CS 714
FALL 2020
HOMEWORK 2

SAMUEL JACKSON

Introduction

I worked with, in no particular order, Lewis Gross, Haley Colgate, Sourav Pal, and Varun Gudibanda on this assignment.

All code for this assignment can be found in the “`homework2`” directory of `https://github.com/stjacks/STJ-work-for-CS-714.git`.

Problem A

A.a.

Proof. Suppose $w_1, w_2, \dots, w_n$ are orthogonal and $v \in \text{span}(w_1, w_2, \dots, w_n)$. We wish to prove

$$v = \sum_{j=1}^{n} \frac{\langle v, w_i \rangle}{||w_j||^2} w_j$$

Since $v \in \text{span}(w_1, \dots, w_n)$, we know

$$v = \sum_{i=1}^{n} a_i w_i = a_1 w_1 + a_2 w_2 + \dots + a_n w_n$$

Consider $\langle v, w_i \rangle$. Then

$$\langle v, w_i \rangle = a_1 \langle w_1, w_i \rangle + \dots + a_n \langle w_n, w_i \rangle = a_i \langle w_i, w_i \rangle = a_i ||w_i||^2$$

Thus, we see

$$v = \sum_{j=1}^{n} \frac{\langle v, w_i \rangle}{||w_j||^2} w_j = \sum_{j=1}^{n} \frac{a_j ||w_j||^2}{||w_j||^2} w_j = \sum_{j=1}^{n} a_j w_j$$

which is precisely the definition of $v \in \text{span}(w_1, w_2, \dots, w_n)$, as desired. □

Date: November 1, 2020.

A.b.i. If our initial guess is the solution, then clearly we converge in $0 < N$ iterations

A.b.ii. We wish to prove that $\langle p_n, p_j \rangle = 0$ for all $0 \le j < n \le n^* - 1$. We will do this via induction.

Base case: $n = 0$
Trivial. Holds for all $0 \le j < n = 0$, i.e. the empty set.

Base case: $n = 1$
Note that

$$p_1 = r_1 - \frac{\langle r_1, p_0 \rangle_A}{||p_0||_A^2} p_0$$

We find

$$\begin{aligned}
\langle p_1, p_0 \rangle_A &= p_1^T A p_0 \\
&= \left(r_1 - \frac{\langle r_1, p_0 \rangle_A}{||p_0||_A^2} p_0 \right)^T A p_0 \\
&= r_1^T A p_0 - \frac{\langle r_1, p_0 \rangle_A}{||p_0||_A^2} p_0^T A p_0 \\
&= r_1^T A p_0 - \frac{\langle r_1, p_0 \rangle_A}{||p_0||_A^2} ||p_0||_A^2 \\
&= r_1^T A p_0 - \langle r_1, p_0 \rangle_A \\
&= r_1^T A p_0 - r_1^T A p_0 \\
&= 0
\end{aligned}$$

as desired.

Inductive step Suppose that $\langle p_k, p_j \rangle_A = 0$ for all $0 \le j < k \le n^* - 1$ and all $k \le n$. Additionally, note that, since A is symmetric,

$$\langle p_k, p_j \rangle_A = \langle p_j, p_k \rangle_A.$$

Then

$$\begin{aligned}
\langle p_{n+1}, p_j \rangle_A &= p_{n+1}^T A p_j \\
&= \left(r_{n+1} - \sum_{i=0}^{n} \frac{\langle r_{n+1}, p_i \rangle_A}{||p_i||_A^2} p_i \right)^T A p_j \\
&= r_{n+1}^T A p_j - \sum_{i=0}^{n} \frac{\langle r_{n+1}, p_i \rangle}{||p_i||_A^2} p_i^T A p_j \\
&= r_{n+1}^T A p_j - \sum_{i=0}^{n} \frac{\langle r_{n+1}, p_i \rangle}{||p_i||_A^2} \langle p_i, p_j \rangle_A \\
&= r_{n+1}^T A p_j - \frac{\langle r_{n+1}, p_j \rangle}{||p_j||_A^2} \langle p_j, p_j \rangle_A && \text{by induction hypothesis and } A \text{ symmetric} \\
&= r_{n+1}^T A p_j - r_{n+1}^T A p_j \\
&= 0
\end{aligned}$$

as desired.

