Math 128a - Midterm - Spring 2002

This exam is open book, open notes, open calculator (you shouldn't need one). The total score is 80 pts. The number of points approximately indicates the number of minutes you should spend on the problem.

1) (20 pts) If you invest $a_j > 0$ dollars each year into an account earning an interest rate of $x\%$, then after 10 years your account will hold $f(x) = \sum_{j=0}^{10} a_j(1 + x/100)^{10-j}$ dollars. You'd like to know the minimum interest rate $x \geq 0$ you need so that at the end of 10 years you have saved at least \$10,000. Assume that $\sum_{j=0}^{10} a_j < 10000$. First, show how to set up this problem as a polynomial zero-finding problem. Second, prove there is a unique nonnegative solution x. Third, show that Newton's method will converge to the desired x from any positive starting point.

Answer: The function $p(x) = f(x) - 10000$ is nonnegative if you have saved at least \$10,000, and negative if you have not. Since

- $p(0) = \sum_{j=0}^{10} a_j - 10000 < 0$ by assumption
- $p(x)$ goes to infinity as x increases, and
- $p(x)$ is a monotonically increasing function of x for $x \geq 0$ ($p'(x) = \sum_{j=0}^{9} \frac{10-j}{100} a_j(1 + x/100)^{9-j} > 0$ if $x > 0$, since all $a_j > 0$)

there is a unique $x_* > 0$ satisfying $p(x_*) = 0$, and this interest rate x_* is the answer. Identifying the polynomial zero-finding problem $p(x) = 0$ as the answer gets you 5 points, and showing that it has a unique nonnegative solution is worth another 5 points.

Since $p''(x) = \sum_{j=0}^{8} \frac{(10-j)(9-j)}{100^2} a_j(1 + x/100)^{8-j}) > 0$ when $x \geq 0$, we see that $p(x)$ is a convex function when $x \geq 0$. By the analysis in class, Newton can behave in one of two ways:

1. If the starting point $x_0 > x_*$, then the sequence of Newton iterates x_i will decrease monotonically and converge to x_*.
2. If the starting point $0 < x_0 < x_*$, then $x_1 > x_*$, and we are in case 1 above.

Either way, Newton will converge from any nonnegative starting point. Even though the function $p(x)$ is not convex for all values of x, it is convex in the region where Newton generates its iterates, so the analysis in class applies. This is worth 10 points. You lose 3 out of the 10 points if you just quote the result in class, since $p(x)$ is not convex for all x.

2) (30 pts) Consider the program below for computing the dot product d of two vectors x and y of floating point numbers:

$d = 0$
for $i = 1$ to n
    $d = d + x_i \cdot y_i$
end

Modify this program so that in addition to computing d, it explicitly computes a floating point number e that is guaranteed to bound the round off error in the computed value of d. In other words the true value of d must be *guaranteed* to lie in the interval $[d-e, d+e]$. (Note: $d+e$ and $d-e$ are not necessarily floating point numbers.) Assume that roundoff error can be described by $fl(a \otimes b) = (a \otimes b)(1+\delta)$ where $\otimes$ means addition or multiplication, and where $|\delta| \leq 2^{-53}$. You should be explicit about any assumptions you may make about the size of n for which your analysis applies. You will get significant partial credit for getting close the best possible e.

Answer: Let $d_i = \sum_{j=1}^{i} x_j \cdot y_j$ so d_n is the true dot product. Let $\hat{d}_i$ be the actual computed value of d_i. To get an expression for $\hat{d}_i$ modify the loop to read

$$\hat{d}_i = (\hat{d}_{i-1} + (x_i \cdot y_i)(1+\delta_{\cdot,i}))(1+\delta_{+,i}) \tag{1}$$

where $|\delta_{\cdot,i}| \leq \epsilon = 2^{-53}$ is the error in the i-th multiplication and $|\delta_{+,i}| \leq \epsilon$ is the error in the i-th addition. We want to compute the smallest upper bound e that we can on $|d_n - \hat{d}_n|$. You get 5 points for this fact. Note that each $\delta_{\cdot,i}$ and $\delta_{+,i}$ can independently take on any value in the range $[-\epsilon, \epsilon]$ (minus 2 points for not getting this).

