Second and Higher-Order Delta-Sigma Modulators

MEAD March 2008

Richard Schreier

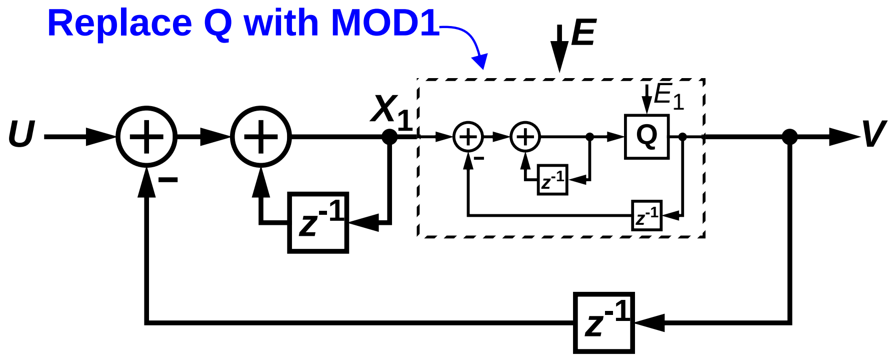
Richard.Schreier@analog.com

ANALOG DEVICES

Overview

1. MOD2 from MOD1

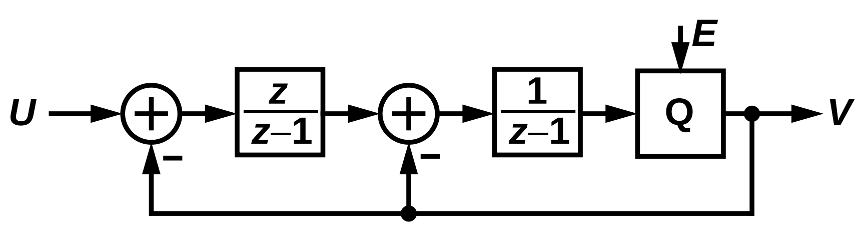
- **Replace the quantizer in MOD1 with another copy of MOD1 in a recursive fashion:**



$$V = U + (1 - z^{-1})E$$
$$V = X_1 + E = X_1 + (1 - z^{-1})E_1$$
$$\Rightarrow E = (1 - z^{-1})E_1$$
$$\Rightarrow \mathbf{V = U + (1 - z^{-1})^2 E_1}$$

Simplified Diagram

- **Combine feedback paths, absorb feedback delay into second integrator…**

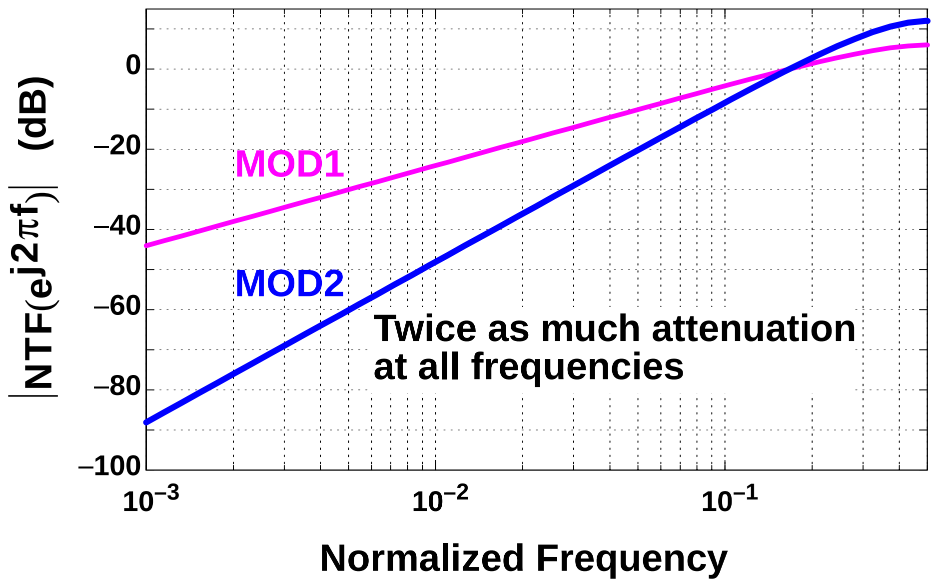


$$V(z) = z^{-1}U(z) + (1 - z^{-1})^2 E(z)$$

- $NTF(z) = (1 - z^{-1})^2$ **and the STF is** $STF(z) = z^{-1}$

- **MOD2's NTF is the *square* of MOD1's NTF**

NTF Comparison



Predicted Performance

- **In-band quantization noise power**

$$IQNP = \int_0^{1/(2\cdot OSR)} \left|NTF(e^{j2\pi f})\right|^2 \cdot S_{ee}(f)\, df$$

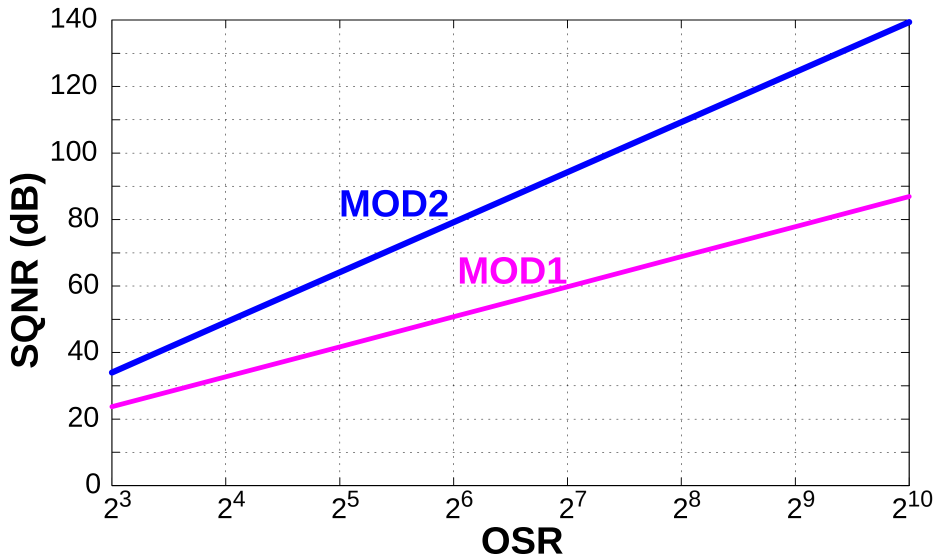
$$\approx \int_0^{1/(2\cdot OSR)} (2\pi f)^4 \cdot 2\sigma_e^2 df$$

$$= \frac{\pi^4\sigma_e^2}{5(OSR)^5} \qquad \ast$$

- **Quantization noise drops as the 5th power of OSR!**

  SQNR increases at 15 dB per octave increase in OSR.

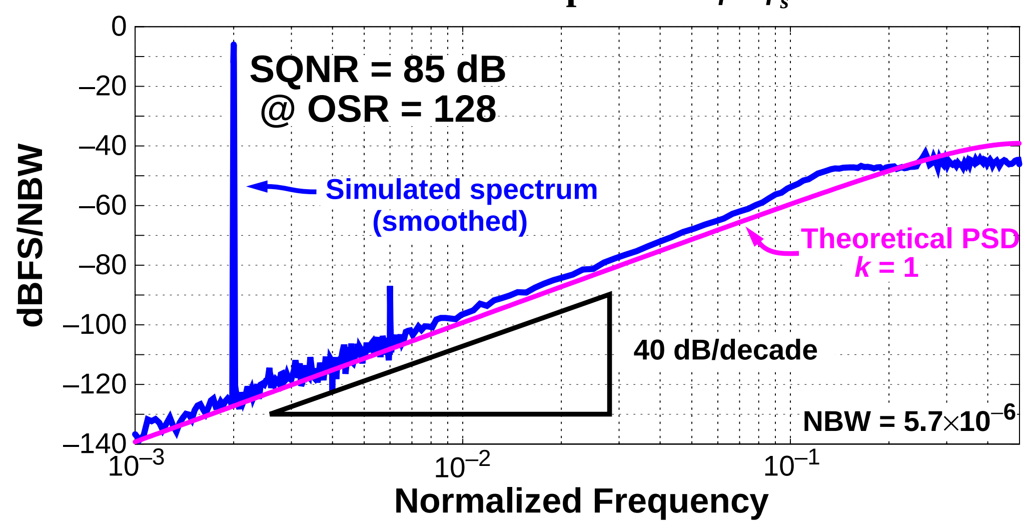
Predicted SQNR vs. OSR



- **For $OSR = 128$ and binary quantization, the predicted SQNR of MOD2 is 94.2 dB**

MOD2 Simulated PSD

Half-scale sine-wave input with $f \approx f_s/500$



- **Observed PSD similar to theory (40 dB/decade slope)**

But 3rd harmonic is visible and in-band PSD is slightly higher.