A.c.i. By problem A.a, we know that any vector w can be written as

$$w = \sum_{i=1}^{N} \frac{\langle w, \phi_i \rangle}{\langle \phi_i, \phi_i \rangle} \phi_i$$

As ϕ_i are orthonormal, this simplifies to to

$$w = \sum_{i=1}^{N} \langle w, \phi_i \rangle \phi_i$$

Thus, we see

$$\begin{aligned}
\langle Av, w \rangle &= v^T A^T w \\
&= v^T A w \\
&= \left(\sum_{i=1}^{N} \langle v, \phi_i \rangle \phi_i^T \right) A \left(\sum_{i=1}^{N} \langle w, \phi_i \rangle \phi_i \right) \\
&= \left(\sum_{i=1}^{N} \langle v, \phi_i \rangle \phi_i^T \right) \left(\sum_{i=1}^{N} \langle w, \phi_i \rangle A \phi_i \right) \\
&= \left(\sum_{i=1}^{N} \langle v, \phi_i \rangle \phi_i^T \right) \left(\sum_{i=1}^{N} \lambda_i \langle w, \phi_i \rangle \phi_i \right) \\
&= \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_j \langle v, \phi_i \rangle \langle w, \phi_j \rangle \langle \phi_i, \phi_j \rangle
\end{aligned}$$

Note that $\langle x, y \rangle = x^T y$ is a scalar, so $x^T y = (x^T y)^T = y^T x = \langle y, x \rangle$. Additionally, as ϕ_i are orthonormal, we find

$$\langle Av, w \rangle = \sum_{i=1}^{N} \lambda_i \langle v, \phi_i \rangle \langle \phi_i, w \rangle$$

as desired.

A.c.ii. As $A \in \mathbb{R}^{N \times N}$ is symmetric positive definite, we know that $x^T A x > 0$ for any vector x. Thus,

$$\begin{aligned}
\phi_i^T A \phi_i &> 0 \\
\phi_i^T (\lambda_i \phi_i) &> 0 \\
\lambda_i \langle \phi_i, \phi_i \rangle &> 0 \\
\lambda_i &> 0
\end{aligned}$$

As this is general for all i, we prove that $\lambda_i > 0$ for $1 \leq i \leq N$.

A.c.iii. Note that $\phi_1, \phi_2, \dots, \phi_n$ form a basis for $\mathbb{R}^N$. Thus, we can write

$$v = a_1\phi_1 + a_2\phi_2 + \cdots + a_N\phi_N$$

for some constants $a_1, \dots, a_N$. Note that

$$\langle v, v\rangle = ||v||^2 = a_1^2\langle \phi_1, \phi_1\rangle + a_2^2\langle \phi_2, \phi_2\rangle + a_N^2\langle \phi_N, \phi_N\rangle$$

We see

$$\begin{aligned}
\langle Av, v\rangle &= v^T A^T v \\
&= (a_1\phi_1 + a_2\phi_2 + \cdots + a_N\phi_N)^T A(a_1\phi_1 + a_2\phi_2 + \cdots + a_N\phi_N) \\
&= (a_1\phi_1 + a_2\phi_2 + \cdots + a_N\phi_N)^T (a_1\lambda_0\phi_1 + a_2\lambda_2\phi_2 + \cdots + a_N\lambda_N\phi_N) \\
&= \lambda_1 a_1^2\langle \phi_1, \phi_1\rangle + \lambda_2 a_2^2\langle \phi_2, \phi_2\rangle + \cdots + \lambda_N a_N^2\langle \phi_N, \phi_N\rangle \\
&\le \lambda_N a_1^2\langle \phi_1, \phi_1\rangle + \lambda_N a_2^2\langle \phi_2, \phi_2\rangle + \cdots + \lambda_N a_N^2\langle \phi_N, \phi_N\rangle \\
&= \lambda_N ||v||^2
\end{aligned}$$

Analogously,

$$\begin{aligned}
\langle Av, v\rangle &= \lambda_1 a_1^2\langle \phi_1, \phi_1\rangle + \lambda_2 a_2^2\langle \phi_2, \phi_2\rangle + \cdots + \lambda_N a_N^2\langle \phi_N, \phi_N\rangle \\
&\ge \lambda_1 a_1^2\langle \phi_1, \phi_1\rangle + \lambda_1 a_2^2\langle \phi_2, \phi_2\rangle + \cdots + \lambda_1 a_N^2\langle \phi_N, \phi_N\rangle \\
&= \lambda_1 ||v||^2
\end{aligned}$$

Thus, we prove

$$\lambda_1 ||v||^2 \le \langle Av, v\rangle \le \lambda_N ||v||^2$$

A.c.iv. As in part iii, we write

$$v = a_1\phi_1 + a_2\phi_2 + \cdots + a_N\phi_N$$

for some constants $a_1, \dots, a_N$. Then

$$\begin{aligned}
||Av||^2 &= \langle Av, Av\rangle \\
&= (Av^T)Av \\
&= v^T AAv \\
&= (a_1\phi_1 + a_2\phi_2 + \cdots + a_N\phi_N)^T AA(a_1\phi_1 + a_2\phi_2 + \cdots + a_N\phi_N) \\
&= (a_1\phi_1 + a_2\phi_2 + \cdots + a_N\phi_N)^T A(a_1\lambda_1\phi_1 + a_2\lambda_2\phi_2 + \cdots + a_N\lambda_N\phi_N) \\
&= (a_1\phi_1 + a_2\phi_2 + \cdots + a_N\phi_N)^T (a_1\lambda_1^2\phi_1 + a_2\lambda_2^2\phi_2 + \cdots + a_N\lambda_N^2\phi_N) \\
&= a_1\lambda_1^2\langle \phi_1, \phi_1\rangle + a_2\lambda_2^2\langle \phi_2, \phi_2\rangle + \cdots + a_N\lambda_N^2\langle \phi_N, \phi_N\rangle \\
&\le a_1\lambda_N^2\langle \phi_1, \phi_1\rangle + a_2\lambda_N^2\langle \phi_2, \phi_2\rangle + \cdots + a_N\lambda_N^2\langle \phi_N, \phi_N\rangle \\
&= \lambda_N^2 ||v||^2
\end{aligned}$$

