Math 128A, Summer 2019
Lecture 4, Thursday 6/27/2019

1 Reviewing Homework 1 Problems

Homework problem 1: $f(x) := \sum_{j=1}^{n} f_j(xy_j)^k$. This problem is very basic in pieces, but it tells us that when we want to do a fourier transform, the fact that we can move the x out to the left means that the variables are separated and x, y live in different worlds. We can do the fast fourier transform on the condition that the x, y can be separated (separation of variables, aka tensor product).

Homework problem 2:

$$s_n := \sum_{k=1}^{n} \frac{1}{k^2} \to \frac{\pi^2}{6} = (((1 + \frac{1}{4}) + \frac{1}{9}) + \frac{1}{16}) + \cdots$$

(L to R) Other way would be right to left (R to L):

$$\left(\left(\frac{1}{n^2} + \frac{1}{(n-1)^2}\right) + \frac{1}{(n-2)^2}\right) + \cdots + 1$$

Neither ways is best, but R to L intuitively should be better beacuse we're adding small terms before larger, so we don't lose as much accuracy.

Method 1 :
Quote a known result.

$$|fl(S_n) - S_n| \leq \frac{\varepsilon}{2} \cdot (|s_2| + |s_3| + \cdots + |s_n|) \tag{1}$$
$$\leq (n-1)\varepsilon \tag{2}$$

This bound may not be optimal, but it does pay attention to order of summation.

We'll reason that the value in the parenthesis is ≤ 2. The crudest thing to say is above in the second line.

Remark: The above is flawed because

$$fl\left(\frac{1}{k^2}\right) = \frac{1}{k^2}(1 + 2\delta_k)$$

But to fix this, we can bridge the LHS and RHS by using the Triangle Inequality.
(1) Sum exact terms in floating point arithmetic into $\hat{s}_n$, to equal $fl\left(\sum_{k=1}^{n} \frac{1}{k^2}\right)$.
(2) Sum the wrong terms exactly:

$$\left|\sum_{k=1}^{n} \frac{1}{k^2}(1 + 2\delta_k) - S_n\right| = \left|\sum_{k=1}^{n} \frac{1}{k^2} \underbrace{2\delta k}_{\varepsilon}\right| \leq 2\varepsilon$$

The floating point of k^2,

$$fl(k^2) = k^2(1+\delta_k)$$

so,

$$\frac{1}{k^2} = \frac{1}{k^2(1+\delta_k)}(1+\delta_k')$$

In the worst case, the numerator and denominator have opposite signs and accumulate error.

$$\begin{aligned}
&\left| fl\left(\sum_{k=1}^{n} \frac{1}{k^2}(1+2\delta_k)\right) \underbrace{-\hat{S}_n + \hat{S}_n}_{\text{bridge}} - Sn \right| \\
&= (n-1)\varepsilon + \frac{\varepsilon}{2}\left(\left|\sum_{k=1}^{2} \frac{1}{k^2}(1+2\delta_k)\right| + \left|\sum_{k=1}^{3} \frac{1}{k^2}(1+2\delta_k)\right| + \cdots + \left|\sum_{k=1}^{n} \frac{1}{k^2}(1+2\delta_k)\right|\right) \\
&\leq 2(n-1)\varepsilon
\end{aligned}$$

Where each of these summation is ≤ 2.
We add as many bridges in to get what we want (things equal to 0). Or multiply something equiv to 1.