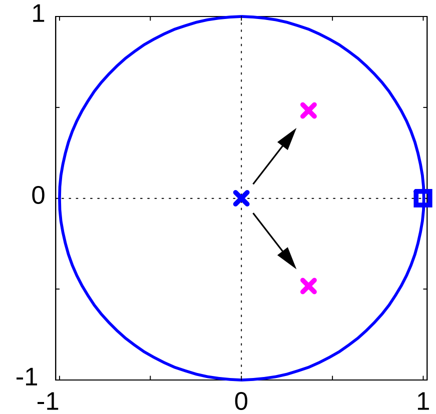
Gain of the Quantizer in MOD2

- **The effective quantizer gain can be computed from the simulation data using**

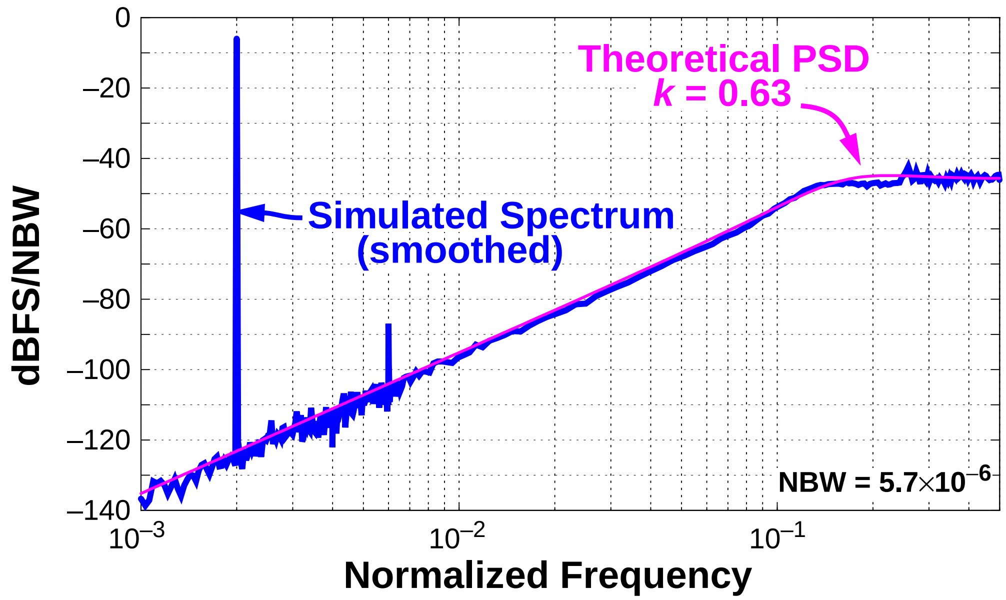
$$k = \frac{\langle v, y\rangle}{\langle y, y\rangle} = \frac{E[|y|]}{E[y^2]} \qquad \text{[S\&T Eq. 2.5]}$$

- **For the preceding simulation, $k = 0.63$.**
- **$k \neq 1$ alters the NTF:**

$$NTF_k(z) = \frac{NTF_1(z)}{k + (1-k)NTF_1(z)}$$



Revised PSD Prediction



- **Agreement is now excellent**

Variable Quantizer Gain

- **When the input is small (below -12 dBFS), the effective gain of the quantizer is $k = 0.75$**
- **The “small-signal NTF” is thus**

$$NTF(z) = \frac{(z-1)^2}{z^2 - 0.5z + 0.25}$$

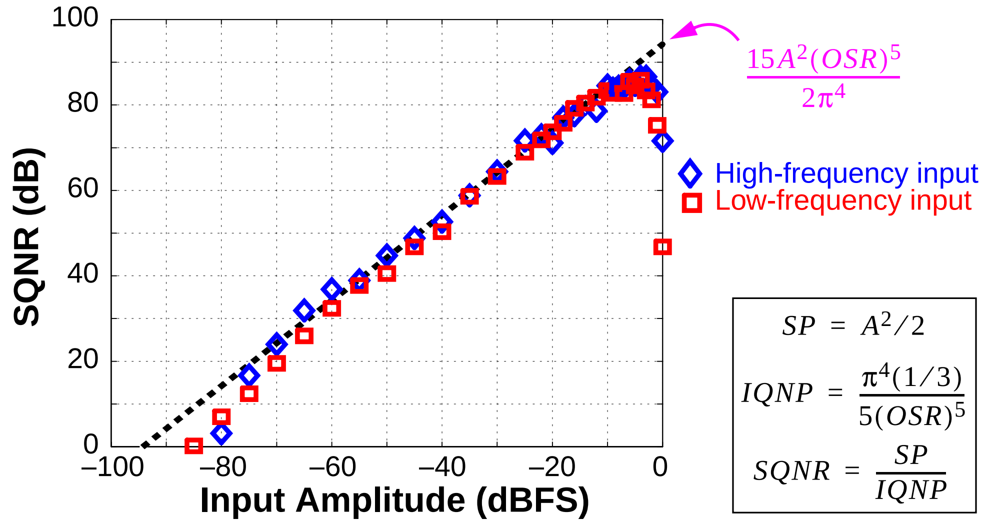
- **This NTF has 2.5 dB les quantization noise suppression than the $(1 - z^{-1})^2$ NTF derived from the assumption that $k = 1$**

  Thus the SQNR should be about 2.5 dB lower than ⚹.

- **As the input signal increases, *k* decreases and the suppression of quantization noise degrades**

  SQNR increases less quickly than the signal power, and eventually the SQNR saturates and then decreases as the signal power is increased.

Simulated SQNR of MOD2



$$SP = A^2/2$$

$$IQNP = \frac{\pi^4(1/3)}{5(OSR)^5}$$

$$SQNR = \frac{SP}{IQNP}$$

- **Well-modeled by ideal formula; less erratic than MOD1**

  Saturation at high signal levels due to decreased quantizer gain and altered NTF. (Worse with low-frequency inputs.)

Stability of MOD2

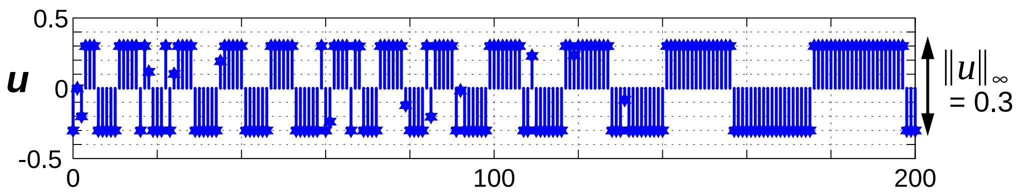
- **Known to be stable with DC inputs up to full-scale, but the state bounds blow up as $|u| \to 1$**

Hein [ISCAS 1991]: $|u| \le 1 \Rightarrow$

$$|x_1| \le |u| + 2 \quad \textbf{(output of 1}^{\textbf{st}}\textbf{ integrator)}$$

$$|x_2| \le \frac{(5-|u|)^2}{8(1-|u|)} \quad \textbf{(output of 2}^{\textbf{nd}}\textbf{ integrator)}$$

