Exercises on the topics of class 6

Exercises with solutions

Ex. 1. Consider < 0 , 10011 , 1100001100 > and < 0 , 01010 , 1011000000 >; multiply them. The obtained result is exact or an approximation of the correct result?

SOLUTION:

The product is positive; its mantissa and exponent (not normalized, in general) are respectively the matissas product and the exponent sum minus the bias. In the example, the result is already normalized:

1,110000110 × 1,101100000 = 10,111110010001
10011 + 01010 - 01111 = 01110

The normalized result is < 0 , 01111 , 0111110010 > where we lost the last bit of the mantissa (approximated representation).

Ex. 2. Sum the following two half precision floating point numbers. What kind of problem do we have in doing their product?

< 1 , 11010 , 1100000000 > < 0 , 11110 , 1111001000 >

SOLUTION:

Let's bring the first number to have the exponent of the second one, thus obtaining

< 1 , 00000 , 0001110000 >

Their sum will be positive, since the mantissa of the positive operand is greater. The resulting mantissa is given by

$$\begin{array}{r} 1{,}1111001000 - \\ 0{,}0001110000 = \\ \hline 1{,}1101011000 \end{array}$$

whereas the exponent is 11110. Thus, the triple representation of the result is

< 0 , 11110 , 1101011000 >

If we multiply the given numbers, the resulting exponent (obtained by summing the exponents and by then subtracting the bias) cannot be represented with 5 bits. Indeed,

11010 + 11110 – 01111 = 101001

that requires 6 bits to be represented (exponent overflow!).

Ex. 3. Multiply the following two half precision IEEE flaoting point numbers. What kind of anomaly do we meet by summing them?

< 1 , 10110 , 1000000000 > < 0 , 01011 , 1100000000 >

SOLUTION:

The product is negative (having the operands different signs). The (non normalized) mantissa is

$$\begin{array}{r} 1,11 \times \\ 1,1 = \\ \hline 111 \\ 111 - \\ \hline 10,101 \end{array}$$

whereas the (non normalized) exponent is 10110 + 01011 - 01111 = 10010.
Thus, the normalized representation of the result is

< 1 , 10011 , 0101000000 >

The anomaly in the sum is that, when we bring the second operand to the exponent of the first one, we obtain < 0, 00000 , 0000000000 >, i.e. we loose the second operand.

Exercises without solutions

Ex. 1. Sum the following two half precision IEEE numbers:

< 1 , 01010 , 1100100000 > < 0 , 01100 , 1101000000 >

Ex. 2. Multiply the following two hal precision IEEE numbers:

< 1, 01010 , 0010000000 > < 0, 10101 , 0100000000 >

Ex. 3. Convert in the half precision IEEE standard and then calculate $(3,\frac{1}{4}) + (5,\frac{3}{4})$ and then verify the result obtained by converting it back to base 10.