Homework 1 Problem 1b-c (b) Find a polynomial $P(x)$ with complex coefficients such that

$$|P(x) - e^{ix}| \leq \epsilon$$

on the interval $|x| \leq 1$.
(c) Design an algorithm for approximating

$$g(x) := \sum_{j=1}^{n} g_j e^{ixy_j}$$

at n points x_i in $O(n)$ operations, with absolute error bounded by

$$\epsilon \sum_{j=1}^{n} |g_j|.$$

KNOWN FACTS:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}, x \in \mathbb{R}$$
$$e^{ix} = \sum_{n=0}^{\infty} \frac{(ix)^n}{n!}, x \in \mathbb{R}$$
$$e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}, z \in \mathbb{C}, z := x + yi$$
$$e^{\alpha+\beta i} = e^{\alpha}(\cos\beta + i\sin\beta)$$

In our problem 1c, recall that we have even and odd parts, where the real parts are even, odd parts are imaginary. To see this, consider $i, i^2 = -1, i^3 = -i, i^4 = 0, i^5 = i, \ldots$. The absolute value (complex number) is bounded to a real number (distance).
So taking the absolute value difference of two complex numbers, we have:

$$\left| e^{ix} - \sum_{n=0}^{p-1} \frac{(ix)^n}{n!} \right| = \left| \sum_{n=p}^{\infty} \frac{(ix)^n}{n!} \right|$$
$$= \sum_{n=part}^{\infty} \left| \frac{(ix)^n}{n!} \right| = \sum_{n=p}^{\infty} \frac{|i|^n x^n}{n!}$$
$$\leq \sum_{n=p}^{\infty} \left(\frac{e}{n}\right)^n |x|^n \leq \sum_{n=p}^{\infty} \left(\frac{e|x|}{n}\right)^n$$

n is going from $p \to \infty$, so we can fix this by replacing n with

$$\leq \sum_{n=p}^{\infty} \left(\frac{e|x|}{p}\right)^n \quad \text{(which is a geometric series)}$$
$$\leq 2 \cdot 2^{-p} \quad (\text{if } |x| \leq 1, p \geq 6)$$

New Example: Like our homework 1.3, consider:

$$a \leq fl\left(\frac{a+b}{2}\right) \leq b, \text{ if } a \leq b$$
$$0 < a \leq fl\left(\sqrt{ab}\right) \leq b$$

The above is the arithmetic mean, whereas the bottom is the geometric mean. The implications of this are applied to within an algorithm, where our numbers are floating point numbers.
Consider:

$$\begin{aligned} a &\leq a \\ a &\leq b \\ 2a &\leq a+b \quad \text{(add ineqs)} \\ fl(2a) &\leq fl(a+b) \quad \text{(what happens when we divide by 2?)} \\ \frac{fl(2a)}{2} &\leq \frac{fl(a+b)}{2} \quad \text{(fine to divide by 2)} \end{aligned}$$

Note rounding is monotone, which helps us know what happens when we perform operations and round (floating point arithmetic).
Consider:

$$\begin{aligned} a^2 &\leq ab \\ \sqrt{a^2} &\leq \sqrt{ab} \\ fl\left(\sqrt{ab}\right) &= \sqrt{ab(1+\delta)} \\ &= \sqrt{ab}(1+\underbrace{\frac{\delta_1}{2}}_{\leq \frac{\varepsilon}{4}}) + O(\varepsilon^2) \end{aligned}$$

Aside:

$$\sqrt{1+\delta} = 1 + \frac{x}{2} - \frac{x^2}{4} + \cdots \quad \text{(Taylor expansion)}$$

Also consider:

$$|\sqrt{1+x}-1| = \left|\frac{(\sqrt{1+x})(\sqrt{1+x}+1)}{\sqrt{1+x}+1}\right| = \left|\frac{x}{1+\sqrt{1+x}}\right| \leq O(|x|)$$

Break time:

2 Bisection

Now, the goal is to solve the following, using bisection (the simplest possible algorithm). Next week we'll talk about variations of Newton and FixedPoint Iteration (which does not guarantee an existing solution), but are higher-dimension.
For Bisection, we apply the intermediate value theorem (IVT). If $f(a) \leq 0 \leq f(b)$, or the other way around $f(b) \leq 0 \leq f(a)$, then this implies

