

# NUM 3

1)  $b \in \mathbb{N}_{\geq 2}$  ... Basis  $t \in \mathbb{N}$ . Mantissenlänge  $e_{\min}, e_{\max} \in \mathbb{Z}$  mit  $e_{\min} < 0 \leq e_{\max}$   
 $\text{IF}(b, t, e_{\min}, e_{\max}) = \{0\} \cup \left\{ \left( \sum_{k=1}^t a_k b^{-k} \right) b^e : a_k \in \{0, \dots, b-1\}, e \in \{\pm 1\}, a_1 \neq 0, e \in \mathbb{Z}, e_{\min} \leq e \leq e_{\max} \right\}$  ... Exponentialschranken

a) zz: Darstellung aller  $x \in \text{IF}$  ist eindeutig

Da  $\left\{ \left( \sum_{k=1}^t a_k b^{-k} \right) b^e : a_k \in \{0, \dots, b-1\}, a_1 \neq 0, e \in \mathbb{Z}, e_{\min} \leq e \leq e_{\max} \right\}$  nur aus

Zahlen  $> 0$  besteht (da  $a_0 = 0$ ) ist für  $x = 0$  die Darstellung klar.

Falls  $x < 0$  muss  $e = -1$  sein, sonst  $e = +1$ , daher o.B.d.A.  $x > 0$

Wir zeigen  $\sum_{k=1}^t a_k b^{-k}$  für  $a_k \in \{0, \dots, b-1\}, a_1 \neq 0$  immer in  $[b^{-1}, 1-b^{-t}]$

liegt durch vollständige Induktion nach  $t$ :

$$t=1 \quad x = \sum_{k=1}^1 a_k b^{-k} = a_1 \cdot b^{-1} \quad \text{Da } a_1 \in \{1, \dots, b-1\} \text{ gilt } 1 \cdot b^{-1} \leq x \leq (b-1) \cdot b^{-1} = 1-b^{-1}$$

$$t+1 \quad x := \sum_{k=1}^{t+1} a_k b^{-k} = \underbrace{\sum_{k=1}^t a_k b^{-k}}_{\in [b^{-1}, 1-b^{-t}]} + \underbrace{a_{t+1} \cdot b^{-(t+1)}}_{\in [0, (b-1) \cdot b^{-t-1}]} \\ \Rightarrow x \in [b^{-1}, 1-b^{-t} + b^{-t-1} - b^{-(t+1)}] = 1-b^{-(t+1)}$$

Seien  $x_1, \dots, x_t, y_1, \dots, y_t \in \{0, \dots, b-1\}$ ,  $x_1 \neq 0, y_1 \neq 0, e_x \in \mathbb{Z}, e_{\min} \leq e_x \leq e_{\max}$ ,  $e_y \in \mathbb{Z}$ ,  $e_{\min} \leq e_y \leq e_{\max}$  zwei Darstellungen mit  $\left( \sum_{k=1}^t x_k b^{-k} \right) b^{e_x} = \left( \sum_{k=1}^t y_k b^{-k} \right) b^{e_y}$

Da  $\sum_{k=1}^t x_k b^{-k}, \sum_{k=1}^t y_k b^{-k} \in [b^{-1}, 1-b^{-t}]$  gilt

$$\left( \sum_{k=1}^t x_k b^{-k} \right) \cdot b^{e_x} \in [b^{e_x-1}, b^{e_x} - b^{e_x+t}] \subseteq [b^{e_x-1}, b^{e_x}] \quad \text{und}$$

$$\left( \sum_{k=1}^t y_k b^{-k} \right) \cdot b^{e_y} \in [b^{e_y-1}, b^{e_y} - b^{e_y+t}] \subseteq [b^{e_y-1}, b^{e_y}]$$