Thus, we see

$$||Av|| \le \lambda_N ||v||$$

and, if we wished, we could prove analogously that

$$\lambda_1 ||v|| \le ||Av||$$

A.d. By applying definitions:

$$\begin{aligned}
p_{n+1} &= (1+\beta_n)p_n - \alpha_n A p_n - \beta_{n-1}p_{n-1} \\
r_{n+1} + \beta_n p_n &= (1+\beta_n)p_n - \alpha_n A p_n - \beta_{n-1}p_{n-1} \\
r_{n+1} &= p_n - \alpha_n A p_n - \beta_{n-1}p_{n-1} \\
r_n - \alpha_n w_n &= p_n - \alpha_n A p_n - \beta_{n-1}p_{n-1} \\
r_n - \alpha_n (A p_n) &= p_n - \alpha_n A p_n - \beta_{n-1}p_{n-1} \\
r_n &= p_n - \beta_{n-1}p_{n-1} \\
r_n &= (r_n + \beta_{n-1}p_{n-1}) - \beta_{n-1}p_{n-1} \\
0 &= 0
\end{aligned}$$

As we reached a true statement, we deduce that the original equality

$$p_{n+1} = (1+\beta_n)p_n - \alpha_n A p_n - \beta_{n-1}p_{n-1}$$

must be true.

A.e.

Proof. Suppose that $A \in \mathbb{R}^{N\times N}$ is non-singular. Recall that the characteristic polynomial of A is

$$p(\lambda) = \det(\lambda I - A)$$

is a monic polynomial of degree n. That is,

$$p(\lambda) = \lambda^n + c_{n-1}\lambda^{n-1} + \cdots + c_1\lambda + c_0\lambda^0$$

Since A is non-singular, the Cayley-Hamilton theorem states that

$$p(A) = 0$$

Thus, we find

$$A^n + c_{n-1}A^{n-1} + \cdots + c_1 A + c_0 I = 0$$
$$A^n = (-c_{n-1})A^{n-1} + \cdots + (-c_1)A + (-c_0)I$$

So A^n can be written as a linear combination of I, A, A^2, …, A^{n-1} as desired. □

A.f.i. Consider

$$\begin{aligned}e_n &= u_n - u\\ &= u_{n-1} + \alpha(f - Au_{n-1}) - u\\ &= (I - \alpha A)u_{n-1} + \alpha f - u\\ &= (I - \alpha A)u_{n-1} + \alpha Au - u\\ &= (I - \alpha A)u_{n-1} - (I - \alpha A)u\\ &= (I - \alpha A)(u_{n-1} - u)\\ &= (I - \alpha A)e_{n-1}\end{aligned}$$

as desired.

A.f.ii. Note:

$$||e_{n+1}|| = ||(1 - \alpha A)e_n||$$

Note that, if we can assume A is sufficiently 'nice' so that $(1 - \alpha A)$ is positive semi-definite, we would be done by A.c.iv. If we can't, then we note the additional inequality

$$||(1 - \alpha A)e_n|| \leq ||(1 - \alpha A)||||e_n||$$

We know that $||(1 - \alpha A)||$ is equal to the spectral radius of $(1 - \alpha A)^1$, which is exactly rho as defined in the problem statement. Thus, we see

$$||e_{n+1}|| \leq ||(1 - \alpha A)|| \cdot ||e_n|| = \rho||e_n||$$

as desired.

A.f.iii. We wish to minimize

$$\rho = \max_{1 \leq j \leq N} |1 - \alpha\lambda_j|$$

Note that for $2 \leq i \leq N - 1$

$$1 - \alpha\lambda_N < 1 - \alpha\lambda_i < 1 - \alpha\lambda_1$$

Thus, the value with the largest magnitude will either be $i = 1$ or $i = N$. We would like to minimize the magnitude of both terms.

Consider $\alpha = \frac{2}{\lambda_1+\lambda_N}$. Then

$$|1 - \alpha\lambda_N| = \left|1 - \frac{2\lambda_N}{\lambda_1 + \lambda_N}\right| = \left|\frac{\lambda_1 - \lambda_N}{\lambda_1 + \lambda_N}\right| = \frac{\lambda_N - \lambda_1}{\lambda_N + \lambda_1}$$

$$|1 - \alpha\lambda_1| = \left|1 - \frac{2\lambda_1}{\lambda_1 + \lambda_N}\right| = \left|\frac{\lambda_N - \lambda_1}{\lambda_1 + \lambda_N}\right| = \frac{\lambda_N - \lambda_1}{\lambda_N + \lambda_1}$$

So if $\alpha = \frac{2}{\lambda_1+\lambda_N}$ then $\rho = \frac{\kappa-1}{\kappa+1}$.

