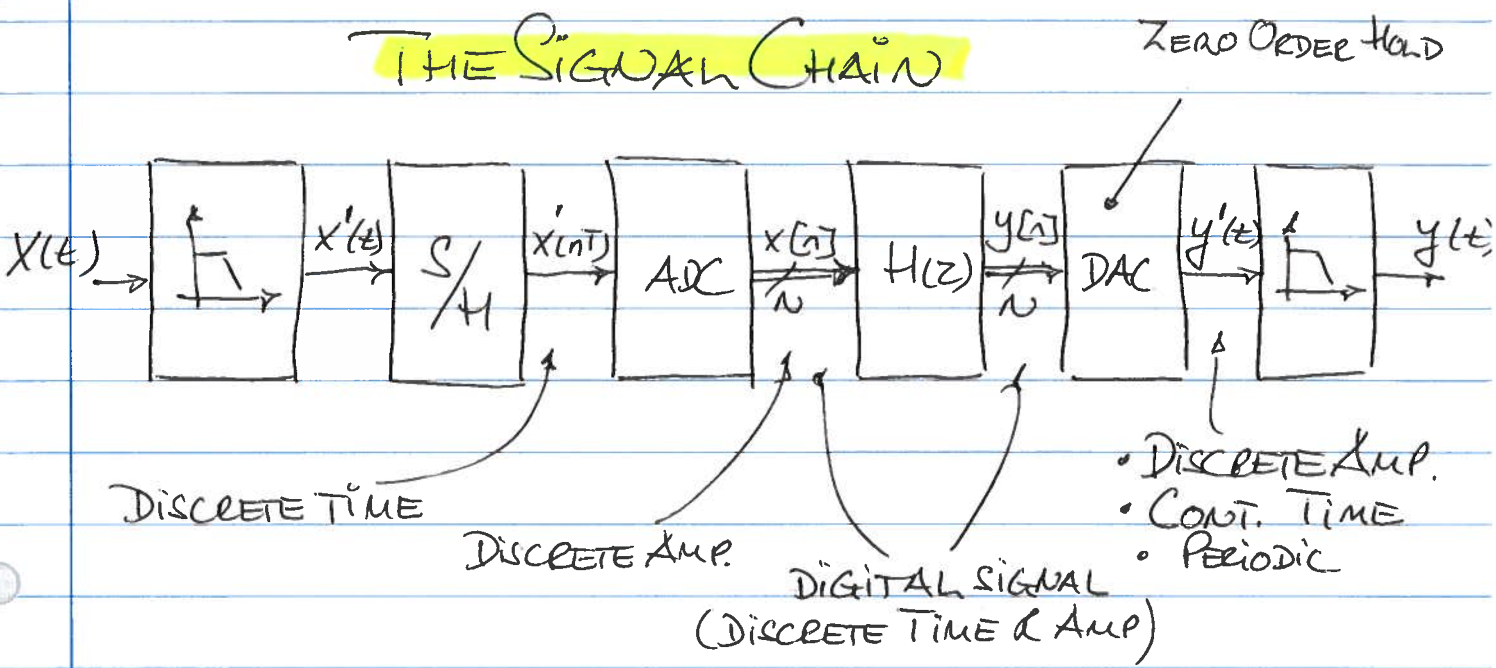
Signalbehandling for computer-ingeniører
COMTEK-5, E22
&
Signalbehandling
EIT-5, E22

10. Realization Structures for Digital Filters, and Finite Word Length Effects

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

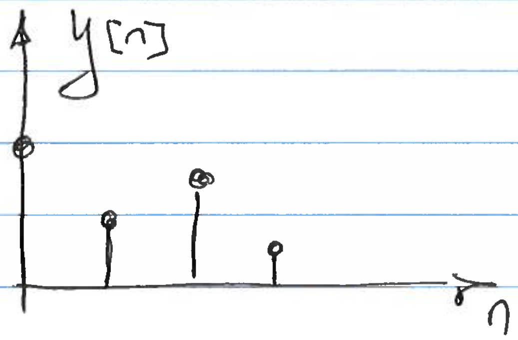
ADC : VARIABLE QUANTIZATION

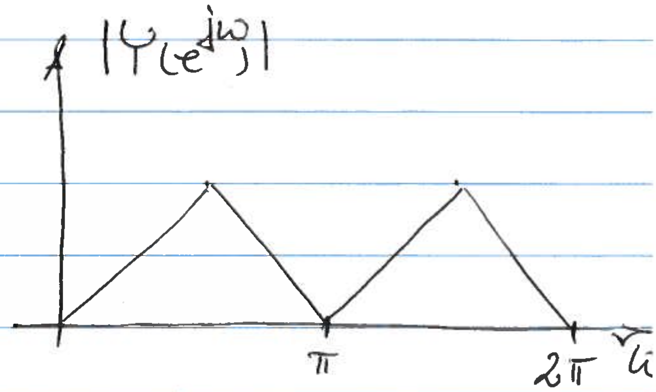
H(z) (COMPUTER) : VARIABLE + COEFFICIENT QUANT.

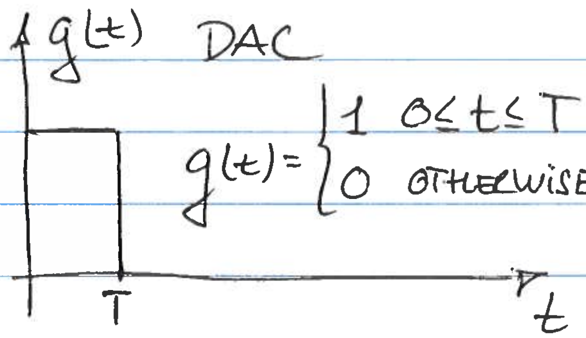
DAC : TIME QUANTIZATION.

Time Quantization

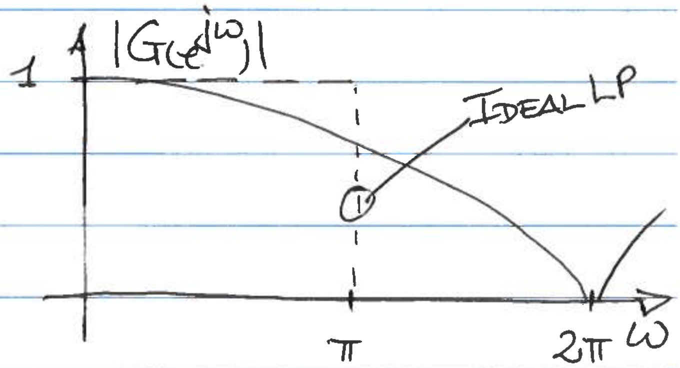
- $f_{sample} > 2 \cdot f_{max,\,analog}$ (Reconstruction).
- Always apply an Anti-Aliasing filter.
- Discrete time $\Rightarrow$ Periodic frequency.
- The DAC is not an ideal L.P. filter

