TURN ON RECORDING!

NUMERICAL METHODS FOR NONCONVEX OPT.

TODAY: assume f is smooth
Lec 13: allow f to be nonsmooth
Looking for LOCAL minimizers.
LINE SEARCH METHODS - today
TRUST REGION METHODS: N+W.

1. Newton's method.

If $H = \nabla^2 f(x)$ is not pos def., then $d = -H^{-1}\nabla f(x)$ might not be a descent dir.

$$\nabla f(x)^T d = -\nabla f(x)^T H^{-1} \nabla f(x)$$

What to do?

(a) compute it anyway
could use a convex comb. of d and $-\nabla f(x)$ which is a descent dir'n.

(b) use Cholesky + if it breaks down, keep adding multiples of I until result is PD.

(c) use "modified Cholesky fact."
of Gill, Murray, Wright
- see N+W

(d) use "eig" to get eigenvalues of H.

(e) when chol breaks, you have $R_1^T R_1 = H_{11}$ $\quad H = \begin{bmatrix} H_{11} & H_{21}^T \\ H_{21} & H_{22} \end{bmatrix}$

Then solve $\begin{bmatrix} H_{11} & 0 \\ 0 & I \end{bmatrix} d = -\nabla f(x_k)$

QUASI-NEWTON + NONLINEAR CONJUGATE GRADIENT.

Need an ARMIJO-WOLFE ~~line~~ search

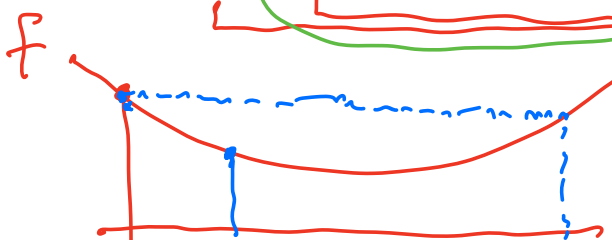
Recall ARMIJO ~~COND~~:

(A) $f(x_k + t d) \le f(x_k) + c_1 t \nabla f(x_k)^T d$, $\quad \nabla f(x_k)^T d < 0$

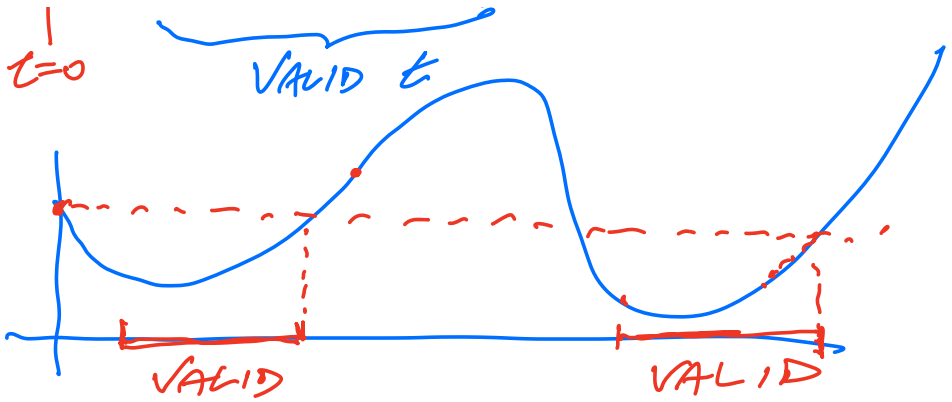
used d before $\in [0, 1/2]$

New: WOLFE COND:

(W) $\nabla f(x_k + t_k d)^T d \ge c_2 \nabla f(x_k)^T d$, $\quad \nabla f(x_k)^T d < 0$



$0 < c_1 < c_2 < 1$



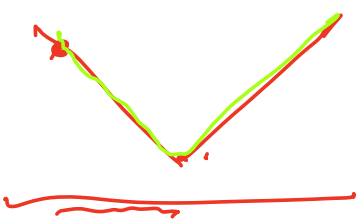
(A) ensures that t NOT too large

(W) " " " " " small

Often a "strong Wolfe"

$$|\nabla f(x_k + t d_k)| \le c_2 |\nabla f(x_k)|$$

Impossible if it turns out that f is nonsmooth



A SIMPLE BRACKETING ARMIJO-WOLFE LINE SEARCH. Assume $\nabla f(x_k)^T d_k < 0$

AWLS

Set $t \leftarrow 1$, done = false, $\alpha \leftarrow 0$, $\beta \leftarrow \infty$

