MATH 714 HW 2

Lewis Gross

Worked with: Varun Gudibanda, Haley Colgate, Sourav Pal, Samuel Jackson

All code can be found on Github

November 1, 2020

A (40 Points)

(a) (2 points) Suppose that $w = \{w_1, w_2, ..., w_n\}$ are orthogonal. Show that if v belongs to the span of w, then

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

We can write v as a linear combination of the vectors of w since v belongs to the span of w, thus

$$v = \sum_{j=1}^{n} c_j w_j \tag{2}$$

To find the coefficients c_j, we will use the orthogonality of the basis vectors of w. Let's take the inner product with w_m

$$\langle v, w_m \rangle = \langle \sum_{j=1}^{n} c_j w_j, w_m \rangle = \sum_{j=1}^{n} c_j \langle w_j, w_m \rangle \tag{3}$$

Since the basis is orthogonal, then $\langle w_j, w_m \rangle = ||w_m||^2 \delta_{jm}$, now

$$\langle v, w_m \rangle = \sum_{j=1}^{n} c_j \langle w_j, w_m \rangle = \sum_{j=1}^{n} c_j ||w_m||^2 \delta_{jm} \tag{4}$$

The only term in the sum that stays is when $j = m$

$$\langle v, w_j \rangle = c_j ||w_j||^2 \qquad \text{THUS} \qquad c_j = \frac{\langle v, w_j \rangle}{||w_j||^2} \tag{5}$$

(b) (4 points) Recall that $p_0 = r_0$ and

$$p_n = r_n - \sum_{j=0}^{n-1} \frac{\langle r_n, p_j \rangle_A}{||p_j||_A^2} p_j \qquad \text{FOR} \qquad 1 \le n \le n^* - 1 \tag{6}$$

where $n^* \le N$ is the number of iterations to converge and $A \in \mathbb{R}^{N \times N}$

i. Explain why the number of iterations to convergence, n^* may be strictly smaller than N

If your solution lives in one of the Kyrlov subspaces, as soon as the basis for the Krylov space spans the space your solution lives in, then we will have an exact solution in that step. This is along the same justification that CG will converge exactly when the full Krylov space is attained (step N), as the error polynomial will be an exact match and the error will be exactly zero.

ii. Prove by induction on n that

$$\langle p_n, p_j \rangle_A = 0 \qquad \text{FOR} \qquad 0 \le j < n \le n^* - 1 \tag{7}$$

First, we will show the base case of $n = 1$, this implies that $j = 0$, since j is an integer and $0 \le j < 1$. Now we wish to show that

$$\langle p_1, p_0 \rangle_A = 0 \tag{8}$$

We know that

$$p_0 = r_0 \qquad \text{AND} \qquad p_1 = r_1 - \frac{\langle r_1, p_0 \rangle_A}{||p_0||_A^2} p_0 \tag{9}$$

Now to show

$$\langle p_1, p_0 \rangle_A = \langle r_1 - \frac{\langle r_1, p_0 \rangle_A}{||p_0||_A^2} p_0, p_0 \rangle_A = 0 \tag{10}$$

$$\langle p_1, p_0 \rangle_A = \langle r_1, p_0 \rangle_A - \langle \frac{\langle r_1, p_0 \rangle_A}{||p_0||_A^2} p_0, p_0 \rangle_A = 0 \tag{11}$$

$$\langle p_1, p_0\rangle_A = \langle r_1, p_0\rangle_A - \frac{\langle r_1, p_0\rangle_A}{\|p_0\|_A^2}\langle p_0, p_0\rangle_A = 0 \tag{12}$$

$$\langle p_1, p_0\rangle_A = \langle r_1, p_0\rangle_A - \frac{\langle r_1, p_0\rangle_A}{\|p_0\|_A^2}\|p_0\|_A^2 = \langle r_1, p_0\rangle_A - \langle r_1, p_0\rangle_A = 0 \tag{13}$$

And thus the base case has been shown. Now, assuming that $\langle p_n, p_j\rangle_A = 0$, we wish to show

$$IP = \langle p_{n+1}, p_j\rangle_A = 0 \tag{14}$$

from (6)

$$p_{n+1} = r_{n+1} - \sum_{i=0}^{n} \frac{\langle r_{n+1}, p_i\rangle_A}{\|p_i\|_A^2} p_i \tag{15}$$

Using this, our inner product for $n+1$ becomes

$$IP = \langle r_{n+1} - \sum_{i=0}^{n} \frac{\langle r_{n+1}, p_i\rangle_A}{\|p_i\|_A^2} p_i, p_j\rangle_A = \langle r_{n+1}, p_j\rangle - \sum_{i=0}^{n} \frac{\langle r_{n+1}, p_i\rangle_A}{\|p_i\|_A^2}\langle p_i, p_j\rangle \stackrel{?}{=} 0 \tag{16}$$

Using the induction hypothesis, we know that $\langle p_n, p_j\rangle = \|p_n\|^2\delta_{nj}$, so

$$IP = \langle r_{n+1}, p_j\rangle - \sum_{i=0}^{n} \frac{\langle r_{n+1}, p_i\rangle_A}{\|p_i\|_A^2}\|p_i\|_A^2\delta_{ij} = \langle r_{n+1}, p_j\rangle - \sum_{i=0}^{n}\langle r_{n+1}, p_i\rangle_A\delta_{ij} \tag{17}$$

Only one term in the sum stays , when $i = j$, thus

$$IP = \langle r_{n+1}, p_j\rangle - \langle r_{n+1}, p_j\rangle_A \equiv 0 \tag{18}$$

And so, we have shown the induction step to be true.

