Numerical Linear Algebra
Lecture 2

NLA notions

- Methods of Matrix Inversion
- Eigenvalues and eigenvectors
- Norms
- Stability of polynomial evaluation

Invertible matrix

- In linear algebra an n-by-n (square) matrix A is called invertible (some authors use nonsingular or nondegenerate) if there exists an n-by-n matrix B such that $\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n$, where I_n denotes the n-by-n identity matrix and the multiplication used is ordinary matrix multiplication. If this is the case, then the matrix B is uniquely determined by A and is called the inverse of A, denoted by A^{-1}. It follows from the theory of matrices that if $\mathbf{AB} = \mathbf{I}$ for finite square matrices A and B, then also $\mathbf{BA} = \mathbf{I}$.
- Non-square matrices (m-by-n matrices which do not have an inverse). However, in some cases such a matrix may have a left inverse or right inverse. If A is m-by-n and the rank of A is equal to n, then A has a left inverse: an n-by-m matrix B such that $BA = I$. If A has rank m, then it has a right inverse: an n-by-m matrix B such that $AB = I$.
- A square matrix that is not invertible is called singular or degenerate. A square matrix is singular if and only if its determinant is 0.

Methods of matrix inversion

- Gaussian elimination
- Gauss-Jordan elimination is an algorithm that can be used to determine whether a given matrix is invertible and to find the inverse.
- An alternative is the LU decomposition which generates upper and lower triangular matrices which are easier to invert. For special purposes, it may be convenient to invert matrices by treating mn-by-mn matrices as m-by-m matrices of n-by-n matrices, and applying one or another formula recursively (other sized matrices can be padded out with dummy rows and columns).
- For other purposes, a variant of Newton's method may be convenient. Newton's method is particularly useful when dealing with families of related matrices: sometimes a good starting point for refining an approximation for the new inverse can be the already obtained inverse of a previous matrix that nearly matches the current matrix. Newton's method is also useful for "touch up" corrections to the Gauss-Jordan algorithm which has been contaminated by small errors due to imperfect computer arithmetic.

Eigendecomposition

Let A be a square $n \times n$ matrix. Let $q_1 ... q_k$ be an eigenvector basis, i.e. an indexed set of k linearly independent eigenvectors, where k is the dimension of the space spanned by the eigenvectors of A. If $k = n$, then A can be written

$$\mathbf{A} = \mathbf{QUQ}^{-1}$$

where Q is the square $n \times n$ matrix whose i-th column is the basis eigenvector q_i of A, and U is the diagonal matrix whose diagonal elements are the corresponding eigenvalues, i.e. $U_{ii} = \lambda_i$.

Properties

Let A be an $n \times n$ matrix with eigenvalues $\lambda_i, i = 1, 2, \ldots, n$. Then

- Trace of A
$$\mathrm{tr}(A) = \sum_{i=1}^{n} \lambda_i = \lambda_1 + \lambda_2 + \cdots + \lambda_n.$$
- Determinant of A
$$\det(A) = \prod_{i=1}^{n} \lambda_i = \lambda_1 \lambda_2 \cdots \lambda_n.$$
- Eigenvalues of A^k are $\lambda_1^k, \ldots, \lambda_n^k$.

  These first three results follow by putting the matrix in upper-triangular form, in which case the eigenvalues are on the diagonal and the trace and determinant are respectively the sum and product of the diagonal.
- If $A = A^H$, i.e., A is Hermitian ($A = \overline{A^T}$), every eigenvalue is real.
- Every eigenvalue of unitary matrix U ($U^* U = U U^* = I$) has absolute value $|\lambda| = 1$.

Example

We take a 2×2 matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 1 & 3 \end{bmatrix}$$

and want it to be decomposed into a diagonal matrix. First, we multiply to a non-singular matrix

$$\mathbf{B} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, [a, b, c, d] \in \mathbb{R}.$$

Then

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix},$$

for some real diagonal matrix

$$\begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix}.$$

Shifting B to the right hand side:

$$\begin{bmatrix} 1 & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix}$$

The above equation can be decomposed into 2 simultaneous equations:

$$\begin{cases} \begin{bmatrix} 1 & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} a \\ c \end{bmatrix} = \begin{bmatrix} ax \\ cx \end{bmatrix} \\ \begin{bmatrix} 1 & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} b \\ d \end{bmatrix} = \begin{bmatrix} by \\ dy \end{bmatrix} \end{cases}$$

