CS 714

HOMEWORK 2

Noah Cohen Kalafut
Computer Science Doctoral Student
University of Wisconsin-Madison
nkalafut@wisc.edu
https://github.com/Oafish1/CSC-714

November 2, 2020

1 Problem A

1.1 Part a

We are given

$$w_1, w_2, \ldots, w_n \text{are orthogonal}$$
$$v \in span\{w_1, w_2, \ldots, w_n\}$$

Then,

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

Since w_i and w_j are orthogonal for $i \neq j$, $\langle w_i, w_j \rangle = 0$ for $i \neq j$ and we can write

$$v = \sum_{j=1}^{n} \frac{\langle c_j w_j, w_j \rangle}{||w_j||^2} w_j = \sum_{j=1}^{n} \frac{\langle c_1 w_1 + c_2 w_2 + \cdots + c_n w_n, w_j \rangle}{||w_j||^2} w_j = \sum_{j=1}^{n} \frac{\langle v, w_j \rangle}{||w_j||^2} w_j$$

1.2 Part b

1.2.1 Part i

Some N may not be linearly independent, meaning that a basis will have strictly fewer than N vectors.

1.2.2 Part ii

We are given

$$p_0 = r_0$$
$$p_n = r_n - \sum_{i=0}^{n-1} \frac{\langle r_n, p_i \rangle_A}{||p_i||_A^2} p_i \text{ for } 1 \leq n \leq n^* - 1 \text{ and symmetric } A$$

We want to prove

$$\langle p_n, p_j \rangle_A = 0 \text{ for } 0 \leq j < n \leq n^* - 1$$

This makes logical sense. What we are doing is taking an original vector r_n and subtracting its projection onto every $p_{j<n}$ – making the result, p_n, orthogonal to those vectors.

Assuming conducive n^*, consider the case where $n = 1$.

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

Then,

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

Suppose $\langle p_n, p_{j_n} \rangle_A = 0$ for $0 \leq j_n < n \leq n^* - 2$.

We can say

$$p_{n+1} = r_{n+1} - \sum_{i=0}^{n} \frac{\langle r_{n+1}, p_i \rangle_A}{||p_i||_A^2} p_i$$

Then, by inductive hypothesis,

$$\begin{aligned}
\left\langle p_{n+1}, p_{j_{n+1}} \right\rangle_A &= \left\langle r_{n+1}, p_{j_{n+1}} \right\rangle_A - \sum_{i=0}^{n} \frac{\langle r_{n+1}, p_i \rangle_A}{||p_i||_A^2} \left\langle p_i, p_{j_{n+1}} \right\rangle_A \\
&= \left\langle r_{n+1}, p_{j_{n+1}} \right\rangle_A - \left\langle r_{n+1}, p_{j_{n+1}} \right\rangle_A = 0
\end{aligned}$$

Thus, $\langle p_n, p_{j_n} \rangle_A = 0$ for $0 \leq j_n < n \leq n^* - 2$ implies $\left\langle p_n, p_{j_{n+1}} \right\rangle_A = 0$ for $0 \leq j_{n+1} < n + 1 \leq n^* - 1$.

So,

$$\langle p_n, p_j \rangle_A = 0 \text{ for } 0 \leq j < n \leq n^* - 1$$

by finite induction.

1.3 Part c

We are given

$A \in \mathbb{R}^{N \times N}$ is symmetric positive definite and has a basis of orthonormal eigenvectors
$\phi_1, \phi_2, \ldots, \phi_N$ with corresponding eigenvalues
$\lambda_1, \lambda_2, \ldots, \lambda_N$ in ascending order

For any $v, w \in \mathbb{R}^N$...

1.3.1 Part i

We are asked to prove

$$\langle Av, w \rangle = \sum_{n=1}^{N} \lambda_n \langle v, \phi_n \rangle \langle \phi_n, w \rangle$$

Observe

$$\sum_{n=1}^{N} \lambda_n \langle v, \phi_n \rangle \langle \phi_n, w \rangle = \sum_{n=1}^{N} \lambda_n v^T \phi_n \phi_n^T w = \sum_{n=1}^{N} \lambda_n \left\langle \phi_n \phi_n^T v, w \right\rangle \tag{1}$$

Because A is symmetric, it is diagonalizable as

$$A = \Phi \Lambda \Phi^T = \sum \lambda_n \phi_n \phi_n^T \tag{2}$$

Where Φ is a matrix with column vectors ϕ_n and Λ has corresponding λ_n on the diagonal.

