The constant K is introduced in order to scale the maximum of $|H(e^{j\omega})|$ into unity. Using Eq. 7.7 in Mitra ($\omega_c = \Omega_c/f_T = 2\pi f_c/f_T$) and values $f_T = 1$ kHz (sampling frequency) and $f_c = 100$ Hz (cut-off frequency),

$$H(z)_{Imp} = \frac{K}{1 - e^{-\omega_c} z^{-1}} = \frac{K}{1 - e^{-\pi/5} z^{-1}}$$

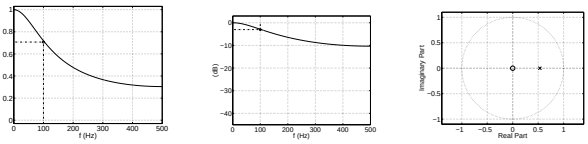
We also know that the maximum is located at zero frequency, because the frequency response of a Butterworth filter is monotonic. Thus we get

$$\frac{K}{1 - e^{-\pi/5}} = 1 \Leftrightarrow K = 1 - e^{-\pi/5}$$

The transfer function of the filter is therefore

$$H(z)_{Imp} = \frac{1 - e^{-\pi/5}}{1 - e^{-\pi/5} z^{-1}} = 0.4665 \cdot \frac{1}{1 - 0.5335 z^{-1}}$$

There is a pole at $z = 0.5335$, see Figure 91 for the amplitude response in linear scale, in desibels and the pole-zero plot.



(a) (b) (c)

Figure 91: Problem 44, the filter $H_{Imp}(z)$ using impulse-invariant method. (a) Amplitude response in linear scale $|H(e^{j\omega})|$ and (b) in desibels $10 \cdot \log_{10}|H(e^{j\omega})|^2$, (c) pole-zero diagram.

c) Transfer function using bilinear transform. Compute the normalized angular discrete-time cut-off frequency ω_c,

$$\omega_c = \frac{2\pi\Omega_c}{\Omega_s} = \frac{2\pi 2\pi f_c}{2\pi f_T} = \frac{2\pi f_c}{f_T} = 0.2\pi$$

and the prewarped cut-off frequency Ω_{pc} ($k = 2/T$):

$$\Omega_{pc} = k \cdot \tan(\frac{\omega_c}{2}) = k \cdot \tan(0.1\pi)$$

The digital filter is obtained through bilinear transform:

$$\begin{aligned} H(z) &= H(s)\big|_{s=k\cdot\frac{1-z^{-1}}{1+z^{-1}},\ \ \Omega_c=\Omega_{pc}=k\cdot\tan(0.1\pi)} \\ &= \frac{\Omega_c}{s+\Omega_c}\Big|_{s=k\cdot\frac{1-z^{-1}}{1+z^{-1}},\ \ \Omega_c=\Omega_{pc}=k\cdot\tan(0.1\pi)} \\ &= \frac{k\cdot\tan(0.1\pi)}{k\cdot\frac{1-z^{-1}}{1+z^{-1}} + k\cdot\tan(0.1\pi)} \qquad \big| \not{k} \\ &= \frac{\tan(0.1\pi)(1+z^{-1})}{(1+\tan(0.1\pi)) - (1-\tan(0.1\pi))z^{-1}} \end{aligned}$$

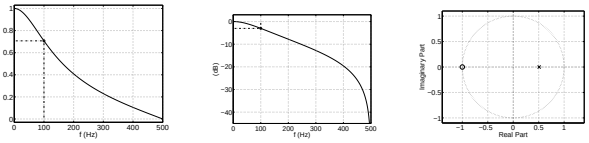
The last task is to normalize the transfer function. The constant term in denominator polynomial should be scaled to 1, and the maximum value of the amplitude response to 1. While this is a Butterworth lowpass filter, the maximum is reached at $\omega = 0$, i.e., $z = e^{j\omega}|_{\omega=0} = 1$.

$$|H(z)_{Bil}|_{max} = \Big|K \cdot \frac{1 + z^{-1}}{1 - \frac{1-\tan(0.1\pi)}{1+\tan(0.1\pi)} z^{-1}}\Big|_{z=1} = 1$$

Finally,

$$H_{Bil}(z) = 0.2452 \cdot \frac{1 + z^{-1}}{1 - 0.5095 z^{-1}}$$

There is a zero at $z = -1$ and a pole at $z = 0.5095$. See Figure 92 for the amplitude response in linear scale, in (power) desibels (`20*log10(A)=10*log10(A^2)`), and the pole-zero plot. Compare also to the filter obtained through the impulse-invariant method in Figure 91.



(a) (b) (c)

Figure 92: Problem 44, the filter $H_{Bil}(z)$ using bilinear transform. (a) Amplitude response in linear scale and (b) in desibels, (c) pole-zero diagram.

45. **Problem:** Use windowed Fourier series method and design a FIR-type (causal) lowpass filter with cutoff frequency $3\pi/4$. Let the order of the filter be 4.

a) Use the rectangular window of length 5.

b) Use the Hamming window of length 5.

c) Compare how the amplitude responses of the filters designed in (a) and (b) differ assuming that the window size is high enough (e.g. $M = 50$).

Solution: Digital FIR filter design with windowed (truncated) Fourier series method. The idea is to find infinite-length impulse response of the ideal filter and truncate it so that a realizable finite-length filter is obtained.

$$h_t[n] = h_d[n] \cdot w[n] \quad \leftrightarrow \quad H_t(z) = H_d(z) \circledast W(z)$$

Now, when cut-off frequency (-3 dB) is at $\omega_c = 3\pi/4$, the infinite-length impulse response of the ideal filter is:

$$h_d[n] = \sin(\frac{3\pi}{4}n)/(\pi n) = (3/4)\,\mathrm{sinc}(\frac{3}{4}n)$$

When computing values, $\sin(x)/x \to 1$, when $x \to 0$, or $\mathrm{sinc}(x) \to 1$, when $x \to 0$. So, we get $h_d[n] = \{\ldots, -0.1592, 0.2251, \underline{0.75}, 0.2251, -0.1592, \ldots\}$.

