## Band - Matrix :-

A matrix $A = (a_{ij})_{n \times n}$ is called a

- band - matrix if $a_{ij} = 0$ if $j < i-k$ or $j > i+l$.

- sparse matrix if more than half of the entries are zero.

$\Rightarrow$ $k$ & $l$ are called the lower & upper bandwidth respectively.

$$\text{bandwidth of } A = \max\{k, l\}.$$

# $A = LU$ is unique if $A$ is non-singular for singular matrices the result may not be true!

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & c & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = L_c U$$

for any $c \in \mathbb{R}$

Exm:

Diagonal matrix if $k = l = 0$

Tri-diagonal " if $k = l = 1$

upper triangular " if $k = 0, \; l = n-1$

## Storage of a band matrix :-

$$A = \begin{pmatrix} 2 & 3 & 0 & 0 \\ 4 & 1 & 3 & 0 \\ 0 & 3 & -1 & 5 \\ 0 & 0 & -1 & 2 \end{pmatrix} \qquad k = 1, \; l = 1$$

The matrix can be stored as

$$A' = \begin{pmatrix} 0 & 2 & 3 \\ 4 & 1 & 3 \\ 3 & -1 & 5 \\ -1 & 2 & 0 \end{pmatrix}_{4 \times 3}$$

## Symmetric band matrix

$$A = \begin{pmatrix} 2 & 3 & 0 & 0 \\ 3 & 1 & 4 & 0 \\ 0 & 4 & 2 & 5 \\ 0 & 0 & 5 & 1 \end{pmatrix}$$ is stored

as $$A^* = \begin{pmatrix} 2 & 3 \\ 1 & 4 \\ 2 & 5 \\ 1 & 0 \end{pmatrix}$$

## Skyline method :

If a sparse band matrix is symmetric and has few zero entries, then it can be stored in single array.

$$A = \begin{pmatrix} 2 & 0 & 3 & 0 & 0 & 0 \\ & 1 & 0 & 5 & 0 & 0 \\ & & 5 & 0 & 0 & 0 \\ & & & 6 & -1 & 0 \\ & & & & 3 & 2 \\ & & & & & 1 \end{pmatrix}_{6\times 6}$$


$$A' = [2\ 1\ 3\ 0\ 5\ 5\ 0\ 6\ -1\ 3\ 2\ 1]$$

$$A'' = [1\ 2\ 5\ 8\ 10\ 12]$$

Where the elements of $A'$ represent the element values of each column up to the diagonal position.

It significantly reduce the computer memory requirement.

$A''$ stores the indexes of the diagonal entries of $A$.

**Theorem** Let, $A$ be a band matrix with lower bandwidth $K$ & upper bandwidth $l$. Let, $A = LU$ be computed without pivoting. Then $L$ has lower bandwidth $K$ and $U$ has upper bandwidth $l$.

**Theorem** Let, $A$ be a band matrix with lower bandwidth $K$ and upper bandwidth $l$. Let, $PA = LU$ is computed by partial pivoting. Then $U$ is banded with upper bandwidth $K+l$ and $L$ is banded with lower bandwidth $l$.

## Vector & Matrix Norm :-

Let, $V \left(\subseteq \mathbb{R}^n \text{ or } \mathbb{C}^n\right)$ be a vector space over $\mathbb{R}$ (or $\mathbb{C}$)

A norm is a function $\|\cdot\| : V \to \mathbb{R}^+$ satisfying the following properties:

a) $\forall\, x \in V$, $\|x\| \geq 0$; $\|x\| = 0$ iff $x = 0$.

b) $\|\lambda x\| = |\lambda| \|x\|$ $\quad \forall\, x \in V$, $\lambda \in \mathbb{F}$ $(\mathbb{R}$ or $\mathbb{C})$.

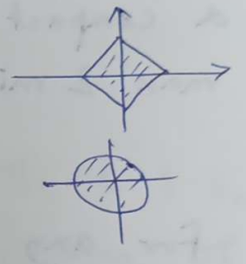
c) $\|x+y\| \leq \|x\| + \|y\|$ $\quad \forall\, x, y \in V$

$\Rightarrow$ We call a vector $x \in V$, a unit vector if $\|x\| = 1$.

