Digital Signal Processing for Music

Part 12: Digital Number Formats

Andrew Beck

Word length and SNR

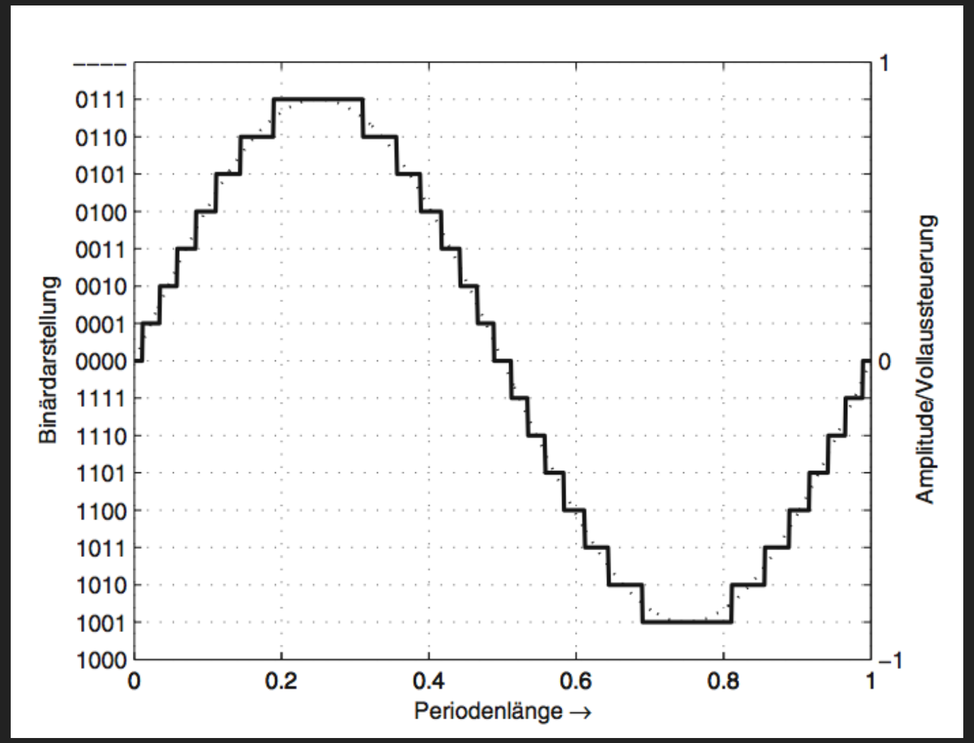
w	Δ	Max. Amp	theo. SNR
8 (Int)	± 1	$0 \ldots 255$	$\approx 48\,\mathrm{dB}$
16 (Int)	± 1	$-32768 \ldots 32767$	$\approx 96\,\mathrm{dB}$
20 (Int)	± 1	$-524288 \ldots 524287$	$\approx 120\,\mathrm{dB}$
24 (Int)	± 1	$-16777216 \ldots 16777215$	$\approx 144\,\mathrm{dB}$
32 (Float)	$\pm 1.175 \cdot 10^{-38}$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

How do we represent this in bits?

Number Formats: Value Range

- **Unnormalized**: $-2^{w-1} \ldots 2^{w-1} - 1$
  - Integer representation
  - Non-symmetric step count for positive and negative values
  - Used for transmission, etc.
- **Normalized**: $-1 \ldots 1$
  - Used for floating point representations
  - Word length independent
  - Used for processing

