

Chapter-6

Numerical considerations

2's complement	Decimal value	Excess-3
101	-3	000
110	-2	001
111	-1	010
000	0	011
001	1	100
010	2	101
011	3	110
100	Reserved pattern	111

FIGURE 6.1: Excess-3 encoding, sorted by excess-3 ordering.

000	0
001	1
010	2
011	3
100	4
101	5
110	6
111	7

FIGURE 6.2: Representable numbers of a 3-bit unsigned integer format.



FIGURE 6.3: Representable numbers of a 3-bit unsigned integer format.

		No-zero		Abrupt underflow		Denorm	
E	M	$S = 0$	$S = 1$	$S = 0$	$S = 1$	$S = 0$	$S = 1$
00	00	2^{-1}	$-(2^{-1})$	0	0	0	0
	01	$2^{-1} + 1 \cdot 2^{-3}$	$-(2^{-1} + 1 \cdot 2^{-3})$	0	0	$1 \cdot 2^{-2}$	$-1 \cdot 2^{-2}$
	10	$2^{-1} + 2 \cdot 2^{-3}$	$-(2^{-1} + 2 \cdot 2^{-3})$	0	0	$2 \cdot 2^{-2}$	$-2 \cdot 2^{-2}$
	11	$2^{-1} + 3 \cdot 2^{-3}$	$-(2^{-1} + 3 \cdot 2^{-3})$	0	0	$3 \cdot 2^{-2}$	$-3 \cdot 2^{-2}$
01	00	2^0	$-(2^0)$	2^0	$-(2^0)$	2^0	$-(2^0)$
	01	$2^0 + 1 \cdot 2^{-2}$	$-(2^0 + 1 \cdot 2^{-2})$	$2^0 + 1 \cdot 2^{-2}$	$-(2^0 + 1 \cdot 2^{-2})$	$2^0 + 1 \cdot 2^{-2}$	$-(2^0 + 1 \cdot 2^{-2})$
	10	$2^0 + 2 \cdot 2^{-2}$	$-(2^0 + 2 \cdot 2^{-2})$	$2^0 + 2 \cdot 2^{-2}$	$-(2^0 + 2 \cdot 2^{-2})$	$2^0 + 2 \cdot 2^{-2}$	$-(2^0 + 2 \cdot 2^{-2})$
	11	$2^0 + 3 \cdot 2^{-2}$	$-(2^0 + 3 \cdot 2^{-2})$	$2^0 + 3 \cdot 2^{-2}$	$-(2^0 + 3 \cdot 2^{-2})$	$2^0 + 3 \cdot 2^{-2}$	$-(2^0 + 3 \cdot 2^{-2})$
10	00	2^1	$-(2^1)$	2^1	$-(2^1)$	2^1	$-(2^1)$
	01	$2^1 + 1 \cdot 2^{-1}$	$-(2^1 + 1 \cdot 2^{-1})$	$2^1 + 1 \cdot 2^{-1}$	$-(2^1 + 1 \cdot 2^{-1})$	$2^1 + 1 \cdot 2^{-1}$	$-(2^1 + 1 \cdot 2^{-1})$
	10	$2^1 + 2 \cdot 2^{-1}$	$-(2^1 + 2 \cdot 2^{-1})$	$2^1 + 2 \cdot 2^{-1}$	$-(2^1 + 2 \cdot 2^{-1})$	$2^1 + 2 \cdot 2^{-1}$	$-(2^1 + 2 \cdot 2^{-1})$
	11	$2^1 + 3 \cdot 2^{-1}$	$-(2^1 + 3 \cdot 2^{-1})$	$2^1 + 3 \cdot 2^{-1}$	$-(2^1 + 3 \cdot 2^{-1})$	$2^1 + 3 \cdot 2^{-1}$	$-(2^1 + 3 \cdot 2^{-1})$
11	Reserved pattern						

FIGURE 6.4: Representable numbers of no-zero, abrupt underflow, and denorm formats.

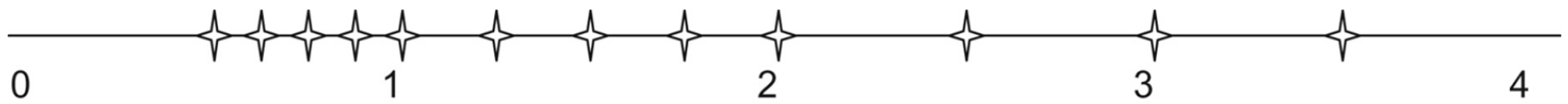


FIGURE 6.5: Representable numbers of the no-zero representation.



FIGURE 6.6: Representable numbers of the abrupt underflow format.



FIGURE 6.7: Representable numbers of a denormalization format.

Exponent	Mantissa	Meaning
11...1	$\neq 0$	NaN
11...1	$= 0$	$(-1)^S * \infty$
00...0	$\neq 0$	denormalized
00...0	$= 0$	0

FIGURE 6.8: Special bit patterns in the IEEE standard format.