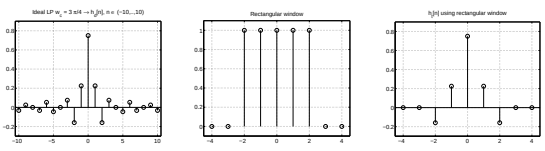
a) Now we are using rectangular window $w_r[n]$ of length 5 (4th order),

$$w_r[n] = \begin{cases} 1, & -2 \le n \le 2 \\ 0, & \text{otherwise} \end{cases}$$

Hence,

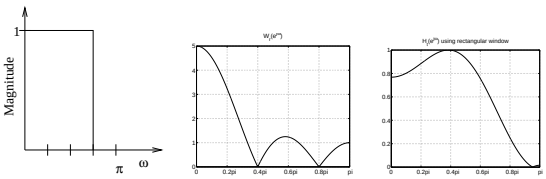
$$h_t[n] = h_d[n] \cdot w_r[n] = \{-0.1592, 0.2251, \underline{0.75}, 0.2251, -0.1592\}$$

If causal filter is needed, then the shift by two is needed
$h_c[n] = h_t[n-2] = \{\underline{-0.1592}, 0.2251, 0.75, 0.2251, -0.1592\}$.
In Figure 93 time-domain view:
(a) $h_d[n]$ (IIR), (b) $w_r[n]$, and (c) $h_t[n] = h_d[n] \cdot w_r[n]$ (FIR).
In Figure 94 the corresponding frequency-domain view:
(a) $H_d(e^{j\omega})$ (ideal, desired), (b) $W_r(e^{j\omega})$, and (c) $H_t(e^{j\omega}) = H_d(e^{j\omega}) \circledast W_r(e^{j\omega})$ (realisable).



(a) (b) (c)

Figure 93: Problem 45(a): time domain view, (a) $h_d[n]$, (b) $w_r[n]$,(c) $h_t[n]$.



(a) (b) (c)

Figure 94: Problem 45(a): frequency domain $(0 \ldots \pi)$, (a) $H_d(e^{j\omega})$, (b) $W_r(e^{j\omega})$, (c) $H_t(e^{j\omega})$.

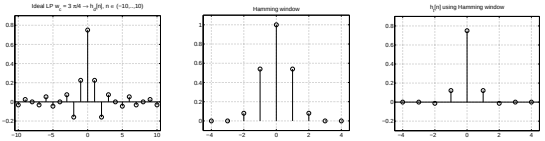
b) Now we are using Hamming window[2] $w_h[n]$ of length 5,

$$w_h[n] = \begin{cases} 0.54 + 0.46\cos(2\pi n/4), & -2 \le n \le 2 \\ 0, & \text{otherwise} \end{cases}$$

Hence,

$$\begin{aligned} h_t[n] &= h_d[n] \cdot w_h[n] = h_d[n] \cdot \big(0.54 + 0.46\cos(2\pi n/(2M))\big) \\ &= \{0.08 \cdot (-0.1592), 0.54 \cdot 0.2251, \underline{0.75}, 0.54 \cdot 0.2251, 0.08 \cdot (-0.15592)\} \\ &= \{-0.0127, 0.1215, \underline{0.75}, 0.1215, -0.0127\} \end{aligned}$$

If causal filter is needed, then
$h_c[n] = h_t[n-2] = \{\underline{-0.0127}, 0.1215, 0.75, 0.1215, -0.0127\}$
In Figure 95 time-domain view:
(a) $h_d[n]$, (b) $w_h[n]$, and (c) $h_t[n] = h_d[n] \cdot w_h[n]$.
In Figure 96 the corresponding frequency-domain view:
(a) $H_d(e^{j\omega})$, (b) $W_h(e^{j\omega})$, and (c) $H_t(e^{j\omega}) = H_d(e^{j\omega}) \circledast W_h(e^{j\omega})$.



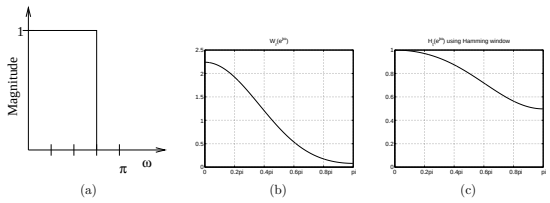
(a) (b) (c)

Figure 95: Problem 45(b): time domain view, (a) $h_d[n]$, (b) $w_h[n]$,(c) $h_t[n]$.

c) **Some examples of window functions:**

i) Rectangular N=11, Figure 97
ii) Rectangular N=65, Figure 98
iii) Hamming N=65, Figure 99

[2]The expression is slightly different from that given in Section 7.6.4 in Mitra, but the same as in Matlab.



(a) (b) (c)

Figure 96: Problem 45(b): frequency domain $(0 \ldots \pi)$, (a) $H_d(e^{j\omega})$, (b) $W_h(e^{j\omega})$, (c) $H_t(e^{j\omega})$.

There are three figures for each item. Top left figure is the window function in time domain $w[n]$. The causal version can be obtained by shifting. Bottom left figure is the window function in frequency domain $W(e^{j\omega})$. The third figure in right is the amplitude frequency of actual filter which is obtained via window function method. The desired lowpass filter $H_d(e^{j\omega})$ is drawn in dashed line, the implemented filter $H_t(e^{j\omega}) = H_d(e^{j\omega}) \circledast W(e^{j\omega})$ is solid line. The cut-off frequency is at 100 Hz, and the sampling frequency is 1000 Hz.

Notice that

i) Rectangular N=11 gives insufficient result.

ii) Rectangular N=65 gives sharp transition band but oscillates (Gibbs phenomenon).

iii) Hamming N=65 is flat both in passband and stopband but the transition band is not as tight as in (ii).

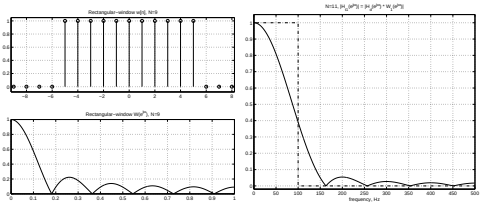


Figure 97: Rectangular window $N = 11$, see the text in Problem 45(c).

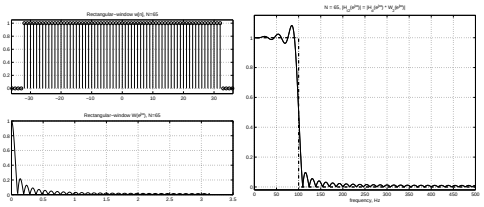


Figure 98: Rectangular window $N = 65$, see the text in Problem 45(c).