Examples :- Let, $x \in \mathbb{C}^n$, $x = (x_1, x_2 \ldots x_n)$

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \qquad (\ell_1 \text{ norm})$$

Unit ball

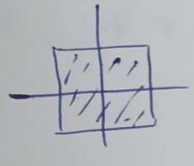


$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}$$

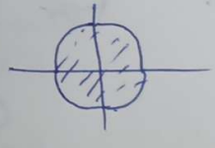
(Euclidean norm)



$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i| \qquad (\ell_\infty \text{ norm})$$



$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, \quad 1 \leq p < \infty$$



$(\ell_p \text{ norm})$

**Theorem** On $\mathbb{R}^n$ (or $\mathbb{C}^n$), all norms are equivalent, i.e. for two norms $\|\cdot\|$ & $\|\cdot\|'$ on $\mathbb{R}^n$, there are two constants $C_1$ & $C_2$ such that

$$C_1 \|x\| \leq \|x\|' \leq C_2 \|x\|, \quad \forall\, x \in \mathbb{R}^n.$$

Hint: This relation is an equivalence relation (Reflexive, Symmtric transitive)

So, sufficient to show, all norms are equivalent to $\|\cdot\|_1$ norm $(\ell_1)$

Let $\|\cdot\|$ be any norm on $\mathbb{R}^n$.

Let, $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$

$$\|x\| = \Big\| \sum_{j=1}^{n} x_j e_j \Big\| \le \sum_{j=1}^{n} |x_j| \, \|e_j\|$$

$$\le \Big(\max_j \|e_j\|\Big) \sum_{j=1}^{n} |x_j|$$

$$= \Big(\max_j \|e_j\|\Big) \|x\|_1$$

$$= M \cdot \|x\|_1 .$$

Consider the function: $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$.

Since, $|\|x\| - \|y\|| \le \|x-y\| \le M \|x-y\|_1$, $x, y \in \mathbb{R}^n$

$\|\cdot\|$ is continuous on $\mathbb{R}^n$ w.r.to $\|\cdot\|_1$ norm.

Since, the unit sphere $\mathbb{D} = \{x \in \mathbb{R}^n ; \|x\|_1 = 1\}$ is a compact set, the cont. function $\|\cdot\|$ attains max & min on it. So, $\exists\ c \le C$ s.t.

$$c \le \|x\| \le C \quad \forall\ x \in K$$

For any $y \in \mathbb{R}^n$, $\dfrac{y}{\|y\|_1} \in \mathbb{D}$.

$$\text{So, } c\|y\|_1 \le \|y\| \le C \|y\|_1$$

---

Exm. For standard $\ell_p$ norms, we have the following:

(i) $\|x\|_2 \le \|x\|_1 \le \sqrt{n}\, \|x\|_2$

(ii) $\|x\|_\infty \le \|x\|_2 \le \sqrt{n}\, \|x\|_\infty$

(iii) $\|x\|_\infty \le \|x\|_1 \le n \|x\|_\infty$

(iv) $\|x\|_\infty \le \|x\|_p \quad p \ge 1$

(v) $\|x\|_p \le \|x\|_q \quad p \ge q$

(vi) $\|x\|_p \le n^{1/p - 1/q} \|x\|_q \quad q > p \ge 1$

(vii) $\|x\|_p \le n^{1/p} \|x\|_\infty \quad p \ge 1$

## Matrix Norm

$\|\cdot\|$ is a matrix norm on $m \times n$ matrices if it is a vector norm on $mn$ dimensional space:

i) $\|A\| \geq 0$ & $\|A\| = 0 \Leftrightarrow A = 0$

ii) $\|\alpha A\| = |\alpha| \|A\|$, $\alpha \in \mathbb{C}$ (or $\mathbb{R}$)

iii) $\|A+B\| \leq \|A\| + \|B\|$

Along with these three axioms, another condition is imposed for consistent norms:

iv) $\|AB\|_{mp} \leq \|A\|_{mn} \|B\|_{np}$, $A \in M_{m \times n}$, $B \in M_{n \times p}$.

Exm. $\|A\|_F := \left(\sum |a_{ij}|^2\right)^{1/2}$ is called the Frobenius norm.

