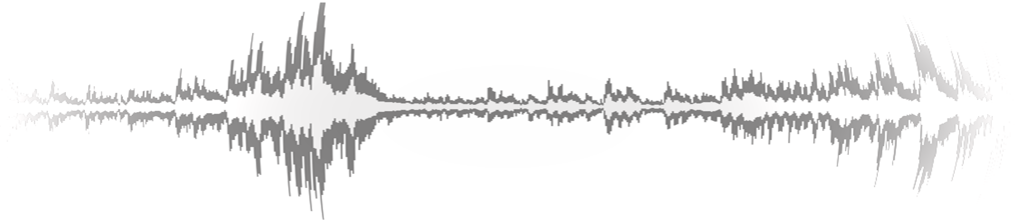
Digital Signal Processing for Music

Part 13: Digital Number Formats

alexander lerch



Georgia Tech Center for Music Technology
College of Design

number formats

word length and SNR

w	**Δ**	**Max. Amp**	**theo. SNR**
8 (Int)	± 1	$0 \dots 255$	$\approx 48\,\text{dB}$
16 (Int)	± 1	$-32768 \dots 32767$	$\approx 96\,\text{dB}$
20 (Int)	± 1	$-524288 \dots 524287$	$\approx 120\,\text{dB}$
24 (Int)	± 1	$-16777216 \dots 16777215$	$\approx 144\,\text{dB}$
32 (Float)	$\pm 1.175 \cdot 10^{-38}$	$\pm 3.403 \cdot 10^{1038}$	$1529\,\text{dB}$
64 (Float)	$\pm 2.225 \cdot 10^{-308}$	$\pm 1.798 \cdot 10^{10308}$	$12318\,\text{dB}$

how do we represent this in bits

number formats

word length and SNR

w	**Δ**	**Max. Amp**	**theo. SNR**
8 (Int)	± 1	$0 \ldots 255$	$\approx$48 dB
16 (Int)	± 1	$-32768 \ldots 32767$	$\approx$96 dB
20 (Int)	± 1	$-524288 \ldots 524287$	$\approx$120 dB
24 (Int)	± 1	$-16777216 \ldots 16777215$	$\approx$144 dB
32 (Float)	$\pm 1.175 \cdot 10^{-38}$	$\pm 3.403 \cdot 10^{1038}$	1529 dB
64 (Float)	$\pm 2.225 \cdot 10^{-308}$	$\pm 1.798 \cdot 10^{10308}$	12318 dB

how do we represent this in bits

number formats

value range

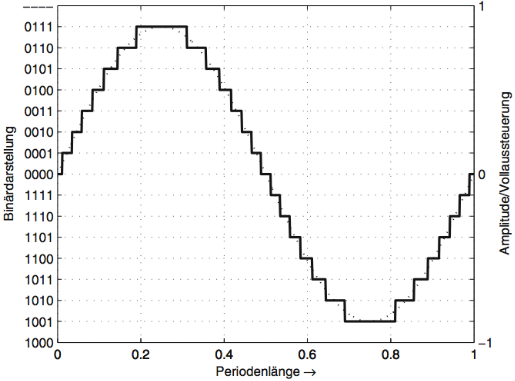
- **unnormalized**[1]: $-2^{w-1} \dots 2^{w-1} - 1$
    - used for transmission etc.

- **normalized** (word length independent): $-1 \dots 1$
    - used for floating point representation
    - used for processing

[1] remember: non-symmetric step count for positive and negative values

number formats

number representation 1/2



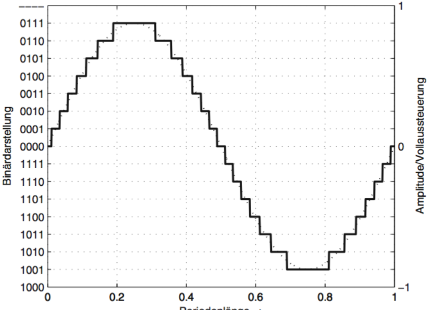
- Least Significant Bit (LSB): b_0 (usually on the right)
- Most Significant Bit (MSB): b_{w-1} (usually on the left)

number formats

number representation 2/2

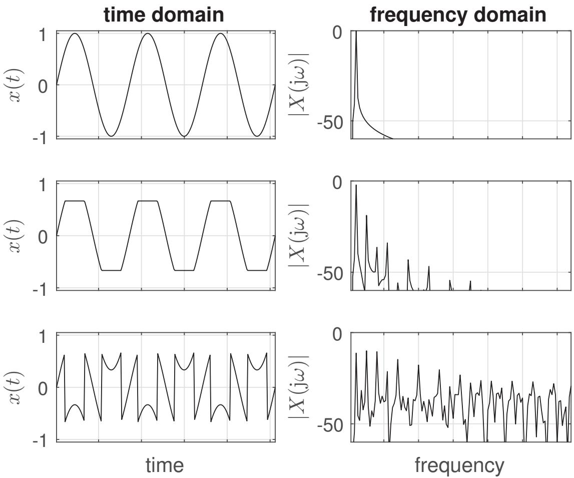
format	amplitude	range (normalized)
2-Complement	$x_Q = -b_{w-1} + \sum\limits_{i=0}^{w-2} b_i 2^{-(w-i-1)}$	$-1 \leq x_Q \leq 1 - 2^{-(w-1)}$
unsigned	$x_Q = \sum\limits_{i=0}^{w-1} b_i 2^{-(w-1)}$	$0 \leq x_Q \leq 1 - 2^{-w}$