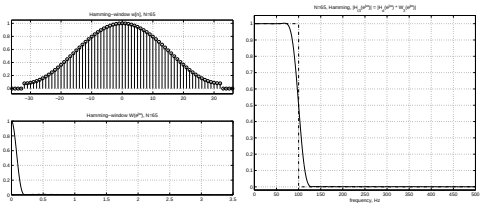


Figure 99: Hamming window $N = 65$, see the text in Problem 45(c).

46. **Problem:** The following transfer functions $H_1(z)$ and $H_2(z)$ representing two different filters meet (almost) identical amplitude response specifications

$$H_1(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}}$$

where $b_0 = 0.1022$, $b_1 = -0.1549$, $b_2 = 0.1022$, $a_1 = -1.7616$, and $a_2 = 0.8314$, and

$$H_2(z) = \sum_{k=0}^{12} h[k] z^{-k}$$

where $h[0] = h[12] = -0.0068$, $h[1] = h[11] = 0.0730$, $h[2] = h[10] = 0.0676$, $h[3] = h[9] = 0.0864$, $h[4] = h[8] = 0.1040$, $h[5] = h[7] = 0.1158$, $h[6] = 0.1201$.

For each filter,

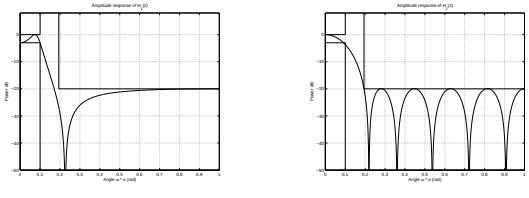
a) state if it is a FIR or IIR filter, and what is the order

b) draw a block diagram and write down the difference equation

c) determine and comment on the computational and storage requirements

d) determine first values of $h_1[n]$

Solution: The transfer functions $H_1(z)$ and $H_2(z)$ have been designed using the same amplitude specifications, see Figure 100.



$|H_1(e^{j\omega})|$ $|H_2(e^{j\omega})|$

Figure 100: Amplitude responses of $H_1(z)$ and $H_2(z)$ in Problem 46.

a) $H_1(z)$ is IIR. There is a denominator polynomial.
$H_2(z)$ is FIR. There is only the nominator polynomial.

b) $H_1(z)$ is an IIR filter. In order to show the feedback in time domain one has to use inverse z-transform:

$$\begin{aligned} H(z) = \frac{Y(z)}{X(z)} &= \frac{b_0 + b_1 z^{-1} + b_2 z^{-1}}{1 + a_1 z^{-1} + a_2 z^{-1}} \\ Y(z)\big(1 + a_1 z^{-1} + a_2 z^{-1}\big) &= X(z)\big(b_0 + b_1 z^{-1} + b_2 z^{-1}\big) \qquad | \; Z^{-1}\{.\} \\ y[n] + a_1 y[n-1] + a_2 y[n-2] &= b_0 x[n] + b_1 x[n-1] + b_2 x[n-2] \end{aligned}$$

From the difference equation the block diagram can be drawn (Figure 101). Note that the same coefficients can be found also in the form of $H_1(z)$.

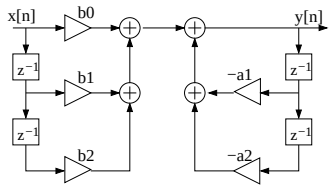


Figure 101: $H_1(z)$ as a block diagram in Problem 46.

The impulse response $h[n]$ of FIR filter $H_2(z)$ is directly seen and its length is 13 (finite impulse response). The block diagram consists only of multipliers and delays (Figure 102).

c) From examination of the two difference equations the computational and storage requirements for both filters are summarized in Table 10.
It is evident that the IIR filter is more econimical in both computational and storage requirements than the FIR filter. However, there are some tricks to improve FIR filter structure (e.g. Sections 6.3.3., 6.3.4 in Mitra).

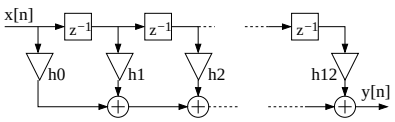


Figure 102: $H_2(z)$ as a block diagram in Problem 46.

	FIR	IIR
Number of multiplications	13	5
Number of additions	12	4
Storage locations (coefficients and data)	26	10

Table 10: Computational and storage requirements of $H_1(z)$ and $H_2(z)$.

d) A simple way to determine the impulse response is to insert an impulse $x[n] = \delta[n]$ into input and compute recursively with difference equation what comes out in $y[n]$. The registers are assumed to be zero in the initial moment. Another way to solve first values of $h_1[n]$ is to apply long division. Unfortunately, both cases are heavy by hands. Inverse z-transform can be used in order to receive exact $h[n]$. Using Matlab,

$$h_1[n] = \{\underline{0.1022},\ 0.0251,\ 0.0615,\ 0.0875,\ 0.1029,\ 0.1086,\ \ldots\}$$

47. **Problem:** Suppose that the calculation of FFT for a one second long sequence, sampled with 44100 Hz, takes 0.1 seconds. Estimate the time needed to compute (a) DFT of a one second long sequence, (b) FFT of a 3-minute sequence, (c) DFT of a 3-minute sequence. The complexities of DFT and FFT can be approximated with $\mathcal{O}(N^2)$ and $\mathcal{O}(N \log_2 N)$, respectively.

Solution: FFT is a computationally effective algorithm for calculating the Discrete Fourier Transform (DFT) of a sequence(Section 8.3.2 in Mitra). The computational complexity of FFT is $\mathcal{O}(N \log N)$ where N is the length of the sequence. The complexity of the basic algorithm for DFT is quadratic to the input length i.e. $\mathcal{O}(N^2)$.

