More on Newton's method

Considerations for optimization

- Suppose we want to minimize $f: \mathbb{R}^n \to \mathbb{R}^1$

Adopt Newton's method by applying it to

$$\nabla f(x) = 0 \qquad (*)$$

Note: solution of $(*)$ may be maximizer, saddle points as well as minimizers or none of those

Later: develop procedures that make convergence to anything but a minimizer unlikely

Note: Newton iteration

$$x_+ = x - \nabla^2 f(x)^{-1} \nabla f(x)$$

where

$$\nabla^2 = \nabla f'(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \ddots & & \vdots \\ \vdots & & \ddots & \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

Essentially always have

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

$\Rightarrow \nabla^2 f(x)$ symmetric

Usually, $\nabla^2 f(x)$ is also positive definite near a local minimum.

Then solving $\nabla^2 f(x) \cdot s = -\nabla f(x)$

In full-matrix case, can solve with Cholesky decomposition

- In sparse case, can use PCG

- Nice property of Newton's method on discretized differential or integral equations: Mesh independence, # of iterations required for given level of accuracy does not increase much as mesh is refined. Often seen in practice. Few theoretical results (Aggomer ref.)

- Another property of Newton's method: Newton's method enjoys a certain scale independence

  Suppose we rescale F:

$$\bar{F} \leftarrow \bar{F} = BF \quad , \quad B \in \mathbb{R}^{n \times n} \text{ nonsingular}$$

  then

$$x_+ = x - \bar{F}'(x)^{-1} \bar{F}(x)$$

$$= x - \left(BF'(x)\right)^{-1} BF(x)$$

$$= x - F'(x)^{-1} F(x)$$

  Suppose we rescale X:

$$\hat{x} = Ax \quad ; \quad A \in \mathbb{R}^{n \times n} \text{ nonsingular}$$

  set $\hat{F}(\hat{x}) = F(A^{-1}\hat{x})$

$$\Rightarrow \hat{F}'(\hat{x}) = F'(A^{-1}\hat{x}) A^{-1}$$

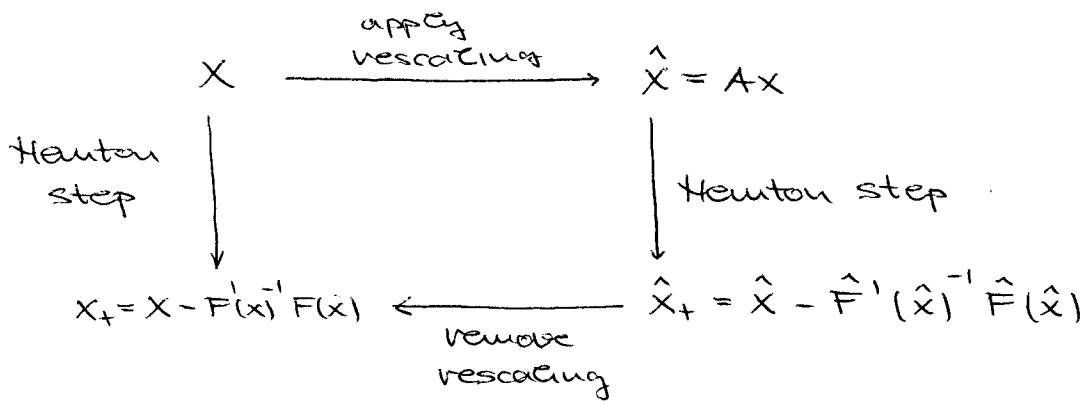
$$\hat{x}_+ = \hat{x} - \hat{F}(\hat{x})^{-1} \hat{F}(\hat{x})$$

$$= \hat{x} - AF'(A^{-1}\hat{x})^{-1} \hat{F}(\hat{x})$$

$$Ax_+ = Ax - AF'(x)^{-1} F(x)$$

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

Schematically



- But scaling may be important to various aspects of practical implementations
- For example, rescale stopping tests

$$\|F(x)\| \leq tol_F \qquad \|s\| \leq tol_x \quad , \quad s = -F'(x)^{-1} F(x)$$

- If components of F or s are badly scaled they may not be effective as desired
- Also, bad scaling may cause problems in solving

$$F'(x) \cdot s = -F(x)$$

Also, bad scaling may cause problems with "globalization" (later)

What to do about bad scaling?