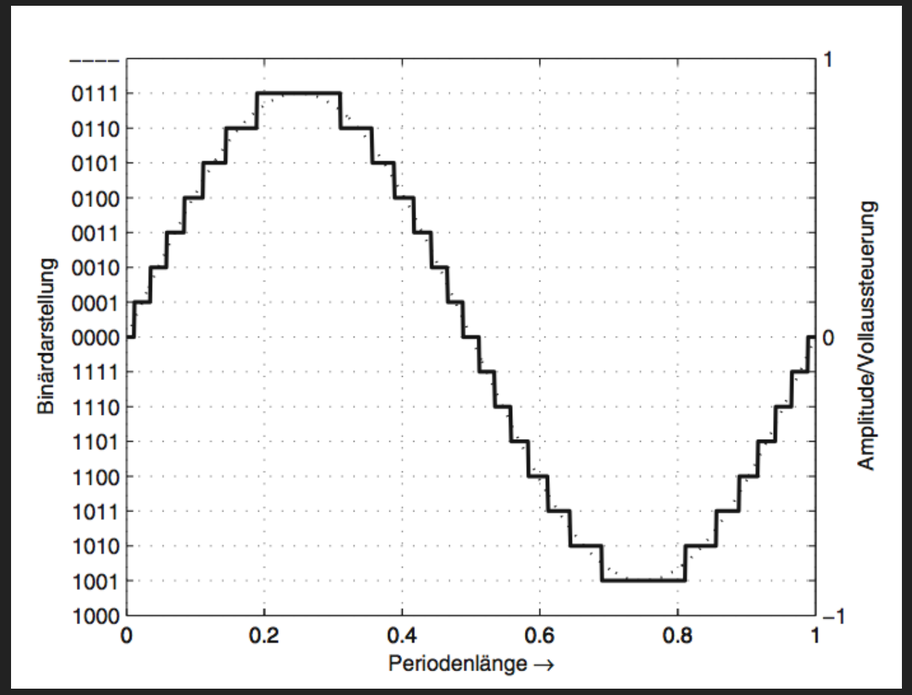
Number Representation



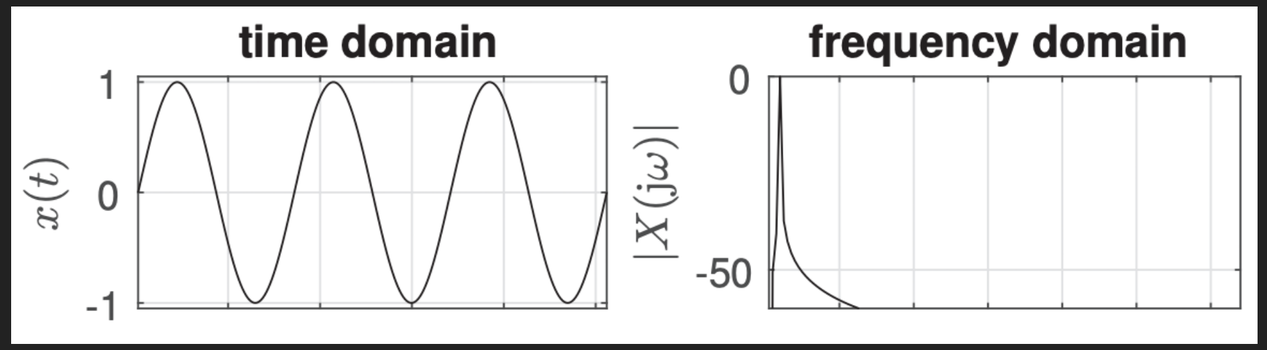
- Least Significant Bit (LSB): b_0 (usually on the right)
- Most Significant Bit (MSB): b_{w-1} (usually on the left)

Format	Amplitude	Range (normalized)
2-Complement	$x_Q = -b_{w-1} + \sum_{i=0}^{w-2} b_i 2^{-(w-i-1)}$	$-1 \leq x_Q \leq 1 - 2^{-(w-1)}$
Unsigned	$x_Q = \sum_{i=0}^{w-1} b_i 2^{-(w-1)}$	$0 \leq x_Q \leq 1 - 2^{-w}$

