Einleitung

> #Definiton 0.1:
> Numerische Mathematik ist die Kunst der fehlerbehafteten rechnerischen Lösung kontinuierlicher Probleme

- "Fehlerbehaftet" Kontrolle über die in der Rechnung entstehenden fehler hat
- "Kontinuierlich" in $\mathbb{R}$ und $\mathbb{C}$
- "Rechnerisch" Verfahren die auf die Benutzung eines Computers zugeschnitten sind

1. Gleitkommazahlen

Zahlendarstellung im Stellenwertsystem zur Basis $\beta \in \mathbb{N}_{\geq 2}$. Für jedes $x \in \mathbb{R}$ gibt es ein Vorzeichen $v \in \{-1, +1\}$ und eine Ziffernfolge $(z_k)_{k \in \mathbb{Z}_{\leq n}}$ mit $z_k \in \{0, \ldots, \beta - 1\}$, wobei $n \in \mathbb{N}$, und $x = v * \sum\limits_{k=-\infty}^{n} z_k \beta^k$. Nicht alle bis auf endliche viele Nachkommastellen gleich $\beta - 1$ sind, diese Darstellung nennt man Festkommadarstellung.

> #Definiton 1.1
> Sei $\beta \in \mathbb{N}_{\geq 2}$. Eine **Gleitkommazahl** zur Basis β mit **Mantissenlänge** l hat die Form m x β^e, mit Exponenten $e \in \mathbb{Z}$ und Mantisse $m \in \mathbb{R}$

Wenn $1 \leq |m| < \beta$, also die Vorkommastelle von 0 verschieden ist, heißt die Gleitkommazahl normalisiert.

Mantisse: Anzahl der gezeigten zahlen ohne Potenz $(1.4 * 10^4$ hat Mantissenlänge 2 (normalisiert))

Beispiele > 1 2

1.1 Technische Gleitkommazahlen

In Computern wird ein Bit für das Vorzeichen verwendet (hidden Bit)

IEEE - 754 - Binärformat
Seien $r, p \in \mathbb{N}^*$ und $B := 2^{r-1} - 1$ der Biaswert. Biased exponent $E := (e_r \ldots e_0)_2$ und $M := (m_p \ldots m_0)_2$

- $0 < E < 2^r - 1$: Die normalisierte Gleitkommazahl
  $(-1)^v * (1 + \frac{M}{2^p})$ x 2^{E-B} der Mantissenlänge p+1 wird dargestellt.
- E = 0: Die subnormale Gleitkommazahlen $(-1)^v * \frac{M}{2^p} * 2^{1-B}$ wird dargestellt. Ihr Mantissenlänge ist $\leq p$. (+0,-0)
- $E = 2^r - 1 : (-1)^v * \infty$ wird dargestellt, falls M = 0. Ist M > 0, so wird keine Zahl dargestellt

Offenbar gibt es prinzipiell für jedes $M \neq 0$ ein anderes NaN, doch Bedeutungsunterschiede zwischen den Nans sind im Standard zum Teil gefragt. NaNs sind die Lösung der Ungleichung $x \neq x$. NaNs werden für nicht-initialisierte Variablen sowie für die undefinierte oder nicht-reele arithmetische Ausdrücke verwenden.
Beispiele > 1 3

Posit-Formate

Eine Abfolge von $n \in \mathbb{N}_{\geq 5}$ Bits $b_{n-1}, \ldots, b_0$ wird im Posit-Format so interpretiert.

- Sonderfälle
  - alle Bits 0 → 0
  - wenn $b_{n-1} = 1$ und alle anderen Bits 0 sind wird NaR (not a Real)
- $v := b_{n-1}$ ist das Vorzeichenbit
- Es sei $i \in \{0, \ldots, n-2\}$ so, dass $b_{n-2} = b_{n-3} = \ldots = b_i$ und $b_{i-1} = 1 - b_i$. Die Bitfolge $b_{n-2}, \ldots, b_{i-1}$ heißt **Regime**
  - es sei $k := i + 1 - n < 0$ falls $b_{n-2} = 1$ und $k := n - 2 - i \geq 0$ sonst
- Es sei E := 0, falls $i \leq 1$. Es sei $E := b_0$, falls i = 2. Ansonsten sei $E := (b_{i-2} b_{i-3})_2$.
- Es sei F := 0, falls $i \leq 3$. Ansonstten sei $F := \frac{(b_{i-3} \ldots b_0)_2}{2^{i-3}}$. F → Nachkommateil

i: die Anzahl von Bits bis das Regime beginnt von rechts
Regime: bei v = 0 → bis das erste mal 0 von links erscheint, v→1 andersrum
k = wenn der b_{n-2} bit von rechts 1 ist → k = n-2-i sonst k = i+1-n
E = bits nach Regime (rechts davon)

Die repräsentierte Zahl ist $(1 - 3v + F) * 2^{(1-2v)*(4k+E+v)}$

Beispiel 1.5