Suppose, for the sake of contradiction, that we choose $\alpha' < \alpha$. Then

$$1 - \alpha\lambda_1 < 1 - \alpha'\lambda_1$$

In particular, since $0 \leq 1 - \alpha\lambda_1$, then

$$|1 - \alpha\lambda_1| < |1 - \alpha'\lambda_1|$$

[1] As per the follow-up discussion to Piazza post 55: `https://piazza.com/class/kexjawqt7y975w?cid=55`

which would be a worse choice for minimizing ρ! Similarly, suppose we choose $\alpha' > \alpha$. Then

$$1 - \alpha\lambda_N > 1 - \alpha'\lambda_N$$

Since $1 - \alpha\lambda_N < 0$, we see

$$|1 - \alpha\lambda_N| < |1 - \alpha'\lambda_N|$$

which, again, would be a worse choice for minimizing ρ. Thus, we see $\alpha = \frac{2}{\lambda_1+\lambda_N}$ is the best choice for minimizing ρ, as desired.

A.f.iv. Suppose we choose $a = \frac{2}{c+C}$ for c and C given in the problem statement. So

$$\begin{aligned} |1 - a\lambda_i| &= \left|1 - \frac{2\lambda_i}{c+C}\right| \\ &= \left|\frac{c+C-2\lambda_i}{c+C}\right| \\ &= \frac{|c+C-2\lambda_i|}{c+C} \end{aligned}$$

Note that if $c + C - 2\lambda_i < 0$, then

$$|1 - a\lambda_i| = \frac{2\lambda_i - C - c}{c+C} \leq \frac{2C - C - c}{c+C} = \frac{C-c}{C+c}$$

Similarly, if $c + C - 2\lambda_i > 0$, then

$$|1 - a\lambda_i| = \frac{C + c - 2\lambda_i}{c+C} \leq \frac{C + c - 2c}{c+C} = \frac{C-c}{C+c}$$

Note that from Problem A.f.iii, since α (as defined in A.f.iii) minimizes the error, we must have

$$\rho \leq \frac{C-c}{C+c} = \frac{\kappa' - 1}{\kappa' + 1} < 1$$

where $\kappa' = C/c$.

A.g.i. Recall that

$$r_1 = r_0 + \alpha_0 A p_0$$

and

$$p_0 = r_0$$

Thus, we find

$$r_1 = r_0 + \alpha_0 A r_0$$

as desired.

A.g.ii. By repeated application of definitions...

$$\begin{aligned}
r_{n+1} &= r_n - \alpha_n A r_n + \frac{\alpha_n \beta_{n-1}}{\alpha_{n-1}}(r_n - r_{n-1}) \\
r_n - \alpha_n w_n &= r_n - \alpha_n A r_n + \frac{\alpha_n \beta_{n-1}}{\alpha_{n-1}}(r_n - r_{n-1}) \\
-\alpha_n A p_n &= -\alpha_n A r_n + \frac{\alpha_n \beta_{n-1}}{\alpha_{n-1}}(r_n - r_{n-1}) \\
-\alpha_n A(r_n + \beta_{n-1} p_{n-1}) &= -\alpha_n A r_n + \frac{\alpha_n \beta_{n-1}}{\alpha_{n-1}}(r_n - r_{n-1}) \\
-\alpha_n \beta_{n-1} A p_{n-1} &= \frac{\alpha_n \beta_{n-1}}{\alpha_{n-1}}(r_n - r_{n-1}) \\
-\alpha_{n-1} A p_{n-1} &= r_n - r_{n-1} \\
-\alpha_{n-1} A p_{n-1} &= (r_{n-1} - \alpha_{n-1} w_{n-1}) - r_{n-1} \\
-\alpha_{n-1} A_{n-1} &= -\alpha_{n-1} A p_{n-1}
\end{aligned}$$

This last line is clearly true, thus, we deduce that the original statement is also true.

