

Working with Fixed Point Arithmetic

Brett Ninness

`brettee.newcastle.edu.au`

School of Electrical Engineering and Computer Science

The University of Newcastle

Background

- Digital hardware is now the primary implementation means:
 - Feedback control;
 - Signal processing.
- Fundamental question:
 - Fixed or floating point hardware?
- Disadvantage of fixed point: Dynamic range small - scaling necessary to avoid under/overflow;
- Advantages of fixed point:
 - Logic circuits simpler, less memory, less processor speed;
 - Implies smaller, faster, cheaper, lower power consumption implementations.

Fixed vs floating

- IEEE-754 32 bit floating point standard:
 - 1 sign bit;
 - 23 bit mantissa;
 - 8 bit exponent.

$$\underbrace{\pm}_{1 \text{ bit}} 0.\underbrace{101101011 \dots 01}_{23 \text{ bits}} \times 2^{0-2^8}$$

- Fixed point - up to programmer

$$\underbrace{\pm}_{1 \text{ bit}} \underbrace{10111 \dots 1}_n \cdot \underbrace{010011 \dots 1}_m$$

Slope-Bias encoding scheme

- Radix point (where decimal point sits) determined by software;
- Location of radix point determines scaling operations;
- Slope-bias encoding scheme is standard:

$$\underbrace{V}_{\text{value}} \approx \underbrace{\tilde{V}}_{\text{approx value}} = \overbrace{S}^{\text{scaling}} \underbrace{Q}_{\text{quantised}} + \underbrace{B}_{\text{bias}}$$