Using 1 and 2, we can see

$$\sum_{n=1}^{N} \lambda_n \langle v, \phi_n \rangle \langle \phi_n, w \rangle = \sum_{n=1}^{N} \lambda_n \left\langle \phi_n \phi_n^T v, w \right\rangle = \langle Av, w \rangle$$

1.3.2 Part ii

We want to prove $\lambda_n > 0$ for $1 \leq n \leq N$.

We know A is positive definite, so we can say

$$z^T A z > 0 \text{ for } z \in \mathbb{R}^N \tag{3}$$

by definition.

Let $z = \phi_n$ for $1 \leq n \leq N$. Then,

$$\phi_n^T A \phi_n = \phi_n^T \lambda_n \phi_n = \lambda_n ||\phi_n||^2 \tag{4}$$

Recall that $\phi_n^T \phi_n = ||\phi_n||^2 > 0$.[1]

Combining 3 and 4, it must be the case that $\lambda_n > 0$ for $1 \leq n \leq N$.

1.3.3 Part iii

We want to prove $\lambda_1 ||v||^2 \leq \langle Av, v\rangle \leq \lambda_N ||v||^2$.

Once again using the fact that A is symmetric, we can say

$$\langle Av, v\rangle = \langle \Phi\Lambda\Phi^T v, v\rangle = v^T \Phi \Lambda^T \Phi^T v = ||\Phi^T v||_\Lambda^2$$

Recall that Λ is diagonal. We can expand

$$||\Phi^T v||_\Lambda^2 = \sum_i \lambda_i (\phi_i v)^2$$

It then directly follows

$$\lambda_1 \sum_i (\phi_i v)^2 \leq \sum_i \lambda_i (\phi_i v)^2 \leq \lambda_N \sum_i (\phi_i v)^2$$
$$\lambda_1 ||\Phi^T v||^2 \leq ||\Phi^T v||_\Lambda^2 \leq \lambda_N ||\Phi^T v||^2$$

Finally, since Φ is orthogonal,

$$||\Phi^T v||^2 = v^T \Phi \Phi^T v = v^T v = ||v||^2$$

Thus,

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

Note that this also implies minimizing or maximizing $\langle Av, v\rangle$ for fixed $||v||^2$ is as simple as setting $v = c\phi_1$ or $c\phi_N$

1.3.4 Part iv

We want to prove $||Av|| \leq \lambda_N ||v||$.

Using the facts that A is symmetric and Φ is orthogonal, we can say

$$||Av|| = ||\Phi\Lambda\Phi^T v|| = \sqrt{v^T \Phi \Lambda^T \Phi^T \Phi \Lambda \Phi^T v} = \sqrt{v^T \Phi \Lambda^T \Lambda \Phi^T v} = ||\Phi^T v||_{\Lambda^T\Lambda}$$

We can continue in much the same way as Part iii, eventually concluding

$$\lambda_1 \left(\sum_i (\phi_i v)^2\right)^{1/2} \leq \left(\sum_i (\lambda_i \phi_i v)^2\right)^{1/2} \leq \lambda_N \left(\sum_i (\phi_i v)^2\right)^{1/2}$$
$$\lambda_1 ||\Phi^T v|| \leq ||\Phi^T v||_{\Lambda^T\Lambda} \leq \lambda_N ||\Phi^T v||$$

Note that the result from Part ii is required to determine these bounds due to the $\Lambda^T \Lambda$ inner product.

Since Φ is orthogonal, $||\Phi^T v|| = ||v||$ and we can conclude

$$\lambda_1 ||v|| \leq ||Av|| \leq \lambda_N ||v||$$

[1] Strictly greater than zero because a basis cannot contain $\vec{0}$

1.4 Part d

We have

$$\begin{aligned} p_{n+1} &= r_{n+1} + \beta_n p_n \\ w_{n+1} &= A p_{n+1} \\ r_{n+1} &= r_n - \alpha_n w_n \end{aligned} \tag{5}$$

Let's begin with substitution $w \to r \to p$

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

Notice, by the first equation of 5, $r_{n+1} = p_{n+1} - \beta_n p_n$. So, we can further substitute

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

1.5 Part e

We are given that $A \in \mathbb{R}^{N \times N}$ is non-singular.

