Globalization

Test: $\quad \text{ared} = t \cdot \text{pred} \geq 0 \quad \cdots \quad (*)$

$$\text{ared} = \|F(x)\| - \|F(x+s)\|$$

$$\text{pred} = \|F(x)\| - \|F(x) + F'(x)\cdot s\|$$

Theorem: If $\{x_k\}$ satisfies $(*)$ and $\sum \frac{\text{pred}_k}{\|F(x_k)\|} = \infty$.

$\Rightarrow \quad F(x_k) = 0.$

If also x_* is a limit point of $\{x_k\}$ such that $F'(x_*)$ nonsingular, then $F(x_*) = 0$ and $x_k \to x_*$

Modify by backtracking: $\quad s \leftarrow \theta s \quad , \theta \in [\theta_{min}, \theta_{max}] \subseteq (0,1)$

Theorem: If x_* is a limit point of $\{x_k\}$ such that $F'(x_*)$ nonsingular, then $F(x_*) = 0$, $x_k \to x_*$ and $s_k = s_k^N \equiv -F'(x_k)^{-1} F(x_k)$

Work toward a practical backtracking algorithm

choose t small $\quad (t = 10^{-4})$

choose $\theta_{min} = \frac{1}{10} \quad , \theta_{max} = \frac{1}{2}$

common choices

How to choose $\theta \in [\theta_{min}, \theta_{max}]$?

crude: $\theta = \frac{1}{2}$ works but there are more refined methods

"Exact line search" ~ choose $\theta \in [\theta_{min}, \theta_{max}]$ to minimize $\|F(x+\theta s)\|$ over $[\theta_{min}, \theta_{max}]$ - too expensive

Good idea: Choose $\theta \in [\theta_{min}, \theta_{max}]$ to minimize a quadratic or cubic $p(\theta)$ that interpolates $\|F(x+\theta s)\|$

- Set $g(\theta) = \|F(x+\theta s)\|^2$ and assume

$\|v\| = \langle v, v\rangle^{1/2}$ for some inner-product $\langle \circ, \circ \rangle$ on $\mathbb{R}^n$

(ex: $\|v\| = \|v\|_2 \qquad \langle v, v\rangle = v^T v$)

- Have

$$g(0) = \|F(x)\|^2$$
$$g(1) = \|F(x+s)\|^2$$

also

$$g'(\theta) = 2 \langle F(x+\theta s), F'(x+\theta s)\cdot s \rangle$$

$$g'(0) = 2 \langle F(x), F'(x)\cdot s \rangle$$
$$= -2\lambda \langle F(x), F(x) \rangle$$
$$= -2\lambda \|F(x)\|^2$$

Note: $g'(0) < 0 \implies s$ is a descent direction for $\|F\|$ at x

Can construct quadratic $p(\theta)$ such that

$$p(0) = g(0)$$
$$p'(0) = g'(0)$$
$$p(1) = g(1)$$

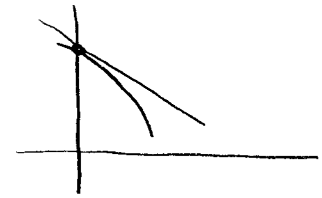
Set $\displaystyle \rho \equiv \frac{\|F(x+s)\|}{\|F(x)\|}$, then for $s = \lambda s^N$, $0 < \lambda < 1$

$$p(\theta) = \|F(x)\|^2 - 2\lambda \|F(x)\|^2 \theta + \|F(x)\| (\rho^2 - 1 + 2\lambda)\, \theta^2$$

Note: $p''(\theta) = 2\|F(x)\|^2 \left(\rho^2 - 1 + 2\lambda\right)$

If $p''(\theta) \leq 0$, take $\theta = \theta_{max}$

$p''(\theta) = 0$, take $\theta = \theta_{max}$



$p''(\theta) > 0$, take global min.

$$\theta_* = \frac{\lambda \|F(x)\|^2}{\|F(x)\|^2\left(\rho^2 - 1 + 2\lambda\right)} = \frac{\lambda}{\rho^2 - 1 + 2\lambda}$$

- If $\theta_* \in [\theta_{min}, \theta_{max}]$ take $\theta = \theta_*$
- otherwise, correct to get $\theta = \theta_{min}$ or $\theta = \theta_{max}$ as appropriate
- At first backtracking step, choose θ in this way
- At subsequent backtracking steps, can do this or can minimize a cubic - see Dennis & Schnabel

Consideration for optimization

- Given $f: \mathbb{R}^n \to \mathbb{R}^1$ want $X_* = \arg\min f(x)$ at least locally
- Adopt Newton to solve $\nabla f(x) = 0$
- How to globalize?
  - Acceptability criteria: Want to reduce f, not $\|\nabla f\|$
  - Then define $ared = f(x) - f(x+s)$
  - Base pred on the <u>local</u> quadratic model of f:

$$q(s) = f(x) + \nabla f(x)^T s + \frac{1}{2} s^T \nabla^2 f(x)\, s$$

see:

$$q(0) = f(x)$$
$$\nabla_s q(0) = \nabla f(x)$$
$$\nabla_s^2 q(0) = \nabla^2 f(x)$$

