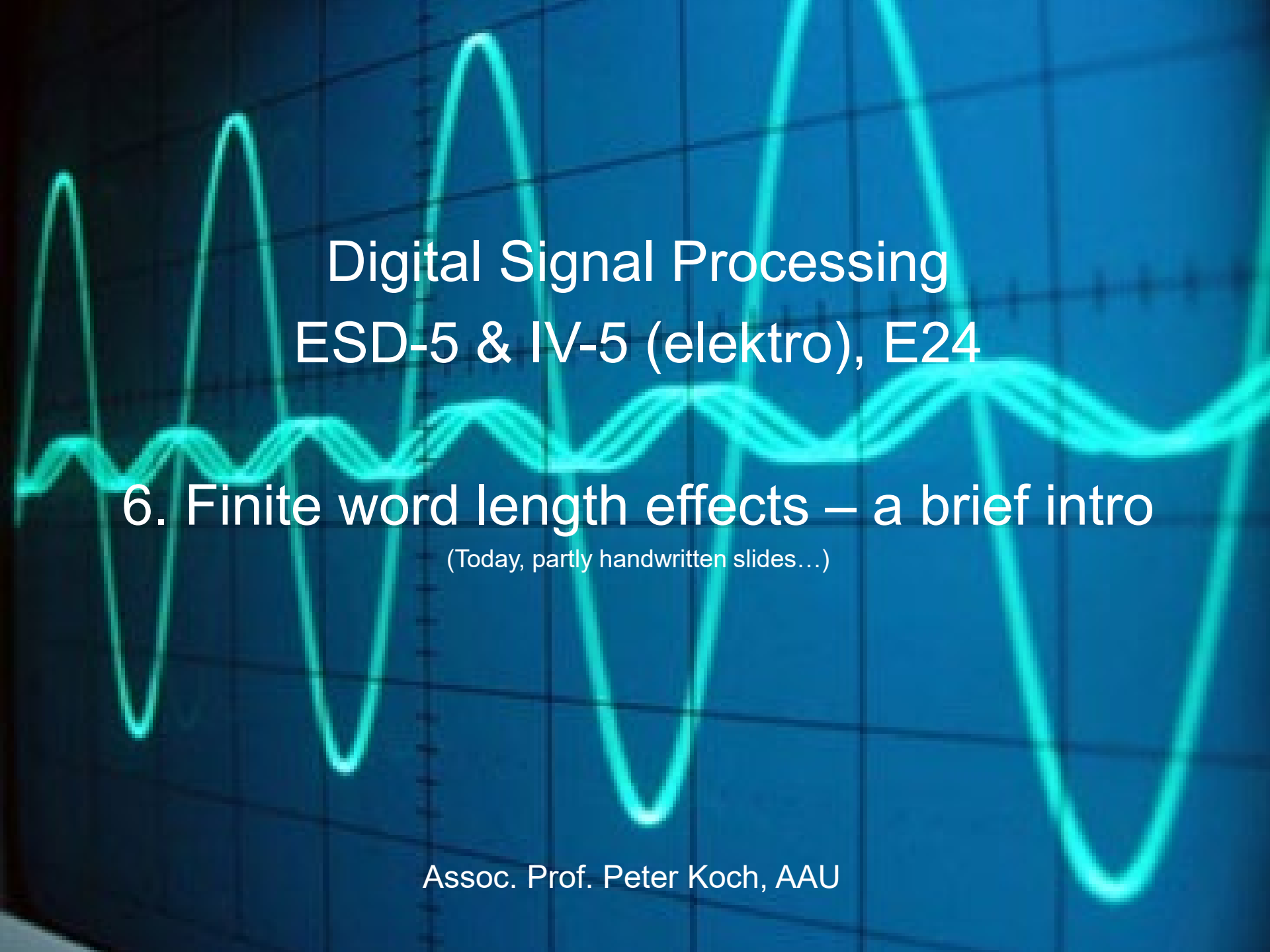


Digital Signal Processing
ESD-5 & IV-5 (elektro), E24

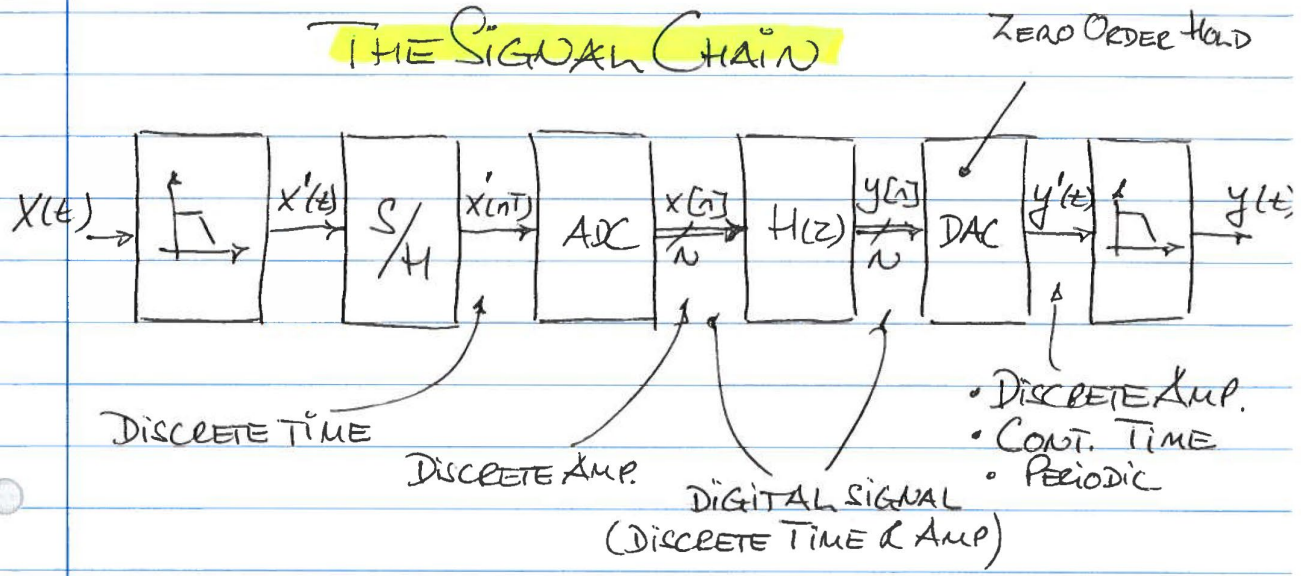
6. Finite word length effects – a brief intro

(Today, partly handwritten slides…)

Assoc. Prof. Peter Koch, AAU

Realisation of Digital Filters

The Signal Chain



Through the signal chain, several errors — or **quantizations** — are introduced.

S/H: **Time** quantization

ADC: **Variable** quantization

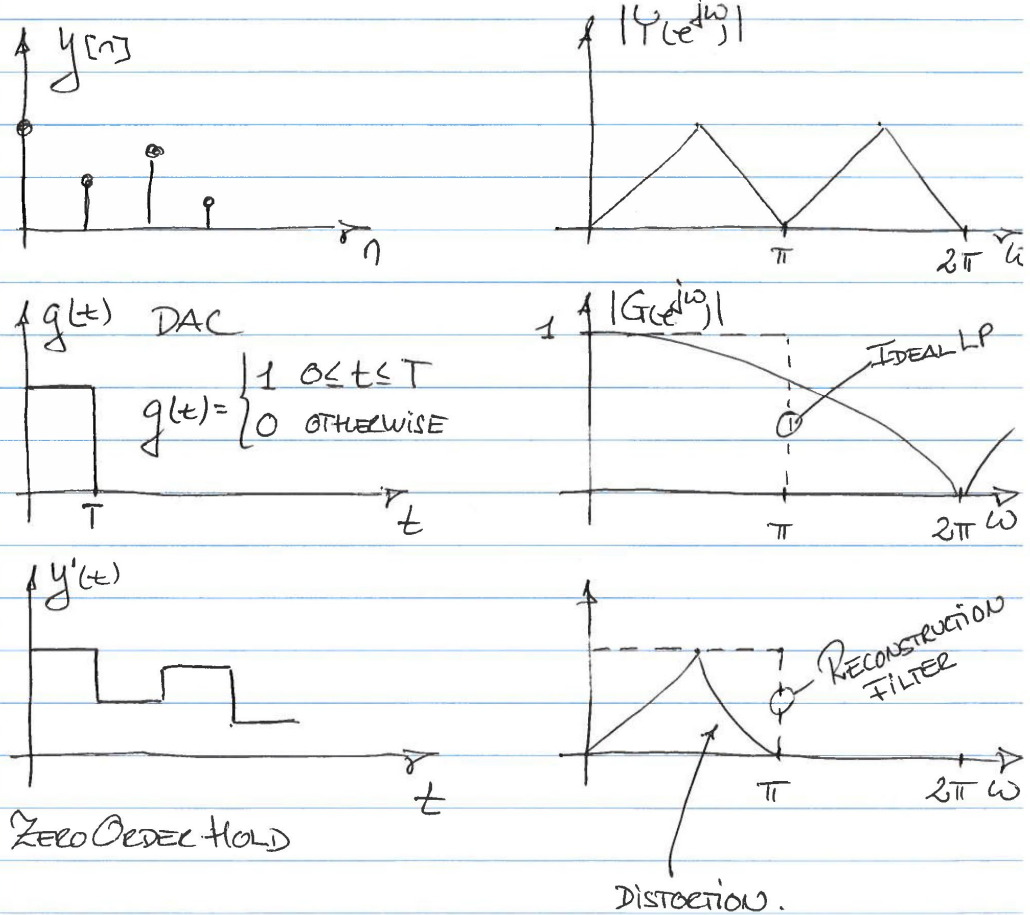
H(z) (computer): Variable + **Coefficient** Quant.

DAC: Time quantization.

