Numerical algebra fundamentals. Part 1

Lecture notes by nazerke.sandibay

1 Floating-point arithmetic, formats

- In fixed point notation, there are a fixed number of digits after the decimal point. A Qm.n number is in the range $[-(2^m), 2^m - 2^{-n}]$, with resolution 2^{-n}.

- **Floating point** represents numbers with relative accuracy and is suitable for the case when numbers in the computations have varying scales (i.e., 10^{-1} and 10^5). Therefore, it is used more often.

Name	Common Name	Base	Digits	Emin	Emax
binary16	half precision	2	11	-14	+ 15
binary32	single precision	2	24	-126	+ 127
binary64	double precision	2	53	-1022	+1023

$$\text{number} = \text{significand} \times \text{base}^{\text{exponent}},$$

Accuracy

The relative accuracy of single precision is $10^{-7} - 10^{-8}$, while for double precision $10^{-14} - 10^{-16}$.

Memory

A float16 takes 2 bytes, float32 takes 4 bytes, float64, or double precision, takes 8 bytes.

Let's calculate the size of a matrix M. Matrix M contains n=width*height elements. Each element is represented in single precision(float32 4 bytes). Size is 4*n bytes. The size of 1000 by 1000 Matrix is 4 Million bytes or 3.8 Megabytes

Application

half precision can be useful in training deep neural network, double precision in computational science and engineering and float on GPU/Data Science

Other formats:

bfloat16 (Brain Floating Point)

Tensor Float from Nvidia

Mixed precision

Possible sources of error:

- rounding-off errors
- check forward/backward stability
- summation error
- subtract two big numbers that are close, the difference will have fewer correct digits

Sum up: there is a trade-off between accuracy and memory, computation time. Inappropriate number representation, operations may cause errors.

Useful source:

Memory Efficient Data Science: Types

https://towardsdatascience.com/memory-efficient-data-science-types-53423d48ba1d

2 Vector norms

The norm should satisfy certain properties:
- $\|\alpha x\| = |\alpha|\|x\|$ (Homogeneity)
- $\|x + y\| \leq \|x\| + \|y\|$ (Triangle inequality)
- If $\|x\| = 0$ then $x = 0$. (Non-Negativity)

The norm of a vector is always non-negative.
p-norms:

$$\|x\|_p = \Big(\sum_{i=1}^{n} |x_i|^p\Big)^{1/p}.$$

$p \geq 1$ and other cases does not satisfy properties of norm.

Note that the L0 norm is not a norm because it does not satisfy the homogeneity property. And 0.5 does not satisfy the triangle inequality property.

Counterexample for p=0: $\|x\|_0 = \sum_{i=1}^{n} |x_i|^0$, **for x=(1,0) and** $\alpha = 2$
$\|\alpha x\|_0 = \|2 * (1,0)\|_0 = 1$ and
$|\alpha|\|x\|_0 = 2 * \|(1,0)\|_0 = 2$

Counterexample for p=0.5. $\|x\|_{0.5} = (\sum_{i=1}^{n} |x_i|^{0.5})^2$ **for x = (1,1) and y = (1,0)**
$\|x + y\|_{0.5} = \|(2,1)\|_{0.5} = (\sqrt{2} + \sqrt{1})^2 = 3+2\sqrt{2}$ and
$\|x\| + \|y\|_{0.5} = \|(1,1)\|_{0.5} + \|(1,0)\|_{0.5} = (\sqrt{1} + \sqrt{1})^2 + (\sqrt{1} + \sqrt{0})^2 = 5$
Special cases:

1. The most well-known and widely used norm is L_2 norm (**Euclidean norm**):

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2},$$

which corresponds to the distance in real life. If the vectors have complex elements, we use their **modulus**.

2. Infinity norm, or **Chebyshev norm** is defined as the maximal element:

$$\|x\|_\infty = \max_i |x_i|$$

The more you increase the p, the more important the largest term becomes and increases exponentially. Similarly, in the case of inf norms, the contribution of the maximum element dominates.

3. L_1 norm (or **Manhattan distance**) which is defined as the sum of modules of the elements of x:

$$\|x\|_1 = \sum_i |x_i|$$

L1 is used in compressed sensing as a surrogate for sparsity.

$$\|x\|_1 = \sum_{i=1}^{m} |x_i|,$$

$$\|x\|_2 = \left(\sum_{i=1}^{m} |x_i|^2 \right)^{1/2} = \sqrt{x^* x},$$

$$\|x\|_\infty = \max_{1 \le i \le m} |x_i|,$$

$$\|x\|_p = \left(\sum_{i=1}^{m} |x_i|^p \right)^{1/p} \quad (1 \le p < \infty).$$

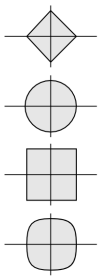

Figure 1: Vector norms

The set of all (x,y) points for which the vector v= (x,y) has a norm equal to 1 are visualized in Figure 1.

To sum up,if we minimize L1 norm , it will try to drive weights to 0, inducing sparsity. L2 norm will reduce all weights but not all the way to 0. L infinity norm will try to decrease maximum element.