Factoring out the eigenvalues x and y:

$$\begin{cases} \begin{bmatrix} 1 & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} a \\ c \end{bmatrix} = x \begin{bmatrix} a \\ c \end{bmatrix} \\ \begin{bmatrix} 1 & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} b \\ d \end{bmatrix} = y \begin{bmatrix} b \\ d \end{bmatrix} \end{cases}$$

Letting

$$\overrightarrow{a} = \begin{bmatrix} a \\ c \end{bmatrix}, \overrightarrow{b} = \begin{bmatrix} b \\ d \end{bmatrix},$$

this gives us two vector equations:

$$\begin{cases} A\overrightarrow{a} = x\overrightarrow{a} \\ A\overrightarrow{b} = y\overrightarrow{b} \end{cases}$$

And can be represented by a single vector equation involving 2 solutions as eigenvalues:

$\mathbf{Au} = \lambda\mathbf{u}$

where λ represents the two eigenvalues x and y, $\mathbf{u}$ represents the vectors $\overrightarrow{a}$ and $\overrightarrow{b}$.

Shifting $\lambda\mathbf{u}$ to the left hand side and factorizing $\mathbf{u}$ out

$(\mathbf{A} - \lambda\mathbf{I})\mathbf{u} = 0$

Since $\mathbf{B}$ is non-singular, it is essential that $\mathbf{u}$ is non-zero. Therefore,

$(\mathbf{A} - \lambda\mathbf{I}) = \mathbf{0}$

Considering the determinant of $(\mathbf{A} - \lambda\mathbf{I})$,

$$\begin{bmatrix} 1-\lambda & 0 \\ 1 & 3-\lambda \end{bmatrix} = 0$$

Thus

$(1 - \lambda)(3 - \lambda) = 0$

Giving us the solutions of the eigenvalues for the matrix $\mathbf{A}$ as $\lambda = 1$ or $\lambda = 3$, and the resulting diagonal matrix from the eigendecomposition of $\mathbf{A}$ is thus
$\begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$.
Putting the solutions back into the above simultaneous equations

$$\begin{cases} \begin{bmatrix} 1 & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} a \\ c \end{bmatrix} = 1 \begin{bmatrix} a \\ c \end{bmatrix} \\ \begin{bmatrix} 1 & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} b \\ d \end{bmatrix} = 3 \begin{bmatrix} b \\ d \end{bmatrix} \end{cases}$$

Solving the equations, we have $a = -2c, a \in \mathbb{R}$ and $b = 0, d \in \mathbb{R}$
Thus the matrix $\mathbf{B}$ required for the eigendecomposition of $\mathbf{A}$ is
$\begin{bmatrix} -2c & 0 \\ c & d \end{bmatrix}$, [c, d]$\in \mathbb{R}$.*i.e.* :

$$\begin{bmatrix} -2c & 0 \\ c & d \end{bmatrix}_{-1} \begin{bmatrix} 1 & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} -2c & 0 \\ c & d \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}, [c, d] \in \mathbb{R}$$

Methods of matrix inversion

- Eigendecomposition

  If matrix A can be eigendecomposed and if none of its eigenvalues are zero, then A is nonsingular and its inverse is given by $\mathbf{A}^{-1} = \mathbf{Q}\Lambda^{-1}\mathbf{Q}^{-1}$.

  Furthermore, because U is a diagonal matrix, its inverse is easy to calculate: $\left[\Lambda^{-1}\right]_{ii} = \frac{1}{\lambda_i}$.

- Cholesky decomposition

  If matrix A is positive definite, then its inverse can be obtained as $\mathbf{A}^{-1} = (\mathbf{L}^*)^{-1}\mathbf{L}^{-1}$, where L is the lower triangular Cholesky decomposition of A.

Methods of matrix inversion

- Analytic solution

  Writing the transpose of the matrix of cofactors, known as an adjugate matrix, can also be an efficient way to calculate the inverse of small matrices, but this recursive method is inefficient for large matrices. To determine the inverse, we calculate a matrix of cofactors:

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|}\left(\mathbf{C}^{\mathrm{T}}\right)_{ij} = \frac{1}{|\mathbf{A}|}\left(\mathbf{C}_{ji}\right) = \frac{1}{|\mathbf{A}|}\begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{21} & \cdots & \mathbf{C}_{n1} \\ \mathbf{C}_{12} & \mathbf{C}_{22} & \cdots & \mathbf{C}_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_{1n} & \mathbf{C}_{2n} & \cdots & \mathbf{C}_{nn} \end{pmatrix}$$

  where $|A|$ is the determinant of A, C_{ij} is the matrix of cofactors, and C^T represents the matrix transpose.