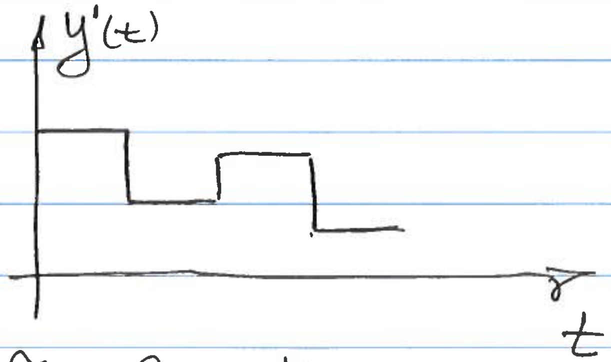




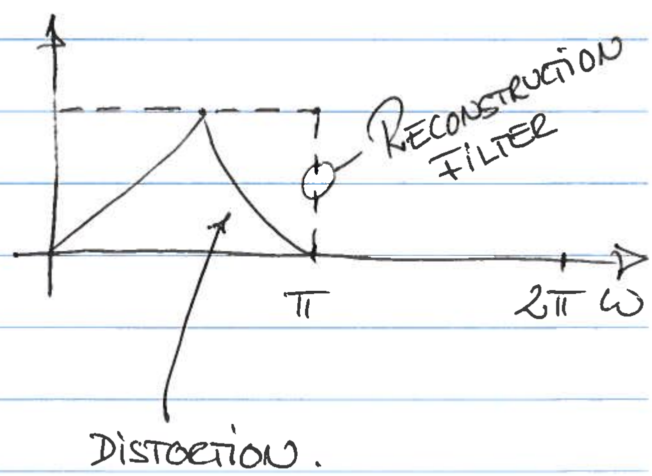


$$g(t) = \begin{cases} 1 & 0 \le t \le T \\ 0 & \text{otherwise} \end{cases}$$

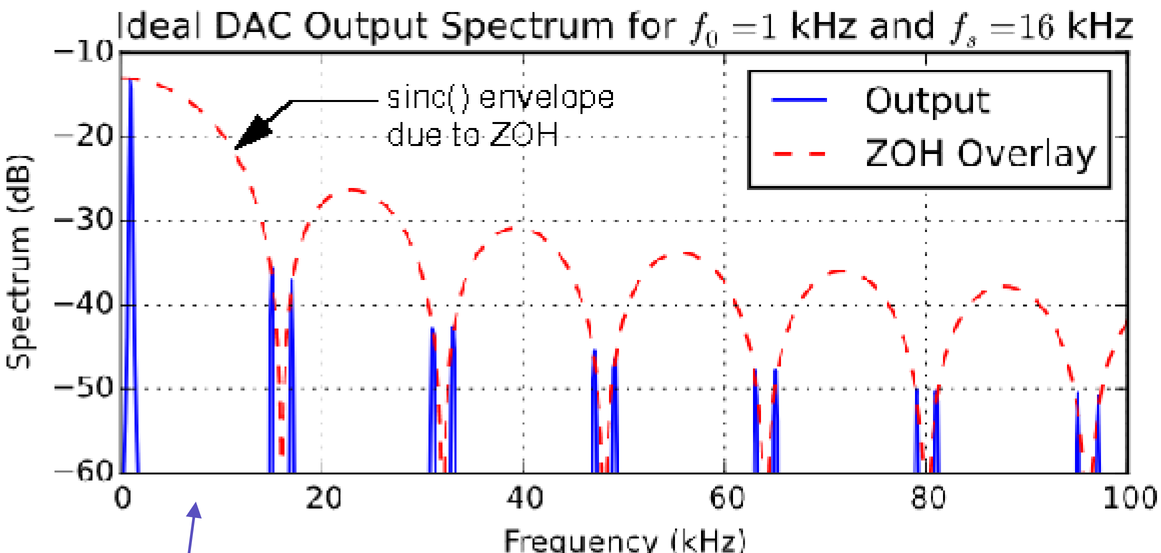




Zero Order Hold



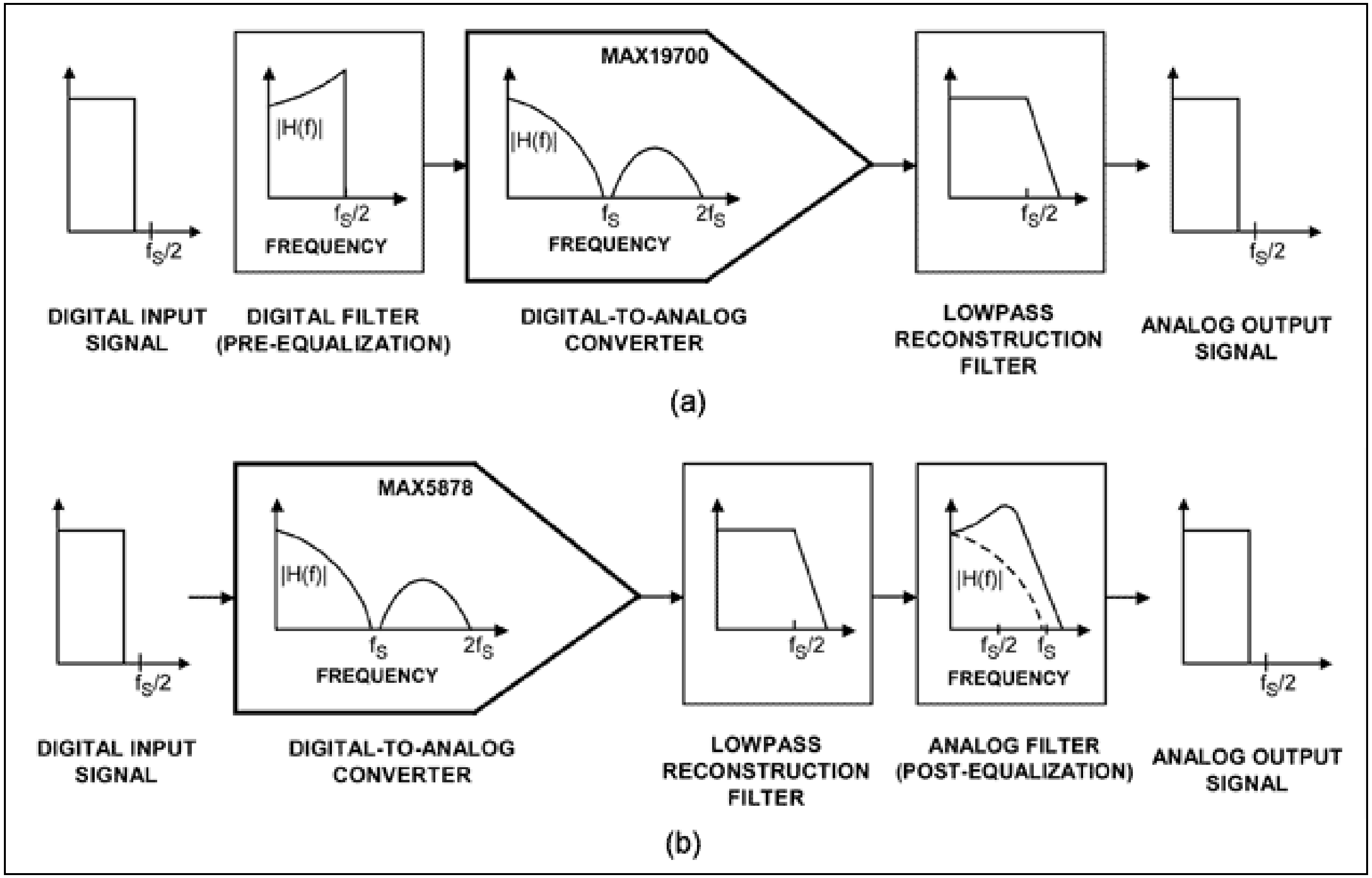
The Sinc shaped ZOH Overlay maintains all the 2π periodic images from the output signal $y[n]$.



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

AALBORG UNIVERSITY
DENMARK

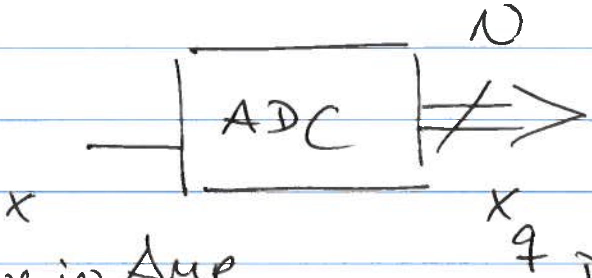
For frequency components close to the Nyquist frequency, the ZOH introduces distortion.
To compensate for this ZOH distortion we may apply either pre- or post-equalization



AALBORG UNIVERSITY
DENMARK

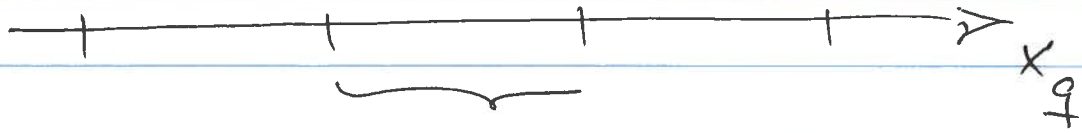
Variable Quantization

- 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)

