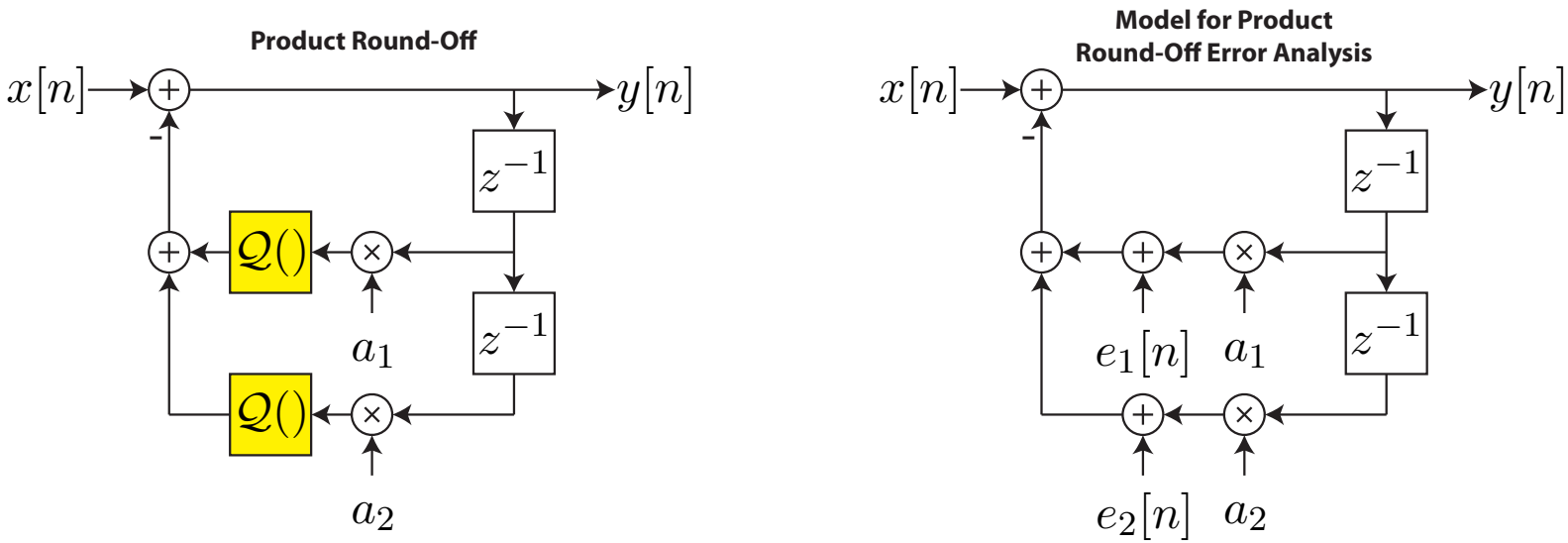
Digital Signal Processing
Effect of Product Roundoff in Finite-Precision Filters

D. Richard Brown III

Product Round-Off: Linear Model

When you take the product of two $B+1$-bit fixed-point numbers, the result requires $2B+2$ bits to store. If we don't have enough bits, we typically round off the least significant bits of the product, which leads to another source of quantization error in finite precision filters.

We will use a linear model for round-off error analysis. For example, consider a second-order all-pole filter:



The main idea is that we insert a quantization noise source after each non-unity product. If an extended precision accumulator is used, then a quantization noise source is placed only after the final sum.

Product Round-Off: Linear Model Assumptions

To facilitate analysis, the product round-off quantization errors are modeled as random sequences just like input quantization errors:

1. Each product round-off error $e_\ell[n]$ is uniformly distributed on $\left[-\frac{\delta}{2}, \frac{\delta}{2}\right]$.
2. The product round-off error $e_\ell[n]$ is independent of $e_\ell[m]$ for all $n \neq m$.
3. The product round-off error $e_\ell[n]$ is independent of $x[m]$ for all n and m.
4. The product round-off error $e_\ell[n]$ is independent of $e_k[m]$ for all $\ell \neq k$.

