Digital Signal Processing

Alle Angaben ohne Gewähr.

Andreas Selig

February 23, 2001

Contents

1 Number Representation and Elementary Operations

1.1 Numbers for Fixed-Point Arithmetic

$$\boxed{z = \sum_{i=n}^{m} \underbrace{a_i}_{digit} \cdot \underbrace{b_i}_{signification}} \tag{1}$$

e.g.: decimal system

$276 = 6 \cdot 10^0 + 7 \cdot 10^1 + 2 \cdot 10^2$

$b_i = 10^i \qquad a_i = \{0, 1, \ldots, 9\}$

In general:

$\boxed{b_i = c^i} \qquad \boxed{a_i = \{0, 1, \ldots, c-1\}}$

$c = 10$ decimal system
$c = 2$ binary system

1.1.1 Binary Coding

$$\boxed{z = \sum_{i=n}^{m} a_i \cdot 2^i} \tag{2}$$

$a_i = \{0, 1\} \qquad z \geq 0$

Example:

$z = \underbrace{1 \cdot 2^2 + 1 \cdot 2^1 + 0 \cdot 2^0}_{integer} + \underbrace{0 \cdot 2^{-1} + 1 \cdot 2^{-2}}_{fraction} = \sum_{n=-2}^{2} a_i \cdot 2^i \qquad$ 5 digits

$\max\{z\} = 2^{m+1} - 2^n$

$\boxed{z \in \{0, 2^n, 2 \cdot 2^n, \ldots, 2^{m+1} - 2^n\}}$ in increments of 2^n

In digital signal processors (DSP's), the number format is fixed:

$$\boxed{m + 1 - n = const.} \tag{3}$$

Common ones are: 8,16,32,64,128 bit

Multiplication of Binary Numbers

- k_1-bit-word $\times$ k_2-bit-word $\quad \rightarrow \quad$ $(k_1 + k_2)$-bit-word
- in processors with limiting wordlenth is necessary $\rightarrow$ rounding, overflow protection
- $(k_2 - 1)$ additions are necessary for the multiplication; Booth's algorithm reduces that.

1.1.2 Two's-Complement Coding

Problem: Binary coding only allows positive values

Solution: Using the two's-complement coding

- negative items are represented by the TC of their absolute value (TC always > 0)
- TC for z with $|z| < 2^m$

$$\boxed{2^{m+1} - |z|} \tag{4}$$

For $|z| < 2^m$

$$\boxed{\tilde{z} = \begin{cases} z & z \geq 0 \\ 2^{m+1} + z & z < 0 \end{cases}} \quad (5)$$

$$\boxed{z = -a_m \cdot 2^m + \sum_{i=n}^{m-1} a_i + 2^i} \quad (6)$$

<u>Complements</u>

For general number base C:

- C´s-complement of z
  $\boxed{\tilde{z} = C^{m+1} - z}$

  Difference to the largest number in the same number system plus 1 LSB

- $(C-1)$´s-complement of z
  $\boxed{\tilde{z} = (c^{m+1} - c^n) - z}$

  Difference to the largest number in the same number system

Note:A complement is again a number in the same number system.

<u>Addition and Subtraction of TC-numbers</u>

$$z = -a_m 2^m + \underbrace{\sum_{i=n}^{m-1} a_i 2^i}_{X} = -a_m 2^m + X$$

$$\begin{array}{llll} \rightarrow & z \geq 0 & a_m = 0 & x = z > 0 \\ & z < 0 & a_m = 1 & x = a_m 2^m - |z| \end{array}$$

Modulo Operations: $z \bmod 2^{m+1}$ is the remainder of z in respect to 2^{m+1}.If there is a overflow into significance 2^{m+1}, then this overflow will be ignored due to the $\bmod 2^{m+1}$ operation.

Theorem: Addition (Subtraction) for any number z_i in TC code is done by:

1. regular addition of binary numbers of all digits
2. build the $\bmod 2^{m+1}$ i.e. ignore the overflow into position (m+1)
   z_i in the form $a_m 2^m + X$

a_m	X

   interpret z_i as a binary number.

   and yields the correct result if $\sum_{i=0}^{g} z_i \quad \in \{-2^m, \ldots, 2^m - 2^n\}$

Proof: 3 types of errors:

1. For each $z_i < 0$
   $a_m 2^m + X - (-a_m 2^m + X) = a_m \cdot 2 \cdot 2^m = 2^{m+1}$
2. binary addition of z_i´s (e.g. two $z_i < 0$)can cause an overflow at $2^m + 1$
   error $= -2^{m+1}$
3. result is now interpreted as a TC number
   $z = \sum_{i=0}^{g} z_i = a_m \underbrace{a_{m-1} \ldots a_n}_{X}$
   if $a_m = 1$ $(z \geq 2^m)$ then the error is
   $(2^m + X) - (-2^m + X) = 2^{m+1}$

Each error is a multiple of $2m + 1 \rightarrow$ the total error is also a multiple of 2^{m+1}

If the correct result is in the valid rage then the errors do not show when all intermediate additions were done $\bmod 2^{m+1}$ and the result is identical to the correct one.

1.1.3 Sign Magnitude Representation

Sign bit + absolute value

$$\boxed{z = (-1)^s |z|} \tag{7}$$

s	a_m	$\cdots$	a_n

$s = 0 \;\rightarrow\; z > 0$
$s = 1 \;\rightarrow\; z < 0$

1.1.4 CSD Coding (Canonic Sign Digit)

$$\boxed{z = \sum_{i=n}^{m} a_i 2^i} \tag{8}$$

$a_i \in \{-1, 0, 1\}$

Conversion of a binary number into a CSD code:

Theorem:

$$\boxed{2^{k+2} - 2^k = 2^{k+1} + 2^k} \tag{9}$$

Application to binary number:

- replace 2 adjancement "1"´s with "1 0 -1"
- CSD codes have no adjacent "± 1"

Conversion by computer:
binary $z \rightarrow \boxed{(z + \frac{1}{2}z) - \frac{1}{2}z}$ in CSD

Properties of CSD code words:

- large number of ´0´-bits
  $\rightarrow$ few additions in multiplication on average $\frac{m}{3}$ bits $\neq 0$ for an m-bit word
- addition of 2 CSD words does not necessarily yield another CSD word $\rightarrow$ conversion necessary

1.2 Floating Point Representation

Reason: Fixed-point representation often causes overflow due to the limited range

Solution: floating-pointing representation and mantissa notation that has

- large dynamic range
- virtually no overflow

$\boxed{z = m \cdot 2^E}$

IEEE-Standard for 32-bit:

$$\boxed{m = (-1)^s \cdot \sum_{i=0}^{23} a_i \cdot 2^{-i}} \tag{10}$$

$a_i \in \{0, 1\} \quad \forall \quad i \in \{1, 2, \ldots, 23\}$
$E = e - 127 \quad\quad e \in \{0, 1, \ldots, 255\}$
s ... sign bit
m ... sign mantissa

$$\boxed{z = (-1)^s \left(\sum_{i=0}^{23} a_i \cdot z^{-i} \right) \cdot 2^{e-127}} \tag{11}$$

Representation/Layout:

s	e	m

mantissa range	$\pm\{1, 1+2^{-23}, \ldots, 2-2^{-23}\}$
exponent range	$E \in \{-126, \ldots, 127\}$
	$E \in \{-127, 128\}$ are not allowed $\rightarrow$ special cases

1.3 Elementary Operations for dip Signal Processing

- addition (subtraction): adder
- multiplication: multiplier
- shifter (multiplication with 2^n)
- delay

2 Digital Multiplication

2.1 Multiplication of Complex Numbers

x, a complex $\qquad x = x_1 + \mathrm{i} \cdot x_2 \qquad a = a_1 + \mathrm{i} \cdot a_2$

$$\boxed{y = x \cdot a = (x_1 + \mathrm{i} \cdot x_2)(a_1 + \mathrm{i} \cdot a_2)} \tag{1}$$

$$\boxed{y = \underbrace{(x_1 a_1 - x_2 a_2)}_{\text{real}} + \mathrm{i}\, \underbrace{(x_2 a_1 + x_1 a_2)}_{\text{imaginary}}} \tag{2}$$

$$\boxed{y = (x_1 - x_2)a_1 + x_2(a_1 - a_2) + \mathrm{i} \cdot [(x_1 + x_2)a_2 + x_2(a_1 + a_2)]} \tag{3}$$