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

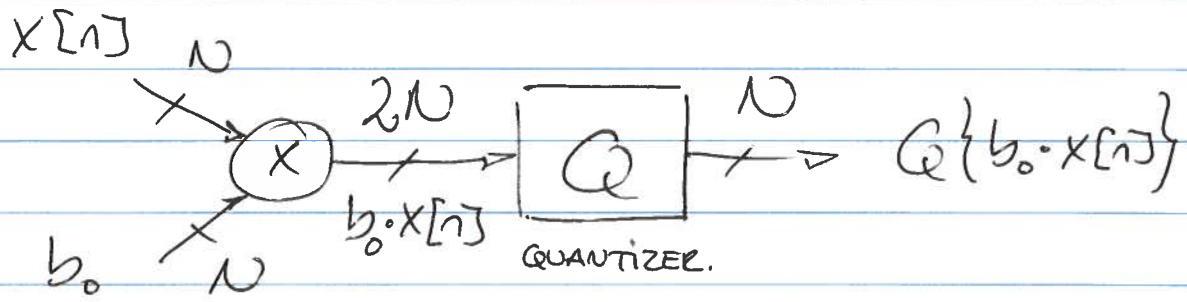
Without proof;

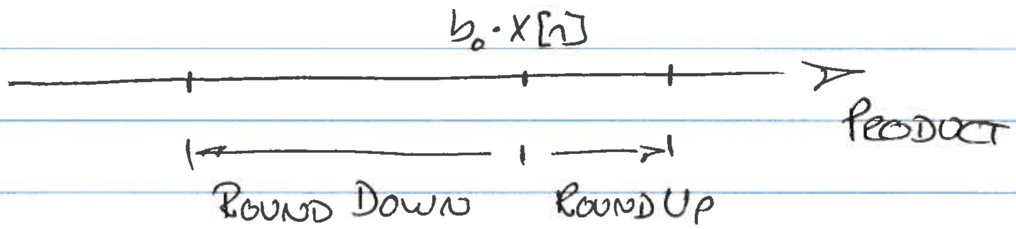
> The SNR on the output of the ADC is increased with 6 dB when N is increased with 1.
>
> $$SNR = 6.02N + 1.76 \text{ dB}$$

- $H(z)$ (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 ;

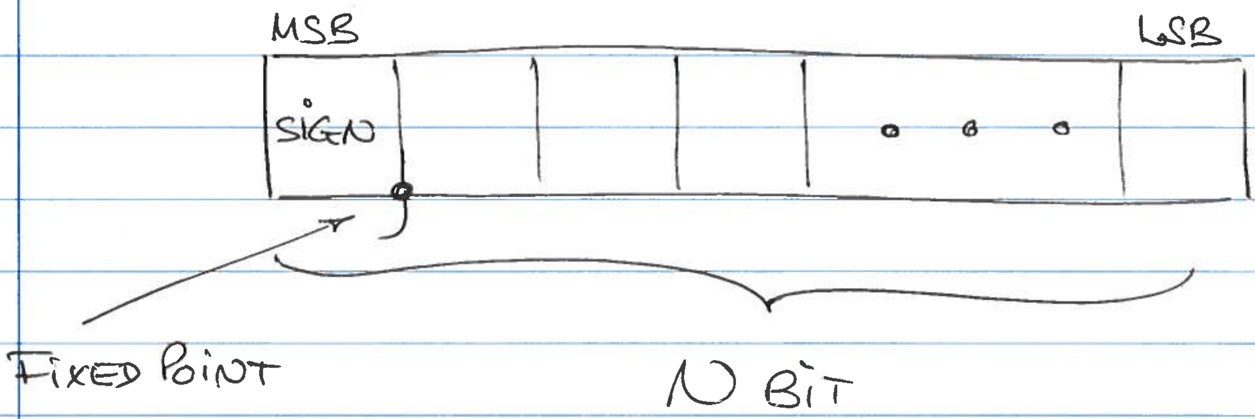
$$\boxed{e[n] = \underbrace{b_0 \cdot x[n]}_{2N \text{ BIT}} - \underbrace{Q\{b_0 \cdot x[n]\}}_{N \text{ BIT}}}$$

THE COMPUTATION ITSELF MAKES NOISE .

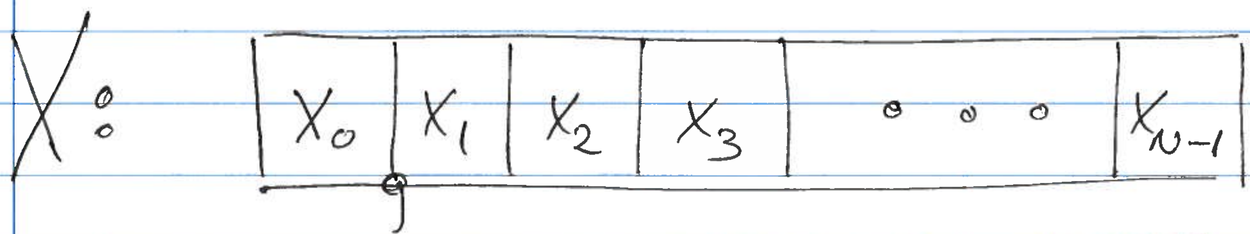
A FEW WORDS ON NUMBER REPRESENTATION.

- 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.



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

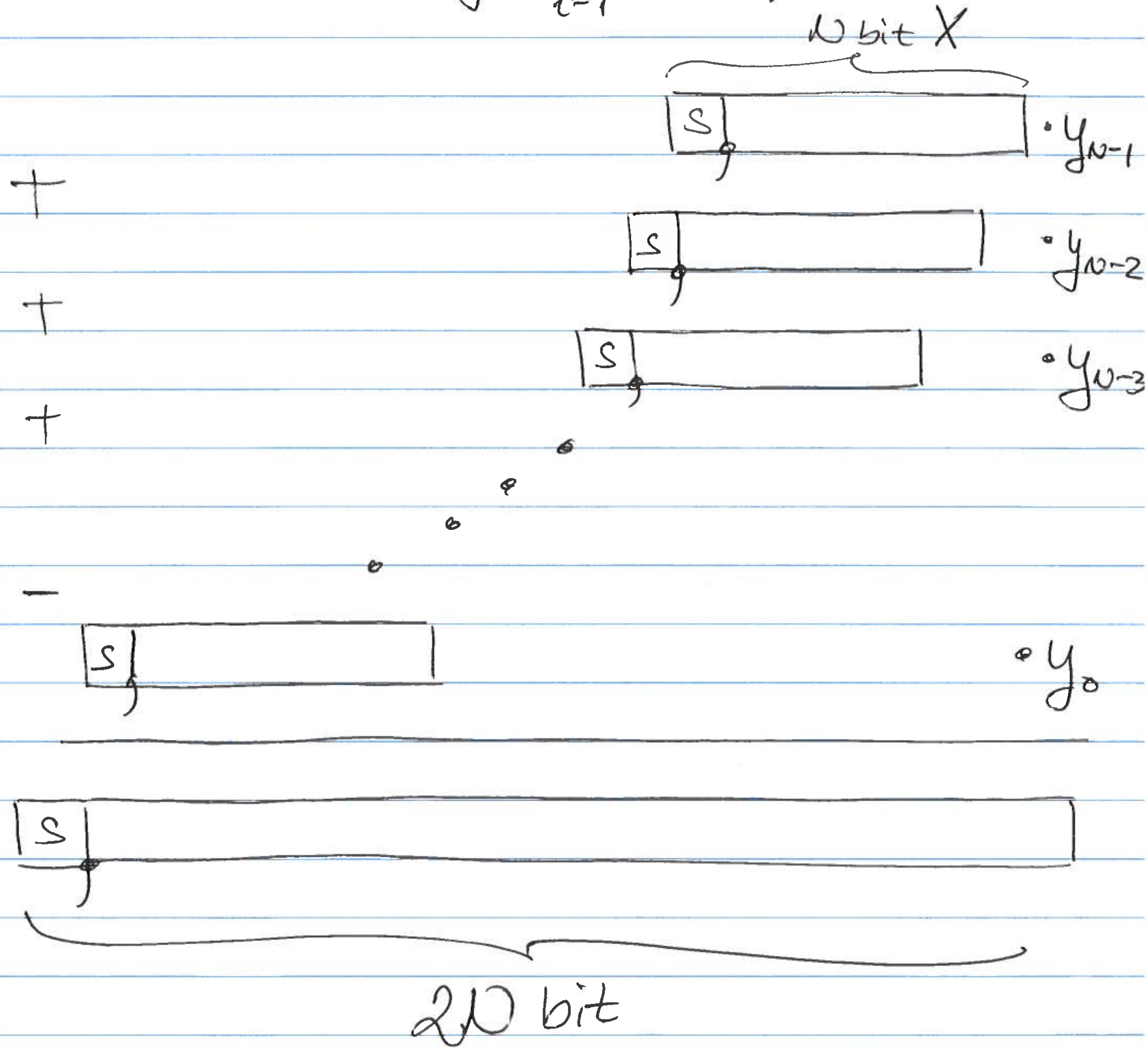


THE NUMERICAL VALUE OF X;