At this point there are (at least) 2 approaches. This is the first approach, which most closely matches the approach in class for a similar problem, computing a sum $\sum_{i=1}^{n} x_i$. Expanding the first few d_i yields the general pattern

$$\begin{aligned}
\hat{d}_0 &= 0 \\
\hat{d}_1 &= x_1 \cdot y_1(1+\delta_{\cdot,1}) \\
&\quad \text{because no error is introduced by adding } \hat{d}_0 = 0 \\
\hat{d}_2 &= x_1 \cdot y_1(1+\delta_{\cdot,1})(1+\delta_{+,2}) \\
&\quad + x_2 \cdot y_2(1+\delta_{\cdot,2})(1+\delta_{+,2}) \\
\hat{d}_3 &= x_1 \cdot y_1(1+\delta_{\cdot,1})(1+\delta_{+,2})(1+\delta_{+,3}) \\
&\quad + x_2 \cdot y_2(1+\delta_{\cdot,2})(1+\delta_{+,2})(1+\delta_{+,3}) \\
&\quad + x_3 \cdot y_3(1+\delta_{\cdot,3})(1+\delta_{+,3}) \\
&\vdots \\
\hat{d}_n &= x_1 \cdot y_1(1+\delta_{\cdot,1})(1+\delta_{+,2}) \cdots (1+\delta_{+,n}) \\
&\quad + x_2 \cdot y_2(1+\delta_{\cdot,2})(1+\delta_{+,2}) \cdots (1+\delta_{+,n}) \\
&\quad \cdots \\
&\quad + x_i \cdot y_i(1+\delta_{\cdot,i})(1+\delta_{+,i}) \cdot (1+\delta_{+,n}) \\
&\quad \cdots
\end{aligned}$$

$$
\begin{aligned}
& +x_n \cdot y_n(1+\delta_{\cdot,n})(1+\delta_{+,n}) \\
= & \sum_{i=1}^{n} x_i y_i E_i
\end{aligned}
$$

where each E_i is the product of $1+\delta$ factors. You get 5 points for displaying this pattern.

Simplifying we get that $\hat{d}_n = \sum_{i=1}^{n}(x_i y_i)E_i = d_n + \sum_{i=1}^{n}(x_i y_i)(E_i - 1)$ so

$$
|d_n - \hat{d}_n| \le |\sum_{i=1}^{n}(x_i y_i)(E_i - 1)| \le \sum_{i=1}^{n}|x_i y_i(E_i - 1)| = \sum_{i=1}^{n}|x_i y_i| \cdot |E_i - 1|
$$

You get 5 points for this inequality. For the importance of the absolute values, which many people missed, see below. To proceed, we need to bound $|E_i - 1|$.

Note that

$$
\begin{aligned}
E_1 &= (1+\delta_{\cdot,1})\prod_{j=2}^{n}(1+\delta_{+,j}) \\
E_i &= (1+\delta_{\cdot,i})\prod_{j=i}^{n}(1+\delta_{+,j}) \text{ if } i \ge 2
\end{aligned}
$$

Thus when $i \ge 2$, E_i is the product of at most $n+2-i$ terms of the form $1+\delta$, and E_1 is the produce of n terms of the form $1+\delta$, where $|\delta| \le \epsilon$. More briefly, we can write that E_i is the product of $\min(n, n-i+2)$ terms of the form $1+\delta$. By a result in class, we can therefore write

$$
|E_i - 1| \le \frac{\min(n, n+2-i)\epsilon}{1 - \min(n, n+2-i)\epsilon} \tag{2}
$$

as long as $n\epsilon < 1$, or $n < \epsilon^{-1} = 2^{53}$, which will certainly be satisfied in practice. You get 5 more points for inequality (2) (minus 1 point for missing the bound on n). You get full credit by replacing $\frac{\min(n,n+2-i)\epsilon}{1-\min(n,n+2-i)\epsilon}$ by the common upper bound $\frac{n\epsilon}{1-n\epsilon}$ or even the (slightly too small) approximation $n\epsilon$.