$\begin{array}{rrcr} 3X & +5Y & +2Z & = 19 \\ 2X & +3Y & + Z & = 11 \\ X & +2Y & +2Z & = 11 \end{array}$ <p style="text-align: center;">Original</p>	$\begin{array}{rrcr} X & +5/3Y & +2/3Z & = 19/3 \\ X & +3/2Y & +1/2Z & = 11/2 \\ X & + 2Y & + 2Z & = 11 \end{array}$ <p style="text-align: center;">Step 1: divide Equation 1 by 3, Equation 2 by 2</p>	$\begin{array}{rrcr} X & +5/3Y & +2/3Z & = 19/3 \\ & -1/6Y & -1/6Z & = -5/6 \\ & 1/3Y & +4/3Z & = 14/3 \end{array}$ <p style="text-align: center;">Step 2: subtract Equation 1 from Equation 2 and Equation 3</p>
$\begin{array}{rrcr} & X & +5/3Y & +2/3Z & = 19/3 \\ & & Y & + Z & = 5 \\ & & Y & +4Z & = 14 \end{array}$ <p style="text-align: center;">Step 3: divide Equation 2 by -1/6 and Equation 3 by 1/3</p>	$\begin{array}{rrcr} & X & +5/3Y & +2/3Z & = 19/3 \\ & & Y & + Z & = 5 \\ & & & + 3Z & = 9 \end{array}$ <p style="text-align: center;">Step 4: subtract Equation 2 from Equation 3</p>	
$\begin{array}{rrcr} & X & +5/3Y & +2/3Z & = 19/3 \\ & & Y & + Z & = 5 \\ & & & Z & = 3 \end{array}$ <p style="text-align: center;">Step5 : divide Equation 3 by 3 Solution for Z!</p>	$\begin{array}{rrcr} & X & +5/3Y & +2/3Z & = 19/3 \\ & & Y & & = 2 \\ & & & Z & = 3 \end{array}$ <p style="text-align: center;">Step 6: substitute Z solution into Equation 2. Solution for Y!</p>	
$\begin{array}{rrcr} & X & & & = 1 \\ & & Y & & = 2 \\ & & & Z & = 3 \end{array}$ <p style="text-align: center;">Step 7: substitute Y and Z into Equation 1. Solution for X!</p>		

FIGURE 6.9: Gaussian elimination and backward substitution for solving systems of linear equations.