(c) (4 points) Given that $A \in \mathbb{R}^{N\times N}$ is a symmetric, positive-definite matrix then $\mathbb{R}^N$ has an orthonormal basis of eigenvectors $\phi_1, \phi_2, ..., \phi_N$:

$$A\phi_n = \lambda_n\phi_n \qquad \text{AND} \qquad \langle \phi_n, \phi_j\rangle = \delta_{nj} \tag{19}$$

Assuming we order the eigenvalues so that $\lambda_1 \le \lambda_2 \le ... \le \lambda_N$, prove the following for all $v, w \in \mathbb{R}^N$

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

Since $v, w \in \mathbb{R}^N$, we can express them as a linear combination of our orthonormal basis vectors

$$v = \sum_{n=1}^{N} c_n\phi_n \qquad \text{AND} \qquad w = \sum_{m=1}^{N} d_m\phi_m \tag{20}$$

Using this and a dummy variable for the second sum,

$$\langle Av, w\rangle = \langle \sum_{n=1}^{N} c_nA\phi_n, \sum_{m=1}^{N} d_m\phi_m\rangle = \langle \sum_{n=1}^{N} c_n\lambda_n\phi_n, \sum_{m=1}^{N} d_n\phi_n\rangle = \sum_{n=1}^{N}\sum_{m=1}^{N} \lambda_n\langle c_n\phi_n, d_n\phi_n\rangle \tag{21}$$

Now, we can use that

$$c_n = \langle v, \phi_n\rangle \qquad \text{AND} \qquad d_m = \langle w, \phi_m\rangle \tag{22}$$

Inserting these gives,

$$\langle Av, w\rangle = \sum_{n=1}^{N}\sum_{m=1}^{N} \lambda_n\langle\langle v, \phi_n\rangle\phi_n, \langle w, \phi_m\rangle\phi_m\rangle = \sum_{n=1}^{N}\sum_{m=1}^{N} \lambda_n\langle v, \phi_n\rangle\langle w, \phi_m\rangle\langle \phi_n, \phi_m\rangle \tag{23}$$

We know that the ϕ are orthonormal, so we can get

$$\langle Av, w\rangle = \sum_{n=1}^{N}\sum_{m=1}^{N} \lambda_n\langle v, \phi_n\rangle\langle w, \phi_m\rangle\delta_{mn} = \sum_{n=1}^{N} \lambda_n\langle v, \phi_n\rangle\langle w, \phi_n\rangle = \sum_{n=1}^{N} \lambda_n\langle v, \phi_n\rangle\langle \phi_n, w\rangle \tag{24}$$

ii. $\lambda_n > 0$ **for** $1 \le n \le N$

Since A is symmetric, positive-definite, we have

$$\langle Ax, x\rangle > 0 \qquad \text{FOR} \qquad x \in \mathbb{R}^N \tag{25}$$

Thus, we can allow $x = \phi_n$, and as such we get

$$\langle A\phi_n, \phi_n\rangle = \langle \lambda_n\phi_n, \phi_n\rangle = \lambda_n\langle \phi_n, \phi_n\rangle = \lambda_n||\phi_n||^2 > 0 \tag{26}$$

We know that $||\phi_n||^2 > 0$, which means that $\lambda_n > 0$.

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

From (i) we know,

$$\langle Av, w\rangle = \sum_{n=1}^{N} \lambda_n\langle v, \phi_n\rangle\langle \phi_n, w\rangle \qquad \text{THUS} \qquad \langle Av, v\rangle = \sum_{n=1}^{N} \lambda_n(\langle v, \phi_n\rangle)^2 \tag{27}$$

Because we order the eigenvalues, we know that

$$\sum_{n=1}^{N} \lambda_1(\langle v, \phi_n\rangle)^2 \le \langle Av, v\rangle = \sum_{n=1}^{N} \lambda_n(\langle v, \phi_n\rangle)^2 \le \sum_{n=1}^{N} \lambda_N(\langle v, \phi_n\rangle)^2 \tag{28}$$

Now, we need another component. By definition and also using our basis

$$||u||^2 = \langle u, u\rangle \qquad \text{AND} \qquad u = \sum_{n=1}^{N} \langle u, \phi_n\rangle\phi_n \tag{29}$$

Using these,

$$||u||^2 = \langle \sum_{n=1}^{N} \langle u, \phi_n\rangle\phi_n, \sum_{m=1}^{N} \langle u, \phi_m\rangle\phi_m\rangle = \sum_{n=1}^{N}\sum_{m=1}^{N} \langle \langle u, \phi_n\rangle\phi_n, \langle u, \phi_m\rangle\phi_m\rangle = \sum_{n=1}^{N}\sum_{m=1}^{N} \langle u, \phi_n\rangle\langle u, \phi_m\rangle\langle \phi_n, \phi_m\rangle \tag{30}$$