A.g.iii. Recall that $p_0 = r_0$. So

$$\begin{aligned}
A q_0 &= \gamma_0 q_0 - \delta_0 q_1 \\
A \frac{r_0}{||r_0||} &= \frac{1}{\alpha_0 ||r_0||} r_0 - \frac{\sqrt{\beta_0} r_1}{\alpha_0 ||r_1||} \\
\frac{1}{||r_0||} A r_0 &= \frac{1}{\alpha_0 ||r_0||} r_0 - \frac{\sqrt{(||r_1||^2 / ||r_0||^2)} r_1}{\alpha_0 ||r_1||} \\
\frac{1}{||r_0||} A r_0 &= \frac{1}{\alpha_0 ||r_0||}(r_0 - r_1) \\
\frac{1}{||r_0||} A r_0 &= \frac{1}{\alpha_0 ||r_0||}(r_0 - (r_0 - \alpha_0 A p_0)) \\
\frac{1}{||r_0||} A r_0 &= \frac{1}{\alpha_0 ||r_0||}(\alpha_0 A p_0) \\
\frac{1}{||r_0||} A r_0 &= \frac{1}{||r_0||} A r_0
\end{aligned}$$

So we deduce that the first claim is correct. For the second, suppose $1 \leq n \leq n^* - 1$. Then

$$\begin{aligned}
Aq_n &= -\delta_{n-1}q_{n-1} + \gamma_n q_n - \delta_n q_{n+1} \\
A\frac{r_n}{||r_n||} &= -\frac{\sqrt{\beta_{n-1}}}{\alpha_{n-1}}\frac{r_{n-1}}{||r_{n-1}||} + \left(\frac{1}{\alpha_n} + \frac{\beta_{n-1}}{\alpha_{n-1}}\right)\frac{r_n}{||r_n||} - \frac{\sqrt{\beta_n}}{\alpha_n}\frac{r_{n+1}}{||r_{n+1}||} \\
A\frac{r_n}{||r_n||} &= -\frac{\sqrt{\beta_{n-1}}}{\alpha_{n-1}}\frac{r_{n-1}}{||r_{n-1}||} + \left(\frac{1}{\alpha_n} + \frac{\beta_{n-1}}{\alpha_{n-1}}\right)\frac{r_n}{||r_n||} - \frac{\sqrt{(||r_{n+1}||^2)/(||r_n||^2)}}{\alpha_n}\frac{r_{n+1}}{||r_{n+1}||} \\
A\frac{r_n}{||r_n||} &= -\frac{\sqrt{\beta_{n-1}}}{\alpha_{n-1}}\frac{r_{n-1}}{||r_{n-1}||} + \left(\frac{1}{\alpha_n} + \frac{\beta_{n-1}}{\alpha_{n-1}}\right)\frac{r_n}{||r_n||} - \frac{1}{\alpha_n}\frac{r_{n+1}}{||r_n||} \\
Ar_n &= -\frac{\sqrt{\beta_{n-1}}||r_n||}{\alpha_{n-1}||r_{n-1}||}r_{n-1} + \left(\frac{1}{\alpha_n} + \frac{\beta_{n-1}}{\alpha_{n-1}}\right)r_n - \frac{1}{\alpha_n}r_{n+1} \\
Ar_n &= -\frac{\sqrt{\beta_{n-1}}||r_n||}{\alpha_{n-1}||r_{n-1}||}r_{n-1} + \left(\frac{1}{\alpha_n} + \frac{\beta_{n-1}}{\alpha_{n-1}}\right)r_n - \frac{1}{\alpha_n}\left(r_n - \alpha_n A r_n + \frac{\alpha_n\beta_{n-1}}{\alpha_{n-1}}(r_n - r_{n-1})\right) \\
0 &= -\frac{\sqrt{\beta_{n-1}}||r_n||}{\alpha_{n-1}||r_{n-1}||}r_{n-1} + \frac{\beta_{n-1}}{\alpha_{n-1}}r_n - \frac{\beta_{n-1}}{\alpha_{n-1}}(r_n - r_{n-1}) \\
0 &= -\frac{\sqrt{\beta_{n-1}}||r_n||}{||r_{n-1}||}r_{n-1} + \beta_{n-1}r_{n-1} \\
0 &= -\beta_{n-1}r_{n-1} + \beta_{n-1}r_{n-1} \\
0 &= 0
\end{aligned}$$

As we reached a true statement, we deduce the second claim is also true.

A.g.iv. We wish to show that

$$AQ_n = Q_nT_n - \delta_{n-1}q_ne_n^T$$

using the definitions of Q_n, T_n, and e_n defined in the problem description.

Consider the columns of the resulting matrix on the right-hand side. The first column corresponds to Aq_0 and is exactly

$$\gamma_0q_0 - \delta_0q_1$$

which, as shown in A.g.iii, is correct. Now consider the column corresponding to Aq_i for $i < i < n$. Then this column is

$$-\delta_{i-1}q_{i-1} + \gamma_iq_i - \delta_iq_{i+1}$$

which, as shown in A.g.iii, is correct. Finally, consider the column corresponding to Aq_{n-1}. Then this column is

$$-\delta_{n-2}q_{n-2} + \gamma_{n-1}q_{n-1} - \delta_{n-1}q_n$$

where the $-\delta_{n-1}q_n$ term comes from the $-\delta_{n-1}q_ne_n^T$ term. This, by A.g.iii, is also correct. As every column in the resulting matrix is correct, we find that

$$AQ_n = Q_nT_n - \delta_{n-1}q_ne_n^T$$

as desired.

