Math 128a - Homework 1 - Due Feb 7 at the beginning of class

1) In class we saw an example showing that in decimal floating point arithmetic, the computed value of $x_{\text{mid}}=\frac{x_{\text{lower}}+x_{\text{upper}}}{2}$ is not necessarily between $x_{\text{lower}}$ and $x_{\text{upper}}$, which would be a problem for the logic in bisection (in 3 decimal digit arithmetic, try $x_{\text{lower}}=.997$ and $x_{\text{upper}}=.999$). We will show that this is impossible in IEEE arithmetic, which is binary. In other words, we will show that in IEEE arithmetic $x_{\text{mid}}=\text{fl}(\text{fl}(x_{\text{lower}}+x_{\text{upper}})/2)$ is in the interval $[x_{\text{lower}},x_{\text{upper}}]$, assuming overflow does not occur when adding $x_{\text{lower}}$ and $x_{\text{upper}}$. (Here $\text{fl}(a \odot b)$ means the floating point result of the operation $a \odot b$.

**Part 1.** Using the fact that IEEE arithmetic is correctly rounded, show that it is monotonic, that is if $a,b,c,d$ and $x$ are IEEE floating point numbers then

$$a \leq b \text{ and } c \leq d \implies \text{fl}(a + c) \leq \text{fl}(b + d)$$

$$a \leq b \text{ and } 0 < x \implies \text{fl}(a/x) \leq \text{fl}(b/x)$$

(Similar facts hold for subtraction and multiplication, but we will not need these here.)

**Answer:** The simplest way to describe how IEEE arithmetic computes $\text{fl}(a \odot c)$ (where $\odot$ is any binary arithmetic operation $+,-,\times$ or $\div$) can be described as follows (although it is not implemented this way!): Take the mathematically exact value of $a \odot c$ and round it to the nearest floating point number. If there is a tie (because $a \odot c$ is exactly half way between two floating point numbers) break the tie by rounding to the nearest floating point number whose bottom bit is zero.

We give two different proofs: first a direct case analysis, and second a proof by contradiction. First suppose $a \odot c = b \odot d$; then the rules above imply that $\text{fl}(a \odot c) = \text{fl}(b \odot d)$. So suppose $a \odot c < b \odot d$. There are two cases: either there is a floating point number $x$ somewhere in the range $a \odot c \leq x \leq b \odot d$ or there is not. If there is, then $x$ is closer to $a \odot c$ than any floating point number exceeding $x$, so $\text{fl}(a \odot c) \leq x$. Similarly $\text{fl}(b \odot d) \geq x$, so $\text{fl}(a \odot c) \leq x \leq \text{fl}(b \odot d)$ as desired. Now suppose there is no floating point number $x$ between $a \odot c$ and $b \odot d$. In other words $x_l < a \odot c < b \odot d < x_u$ where $x_l$ and $x_u$ are adjacent floating point numbers. Then the nearest floating point number to either $a \odot c$ or $b \odot d$ must be either $x_l$ or $x_u$. Now there are 3 possibilities: $a \odot c < (x_l + x_u)/2$, $a \odot c = (x_l + x_u)/2$ or $a \odot c > (x_l + x_u)/2$. In the first case $\text{fl}(a \odot c) = x_l$, which must be less than or equal to $\text{fl}(b \odot d)$ (which is either $x_l$ or $x_u$). In the second case $\text{fl}(b \odot d) = x_u$, which must be greater than or equal to $\text{fl}(a \odot b)$ (which is either $x_l$ or $x_u$). In the third case $\text{fl}(a \odot b) = x_u = \text{fl}(c \odot d)$.

Now we do a proof by contradiction. As before, if $a \odot c = b \odot d$ then $\text{fl}(a \odot c) = \text{fl}(b \odot d)$, so it suffices to consider the case $a \odot c < b \odot d$. Suppose for the sake of contradiction that $\text{fl}(a \odot c) > \text{fl}(b \odot d)$. Because we round to the nearest floating point number, there can’t be any floating point numbers between $a \odot c$ and $\text{fl}(a \odot c)$, so in particular $\text{fl}(b \odot d) < a \odot c$. Similarly $b \odot d < \text{fl}(a \odot c)$. Altogether then $\text{fl}(b \odot d) < a \odot c < b \odot d < \text{fl}(a \odot c)$. But this implies

$$|(b \odot d) - \text{fl}(b \odot d)| + |\text{fl}(a \odot c) - (a \odot c)| = (b \odot d) - \text{fl}(b \odot d) + \text{fl}(a \odot c) - (a \odot c) \geq (a \odot c) - \text{fl}(b \odot d) + \text{fl}(a \odot c) - (b \odot d) = |(a \odot c) - \text{fl}(b \odot d)| + |\text{fl}(a \odot c) - (b \odot d)|$$

