Floating-point arithmetic

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

Two examples with MATLAB

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

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

- 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 $\kappa_r := \sup_{\delta x} \frac{\|f(x+\delta x)-f(x)\|}{\|f(x)\|} / \frac{\|\delta x\|}{\|x\|}$)

- (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.

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 $\kappa_r := \sup_{\delta x} \frac{\|f(x+\delta x)-f(x)\|}{\|f(x)\|} / \frac{\|\delta x\|}{\|x\|}$)
- (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.

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

- 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
  (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 "

- 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

- 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])

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 $\|A^{-1}\Delta A\| \ll 1$),

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

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 $\|A^{-1}\Delta A\| \ll 1$),

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

'proof': By Neumann series

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

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

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

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)

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

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

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

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

- Well-conditioned linear system $Ax = b$, $\kappa_2(A) \approx 1$
- 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 $\sigma_i(A + E) \in \sigma_i(A) + [-\|E\|_2, \|E\|_2]$

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

(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)

(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. $\begin{bmatrix} \epsilon & 1 \\ 1 & 1 \end{bmatrix}$) 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

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 $A = R^T R$ for $A \succ 0$

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

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

(In)stability of Gram-Schmidt

- Gram-Schmidt is subtle
  - plain (classical) version: $\|Q^T Q - I\| \leq \epsilon \kappa_2(A))^2$
  - modified Gram-Schmidt (orthogonalise 'one vector at a time'): $\|Q^T Q - I\| \leq \epsilon \kappa_2(A)$
  - Gram-Schmidt twice (G-S again on computed Q): $\|Q^T Q - I\| \leq \epsilon$

Stability of Householder QR

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

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

and (of course) R upper triangular.

Rough proof

- Each reflector satisfies $fl(H_iA) = H_iA + \epsilon_i\|A\|$
- Hence $(\hat{R} =)fl(H_n \cdots H_1A) = H_n \cdots H_1A + \epsilon\|A\|$
- $fl(H_n \cdots H_1) =: \hat{Q}^T = H_n \cdots H_1 + \epsilon$,
- Thus $\hat{Q}\hat{R} = A + \epsilon\|A\|$

Stability of Householder QR

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

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

and (of course) R upper triangular.

Rough proof

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

Notes:

- This doesn't mean $\|\hat{Q} - Q\|, \|\hat{R} - R\|$ are small at all! Indeed Q, R are as ill-conditioned as A
- $Ax = b$ via QR, least-squares stable (please see arXiv:2009.11392 for application in low-rank approximation)

Orthogonality matters for stability

With orthogonal matrices Q,

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

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