## Lecture 3.5: Contents

**Round-off errors in Linear Algebra**

- Catastrophic cancellation. Floating Point arithmetic.
- Forward and Backward analysis.
- Gauss transformations and Orthogonal matrices.
- Analysis of Gaussian Elimination.
- Iterative refinement.

**Special Linear Systems**

- Symmetric and Positive definite matrices.
- Indefinite matrices. Toeplitz matrices.
- Perturbation results. Sherman-Morrison.



## Catastrophic Cancellation

**Example** Suppose $x = 101 \pm 1$ and $y = 100 \pm 1$. Compute $z = x - y$ with error bounds. The result is $z = 1 \pm 2$.

**Observation** The error bound $\Delta z$ is large relative to the result.

**Definition** Loss of accuracy during addition or subtraction of floating point numbers is called *cancellation*.

This sometimes occurs during plus or minus operations. Never during multiply or division.

**Remark** A matrix-vector multiply $y = Ax$ consists of many potentially bad operations $y_i := y_i + a_{ij}x_j$. Can we trust the results?



## Floating Point Numbers

**Definition** A floating point number system is defined by its base $\beta$, exponent range $[L, U]$, and precision $t$. A number $x$ in the floating point system can be written

$$x = d_1.d_2d_3\dots d_t \times \beta^e,$$

where

$$0 \le d_i < \beta, \quad d_1 \ne 0, \quad L \le e \le U.$$

**Remark** zero cannot be written this way but is included in a floating point system.

**Example** The system $(10, 3, -4, 4)$ includes $x = 8.765 \times 10^2$. Most common is the IEEE double precision system $(2, 52, -1022, 1023)$.



**Definition** Suppose $m \le |x| \le M$, i.e. $x$ is within the range of the floating point system. By $\mathrm{fl}(x)$ we means the closest floating point number to $x$.

**Lemma** Let $u = \frac{1}{2}\beta^{1-t}$. Then $\mathrm{fl}(x) = x(1+\epsilon)$, $|\epsilon| \le u$.

Arithmetic operations are also assumed to satisfy the same bound, i.e.

$$\frac{|fl(a \operatorname{op} b) - a \operatorname{op} b|}{|a \operatorname{op} b|} \le u, \quad a \operatorname{op} b \ne 0.$$

Holds for $+, -, \cdot, /, \sqrt{x}, e^x, \dots$.

## Round-Off Errors and Scalar Products

Compute $x^T y$ by the following code

```
s=0;
for i=1:n
    s:=s+x(i)*y(i)
end
```

**Lemma** If $nu \leq 0.01$ then $|\text{fl}(x^T y) - x^T y| \leq 1.01nu|x|^T|y|$.

**Remark** If $x^T y << |x|^T|y|$ the relative error in the result may be large.



**Corollary** $\text{fl}(AB) = AB + E, \quad |E| \leq nu|A||B| + \mathcal{O}(u^2)$.

**Remark** Each element of $AB$ is computed as a scalar product.

The result is quite bad if $|AB| << |A||B|$.

**Note** Worst case error bounds are rarely very sharp. Statistical methods often give a better understanding of the actual errors.



## Orthogonal Matrices

**Lemma** If $Q$ is orthogonal then $\text{fl}(QA) = Q(A + F)$, where $\|F\|_2 \leq \mathcal{O}(\mathbf{u})\|A\|_2$.

**Remark** This means that multiplication by an orthogonal matrix is backwards stable. The same is true for a sequence of orthogonal matrices.

Important for computing eigenvalues and solving least squares problems.



## Gauss transformations and Round-off errors

**Lemma** Suppose $M$ is the Guass transformation that zeroes the first column of a matrix $A$. Then

$$\text{fl}(MA) = MA + E, \quad |E| \leq 3\mathbf{u}(|A| + |m||A(1,:)|) + \mathcal{O}(u^2),$$

where $m$ is the vector of multipliers.

**Remarks** Partial pivoting means $|m| \leq 1$.

Note that $|m||A(1,:)|$ is an outer-product. The error is zero in the first row and the first column of $MA$.