Property: $[\alpha, \beta]$ always contains a valid A-W step t

while not done

$x \leftarrow x_k + t d$

if $f(x) > f(x_k) + c_1 t \nabla f(x_k)^T d$

$\beta \leftarrow t$: (A) is violated, so t is too large.

elseif $\nabla f(x)^T d < c_2 \nabla f(x)^T d$

$\alpha \leftarrow t$: (W) is violated, so t is too small

else $\alpha \leftarrow t$, $\beta \leftarrow t$, done $\leftarrow$ true

end

% set up next t evaluation

if $\beta < \infty$

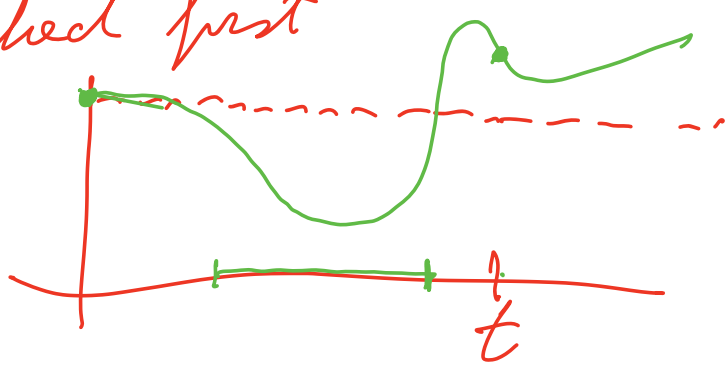
$t \leftarrow \frac{\alpha+\beta}{2}$ (bisection)

else

$t = 2t$ (double)

end

Essential that violation of (A) is checked first



Thm If f is C^1 and bounded below, on $\{x_k + td : t \geq 0\}$

then the A(W) terminates with a valid A-W step.

For an analysis under weaker assumptions on f, see Lewis+Overton 2013.

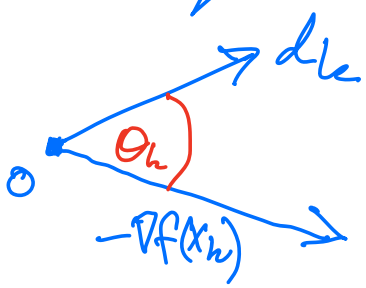
Code: linesearch_ww.m
"weak Wolfe" ≡ "armijo Wolfe"
in HANSO.

ZOUTENDIJK'S THEOREM

Assume f is bd below, C^1, $\nabla f(x)$ is Lipschitz on $\{x : f(x) \le f(x_0)\}$

$$x_{k+1} = x_k + t_k d_k$$

where t_k satisfies the (A-W) conditions and



$$\cos\theta_k = -\frac{\nabla f(x_k)^T d_k}{\|\nabla f(x_k)\| \, \|d_k\|}$$

Then $\sum_{k=0}^{\infty} (\cos\theta_k)^2 \, \|\nabla f(x_k)\|^2 < \infty$ (*)

Pf

11 Write g_k for $\nabla f(x_k)$, t_k for $f(x_k)$

We have from (W)

$$g_{k+1}^T d_k \geq c_2 g_k^T d_k$$

or $(g_{k+1} - g_k)^T d_k \geq (c_2 - 1) g_k^T d_k$

or $L \|t_k d_k\| \|d_k\| \geq (c_2 - 1) g_k^T d_k$

or $t_k \geq \frac{c_2 - 1}{L} \frac{g_k^T d_k}{\|d_k\|^2}$ ←

subst. into (A):

$$f_{k+1} \leq f_k + t_k c_1 g_k^T d_k$$

$$f_{k+1} \leq f_k + \frac{c_2 - 1}{L} \frac{g_k^T d_k}{\|d_k\|^2} c_1 g_k^T d_k \frac{\|g_k\|^2}{\|g_k\|^2}$$

$$f_{k+1} \leq f_k - K (\cos\theta_k)^2 \|g_k\|^2$$

$K = c_1(1 - c_2)/L$

Sum over $j \leq k$

$$f_{k+1} \leq f_0 - K \sum_{j=0}^{k} (\cos\theta_j)^2 \|g_j\|^2$$

Since f is bd below,

let $k \to \infty$

Apply to any Newton-like method

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

as long as H_k is UNIFORMLY POS DEF

QUASI-NEWTON METHODS

("secant methods")

Motivation:

Newton's method is $O(n^3)$ work

Instead: update an approximation

$\leftarrow$ FACTORIZATION or INVERSE of $\nabla^2 f(x_k)$ is $O(n^2)$ time.