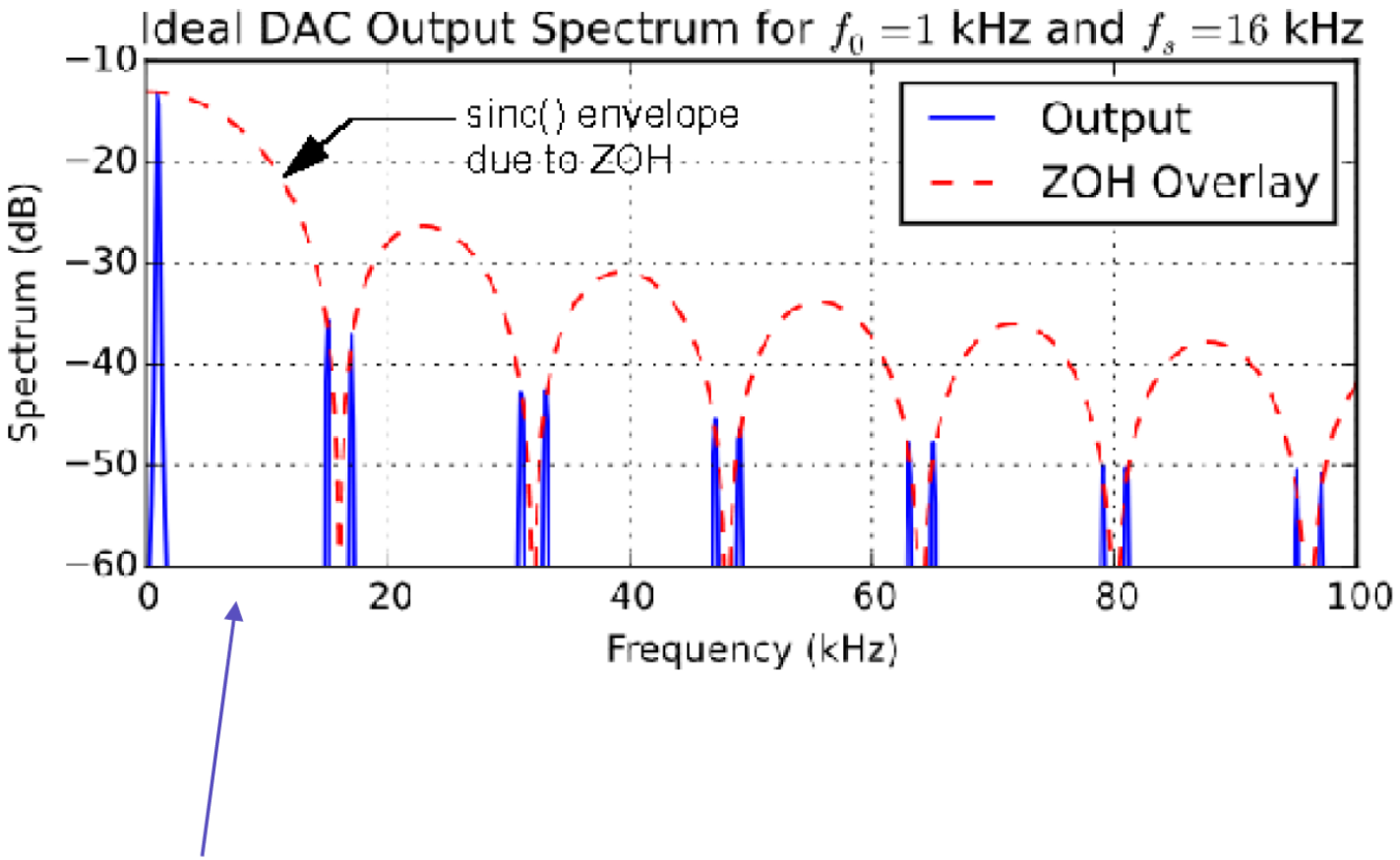
Time Quantization

Time Quantization

- $f_{sample} > 2 \cdot f_{max,analog}$ (Reconstruction).
- Always apply an Anti-Aliasing filter.
- Discrete time $\Rightarrow$ Periodic frequency.
- The DAC is NOT an ideal L.P. filter

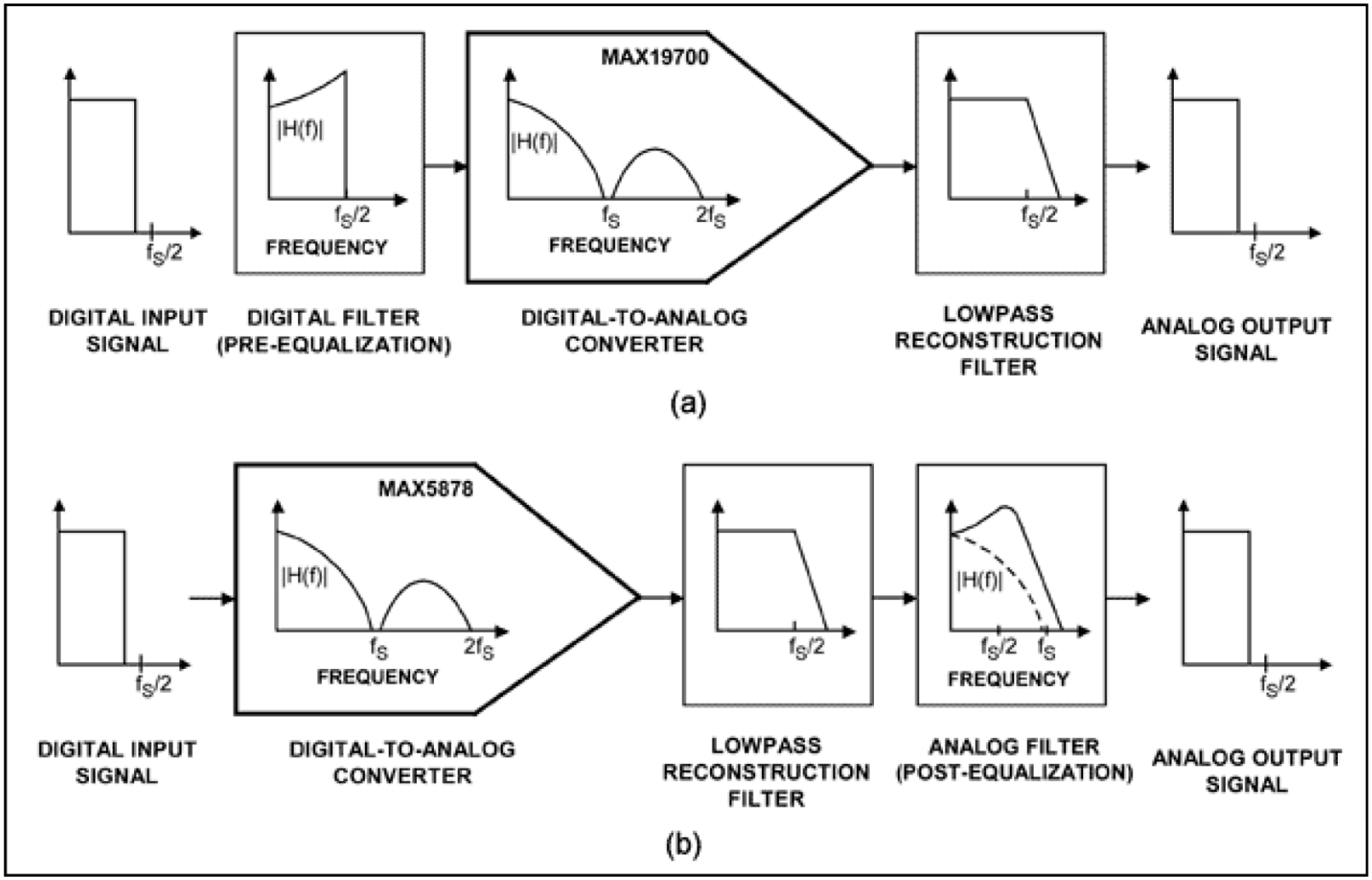


The Sinc shaped ZOH Overlay maintains all the 2π periodic images from the output signal $y[n]$.



Here we see clearly the need for a reconstruction LP-filter with a cut-off frequency at $f_s/2$

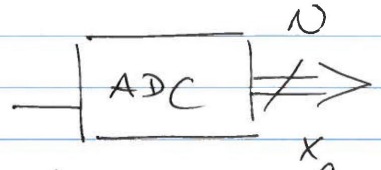
For frequency components close to the Nyquist frequency, the ZOH introduces distortion.
To compensate for this ZOH distortion we may apply either pre- or post-equalization



Variable Quantization

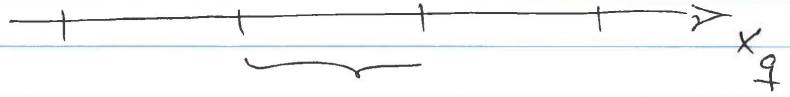
Variable Quantization

- ADC



Continuous in Amp — Discrete in Amp

N bit ADC $\Rightarrow 2^N$ possible values for x_q



Quantization Step

$\Delta = 1/2^N$

So, at the ADC output we have a finite Signal-to-Noise Ratio (SNR)

$$SNR \triangleq 20 \cdot \log \frac{\text{Signal (RMS)}}{\text{Noise (RMS)}}$$

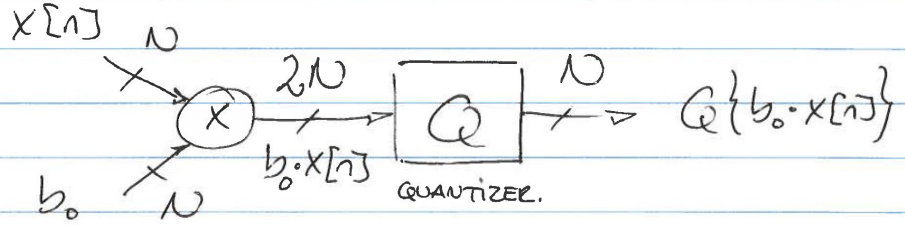
Without proof;