Consider the characteristic polynomial of A

$$p(A) = \det(\lambda I_N - A) = c_0 + c_1 \lambda + c_2 \lambda^2 + \cdots + c_N \lambda^N$$

By the Cayley-Hamilton Theorem, replacing λ with A provides $p(A) = 0$. So, we can solve for A^N.

$$\begin{aligned} p(A) = c_0 I_N + c_1 A + c_2 A^2 + \cdots + c_N A^N &= 0 \\ A^N &= -\frac{1}{c_N}(c_0 I_N + c_1 A + c_2 A^2 + \cdots + c_{N-1} A^{N-1}) \\ A^N = k_0 I_N + k_1 A + k_2 A^2 + \cdots + k_{N-1} A^{N-1} &\text{ for } k_n = -\frac{c_n}{c_N} \end{aligned}$$

Thus, A^N can be represented as a linear combination of $I_N, A, A^2, \ldots, A^{N-1}$.

1.6 Part f

We are given

$$u_{n+1} = u_n + \alpha(f - A u_n) \tag{6}$$

1.6.1 Part i

We are given

$$e_n = u_n - u$$

Observe, through substitution

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

Then, using the *Richardson Iteration* from 6,

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

Thus,

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

1.6.2 Part ii

We know from Part i that

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

So,

$$\begin{aligned} ||e_{n+1}|| &= ||(I - \alpha A)e_n|| \\ &= \left(e_n^T (I - \alpha A)^T (I - \alpha A)e_n\right)^{1/2} \end{aligned}$$

Since A is symmetric, we can write

$$\begin{aligned} e_n^T (I - \alpha A)^T (I - \alpha A)e_n &= e_n^T (I - \alpha\Phi\Lambda\Phi^T)^T (I - \alpha\Phi\Lambda\Phi^T)e_n \\ &= (e_n^T - \alpha e_n^T \Phi\Lambda^T\Phi^T)(e_n - \alpha\Phi\Lambda\Phi^T e_n) \\ &= e_n^T e_n - \alpha e_n^T \Phi\Lambda\Phi^T e_n - \alpha e_n^T \Phi\Lambda^T\Phi^T e_n + \alpha^2 e_n^T \Phi\Lambda^T\Phi^T\Phi\Lambda\Phi^T e_n \\ &= ||e_n||^2 - 2\alpha||\Phi^T e_n||_\Lambda^2 + \alpha^2||\Phi^T e_n||_{\Lambda^2}^2 \end{aligned}$$

Following previous work in 1.3, we can state

$$\begin{aligned} e_n^T (I - \alpha A)^T (I - \alpha A)e_n &\leq \max_{1\leq j\leq N} \left(||e_n||^2 - 2\alpha\lambda_j||e_n||^2 + \alpha^2\lambda_j^2||e_n||^2\right) \\ &= \max_{1\leq j\leq N} (1 - 2\alpha\lambda_j + \alpha^2\lambda_j^2)||e_n||^2 \end{aligned}$$

By taking the square root of both sides, we obtain

$$\begin{gathered} ||e_{n+1}|| \leq \rho||e_n||^2 \\ \rho = \max_{1\leq j\leq N} |1 - \alpha\lambda_j| \end{gathered}$$

1.6.3 Part iii

In Part ii, we defined ρ as

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

Minimizing this would raise our estimated rate of convergence.
Consider

$$\alpha = \frac{2}{\lambda_1 + \lambda_N}$$

Notice that, since all λ_j are strictly positive, $|\lambda_N - \lambda_1| < |\lambda_N + \lambda_1|$ and $|1 - \alpha\lambda_1| < 1$

Also notice that, for the above α, $1 - \alpha\lambda_1 = \frac{-\lambda_1+\lambda_N}{\lambda_1+\lambda_N} = -\left(\frac{\lambda_1-\lambda_N}{\lambda_1+\lambda_N}\right) = -(1 - \alpha\lambda_N)$.

This centers the range of $\alpha\lambda_j$ for $\lambda_1, \lambda_2, \ldots, \lambda_N$ around 1, thereby making $\max_{1\leq j\leq N} |1 - \alpha\lambda_j|$ as close to 0 as possible.