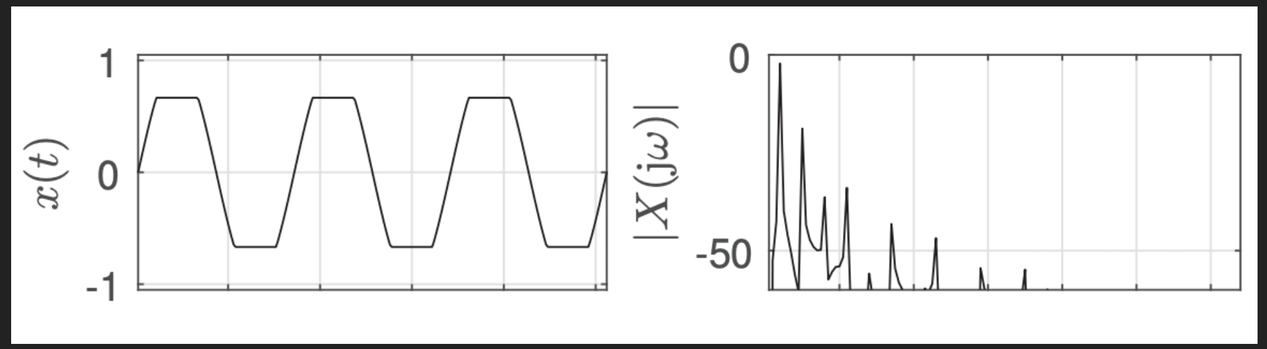
- w: word length
- b_i: ith bit



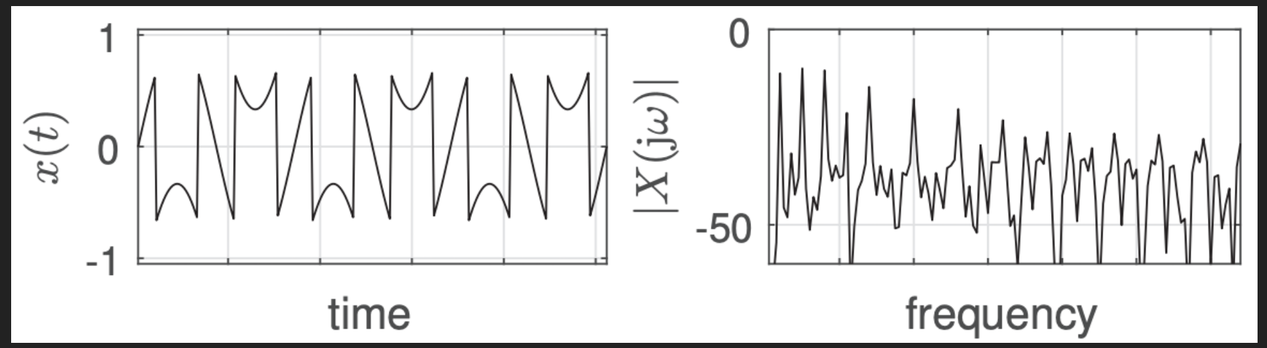
Clipping & Wrap-Around



0:00 / 0:00



0:00 / 0:00



0:00 / 0:00

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

Floating Point

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

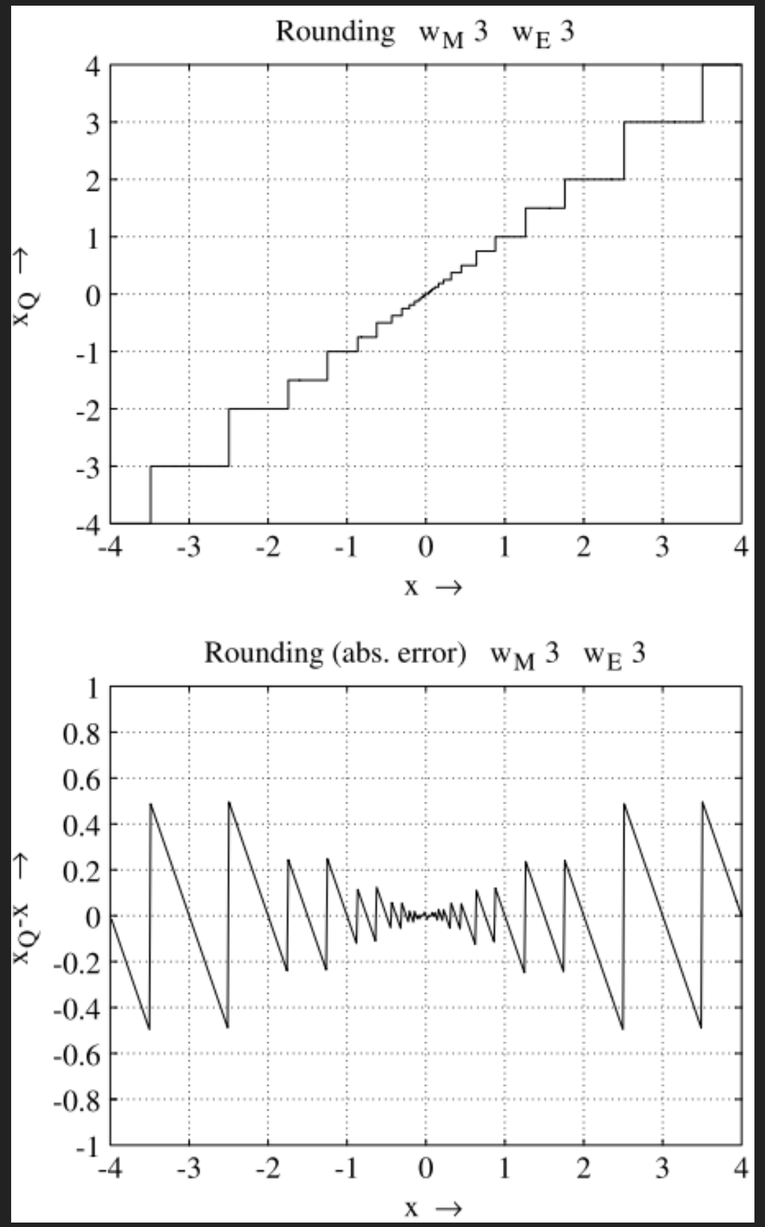
- M_G: Normalized Mantissa $0.5 \leq M_G < 1$
- E_G: Exponent

32 Bit IEEE 754 Floating Format

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

Exceptions

Type	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



- **High Exponent**:

  Large quantization error energy
- **Low Exponent**:

  Small quantization error energy
- **Linear quantization**:

  Within one exponent

Summary

- Most common number representations
  - 2-Complement for high quality audio storage
  - Floating point for high quality audio processing (non-linear quantization)