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 xmid=(xlower+xupper)/2 is not necessarily between xlower and xupper, which would be a problem for the logic in bisection (in 3 decimal digit arithmetic, try xlower = .997 and xupper = .999). We will show that this is impossible in IEEE arithmetic, which is binary. In other words, we will show that in IEEE arithmetic xmid $= fl(fl(\text{xlower} + \text{xupper})/2)$ is in the interval [xlower,xupper], assuming overflow does not occur when adding xlower and xupper. (Here $fl(a \text{ op } b)$ means the floating point result of the operation a op 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

$$\begin{array}{lll} a \le b \text{ and } c \le d & \text{implies} & fl(a+c) \le fl(b+d) \\ a \le b \text{ and } 0 < x & \text{implies} & fl(a/x) \le fl(b/x) \end{array}$$

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

Answer: The simplest way to describe how IEEE arithmetic computes $fl(a \otimes c)$ (where $\otimes$ 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 $a \otimes c$ and round it to the nearest floating point number. If there is a tie (because $a \otimes c$ 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 $a \otimes c = b \otimes d$; then the rules above imply that $fl(a \otimes c) = fl(b \otimes d)$. So suppose $a \otimes c < b \otimes d$. There are two cases: either there is a floating point number x somewhere in the range $a \otimes c \le x \le b \otimes d$ or there is not. If there is, then x is closer to $a \otimes c$ than any floating point number exceeding x, so $fl(a \otimes c) \le x$. Simililary $fl(b \otimes d) \ge x$, so $fl(a \otimes c) \le x \le fl(b \otimes d)$ as desired. Now suppose there is no floating point number x between $a \otimes c$ and $b \otimes d$. In other words $x_l < a \otimes c < b \otimes d < x_u$ where x_l and x_u are adjacent floating point numbers. Then the nearest floating point number to either $a \otimes c$ or $b \otimes d$ must be either x_l or x_u. Now there are 3 possibilities: $a \otimes c < (x_l + x_u)/2$, $a \otimes c = (x_l + x_u)/2$ or $a \otimes c > (x_l + x_u)/2$. In the first case $fl(a \otimes c) = x_l$, which must be less than or equal to $fl(b \otimes d)$ (which is either x_l or x_u). In the second case $fl(b \otimes d) = x_u$, which must be greater than or equal to $fl(a \otimes b)$ (which is either x_l or x_u). In the third case $fl(a \otimes b) = x_u = fl(c \otimes d)$.

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

$$\begin{array}{rcl} |(b \otimes d) - fl(b \otimes d)| + |fl(a \otimes c) - (a \otimes c)| & = & (b \otimes d) - fl(b \otimes d) + fl(a \otimes c) - (a \otimes c) \\ & > & (a \otimes c) - fl(b \otimes d) + fl(a \otimes c) - (b \otimes d) \\ & = & |(a \otimes c) - fl(b \otimes d)| + |fl(a \otimes c) - (b \otimes d)| \end{array}$$

But

$$|(b \otimes d) - fl(b \otimes d)| \le |fl(a \otimes c) - (b \otimes d)|$$

since $fl(b \otimes d)$ is the closest floating point number to $b \otimes d$, and

$$|fl(a \otimes c) - (a \otimes c)| \le |(a \otimes c) - fl(b \otimes d)|$$

since $fl(a \otimes c)$ is the closest floating point number to $a \otimes c$, so we get

$$\begin{aligned} X \equiv |(b \otimes d) - fl(b \otimes d)| + |fl(a \otimes c) - (a \otimes c)| &> |(a \otimes c) - fl(b \otimes d)| + |fl(a \otimes c) - (b \otimes d)| \\ &\quad \text{(from before)} \\ &\ge |fl(a \otimes c) - (a \otimes c)| + |fl(b \otimes d) - (b \otimes d)| \\ &= X \end{aligned}$$

or $X > X$, a contradiction.

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

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

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

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

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

Answer: Dividing 2*xlower $\le$ fl(xlower+xupper) $\le$ 2*xupper by $x = 2$ and applyi ng part 2 of Part 1 yields fl(2*xlower/2) $\le$ fl(fl(xlower+xupper)/2) $\le$ fl(2*xupper/2). But (2*xlower)/2 = xlower is an exact floating point number, so fl((2*xlower)/2) = xlower. Similarly fl((2*xupper)/2) = xupper.

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

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

Part 6. What happens if xlower and xupper are adjacent IEEE floating point numbers?

