Mathematics 373 Workshop 1 Solutions

Bisection

Fall 2003

Problem 1 Consider $f(x) = x - \cos 2x$. Our goal is to solve $f(x) = 0$. A sketch of $y = f(x)$ could tell you how many solutions there are and where to look for them. However, you must make some choices before you can get a calculator to draw a **useful** graph. In many cases, the analysis to decide what should be graphed also tells you what the graph will show. In such cases, the graph serves as a check of your analysis.

A key tool in analyzing expressions is to see if one term **dominates** the expression for certain values of x. This uses the **lower bound form of the triangle inequality**

$$\big|A \pm B\big| \geq \big|\,|A| - |B|\,\big|$$

which is easily proved by considering the four cases given by the choices of signs of A and B (the use of $\pm$ to connect the terms is meant to emphasize that the sign is not important here). If you have a lower bound on $|A|$ on an interval and a upper bound on $|B|$ on **the same interval**, then $A + B$ will have the same sign as A, so it cannot be zero on that interval.

1a Statement What is an upper bound on $|\cos 2x|$ for all x? What does this say about possible locations of solutions of $f(x) = 0$?

1a Solution Since $\cos 2x$ is a value of the cosine function, we must have $-1 \leq \cos 2x \leq 1$, or $|\cos 2x| \leq 1$. Thus, if $f(x) = 0$, $x - 1 \leq f(x) \leq x + 1$. In particular, at any point where $f(x) = 0$, we have $x - 1 \leq 0$ and $0 \leq x + 1$. Writing these inequalities as a chain defining an interval gives

$$-1 \leq x \leq 1.$$

Note that negative values of x must be considered at this stage. All variables are assumed to represent **arbitrary real numbers** — there is no distinction between positive and negative numbers in algebra or calculus.

1b Statement Compute $f'(x)$. Use it to identify all maxima and minima of f. Are any in the interval containing solutions of $f(x) = 0$ found in (a)? An interval on which a function is everywhere increasing or everywhere decreasing can contain at most one point where the function is zero, and the function must have opposite signs at the endpoints of such an interval for the interval to be zero anywhere on the interval. How many solutions are there to $f(x) = 0$?

1b Solution Elementary calculus gives

$$f'(x) = 1 + 2\sin 2x.$$

The extreme values of $f(x)$ occur where $f'(x) = 0$, which can be expressed as $\sin 2x = -1/2$. We know that $\sin(\pi/6) = 1/2$, $\sin(-\theta) = -\sin\theta$ and $\sin(\theta + \pi) = -\sin\theta$. This gives

$$2x = \begin{cases} -\dfrac{\pi}{6} + 2n\pi \\ \dfrac{7\pi}{6} + 2n\pi \end{cases} \quad \text{or} \quad x = \begin{cases} \dfrac{(12n-1)\pi}{12} \\ \dfrac{(12n+7)\pi}{12} \end{cases}$$

for some integer n. The only one of these to fall in the interval $[-1, 1]$ is $-\pi/12 \approx -0.2618$. Consideration of the sign of $f'(x)$ shows that $f(x)$ decreases between $x = -1$ and $x = -\pi/12$ and increases between $x = -\pi/12$ and $x = 1$. In particular, to 6 decimal places, $f(-1) = -.583853$, $f(-\pi/12) = -1.127825$ and $f(1) = 1.416147$. This shows that $f(x)$ is negative between -1 and $-\pi/12$, so there is **only one** solution of $f(x) = 0$.

1c Statement The usual treatment of the bisection method refers to the interval on which the function changes sign as $[a, b]$ with $a < b$. This requires that you distinguish two cases depending on the sign of $f(a)$. Alternatively, the names a and b can be used for values for which you have verified that $f(a) < 0 < f(b)$. The bisection algorithm uses any such pair to produces another with $|a - b|$ as small as you like (as long as you can compute $f(x)$ accurately enough to perform the test) by the following method: let $m = (a + b)/2$; compute $T = f(m)$; if $T < 0$, replace a with m, else replace b with m; repeat. Since the $|a - b|$ is cut in half at each pass through the loop, you can determine in advance the number of steps to make $|a - b|$ small enough to meet your needs and just do that many passes through the loop. The algorithm works for any a and b, but it is easier (especially on a binary computer) to start with $a - b = \pm 1$.

Give initial values of a and b, and determine the number of steps to find an interval of length at most 10^{-6} on which f changes sign. Then do those steps and report the interval.

