Problems with Newton's method

1) Calculating & Storing Hessian is expensive for large problems
2) solving the Newton systems can be expensive
3) We only get local convergence: initial point needs to be suff close to true solution!

1,2: Quasi Newton methods
3: Globalization strategies (line search and trust region).

Line search methods

Def (descent direction) Let f cont diff'ble, $p \in \mathbb{R}^n$ is a descent dir if

$$\nabla f(x)^T p < 0$$

For descent directions: $\exists \bar{t} > 0 \quad \forall t \in [0, \bar{t}] \quad f(x+tp) \leq f(x)$

i.e. we can improve objective function if we look in dir p.

Examples:

$p = -\nabla f(x)$

$p = -H^{-1} \nabla f(x)$ where H pos def ($H = \nabla^2 f(x) \Rightarrow$ Newton)

Plan:

- Compute descent direction p_k given x_k
- compute step size $t_k > 0$
- update $x_{k+1} = x_k + t_k p_k$

How to choose t_k?

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

1): Simple decrease: $f(x_k + t_k p_k) < f(x_k)$ is not enough:

Example: could imagine method that enforces:

$$f(x_k + t_k p_k) \geq f(x_k) - \frac{1}{2^k}$$

$$f(x_{k+1}) \geq f(x_k) - \frac{1}{2^k} \cdots \geq f(x_0) - \underbrace{\sum_{j=0}^{k} \frac{1}{2^j}}_{\leq 2}$$

such a method is not good as it decreases f values by at most 2!

Sufficient decrease condition (SDC)

$$f(x_k + t_k p_k) \leq f(x_k) + c_1 t_k \overbrace{\nabla f(x_k)^T p_k}^{<0}$$

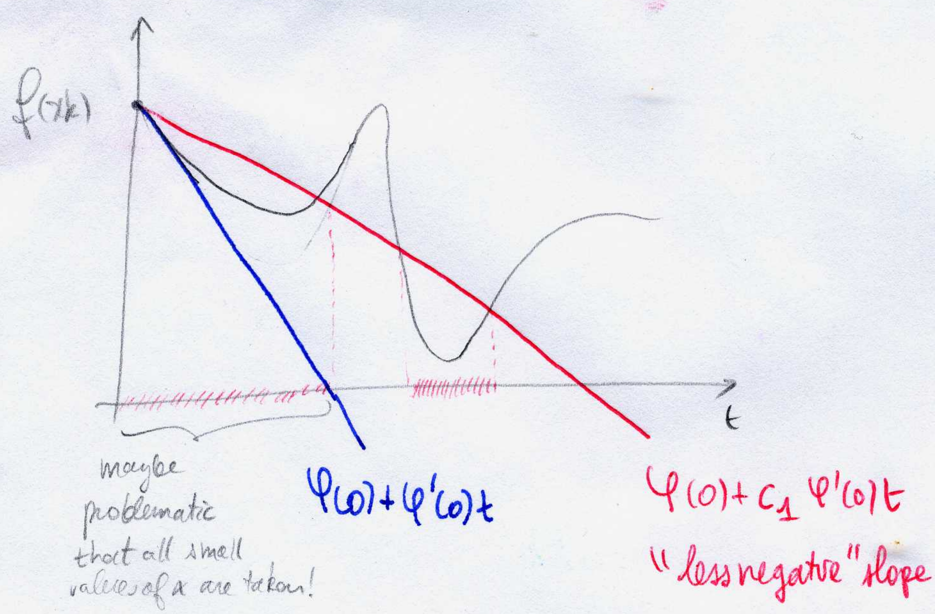
where $\quad 0 < c_1 < 1, \quad c_1 \sim 10^{-4}$ typical.

motivation: $\varphi(t) = f(x_k + t p_k) \Rightarrow \varphi(0) = f(x_k)$

$\varphi'(t) = \nabla f(x_k + t p_k)^T p_k$

(SDC) $\Leftrightarrow \quad \varphi(t_k) \leq \varphi(0) + c_1 t_k \varphi'(0)$

$\Leftrightarrow$ decrease is a fraction of decrease predicted by Taylors theorem.