Idea: Want components of F and components of x to be about the same in magnitude

May be able to find diagonal matrices D_x, D_F with diagonal entries that are typical of scales of x- and F- components

Then replace x, F by $D_x^{-1} \cdot x$ and $D_F^{-1} \cdot F$

Some variants of Newton's method

- Finite difference Newton's method

- Often evaluating $F'(x)$ is undesirable or even impossible

Idea: approximate $F'(x)$ with finite differences

Forward difference approximation:

j^{th} column of $F'(x)$

$$F'(x) = \frac{\partial F(x)}{\partial x_j} = \frac{1}{h}\left\{F(x+he_j) - F(x)\right\} + \mathcal{O}(h)$$

Central difference approximation

$$F'(x) = \frac{\partial F(x)}{\partial x_j} = \frac{1}{2h}\left\{F(x+he_j) - F(x-he_j)\right\} + \mathcal{O}(h^2)$$

- Forward diff. is cheaper, central diff. is more accurate for small h

see: approximating $F'(x)$ requires n F-evaluations (forward diff) or $2n$ F-evaluations (central diff.)

- If F' is sparse, can often greatly reduce F-evaluations with Curtis-Powell-Reid "trick"

Illustrate: Suppose

$$F'(x) = \begin{pmatrix} x & x & & & \\ x & \ddots & \ddots & & \\ & \ddots & \ddots & x & \\ & & x & x & \end{pmatrix} \quad \text{tridiagonal}$$

$\Leftrightarrow$ $F_i(x)$ depends only on x_{i-1}, x_i, x_{i+1}

see: columns 1, 4, 7, 10, ... of F' have no common nonzero components

columns 2, 5, 8, 11, ... ——— " ———

columns 3, 6, 9, 12, ... ——— " ———

So Forming

$$\frac{1}{h}\left\{ F(x+h(e_1+e_4+e_7+\dots)) - F(x) \right\}$$

gives non-zero components of

1st column of F' in components 1-2

4th column ——————"—————— 3-5

7th column ——————"—————— 6-8

Forming

$$\frac{1}{h}\left\{ F(x+h(e_2+e_5+e_8+\dots) - F(x) \right\}$$

gives non-zero components of

2nd column of F' in components 1-3

5th ——————"—————— 4-6

8th ——————"—————— 7-9

Finally form

$$\frac{1}{h}\left\{ F(x+h(e_3+e_6+e_9+\dots)) - F(x) \right\}$$ to get

non-zero entries of columns 3, 6, 9

So get F.D. approximation of all $F'(x)$ with just 3 F-eval.

Aside: These days automatic differentiation may be useful in evaluating $F'(x)$ when it is otherwise unavailable

What about convergence of FD Newton?

Look at convergence of very general Newton-Like Method

given x_0 and B_0

for $k = 0, 1, \dots$

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

determine B_{k+1}

Standard Assumption: Suppose that F is continuously differentiable near x_* such that $F(x_*)=0$ and $F'(x_*)$ is nonsingular

Theorem: Suppose the standard assumption holds and that $\{x_k\}$ is produced by NLM with $\{B_k\}$ such that for all k

① $\|B_k^{-1}\| \leq m$

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

Then for any ϵ such that $e_{max} < \epsilon < 1$ there is a $\delta > 0$ such that if

$$\|x_0 - x_*\| \leq \delta$$

then $x_k \longrightarrow x_*$ with

$$\|x_{k+1} - x_*\| \leq \epsilon \|x_k - x_*\|$$

Moreover if

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

the convergence is superlinear.