Thus, $\alpha = \frac{2}{\lambda_1+\lambda_N}$ minimizes $\rho = \frac{\lambda_N-\lambda_1}{\lambda_1+\lambda_N} = \frac{\kappa-1}{\kappa+1} < 1$ where $\kappa = \frac{\lambda_N}{\lambda_1}$.

1.6.4 Part iv

We can perform Part iii for bounded eigenvalues, just in a less optimal manner.
Given $0 < c \leq \lambda_1 \leq \lambda_N \leq C < \infty$, we can choose the (potentially sub-optimal) $\alpha = \frac{2}{c+C}$.

Again notice that $1 - \alpha c = \alpha C - 1$, centering our ρ estimation, $\hat{\rho} = 1 - \alpha\hat{\lambda}_j$ about 0 for the range $\hat{\lambda}_j \in [c, C]$.
It is directly evident that, if $\lambda_1 > c$ and $\lambda_N < C$, then $\rho < \hat{\rho}$.
Also notice that, since $C \geq c > 0$, $\hat{\rho} = \frac{C-c}{C+c} < 1$.

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

1.7 Part g

1.7.1 Part i

In the CG algorithm, we define

$$p_0 = r_0$$
$$w_n = Ap_n$$
$$r_n = r_{n-1} - \alpha_{n-1} w_{n-1}$$

Then, through substitution, we can say

$$w_0 = Ar_0$$
$$r_1 = r_0 - \alpha_0 w_0 = r_0 - \alpha_0 A r_0$$

1.7.2 Part ii

In the CG algorithm, we define

$$p_n = r_n + \beta_{n-1} p_{n-1}$$
$$w_n = Ap_n$$
$$r_n = r_{n-1} - \alpha_{n-1} w_{n-1} \text{ for } 1 \leq n \leq n^* - 1$$

Then,

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

From our givens, we know

$$w_{n-1} = -\frac{r_n - r_{n-1}}{\alpha_{n-1}}$$

Then,

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

for $1 \leq n \leq n^* - 1$.

1.7.3 Part iii

We have

$$r_1 = r_0 - \alpha_0 A r_0$$
$$r_{n+1} = r_n - \alpha_n A r_n + \frac{\alpha_n \beta_{n-1}}{\alpha_{n-1}} (r_n - r_{n-1})$$
$$\beta_{n-1} = \frac{r_n^T r_n}{r_{n-1}^T r_{n-1}}$$
$$\gamma_0 = \frac{1}{\alpha_0}$$
$$\gamma_n = \frac{1}{\alpha_n} + \frac{\beta_{n-1}}{\alpha_{n-1}}$$
$$\delta_n = \frac{\sqrt{\beta_n}}{\alpha_n} \text{ for } 1 \leq n \leq n^* - 1$$

Consider,

$$\begin{aligned}
Aq_0 &= \gamma_0 q_0 - \delta_0 q_1 \\
A\frac{r_0}{||r_0||} &= \frac{1}{\alpha_0}\frac{r_0}{||r_0||} - \frac{\sqrt{\beta_0}}{\alpha_0}\frac{r_1}{||r_1||} \\
\alpha_0 A r_0 \frac{1}{||r_0||} &= \frac{r_0}{||r_0||} - \sqrt{\beta_0}\frac{r_1}{||r_1||}
\end{aligned}$$

Using our equation from Part i

$$\begin{aligned}
(r_0 - r_1)\frac{1}{||r_0||} &= \frac{r_0}{||r_0||} - \sqrt{\beta_0}\frac{r_1}{||r_1||} \\
\frac{r_1}{||r_0||} &= \sqrt{\frac{r_1^T r_1}{r_0^T r_0}}\frac{r_1}{||r_1||} \\
\frac{r_1}{||r_0||} &= \frac{||r_1||}{||r_0||}\frac{r_1}{||r_1||} \\
\frac{r_1}{||r_0||} &= \frac{r_1}{||r_0||}
\end{aligned}$$

Then, consider

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

$$\alpha_n A r_n + \frac{\alpha_n\beta_{n-1}}{\alpha_{n-1}}(r_n - r_{n-1}) = r_n - r_{n+1}$$

$$r_{n+1} = r_n - \alpha_n A r_n - \frac{\alpha_n\beta_{n-1}}{\alpha_{n-1}}(r_n - r_{n-1})$$

We know this to be true from Part ii. Thus,

$$\begin{aligned}
Aq_0 &= \gamma_0 q_0 - \delta_0 q_1 \\
Aq_n &= -\delta_{n-1}q_{n-1} + \gamma_n q_n - \delta_n q_{n+1}
\end{aligned}$$

for $1 \leq n \leq n^* - 1$.