A.g.v. Note that q_i are orthogonal unit vectors, so q_i are orthonormal. From A.g.iv, we know that

$$AQ_n = Q_n T_n - \delta_{n-1} q_n e_n^T$$

So

$$Q_n^T A Q_n = Q_n^T Q_n T_n - \delta_{n-1} Q_n^T q_n e_n^T$$

Since q_i are orthonormal, we know that $Q_n^T Q_n = I$, the identity matrix. Further, notice that $q_n e_n^T = [0\ 0\ \ldots\ q_n]$. So $Q_n^T q_n e_n^T$ will have zeroes in every column except the last, where it will be equal to $[q_0^T q_n\ q_1^T q_n\ \ldots\ q_{n-1}^T q_n]^T$, which is also zero by orthogonality. Thus, if Z denotes the zero matrix, we have

$$Q_n^T A Q_n = I T_n - \delta_{n-1} Z = T_n$$

as desired.

Problem B

We have the function

$$f(x) = e^{-400(x-0.5)^2}$$

Note

$$f'(x) = -800(x-0.5)e^{-400(x-0.5)^2}$$

Suppose we have a line $y = mx + c$ that intersects $f(x)$ at points a and b. We'd like to find the point between a and b that maximizes the difference between $f(x)$ and y. To do this, we need to find the critical points of

$$f(x) - y$$

which corresponds to solving

$$f'(x) - y' = 0$$

So,

$$-800(x-0.5)e^{-400(x-0.5)^2} - m = 0$$

Unfortunately, I don't know how to solve this. Luckily for us, we can ask Matlab to solve this! While this will be a numerical solution and involve a small error, this error will be extremely small due to the accuracy of `fzero` and will not effect our 10^{-2} bound. Unfortunately `fzero` occasionally doesn't return a critical point. In this case we have a second approach. Note that

$$f'(x) = -800(x-0.5)e^{-400(x-0.5)^2} \leq -400e^{-400(x-0.5)^2} \leq -400$$

In particular, this means that if an approximation of the critical point is off by h from the actual critical point, then the maximum amount of error when evaluating that critical point is is $h \cdot |-400| + h \cdot |m| = h(400 + |m|)$. I.e., we assume the linear function and $f(x)$ grow further apart by their maximal slopes. Since we want 10^{-2} accuracy, if we sample such that

$$h(400 + |m|) \leq 10^{-4}$$

then we're probably fine. In particular,

$$h \leq \frac{1}{10^4 \cdot (400 + |m|)}$$

We do this only when `fzero` doesn't give an answer (and only when double-checking our answer, see code comments).

We see that $N = 100$ appears to be the first value of N with maximum error $< 10^{-2}$. Further, since the maximum error detected when $N = 100$ was 0.0097, we know an additional error of 10^{-4} at some point wouldn't cause the overall error to increase above 10^{-2}. Mission accomplished!

Problem C

C.a. We have the following stencil:

$$\frac{u(x,y,t-1)-2u(x,y,t)+u(x,y,t+1)}{(\Delta t)^2}$$

$$=\frac{u(x+1,y,t)+u(x-1,y,t)+u(x,y+1,t)+u(x,y-1,t)-4u(x,y,t)}{(\Delta x)^2}$$

So,

$$u(x,y,t+1)=-u(x,y,t-1)+2u(x,y,t)+(\Delta t)^2(\text{five-point Laplacian})$$

And at the edges of the x/y grid we can use the 2nd-order forward/backward difference formula in x and y for the second derivative.

Of course, this is a two-step method in time so we need to bootstrap the first step. We will do this by Forward Euler. So

$$u(x,y,t+1)=u(x,y,t)+\Delta tu'(x,y,t)$$

Since we are dealing with the case where t=0, this simplifies to

$$u(x,y,\Delta t)=u(x,y,0)+\Delta tu'(x,y,0)$$

$$u(x,y,\Delta t)=0+\Delta t(f(x)f(y))$$

Since $u(x,y,0)=0$, we don't have to worry about incorporating the 5-point Laplacian in x/y (since it will be 0). Since this is the "simplest" numerical method as directed by the problem, we also don't need to worry about sacrificing our 2nd order accuracy by using a 1st order accurate method for the first step (see section 5.9.3 LeVeque 2007 for additional details).

We will compare error by making a large grid and using that as the 'correct' value. Since we want to observe convergence as a function of Δx and not Δt, we fix a very small Δt. Unfortunately the script `ProblemCa.m` takes a very long time to run (approximately 4 hours on the CS lab computer `snares-01.cs.wisc.edu`). Additionally, the memory constraints of the matrices in question are so large that it's difficult to run on either my personal laptop or the remote CS machines. Together, this is why we use only four different values of Δx and, further, only iterate our solution for 100 time steps. Since we fixed Δt to be extremely small, we aren't iterating very far at all in time, only to time $100\cdot\Delta t$. With more computing power we would be able to both test more values of Δx and iterate our solution for more timesteps.

