Digital Signal Processing
Effect of Coefficient Quantization on Cascaded IIR Filters

D. Richard Brown III

Quantized IIR Filter Coefficients

Consider $H(z) = \frac{b_0+b_1z^{-1}+\cdots+b_8z^{-8}}{1+a_1z^{-1}+\cdots+a_8z^{-8}}$ with unquantized and 8-bit quantized coefficients:

i	a_i	$\hat{a}_i$	b_i	$\hat{b}_i$
0	1.0000	1.0000	0.0039	0.0039
1	-3.7597	-3.7500	0.0093	0.0093
2	8.1976	8.2500	0.0198	0.0200
3	-11.8524	-11.8750	0.0276	0.0273
4	12.3314	12.3750	0.0317	0.0317
5	-9.2974	-9.2500	0.0276	0.0273
6	4.9767	5.0000	0.0198	0.0200
7	-1.7419	-1.7500	0.0093	0.0093
8	0.3172	0.3750	0.0039	0.0039

The numerator coefficients have $q_b = 11$ fractional bits and the denominator coefficients have $q_a = 3$ fractional bits. Coefficient quantization has made this IIR filter unstable.

We could fix this problem by using more bits to quantize the coefficients. Or, we can **change the realization structure** and get very good results with even less bits.

Cascaded Second Order Sections Filter Realization

We can convert the original filter to cascaded second-order direct form II sections and quantize the coefficients in each section separately. Here is one possible realization:

section 1

i	a_i	$\hat{a}_i$	b_i	$\hat{b}_i$
0	1.0000	1.0000	0.2498	0.2500
1	-0.7400	-0.7500	0.0213	0.0156
2	0.8610	0.8750	0.2498	0.2500

section 2

i	a_i	$\hat{a}_i$	b_i	$\hat{b}_i$
0	1.0000	1.0000	0.2498	0.2500
1	-1.0375	-1.0625	0.1724	0.1719
2	0.7062	0.6875	0.2498	0.2500

section 3

i	a_i	$\hat{a}_i$	b_i	$\hat{b}_i$
0	1.0000	1.0000	0.2498	0.2500
1	-1.3740	-1.3750	0.4356	0.4375
2	0.5431	0.5625	0.2498	0.2500

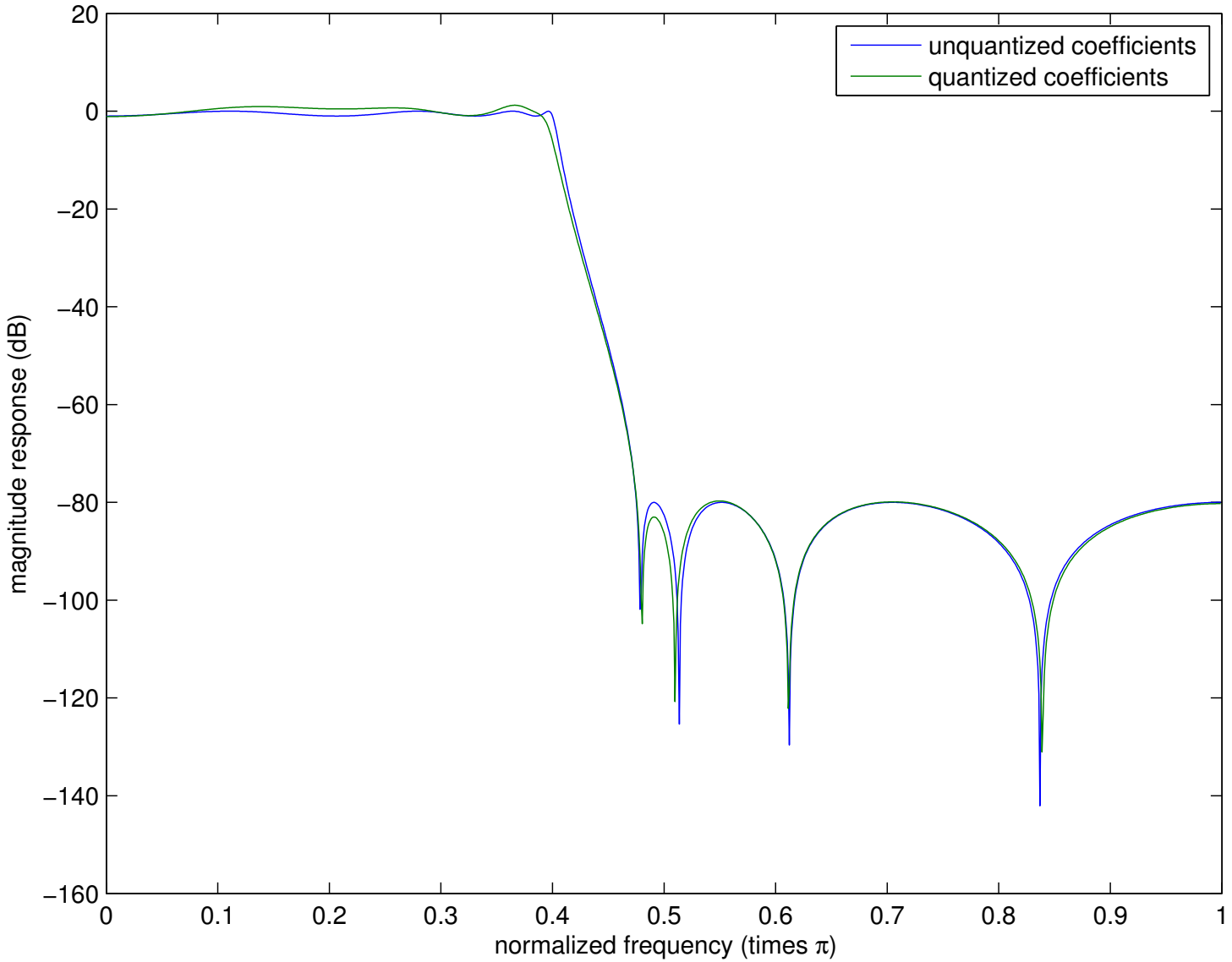
section 4

i	a_i	$\hat{a}_i$	b_i	$\hat{b}_i$
0	1.0000	1.0000	0.2498	0.2500
1	-0.6083	-0.6250	-0.0331	-0.0312
2	0.9605	0.9375	0.2498	0.2500