1.7.4 Part iv

Consider

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

Notice, by our conclusion from Part iii,

$$AQ_n = [Aq_0 \quad Aq_1 \quad \dots \quad Aq_{n-1}]$$

$$\begin{aligned}
Q_n T_n &= [\gamma_0 q_0 - \delta_0 q_1 \quad -\delta_0 q_0 + \gamma_1 q_1 - \delta_1 q_2 \quad \dots \quad -\delta_{n-2}q_{n-2} + \gamma_{n-1}q_{n-1}] \\
&= [Aq_0 \quad Aq_1 \quad \dots \quad Aq_{n-1} + \delta_{n-1}q_n] \\
&= [Aq_0 \quad Aq_1 \quad \dots \quad Aq_{n-1}] + \delta_{n-1}q_n e_n^T
\end{aligned}$$

Thus,

$$Q_n T_n - \delta_{n-1} q_n e_n^T = [Aq_0 \quad Aq_1 \quad \dots \quad Aq_{n-1}] = AQ_n$$

1.7.5 Part v

Take our conclusion from Part iv,

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

We can manipulate this. Keep in mind that Q is orthogonal and $q_n \notin Q$.

$$\begin{aligned}
AQ_n &= Q_nT_n - \delta_{n-1}q_ne_n^T \\
Q_n^TAQ_n &= Q_n^TQ_nT_n - Q_n^T\delta_{n-1}q_ne_n^T \\
Q_n^TAQ_n &= T_n - \delta_{n-1}Q_n^Tq_ne_n^T \\
Q_n^TAQ_n &= T_n
\end{aligned}$$

2 Problem B

Define

$$f(x) = e^{-400(x-.5)^2}$$

We can determine the derivative of f

$$\frac{\delta f}{\delta x} = -800(x-.5)e^{-400(x-.5)^2}$$

For any two sample points x_j, x_{j+1}, we know that, when $\frac{\delta f}{\delta x} = \frac{x_{j+1}-x_j}{2}$ for $x_j < x < x_{j+1}$, [2] $f(x)$ is the furthest from its linear interpolant.

Guessing x for $\frac{\delta f}{\delta x} = \frac{x_{j+1}-x_j}{2}$ through a search algorithm, we can construct an upper bound of the error, $e \geq |f(x) - f(\hat{x})|$ based on the accuracy of our estimation $\hat{x} = x + \epsilon$. Observe,

$$\begin{aligned}
&|e^{-400(x-.5)^2} - e^{-400(x+\epsilon-.5)^2}| \\
&= |e^{-400(x-.5)^2} - e^{-400(x-.5)^2}e^{-400(\epsilon^2+\epsilon(x-.5)}| \\
&= |e^{-400(x-.5)^2}\left(1 - e^{-400(\epsilon^2+\epsilon(x-.5)}\right)| \\
&\leq |1 - e^{-400(\epsilon^2 \pm .5\epsilon)}|
\end{aligned}$$

Note that we have an upper bound on the magnitude of ϵ when using a search algorithm such as binary search. Thus, we can calculate $e = |1 - e^{-400(\epsilon^2 \pm .5\epsilon)}| \geq |f(x) - f(\hat{x})|$.

So, we can take $N+1$ samples of f and create a linear interpolant. Then, for each interval (x_j, x_{j+1}), an upper bound of the uniform norm can be found $|f(\hat{x}) - lin(\hat{x})| + e$. By the definition of the uniform norm, the largest of these will then be equivalent to the uniform norm for $x \in [0, 1]$.

Doing this in *MatLab* provides $N = 100$ for ϵ sufficiently small.