2.2 Booth´s Algorithm

<u>Goal:</u> reduce the number of additions in the multiplication

Consider the multiplication $y = a \cdot x \qquad a, x$ are in the TC coding

$$\boxed{a = -a_0 + \sum_{i=1}^{n-1} a_i \cdot 2^{-i}} \tag{4}$$

$$a = -a_0 + \sum_{i=1;\text{odd}}^{n-1} a_i \cdot 2^{-i} + \sum_{i=1;\text{even}}^{n-2} a_i \cdot 2^{-i} + \sum_{i=1;\text{even}}^{n-2} a_i \cdot 2^{-i} - \sum_{i=1;\text{even}}^{n-2} a_i \cdot 2^{-i}$$

$$a = -a_0 + \sum_{i=1;\text{odd}}^{n-1} a_i \cdot 2^{-i} + \sum_{i=2;\text{even}}^{n-2} a_i \cdot 2^{-i} - 2 \sum_{i=1;\text{even}}^{n-2} a_i \cdot 2^{-i}$$

Substitution:

$$a = -a_0 + \sum_{i=1;\text{odd}}^{n-1} a_i \cdot 2^{-i} + \sum_{i=1;\text{odd}}^{n-3} a_{i+1} 2^{-i} - 2 \sum_{i=3}^{n-1} a_{i-1} 2^{-i}$$

$$\boxed{a = \sum_{i=1;\text{odd}}^{n-1} (a_i + a_{i+1} - 2a_{i-1}) \cdot 2^{-i}} \tag{5}$$

the number of additions is about half:

- all z_i can be computed beforehand for the fixed multiplier

<u>Booth´s coding:</u>

$z_i = a_i + a_{i+1} - 2a_{i-1} \quad \forall \quad i \in \{1, 3, \ldots, n-1\}$

$\qquad z_i \in \{-2, -1, 0, 1, 2\}$

$$x \cdot a = \sum_{i=1;\text{odd}}^{n-1} x \cdot z_i \cdot 2^{-i}$$

$\rightarrow$ only shifts of the signal x are necessary (no "real" additions are necessary)

2.3 CSD Coding of the Multiplier

Conversion → CSD

Result:

- n bit CSD word with on average $\frac{n}{3}$ bits that are non-zero
- $\frac{n}{3} - 1$ additions necessary
- even less that the Booth´s algorithm , that needs a CSD ALU (2-valued) and complement operation

2.4 Factorization of the Multiplier

Goal:

$$\boxed{a = f_0 \cdot \prod_i f_i = \lambda_0 \cdot 2^{\nu_0} \cdot \prod_i \underbrace{(1 + \lambda_i \cdot 2^{\nu_i})}_{f_i}} \tag{6}$$

Algorithm:

1. Normalization:

$$\boxed{a_0 = \frac{a}{\lambda_0 \cdot 2^{\nu_0}}} \qquad \in [0.5, 1] \tag{7}$$

   → unique $\lambda_0, \nu_0 \rightarrow f_0 = \lambda_0 \cdot 2^{\nu_0}$

2. Expand a_0

$$\boxed{a_0 = \underbrace{(1 + \lambda_1 \cdot 2^{\nu_1})}_{f_i} + e_1} \tag{8}$$

   for smallest error $|e_1| \rightarrow e_1 = a_0 - f_1$ from a table (page 6)

3. Expande$_1$

$$\boxed{e_1 = f_1 \cdot \lambda_2 \cdot 2^{\nu_2} + e_2} \tag{9}$$

   for smallest error $|e_2|$
   → λ_2, ν_2 from a table $\qquad e_2 = e_1 - f_1 \cdot \lambda_2 \cdot 2^{\nu_2}$
   Continue until e_n small enough

2.5 Limiting the World length after Multiplication

There are two methods to limit the wordlenth of a N bit word, quantization by truncation or by rounding.

Number format , for $N = m + 1 - n$; $m \geq 0, n \leq 0$

a_m	a_{m-1}	$\cdots$	a_{n+1}	a_n

2.5.1 Rounding

- rounding to a_n with significance $2^n = q$ (q: quantization level)
  1. add $\frac{q}{2}$
  2. truncate after bit a_n at significance $2^{n-1} = \frac{q}{2}$
- Quantization characteristic (for y in TC coding)
  non-linear characteristic , quantization error

$$\boxed{e = Q\{y\} - y} \tag{10}$$

  → **additive quantization noise model:**
  the value e is the result of a random process (→ noise)
  Assumption: e is uniformly distributed in $[-\frac{q}{2}, \frac{q}{2}]$
  Note: this is only true if y is uniformly distributed (e.g. speech, music)

2.5.2 Truncating

- truncation all bits to the right of a_n up to the significance $\frac{q}{2} = a^{n-1}$ and smaller
- quantization error is uniformly distributed in $[-q, 0]$

3 Sampling Rate Conversion

In System Theory:

$$\boxed{f(t) \circ\!\!-\!\!\bullet F(\omega)} \tag{1}$$

$$\boxed{f_s(t) = \sum_{n=-\infty}^{\infty} f(nT_s) \cdot \delta(t - nT_s)} \tag{2}$$

$$\boxed{F_s(t) = \frac{1}{T_S} \sum_{n=-\infty}^{\infty} F(\omega - n\omega_s)} \tag{3}$$

with:
T_S ... sampling period $\rightarrow \omega_s = \frac{2\pi}{T_s}$
ω_s ... sampling frequency $\omega_s \geq 2\omega_0$
r_s ... sampling rate $r_s = \frac{1}{T_s}$

$$F_s = \sum_{n=-\infty}^{\infty} f(nT_s) \cdot e^{-i\omega T_s n} = \sum_{n=-\infty}^{\infty} f_n \cdot z^{-n}|_{z=e^{i\omega T_s}} = \tilde{F}(z)|_{z=e^{i\omega T_s}}$$

$$\boxed{F_s = \tilde{F}(z)|_{z=e^{i\omega T_s}}} \tag{4}$$

$\tilde{F}(z)$ is the z-transformed (ZT) of the series f_n (complex-valued function of a complex-valued value).
In real world systems, we have different sampling rates and want to connect theses $\rightarrow$ sampling rate conversion.

3 Types of Sampling Rate Conversions

1. $T_s' = M \cdot T_s \quad M \in \mathbb{N}$ Down sampling, Decimation
   period increases
   frequency decreases $\omega_s' = \frac{\omega_s}{M}$
   rate decreases $r_s' = \frac{r_s}{M}$
2. $T_s' = \frac{T_s}{L} \quad L \in \mathbb{N}$ Up-sampling, Interpolation
   period decreases
   frequency increases $\omega_s' = L \cdot \omega_s$
   rate increases $L \cdot r_s$
3. $T_s' = \frac{M}{L} \cdot T_s \quad M, L \in \mathbb{N} \quad \frac{M}{L}$ is rational
   $\omega_s' = \frac{L}{M} \cdot \omega_s \quad r_s' = \frac{L}{M} \cdot r_S$

3.1 Down-Sampler

$$\boxed{T_s' = M \cdot T_s} \tag{5}$$

$$\boxed{\omega_s' = \frac{\omega_s}{M}} \tag{6}$$

$z = e^{i\omega T_s}$
$z' = e^{i\omega T_s'} = z^M$

Note: $\omega_s' < \omega_s$
$\rightarrow$ the spectra in $F_s(\omega)$ move closer together
$\rightarrow$ danger of aliasing
$\rightarrow F(\omega) = 0$ for $|\omega| > \frac{\omega_s'}{2}$

Condition for time-invariance:
If for $y(t) = H\{x(t)\}$follows that $y(t - t_0) = H\{x(t - t_0)\}$, than the system $H\{.\}$ is time-invariant.

3.2 Decimation

see supplement page

3.3 Up-Sampler

$$\boxed{T_s' = \frac{T_s}{L} T_s} \tag{7}$$

$$\boxed{\omega_s' = L \cdot \omega_s} \tag{8}$$

$z = e^{i\omega T_s}$
$z' = e^{i\omega T_s'} = e^{i\omega \frac{T_s}{L}} = z^{\frac{1}{L}}$

$$\boxed{F_s = \frac{1}{T_s} \sum_{n=-\infty}^{\infty} F(\omega - n\omega_s)} \tag{9}$$

$$\boxed{F_s' = \frac{1}{T_s} \sum_{n=-\infty}^{\infty} F^*(\omega - n\omega_s')} \tag{10}$$

$$\boxed{F^* = \sum_{k=-L_0}^{L_0} F(\omega - k\omega_s)} \tag{11}$$

$L_0 = \frac{L-1}{2}$ for$L$ odd (more complicated for L even)
The interpolation low pass amplifies by L to make the value $y_m = x(m \cdot T_s')$ as large as x_m for $m = n \cdot L$ after filtering.