Inversion of 2×2 matrices

The cofactor equation listed above yields the following result for 2×2 matrices. Inversion of these matrices can be done easily as follows:

$$\mathbf{A}^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{\det(\mathbf{A})} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

This is possible because $1/(ad - bc)$ is the reciprocal of the determinant of the matrix in question, and the same strategy could be used for other matrix sizes.

Inversion of 3×3 matrices

A computationally efficient 3×3 matrix inversion is given by

$$\mathbf{A}^{-1} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & k \end{bmatrix}^{-1} = \frac{1}{\det(\mathbf{A})} \begin{bmatrix} A & B & C \\ D & E & F \\ G & H & K \end{bmatrix}^{T} = \frac{1}{\det(\mathbf{A})} \begin{bmatrix} A & D & G \\ B & E & H \\ C & F & K \end{bmatrix}$$

where the determinant of A can be computed by applying the rule of Sarrus as follows:
$\det(\mathbf{A}) = a(ek - fh) - b(kd - fg) + c(dh - eg)$.
If the determinant is non-zero, the matrix is invertible, with the elements of the above matrix on the right side given by

$$\begin{array}{lll} A = (ek - fh) & D = (ch - bk) & G = (bf - ce) \\ B = (fg - dk) & E = (ak - cg) & H = (cd - af) \\ C = (dh - eg) & F = (gb - ah) & K = (ae - bd). \end{array}$$

Eigenvalues and eigenvectors

- The vector x is an eigenvector of the matrix A with eigenvalue λ (lambda) if the following equation holds: $\mathbf{Ax} = \lambda\mathbf{x}$.

  If the eigenvalue $\lambda > 1, x$ is stretched by this factor. If $\lambda = 1$, the vector x is not affected at all by multiplication by A. If $0 < \lambda < 1$, x is shrunk (or compressed). The case $\lambda = 0$ means that x shrinks to a point (represented by the origin), meaning that x is in the kernel of the linear map given by A. If $\lambda < 0$ then x flips and points in the opposite direction as well as being scaled by a factor equal to the absolute value of λ.

- As a special case, the identity matrix I is the matrix that leaves all vectors unchanged: $I\mathbf{x} = 1\mathbf{x} = \mathbf{x}.$,

- Every non-zero vector x is an eigenvector of the identity matrix with eigenvalue $\lambda = 1$.

Eigenvalues and eigenvectors

- The eigenvalues of A are precisely the solutions λ to the equation $\det(A - \lambda I) = 0$.

  Here *det* is the determinant of the matrix formed by $A - \lambda I$. This equation is called the characteristic equation of A. For example, if A is the following matrix (a so-called diagonal matrix):

$$A = \begin{bmatrix} a_{1,1} & 0 & \cdots & 0 \\ 0 & a_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & a_{n,n} \end{bmatrix},$$

  then the characteristic equation reads

$$\det(A-\lambda I)=\det\left(\begin{bmatrix} a_{1,1} & 0 & \cdots & 0 \\ 0 & a_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & a_{n,n} \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\right)$$

$$= \det \begin{bmatrix} a_{1,1}-\lambda & 0 & \cdots & 0 \\ 0 & a_{2,2}-\lambda & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & a_{n,n}-\lambda \end{bmatrix}$$

$$= (a_{1,1}-\lambda)(a_{2,2}-\lambda)\cdots(a_{n,n}-\lambda)=0.$$

The solutions to this equation are the eigenvalues $\lambda i = ai, i(i = 1, ..., n)$.
The eigenvalue equation for a matrix A can be expressed as
$A\mathbf{x} - \lambda I\mathbf{x} = \mathbf{0}$,
which can be rearranged to $(A - \lambda I)\mathbf{x} = \mathbf{0}$.

A criterion from linear algebra states that a matrix (here: $A - \lambda I$) is non-invertible if and only if its determinant is zero, thus leading to the characteristic equation.

