Numerical Analysis

Daniel Mao

Contents

Chapter 1

Error

1.1 Error

Definition (Absolute Error)**.** *Let x be a result. Let $\tilde{x}$ be an approximation of the result x. We define the* ***absolute error****, denoted by Δx. to be a number given by*

$$\Delta x = |x - \tilde{x}|.$$

Definition (Relative Error)**.** *Let x be a result. Let $\tilde{x}$ be an approximation of the result x. We define the* ***relative error****, denoted by δx. to be a number given by*

$$\delta x = \left| \frac{x - \tilde{x}}{x} \right|,$$

assuming $x \neq 0$.

1.2 Condition of a Mathematical Problem

Definition (Absolute Condition Number)**.** *Let P be a problem with input $\vec{x}$ and output $\vec{z}$. We define the* ***absolute condition number*** *of P, denoted by κ_A, to be a number given by*

$$\kappa_A := \frac{\|\Delta \vec{z}\|}{\|\Delta \vec{x}\|}.$$

Definition (Relative Condition Number)**.** *Let P be a problem with input $\vec{x}$ and output $\vec{z}$. We define the* ***relative condition number*** *of P, denoted by κ_R, to*

be a number given by

$$\kappa_R := \frac{|\delta z|}{|\delta x|} = \frac{\|\Delta \vec{z}\|/\|\vec{z}\|}{\|\Delta \vec{x}\|/\|\vec{x}\|}.$$

Definition (Well-Conditioned)**.** *Let P be a problem with input $\vec{x}$ and output $\vec{z}$. We say that P is **well-conditioned** if small changes in the input $\vec{x}$ result in small changes in the output $\vec{z}$. i.e., if κ_A and κ_R are small.*

Definition (Ill-Conditioned)**.** *Let P be a problem with input $\vec{x}$ and output $\vec{z}$. We say that P is **ill-conditioned** if small changes in the input $\vec{x}$ result in large changes in the output $\vec{z}$. i.e., if κ_A and κ_R are large.*

Chapter 2

Floating Point Arithmetic

Definition (Floating Point Representation)**.**

exponent
mantissa

Definition (Rounding)**.** *Given a number x, its binary representation $(1_{a_1a_2...a_na_{n+1}})_2$, and n spaces mantissa;*

(1) Chopping: $fl(x) = 1.a_1a_2...a_n \times 2^e$

(2) Rounding: $fl(x) = \begin{cases} 1.a_1a_2...a_n \times 2^e, & \text{if } a_{n+1} = 0 \\ 1.a_1a_2...(a_n+1) \times 2^e, & \text{otherwise.} \end{cases}$

Definition (Machine Epsilon)**.** *We define the* ***machine epsilon*** ε_{mach} *to be the distance between 1 and the next larger number.*

Proposition 2.0.1. *The machine epsilon is*

$$\varepsilon_{mach} = b^{1-m}$$

where m is the number of spaces available for the mantissa.

Definition (Unit Roundoff)**.** *We define the* ***unit roundoff****, denoted by μ, to be the smallest positive number such that $fl(1+\mu) > 1$.*

Proposition 2.0.2. *The unit roundoff is*

$$\mu = \begin{cases} \varepsilon_{mach}, & \text{if chopping is used} \\ \varepsilon_{mach}/2, & \text{if rounding is used.} \end{cases}$$

Theorem 1. *Let x be a number. For any floating point system,*

$$|\delta x| = \left| \frac{x - fl(x)}{x} \right| \leq \mu.$$

Proposition 2.0.3. *For any floating point system F under chopping,*

$$fl(x) = x(1 + \eta)$$

for some $\eta \in \mathbb{R}$ such that $|\eta| \leq \varepsilon_{mach}$.

Remark.
- *Chopping error is always positive.*
- *Rounding error can be positive or negative.*

2.1 Floating Point Operations

Definition (Floating Point Operations)**.** *We define*

(1) ***floating point addition****, denoted by $\oplus$, as*

$$\forall x, y \in \mathbb{R}, x \oplus y = fl(fl(x) + fl(y)),$$

(2) ***floating point subtraction****, denoted by $\ominus$, as*

$$\forall x, y \in \mathbb{R}, x \ominus y = fl(fl(x) - fl(y)),$$

(3) ***floating point multiplication****, denoted by $\otimes$, as*

$$\forall x, y \in \mathbb{R}, x \otimes y = fl(fl(x) \times fl(y)),$$

(4) ***floating point division****, denoted by $\oplus$, as*

$$\forall x, y \in \mathbb{R}, x \oplus y = fl(fl(x)/fl(y)).$$

Proposition 2.1.1. *For any floating point system F and any $a, b \in \mathbb{R}$,*

$$a \oplus b = (fl(a) + fl(b))(1 + \eta)$$

for some $\eta \in \mathbb{R}$ such that $|\eta| \leq \mu$, where μ is the unit roundoff; and

$$a \oplus b = (a(1 + \eta_1) + b(1 + \eta_2))(1 + \eta)$$

for some $\eta, \eta_1, \eta_2 \in \mathbb{R}$ such that $|\eta|, |\eta_1|, |\eta_2| \leq \mu$, where μ is the unit roundoff.

Remark. *Floating point addition is not necessarily associative. i.e.,* $\exists a, b, c \in \mathbb{R}, (a \oplus b) \oplus c \neq a \oplus (b \oplus c)$.

Remark. *Floating point multiplication is not necessarily distributive over floating point addition.*

Definition (Underflow)**.** *If the result of a computation is less than the smallest positive normal number in the floating point system, then it is rounded down to zero.*

Definition (Overflow)**.** *If the result of a computation is larger (smaller) than the largest positive (smallest negative) number in the floating point system, then Inf (-Inf) is returned as the result.*