$$\boxed{X = -x_0 + \sum_{i=1}^{N-1} x_i \cdot 2^{-i}}$$

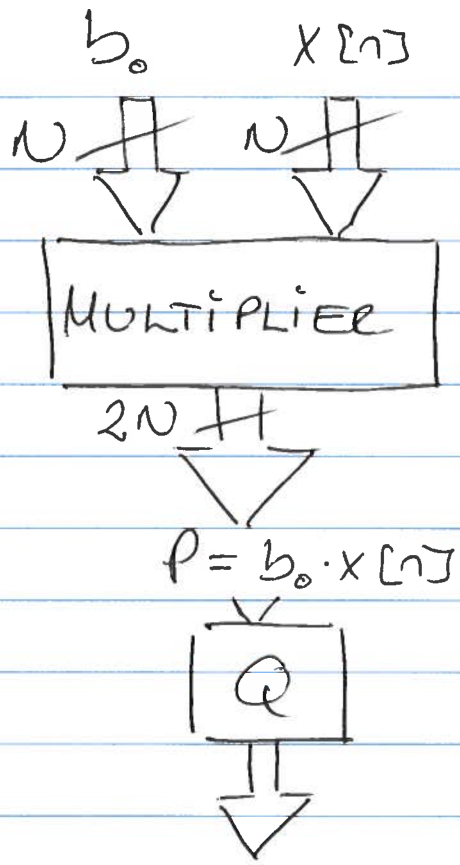
MULTIPLYING TWO N BIT NUMBERS;

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



- 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.



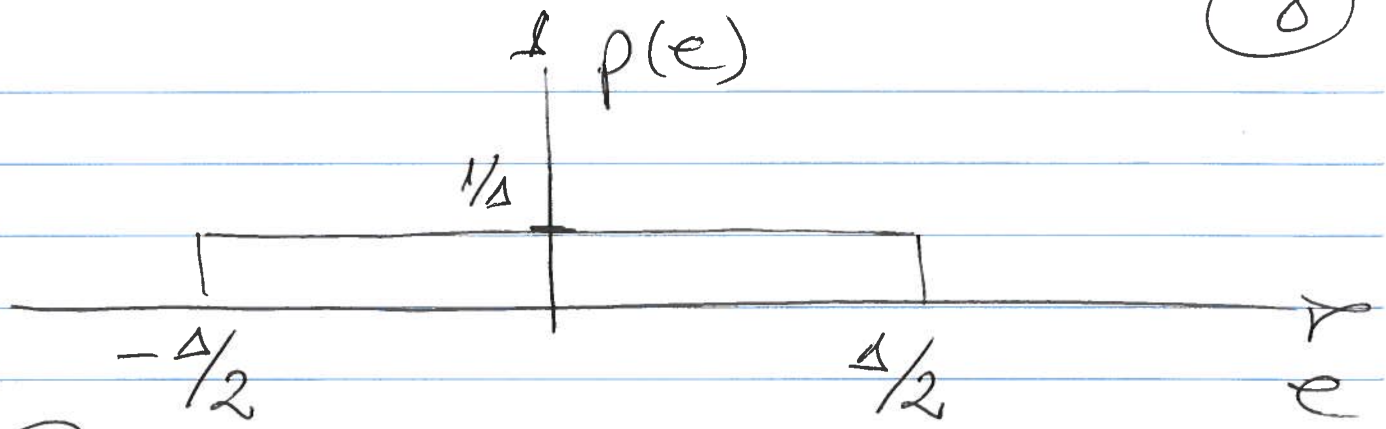
$$Q\{P\} = P - e[n]$$

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

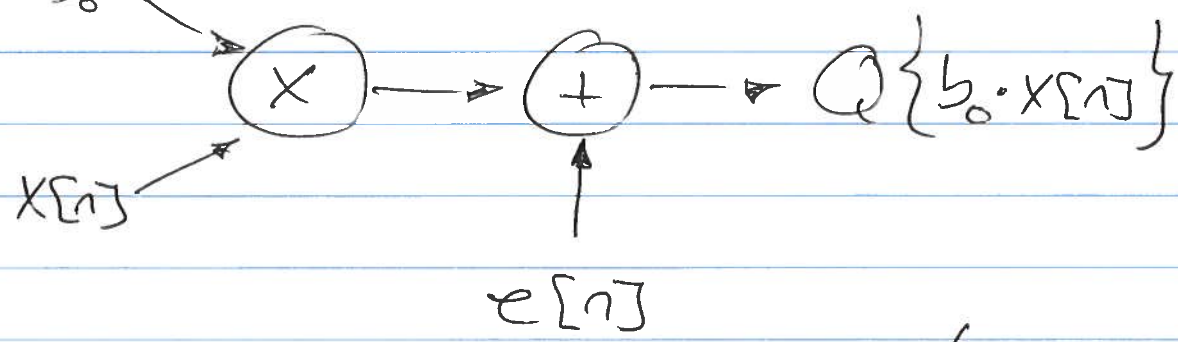
- MEAN

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

- 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

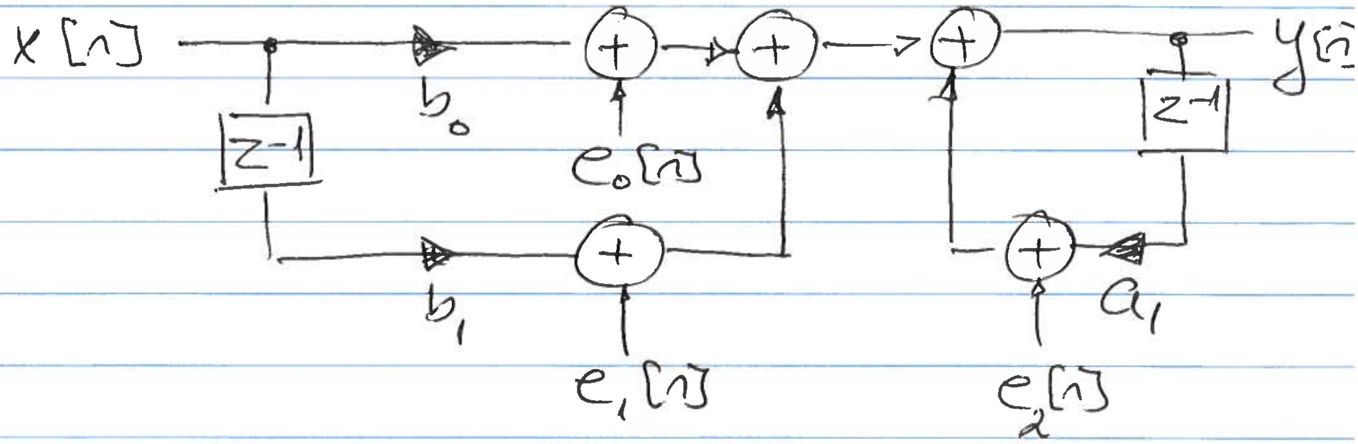


WHERE $e[n]$ IS A WHITE SIGNAL (μ_e, σ_e^2)