3.4 Interpolation

See supplement page

3.5 Sampling Rate Conversion by Rational Factor

$M, L \in \mathbb{N}$

$$\boxed{T_s' = \frac{M}{L} \cdot T_s} \tag{12}$$

$$\boxed{\omega_s' = \frac{L}{M} \cdot \omega_s} \tag{13}$$

Procedure:

1. Interpolation $T_{s1} = \frac{1}{L} \cdot T_s$
2. Decimation $T_s' = M \cdot T_{s1}$

Sampling theorem has to be met at all times $\rightarrow \omega_s, \omega_{s1}, \omega_s' \geq 2\omega_c$

3.6 Equivalences

see supplement page

4 Synthesis of digital Filter

The system is often realized by a digital system, i.e. often quantization and coding, e.g. by a DSP-program or a digital system

- h_n: unit pulse response
  response to $x_n = \delta_n = \begin{cases} 1 & \text{for} \quad n = 0 \\ 0 & \text{otherwise} \end{cases}$
- $H(z)$: (pulse)transfer function
  $H(z) = \frac{Y(z)}{X(z)}$

- $x_n \quad \circ\!\!-\!\!\bullet \quad X(z)$

$$\boxed{X(z) = \mathcal{Z}\{x_n\} = \sum_{n=-\infty}^{\infty} x_n \cdot z^{-n}} \tag{1}$$

  the sum converges only in $|z| > R_x$ (radius of convergence)

- convolution
  $Y(z) = X(z) \cdot H(z) \qquad \rightarrow \qquad y_n = \sum_{k=0}^{n} x_n h_{n-k}$

4.1 Types of Filter

4.1.1 FIR Filters

- Impulse response

$$\boxed{h_n = \begin{cases} c_n & \text{for} \quad n = 1, 2, ..., N \\ 0 & \text{otherwise} \end{cases}} \tag{2}$$

N	...	degree of the filter
$(N+1)$	...	length

- transfer function

$$\boxed{H(z) = \sum_{n=0}^{N} c_n \cdot z^{-n} = \frac{c_0 z^N + c_1 z^{N-1} + \cdots + c_N}{z^N}} \tag{3}$$

zeros	...	determined by numerator
poles	..	N poles at$z = 0$

- difference equation for y_n

$$\boxed{y_n = h_0 \cdot x_n + h_1 \cdot x_{n-1} * \cdots + h_N \cdot x_{n-N}} \tag{4}$$

  $\rightarrow$ non-recursive $\rightarrow$ non-recursive filter $\rightarrow$ FIR (finite impulse response)

4.1.2 FIR Filters with linear Phase

Linear Phase:

- continuous-time domain:
  $H(\omega) = |H(\omega)| \cdot e^{-i\varphi(\omega)}$ is linear phase, if $\varphi(\omega) = -\omega t_0$ for $t_0 \in \mathbf{R} \rightarrow$ sinusoids over all frequencies are delayed by the same amount t_0 (in reality not exactly possible)

- discrete-time domain:

$$\boxed{H(z)|_{z=e^{i\omega T_S}} = H(e^{i\omega T_S}) = |H(e^{i\omega T_S})| \cdot e^{i\varphi(\omega)}} \tag{5}$$

  is linear phase if $\varphi(\omega) = -\omega \cdot k \cdot \frac{T_S}{2}$ for $k \in \mathbb{N} \rightarrow$ can be realized with special FIR filter
  $\rightarrow$ yields conditions for c_n

Impulse responses for linear phase FIR filter of degree N :

1. h_n is symmetric to $\frac{N}{2}$, N even

$$\boxed{h_n = H_{N-n}} \qquad \forall \quad n \in \{0, 1, \ldots, N\} \tag{6}$$

$$\boxed{H(e^{i\omega T_s}) = e^{-i\omega T_s \frac{N}{2}} \underbrace{[h_{\frac{N}{2}} + 2 \sum_{m=1}^{\frac{N}{2}} h_{\frac{N}{2}-m} \cdot \cos(\omega T_s m)]}_{\text{real}, M(\omega)}} \tag{7}$$

$$\boxed{\varphi(\omega) = \text{arc}\{H(e^{i\omega T_s})\} = -\omega T_s \frac{N}{2} = \begin{cases} 0 & \text{for} \quad M(\omega) \geq 0 \\ \pi & \text{for} \quad M(\omega) < 0 \end{cases}} \tag{8}$$

   $\rightarrow$ linear phase, except for sign changes of $M(\omega)$

2. h_n is symmetric to $\frac{N}{2}$ N odd

$$\boxed{H(e^{i\omega T_s}) = e^{-i\omega T_s \frac{N}{2}} \cdot 2 \sum_{m=1}^{\frac{N+1}{2}} h_{\frac{N+1}{2}-m} \cdot \cos\left[(m-\tfrac{1}{2})\omega T_s\right]} \tag{9}$$

3. h_n is anti-symmetric to $\frac{N}{2}$, N even

$$\boxed{H(e^{i\omega T_s}) = e^{-i\omega T_s \frac{N}{2}} \cdot i \cdot 2 \underbrace{\sum_{m=1}^{\frac{N}{2}} h_{\frac{N}{2}-n} \sin(m\omega T_s)}_{M(\omega)}} \tag{10}$$

$$\boxed{\varphi(\omega) = \text{arc}\left\{H(e^{i\omega T_s})\right\} = -\omega T_s \frac{N}{2} + \frac{\pi}{2} + \begin{cases} 0 & \text{for} \quad M(\omega) \geq 0 \\ \pm\pi & \text{otherwise} \end{cases}} \tag{11}$$

4. h_n is anti-symmetric to $\frac{N}{2}$, N odd

$$\boxed{H(e^{i\omega T_s}) = e^{-i\omega T_s \frac{N}{2}} \cdot i \cdot 2 \sum_{m=1}^{\frac{N+1}{2}} h_{\frac{N+1}{2}-m} \sin[(n-\tfrac{1}{2})\omega T_s]} \tag{12}$$

Zero Location at FIR-Filters (→ Pole-Zero-Plot)

$$\boxed{H(z) = \sum_{n=0}^{N} h_n z^{-n}} \qquad \boxed{h_n = \pm h_{N-m}} \tag{13}$$

$$\boxed{H(z) = z^{-n} \cdot \sum_{n=0}^{N} h_n \cdot z^{N-n}} \tag{14}$$

$H(z)$ is polynomial in $z \rightarrow N$ zeros: z_ν

For z_ν not real but complex :
$\rightarrow \overline{z_\nu}$ (complex conjugated) is also a zero of $H(\omega)$ because h_n is real $\forall n$

<u>Consider:</u> $H(z^{-1})$

$$\boxed{H(z^{-1}) = \sum_{n=0}^{N} h_n \cdot z^n} \tag{15}$$

must have z_ν^{-1} as a zero and $\overline{z_\nu^{-1}} = (\overline{z}_\nu)^{-1}$ is also a zero

Substitution $m = N - n$
$$H(z^{-1}) = \sum_{m=0}^{N} h_{N-m} \cdot z^{N-m} = z^N \sum_{m=0}^{N} h_{N-m} \cdot z^{-m} = z^N \cdot \sum_{m=0}^{N} (\pm h_m) z^{-m} = z^N \cdot (\pm 1) \underbrace{\sum_{m=0}^{N} h_m z{-}m}_{H(z)}$$

$$\boxed{H(z^{-1}) = \pm z^N \cdot H(z)} \tag{16}$$

→ the zeros of $H(z)$ and $H(z^{-1})$ are the same → $\quad z_\nu, \overline{z}_\nu, z_\nu^{-1}, \overline{z}_\nu^{-1}$ are zeros of $H(z)$

Summary:

$$\left.\begin{array}{ll} \text{time domain} & h_n = \pm h_{N-n} \\ \text{z-domain (z-plain):} & \text{symmetry of zeros} \end{array}\right\} \rightarrow \text{both characterize linear phase FIR filter}$$

4.1.3 IIR-Filters

- Transfer function:

$$\boxed{H(z) = \frac{a_0+a_1\cdot z^{-1}+\dots+a_N\cdot z^{-N}}{1+b_1\cdot z^{-1}+\dots+b_N\cdot z^{-N}}} \tag{17}$$

  - degree N ; N poles , N zeros
  - we can split it up into transfer functions of degree ≤ 2
  - cascading filters of degree ≤ 2

$$\boxed{H(z) = \frac{a_0+a_1\cdot z^{-1}+a_2\cdot z^{-2}}{1+b_1\cdot z^{-1}+b_2\cdot z^{-2}}} \tag{18}$$