How? Exploit gradient info.

After line search we have

x_k $\quad g_k = \nabla f(x_k)$

x_{k+1} $\quad g_{k+1} = \nabla f(x_{k+1})$

Let $s_k = x_{k+1} - x_k$

$y_k = g_{k+1} - g_k$.

$$y_k = \int_0^1 \nabla^2 f(x_k + \tau s_k)\, s_k \, d\tau$$

$$= \int_0^1 \left(\nabla^2 f(x_k + \tau_k s_k)\, d\tau\right) s_k$$

G_k "average Hessian along s_k"

So this motivates requiring our new approx to $\nabla^2 f(x_{k+1})$, say B_{k+1} to satisfy

$$B_{k+1} s_k = y_k$$

the SECANT (or QN) EQN.

or, if we approx $\nabla^2 f(x_{k+1})^{-1}$, say, H_{k+1}

$$H_{k+1} y_k = s_k$$

Famous choices: PSB, DFP

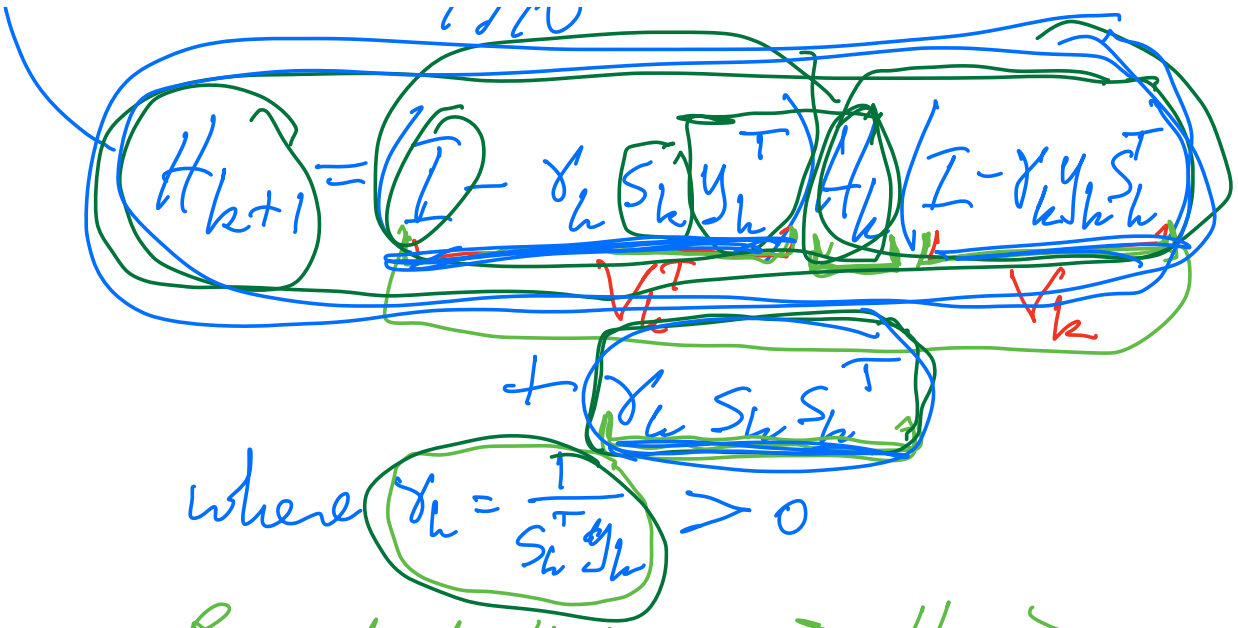
Davidon 1959

Fletcher Powell 1963

BFGS

Broyden, Fletcher, Goldfarb, Shanno

1970



Can check $H_k \succ 0 \Rightarrow H_{k+1} \succ 0$

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

$(y_k - \gamma_k y_k s_k^T y_k) = 0$

$= s_k$

Computing H_{k+1} is $O(n^2)$

Setting $d_k = H_{k+1} \nabla f(x_k)$ $O(n^2)$

If we approx $B_{k+1} \approx \nabla f(x_k)$ instead, then we have to solve linear system.
Or, could update Cholesky fact.