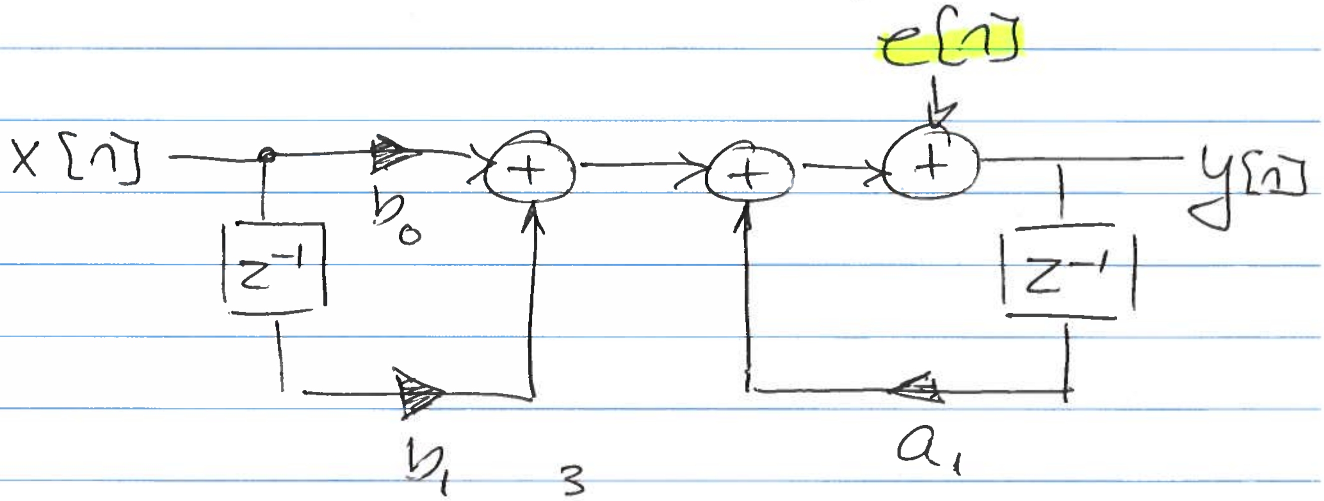
EXAMPLE: 1^{ST} ORDER IIR FILTER.

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

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 <u>NOT</u> INFINITE

COEFFICIENT QUANTIZATION (in $H(z)$)

$$x[n] \longrightarrow \bigotimes_{\downarrow\, b_0 + \Delta b_0} \longrightarrow \underbrace{b_0 \, x[n]}_{\text{EXACT}} + \underbrace{\Delta b_0 \cdot x[n]}_{\text{QUANTIZATION ERROR.}}$$

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

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

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).

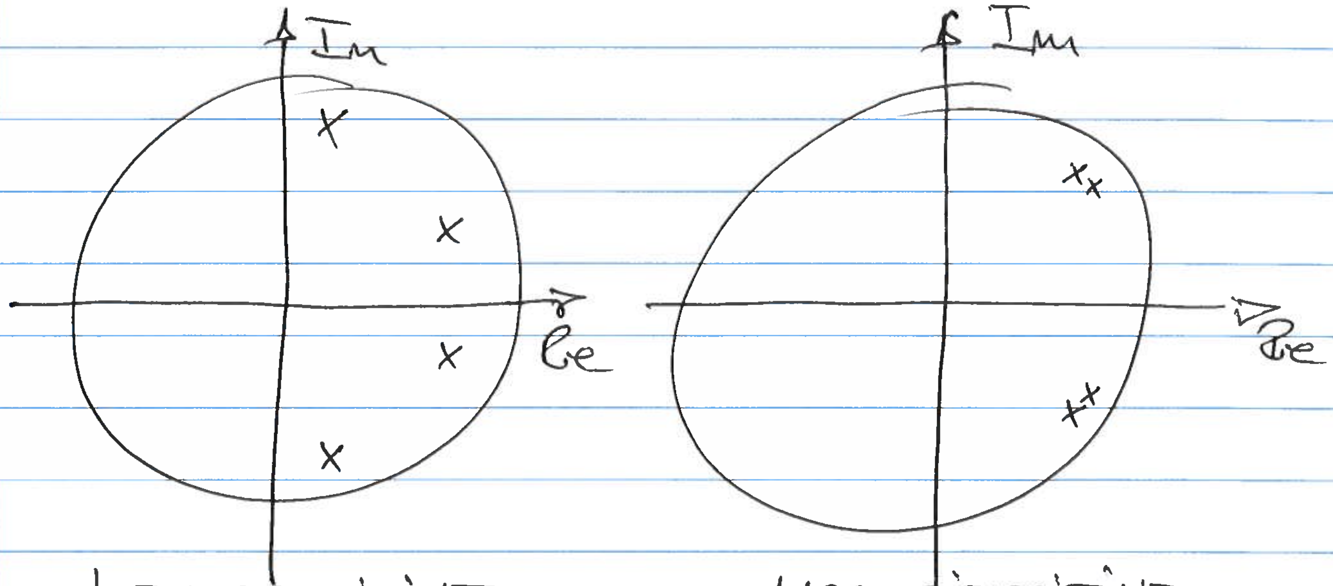
Basically, we want to know "the change in the poles as related to the change in the filter coefficients";

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

(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;

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

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

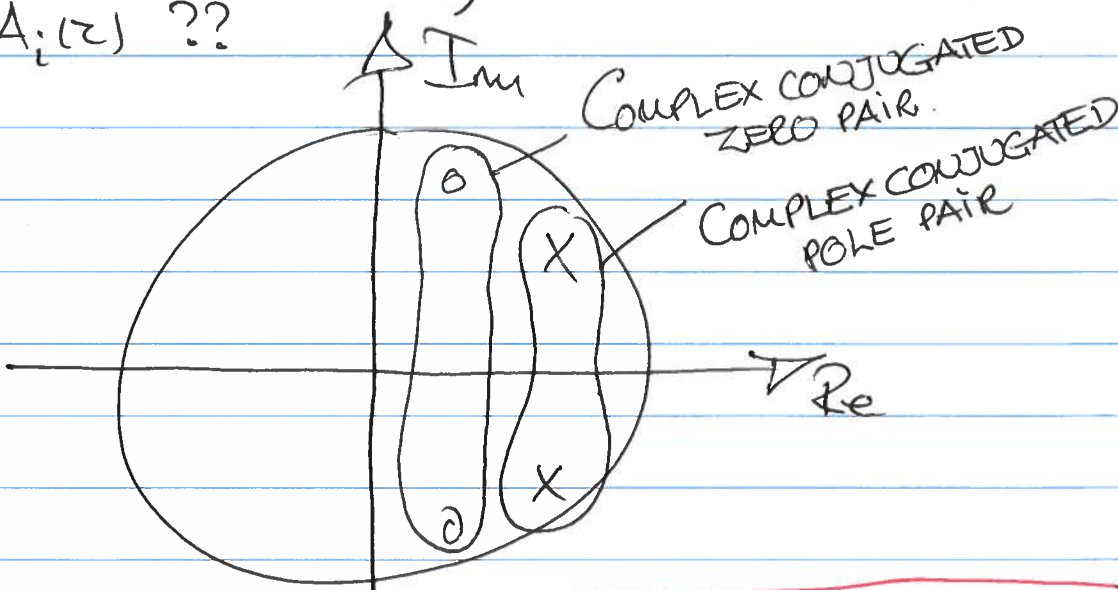
SO, THE TRICK IS TO RE-ORGANIZE $H(z)$ INTO A CASCADE OF 2^{ND} ORDER SECTIONS (EVT. ALSO ONE 1^{ST} ORDER SECTION).

$$\underbrace{- \boxed{H_1(z)} - \boxed{H_2(z)} - \boxed{H_3(z)} - \cdots - \boxed{H_{N/2}(z)} -}_{H(z)}$$

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

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

NOW, QUESTION IS: WHICH ZERO'S 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 $— \boxed{H_1(z)} — \boxed{H_2(z)} —$ IDENTICAL TO

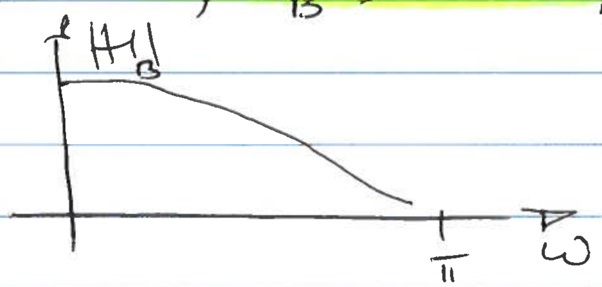
$— \boxed{H_2(z)} — \boxed{H_1(z)} —$?

