Nonlinear Least-squares problems

So far $F(x_*) = 0$, $F: \mathbb{R}^n \to \mathbb{R}^n$

Will consider $F: \mathbb{R}^n \to \mathbb{R}^m$

Say
- $m > n$ $\Rightarrow$ overdetermined
- $m < n$ $\Rightarrow$ underdetermined

} regardless of existance uniqueness of solution

Tonight, overdetermined case ~ nonlinear LS problem

Assume $\|\cdot\| = \|\cdot\|_2$ throughout

Linear LS problems:

- suppose s depends on $\alpha_1, \dots, \alpha_n$
- postulate linear model $s = x_1\alpha_1 + \dots + x_n\alpha_n$
- want to determine x_i's from observations

$i = 1, \dots m$
- $b_i = i^{th}$ observed value of s
- $a_{ij} = i^{th}$ observed value of α_j $\quad 1 \le j \le n$

Method of least squares: Determine x_i's to

$$\text{minimize} \sum_{i=1}^{m} \left(b_i - \sum_{j=1}^{n} a_{ij} x_j \right)^2 = \|b - Ax\|^2$$

where $b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}$ $\quad A = (a_{ij})$ $\quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$

$b - Ax$ ~ residual (vector)

$\|b - Ax\|^2$ ~ sum of squared residuals

solution is characterized by $A^TAx = A^Tb$ ..... $(*)$

normal equation of least-squares

$\Leftrightarrow$ $A^T(b - Ax) = 0$ i.e. residuals are orthogonal (normal) to range (A)

(*) always has a solution - unique $\Leftrightarrow$ rank $(A) = n$

Also A^TA is symmetric and positive semidefinite

(SPD) $\Leftrightarrow$ rank $(A) = n$

- If A is full rank, can form A^TA A^Tb and solve using Cholesky. But there are alternatives base on QR decomposition or SVD of A
- Extend to nonlinear LS problem.

Obvious extension: nonlinear model

$$y = f(x, t) \qquad \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \text{ independent variables}$$

$$\begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix} \sim \text{parameters}$$

Then given observations on y and x choose t to

$$\text{minimize } \| b - g(a, t) \|^2$$

where

$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \qquad g = \begin{pmatrix} f(a_1, t) \\ \vdots \\ f(a_m, t) \end{pmatrix}$$

$b_i = i^{th}$ observation on y $\qquad$ $a_i = i^{th}$ observation of x

Example: $y = t_1 e^{t_2 x}$

Consider even more general problem

$$\text{minimize } f(x) = \frac{1}{2} \| F(x) \|^2 \qquad F: \mathbb{R}^n \to \mathbb{R}^m$$

Natural: apply Newton-based methods to find x_* such that $\nabla f(x_*) = 0$

This problem has <u>special structure</u>

Can show: $\nabla f = J^T F$, $J = F'$ (are evaluated at "current" x)

$$\nabla^2 f = J^T J + \sum_{i=1}^{n} F_i \nabla^2 F_i$$, where F_i = i^{th} component of F

See: $J^T J$ is symmetric positive semidefinite always, and positive definite $\Leftrightarrow$ rank $(J) = n$
Also once J is evaluated to form ∇f, have $J^T J$ with some additional arithmetic (and no additional function evaluations)

$\sum_{i=1}^{m} F_i \nabla^2 F_i$ requires additional evaluations of m $n \times n$ Hessians (usually infeasible)

Also $\sum F_i \nabla^2 F_i$ may be such that $\nabla^2 f$ is indefinite $\Rightarrow$ Newton step may not be a descent direction

Gauss-Newton method

given an initial x
until termination do
solve $J^T J s = -J^T F$ $(**)$
update $x \leftarrow x + s$

Remarks:
(1) $(**)$ ~ normal equation for linear LS problem

$$\min \|F - Js\|^2$$

usually solve using QR or SVD, sometimes form $(**)$ and solve using Cholesky

(2) Gauss-Newton step $s = -(J^TJ)^{-1}J^TF$ is a descent direction for f if rank $(J) = n$

$$\nabla f^T s = (J^TF)^T \left(-(J^TJ)^{-1}J^TF \right)$$

$$= -(J^TF)^T (J^TJ)^{-1} (J^TF) < 0$$

for $J^TF \neq 0$,

since

rank $(J) = n \Rightarrow J^TJ$ is SPD $\Rightarrow (J^TJ)^{-1}$ is SPD

(3) termination? As before

stop if iteration number $\geq$ itmax

stop if $\|s\| \leq \epsilon_x$

can't expect $\|F(x_*)\| = 0$, so stop if

$$\|\nabla f(x)\| = \|J(x)^TF(x)\| \leq \tau_F$$

Convergence

Recall: Given $F: \mathbb{R}^n \to \mathbb{R}^n$, consider

General iteration

Given x_0 and B_0

for $k = 0, 1, \ldots$

$x_{k+1} = x_k - B_k^{-1}F(x_k)$

determine B_{k+1}

Theorem: Suppose F is continuously differentiable near x_* such that $F(x_*) = 0$ and $F'(x_*)$ nonsingular. Suppose that the general iteration produces $\{x_k\}$ with $\{B_k\}$ such that for all k

