Floating-point arithmetic

Higham 2002 (A.S.N.A)
2nd ed.

- Computers store number in base 2 with finite/fixed memory (bits)
- Irrational numbers are stored inexactly, e.g. $1/3 \approx \boxed{0.333...}$
- Calculations are rounded to nearest floating-point number (rounding error)
- Thus the accuracy of the final error is nontrivial

(base 2)

$\sqrt{3} = 1.011000 \times 2^{1}$

$-1000 \quad 1000$

$10^{-300} \sim 10^{300}$

$1. \leftarrow \left(\frac{1}{3} \cdot 3\right) = 0.99...$

Two examples with MATLAB

- $((\texttt{sqrt}(2))^2 - 2) * \boxed{\texttt{1e15}} = 0.4441$ (should be 1..)
- $\sum_{n=1}^{\infty} \frac{1}{n} \approx 30$ (should be ∞..)

$+ 10^{-16}$

$n = 10^{16}$

$fl(30 + \boxed{10^{-4}}) = 30$

rounded off.

An important (but not main) part of numerical analysis/NLA is to study the effect of rounding errors

Canonical reference: Higham's book (2002)

Conditioning and stability

$Ax = b$

- Conditioning is the sensitivity of a problem (e.g. of finding $y = f(x)$ given x) to perturbation in inputs, i.e., how large $\kappa := \sup_{\delta x} \|f(x+\delta x) - f(x)\|/\|\delta x\|$ is in the limit $\delta x \to 0$.

  $y+\delta y$ $\quad$ y

  (this is *absolute* condition number; equally important is *relative* condition number $\kappa_r := \sup_{\delta x} \frac{\|f(x+\delta x)-f(x)\|}{\|f(x)\|} / \frac{\|\delta x\|}{\|x\|}$)

  ill-conditioned $\Leftrightarrow$ κ ($\gg 1$) big $\Leftrightarrow \kappa \gg 1$

  well-cond. κ small $\Leftrightarrow$ $\kappa = O(1)$

- (Backward) Stability is a property of an algorithm, which describes if the computed solution $\hat{y}$ is a 'good' solution, in that it is an exact solution of a nearby input, that is, $\hat{y} = f(x + \Delta x)$ for a small Δx.

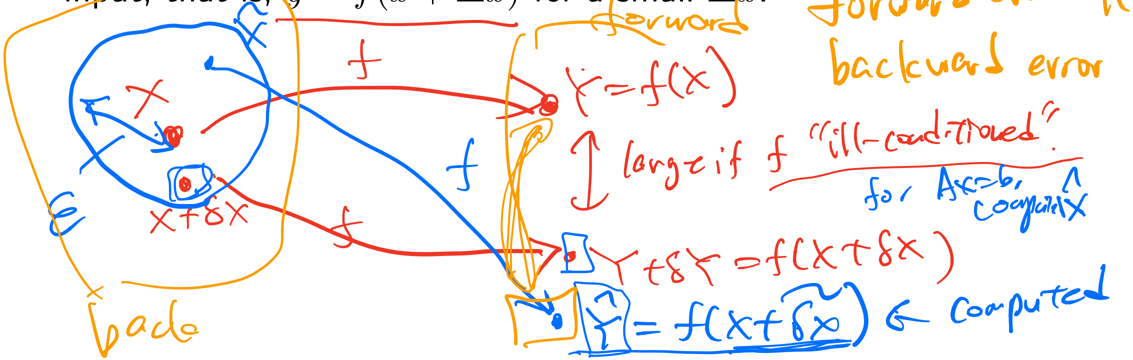
forward error $\|y - \hat{y}\|$

backward error $\|x - \hat{x}\|$

$y = f(x)$, $\hat{y} = f(\hat{x})$.

$(A + \Delta A)\hat{x} = (b + \Delta b)$

w/ $\|\Delta A\|$ small.



Conditioning and stability

- Conditioning is the sensitivity of a problem (e.g. of finding $y = f(x)$ given x) to perturbation in inputs, i.e., how large $\kappa := \sup_{\delta x} \|f(x+\delta x) - f(x)\|/\|\delta x\|$ is in the limit $\delta x \to 0$.
  (this is *absolute* condition number; equally important is *relative* condition number $\kappa_r := \sup_{\delta x} \frac{\|f(x+\delta x)-f(x)\|}{\|f(x)\|} / \frac{\|\delta x\|}{\|x\|}$)

  $(A+\Delta A)\hat{x} = b$

  $\boxed{A}x = b$

  not exact anyway!