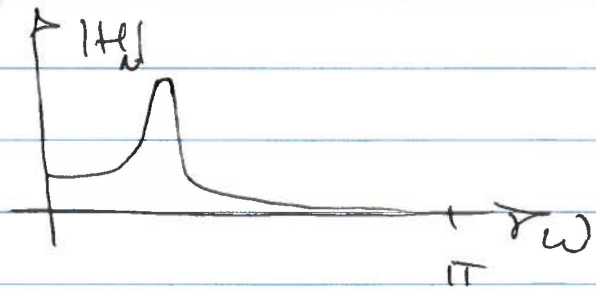
IF WE DO THE COMPUTATION IN INFINITE WORDLENGTH (FLOATING POINT), THEN THE ANSWER IS „YES".

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

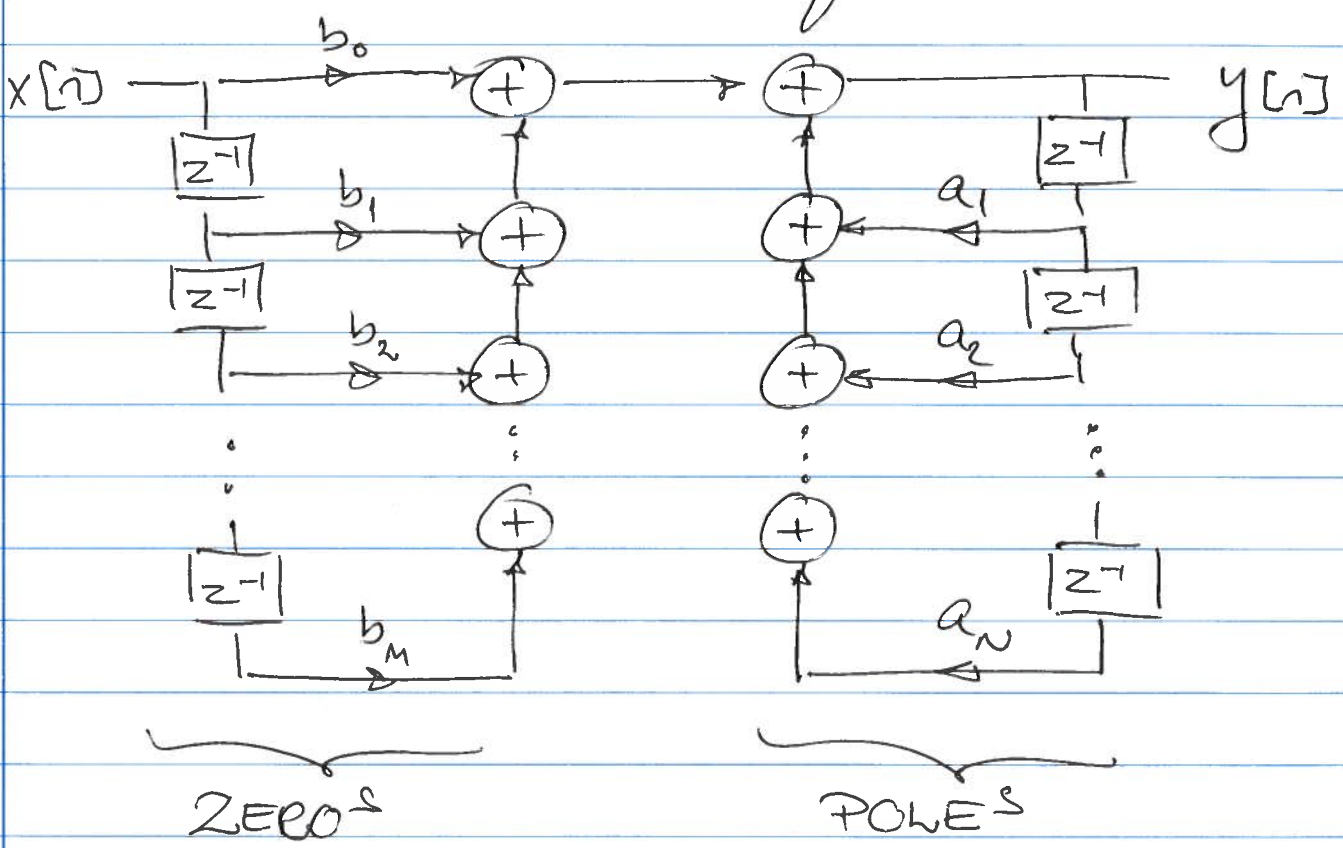
RECOMMENDED

REALISATION STRUCTURES

DIFFERENCE EQUATION ;

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

GRAPHICAL REPRESENTATION



THIS STRUCTURE IS KNOWN AS THE

DIRECT FORM I

BECAUSE IT IS A DIRECT REALISATION OF THE DIFFERENCE EQUATION

NOW, THE DF-I STRUCTURE CAN BE RE-ORGANIZED INTO WHAT IS KNOWN AS DF-II

$$X(z) - \boxed{H(z)} - Y(z) \Rightarrow X(z) - \underbrace{\boxed{H_1(z)}}_{\text{POLES}} \xrightarrow{W(z)} \underbrace{\boxed{H_2(z)}}_{\text{ZEROS}} - Y(z)$$

$$H(z) = H_1(z) \cdot H_2(z) = \frac{W(z)}{X(z)} \cdot \frac{Y(z)}{W(z)}$$

$\Downarrow$

$$H(z) = \left\{ \frac{1}{1 - \sum_{k=1}^{N} a_k z^{-k}} \right\} \cdot \left\{ \sum_{l=0}^{M} b_l z^{-l} \right\}$$

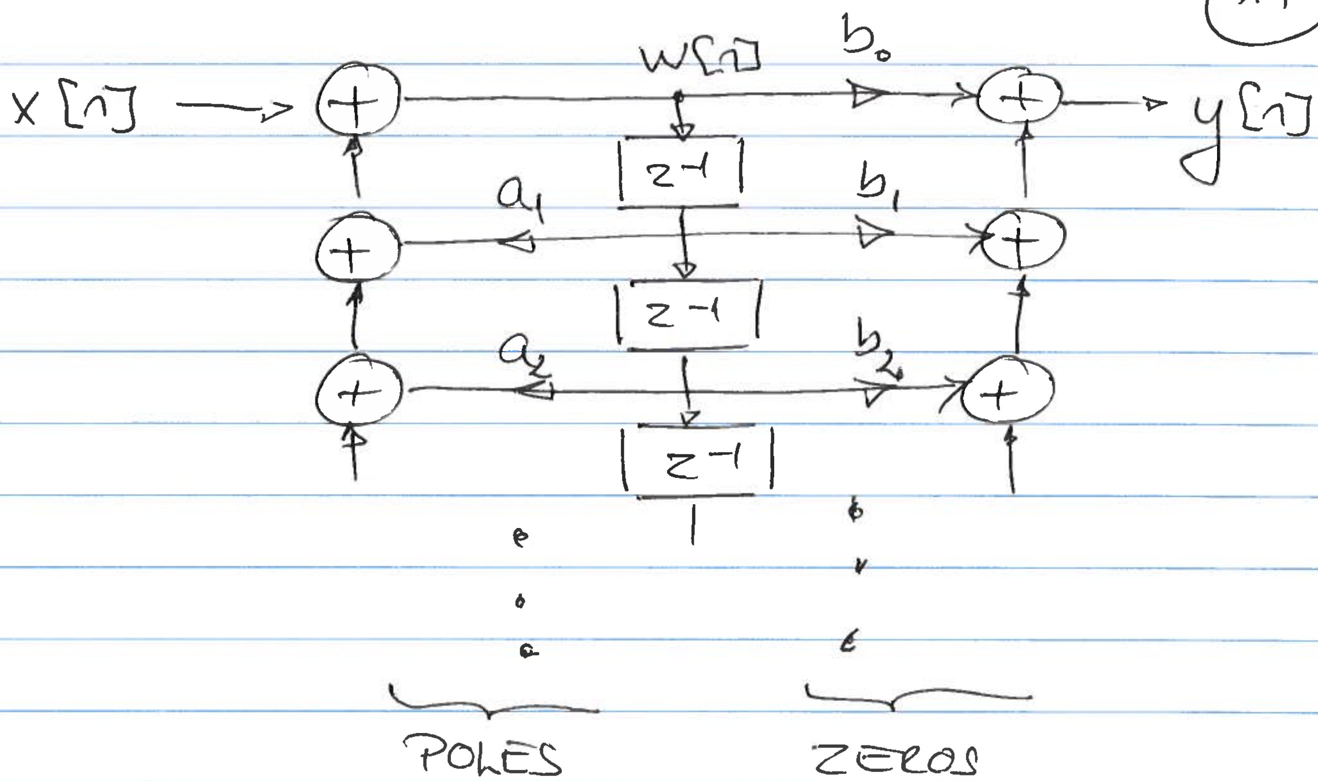
$\Downarrow$

$$\begin{cases} W(z) \cdot \left(1 - \sum_{k=1}^{N} a_k z^{-k}\right) = X(z) \\ W(z) \cdot \sum_{l=0}^{M} b_l \cdot z^{-l} = Y(z) \end{cases}$$