Here, it is supposed that the calculation of FFT for a one second long sequence, sampled with 44100 Hz, takes 0.1 seconds. Thus, the length of the sequence is $N = 1\text{ s} \times 44100\text{ Hz} = 44100$ samples and we can approximate the number of operations needed for the calculation as $N \log_2 N$ (using the base-2 logarithm). Since performing these operations takes 0.1 seconds, we get the (average) execution time for a single operation:

$$t = \frac{0.1\text{ s}}{44100 \log_2(44100)} \approx 147\text{ ns}$$

a) The time needed to compute DFT of a one second long sequence is estimated as the number of operations needed times the execution time for a single operation:

$$N^2 t = 44100^2 \times 147\text{ ns} \approx 300\text{ s} \approx 5\text{ min}$$

b) A 3-minute sequence, sampled with 44100 Hz, consists of $N' = 180\text{ s} \times 44100\text{ Hz} = 7938000$ samples. Calculating FFT for N' takes approximately:

$$N' \log_2(N') t = 7938000 \log_2(7938000) \times 147\text{ ns} \approx 30\text{ s}$$

c) Calculating DFT for N' takes approximately:

$$(N')^2 t = 7938000^2 \times 147\text{ ns} \approx 9 \cdot 10^6\text{ s} \approx 100\text{ d}$$

It should be noted that these are only very crude approximations of the actual time it takes to calculate the FFT and DFT algorithms with different sizes of input sequences. The $\mathcal{O}(\cdot)$ notation omits all additive constants and constant coefficients of the complexity and concerns only the asymptotic behavior of complexity when N grows without limit. In addition, the length of N is assumed to be a power of 2 in FFT algoritms.

48. **Problem:** Express the desimal number -0.3125 as a binary number using sign bit and four bits for the fraction in the format of (a) sign-magnitude, (b) ones' complement, (c) two's complement. What would be the value after truncation, if only three bits are saved.

Solution: The binary number representation is discussed in Section 8.4 in Mitra. Now, $-0.3125 = -5/16$. We can express it in fixed-point representation using a sign bit s and four bits for the fraction.

There are three different forms for negative numbers, for which all the sign bit is 0 for a positive number and 1 for a negative number.

a) Sign-magnitude format: $1_\Delta 0101$.
b-bit fraction is always $\sum_{i=1}^{b} a_{-i} 2^{-i}$. For a negative number $s = 1$:
$S = -(0 \cdot 2^{-1} + 1 \cdot 2^{-2} + 0 \cdot 2^{-3} + 1 \cdot 2^{-4}) = -0.3125$.

b) Ones' complement: $1_\Delta 1010$.
Decimal number $S = -s(1 - 2^{-b}) + \sum_{i=1}^{b} a_{-i} 2^{-i}$. The negative number can also be achieved by complementing all bits of the corresponding positive value ($+0.3125 \triangleq 0_\Delta 0101 \rightarrow 1_\Delta 1010 \triangleq -0.3125$).
$S = -1(1 - 2^{-4}) + (1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} + 0 \cdot 2^{-4})$
$= -0.9375 + 0.625 = -0.3125$

c) Two's complement: $1_\Delta 1011$.
Decimal number $S = -s + \sum_{i=1}^{b} a_{-i} 2^{-i}$. It can also be achieved by complementing all bits and adding 1 to the least-significant bit (LSB) ($+0.3125 \triangleq 0_\Delta 0101 \rightarrow 1_\Delta 1010 + 1 = 1_\Delta 1011 \triangleq -0.3125$).
$S = -1 + (1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} + 1 \cdot 2^{-4})$
$= -1 + 0.6875 = -0.3125$

The two's complement is normally used in DSP chips.

After truncation

a) $1_\Delta 0101 \rightarrow 1_\Delta 01 \triangleq -0.25$

b) $1_\Delta 1010 \rightarrow 1_\Delta 10 \triangleq -0.25$

c) $1_\Delta 1011 \rightarrow 1_\Delta 10 \triangleq -0.5$

it can be seen that in this case truncation of (a) and (b) produced a bigger number, but (c) a smaller. The analysis of quantization (truncation) process (Mitra, Section 9.1) results to quantization errors depicted in Problem 50.

49. **Problem:** In the following Figure 103, some error probability density functions of the quantization error are depicted.

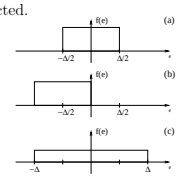


Figure 103: Problem 49: Error density functions, also at page 18.

(a) Rounding

(b) Two's complement truncation

(c) Magnitude (one's complement) truncation

is used to truncate the intermediate results. Calculate the expectation value of the quantization error m_e and the variance σ_e^2 in each case.

Solution: In this problem we are analysing different types of quantization methods. Δ here means the quantization step, $\Delta = 2^{-B}$. For example, if we are using $(B+1) = (4+1)$ bits and fixed-point numbers with two's complement representation, possible $2^{B+1} = 32$ quantized values are
$\{-1, -15/16, -14/16, \ldots, 14/16, 15/16\}$.

The area (integral) of the propability density function $f(e)$ is always one. All the distributions are uniform. Hence, $f(e)$ (height of the box) of each pdf is easily computed. We first compute $\mathrm{E}[E] = m_e$ and $\mathrm{Var}[E] = \mathrm{E}[(E - \mathrm{E}[E])^2] = \sigma_e^2$ for general uniform distribution (see Figure 104).

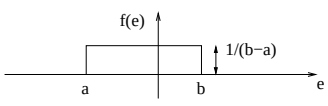


Figure 104: Computing the mean and variance of general uniform distribution in Problem 49.

$$f(e) = \begin{cases} \dfrac{1}{b-a} & a \le e \le b \\ 0 & e < a \vee e > b \end{cases}$$

$$m_e = \int_{-\infty}^{\infty} e f(e) de = \int_a^b e \frac{1}{b-a} de = \frac{1}{b-a} \Big/_a^b \frac{1}{2} e^2$$

$$= \frac{1}{2} \frac{1}{b-a} (b^2 - a^2) = \frac{1}{2} \frac{1}{b-a} (b-a)(b+a) = \frac{1}{2}(b+a)$$

$$\sigma_e^2 = \int_{-\infty}^{\infty} (e - m_e)^2 f(e) de = \int_a^b \left[e - \frac{1}{2}(a+b) \right]^2 \frac{1}{b-a} de$$

$$= \frac{1}{b-a} \Big/_a^b \frac{1}{3} \left[e - \frac{1}{2}(a+b) \right]^3$$

$$= \frac{1}{3} \frac{1}{b-a} \left\{ \left[b - \frac{1}{2}(a+b) \right]^3 - \left[a - \frac{1}{2}(a+b) \right]^3 \right\}$$

$$= \frac{1}{3} \frac{1}{b-a} \left\{ \left[\frac{1}{2} b - \frac{1}{2} a \right]^3 - \left[\frac{1}{2} a - \frac{1}{2} b \right]^3 \right\}$$

$$= \frac{1}{12} \frac{1}{b-a} (b-a)^3 = \frac{1}{12} (b-a)^2$$

Computation of mean and variance for each tree cases in the exercise paper, (a) rounding, (b) two's complement truncation, and (c) magnitude truncation.

