FLOATING POINT NUMBERS AND ARITHMATIC

A COMPUTER MUST USE A DISCRETE REPRESENSTATION OF $\mathbb{R}$.

(a) THERE MUST BE A LARGEST AND SMALLEST POSITIVE NUMBER.

ON DOUBLE PRECISION MACHINE

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

$$N_{MIN} \approx 2.23 \times 10^{-308}$$

THIS IS TYPICALLY NOT THE ISSUE.

(b) GAPS BETWEEN ADJACENT NUMBERS.

ON DOUBLE PRECISION MACHINE.

$[1, 2]$

$$1, \; 1 + 2^{-52}, \; 1 + 2 \times 2^{-52}, \; \ldots, \; 2$$

NEXT INTERVAL

$[2, 4]$

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

IN GENERAL, THE INTERVAL $[2^j, 2^{j+1}]$ IS REPRESENTED AS 2^j TIMES THE NUMBERS REPRESENTING THE INTERVAL $[1, 2]$.

IN THE FLOATING POINT REPRESENTATION THE GAPS BETWEEN ~~SUCCESSIVE~~ ADJACENT NUMBERS SCALE WITH THEIR SIZE.

CALL SET OF FLOATING POINT NUMBERS $\mathbb{F} \subset \mathbb{R}$.

CALL ϵ ("MACHINE EPSILON") IS THE RESOLUTION OF THE FLOATING POINT NUMBERS. AND IS HALF THE DISTANCE ~~BET~~ BETWEEN 1 AND THE ~~N~~ ADJACENT NUMBER.

FOR DOUBLE PRECISION

$$\epsilon = 2^{-53} \approx 1.11 \times 10^{-16}$$

~~WE KNOW~~ AS A RESULT

$$\forall x \in \mathbb{R} \quad \exists \quad x' \in \mathbb{F}$$

$$\text{S.T.} \quad |x - x'| \leq \epsilon |x|$$

LET $fl : \mathbb{R} \to \mathbb{F}$ BE THE FUNCTION THAT ROUNDS $x \in \mathbb{R}$ TO THE NEAREST FLOATING POINT.

<u>FLOATING POINT AXIOM</u> I (FPA I)

$$\forall x \in \mathbb{R} \quad \exists \quad \epsilon' \text{ WITH}$$

$$|\epsilon'| \leq \epsilon \quad \text{S.T.} \quad fl(x) = x(1 + \epsilon')$$

FLOATING POINT ARITHMATIC

$+, -, \times, \div$ ON $\mathbb{R}$

HAVE ANALOGUES

$\oplus, \ominus, \otimes, \oslash$ ON $\mathbb{F}$

CONSTRUCTED S.T.

$$x \circledast y = fl(x * y)$$

FOR $x, y \in \mathbb{F}$

WHERE $*$ IS $+, -, \times,$ OR $\div$.

FUNDAMENTAL AXIOM OF FLOATING POINT ARITHMATIC (FPA II)

$\forall\, x, y \in \mathbb{F}$ $\exists\, \epsilon'$ WITH

$|\epsilon'| \leq \epsilon$ S.T.

$$x \circledast y = (x * y)(1 + \epsilon')$$

STABILITY

- STABILITY ~~PER~~ PERTAINS TO THE PERTURBATION BEHAVIOUR OF THE ALGORITHM USED TO SOLVE THE PROBLEM ON A COMPUTER.
- ALGORITHM: $\tilde{f} : X \to Y$ BETWEEN SAME SPACES AS THE PROBLEM.

FIX: (i) PROBLEM f

(ii) FLOATING PT. COMPUTER

(iii) AN ALGORITHM FOR f

(iv) IMPLEMENTATION OF THE ALGORITHM.

$x \in X$ IS ROUNDED $x' = fl(x)$

THEN SUPPLIED TO THE

PROGRAM.

THE PROGRAM IS RUN

AND THE ~~ARE~~ RESULT IS

$\tilde{f}(x) \in Y$.

DESPITE THE COMPLEXITY,

WE CAN MAKE CLEAN

STATEMENTS ABOUT $\tilde{f}(x)$

USING FPA I & II.

<u>ACCURACY</u>

- ABSOLUTE ERROR

$$\| \tilde{f}(x) - f(x) \|$$

- RELATIVE ERROR

$$\frac{\| \tilde{f}(x) - f(x) \|}{\| f(x) \|}$$

AN ALGORITHM IS ACCURATE

IF FOR EACH $x \in X$

$$\frac{\| \tilde{f}(x) - f(x) \|}{\| f(x) \|} = O(\epsilon)$$

ERROR IS ON THE ORDER

OF MACHINE EPSILON.

MORE PRECISELY, $\exists$ CONSTANT C

S.T. $\forall \; x \in X$

$$\frac{\| \tilde{f}(x) - f(x) \|}{\| f(x) \|} \leq C \epsilon$$

AS $\epsilon \to 0$.

STABILITY

AN ALGORITHM $\tilde{f}$ FOR PROBLEM f IS STABLE IF FOR EACH $x \in X$

$$\frac{\| \tilde{f}(x) - f(\tilde{x}) \|}{\| f(\tilde{x}) \|} = O(\epsilon)$$

FOR SOME $\tilde{x} \in X$

$$\frac{\| \tilde{x} - x \|}{\| x \|} = O(\epsilon)$$

A STABLE ALGORITHM GIVES NEARLY THE RIGHT ANSWER TO NEARLY THE RIGHT QUESTION.

BACKWARD STABILITY

THE ALGORITHM IS BACKWARD STABLE IF FOR EACH $x \in X$

$$\tilde{f}(x) = f(\tilde{x})$$

FOR SOME $\tilde{x} \in X$ WITH

$$\frac{\| \tilde{x} - x \|}{\| x \|} = O(\epsilon)$$

A BACKWARD STABLE ALGORITHM GIVES EXACTLY THE RIGHT ANSWER TO NEARLY THE RIGHT QUESTION.