(SDC) is not enough to guarantee convergence: need cdt° which ensure t_k is suff large.

4 type of cdt°: Wolfe, Strong Wolfe, Goldstein & Backtracking

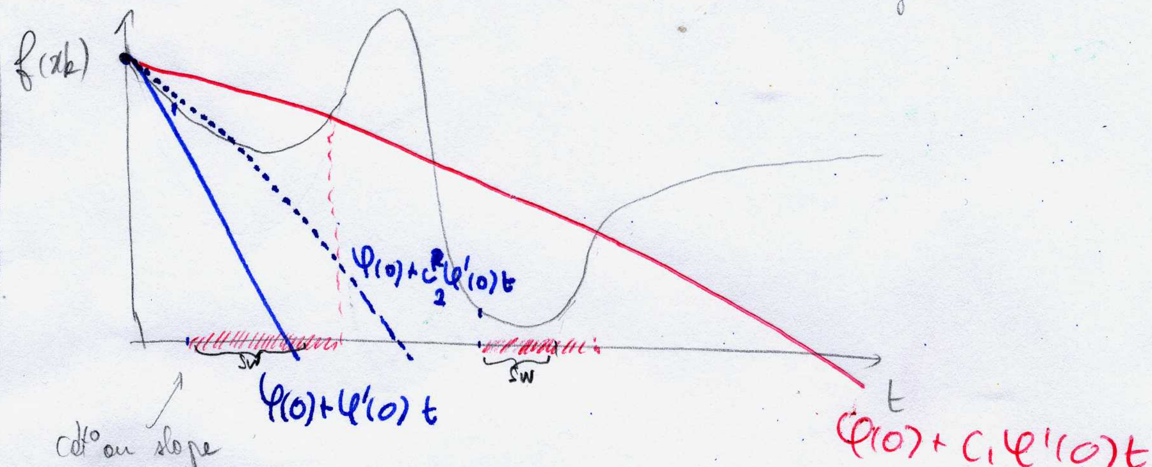
1) Wolfe conditions:

$$f(x_k+t_k p_k) \le f(x_k) + c_1 t_k \nabla f(x_k)^T p_k \qquad (W1) = (SDC)$$

$$\nabla f(x_k+t_k p_k)^T p_k \geq c_2 \underbrace{\nabla f(x_k)^T p_k}_{\text{cond on tangent}} \quad \text{where } 0<c_1<c_2<1 \qquad (W2)$$

$c_2 \sim 0.9$ typical

$(W2) \Leftrightarrow \varphi'(t_k) \geq c_2 \varphi'(0)$



cdt° on slope removed stepsizes that were too small

2) Strong Wolfe conditions

Wolfe conditions may let through step sizes that are far away from minimizers

$\Rightarrow$ limit $\varphi'(t_k)$ by above

$$\begin{cases} f(x_k+t_k p_k) \le f(x_k) + c_1 t_k \nabla f_k^T p_k & (SW1) \\ |\nabla f(x_k+t_k p_k)^T p_k| \le c_2 |\underbrace{\nabla f_k^T p_k}_{-\varphi'(0)}| & (SW2) \end{cases}$$

meaning: avoid too positive slopes $\rightsquigarrow$ restrict to things that look like extremum.

$$(SW2) \quad \underbrace{c_2 \varphi'(0)}_{<0} \le \varphi'(t_k) \le -c_2 \varphi'(0)$$