Answer: The argument that xlower $\le$ xmin $\le$ xupper is still true, so either xmid = xlower or xmin = xupper.

2) Suppose x is the exact answer to a problem, and $\widehat{x}$ is our approximate answer. In class we defined the absolute error in $\widehat{x}$ as $|x-\widehat{x}|$ and the relative error in $\widehat{x}$ as $|x-\widehat{x}|/|x|$. In this problem we will explore some simple properties of these error measures.

Write the base β expansion of $x>0$ as $x = .x_1x_2\cdots x_n \cdot \beta^{e_x}$, and the base β 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 β 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 $\widehat{x}$ as a base β number. Show that if the relative error $|x-\widehat{x}|/|x| < 1$, then the leading $\lfloor \log_\beta \frac{|x|}{|x-\widehat{x}|} \rfloor - 1$ nonzero base β digits of $\widehat{x}$ are correct, i.e. x and $\widehat{x}$ agree to that many digits. ($\lfloor x \rfloor$ is the *floor of* x, the largest integer less than or equal to x.)

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

$$
\begin{aligned}
k = \lfloor \log_\beta \frac{|x|}{|x-\widehat{x}|} \rfloor & \text{ implies} \\
\beta^k \leq \frac{|x|}{|x-\widehat{x}|} & \text{ implies} \\
|x-\widehat{x}| &\leq \beta^{-k}|x| \\
&< \beta^{e_x-k} \\
&\leq \frac{1}{2}\beta^{e_x-k+1} \\
&= \frac{1}{2}\beta^{e_x-(k-1)}
\end{aligned}
$$

Part 2. Suppose you have solved your problem and gotten $\widehat{x}$, and also a bound $e_{abs} \geq |x-\widehat{x}|$ on the absolute error (perhaps using rounding error analysis as described in class). You would like a bound $e_{rel} \geq |x-\widehat{x}|/|x|$ on the relative e rror. 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 $e_{rel} = e_{abs}/|\widehat{x}|$. Show that it is ok to use $e_{abs}/|\widehat{x}|$ instead of $e_abs/|x|$ by showing that

$$\frac{\frac{|x-\widehat{x}|}{|\widehat{x}|}}{1+\frac{|x-\widehat{x}|}{|\widehat{x}|}} \leq \frac{|x-\widehat{x}|}{|x|} \leq \frac{\frac{|x-\widehat{x}|}{|\widehat{x}|}}{1-\frac{|x-\widehat{x}|}{|\widehat{x}|}}$$

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

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

$$\frac{|x-\widehat{x}|}{|\widehat{x}|+|x-\widehat{x}|} \leq \frac{|x-\widehat{x}|}{|x|} \leq \frac{|x-\widehat{x}|}{|\widehat{x}|-|x-\widehat{x}|}$$

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

$$|\widehat{x}| + |x - \widehat{x}| \geq |x| \geq |\widehat{x}| - |x - \widehat{x}|$$

which follows from the triangle inequality:

$$\begin{aligned} |x| &= |x - \widehat{x} + \widehat{x}| \leq |\widehat{x}| + |x - \widehat{x}| \quad \text{and:} \\ |\widehat{x}| &= |x - \widehat{x} + x| \leq |x - \widehat{x}| + |x| \\ |x| &\geq |\widehat{x}| - |x - \widehat{x}| \end{aligned}$$

So, if $e_{rel} \leq .1$ then $1 - e_{rel} \geq 0.9$ so $\frac{e_{rel}}{1-e_{rel}} \leq \frac{e_{rel}}{0.9} \leq 1.2e_{rel}$, so that the actual relative error is less than or equal to $1.2e_{rel}$. Similarly $.8e_{rel} \leq \frac{e_{rel}}{1.1} \leq \frac{|x-\widehat{x}|}{|x|}$.

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 ϵ in between these values, we will show that the derivative of $1 - 2n\epsilon$ with respect to ϵ 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 ϵ increases from 0 to $1/(2n)$. In other words we have to show

$$\begin{aligned} -2n &\leq -n(1-\epsilon)^{n-1}(1-n\epsilon) - n(1-\epsilon)^n \\ &= -n(1-\epsilon)^{n-1}(2-(n+1)\epsilon) \ \text{ or} \\ 2(1-(1-\epsilon)^{n-1}) &\geq -(1-\epsilon)^{n-1}(n+1)\epsilon \end{aligned}$$

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 $\frac{1}{1-n\epsilon} \leq \frac{1}{0.9} \leq 1.2$, 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 decial 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.