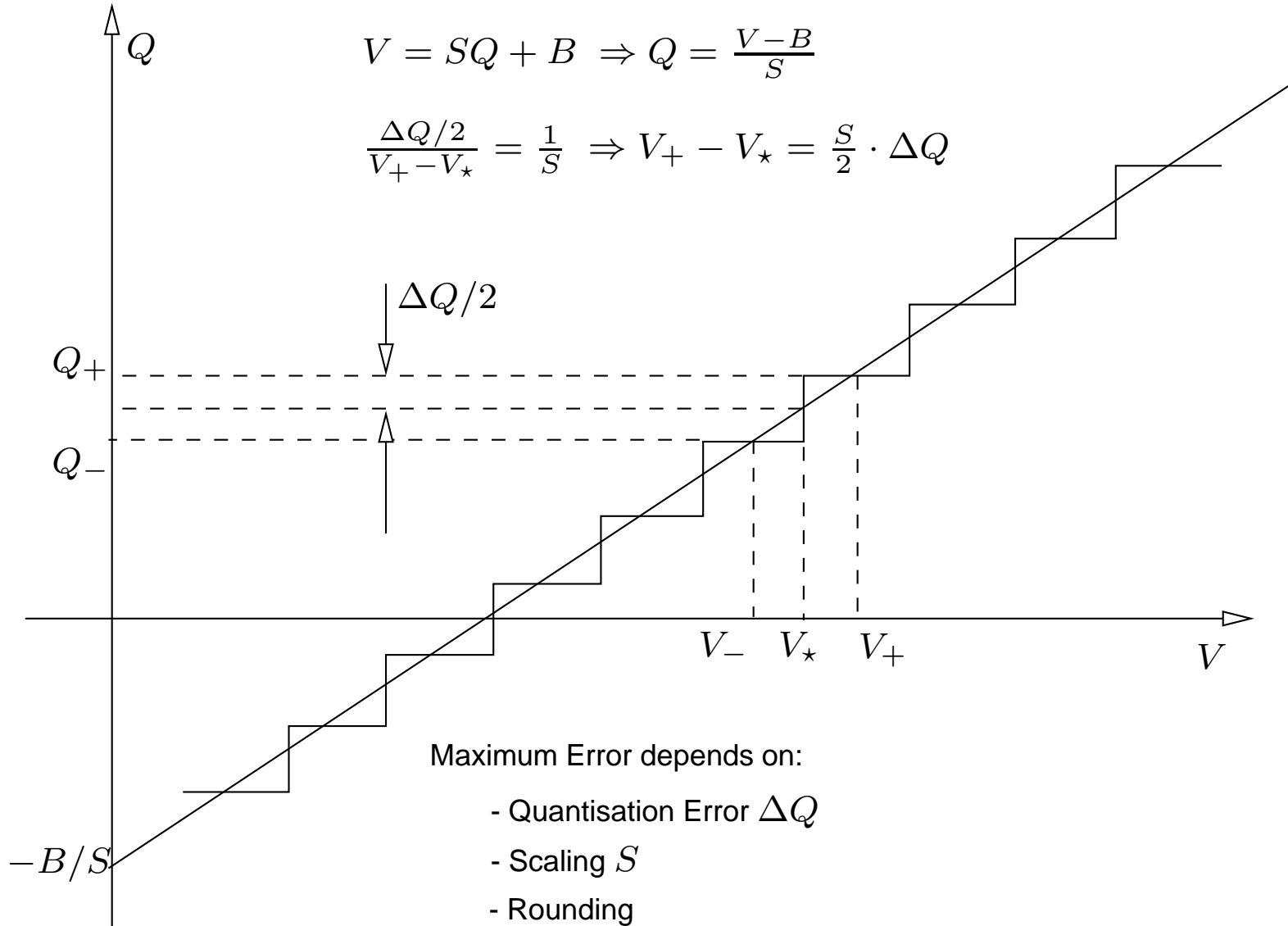
Example-Water Temperature

- Measured using 8 bit unsigned char;
- Option 1: $0 - 100^{\circ}\text{C} = (0 - 100)$ bits;
 - Good point: Easy;
 - Bad point: 60% of number range (101 – 255) unused;
- Option 2: $0 - 100^{\circ}\text{C} = (0 - 255)$ bits;
 - Better accuracy;

$$.^{\circ}\text{C} = \frac{\text{char}}{2.55} \left. \vphantom{\frac{\text{char}}{2.55}} \right\} \text{Expensive computation}$$

- Option 3: $0 - 100^{\circ}\text{C} = (0 - 200)$ bits;
 - Better accuracy than option 1;
 - Easy conversion arithmetic;
 - (201 – 255) wasted.

In General



Example of Bias Use

- Electronically controlled engine;
- Maintain air/fuel ratio;
- Infer flow from pressure and temperature measurements using ideal gas equation $P \cdot V = \alpha \cdot T$;
- Implies division by $^{\circ}\text{K}$.
- Temperature range 222 – 380 $^{\circ}\text{K}$;
- Quantisation: 3 bit unsigned int.

Key point: slope S required to span total range and ΔQ may both depend on bias B .

Example: Value from quantisation

● $S = F \cdot 2^{-E}$ $F \in [1, 2]$ considered from now on;

● Example:

$$Q = 10110101, \quad F = 1, E = 4, B = 0$$

$$\begin{aligned}\tilde{V} &= F2^{-E}Q \\ &= 2^{-4}Q \\ &= 2^{-4} (1 \cdot 2^7 + 0 \cdot 2^6 + 1 \cdot 2^5 + 1 \cdot 2^4 + 0 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0) \\ &= 8 + 2 + 1 + 0.25 + 0.075 = 11.3125.\end{aligned}$$

Implicit radix point 1011.0101.

Notes

- Quantisation ΔQ may be manipulated in software by choice of radix point;
- S and B are constants only used when final physical value needs to be realised;
- Scaling choices based on
 - Maximum precision;
 - Minimum number of arithmetic operations;
 - S may be inherited by hardware (A/D, D/A);
 - S may be a design choice.

Addition

$$V_a = F_a 2^{-E_a} Q_a + B_a \quad V_b = F_b 2^{-E_b} Q_b + B_b$$

$$V_c = F_c 2^{-E_c} Q_c + B_c$$

Then $V_a = V_b + V_c$ implies

$$F_a 2^{-E_a} Q_a + B_a = F_b 2^{-E_b} Q_b + B_b + F_c 2^{-E_c} Q_c + B_c.$$

That is

$$Q_a = \frac{F_b}{F_a} 2^{-(E_b - E_a)} \cdot Q_b + \frac{F_c}{F_a} 2^{-(E_c - E_a)} \cdot Q_c + 2^{E_a} \frac{(B_b + B_c - B_a)}{F_a}$$

Therefore, in general:

$$Q_a \neq Q_b + Q_c$$

Scaling for Speed

$$F_b = F_a = F_c, \quad B_a = B_b + B_c.$$

$$Q_a = 2^{-(E_b - E_a)} Q_b + 2^{-(E_c - E_a)} Q_c$$

Software

$$a = b \gg (E_b - E_a) + c \gg (E_c - E_a);$$

Note

$$\begin{aligned} 2^{-n} Q &= 2^{-n} (b_m 2^m + b_{m-1} 2^{m-1} + \dots + b_1 2 + b_0) \\ &= b_m 2^{m-n} + b_{m-1} 2^{m-1-n} + \dots + b_1 2^{1-n} + b_0 2^{-n} \\ &= 101101 \dots 1 \longrightarrow n \end{aligned}$$