- Impulse response: $h_n \circ\!\!-\!\!\bullet \; H(z)$
  h_n has an infinite number of non-zero values $\rightarrow$ infinite impulse response (IIR)

- Difference equation for y_n:

$$\boxed{H(z) = \frac{Y(z)}{X(z)}} \tag{19}$$

$Y(z)\cdot(1+b_1\cdot z^{-1}+b_2\cdot z^{-2}) = X(z)\cdot(a_0+a_1\cdot z^{-1}+a_2\cdot z^{-2})\ Y(z) \;\bullet\!\!-\!\!\circ\; y_n$

$$\boxed{y_n = \underbrace{a_0\cdot z_n + a_1\cdot x_{n-1} + a_2\cdot x_{n-2}}_{\text{feed forward}} \quad \underbrace{-b_1\cdot y_{n-1} - b_2\cdot y_{n-2}}_{\text{feedback}}} \tag{20}$$

4.2 Transformation of Continuous-Time Filter into Discrete-Time Filter

Given: A continuous-time filter that satisfies the design specification

- given $h_0(t) \quad \rightarrow \quad h_n \quad \rightarrow \quad H(z)$
- given $H_0(\omega) \quad \rightarrow \quad H(z) \quad \rightarrow \quad h_n$

4.2.1 Invariant Impulse Response Method

Given: $h_0(t)$ impulse response of an analog (continuous-time) filter that satisfies the specification

Sampling $h_0(t)$ yields

$$\boxed{h_n = T_s\cdot h_0(n\cdot T_s)} \tag{21}$$

Transfer function:

$$\boxed{H(z)|_{z=e^{i\omega T_s}} = H(e^{i\omega T_s})} \tag{22}$$

System theory:

$$\boxed{h_0(t) = \sum_{n=-\infty}^{\infty} T_s\cdot h_0(n\cdot T_s)\cdot\delta(t-nT_s)} \tag{23}$$

$$\rightarrow \boxed{H(e^{-i\omega T_s}) \sum_{n=-\infty}^{\infty} H_0(i(\omega-n\omega_s))} \tag{24}$$

$$\boxed{h_n = T_s\cdot h_0(nT_s)} \tag{25}$$

$$\boxed{H(e^{-i\omega T_s}) = \sum_{n=-\infty}^{\infty} H_0(i(\omega-\omega_s))} \tag{26}$$

aliasing $\rightarrow \omega_s$ has to be chosen approximaly; not all $H_0(\omega)$ will be yield a working $H(e^{i\omega T_s})$

Realization of $h_n, H(z)$:
PFE of

$$\boxed{H_0(i\omega) = \sum_{k=1}^{M} \frac{A_k}{i\omega - p_k}} \tag{27}$$

with p_k pole of $H_0(p)$

$$\boxed{H_0(i\omega) \;\bullet\!\!-\!\!\circ\; h(t) = s(t)\cdot\sum_{k=1}^{M} A_k\cdot e^{p_x t}} \tag{28}$$

$$\rightarrow \boxed{h_n = T_s \cdot h_0(nT_s) = \begin{cases} T_s \cdot \sum\limits_{k=1}^{M} A_k \cdot e^{p_k n T_s} & n \geq 0 \\ 0 & n < 0 \end{cases}} \tag{29}$$

The transfer function

$$\boxed{\left.\begin{array}{ll} e^{p_x T_s n} & n \geq 0 \\ 0 & n < 0 \end{array}\right\} \circ\!\!-\!\!\bullet \quad \frac{1}{1-e^{p_x T_s} z^{-1}}} \tag{30}$$

$$\rightarrow \boxed{H(z) = \sum_{k=1}^{M} T_s \cdot A_k \cdot \frac{1}{1-e^{p_k T_s} z^{-1}}} \tag{31}$$

Note:

- aliasing has to be considered into the design
- only filters with finite bandwidth are possible
- IIR-filters

4.2.2 Bilinear Transform

Given: $H_0(i\omega) = H_0(p)|_{p=i\omega}$ transfer Funktion of a continuous-time filter

Apply bilinear transform to $H_0(p)$

$$\boxed{H(z) = H_0(p)\big|_{p=\omega_s \frac{1-z^{-1}}{1+z^{-1}}}} \tag{32}$$

Investigation of the conformal mapping $p \rightarrow z$

$$\boxed{p = \omega_s \cdot \frac{1-z^{-1}}{1+z^{-1}}} \tag{33}$$

$$\boxed{z = \frac{\omega_s + p}{\omega_s - p}} \tag{34}$$

$\rightarrow$ it maps the complex p-plane into the complex z-plane
p : continuous-time $p = \alpha + i\lambda$
z : discrete-time

$z = -\frac{p+\omega_s}{p-\omega_s} = -\frac{p-\omega_s+2\omega_s}{p-\omega_s} = -1 + \frac{2\omega_s}{\omega_s-\alpha-i\lambda}$

Summary:

- the whole imaginary axis in p gets mapped onto the unit circle in z:
  $p = i\lambda \rightarrow z = e^{i\omega T_s}$
- the left half-plane of p gets mapped into the inside of the unit circle in z:
  $p = i\lambda = \omega_s \cdot \frac{1-e^{-i\omega T_s}}{1+e^{-i\omega T_s}} = i\omega_s \cdot \frac{\sin(\omega \frac{T_s}{2})}{\cos(\omega \frac{T_S}{2})} = i\omega_s \cdot \tan(\omega \frac{T_s}{2})$

$$\boxed{\lambda = \omega_s \cdot \tan(\omega \tfrac{T_s}{2}) = \omega_s \tan(\pi \tfrac{\omega}{\omega_s})} \tag{35}$$

The frequency band $\lambda \in (-\infty; \infty)$ is mapped into the band $\omega \in \left[-\frac{\omega_s}{2}; \frac{\omega_s}{2}\right)$ and repeated with the period ω_s.

- (32) says that the amplitudes $H_0(p)$ and $H(z)$ are the same, only the frequency axis is distorted (35)
- no aliasing as in § 4.2.1

Filter design:

1. given: specification plot for $\left|H(e^{i\omega T_s})\right|^2$ with ω_{ST}, ω_P and the passband and stop band ripple
2. transfer the specs into the λ-domain with (35)
3. analog filter synthesis, e.g. by table $\rightarrow H_0(p)$
4. transfer $H_0(p) \rightarrow H(z)$ with (32) rational function $\rightarrow$ rational function
5. realization of $H(z)$ by first and second order section (IIR-filter)

4.3 Discrete-Time-Domain Filter Design

4.3.1 Window Method (FIR-Filter)

Start with an ideal filter and its transfer function

$$\boxed{H_{id}(e^{i\omega T_s} = \sum_{n=-\infty}^{\infty} \mathrm{rect}_{\omega_c}(\omega - \omega_s)} \tag{36}$$

ideal low pass with cutoff $\omega_c \leq \frac{\omega_s}{2}$
We know:

$$\boxed{H(z)|_{z=e^{i\omega T_s}} = H(e^{i\omega T_s}) = H_s(i\omega) = \sum_{n=-\infty}^{\infty} H_0(i(\omega - n\omega_s))} \tag{37}$$

$H_0(i\omega)$	...	spectrum of continuous-time $h(t)$
$H_s(i\omega)$	...	spectrum of sampled $T_s \cdot h(t)$ (discrete-time)

$$\boxed{h_n = T_s \cdot h(nT_s) = T_s \cdot \frac{1}{2\pi} \int_{-\infty}^{\infty} H_0(i\omega) \cdot e^{i\omega t} d\omega} \tag{38}$$

From (37) we can see that $H_0(i\omega) = H_s(\omega) \cdot \mathrm{rect}_{\frac{\omega_s}{2}}(\omega)$

$h_n = \frac{T_s}{2\pi} \int_{-\infty}^{\infty} \mathrm{rect}_{\frac{\omega_s}{2}} \cdot H_s(i\omega) \cdot e^{i\omega T_s} d\omega$

$h_n = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} H_s(i\omega) \cdot e^{i\omega T_s} d\omega$

$$\boxed{h_n = \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} H(e^{i\omega T_s}) \cdot e^{i\omega T_s} d\omega} \tag{39}$$

$$\boxed{h_n \quad \circ\!-\!\bullet \quad H(e^{i\omega T_s}) = \sum_{n=-\infty}^{\infty} h_n e^{-i\omega T_s} = H(z)|_{z=e^{i\omega T_s}}} \tag{40}$$