Induced norm:

Given vector norms $\|\cdot\|_n$ & $\|\cdot\|_m$, the induced norm or subordinate matrix norm of $A \in M_{m \times n}$ is the smallest ~~norm~~ $c$ such that $\forall x \in \mathbb{C}^n$

$$\|Ax\|_m \leq c \|x\|_n \quad \text{holds.}$$

In other word,

$$\|A\|_{m,n} = \sup_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Ax\|_m}{\|x\|_n}$$

$$= \sup_{\substack{x \in \mathbb{C}^n \\ \|x\|_n = 1}} \|Ax\|_m$$

So, $\frac{\|Ax\|_m}{\|x\|_n} \leq \|A\|_{m,n}$.

$\therefore$ $\|Ax\|_m \leq \|A\|_{m,n} \|x\|_n$

Note:

$$\|I\|_{n,n} = \sup_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Ax\|_n}{\|x\|_n} = 1.$$

Whereas, $\|I\|_F = \sqrt{n}$. So, $\|\cdot\|_F$ is not an induced nor

**Result:**

Let, $A \in \mathbb{R}^{m \times n}$ Then,

$$\|A\|_1 = \sup_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \frac{\|Ax\|_1}{\|x\|_1} = \max_j \sum_{i=1}^{m} |a_{ij}|$$

= max absolute column sum

$$\|A\|_\infty = \sup_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \frac{\|Ax\|_\infty}{\|x\|_\infty} = \max_i \sum_{j=1}^{n} |a_{ij}|$$

= max absolute row sum.

**Proof:**

Let, $A = [a_1\ a_2 \dots a_n]$, $a_i$'s are n columns of A.

$$\|Ax\|_1 = \left\| \sum_{j=1}^{n} x_j a_j \right\|_1 \leq \sum_{j=1}^{n} |x_j| \|a_j\|_1$$

$$\leq \left(\max_j \|a_j\|_1\right) \|x\|_1$$

So, $x \neq 0 \Rightarrow \sup_{x \in \mathbb{R}^n} \frac{\|Ax\|_1}{\|x\|_1} \leq \max_j \|a_j\|_1$

$$= \max_j \left(\sum_{i=1}^{m} |a_{ij}|\right)$$

Now, Choose $x = e_j$, where $j$ maximizes $\|a_j\|_1$

$$\|Ae_j\|_1 = \|a_j\|_1 \|e_j\|_1$$

So, $$\|A\|_1 = \max_j \|a_j\|_1$$

2nd $$\|A\|_\infty = \max_{\|x\|_\infty = 1} \|Ax\|_\infty$$

$$= \max_{\|x\|_\infty = 1} \max_{1 \leq i \leq m} \left| \sum_{j=1}^{n} a_{ij} x_j \right|$$

$$\leq \max_{\|x\|_\infty = 1} \max_i \sum_{j=1}^{n} |a_{ij}| |x_j|$$

$$\leq \max_i \sum_{j=1}^{n} |a_{ij}| = \sum_{j=1}^{n} |a_{kj}| \text{ for some } k.$$

To show the upper bound is attained, we choose

$$x_j = \begin{cases} \frac{a_{kj}}{|a_{kj}|}, & a_{kj} \neq 0 \\ 0 & \text{else} \end{cases}$$

and $x_0 = \begin{pmatrix} x_1 \\ \vdots \\ x_j \\ \vdots \\ x_n \end{pmatrix}$

Then, $\max_{\|x\|_\infty = 1} \|Ax_0\|_\infty = \max_i \left| \sum_{j=1}^{n} a_{ij} x_j \right|$

$$= \sum_{j=1}^{n} |a_{kj}|$$

$$\|Ax_0\|_\infty = \max_i \left| \sum_{j=1}^{n} a_{ij} x_j \right|$$

$$= \sum_{j=1}^{n} |a_{kj}|$$

Condition Number :- The Number $K(A) = \|A\|\|A^{-1}\|$ is called the condition number of the invertible matrix $A \in \mathbb{R}^{n \times n}$.

$$\|x\| = \|Ix\| \leq \|I\|\|x\|$$

i.e. $\|I\| \geq 1$.