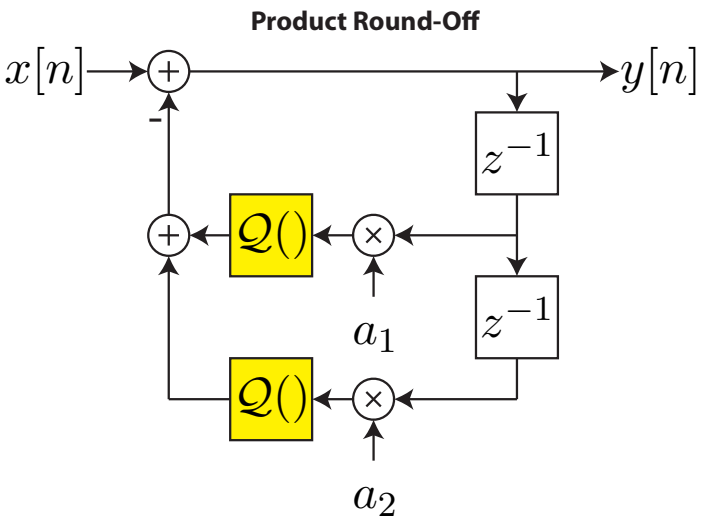
Procedure: Denote the round-off noise variance σ_e^2 (assumed the same for all products). For each product round-off error source $\ell = 1, \ldots, L$:

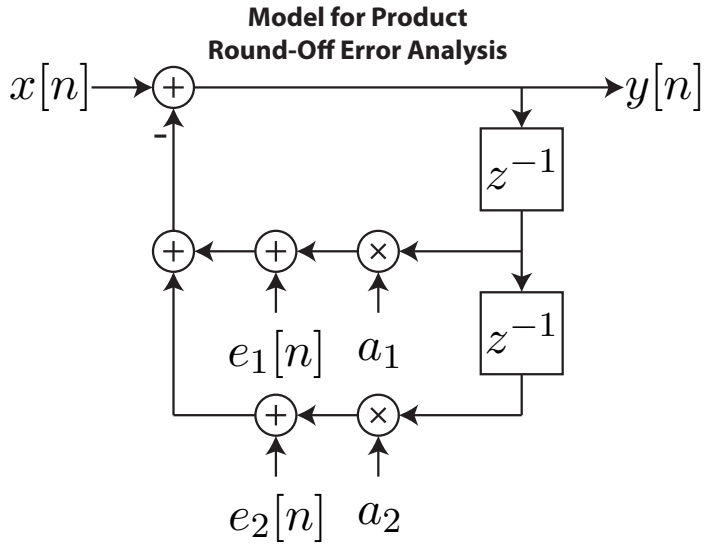
1. Determine the "noise transfer function" $G_\ell(z) = \frac{Y(z)}{E_\ell(z)}$.
2. Compute the output noise variance σ_ℓ^2 caused by roundoff error $e_\ell[n]$ as

$$\sigma_\ell^2 = \sigma_e^2 \cdot \frac{1}{2\pi} \int_{-\pi}^{\pi} |G_\ell(e^{j\omega})|^2 \, d\omega = \sigma_e^2 \cdot \sum_{n=-\infty}^{\infty} |g_\ell[n]|^2$$

The total round-off noise variance at output is then $\sigma_{tot}^2 = \sum_{\ell=1}^{L} \sigma_\ell^2$.

