Newton's Method in Mixed Precision

C. T. Kelley
NC State University
tim_kelley@ncsu.edu
Supported by NSF, ARO

RTG, September 2019

Outline

Outline

Nonlinear Equations

Objective: solve

$$\mathbf{F}(\mathbf{x}) = 0$$

where

$$\mathbf{F} = (f_1, f_2, \dots, f_N)^T.$$

Newton's method is

$$\mathbf{x}_+ = \mathbf{x}_c - \mathbf{F}'(\mathbf{x}_c)^{-1}\mathbf{F}(\mathbf{x}_c).$$

Jacobian:

$$(\mathbf{F}')_{ij} = \partial f_i / \partial x_j$$

Local Convergence to distinguished root $\mathbf{x}^*$

Standard assumptions for local convergence:
There is $\mathbf{x}^* \in D$ such that

- $\mathbf{F}(\mathbf{x}^*) = 0$,
- $\mathbf{F}'(\mathbf{x}^*)$ is nonsingular, and
- $\mathbf{F}'(\mathbf{x})$ is Lipschitz continuous with Lipschitz constant γ, i. e.

$$\|\mathbf{F}'(\mathbf{x}) - \mathbf{F}'(\mathbf{y})\| \leq \gamma \|\mathbf{x} - \mathbf{y}\|,$$

for all $\mathbf{x}, \mathbf{y} \in D$.

Rules for talking about Newton's method

- $\mathbf{x}^*$ is the solution in SA
  which may not be the one you want
- $\mathbf{e} = \mathbf{x} - \mathbf{x}^*$ is the error
- Convergence theorems in terms of change from
  - current iteration $\mathbf{x}_c$ to
  - next iteration $\mathbf{x}_+$

Famous local convergence theorem

Assume that the standard assumptions hold, $\mathbf{x}_c \in D$, and that

$$\|\mathbf{e}_c\| \le \frac{1}{2\|\mathbf{F}'(\mathbf{x}^*)^{-1}\|\gamma}.$$

Then

$$\|\mathbf{F}'(\mathbf{x}^*)^{-1}\|/2 \le \|\mathbf{F}'(\mathbf{x}_c)^{-1}\| \le 2\|\mathbf{F}'(\mathbf{x}^*)^{-1}\|.$$

Moreover, if $\mathbf{e}_+$ is the Newton iterate from $\mathbf{x}_c$ then

$$\|\mathbf{e}_+\| \le \gamma\|\mathbf{F}'(\mathbf{x}^*)^{-1}\|\|\mathbf{e}_c\|^2 \le \|\mathbf{e}_c\|/2.$$

For the entire iteration . . .

Corollary: Assume that the standard assumptions hold, $\mathbf{x}_0 \in D$, and that

$$\|\mathbf{e}_0\| \leq \frac{1}{2\|\mathbf{F}'(\mathbf{x}^*)^{-1}\|\gamma}.$$

Then the

- Newton iteration exists (i. e. $\mathbf{F}'(\mathbf{x}_n)$ is nonsingular for all n),
- converges to $\mathbf{x}^*$, and
- the convergence is q-quadratic

$$\|\mathbf{e}_{n+1}\| = O(\|\mathbf{e}_n\|^2)$$

What does this mean?

In an ideal world where

- precision is infinite,
- derivatives are analytic,
- linear solvers are exact,

Newton's method works great with good initial data.
But . . .

. . . you'll be doing it wrong.

In practice, you get

$$\mathbf{x}_+ = \mathbf{x}_c - \mathbf{J}_c^{-1}(\mathbf{F}(\mathbf{x}_c) + \mathbf{E}_c)$$

where

- $\mathbf{J}_c \approx \mathbf{F}'(\mathbf{x}_c)$ (maybe badly)
- $\mathbf{E}_c$ is the (usually small) error in $\mathbf{F}$

A less famous theorem

Same assumptions as for Newton plus

$$\|\mathbf{F}_c - \mathbf{F}'(\mathbf{x}_c)\| \le \frac{1}{4\|F'(x^*)^{-1}\|}.$$

Then J_c is nonsingular and $\mathbf{x}_+$ satisfies

$$\|e_+\| \le \|\mathbf{F}'(\mathbf{x}^*)^{-1}\| \left(\gamma \|e_c\|^2 + 6\|\mathbf{J}_c - \mathbf{F}'(\mathbf{x}_c)\| \|e_c\| + 8\|\mathbf{E}_c\| \right).$$