New: for ideal low pass

$$\boxed{h_{id,n} = T_s \cdot \frac{\omega_c}{\pi} \cdot \mathrm{sinc}(\omega_s n T_s) = 2\frac{\omega_c}{\omega_s}\mathrm{sinc}(2\pi n \frac{\omega_c}{\omega_s})} \tag{41}$$

This impulse response is infinite in both directions and non-causal

- cannot be realized
- windowing $\rightarrow$ finite length

i.e. multiplication with a window function w_n and shifting it to the right to make it causal

$$\boxed{h_n = W_{n-\frac{N}{2}} \cdot h_{id,n-\frac{N}{2}}} \qquad \forall n \in \{0, \ldots, N\} \text{ non} - \text{zero} \tag{42}$$

Note:

- N degree of the FIR-filter
- only causal filters are realized
- if $w_{-n} = \pm w_n$ (window function symmetric or anti-symmetric) then h_n is a linear phase FIR-filter
- effect of windowing on the transfer function:
  multiplication in the time-domain $\rightarrow$ convolution in the ω-domain
  $H(e^{i\omega T_s}) = e^{i\omega \frac{N}{2} T_s} \cdot \frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} W(e^{i(\omega - y)T_s}) \cdot H_{id}(e^{iyT_s}) \cdot dy$

$$\boxed{H(e^{i\omega T_s}) = \frac{e^{i\omega \frac{N}{2} T_s}}{\omega_s} \cdot \int_{-\omega_c}^{\omega_s} W(e^{i(\omega - y)T_s}) \cdot dy} \tag{43}$$

Common window functions:

1. Rectangular window

$$\boxed{w_{R,N} = \begin{cases} 1 & n = -\frac{N}{2}, \ldots, 0, \ldots, \frac{N}{2} \\ 0 & \text{otherwise} \end{cases}} \tag{44}$$

   with substitution $m = n + \frac{N}{2}$ and geometric series:

$$\boxed{W_{R,N}(e^{i\omega T_s}) = \frac{\sin\left(\frac{N-1}{2}\omega T_s\right)}{\sin\left(\frac{1}{2}\omega T_s\right)}} \tag{45}$$

   periodic in ω

   - nominator zeros
     $\omega = \nu \frac{\omega_s}{N-1}$
   - denominator zeros
     $\omega = \nu\omega_s$
     $\rightarrow$ the denominator zeros are all canceled by zeros in the numerator (at $\omega = \nu\omega_s$)
   - value at $\omega = 0$ (and $\omega = \nu\omega_s$)
     $\lim\limits_{\omega\to 0} \frac{\sin\left(\frac{N+1}{2}\omega T_s\right)}{\sin\left(\frac{1}{2}\omega T_s\right)} = N+1$ (L´Hospital)

   <u>Filter design:</u>

   (a) given: specification plot / tolerance shame
   (b) $\omega_c = \frac{1}{2}(\omega_P - \omega_{ST})$
       $N = 2 \cdot \frac{\omega_s}{\Delta\omega} - 1 = 2 \cdot \frac{\omega_S}{\omega_{ST}-\omega_P}$
   (c) problems: Gibbs´ phenomenon
       $\delta_P, \delta_{ST} \geq 0.16$
       solution:
       - other window function
       - smaller maximum $\rightarrow$ smaller δ_p, δ_{ST} possible

   Criteria for good filter design:

   - narrow main lobe $\rightarrow$ good $\Delta\omega$
   - small area of the side lobes (small ripple)

2. Triangular window (Bartlett window)

$$\boxed{w_{T,N} = \begin{cases} 1 - \frac{|n|}{\frac{N}{2}+1} & n = -\frac{N}{2}, \ldots, 0, \ldots, \frac{N}{2} \\ 0 & \text{otherwise} \end{cases}} \tag{46}$$

$$\boxed{W_{T,N}(e^{i\omega T_s}) = \frac{2}{N+2}\left(\frac{\sin\left(\frac{N+2}{4}\omega T_s\right)}{\sin\left(\frac{1}{2}\omega T_s\right)}\right)^2} \tag{47}$$

   - small side lobes $\rightarrow$ small ripples
   - wider main lobes $\rightarrow$ wider $\Delta\omega$ (bad)

3. Raised cosine (Henning, Hamming)

$$\boxed{w_{H,N} = \begin{cases} \alpha + (1-\alpha)\cos\left(2\pi\frac{n}{N+1}\right) & n = -\frac{N}{2}, \ldots, 0, \ldots, \frac{N}{2} \\ 0 & \text{otherwise} \end{cases}} \tag{48}$$

$$\boxed{\alpha = \left\{ \begin{array}{ll} 0.5 & \text{Henning} \\ 0.54 & \text{Hamming} \end{array} \right\} \text{window}} \tag{49}$$

   smooth towards the end-points $\rightarrow$ small side lobes

4. Kaiser window

$$\boxed{w_{K,N} = \begin{cases} \frac{1}{\mathrm{I}_0(\alpha)}\mathrm{I}_0\left(\alpha\sqrt{1-\left(\frac{n}{N/2}\right)^2}\right) & n = -\frac{N}{2},\ldots,0,\ldots,\frac{N}{2} \\ 0 & \text{otherwise} \end{cases}} \tag{50}$$

I_0 ... modified zeroth-order Bessel function of first kind
$\alpha \geq 0$ ... parameter to adjust the window $\rightarrow$ it determines the ripple for a given N

Filter design:

(a) given: specification plot / tolerance shame
(b)
- choose $\min(\delta_P, \delta_{ST}) = \delta$
- $\Delta\omega = \omega_{ST} - \omega_P$
- numerical define α, N

$$\boxed{\alpha = \begin{cases} 0.1102 \cdot (A - 8.7) & A > 50 \\ 0.5842 \cdot (A-21)^{0.4} + 0.07886 \cdot (A - 21) & 21 < A < 50 \\ 0 & \text{otherwise} \end{cases}} \tag{51}$$

$$\boxed{N = \frac{A - 7.95}{14.6 \cdot \frac{\Delta\omega}{\omega_s}}} \tag{52}$$

N must be integer
or with our tables :$N = 4\mathrm{Hz} \cdot \frac{\omega_s}{\Delta\omega}$

(c)
- determine $w_{W,N}$ with (50) and table
- compute $\omega_c = \frac{1}{2}(\omega_P - \omega_{ST}) \quad \rightarrow \quad h_{id,n}$ with (41)
- compute $h_n = w_{K,n-\frac{N}{2}} \cdot h_{id,n-\frac{N}{2}}$ with (42)

$\rightarrow$ FIR filter

4.3.2 Remez Multiple Exchange Algorithm

$$\boxed{H(\Omega) = F(\Omega) \cdot G(\Omega)} \tag{53}$$

$$\boxed{G(\Omega) = \sum_{n=0}^{M} a_n \cos(n\Omega)} \tag{54}$$

- $F(\Omega)$ is fixed for given filter type and given N
- $G(\Omega)$ has all the $(M+1)$ available parameters
- find the optimal $G(\Omega)$ which $H(\Omega)$ follows

Approximation problem:
find the original $G(\Omega)$ to minimize the error ϵ

$$\boxed{\epsilon = \max_{\Omega \in X} |E(\Omega)|} \tag{55}$$

X ... closed subset of $[0, \pi]$

with: $E(\Omega) - W(\Omega)[G(\Omega) - D(\Omega)]$ (see hand out)

Altnation theorem:
$G(\Omega)$ in (53) is the unique best solution to minimize ϵ, if and only if there exists at least $M+2$ points $\Omega_0, \ldots, \Omega_{M+1}$

$$\boxed{\Omega_0 < \Omega_1 < \ldots < \Omega_{M+1}} \tag{56}$$

$$\boxed{\left.\begin{array}{l} E(\Omega_{k+1}) = -E(\Omega_k) \\ |E(\Omega_k)| = \epsilon \end{array}\right\} \forall k = 0, 1, \ldots, M+1} \tag{57}$$

The error values of each of the $M+2$ points alternate between $\pm\epsilon$ i.e. they reach the maximum exactly

$$\boxed{E(\Omega_k) = W(\Omega_k)[G(\Omega_k - D(\Omega_k))] = (-1)^k \cdot \epsilon} \tag{58}$$

There are $M+2$ linear equations in a_n and ϵ; $M+2$ free parameters $\rightarrow$ unique solution.