$$\exists_{x \in [a,b]} f(x) = 0,$$

assuming $f \in C[a,b]$; that is, assuming f is continuous over $[a,b] \in \mathbb{R}$.
Our plan:
Step 0: Bracket the root; find a, b with $f(a)f(b) < 0$ (opposite signs).
$f(x) := \ln x$, $\frac{1}{e} < x < e$, $\ln \in (-1,1), x = 1$.
Then we refine the bracket:
Take $m = \text{middle}(a,b)$, with one of the following definitions:

$$\begin{aligned} \text{mid}(a,b) &= \frac{a+b}{2} \quad \text{(tradition but stupid)} \\ &= \sqrt{ab} \quad \text{(better)} \\ &= \max\{a, \min\{b, \sqrt{ab}\}\} \quad \text{(if } a,b > 0\text{)} \end{aligned}$$

2.1 Pseudocode / Algorithm

If $f(m) = 0$, then stop.
If sign $f(m) =$ sign $f(a)$, replace $a := m$, then $f(a) := f(m)$. Otherwise, $b := m$.
Recursion.

Stopping Criterion:
What if we're given a user-specified tolerance `tol`? And let $|b-a| \leq$ `tol`?
How about stopping when $|b-a| \leq \varepsilon \min\{|b|, |a|\}$?
How about stopping when the floating point numbers are equal, $m = a$ or $m = b$? This isn't too bad.
Some people also stop when $|f(m)| < \varepsilon \cdot \min\{|f(a)|, |f(b)|\}$.

Aside: anything like user-specified or provided by the user is never a good idea (it won't sell if it needs the user to perform the hard work).

How to bracket $[a,b]$?
- Randomness if ignorant
- Structural if knowledgeable about f

2.2 Bisection in Real (Infinite) Arithmetic

$$(b-a) \to \frac{1}{2}(b-a) \to \frac{1}{4}(b-a) \to \cdots$$

2.3 Bisection in Floating Point Arithmetic

Stopping criterion of $m = a \cup m = b$. In computer arithmetic, we are computing some $64 - bit$ number. Bisection should produce 1-bit per step, so ≤ 64 steps.
Theoretically, trying to find some 64-bit number, asking 64 Yes/No questions should find us what we want.
In our homework, we are converging down to 0, and there are actually a lot of floating point numbers at 0. Turns out it takes like 1000 steps to underflow to get this stopping criterion (equality), because $2^{-1074} =$ smallest nonzero floating point number.

Example: Let $f(x) = x = 0$, with $[a,b] := [-1, 2]$, and $m := \frac{a+b}{2}$.
So we have:
$m = \frac{-1+2}{2} = \frac{1}{2}$, then:

1: $[a, b] = [-1, \frac{1}{2}]; m = \frac{-1+1/2}{2} = \frac{-1}{4}$
2: $[a, b] = [-\frac{1}{4}, \frac{1}{2}]$
$\vdots$
around $k \approx 1074$: we get $[-2^{-k}, 2^{-k+1}]$, until 2^{-k} underflows.
Aside: our homework problem is designed to convince us that using the Arithmetic mean can take about 20 times as long as using the geometric mean.

2.4 Geometric mean

The good part is: $[a, b] = [2^{-1000}, 1]$ with $f(x) := x - 2^{-900}$.
$m := \sqrt{ab} = \sqrt{2^{-1000}} = 2^{-500}$
Now we're 500 orders of magnitude closer to the solution, and better yet, we have 1 bit of the solution. What if $a = 0$? What if $a < 0 < b$?
So we need a little more detail in the logic. Our first observation is

$$\begin{aligned}
&0 < a \le b \quad \text{then let } m := \sqrt{ab} \\
&a \le b < 0 \quad \text{then let } m := -\sqrt{ab} \\
&0 = a \le b \quad \text{then let } m := \texttt{realmin} = 2^{-1074} \\
&a \le b = 0 \quad \text{then let } m := -\texttt{realmin} = -2^{-1074} \\
&a < 0 < b \quad \text{then let } m := 0
\end{aligned}$$

Remark: In general, we'd rather take a few extra steps via GM than take potentially thousands more steps via AM. (Geometric mean, Arithmetic mean).