Multiplication $V_a = V_b * V_c$

$$\begin{aligned} F_a 2^{-E_a} Q_a + B_a &= (F_b 2^{-E_b} Q_b + B_b) (F_c 2^{-E_c} Q_c + B_c) \\ &= F_b F_c 2^{-(E_b+E_c)} Q_b Q_c + F_b 2^{-E_b} B_c Q_b + \\ &\quad F_c 2^{-E_c} B_b Q_c + B_b B_c. \end{aligned}$$

Therefore

$$\begin{aligned} Q_a &= \frac{F_b F_c}{F_a} \cdot 2^{-(E_b+E_c-E_a)} Q_b Q_c + \frac{F_b}{F_a} \cdot 2^{-(E_b-E_a)} Q_b B_c + \\ &\quad \frac{F_c}{F_a} \cdot 2^{-(E_c-E_a)} Q_c B_b + \frac{2^{E_a} (B_b B_c - B_a)}{F_a}. \end{aligned}$$

Consequently, in general $Q_a \neq Q_b * Q_c$

Option: Scale for Speed

$$F_b = F_a = F_c = 1, \quad B_a = B_b = B_c = 0.$$

Then

$$Q_a = 2^{-(E_b + E_c - E_a)} Q_b \cdot Q_c$$

Code:

$$a = (b * c) >> (E_b + E_c - E_a);$$

Note

$$E_b = E_c = E_a = n$$

implies

$$a = (b * c) >> n;$$

Example

$$\begin{aligned}10 \times 4 &= 2^{-n}Q_a \times 2^{-n}Q_b \\&= 1010.0 \times 0100.0 \\&= 20 \times 8 \\&= 160 \\&= 10100000 \xrightarrow{\gg 1} \underbrace{01010000}_{40} \cdot 0.\end{aligned}$$

$$c = a * b \gg n;$$

Division $V_a = V_b/V_c$

$$F_a 2^{-E_a} Q_a + B_a = \frac{F_b 2^{-E_b} Q_b + B_b}{F_c 2^{-E_c} Q_c + B_c}$$

Therefore

$$Q_a = \frac{F_b 2^{-(E_b-E_a)} Q_b + 2^{E_a} B_b}{F_a F_c 2^{-E_c} Q_c + F_a B_c} - \frac{B_a 2^{E_a}}{F_a}.$$

Therefore, in general

$$Q_a \neq \frac{Q_b}{Q_c}.$$

Division scaling for speed

$$B_a = B_b = B_c = 0, \quad F_a = F_b = F_c = 1;$$

Then

$$Q_a = 2^{-(E_b - E_a - E_c)} \frac{Q_b}{Q_c}$$

hence

$$a = (b/c) \gg (E_b - E_a - E_c) ;$$

Special case of

$$E_a = E_b = E_c = n;$$

$$a = (b/c) \ll n ;$$

Example

$$\begin{aligned}\frac{10}{4} &= \frac{2^{-n}Q_a}{2^{-n}Q_b} \\ &= \frac{1010.0}{0100.0} \\ &= \frac{20}{8} \\ &= 2 \\ &= 00010 \xrightarrow{\ll 1} \underbrace{0010}_2 \cdot 0.\end{aligned}$$