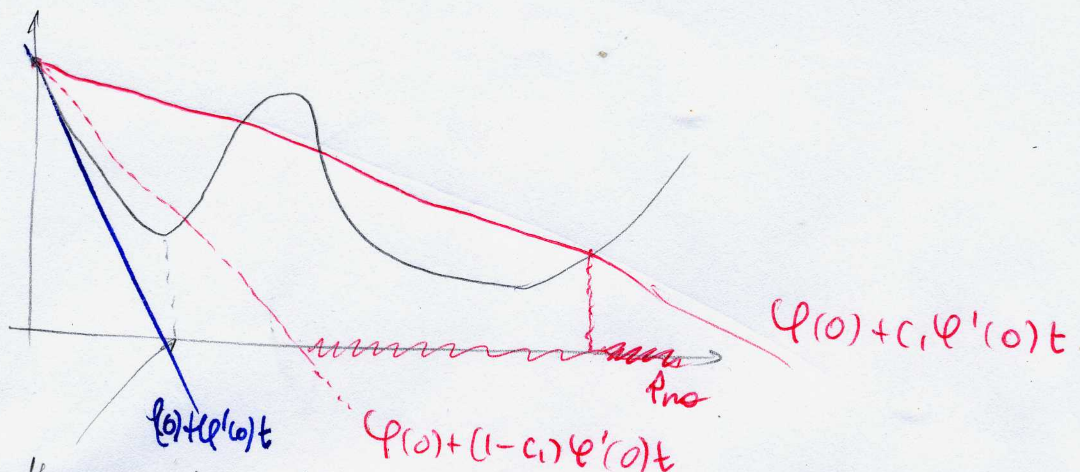
problem: requires derivative eval at trial pt $x_k + t_k p_k$

can be expensive

3] Goldstein cdt°

$$\begin{cases} f(x_k + t_k p_k) \leq f(x_k) + c_1 t_k \nabla f_k^T p_k \\ f(x_k + t_k p_k) \geq f(x_k) + (1-c_1) t_k \nabla f_k^T p_k \end{cases}$$

where $c_1 \in (0, \frac{1}{2})$ for two cdt° to be simult satisfie



this min got missed!!

Goldstein & Wolfe cdt° are similar in convergence proof

4) Backtracking (simple)

Let $0 < \beta_1 \leq \beta_2 < 1$

Set $t_k^{(0)}$

for $i = 0 \ldots$

| if $t_k^{(i)}$ satisfies (SDC) STOP
| else select $t_k^{(i+1)} \in [\beta_1 t_k^{(i)}, \beta_2 t_k^{(i)}]$

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

Thm If the step size t_k satisfies the Wolfe conditions or SWolfe or Goldstein or Backtracking then:

$\exists\ c > 0$ (indep of x) s.t.

$$t_k \geq -c \frac{\nabla f_k^T p_k}{\|p_k\|^2} \quad (*)$$

Thm (Convergence of line search methods)

If f is bounded below then:

$$-\sum_{k=0}^{\infty} t_k \nabla f(x_k)^T p_k < \infty \quad \text{(Convergent series)}$$

$$\Rightarrow \sum_{k=0}^{\infty} \underbrace{\frac{(\nabla f(x_k)^T p_k)^2}{\|\nabla f(x_k)\|^2 \|p_k\|^2}}_{\cos^2\theta_k} \|\nabla f(x_k)\|^2 < \infty$$

$\cos^2\theta_k$; θ_k = angle between $\nabla f(x_k)$ and p_k

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

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

If $p_k = -H_k^{-1}\nabla f(x_k)$

upper bd for num

$$|\nabla f(x_k)^T p_k| = \nabla f(x_k)^T H_k \nabla f(x_k) \geq \frac{1}{\lambda_{max}(H_k)} \|\nabla f(x_k)\|_2^2$$

low bd for denum

$$\|p_k\|^2 = \|H_k^{-1}\nabla f_k\|^2 = \nabla f_k^T H_k^{-2} \nabla f_k \leq \frac{1}{[\lambda_{min}(H_k)]^2} \|\nabla f(x_k)\|_2^2$$

$$\Rightarrow \cos^2\theta_k = \frac{\nabla f_k^T p_k}{\|\nabla f_k\|^2 \|p_k\|^2} \geq \left(\frac{\lambda_{min}(H_k)}{\lambda_{max}(H_k)}\right)^2 = \text{cond}(H_k)^{-2}$$

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

$$\text{cond}(H_k) \leq \sqrt{\gamma^{-1}}$$

Transition to local convergence (p 46)

If:

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

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

iii) t_k chosen according to Wolfe, Goldstein

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

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:

$$(\nabla^2 f(x_k) + \mu_k I)\, p_k = -\nabla f(x_k) \qquad \mu_k \geq 0, \quad \mu_k \to 0$$

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).