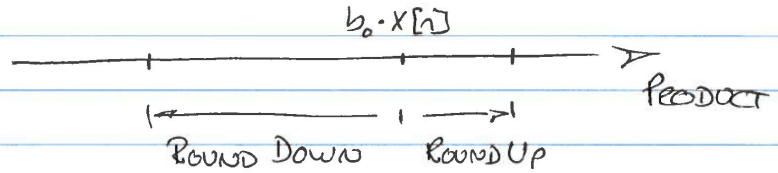
> The SNR on the output of the ADC is increased with 6dB when N is increased with 1.
>
> $$SNR = 6.02N + 1.76 \text{ dB}$$

- $H(z)$ (COMPUTER)

MULTIPLICATION BETWEEN SIGNAL AND COEF.

$$y[n] = \sum_{k=1}^{N} a_k \cdot y[n-k] + \sum_{l=0}^{M} b_l x[n-l]$$





SO, THE QUANTIZER WHICH TAKES US BACK FROM 2N BIT TO N BIT, INTRODUCES AN ERROR ;

$$e[n] = b_0 \cdot x[n] - Q\{b_0 \cdot x[n]\}$$

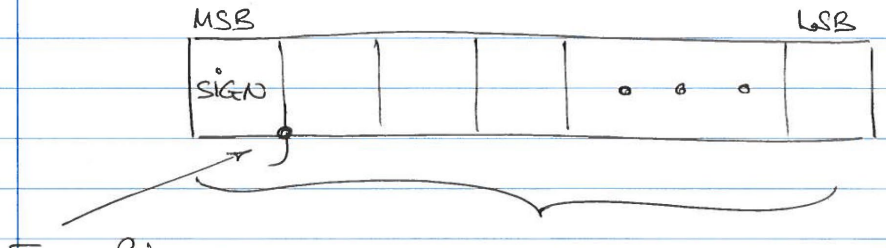
2N BIT N BIT

THE COMPUTATION ITSELF MAKES NOISE

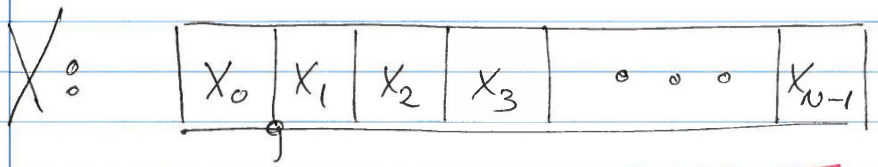
A FEW WORDS ON NUMBER REPRESENTATION.

- FLOATING POINT → LARGE DYNAMIC RANGE ✓
  → TIME, AREA, POWER ✗
- FIXED POINT → LARGE DYNAMIC RANGE ✗
  → TIME, AREA, POWER ✓

SIGNIFICANTLY USED IN SIGNAL PROCESSING IN TERMS OF 2's COMPLEMENT NUMBER REPRESENTATION.



$$\text{SIGN} = \begin{cases} 0 \rightarrow \text{POS. NUMBER} \\ 1 \rightarrow \text{NEG. NUMBER.} \end{cases}$$

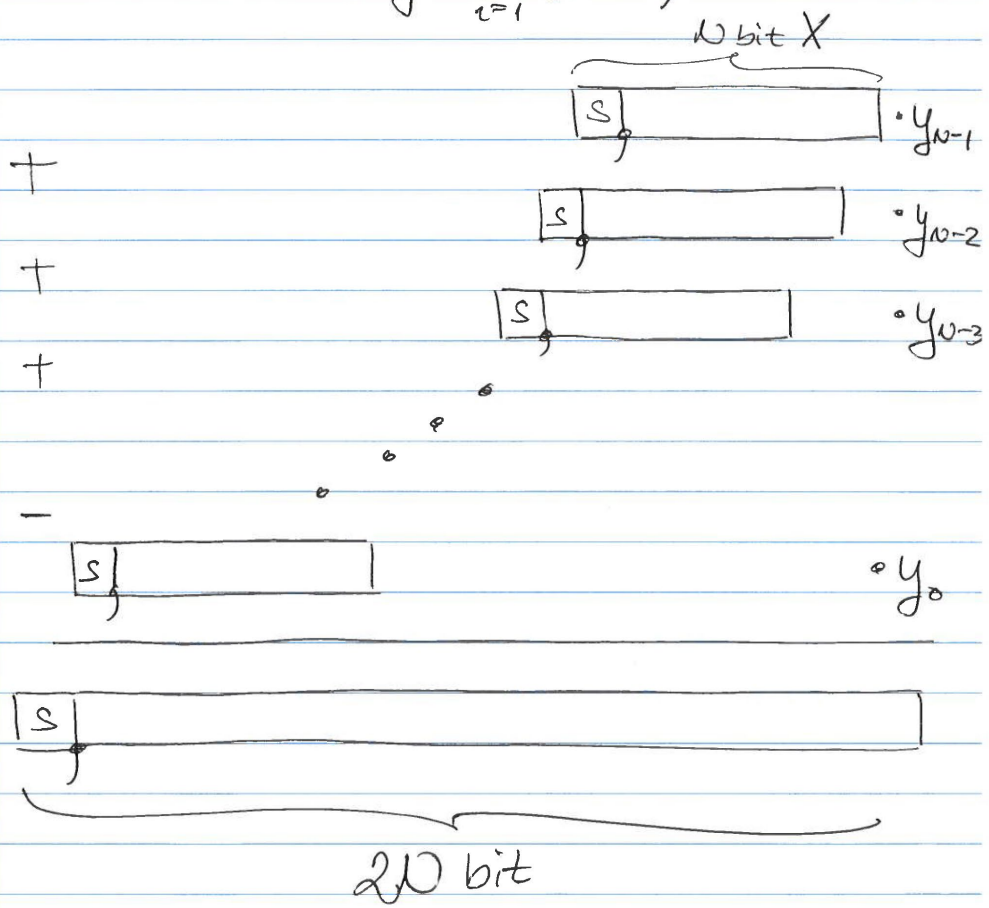


THE NUMERICAL VALUE OF X;

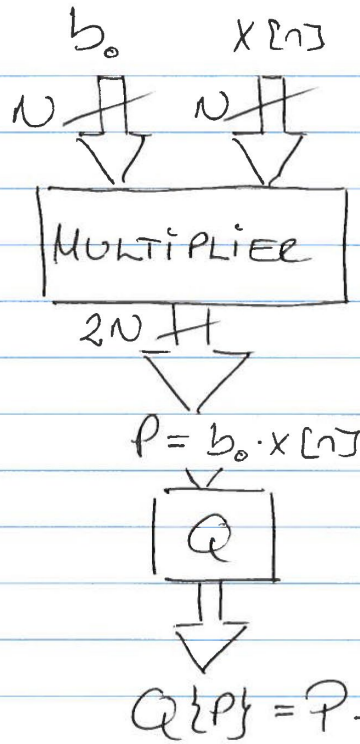
$$X = -X_0 + \sum_{i=1}^{N-1} X_i \cdot 2^{-i}$$

Multiplying two N bit numbers;

$$P = X \cdot Y = X \cdot \left(-y_0 + \sum_{i=1}^{N-1} y_i \cdot 2^{-i}\right)$$



- Actually, the product P only has 2N-1 bit which is compensated for via left-shift (0 into LSB position)
- Furthermore, sign-extension is need for the addition.

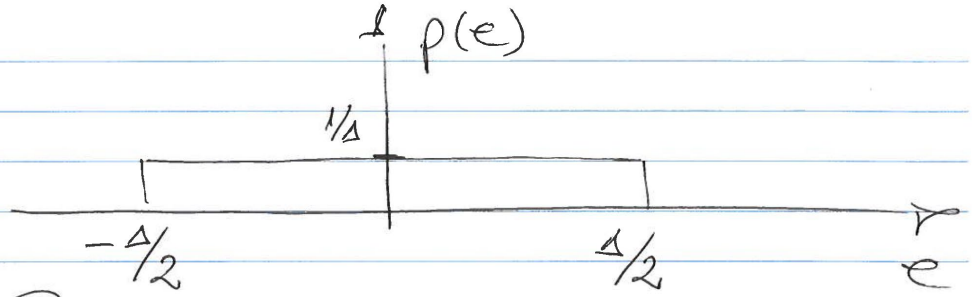


CAN WE SAY ANYTHING ABOUT THIS ERROR ??

DIFFICULT, BUT WE CAN MAKE SOME ASSUMPTIONS

- $e[n]$ IS A STATIONARY WHITE NOISE PROCESS
- THE AMPLITUDE OF $e[n]$ IS EVENLY DISTRIBUTED OVER ONE QUANTIZATION STEP.
- $e[n]$ AND $x[n]$ ARE UN-CORRELATED.