Also, $1 \leq \|I\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\| = K(A)$

So, $K(A) \geq 1 \quad \forall A \in \mathbb{R}^{n \times n}$.

Perturbation Theory :-

Suppose $Ax = b$ & $(A + \delta A)\hat{x} = b + \delta b$.

Our goal is to bound the norm of $\delta x = \hat{x} - x$.

$$(A + \delta A)(x + \delta x) = b + \delta b$$

$$Ax = b$$

$$\delta A\, x + (A + \delta A)\delta x = \delta b$$

$$\Rightarrow \quad \delta x = A^{-1}(\delta b - \delta A \cdot x)$$

Taking norms,

$$\|\delta x\| \le \|A^{-1}\| \left( \|\delta A\| \|\hat{x}\| + \|\delta b\| \right)$$

(Used consistency of matrix norm)

$$\text{So, } \frac{\|\delta x\|}{\|\hat{x}\|} \le \|A^{-1}\|\|A\| \left( \frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\|\|\hat{x}\|} \right)$$

$$= K(A) \left( \frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\|\|\hat{x}\|} \right)$$

The relative change $\frac{\|\delta x\|}{\|\hat{x}\|}$ in the answer is depends on the relative change in the data.

---

**Lemma**: If $\|X\| < 1$, then $I - X$ is invertible, $(I-X)^{-1} = \sum_{i=0}^{\infty} X^i$

and $\|(I-X)^{-1}\| \le \frac{1}{1-\|X\|}$.

**Proof**: Let, $(I-X)$ is singular. Then, $\exists\ x_0 \neq 0$

s.t. $(I-X)x_0 = 0$

$$0 \neq \|x_0\| = \|Ix_0 - (I-X)x_0\| = \|Xx_0\| \le \|X\|\|x_0\| < \|x_0\|$$

which is absurd.

Consider $S_n = \sum_{i=0}^{n} X^i$

$$(I-X)S_n = I - X^{n+1} \to I \quad \text{as } n \to \infty$$

as $\|X^i\| \le \|X\|^i \to 0$ as $i \to \infty$

Therefore, $(I-X)S = I$.

$\therefore S = (I-X)^{-1}$

i.e. $\sum_{i=0}^{\infty} X^i = (I-X)^{-1}$

Also, $\|(I-X)^{-1}\| = \|\sum_{i=0}^{\infty} X^i\| \le \sum_{i=0}^{\infty} \|X\|^i = \frac{1}{1-\|X\|}$.

Perturbed equation.

$$(A+\delta A)(x+\delta x) = b+\delta b$$

$$\Rightarrow \quad \delta A\, x + (A+\delta A)\delta x = \delta b$$

$$\therefore \quad \delta x = (A+\delta A)^{-1}(-\delta A x + \delta b)$$

$$= \left[A(I+A^{-1}\delta A)\right]^{-1}(-\delta A x + \delta b)$$

$$= (I+A^{-1}\delta A)^{-1} A^{-1}(-\delta A x + \delta b)$$

Note $\|A^{-1}\delta A\| \le \|A^{-1}\|\,\|\delta A\| < 1$ $\because$ $\delta A$ is small enough.

So, $$\frac{\|\delta x\|}{\|x\|} \le \|(I + A^{-1}\delta A)^{-1}\|\,\|A^{-1}\|\left(\|\delta A\| + \frac{\|\delta b\|}{\|x\|}\right)$$

$$\le \frac{\|A^{-1}\|}{1-\|A^{-1}\|\,\|\delta A\|}\left(\|\delta A\| + \frac{\|\delta b\|}{\|x\|}\right)$$

$$= \frac{\|A^{-1}\|\|A\|}{1-\|A^{-1}\|\|A\|\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\|\,\|x\|}\right)$$

$$\le \frac{K(A)}{1-K(A)\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right)$$

$\because$ $\|b\| = \|Ax\| \le \|A\|\,\|x\|$

$$\frac{K(A)}{1-K(A)\frac{\|\delta A\|}{\|A\|}} \simeq K(A) \text{ if } \|\delta A\| \text{ is small.}$$

i.e. relative input error is amplified by $K(A)$ in the relative output error.

