

1. First,  $x = (1011.\overline{010})_2$  has integer part  $x_{\text{int}} = (1011.0)_2 = 11$  and fractional part  $x_{\text{frac}} = (0.\overline{010})_2$ .

$$8x_{\text{frac}} = (010.\overline{010})_2 \implies 7x_{\text{frac}} = (010.0)_2 = 2 \implies \boxed{x_{\text{frac}} = 2/7} \implies \boxed{(1011.\overline{010})_2 = 11\frac{2}{7} = \frac{79}{7}}$$

Second,  $x = (11.101\overline{01})_2$  has integer part  $x_{\text{int}} = (11.0)_2 = 3$  and fractional part  $x_{\text{frac}} = (0.101\overline{01})_2$ .

$$8x_{\text{frac}} = (101.\overline{01})_2 \text{ and } 32x_{\text{frac}} = (10101.\overline{01})_2 \implies 24x_{\text{frac}} = (10101.0)_2 - (101.0)_2 = 21 - 5 = 16.$$

Therefore,  $\boxed{x_{\text{frac}} = 16/24 = 2/3}$  and  $\boxed{(11.101\overline{01})_2 = 3\frac{2}{3} = \frac{11}{3}}$ .

2. Clearly  $2.75 = (10.11)_2 = (1.011)_2 \times 2^1$ . Therefore, since 2.75 is associated with a mantissa that is exactly represented in 52 binary digits (indeed, exactly in only 3 binary digits),  $\text{fl}(2.75) = 2.75$  exactly, in which case

$$\left| \frac{\text{fl}(2.75) - 2.75}{2.75} \right| = 0 < \frac{1}{2}\epsilon_{\text{mach}}$$

clearly holds. To get the machine representation, consider

$$(1.011)_2 = +1. \underbrace{0110 \ 0000 \ 0000 \ 0000 \ 0000 \ 0000 \ 0000 \ 0000 \ 0000 \ 0000 \ 0000 \ 0000 \ 0000 \ 0000}_{52 \text{ bits}} \times 2^1.$$

The sign bit is 0, because 2.75 is positive. The true exponent is  $1 = F - 1023$ , so  $F = 1024 = (10000000000)_2$ , and therefore, the machine representation is

$$\boxed{0 \mid 100 \ 0000 \ 0000 \mid 0110 \ 0000 \ 0000 \ 0000 \ 0000 \ 0000 \ 0000 \ 0000 \ 0000 \ 0000 \ 0000 \ 0000 \ 0000 \ 0000 \ 0000 \ 0000},$$

that is 4006000000000000 in hex format (verify in MATLAB).

3. Let  $a = 3344556600$  and  $b = 1.2222222$  be the lengths of the right triangle's sides. Its hypotenuse is then

$$c = \sqrt{a^2 + b^2} = \sqrt{3344556600^2 + 1.2222222^2}$$

At the command line in OCTAVE we find

```
octave:10> a=3344556600
a = 3.344556600000000e+09
octave:11> b = 1.2222222
b = 1.222222200000000e+00
octave:12> sqrt(a^2+b^2)-a
ans = 0.000000000000000e+00,
```

that is no digits of accuracy for the naive evaluation. However, notice that

$$c - a = \sqrt{a^2 + b^2} - a = \frac{b^2}{\sqrt{a^2 + b^2} + a} = \frac{b^2}{c + a},$$

or from the OCTAVE command line

```
octave:13> b^2/(sqrt(a^2+b^2)+a)
ans = 2.23322144731059e-10
```

This answer is accurate to essentially all reported digits.

