Math 128A, Summer 2019
Lecture 5, Monday 7/1/2019

Announcements: Homework due in class Wednesday.

1 Homework 1 Review (due Wednesday in Class)

1.1 Problem 1

$(Fg)_j = \sum_{k=1}^{n} \underbrace{e^{ix_j y_k}}_{F_{ij}} g_k$. We denote F_{jk} for fourier, although this acts like Laplace transform, for things like decay rate. (edit: nevermind, we inserted the i, so it's legal to call this F for fourier transform.)

We need $O(n^2)$ cost to get the matrix elements of F_{jk} because each x_j and y_k has to talk to each other. To see this, recall that matrix multiplication is like dotting two vectors; n multiplications and n additions, per row. So in total for the matrix, we need $O(n^2)$.

The essence of the problem is that we can do this faster if we exploit the fact that we can bring out a term. Like in (a) where we let

$$\sum_{j=1}^{n} (x_i y_j)^k f_j = x_i^k.$$

So we have:

$$\begin{bmatrix} e^{ix_1 y_i} & \dots & e^{ix_1 y_n} \\ \vdots & \ddots & \vdots \\ e^{ix_n y_1} & \dots & e^{ix_n y_n} \end{bmatrix}$$

This problem is about a "miracle" that rank $B \leq r$. Suppose our matrix C is $r \times n$, with $r = 15 << n$; thus $O(15n) = O(rn) = O(n) << O(n^2)$. 15 (or the actual number) is a lot, but it is surely less than n^2.

Hence r depends on our accuracy ε and not on n. So for the matrices $F, B := CD, E =$ error, we have:

$$\mathcal{M}(F) = \underbrace{\mathcal{M}(CD)}_{\mathcal{M}(B)} + \mathcal{M}(E)$$

$$\text{rank } [(n \times r)(r \times n)] = \min\{(n \times r), (r \times n)\}$$

So to ensure the error bound of each element, in this problem we use:

$$\begin{aligned} e^{ix_j y_k} &= \sum_{p=0}^{r-1} \frac{(ix_j y_k)^p}{p!} + \overbrace{O\left(\frac{1}{r!}\right)}^{E} \\ &= \sum_{p=0}^{r-1} \left(\frac{ix_j^p}{\sqrt{p!}}\right)\left[\frac{y_k^p}{\sqrt{p!}}\right] \quad \text{(or equivalently can split by } 1 \cdot p!) \\ &= \sum_{p=0}^{r-1} (C_{jp}) \cdot [D_{pk}] \end{aligned}$$

The miracle is that we usually, in a fourier transform, have:

$$U(x,t) = \sum_{j=0}^{\infty} f_j(t) g_j(x)$$

where we need infinite sums because one of f_j or g_j is not enough.

1.2 Problem 2

(in class) The maximum accuracy achievable (in computing $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$):

$$fl\left(\sum_{k=1}^{n} \frac{1}{k^2}\right) = s_n\left(1 + 2n\delta\right), \qquad |\delta| \leq \varepsilon$$

There are two competing errors: truncation error on the left, and the floating point error in summing to infinity. Our question is to find the minimum error in between these two.
So the maximum accuracy at the bottom of this curve will be $O(n\varepsilon) + O\left(\frac{1}{n}\right)$, where $\sqrt{1/\varepsilon} = 2^{26}$. The best thing would be to take 2^{26}, and our value of accuracy will only be 2^{-26}.

2 Fixed Point Iteration

- Bisection (bicycle; low cost, doesn't smell)
- Fixed point iteration (car)
- Newton's Method (maseratti; good, but requires maintenance or breaks down)

Step 0: Convert to fixed point form. That is, convert $f(x) = 0 \to x = g(x)$. For example,

$$\begin{aligned} f(x) = 0 &\to x = x - f(x) = g(x) \\ &\to x = x + \alpha f(x) \\ &\to x = x + h(x) f(x) \end{aligned}$$

For choosing, we question if we want to introduce extraneous solutions and how fast we reach convergence