Thus we can use

$$
|d_n - \hat{d}_n| \le e = \sum_{i=1}^{n}|x_i y_i|\frac{\min(n, n+2-i)\epsilon}{1 - \min(n, n+2-i)\epsilon}
$$

You lose 5 points if you miss the absolute values. So the simplest program is thus

```
d = 0
e = 0
for i = 1 to n
      d = d + x_i · y_i
      e = e + abs(x_i · y_i) min(n,n+2−i)ε / (1−min(n,n+2−i)ε)
end
```

or even

$d = 0$
$e = 0$
for $i = 1$ to n
    $d = d + x_i \cdot y_i$
    $e = e + \mathrm{abs}(x_i \cdot y_i)$
end
$e = e \cdot \frac{n\epsilon}{1-n\epsilon}$

This program (or a variant with a different reasonable bound on $|E_i - 1|$) gets 4 points.

However, to get a *guaranteed* upper bound on $|d_n - \hat{d}_n|$ we would need to take the roundoff error in the computation of e into account as well. One can use the same approach as above to get a bound on the error in e. Since e is a sum of positive numbers it turns out that the computed value of e cannot be much smaller than the true value of e. Without doing the details, let's just say that we should multiply the computed value of e above by a final factor of $(1-(n+2)\epsilon)(1+4\epsilon)/(1-(n+2)2\epsilon))$, which is a tiny bit larger than 1, to get a guaranteed upper bound as long as $(n+2)2\epsilon < 1$, or $n < 2^{52} - 2$. If you take all this into account, you get 1 more point.

Here is a second, simpler approach. From (1) we get that

$$\hat{d}_i = \hat{d}_{i-1} + x_i \cdot y_i + x_i \cdot y_i(\delta_{\cdot,i} + \delta_{+,i} + \delta_{\cdot,i} \cdot \delta_{+,i}) + \delta_{+,i}\hat{d}_{i-1} = \hat{d}_{i-1} + x_i \cdot y_i + r_i$$

where $r_i = x_i \cdot y_i(\delta_{\cdot,i} + \delta_{+,i} + \delta_{\cdot,i} \cdot \delta_{+,i}) + \delta_{+,i}\hat{d}_{i-1}$. Thus

$$|r_i| \le |x_i \cdot y_i|(\delta_{\cdot,i} + \delta_{+,i} + \delta_{\cdot,i} \cdot \delta_{+,i}) + \delta_{+,i}|\hat{d}_{i-1}| \le (2\epsilon + \epsilon^2)|x_i \cdot y_i| + \epsilon|\hat{d}_{i-1}| \quad .$$

Then we write a recurrence for the error $|d_i - \hat{d}_i|$ as follows:

$$\begin{aligned} |d_i - \hat{d}_i| &= |(d_{i-1} + x_i \cdot y_i) - (\hat{d}_{i-1} + x_i \cdot y_i + r_i)| \\ &= |d_{i-1} - \hat{d}_{i-1} - r_i| \\ &\le |d_{i-1} - \hat{d}_{i-1}| + |r_i| \\ &\le |d_{i-1} - \hat{d}_{i-1}| + (2\epsilon + \epsilon^2)|x_i \cdot y_i| + \epsilon|\hat{d}_{i-1}| \end{aligned}$$

You get 15 points for this recurrence (minus 5 for missing the absolute values). Thus an error bound (5 more points) is given by $e = \sum_{i=1}^{n}(2\epsilon+\epsilon^2)|x_i \cdot y_i| + \epsilon|\hat{d}_{i-1}|$. The corresponding program (4 points) is

$d = 0$
$e = 0$
for $i = 1$ to n
    $e = e + \mathrm{abs}(x_i \cdot y_i) \cdot (2\epsilon + \epsilon^2) + \epsilon \cdot \mathrm{abs}(d)$
    $d = d + x_i \cdot y_i$
end

As before, a complete solution would take the (tiny) rounding error in the computation of e into account as well (1 point).