Exm

$$\begin{pmatrix} 8 & 6 & 4 & 1 \\ 1 & 4 & 5 & 1 \\ 8 & 4 & 1 & 1 \\ 1 & 4 & 3 & 6 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 19 \\ 11 \\ 14 \\ 14 \end{pmatrix} \Rightarrow x = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

If we slightly modify the right hand side:

$$A\hat{x} = \begin{pmatrix} 19.01 \\ 11.05 \\ 14.07 \\ 14.05 \end{pmatrix}, \quad \hat{x} = \begin{pmatrix} -2.34 \\ 9.745 \\ -4.85 \\ -1.34 \end{pmatrix}$$

$\Rightarrow$ Stability Issue:

$\text{Cond}_2(A) \simeq 5367$, this amplies the relative error.

## Floating Point Arithmatic:-

**Theorem** Let, V & W be normed spaces. If V is finite dimensional, then all linear transformations from V to W are bounded.

or. $S = \{ \|Av\| : \|v\|_V = 1 \}$ is bounded.

**Proof:** $v_1, v_2, \ldots, v_n$ be a basis of V.

Then, for any $v = \sum_{j=1}^{n} \alpha_j v_j$

$$\therefore \|Av\|_W = \|\sum \alpha_j Av_j\|_W$$

$$\leq \sum_j |\alpha_j| \|Av_j\|_W$$

$$\leq \left(\max_j |\alpha_j|\right) \sum_{j=1}^{n} \|Av_j\|_W$$

Define $\|v\|_N = \max_j |\alpha_j|$, this is a norm in V.

Now as all norms in V are equivalent,

$\exists\, c$ s.t.

$$\|v\|_N \leq c \|v\|_V$$

So, we get

$$\|Av\|_W \leq C_1 \|v\|_V$$

(where, $C_1 = C \sum_{j=1}^{n} \|Av_j\|_W$, finite quantity)

---

Floating Point Arithmatic :-

Since computers use a finite no of bits to represent a real number, they can represent only a finite subset of $\mathbb{R}$. So, there are two difficulties:

1) the ~~right~~ represented no. can not be arbitrarily large or small.

2) there must be gaps between them.

(1) is not an issue with modern computers. IEEE double precision system permits numbers upto $1.79 \times 10^{308}$ and $2.23 \times 10^{-308}$.

(2) is a serious concern for scientific computing.

The interval $[1,2]$ is represented by

$$1,\ 1+2^{-52},\ 1+2\times 2^{-52},\ 1+3\times 2^{-52},\ \ldots,\ 2 \quad \ldots(*)$$

The interval $[2,4]$ is represented by same numbers multiplied by 2.

$$2,\ 2+2^{-51},\ 2+2\times 2^{-51},\ \ldots,\ 4$$

In general, $[2^j, 2^{j+1}]$ is represented by $(*) \times 2^j$.

The gaps between adjacent numbers are never larger than $2^{-52} \simeq 2.22 \times 10^{-16}$

In fixed point representation, the gaps are all of the same size.

## Fixed point Representation :-

The representation has fixed no of bits for integer part & fractional part.

Signed fixed pt | Sign | Integer | Fraction |

### Sign - Magnitude :

$-43.625$ is represented as follows:

~~fixed <32,16>~~ | 1 | 0000000000101011 | 1010000000000000 | Q<15,16>

Smallest # 0|00 -- 0|0 .... 1| $= 2^{-16} \simeq 0.000015$

Largest # 0|11 ... 1|1 ... 1| $= (2^{15}-1) + (1-2^{-16})$

$= 32768$

Gap between numbers $= 2^{-16}$.

### 1's complement form :

By complementing each bit in a signed binary integer, the 1's complement of a no. is derived.

$55_{10} = 0|0110111_2$ in 8 bit

$-55_{10} = 11001000_2$

### 2's complement form :

By adding 1 to the signed binary no.'s 1's complement, 2's complement of a no is derived.