Effect of Product Round-Off: Simple Example





We denote the round-off noise variance for both products as σ_e^2. We have

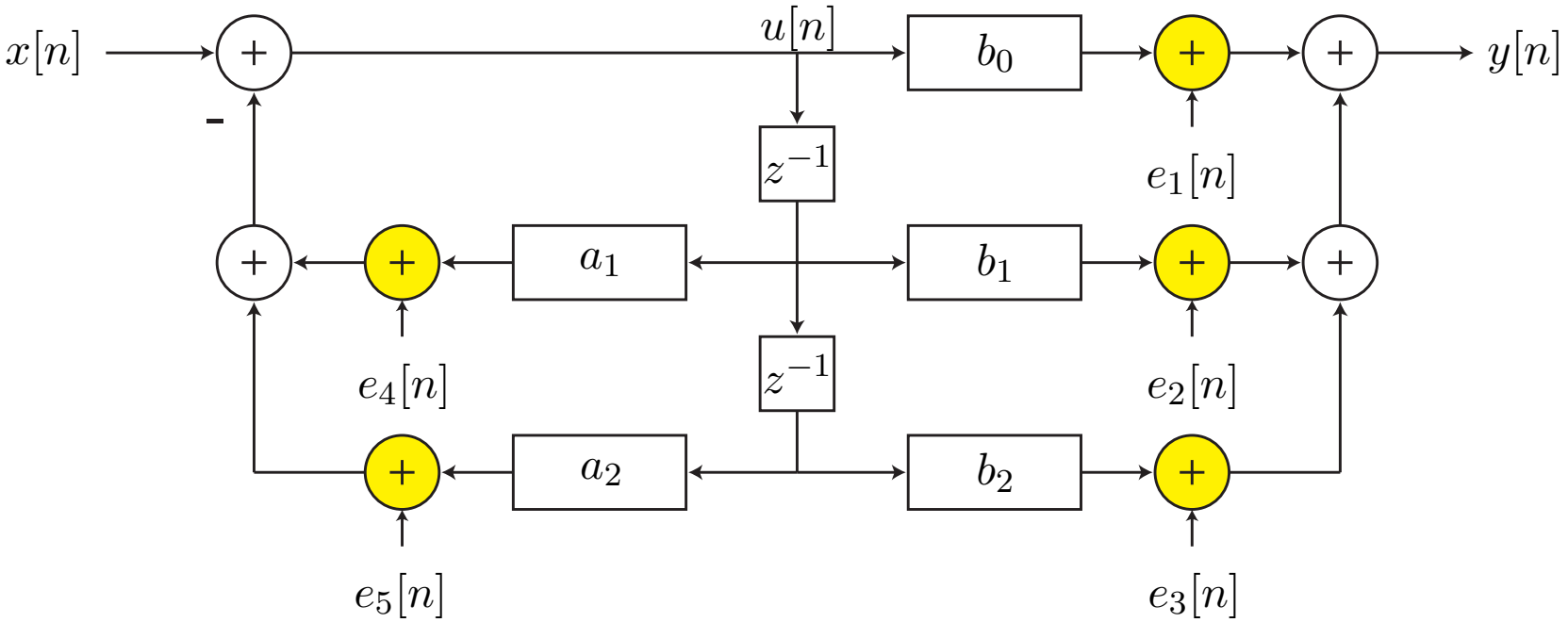
$$G_1(z) = G_2(z) = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2}} = H(z)$$

Hence, assuming a_1 and a_2 are such that $H(z)$ is stable,

$$\sigma_{tot}^2 = 2\sigma_e^2 \frac{1}{2\pi} \int_{-\pi}^{\pi} |G_1(e^{j\omega})|^2 \, d\omega = 2\sigma_e^2 \left(\frac{1 + a_2}{1 - a_2} \right) \left(\frac{1}{1 + 2a_2 + a_2^2 - a_1^2} \right)$$

Product Roundoff Noise: DF-II Second Order Section

Given the realization structure with product roundoff errors:



The roundoff errors appear at the output as

$$Y(z) = E_1(z) + E_2(z) + E_3(z) + H(z)(E_4(z) + E_5(z))$$

where $H(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}}$. Hence the variance of the product roundoff noise at the output is

$$\sigma_y^2 = 3\sigma_e^2 + 2\sigma_e^2 \cdot \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(e^{j\omega})|^2 \, d\omega.$$

Product Roundoff Noise: General DF-I and DF-II

Assume $H(z) = \frac{B(z)}{A(z)}$ with N non-unity denominator coefficients and $M+1$ non-unity numerator coefficients.