Für  $e_x \neq e_y$  gilt  $z \in [b^{e_x-1}, b^{e_x}]$  und  $z \in [b^{e_y-1}, b^{e_y}]$ , aber auch  $[b^{e_x-1}, b^{e_x}] \cap [b^{e_y-1}, b^{e_y}] = \emptyset$

$$\Rightarrow e_x = e_y =: e \quad \text{also ist nun zu zeigen } \sum_{k=1}^t x_k b^{-k} = \sum_{k=1}^t y_k b^{-k} \Rightarrow x_k = y_k \quad \forall k$$

Angenommen  $\exists j \in \{1, \dots, t\}$  mit  $x_j \neq y_j$  o.B.d.A.  $\forall k < j : x_k = y_k$  (also minimal)

$$\text{o.B.d.A. } x_j > y_j \Rightarrow \sum_{k=1}^t x_k b^{-k} = \sum_{k=1}^{j-1} x_k b^{-k} + x_j b^{-j} + \sum_{k=j+1}^t x_k b^{-k} = \sum_{k=1}^{j-1} y_k b^{-k} + x_j b^{-j} + \sum_{k=j+1}^t x_k b^{-k}$$

$$\dots = \sum_{k=1}^{j-1} y_k b^{-k} + y_j b^{-j} + (x_j - y_j) b^{-j} + \sum_{k=j+1}^t x_k b^{-k}$$

$$\sum_{k=1}^t y_k b^{-k} = \sum_{k=1}^{j-1} y_k b^{-k} + y_j b^{-j} + \sum_{k=j+1}^t x_k b^{-k} \Rightarrow \sum_{k=j+1}^t y_k b^{-k} = (x_j - y_j) b^{-j} + \sum_{k=j+1}^t x_k b^{-k}$$

$$\sum_{k=j+1}^t y_k b^{-k} \leq b^{-(j+2)} - b^{-(t+1)} < b^{-(j+2)} \quad \text{und } (x_j - y_j) b^{-j} + \sum_{k=j+1}^t x_k b^{-k} \geq b^{-j} \Rightarrow \text{Darstellung ist eindeutig}$$

## NUM 01

1) b) zz:  $x_{\min} = \min \{x \in F : x > 0\} = b^{e_{\min}-1}$

$$x_{\max} = \max \{x \in F : x > 0\} = b^{e_{\max}}(1 - b^{-t})$$

Von vorher wissen wir  $\sum_{k=1}^t a_k b^{-k} \in [b^{-1}, 1 - b^{-t}]$

$$\Rightarrow \left( \sum_{k=1}^t a_k b^{-k} \right) \cdot b^e \in [b^{e-1}, b^e(1 - b^{-t})]$$

Offensichtlich gilt also  $b^{e_{\min}-1} \leq x_{\min}$  und  $b^{e_{\max}}(1 - b^{-t}) \geq x_{\max}$

Wählen wir  $a_1 = 1, a_j = 0 \forall j \geq 1, e = e_{\min}$

$$\Rightarrow \left( \sum_{k=1}^t a_k b^{-k} \right) \cdot b^e = b^{-1} \cdot b^{e_{\min}} = b^{e_{\min}-1}$$

Wählen wir  $a_j = (b-1) \forall j \geq 1, e = e_{\max}$

$$\begin{aligned} \Rightarrow \left( \sum_{k=1}^t a_k b^{-k} \right) \cdot b^e &= \left( \sum_{k=1}^t (b-1) b^{-k} \right) \cdot b^{e_{\max}} = (b-1) \cdot b^{e_{\max}} \cdot \sum_{k=1}^t b^{-k} \\ &= b^{e_{\max}} \cdot (b-1) \cdot \frac{b^{-t}(b^{t-1}-1)}{(b-1)} = b^{e_{\max}} \cdot (1 - b^{-t}) \end{aligned}$$

c) ges: #F abhängig von  $b, t, e_{\min}, e_{\max}$

Da die Darstellung eindeutig ist Frage äquivalent zu:

Wie viele Arten gibt es  $a_k, e$  und  $b$  zu wählen?