a) Rounding: $a = -\frac{\Delta}{2}$, $b = \frac{\Delta}{2}$

$$m_e = \frac{1}{2}(-\frac{\Delta}{2} + \frac{\Delta}{2}) = 0$$

$$\sigma_e^2 = \frac{1}{12} \left[\frac{\Delta}{2} - \left(-\frac{\Delta}{2} \right) \right]^2 = \frac{\Delta^2}{12}$$

b) Two's complement truncation: $a = -\Delta$, $b = 0$

$$m_e = \frac{1}{2}(-\Delta + 0) = -\frac{\Delta}{2}$$

$$\sigma_e^2 = \frac{1}{12}[0 - (-\Delta)]^2 = \frac{\Delta^2}{12}$$

c) Magnitude truncation: $a = -\Delta$, $b = \Delta$

$$m_e = \frac{1}{2}(-\Delta + \Delta) = 0$$

$$\sigma_e^2 = \frac{1}{12}[\Delta - (-\Delta)]^2 = \frac{\Delta^2}{3}$$

50. **Problem:** In this problem we study the roundoff noise in direct form FIR filters. Consider an FIR filter of length N having the transfer function

$$H(z) = \sum_{k=0}^{N-1} h[k] z^{-k}.$$

Sketch the direct form realization of the transfer function.

a) Derive a formula for the roundoff noise variance when quantization is done before summations.

b) Repeat (a) for the case where quantization is done after summations, i.e. a double precision accumulator is used.

Solution: Direct form realization of the filter. Quantization blocks are marked by Q in Figure 105.

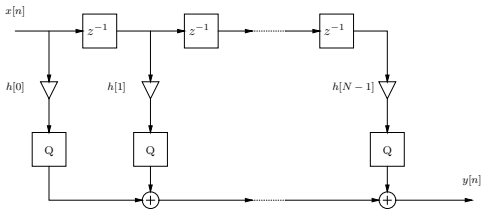


Figure 105: Filter with finite wordlength in Problem 50.

a) The roundoff noise model ($e_i[n]$:s are error sources), when quantization is done before summations, is depicted in Figure 106.

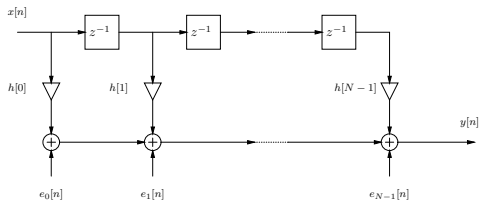


Figure 106: Roundoff noise model with N quantization points in Problem 50.

It is assumed that the quantization is done using rounding. $B+1$ bits are used in the coefficient quantization ($\Delta = 2^{-B}$):

$$\Rightarrow \sigma_e^2 = \frac{2^{-2B}}{12}, \quad m_e = 0 \text{ for all } e_i[n], \quad i = 0, \ldots, N-1.$$

Transfer functions from noise sources to the output are equal to unity. Total output noise is thus

$$e[n] = \sum_{i=0}^{N-1} e_i[n] \ .$$

The variance of the noise is

$$\begin{aligned}
\sigma_{e,tot}^2 &= E\left[e^2[n]\right] - \underbrace{E[e[n]]}_{=0 \text{ (rounding)}}{}^2 \\
&= E\left[\left(\sum_{i=0}^{N-1} e_i[n]\right)^2\right] \qquad [E[e_i[n]e_j[n]] = 0, i \neq j] \\
&= \sum_{i=0}^{N-1} E\left[e_i^2[n]\right] = \sum_{i=0}^{N-1} \sigma_e^2 = N\sigma_e^2 = N\frac{2^{-2B}}{12}
\end{aligned}$$

b) The model, when quantization is done after summations, is drawn in Figure 107. Now there is only one quantization point, i.e., there is only one noise source, $e[n]$.

$$\Rightarrow \sigma_{e,tot}^2 = \sigma_e^2 = \frac{2^{-2B}}{12}.$$

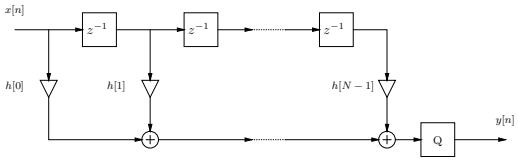


Figure 107: Filter with only one quantization point in Problem 50.

51. **Problem:** The quantization errors produced in digital systems may be compensated by error-shaping filters (Section 9.10 in Mitra). The error components are extracted from the system and processed e.g. using simple digital filters. This way the noise at the output of the system can be reduced.

Consider a lowpass DSP system with a second-order noise reduction system in Figure 108(a).

a) What is the transfer function of the system if infinite wordlength is used?

b) Derive an expression for the transform of the quantized output, $Y(z)$, in terms of the input transform, $X(z)$, and the quantization error, $E(z)$, and hence show that the error feedback network has no adverse effect on the input signal.

c) Deduce the expression for the error feedback function.

d) What values k_1 and k_2 should have in order to work as an error-shaping system?

Solution: First-order and second-order feedback structures are introduced in Sections 9.10.1 and 9.10.2 in Mitra. Consider first the block diagram shown in Figure 108(a) and its round-off noise model in Figure 108(b).