USING THESE ASSUMPTIONS WE CAN SAY SOMETHING DEFINITELY ABOUT THE STATISTICS OF $e[n]$



PROBABILITY DENSITY FUNCTION

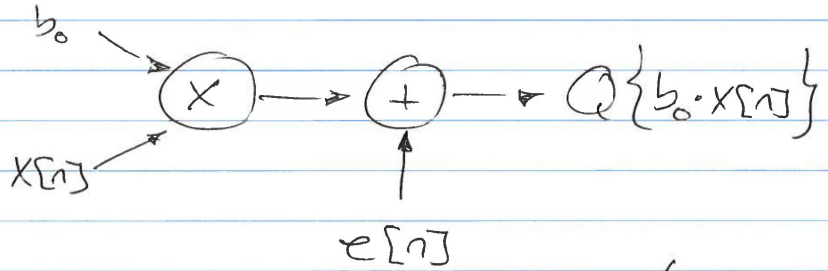
- MEAN

$$\mu_e = \int_{-\infty}^{\infty} e \cdot p(e) \cdot de = \frac{1}{\Delta}\left[\frac{e^2}{2}\right]_{-\Delta/2}^{\Delta/2} = 0$$

- VARIANCE

$$\sigma_e^2 = \int_{-\infty}^{\infty} e^2 \cdot p(e) \cdot de = \frac{1}{\Delta}\left[\frac{e^3}{3}\right]_{-\Delta/2}^{\Delta/2} = \frac{\Delta^2}{12}$$

USING THESE STATISTICAL VALUES, WE CAN NOW DERIVE A MODEL FOR THE QUANTIZATION NOISE

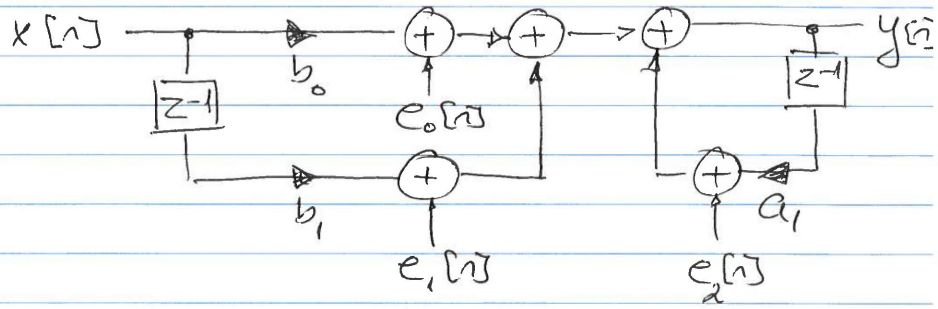


WHERE $e[n]$ IS A WHITE SIGNAL (μ_e, σ_e^2)

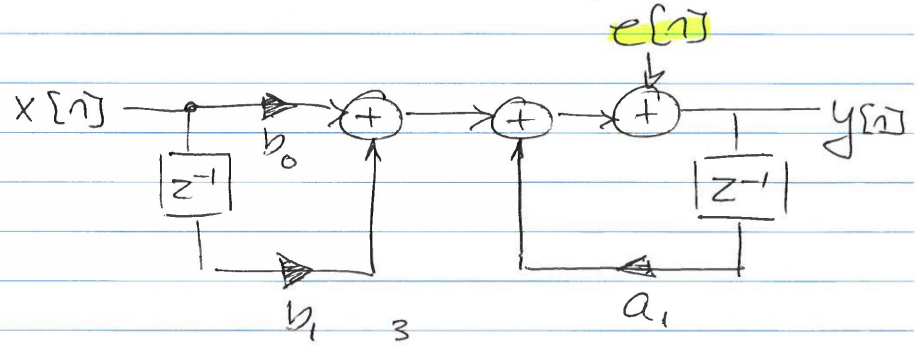
EXAMPLE: 1ST ORDER IIR FILTER.

$$y[n] = a_1 \cdot y[n-1] + b_0 \cdot x[n] + b_1 \cdot x[n-1]$$

GRAPHICAL REPRESENTATION;



SINCE THESE ARE ALL LINEAR OPERATIONS, WE CAN RE-ORGANIZE THE GRAPH;

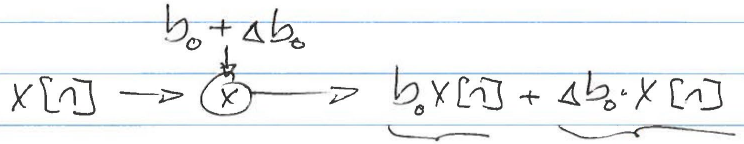


WHERE $e[n] = \sum_{i=1}^{3} e_i[n]$, AND THUS;

- THE TOTAL NOISE IMPACTS DIRECTLY THE OUTPUT.
- AT THE OUTPUT THE SNR IS NOT INFINITE

Coefficient Quantization

COEFFICIENT QUANTIZATION (in $H(z)$)



WHERE b_0 IS THE EXACT COEFFICIENT (CALCULATED WITH MANY DIGITS) AND Δb_0 IS THE ERROR INTRODUCED ON THE COEFFICIENT WHEN IT IS REPRESENTED IN N BIT

$$\hat{H}(z) = \frac{\sum_{k=0}^{M} \tilde{b}_k z^{-k}}{\sum_{l=1}^{N} \tilde{a}_l z^{-l}} = \frac{\tilde{B}(z)}{\tilde{A}(z)}$$

SINCE, GENERALLY, $\tilde{b}_k \neq b_k$ AND $\tilde{a}_l \neq a_l$ WE WILL SEE SOME CHANGES IN THE POLE/ZERO LOCATIONS.

ONE POSSIBLE METHOD TO INVESTIGATE HOW SEVERE THIS PROBLEM IS, IS TO MAKE AN ANALYSIS OF THE POLE-SENSITIVITY

Sensitivity Analysis (without proof)

$$A(z) = 1 - \sum_{k=1}^{N} a_k z^{-k} = \prod_{k=1}^{N} (1 - d_k z^{-1})$$

where d_k are the roots in $A(z)$ (poles).

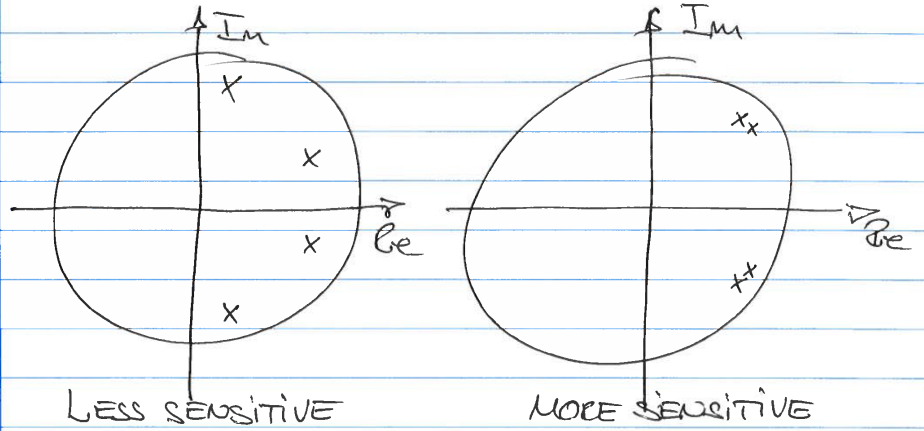
Basically, we want to know "the change in the poles as related to the change in the filter coefficients";

$$\frac{\partial d_i}{\partial a_k} = \frac{d_i^{N-k}}{\prod_{\substack{j=1 \\ j \neq i}}^{N} (d_i - d_j)}$$

(this equation is not show in the book)

So, what we see is that a change in the pole-location as related to a change in the coefficients (due to quantization) is inverse proportional to the distance between the poles

So, THE CLOSER TOGETHER THE POLES ARE LOCATED, THE MORE SENSITIVE THE POLE LOCATION IS DUE TO A CHANGE IN THE FILTER COEFFICIENTS



LESS SENSITIVE MORE SENSITIVE

CAN WE DO SOMETHING ABOUT THIS PROBLEM?

YES WE CAN.

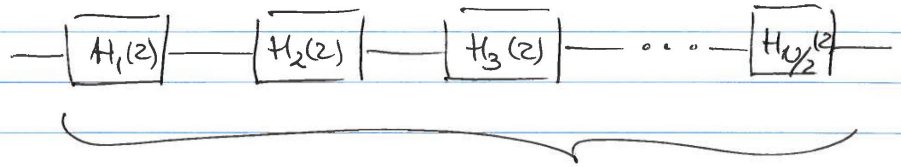
WE MAY RE-ORGANIZE $H(z)$ IN SUCH A WAY THAT CLOSELY SPACES POLES ARE DE-COUPLED INTO INDIVIDUAL FILTER SECTIONS

TAKE THE EXAMPLE ABOVE;

$$H(z) = \frac{B(z)}{A(z)} = \frac{B(z)}{1 - \sum_{k=1}^{4} a_k z^{-k}} = \frac{B(z)}{\left(1 - \sum_{l=1}^{2} a'_l z^{-l}\right)\left(1 - \sum_{m=1}^{2} a''_m z^{-m}\right)}$$

$$= \frac{B(z)}{A_1(z) \cdot A_2(z)}$$

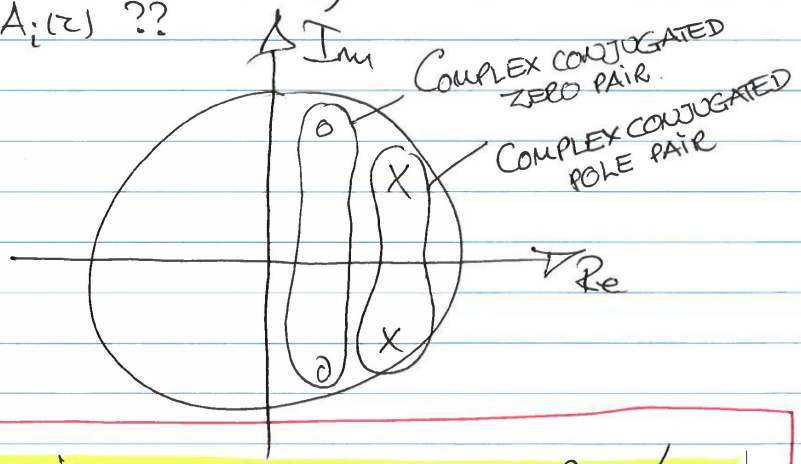
SO, THE TRICK IS TO RE-ORGANIZE $H(z)$ INTO A CASCADE OF 2ND ORDER SECTIONS (EVT. ALSO ONE 1ST ORDER SECTION).



$$H(z) = \prod_{i=0}^{N/2} H_i(z) \quad \text{(FACTORIZATION)}$$

WHERE $H_i(z) = \frac{B_i(z)}{A_i(z)}$

NOW, QUESTION IS; WHICH ZEROS FROM $H(z)$ (I.E., ROOTS FROM $B(z)$) SHOULD BE PAIRED WITH $A_i(z)$??