1
But
\[ |(b \otimes d) - fl(b \otimes d)| \leq |fl(a \otimes c) - (b \otimes d)| \]
since \( fl(b \otimes d) \) is the closest floating point number to \( b \otimes d \), and
\[ |fl(a \otimes c) - (a \otimes c)| \leq |(a \otimes c) - fl(b \otimes d)| \]
since \( fl(a \otimes c) \) is the closest floating point number to \( a \otimes c \), so we get
\[
X \equiv |(b \otimes d) - fl(b \otimes d)| + |fl(a \otimes c) - (a \otimes c)| > |(a \otimes c) - fl(b \otimes d)| + |fl(a \otimes c) - (b \otimes d)|
\]
(from before)
\[
\geq |fl(a \otimes c) - (a \otimes c)| + |fl(b \otimes d) - (b \otimes d)|
\]
\[= X \]
or \( X > X \), a contradiction.

**Part 2.** Show that \( fl(2 \times x) = 2 \times x \) exactly, assuming overflow does not occur.

**Answer:** If \( x \) is an exact floating point number, so is \( 2 \times x \) (barring overflow), since multiplying by two just increases the exponent by one. So \( fl(2 \times x) = 2 \times x \).

**Part 3.** Show that \( 2 \times \text{xlower} \leq fl(\text{xlower} + \text{xupper}) \leq 2 \times \text{xupper} \).

**Answer:** If \( \text{xlower} \leq \text{xupper} \) are floating point numbers, we have \( \text{xlower} + \text{xlower} \leq \text{xlower} + \text{xupper} \), so by the first part of Part 1 \( fl(\text{xlower} + \text{xlower}) \leq fl(\text{xlower} + \text{xupper}) \), and by Part 2 we get \( 2 \times \text{xlower} \leq fl(\text{xlower} + \text{xupper}) \). Similarly, \( fl(\text{xlower} + \text{xupper}) \leq 2 \times \text{xupper} \).

**Part 4.** Conclude that \( \text{xlower} \leq fl(fl(\text{xlower} + \text{xupper})/2) \leq \text{xupper} \).

**Answer:** Dividing \( 2 \times \text{xlower} \leq fl(\text{xlower} + \text{xupper}) \leq 2 \times \text{xupper} \) by \( x = 2 \) and applying part 2 of Part 1 yields \( fl(2 \times \text{xlower}/2) \leq fl(fl(\text{xlower} + \text{xupper})/2) \leq fl(2 \times \text{xupper}/2) \). But \( (2 \times \text{xlower})/2 = \text{xlower} \) is an exact floating point number, so \( fl((2 \times \text{xlower})/2) = \text{xlower} \). Similarly \( fl((2 \times \text{xupper})/2) = \text{xupper} \).

**Part 5.** Where does this argument fail for correctly rounded decimal arithmetic?

**Answer:** This argument fails for decimal arithmetic because \( fl(2 \times x) \) does not have to equal \( 2 \times x \) exactly. (In decimal arithmetic, the formula \( \text{xmin} = (\text{xlower} + \text{xupper})/2 \) could be replaced by \( \text{xmin} = \max( \text{xlower}, \min( \text{xupper}, (\text{xupper} + \text{xlower})/2 ) ) \) to guarantee that \( \text{xlower} \leq \text{xmid} \leq \text{xupper} \).

**Part 6.** What happens if \( \text{xlower} \) and \( \text{xupper} \) are adjacent IEEE floating point numbers?

**Answer:** The argument that \( \text{xlower} \leq \text{xmin} \leq \text{xupper} \) is still true, so either \( \text{xmid} = \text{xlower} \) or \( \text{xmin} = \text{xupper} \).
2) Suppose $x$ is the exact answer to a problem, and $\hat{x}$ is our approximate answer. In class we defined the absolute error in $\hat{x}$ as $|x - \hat{x}|$ and the relative error in $\hat{x}$ as $|x - \hat{x}|/|x|$. In this problem we will explore some simple properties of these error measures.