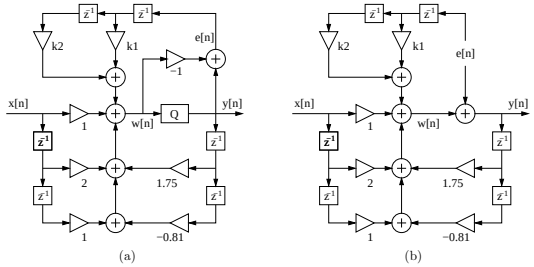


Figure 108: (a) Second-order direct form I system with second-order noise reduction, (b) and its noise model in Problem 51.

a) If infinite precision is used, the quantization is not needed and $e[n] \equiv 0$ (see Figure 108(b) with $e[n] = 0$). In that case, the system function is

$$H(z) = \frac{1 + 2z^{-1} + z^{-2}}{1 - 1.75z^{-1} + 0.81z^{-2}}$$

Computing zeros and poles we get a pole-zero diagram from which it can be derived that the filter is lowpass (Figure 109).

b) From Figure 108(a) it can be obtained the following difference equations:

$$\begin{aligned}
e[n] &= y[n] - w[n] \\
w[n] &= (x[n] + 2x[n-1] + x[n-2]) \\
&\quad + (1.75y[n-1] - 0.81y[n-2]) \\
&\quad + (k_1 e[n-1] + k_2 e[n-2])
\end{aligned}$$

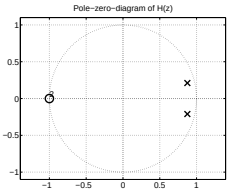


Figure 109: The pole-zero plot of $H(z) = (1+2z^{-1}+z^{-2})/(1-1.75z^{-1}+0.81z^{-2})$ in Problem 51.

After z-transform,

$$\begin{aligned}
Y(z) &= \left[\frac{1 + 2z^{-1} + z^{-2}}{1 - 1.75z^{-1} + 0.81z^{-2}}\right] X(z) + \left[\frac{1 + k_1 z^{-1} + k_2 z^{-2}}{1 - 1.75z^{-1} + 0.81z^{-2}}\right] E(z) \\
&= H(z)X(z) + H_e(z)E(z)
\end{aligned}$$

It can be observed that the noise transfer function $H_e(z)$ modifies only the quantization error.

c) The noise transfer function is

$$H_e(z) = \frac{1 + k_1 z^{-1} + k_2 z^{-2}}{1 - 1.75z^{-1} + 0.81z^{-2}} = H_{eu}(z)\, H_{es}(z)$$

Notice that without error-shaping feedback structure, i.e., $k_1 = 0$ and $k_2 = 0$, the noise transfer function is (u = unshaped)

$$H_{eu}(z) = \frac{1}{1 - 1.75z^{-1} + 0.81z^{-2}}$$

So, the error-feedback circuit is actually shaping the error spectrum by (s = shaping)

$$H_{es}(z) = 1 + k_1 z^{-1} + k_2 z^{-2}$$

d) Without error-shaping the quantized output spectrum is

$$Y_u(z) = H(z)X(z) + H_{eu}(z)E(z)$$

Error-shaping filter $H_{es}(z)$ should efficiently discard the effects of the poles of $H_{eu}(z)$. Error-feedback coefficients are chosen to be simple integers or fractions ($k_i = 0, \pm 0.5, \pm 1, \pm 2$), so that the multiplication can be performed using a binary shift operation and it will not introduce an additional quantization error. Choosing $k_1 = -2$, $k_2 = 1$, $H_{es}(z) = 1 - 2z^{-1} + z^{-2}$ is a highpass filter with two zeros at $z = 1$.
The error shaping structure lowers the noise in the passband by pushing it into the stopband of the filter (see Figure 9.45 in Mitra).

52. **Problem:** Consider a cosine sequence $x[n] = \cos(2\pi(f/f_s)n)$ where $f = 10$ Hz and $f_s = 100$ Hz as depicted in the top left in Figure 110. While it is a pure cosine, its spectrum is a peak at the frequence $f = 10$ Hz (top middle) or at $\omega = 2\pi f/f_s = 0.2\pi$ (top right).

a) Sketch the output sequence $x_u[n]$ and its spectra using up-sampler with up-sampling factor $L = 2$.

b) Sketch the output sequence $x_d[n]$ and its spectra using down-sampler with factor $M = 2$.

Solution: Sometimes it is necessary or useful to change the sampling frequency f_s. Consider music formats DAT (48 kHz) and CD (44.1 kHz).

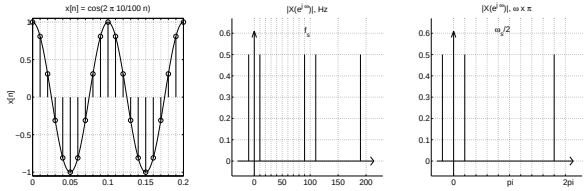


Figure 110: Problem 52(a). The original sequence of a cosine of $f = 10$ Hz and its spectrum. The angular frequency $\omega = 2\pi(f/f_s) = 2\pi(10/100) = 0.2\pi$.

a) Up-sampling with factor $L = 2$. In the time domain there will be $L-1$ zeros between the original samples, see Figure 111(a).

$$x_u[n] = \begin{cases} x[n/L], & n = 0, \pm L, \pm 2L, \ldots \\ 0, & \text{otherwise} \end{cases}$$
$$= \begin{cases} x[n/2], & n = 0, \pm 2, \pm 4, \ldots \\ 0, & \text{otherwise} \end{cases}$$

In the frequency domain the sampling frequency is multiplied by L, hence, the new sampling frequency is 200 Hz. $L-1$ images from the original spectrum are emerged equivalently between 0 and $f_{s,new}$.

$$X_u(e^{j\omega}) = X(e^{j\omega L}) = X(e^{j2\omega})$$

Each cosine is a peak pair ($\pm f$) in the spectrum. The original peaks are at $f = 10$ and $f = 200 - 10 = 190$ Hz, and after up-sampling new images at $f = 90$ and $f = 110$ Hz, as shown in Figure 111(b). The same with angular frequencies is shown in Figure 111(c).

Notice that if you ideally convert the sequence $x_u[n]$ into continuous-time $x_u(t)$ you will find also a high frequency component, an image component. Normally images are filtered out using a lowpass filter (see anti-imaging and anti-aliasing filters). See Figure 112.

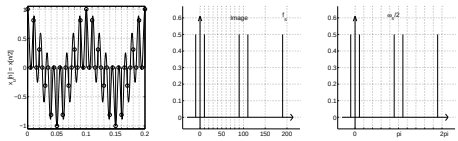


Figure 111: Problem 52(a). Up-sampled signal $x_u[n]$, factor $L = 2$. The sampling frequency is increased to 200 Hz, and there is an image spectrum.