As the experiments were both a) run remotely and b) ran for a long time, I created a second script `ProblemCaPlot.m` to plot the results. `ProblemCa.m` prints the error vector, while `ProblemCaPlot.m` takes this vector and displays the log-log plot asked for. Figure 1 is the resulting plot. We see that the error is indeed second-order accurate, as desired.

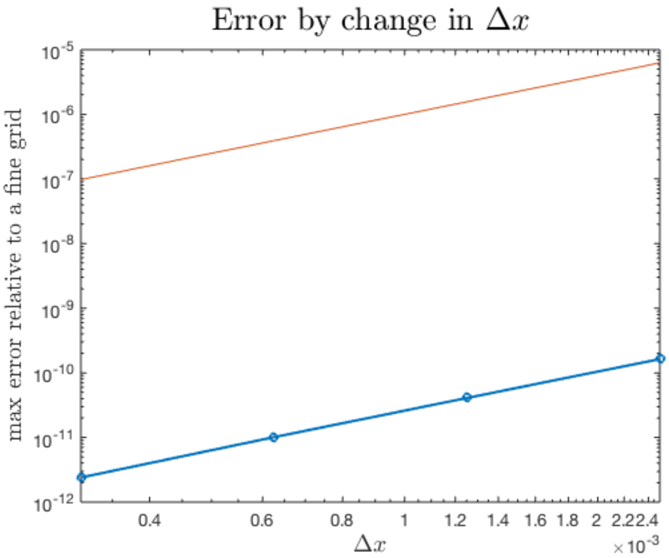

FIGURE 1. Graph shown after running `ProblemCa.m`. The blue line is the error we observed while the red line is a reference of what the slope should be if we observe 2nd order error

C.B

We have $y''(t) = \lambda y(t)$. Using the usual centered difference formula (and using the notation $y(t) = y^t$), we get

$$\frac{y^{t+1} - 2y^t + y^{t-1}}{(\Delta t)^2} = \lambda y^t$$
$$y^{t+1} = (2 + \lambda(\Delta t)^2)y^t - y^{t-1}$$
$$y^{t+1} - (2 + \lambda(\Delta t)^2)y^t + y^{t-1} = 0$$

To find the region of stability, we need to find ρ that satisfy

$$\rho^2 - (2 + \lambda(\Delta t)^2)\rho + 1 = 0$$

Using the quadratic formula,

$$\rho = \frac{(2 + \lambda(\Delta t)^2) \pm \sqrt{(2 + \lambda(\Delta t)^2)^2 - 4}}{2}$$

And, letting $x = \lambda(\Delta t)^2$,

$$\rho = \frac{(2 + x) \pm \sqrt{(2 + x)^2 - 4}}{2}$$

Let

$$\rho_+ = 1 + \frac{x}{2} + \frac{\sqrt{(2 + x)^2 - 4}}{2}$$
$$\rho_- = 1 + \frac{x}{2} - \frac{\sqrt{(2 + x)^2 - 4}}{2}$$

We see that $re(x) \leq 0$, otherwise $|\rho_+| > 1$. Further, $re(x) \geq -4$, otherwise $|\rho_-| > 1$. If we plug in $x = a + bi$ and try to solve the inequality directly for constraints on a and b, however, this will be incredibly messy. Instead, we'll simply use MATLAB to sample a fine grid and plot the region such that $|\rho_+|$ and $|\rho_-|$ are ≤ 1.

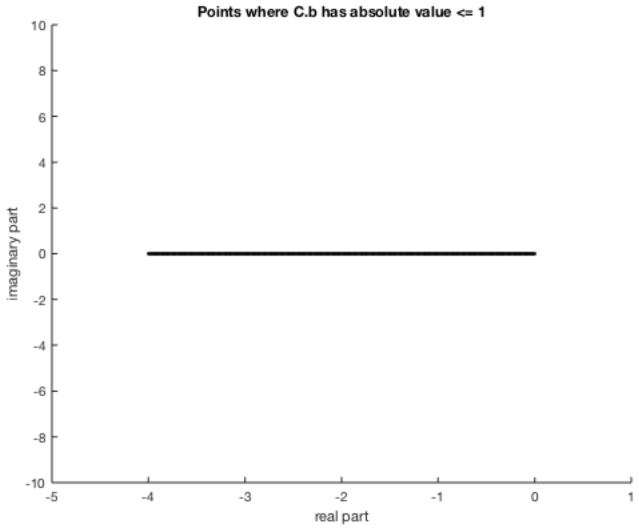

FIGURE 2. Graph shown after running `ProblemCb.m`, i.e. the stability region

As shown in Figure 2, ρ_+ and ρ_- have magnitude ≤ 1 when they are both real and both between -4 and 0

C.c. We would like to use the method of lines. Define $\Delta_h = I \otimes A + A \otimes I$, where A is a tri-diagonal matrix with $-1/h^2$ on the lower and upper diagonals and $2/h^2$ on the diagonal (the usual matrix for the centered finite difference for the 2nd derivative in time). This is the discretization in space of the LHS of our PDE. We know

$$\lambda_k(A) = \frac{-4}{h^2} \sin^2\left(\frac{k\pi h}{2}\right)$$

for $k = 1, \dots, N+1$. Thus, we know

$$\lambda_{k,l}(\Delta_h) = \frac{-4}{h^2}\left(\sin^2\left(\frac{k\pi h}{2}\right) + \sin^2\left(\frac{l\pi h}{2}\right)\right)$$

for $1 \leq k, l \leq N+1$.