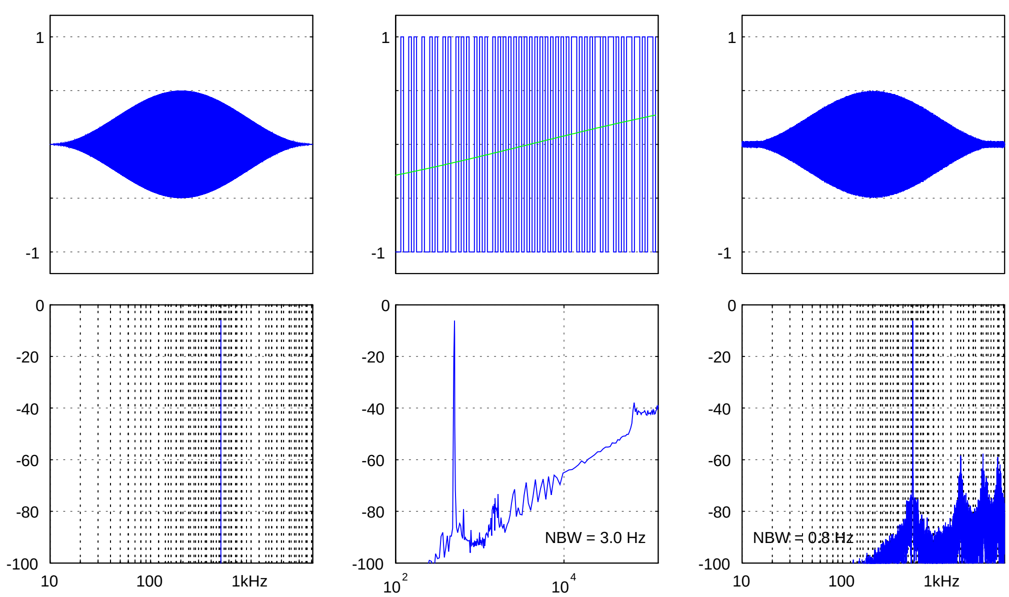
- **However, with a hostile input (whose magnitude is less than 30% of full-scale) MOD2 can be driven unstable!**



- **As a result of this input-dependent stability, it is wise to keep the input below 70-90% of full-scale**

Deadbands, Distortion & Tones

Audio Comparison of MOD1 and MOD2



Observations

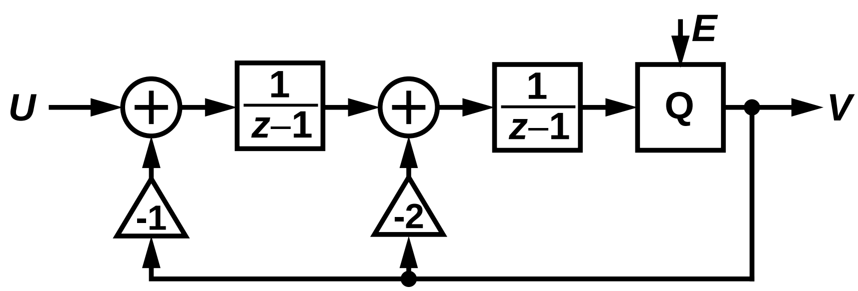
Tones

- **Quantization noise of MOD1 is distinctly non-white**
  Audible tones when input is near zero, or near other simple rational fractions of full-scale.
- **MOD2 is better than MOD1 in terms of its tendency toward tonal behavior**

Dead-bands

- **MOD1 has dead-bands whose widths are proportional to $1/A$, where A is the gain of the internal op-amp**
- **MOD2 has dead-bands whose widths are proportional to $1/A^2$**
- **Dead-band behavior is less problematic in MOD2**

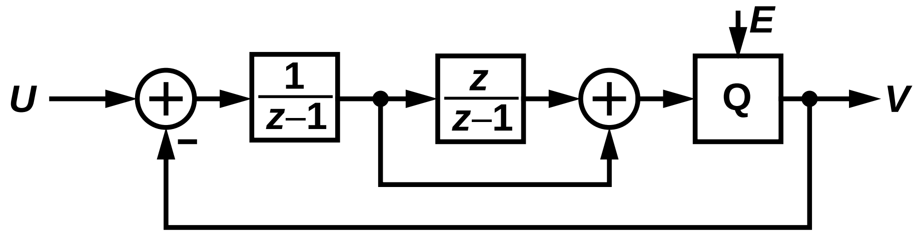
Topological Variant– Delaying Integrators



$$NTF(z) = (1 - z^{-1})^2 \qquad STF(z) = z^{-2}$$

+ **Delaying integrators reduce the settling requirements**
  Can decouple the integration phase from the driving phase

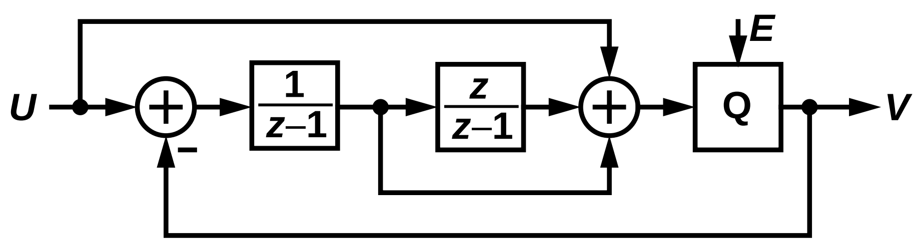
Topological Variant–Feed-Forward



$$NTF(z) = (1 - z^{-1})^2 \qquad STF(z) = 2z^{-1} - z^{-2}$$

- **+ Output of first integrator has no DC component**
  Dynamic range requirements of this integrator are relaxed.
- **– Although $|STF| \approx 1$ near $\omega = 0$, $|STF| = 3$ for $\omega = \pi$**
  Instability is more likely.

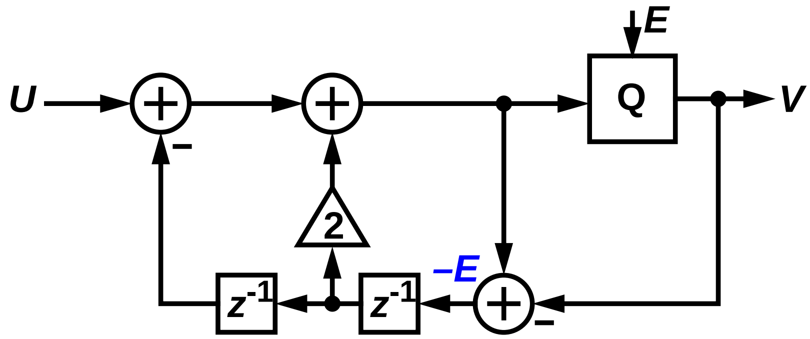
Topological Variant–Feed-Forward with Extra Input Feed-In



$$NTF(z) = (1 - z^{-1})^2 \qquad STF(z) = 1$$

- **+ No DC component in either integrator's output**
  Reduced dynamic range requirements in both integrators.
- **+ Perfectly flat STF**
  No increased risk of instability.

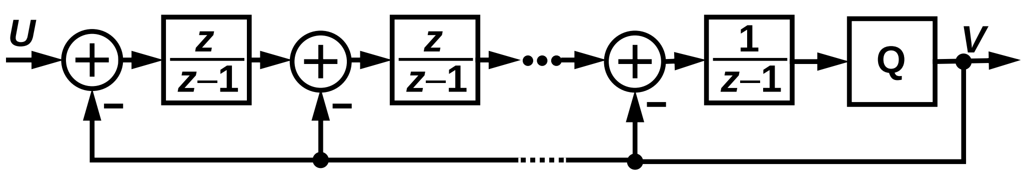
Topological Variant– Error Feedback



$$NTF(z) = (1 - z^{-1})^2 \qquad STF(z) = 1$$

- **+ Simple**
- **– Very sensitive to gain errors**
  Only suitable for digital implementations.

2. MOD*N* from MOD2



$$V = E + \frac{1}{z-1}\left(-V + \frac{z}{z-1}\left(-V + \dots \frac{z}{z-1}(-V+U)\right)\right)$$

$$(1-z^{-1})^N V = (1-z^{-1})^N E - ((1-z^{-1})^{N-1} + (1-z^{-1})^{N-2} + \dots + 1)z^{-1}V + z^{-1}U$$