$b$ ... zwei Möglichkeiten  $e \dots (e_{\max} - e_{\min} + 1)$  Möglichkeiten

$a_1 \dots b-1$  Möglichkeiten  $a_j, j \neq 1 \dots b$  Möglichkeiten

$$\Rightarrow \#F = 1 + 2 \cdot (e_{\max} - e_{\min} + 1) \cdot (b-1) \cdot b^{t-1}$$

$\uparrow \quad \uparrow \quad \uparrow \quad \uparrow \quad \uparrow$   
für 0    G    e     $a_1$      $a_j$

□

# NUM Ü1

2)  $F = F(b, t, e_{\min}, e_{\max})$

a)  $e \in \mathbb{Z}, e_{\min} \leq e \leq e_{\max} \quad M := b^t - b^{t-1} \in N$

$$\text{zz: } F \cap [b^{e-1}, b^e] = \{(b^{-1} + j \cdot b^{-t}) \cdot b^e : j = 0, \dots, M-1\}$$

Von 1) wissen wir  $\sum_{k=1}^t a_k b^{-k} \in [b^{-1}, 1-b^{-t}]$

$$\Rightarrow \left\{ \left( \sum_{k=1}^t a_k b^{-k} \right) \cdot b^e : a_k \in \{0, \dots, b-1\}, a_t \neq 0 \right\} \subseteq [b^{e-1}, b^e - b^{e-t}]$$

$$\text{Für alle } f \neq e \text{ gilt } \left\{ \left( \sum_{k=1}^t a_k b^{-k} \right) \cdot b^f : \dots \right\} \subseteq [b^{f-1}, b^f]$$

und somit nicht Elemente aus  $[b^{e-1}, b^e]$

$$\Rightarrow F \cap [b^{e-1}, b^e] = \left\{ \left( \sum_{k=1}^t a_k b^{-k} \right) \cdot b^e : \dots \right\}$$

$$\text{Nun ist nur mehr zu zeigen } \left\{ \sum_{k=1}^t a_k b^{-k} : \dots \right\} = \{b^{-1} + j \cdot b^{-t} : j = 0, \dots, M-1\}$$

Vollständige Induktion nach  $t$ :

$$t=1: \left\{ \sum_{k=1}^t a_k b^{-k} : \dots \right\} = \{a_1 \cdot b^{-1} : a_1 \in \{1, \dots, b-1\}\}$$

$$\{b^{-1} + j \cdot b^{-t} : \dots\} = \{b^{-1} + j \cdot b^{-1} : j \in \{0, \dots, b^1 - b^0 - 1\}\} = \{b^{-1} (j+1) : j \in \{0, \dots, b-2\}\}$$

$$= \{b^{-1} j : j \in \{1, \dots, b-1\}\}$$

$$t \Rightarrow t+1: \text{Angenommen } \left\{ \sum_{k=1}^t a_k b^{-k} : \dots \right\} = \{b^{-1} + j \cdot b^{-t} : j = 0, \dots, b^t - b^{t-1} - 1\}$$

$$\left\{ \sum_{k=1}^{t+1} a_k b^{-k} : \dots \right\} = \left\{ \sum_{k=1}^t a_k b^{-k} + a_{t+1} b^{-(t+1)} : \dots \right\}$$

$$= \left\{ x + y : x \in \left\{ \sum_{k=1}^t a_k b^{-k} : \dots \right\}, y \in \{a_{t+1} b^{-(t+1)} : a_{t+1} \in \{0, \dots, b-1\}\} \right\}$$

$$= \left\{ x + y : x \in \{b^{-1} + j \cdot b^{-t} : j \in \{0, \dots, b^t - b^{t-1} - 1\}\}, y \in \{a_{t+1} b^{-(t+1)} : a_{t+1} \in \{0, \dots, b-1\}\} \right\}$$

