac{\nabla f(x_k)^T p_k}{\|\nabla f(x_k)\|^2} \|p_k\|^2 < \infty \quad \text{(convergent series)}
$$

$$
\sum_{k=0}^{\infty} \left( \frac{\nabla f(x_k)^T p_k}{\|\nabla f(x_k)\|^2} \|p_k\|^2 \right) < \infty
$$

$$
\cos^2 \theta_k \quad \theta_k = \text{angle between } \nabla f(x_k) \text{ and } p_k
$$

In order to get $\nabla f(x_k) \to 0$ we need to ensure:

$$
\cos^2 \theta_k > \gamma > 0
$$

If $p_k = -H_{k-1} \nabla f(x_k)$

$$
\frac{\nabla f(x_k)^T p_k}{\|\nabla f(x_k)\|^2} \geq \frac{1}{\lambda_{\max}(H_k)}
$$

$$
\|p_k\|^2 = \|H_{k-1} \nabla f(x_k)^T \nabla f(x_k) \| \leq \frac{1}{(\lambda_{\min}(H_k))^2}
$$

$$
\cos^2 \theta_k = \frac{\nabla f(x_k)^T p_k}{\|\nabla f(x_k)\|^2 \|p_k\|^2} \geq \left( \frac{\lambda_{\min}(H_k)}{\lambda_{\max}(H_k)} \right)^2 = \frac{1}{\operatorname{cond}(H_k)^2}
$$

need $\operatorname{cond}(H_k)^{-2} \geq \gamma$ for all $k$

$$
\operatorname{cond}(H_k) \leq \sqrt{\gamma^{-1}}
$$
Transition to local convergence \((p.46)\)

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

\(i)\) \(x_k \to x^*\) which 2nd order inf optimization holds

\(ii)\) \(f\) chosen according to Wolfe, Goldstein

\(\Rightarrow\) \(\exists\ IM\) s.t. \(A_k \geq 0\) \(\Rightarrow k=1\)

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

\(\Rightarrow\) Globally convergent method.

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

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

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

Cholesky: \(A = LL^T\) \((L, U \text{ factors for symm matrices, very robust factor})\)