2.1 Example of Fixed Point Iteration in Action (and Error Bounding via MVT)

$$\underbrace{xe^x - 1 = 0}_{f(x)}, x = w(1) \quad \text{(Lambert's something)}$$

$$xe^x = 1$$
$$x = e^{-x} = g_A(x)$$

Alternatively, we could use $e^x = \frac{1}{x}$ or $x = \ln\left(\frac{1}{x}\right)$.
Step 1: Guess some x_0
Step 2: Let $x_1 := g(x_0)$; that is, project $g(x_0)$ to $y = x$ line and set it as the new input.
Step 3: Keep going with $x_{k+1} := g(x_k)$.
Or shown differently:

$$x_0 \mapsto x_1 = g(x_0) \mapsto x_2 = g(x_1) \mapsto x_3 = g(x_2) \mapsto \cdots \mapsto x_{n+1} = g(x_n)$$

Remark: These seem to converge (spiral inwards), but how do we conduct (numerical) analysis?

(0) Correct limit?
For example, $x_{n+1} = g(x_n)$. So suppose $x_n \to x$. Then $g(x_n) \to g(x)$ because g is continuous. Also, $x_{n+1} \to x$. Then $x \leftarrow x_{n+1} = g(x_n) \to g(x)$.

So if $x_n \to x$, as $n \to \infty$, then $x = g(x)$ is a solution. So if $x = e^x$, then we reverse the steps:

$$xe^x = 1 \implies xe^x - 1 = f(x) = 0$$

is a solution. We cannot blanketly say this is the unique solution, because our original equation may have multiple (or extraneous) solutions. But we can be sure that our algorithm (fixed point iteration) does converge to the correct thing, provided the limit exists.

(1) Rate of convergence?
Assuming (or given) g differentiable:

$$x_{n+1} = g(x_n)$$
$$x = g(x)$$
$$\underbrace{x_{n+1} - x}_{e_{n+1}} = g(x_n) - g(x) \quad \text{(subtracting these equations, we get the error)}$$
$$\neq g(e_n)$$
$$= g'\underbrace{(\xi_n)}(x_n - x)$$

where this ξ_n is unknown, given by MVT between x_n and x

$$\implies |e_{n+1}| \leq |g'(\xi_n)||e_n|$$

This is an optimistic bound; usually, the error is not (much) lower than this upper bound. We use this to estimate:

$$|e_{n+1}| \leq |g'(\xi_n)||e_n| \leq \max|g'(x)||e_n|$$

Consider:

$$\begin{aligned} g(x) &= e^{-x} \\ g'(x) &= -e^{-x} \\ |g'(x)| &= |-e^{-x}| = e^{-x} \leq \frac{1}{2}; \quad x \geq \ln 2 \approx 0.6931 \end{aligned}$$

However, $x = e^{-x} \leq \frac{1}{2}$. We call this "slightly slow convergence". (To estimate this precisely, we need to know the solution; but of course, we are trying to *find* the solution, so this is non-optimal.)

So we try something else:

$$\begin{aligned} xe^x - 1 &= 0 \\ x + xe^x - 1 &= x \\ x(1 + e^x) &= 1 + x \\ x &= \frac{1+x}{1+e^x} =: g(x) \end{aligned}$$

This should be equivalent because we didn't divide by 0. So for the derivative,

$$\begin{aligned} g'(x) &= \frac{(1+e^x)\cdot 1 - (1+x)\cdot e^x}{(1+e^x)^2} \\ &= \frac{e^x + 1 - e^x - xe^x}{(1+e^x)^2} \\ &= \frac{1 - xe^x}{(1+e^x)^2} \end{aligned}$$

So $g'(x) = 0$ at the solution; so if we get close enough within $\pm\frac{1}{2}$ of the solution, it is guaranteed that we will find the solution.