$55_{10} = 0\cdot0110111_2$ ~~fixed <8,0>~~ UQ <8,0>

$-55_{10} = 11001001_2$

$2.5_{10} = 0101_2$

$-2.5_{10} = 1011_2$

~~fixed <8,3>~~ : $0|0010|110_2 = 2.75$ $Q<5,3>$

fixed <8,5> : $0|00|10110_2 = 0.6875$ $Q<3,5>$

Shifting: $UQ<6,3>$ ~~fixed <9,3>~~ without sign:

$11\,0101|000_2 = 53_{10}$

$\rightarrow$

$011010|100_2 = 26.5_{10}$

Disadvantage: Loss of range & precision.

~~fixed <8,1>~~ $Q<7,1>$, the fractional part is only precise to 0.5

We can~~to~~not represent 0.75

We can represent 0.75 in ~~fixed <8,2>~~ $Q<6,2>$, but, then we loose range on integral part.

$Q<a,b>$ : signed 2's complement fixed point number with a integer bits and b fractional bits.

Exn. Compute $0.75 + (-0.625)$ using $Q<4,4>$.

$0.75_{10} = 0000|1100_2$

$0.625_{10} = 0000|1010_2$

1st Complement: $1111|0101$

$+1$

$1111\,0110$ ← 2's complement

$$\begin{array}{r} 0.75 \\ -\ 0.625 \\ \hline 0.125 \end{array} \qquad \begin{array}{r} 0000.1100 \\ 1111.0110 \\ \hline 0000.0010 \end{array}$$

Fixed point representation is used in digital signal processing, machine learning as the computation is faster and consume less power.

Sign - Magnitude :

8-bit : Q<4,4>

$0000|0000 = 0_{10}$

$1000|0000 = -0_{10}$

1's complement : $0000|0000 = 0_{10}$

$1111|1111$

Complement is $000\ 0000 = 0$

$= -0_{10}$

2's complement :- $1000|0001 = -127_{10}$

1's complement : $111\ 1110$

$+\ 1$

$111\ 1111 = 127_{10}$

$0000|0000 = 0_{10}$

$1111|1111 = -ve$

1's complement : $000\ 0000$

$+\ 1$

$000\ 0001 = 1$

$= -1.$

So, the range is $-2^{+p}$ to $2^{p}-1$ for p bits

Problem:

$127_{10} = 0111\ 1111_2$

$2_{10} = 00000010_2$

$10000001_2 = -127_{10}$ (overflow)

# Floating point representations :-

**Defn :-** Let, $b \in \mathbb{N}$ & $b \geq 2$ Any real no can be represented exactly in base $b$ as:

$$(-1)^s \times (.d_1 d_2 \ldots d_n)_b \times b^e$$

where $d_i \in \{0, 1, \ldots b-1\}$ with $d_1 \neq 0$ or $d_1 = d_2 = \ldots = 0$, $s = 0$ or $1$ and an appropriate integer $e$ is called the exponent.

$(.d_1 d_2 \ldots d_n \text{---})_b = \sum_{i=1}^{n} d_i b^{-i}$ is called the mantissa. The exponent $e$ has range: $e_{min} \leq e \leq e_{max}$

This is called n-digit floating point representation.

**Exm.**

$$6.238 = (-1)^0 \times 0.6238 \times 10^1$$

$b = 10, \; s = 0.$

$$-0.0014 = (-1)^1 \times 0.14 \times 10^{-2}$$

$\Rightarrow$ If the exponent $e$ lies outside $[e_{min}, e_{max}]$ range, it is called underflow and overflow resp.

IEEE 754 Standard system:

Single precision: $b = 2$, $n = 23$, 8 bits for exponent (32 bits)

double precision: $b = 2$, $n = 52$, 11 bits for exponent (64 bits)

IEEE single precision format : $(-1)^s \, 2^{e-127} \, (1+f)$

IEEE double precision format : $(-1)^s \, 2^{e-1023} \, (1+f)$

## Machine Precision :-

This number is half the distance between 1 and the next larger floating point number.

For all $x \in R$, $\exists x' \in$ floating point numbers such that

$$\frac{|x - x'|}{|x|} \leq eps.$$

IEEE single precision: $eps = 2^{-24} = 5.96 \times 10^{-8}$

double " : $eps = 2^{-53} = 1.11 \times 10^{-16}$.

| Sign | Exponent | Mantissa |
|---|---|---|

Biased exponent: The exponent needs to represent both +ve & -ve numbers. A bias is added to the actual exponent in order to get the stored exponent.