Z^{-1} $\Downarrow$

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

SO, THE INTERMEDIA VARIABLE $w[n]$ IS GENERATED AT THE INPUT AND IS USED TO GENERATE THE OUTPUT



DIRECT FORM II

- THE SAME NUMBER OF MULTIPLICATIONS
- THE SAME NUMBER OF ADDITIONS
- HALF THE NUMBER OF MEMORIES

… AS COMPARED TO DF-I

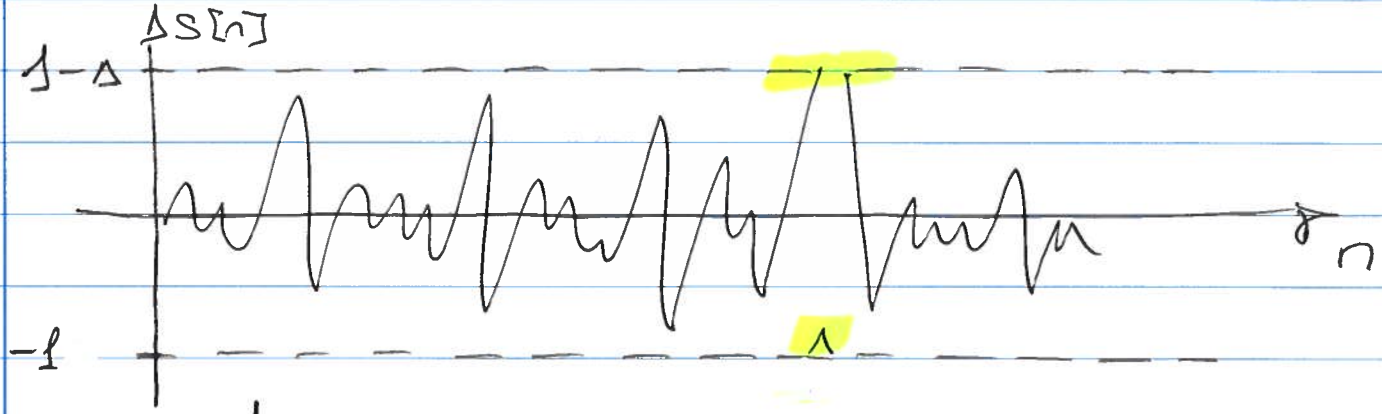
WHAT ABOUT THE NUMERICAL PROPERTIES OF THESE TWO FORMS ??

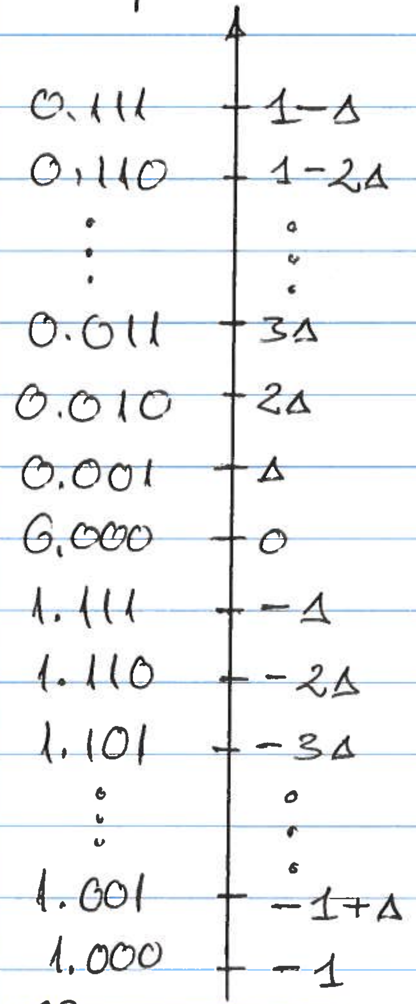
I'LL TELL YOU ABOUT THAT ON THE 6TH SEMESTER

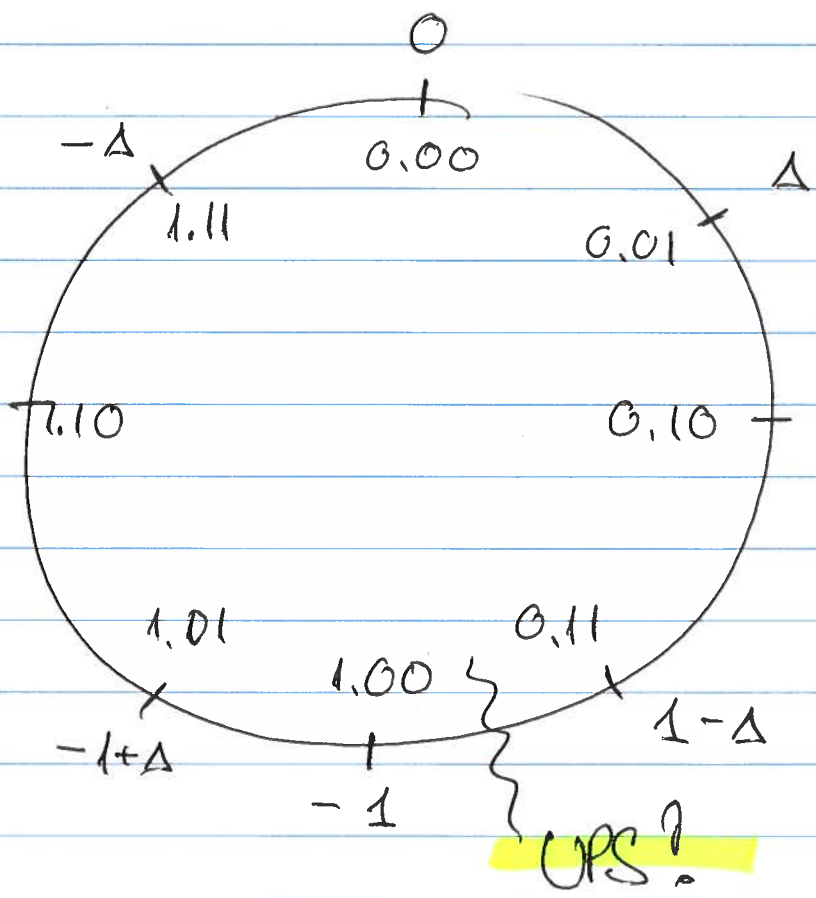
SCALING

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?