4.4 Frequency Transformation Low pass

1. Analog frequency transformation
   analog low pass $\rightarrow$ analog filter type $\rightarrow$ analog-digital-transformation

2. Digital frequency transformation
   digital low pass $\rightarrow$ digital filter type
   general transformation of $z^{-1} \rightarrow \tilde{z}^{-1} = f(z)$

$$\boxed{\tilde{z}^{-1} = f(z) = \pm \prod_{k=1}^{K} \frac{z^{-1} - \alpha k}{1 - \alpha_k z^{-1}}} \tag{59}$$

Each term in (59) does a conformal mapping on the unit circle:

$$\boxed{z = e^{i\omega T_s} \qquad \rightarrow \qquad \tilde{z} = e^{i\tilde{\omega} T_s}} \tag{60}$$

$$\left.\frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}\right|_{z = e^{i\omega T_s}} = \frac{(1-\alpha)\cos(\omega\frac{T_s}{2}) - i(1+\alpha)\sin(\omega\frac{T_s}{2})}{(1-\alpha)\cos(\omega\frac{T_s}{2}) + i(1+\alpha)\sin(\omega\frac{T_s}{2})} = \frac{\underline{z}_0}{\bar{z}_0} = e^{2i\varphi}$$

$$\boxed{\tilde{z}^{-1} = e^{-i\tilde{\omega} T_s} = e^{2i\varphi}} \tag{61}$$

$$\tilde{\omega} T_s = -2 \cdot \mathrm{arc}(\underline{z}_0) = 2 \cdot \arctan\left\{\frac{(1+\alpha)\sin(\omega\frac{T_s}{2})}{(1-\alpha)\cos(\omega\frac{T_s}{2})}\right\}$$

$$\boxed{\tilde{\omega} = \frac{2}{T_s} \cdot \arctan\left\{\frac{1+\alpha}{1-\alpha}\tan\left(\omega\frac{T_s}{2}\right)\right\}} \tag{62}$$

$$\boxed{\alpha = \frac{\sin\left((\tilde{\omega} - \omega)\frac{T_s}{2}\right)}{\sin\left((\tilde{\omega} + \omega)\frac{T_s}{2}\right)}} \tag{63}$$

5 Finite World length Effects in Digital Systems

Given: e.g. fixed point format for coefficients and signals

a_0	a_1	a_2	$\cdots$	a_n

N-bit number

$\rightarrow$ LSB (smallest possible number): $2^{-(N-1)}$
$\rightarrow$ finite precision; quantization

5.1 Pole Grid (Possible Pole Location)

The coefficient quantization causes the poles and zeros to be quantized as well; i.e. they can only be related in discrete steps. These possible locations in the z-plane are called the **pole grid** (zero grid). The pole grid depends very much on the actual filter structure that is used to implement $H(z)$.
$\rightarrow$ different pole grids for different structures
$\rightarrow$ example in exercise 9

5.2 Overflow-Scaling

Problem: overflow number format is not sufficient for multiplication and adding results

Given:

$$\boxed{|x_n| \leq 1} \tag{1}$$

input signal is bound

Requirement: overflow may *not* occur on any inner node and on the output

$$\boxed{|y_{i,n} \leq 1|} \quad \text{and} \quad \boxed{|y_n| \leq 1} \tag{2}$$

output signals are bound
(2) $\rightarrow$ with discrete condition:

$$\boxed{|y_{i,n}| = S \cdot \left|\sum_{k=0}^{\infty} h_{i,k} \cdot x_{n-k}\right|} \tag{3}$$

scaling multiplier is in general necessary to prevent overflow!

$$\boxed{|y_n| = \left|\sum_{k=0}^{\infty} h_k \cdot x_{n-k}\right| \leq 1} \tag{4}$$

this must be scaled already from the filter design process.
Fixed-point arithmetic with normal format:

- all signal and multipliers values are $\in [-1, 1-$ LSB$]$ with two´s-complement-coding
- all intermediate results may have overflows without affecting the end result see cap. 1.1.2
- only input nodes to non-linear multipliers have to be protected

5.2.1 Sum Scaling - Absolute Scaling

$$|y_{i,n}| = S_i \cdot \left|\sum_{k=0}^{\infty} h_{i,k} \cdot x_{n-k}\right| \leq S_i \cdot \sum_{k=0}^{\infty} |h_{i,k}| \cdot \underbrace{|x_{n-k}|}_{\leq 1} \leq S_i \sum_{k=0}^{\infty} |h_{i,k}| \leq 1$$

$$\boxed{S_i \leq \frac{1}{\sum\limits_{k=0}^{\infty} |h_{i,k}|}} \tag{5}$$

Note: choose largest S_i to return the largest SNR! Thus S_i is the required scaling factor for node i, the scaling for the whole filter has to use the smallest S_i of all the nodes.

$$\boxed{S \leq \min S_i = \min \frac{1}{\sum\limits_{k=0}^{\infty} |h_{i,k}|}} \tag{6}$$

$\rightarrow$ overflow protection for all inner nodes is guaranted!

5.2.2 L_∞ – Norm – Scaling

Def: L_P-Norm of $F(e^{i\omega T_s})$

$$\boxed{\left|\left|F(e^{i\omega T_s})\right|\right|_P = \left[\frac{1}{\omega_s}\int_0^{\infty} \left|F(e^{i\omega T_s})\right|^P d\omega\right]^{\frac{1}{P}}} \tag{7}$$

Inverse FT for discrete-time signals

$$\boxed{y_{i,n} = \frac{1}{\omega_s}\int_0^{\omega_s} Y_i(e^{i\omega T_s}) \cdot e^{in\omega T_s} d\omega} \tag{8}$$

$$\begin{aligned}
|Y_{i,n}| &= S_i \cdot \frac{1}{\omega_s}\left|\int_0^{\omega_s} H_i(e^{i\omega T_s}) \cdot X(e^{i\omega T_s}) d\omega\right| \\
&\leq S_i \cdot \frac{1}{\omega_s}\int_0^{\omega_s} \left|H_i(e^{i\omega T_s})\right| \cdot \left|X(e^{i\omega T_s})\right| d\omega \\
&\leq \underbrace{S_i \max_{\omega} \left|H_i(e^{i\omega T_s})\right|}_{||H_i(e^{i\omega T_s}||_\infty} \cdot \underbrace{\frac{1}{\omega_s}\int_0^{\omega_s} \left|X(e^{i\omega T_s})\right| d\omega}_{||X(e^{i\omega T_s})||_1}
\end{aligned}$$

$$\boxed{S_i \leq \frac{1}{||H_i(e^{i\omega T_s}||_\infty \cdot ||X(e^{i\omega T_s})||_1}} \tag{9}$$

Now for the input signal that fulfills $\left|\left|X(e^{i\omega T_s})\right|\right|_1 \leq 1$ we get

$$\boxed{S \leq \frac{1}{||H_i(e^{i\omega T_s}||_\infty}} \tag{10}$$

Note: sinusoidal signals with amplitudes of ≤ 1 meet the condition $\left|\left|X(e^{i\omega T_s})\right|\right|_1 \leq 1$
$\rightarrow$ "L_∞-Norm-Scaling protects sinusoidal signals "

5.2.3 L_2 – Norm – Scaling

From (8) follows

$$
\begin{aligned}
|Y_{i,n}|^2 &= S_i^2 \cdot \frac{1}{\omega_s}\left|\int_0^{\omega_s} X(e^{i\omega T_s}) \cdot H_i(e^{i\omega T_s}) \cdot e^{in\omega T_s} \cdot d\omega\right| \\
&\leq S_i^2 \underbrace{\frac{1}{\omega_s}\int_0^{\omega_s} |X(e^{i\omega T_s})|^2 d\omega}_{\left|\left|X(e^{i\omega T_S})\right|\right|_2^2} \underbrace{\frac{1}{\omega_s}\int_0^{\omega_s} |H_i(e^{i\omega T_s})|^2 d\omega}_{\left|\left|H_i(e^{i\omega T_S})\right|\right|_2^2}
\end{aligned}
$$

$$\boxed{S_i \leq \frac{1}{\left|\left|X(e^{i\omega T_S})\right|\right|_2 \cdot \left|\left|H_i(e^{i\omega T_S})\right|\right|_2}} \tag{11}$$

Now, the specific input signals:

$$\boxed{\left|\left|X(e^{i\omega T_S})\right|\right|_2^2 = \sum_{n=-\infty}^{\infty} |x_n|^2} \tag{12}$$

For signals with banded signal energy

$$\boxed{\left|\left|X(e^{i\omega T_S})\right|\right|_2^2 \leq 1} \tag{13}$$

the L_2-norm-scaling

$$\boxed{S_i \leq \frac{1}{\left|\left|H_i(e^{i\omega T_S})\right|\right|_2}} \tag{14}$$

Note: direct from the unit-pulse response

$$\boxed{\left|\left|H_i(e^{i\omega T_S})\right|\right|_2 = \sqrt{\sum_0^{\infty} |h_{i,n}|^2}} \tag{15}$$

L_2-norm-scaling protects from overflow for bounded signal energy signals.