For a general direct-form II filter, similar analysis techniques can be used to show

$$\sigma_y^2 = \underbrace{N\sigma_e^2 \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(e^{j\omega})|^2 \, d\omega}_{\text{noises that propagate through filter}} + \underbrace{(M+1)\sigma_e^2}_{\text{noises directly connected to output}}$$

Note that $\sigma_y^2 = \sigma_e^2 \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(e^{j\omega})|^2 \, d\omega + \sigma_e^2$ if an extended-precision accumulator is used.

For a direct-form I filter, inspection of the signal flow graph shows that all of the roundoff noises propagate through $G_i(z) = \frac{1}{A(z)}$. Hence,

$$\sigma_y^2 = \begin{cases} (M+1+N)\sigma_e^2 \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{|A(e^{j\omega})|^2} \, d\omega & \text{standard accumulation} \\ \sigma_e^2 \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{|A(e^{j\omega})|^2} \, d\omega & \text{extended-precision accumulation.} \end{cases}$$

2nd Order DF-I vs. DF-II Example

Suppose $\sigma_e^2 = 1$ and

$$H(z) = \frac{0.6 + 0.54z^{-1} + 0.108z^{-2}}{1 - 1.3z^{-1} + 0.4z^{-2}}$$

with ROC $|z| > 0.8$ (causal and stable).

With these numbers, we can compute the relevant integrals (using, for example, the algebraic technique) to be

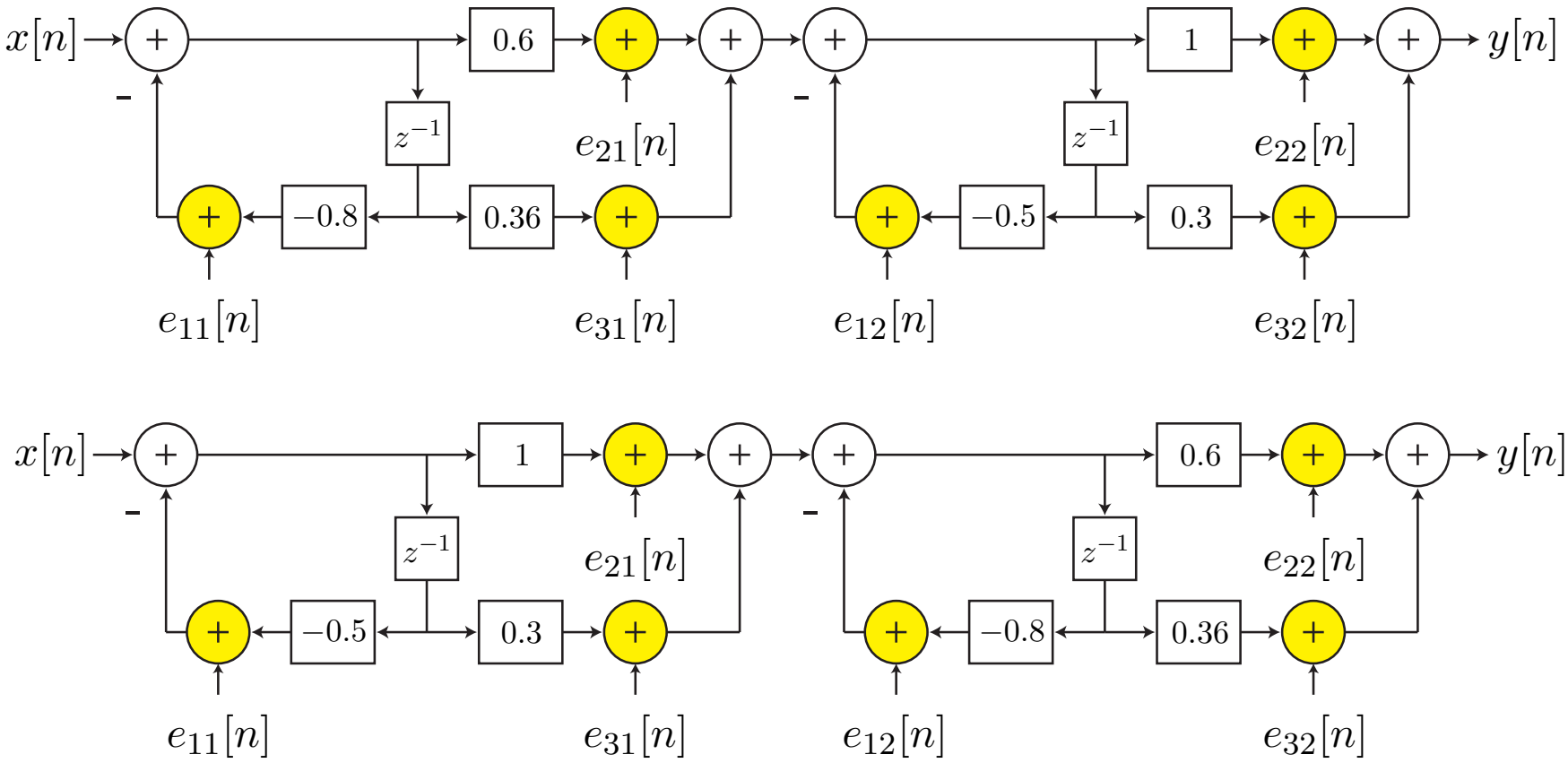
$$\frac{1}{2\pi}\int_{-\pi}^{\pi} |H(e^{j\omega})|^2 \, d\omega \approx 12.7719 \qquad \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{1}{|A(e^{j\omega})|^2} \, d\omega \approx 8.6420$$