Example

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

The characteristic equation of this matrix reads

$$\det(A - \lambda I) = \det \begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix} = 0.$$

Calculating the determinant, this yields the quadratic equation $\lambda^2 - 4\lambda + 3 = 0$, whose solutions (also called roots) are $\lambda = 1$ and $\lambda = 3$. The eigenvectors for the eigenvalue $\lambda = 3$ are determined by using the eigenvalue equation, which in this case reads

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 3 \begin{bmatrix} x \\ y \end{bmatrix}.$$

This equation reduces to a system of the following two linear equations:

$$2x + y = 3x,$$
$$x + 2y = 3y.$$

Example

Both equations reduce to the single linear equation $x = y$. Or any vector of the form (x, y) with $y = x$ is an eigenvector to the eigenvalue $\lambda = 3$. However, the vector $(0, 0)$ is excluded. A similar calculation shows that the eigenvectors corresponding to the eigenvalue $\lambda = 1$ are given by non-zero vectors (x, y) such that $y = -x$. For example, an eigenvector corresponding to $\lambda = 1$ is

$$\begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

whereas an eigenvector corresponding to $\lambda = 3$ is

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Singular values

The singular values of a compact operator $T : X \to Y$ acting between Hilbert spaces X and Y, are the square roots of the eigenvalues of the nonnegative self-adjoint operator $T^*T : X \to X$ (where T^* denotes the adjoint of T). For the case of a matrix A, singular values are computed as: $\sigma = \sqrt{\lambda(A^*A)}$.
The singular values are nonnegative real numbers, usually listed in decreasing order $(\sigma_1(T), \sigma_2(T), ...)$. If T is self-adjoint, then the largest singular value $\sigma_1(T)$ is equal to the operator norm of T.
In the case of a normal matrix A (or $A^*A = AA^*$, when A is real then $A^TA = AA^T$), the spectral theorem can be applied to obtain unitary diagonalization of A as $A = U\Lambda U^*$. Therefore, $\sqrt{A^*A} = U|\Lambda|U^*$ and so the singular values are simply the absolute values of the eigenvalues.

Norm of a vector

We are going to use following such called L_p-norms which are usually denoted by $\|\cdot\|_p$:

$$\|x\|_p = (\sum_{k=1}^{n} |x_k|^p)^{1/p}, \quad p \geq 1.$$

- $p = 1, \|x\|_1 = \sum_{k=1}^{n} |x_k|$, one-norm
- $p = 2, \|x\|_2 = (\sum_{k=1}^{n} |x_k|^2)^{1/2}$, two-norm
- $p = \infty, \|x\|_\infty = \max_{1 \leq k \leq n} |x_k|$, max-norm or infinity-norm.

Norm of a vector

All these norms, as all other vector norms, has following properties:

- $x \neq 0 \rightarrow \|x\| > 0$ (positivity), $\|0\| = 0$.
- $\|\alpha x\| = |\alpha| \|x\|$ for all $\alpha \in \mathbb{R}$ (homogeneity)
- $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality)

Norm of a vector

Norms are defined differently, but they can be compared. Two norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ on a vector space V are called equivalent if there exist positive real numbers C and D such that for all x in V

$$C\,\|x\|_\alpha \leq \|x\|_\beta \leq D\,\|x\|_\alpha\,.$$

For example, let $x = (x_1, ..., x_n)^T$

$$\alpha\|x\|_1 \leq \|x\|_2 \leq \beta\|x\|_1$$

In this case we have: $\alpha = 1/\sqrt{n}, \beta = 1$.

Norm of a vector

Other examples:

- $$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\,\|x\|_2$$
- $$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\,\|x\|_\infty$$
- $$\|x\|_\infty \leq \|x\|_1 \leq n\,\|x\|_\infty\,,$$
- $$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\,\|x\|_2 \leq n\,\|x\|_\infty$$

If the vector space is a finite-dimensional real or complex one, all norms are equivalent. On the other hand, in the case of infinite-dimensional vector spaces, not all norms are equivalent.

Inner product

The dot product of two vectors $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ is defined as:

$x \cdot y = (x, y) = \sum_{k=1}^{n} x_k y_k = x^T y, \quad \|x\|_2 = \sqrt{x^T x}.$
We note that $x^T y$ is a scalar, but xy^T is a matrix.

Example

$$x^T y = [-1, 2, 3] \cdot \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} = (-1) \cdot (3) + 2 \cdot 2 + 3 \cdot 1 = 4.$$

$$xy^T = \begin{bmatrix} -1 \\ 2 \\ 3 \end{bmatrix} \cdot [3, 2, 1] = \begin{bmatrix} -3 & -2 & -1 \\ 6 & 4 & 2 \\ 9 & 6 & 3 \end{bmatrix}$$

Norm of a vector

Example

$$x = \begin{bmatrix} -1 \\ 2 \\ 3 \\ -5 \end{bmatrix}; \qquad \begin{aligned} ||x||_1 &= |-1| + |2| + |3| + |-5| = 11 \\ ||x||_2 &= \sqrt{(-1)^2 + 2^2 + 3^2 + (-5)^2} = \sqrt{39} \end{aligned}$$

$||x||_\infty = max(|-1|, |2|, |3|, |-5|) = 5.$

Norm of a vector

A vector x is normalized if $||x|| = 1$. If $x \neq 0$ then $\frac{x}{||x||}$ is normalized vector.