In general:

$$\boxed{0 < S_{\text{abs}} \leq S_\infty \leq S_2 \leq 1} \tag{16}$$

5.3 Limit Signals

Definition: A non-zero output signal that does not vanish for $n \to \infty$ when the input is $x_n = 0\, \forall n \geq n_0$, is called limit cycle.

Note: In digital systems, limit cycles are periodic. Linear systems do not have limit cycles.
$\rightarrow$ non-linearity is necessary, quantization rounding usual: rounding/truncating after multiplication is a type of quantization and the cause for limit cycles.

5.4 Quantization Noise

5.4.1 Discrete-Time Random Process

- auto-correlation

$$\boxed{R_{xx}(n) = \lim_{k\to\infty} \frac{1}{2k+1} \sum_{l=-k}^{k} x_l \cdot x_{n+l}} \tag{17}$$

- effective value

$$\boxed{x_{eff}^2 = \overline{x_n}^2 = R_{xx}(0)} \tag{18}$$

- power spectral density

$$\boxed{S_{xx}(z) = \sum_{n=-\infty}^{\infty} R_{xx}(n) \cdot z^{-n}} \tag{19}$$

- output signal

$$\boxed{y_n = \sum_{l=-\infty}^{\infty} h_l \cdot x_{n-l}} \tag{20}$$

and

$$\boxed{R_{yy}(n) = \sum_{l=-\infty}^{\infty} h_l \cdot \sum_{m=-\infty}^{\infty} h_m \cdot R_{xx}(n+l-m)} \tag{21}$$

$$\boxed{R_{yy}(n) \quad \circ\!\!\stackrel{\mathrm{Z}}{\longrightarrow}\!\!\bullet \quad S_{yy}(z) = H(z^{-1}) \cdot H(z) \cdot S_{xx}(z)} \tag{22}$$

on the unit-circle

$$\boxed{S_{yy}(e^{i\omega T_s}) = \left|H(e^{i\omega T_s})\right|^2 S_{xx}(e^{i\omega T_s})} \tag{23}$$

$$\boxed{R_{yy}(n) = \frac{1}{\omega_s} \int_0^{\omega_s} \left|H(e^{i\omega T_s})\right|^2 \cdot S_{xx}(e^{i\omega T_s}) e^{i\omega T_s}} \tag{24}$$

for $n = 0$

$$\boxed{y_{eff}^2 = \overline{y_n}^2 = R_{yy}(0) = \frac{1}{\omega_s} \int_0^{\omega_s} \left|H(e^{i\omega T_s})\right|^2 S_{xx}(e^{i\omega T_s}) d\omega} \tag{25}$$

5.4.2 Application to Quantization Noise

e.g rounding after multiplication

- quantization error $e_n \in [-\frac{q}{2}; \frac{q}{2}]$ uniformly distributed
- ergodic proceed; here:

$$\boxed{R_{ee}(0) = e_{\text{eff}}^2 = \frac{1}{q} \int_{-\frac{q}{2}}^{\frac{q}{2}} e_n^2 de_n = \frac{q^2}{12}} \tag{26}$$

$\rightarrow$ Assumption: e_n is white noise:

$$\boxed{R_{ee}(n) = \begin{cases} \frac{q^2}{12} & n = 0 \\ 0 & \text{otherwise} \end{cases}} \tag{27}$$

$$\boxed{S_{ee}(z) = R_{ee}(0) = \frac{q^2}{12}} \tag{28}$$

$$\begin{aligned} R_{yy} &= \sum_{l=-\infty}^{\infty} h_l \cdot \sum_{m=-\infty}^{\infty} h_m \cdot R_{ee}(l-m) \\ &= R_{ee}(0) \sum_{l=-\infty}^{\infty} h_l^2 \\ &= \frac{q^2}{12} \sum_{n=-\infty}^{\infty} h_n^2 \end{aligned}$$

$$\boxed{R_{yy}(0) = \frac{q^2}{12} \frac{1}{\omega_s} \int_0^{\omega_s} |H(e^{i\omega T_S})|^2 d\omega = \frac{q^2}{12} ||H(e^{i\omega T_s})||_2^2} \tag{29}$$

- Now for all inner nodes where quantization noise occurs:
  assumption: all noise sources are independent

$$\boxed{y_{\text{eff,total}} = \frac{q^2}{12} \sum_{j=1}^{k} ||H_j(e^{i\omega T_s})||_2^2 = \frac{q^2}{12} \sum_{i=1}^{k} \sum_{n=-\infty}^{\infty} h_{i,n}^2} \tag{30}$$

6 Discrete and Fast Fourier Transformation

6.1 Discrete Fourier Transformation

- x_n: sequence of length $N : n = 0, 1, \ldots, N-1$ (sometimes also considered periodic with N)

$$\boxed{\mathcal{X}(z) = \sum_{n=0}^{N-1} x_n \cdot z^{-n}} \tag{1}$$

- evaluating this only at N discrete, equally spaced points on the unit circle $\quad \forall m = 1, 2, \ldots, N-1$

$$\boxed{z_m = e^{i \cdot m \cdot \frac{2\pi}{N}}} \tag{2}$$

$$\boxed{\mathcal{X}(m \cdot z) = \sum_{n=0}^{N-1} x_n \cdot e^{-i \cdot m \frac{2\pi}{N} := \mathcal{X}_m}} \tag{3}$$

$\mathcal{X}_m$ is called the DFT of $x_n \quad \forall m = 0, 1, \ldots, N-1$

- the DFT is a transformation (linear) of N vectors into N vectors
- twiddle factor

$$\boxed{w_n = e^{-i \frac{2\pi}{N}}} \tag{4}$$

$\forall m = 0, 1, \ldots, N-1$

$$\boxed{\mathcal{X}_m = \sum_{n=0}^{N-1} x_n \cdot w_n^{m \cdot n}} \tag{5}$$

- find the inverse by building

$$\begin{aligned}
\sum_{m=0}^{N-1} \mathcal{X}_m \cdot e^{imn\frac{2\pi}{N}} &= \sum_{k=0}^{N-1} x_k \cdot \sum_{m=0}^{N-1} e^{im\frac{2\pi}{N}(n-k)} \\
&= x_n \cdot N \\
\Rightarrow x_n &= \frac{1}{N} \sum_{m=0}^{N-1} \mathcal{X}_m \cdot e^{imn\frac{2\pi}{N}}
\end{aligned}$$

$$\boxed{x_n = \frac{1}{N} \sum_{m=0}^{N-1} \mathcal{X}_m \cdot w_n^{-mn}} \tag{6}$$

is called the IDFT

- corresponding frequency samples

$$z_m = e^{im\frac{2\pi}{N}} \overset{!}{=} e^{i2\pi\frac{\omega_m}{\omega_s}}$$

$$\boxed{\omega_m = m \cdot \frac{\omega_s}{N} = \frac{m}{N} \cdot \omega_s} \tag{7}$$

- (5) and (6) build the DFT, IDFT pair: both are discrete in time <u>and</u> frequency
- properties:

  1. $\mathcal{X}_m$ is periodic in N

$$\begin{aligned}
\mathcal{X}_{m+\nu N} &= \sum_{n=0}^{N-1} x_n \cdot w_N^{n(m+\nu N)} \\
&= \sum_{n=0}^{N-1} x_n \cdot w_N^{nm} \cdot w_N^{\nu \cdot nN} \\
&= \sum_{n=0}^{N-1} x_n \cdot w_N^{mn} \\
&= \mathcal{X}_m
\end{aligned}$$

2. so far, only values of $n = 0, 1, \ldots, N-1$ were considered; now compute:

$$\begin{aligned} x_{n+\nu N} &= \frac{1}{N} \sum_{m=0}^{N-1} \mathcal{X}_m w_N^{-m(n+\nu N)} \\ &= \frac{1}{N} \sum_{m=0}^{N-1} \mathcal{X}_m \cdot w : N^{-mn} \cdot w_N^{-\nu \cdot mN} \\ &= x_n \end{aligned}$$

$\Rightarrow$ the discrete-time signal is extended periodically with period N outside of our given range