THE IDEA IS THAT CLOSELY SPACED POLE/ZERO PAIRS ARE MERGED INTO ONE SECTION $H_i(z)$

NEXT QUESTION ;

IS —[$H_1(z)$]—[$H_2(z)$]— IDENTICAL TO

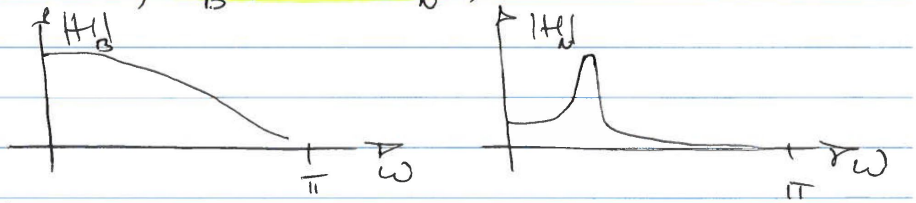
—[$H_2(z)$]—[$H_1(z)$]— ?

IF WE DO THE COMPUTATION IN INFINITE WORDLENGTH (FLOATING POINT), THEN THE ANSWER IS „YES".

HOWEVER, FOR FIXED POINT COMPUTATION, THE ANSWER MOST LIKELY IS „NO".

HENCE, WHICH ONE TO CHOOSE ?

TO ANSWER THIS QUESTION, WE DISTINGUIS BETWEEN BROAD-BAND AND NARROW-BAND SECTIONS; $H_B(z)$ AND $H_N(z)$



1) $H_N(z)$ FIRST — REDUCTION OF NOISE SPECTRUM THROUGH THE CASCADE

2) $H_N(z)$ LAST — THE AMPLITUDE LEVEL IS REDUCED ONLY AT THE OUTPUT OF THE FILTER

RECOMMENDED

Pole Locations in a 2^{nd} Order Section

LET'S HAVE A LOOK AT A 2nd ORDER SYSTEM;

$$H(z) = \frac{1}{A(z)} = \frac{1}{\sum_{l=0}^{2} a_l \cdot z^{-l}}$$

$$= \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2}}$$

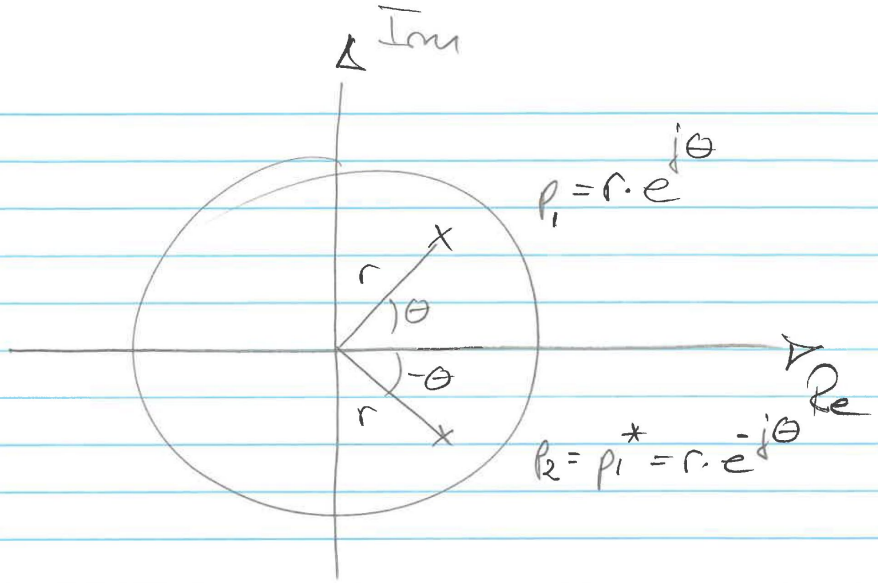
NOTE THE NORMALIZED POLYNOMIUM, $a_0 = 1$.

$$= \frac{z^2}{z^2 + a_1 z + a_2}$$

$$= \frac{z^2}{(z - p_1)(z - p_2)}$$

$$= \frac{1}{(1 - p_1 z^{-1})(1 - p_2 z^{-2})}$$

WHERE $p_2 = p_1^*$ (OTHERWISE WE GET COMPLEX COEFFICIENTS a_i)



USING EULER WE CAN REWRITE THE POLES INTO j

$$p_1 = r(\cos\theta + j\sin\theta)$$

$$p_2 = r(\cos\theta - j\sin\theta)$$

$$H(z) = \frac{z^2}{(z-p_1)(z-p_2)}$$

$$= \frac{z^2}{(z-r(\cos\theta+j\sin\theta))(z-r(\cos\theta-j\sin\theta))}$$

$$= \frac{z^2}{z^2 + r^2\cos^2\theta - 2zr\cos\theta + r^2\sin^2\theta}$$

$$= \frac{z^2}{z^2 + r^2 - 2zr\cos\theta}$$

$$= \frac{1}{1 - 2r\cos\theta \cdot z^{-1} + r^2 z^{-2}} = \frac{Y(z)}{X(z)}$$

$$\Downarrow$$

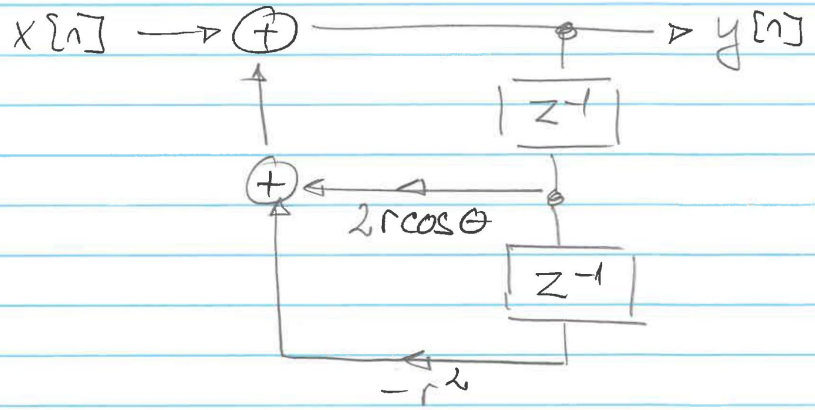
$$Y(z) = X(z) + Y(z)\left(2r\cos\theta\, z^{-1} - r^2 z^{-2}\right)$$

$$\Downarrow \; z^{-1}$$

$$y[n] = x[n] + 2r\cos\theta\, y[n-1] - r^2 y[n-2]$$

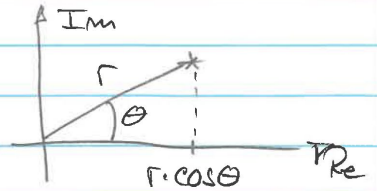
$$y[n] = x[n] + 2r\cos\theta\, y[n-1] - r^2 y[n-2]$$

LET'S NOW USE A DIRECT FORM II



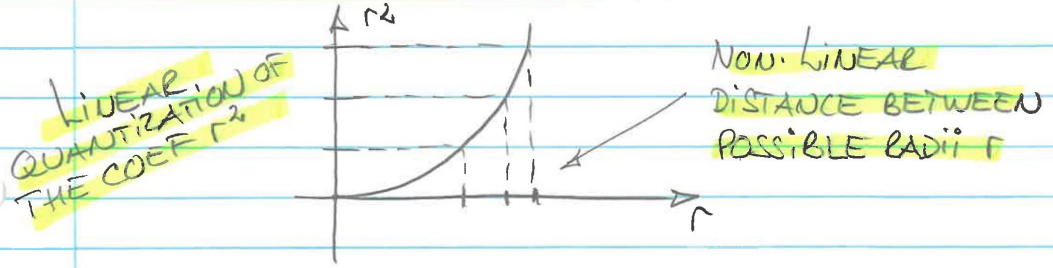
IN THIS STRUCTURE, THE COEFFICIENTS a_1 AND a_2 ARE $2r\cos\theta$ AND $-r^2$ WHICH NOW MUST BE QUANTIZED

- $r\cdot\cos\theta = \mathrm{Re}\{p_1\} = \mathrm{Re}\{p_2\}$

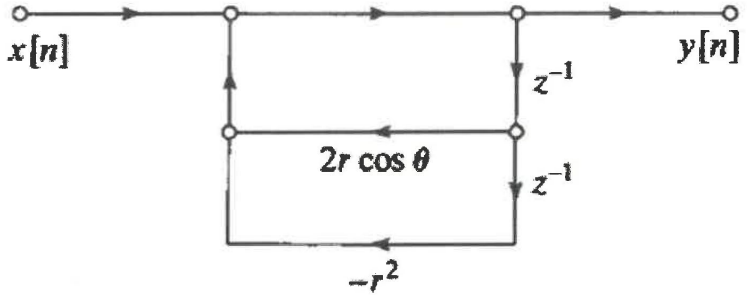


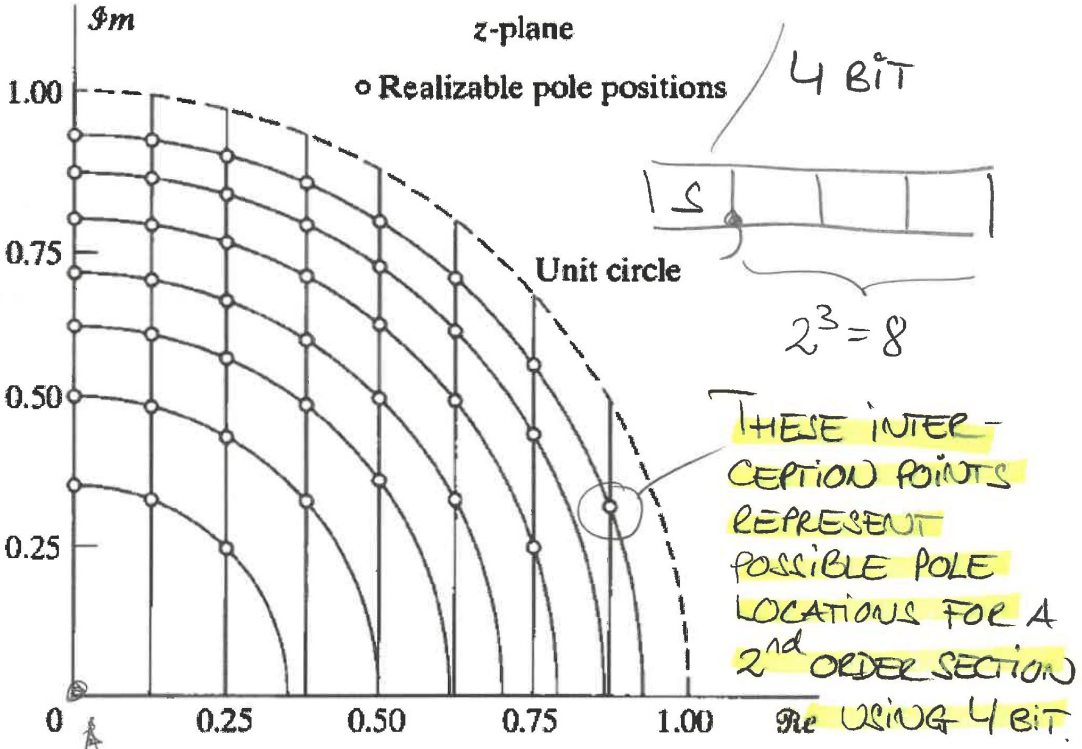
GIVEN N BIT, WE CAN QUANTIZE THE REAL AXIS (POSITIVE AND NEGATIVE) INTO 2^N STEPS, EACH EQUAL TO 2^{-N}

