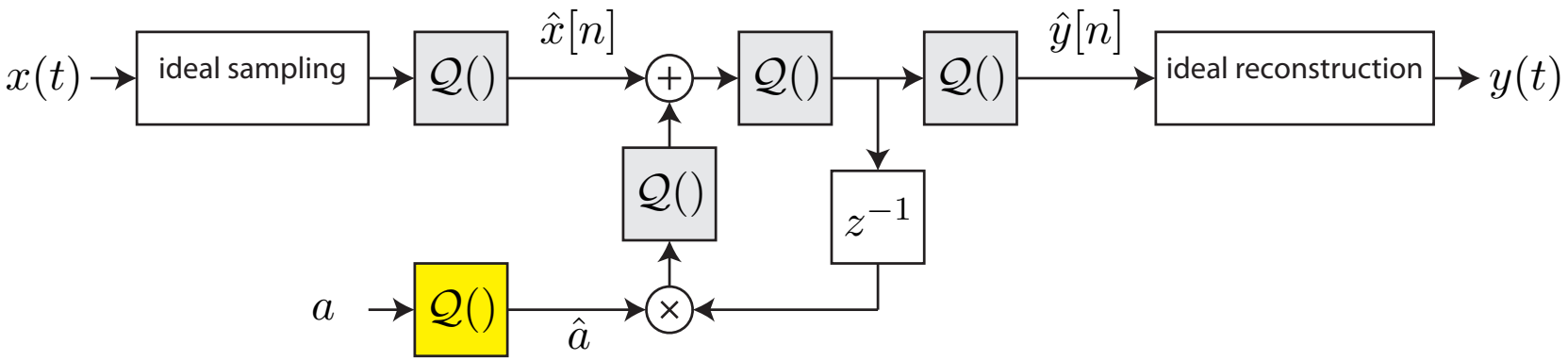
Digital Signal Processing
Effect of Coefficient Quantization on FIR Filters

D. Richard Brown III

Effect of Coefficient Quantization

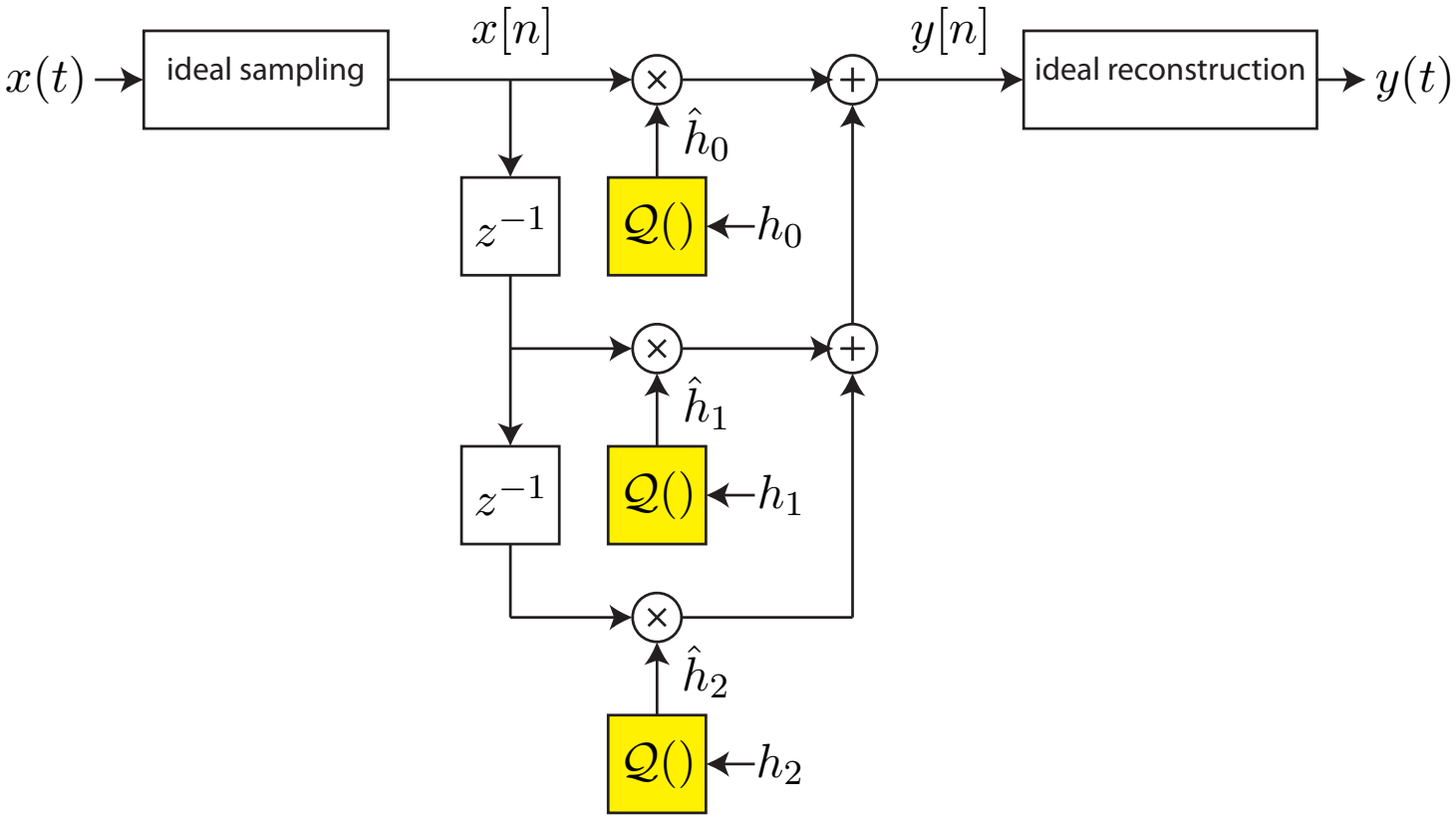
General model accounting for finite-precision effects:



Here, we focus on the effect of coefficient quantization and ignore the other sources of quantization error.

FIR Filter Coefficient Quantization

For a direct form FIR filter, we have the realization structure



with $\hat{h}_n = h_n + e_n$.

Coefficient Quantization Example (4 bits)

Suppose we quantize the coefficients $\{h_n\}$ using a 4-bit quantizer (including a sign bit) and denote the number of fractional bits as q.

	$q=0$ ($\Delta=1$)		$q=1$ ($\Delta=\frac{1}{2}$)		$q=2$ ($\Delta=\frac{1}{4}$)		$q=3$ ($\Delta=\frac{1}{8}$)	
h_n	$\hat{h}_n$	e_n	$\hat{h}_n$	e_n	$\hat{h}_n$	e_n	$\hat{h}_n$	e_n
0.13	0	-0.13	0	-0.13	0.25	-0.12	0.125	-0.005
-0.8	-1	-0.2	-1	-0.2	-0.75	0.05	-0.75	0.05
1.3	1	-0.3	1.5	-0.2	1.25	-0.05	0.875	-0.425

Remarks:

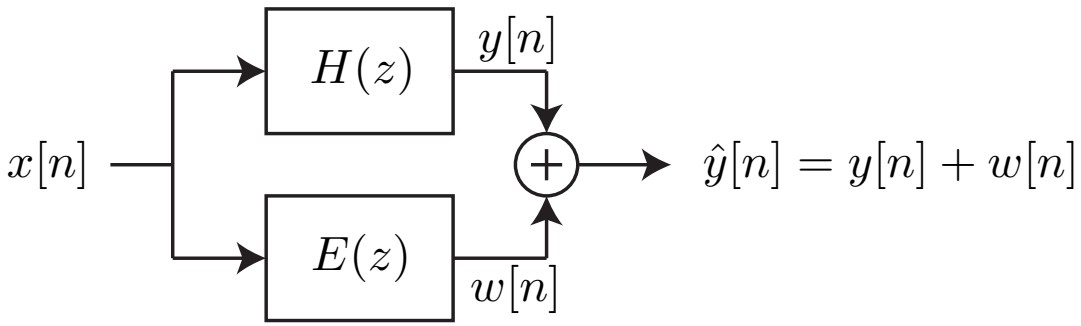
1. In most implementations, we use the same number of fractional bits for all of the coefficients.
2. The idea is to minimize the quantization errors by maximizing the number of fractional bits while avoiding overflow.
3. Problems can occur when you have both very large coefficients and very small coefficients (large coefficient dynamic range).

FIR Filter Coefficient Quantization: Equivalent System

For a causal FIR filter with N quantized coefficients, we have

$$\begin{aligned}\hat{H}(z) &= \sum_{n=0}^{N-1} \hat{h}_n z^{-n} = \sum_{n=0}^{N-1} (h_n + e_n) z^{-n} \\ &= \sum_{n=0}^{N-1} h_n z^{-n} + \sum_{n=0}^{N-1} e_n z^{-n} = H(z) + E(z)\end{aligned}$$

Hence, the quantized FIR filter $\hat{H}(z)$ is equivalent to a parallel connection of $H(z)$ and $E(z)$:



Note $E(z)$ is FIR and causal.