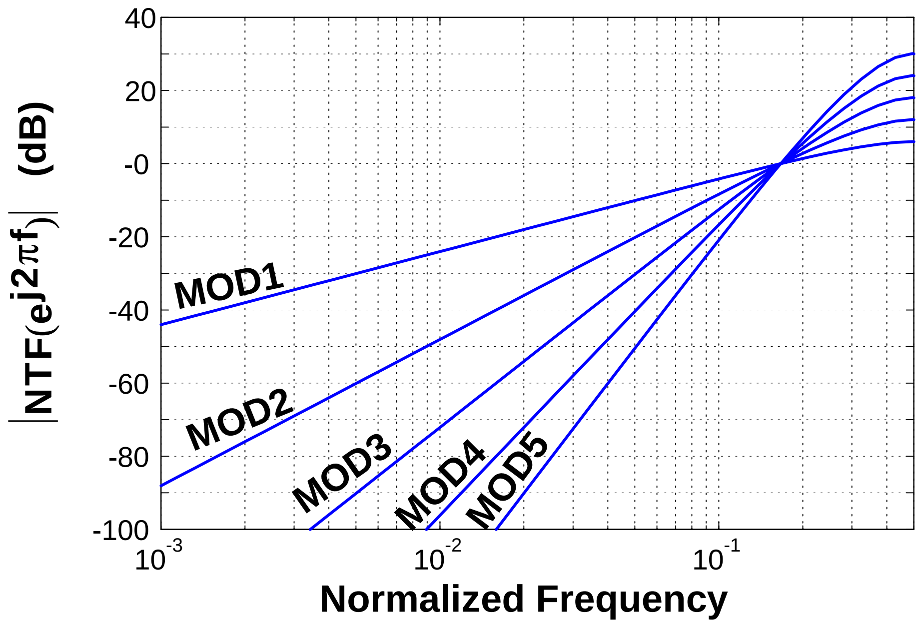
$$(1-z^{-1})^N V = (1-z^{-1})^N E - \left(\frac{(1-z^{-1})^N - 1}{1-z^{-1}-1}\right)z^{-1}V + z^{-1}U$$

$$(1-z^{-1})^N V = (1-z^{-1})^N E - \left(\frac{(1-z^{-1})^N - 1}{-\cancel{z^{-1}}}\right)\cancel{z^{-1}}V + z^{-1}U$$

$$\therefore \boldsymbol{V(z) = z^{-1}U(z) + (1 - z^{-1})^N E(z)}$$

- **NTF of MOD*N* is the *N*th power of MOD1’s NTF**

NTF Comparison



Predicted Performance

- **In-band quantization noise power**

$$IQNP = \int_0^{1/(2 \cdot OSR)} |NTF(e^{j2\pi f})|^2 \cdot S_{ee}(f)\, df$$

$$\approx \int_0^{1/(2 \cdot OSR)} (2\pi f)^{2N} \cdot 2\sigma_e^2 df$$

$$= \frac{\pi^{2N}}{(2N+1)(OSR)^{2N+1}} \sigma_e^2$$

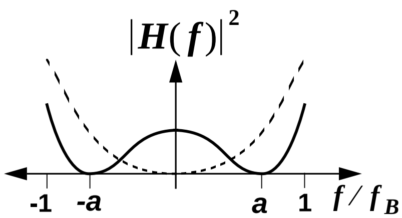
- **Quantization noise drops as the (2N+1)th power of OSR!**

SQNR increases at (6N+3) dB per octave increase in OSR.

Improving NTF Performance– Zero Optimization

- **Minimize the rms in-band value of H by finding the a_i which minimize the integral of $|H|^2$ over the passband.**

  Normalize passband edge to 1 for ease of calculation.



i.e. Find the a_i which minimize the integral

$$\int_{-1}^{1}(x^2-a_1^2)^2dx\,, \qquad n = 2$$

$$\int_{-1}^{1}x^2(x^2-a_1^2)^2dx\,, \qquad n = 3$$

$$\int_{-1}^{1}(x^2-a_1^2)^2(x^2-a_2^2)^2dx\,, \qquad n = 4$$

$$\vdots$$

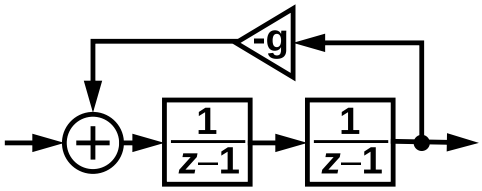
Solutions Up to Order = 8

Order	Optimal Zero Placement Relative to f_B	SQNR Improvement
1	0	0 dB
2	$\pm\frac{1}{\sqrt{3}}$	3.5 dB
3	$0, \pm\sqrt{\frac{3}{5}}$	8 dB
4	$\pm\sqrt{\frac{3}{7}\pm\sqrt{\left(\frac{3}{7}\right)^2-\frac{3}{35}}}$	13 dB
5	$0, \pm\sqrt{\frac{5}{9}\pm\sqrt{\left(\frac{5}{9}\right)^2-\frac{5}{21}}}$ [Y. Yang]	18 dB
6	±0.23862, ±0.66121, ±0.93247	23 dB
7	0, ±0.40585, ±0.74153, ±0.94911	28 dB
8	±0.18343, ±0.52553, ±0.79667, ±0.96029	34 dB

Topological Implication

- **Apply feedback around pairs integrators:**

2 Delaying Integrators



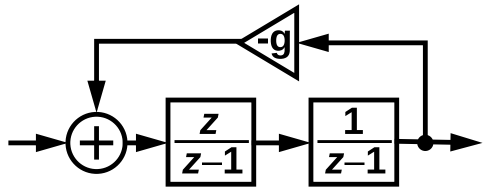
Poles are the roots of

$$1 + g\left(\frac{1}{z-1}\right)^2 = 0$$

i.e. $z = 1 \pm j\sqrt{g}$

Not quite on the unit circle, but fairly close if g<<1.

Non-delaying + Delaying Integrators (LDI Loop)



Poles are the roots of

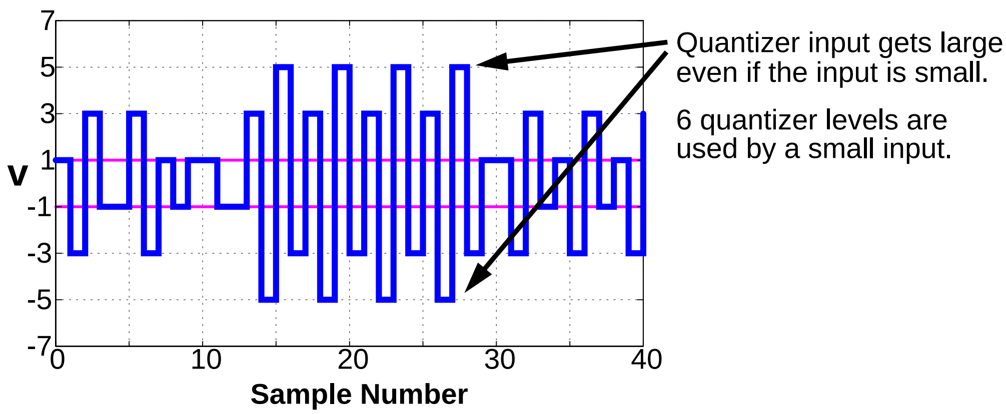
$$1 + \frac{gz}{(z-1)^2} = 0$$

i.e. $z = e^{\pm j\theta}$, $\cos\theta = 1 - g/2$

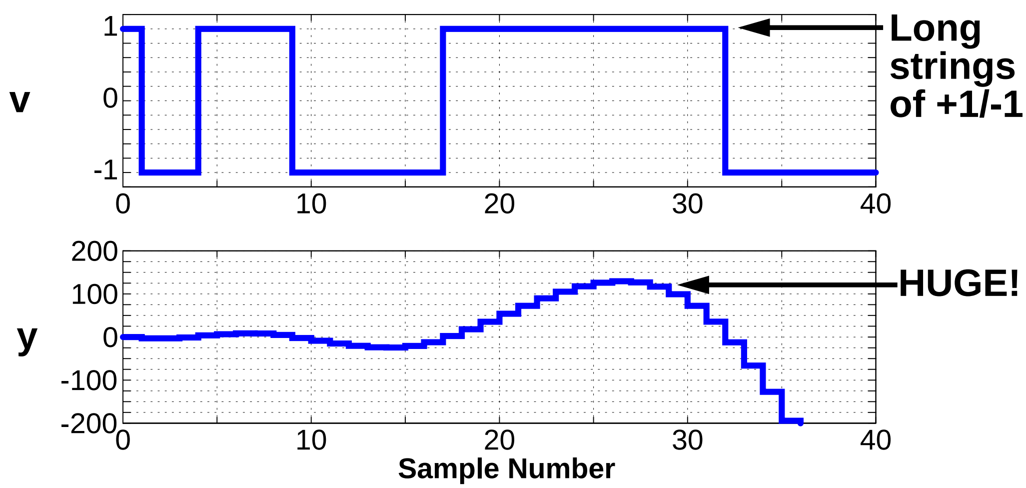
Precisely on the unit circle, regardless of the value of g.

Problem: High-Order Modulators Want Multi-bit Quantizers

e.g. a 3rd-Order Modulator with an Infinite Quantizer and Zero Input



The 3rd-Order Modulator is Unstable with a Binary Quantizer



- **The quantizer input grows without bound**

  The modulator is unstable, even with an arbitrarily small input.