- r^2 IS THE POLE RADII SQUARED



DIRECT FORM STRUCTURE





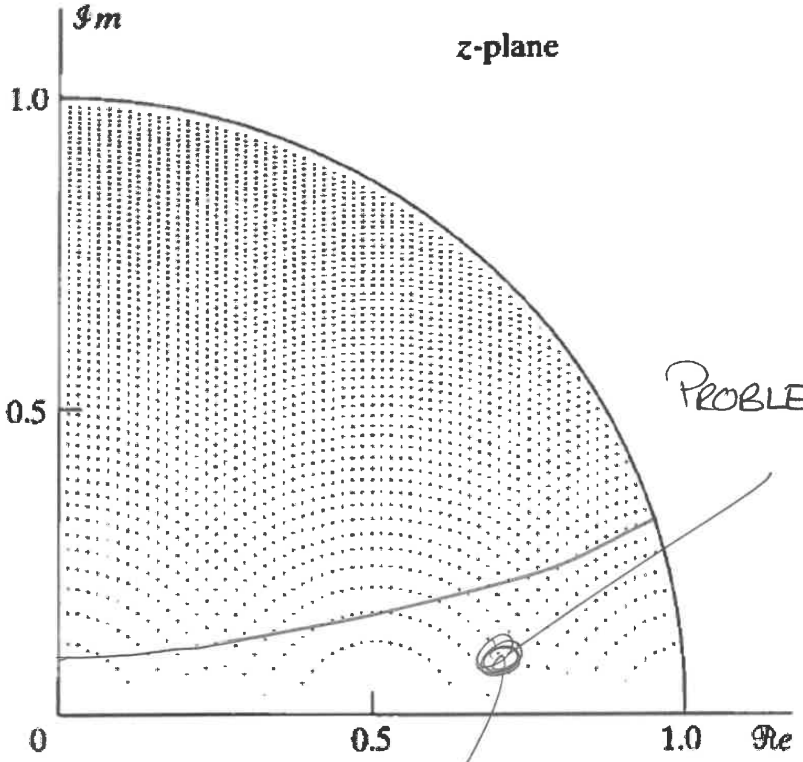
4 BIT

$2^3 = 8$

THESE INTER-CEPTION POINTS REPRESENT POSSIBLE POLE LOCATIONS FOR A 2nd ORDER SECTION USING 4 BIT.

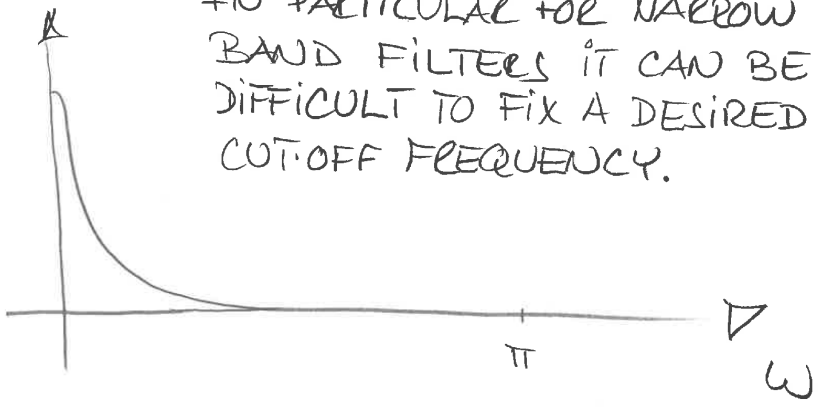
ALTHOUGH NOT RELEVANT FOR AN IIR-FILTER THE POLE CAN ALSO BE LOCATED IN Z=0.

POSSIBLE POLE LOCATIONS FOR 7 BIT

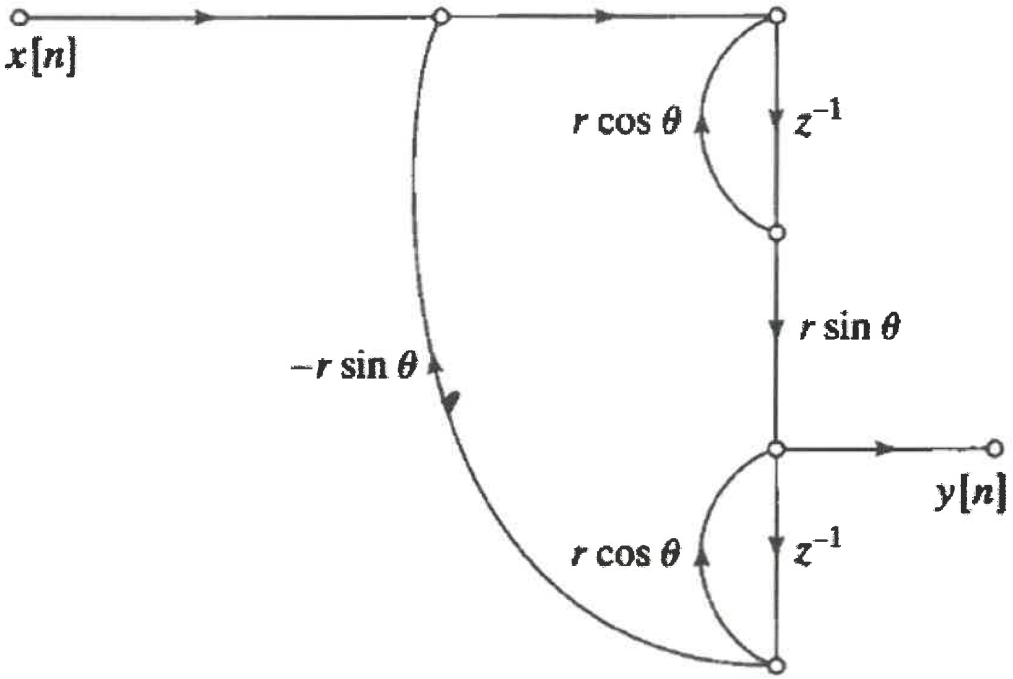


PROBLEMATIC REGION.

IN PARTICULAR FOR NARROW BAND FILTERS IT CAN BE DIFFICULT TO FIX A DESIRED CUT-OFF FREQUENCY.



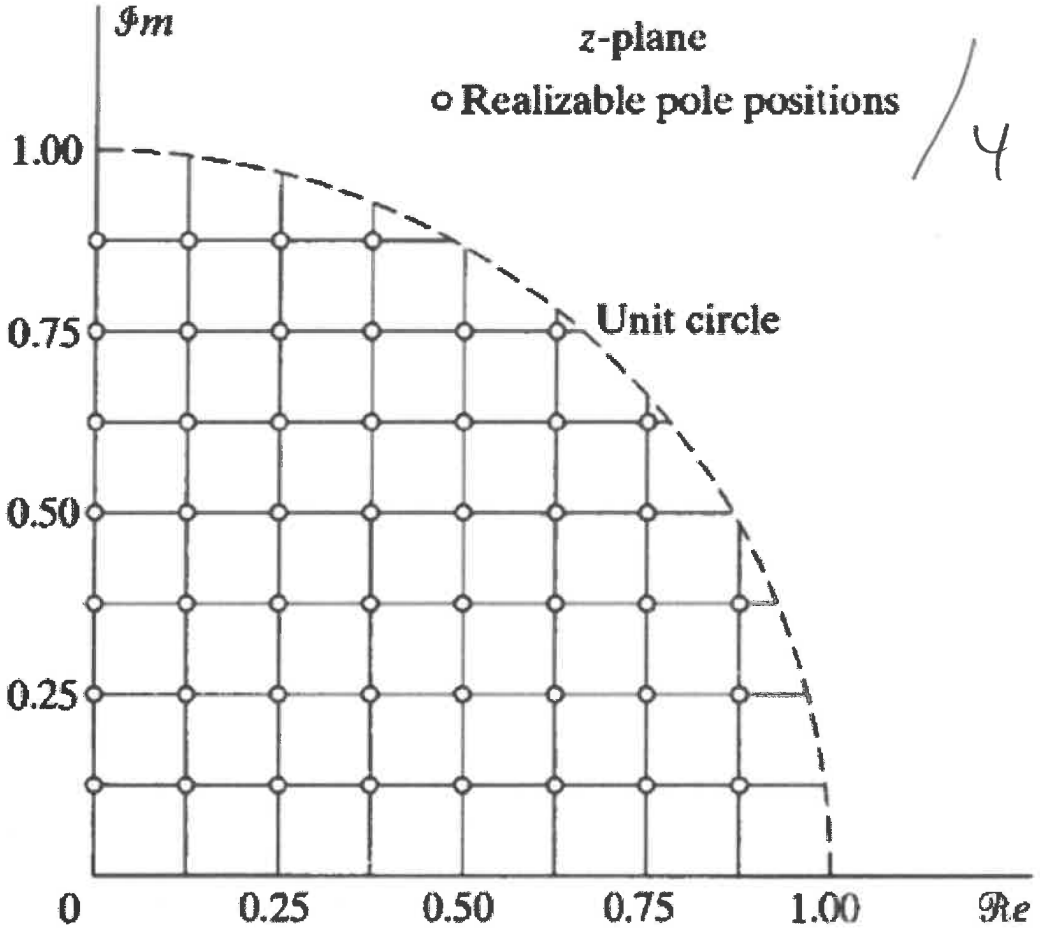
COUPLED FORM



- EXERCISE

SHOW THAT THE POLES (FLOATING POINT) ARE IDENTICAL IN THE DIRECT FORM AND IN THE COUPLED FORM.