- w : word length
- b_i : ith Bit



number formats

quantization: clipping & wrap-around



number formats

fixed point and floating point

number formats and their most frequent uses

- **unsigned format**: small word lengths (4...8 Bit)
- **2's complement**: file formats with higher word lengths (16...24 Bit), some DSPs
- **floating point**: internal representation for processing

number formats

fixed point and floating point

number formats and their most frequent uses

- **unsigned format**: small word lengths (4. . . 8 Bit)
- **2’s complement**: file formats with higher word lengths (16. . . 24 Bit), some DSPs
- **floating point**: internal representation for processing

number formats

fixed point and floating point

number formats and their most frequent uses

- **unsigned format**: small word lengths (4. . . 8 Bit)
- **2's complement**: file formats with higher word lengths (16. . . 24 Bit), some DSPs
- **floating point**: internal representation for processing

number formats

fixed point and floating point

number formats and their most frequent uses

- **unsigned format**: small word lengths (4. . . 8 Bit)
- **2's complement**: file formats with higher word lengths (16. . . 24 Bit), some DSPs
- **floating point**: internal representation for processing

number formats

floating point 1/2

$$x_Q = M_G \cdot 2^{E_G}$$

- M_G: Normalized Mantissa $0.5 \leq M_G < 1$
- E_G: Exponent

32 Bit IEEE 754 Floating Format:

Bit 31: Sign	**Bits 30-23: Exponent**	**Bits 22-0: Mantissa**
s	$e_7 \ldots e_0$	$m_{22} \ldots m_0$

Exceptions

Typ	E_G	M_G	**Value**
normal	$1 \leq E_G \leq 254$	any	$(-1)^s(0.m)2^{E_G-127}$
NAN (not a number)	255	$\neq 0$	undefined
Infinity	255	$= 0$	∞
Zero	0	0	0

number formats

floating point 1/2

$$x_Q = M_G \cdot 2^{E_G}$$

- M_G: Normalized Mantissa $0.5 \leq M_G < 1$
- E_G: Exponent

32 Bit IEEE 754 Floating Format:

Bit 31: Sign	Bits 30-23: Exponent	Bits 22-0: Mantissa
s	$e_7 \; \ldots \; e_0$	$m_{22} \; \ldots \; m_0$

Exceptions

Typ	E_G	M_G	Value
normal	$1 \leq E_G \leq 254$	any	$(-1)^s(0.m)2^{E_G-127}$
NAN (not a number)	255	$\neq 0$	undefined
Infinity	255	$= 0$	∞
Zero	0	0	0

number formats
floating point 1/2

$$x_Q = M_G \cdot 2^{E_G}$$

- M_G: Normalized Mantissa $0.5 \leq M_G < 1$
- E_G: Exponent

32 Bit IEEE 754 Floating Format:

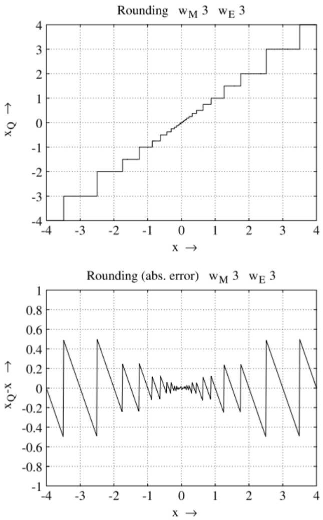
Bit 31: Sign	Bits 30-23: Exponent	Bits 22-0: Mantissa
s	e_7 … e_0	m_{22} … m_0

Exceptions

Typ	E_G	M_G	Value
normal	$1 \leq E_G \leq 254$	any	$(-1)^s(0.m)2^{E_G-127}$
NAN (not a number)	255	$\neq 0$	undefined
Infinity	255	$= 0$	∞
Zero	0	0	0

number formats

floating point 2/2

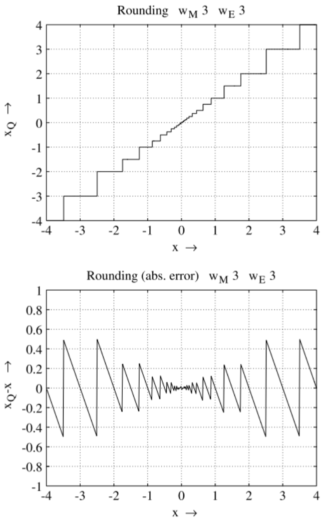


- **high exponent**: large quantization error energy
- **low exponent**: small quantization error energy

- **linear quantization** within one exponent

number formats

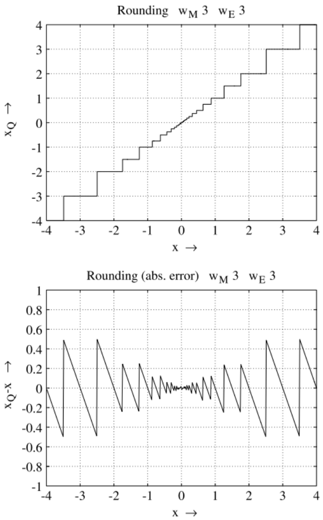
floating point 2/2



- **high exponent**: large quantization error energy
- **low exponent**: small quantization error energy
- **linear quantization** within one exponent

number formats

floating point 2/2



- **high exponent**: large quantization error energy
- **low exponent**: small quantization error energy

- **linear quantization** within one exponent

number formats

quantization: summary

- most common number representations
  - 2-complement for high quality audio storage
  - floating point for high quality audio processing (non-linear quantization)