Solutions to the Stability Problem

Historical Order

1. **Use multi-bit quantization**

   Originally considered undesirable because the inherent linearity of a 1-bit DAC is lost when a multi-bit quantizer is used.

   Less of an issue now that mismatch-shaping is available.

2. **Use a more general NTF (not pure differentiation)**

   Lower the NTF gain so that quantization error is amplified less.

   A common rule of thumb is to limit the maximum NTF gain to ~1.5.

   Unfortunately, limiting the NTF gain reduces the amount by which quantization noise is attenuated.

3. **Use a multi-stage (MASH) architecture**

   More on this later in the course.

- **Combinations of the above are possible**

Multi-bit Quantization

- Can show that a modulator with NTF H and unity STF is guaranteed to be stable if $|u| < u_{max}$ at all times, where $u_{max} = nlev + 1 - \|h\|_1$ and $\|h\|_1 = \sum_{i=0}^{\infty} |h(i)|$
- In MODN,
  $H(z) = (1 - z^{-1})^N$, so
  $h(n) = \{1, -a_1, a_2, -a_3, \ldots (-1)^N a_N, 0 \ldots\}$, where $a_i > 0$
  and thus $\|h\|_1 = H(-1) = 2^N$.
- Thus $nlev = 2^N$ implies $u_{max} = nlev + 1 - \|h\|_1 = 1$.
  - MODN is guaranteed to be stable with an n-bit quantizer if the input magnitude is less than $\Delta/2$.
  - This result is extremely conservative.
- Similarly, $nlev = 2^{N+1}$ guarantees the modulator is stable for inputs up to 50% of full-scale.

Proof of $\|h\|_1$ Criterion

By Induction

- Assume STF = 1 and $(\forall n)(|u(n)| \le u_{max})$.
- Assume $|e(i)| \le 1$ for $i < n$. **[Induction Hypothesis]**

Then

$$|y(n)| = \left|u(n) + \sum_{i=1}^{\infty} h(i)e(n-i)\right|$$
$$\le u_{max} + \sum_{i=1}^{\infty} |h(i)||e(n-i)|$$
$$\le u_{max} + \sum_{i=1}^{\infty} |h(i)| = u_{max} + \|h\|_1 - 1$$

- Thus
  $(u_{max} \le nlev + 1 - \|h\|_1) \Rightarrow (|y(n)| \le nlev) \Rightarrow (|e(n)| \le 1)$
- And by induction $|e(i)| \le 1$ for all $i > 0$. **QED**

The Lee Criterion for Stability in a 1-bit Modulator: $\|H\|_\infty \le 2$

[Wai Lee, 1987]

- **The measure of the "gain" of *H* is the maximum magnitude of *H* (over frequency), otherwise known as the *infinity-norm* of *H*:**
  $$\|H\|_\infty \equiv \max_{\omega \in [0, 2\pi]} (H(e^{j\omega}))$$

Q: Is the Lee criterion <u>necessary</u> for stability?

For MOD2, $H(z) = (1 - z^{-1})^2$ and so $\|H\|_\infty = H(-1) = 4$.

Since MOD2 is known to be stable, the Lee criterion is not necessary.

Q: Is the Lee criterion <u>sufficient</u> to ensure stability?

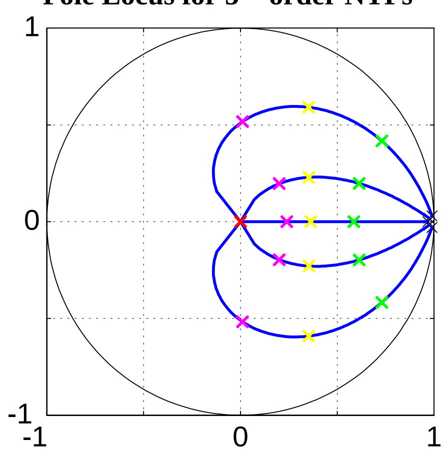
No. There are lots of counter-examples, but $\|H\|_\infty \le 1.5$ often works.

- **Let's look at some examples**

The NTF Family Used by the ΔΣ Toolbox

- **Poles chosen such that $1/\mathrm{den}(H(z))$ is a maximally flat transfer function.**

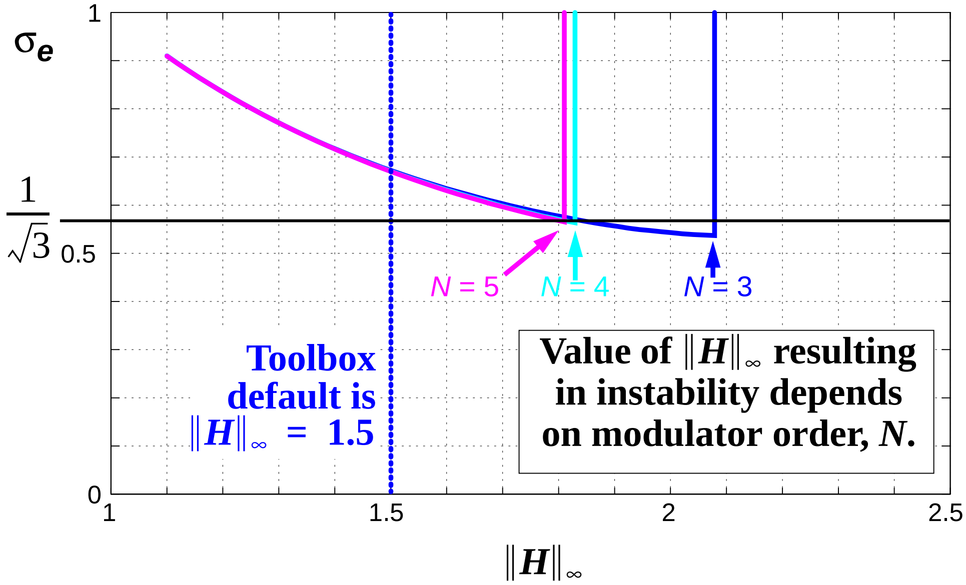
Pole Locus for 5th-order NTFs



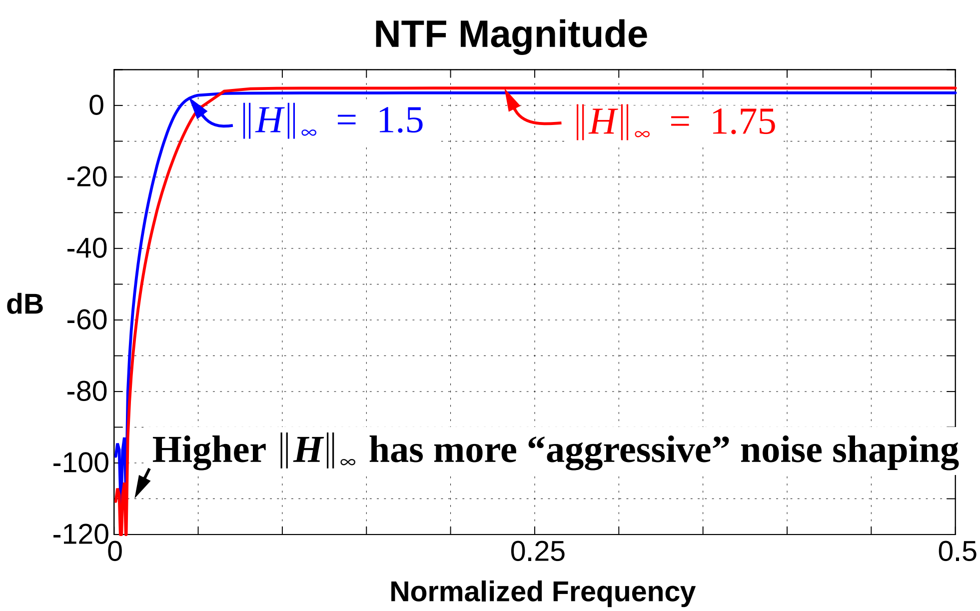
For lowpass modulators, the pole placement is similar to a Butterworth transfer function. Yields a flat STF for both lowpass and bandpass modulators employing the CRFB topology with one feed-in.