Example

$$x^T x = [-1, 2, 3, -5] \cdot \begin{bmatrix} -1 \\ 2 \\ 3 \\ -5 \end{bmatrix} = (-1)\cdot(-1) + 2\cdot 2 + 3\cdot 3 + (-5)^2 = 39$$

$$||x||_2 = \sqrt{39}$$

$$V = \frac{x}{||x||} = \left[\frac{-1}{\sqrt{39}}, \frac{2}{\sqrt{39}}, \frac{3}{\sqrt{39}}, \frac{-5}{\sqrt{39}}\right]^T \Longrightarrow ||V||_2 = 1$$

Matrix norm. Definition

Let K will denote the field of real or complex numbers. Let $K^{m\times n}$ denote the vector space containing all matrices with m rows and n columns with entries in K. Let A^* denotes the conjugate transpose of matrix A.
A matrix norm is a vector norm on $K^{m\times n}$. That is, if $\|A\|$ denotes the norm of the matrix A, then,

- $\|A\| > 0$ if $A \neq 0$ and $\|A\| = 0$ if $A = 0$.
- $\|\alpha A\| = |\alpha|\|A\|$ for all α in K and all matrices A in $K^{m\times n}$.
- $\|A + B\| \leq \|A\| + \|B\|$ for all matrices A and B in $K^{m\times n}$.

Matrix norm. Definition

In the case of square matrices (thus, $m = n$), some (but not all) matrix norms satisfy the following condition, which is related to the fact that matrices are more than just vectors:
$\|AB\| \leq \|A\|\|B\|$ for all matrices A and B in $K^{n\times n}$.
A matrix norm that satisfies this additional property is called a sub-multiplicative norm. The set of all n-by-n matrices, together with such a sub-multiplicative norm, is an example of a Banach algebra.

Matrix norm. Induced norm

If vector norms on K_m and K_n are given (K is field of real or complex numbers), then one defines the corresponding induced norm or operator norm on the space of m-by-n matrices as the following maxima:

$$\|A\| = \max\{\|Ax\| : x \in K^n \text{ with } \|x\| = 1\}$$
$$= \max\left\{\frac{\|Ax\|}{\|x\|} : x \in K^n \text{ with } x \neq 0\right\}.$$

If $m = n$ and one uses the same norm on the domain and the range, then the induced operator norm is a sub-multiplicative matrix norm.

Matrix norm. p-norm

The operator norm corresponding to the p-norm for vectors is:

$$\|A\|_p = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}.$$

In the case of $p = 1$ and $p = \infty$, the norms can be computed as:

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{m} |a_{ij}|,$$

which is simply the maximum absolute column sum of the matrix.

$$\|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}|,$$

which is simply the maximum absolute row sum of the matrix

Matrix norm. Example

For example, if the matrix A is defined by

$$A = \begin{bmatrix} 3 & 5 & 7 \\ 2 & 6 & 4 \\ 0 & 2 & 8 \end{bmatrix},$$

then we have $||A||_1 = max(5, 13, 19) = 19$ and
$||A||_\infty = max(15, 12, 10) = 15$. Consider another example

$$A = \begin{bmatrix} 2 & 4 & 2 & 1 \\ 3 & 1 & 5 & 2 \\ 1 & 2 & 3 & 3 \\ 0 & 6 & 1 & 2 \end{bmatrix},$$

where we add all the entries in each column and determine the greatest value, which results in $||A||_1 = max(6, 13, 11, 8) = 13$.
We can do the same for the rows and get $||A||_\infty = max(9, 11, 9, 9) = 11$.
Thus 11 is our max.

Matrix norm. Example

In the special case of $p = 2$ (the Euclidean norm) and $m = n$ (square matrices), the induced matrix norm is the spectral norm. The spectral norm of a matrix A is the largest singular value of A i.e. the square root of the largest eigenvalue of the positive-semidefinite matrix A^*A:

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^* A)} = \sigma_{\max}(A)$$

where A^* denotes the conjugate transpose of A.

