Numerical Methods for NonConvex Optimization

- that are applicable when f is smooth but nonconvex: TODAY
- that " " " f is nonsmooth + nonconvex: NEXT WEEK

Assume f smooth (but not nec. convex).
Two classes of methods.
LINE SEARCH (Armijo + Weak or Strong Wolfe.)
TRUST REGION — won't discuss. See N+W. (Nocedal + Wright)

Line Search Methods

1. Newton's Method.
Can continue to use backtracking line search.
but need to modify $H = \nabla^2 f(x)$ if it is not positive definite — because otherwise $d = -H^{-1}\nabla f(x)$ may not be a descent direction: $d^T \nabla f(x) = -d^T H d$.

How?

(a) Use chol, repeat with adding multiples of I to H if necessary

(b) "Modified Cholesky" factorization of GMW.
See N+W. A bit complicated, not built-in to Matlab.

(c) Use eig. A bit expensive.

(d) Use symmetric indef. fact $H = LBL^T$ where B has 1×1 and 2×2 blocks.
(Bunch-Parlett-Kaufman)

2. QN and CG Methods (Quasi-Newton + Conjugate Gradient)
We need a more sophisticated line search.

Weak Wolfe line search

Much less complicated than "strong Wolfe"! *

And, useful for nonsmooth functions as well as smooth.

Recall we want t such that (ARMIJO CONDITION)

$$f(x_0+td) \le f(x_0) + c_1 t \nabla f(x_0)^T d$$ (SO DON'T GO TOO FAR)

New condition (WEAK WOLFE)

$$\nabla f(x_0+td)^T d \ge c_2 \underbrace{\nabla f(x_0)^T d}_{<0}$$ (SO GO FAR ENOUGH)

One scenario:

another scenario:

To ensure existence of such points in smooth case, we assume $c_2 > c_1 > 0$.

c_2: normally take NOT too small, otherwise too demanding.

c_1: normally take small

For nonsmooth case, might like to take $c_2=0$, so that $\nabla f(\cdot)^T d$ changes sign.
But this is too restrictive in smooth case — for example, can prevent superlinear convergence of BFGS.

We will update an interval $[\alpha, \beta]$ that brackets a point satisfying the W.W. conditions.

Initially $\alpha=0$, $\beta=\infty$.

[* Strong Wolfe condition is $|\nabla f(x_0+td)| \le c_2 |\nabla f(x_0)^T d|$. See N+W.]

<u>Weak Wolfe LS Alg.</u> $\leftarrow$ $t \leftarrow 1$

while not done

$x \leftarrow x_0 + td$

if $f(x) > f(x_0) + c_1 t \nabla f(x_0)^T d$

$\beta \leftarrow t$ % 1^{ST} condition violated, gone too far

elseif $\nabla f(x)^T d < c_2 \nabla f(x_0)^T d$

$\alpha \leftarrow t$ % 2^{nd} condition violated, not gone far enough

else

$\alpha \leftarrow t$

$\beta \leftarrow t$

STOP

end.

% set up next function evaluation

if $\beta < \infty$

$t \leftarrow (\alpha + \beta)/2$

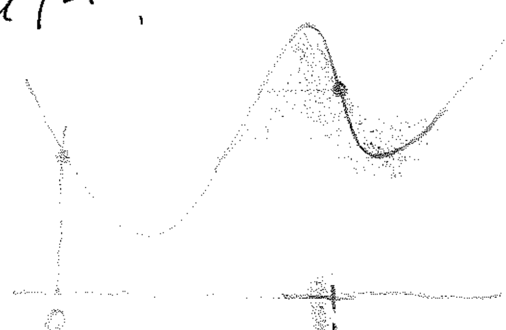
else

$t \leftarrow 2t$

end.

It is crucial that violation of the $\underline{1^{ST}}$ condition is checked 1^{ST}.

e.g.



Both conditions violated.
Update β, NOT α.

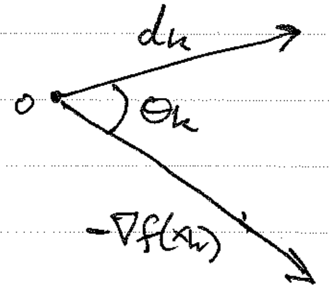
ZOUTENDIJK'S THEOREM

Assume f is bdd below, $f \in C^1$ and ∇f is Lipschitz on $\{x : f(x) \le f(x_0)\}$. (Lip const. L)

Define a descent algorithm:

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

e.g. $d_k = -\nabla f(x_k)$
$\theta_k = 0$.



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

where t_k satisfies the Weak Wolfe + Armijo conditions

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

so, if $\cos\theta_k > \tau > 0$ for all k
$$(\theta_k < \psi < \tfrac{\pi}{2})$$

we have $\nabla f(x_k) \to 0$.

Pf Write g_k for $\nabla f(x_k)$, f_k for $f(x_k)$.

We have from the Wolfe condition that

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

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

so $$L \|t_k d_k\| \, \|d_k\| \ge (c_2 - 1) g_k^T d_k$$

so $$t_k \ge \frac{c_2 - 1}{L} g_k^T d_k / \|d_k\|^2 \qquad \text{(product of 2 neg. numbers)}$$

Substitute this into the Armijo condition:

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

so

$$f_{k+1} \le f_k \underbrace{-K}_{\frac{c_1(1-c_2)}{L}} (\cos\theta_k)^2 \|g_k\|^2$$

Sum over $j \le k$

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

Since f is bd below, let $k \to \infty \Rightarrow (*)$.