Then define:

$$pred = f(x) - g(s) = -\nabla f(x)^T s - \frac{1}{2} s^T \nabla^2 f(x) s$$

Then accept s if $ared \geq t\, pred > 0$ as before

- In backtracking, start with $s = s^N \equiv -\nabla^2 f(x)^{-1} \nabla f(x)$ as before

- At each backtrack, choose $\theta \in [\theta_{min}, \theta_{max}]$ to minimize a quadratic or cubic that interpolates $f(x+\theta s)$

- See: $g(\theta) = f(x+\theta s)$

  have
$$\left.\begin{aligned} g(0) &= f(x) \\ g(1) &= f(x+s) \\ g'(0) &= \nabla f(x)^T s \end{aligned}\right\} \Rightarrow \text{ can construct quadratic}$$

As before first step-reduction have an additional value of g and can construct <u>cubic</u>

Note: For $s^N = -\nabla f^2(x)^{-1} \nabla f(x)$

have $\nabla f(x)^T s^N = -\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) < 0$

if $\nabla^2 f(x)$ is SPD

(near minimizer should be PD)

so if $\nabla^2 f(x)$ is SPD, s^N is a descent direction for f at x, and $s = \lambda s^N$ will reduce f for sufficiently small $\lambda > 0$

More-Sorensen and Schultz-Schnabel-Byrd have studied ared/pred conditions for minimization in the context of trust-region-methods

Alternative acceptability criteria

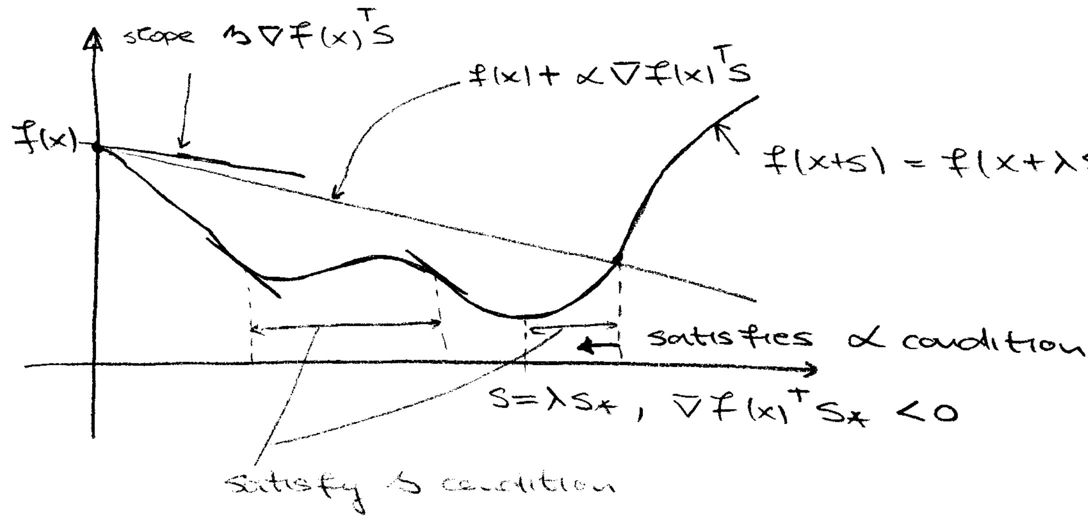
Goldstein-Armijo conditions (GA conditions)

(a.k.a. weak Wolfe conditions, see Nocedal-Wright)

$$f(x+s) \le f(x) + \alpha \nabla f(x)^T s \qquad \sim \text{alpha conditions}$$

$$\nabla f(x+s)^T s \ge \beta \nabla f(x)^T s \qquad \sim \text{beta conditions}$$

$$0 < \alpha < \beta < 1$$



α - condition keep steps from being too long

β - condition keep steps from being too short

<u>Theorem:</u> (Wolfe 1969, 1971) Suppose f is continuously differentiable and $\{x_k\}$ such that each $s_k \equiv x_{k+1} - x_k$ is a descent-direction for f at x_k satisfy the GA conditions. Suppose also that

$$\| \nabla f(x_{k+1}) - \nabla f(x_k) \| \le \lambda \| s_k \| \quad \text{for each } k$$

Then $f(x_k) \to -\infty$ or $\nabla f(x_k)^T \left(\frac{s_k}{\|s_k\|} \right) \to 0 \qquad (*)$

Note: Each s_k must be a descent direction (more later)

Also, will later construct algorithms that keep

$$-\left(\frac{\nabla f_k}{\|\nabla f_k\|} \right)^T \left(\frac{s}{\|s_k\|} \right) \ge \epsilon > 0 \quad --- \quad (**)$$