of B_{k+1} is $O(n^2)$

THM (Powell, 1976)

Suppose (i) $f \in C^2$, convex

(ii) $\Omega = \{x : f(x) \le f(x_0)\}$ is convex

(iii) $\exists\, m, M$ s.t.

$$m\|z\|^2 \le z^T \nabla^2 f(x) z \le M\|z\|^2$$

strongly convex

$z \in \mathbb{R}^n$, $x \in \Omega$

Then $\{x_k\}$ from BFGS with A-W line search, satisfies

$x_k \to$ unique local minimizer of f, x^*

Here H_0 is any pos. def. matrix.

Why is $s_k^T y_k > 0$?

$$g_{k+1}^T d_k \ge c_2 g_k^T d_k \quad (W)$$

$$y_k^T d_k = g_{k+1}^T d_k - g_k^T d_k \ge (c_2 - 1) g_k^T d_k > 0$$

$s_k = t_k d_k$.

$$s_k^T y_k = y_k^T s_k \geq (c_2 - 1) g_k^T s_k > 0.$$

Superlinear Convergence (Dennis+Moré)

If also assume $\nabla^2 f$ is Lipschitz on Ω then $x_k \to x^*$ superlinearly, i.e.,

$$\lim_{k \to \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0$$

LIMITED MEMORY BFGS

L-BFGS.

(N+W, Sec 7.2)

Use $O(n)$ storage and $O(n)$ work.

Let $m \ll n$

At iter k, we have x_k, $\nabla f(x_k)$ and we have SAVED from previous iterates

$$\{s_i, y_i\}, \quad i = k-m, \ldots, k-1$$

Choose H_k^0, usually $\gamma_k I$, + set

$$H_k = V_{k-1}^T \cdots V_{k-m+1}^T V_{k-m}^T H_k^0 V_{k-m} V_{k-m+1} \cdots V_{k-1}$$
$$+ \gamma_{k-m} V_{k-1}^T \cdots V_{k-m+1}^T s_{k-m} s_{k-m}^T V_{k-m+1} \cdots V_{k-1}$$
$$+ \cdots$$
$$+ \gamma_{k-1} s_{k-1} s_{k-1}^T$$

Usually use "scaling": $H_k^0 = \tau_k I$

where $\tau_k = \dfrac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}}$

Can show L-BFGS has linear convergence to minimizer when f is smooth + convex

OTHER $O(n)$ METHODS

NONLINEAR CONJUGATE GRADIENT

$$d_0 = -g_0 \equiv -\nabla f(x_0)$$

for $k = 1, 2, \ldots$

$$x_{k+1} = x_k + t_k d_k$$

$$d_{k+1} = -g_{k+1} + \beta_{k+1} d_k$$

FLETCHER–REEVES

$$\beta_{k+1}^{FR} = \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k}$$

POLAK–RIBIERE

$$\beta_{k+1}^{PR} = \frac{g_{k+1}^T (g_{k+1} - g_k)}{g_k^T g_k}$$

In both cases, reduces to "linear CG" if $Ax = b$ $A \succ 0$

i.e. $\min \frac{1}{2} x^T A x - b^T x$

But so does BFGS.

A variant:

β_{k+1}^{FRPR} is projection of β_{k+1}^{PR} onto $\left[-\beta_{k+1}^{FR}, \beta_{k+1}^{FR}\right]$

CONSTRAINED NONCONVEX OPTIMIZATION

Two excellent codes

SNOPT "successive quadratic programming"

IPOPT "interior point"

$A = U \Sigma V^T$

$A^T A = V \Sigma \cancel{U^T U} \Sigma V^T = V \Sigma^2 V^T$