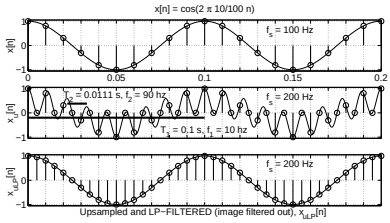


Figure 112: A closer look at up-sampling. Top, original sequence. Middle $L = 2$, $L - 1 = 1$ zeros added between the original samples. Bottom, using (ideal) LP-filter to remove the image, i.e., 90 Hz component. The continuous curve is plotted only for better visual view. See the text in Problem 52(a).

b) Down-sampling with factor $M = 2$ means taking only every second sample.

$$x_d[n] = x[nM] = x[2n]$$

A possible effect is losing information. However, in this case, this does not occur because $f = 10$ Hz $< f_{s,new}/2 = 25$ Hz. See Figure 113(a).

In the frequency domain the sampling frequency is decreased to 50 Hz. See Figures 113(b)-(c).

$$X_d(e^{j\omega}) = \frac{1}{M}\sum_{k=0}^{M-1} X(e^{j(\omega - 2\pi k)/M})$$

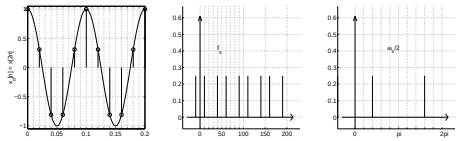


Figure 113: Problem 52(b). Down-sampled signal $x_d[n]$, factor $M = 2$. The sampling frequency is decreased to 50 Hz.

53. **Problem:** Express the output $y[n]$ of the system shown in Figure 114 as a function of the input $x[n]$.

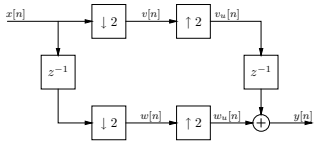


Figure 114: Multirate system of Problem 53.

Solution: Consider an input signal $x[n]$ with the corresponding z-transform $X(z)$. After factor-of-L up-sampling, the z-transform of the signal $x_u[n]$ is

$$X_u(z) = X(z^L)$$

and after factor-of-M down-sampling, the z-transform of the signal $x_d[n]$ is

$$X_d(z) = \frac{1}{M}\sum_{k=0}^{M-1} X(z^{1/M}W_M^{-k})$$

where $W_M = e^{-j2\pi/M}$. See Section 10.1.2 in Mitra for the derivation of these equations.

Using these equations, let us derive the z-transforms of the intermediate signals $v[n]$, $v_u[n]$, $w[n]$, and $w_u[n]$ and finally the z-transform of the output $y[n]$. Let us denote the delayed version of the input as $X'(z) = z^{-1}X(z)$. Furthermore, note that $W_2^{-1} = e^{j2\pi/2} = -1$.

$$V(z) = \frac{1}{2}\sum_{k=0}^{1} X(z^{1/2}W_2^{-k}) = \frac{1}{2}X(z^{1/2}) + \frac{1}{2}X(-z^{1/2})$$

$$W(z) = \frac{1}{2}\sum_{k=0}^{1} X'(z^{1/2}W_2^{-k}) = \frac{1}{2}z^{-1/2}X(z^{1/2}) - \frac{1}{2}z^{-1/2}X(-z^{1/2})$$

$$V_u(z) = V(z^2) = \frac{1}{2}X(z) + \frac{1}{2}X(-z)$$

$$W_u(z) = W(z^2) = \frac{1}{2}z^{-1}X(z) - \frac{1}{2}z^{-1}X(-z)$$

$$Y(z) = z^{-1}V_u(z) + W_u(z) = z^{-1}X(z)$$

or $y[n] = x[n-1]$ in time-domain (derive the same in time-domain!).

54. **Problem:** Show that the factor-of-L up-sampler $x_u[n]$ and the factor-of-M down-sampler $x_d[n]$ defined as in Problem 52 are linear systems.

Solution: First, consider the up-sampler. Let $x_1[n]$ and $x_2[n]$ be two arbitrary inputs with $y_1[n]$ and $y_2[n]$ as the corresponding outputs. Now,

$$y_1[n] = \begin{cases} x_1[n/L] & : \quad n = 0, \pm L, \pm 2L, \ldots \\ 0 & : \quad \text{otherwise} \end{cases}$$
$$y_2[n] = \begin{cases} x_2[n/L] & : \quad n = 0, \pm L, \pm 2L, \ldots \\ 0 & : \quad \text{otherwise} \end{cases}$$

Let us now apply the input $x_3[n] = \alpha x_1[n] + \beta x_2[n]$ with the corresponding output $y_3[n]$ as

$$y_3[n] = \begin{cases} \alpha x_1[n/L] + \beta x_2[n/L] & : \quad n = 0, \pm L, \pm 2L, \ldots \\ 0 & : \quad \text{otherwise} \end{cases}$$
$$= \begin{cases} \alpha x_1[n/L] & : \quad n = 0, \pm L, \pm 2L, \ldots \\ 0 & : \quad \text{otherwise} \end{cases} + \begin{cases} \beta x_2[n/L] & : \quad n = 0, \pm L, \pm 2L, \ldots \\ 0 & : \quad \text{otherwise} \end{cases}$$
$$= \alpha y_1[n] + \beta y_2[n]$$

Thus, the up-sampler is a linear system.

Now, consider the down-sampler with the inputs $x_1[n]$ and $x_2[n]$ and the corresponding outputs $y_1[n]$ and $y_2[n]$. Now, $y_1[n] = x_1[nM]$ and $y_2[n] = x_2[nM]$. By applying the input $x_3[n] = \alpha x_1[n] + \beta x_2[n]$ we get the corresponding output $y_3[n] = x_3[nM] = \alpha x_1[nM] + \beta x_2[nM]$. Hence, the down-sampler is also a linear system.

It should also be noted, that both the up-sampler and the down-sampler are time-varying, i.e. not LTI systems.

55. **Problem:** Consider the multirate system shown in Figure 115 where $H_0(z)$, $H_1(z)$, and $H_2(z)$ are ideal lowpass, bandpass, and highpass filters. Sketch the Fourier transforms of the outputs $y_0[n]$, $y_1[n]$, and $y_2[n]$ if the Fourier transform of the input is as shown in Figure 116(a).