- (Backward) Stability is a property of an algorithm, which describes if the computed solution $\hat{y}$ is a 'good' solution, in that it is an exact solution of a nearby input, that is, $\hat{y} = f(x + \Delta x)$ for a small Δx.

  $\|x\|_\infty \le \|x\|_2 \le \|x\|_\infty \sqrt{n}$

If problem is ill-conditioned $\kappa \gg 1$, then blame the problem not the algorithm

Notation/convention: $\hat{x}$ denotes a computed approximation to x (e.g. of $x = A^{-1}b$)
ϵ denotes a small term $O(u)$, on the order of unit roundoff/working precision; so we write e.g. $u, 10u, (m+n)u, mnu$ all as ϵ

$\kappa\left(\begin{bmatrix} 10^{10} & 1 \\ & 10^{10} \end{bmatrix}\right) > 10^{10}$

- Consequently (in this lecture/discussion) norm choice does not matter

Numerical stability: backward stability

For computational task $Y = f(X)$ and computed approximant $\hat{Y}$,

- Ideally, error $\|Y - \hat{Y}\|/\|Y\| = \epsilon$: seldom true
  (u: unit roundoff, $\approx 10^{-16}$ in standard double precision)
- Good alg. has Backward stability $\hat{Y} = f(X + \Delta X)$, $\frac{\|X - \hat{X}\|}{\|X\|} = \epsilon$ "exact solution of slightly wrong input "

Numerical stability: backward stability

For computational task $Y = f(X)$ and computed approximant $\hat{Y}$,

- Ideally, error $\|Y - \hat{Y}\|/\|Y\| = \epsilon$: ~~seldom true~~ "forward stability"

  (u: unit roundoff, $\approx 10^{-16}$ in standard double precision)

- Good alg. has Backward stability $\hat{Y} = f(X + \Delta X)$, $\frac{\|X - \hat{X}\|}{\|X\|} = \epsilon$ "exact solution of slightly wrong input "

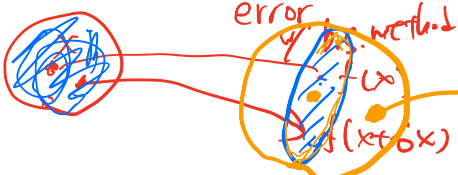
  $x+\Delta x$ x x_* $f(x)$ $\boxed{f(x+\Delta x)}$

- Justification: Input (matrix) is usually inexact anyway! $f(X + \Delta X)$ is just as good at $f(X)$ at approximating $f(X_*)$ where $\|\Delta X\| = O(\|X - X_*\|)$
  We shall 'settle with' such solution, though it may not mean $\hat{Y} - Y$ is small

- $\boxed{\text{Forward stability}}$ "$\|Y - \hat{Y}\|/\|Y\| = O(\kappa(f)u)$" "error is as small as backward stable alg." (sometimes used to mean small error; we follow Higham's book [2002])

error of method $\leq (\kappa(f) \cdot u)$

$\|Y - \hat{Y}\| \leq \kappa(f) \cdot u$



$\tilde{f} = f(x + \Delta\tilde{x})$ may be large.

Matrix condition number

$$\kappa_2(A) = \boxed{\frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} (\geq 1)}$$

e.g. for linear systems. A backward stable soln for $Ax = b$, s.t. $(A + \Delta A)\hat{x} = b$ satisfies, assuming backward stability $\|\Delta A\| \leq \epsilon\|A\|$ and $\kappa_2(A) \ll \epsilon^{-1}$ (so $\|A^{-1}\Delta A\| \ll 1$),

$$\boxed{\frac{\|\hat{x} - x\|}{\|x\|} \lesssim \epsilon\kappa_2(A)}$$

in practice $\kappa_2(A) < 10^4$

2-norm cond num. κ_∞

Matrix condition number

$$\kappa_2(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} (\geq 1)$$

e.g. for linear systems. A backward stable soln for $Ax = b$, s.t. $(A + \Delta A)\hat{x} = b$ satisfies, assuming backward stability $\|\Delta A\| \leq \epsilon\|A\|$ and $\kappa_2(A) \ll \epsilon^{-1}$ (so $\|A^{-1}\Delta A\| \ll 1$),

$\hat{x} = \underline{(A+\Delta A)^{-1} b}$ $(b + \Delta b)$

$$\frac{\|\hat{x} - x\|}{\|x\|} \lesssim \epsilon \boxed{\kappa_2(A)}$$

useful when $\|X\| < 1$

'proof': By Neumann series $(I - \underbrace{X}_{\in \mathbb{R}^{n\times n}})^{-1} = I + X + X^2 + \cdots = \sum_{i=0}^{\infty} X^i$ $(\|X\| < 1)$ any norm.