$$= \{b^{-1} + b^{-(t+1)} (x+y) : x \in \{j \cdot b : j \in \{0, \dots, b^t - b^{t-1} - 1\}\}, y \in \{0, \dots, b-1\}\}$$

$$= \{b^{-1} + b^{-(t+1)} (x+y) : x \in \{0, b, 2b, 3b, \dots, b^{t+1} - b^t - b\}, y \in \{0, \dots, b-1\}\}$$

$$= \{b^{-1} + b^{-(t+1)} x : x \in \{0, 1, 2, \dots, b-1, b, b+1, \dots, 2b-1, 2b, \dots, b^{t+1} - b^t - b + b - 1\}\}$$

$$= \{b^{-1} + b^{-(t+1)} x : x \in \{0, \dots, b^{t+1} - b^t - 1\}\}$$

## NUM Ü1

2) b)  $x \in \mathbb{R} \setminus \{0\}$  Rundung  $rd(x) \in F$  definiert durch  $|x - rd(x)| = \min_{z \in F} |x - z|$

wobei  $rd(x)$  das behaggrößere Element ist falls nicht eindeutig

$$x_{\min} \leq |x| \leq x_{\max}$$

$$\text{zz: } \frac{|x - rd(x)|}{|x|} \leq \frac{1}{2} b^{e-t} =: \text{eps} \dots \text{heißt Maschinenpräzision}$$

$$\exists e \in \mathbb{Z}, e_{\min} \leq e \leq e_{\max}: x \in [b^{e-1}, b^e), \text{ da o.B.d.A. } x > 0$$

Die Zahl  $z \in F$ , die  $|x - z|$  minimiert muss in  $[b^{e-1}, b^e]$  liegen.

Wie in a) gezeigt gilt  $F \cap [b^{e-1}, b^e] = \{b^e\} \cup \{(b^{-1} + j)b^{-t}\} : j=0, \dots, b^t - b^{t-1} - 1\}$

Also gibt es  $b^t - b^{t-1} + 1$  Werte mit equidistantem Abstand  $b^{-t}$  zueinander.

$$\Rightarrow |x - rd(x)| \leq \frac{1}{2} b^{e-t}, \text{ da auch } |x| \geq |b^{e-1}| > 0 \text{ gilt folgt}$$

$$\frac{|x - rd(x)|}{|x|} \leq \frac{1}{2} b^{e-t} \cdot |b^{-(e-1)}| = \frac{1}{2} b^{e-t-e+1} = \frac{1}{2} b^{-1-t}$$

□

# NUM Ü1

3.)  $F = F(2, 3, -1, 1) = \{0\} \cup \left\{ \left( 6 \sum_{k=1}^3 a_k 2^{-k} \right) \cdot 2^e : 6 \in \{-1, 1\}, a_k \in \{0, 1\}, a_1 = 1, e \in \mathbb{Z}, -1 \leq e \leq 1 \right\}$

a) ges: alle Gleitkommazahlen aus  $F$

$$F = \{0,$$

$$\frac{1}{4}, \frac{5}{16}, \frac{3}{8}, \frac{7}{16},$$

$$\frac{1}{2}, \frac{5}{8}, \frac{3}{4}, \frac{7}{8},$$

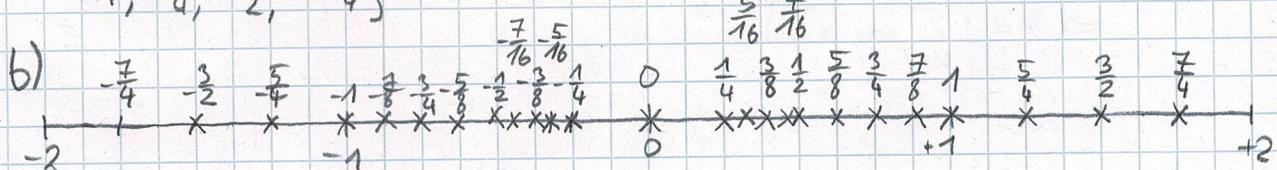
$$1, \frac{5}{4}, \frac{3}{2}, \frac{7}{4},$$

