Trust region methods

Have trust region

$$H_\delta(x) = \{ y : \| y - x \| \le \delta \}$$

choose

$$s = \underset{\|w\| \le \delta}{\operatorname{argmin}} \| F + Jw \| , \quad F = F(x), \quad J = F'(x)$$

- x is a stationary point for F if $\|F + Js\| \ge \|F\|$ for all s
- Saw: $\|s^N\| \le \delta \Rightarrow s = s^N$

  $\|s^N\| > \delta \Rightarrow \|s\| = \delta$

  etc.
- If J nonsingular, then $s = s(\mu) = -[J^TJ + \mu I]^{-1} J^T F$ for unique $\mu \ge 0$

  $\|s^N\| \le \delta \Rightarrow \mu = 0$

  $\|s^N\| > \delta \Rightarrow \mu > 0$ and such that $\|s(\mu)\| = \delta$

Also, $s(\mu)$ and $\varphi(\mu) = \|s(\mu)\|^2$ are differentiable and

$\varphi'(\mu) < 0 \Rightarrow \|s(\mu)\|$ monotone decreasing

See: $\mu = 0 \Rightarrow s(\mu) = s^N$

for μ large, $s(\mu) \approx -\frac{1}{\mu} J^T F$ - short step in steepest descent direction

$$f(x) = \tfrac{1}{2}\|F(x)\|^2 = \tfrac{1}{2} F^T F$$

$$\nabla f = J^T F$$

$$f(x+s) = \tfrac{1}{2} F(x+s)^T F(x+s)$$

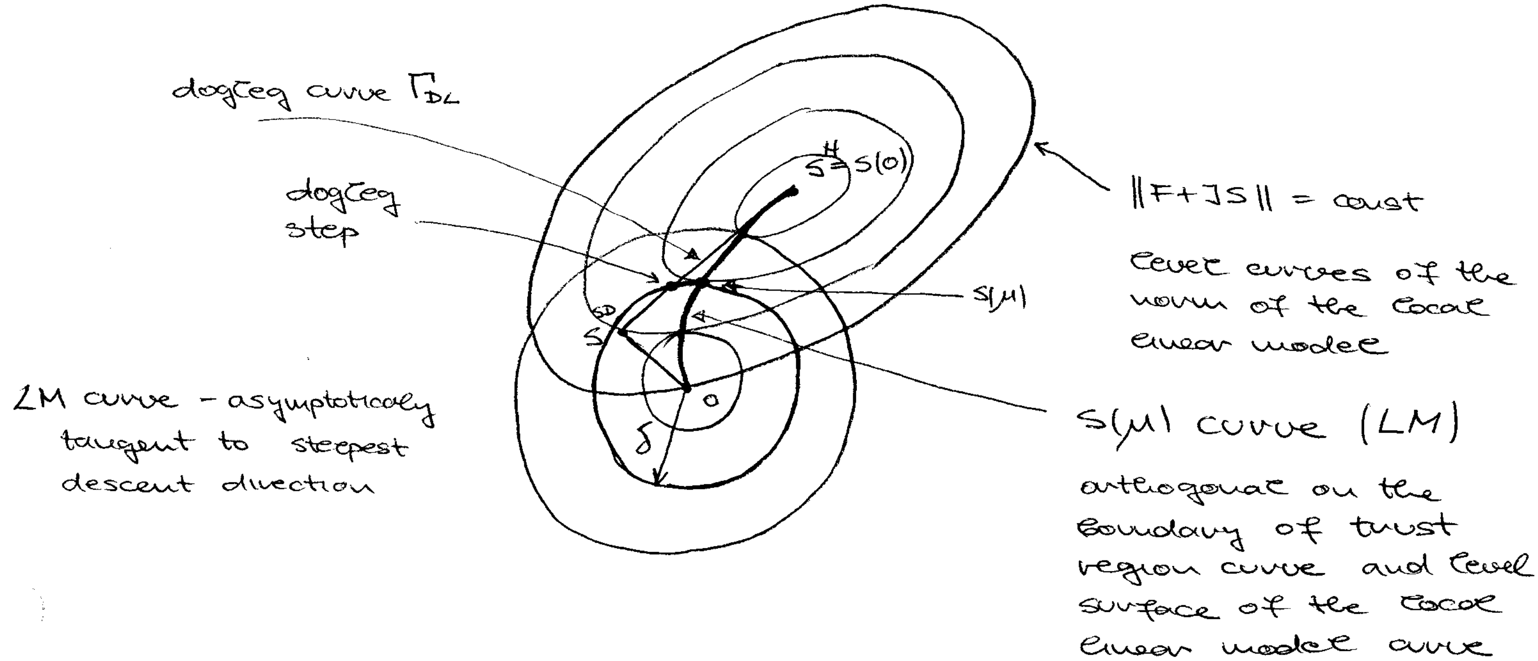
$$= \tfrac{1}{2}\|F\|^2 + \underbrace{F^T J s}_{(J^TF)^T s} + O(s)$$