Write the base $\beta$ expansion of $x > 0$ as $x = x_1x_2 \cdots x_n \cdot \beta^{e_x}$, and the base $\beta$ expansion of $y > 0$ as $y = y_1y_2 \cdots y_n \cdot \beta^{e_y}$. We will say that $x$ and $y$ agree to their leading $d$ base $\beta$ digits if $|x - y| < \frac{1}{2} \beta^{\max(e_x,e_y)-d}$. For example, .1230 and .1226 agree to 3 decimal digits, as do 1.00 and .996, or .1233 and .1237.

**Part 1.** Suppose you print out $\hat{x}$ as a base $\beta$ number. Show that if the relative error $|x - \hat{x}|/|x| < 1$, then the leading $[\log_\beta |x|/|x - \hat{x}|] - 1$ nonzero base $\beta$ digits of $\hat{x}$ are correct, i.e. $x$ and $\hat{x}$ agree to that many digits. ($|x|$ is the floor of $x$, the largest integer less than or equal to $x$.)

**Answer:** Let $k = [\log_\beta |x|/|x - \hat{x}|]$. The assumption that $|x - \hat{x}|/|x| < 1$ tells us that $k \geq 0$ and that $x$ and $\hat{x}$ have the same sign (if they have opposite signs then $|x - \hat{x}| = |x| + |\hat{x}|$). Since they have the same sign, w.l.o.g. we will assume they are both positive. We will show that $|x - \hat{x}| \leq \frac{1}{2} \beta^{e_x - (k - 1)}$, which means that $x$ and $\hat{x}$ agree to $k - 1$ digits:

$$
|x - \hat{x}| \leq \frac{1}{2} \beta^{e_x - (k - 1)}
$$

**Part 2.** Suppose you have solved your problem and gotten $\hat{x}$, and also a bound $e_{abs} \geq |x - \hat{x}|$ on the absolute error (perhaps using rounding error analysis as described in class). You would like a bound $e_{rel} \geq |x - \hat{x}|/|x|$ on the relative error. One obvious candidate is $e_{rel} = e_{abs}/|x|$, but of course you can’t compute this because you don’t know $x$ (otherwise we wouldn’t need an error bound!). So instead you try $e_{rel} = e_{abs}/|\hat{x}|$. Show that it is ok to use $e_{abs}/|\hat{x}|$ instead of $e_{abs}/|x|$ by showing that

$$
\frac{|x - \hat{x}|}{|x|} \leq \frac{|x - \hat{x}|}{|\hat{x}|} \leq \frac{|x - \hat{x}|}{|x|} \leq 1.2 e_{rel}
$$

Conclude that if $e_{rel} \leq .1$, then the actual relative error satisfies $.8 e_{rel} \leq |x - \hat{x}|/|x| \leq 1.2 e_{rel}$.

**Answer:** Multiplying numerator and denominator of both ends of the inequality we want to prove by $|\hat{x}|$ shows that the inequality is equivalent to:

$$
\frac{|x - \hat{x}|}{|\hat{x}| + |x - \hat{x}|} \leq \frac{|x - \hat{x}|}{|x - \hat{x}|} \leq \frac{|x - \hat{x}|}{|\hat{x} - x|}
$$
We need to assume that \(|x - \hat{x}| < 1\) so that the right-hand side is positive. Then we can take the reciprocal of everything and divide everything by \(|x - \hat{x}|\) to see that the statement we need to prove is equivalent to:

\[
|x| + |x - \hat{x}| \geq |x| \geq |\hat{x}| - |x - \hat{x}|
\]

which follows from the triangle inequality:

\[
|x| = |x - \hat{x} + \hat{x}| \leq |\hat{x}| + |x - \hat{x}| \quad \text{and:} \quad |\hat{x}| = |x - \hat{x} + x| \leq |x - \hat{x}| + |x| \quad |x| \geq |\hat{x}| - |x - \hat{x}|
\]

So, if \(e_{\text{rel}} \leq .1\) then \(1 - e_{\text{rel}} \geq 0.9\) so \(\frac{e_{\text{rel}}}{1 - e_{\text{rel}}} \leq \frac{e_{\text{rel}}}{0.9} \leq 1.2e_{\text{rel}},\) so that the actual relative error is less than or equal to \(1.2e_{\text{rel}}\). Similarly \(0.8e_{\text{rel}} \leq \frac{e_{\text{rel}}}{1.1} \leq \frac{|x - \hat{x}|}{|x|}.$$
3) Let \( 1 + r = \prod_{i=1}^{n} (1 + \delta_i) \), where \(|\delta_i| \leq \epsilon < 1\).