Using orthonormality,

$$||u||^2 = \sum_{n=1}^{N}\sum_{m=1}^{N} \langle u, \phi_n\rangle\langle u, \phi_m\rangle\langle \phi_n, \phi_m\rangle = \sum_{n=1}^{N}\sum_{m=1}^{N} \langle u, \phi_n\rangle\langle u, \phi_m\rangle\delta_{mn} = \sum_{n=1}^{N} (\langle u, \phi_n\rangle)^2 \tag{31}$$

Now, we can use $||u||^2 = \sum_{n=1}^{N}(\langle u, \phi_n\rangle)^2$ in (32), giving the desired result.

$$\lambda_1||v||^2 \le \langle Av, v\rangle \le \lambda_N|v||^2 \tag{32}$$

iv. $||Av|| \le \lambda_N||v||$

We start by squaring both side $||Av||^2 \le \lambda_N^2||v||^2$. We will try and show the squared version to be true

$$||Av||^2 = \langle Av, Av\rangle = \langle A^T Av, v\rangle = \langle A^2v, v\rangle \tag{33}$$

From before, we have

$$\langle Bv, v\rangle \le \lambda_{max}(B)||v||^2 \tag{34}$$

Applying this

$$\langle A^2v, v\rangle \le \lambda_{max}(A^2)||v||^2 \tag{35}$$

We know that if $A\phi = \lambda\phi$, then $A^2\phi = \lambda^2\phi$ and the max eigenvalue of A^2 will then be λ_N^2

$$||Av||^2 = \langle A^2v, v\rangle \le \lambda_N^2||v||^2 \qquad \text{THUS} \qquad ||Av|| \le \lambda_N||v|| \tag{36}$$

And we've now shown the desired outcome.

(d) (6 points) Deduce from the update formulas for p_n,w_n, and r_n that

$$p_{n+1} = (1+\beta_n)p_n - \alpha_n A p_n - \beta_{n-1}p_{n-1} \qquad \text{FOR} \qquad 1 \le n \le n^* - 2 \tag{37}$$

We have that $p_0 = r_0 = b - Ax_0$, and then

$$w_{k-1} = Ap_{k-1} \tag{38}$$

$$r_k = r_{k-1} - \alpha_{k-1}w_{k-1} \tag{39}$$

$$p_k = r_k + \beta_{k-1}p_{k-1} \tag{40}$$

First, we insert (39) into (40) to get

$$p_k = r_{k-1} - \alpha_{k-1}w_{k-1} + \beta_{k-1}p_{k-1} \tag{41}$$

We desire $k = n+1$, so doing this gives

$$p_{n+1} = r_n - \alpha_n w_n + \beta_n p_n \tag{42}$$

Next, we use (38) for w_n, giving

$$p_{n+1} = r_n - \alpha_n A p_n + \beta_n p_n \tag{43}$$

Last, we use (40) for $k = n$, but rearranged and solved for r_n

$$p_{n+1} = p_n - \beta_{n-1}p_{n-1} - \alpha_n A p_n + \beta_n p_n = (1+\beta_n)p_n - \alpha_n A p_n - \beta_{n-1}p_{n-1} \tag{44}$$

And we have done it!

(e) (6 points) Prove that if $A \in \mathbb{R}^{N\times N}$ is non-singular, then A^N is a linear combination of $\{I, A, A^2, ..., A^{N-1}\}$. Use the Cayley-Hamilton theorem.

The Cayley-Hamilton theory states that for a non-singular matrix $A \in \mathbb{R}^{N\times N}$, with characteristic polynomial $p(\lambda) = \det(A - \lambda I)$, the matrix equation $P(A) = 0$. The characteristic polynomial has coefficients $\{c_n\}$,

$$p(\lambda) = \lambda^n + c_{n-1}\lambda^{n-1} + c_{n-2}A^{n-2} + ... + c_2 A^2 + c_1\lambda + (-1)^n \det(A) \tag{45}$$

Now, the theorem tells us that that $p(A) = 0$, giving

$$p(A) = A^n + c_{n-1}A^{n-1} + c_{n-2}A^{n-2} + ... + c_2 A^2 + c_1 A + (-1)^n \det(A) I_n = 0 \tag{46}$$

We can solve for A^n, and this gives that A is a linear combination of $\{I, A, A^2, ..., A^{N-1}\}$

$$A^n = c_{n-1}A^{n-1} + c_{n-2}A^{n-2} + ... + c_2 A^2 + c_1 A + (-1)^n \det(A) I_n \tag{47}$$

(f) (8 points) For any $\alpha \neq 0$, the linear equation $Au = f$ is equivalent to

$$u = u + \alpha(f - Au) \tag{48}$$

In this context, the Richardson iteration is defined by

$$u_{n+1} = u_n + \alpha(f - Au_n) \tag{49}$$

i. Show that the error $e_n = u_n - u$ satisfies $e_{n+1} = (I - \alpha A)e_n$

If $e_n = u_n - u$, then $e_{n+1} = u_{n+1} - u$. They can be re-arranged to solve for u_n and u_{n+1}. Inserting these,

$$(e_{n+1} - u) = (e_n - u) + \alpha(f - A(e_n - u)) \tag{50}$$

We get some cancellation, also recall that $Au = f$,

$$e_{n+1} = e_n + \alpha(f - Ae_n - f) = e_n - \alpha A e_n = (I - \alpha A)e_n \tag{51}$$

ii. Deduce that $\|e_{n+1}\| \leq \rho \|e_n\|$, where the error reduction factor is

$$\rho = \max_{1 \leq j \leq N} |1 - \alpha \lambda_j| \tag{52}$$

It follows that $\|e_{n+1}\| \leq \rho^n \|e_0\|$, so if $\rho < 1$, then $e_n \to 0$ and hence the Richardson iteration u_n converges to the solution u.