Matrix norm. Example

Example

$$A = \begin{bmatrix} 1 & -2 & -3 \\ 6 & 4 & 2 \\ 9 & -6 & 3 \end{bmatrix}$$

$$||A||_2 = \max \sqrt{\lambda(A^T A)}$$

$$A^T = \begin{bmatrix} 1 & 6 & 9 \\ -2 & 4 & -6 \\ -3 & 2 & 3 \end{bmatrix}, \; A^T A = \begin{bmatrix} 118 & -32 & 36 \\ -32 & 56 & -4 \\ 36 & -4 & 22 \end{bmatrix}$$

$$\lambda(A^T A) = \begin{bmatrix} 8.9683 \\ 45.3229 \\ 141.7089 \end{bmatrix}; \quad \max \sqrt{\lambda(A^T A)} = \max(2.9947, 6.7322, 11.9042) = 11.9042$$

Matrix norm. Example

Example

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \; A^T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \; A^T A - \lambda I = \begin{bmatrix} 1-\lambda & 0 \\ 0 & 1-\lambda \end{bmatrix} = 0;$$

$$\lambda_1 = 1, \; \lambda_2 = 1; \; ||A||_2 = \max \sqrt{\lambda(A^T A)} = \max(1,1) = 1$$

Matrix norm

Any induced norm satisfies the inequality

$$\|A\| \geq \rho(A),$$

where $\rho(A) := \max\{|\lambda_1|, ..., |\lambda_m|\}$ is the spectral radius of A. For a symmetric or hermitian matrix A, we have equality for the 2-norm, since in this case the 2-norm is the spectral radius of A. For an arbitrary matrix, we may not have equality for any norm.

Example

Take

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},$$

the spectral radius $\rho(A)$ of A is 0, but A is not the zero matrix, and so none of the induced norms are equal to the spectral radius of A:

$$\|A\|_1 = 1, \|A\|_\infty = 1, \|A\|_2 = 1.$$

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^* A)} = \sigma_{\max}(A); A^* A = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

Matrix norm. "Entrywise" norms

For square matrices we have the spectral radius formula:

$$\lim_{r\to\infty} \|A^r\|^{1/r} = \rho(A).$$

These vector norms treat an $m \times n$ matrix as a vector of size $m \cdot n$, and use one of the familiar vector norms.
For example, using the p-norm for vectors, we get:

$$\|A\|_p = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^p \right)^{1/p}.$$

The special case $p = 2$ is the Frobenius norm, and $p = \infty$ yields the maximum norm.

Matrix norm. Frobenius norm.

For $p = 2$, this is called the Frobenius norm or the Hilbert - Schmidt norm, though the latter term is often reserved for operators on Hilbert space. This norm can be defined in various ways:

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} = \sqrt{\text{trace}(A^*A)} = \sqrt{\sum_{i=1}^{\min\{m,n\}}\sigma_i^2}$$

where A^* denotes the conjugate transpose of A, σ_i are the singular values of A, and the trace function is used. The Frobenius norm is very similar to the Euclidean norm on K_n and comes from an inner product on the space of all matrices. The Frobenius norm is sub-multiplicative and is very useful for numerical linear algebra. This norm is often easier to compute than induced norms and has the useful property of being invariant under rotations.

Matrix norm. Max norm.

The max norm is the elementwise norm with $p = \infty$:

$$\|A\|_{\max} = \max\{|a_{ij}|\}.$$

This norm is not sub-multiplicative.

Matrix norm. Equivalence of norms.

For any two vector norms $||\cdot||_\alpha$ and $||\cdot||_\beta$, we have

$$r\,\|A\|_\alpha \leq \|A\|_\beta \leq s\,\|A\|_\alpha$$

for some positive numbers r and s, for all matrices A in $K^{m\times n}$. In other words, they are equivalent norms; they induce the same topology on $K^{m\times n}$. This is a special case of the equivalence of norms in finite-dimensional Normed vector spaces.

Examples of norm equivalence

For matrix $A \in \mathbb{R}^{m \times n}$ the following inequalities hold:

- $\|A\|_2 \leq \|A\|_F \leq \sqrt{r}\|A\|_2$, where r is the rank of A
- $\|A\|_{\max} \leq \|A\|_2 \leq \sqrt{mn}\|A\|_{\max}$
- $\frac{1}{\sqrt{n}}\|A\|_\infty \leq \|A\|_2 \leq \sqrt{m}\|A\|_\infty$
- $\frac{1}{\sqrt{m}}\|A\|_1 \leq \|A\|_2 \leq \sqrt{n}\|A\|_1$.