Assuming $\|.\| = \|.\|_2$ througbout

Fundamental practical difficulty:

can't compute $s(\mu)$ such that $\|s(\mu)\| = \delta$ both accurately and efficiently

Approach 1 (the Levenberg-Marquardt approach)



Compute $s(\mu)$ <u>exactly</u> on LM curve such that $\|s(\mu)\|$ <u>approximately</u> equals δ

Set $\Phi(\mu) = \|s(\mu)\| - \delta$

want μ_* such that $\Phi(\mu_*) = 0$

Problem: Computing $s(\mu) = -[J^TJ + \mu I]^{-1} J^TF$

cost up to $O(n^3)$ arithmetic to factor $J^TJ + \mu I$ each time we change μ

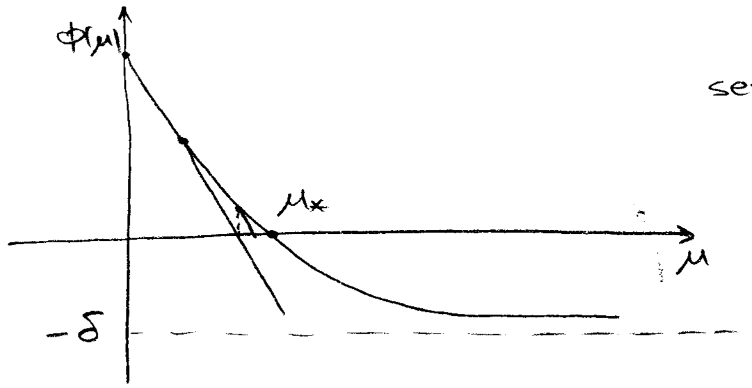
Consider Newton's method. Have

$$\Phi'(\mu) = \frac{-s(\mu)^T [J^TJ + \mu I]^{-1} s(\mu)}{\|s(\mu)\|}$$

In computing $s(\mu)$ already have factored $J^TJ+\mu I$ so can evaluate $[J^TJ+\mu I]^{-1}s(\mu)$ and $\phi'(\mu)$ in $O(n^2)$ arithmetic.

- Already know $\|s(\mu)\|$ is monotone decreasing
  $\Rightarrow$ $\Phi(\mu)$ is monotone decreasing

Also, can show $|\Phi'(\mu)|$ is monotone decreasing



see: Newton step always undershoots μ_*

Newton step takes local linear model $\Phi + \Phi s$ to zero

Consider new model

$$m(\sigma) = \frac{\alpha}{\beta+\mu+\sigma} - \delta$$

choose α, β such that

$$m(0) = \phi(\mu)$$
$$m'(0) = \phi'(\mu)$$

Get:

$$\alpha = -\frac{[\phi(\mu)+\delta]^2}{\phi'(\mu)} \qquad \beta = \frac{-\phi(\mu)+\delta}{\phi'(\mu)} - \mu$$

choose σ such that $m(\sigma)=0$

Get

$$\sigma = \frac{\alpha}{\delta} - (\beta+\mu) = \frac{\|s(\mu)\|}{\delta}\underbrace{\left\{-\frac{\phi(\mu)}{\phi'(\mu)}\right\}}_{\text{Newton step } \sigma^N}$$

see this has desired effect of

$$\delta = \begin{cases} \text{longer than } \delta^H & \text{if } \mu < \mu^* \\ \text{shorter than } \delta^H & \text{if } \mu > \mu^* \end{cases}$$

Also as $\mu \to \mu_*$ $\quad \frac{\|s(\mu)\|}{\delta} \to 1$ and $\delta \to \delta^H$

can show quadratic convergence to μ_*

- Dennis & Schnabel recommend stopping once

$$\frac{3}{4}\delta \leq \|s(\mu)\| \leq \frac{3}{2}\delta$$

Approach 2 (The dogleg approach)

- Find step s such that $\|s\| = \delta$ exactly but s is only approximately on the $s(\mu)$ - curve
- Construct dogleg curve connecting $s=0$, $s=s^{SD}$, $s=s^H$