1c Solution We have found that $[-1, 1]$ has $f(-1) < 0 < f(1)$. This is a suitable initial interval and the first bisection step will give an interval of length 1 that we said was desirable. With $a = -1$ and $b = 1$, the length of the interval after n steps will be 2^{1-n}. To get an interval of length less than 10^{-6} requires $2^{1-n} < 10^{-6}$. Taking logarithms, this is equivalent to $(1 - n)\log 2 < -6 \log 10$ or $(n - 1) > 6 \log 10/\log 2 \approx 19.93$. Thus n must be an integer greater than 20.93. The smallest acceptable value of n (starting from an interval of length 2) is 21. A full table of these steps with entries rounded to 7 decimal places is shown in **Table 1**. The new endpoint on each line is **emphasized** to show the progress of the method.

The bisection algorithm uses only the sign of $f(m)$, but the size of this quantity helps to check the method. For one thing, the **mean value theorem** says that $f(m)$ is equal to the difference between m and the root of $f(x) = 0$ times the derivative of f somewhere between those values. When $[a, b]$ is a short interval, the derivative doesn't change much, so $f(m)$ is essentially proportional to the distance to the root. In particular, $|f(m)|$ decreases whenever the sign of $f(m)$ remains the same, but it may increase if the sign is different reflecting the fact that $f(x)$ is an increasing function on $[a, b]$ after the first step. Also $f(m)$ should be comparable to the desired accuracy when the algorithm stops. A calculation not showing these features is probably wrong, with the cause of the error depending on the implementation of the algorithm.

To perform a bisection step, you need to determine the sign of $f(m)$, but this may not be possible if $f(m)$ is close to zero. This is usually only a minor inconvenience since it signifies that you have found m close to the root. However, if $\left|f'(x)\right|$ is very small, you may need extended precision in the calculation of $f(m)$ to detect changes over an interval of the desired accuracy. At the other extreme, if $\left|f'(x)\right|$ is very large, it is likely that large values will appear in the calculation of $f(m)$ even when the result is small. Again, extended precision may be needed for accurate computation.

Problem 2

Consider the polynomial

$$g(x) = x^4 - 172x^3 + 11084x^2 - 317169x + 3400321.$$

How can we find where $g(x) = 0$?

a	b	m	$f(m)$
$-1.$	1.	0.	$-1.$
0.	1.	0.5000000	-0.0403023
0.5000000	1.	0.7500000	0.6792628
0.5000000	**0.7500000**	0.6250000	0.3096776
0.5000000	**0.6250000**	0.5625000	0.1313235
0.5000000	**0.5625000**	0.5312500	0.0445603
0.5000000	**0.5312500**	0.5156250	0.0018782
0.5000000	**0.5156250**	0.5078125	-0.0192764
0.5078125	0.5156250	0.5117187	-0.0087155
0.5117187	0.5156250	0.5136719	-0.0034221
0.5136719	0.5156250	0.5146485	-0.0007729
0.5146485	0.5156250	0.5151367	0.0005519
0.5146485	**0.5151367**	0.5148926	-0.0001106
0.5148926	0.5151367	0.5150147	0.0002207
0.5148926	**0.5150147**	0.5149537	0.0000551
0.5148926	**0.5149537**	0.5149231	-0.0000278
0.5149231	0.5149537	0.5149384	0.0000141
0.5149231	**0.5149384**	0.5149308	-0.0000063
0.5149308	0.5149384	0.5149346	0.0000035
0.5149308	**0.5149346**	0.5149327	-0.0000019
0.5149327	0.5149346	0.5149337	8×10^{-7}
0.5149327	**0.5149337**		

Table 1

Without some analysis, a graph is likely to only show that $g(x)$ is usually large and positive. The first step must be to isolate regions where $g(x)$ is close to zero?

First, note that all terms are positive is $x < 0$, so we can confine attention to positive values of x. Then, notice that for $x > 172$, the fourth degree term is larger than the third degree term, the second degree term is larger than the first degree term, and the constant term is positive, so $g(x) > 0$ for all such values.

A useful way to visualize the size of the term $a_k x^k$ is to plot the point $(k, \ln |a_k|)$. For each x, the line of slope $m = -\ln |x|$ through this point meets the y-axis at $\ln \left|a_k x^k\right|$. Since lines of a fixed slope only rarely contain more than one of these points, this shows that a single term will usually dominate the expression. The difference between two of these intercepts shows the **ratio** of the terms. **Figure 2** is a picture of this figure for the polynomial $g(x)$.