**Part 1.** Show that if \( n\epsilon < 1 \), then \(|r| \leq n\epsilon/(1 - n\epsilon)\).

**Answer:** Note that each term in the product is positive, so

\[
(1 - \epsilon)^n \leq 1 + r = \prod_{i=1}^{n} (1 + \delta_i) \leq (1 + \epsilon)^n
\]

and so

\[
(1 - \epsilon)^n - 1 \leq r \leq (1 + \epsilon)^n - 1
\]

We first show \((1 + \epsilon)^n - 1 \leq n\epsilon/(1 - n\epsilon)\) by induction, or equivalently \((1 + \epsilon)^n \leq 1/(1 - n\epsilon)\). We need to show the same expression is true with \(n + 1\) in place of \(n\). The base case \(n = 0\) is trivial. Multiply through by \(1 + \epsilon\) to get \((1 + \epsilon)^{n+1} \leq \frac{1 + \epsilon}{1 - n\epsilon}\). We need to show \(\frac{1 + \epsilon}{1 - n\epsilon} \leq \frac{1}{1 - (n+1)\epsilon}\), or \((1 + \epsilon)(1 - (n + 1)\epsilon) \leq (1 - n\epsilon), or \(1 - n\epsilon - (n + 1)\epsilon^2 \leq 1 - n\epsilon, which is true.

We take a different approach to showing \((1 - \epsilon)^n - 1 \geq -n\epsilon/(1 - n\epsilon), or equivalently \(1 - 2n\epsilon \leq (1 - \epsilon)^n(1 - n\epsilon)\). This is clearly true for \(1 > n\epsilon \geq .5\) and at \(\epsilon = 0\). To show it for \(\epsilon\) in between these values, we will show that the derivative of \(1 - 2n\epsilon\) with respect to \(\epsilon\) is always less than the derivative of \((1 - \epsilon)^n(1 - n\epsilon)\), so they start equal to one at \(\epsilon = 0\), and then \(1 - 2n\epsilon\) decreases faster as \(\epsilon\) increases from 0 to \(1/(2n)\). In other words we have to show

\[
-2n \leq -n(1 - \epsilon)^n - n(1 - \epsilon)^n - (1 - \epsilon)^n(2 - (n + 1)\epsilon) \text{ or } 2(1 - (1 - \epsilon)^{n-1}) \geq -(1 - \epsilon)^{n-1}(n + 1)\epsilon
\]

which is clearly true as desired.

**Part 2.** Show that if \(n\epsilon \leq .1\), then \(r \leq 1.2n\epsilon\).

**Answer:** If \(n\epsilon \leq .1\) then \(\frac{1}{1 - n\epsilon} \leq \frac{1}{.9} \leq 1.2\), so \(r \leq \frac{n\epsilon}{1 - n\epsilon} \leq 1.2n\epsilon\).

**Part 3.** In IEEE double precision, how big can \(n\) be and satisfy \(n\epsilon \leq .1\)?

**Answer:** In IEEE double precision, \(\epsilon = 2^{-53}\) so we solve \(2^{-53}n \leq .1\) to get \(n \leq (0.1)2^{53} \approx 9 \times 10^{14}\).

**Part 4.** If you compute \(p = \prod_{i=1}^{n} x_i\) in floating point arithmetic, and no over/underflow occurs, and \(n\epsilon \leq .1\), about how many leading decimal digits of the computed value of \(p\) are correct when using IEEE double precision arithmetic with \(n = 10\)? \(n = 100\) \(n = 1000\) \(n = 10000\)?

**Answer:** If we compute \(p = \prod_{i=1}^{n} x_i\), the relative error is \(r \leq 1.2n\epsilon\), so, by problem 1 we expect \(\log_{10}(1/r) - 1 \geq \log_{10}(1/(1.2n\epsilon)) - 1 = -\log_{10}(1.2\epsilon) - \log_{10}(n) - 1\) digits to be correct. In IEEE arithmetic, \(-\log_{10}(1.2\epsilon) > 15\), so we expect at least \(14 - \log_{10}(n)\) correct digits. Thus if \(n = 10\) we expect 13, if \(n = 100\) we expect 12, if \(n = 1000\) we expect 11 and if \(n = 10000\) we expect 10 correct digits.

5