$$-\frac{1}{4}, -\frac{5}{16}, -\frac{3}{8}, -\frac{7}{16},$$

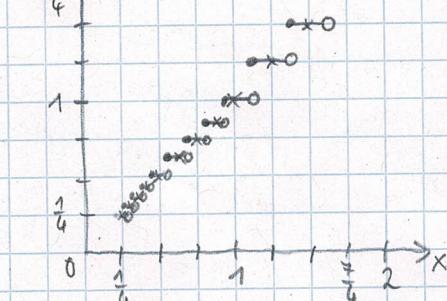
$$-\frac{1}{2}, -\frac{5}{8}, -\frac{3}{4}, -\frac{7}{8},$$

$$-1, -\frac{5}{4}, -\frac{3}{2}, -\frac{7}{4}\}$$

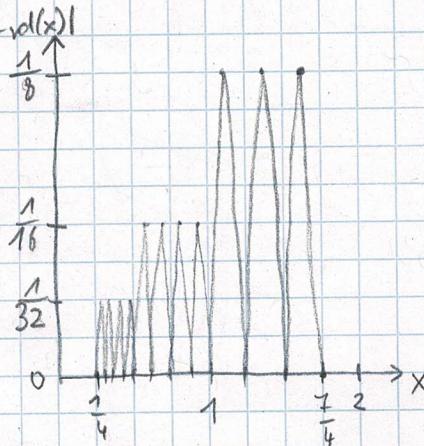
b)



$$vd(x)$$



$$|x - vdl(x)|$$



## NUM Ü1

4.) a)  $p \in \mathbb{R}_{\geq 1}$   $F(p, x) := x^2 - 2px + 1 = 0$

$\Phi : \mathbb{R}_{\geq 1} \rightarrow \mathbb{R}^2$   $\Phi(p) := (x_+, x_-)$  wobei  $x_{\pm}$  die Lösungen von  $F(p, x) = 0$  sind

ges:  $K_{rel}(p) = \frac{\|\Phi'(p)\| \cdot |p|}{\|\Phi(p)\|}$

$$\Phi(p) = (p + \sqrt{p^2 - 1}, p - \sqrt{p^2 - 1})$$

$$\text{für } p \geq 1: \Phi(p) = \left(1 + \frac{p}{\sqrt{p^2 - 1}}, 1 - \frac{p}{\sqrt{p^2 - 1}}\right)$$

$$\begin{aligned}\|\Phi(p)\| &= \sqrt{(p + \sqrt{p^2 - 1})^2 + (p - \sqrt{p^2 - 1})^2} = \\ &= \sqrt{p^2 + 2p\sqrt{p^2 - 1} + p^2 - 1 + p^2 - 2p\sqrt{p^2 - 1} + p^2 - 1} = \sqrt{2(2p^2 - 1)}\end{aligned}$$

$$\begin{aligned}\|\Phi'(p)\| &= \sqrt{\left(1 + \frac{p}{\sqrt{p^2 - 1}}\right)^2 + \left(1 - \frac{p}{\sqrt{p^2 - 1}}\right)^2} = \\ &= \sqrt{1 + 2 \frac{p}{\sqrt{p^2 - 1}} + \frac{p^2}{p^2 - 1} + 1 - 2 \frac{p}{\sqrt{p^2 - 1}} + \frac{p^2}{p^2 - 1}} = \sqrt{2 \left(1 + \frac{p^2}{p^2 - 1}\right)}\end{aligned}$$