1) $\|B_k^{-1}\| \leq M$

2) $\|I - B_k^{-1}F(x_k)\| \leq \epsilon_{max} < 1$

Then, given any ϵ, $0 < \epsilon_{max} < 1$ there is a $\delta > 0$ such that $\|x_0 - x_*\| \le \delta$, then $x_k \to x_*$ with

$\|x_{k+1} - x_*\| \le \epsilon \|x_k - x_*\|$ for all k

Apply to Gauss-Newton:

$$F \leftarrow \nabla f = J^T F \quad \text{(not same } F\text{'s)}$$

$$B_k \leftarrow J(x_k)^T J(x_k)$$

Assume $f = \frac{1}{2}\|F\|^2$ is twice continuously differentiable near x_* such that $\nabla f(x_*) = J^T(x_*) F(x_*) = 0$ and $J(x_*)$ is full-rank

see: $J(x_*)^T J(x_*)$ is SPD $\Rightarrow$ $J(x)^T J(x)$ and $\left(J(x)^T J(x)\right)^{-1}$ exist are continuous and are SPD and

$\|\left(J(x)^T J(x)\right)^{-1}\| \le M$ for x near x_*

So (1) holds

Assume: $\|I - \left(J(x_*)^T J(x_*)\right)^{-1} \nabla^2 f(x)\| \le \epsilon_* < 1$

then for any ϵ_{max} in $(\epsilon_*, 1)$ have

$$\|I - \left(J(x)^T J(x)\right)^{-1} \nabla^2 f(x)\| \le \epsilon_{max} < 1 \quad \text{---} \quad (***)$$

for x sufficiently near x_* $\Rightarrow$ (2) also holds for x_k near x_*

Then Theorem $\Rightarrow$ For any ϵ, $\epsilon_{max} < \epsilon < 1$, $\exists \delta > 0$ such that

$$\|x_0 - x_*\| \le \delta \Rightarrow x_k - x_*$$

with $\|x_{k+1} - x_*\| \le \epsilon \|x_k - x_*\|$ for all k

Theorem concludes with $x_k \to x_*$ superlinearly if

$$\lim_{k\to\infty} \| I - B_k^{-1} F'(x_k) \| = 0$$

Translates to $\| I - (J(x_*)^T J(x_*))^{-1} \nabla^2 f(x_*) \| = 0$

see (***) is $\epsilon_{max} \geq \| I - (J^T J)^{-1} \nabla^2 f \|$

$$= \| I - (J^T J)^{-1} \left(J^T J + \sum_{i=1}^{m} F_i \nabla^2 F_i \right) \|$$

cancels

$$= \| (J^T J)^{-1} \left(\sum_{i=1}^{m} F_i \nabla^2 F_i \right) \|$$

this is small in the "small" residual case (all $F_i(x_*)$'s small) or when F is not very nonlinear (all $\nabla^2 F_i(x_*)$'s small)

- from last statement of theorem, get <u>superlinear</u> convergence to x_* in "zero" residual case. (all $F_i(x_*) = 0$)
- can do better

Theorem: Suppose x_* is such that $F(x_*) = 0$, $J(x_*)$ is full rank and

$$\| J(x) - J(x_*) \| \leq \gamma \| x - x_* \| \quad \text{for } x \text{ near } x_*$$

Then $\exists\, \delta > 0$ such that if $\| x_0 - x_* \| \leq \delta$ then the Gauss-Newton iterates are well-defined and converge quadratically to x_*

proof: Have

$$\| (J(x_*)^T J(x_*))^{-1} \sum_{i=1}^{m} F_i(x_*) \nabla^2 F_i(x_*) \| = 0$$

($F_i(x_*) = 0$)

and have local superlinear convergence

so can assume $x_k \to x_*$

For x near x_* get

$$x_+ = x - (J^T J)^{-1} J^T F$$

set $e = x - x_*$

$e_+ = x_+ - x_*$

Have $e_+ = e - (J^T J)^{-1} J^T F$

Also $F(x) = \cancel{F(x_*)}^{0} + \int_0^1 \frac{d}{dt} F(x_* + te)\, dt$

$$= \left\{ \int_0^1 J(x_* + te)\, dt \right\} e \pm J(x_*) e$$

$$= J(x_*) e + \underbrace{\left\{ \int_0^1 [J(x_* + te) - J(x_*)]\, dt \right\}}_{} e$$