Many people forgot the absolute values when computing the upper bound, resulting in a bound something like $n2^{-53}|\sum_{i=1}^{n} x_i y_i|$ instead of $n2^{-53}\sum_{i=1}^{n}|x_i y_i|$, i.e. that the relative error in the computed d is always small. This is not true: consider the dot product $1 \cdot 2^{53} + 1 \cdot 1 + (-1) \cdot 2^{53}$. The true value is 1, but the computed value is 0, which is a large relative error. This error (=1) is not bounded by $n2^{-53}|\sum_{i=1}^{n} x_i y_i| = 0$ but is bounded by $n2^{-53}\sum_{i=1}^{n}|x_i y_i| \approx 6$.

Parts 2 and 3 were stated erroneously in the exam as given: the upper bounds you were asked to prove should have been larger by a factor of 2. The statements and answers given below are correct. I apologize for this error. We took it into account in the grading, where in fact it made little difference in people's answers.

3) (30 pts) Let $x_i = \cos\frac{(2i+1)\pi}{2n}$, $0 \le i \le n-1$, be the zeros of the Chebyshev polynomial $T_n(x)$. Let $p(x)$ be the polynomial that interpolates the function $f(x) = 2\sin(x) - \sqrt{3}\cos(2x)$ at $x_0, ..., x_{n-1}$.

1. (5 pts) Write down an expression for the interpolation error $f(x) - p(x)$, in terms of the n-th derivative of $f(x)$ and $T_n(x)$.

   Answer: From the book or class notes, $f(x) - p(x) = \frac{f^{(n)}(z)}{n!}2^{1-n}T_n(x)$. Stating the theorem gets 3 points, and applying it correctly gets 2 points.

2. (10 pts) Show that the interpolation error $|f(x) - p(x)| \le \frac{2^{2-n}+2\sqrt{3}}{n!}$ as long as $|x| \le 1$.

   Answer: $f^{(n)}(z) = 2f_1(z) + \sqrt{3}\cdot 2^n f_2(z)$, where $|f_1(z)| = |\pm\cos(z)| \le 1$ or $|f_1(z)| = |\pm\sin(z)| \le 1$ and where $|f_2(z)| = |\pm\cos(2z)| \le 1$ or $|f_2(z)| = |\pm\sin(2z)| \le 1$, so $|f^{(n)}(z)| \le 2 + \sqrt{3}\cdot 2^n$. Furthermore, $|T_n(x)| \le 1$ when $|x| \le 1$. Plugging these bounds into the answer to part 1 yields the result.

   Bounding $|f^{(n)}(z)|$ gets 4 points, bounding $|T_n(x)|$ gets 4 points, and putting it all together gets 2 points.

3. (15 pts) Show that the interpolation error $|f'(x_i) - p'(x_i)| \le \frac{4(2^{1-n}+\sqrt{3})}{(n-1)!}$ as long as $|x_i| \le \frac{1}{\sqrt{2}}$.

   Answer: From part 1,

$$f'(x) - p'(x) = \frac{f^{(n+1}(z(x))\cdot z'(x)}{n!}\cdot 2^{1-n}T_n(x) + \frac{f^{(n)}(z(x))}{n!}\cdot 2^{1-n}T_n'(x)$$

   Plugging in $x = x_i$ yields $f'(x_i) - p'(x_i) = \frac{f^{(n)}(z(x_i))}{n!}2^{1-n}T_n'(x_i)$. The factor $f^{(n)}(z(x))$ is bounded by $2 + \sqrt{3}\cdot 2^n$ as before. We differentiate

$$T_n'(x) = \frac{d}{dx}\cos(n\arccos x) = -\sin(n\arccos x)\cdot\frac{n}{-\sqrt{1-x^2}}$$

   so that if $|x| \le \frac{1}{\sqrt{2}}$ then $|T_n'(x)| \le 1\cdot\frac{n}{1/2} = 2n$. Altogether then

$$|f'(x_i) - p'(x_i)| \le \frac{2n(2^{1-n}+\sqrt{3})\cdot 2}{n!} = \frac{4(2^{1-n}+\sqrt{3})}{(n-1)!}$$

   as desired.

   Differentiating the bound gets 5 points, evaluating and simplifying it for x_i gets 2 points, bounding $|T_n'(x_i)|$ gets 5 points (2 for using $T_n(x) = \cos(n\arccos x)$, 2 for differentiating this formula, and 1 for bounding it), bounding $|f^{(n)}(x)|$ gets 1 points, and putting it all together gets 2 points.