$$(A + \Delta A)^{-1} = (A(I \boxed{+ A^{-1}\Delta A}))^{-1} = (I - A^{-1}\Delta A + O(\|A^{-1}\Delta A\|^2))A^{-1}$$

So $\boxed{\hat{x}} = (A + \Delta A)^{-1}b = A^{-1}b - A^{-1}\Delta A\underbrace{A^{-1}b}_{x} + O(\|A^{-1}\Delta A\|^2) =$
$= x - \boxed{A^{-1}\Delta A x} + \boxed{O(\|A^{-1}\Delta A\|^2)}$, Hence

$$\|x - \hat{x}\| \lesssim \|A^{-1}\Delta A x\| \leq \boxed{\|A^{-1}\|\|\Delta A\|}\|x\| = \kappa_2(A)\|x\|$$

p

Backward stable+well conditioned=accurate solution

Suppose

- $Y = f(X)$ computed backward stably i.e., $\hat{Y} = f(X + \Delta X)$, $\|\Delta X\| = \epsilon$.
- Conditioning $\|f(X) - f(X + \Delta X)\| \lesssim \kappa \|\Delta X\|$ $\left(+\|(\Delta X)^2\|\right)$

$\|\Delta X\|$

Then (relative version possible)

$$\|\hat{Y} - Y\| \lesssim \kappa \epsilon$$

Backward stable+well conditioned=accurate solution

Suppose

- $Y = f(X)$ computed backward stably i.e., $\hat{Y} = f(X + \Delta X)$, $\|\Delta X\| = \epsilon$.
- Conditioning $\|f(X) - f(X + \Delta X)\| \lesssim \kappa \|\Delta X\|$

Then (relative version possible)

$$\|\hat{Y} - Y\| \lesssim \kappa\epsilon$$

'proof': $\|\hat{Y} - Y\| = \|f(X + \Delta X) - f(X)\| \lesssim \kappa \underbrace{\|\Delta X\|}_{\varepsilon} \cancel{\|f(X)\|} = \kappa\epsilon$

Backward stable+well conditioned=accurate solution

Suppose

- $Y = f(X)$ computed backward stably i.e., $\hat{Y} = f(X + \Delta X)$, $\|\Delta X\| = \epsilon$.
- Conditioning $\|f(X) - f(X + \Delta X)\| \lesssim \kappa \|\Delta X\|$

Then (relative version possible)

$$\|\hat{Y} - Y\| \lesssim \kappa\epsilon$$

'proof': $$\|\hat{Y} - Y\| = \|f(X + \Delta X) - f(X)\| \lesssim \kappa \|\Delta X\| \cancel{\|f(X)\|} = \kappa\epsilon$$

$\kappa(Q) = \frac{\sigma_{max}}{\sigma_{min}} = 1$

If well-conditioned $\boxed{\kappa = O(1),}$ good accuracy! Important examples:

- Well-conditioned linear system $Ax = b$, $\kappa_2(A) \approx 1$ incl. $\underbrace{Q}_{\text{orth.}}x = b$
- Eigenvalues of symmetric matrices (via Weyl's bound $\lambda_i(A + E) \in \lambda_i(A) + [-\|E\|_2, \|E\|_2]$)
- Singular values of any matrix $\sigma_i(A + E) \in \sigma_i(A) + [-\|E\|_2, \|E\|_2]$

as b.s. alg
$|\hat{\lambda}_i(A) - \lambda_i(A)| \le \epsilon\|A\|$
$|\sigma_i(A) - \hat{\sigma}_i(A)| \le \epsilon\|A\|$

eigvec(I)

Note: eigvecs/singvecs can be highly ill-conditioned

Backward stability of triangular systems

Recall $Ax = b$ via $Ly = b$, $Ux = y$ (triangular systems).