2a Statement When $\ln x$ is close to the negative of the slope of the line joining the points with $k = 1$ and $k = 3$ (the negative terms), the largest term is the positive term with $k = 2$, but each of the negative terms is almost as large, and the other terms are much smaller. The sum of these negative terms

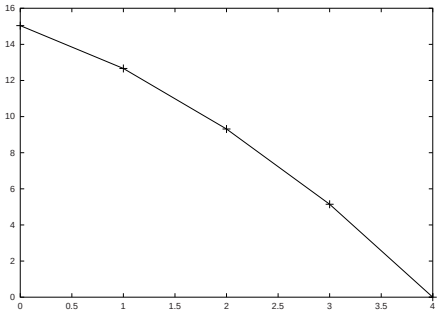

Figure 2

have the best chance to dominate near this value. Try a value of x suggested by this. Do you get a negative value of $g(x)$?

(This approach can be refined, but we will not pursue it further. Properties of the derivatives $g'(x)$ and $g''(x)$ used as an aid to graph sketching in Calculus usually give more precise results. They show that $g(x) = 0$ has two real solutions.)

2a Solution The slope of the line joining $(1, \ln 317169)$ and $(3, \ln 172)$ is

$$\frac{\ln 317169 - \ln 172}{1-3} = \frac{12.66719 - 5.147494}{-2} = -\frac{7.519696}{2} = -3.759848.$$

The exponential of the negative of this quantity is close to 43, and $g(43) = -33$. We **have** found a value of x where $g(x)$ is negative. Since $g(x)$ has even degree, it is positive if $|x|$ is large, so we have shown that there are at least two real solutions of $g(x) = 0$.

2b Statement Evaluating $g(x)$ in the given form needs to be done with care. The individual terms are quite large (about 2×10^7), yet they combine to give a small value. How much accuracy is needed in computation to give answers accurate to within 10^{-6}?

2b Solution One piece of information was omitted from the statement and needs to be found by investigation: we need to know the size of $g(x)$ when x is within 10^{-6} of a root. A study of $g'(x)$ for x between 39 and 47 shows values that get almost as large as 200 in absolute value. When the roots are found, it will turn out that g' is a little more than 100 in absolute value at each root. The factor of 2 between these values is only **one binary digit** of accuracy, which would not be significant in describing **decimal digits** of accuracy. What the size of the derivative means is that we must be able to compute $g(x)$ to within 10^{-8} in order to estimate the distance to the root to within 10^{-6}. If the individual terms of the polynomial are computed before being added together, this accuracy must be available in a number greater than 2×10^7, which means that at least 15 accurate digits must be present. To guard against underestimating our needs, at least 16 decimal digits will be needed for reliable computation. If less accuracy is used, the sign of $g(x)$ may not be correctly determined because of difficulty computing $g(x)$ and bisection will begin looking in an incorrect interval.

2c Statement Find the roots by bisection.

2c Solution Before starting the bisection process, $g(x)$ should be computed at some integers near 43. This leads to the discovery of a root between 39 and 40 and a second root between 46 and 47. These

are the values that should be used at the start of bisection. Then, 20 steps will give an interval of length less than 10^{-6}.

Initially, a bisection routine written in Maple was used. The default precision of Maple is 10 decimal digits. The values of $g(x)$ were shown to only **three** decimal places, reflecting the fact that numbers greater than 10^7 appeared in the computation of these quantities. The final intervals of [39.48135376, 39.48135472] and [46.58429525, 46.58429620] do **not** contain the values 39.48131475 and 46.58435067 found by the `fsolve` function. In particular, in step number 14 on the first interval, $g(39.48132325)$ was found to be $+0.014$ although it should be negative is `fsolve` is correct, and in step number 13 on the second interval, $g(46.58435057)$ was found to be 0, which was treated as positive in my bisection program, although `fsolve` believes that it should be negative.

Repeating the calculation after saying `Digits:=20` to get additional accuracy gave

$$g(39.481315612792968750) = -0.0000906180186$$
$$g(46.584350585937500000) = -0.000009447080$$

leading to intervals

$$[39.481314659118652343, 39.481315612792968750]$$
$$[46.584350585937500000, 46.584351539611816406]$$

(By contrast, the values reported by `fsolve` showed 10 more digits, but no changes in the 10 places obtained originally when the calculation was repeated with additional accuracy).