Here s^{SD} - "steepest descent step"

s^{SD} minimizer of $\|F+Js\|$ for $s = \lambda\left(-\nabla \frac{1}{2}\|F\|^2\right)$

$$= -\lambda \underbrace{J^T F}_{\substack{\text{steepest} \\ \text{descent} \\ \text{dir. for } \|F\|}}$$

see: $\psi(\lambda) = \frac{1}{2}\|F + J(\lambda s)\|^2$, $s = -J^T F$

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

$$0 = \psi'(\lambda) = F^T J s + \lambda \|Js\|^2$$

$$\Rightarrow \lambda = -\frac{F^T J s}{\|Js\|^2} \sim \text{easy to compute}$$

see: Γ_{DL} ends at s^H

- A short step (corresponding to small δ) along Γ_{DL} is in steepest-descent direction (similar to $s(\mu)$ for large μ)

Can show, as s traverses Γ_{DL} from 0 to s^{SD} to s^{N}

(1) $\|s\|$ is monotone increasing

(2) $\|F+Js\|$ is monotone decreasing

So $\exists$ unique point s^{DL} (the dogleg step) where Γ_{DL} intersects the TR boundary and

$$s^{DL} = \underset{\substack{s \in \Gamma_{DL} \\ \|s\| \leq \delta}}{\arg\min} \|F+Js\|$$

Computing s^{DL}

(1) if $\|s^{N}\| \leq \delta \Rightarrow s^{DL} = s^{N}$

(2) if $\|s^{N}\| > \delta$, compute s^{SD}

(3) if $\|s^{SD}\| \geq \delta$, then $s^{DL} = \dfrac{\delta}{\|s^{SD}\|} s^{DL}$

(4) if $\|s^{SD}\| < \delta$, then compute $s^{DL} = s^{SD} + \tau(s^{N} - s^{SD})$ for unique τ in $(0,1)$ such that $\|s^{DL}\| = \delta$

Finding τ: want

$$\|s^{SD} + \tau(s^{N} - s^{SD})\|^2 - \delta^2 = 0$$

$$\left(\|s^{SD}\|^2 - \delta^2\right) + 2 s^{SD^T}(s^{N} - s^{SD})\tau + \|s^{N} - s^{SD}\|\tau^2 = 0$$

or

$$a\tau^2 + 2b\tau + c = 0$$

see

$c < 0$, $b > 0$ since $\|s\|$ is monotone decreasing as s goes from 0 to s^{SD} to s^{N}

and $a > 0$

So $$\tau_{\pm} = \frac{-b \pm \sqrt{b^2 - ac}}{a}$$

want $\tau = \tau_+$ since we want $0 < \tau < 1$

$$\tau = \frac{-b + \sqrt{b^2 - ac}}{a}$$

to avoid cancellation, use

$$\tau = \frac{-b + \sqrt{b^2-ac}}{a} \cdot \frac{b + \sqrt{b^2-ac}}{b+\sqrt{b^2-ac}} = \frac{\cancel{-b^2}+\cancel{b^2} - ac}{\cancel{a}\left(b+\sqrt{b^2-ac}\right)}$$

$$\tau = \frac{-c}{b + \sqrt{b^2-ac}}$$

Considerations for optimization

- Want to minimize $f: \mathbb{R}^n \to \mathbb{R}^1$

  Use Newton + globalization to solve $\nabla f(x) = 0$

- Want to reduce f at each step so base acceptability etc. on local quadratic model

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

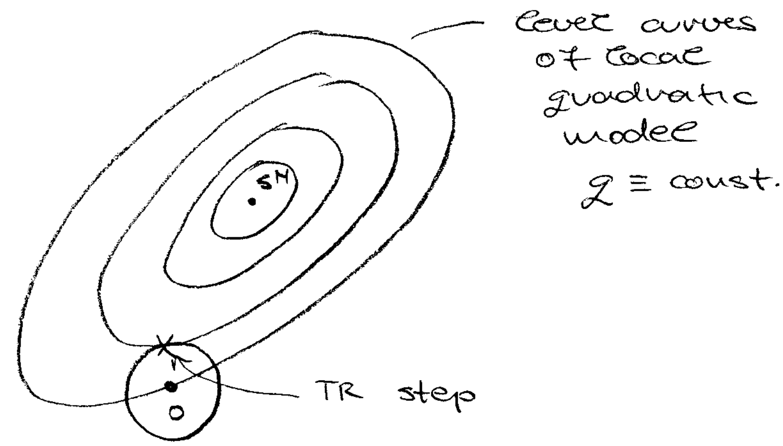
recall: $s^N = -\nabla^2 f \nabla f$ is guaranteed to be descent step (for local quadratic model) only if $\nabla^2 f$ is SPD. If necessary perturbe $\nabla^2 f$ to $B = \nabla^2 f + \mu I$ where $\mu > 0$ such that B is SPD and well-conditioned