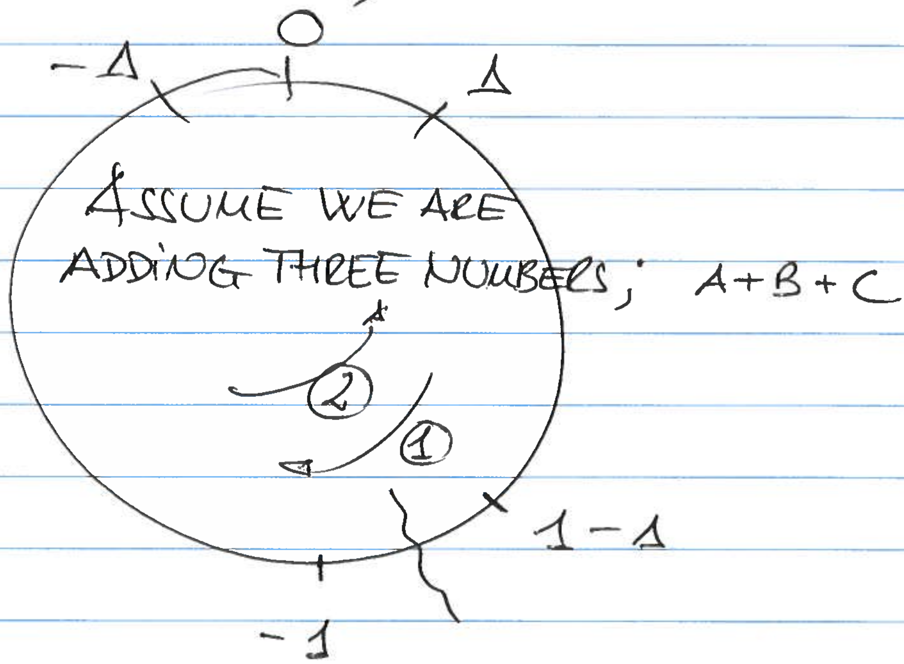




2^S COMPLEMENT.

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

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



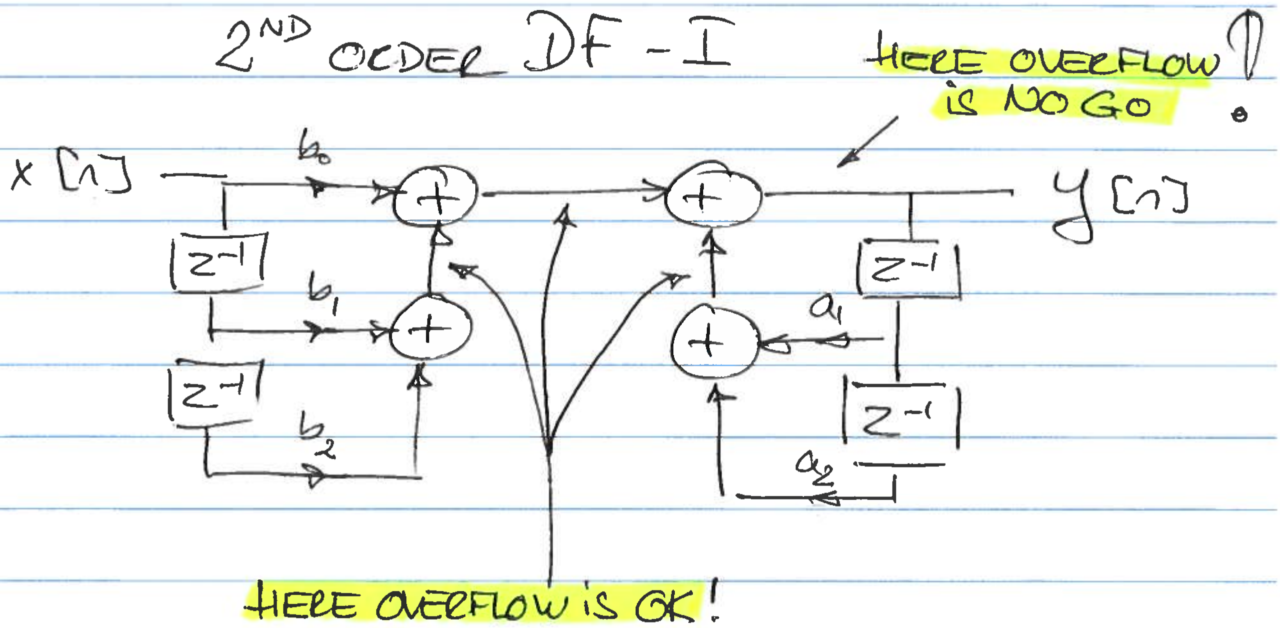
(1) $|A+B| \geq 1 \Rightarrow$ OVERFLOW

(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



" ... 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 $w_k[n]$;

$$x[n] \longrightarrow \boxed{h_k[n]} \longrightarrow w_k[n]$$

WHERE $h_k[n]$ IS THE IMPULSE RESPONSE FROM THE 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}

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

$$\Downarrow$$

$$\left| w_k[n] \right| \leq x_{max} \left| \sum_{m=-\infty}^{\infty} h_k[m] \right| \leq x_{max} \cdot \sum_{m=-\infty}^{\infty} \left| h_k[m] \right| < 1 \;!$$

REQUIREMENT.
(NO OVERFLOW)

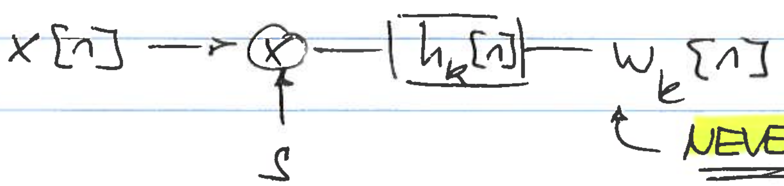
$$\Downarrow$$

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

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)

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



NEVER OVERFLOW

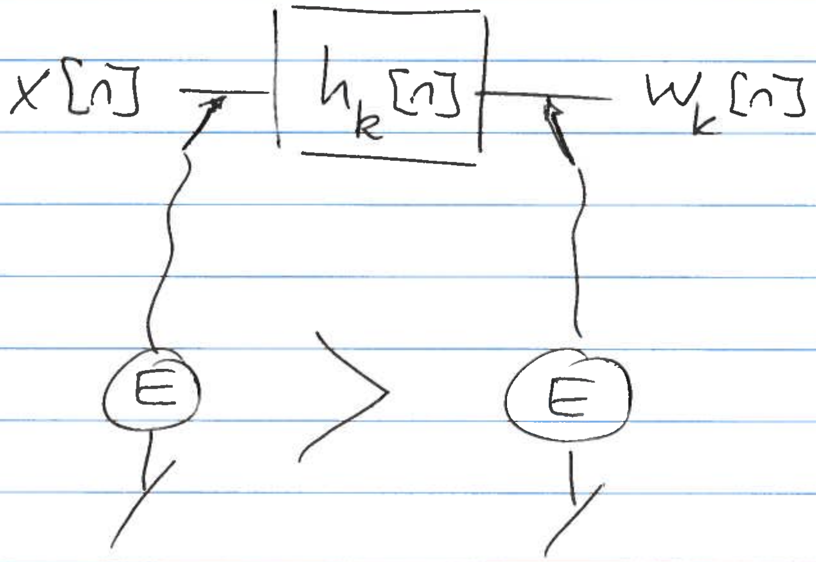
AN ISSUE WITH THE MAX-VALUE SCALING STRATEGY IS THAT s MAY BECOME SO SMALL THAT IT COMPROMISES THE SNR OF THE FILTER.

2) SINUSOID SCALING

MAINLY USE FOR NARROW BAND SIGNALS
— SELF STUDY.

3) VARIANCE SCALING

THE IDEA HERE IS TO ENSURE THAT THE ENERGY IN ALL VARIABLE WHERE OVERFLOW IS NOT ALLOWED, IS LESS THAN THE ENERGY IN THE INPUT SIGNAL.



THE ENERGY (OR VARIANCE) IS THE POWER INTEGRATED OVER TIME.

THAT CAN BE EXPRESSED BY PARSEVAL

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

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

$$\Downarrow$$

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

$$\Downarrow$$

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

EQU 12.5
p. 464.

HERE WE HAVE APPLIED; $|\underline{a} \cdot \underline{b}|^2 \leq |\underline{a}|^2 \cdot |\underline{b}|^2$

NOW, OUR REQUIREMENT WAS;

$$\sum_{n=-\infty}^{\infty} |w_k[n]|^2 \leq \sum_{n=-\infty}^{\infty} |x[n]|^2$$

$$\Downarrow$$

$$\frac{\sum_{n=\infty}^{\infty} |w_k[n]|^2}{\sum_{n=-\infty}^{\infty} |x[n]|^2} \leq 1$$

WE CAN ALWAYS FULFILL THIS REQ. IF THE INPUT SIGNAL IS SCALED PROPERLY.

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

$\Downarrow$

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

SEE p. 24

$\Downarrow$

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

$\Downarrow$

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

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

$\Downarrow$

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