Equivalence of the norms

All norms are equivalent in the sense that

$$C_1 \|x\|_* \le \|x\|_{**} \le C_2 \|x\|_*$$

for some positive constants $C_1(n), C_2(n)$, $x \in \mathrm{R}^n$ for any pairs of norms $\|\cdot\|_*$ and $\|\cdot\|_{**}$. The equivalence of the norms basically means that if the vector is small in one norm, it is small in another norm.

Some examples:
$\|x\|_2 \le \|x\|_1 \le \sqrt{n}\|x\|_2$
$\|x\|_\infty \le \|x\|_2 \le \sqrt{n}\|x\|_\infty$
$\|x\|_\infty \le \|x\|_1 \le n\|x\|_\infty$
In one line:
$\|x\|_\infty \le \|x\|_2 \le \|x\|_1 \le \sqrt{n}\|x\|_2 \le n\|x\|_\infty$

- Typical choice of $\|x\| = \|x\|_2$ leads to the **linear least squares problem** (and has been used for ages).
- The choice $\|x\| = \|x\|_1$ leads to the **compressed sensing**
- It typically yields the **sparsest solution**

$$\|x\| \to \min_x$$

$$\text{subject to } Ax = f$$

Forward and backward stability

- Let x be an object (for example, a vector)
- Let $f(x)$ be the function (functional) you want to evaluate

You also have a **numerical algorithm** alg(x) that actually computes **approximation** to $f(x)$.

The algorithm is called **forward stable**, if (The error is very small)

$$\|alg(x) - f(x)\| \leq \varepsilon$$

The algorithm is called **backward stable**, if for any x there is a close vector $x + \delta x$ such that

$$alg(x) = f(x + \delta x)$$

3 Matrix. Terminology

Given matrix A

Range of A - the linear space, a set of all y such that: $y = Ax$. It is also called an image and denoted by range(A) (or im(A))

Rank of A - a maximal number of linearly independent *columns* in a matrix A, or the **dimension of its column space** $= \dim \mathrm{im}(A)$.

Nullspace of A - the set of all solutions to the homogeneous equation $\sum_i a_i x_i = 0$, or in the matrix form: $Ax = 0, \quad \|x\| \neq 0$.

Non-singular matrix (or regular matrix) can be defined as a matrix whose determinant is a non-zero value and the non-singular matrix property is to be satisfied to find the inverse of a matrix.

4 Matrix Norms

$\|\cdot\|$ is called a **matrix norm** if it is a vector norm on the vector space of $n \times m$ matrices:

- $\|A\| \geq 0$ and if $\|A\| = 0$ then $A = 0$
- $\|\alpha A\| = |\alpha| \|A\|$
- $\|A + B\| \leq \|A\| + \|B\|$

Additionally, some norms can satisfy the **submultiplicative** property

$$\|AB\| \leq \|A\| \|B\|$$

These norms are called **submultiplicative** norms. Example of a non-submultiplicative norm is **Chebyshev** norm

$$\|A\|_C = \max_{i,j} |a_{ij}|$$

Important matrix norms

- Frobenuis norm, denoted by $\|A\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} (a_{ij})^2}$

Matrix p-norms

Important case of operator norms are matrix p-norms, which are defined for $\|\cdot\|_* = \|\cdot\|_{**} = \|\cdot\|_p$.

Among all p-norms three norms are the most common ones:

- $p = 1, \quad \|A\|_1 = \max_j \sum_{i=1}^{n} |a_{ij}|$.

Note that it is a maximum of absolute value, we sum the absolute values down each column and then take the biggest answer

- $p = 2$, spectral norm, denoted by $\|A\|_2$.It is directly related to the **singular value decomposition** (SVD) of the matrix. It holds

$$\|A\|_2 = \sigma_1(A) = \sqrt{\lambda_{\max(A^* A)}}$$

where $\sigma_1(A)$ is the largest singular value of the matrix A and * is a *conjugate transpose* of the matrix.

it can also be solved by optimization problem

- $p = \infty$, $\|A\|_\infty = \max_i \sum_{j=1}^{m} |a_{ij}|$.

Example:

$$A = \begin{bmatrix} 1 & 2 \\ 0 & 2 \end{bmatrix}, \|A\|_1 = 4, \|A\|_2 = 2.92, \|A\|_\infty = 3, \|A\|_F = 3$$

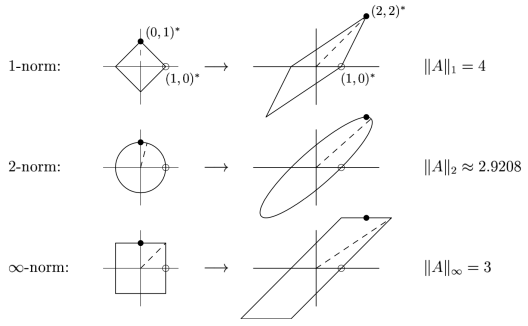

Figure 2: Matrix norms

Matrix norms are visualized in Figure 2. On the left, is the area covered by vectors with the unit norm. On the right, their images under matrix A. Dashed lines mark the vectors that are amplified most by A in each norm.