The computed solution $\hat{x}$ for a (upper/lower) triangular linear system $Rx = b$ solved via back/forward substitution is backward stable, i.e., it satisfies

$$(R + \Delta R)\hat{x} = b, \qquad \|\Delta R\| = O(\epsilon\|R\|).$$

Proof: Trefethen-Bau or Higham (nonexaminable but interesting)

- backward error can be bounded componentwise
- this means $\|\hat{x} - x\|/\|x\| \leq \epsilon\kappa_2(R)$
  - (unavoidably) poor worst-case (and attainable) bound when ill-conditioned
  - often better with triangular systems

pessimistic but attainable

(In)stability of $Ax = b$ via LU with pivots

Fact (proof nonexaminable): Computed $\hat{L}\hat{U}$ satisfies $\boxed{\frac{\|\hat{L}\hat{U} - A\|}{\|L\|\|U\|} = \epsilon}$

(note: not $\frac{\|\hat{L}\hat{U} - A\|}{\|A\|} = \epsilon$)

- If $\|L\|\|U\| = O(\|A\|)$, then $\boxed{(L + \Delta L)(U + \Delta U)\hat{x} = b}$
  $\Rightarrow$ $\hat{x}$ backward stable solution (exercise)

$(A + \Delta A)\hat{x} = b$

b.s. if

$\hat{L}\hat{U} = A + \Delta A$

$\|\Delta A\| \le \epsilon \|A\|$

$\|\Delta A\| \le \epsilon \|\hat{L}\|\|\hat{U}\|$

may or may not

$\Rightarrow$ b.s. of LU.

possible

$\|\hat{L}\| \cdot \|\hat{U}\| >> \|A\|$

$A \sim \hat{L}\hat{U}$

$\| \quad \| \quad \| \quad \| \le \|\hat{L}\| \|\hat{U}\|$

(In)stability of $Ax = b$ via LU with pivots

Fact (proof nonexaminable): Computed $\hat{L}\hat{U}$ satisfies $\frac{\|\hat{L}\hat{U}-A\|}{\|L\|\|U\|} = \epsilon$

(note: not $\frac{\|\hat{L}\hat{U}-A\|}{\|A\|} = \epsilon$)

- If $\|L\|\|U\| = O(\|A\|)$, then $(L+\Delta L)(U+\Delta U)\hat{x} = b$ $\Rightarrow$ $\hat{x}$ backward stable solution (exercise)

$\|\hat{L}\|\|\hat{U}\| = O(\|A\|)$

Question: Does $LU = A + \Delta A$ or $LU = PA + \Delta A$ with $\|\Delta A\| = \epsilon\|A\|$ hold?

Without pivot $(P = I)$: $\|L\|\|U\| \gg \|A\|$ unboundedly (e.g. $\left[\begin{smallmatrix}\epsilon & 1\\ 1 & 1\end{smallmatrix}\right]$) unstable

(In)stability of $Ax = b$ via LU with pivots

Fact (proof nonexaminable): Computed $\hat{L}\hat{U}$ satisfies $\frac{\|\hat{L}\hat{U}-A\|}{\|L\|\|U\|} = \epsilon$

(note: not $\frac{\|\hat{L}\hat{U}-A\|}{\|A\|} = \epsilon$)

- If $\|L\|\|U\| = O(\|A\|)$, then $(L + \Delta L)(U + \Delta U)\hat{x} = b$
  $\Rightarrow$ $\hat{x}$ backward stable solution (exercise)

Question: Does $LU = A + \Delta A$ or $LU = PA + \Delta A$ with $\|\Delta A\| = \epsilon\|A\|$ hold?

Without pivot ($P = I$): $\|L\|\|U\| \gg \|A\|$ unboundedly (e.g. $\left[\begin{smallmatrix}\epsilon & 1\\ 1 & 1\end{smallmatrix}\right]$) unstable

$\frac{\|L\|\|U\|}{\|A\|} \sim 2^n.$

With pivots:

- Worst-case: $\|L\|\|U\| \gg \|A\|$ grows exponentially with n, unstable
  - growth governed by that of $\|L\|\|U\|/\|A\| \Rightarrow \|U\|/\|A\|$
- In practice (average case): perfectly stable
  - Hence this is how $Ax = b$ is solved, despite alternatives with guaranteed stability exist (but slower; e.g. via SVD, or QR (next))

Resolution/explanation: among biggest open problems in numerical linear algebra!

Stability of Cholesky for $A \succ 0$

Cholesky $A = \boxed{R^T R}$ for $A \succ 0$

- succeeds without pivot (active matrix is always positive definite)
- R never contains entries $> \|A\|_2$

$\Rightarrow$ backward stable! Hence positive definite linear system $\boxed{Ax = b \text{ stable via Cholesky}}$

$A \to (A - \;\; L^T)$

C-F

$$\hat{R}^T \hat{R} = A + \epsilon \|A\|_2$$

$$\epsilon = O(u)$$

(caveat;
if $\boxed{\kappa_2(A) \approx \frac{1}{u}}$
then Cholesky may break down;
fails to be positive due to num. errors.