Starting with the result

$$e_{n+1} = (I - \alpha A) e_n \qquad \text{THUS} \qquad \|e_{n+1}\|_2 = \|(I - \alpha A) e_n\|_2 \leq \|(I - \alpha A)\|_2 \, \|e_n\|_2 \tag{53}$$

Now, we use that the two norm is the max eigenvalue (for SPD matrices)

$$\rho = \|(I - \alpha A)\|_2 = \max \lambda (I - \alpha A) = \max |1 - \alpha \lambda(A)| = \max_{i \leq j \leq N} |1 - \alpha \lambda_j| \tag{54}$$

iii. Prove that ρ is minimized by choosing

$$\alpha = \frac{2}{\lambda_1 + \lambda_N} \tag{55}$$

in which case

$$\rho = \frac{\lambda_N - \lambda_1}{\lambda_N + \lambda_1} = \frac{\kappa - 1}{\kappa + 1} < 1 \qquad \text{WHERE} \qquad \kappa = \frac{\lambda_N}{\lambda_1} \tag{56}$$

To find the α that minimizes ρ, we could search all the eigenvalues for the optimal one. Assuming we know nothing about the matrix, We know that this will need to occur at either the largest or smallest eigenvalue (depending on the sign it would either be the largest or smallest eigenvalue). Since we have ordered them in increasing fashion. Thus

$$\rho = \max_{i \leq j \leq N} |1 - \alpha \lambda_j| = \max\{|1 - \alpha \lambda_1|, |1 - \alpha \lambda_N|\} \tag{57}$$

Define α_0 as the minimizer of ρ, we desire $\alpha_0 = \text{argmin}_\alpha \max\{|1 - \alpha \lambda_1|, |1 - \alpha \lambda_N|\}$. It will help to look at a test function similar to this. Define $t(x, y) = \max\{x, y\}$, it is clear that to minimize t, we need $x = y$, otherwise the max will return the larger of the 2. This implies that we can find α_0, by setting equal the two possible maxima of ρ. This gives $|1 - \alpha \lambda_1| = |1 - \alpha \lambda_N|$, which creates two cases

$$1 - \alpha \lambda_1 = 1 - \alpha \lambda_N \qquad \text{OR} \qquad 1 - \alpha \lambda_1 = -(1 - \alpha \lambda_N) \tag{58}$$

The first case can only be true when $\lambda_1 = \lambda_N$, but this would require a matrix with all equal eigenvalues, which is not an assumption we can make in general for A. The other case gives the desired result as,

$$2 = \alpha(\lambda_1 + \lambda_N) \qquad \text{THUS} \qquad \alpha = \frac{2}{\lambda_1 + \lambda_N} \tag{59}$$

Now to show the desired result for ρ

$$\rho = \left| 1 - \frac{2}{\lambda_1 + \lambda_N} \lambda_1 \right| \qquad \text{AND} \qquad \rho = \left| 1 - \frac{2}{\lambda_1 + \lambda_N} \lambda_N \right| \tag{60}$$

We will use the first expression (though by design thte two are equal),

$$\rho = \left| \frac{\lambda_1 + \lambda_N}{\lambda_1 + \lambda_N} - \frac{2}{\lambda_1 + \lambda_N} \lambda_1 \right| = \left| \frac{\lambda_1 + \lambda_N - 2\lambda_1}{\lambda_1 + \lambda_N} \right| = \left| \frac{\lambda_N - \lambda_1}{\lambda_N + \lambda_1} \right| = \frac{\lambda_N - \lambda_1}{\lambda_N + \lambda_1} = \frac{\frac{\lambda_N}{\lambda_1} - 1}{\frac{\lambda_N}{\lambda_1} + 1} = \frac{\kappa - 1}{\kappa + 1} \tag{61}$$

The absolute value can be removed after the fourh equality above because explicitly λ_N is greater than λ_1 due to our ordering of eigenvalues. And we have shown the result

iv. Suppose that we do not know the exact values of the extremal eigenvalues, but only for some one-sided bounds

$$0 < c \leq \lambda_1 \leq \lambda_N \leq C < \infty \tag{62}$$

For the (sub-optimal) choice $\alpha = \frac{2}{c+C}$, show

$$\rho \leq \rho' = \frac{C-c}{C+c} = \frac{\kappa'-1}{\kappa'+1} < 1 \qquad \text{WHERE} \qquad \kappa' = \frac{C}{c} \tag{63}$$

First, we will show $\rho' < 1$. If

$$\frac{C-c}{C+c} < 1 \qquad \text{THEN} \qquad C-c < C+c \tag{64}$$

Which is true, as the C cancel on both sides and we get $-c < c$, which is also true because $0 < c < \infty$. Next, we will aim show that $\rho < \rho'$. We will start by assuming it is true and working backwards to try and get a true statement

$$\rho = \frac{\lambda_N - \lambda_1}{\lambda_N + \lambda_1} < \frac{C-c}{C+c} = \rho' \tag{65}$$