σ_e vs. $\|H\|_\infty$

Binary modulators of order N = 3, 4, 5 with a small input

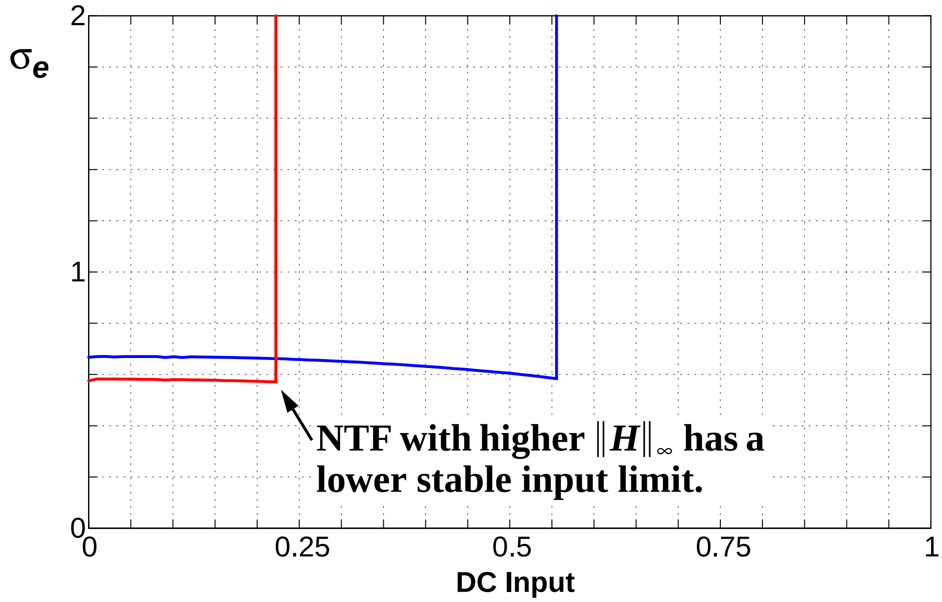


Two 5th-Order Binary Modulators



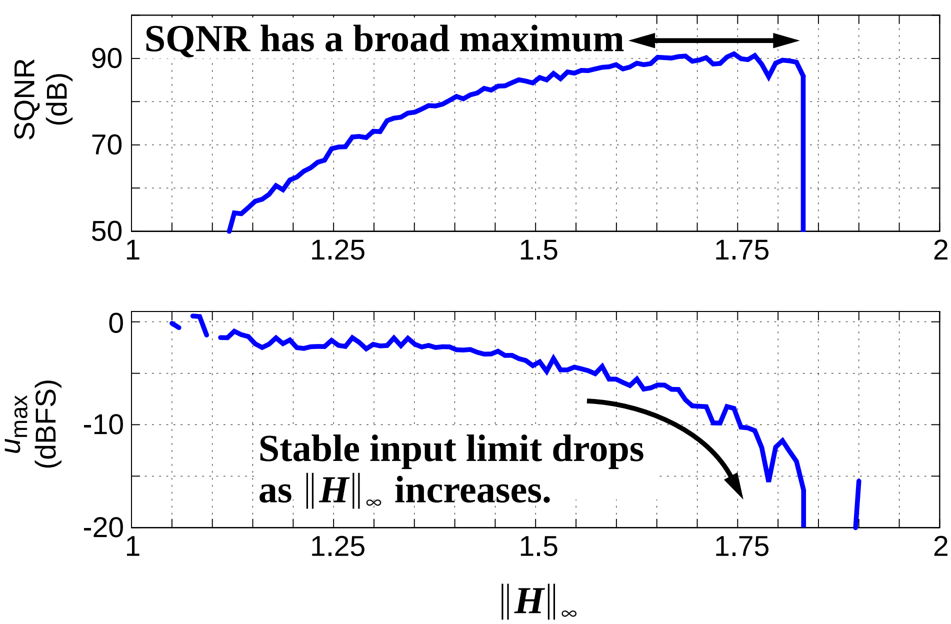
σ_e vs. u_{DC}

For the two 5th-order modulators

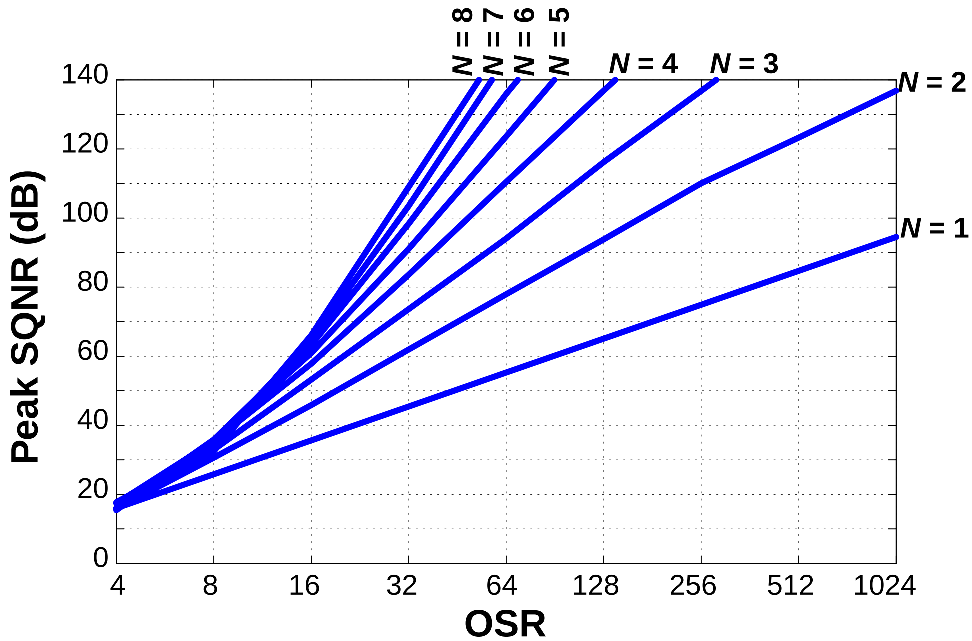


SQNR vs. $\|H\|_\infty$

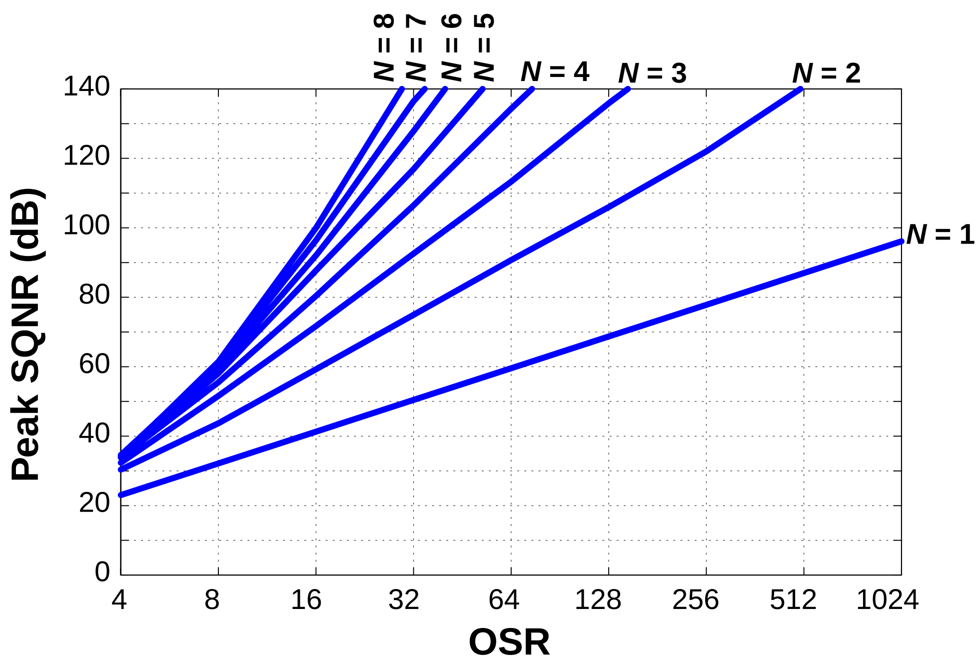
5th-Order NTFs, Binary Quantization, OSR = 32



