IEEE Arithmetik – Standard 754 (1985)

Datentyp	b	t	e_{min}	e_{max}	eps	x_{min}	x_{max}
`single`	2	23+1	-126	127	$\approx 5.96 \cdot 10^{-8}$	$\approx 10^{-38}$	$\approx 10^{38}$
`double`	2	52+1	-1022	1023	$\approx 1.11 \cdot 10^{-16}$	$\approx 10^{-308}$	$\approx 10^{308}$

Exception Handling

Flag	Beispiel	Ergebnis
invalid	$0/0$, $0 \cdot \infty$, $\sqrt{-1}$, ∞/∞, $+\infty + (-\infty)$	`NaN` ("not a number")
overflow	$x_{max} * x_{max}$	$\pm\infty$ (in MATLAB: `Inf`)
division by zero	$x/0$ for $x \neq 0$	$\pm\infty$
underflow	x_{min}/b^s, $1 < s < t$	subnormale Zahlen

Speicherung von `single`- bzw. `double`-Variablen:

`single`: (32 bit)

```
 S EEEEEEEE MMMMMMMMMMMMMMMMMMMMMMM
 0 1      8 9                     31
```

Beispiele:

`0 11111111 00000000000000000000000`	=	$+\infty$	
`1 11111111 00000000000000000000000`	=	$-\infty$	
`0 11111111 00000100000000000000000`	=	`NaN`	
`1 11111111 00100010001001010101010`	=	`NaN`	
`0 10000000 00000000000000000000000`	=	$+1.0 * 2^{128-127} = 2$	
`0 10000001 10100000000000000000000`	=	$+1.101 * 2^{129-127} = 6.5$	
`1 10000001 10100000000000000000000`	=	$-1.101 * 2^{129-127} = -6.5$	
`0 00000001 00000000000000000000000`	=	$+1.0 * 2^{1-127} = 2^{-126} = x_{min}$	
`0 00000000 10000000000000000000000`	=	$+0.1 * 2^{-126} = 2^{-127}$	
`0 00000000 00000000000000000000001`	=	$+0.0...01 * 2^{-126} = 2^{-149}$	
	=	kleinste darstellbare Zahl	

`double`: (64 bit)

```
 S EEEEEEEEEEE MMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMM
 0 1         11 12                                          63
```