3 Problem C

3.1 Part a

3.1.1 Main Method

For a second-order centered approximation of u_{tt}, we can say [3]

$$u_{tt} = \frac{u_{t-1} - 2u_t + u_{t+1}}{h_t^2} + O(h^2) \tag{7}$$

[2] Strictly less than and less than or equal to provide the same result since, at x_j, x_{j+1}, the linear interpolant is, by definition, equivalent to the original function.

[3] The subscript of u, if not indicating a derivative, represents its offset assuming all non-present variables are fixed. For example, $u_{t+1} = u(x, y, t+1)$.

Then, keeping in mind that $h_x = h_y$, for a 5-point second-order approximation of the laplacian,

$$\Delta u = u_{xx} + u_{yy} = \frac{-4u_{x,y} + u_{x-1,y} + u_{x+1,y} + u_{x,y-1} + u_{x,y+1}}{h_{xy}^2} + O(h^2)$$

Using the wave equation, we can come up with an approximate solution for $u_{x,y,t+1}$.

$$\frac{u_{t-1} - 2u_t + u_{t+1}}{h_t^2} \approx \frac{-4u_{x,y} + u_{x-1,y} + u_{x+1,y} + u_{x,y-1} + u_{x,y+1}}{h_{xy}^2}$$

$$u_{t+1} \approx 2u_t - u_{t-1} + \frac{h_t^2}{h_{xy}^2}\left(-4u_{x,y} + u_{x-1,y} + u_{x+1,y} + u_{x,y-1} + u_{x,y+1}\right)$$

In general,

$$u_{t+1} \approx 2u_t - u_{t-1} + h_t^2 \Delta u \tag{8}$$

For the boundaries of x, y, however, a different second-order derivative approximation may need to be used to calculate Δu. We will use a second-order one-sided second derivative approximation. As an example,

$$u''_{x=0} = \frac{1}{h_x^3}\left(2u_0 - 5u_1 + 4u_2 - u_3\right)$$

In practice, however, these approximations are prone to blowing up.

3.1.2 Boundary Conditions

As for the boundary conditions, we want

$$u(x, y, 0) = 0$$
$$u_t(x, y, 0) = f(x)f(y)$$

From the second equation, we can derive the following second-order forward derivative approximation

$$\frac{1}{2h_t}(u_1 - u_{-1}) \approx f(x)f(y)$$

We can then say

$$u_{-1} \approx u_1 - 2f(x)f(y)h_t$$

3.1.3 Final Scheme

So, we have our scheme. First, implement the Neumann boundary condition representing initial velocity. After, calculate the laplacian for a time step. Then, calculate a new value for all $u_{x,y,c}$ using 8. Repeat. In practice, you may want to add a scaling value c^2 before the addition of the laplacian.

The scheme initiates as a 3D matrix $\mathbb{R}^{(N+1)\times(N+1)\times(N_t+1)}$ filled with all zeros. The non-zero values come from our Neumann boundary condition above.

The error can be seen below in figure 1 along with a couple snapshots of the scheme in practice in figure 2. The error was calculated against a fine estimation of the actual solution using the scheme with $N = 1000$.

Figure 1: Log-log h vs Error

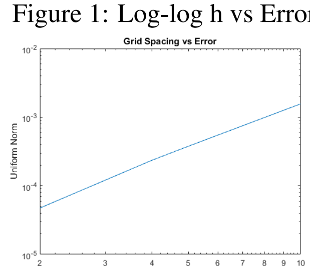
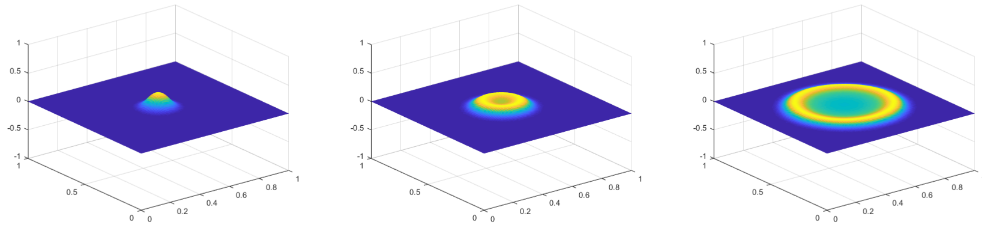


Figure 2: 2D wave FD scheme at $t = 30, 70, 150$ where $N = 100$. Here, $c = .2$.