## Round-Off errors and the LU Decomposition

**Ideal situation** The only error is when $A$ and $b$ are stored in memory.

Suppose $(A+E)\hat{x} = (b+e)$, where

$$\|E\|_\infty \le u\|A\|_\infty, \; \|e\|_\infty \le u\|b\|_\infty,$$

holds and also that $u\kappa_\infty(A) \le 1/2$. Then

$$\frac{\|x-\hat{x}\|_\infty}{\|x\|_\infty} \le 4u\kappa_\infty(A).$$

**Remark** It is not possible to prove a better error bound.



**Theorem** Suppose no pivoting occurs during the $LU$ decomposition then the computed matrices $\hat{L}$ and $\hat{U}$ satisfy

$$\hat{L}\hat{U} = A + H = LU + H,$$

where

$$|H| \le 3(n-1)u(|A| + |\hat{L}||\hat{U}|) + \mathcal{O}(u^2).$$

**Remark** Pivoting doesn't change the analysis.

This is an example of **Backwards error analysis**. The computed decomposition $\hat{L}\hat{U}$ is the exact $LU$ decomposition of a matrix $\hat{A}$ which is close to $A$.



**Theorem** Let $\hat{L}$ and $\hat{U}$ be the computed LU factors and that we compute the solution $\hat{L}\hat{U}\hat{x} = b$. Then $(A+E)\hat{x} = b$ with

$$|E| \le nu(3|A| + 5|\hat{L}||\hat{U}|) + \mathcal{O}(u^2).$$

**Remark** If the factor $|\hat{L}||\hat{U}|$ is small then this would be comparable to the ideal situation. Pivoting makes $l_{ij} \le 1$ and typically $|\hat{U}|$ is comparable in size to $|A|$.

The growth of elements $u_{ij}$ during Guassian elimination has been studied extensively. Usually the growth rate is very small in practice.



## Iterative Refinement

**Observation** If $\widehat{x}$ is an approximate solution to $Ax = b$, and $\widehat{r} = b - A\widehat{x}$, then the error $e = \widehat{x} - x$ satisfies

$$Ae = Ax - A\widehat{x} = b - A\widehat{x} = \widehat{r}.$$

**Idea** Compute $\widehat{r}$, solve for $e$, and update $\widehat{x}^{(1)} = \widehat{x} + e$.

This is called *Iterative refinement*. Can repeat the process if needed.

**Algorithm** One step iterative refinement

1. Compute the decomposition $PA = LU$.
2. Solve $Ax = b$ to obtain $\widehat{x}$.
3. Compute residual $\widehat{r} = b - A\widehat{x}$ in extended precision.
4. Solve $Ae = \widehat{r}$ using the $LU$ decomposition.
5. Update $\widehat{x} := \widehat{x} + e$

**Remark** Requires $\mathcal{O}(n^2)$ additional work. If the computed residuals have a few correct digits then usually the error is reduced.

**Remark** Most problems involve inexact data. It doesn't make sense to work to obtain a higly accurate solution to an imprecise problem.



## Symmetric and positive definite matrices

**Definition** A matrix $A \in \mathbb{R}^{n\times n}$ is *positive definite* if $x^TAx > 0$ for all non-zero $x \in \mathbb{R}^n$. If $x^TAx \geq 0$ the matrix is *positive semi-definite*.

**Theorem** If $A \in \mathbb{R}^{n\times n}$ is *positive definite* and $X_k \in \mathbb{R}^{n\times k}$ has rank $k$ then $B = X_k^TAX_k$ is also positive definite.

**Remark** Theory often says a matrix is positive definite. Examples are covariance matrices and finite element discretizations of elliptic equations.



**Lemma** Suppose $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$, and $a \in \mathbb{R}$. Then

$$B = \begin{pmatrix} A & b \\ b^T & a \end{pmatrix}$$

is *positive definite* if and only if $A$ is positive difinite and $b^TA^{-1}b < a$. In this case

$$\det(B) = \det(A)\det(a - b^TA^{-1}b).$$

**Remark** This is the basis of a recursive algorithm for computing the Cholesky decomposition.



**Theorem** If $A$ is *symmetric* and *positive definite* then there exists a unique upper triangular matrix $R$, with positive diagonal elements, such that

$$A = R^TR.$$

**Remark** This is the *Cholesky factorization*. Requires about half the work compared to regular $LU$ factorization.

The analysis of Cholesky is the same as for the $LU$ decompositon, except $|R|^T|R| \approx |A|$ since the largest elements of $R$ are positive. Thus the results are much better.

