Math 202-02
Some Additional Problems on Material That Might be Covered on Midterm 1

Spring, 2016

1. A fixed point system uses six decimal digits to represent values:
   - There is no sign digit. So all values are nonnegative.
   - The rightmost two digits represent the fractional part of the input value (between 0 and 1).
   - The remaining four digits represent the part of the value that’s greater than 1.

Suppose rounding is used to approximate values that can’t be represented exactly. So for example 28.7348 would be stored as 28.73, and 28.7351 would be stored as 28.74.

(a) What is the smallest positive value that can be stored exactly in this system?
(b) What is the largest positive value that can be stored exactly in this system?

Overflow occurs when a value is too large to be represented.

(c) Are there two real numbers $x$ and $y$ with the following three properties:
   i. $x$ and $y$ can be represented exactly in the system,
   ii. $x + y$ doesn’t cause overflow, and
   iii. $x + y$ cannot be represented exactly in the system.
Solutions:

(a) 0.01
(b) 9999.99
(c) No. If \( x + y \) doesn’t overflow, then it will have two digits to the right of the decimal point, and no more than four digits to the left.

2. A floating point system also uses six decimal digits to represent values:

- One digit represents the sign (0 is positive, 1 is negative)
- The mantissa or significand uses three decimal digits.
- The exponent uses two decimal digits: one digit for the sign, and one digit for the absolute value of the exponent.

So, for example, 28.7 would be represented as \(+.287 +2\), since \(28.7 = 0.287 \times 10^2\). Input values that need more than 3 digits to represent the significand are rounded to 3 digits. For example, 28.748 would be represented as \(+.287 +2\), and 28.751 would be represented as \(+.288 +2\).

(a) What is the smallest positive value that can be stored exactly in this system?
(b) What is the largest positive value that can be stored exactly?

Arithmetic operations compute the exact result, and then round to the nearest value with a 3-digit significand.

(c) Show the result of the floating point operation \((+.287 +2) + (+.353 +4)\).
(d) Show the result of the floating point operation \((+.287 +2) \times (+.353 +4)\).
(e) Show the result of the floating point operation \((+.287 +5) \times (+.353 +6)\).

Solutions:

(a) Technically, this should be \(+0.001 -9 = 10^{-12}\). However, some specifications of floating point systems require that the significand
have a nonzero digit immediately to the right of the decimal point. For such systems, the smallest positive value would be $+0.100 -9 = 10^{-10}$.

(b) The largest positive value is $+0.999 +9 = 9.99 \times 10^8 \approx 10^9$.

(c) In exact arithmetic

\[+.287 +2 + .353 +4 = +.35587 +4\]

This is rounded to a 3-digit mantissa. So the result is $+.356 +4$.

(d) In exact arithmetic

\[+.287 +2 \times +.353 +4 = +.101311 +6\]

This is rounded to a 3-digit mantissa. So the result is $+.101 +6$.

(e) In exact arithmetic

\[+.287 +5 \times +.353 +6 = +.101311 +11\]

This is overflow: the result is too large to be stored as a floating point value, since the exponent is greater than $+9$, the largest exponent available.

3. If possible, carry out the indicated operations. If an operation isn’t possible, explain why not.

   (a) 

\[
\begin{bmatrix}
1 & 2 \\
3 & 4
\end{bmatrix}
\begin{bmatrix}
4 & 3 \\
2 & 1
\end{bmatrix}
\]

   (b) 

\[
\begin{bmatrix}
1 & 2 & -1 \\
3 & 4 & 1
\end{bmatrix}
\begin{bmatrix}
4 & 3 \\
2 & 1
\end{bmatrix}
\]

   (c) 

\[
\begin{bmatrix}
4 & 3 \\
2 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 2 & -1 \\
3 & 4 & 1
\end{bmatrix}
\]

Solutions:
(a) \[
\begin{bmatrix}
1 & 2 \\
3 & 4 \\
\end{bmatrix}
\begin{bmatrix}
4 & 3 \\
2 & 1 \\
\end{bmatrix}
= 
\begin{bmatrix}
8 & 5 \\
20 & 13 \\
\end{bmatrix}
\]

(b) This isn’t possible since the number of columns in the matrix on the left doesn’t equal the number of rows in the matrix on the right.

\[
\begin{bmatrix}
1 & 2 & -1 \\
3 & 4 & 1 \\
\end{bmatrix}
\begin{bmatrix}
4 & 3 \\
2 & 1 \\
\end{bmatrix}
\]

(c) \[
\begin{bmatrix}
4 & 3 \\
2 & 1 \\
\end{bmatrix}
\begin{bmatrix}
1 & 2 & -1 \\
3 & 4 & 1 \\
\end{bmatrix}
= 
\begin{bmatrix}
13 & 20 & -1 \\
5 & 8 & -1 \\
\end{bmatrix}
\]