Local Improvement Theorem

Same assumptions as for Newton and, for all n,

$$\|\mathbf{J}_n - \mathbf{F}'(\mathbf{x}_n)\| \le \frac{1}{4\|F'(x^*)^{-1}\|}.$$

and

$$\|\mathbf{E}_n\| \le \epsilon_F.$$

Then

$$\|\mathbf{e}_{n+1}\| = O(\|\mathbf{e}_n\|^2 + \|\mathbf{J}_n - \mathbf{F}'(\mathbf{x}_n)\|\|\mathbf{e}_n\| + \epsilon_F).$$

The theorem does not predict convergence, rather stagnation.

Examples

- $\epsilon_F = 0$, $\mathbf{J}_n = \mathbf{F}(\mathbf{x}_n)$: Newton
- $\epsilon_F > 0$, floating point error: Newton in practice
- $\epsilon_F > 0$, $\mathbf{J}_n$ finite difference Jacobian, step h
  - Use optimal $h = \sqrt{\epsilon_F}$ and
  - $\|\mathbf{e}_{n+1}\| = O(\|\mathbf{e}_n\|^2 + h\|\mathbf{e}_n\| + \epsilon_F)$
  - Same behavior as Newton until stagnation.
- $\epsilon_F > 0$, $\mathbf{J}_n = \mathbf{F}'(\mathbf{x}_0)$, chord method

Implementation: ignore ϵ_F

Intialize $\mathbf{x}_0$, $n = 0$, termination criteria
while Not happy **do**
  Evaluate $\mathbf{F}(\mathbf{x}_n)$; terminate?
  Evaluate $\mathbf{J}_n \approx \mathbf{F}'(\mathbf{x}_n)$
  Solve $\mathbf{J}_n \mathbf{s} = -\mathbf{F}(\mathbf{x}_n)$
  $\mathbf{x}_{n+1} = \mathbf{x}_n + \mathbf{s}$
end while

Genius Idea!

- Store **J** in reduced precision.
- Solve in reduced precision.
  - Cut $O(N^2)$ storage by factor of 2 (single)
  - Cut $O(N^3)$ work by factor of 2 (single)
- How can you lose? Why isn't this in all the books?

The case in this talk

ϵ_F floating point double precision roundoff

- ϵ_F floating point double precision roundoff
- $\mathbf{J}_c = \mathbf{J}_N + \mathbf{J}_{be}$ where
- Solver is double, single, or half precision LU
  - $\mathbf{J}_N$ is the nominal approximation you give the linear solver $\mathbf{F}'(\mathbf{x}_c)$ in double or finite-difference approximation
  - The solver returns the solution of $\mathbf{J}_{be}\mathbf{s} = -\mathbf{F}(\mathbf{x}_c) - \mathbf{E}_c$

So the less famous theorem says ...

$$\|\mathbf{e}_{n+1}\| = O\left(\|\mathbf{e}_n\|^2 + (\|\mathbf{J}_{Nn} - \mathbf{F}'(\mathbf{x}_n)\| + \|\mathbf{J}_{Nn} - \mathbf{J}_{be}\|)\|\mathbf{e}_n\| + \epsilon_F\right).$$

The Jacobian you think you have is harmless

- Analytic Jacobian: $\|\mathbf{J}_{Nn} - \mathbf{F}'(\mathbf{x}_n)\| + O(\epsilon_F)$
- Difference Jacobian: $\|\mathbf{J}_{Nn} - \mathbf{F}'(\mathbf{x}_n)\| + O(\epsilon_F^{1/2})$
- But what about the backward error?
- Large backward error $\rightarrow$ slow nonlinear convergence.
  Can we see this numerically?

What is that backward error?

The standard thing you get in school is from, for example

- J. W. Demmel, Applied Numerical Linear Algebra, SIAM, Philadelphia, 1997.

If you're solving $\mathbf{Ax} = \mathbf{b}$ and the solver shows up with

$$(\mathbf{A} + \delta\mathbf{A})\mathbf{x} = \mathbf{b}$$

then page 49 says $\|\delta\mathbf{A}\|_\infty \le 3 g_{PP} N^3 \epsilon_S \|\mathbf{A}\|_\infty$, where

- g_{PP} is the growth factor and
- ϵ_S is the unit roundoff in the precision of the solver.