$$c = (a/b) \ll n;$$

Example-Improved Precision

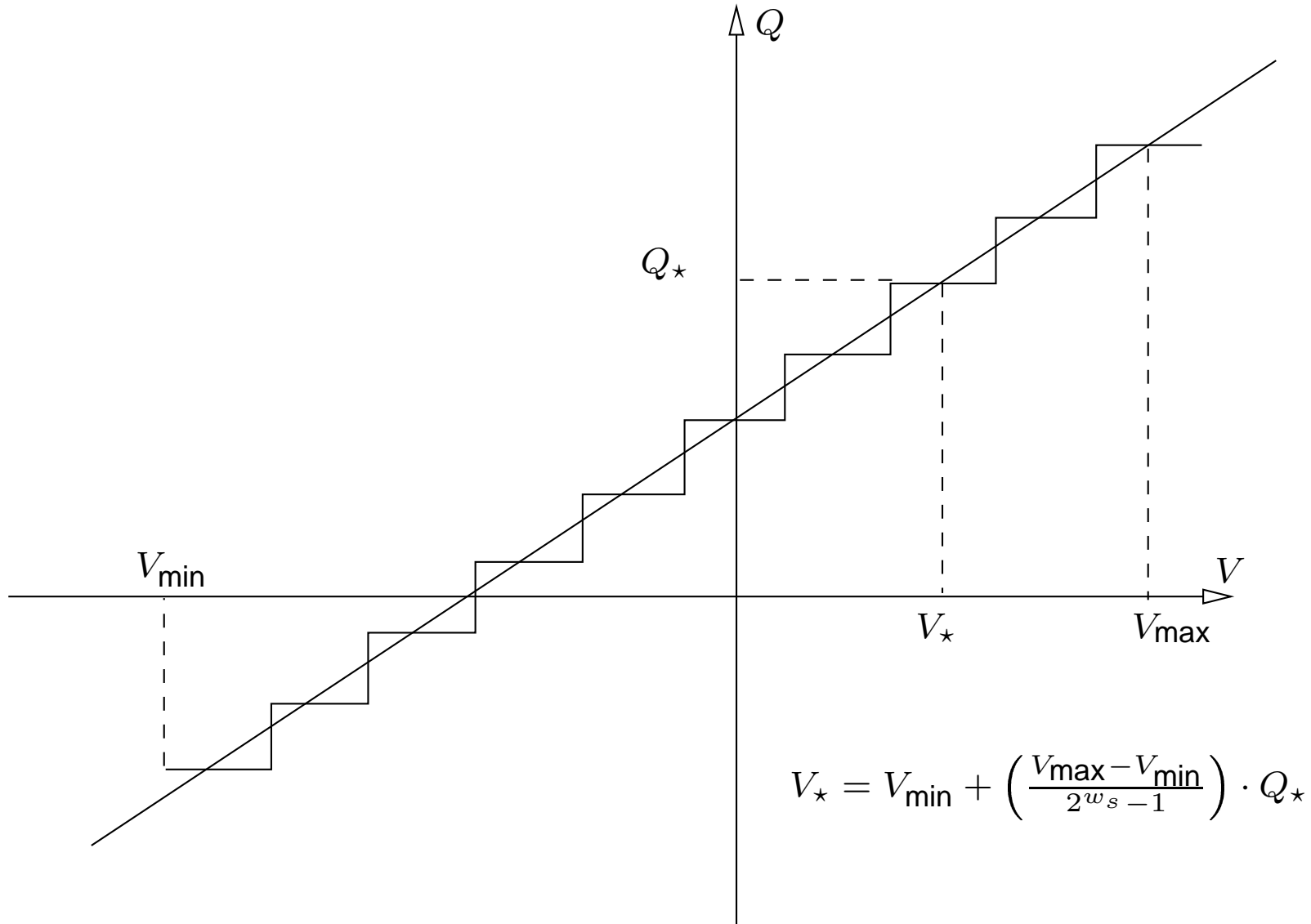
$$2^{-n}Q_a = \frac{2^{-n}Q_b}{2^{-n}Q_c} \Rightarrow Q_a = \frac{2^n Q_b}{Q_c}$$