Operator norms

- The most important class of the matrix norms is the class of **operator norms**. They are defined as

$$\|A\|_{*,**} = \sup_{x \neq 0} \frac{\|Ax\|_*}{\|x\|_{**}},$$

  where $\|\cdot\|_*$ and $\|\cdot\|_{**}$ are **vector norms**.

- **Frobenius norm** is a matrix norm, but not an operator norm, i.e. you can not find $\|\cdot\|_*$ and $\|\cdot\|_{**}$ that induce it. Think about identity matrix.

Scalar(inner) product

While the norm is a measure of distance, the **scalar product** takes angle into account. It is defined as

- **For vectors:**

$$(x, y) = x^* y = \sum_{i=1}^{n} \overline{x}_i y_i,$$

where $\overline{x}$ denotes the *complex conjugate* of x. The Euclidean norm is then

$$\|x\|_2 = \sqrt{(x, x)},$$

or it is said the norm is **induced** by the scalar product.

- **For matrices** (Frobenius scalar product):

$$(A, B)_F = \sum_{i=1}^{n} \sum_{j=1}^{m} \overline{a}_{ij} b_{ij} \equiv \mathrm{trace}(A^* B),$$

where $\mathrm{trace}(A)$ denotes the sum of diagonal elements of A. One can check that $\|A\|_F = \sqrt{(A, A)_F}$.

The angle between two vectors is defined as

$$\cos \phi = \frac{(x, y)}{\|x\|_2 \|y\|_2}.$$

Similar expression holds for matrices.

Cauchy-Schwarz-Bunyakovsky inequality

An important property of the scalar product is the **Cauchy-Schwarz-Bunyakovski inequality**:

$$|(x, y)| \leq \|x\|_2 \|y\|_2,$$

and thus the angle between two vectors is defined properly.

5 Matrix types and operations with them

Unitary matrix

Complex $n \times n$ square matrix is called **unitary** if

$$U^*U = I_n,$$

Properties:

- $\|Uz\|_2 = \|z\|_2$ for all z meaning it preserve norms
- Columns and rows of unitary matrices both form orthonormal basis in C^n
- $U^*U = I_n$ — left unitary for $m > n$
- $UU^* = I_m$ — right unitary for $m < n$
- Important property: **a product of two unitary matrices is a unitary matrix:**

$$(UV)^*UV = V^*(U^*U)V = V^*V = I,$$

- Unitary matrices also do not change matrix norms $\|\cdot\|_2$ and $\|\cdot\|_F$, i.e. for any square A and unitary U,V:

$$\|UAV\|_2 = \|A\|_2 \qquad \|UAV\|_F = \|A\|_F.$$

Important examples

- Permutation matrix P whose rows (columns) are permutation of rows (columns) of the identity matrix.
- Fourier matrix

$$F_n = \frac{1}{\sqrt{n}} \left\{ e^{-i\frac{2\pi kl}{n}} \right\}_{k,l=0}^{n-1}$$

- Householder matrices
- Givens (Jacobi) matrices

  Householder and Givens are two important classes of unitary matrices, using composition of which we can construct any unitary matrix

Singular Value Decomposition

Any matrix $A \in \mathbb{C}^{n \times m}$ can be written as a product of three matrices:

$$A = U\Sigma V^*,$$

where - U is an $n \times n$ unitary matrix - V is an $m \times m$ unitary matrix - Σ is a diagonal matrix with non-negative elements $\sigma_1 \geq \ldots, \geq \sigma_{\min(m,n)}$ on the diagonal.

Moreover, if $\text{rank}(A) = r$, then $\sigma_{r+1} = \cdots = \sigma_{\min(m,n)} = 0$.

If one truncates (replace by 0) all singular values except for r first, then the resulting matrix yields best rank-r approximation both in $\|\cdot\|_2$ and $\|\cdot\|_F$.

$\min\limits_{\text{rank}(A_r)=r} \|A - A_r\|$ — finding best rank-r approximation.

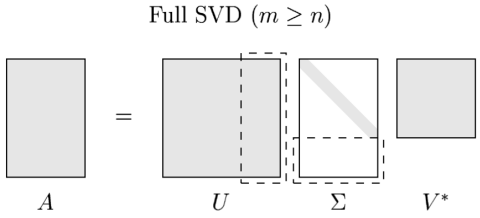

Figure 3: SVD visualization

6 Other important notes

Apart from above mentioned points, errors may be specific to the algorithm itself. For example, Nyquist–Shannon sampling theorem (the sampling rate should be twice higher then signal's highest frequency). Also, grid size cause errors due to approximation. The linearized version of dynamic systems around some specific point causes errors around regions far from the point of linearization.

The figures are taken from "Numerical Linear Algebra" by Lloyd N. Trefethen and David Bau.

Materials are summarized from lecture notes from "Numerical Linear Algebra" course by I.Oseledets and "Numerical Linear Algebra" book by Lloyd N. Trefethen and David Bau.