3. IDFT with the help of DFT

$$\begin{aligned} \overline{x}_n &= \frac{1}{N} \sum_{m=0}^{N-1} \overline{\mathcal{X}}_m \cdot \overline{w}_N^{-mn} \\ &= \frac{1}{N} \sum_{m=0}^{N-1} \overline{\mathcal{X}}_m \cdot w_N^{mn} \end{aligned}$$

$$\boxed{x_n = \overline{\left[\tfrac{1}{N} \cdot \mathrm{DFT}\{\overline{\mathcal{X}}_m\}\right]}} \qquad (8)$$

4. Complexity
   - additions: $N(N-1)$ of additions
   - multiplications: $N \cdot N = N^2$ of multiplication

6.2 Fast Fourier Transformation

<u>here:</u> radix-2 FFT: $N = 2^\nu$, $\quad \nu \in \mathbb{N}$

- Method to reduce the number of multiplications ("divide and conquer")
  - split the N-point DFT into two $\frac{N}{2}$-point DFTs
  - combine the two correctly with "Butterfly" for radix-2 FFTs
  - repeat to split the $\frac{N}{2}$-point DFTs again and again
- mathematically (N even)
  - $\forall m = 0, 1, \ldots, N-1$

$$\boxed{F_m = \sum_{n=0}^{N-1} f_m \cdot w_N^{mn}} \qquad (9)$$

  - split into even and odd parts of n

$$\begin{aligned} F_m &= \underbrace{\sum_{n=0}^{\frac{N}{2}-1} f_{2n} \cdot w_N^{m2n}}_{\text{even}} + \underbrace{\sum_{n=0}^{\frac{N}{2}-1} f_{2n+1} w_N^{m(2n+1)}}_{\text{odd}} \\ &= \underbrace{\sum_{n=0}^{\frac{N}{2}-1} f_{2n} w_{\frac{N}{2}}^{mn}}_{\text{DFT of } \frac{N}{2}} + w_N^m \underbrace{\sum_{n=0}^{\frac{N}{2}-1} f_{2n+1} w_{\frac{N}{2}}^{mn}}_{\text{DFT of } \frac{N}{2}} \end{aligned}$$

  - split into first and second part of m:

$$\begin{aligned} F_m &= \sum_{n=0}^{\frac{N}{2}-1} f_{2n} \cdot w_{\frac{N}{2}}^{mn} + w_N^m \sum_{n=0}^{\frac{N}{2}-1} f_{2n+1} \cdot w_{\frac{N}{2}}^{mn} \\ F_{m+\frac{N}{2}} &= \sum_{n=0}^{\frac{N}{2}-1} f_{2n} \cdot w_{\frac{N}{2}}^{n(m+\frac{N}{2})} + w_N^{m+\frac{N}{2}} \sum_{n=0}^{\frac{N}{2}-1} f_{2n+1} \cdot w_{\frac{N}{2}}^{n(m+\frac{N}{2})} \end{aligned}$$

with $w_{\frac{N}{2}}^{n(m+\frac{N}{2})} = w_{\frac{N}{2}}^{m}$ and $w_N^{m+\frac{N}{2}} = -w_N^m$

$$F_m = \sum_{n=0}^{\frac{N}{2}-1} f_{2n} \cdot w_{\frac{N}{2}}^{mn} + w_N^m \sum_{n=0}^{\frac{N}{2}-1} f_{2n+1} \cdot w_{\frac{N}{2}}^{mn}$$

$$F_{m+\frac{N}{2}} = \sum_{n=0}^{\frac{N}{2}-1} f_{2n} \cdot w_{\frac{N}{2}}^{mn} - w_N^m \sum_{n=0}^{\frac{N}{2}-1} f_{2n+1} \cdot w_{\frac{N}{2}}^{mn}$$

$$\boxed{F_m = A_m + w_N^m \cdot B_m} \tag{10}$$

$$\boxed{F_{m+\frac{N}{2}} = A_m - w_N^m \cdot B_m} \tag{11}$$

A_m and B_m are $\frac{N}{2}$-point DFTs of a disjoint set of f_n

A_m : only even $f_n \to f_{2n}$

B_m : only odd $f_n \to f_{2n+1}$

- Complexity of FFTs
  - additions:
    $\frac{2}{\text{butterfly}} \cdot \frac{N}{2} \frac{\text{butterflies}}{\text{split}} \cdot \ln(N)\,\text{splits} = N \cdot \ln N$
  - multiplications:
    $\frac{1}{\text{butterfly}} \cdot \frac{N}{2} \frac{\text{butterflies}}{\text{split}} \cdot \ln(N)\,\text{splits} = \frac{N}{2} \ln N$

  $\to$ much better (faster) than DFT ($\sim N^2$)
- Note:
  - we have shown the radix-2-FFT in decimation in frequency $\to$ DIT-radix-2-FFT
  - reverse architecture decimation in frequency $\to$ DIF-radix-2-FFT
  - other types: split-radix, Winograd are even better

6.3 DFT of two real series in on step

Given: u_n, v_n real series $\to \mathcal{U}_m, \mathcal{V}_m$

Faster method than just two N-point FFTs

$$\boxed{x_n = u_n + i \cdot v_n} \tag{12}$$

one signal N-point FFT $\to \mathcal{X}_m$

7 Discrete Convolution and Correlation

7.1 Linear discrete Convolution

Given:

- x_n, h_n for $n \geq 0$ (0 otherwise)
- for $n \geq 0$

$$\boxed{y_n = x_n * h_n = \sum_{k=0}^{\infty} x_k \cdot h_{n-k}} \tag{1}$$

- convolution theorem

$$\boxed{Y(z) = X(z) \cdot H(z) = \sum_{n=0}^{\infty} x_n \cdot z^{-1} \cdot \sum_{n=0}^{\infty} h_n \cdot z^{-1}} \tag{2}$$

- numerical complexity
  - multiplications:
    $L_x \cdot L_h$ (each value of x_n with each value of h_n)
  - additions:
    $L_x \cdot L_h - \underbrace{(L_x + L_h - 1)}_{L_y}$ $\qquad [-(L_x - 1)(L_h - 1)]$

7.2 Cyclic discrete Convolution

Given: x_n, h_n periodic with N (compare IDFT) unique in $n = 0, 1, \dots, N-1$

- periodic with N

$$\boxed{y_n = x_n \otimes h_n = \sum_{k=0}^{N-1} x_k \cdot h_{n-k} = y_{n+\nu N}} \tag{3}$$

- convolution theorem

$$\boxed{Y_m = X_m \cdot H_n = \sum_{n=0}^{N-1} x_n \cdot w_N^{mn} \cdot \sum_{n=0}^{N-1} h_n \cdot w_N^{mn}} \tag{4}$$

- numerical complexity
  - multiplication:
    $N \cdot N = N^2$
  - additions:
    $(N-1) \cdot N = N^2 - N$
- fast convolution $\rightarrow$ FFT-convolution
  1. FFT of x_n and $h_n \rightarrow \mathcal{X}_m, \mathcal{H}_m$
     multiplications: $N \cdot \ln N$
     additions: $2N \cdot \ln N$
  2. multiplications in frequency domain:

     $$\boxed{\mathcal{Y}_m = \mathcal{X}_m \cdot \mathcal{Y}_m} \tag{5}$$

     multiplications: N
     additions: 0
  3. IFFT of $\mathcal{Y}_m \rightarrow y_n$

     $$\boxed{\overline{y}_n = \frac{1}{N} \cdot \sum_{m=0}^{N-1} \overline{\mathcal{X}_m \cdot \mathcal{H}_m} \cdot w} \tag{6}$$

     $\frac{N}{2} \operatorname{ld} N$ multiplications and $N \cdot \operatorname{ld} N$ additions
- compute the linear discrete convolution with a cyclic convolution for fixed x_n and h_n (goal: fast convolution)
  1. shift both x_n, h_n so, that their non-zero values are in range $\{0, 1, \dots, \begin{matrix} L_x - 1 \\ L_h - 1 \end{matrix}\}$
  2. choose N large enough to prevent aliasing of y_n

     $$\boxed{N \geq L_y = L_x + L_h - 1} \tag{7}$$

     (usually the next larger power of 2) reason:
     - all 3 sequences have to fit in N
     - y_n is always the largest
  3. extend x_n, h_n to $N-1$ with zeros (make them N large)
  4. cyclic convolution with (3) or even the fast convolution $\rightarrow \tilde{y}_n$
  5. take the values $n \in \{0, 1, \dots, N-1\}$ (everything else is 0)and back shift by the sum of both shifts from step 1 to get y_n aligned correctly