Suggests

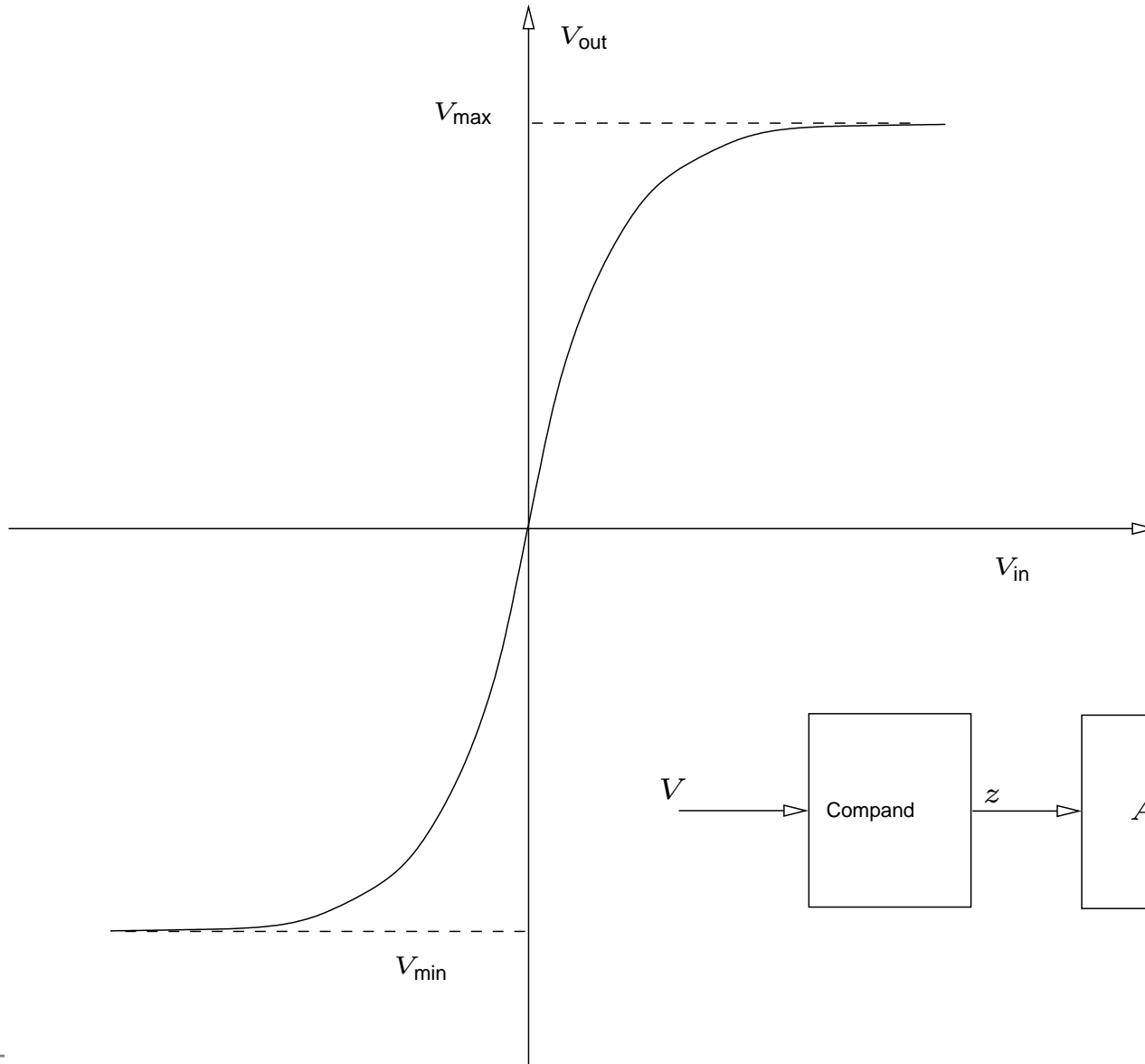
$$c = (a \ll n) / b;$$

$$\begin{aligned} \frac{10}{4} &= \frac{01010.0}{00100.0} \\ &\Rightarrow \frac{101000}{001000} \\ &= \frac{40}{8} \\ &= 5 \\ &= 00010 \cdot 1 = 2.5 \end{aligned}$$

Analog to Digital Conversion



Comanding



Comanding

- μ Law

$$z = \text{sgn}(V) \cdot \frac{\log(1 + \mu|V|)}{\log(1 + \mu)}$$

- A Law

$$z = \begin{cases} \frac{AV}{1 + \log A} & ; V \in [0, 1/A] \\ \frac{1 + \log(AV)}{1 + \log A} & ; V \in [1/A, 1] \end{cases}$$

- Worse SNR than μ law;
- Better dynamic range than μ law.

Ordering of Operations

$$y[k] = a * y[k - 1] + u[k], \quad a = (1.0156)_{10} = (\underbrace{01.000001}_{6\text{bits}})_2$$

Therefore, if only 6 bits of precision are available the above is equivalent to

$$y[k] = y[k - 1] + u[k] \quad \times!$$

Better ordering

$$y[k] = y[k - 1] + \underbrace{0.0156}_{=(0.000001)_2=6\text{bits}} * y[k] + u[k];$$

Limit Cycle Oscillations

$$y[k] = Q(a * y[k - 1]) + u[k].$$

- Suppose quantisation involves 5 bits made up of 4 bits of magnitude plus one sign bit, and suppose $a = -1/2$.
- Suppose we start with $u[0] = 15/16 = 0.1111$.

	k	$y[k]$	value
	0	0.1111	15/16
	1	1.1000	-8/16
Then	2	0.0100	4/16
	3	1.0010	-2/16
	4	0.0001	1/16
	5	1.0001	-1/16