$$\begin{aligned}K_{rel}(p) &= \frac{\|\Phi'(p)\| \cdot |p|}{\|\Phi(p)\|} = \frac{\sqrt{2} \cdot \sqrt{1 + \frac{p^2}{p^2 - 1}} \cdot \sqrt{p^2}}{\sqrt{\frac{p^2 - 1 + p^2}{p^2 - 1} \cdot p^2}} = \frac{\sqrt{2} \cdot \sqrt{p^2}}{\sqrt{2p^2 - 1}} \\ &= \sqrt{\frac{p^2}{\frac{p^2 - 1}{2p^2 - 1}}} = \sqrt{\frac{p^2}{\frac{p^2}{p^2 - 1}}} = p \cdot \sqrt{\frac{1}{p^2 - 1}}\end{aligned}$$

$$\begin{aligned}\lim_{p \rightarrow 1^+} K_{rel}(p) &= \lim_{p \rightarrow 1^+} \sqrt{\frac{p^2}{p^2 - 1}} = \sqrt{\lim_{p \rightarrow 1^+} \frac{p^2}{p^2 - 1}} = \sqrt{\lim_{p \rightarrow 1^+} p^2 \cdot \lim_{p \rightarrow 1^+} \frac{1}{p^2 - 1}} \\ &= \sqrt{1 \cdot \lim_{p \rightarrow 1^+} \frac{1}{p^2 - 1}} = \infty\end{aligned}$$

$\Rightarrow$  für  $p$  nahe an 1 ist das Problem schlecht konditioniert

b)  $t \in \mathbb{R}_{\geq 1}$   $G(t, x) := x^2 - \frac{1+t^2}{t} x + 1 = 0$

$$F(p, x) = G(p + \sqrt{p^2 - 1}, x)$$

$$x^2 - 2px + 1 = 0$$

$$x^2 - \frac{1 + (p + \sqrt{p^2 - 1})^2}{p + \sqrt{p^2 - 1}} x + 1 = 0$$

$$x^2 - \frac{1 + p^2 + 2p\sqrt{p^2 - 1} + p^2 - 1}{p + \sqrt{p^2 - 1}} x + 1 = 0$$

$$x^2 - 2p \cdot \frac{p + \sqrt{p^2 - 1}}{p + \sqrt{p^2 - 1}} x + 1 = 0$$

$$x^2 - 2px + 1 = 0 \Rightarrow \text{Unformung} \dots$$

# NUM 01

4) b) ...  $\Psi: \mathbb{R}_{\geq 1} \rightarrow \mathbb{R}^2$  ... Lösungen von  $h(t, x)$

$$\Psi(t) = \left( t, \frac{1}{t} \right), \text{ da } \left( x^2 - \frac{1+t^2}{t} x + 1 \right)(t) =$$

$$t^2 - \frac{1+t^2}{t} t + 1 = \frac{t^3}{t} - \frac{t+t^3}{t} + 1 = \frac{t^3-t-t^3}{t} + 1 = -1+1=0 \quad \text{und}$$

$$\left( x^2 - \frac{1+t^2}{t} x + 1 \right) \left( \frac{1}{t} \right) = \frac{1}{t^2} - \frac{1+t^2}{t} \cdot \frac{1}{t} + 1 = \frac{1-t-t^2}{t^2} + 1 = -1+1=0$$

$$\Psi'(t) = \left( 1, -\frac{1}{t^2} \right) \quad \| \Psi'(t) \| = \sqrt{1 + \frac{1}{t^4}}$$

$$\| \Psi(t) \| = \sqrt{t^2 + \frac{1}{t^2}}$$

$$K_{\text{rel}}(t) = \frac{\| \Psi'(t) \| \cdot |t|}{\| \Psi(t) \|} = \frac{\sqrt{1 + \frac{1}{t^4}} \cdot \sqrt{t^2}}{\sqrt{t^2 + \frac{1}{t^2}}} = \sqrt{\frac{t^2 + \frac{1}{t^2}}{t^2 + \frac{1}{t^2}}} = 1$$

$\Rightarrow$  für alle  $t (\geq 1)$  gilt, dass das Problem gut konditioniert ist