Cross multiplying gives

$$(\lambda_N - \lambda_1)(C+c) < (\lambda_N + \lambda_1)(C-c) \tag{66}$$

$$\lambda_N C - \lambda_1 C + \lambda_N c - \lambda_1 c < \lambda_N C + \lambda_1 C - \lambda_N c - \lambda_1 c \tag{67}$$

Some cancellation occurs giving,

$$-\lambda_1 C + \lambda_N c < \lambda_1 C - \lambda_N c \tag{68}$$

We can rearrange this to get

$$\lambda_N c < \lambda_1 C \qquad \text{OR} \qquad \frac{\lambda_N}{\lambda_1} < \frac{C}{c} \tag{69}$$

This second form does it, but we will show so by a small related example/lemma. If $a, b, x, y > 0$ and $a > b$ and $x < y$, then

$$\frac{a}{x} > \frac{b}{y} \tag{70}$$

We know this because $\frac{1}{x} > \frac{1}{y}$, and thus, the product of a and $\frac{1}{x}$ is greater than the product of b and $\frac{1}{y}$. Essentially, since both factors are strictly greater than the other two factors, the products will have the strictly greater than hold. Back to our problem, $C > \lambda_N$ and $c < \lambda_1$, so we by the lemma just demonstrated, the condition holds and the desired $\rho < \rho' < 1$ has been shown.

(g) (10 Points) Define the normalized CG residual

$$q_n = \frac{r_n}{||r_n||} \qquad \text{FOR} \qquad 0 \leq n \leq n^* - 1 \tag{71}$$

so that $\{q_0, ..., q_{n-1}\}$ is an orthonormal basis for the Krylov space $\mathcal{K}_n$

i. Show that $r_1 = r_0 - \alpha_0 A r_0$

ii. Show that

$$r_{n+1} = r_n - \alpha_n A r_n + \frac{\alpha_n \beta_{n-1}}{\alpha_{n-1}}(r_n - r_{n-1}) \qquad \text{FOR} \qquad 1 \leq n \leq n^* - 1 \tag{72}$$

iii. Define

$$\gamma_0 = \frac{1}{\alpha_0} \qquad \text{AND} \qquad \gamma_n = \frac{1}{\alpha_n} + \frac{\beta_{n-1}}{\alpha_{n-1}} \qquad \text{FOR} \qquad 1 \leq n \leq n^* - 1 \tag{73}$$

and

$$\delta_n = \frac{\sqrt{\beta_n}}{\alpha_n} \tag{74}$$

Deduce from parts i. and ii. that

$$Aq_0 = \gamma_0 q_0 - \delta_0 q_1 \tag{75}$$

$$Aq_n = -\delta_{n-1} q_{n-1} + \gamma_n q_n - \delta_n q_{n+1} \qquad \text{FOR} \qquad 1 \leq n \leq n^* - 1 \tag{76}$$

iv. Show that

$$AQ_n = Q_n T_n - \delta_{n-1} q_n e_n^T \tag{77}$$

where $Q = [q_0|q_1|...|q_{n-1}] \in \mathbb{R}^{N\times n}$ **is orthogonal**

$$T_N = matrix \tag{78}$$

is tridiagonal and $e_n = [0, 0, ..., 1]^T \in \mathbb{R}^n$

v. Deduce that $Q_n^T A Q_n = T_n$

I ran out of time. I am sorry to disappoint you.

B (10 points) Consider the function

$$f(x) = \exp(-400(x-0.5)^2) \qquad \text{FOR} \qquad x \in [0,1] \tag{79}$$

Sample it on a grid $x_j = (j-1)h$ **with** $h = \frac{1}{N-1}$ **and** $1 \le j \le N$ **for some** N **to be determined. Consider the linear interpolant of** f **computed from the** N **samples** $f(x_j)$**. Numerically, find the smallest value of** N **such that** f **differs from its linear interpolant by at most** 10^{-2} **in the uniform norm. MATLAB has interp1.m for 1D linear interpolation**

Define $\hat{f}(x)$ as the linear interpolant, which satisfies for $j \in [1, N]$

$$\frac{\hat{f}(x) - f(x_j)}{f(x_{j+1}) - f(x_j)} = \frac{x - x_j}{x_{j+1} - x_j} \qquad \text{OR} \qquad \hat{f}(x) = f(x_j) + \frac{f(x_{j+1}) - f(x_j)}{x_{j+1} - x_j}(x - x_j) \tag{80}$$

Define $g(x) = f(x) - \hat{f}(x)$, we want to pick N such that $||g(x)||_\infty < 10^{-2}$. There exists one point in each interval where the maximum of g occurs, to find this point, we will require $0 = g'(x_*)$

$$g'(x) = f'(x) - \hat{f}'(x) = (400 - 800x)\exp(-400(x-0.5)^2) - \frac{f(x_{j+1}) - f(x_j)}{x_{j+1} - x_j} \tag{81}$$

$$g'(x) = (400 - 800x)\exp(-400(x-0.5)^2) - \frac{f(x_{j+1}) - f(x_j)}{x_{j+1} - x_j} \tag{82}$$