Solution: First, let us denote the down-sampled signal as $x_d[n]$ and the again up-sampled signal as $x_u[n]$, shown in Figure 115.

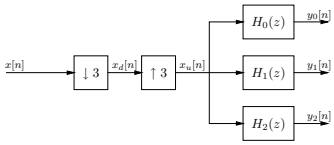


Figure 115: The multirate system in Problem 55.

The corresponding Fourier transforms (spectra) $X_d(z)$ and $X_u(z)$ are as follows (notice the reduced amplitude) in Figure 116.

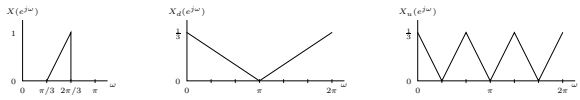


Figure 116: Original, upsampled and downsampled spectrum in Problem 55.

Now, the Fourier transforms of the outputs $Y_0(z)$, $Y_1(z)$, and $Y_2(z)$, are obtained by (ideally) filtering $X_u(z)$. The output spectra are in Figure 117.

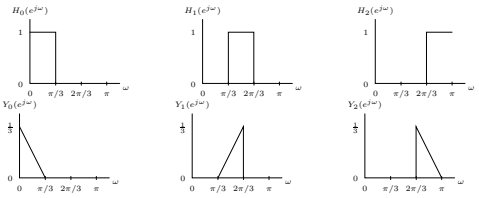


Figure 117: Bandpass filters in top row, and corresponding Output spectra in bottom row in Problem 55.

Formulas

Z-transform

$$X(z) = \sum_{n=-\infty}^{\infty} x[n]z^{-n}$$

Summary of some important z-transform pairs and properties

Discrete-time function		Z-transform	ROC
Unit impulse	$\delta[n]$	1	all z
Unit step	$\mu[n]$	$\frac{1}{1-z^{-1}}$	$\lvert z\rvert > 1$
Exponential	$a^n\mu[n]$	$\frac{1}{1-az^{-1}}$	$\lvert z\rvert > \lvert a\rvert$
Damped cosine wave	$b^n\cos(\theta n)\mu[n]$	$\frac{1-b\cos(\theta)z^{-1}}{1-2b\cos(\theta)z^{-1}+b^2z^{-2}}$	$\lvert z\rvert > \lvert b\rvert$
Damped sine wave	$b^n\sin(\theta n)\mu[n]$	$\frac{b\sin(\theta)z^{-1}}{1-2b\cos(\theta)z^{-1}+b^2z^{-2}}$	$\lvert z\rvert > \lvert b\rvert$
Linear combination	$ax[n]+by[n]$	$aX(z)+bY(z)$	
Time shift	$x[n\pm n_0]$	$z^{\pm n_0}X(z)$	
Exponential weighting	$a^nx[n]$	$X\left(\frac{z}{a}\right)$	
Linear weighting	$nx[n]$	$-z\frac{dX(z)}{dz}$	
Convolution	$x[n]\circledast y[n]$	$X(z)Y(z)$	
Product	$x[n]y[n]$	$\frac{1}{2\pi j}\oint_C X(v)Y\left(\frac{z}{v}\right)\frac{dv}{v}$	
Even	$x[n]=x[-n]$	Real $X(e^{j\omega})$	
Odd	$x[n]=-x[-n]$	Imaginary $X(e^{j\omega})$	
Real	$x[n]$	Even $\lvert X(e^{j\omega})\rvert$ and odd $\arg X(e^{j\omega})$	

Fourier Series

Continuous-time signals

$$x(t) = \sum_{k=-\infty}^{\infty} c_k e^{jk\Omega t}$$
$$c_k = \frac{1}{T}\int_T x(t)e^{-jk\Omega t}$$

Discrete-time sequences

$$x[n] = \sum_{k=\langle N\rangle} c_k e^{jk\omega n}$$
$$c_k = \frac{1}{N}\sum_{n=\langle N\rangle} x[n]e^{-jk\omega n}$$

Fourier Transform

Continuous-time signals

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(j\Omega)e^{j\Omega t}\mathrm{d}\Omega$$
$$X(j\Omega) = \int_{-\infty}^{\infty} x(t)e^{-j\Omega t}\mathrm{d}t$$

Discrete-time sequences

$$x[n] = \frac{1}{2\pi}\int_{2\pi} X(e^{j\omega})e^{j\omega n}\mathrm{d}\omega$$
$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x[n]e^{-j\omega n}$$

Connection to z-transform: $z \leftrightarrow e^{j\omega}$.

Convolution

Continuous-time signals

$$y(t) = h(t)\circledast x(t) = x(t)\circledast h(t)$$
$$= \int_{-\infty}^{\infty} h(\tau)x(t-\tau)\mathrm{d}\tau$$

Discrete-time sequences

$$y[n] = h[n]\circledast x[n] = x[n]\circledast h[n]$$
$$= \sum_{k=-\infty}^{\infty} h[k]x[n-k]$$

Filter Design

$s = k\cdot(1-z^{-1})/(1+z^{-1}),\quad k=1 \text{ or } k=2/T$

$\Omega_{prewarp,c} = k\cdot\tan(\omega_c/2),\quad k=1 \text{ or } k=2/T$

$\omega_c = 2\pi f_c/f_s$

$$H(e^{j\omega}) = \begin{cases}1, & |\omega|<\omega_c\\ 0, & |\omega|\geq\omega_c\end{cases} \quad\leftrightarrow\quad h[n] = \frac{\sin(\omega_c n)}{\pi n} = \frac{\omega_c}{\pi}\mathrm{sinc}(\frac{\omega_c n}{\pi})$$

$h_{FIR}[n] = h_{ideal}[n]\cdot w[n]$

$H_{FIR}(e^{j\omega}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} H_{ideal}(e^{j\theta})W(e^{j(\omega-\theta)})\mathrm{d}\theta$

Multirate Systems

$x_u[n] = x[n/L], n=0,\pm L,\pm 2L,\ldots;\quad x_u[n]=0,\ \text{otherwise}$

$x_d[n] = x[nM]$

$X_u(z) = X(z^L)$

$X_d(z) = (1/M)\sum_{k=0}^{M-1} X(z^{1/M}W_M^{-k})$

$X_d(e^{j\omega}) = (1/M)\sum_{k=0}^{M-1} X(e^{j(\omega-2\pi k)/M})$