3.2 Part b

We are given

$$y''(t) = \gamma y$$

Note,

$$y_{t-1} - 2y_t + y_{t+1} = \lambda(\Delta t)^2 y_t$$

We can construct the stability polynomial

$$\begin{aligned}
\pi(\xi, z) &= \rho(\xi) - \lambda(\Delta t)^2\sigma(\xi) \\
&= \xi^2 - 2\xi + 1 - \lambda(\Delta t)^2\xi \\
&= \xi^2 - (2 + \lambda(\Delta t)^2)\xi + 1 \\
&= \xi^2 - (2 + z)\xi + 1
\end{aligned}$$

Solving for the roots provides

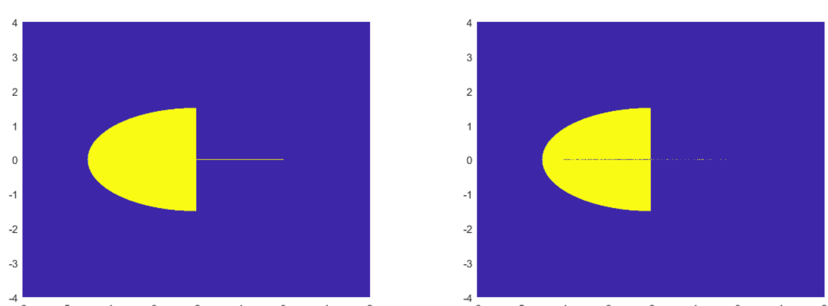
$$\begin{aligned}
\xi &= \frac{2 + z \pm \sqrt{(2+z)^2 - 4}}{2} \\
&= \frac{2 + z \pm \sqrt{z^2 + 4z}}{2}
\end{aligned}$$

The ODE is absolutely stable if and only if $|r| < 1$ for all roots r excepting $|r| = 1$ for simple roots r. So,

$$\left| \frac{2 + z \pm \sqrt{z^2 + 4z}}{2} \right| \leq 1$$

This can be seen below in figure 3. Notice the scattered line along the real axis. This is caused by some real values of z providing both roots equal to 1 – which does not imply absolute stability.

Figure 3: Possible $z = \lambda(\Delta t)^2$ highlighted in yellow



3.3 Part c

We can apply MOL to the problem

$$u_{tt} = \frac{-4u_{x,y} + u_{x-1,y} + u_{x+1,y} + u_{x,y-1} + u_{x,y+1}}{h_{xy}^2} + O(h^2)$$

Then, we know

$$U''(t) = AU(t) + g(t) \text{ where } A = \begin{bmatrix} 0 & 1 & \\ 1 & \ddots & \ddots \\ & \ddots & \end{bmatrix} \oplus \begin{bmatrix} -4 & 1 & \\ 1 & \ddots & \ddots \\ & \ddots & \end{bmatrix}$$

A will have no complex eigenvalues. Additionally, all of the eigenvalues will be strictly negative such that there is a value $0 < c \leq -\lambda$ that holds for all eigenvalues λ no matter the size of the matrix. This was also shown on the previous homework using the matrix above, but negative. Given our answer from Part b, our method from Part a can be absolutely stable.

As for the CFL, we will have $\frac{2h_t}{h_{xy}}$ which implies we would need $h_t < \frac{1}{2}h_{xy}$ since a 5-point stencil is used.

3.4 Part d

We can mainly follow along with [1].