$$= J(x_*) e + E(x) e$$

Note: $\|E(x)\| \le \int_0^1 \|J(x_* + te) - J(x_*)\|\, dt$

$\|\cdot\| \le t\gamma \|e\|$

$$= \frac{\gamma}{2} \|e\|$$

So

$$e_+ = e - (J^T J)^{-1} J^T \left[Je + (J_* - J) e + Ee \right]$$

$$= e - e - \underbrace{(J^T J)^{-1} J^T}_{\text{bounded near } x_* \text{ (say } \le M)} \left[\underbrace{(J_* - J) e}_{\|\cdot\| \le \gamma \|e\|} + \underbrace{Ee}_{\|\cdot\| \le \frac{\gamma}{2}\|e\|} \right]$$

$$\|e_+\| \le M \frac{3\gamma}{2} \|e\|^2 \implies x_k \to x_* \text{ quadratically}$$

What about generalization ?

Criterion for accepting steps is as before :

$$\text{ared} \geq t \, \text{pred}$$

where

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

$$\text{pred} = f(x) - q(s)$$

where

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

not $\nabla^2 f$

$$\text{pred} = f(x) - q(s)$$

$$= -\nabla^2 f s - \frac{1}{2} s^T J^T J s$$

note :

$$q(s) = \frac{1}{2} \|F\|^2 + F^T J s + \frac{1}{2} \|Js\|^2$$

$$= \frac{1}{2} \|F + Js\|^2$$

$$\Rightarrow \text{pred} = \frac{1}{2} \left(\|F\|^2 - \|F + Js\|^2 \right)$$

and have $\text{ared} \geq t \, \text{pred}$

$$\|F\|^2 - \|F(x+s)\|^2 \geq t \left(\|F\|^2 - \|F + Js\|^2 \right)$$

(almost like $\|F\| - \|F(x+s)\| \geq t \left(\|F\| - \|F + Js\| \right)$ when $u = u$)

How to modify steps if necessary ?

Know GN step is a descent direction for f, so can implement backtracking as before.

Can also implement trust region methods as before

Have $q(s) = \frac{1}{2} \|F + Js\|^2$

so given TR radius $\delta > 0$

$$\text{TR step} = \underset{\|s\| \leq \delta}{\operatorname{argmin}} \, q(s) = \underset{\|s\| \leq \delta}{\operatorname{argmin}} \, \|F + Js\|$$

and as before have

$$\text{TR step} = s(\mu) = -\left[J^TJ + \mu I\right]^{-1} J^TF \qquad \text{---} \quad (****)$$

for a unique $\mu \geq 0$, as follows

$$\|s^{GN}\| \leq \delta \Rightarrow \text{TR step} = s^{GN} \longleftarrow -(J^TJ)^{-1}J^TF$$

$$\|s^{GN}\| > \delta \Rightarrow \text{TR step} = s(\mu) \text{ for the unique } \mu > 0$$
$$\text{such that } \|s(\mu)\| = \delta$$

Remark: Methods for nonlinear LS with step $(****)$ are called Levenberg-Marquardt methods and date back to Levenberg (1944) and Marquardt (1963). In the old days LM algorithms controlled μ. Now considered much better to control δ

Remark: Can find TR step (or dogleg approximation) as before. Note that model Hessian is J^TJ not $\nabla^2 f$ so can't see & exploit negative curvature

What if (e.g. in "large residual" or "highly nonlinear case")

$$\sum F_i \nabla^2 F_i \quad \text{is not small relative to } J^TJ$$

Evaluating $\sum F_i \nabla^2 F_i$ to get $\nabla^2 f$ is very likely infeasible

Idea: Use $J^TJ + A$, where A is maintained by quasi-Newton (secant) updating.

Usual situation: Have B and step s (possibly $-B^{-1}\nabla f$) at "current" x

set $x_+ = x+s$ and B_+ "least change" update satisfying $B_+ s = y \equiv f(x_+) - f(x)$

Now have

$$B = J^TJ + A$$

and want

$$B_+ = J_+^T J_+ + A_+$$

where

A_+ = "least change" update satisfying an appropriate secant condition (?)

straight forward condition: want

$$B_+ s = f(x_+) - f(x) \equiv y$$

so require

$$A_+ s = y - J_+^T J_+ s$$

Dennis - Schnabel suggest replacing y by $y^\#$

$$y^\# = J(x_+)^T F(x_+) - J(x)^T F(x)$$

Also suggest symmetry - preserving update analogous to Davidon - Fletcher - Powell update

$$A_+ = A + \frac{(y^\# - As)y^T + y(y^\# - As)^T}{y^T s} - \frac{(y^\# - As)^T s \, yy^T}{(y^T s)^2}$$

DGW $y^\#$ is "slight but clear winner" among several choices that all give local superlinear convergence (D-S page 232)

More recent updates analogous to Broyden - Fletcher - Goldfarb - Shanno update are now considered somewhat better (Al Bali - Fletcher)