For a TR method, choose

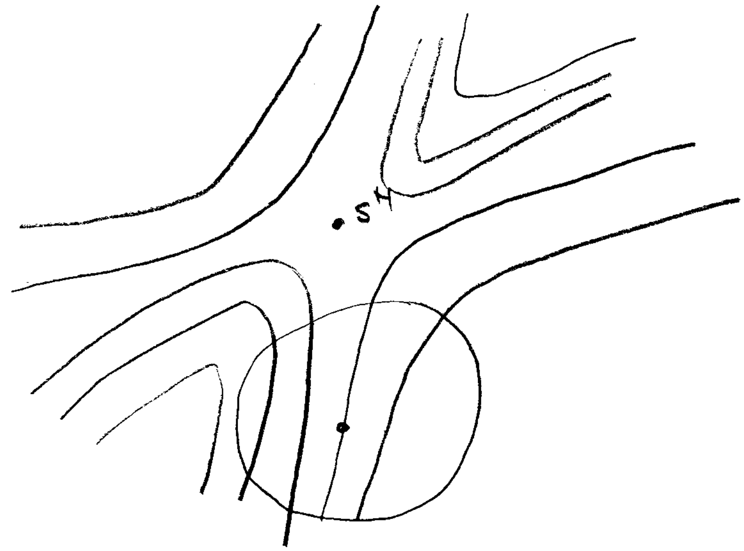
$$s = \underset{\|w\| \le \delta}{\operatorname{argmin}}\; q(w)$$

New issue: If $\nabla^2 f$ indefinite (negative curvature - - some of the eigenvalues are negative) then q has no global minimizer and local minimization within TR may be complicated

$\nabla^2 f$ - SPD



$\nabla^2 f$ - has one negative
and one positive
eigenvalue



Can show: TR step $= -\left(\nabla^2 f + \mu I\right)^{-1} \nabla f$

where $\mu \geq 0$ is such that $\nabla^2 f + \mu I$ is SPD

Traditional approach (see Dennis & Schnabel)

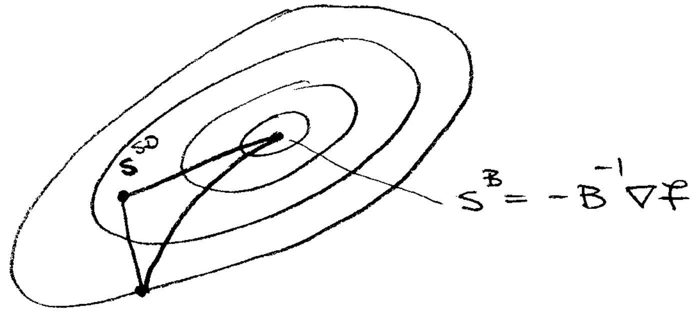
- Perturb if necessary to get $B = \nabla^2 f + \mu I$, $\mu \geq 0$
  that is "safely" SPD.

  Then take $q(s) = f + \nabla f^T s + \frac{1}{2} s^T B s$

  Then have as before

  LM step $s(\mu) = -\left(B + \mu I\right)^{-1} \nabla f$

steepest descent
step does not
depend on Hessian
just on first
order derivative



Dogleg curve: $0 \to s^{SD} = \underset{\substack{s=-\lambda\nabla f \\ \text{for } \lambda\geq 0}}{\arg\min}\ \mathcal{q}(s)$

$\to s^{B}$

(same important properties as before)

So as before, can compute

- LM step $s(\mu)$ such that $\|s(\mu)\| \approx \delta$
- dogleg step: s on dogleg curve with $\|s\| = \delta$

But this $(\nabla^2 f \leftarrow B = \nabla^2 f + \mu I)$ doesn't take advantage of negative curvature

How to do this?

- still have $s = s(\mu) = -(\nabla^2 f + \mu I)^{-1} \nabla f$ for $\mu \geq 0$ such that $\nabla^2 f + \mu I$ is SPD and $\|s(\mu)\| = \delta$
- since $\nabla^2 f$ is symmetric, have $\nabla^2 f = U \Lambda U^T$ where $\Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$; $U^T U = I$ and $\lambda_1 \geq \ldots \geq \lambda_n$

λ_n - the most negative eigenvalue

Have $\nabla^2 f + \mu I = U \Lambda U^T + \mu \underbrace{U U^T}_{I}$

$= U(\Lambda + \mu I) U^T$

$$\Rightarrow \quad (\nabla^2 f + \mu I)^{-1} = U \begin{pmatrix} \frac{1}{\lambda_1+\mu} & & O \\ & \ddots & \\ O & & \frac{1}{\lambda_n+\mu} \end{pmatrix} U^T$$