What does this mean?

Suppose $g_{PP} = 1$, you are still in trouble if N is large.
$N^3 \epsilon_S = O(1)$ if

- (double): $\epsilon_S = 10^{-16}$, $N \approx 2 \times 10^5$
- (single): $\epsilon_S = 10^{-8}$, $N \approx 5 \times 10^2$
- (half): $\epsilon_S = 10^{-4}$, $N \approx 22$

These results are clearly silly. What's up?

Details

- Nicholas J. Higham, Accuracy and Stability of Numerical Algorithms, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1996.

Page 175-177: Componentwise backward error (ignore permutation matrix)

$$|\delta \mathbf{A}| \le 2\gamma_N |\hat{L}||\hat{U}|$$

where $\hat{\mathbf{L}}\hat{\mathbf{U}} = \mathbf{A} + \delta \mathbf{A}$ and

$$\gamma_N = \frac{N\epsilon_S}{1 - N\epsilon_S}$$

Did the N^3 go away?

Nope!
The growth factor part is

$$\hat{g}_{PP} = \frac{\max |\hat{\mathbf{U}}_{ij}|}{\max |\mathbf{A}_{ij}|}$$

So

- $|L_{ij}| \le 1$ implies (worst case) $\|\hat{L}\|_\infty \le N$
- Define $\hat{g}_{PP}$ by $|\hat{U}_{ij}| \le \hat{g}_{PP} |\mathbf{A}_{ij}|$ and so

$$\|\hat{U}\|_\infty \le \hat{g}_{PP} N \|\mathbf{A}\|_\infty.$$

More N^3

- Bottom line:

$$\|\mathbf{J}_N - \mathbf{J}_{be}\|_\infty \le 2N^2 \gamma_N \hat{g}_{PP} \|\mathbf{J}_{be}\|_\infty.$$

- The N^3 is from

$$N^2 \gamma_N = \frac{N^3 \epsilon_S}{1 - N\epsilon_S}$$

Are we doomed?

Nope!

In many cases $|\hat{\mathbf{L}}||\hat{\mathbf{U}}| \le C|A|$

- $\mathbf{A}$ symmetric
- Totally positive $\mathbf{A}$ (so $L_{ij} \ge 0$ and $U_{ij} \ge 0$)

So, in the perfect world where

- $|\hat{\mathbf{L}}||\hat{\mathbf{U}}| \le C|A|$ and
- $g_{PP} = O(1)$,

$$\|\mathbf{J}_N - \mathbf{J}_{be}\|_\infty = O(N\epsilon_S)?$$

Probably even better ...

- N. J. Higham and T. Mary, A new approach to probabilistic rounding error analysis, Tech. Report 2018.33, Manchester Institute for Mathematical Sciences, School of Mathematics, The University of Manchester, 2018.
- I. C. F. Ipsen and H. Zhou, Probabilistic error analysis for inner products, 2019.

Big assumption: rounding errors are independen

Some people in the room do not believe this.

Higham-Mary results: Lots of notation

Define

$$\tilde{\gamma}(\lambda) = \exp\left(\lambda\sqrt{N}\epsilon_S + \frac{N\epsilon_S^2}{1-\epsilon_S}\right) - 1$$

$$P(\lambda) = 1 - 2\exp\left(-\frac{\lambda^2(1-\epsilon_S)^2}{2}\right)$$

and

$$Q(\lambda, N) = 1 - N(1 - P(\lambda))$$

Limiting cases

- $N\epsilon_S$ small $\rightarrow \tilde{\gamma}(\lambda) \approx \lambda\sqrt{N}\epsilon_S$
- ϵ_S small, λ large $\rightarrow P(\lambda) \approx 1$
- N large and λ large and curated $\rightarrow Q(\lambda, N^3) \approx 1$ independeltly of N

At last, a theorem!

Theorem:
Use Gaussian elimination for $\mathbf{A}\mathbf{x} = \mathbf{b}$. The the computed LU factors $\hat{\mathbf{L}}$ and $\hat{\mathbf{U}}$ satisfy

$$\mathbf{A} + \delta\mathbf{A} = \hat{\mathbf{L}}\hat{\mathbf{U}} \text{ and } |\delta\mathbf{A}| \le (3\tilde{\gamma}(\lambda) + \tilde{\gamma}(\lambda)^2)|\hat{\mathbf{L}}||\hat{\mathbf{U}}|$$

with probability at least $Q(\lambda, N^3/3 + 3N^2/2 + 7N/6)$.
Wait! What? Is this good?

