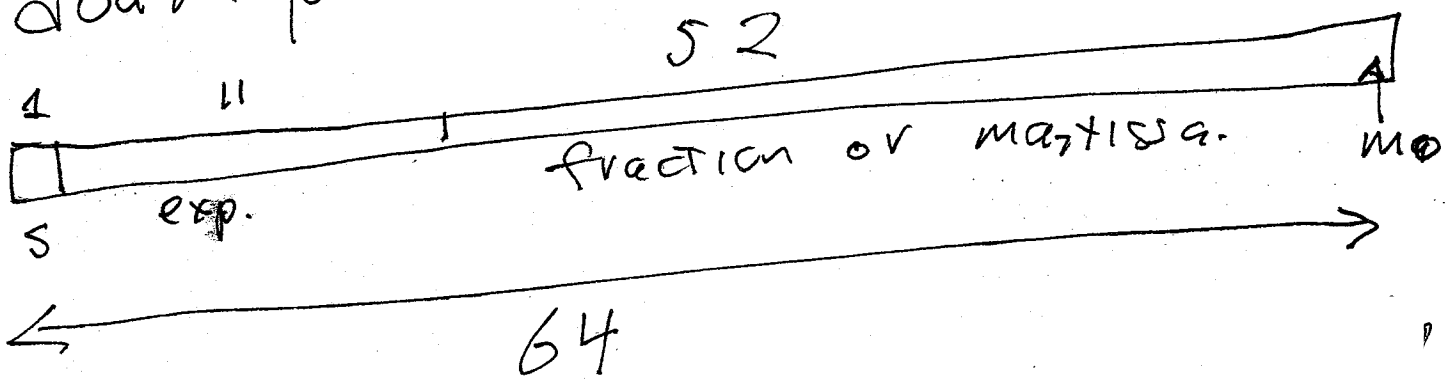
IEEE 754 = Binary 64

double precision floating point representation

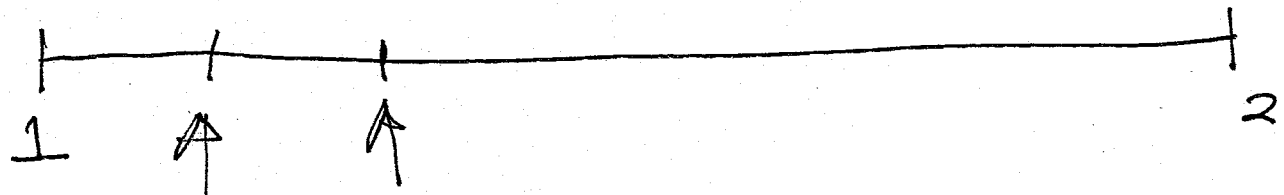


$$X = (-1)^s (1.m_{51} \ldots m_0)_2 \times 2^{e-1023}$$

$$0 \leq e \leq 2047$$

$$X_{MAX} \approx 1.79 \times 10^{308}$$

$$0 < X_{min} \approx 2.23 \times 10^{-309}$$

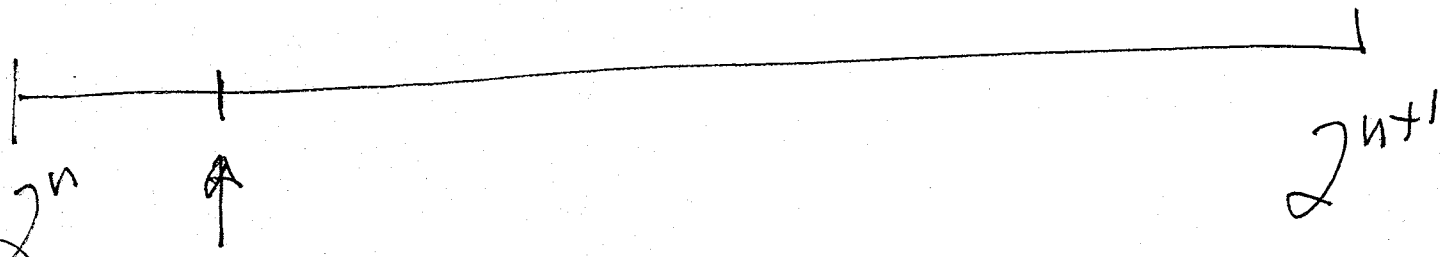


$1+2^{-52}$ $1+2\cdot 2^{-52}$ . .

$\longleftrightarrow$ 2^{-52}

next floating point number

base 10 - gap is $2^{-52} \approx 2.22 \times 10^{-16}$



$1+2^{n-52}$.