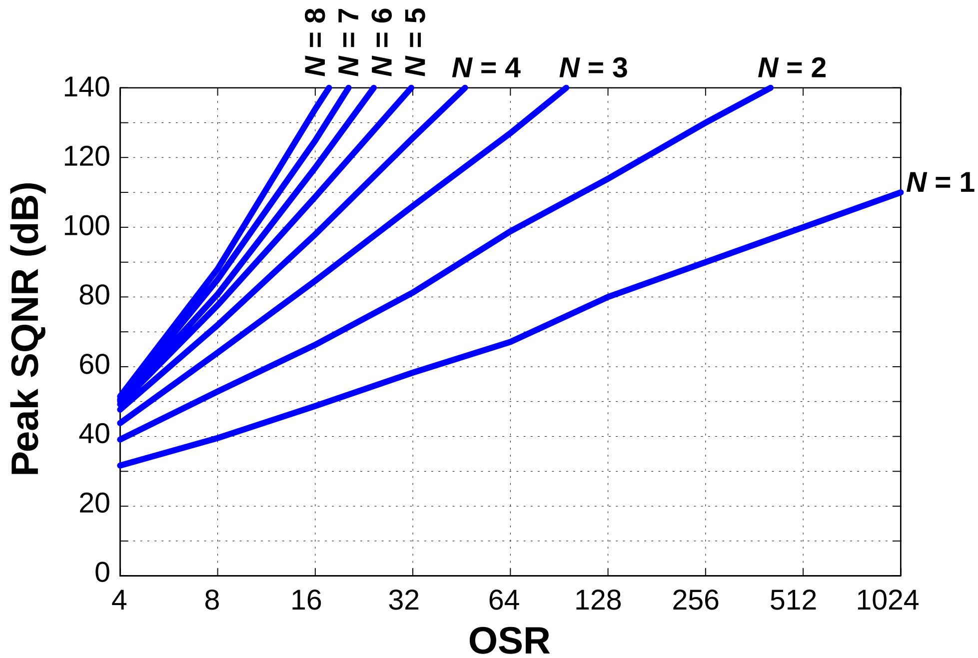
SQNR Limits for Binary Modulators



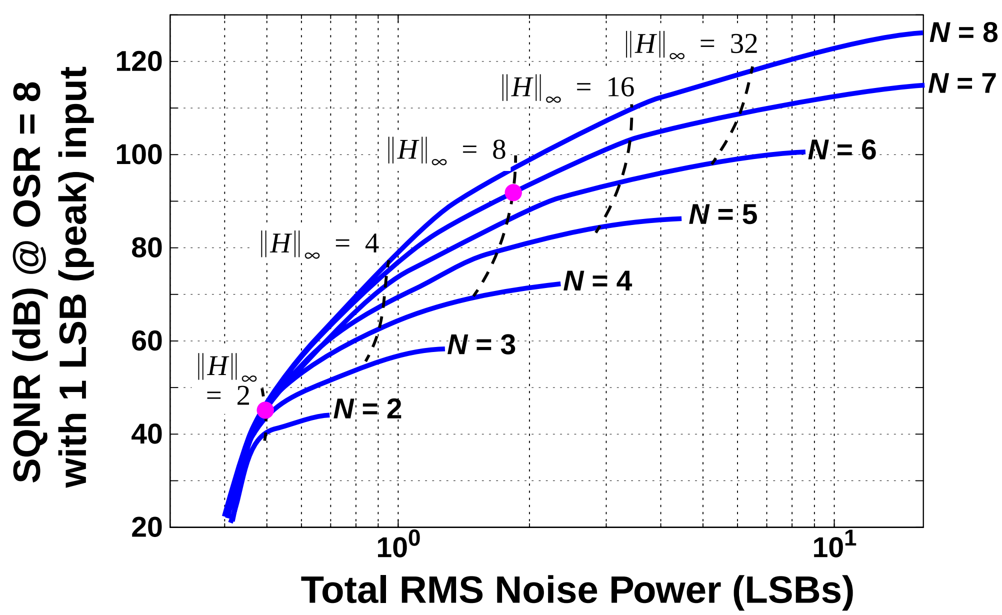
SQNR Limits for 2-bit Modulators



SQNR Limits for 3-bit Modulators

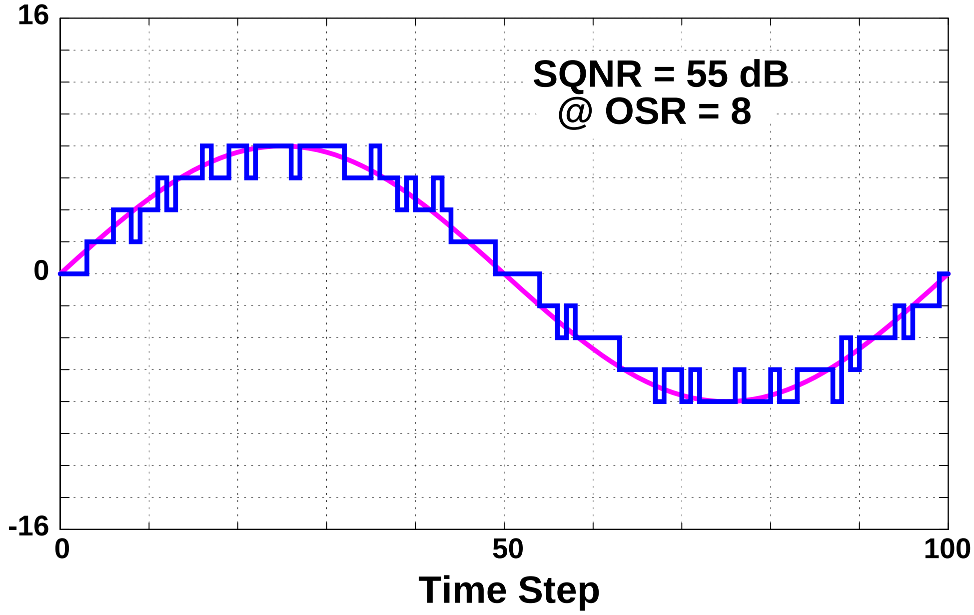


Theoretical SQNR Limits for Multi-Bit Modulators



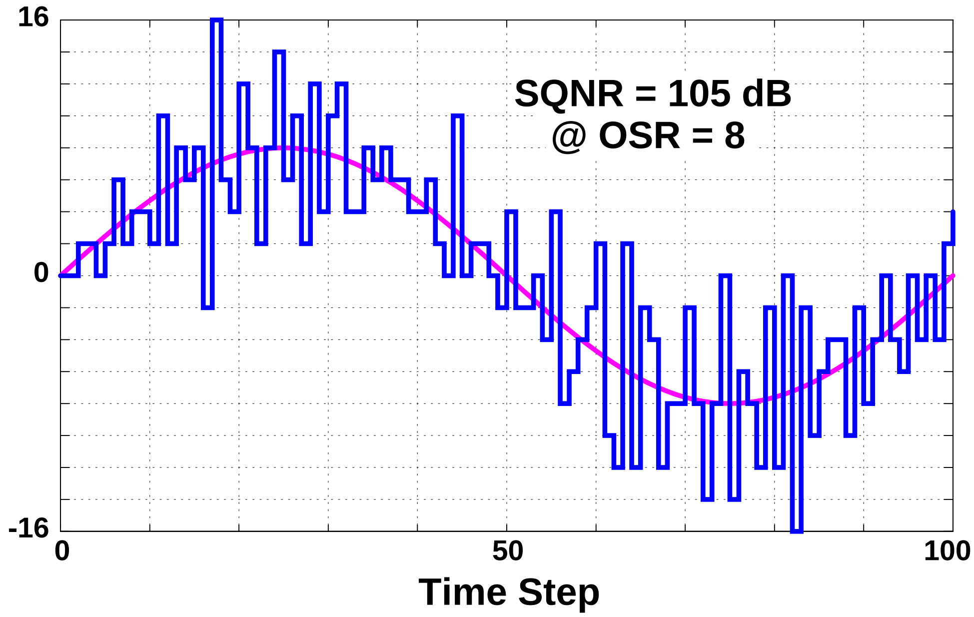
Example Waveforms

7th-order 17-level modulator, $\|H\|_\infty = 2$

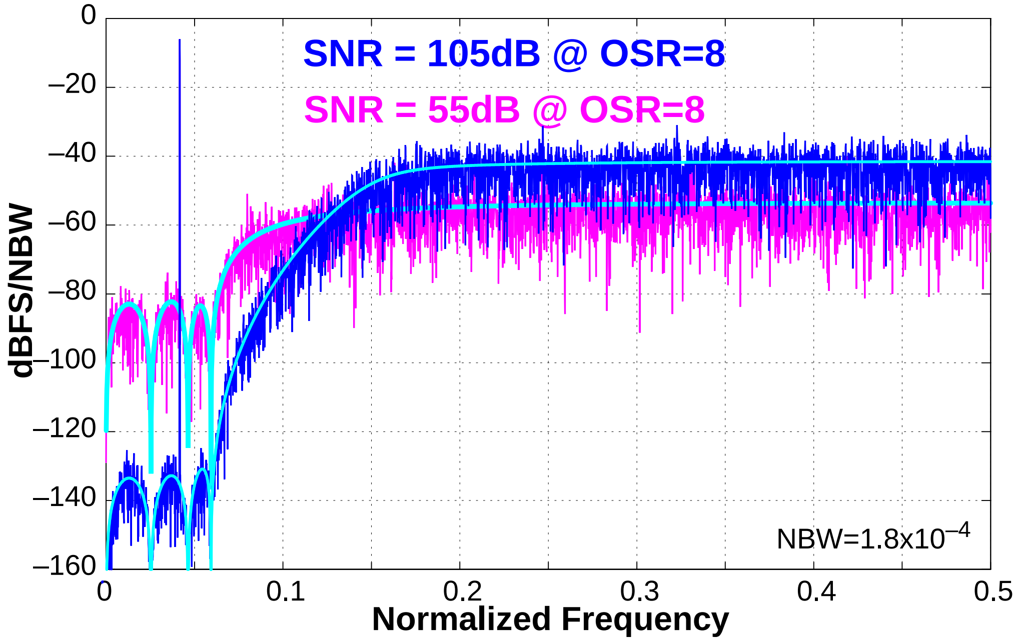


Example Waveforms

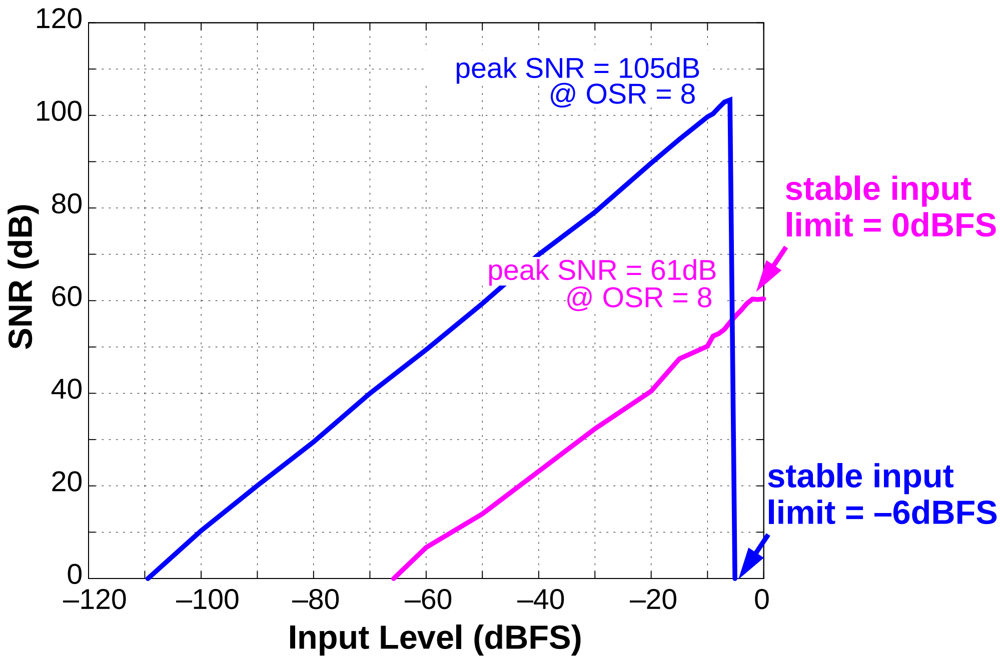
7th-order 17-level modulator, $\|H\|_\infty = 8$



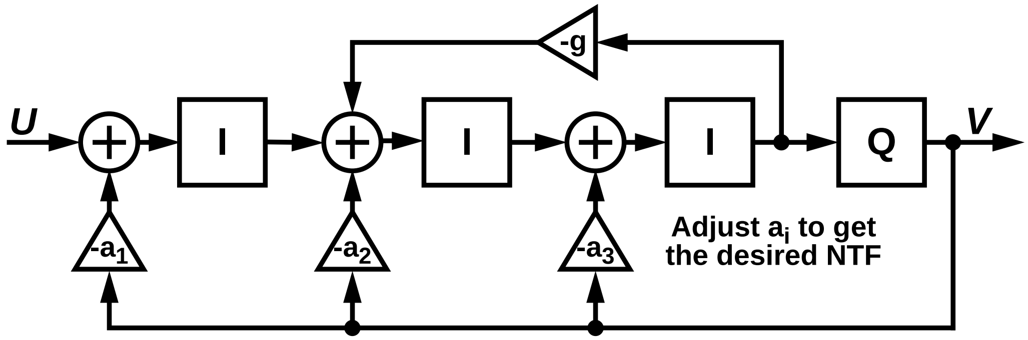
Spectra



SNR Curves



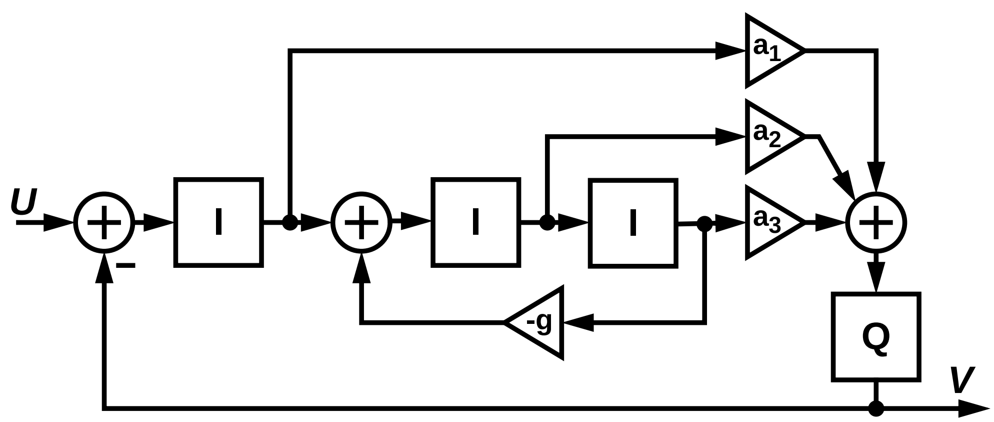
Example Topology– Feedback



- **N integrators precede the quantizer**
- **Feedback from the quantizer to the input of each integrator (via a DAC)**
- **Local feedback around pairs of integrators is possible**
- **Multiple input feed-in branches are also possible**

R. SCHREIER ANALOG DEVICES, INC.

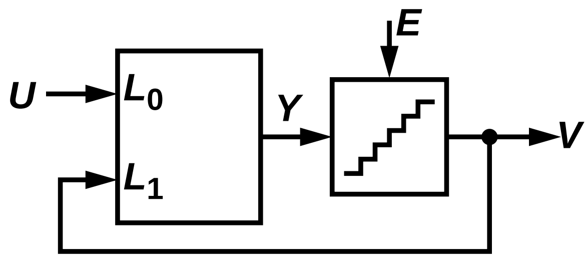
Example Topology– Feedforward



- **N integrators in a row**
- **Each integrator output is fed forward to the quantizer**
- **Local feedback around pairs of integrators is possible**
- **Multiple input feed-in branches are also possible**

General Single-Quantizer ΔΣ Modulator

- **The input to the quantizer is some linear combination of the input to the modulator and the fed-back output**



$$Y = L_0 U + L_1 V$$
$$V = Y + E$$
$$\Rightarrow$$
$$V = GU + HE, \text{ where}$$
$$H = \frac{1}{1 - L_1} \ \ \& \ \ G = L_0 H$$

Inverse Relations:

$$L_1 = 1 - 1/H, \ L_0 = G/H$$

Summary

- **MOD2 is better than MOD1**
  - **Higher SQNR**
  - **Whiter quantization noise**
  - **Smaller deadbands**
- **MOD*N* is better than MOD2**
  - **Even higher SQNR**
  - **Tonal behavior unlikely**
  - **Deadbands virtually eliminated**
- **BUT high-order modulators must deal with instability**
  - **Modify the NTF, reduce the input range, and/or use multi-bit quantization**