We start by assuming

$$U_j^n = e^{ijh\xi}$$
$$U_j^{n+1} = g(\xi)e^{ijh\xi}$$

Then, using our final equation from Part a, 8,

$$\begin{aligned}
g(\xi)e^{ikjh\xi} &= 2e^{ikjh\xi} - g(\xi)^{-1}e^{ikjh\xi} + \frac{h_t^2}{h^2}\left(e^{i(k+1)jh\xi} + e^{i(k-1)jh\xi} + e^{ik(j-1)h\xi} + e^{ik(j+1)h\xi} - 4e^{ikjh\xi}\right) \\
g(\xi) &= 2 - g(\xi)^{-1} + \frac{h_t^2}{h^2}\left(e^{ijh\xi} + e^{-ijh\xi} + e^{-ikh\xi} + e^{ikh\xi} - 4\right) \\
0 &= g(\xi)^2 - g(\xi)\left(2+z\right) + 1 \text{ for } z = \frac{h_t^2}{h^2}\left(e^{ijh\xi} + e^{-ijh\xi} + e^{-ikh\xi} + e^{ikh\xi} - 4\right) \\
g(\xi) &= \frac{2 + z \pm \sqrt{((2+z)^2 - 4)}}{2} \text{ for } z = \frac{h_t^2}{h^2}\left((e^{\frac{1}{2}ijh\xi} - e^{-\frac{1}{2}ijh\xi})^2 + (e^{\frac{1}{2}ikh\xi} - e^{-\frac{1}{2}ikh\xi})^2\right) \\
g(\xi) &= \frac{2 + z \pm \sqrt{(z^2 + 4z)}}{2} \text{ for } z = -4\frac{h_t^2}{h^2}\left(\sin^2(\frac{1}{2}jh\xi) + \sin^2(\frac{1}{2}kh\xi)\right)
\end{aligned}$$

z will be negative and real. The magnitude depends on various factors such as h_t, h_{xy}, i, j. This does make it appear, however, that the method from Part a can be absolutely stable. This is the same conclusion as in Part c.

The CFL also shares the same form as in Part c with one key addition: There is a lower bound $-8\frac{h_t^2}{h_{xy}^2}$.

3.5 Part e

We have the numerical method

$$U^{n+1} = 2U - U^{n-1} + \frac{h_t^2}{h^2}\left(U_{j-1} + U_{j+1} + U_{k-1} + U_{k+1} - 4U\right)$$

Expanding this provides

$$\begin{aligned}
2U + U_{tt}h_t^2 + U_{tttt}\frac{h_t^4}{12}\cdots &= 2U + \frac{h_t^2}{h^2}\left(2U + U_{xx}h^2 + U_{xxxx}\frac{h^4}{12}\cdots + 2U + U_{yy}h^2 + U_{yyyy}\frac{h^4}{12}\cdots - 4U\right) \\
U_{tt} + U_{tttt}\frac{h_t^2}{12}\cdots &= U_{xx} + U_{xxxx}\frac{h^2}{12}\cdots + U_{yy} + U_{yyyy}\frac{h^2}{12}\cdots
\end{aligned}$$

Using our original PDE, $U_{tt} = U_{xx} + U_{yy}$ and

$$U_{tttt}\frac{h_t^2}{12}\cdots = U_{xxxx}\frac{h^2}{12}\cdots + U_{yyyy}\frac{h^2}{12}\cdots$$

We then obtain a new PDE

$$U_{tttt}h_t^2 = U_{xxxx}h^2 + U_{yyyy}h^2$$

Let's apply the 2D fourier transform

$$\begin{aligned}
\hat{U}_{tttt}h_t^2 &= \int_\infty \int_\infty \left(U_{xxxx}h^2 + U_{yyyy}h^2\right) e^{-i(\xi x+\omega y)} dydx \\
\hat{U}_{tttt}h_t^2 &= \int_\infty \left(i\omega U_{xxxx}h^2 e^{-i(\xi x+\omega y)} + \omega^4 \hat{U}(\omega)h^2\right) dx \\
\hat{U}_{tttt} &= \frac{h^2}{h_t^2}\left(i\omega\xi^4\hat{U}(\xi) + i\xi\omega^4\hat{U}(\omega)\right)
\end{aligned}$$

4 Problem D

Say we have an ODE

$$u_{tt} = u_t + \beta$$

Then, we can write

$$\begin{aligned}
\frac{1}{\Delta t}\left(u^{n+1} - 2u^n + u^{n-1}\right) &= u_t + \beta \\
u^{n+1} &= 2u^n - u^{n-1} + \Delta t u_t + \beta
\end{aligned}$$

as a second-order accurate approximation of the second derivative with respect to time.

We can assign a finite difference scheme to u_t such that

$$U^{n+1} = BU^n + \beta$$

Which is Lax-Richtmyer stable since $||B(k)^n|| < C_T$ for all $k > 0$, $nk \leq T$ and, therefore, converges.

Thus, for an ODE of the form $u_{tt} = u_t + \beta$, there exists some convergent method for approximating u over time.

References

[1] Randall J. LeVeque. *Diffusion Equations and Parabolic Problems*, chapter 9, pages 181–200.