So the x that gives the max difference occurs at $g'(x_*) = 0$. Recall, $x_{j+1} - x_j = h = \frac{1}{N-1}$, so

$$0 = (400 - 800x)\exp(-400(x-0.5)^2) - (N-1)(f(x_{j+1}) - f(x_j)) \tag{83}$$

This equation does not have a closed form for the root, but the root x_* can be computed numerically. We will compute X_* and $g(x_*)$ in every interval in our grid and then take the maximum $|g(x_*)|$ as the uniform norm error. We will repeat this process for increasing N until the max $|g(x_*)| < 10^{-2}$. Here is the code that computes N

```
tol = 10^(-2); error = tol  + 1; N = 10;
while(error>tol)
    N=N+1;
    h = 1/(N-1) ;
    xgrid = [0:h:1] ;
    errors = zeros(size(xgrid));
    for j = 1:length(xgrid)-1
        xguess = (xgrid(j)+xgrid(j+1))/2;
        xstar = fzero( @(x) gprime( x , xgrid(j+1) , xgrid(j) , N) ,
            xguess );
        interp_star = veloc(xgrid(j)) + (N-1)*(veloc(xgrid(j+1)) -veloc
            (xgrid(j))) *( xstar - xgrid(j) ) ;
```

```
        errors(j) = veloc(xstar) - interp_star;
    end
    error = max(abs(errors));
end
disp(['N is ',num2str(N)]);
```

It calls the grprime.m, to compute $g'(x)$, which is given by

```
function g = gprime(x,xhigh,xlow,N)
% x is the position, xhigh is the upper bound for the interval, xlow is
    the lower bound for the interval, N is the number of grid points in
    the mesh
    g  = (400-800*x).*exp(-400.*(x-0.5).^2 ) - (N-1).*(veloc(xhigh) -
       veloc(xlow));
end
```

The function veloc.m is the function f of this problem (to be used in C with the same name).

```
function f = veloc(x)
    f = exp(-400.*(x-0.5).^2);
end
```

The code resulted in $N = 101$.

C (50 Points)

Consider the 2D wave equation

$$u_{tt} = u_{xx} + u_{yy} \qquad \text{FOR} \qquad 0 \le x, y \le 1 \tag{84}$$

with homogeneous Dirichlet boundary conditions. Fix the initial conditions to be

$$u(x, y, 0) = 0 \qquad u_t(x, y, 0) = f(x)f(y) \tag{85}$$

where f was defined in B. Consider a spatial grid $\mathbf{x_j} = (x_{j_1}, x_{j_2}) = (j_1 h, j_2 h)$ with h small enough to resolve the initial condition, in the sense of problem B (i.e use the critical N in problem B)

(a) (30 points) Implement and test the "simplest" numerical method, which uses the 3-point formula for the second derivative in time, and the 5-point Laplacian at a time t_n. It results in a two-step method. Explain how you initialize your scheme. Show a log-log plot of the error vs the grid spacing Δx and check from this plot that your method is second-order accurate.

We will have a discretization in $(x_i, y_j, t_n) = (i\Delta x, j\Delta y, n\Delta t)$, with $i, j, n \in \mathbb{Z}_{0+}$ Using the requested derivative approximations, the wave equation becomes

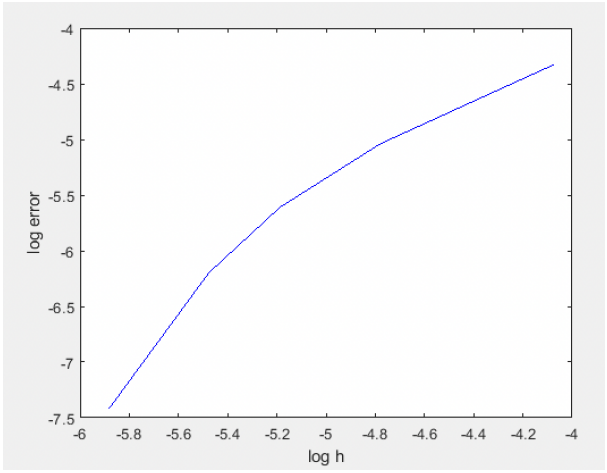
$$\frac{U_{ij}^{n+1} - 2U_{ij}^{n} + U_{ij}^{n-1}}{\Delta t^2} = \frac{U_{i+1j}^{n} + U_{i-1j}^{n} + U_{ij+1}^{n} + U_{ij-1}^{n} - 4U_{ij}^{n}}{h^2} \tag{86}$$

Now, we will solve this equation for U_{ij}^{n+1}

$$U_{ij}^{n+1} = 2U_{ij}^{n} + \frac{\Delta t^2}{h^2}\left(U_{i+1j}^{n} + U_{i-1j}^{n} + U_{ij+1}^{n} + U_{ij-1}^{n} - 4U_{ij}^{n}\right) - U_{ij}^{n-1} \tag{87}$$

$$U_{ij}^{n+1} = \frac{\Delta t^2}{h^2}\left(U_{i+1j}^{n} + U_{i-1j}^{n} + U_{ij+1}^{n} + U_{ij-1}^{n} + (2\frac{h^2}{\Delta t^2} - 4)U_{ij}^{n}\right) - U_{ij}^{n-1} \tag{88}$$

This is the basis for a scheme to be implemented in MATLAB. See github for the wave equation driver script. Here is the resulting log logplot



The linear polyfit of the data gave a coefficient of 1.6, which is close to the 2 needed for quadratic convergence. The coeffs can be found by running processing_C_a.m

```
coeffs =

    1.6454    2.6403
```

(b) (5 points) Consider the ODE $y''(t) = \lambda y$ and the 3 point rule for y'' as a two step explicit time integrator. Find the region of stability of this ODE solver in terms of $\lambda\Delta t^2$, and plot it in the complex plane.