Here, $||\cdot||_p$ refers to the matrix norm induced by the vector p-norm. Another useful inequality between matrix norms is

$$\|A\|_2 \leq \sqrt{\|A\|_1\|A\|_\infty}.$$

Stability of polynomial evaluation

We will discuss stability of polynomial evaluation by Horner's rule. Let the polynomial is given by

$$p(x) = \sum_{i=0}^{d} c_i x^i,$$

where c_i are coefficients of the polynomial, d is its degree.
To compute roots of this polynomial we can use Horner's rule, see alg.below. This rule can be programmed as the following iterative algorithm for every mesh point $x_j \in [x_{left}, x_{right}], j \in 1, 2, ...N$, where N is the total number of the discretization points:
Algorithm 1

- Step 0. Set counter $i = d - 1$ and initialize $p_d = c_d$.
- Step 1. Compute $p_i = x_j \cdot p_{i+1} + c_i$
- Step 2. Set $i := i - 1$ and go to step 1. Stop if $i = 0$.

In our comp.ex. 1 we need evaluate of roots of polynomial. Let us take
$p(x) = (x-9)^9 = x^9 - 81x^8 + 2916x^7 - 61236x^6 + 826686x^5 - 7440174x^4 + 44641044x^3 - 172186884x^2 + 387420489x^1 - 387420489$
together with upper and lower bounds for this solution.
To compute bounds of the solution we insert term with error $1 + (\sigma_{1,2)_i}$ for every floating point iteration in Algorithm 1 to get following algorithm:

Algorithm 2

- Step 0. Set counter $i = d - 1$ and initialize $p_d = c_d$.
- Step 1. Compute
  $p_i = (x_j \cdot p_{i+1}(1 + (\sigma_1)_i) + c_i)(1 + (\sigma_2)_i), \quad |(\sigma_1)_i|, |(\sigma_2)_i| \leq \varepsilon$
- Step 2. Set $i := i - 1$ and go to step 1. Stop if $i = 0$.

Floating-point numbers

We can represent a floating-point number as:

$$x = \pm \left(d_0 + \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + ... + \frac{d_{t-1}}{\beta^{t-1}} \right) \beta^e,$$

where

$$0 \leq d_k \leq \beta - 1, L \leq e \leq U,$$

Here we have:

- β is base or radix
- e exponent
- t precision
- $[L, U]$ is exponent range
- $d_k, k = 0, ..., t-1$ mantissa (integer)

Most of computers now use binary ($\beta = 2$) arithmetics.

base	t	L	U	
2	24	-126	127	32 bit
2	53	-1022	1023	64 bit

In the algorithm above ε is the machine epsilon and we define it as the maximum relative representation error $0.5 \cdot \beta^{1-p}$ which is measured in a floating point arithmetic with the base β and with precision $p > 0$. Now the following values of machine epsilon apply to standard floating point formats:

Table 1. *Values of machine epsilon in standard floating point formats. Notation $*$ means that one bit is implicit in precision p. Machine epsilon ε_1 is computed accordingly to Demmel.*

EEE 754 - 2008	description	Base, b	Precision, p	Machine eps.1 $\varepsilon_1 = 0.5 \cdot b^{-(p-1)}$
binary16	half precision	2	11^*	$2^{-11} = 4.88e-04$
binary32	single precision	2	24^*	$2^{-24} = 5.96e-08$
binary64	double precision	2	53^*	$2^{-53} = 1.11e-16$
binary80	extended precision	2	64	$2^{-64} = 5.42e-20$
binary128	quad. precision	2	113^*	$2^{-113} = 9.63e-35$
decimal32	single prec. decimal	10	7	5×10^{-7}
decimal64	double prec. decimal	10	16	5×10^{-16}
decimal128	quad. prec. decimal	10	34	5×10^{-34}

Expanding expression for p_i in the algorithm 2 we can get the final value p_0 as

$$p_0 = \sum_{i=0}^{d-1} \left((1 + (\sigma_2)_i) \prod_{k=0}^{i-1} (1 + (\sigma_1)_k)(1 + (\sigma_2)_k) \right) c_i x^i + \left(\prod_{k=0}^{d-1} (1 + (\sigma_1)_k)(1 + (\sigma_2)_k) \right) c_d x^d \tag{1}$$

