Otto-von-Guericke-University Magdeburg
Max Planck Institute for Dynamics of Complex Technical Systems
Computational Methods for Systems and Control Theory

Dr. Jens Saak, Dipl.-Math. Martin Köhler
Website: http://www.mpi-magdeburg.mpg.de/csc/teaching/16ws/sc1/

Winterterm 2016/2017

Scientific Computing 1
Handout 6
November 12, 2016

The IEEE 754 Standard

precision	p	t	$e_{\min}$	$e_{\max}$	$\mathbf{u}$	$x_{\min}$	$x_{\max}$
`half`	2	$10+1$	-13	16	$\approx 4.88 \cdot 10^{-4}$	$\approx 6 \cdot 10^{-5}$	$\approx 1 \cdot 10^{5}$
`single`	2	$23+1$	-125	128	$\approx 5.96 \cdot 10^{-8}$	$\approx 1 \cdot 10^{-38}$	$\approx 3 \cdot 10^{38}$
`double`	2	$52+1$	$-1\,021$	$1\,024$	$\approx 1.11 \cdot 10^{-16}$	$\approx 10^{-308}$	$\approx 10^{308}$
`quad`	2	$112+1$	$-16\,381$	$16\,384$	$\approx 9.63 \cdot 10^{-35}$	$\approx 10^{-4\,932}$	$\approx 10^{4\,932}$

Table 1: IEEE standard 754-2008, data types.

Storage patterns for `half`, `single` and `double` precision variables:

`half`: (16 bit)

```
S EEEE MMMMMMMMMMMMMMMM
0 1  4 5               15
```

`single`: (32 bit)

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

`double`: (64 bit)

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

Flag	Example	Result
invalid	$0/0,\ \ 0\cdot\infty,\ \ \sqrt{-1},$ $\infty/\infty,\ \ +\infty+(-\infty)$	`NaN` (“not a number”)
overflow	$x_{\max} * x_{\max}$	$\pm\infty$ usually denoted: $\pm$`Inf`
division by zero	$x/0$ for $x \neq 0$	$\pm\infty$
underflow	$x_{\min}/p^s, 1 < s < t$	subnormal number
inexact	$\mathrm{rd}(x \circ y) \neq x \circ y$	correctly rounded result

Table 2: IEEE Standard 754, Exception Handling.

Examples:

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	=	$+0.0\ldots01 * 2^{-126} = 2^{-149}$
	=	smallest representable number
0 00000000 00000000000000000000000	=	$+0$
1 00000000 00000000000000000000000	=	-0
0 01111111 00000000000000000000000	=	$1.0 * 2^{127-127} = 1.0$
1 01111111 00000000000000000000000	=	$-1.0 * 2^{127-127} = -1.0$