The three point rule gives

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{\Delta t^2} = \lambda y_n \qquad \text{OR} \qquad y_{n+1} - (\lambda\Delta t^2 + 2)y_n + y_{n-1} = 0 \tag{89}$$

We then use the form

$$y^n = \sum_k c_k \rho_k^n \qquad \text{THUS} \qquad y^{n+1} = \sum_k c_k \rho_k^{n+1} \qquad \text{AND} \qquad y^{n-1} = \sum_k c_k \rho_k^{n-1} \tag{90}$$

This gives

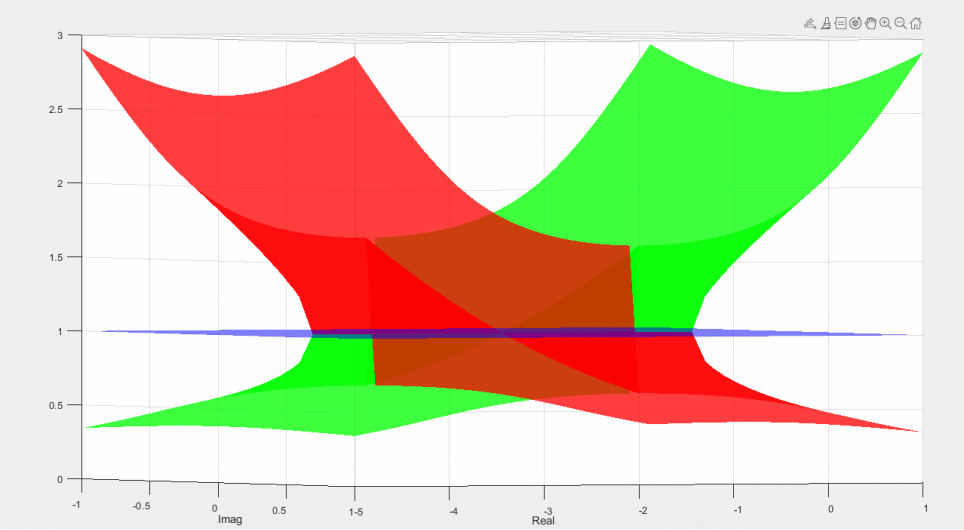
$$0 = \rho^2 - (2 + \Delta t^2\lambda)\rho + 1 \tag{91}$$

$$\rho = \frac{2 + \Delta t^2\lambda}{2} \pm \frac{\sqrt{(2 + \Delta t^2\lambda)^2 - 4}}{2} = 1 + \frac{\Delta t^2\lambda}{2} \pm \sqrt{(1 + \frac{\Delta t^2\lambda}{2})^2 - 1} \tag{92}$$

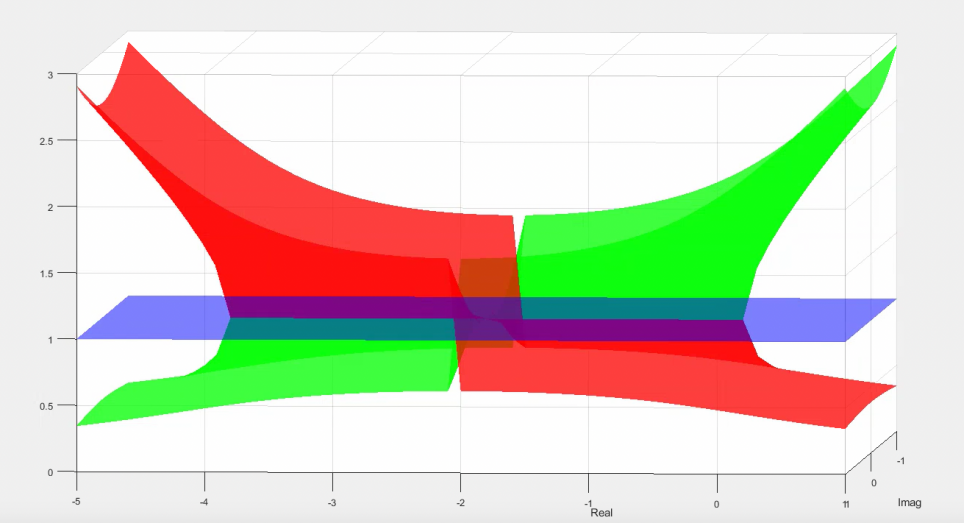
First, let's denote $\alpha = \Delta t^2\lambda$ Now, we require that $|\rho| < 1$

$$\left|1 + \frac{\alpha}{2} \pm \sqrt{(1 + \frac{\alpha}{2})^2 - 1}\right| \leq 1 \tag{93}$$

Here are some plots of the magnitude of both roots compared to unity. In order to have stability, the magnitude of both roots must be lesser than equal to unity. In the following figures, the green surface corresponds to the positive roots magnitude and the red surface corresponds to the negative roots magnitude.



This image shows that there is a region where the planes slice the unity plane and there are parts where the green surface is below and the red surface is above and vice versa. Since both roots need to be below, most of the complex plain will fail to satisfy the CFL condition.