4. For (a) the MATLAB function is as follows.

```
% Function obtains roots of
%      a x^2 + b x + c = 0
% by direct use of quadratic formula.
% function [xm xp] = SimpleQuadratic(a,b,c)
% function [xm xp] = SimpleQuadratic(a,b,c)
xm = 0.5*(-b - sqrt(b^2 - 4*a*c))/a;
xp = 0.5*(-b + sqrt(b^2 - 4*a*c))/a;
```

For (i) the roots are  $x_{\mp} = -\frac{3}{4} \mp \frac{1}{4}\sqrt{9-8} = -1, -\frac{1}{2}$ . Using MATLAB, we find the following.

```
>> a = 2; b = 3; c = 1;
>> [xm xp] = SimpleQuadratic(a,b,c);
>> disp([xm xp])
-1.0000000000000000e+00    -5.0000000000000000e-01
```

For (ii) the roots are  $x_{\mp} = -\frac{3}{2} \mp \frac{1}{2}\sqrt{9-16} = -\frac{3}{2} \mp i\frac{\sqrt{7}}{2}$ . Using MATLAB, we find the following (some spaces removed from command window output).

```
>> a = 1; b = 3; c = 4;
>> [xm xp] = SimpleQuadratic(a,b,c);
>> disp([xm xp])
Column 1
-1.5000000000000000e+00 - 1.322875655532295e+00i
Column 2
-1.5000000000000000e+00 + 1.322875655532295e+00i
```

Note  $\text{sqrt}(7)/2 = 1.322875655532295e+00$ . So the simple quadratic formula yields accurate roots for (i) and (ii).

For (b) and case (iii) the exact roots are

$$x_{\mp} = -\frac{3}{2} \mp \frac{1}{2}\sqrt{9 - 4 \times 8^{-14}}.$$

We will use MATLAB's root function to define the "exact" roots and compare our answers with those obtained with this internal function. We use both the function SimpleQuadratic.m given above, as well as the following one.

```
% Function obtains roots to quadratic equation
%      a x^2 + b x + c = 0
% by quadratic formula, but with modification to improve
% accuracy. Here the modification is relevant for b > 0
% with sqrt(b^2 - 4*a*c) close in size to b. Namely,
%
% xm = -(b + sqrt(b^2 - 4ac))/(2a) ... usual formula.
% xp = c/(a*xm) ..... modified formula.
%
% Note modified formula for xp avoids subtraction of
% similar size numbers which is prone to accuracy loss.
% function [xm xp] = ModifiedQuadratic(a,b,c)
% function [xm xp] = ModifiedQuadratic(a,b,c)
xm = -0.5*(b + sqrt(b^2 - 4*a*c))/a;
xp = c/(a*xm);
```

To generate and compare the outputs, we have used the following script.

```
% Script: Set2Problem4b
% Makes table required by Problem 4b of Homework Set 2.

a = 1; b = 3; c = 8^(-14);
```

```

% Matlab computation of the roots.
tmp = roots([a b c]);
xm_matlab = tmp(1); xp_matlab = tmp(2);

% Roots obtained via naive use of quadratic formula.
[xm_simple xp_simple] = SimpleQuadratic(a,b,c);

% Roots obtained with modification of quadratic formula.
[xm_modified xp_modified] = ModifiedQuadratic(a,b,c);

FID = fopen('Set2Problem4b-Table.txt','w');
fprintf(FID,'          SimpleQuadratic          ModifiedQuadratic          roots([1 3 8^(-14)])\n');
fprintf(FID,'-----\n');
fprintf(FID,'x_m: %1.15e %1.15e %1.15e\n',xm_simple,xm_modified,xm_matlab);
fprintf(FID,'x_p: %1.15e %1.15e %1.15e\n',xp_simple,xp_modified,xp_matlab);
fclose(FID);

```

The script outputs the table

	SimpleQuadratic	ModifiedQuadratic	roots([1 3 8 <sup>-14</sup> ])
x_m:	-2.999999999999925e+00	-2.999999999999925e+00	-2.999999999999924e+00
x_p:	-7.571721027943568e-14	-7.579122514774592e-14	-7.579122514774593e-14

to the file `Set2Problem4b-Table.txt`. Our modified algorithm yields a second root in excellent agreement with MATLAB's answer, while the naive formula for  $x_+$  does not.

If student made a mistake, using instead of  $8^{14}$  value  $8e-14$ , please, do not subtract points. In this case right roots are:

```

>> roots([1 3 8e-14])

ans =

    -2.999999999999973e+00
    -2.666666666666690e-14

```