Lecture-4

Numerical Representation in Computers.

Int → Int8, Int16, Int64, Int32, Uint8.

Float → Float16, Float32, Float64, BigFloat

Int64 → -2^{63} to $2^{63}-1$

Float 64

IEEE 754

1 + 11 + 52

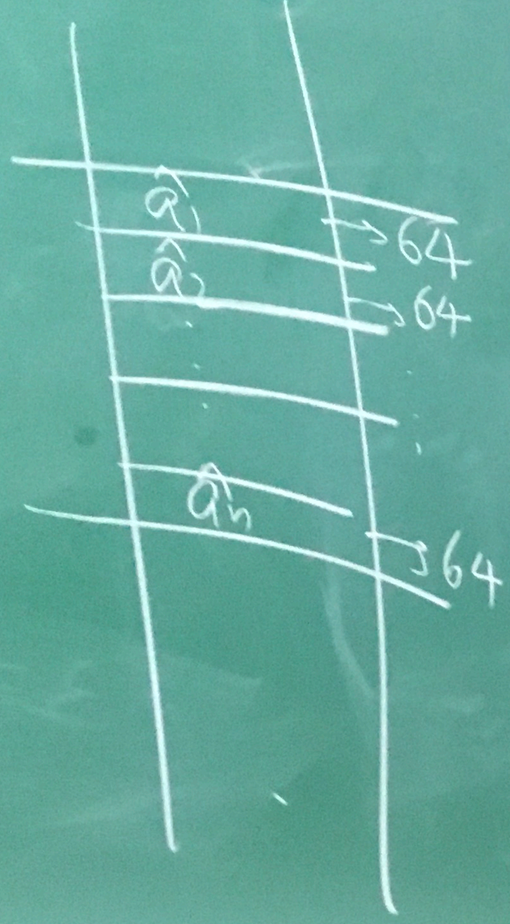
Sign (s) → 0 → +ve, 1 → -ve

Exponent (e) → an integer → -1023 to 1023 / 1022

mantissa: $b_0 b_1 \ldots b_{51}$ → $1.b_0 b_1 \ldots b_{51}$

pt
ewtation

version of a).



nt

$$1.b_0 b_1 \ldots b_{51}$$

$$m = 1 + \sum_{k=0}^{51} b_k \, 2^{-k}$$

$$S \times 2^{e} \times m$$

Flow

Sign
(S)

0 -> +ve
1 -> -ve

Real

$a \longrightarrow$ Floating Pt representation $\hat{a}$

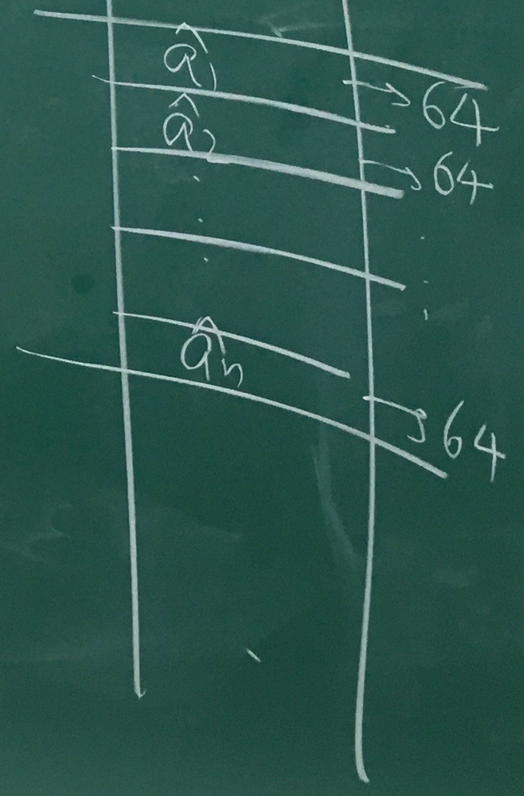
(quantized version of a).

