WS 23/24 Numerics Notes

Igor Dimitrov

2023-10-30

Table of contents

Preface

Notes for the lecture "WS 23/24 Numerics 0" at Uni Heidelberg.

1 Floating Point Numbers

1.1 ANSI/IEEE 64 Bit

Let $\tilde{a}$ be a 64 bit IEEE floating point number. $\tilde{a}$ is represented as

```
S E ... E M ... M
```

Where S is the sign bit, 11 E's are the exponent bits and 52 M's are mantissa bits. Interpretation (Case analysis on value of E):

1. S | 0 … 0 | M:
    1. $M = 0 \Rightarrow \tilde{a} = (-1)^S 0$
    2. $M \neq 0 \Rightarrow \tilde{a} = (-1)^S \times 2^{-1022} \times 0.M$ (**subnormal range**)
2. $1 \leq E \leq 2046 \Rightarrow \tilde{a} = (-1)^S \times 2^{E-1023} \times 1.M$ (**normal range**)
3. S | 1 … 1 | M:
    1. $M = 0 \Rightarrow \tilde{a} = (-1)^S$`inf`
    2. $M \neq 0 \Rightarrow \tilde{a} =$ `NaN`(Not a Number) (**exceptions**)

1.2

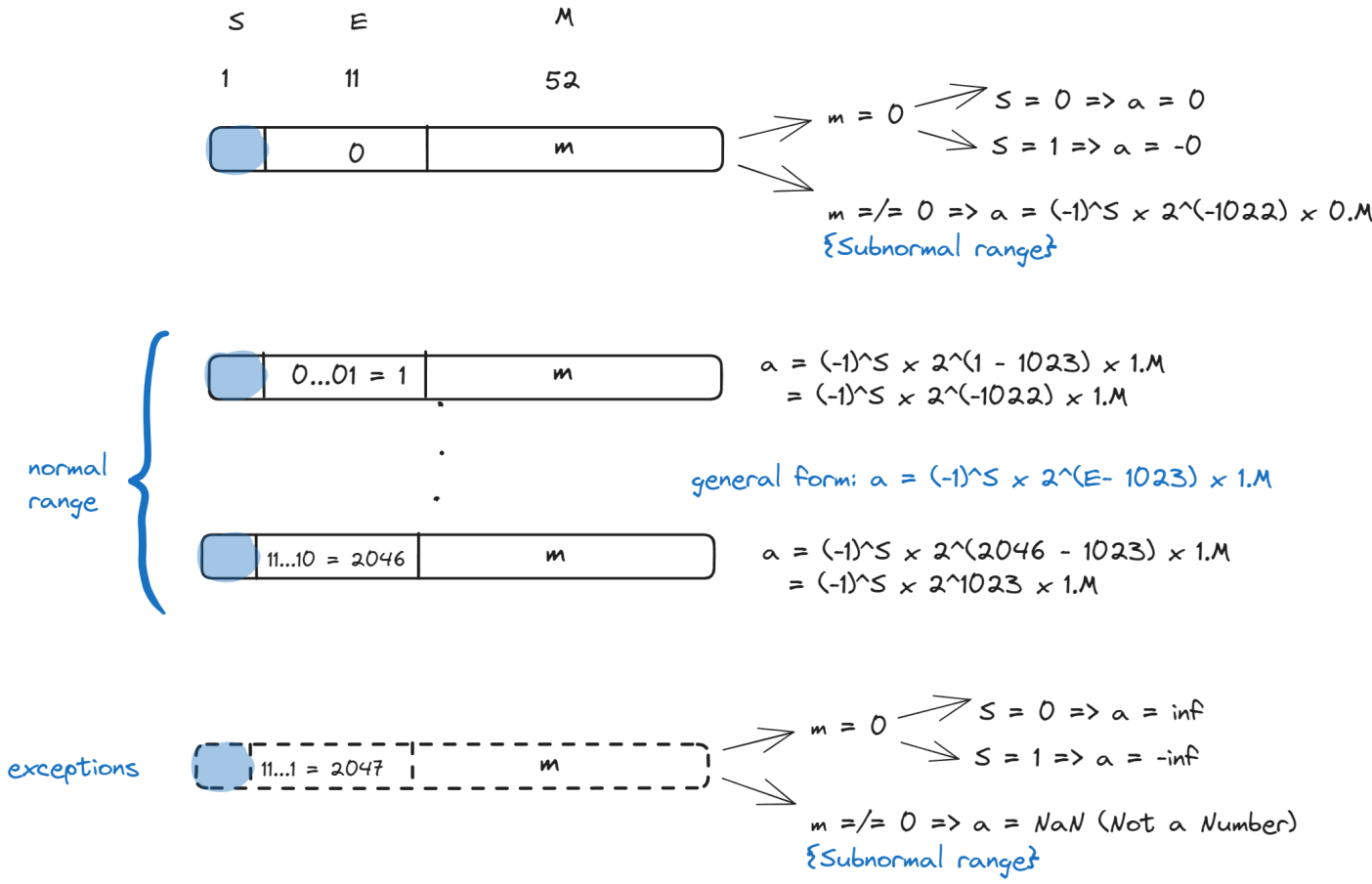

Figure 1.1: floating-point