Goodness of results

Remember, we get to pick λ to make things look good.

- $N\epsilon_S$ small so $(3\tilde{\gamma}(\lambda) + \tilde{\gamma}(\lambda)^2) = O(\epsilon_S \sqrt{N})$
  - Much better than $O(N)$
- Grow $\lambda \approx \sqrt{\log(N)}$ and $Q(\lambda, N^3/3 + 3N^2/2 + 7N/6) \approx 1$

So you can use $\sqrt{N}$ with confidence(?)

What should we observe if $\sqrt{N}$ is the right thing?

- Trouble (slow nonlinear convergence) when $\sqrt{N}\epsilon_S \geq .1$
  - Double: $N \approx 10^{30}$. Not on my computer.
  - Single: $N \approx 10^{14}$. Not on my computer.
  - Half: $N \approx 10^{6}$. Maybe if we push it.
- Expectation: Single just as good as double.
- Expect to see deterioration with N for half.

Chandrasekhar H-equation

Midpoint rule discretization

$$\mathcal{F}(H)(\mu) = H(\mu) - \left(1 - \frac{c}{2}\int_0^1 \frac{\mu H(\mu)}{\mu+\nu}\,d\nu\right)^{-1} = 0.$$

- Defined on $C[0,1]$
- $\mathcal{F}'$ nonsingular for $0 \le c < 1$.
  Simple fold singularity at $c = 1$.
- Any sensible discretization inherits the singularity structure.

Discrete Problem

$$\mathbf{F}(\mathbf{u})_i \equiv u_i - \left(1 - \frac{c}{2N}\sum_{j=1}^{N}\frac{u_j \mu_i}{\mu_j + \mu_i}\right)^{-1} = 0.$$

Midpoint rule says

$$\frac{c}{2N}\sum_{j=1}^{N}\frac{u_j \mu_i}{\mu_j + \mu_i} = \frac{c(i-1/2)}{2N}\sum_{j=1}^{N}\frac{u_j}{i+j-1}.$$

so can evaluate $\mathbf{F}$ in $O(N\log(N))$ work with FFT.

Analytic Jacobian

Define $\mathbf{M}$ by

$$\mathbf{M}(\mathbf{u})_i = \frac{c(i-1/2)}{2N} \sum_{j=1}^{N} \frac{u_j}{i+j-1}$$

and compute the Jacobian analytically as

$$\mathbf{F}'(\mathbf{u}) = \mathbf{I} - \mathrm{diag}(\mathbf{G}(\mathbf{u}))^2 \mathbf{M}$$

where

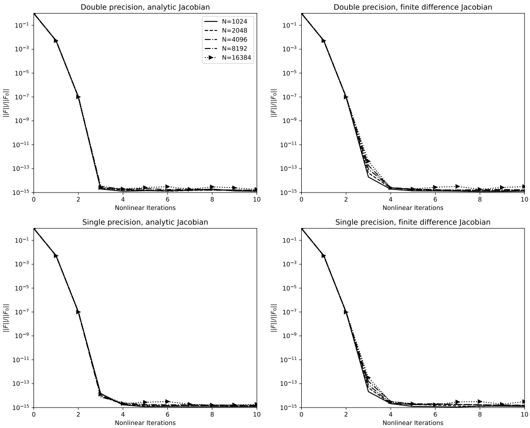
$$\mathbf{G}(\mathbf{u})_i = \left(1 - \frac{c}{2N} \sum_{j=1}^{N} \frac{u_j \mu_i}{\mu_j + \mu_i}\right)^{-1}.$$

Takes $O(N^2)$ work.

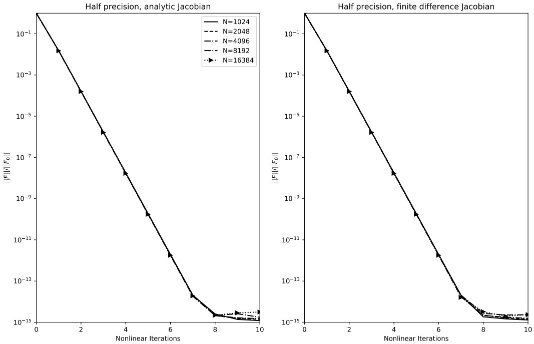
Experiments

- $c = .5, .99, 1.0$ (no theory for $c = 1.0$)
- Analytic and forward difference Jacobians
  Theory predicts single as good as double
- Double, single, and half precision factor/solve
- Everything else in double
- $N = 2^p$, $p = 10, \dots, 14$, $2^14 = 16384$
  Larger N took far to long in half.