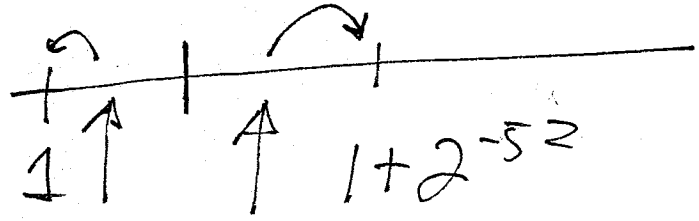
$2^{52} \longrightarrow 1+2$

2^{53}

Rel size of the gaps is 2^{-52}

$\mathbb{F}$ = floating point numbers – finite set

Given $x \in \mathbb{R}$, $fl(x)$ is the nearest floating point number via rounding



Basic Axiom – for all $x \in \mathbb{R}$, $fl(x)$ satisfies

$$\frac{|x - fl(x)|}{|x|} \leq \underbrace{\frac{1}{2}2^{-52}}_{\text{machine epsilon} = \varepsilon_{mac}} \approx 1.11 \times 10^{-16}$$

WARNING: IBM Summit

200 PLOPS = 200 peta floating point operations / second

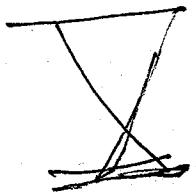
$200 \times 10^{+15}$ / 2×10^{17}

each flop can yield 1×10^{-16} error

1 second $(2 \times 10^{17})(1 \times 10^{-16})$

= 20 error

Stability



inputs

$\xrightarrow{f}$ y

soln.

$(A,b) \mapsto x$ w/ $Ax=b$

F ⇐ into floats.

specific algorithm.

$F(X) \xrightarrow{F(f)} F(y)$

Stability = fairly accurate translation

Least squares

$m \times n$

$m > n$

solve $A\vec{x} = \vec{b}$

$$A_{11} x_1 + A_{12} x_2 + \dots + A_{1n} x_n = b_1$$

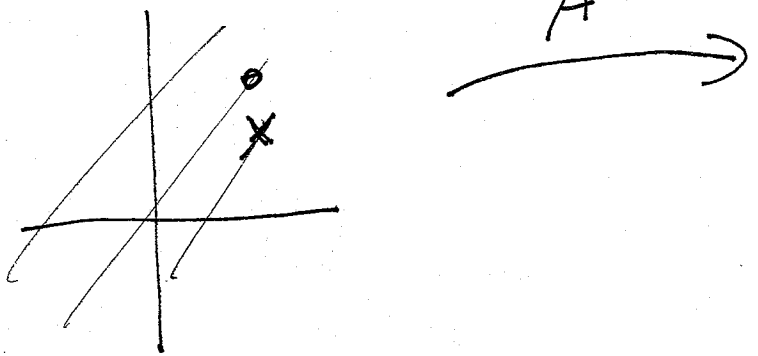
$$\vdots$$

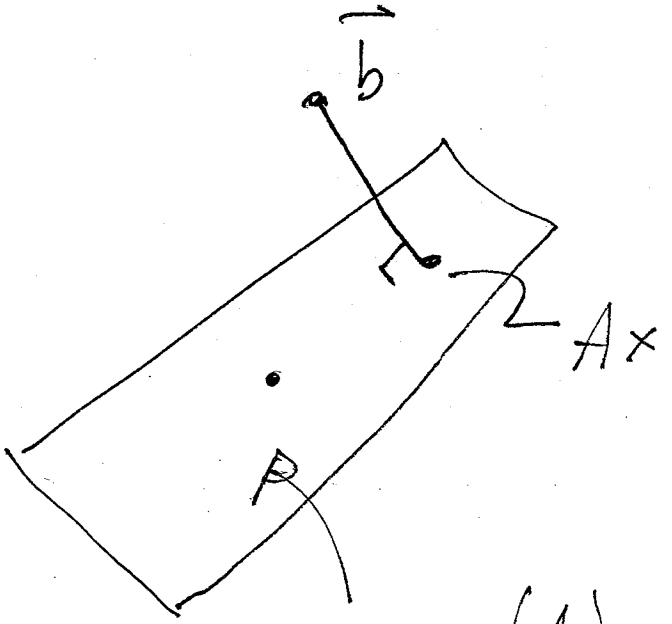
$$A_{m1} x_1 + A_{m2} x_2 + \dots + A_{mn} x_n = b_m.$$

more equations than unknowns.

As a map $A: \mathbb{R}^n \to \mathbb{R}^m$ $m \geq n$.