which means that

$$\sigma_y^2 = \begin{cases} 3 + 2 \cdot 12.7719 = 28.5438 & \text{DF-II} \\ 5 \cdot 8.6420 = 43.2099 & \text{DF-I} \end{cases}$$

Cascaded DF-II Realization Structure

Now, what if we split this realization structure up into a cascade of two first order sections? There are (at least) four possibilities. Here are two:



Note both realizations have the same $H(z) = H_1(z)H_2(z) = H_2(z)H_1(z)$. Is there any difference?

Analysis of Cascaded DF-II Realization Structure

In each realization structure, some of the noises propagate through $H_1(z)$, some propagate through $H_2(z)$, some propagate through $H(z) = H_1(z)H_2(z)$, and some are directly connected to the output. Given

$$H_1(z) = \frac{0.6 + 0.36z^{-1}}{1 - 0.8z^{-1}}$$

$$H_2(z) = \frac{1 + 0.3z^{-1}}{1 - 0.5z^{-1}}$$

we can use algebraic methods (or the usual series convergence results) to compute

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} |H_1(e^{j\omega})|^2 \, d\omega \approx 2.32$$

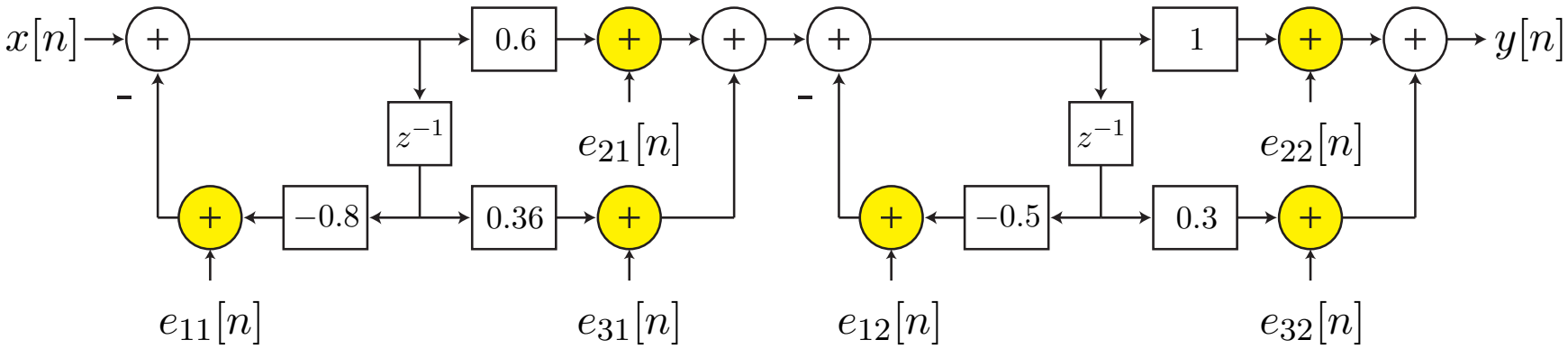
$$\frac{1}{2\pi}\int_{-\pi}^{\pi} |H_2(e^{j\omega})|^2 \, d\omega \approx 1.8533$$

