Lecture 22

Details of a practical Newton iteration C con

- Convergence with strong Wolfe line search
- Theorem 3.2 in Nocedal + Wright
- Zoutendijk's condition

- What to do with indefinite Hessian's in Newton?
- Modified Cholesky ✓
- Proof of convergence

Reminder: Report 1 due Monday.

Math in Industry workshop:

https://sites.google.com/view/iwpd2025/home

Pre workshop event this afternoon on Matlab.

Back to line
Search methods

minimize $f(\underline{x})$

Assume $g(\underline{x})$ is available.

Alg: While $\|g_k\| \geq tol$

- pick a direction p_k as a descent dir.
- find α_k to satisfy the Strong Wolfe Condition or Backtracking –
- Set $\underline{x}_{k+1} = \underline{x}_k + \alpha_k p_k$

|T|

LOTS OF CONDITIONS

$\|g_k\| \to 0$ w/ Backtracking line Search.

then

Theorem 3.2 to Nocedal + Wright shows the same thing w/ Strong Wolfe.

- lots of assumptions
  - ↳ f bounded below
  - ↳ differentiable
  - ↳ Lipschitz
  - etc.

Then

$$\sum_{k\geq 0}^{\infty} \cos^2\theta_k \|g_k\|^2 < \infty$$

$$\cos\theta_k = \frac{-g_k^T p_k}{\|g_k\| \cdot \|p_k\|}$$

Called Zoutendijk's Conditio

$$\sum_{k\geq 0}^{\infty} \cos^2\theta_k \|g_k\|^2 < \infty$$

So if $\cos^2\theta_k \geq \mu$

then $\|g_k\| \to 0$.

Newton for Optimization

"Simple idea"

$$p_k = -H_k^{-1} g_k$$

Why?

$$f(x_k + p) \simeq M_k(p) =$$

$$f_k + p^T g_k + \frac{1}{2} p^T H_k p$$

Quiz

What is

$$\min_{p} \; M_k(p)$$

(or argmin if you prefer).

If $H_K \succ 0$, then the function $M_K(p)$ is strongly convex, so

$$\partial M_K / \partial p = g + H_K \underline{p} = 0$$

$$\underline{p} = -H_K^{-1} g$$

If H_K is indefinite, or neg. def. then...

$$M_K(p) \to -\infty.$$

So pick $\underline{p}$ as any direction when $p^T H_K p < 0$.

If H_K is positive semidefinite, then $p_K = -H^+ g$ (pseudo-inverse) to get a soln.

There are others too!

Br

1.

2.

3.

In Newton

if $H_k \succ 0$, then great!

ut What if not?

Make it SPD!

↳ Normal Eq.

↳ Low-rank approx.

Use Grad. Desc. instead.

Enforce a max. Step length.

Choose neg. eigenvector direction)

Or pick a different model.

If H_k is not pos. def. then we pick C_k s.t.

$$B_k = H_k + C_k$$

is pos. def.

Then use $p = -B_k^{-1} g$ as a search dir.

How to pick c_k?

if $B_k = H_k + \tau I$

and $\tau > |\lambda_{min}(H_k)|$.

Then $B_k \succ 0$.

How to test if

A, H_k, etc is

Sym. pos. def?

1. A symmetric matrix A is spd. iff $A = FF^T$ for some F.

2. Cholesky factorization will give you F.

We can use this w/ Bisection/Doubling to find a τ s.t.

$H_k + \tau I$ is spd.

How does Cholesky work?

Why does it fail?

$A = FF^T \leftarrow$ F is lower tri.

$A = LDL^T \leftarrow$ L is lower tri w/ 1s on diag.

$\uparrow$ D is diag.

$\boxed{\text{D is all positive!}}$

$A = \begin{bmatrix} \alpha \\ a_1 \end{bmatrix}$

A(2:end,1)

Then Chole

Start w/

$A = \begin{bmatrix} 1 & 0 \\ \ell & R \end{bmatrix}$

$\alpha = \gamma_1$

$RDR^T =$

$A(1,1)$

$a_1^T \leftarrow A(1,2{:}\text{end})$

$\underline{a}_1$

$B \leftarrow A(2{:}\text{end}, 2{:}\text{end})$

oky factorn

$$\left[\begin{bmatrix}\gamma & 0\\ 0 & D\end{bmatrix}\begin{bmatrix}1 & \ell^T\\ 0 & R^T\end{bmatrix}\right]$$

$\underline{\ell} = \underline{a}_1/\gamma$

$B - \gamma \underline{\ell}\,\underline{\ell}^T$

If A is spd.
then $\gamma > 0$
for all steps $\Longleftrightarrow$
all $D_{ii} > 0$.

Idea: if
$\gamma < 0$,

...

Use $\gamma > 0$
anyway by Fiat!
Force!
Will.

This will give
Cholesky of
$A + E_K$
$\uparrow$
perturbation.

Theory

If $B_k = H_k + E_k$

where E_k comes from a modified Cholesky factorization of H_k.

Then we can pick choices so that $B_k \succeq 0$, and

$$\boxed{\|B_k\| \cdot \|B_k^{-1}\| \le K}$$

Also, if $p_k = -B_k^{-1} g_k$, then

$$\cos(p_k, g_k) \ge \frac{1}{K}.$$

More + Sorensen 1982, old paper.

How does this give bound on cos?

$$\cos(\underline{a}, \underline{b}) = \frac{\underline{a}^T \underline{b}}{\|\underline{a}\| \cdot \|\underline{b}\|}.$$

w/ $M \succeq 0$, $M = M^T$,

$M\underline{a} = \underline{b} \Leftrightarrow \underline{a} = M^{-1}\underline{b}$.

$$\frac{1}{K} = \frac{1}{\frac{\lambda_{\max}(M)}{\lambda_{\min}(M)}} = \frac{\lambda_{\min}(M)}{\lambda_{\max}(M)}$$

$\|a\| =$

$$\frac{a^T b}{\|a\| \cdot \|b\|} = \frac{a^T M a \, \|a\|}{a^T a \, \|b\|} \geq \lambda_{\min}(M) \cdot \frac{\|a\|}{\|b\|} \geq \frac{\lambda_{\min}(M)}{\lambda_{\max}(M)}.$$

$$\sqrt{a^T a}$$

$$\frac{\|a\|}{\|b\|} = \frac{\|M^{-1} b\|}{\|b\|} \geq \sigma_{\min}(M^{-1}) \geq \frac{1}{\lambda_{\max}(M)}$$

for $M = M^T$,
$M \succ 0$

$$\lambda_{\min}(M^{-1}) = \frac{1}{\lambda_{\max}(M)}$$