As our method of lines discretization is that $u_{tt} = \Delta_h u$, we know from C.b that the region of stability requires $\lambda(\Delta_h)(\Delta t)^2$ to be real and between -4 and 0 to guarantee convergence. So we need

$$-4 \leq \frac{-4(\Delta t)^2}{h^2}\left(\sin^2\left(\frac{k\pi h}{2}\right) + \sin^2\left(\frac{l\pi h}{2}\right)\right) \leq 0$$

or, simplifying,

$$0 \leq \frac{(\Delta t)^2}{h^2}\left(\sin^2\left(\frac{k\pi h}{2}\right) + \sin^2\left(\frac{l\pi h}{2}\right)\right) \leq 1$$

Note that $0 \leq \sin^2(k\pi h/2) + \sin^2(l\pi h/2) \leq 2$, so our expression is always ≥ 0. Thus, to satisfy the condition, we need to guarantee

$$2\frac{(\Delta t)^2}{h^2} \leq 1$$
$$\frac{(\Delta t)^2}{h^2} \leq \frac{1}{2}$$

which is our CFL condition.

C.d. Let $h = \Delta x$ and $g = g(k_1, k_2)$. Suppose that $u^n_{j,l} = \rho^n e^{ik_1 jh} e^{ik_2 lh}$. Without loss of generality let $\rho^{n-1} = 1$ and $g = \rho^n$. Then using the same equations we used in C.a, we have

$$g^2 e^{ik_1jh}e^{ik_2lh} = \Bigg(-e^{ik_1jh}e^{ik_2lh} + 2ge^{ik_1jh}e^{ik_2lh}$$

$$+ (\Delta t)^2(1/h)^2 g(e^{ik_1(j-1)h}e^{ik_2lh} + e^{ik_1(j+1)h}e^{ik_2lh} + e^{ik_1jh}e^{ik_2(l-1)h} + e^{ik_1jh}e^{ik_2(l+1)h} - 4e^{ik_1jh}e^{ik_2lh})\Bigg)$$

Simplifying, we have

$$g^2 = -1 + 2g + (\Delta t)^2(1/h)^2 g(e^{-ik_1h} + e^{ik_1h} + e^{-ik_2h} + e^{ik_2h} - 4)$$

Using Euler's formula and the half-angle formula and simplifying, we find

$$g^2 = -1 + g\left(2 + (\Delta t)^2(1/h)^2(2\cos(k_1h) + 2\cos(k_2h) - 4)\right)$$
$$g^2 = -1 + g\left(2 + (\Delta t)^2(1/h)^2 2(\cos(k_1h) - 1 + \cos(k_2h) - 1)\right)$$
$$g^2 = -1 + g\left(2 - (\Delta t)^2(1/h)^2 4(\sin^2(k_1h/2) + \sin^2(k_2h/2))\right)$$

So

$$g^2 - g(2 - 4(\Delta t)^2(1/h)^2(\sin^2(k_1h/2) + \sin^2(k_2h/2))) + 1 = 0$$

Notice that if we let $x = -4(\Delta t)^2(1/h)^2(\sin^2(k_1h/2) + \sin^2(k_2h/2))$ then we have the equation

$$g^2 - g(2 - x) + 1 = 0$$

which is precisely the equation we solved in C.b. We know that $|g| \leq 1$ for $-4 \leq x \leq 0$. Thus, our constraints are

$$-4 \leq -4\left(\frac{\Delta t}{h}\right)^2(\sin^2(k_1h/2) + \sin^2(k_2h/2)) \leq 0$$
$$0 \leq \left(\frac{\Delta t}{h}\right)^2(\sin^2(k_1h/2) + \sin^2(k_2h/2)) \leq 1$$

Note that $0 \leq (\sin^2(k_1h/2) + \sin^2(k_2h/2)) \leq 2$. So clearly $(\Delta t)^2(1/h)^2(\sin^2(k_1h/2) + \sin^2(k_2h/2))$ is bounded below by 0. Further,

$$\left(\frac{\Delta t}{h}\right)^2(\sin^2(k_1h/2) + \sin^2(k_2h/2)) \leq 2\left(\frac{\Delta t}{h}\right)^2$$

So to get our CFL condition we need

$$2\left(\frac{\Delta t}{h}\right)^2 \leq 1$$
$$\left(\frac{\Delta t}{h}\right)^2 \leq \frac{1}{2}$$

which is precisely the same condition we reached in C.c.

University of Wisconsin, Fall 2020