Vector $\underline{a} \in \mathbb{R}^n$

8n bytes

$\leftarrow \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \rightarrow \begin{bmatrix} \hat{a}_1 \\ \vdots \\ \hat{a}_n \end{bmatrix}$

64n bits



1.b

$m = 1 -$

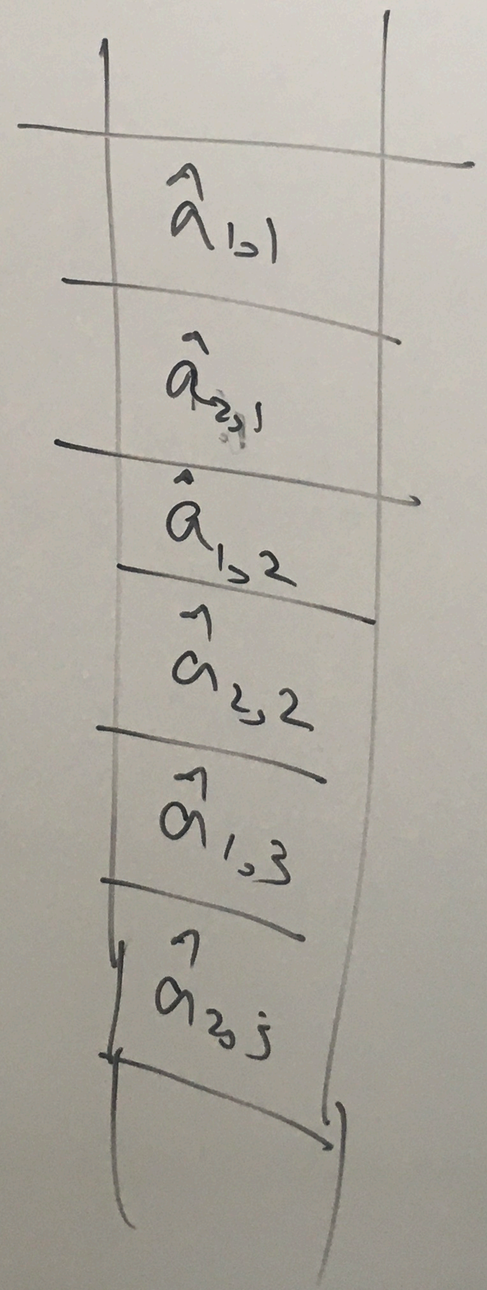
$S \times 2^e \times$

Representing a real matrix (m×n) in Float64

We use 64mn bits or 8mn bytes of memory.

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \end{bmatrix}$$

2×3

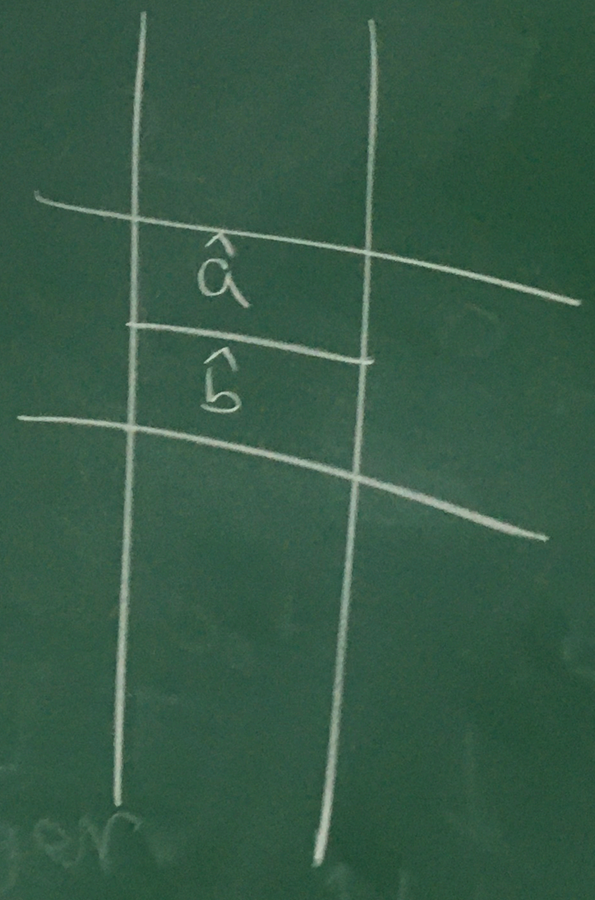


A[1,2] or

A[3]

Representation of complex numbers using Float64.