proof idea:

set $e_k = x_k - x_*$, for each k, $J_* = F'(x_*)$; $J(x) = F'(x)$

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

$$\Rightarrow e_{k+1} = e_k - B_k^{-1}\Big\{ F(x_*) + \underbrace{J_* e_k}_{[J(x_k) + J_* - J(x_k)]} + \theta(e_k) \Big\}$$

$$= \Big[I - B_k^{-1} J(x_k)\Big] e_k + B_k^{-1}\Big[J(x_k) - J_*\Big] e_k + B_k^{-1}\theta(e_k)$$

$$\|e_{k+1}\| \leq \underbrace{\|\Big[I - B_k^{-1} J(x_k)\Big] e_k\|}_{\leq e_{max}\|e_k\|} + \underbrace{\|B_k^{-1}\Big[J(x_k) - J_*\Big] e_k\|}_{\downarrow} + \underbrace{\|B_k^{-1}\theta(e_k)\|}_{M\frac{\|\theta(e_k)\|}{\|e_k\|}\|e_k\|} \quad \cdots (*)$$

$M\theta(\|e_k\|)\|e_k\|$

so get

$$\|e_{k+1}\| \leq \left\{ \epsilon_{max} + M\,\Theta(\|e_k\|) + M\,\frac{\|\Theta(e_k)\|}{\|e_k\|} \right\} \|e_k\|$$

Pick $\delta > 0$ so small that $\{\ldots\} \leq \epsilon$ whenever

$$\|x_k - x_*\| \leq \delta$$

Easy induction: if $\|x_0 - x_*\| \leq \delta$ then $\|x_k - x_*\| \leq \delta$ for all k and $\|x_{k+1} - x_*\| \leq \epsilon \|x_k - x_*\|$

from (*), see $\overline{\lim_{k\to\infty}} \dfrac{\|e_{k+1}\|}{\|e_k\|} = \overline{\lim_{k\to\infty}} \|I - B_k^{-1} F'(x_k)\|$

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

then

$$\lim_{k\to\infty} \frac{\|e_{k+1}\|}{\|e_k\|} = 0 \qquad \text{(superlinear convergence)}$$

Apply Theorem to convergence of FD Newton

Have B_k = FD approximation of $F'(x_k)$

Disregarding rounding and function evaluation errors can make $\|I - B_k^{-1} F'(x_k)\|$ arbitrarily small by taking h sufficiently small.

Theorem: Suppose standard assumptions hold. Then for $\epsilon > 0$ $\exists\, \delta > 0$ and $\eta > 0$ such that if $\|x_0 - x_*\| < \delta$ and $|h| < \eta$ in all FD approximations of F', then the FD Newton iterates x_k converge to x_* with

$$\|x_{k+1} - x_*\| \leq \epsilon \|x_k - x_*\|$$

for each k, if h_k = difference step at k^{th} iteration $\longrightarrow 0$ as $k \to \infty$ then convergence is superlinear.

In practice, pick some small h and use it for all iterations

What should h be?

Taking h smaller makes truncation error smaller

(truncation error

$O(h)$ - for forward diff.

$O(h^2)$ - for central diff.)

But taking h too small makes larger the errors from F-evaluations and roundoff.