Then $(*)$ gives

$$\underbrace{\|\nabla f_k\| \left(- \frac{\nabla f_k}{\|\nabla f_k\|}\right)^T \left(\frac{s_k}{\|s_k\|}\right)}_{\geq \epsilon} = - \nabla f_k^T \left(\frac{s_k}{\|s_k\|}\right) \to 0$$

$\Rightarrow \quad \|\nabla f_k\| \to 0$

can obtain convergence statements from Theorem by making assumptions on f

1) $L(x_0) = \{ x : f(x) \leq f(x_0) \}$ – bounded

2) $\nabla^2 f \sim$ SPD in $L(x_0)$

Can show:

Theorem: Suppose hypothesis of previous theorem hold and f is twice continuously differentiable. If $\{x_k\}$ has a limit point x_* such that $\nabla^2 f(x_*)$ is SPD and $(**)$ holds for some $\epsilon > 0$ whenever x_k is sufficiently near x_* then x_* is a local minimizer of f.
If also there is a c such that

$$\|s_k\| \leq c \, \|\nabla f(x_k)\|$$

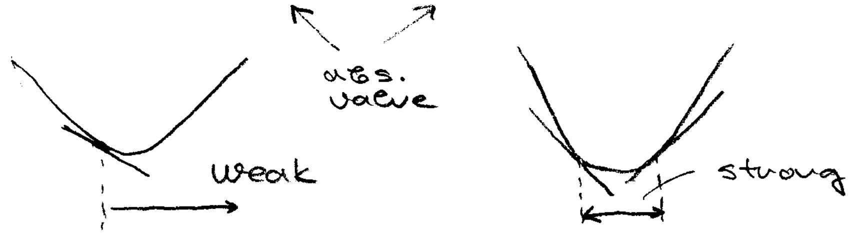
whenever x_k is sufficiently near x_*, then $x_k \to x_*$

Remark: Strong Wolfe conditions

$f(x+s) \leq f(x) + \alpha \nabla f(x)^T s$ as before

$|\nabla f(x+s)^T s| \leq \beta |\nabla f(x)^T s|$

abs. value



- Strong Wolfe conditions are implemented in a widely-used code by More and Thuente (MT)

- Additional points about backtracking with these conditions
  - need $0 < \alpha < \frac{1}{2}$ so that full Newton steps will be acceptable near a minimizer (see Dennis-Schnabel)
  - In practice, β-condition is often not used (starting with Newton step before backtracking keep steps from being too short
  - But MT algorithm does use β-condition
  - At each backtrack, need only re-evaluate f (and not ∇f)
  - <u>very important:</u> Need each $s^N = -\nabla^2 f(x)^{-1} \nabla f(x)$ to be descent direction for f

    guaranteed $\Longleftrightarrow$ $\nabla^2 f(x)$ is SPD

    can expect this to hold near a minimizer but can't expect this away from a minimizer

  - what to do if $\nabla^2 f(x)$ is not SPD

    Traditional approach (a la Dennis-Schnabel)
    - start Cholesky on $\nabla^2 f(x)$
    - If successful, then $\nabla^2 f(x)$ is SPD
    - If not, add as necessary positive d_i to the diagonal to obtain

$$LL^T = \nabla^2 f(x) + D, \quad D = \begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_n \end{pmatrix}$$

    where $d_i \geq 0$ and LL^T is well-conditioned

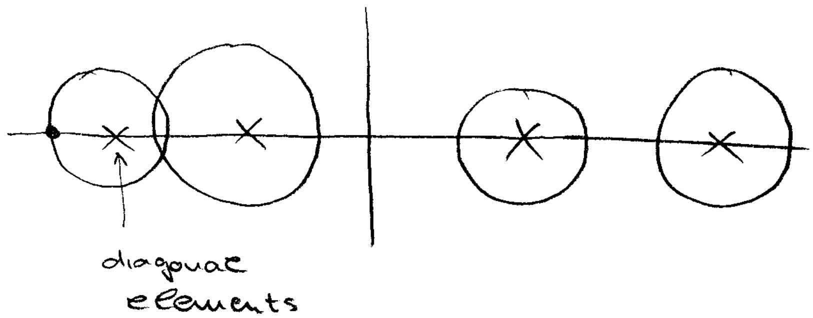
  - use Gerschgorin theorem to compute δ less than the smallest eigenvalue of the Hessian

Gerschgorin Theorem: $A\ (a_{ij})$

set $r_i = \sum_{j \neq i} |a_{ij}|$

$C_i = \{ z \in \mathbb{C} : |z - a_{ii}| \leq r_i \}$

$\Rightarrow$ all eigenvalues are in $\bigcup_{i=1,\dots,n} C_i$



- Take $\mu = \min\{ |\sigma|, \max\{d_i\}\}$
- set $B = \nabla^2 f(x) + \mu I$ (SPD)
- solve $B \cdot s = -\nabla f(x)$

claim: s is a descent direction (but not a Newton step)

$$\nabla f^T s = \nabla f^T [-B^{-1} \nabla f] = -\nabla f^T B^{-1} \nabla f < 0$$