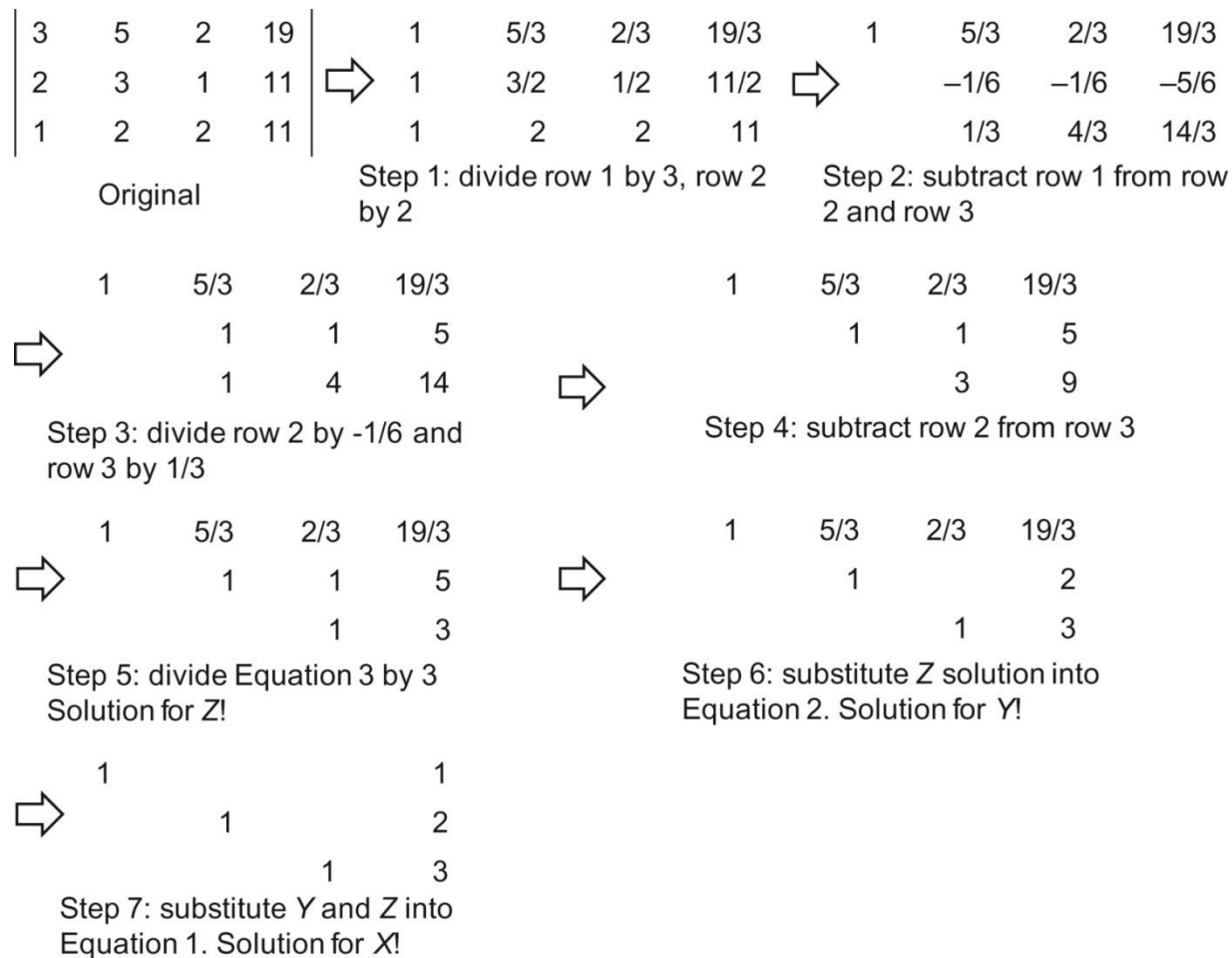


FIGURE 6.10: Gaussian elimination and backward substitution in matrix view

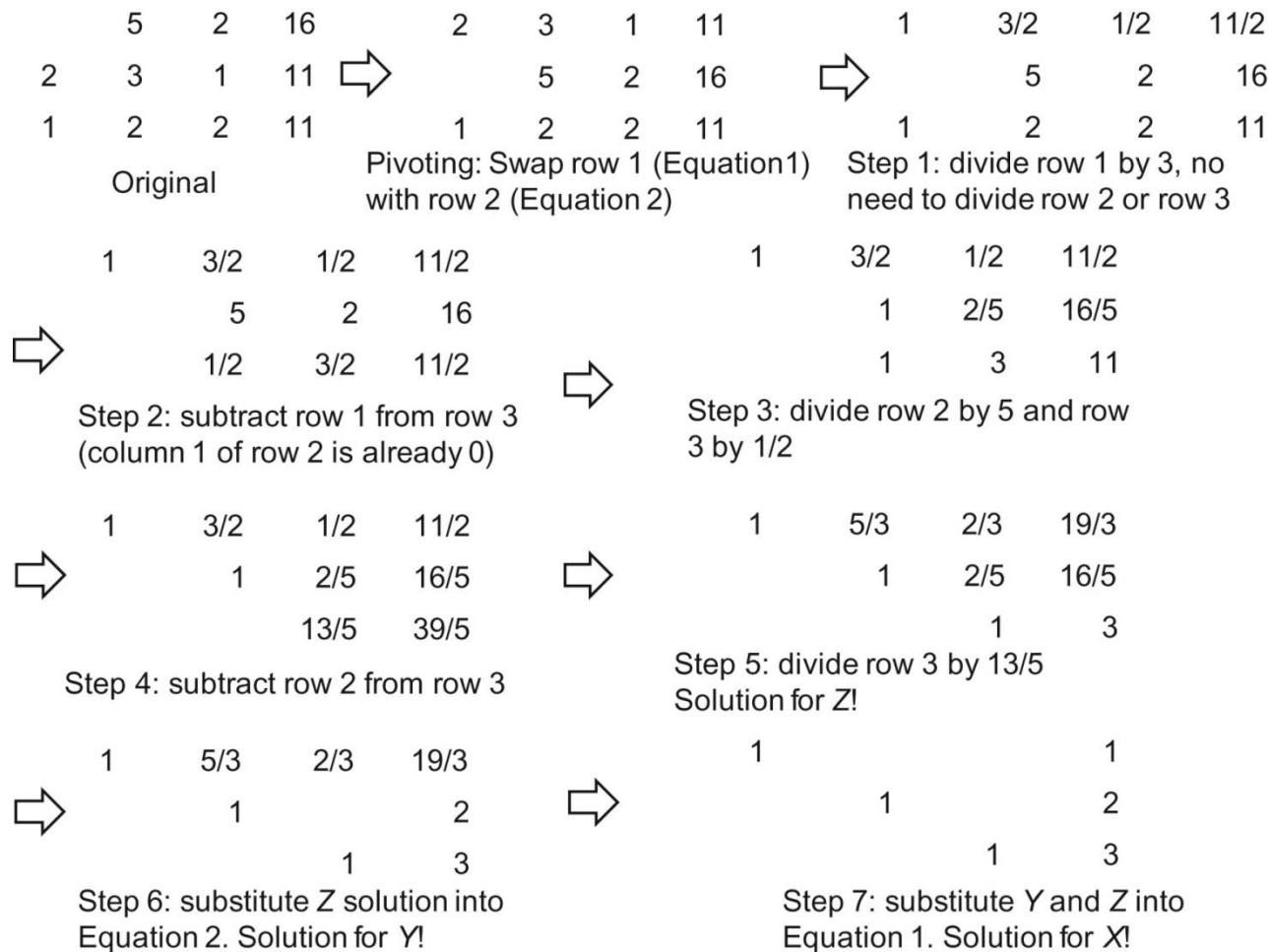


FIGURE 6.11: Gaussian elimination with pivoting.