Next, we will write upper and lower bounds for products of $\sigma := \sigma_{1,2}$ provided that $k\varepsilon < 1$:

$$(1 + \sigma_1) \cdot \ldots \cdot (1 + \sigma_k) \leq (1 + \varepsilon)^k \leq 1 + k\varepsilon + O(\varepsilon^2),$$
$$1 - k\varepsilon \leq (1 - \varepsilon)^k \leq (1 + \sigma_1) \cdot \ldots \cdot (1 + \sigma_k) \tag{2}$$

Applying estimate above we can get the following approximation

$$1 - k\varepsilon \leq (1 + \sigma_1) \cdot \ldots \cdot (1 + \sigma_k) \leq 1 + k\varepsilon. \tag{3}$$

Using the estimate above we can rewrite (4) assuming $|\tilde{\sigma}_i| \leq 2d\varepsilon$ as

$$p_0 \approx \sum_{i=0}^{d} (1 + \tilde{\sigma}_i) c_i x^i = \sum_{i=0}^{d} \tilde{c}_i x^i \tag{4}$$

We also can write formula for the computing error in the polynomial:

$$|p_0 - p(x)| = \left| \sum_{i=0}^{d} (1 + \tilde{\sigma}_i) c_i x^i - \sum_{i=0}^{d} c_i x^i \right| = \left| \sum_{i=0}^{d} \tilde{\sigma}_i c_i x^i \right| \leq 2d\varepsilon \sum_{i=0}^{d} |c_i x^i| \tag{5}$$

If we will choose $\tilde{\sigma}_i = \varepsilon \cdot \mathrm{sign}(\mathrm{c_i x^i})$ then the error bound above can be attained within the factor $2d$. In this case we can take $\frac{\sum_{i=0}^{d} |c_i x^i|}{|\sum_{i=0}^{d} c_i x^i|}$ as the relative condition number for the case of polynomial evaluation. The following algorithm computes the lower and upper bound bp in the polynomial evaluation at every point x_j on the interval $[p - bp, p + bp]$.

Algorithm 3

- Step 0. Set counter $i = d - 1$ and initialize $p = c_d$, $bp = |c_d|$.
- Step 1. Compute $p = x_j \cdot p + c_i$, $bp = |x_j| \cdot bp + |c_i|$.
- Step 2. Set $i := i - 1$ and go to step 1. Stop if $i = 0$.
- Step 3. Set $bp = 2 \cdot d \cdot \varepsilon \cdot bp$ as error bound at the point $|x_j|$.

Computation of roots of the polynomial

We want to compute roots of $(x-1)^5 = 0$ in Matlab. To do this we need to rewrite it as:

$$(x-1)^5 = x^5 - 5x^4 + 10x^3 - 10x^2 + 5x - 1$$

We already see that all roots should be $x = 1$. But in Matlab we have:
$r = roots([1\ \ -5\ 10\ \ -10\ 5\ \ -1])$
and computed roots are:
$r = 1.0008 + 0.0006i$
$1.0008 - 0.0006i$
$0.9997 + 0.0009i$
$0.9997 - 0.0009i$
0.9990
Error:
$disp(abs(r-1)')$
1.1322e-03 1.1322e-03 1.1326e-03 1.1326e-03 1.1328e-03

Computation of roots of the polynomial

Why ? Let us analyze the problem

$$(x-1)^5 = \varepsilon,$$

Then we can find that

$$x = 1 + \varepsilon^{1/5}$$

If $\varepsilon = 10^{-15}$ then we will have $\varepsilon^{1/5} = 10^{-15/5} = 10^{-3}$. We observe that zeros of polynomial $(x-1)^5$ are very sensitive to the changes in the coefficients. Is it always like that?

For coefficients

c = [1 -15 85 -225 274 -120];

exact roots are: 1,2,3,4,5:

r = roots(c);

fel = sort(r) - (1:5)'

And the error is very small now

fel = -4.9960e-15 6.6613e-14 -1.5010e-13 9.6811e-14 -8.8818e-16

Condition number

We can see the roots r of the polynomial as functions $f(c)$ depending on the coefficients c .

$$r = f(c).$$

When we perturbate coefficients $c + \delta c$, we also perturbate roots $r + \delta r$.
If small changes in data $|\delta c|/|c|$ gives small changes in the output data, or result $|\delta r|/|r|$, then the problem is called well-posed.
Otherwise, the problem is ill-posed. Condition number is defined as

$$k = \frac{|\delta r|/|r|}{|\delta c|/|c|}$$

It is not always possible compute this number, but often is possible compute estimate for k:

$$|\delta r|/|r| \leq k|\delta c|/|c|.$$