Zoutendijk: Thm applies to any "Newton-like" method

$d_k = H_k^{-1} \nabla f(x_k)$ as long as H_k is UNIFORMLY POSITIVE DEF.

However, hard to prove this for QN methods — and not true for CG methods.

Rate of convergence of Gradient Method: as in convex case, slow.

______ of Newton's Method: as before, quadratic under regularity assumption, but no guarantees can be made re # iters.

QUASI-NEWTON METHODS.

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

Want to update an approx. to [FACTORIZATION OR INVERSE] of $\nabla^2 f(x)$ in $\underline{O(n^2)}$ time.

How? Make better use of gradient info.

After line search, we have:

$$x_k \qquad g_k \equiv \nabla f(x_k)$$
$$x_{k+1} \qquad g_{k+1} = \nabla f(x_{k+1}).$$

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

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

From Fund. Thm. of Calc.,

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

$$= \underbrace{\left[\int_0^1 \nabla^2 f(x_k + \tau s_k)\, d\tau\right]}_{G_k \text{ "average Hessian along } s_k\text{"}} s_k.$$

So it seems reasonable that our new approx to $\nabla^2 f(x_{k+1})$, say B_{k+1}, should satisfy

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

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

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

THE SECANT EQUATION.

Davidon–Fletcher–Powell
1959 1963.

Various choices known: PSB, DFP, BFGS

Broyden–Fletcher–Goldfarb–Shanno
1970

BFGS:

$$C_{k+1} = (I - \gamma_k s_k y_k^T) C_k (I - \gamma_k y_k s_k^T) + \gamma_k s_k s_k^T$$

where $\gamma_k = \dfrac{1}{s_k^T y_k}$.

Check that $C_{k+1} y_k =$

$$(I - \gamma_k s_k y_k^T) C_k \underbrace{(y_k - \cancel{\gamma_k} y_k \cancel{s_k^T y_k})}_{0} + \cancel{\gamma_k} s_k \cancel{s_k^T y_k}$$

✓.

How much work is needed to compute C_{k+1}? ASK

⟹ HOW DO WE KNOW $s_k^T y_k > 0$? DIRECTLY FROM THE WOLFE CONDITION (see next page)

THM (Powell, 1976)

Suppose 1. $f \in C^2$

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

3. $\exists\, m, M$ s.t.

STRONG CONVEXITY

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

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

Then $\{x_k\}$ generated by BFGS with Armijo-Wolfe line search satisfies $x_k \to$ UNIQUE LOCAL MINIMIZER OF f (x^*) (existing).

Pf: beautiful, 2 pages + 3 outstanding. See Nocedal + Wright.

HARD PART IS SHOWING EIGS(C_k) REMAIN BOUNDED AND $\ge \delta > 0$.

<u>Clarifying why $s_k^T y_k > 0$</u>

$$g_{k+1}^T d_k \geq c_2 \underbrace{g_k^T d_k}_{<0} \quad \text{(Wolfe)}$$

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

$$s_k = t_k d_k$$

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

Superlinear Convergence (Dennis + Moré).

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

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

Pf: See N+W.

NONCONVEX CASE: Nothing known in theory. But in practice, always works! Like st. desc. $\|\nabla f(x_k)\| \to 0$ CLUSTER POINTS ARE STATIONARY but UNLIKE ST. D. convergence is superlinear.

Methods that are $O(n)$

LIMITED MEMORY BFGS — far more efficient: $O(n)$ see Noc + Wri. But not sup. conv.

NONLINEAR CG Set $d_0 = -\nabla f(x_0) = -g_0$.

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

$$d_{k+1} = -\underset{(-\nabla f(x_{k+1}))}{g_{k+1}} + \beta_{k+1} d_k \qquad \text{AS IN "LINEAR" CG}$$

~~Fletcher Reeves~~ $\beta^{FR}_{k+1} = \dfrac{g_{k+1}^T g_{k+1}}{g_k^T g_k}$

Polak-Ribière

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

In both cases reduce to "linear" CG when f is quadratic.

"Linear CG" $\equiv$ solve $Ax=b$, $A \succ 0$

<u>Theorem</u> If $f(x) = x^T A x - b^T x$ $\quad A \succ 0$
and we use $t_k \leftarrow$ exact line search (minimum of quadratic along line)
then CG $\equiv$ BFGS
and terminates in n steps. (More in presence of rounding & ill-conditioning of A).
$\uparrow$ OR FEWER: #DISTINCT EIGS(A).
When n is large, want <u>good approx</u> in much fewer than n steps.

Convergence is like $\left(1-\sqrt{\frac{m}{M}}\right)^k$ compared to $\left(1-\frac{m}{M}\right)^k$ for steepest descent.

<u>Nonlinear Case</u> (Nonquadratic)
To get convergence result for Nonlinear CG, need to use a "strong Wolfe" line search to prove that FR converges — but it's generally inferior to PR. (See NocWri)

A variant: CGPRFR

$$\beta_{k+1} = \begin{cases} -\beta_{k+1}^{FR} & \text{if } \beta_{k+1}^{PR} < -\beta_{k+1}^{FR} \\ \beta_{k+1}^{PR} & \text{otherwise} \\ \beta_{k+1}^{FR} & \text{if } \beta_{k+1}^{PR} > \beta_{k+1}^{FR} \end{cases}$$

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

is a good compromise: as good or better than PR in practice, same conv. theory as FR.
Recent work: variants that use only weak Wolfe.

Nonlinearly Constrained Optimization

A huge area.
Two big classes of algs.

SQP
IP.

See N+W.
Software SNOPT
IPOPT.

Margaret Wright will teach a course that treats these in depth in Fall 2016.