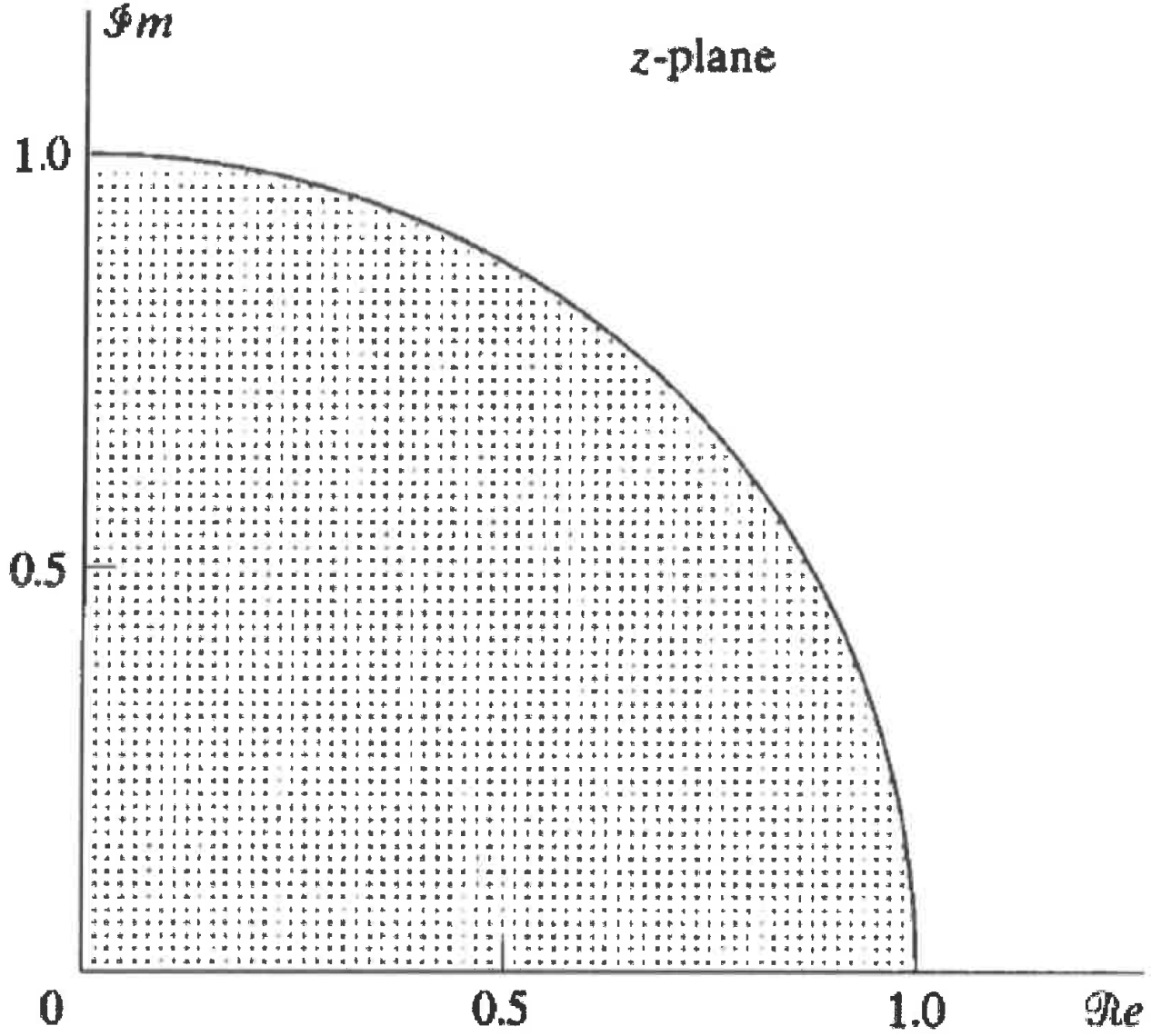
COUPLED FORM



4 BIT

NOW POLE LOCATIONS IN A REGULAR GRID WITHIN THE UNIT CIRCLE

COUPLED FORM / 7 BIT

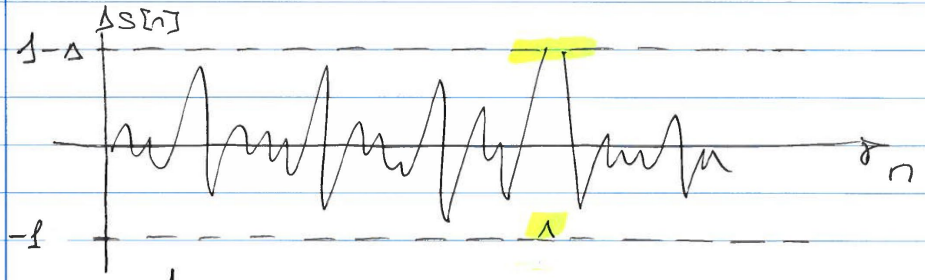


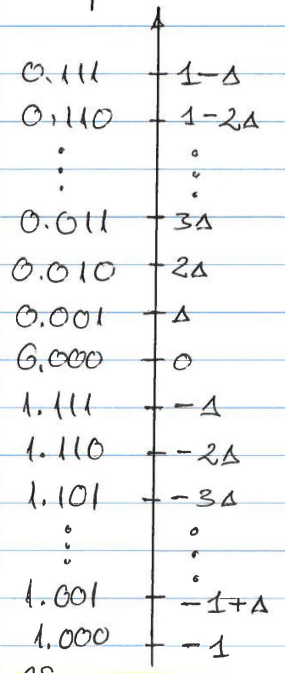
Scaling

Scaling

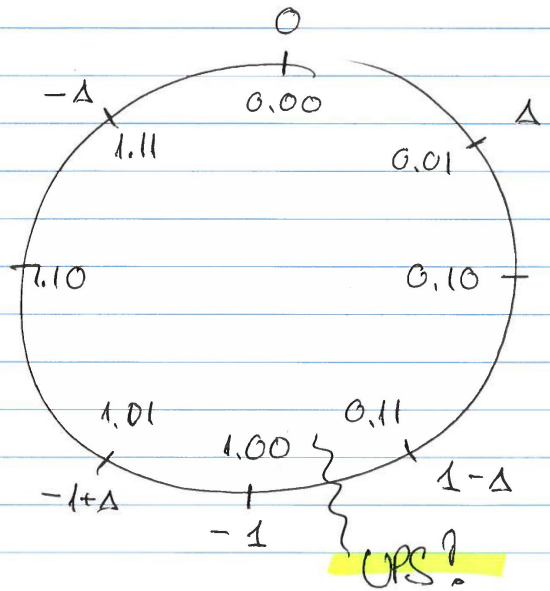
No matter whether we are using DF-I or DF-II, we have to ensure that our implementation of the filter provides the largest possible SNR, without Overflow

What is Overflow?



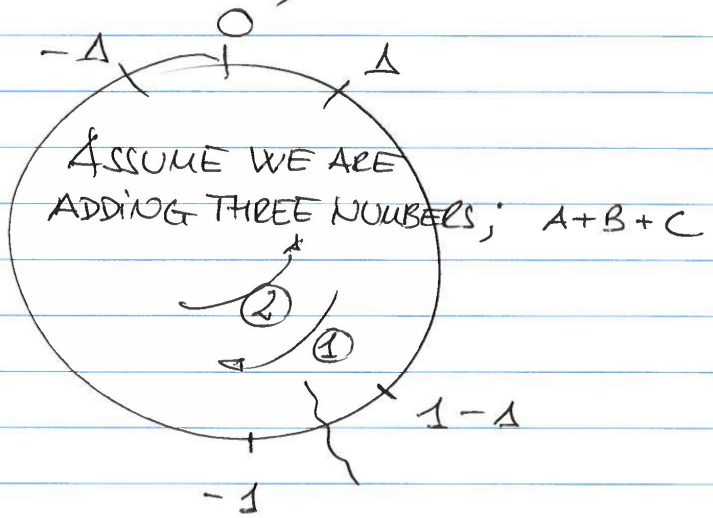


2's complement.



OVERFLOW IS POSSIBLE ONLY WHEN ADDING TWO NUMBERS WITH THE SAME SIGN

THEREFORE, 2's COMPLEMENT IS VERY USEFUL COMBATING OVERFLOW;



① $|A+B| \geq 1 \Rightarrow$ OVERFLOW

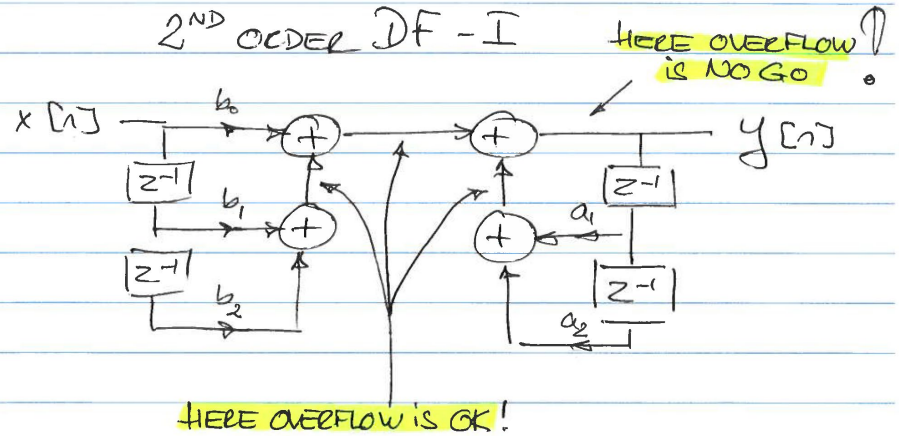
② $|(A+B)+C| < 1 \Rightarrow$ NO OVERFLOW.