**Corollary** The computed Cholesky factor $\hat{R}$ satisfies
$$\hat{R}^T\hat{R} = A + H = R^T R + H,$$
where
$$|H| \leq 3(n-1)u(|A| + |\hat{R}|^T|\hat{R}|) + \mathcal{O}(u^2).$$

**Corollary** Let $\hat{R}$ be the computed Cholesky factor and suppose that we compute the solution $\hat{R}^T\hat{R}\hat{x} = b$. Then $(A+E)\hat{x} = b$ with
$$|E| \leq nu(3|A| + 5|\hat{R}^T||\hat{R}|) + \mathcal{O}(u^2).$$



**Theorem** If $A$ is *symmetric* and *positive semi-definite* then $A = LDL^T$, with $d_{ii} \geq 0$.

**Remark** If $\widetilde{a}_{ii} = 0$ then *symmetric pivoting*, $A := PAP^T$, can be used to move a non-zero diagonal element to the pivoting position. If no such element exists the factorization is complete.

**Corollary** A *symmetric* and *indefinite* matrix $A$ can be factored $PAP^T = LDL^T$.

**Example** Suppose
$$A = \begin{pmatrix} C & B \\ B^T & 0 \end{pmatrix}$$
where $C$ is symmetric and positive definite and $B$ has full rank.



# Toeplitz matrices

**Definition** A matrix $T$ has *Toeplitz structure* if there exists scalars $\{r_k\}$ such that $a_{ij} = r_{j-i}$.

**Example** The matrix
$$T = \begin{pmatrix} r_0 & r_1 & r_2 & r_3 \\ r_{-1} & r_0 & r_1 & r_2 \\ r_{-2} & r_{-1} & r_0 & r_1 \\ r_{-3} & r_{-2} & r_{-1} & r_0 \end{pmatrix}$$
has *Toeplitz* structure.

**Definition** A matrix is *persymmetric* if $B = EB^TE^T$, where $E = (e_n, \ldots, e_1)$.



# Symmetric Toeplitz matrices

Suppose $T$ is a symmetric Toeplitz matrix, with diagonals $\{r_k\}$. The Yule-Walker system is $T_n y = -r = -(r_1, \ldots, r_n)$, or
$$\begin{pmatrix} T_{n-1} & E_{n-1}r \\ r^T E_{n-1} & r_0 \end{pmatrix} \begin{pmatrix} \nu \\ \mu \end{pmatrix} = -\begin{pmatrix} r_{n-1} \\ r_{k+1} \end{pmatrix}$$

**Remark** Durbin's algorithm solves the Yule-Walker equations in $\mathcal{O}(n^2)$ operations.

**Lemma** The system $Tx = b$, where $T$ is a symmetric Toeplitz matrix, can be solved using *Levinsons* algorithm in $\mathcal{O}(n^2)$ operations. The inverse $T^{-1}$ can be computed using *Trench's* algorithm in $\mathcal{O}(n^2)$ operations.

## Unsymmetric Toeplitz matrices

Suppose we want to solve a system of the form

$$T = \begin{pmatrix} 1 & r_1 & r_2 & r_3 \\ p_1 & 1 & r_1 & r_2 \\ p_2 & p_1 & 1 & r_1 \\ p_3 & p_2 & p_1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix}$$

This can be done in $\mathcal{O}(n^2)$ operations.

**Remark** This means *Toeplitz* matrices can be used as preconditioners for linear systems derived from finite difference approximations of *PDE*s.



## Perturbation Results

**Lemma** $B^{-1} = A^{-1} - B^{-1}(B - A)A^{-1}$.

**Remark** The special case of a rank 1 update $B = A + uv^T$ is called the Sherman-Morrison formula

$$(A + uv^T)^{-1} = A^{-1} - A^{-1}u(1 + v^T A^{-1}u)^{-1}v^T A^{-1}.$$

Special structures, e.g. Toeplitz or Banded, makes $A^{-1}$ easy to compute. Update formulas matrices that are "close" to a special structure cheaper to invert.



**Example** Suppose we have a decomposition $PA = LU$ and want to solve $\widehat{A}x = b$, where $A$ and $\widehat{A}$ only differs on one row.

How to organize the computation? How much work is required?