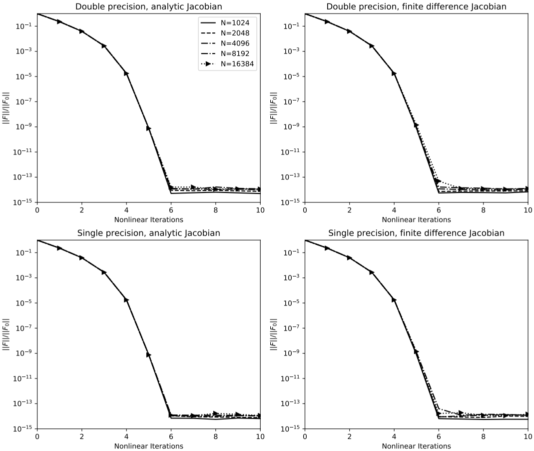
$c = .5$, double and single



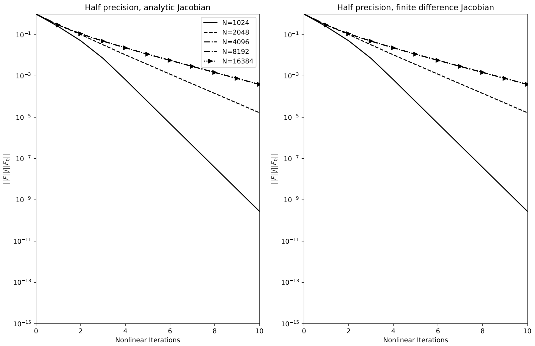
$c = .5$, half


Half precision, analytic Jacobian
Half precision, finite difference Jacobian
N=1024
N=2048
N=4096
N=8192
N=16384
||F||/||F0||
Nonlinear Iterations

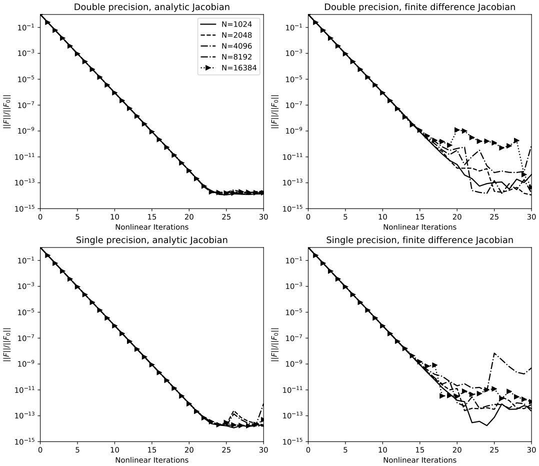
$c = .99$, double and single



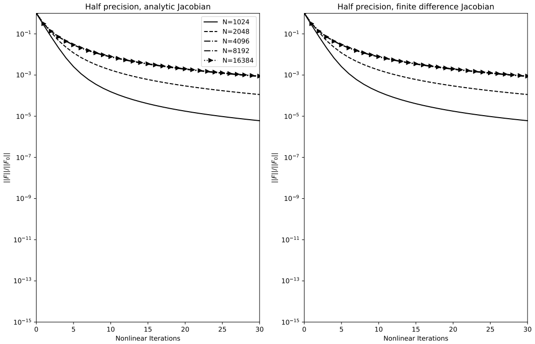
$c = .99$, half


Half precision, analytic Jacobian
Half precision, finite difference Jacobian
N=1024
N=2048
N=4096
N=8192
N=16384
$||F||/||F_0||$
Nonlinear Iterations
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
10^{-11}
10^{-13}
10^{-15}
0
2
4
6
8
10

$c = 1.0$, double and single, theory not from this talk



$c = 1.0$, half, DOOM! Some theory out there



Summary

- Low quality linear solvers are just fine
  - Single precision $\rightarrow$ same nonlinear results
  - Half precision $\rightarrow$ not great
- The precision for you is 32!
- $c = 1.0$ is different