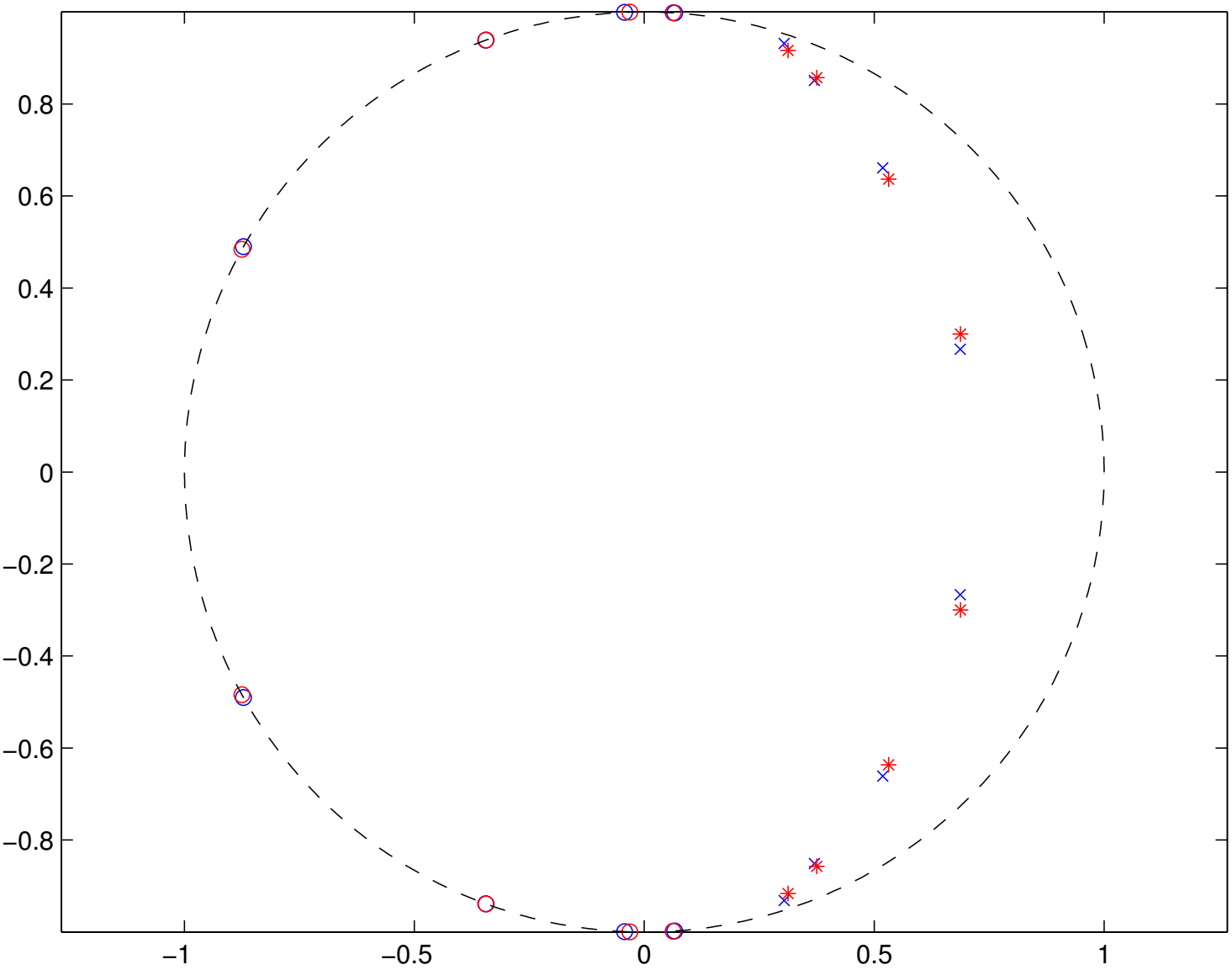
We have assumed all of the sections to have 6-bit coefficients (including the sign bit). We have also assumed all numerator coefficients to have the same number of fractional bits ($q_b = 6$ here) and all denominator coefficients to have the same number of fractional bits ($q_a = 4$ here).

Dynamic range of denominator coefficients: $\frac{1.3740}{0.5431} \approx 2.53$.

IIR Coeff. Quantization Example (6-bit Cascaded DF-II)



IIR Coeff. Quantization Example (6-bit Cascaded DF-II)



Remarks

- By using cascade form, we can reduce the dynamic range of the coefficients and improve the accuracy of the coefficient quantization.
- Each section is quantized separately.
- Cascaded form is usually preferred to parallel form because the quantization in each section only affects the zeros of that section.
- Direct IIR forms are rarely used for implementing anything other than second order systems.
- Lattice structures are also usually insensitive to coefficient quantization, but require more computation than cascaded IIR DF-II.