What happens if we try to bound:

$$\left|\frac{1-xe^x}{(1+e^x)^2}\right| \leq \begin{cases} \frac{1}{4} & x = 0 \\ \text{maybe } \frac{1}{2}? & 0 < x < 1 \\ \frac{1}{5} & x = 1 \end{cases}$$

To prove this, suppose $0 < x_0 < 1$. Then

$$0 \leq g(x_0) = x_1 = \frac{1+x_0}{1+e^{x_0}} \leq 1$$

which we verify by comparing the graphs of $e^x > x$ in this interval, and knowing that e^x is positive.

Now consider:

$$|g'(x)| = \underbrace{\left|\frac{1-xe^x}{(1+e^x)^2}\right|}_{\text{find upper bound}} \leq \frac{1+xe^x}{(1+e^x)^2} \leq \frac{1+e}{4} \leq \frac{3.7}{4} < 1$$

This isn't good enough, so consider that the minimum of the denominator is 2^2 (at $x = 0$).

For the numerator, we check endpoints and where the derivative of the numerator is 0:

$$|1 - xe^x| \leq \begin{cases} 1 & x = 0 \\ e - 1 \approx 1.7 & x = 1 \end{cases}$$
$$\implies -1 \cdot e^x - xe^x < 0$$

So this shows that $|1 - xe^x| \leq 1.7$, so we showed our desired expression is bounded by $\frac{1.7}{4} < \frac{1}{2}$.

Conclusion. This $g(x)$ is better than $g(x) = e^{-x}$.

Theorem 2.1. Suppose $x_{n+1} = g(x_n)$, with g continuous.
1. $a \leq x \leq b \implies a \leq g(x) \leq b$; Invariance
2. $a \leq x \leq b \implies |g'(x)| \leq |\frac{1}{2}|$; Contraction

then $|x_n - x| \leq 2^{-n}|x_0 - x|$ and $x_n \to x$ as $n \to \infty$

Famous example.
The square root $\sqrt{}$.

Step 0:

$x_{n+1} := \frac{1}{2}\left(x_n + \frac{a}{x_n}\right) = g(x_n)$. If $x_n \to x$, then $x = g(x)$.

$$x = \frac{1}{2}\left(x + \frac{a}{x}\right)$$
$$2x = x + \frac{a}{x}$$
$$x = \frac{a}{x}$$
$$x^2 = a \implies x^2 - a = 0.$$

Remark: Usually easier to check contraction first, then invariance after.

Step 1: Contraction

$$|g'(x)| = \left|\frac{1}{2}\left(1 - \frac{a}{x^2}\right)\right| \leq \frac{1}{2}$$
$$\implies \left|1 - \frac{a}{x^2}\right| \leq 1$$
$$\implies -1 \leq 1 - \frac{a}{x^2} \leq 1.$$

UB is obvious, so we check LB:

$$-1 \leq 1 - \frac{a}{x^2}$$
$$-2 \leq -\frac{a}{x^2}$$
$$2 \geq \frac{a}{x^2}$$
$$x^2 \geq \frac{a}{2}$$
$$x \geq \sqrt{\frac{a}{2}} = \frac{\sqrt{a}}{\sqrt{2}}$$

We can then assume $1 \leq a \leq 4$; as long as we start as bigger than $\sqrt{a}$, for example a, then we are good. So we have shown contraction.
Step 2: Invariance

$$g(x) = \frac{1}{2}\left(x + \frac{a}{x}\right)$$

To prove this is invariant, let's try the interval $1 \leq x \implies g(x) \geq \frac{1}{2}\left(1 + \frac{a}{a}\right) = 1$. Then $x \leq a \implies g(x) \leq \frac{1}{2}\left(a + \frac{a}{1}\right) = a$, so the interval $[1, a]$ is invariant (quite beautifully).
Alternatively, we could use calculus with $g'(x) = \frac{1}{2}\left(1 - \frac{a}{x^2}\right) = 0$ at $x = \sqrt{a}$.

Remark: So $[1, a]$ is invariant and $|g'(x)| \leq \frac{1}{2}$ over this interval, so $|x_n - x = \sqrt{a}| \leq 2^{-n}|a - 1|$.

Lecture ends here. Next time we'll talk more about Newton's method and fixed point iteration.