THUS; THERE MAY BE SOME INTERNAL VARIABLES IN OUR FILTER STRUCTURE WHERE THE PARTIAL SUMS ARE ALLOWED TO OVERFLOW, AS LONG AS WE ENSURE THAT THE TOTAL SUM DOES NOT OVERFLOW

AN ALTERNATIVE FORMULATION;

WE WANT TO DESIGN OUR FILTER WITH BEST POSSIBLE UTILIZATION OF THE DYNAMIC RANGE (I.E., BEST SNR) OF THE VARIABLES, AND AT THE SAME TIME REDUCE THE RISC OF OVERFLOW IN THOSE VARIABLES WHERE OVERFLOW IS NOT ALLOWED

2^{ND} ORDER DF-I



" ... REDUCE THE RISC OF OVERFLOW ... "

OUR PROBLEM IS THAT 0% RISC AND A HIGH SNR (FULL UTILIZATION OF THE DYNAMIC RANGE) ARE CONFLICTING REQUIREMENTS

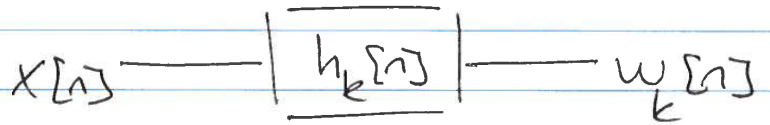
THREE POSSIBLE SCALING STRATEGIES

PP. 463-465

1) MAX-VALUE SCALING

THE IDEA IS TO SCALE THE INPUT SIGNAL SUCH THAT OVERFLOW DOES NOT OCCUR IN THOSE VARIABLE WHERE OVERFLOW IS NOT ALLOWED.

— SUCH A VARIABLE IS DENOTED w_k AND THE SEQUENCE IN THIS VARIABLE IS $w_k[n]$;



WHERE $h_k[n]$ IS THE IMPULSE RESPONSE FROM THE k INPUT TO THE VARIABLE w_k.

$$|w_k[n]| = \left| \sum_{m=-\infty}^{\infty} h_k[m]\, x[n-m] \right|$$

NOW, ASSUME THAT WE KNOW THE NUMERICAL LARGEST VALUE IN $x[n]$ — x_{max}

$$|w_k[n]| = \left| \sum_{m=-\infty}^{\infty} h_k[m] \cdot x[n-m] \right|$$

$$\Downarrow$$

$$|w_k[n]| \le x_{max} \left| \sum_{m=-\infty}^{\infty} h_k[m] \right| \le x_{max} \cdot \sum_{m=-\infty}^{\infty} |h_k[m]|$$

$$< 1 \quad \text{Q}$$

REQUIREMENT: (NO OVERFLOW)

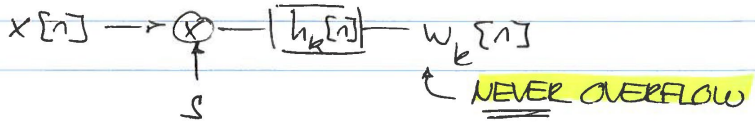
$$\Downarrow$$

$$x_{max} < \frac{1}{\sum_{m=-\infty}^{\infty} |h_k[m]|}$$

WHICH SHOULD BE TRUE FOR ALL VARIABLES w_k WHERE OVERFLOW IS NOT ALLOWED

IF THIS IS NOT THE CASE, THEN WE CAN SCALE THE INPUT SIGNAL (MULTIPLY WITH S)

$$S \cdot x_{max} < \frac{1}{\max_k \left\{ \sum_{-\infty}^{\infty} |h_k[m]| \right\}}$$

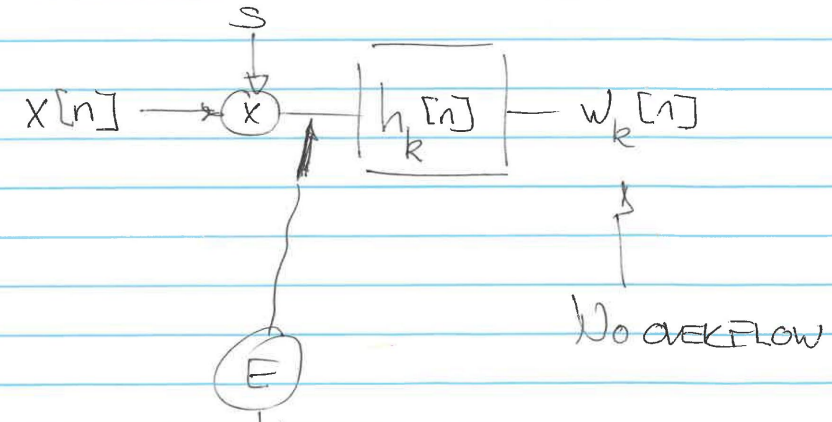


AN ISSUE WITH THE MAX-VALUE SCALING STRATEGY IS THAT S MAY BECOME SO SMALL THAT IT COMPROMISES THE SNR OF THE FILTER.

2) SINUSOID SCALING (USED FOR NARROW BAND SIGNALS — SELF STUDY)

3) VARIANCE SCALING

THE IDEA HERE IS TO ADJUST THE ENERGY (VARIANCE) IN THE INPUT SIGNAL SUCH THAT THERE IS NO OVERFLOW IN THE VARIABLE WHERE OVERFLOW IS NOT ALLOWED



THE ENERGY IS THE POWER INTEGRATED OVER TIME.

$$w_k[n] = h_k[n] * x[n]$$

$$\Downarrow$$

$$W(e^{j\omega}) = H_k(e^{j\omega}) \cdot X(e^{j\omega})$$

WE CAN NOW USE THE IDTFT ON THE RIGHT-HAND SIDE :

$$w_k[n] = \text{IDTFT}\left\{W(e^{j\omega})\right\} = \text{IDTFT}\left\{H_k(e^{j\omega}) \cdot X(e^{j\omega})\right\}$$

$$\Downarrow$$

$$w_k[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} H_k(e^{j\omega}) \cdot X(e^{j\omega}) \cdot e^{j\omega n}\, d\omega$$

$$\Downarrow$$

$$|w_k[n]|^2 = \left|\frac{1}{2\pi}\int_{-\pi}^{\pi} H_k(e^{j\omega}) \cdot X(e^{j\omega}) \cdot e^{j\omega n}\, d\omega\right|^2$$

$$\Downarrow$$

$$|w_k[n]|^2 \leq \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi} \left|H_k(e^{j\omega}) \cdot X(e^{j\omega})\, e^{j\omega n}\right|^2 d\omega$$

SCHWARTZ

$$= \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi} \left|H_k(e^{j\omega})\right|^2 \left|X(e^{j\omega})\right|^2 d\omega$$

$$\leq \left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \left|H_k(e^{j\omega})\right|^2 d\omega\right) \cdot \left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \left|X(e^{j\omega})\right|^4 d\omega\right)$$

Equ. 125 p. 464 (without proof here)

(*)

$$|W_k[n]|^2 \leq \left(\frac{1}{2\pi}\int_{-\pi}^{\pi} |H_k(e^{j\omega})|^2 \, d\omega\right) \cdot \left(\frac{1}{2\pi}\int_{-\pi}^{\pi} |X(e^{j\omega})|^2 \, d\omega\right)$$

NOW, LET'S INTRODUCE PARSEVAL'S THEOREM;

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

WE NOW USE PARSEVAL IN (*)

$$|W_k[n]|^2 \leq \sum_{n=-\infty}^{\infty} |x[n]|^2 \cdot \frac{1}{2\pi}\int_{-\pi}^{\pi} |H_k(e^{j\omega})|^2 \, d\omega$$

OUR REQUIREMENT WAS;

"ADJUST THE ENERGY IN THE INPUT-SIGNAL $x[n]$ SUCH THAT THERE IS NO OVERFLOW IN $W_k[n]$"

$$\Downarrow$$

$$|W_k[n]|^2 < 1$$

$$\Downarrow$$

$$\sum_{n=-\infty}^{\infty} |s \cdot x[n]|^2 \cdot \frac{1}{2\pi}\int_{-\pi}^{\pi} |H_k(e^{j\omega})|^2 \, d\omega < 1$$

$$s^2 \sum_{n=-\infty}^{\infty} |x[n]|^2 \cdot \frac{1}{2\pi} \int_{-\pi}^{\pi} |H_k(e^{j\omega})|^2 \, d\omega < 1$$

$$s^2 \cdot E_x < \frac{1}{\frac{1}{2\pi} \int_{-\pi}^{\pi} |H_k(e^{j\omega})|^2 \, d\omega} = \frac{1}{\sum_{n=-\infty}^{\infty} |h_k[n]|^2}$$

WHICH SHOULD HOLD FOR ALL VARIABLES WHERE OVERFLOW IS NOT ALLOWED.

IF WE ASSUME THAT THE ENERGY IN THE INPUT SIGNAL $E_x = 1$ THEN

$$s < \frac{1}{\sqrt{\sum_{n=-\infty}^{\infty} |h_k[n]|^2}}$$

$$\Downarrow$$

$$s = \frac{1}{\max_k \left\{ \sqrt{\sum_{n=-\infty}^{\infty} |h_k[n]|^2} \right\}}$$

EQU. 12.6 p. 464.