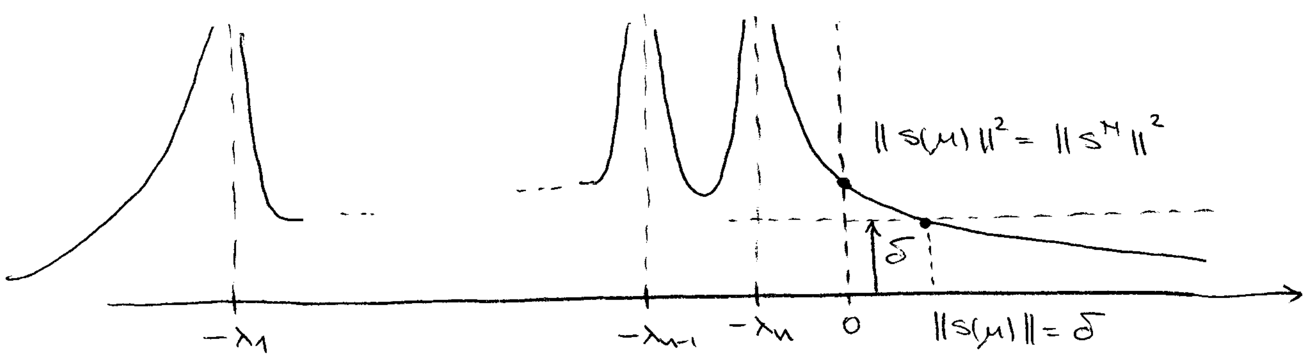
$$\Rightarrow \quad s(\mu) = -(\nabla^2 f + \mu I)^{-1} \nabla f = U \begin{pmatrix} \frac{\gamma_1}{\lambda_1+\mu} \\ \vdots \\ \frac{\gamma_n}{\lambda_n+\mu} \end{pmatrix}$$

where $\begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_n \end{pmatrix} = -U^T \nabla f$

so γ_i is component of gradient in direction of i^{th} column of U

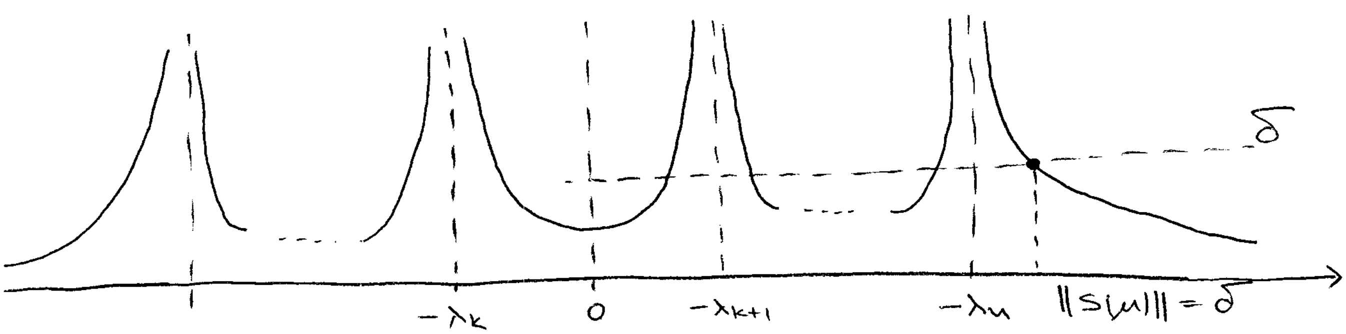
$$\|s(\mu)\|^2 = \left\| U \begin{pmatrix} \frac{\gamma_1}{\lambda_1+\mu} \\ \vdots \\ \frac{\gamma_n}{\lambda_n+\mu} \end{pmatrix} \right\|^2 = \sum_{i=1}^{n} \frac{\gamma_i^2}{(\lambda_i+\mu)^2}$$

Suppose $\nabla^2 f$ is SPD. Then $\lambda_1 > \dots > \lambda_n > 0$ and $\|s(\mu)\|^2$ is a nice (analytic) function of μ for $\mu \geq 0$ but has poles at $-\lambda_n \dots -\lambda_1$



Suppose $\nabla^2 f$ is indefinite

$$\lambda_1 > \dots > \lambda_k > 0 > \lambda_{k+1} > \dots > \lambda_n$$



Can compute $s(\mu)$ accurately but may be undesirably expensive