$n=2$ $m=3$.





range(A) = col(A)

(1) Find x which minimizes $\|A\vec{x} - \vec{b}\|_2$

(2) Ax will be the perpendicular projection of $\vec{b}$ onto col(A) so $Ax = P\vec{b}$

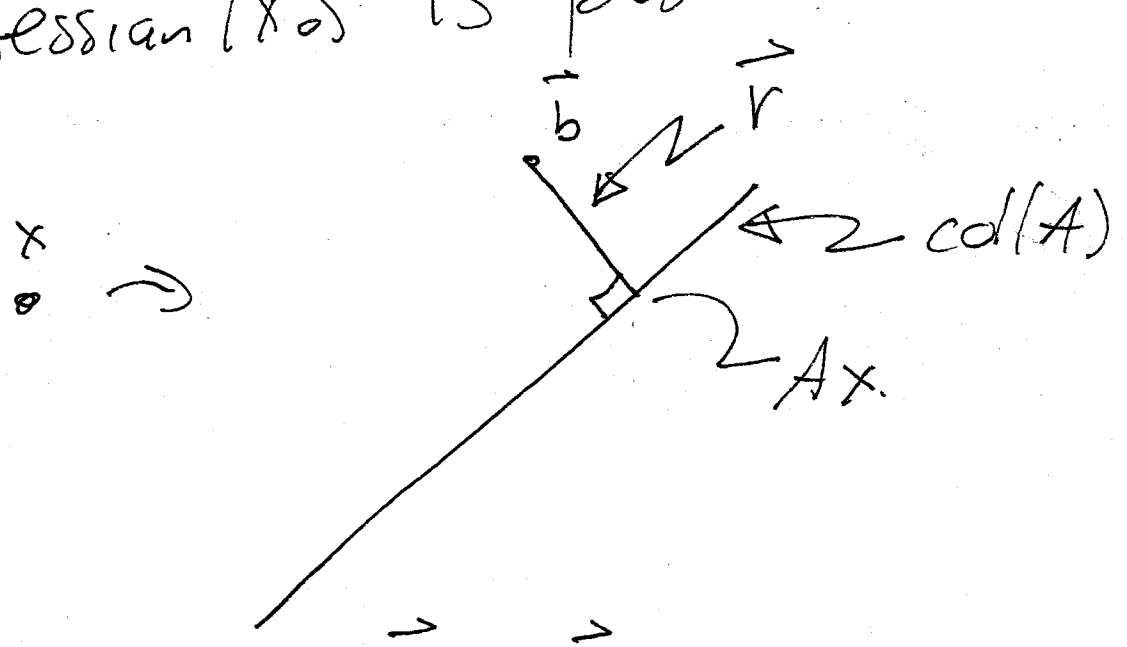
For (1) Let $\Phi(x) = \|Ax-b\|_2^2 \Rightarrow$

minimize Φ via calculus ie. Find x_0

with $\nabla\Phi(x_0) = 0$. turns out that

at x_0 the Hessian(x_0) is pos def so min.

(2)



$$\text{residual} = \vec{r} = \vec{b} - A\vec{x}$$

For minimum, $\vec{r} \perp \text{col}(A)$

so if $A = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}$

$\vec{r} \perp \text{col}(A)$

$\vec{r} \perp a_j$ all j

$\vec{a}_j^T \vec{r} = 0$ all j

$$A^T \vec{r} = \begin{bmatrix} \vec{a}_1^T \\ a_2^T \\ \vdots \\ a_n^T \end{bmatrix} \vec{r} = \begin{bmatrix} a_1^T \vec{r} \\ \vdots \\ a_n^T \vec{r} \end{bmatrix} = 0$$

So minimum of $\vec{r}$ is characterized by $A^T \vec{r} = \vec{0}$

$$A^T(\vec{b} - A\hat{x}) = 0$$

$$\boxed{A^T \vec{b} = A^T A x}$$ normal equations

$$\boxed{A^T A \vec{x} = A^T \vec{b}}$$

first try, compute $(A^T A)^{-1}$

$\Rightarrow \vec{x} = (A^T A)^{-1} A^T \vec{b}$ unique soln.

Theorem: $m \geq n$, A^TA is invertible if and only if A has full rank i.e. rank $= n$
i.e. lin ind col.

We prove the contrapositive.

A^TA is singular $\iff$ A is rank deficient

Assume A^TA sing $\Rightarrow$ there is a $\vec{v} \neq 0$ s.t.

$(A^TA)\vec{v} = \vec{0} \Rightarrow v^TA^TA\vec{v} = \vec{0}$

So $(A\vec{v})^T(A\vec{v}) = 0$ or $\|Av\|_2^2 = 0$

$\Rightarrow A\vec{v} = 0$

so

A is rank def since $\vec{v} \neq 0$.

Converse $\text{rank}(A) < n \Rightarrow$ There is a $\vec{v} \neq 0$ with $A\vec{v} = 0 \Rightarrow$

$A^T A \vec{v} = 0 \Rightarrow A^T A$ is singular since $\vec{v} \neq 0$.

If $m \geq n$ and $\text{rank}(A) = n$

$\Rightarrow A^+ = (A^T A)^{-1} A^T$ ≙ pseudo inverse

$\Rightarrow$ soln to the normal eq. is

$\vec{x} = A^+ \vec{b}$.

Lemma If $m = n$ and A is invertible

then $A^+ = A^{-1}$

Pf

$$\left((A^T A)^{-1} A^T\right) A$$

$$= A^{-1} (A^T)^{-1} A^T A$$

$$= A^{-1} A = I.$$