$C = a + ib$



format"

]

Index	float64 Value.
	$\sqrt{2}$
	-11

Sparse vectors & matrices

A vector or a matrix is said to be sparse if the fraction of non-zero entries is small.

$nnz(A)$ = number of non-zero elements of A.

$A_{m \times n} \rightarrow \frac{nnz(A)}{mn}$

Represen

numbers

C = a

COO format

"Coordinate list format"

Ex $A = \begin{bmatrix} 0 & 0 & \sqrt{2} \\ -11 & 0 & 0 \end{bmatrix}$

UInt	UInt	Float64
Row-Index	Column Index	Value.
1	3	$\sqrt{2}$
2	1	-11

Sp

A v

to be

of n

nn

A m

Floating point operation (flop)

Example: $\alpha \in \mathbb{R}$, $A \in \mathbb{R}^{m \times n}$

$B = \alpha A$. → complexity is mn.

$$= \begin{bmatrix} \alpha a_{1,1} & \cdots \\ \vdots & \ddots \\ \cdots & \alpha a_{m,n} \end{bmatrix}$$

$\min\{\text{nnz}(x), \text{nnz}(y)\}$
flops.

=

ordinates i

$\neq 0$ & $y_i \neq 0$

② $x, y \in \mathbb{R}^n$

$z = x + y$

$\rightarrow n$ flops.

③ $a, b \in \mathbb{C}$ $\Big\}$ $\rightarrow$ 6 flops.

$c = ab$

$= (Re(a)Re(b) - Im(a)Im(b))$

$+ i(Re(a)Im(b) + Im(a)Re(b))$

$\mathbb{R}^n$

rse.

$\underline{y}$ is sparse)

$nnz(\underline{y})$ flops

④ Adding two sparse vectors

$\underline{x}, \underline{y} \in \mathbb{R}^n$.

$\underline{z} = \underline{x} + \underline{y}$ $\Big\}$ $\leq \min\{nnz(\underline{x}), nnz(\underline{y})\}$ flops.

Number additions = number of co-ordinates i such that $x_i \neq 0$ & $y_i \neq 0$

②

$\underline{z}$ =

③

a, b

c =

=(

+ i(R

(5) Say $\underline{x}, \underline{y} \in \mathbb{R}^n$

and $\underline{y}$ is sparse.

($\underline{x}$ is "dense", $\underline{y}$ is sparse)

$\underline{z} = \underline{x} + \underline{y}$

Complexity: $nnz(\underline{y})$ flops

(4)

$\underline{z} =$

Number

nun

Such

Multiplication of Matrices & Vectors.

(1) $\underline{a}, \underline{b} \in \mathbb{R}^n$

$c \in \mathbb{R}$, $c = \underline{a}^T \underline{b} = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = \underline{b}^T \underline{a}$

$= a_1 b_1 + \cdots + a_n b_n$

$= \sum_{i=1}^{n} a_i b_i$

Complexity = $2n-1$ flops $\approx 2n$ flops

$a_1 b_n$

$a_m b_n$

$b_n a$

Complexity =

mn

flops

(2) Outer Product

$\underline{a} \in \mathbb{R}^m , \underline{b} \in \mathbb{R}^n$

$C = \underline{a}\,\underline{b}^T$

C: $m \times n$; $\underline{a}$: $m \times 1$; $\underline{b}^T$: $1 \times n$

$= \begin{bmatrix} a_1 \\ a_m \end{bmatrix} \begin{bmatrix} b_1 & b_2 & \cdots & b_n \end{bmatrix}$

(1)

R^n

rse.

y is sparse)

$nnz(y)$ flops

$$C = \begin{bmatrix} a_1 b_1 & a_1 b_2 & \cdots & a_1 b_n \\ \vdots & & & \\ a_m b_1 & a_m b_2 & \cdots & a_m b_n \end{bmatrix}$$

$$= \begin{bmatrix} b_1 \underline{a} & b_2 \underline{a} & \cdots & b_n \underline{a} \end{bmatrix}$$

$$= \begin{bmatrix} a_1 \underline{b}^T \\ a_2 \underline{b}^T \\ \vdots \\ a_m \underline{b}^T \end{bmatrix}$$

Complexity = mn flops.

(2) $\underline{a} \in \mathbb{R}^n$

C = ↓ $m \times n$

$m \times$

=