FIR Filter Coefficient Quantization: Error Bounds

By definition, for causal FIR $E(z)$ with N coefficients, we have a ROC $|z| > 0$ and

$$E(e^{j\omega}) = \sum_{n=0}^{N-1} e_n e^{-j\omega n}$$

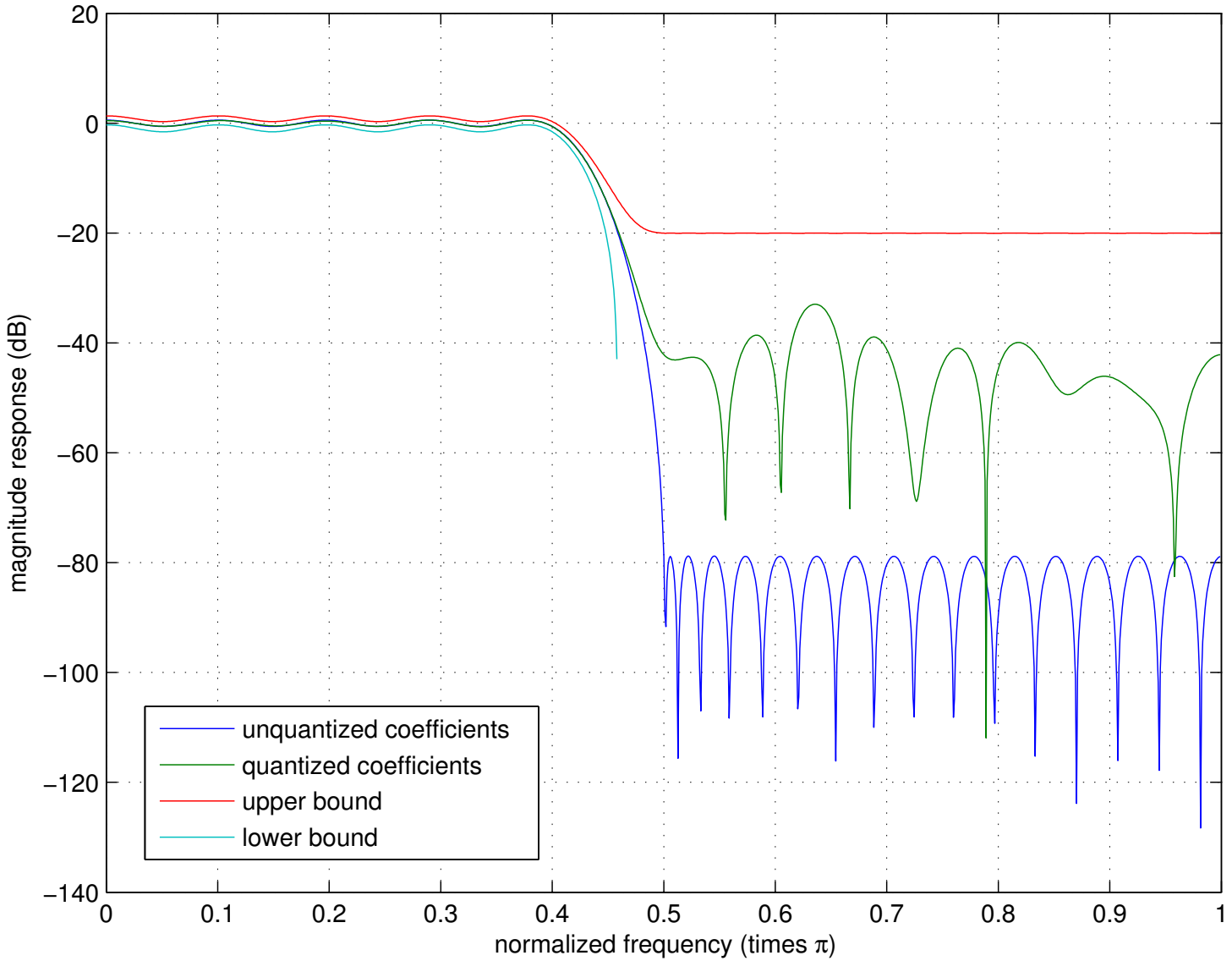
With rounding quantization, each e_n is bounded between $-\Delta/2$ and $\Delta/2$. Hence,

$$|E(e^{j\omega})| = \left| \sum_{n=0}^{N-1} e_n e^{-j\omega n} \right| \leq \sum_{n=0}^{N-1} \left| e_n e^{-j\omega n} \right| \leq \sum_{n=0}^{N-1} \frac{\Delta}{2} = \frac{N\Delta}{2}$$

and we can write

$$\left(|H(e^{j\omega})| - \frac{N\Delta}{2} \right)^+ \leq |H(e^{j\omega})| \leq |H(e^{j\omega})| + \frac{N\Delta}{2}.$$

FIR Coefficient Quantization Example ($\Delta = 2^{-8}$, $N = 51$)



Final Remarks on Fixed-Point FIR Filtering

- Direct form realizations of fixed-point FIR filters are commonly used since they tend to provide acceptable performance in most cases.
- Symmetry properties of direct form FIR filters are not affected by quantization. Hence, direct form linear phase filters still have linear phase after coefficient quantization.
- Cascade/parallel forms can provide better performance by reducing the dynamic range of the coefficients, but are not often used for FIR filters unless the zeros of $H(z)$ are tightly clustered.
- Some additional care is usually needed to preserve linear phase with quantized coefficients in cascade form.