Also recall our earlier result

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} |H(e^{j\omega})|^2 \, d\omega \approx 12.7719.$$

Analysis of First Cascaded DF-II Realization Structure

For the first realization:



- $e_{11}[n]$ propagates through $H(z)$ to get to the output.
- $e_{21}[n]$, $e_{31}[n]$, and $e_{12}[n]$ propagate through $H_2(z)$ to get to the output.
- $e_{22}[n] = 0$.
- $e_{23}[n]$ is directly connected to the output.

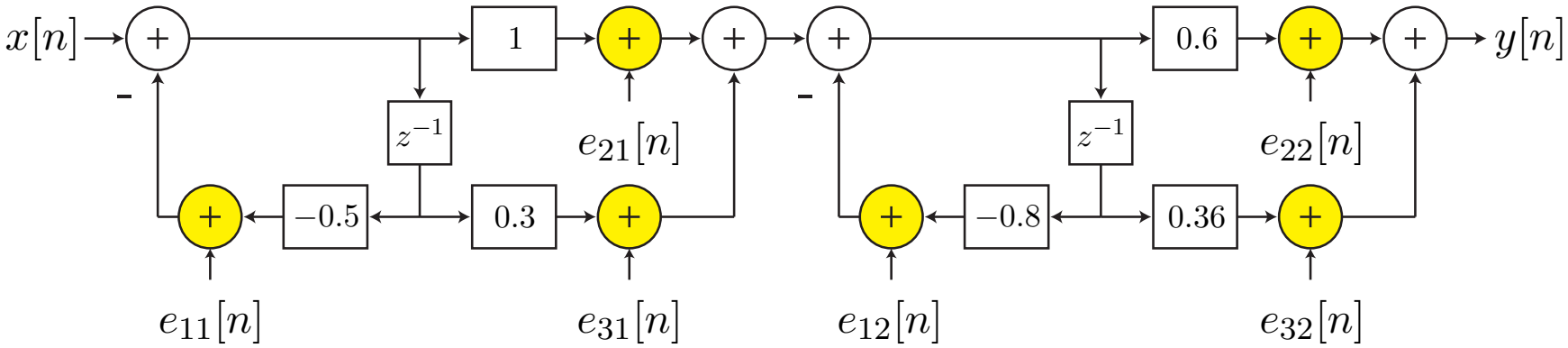
Hence, assuming $\sigma_e^2 = 1$, we have

$$\sigma_y^2 = 12.7719 + 3 \cdot 1.8533 + 0 + 1 \approx 19.33$$

which is better than the regular DF-II realization with $\sigma_y^2 \approx 28.54$.

Analysis of Second Cascaded DF-II Realization Structure

In the second realization:



- $e_{11}[n]$ propagates through $H(z)$ to get to the output.
- $e_{21}[n] = 0$.
- $e_{31}[n]$, and $e_{12}[n]$ propagate through $H_1(z)$ to get to the output.
- $e_{22}[n]$ and $e_{23}[n]$ are directly connected to the output.

Hence, assuming $\sigma_e^2 = 1$, we have

$$\sigma_y^2 = 12.7719 + 0 + 2 \cdot 2.32 + 0 + 2 \approx 19.41$$

which is slightly worse than the first ordering.