(In)stability of Gram-Schmidt

$A = QR$

- Gram-Schmidt is subtle
  - plain (classical) version: $\|Q^TQ - I\| \le \epsilon\kappa_2(A))^2$

    10^{-5}

    1 not 2
  - modified Gram-Schmidt (orthogonalise 'one vector at a time'): $\|Q^TQ - I\| \le \epsilon\kappa_2(A)$
  - Gram-Schmidt twice (G-S again on computed Q): $\|Q^TQ - I\| \le \epsilon$

$GS(\hat{Q}) = Q_2 R_2$

Stability of Householder QR

With Householder QR, the computed $\hat{Q}, \hat{R}$ satisfy

$$\|\hat{Q}^T\hat{Q} - I\| = O(\epsilon), \quad \|A - \hat{Q}\hat{R}\| = O(\epsilon\|A\|),$$

best possible!

and (of course) $\hat{R}$ upper triangular.

Rough proof

- Each reflector satisfies $fl(H_i A) = H_i A + \epsilon_i \|A\|$
- Hence $(\hat{R} =) fl(H_n \cdots H_1 A) = H_n \cdots H_1 A + \epsilon\|A\|$
- $fl(H_n \cdots H_1) =: \hat{Q}^T = H_n \cdots H_1 + \epsilon\|\cancel{A}\|$,
- Thus $\hat{Q}\hat{R} = A + \epsilon\|A\|$

$\sim \sqrt{n} \cdot \epsilon_i$

10^{-15}

Stability of Householder QR

With Householder QR, the computed $\hat{Q}, \hat{R}$ satisfy

$$\boxed{\hat{Q}^T\hat{Q} - I = O(\epsilon)}, \quad \boxed{\|A - \hat{Q}\hat{R}\| = O(\epsilon\|A\|)},$$

and (of course) R upper triangular.

Rough proof

- Each reflector satisfies $fl(H_iA) = H_iA + \epsilon_i\|A\|$
- Hence $(\hat{R} =) fl(H_n \cdots H_1 A) = H_n \cdots H_1 A + \epsilon\|A\|$
- $fl(H_n \cdots H_1) =: \hat{Q}^T = H_n \cdots H_1 + \epsilon\|A\|$,
- Thus $\hat{Q}\hat{R} = A + \epsilon\|A\|$

$A = QR$

Notes:

- This doesn't mean $\boxed{\|\hat{Q} - Q\|, \|\hat{R} - R\|}$ are small at all! Indeed $\hat{Q}, R$ are as ill-conditioned as A

  $\approx O(\kappa_2(A)\epsilon)$
- $Ax = b$ via QR, least-squares stable
  (please see arXiv:2009.11392 for application in low-rank approximation)

Orthogonality matters for stability

With orthogonal matrices Q,

$$\boxed{\frac{\|fl(QA) - QA\|}{\|QA\|} \leq \epsilon}, \qquad \frac{\|fl(AQ) - AQ\|}{\|AQ\|} \leq \epsilon$$

Orthogonality matters for stability

With orthogonal matrices Q,

$$\frac{\|fl(QA) - QA\|}{\|QA\|} \le \epsilon, \qquad \frac{\|fl(AQ) - AQ\|}{\|AQ\|} \le \epsilon$$

whereas in general, $\|fl(AB) - AB\| \le \epsilon\|A\|\|B\|$, so

$\|fl(AB) - AB\|/\|AB\| \le \epsilon \min(\kappa_2(A), \kappa_2(B))$ (why? exercise)

Orthogonality matters for stability

With orthogonal matrices Q,

$$\frac{\|fl(QA) - QA\|}{\|QA\|} \leq \epsilon, \qquad \frac{\|fl(AQ) - AQ\|}{\|AQ\|} \leq \epsilon$$

whereas in general, $\|fl(AB) - AB\| \leq \epsilon\|A\|\|B\|$, so
$\|fl(AB) - AB\|/\|AB\| \leq \epsilon \min(\kappa_2(A), \kappa_2(B))$ (why? exercise)
Hence algorithms involving ill-conditioned matrices are unstable (e.g. eigenvalue decomposition of non-normal matrices, Jordan form, etc),
whereas those based on orthogonal matrices are stable, e.g.

- Householder QR factorisation
- **QR algorithm** for $Ax = \lambda x$
- **Golub-Kahan** algorithm for $A = U\Sigma V^T$
- **QZ algorithm** for $Ax = \lambda Bx$

We next turn to the algorithms in boldface