To illustrate this

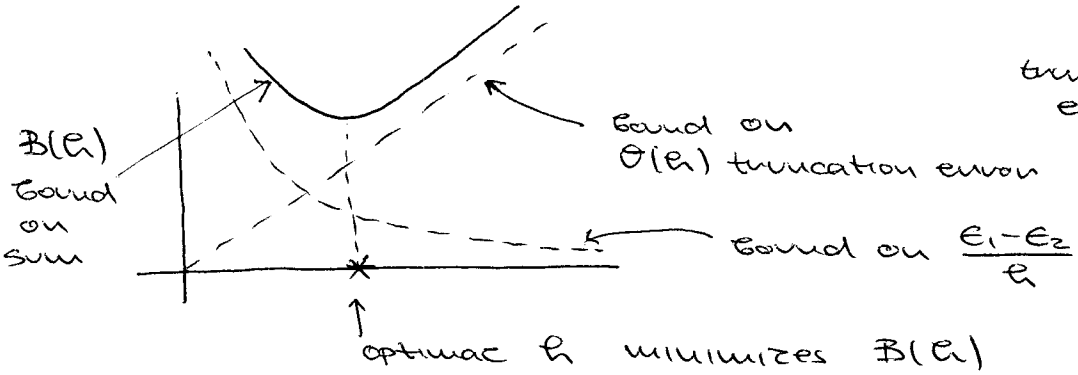
$$\frac{F(x+he_j) - F(x)}{h}$$

errors in F evaluation

$$\frac{F(x+he_j) + \epsilon_1 - F(x) - \epsilon_2}{h} = F'(x) + O(h) + \frac{\epsilon_1 - \epsilon_2}{h}$$

truncation error

error from inaccurate F-eval.



For scalar function can derive a bound on error that is minimized by "optimal h"

Extending to vector-valued function is "shaky"

I like the following (from Pernice-Walker 1997)

ϵ_F - bound on relative error in computed F-values (function epsilon)

$F_{computed} = F + e$ $\qquad \frac{\|e\|}{\|F\|} \leq \epsilon_F$

$h = \sqrt{(1+\|x\|)\,\epsilon_F}$ for forward differences

$h = \sqrt[3]{(1+\|x\|)\,\epsilon_F}$ for central differences

main assumption F, F' and F'' are have the same scale

If $\epsilon_F \approx 10^{-k}$ $\approx k$ accurate digits in computed F

expect $\approx \frac{k}{2}$ accurate digits in forward diff. approx.

$\approx \frac{2k}{3}$ ——//—— in central ——//——

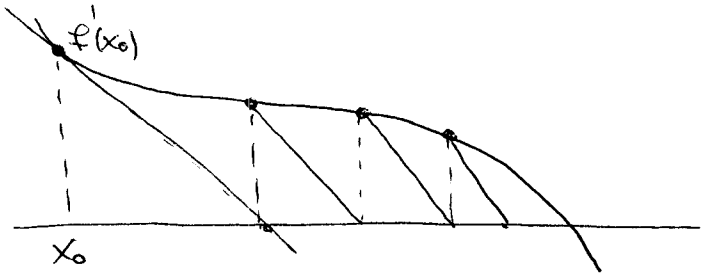
Often just take

$$h = \begin{cases} \epsilon^{1/2} & \text{for forward. diff.} \\ \epsilon^{1/2} & \text{for central diff.} \end{cases}$$

If F evaluation is unlikely to incur much error

- Usually, FD Newton converges about as rapidly as Newton (but not always)

Chord method



given initial x
evaluate $F(x)$ stop ?
evaluate $J = F'(x)$ $\qquad (*)$
until termination do
$\quad x \leftarrow x - J^{-1}F(x)$
$\quad$ evaluate $F(x) \longrightarrow$ stop

Chords advantage:

- only one F' evaluation
- also if using direct linear algebra method then only factor J once at (*) and re-use factors at (**)

Theorem: Suppose standard assumption holds. Then for any $\epsilon > 0$ $\exists \delta > 0$ such that if $\| x_0 - x_* \| \leq \delta$ then Chord method iterates converge to x_* with

$$\| x_{k+1} - x_* \| \leq \epsilon \| x - x_* \|$$

for each k

proof: Follows from convergence theorem for NLM

Chord method (a.k.a. "modified Newton" method) very successful in contexts with <u>good</u> initial guess (e.g. stiff ODE solvers)

Shamanski's method

```
given x
evaluate F(x) stop
until termination do
    evaluate J = F'(x)
    for j = 1, ... k do
        x ← x - J^{-1} F(x)
        evaluate F(x) stop
```

- get local superlinear convergence from Theorem for NLM