Normalised Mantissa: A floating point number consists of its significant digits.

A normalized mantissa is one with 1 to the left of the decimal.

Exm 85.125

$85 = 1010101$

$0.125 = 001$

So, $85.125 = 1010101.001$

$= 1.01010101 \times 2^6$

Sign = 0

Single preision:

biased exponent = $6 + 127 = 133 = 10000101$

| | |
|---|---|
| 2 \| 85 | |
| 2 \| 42 | 1 |
| 2 \| 21 | 0 |
| 2 \| 10 | 1 |
| 2 \| 5 | 0 |
| 2 \| 2 | 1 |
| 2 \| 1 | 0 |
| 0 | 1 |

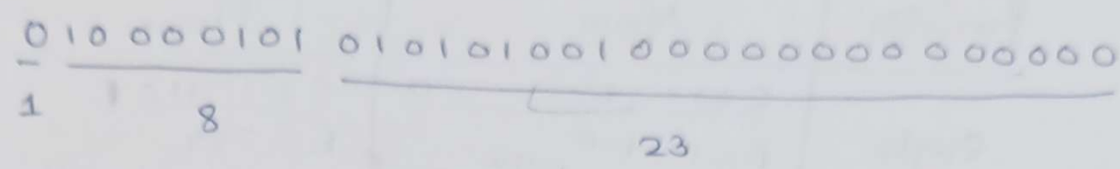
$0.125 \times 2 = 0.25 \rightarrow 0$

$0.25 \times 2 = 0.5 \rightarrow 0$

$0.5 \times 2 = 1.0 \rightarrow 1$

Normalized mantissa : 01 0101001

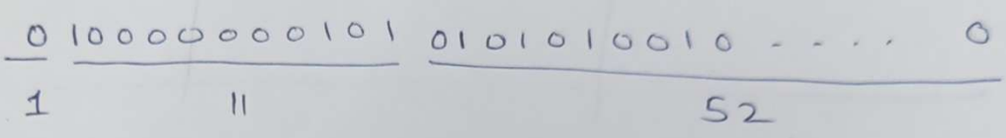
IEEE single precision:




Double precision:

biased exponent = 6 + 1023 = 1029 = 10000000101

In double precision :




Special cases

| exponent | mantissa | value |
|---|---|---|
| 0 | 0 | 0 |
| 255 | 0 | infinity |
| 0 | not 0 | denormalized |
| 255 | not 0 | NAN. |

Exm. IEEE-754 32-bit floating point representation pattern is 1 01111110 10000000000000000000000

S = 1 -ve

E = 01111110 = 126

F = 1.1 = $1 + 2^{-1}$ = 1.5

The number is = $-1.5 \times 2^{126-127}$ = −0.75.

## Minimum & Maximum numbers

| Precision | Nmin | Nmax |
|---|---|---|
| Single | 0 00 … 1 0 … 0<br>E = 1 , F = 0<br>$N_{min} = 1 \times 2^{-126} = 1.17 \times 10^{-38}$ | 0 1111110 …<br>E = 254<br>$N_{max} = 1.1 \ldots 1 \times 2^{127} = (2 - 2^{-23}) \times 2^{127} = 3.$ |
| Double | 0 0 … 1 0 … 0<br>$N_{min} = 1 \times 2^{-1022} = 2.22 \times 10^{-308}$ | 0 1 … 0 1<br>$N_{max} = 1.1 \ldots 1 \times 2^{1023} = (2 - 2^{-52}) \times 2^{1023} = 1.798 \times 10^{308}$ |

$$N = (-1)^{sign} \times (1 + significand) \times 2^{exponent - bias}$$

$-0.75$ : $S = 1$.

$0.75 \times 2 = 1.50$

$0.50 \times 2 = 1.0$ ↓ 1

$0.11 = 1.1 \times 2^{-1} = 1.1 \times 2^{126-127}$

1 01111110 100 … 0

$+20 = 10100 \times 2^0 = 1.0100 \times 2^4$

exponent $= 4 + 1023 = 1027 = 1024 + 3$

$= 10000000011$

0 10000000011 0100 … 0