This plot shows that there is a slice along the real axis from -4 to 0 where both surfaces intersect the unity plane. The region of stability is $-4 < \alpha < 0$, for $\alpha \in \mathbb{R}$. The following code was worked on with Varun Gudibanda.

```
%% Code based off code from Varun Gudibanda
% Find region of Stability for C(b)on HW2
f = @(x,y) 1+(x+i*y)/2 + sqrt((1+(x+i*y)/2).^2 - 1);
g = @(x,y) 1+(x+i*y)/2 - sqrt((1+(x+i*y)/2).^2 - 1);
[X,Y] = meshgrid(-5:0.1:1,-1:0.1:1);
h = ones(size(X));

figure(1)
surf(X, Y, abs(f(X,Y)), 'FaceColor','g', 'FaceAlpha',0.75, 'EdgeColor',
    'none')
hold on
```

```
surf(X, Y, abs(g(X,Y)), 'FaceColor','r', 'FaceAlpha',0.75, 'EdgeColor',
    'none')
surf(X, Y, h, 'FaceColor', 'b', 'FaceAlpha', 0.5, 'EdgeColor', 'none')
xlabel('Real')
ylabel('Imag')
hold off
```

(c) (5 points) From your answer to (b), and your knowledge of the spectrum of the discrete Laplacian, perform the "method of lines" stability analysis for the method in (a). What CFL condition does this analysis result in?

In the last question, we showed that $-4 < \lambda\Delta t^2 < 0$ is the region of stability. Now, with the method of lines, we have the semi-discrete system

$$y''(t) = \Delta_h y(t) \tag{94}$$

If the matrix $A_h \in \mathbb{R}^{N\times N}$ is the 1D Dirichlet second derivative difference matrix, then we know that

$$\Delta_h = I \otimes A_h + A_h \otimes I \tag{95}$$

We know the eigenvalues of A_h are

$$\lambda(A_h) = -\frac{4}{h^2}\sin^2\left(\frac{k\pi h}{2}\right) \qquad \text{FOR} \qquad k = 1, .., N \tag{96}$$

From the last homework, we know that the eigenvalues of Δ_h are all possible sums of the eigenvalues of A and A, giving

$$\lambda(\Delta_h) = -\frac{4}{h^2}\sin^2\left(\frac{i\pi h}{2}\right) - \frac{4}{h^2}\sin^2\left(\frac{i\pi h}{2}\right) \qquad \text{FOR} \qquad i = 1, .., N \qquad j = 1, ..., N \tag{97}$$

This gives a CFL condition of

$$-4 < \Big(-\frac{4}{h^2}\sin^2\left(\frac{i\pi h}{2}\right) - \frac{4}{h^2}\sin^2\left(\frac{i\pi h}{2}\right)\Big)\Delta t^2 < 0 \qquad \text{FOR} \qquad i = 1, .., N \qquad j = 1, ..., N \tag{98}$$

$$4 > \frac{4}{h^2}\Big(\sin^2\left(\frac{i\pi h}{2}\right) + \sin^2\left(\frac{i\pi h}{2}\right)\Big)\Delta t^2 > 0 \qquad \text{FOR} \qquad i = 1, .., N \qquad j = 1, ..., N \tag{99}$$

We know the sum of those two sines will be maxed out when both reach a maximum of 1, so to be safe for every i and k, we can use

$$4 > \frac{4}{h^2}(2)\Delta t^2 > 0 \qquad \text{THUS} \qquad 2 > \frac{h^2}{\Delta t^2} > 0 \tag{100}$$

and this is our CFL condition!

(d) (5 points) Perform the von Neumann stability analysis for the method in (a) and check if the resulting CFL condition agrees with what you found in the previous question. Since this is a 2D problem, a plane wave is $\exp(ik_1j_1\Delta x)\exp(ik_2j_2\Delta y)$

To perform the von Neumann stability analysis, we return to (88). Since the plane wave depends on the imaginary number i, we will change our indexing from (i, j) to (l, m) so that it is not confused

$$U_{lm}^{n+1} = \frac{\Delta t^2}{h^2}\Big(U_{l+1m}^n + U_{l-1m}^n + U_{lm+1}^n + U_{lm-1}^n + (2\frac{h^2}{\Delta t^2} - 4)U_{lm}^n\Big) - U_{lm}^{n-1} \tag{101}$$

Since we have $h = \Delta x = \Delta y$, we can assume that $k = k_1 = k_2$. Using this and our new indexing scheme, we have

$$U_{lm} \propto \exp(ik_xlh)\exp(ik_ymh) \tag{102}$$

The von Neumann analysis imposes

$$U^n = g(k_x, k_y)U^{n-1} \qquad \text{AND} \qquad U^{n+1} = g(k_x, k_y)^2U_{n-1} \tag{103}$$

For now, we will use the shorthand of $g = g(k_x, k_y)$ Using all these,

$$\exp(ik_x lh)\exp(ik_y mh)g^2 = g\frac{\Delta t^2}{h^2}\big(\exp(ik_x(l+1)h)\exp(ik_y mh) + \exp(ik_x(l-1)h)\exp(ik_y mh) + \exp(ik_x lh)\exp(ik_y(m+1)h) + \exp(ik_x lh)\exp(ik_y(m-1)h) + (2\frac{h^2}{\Delta t^2} - 4)\exp(ik_x lh)\exp(ik_y mh)\big) - \exp(ik_x lh)\exp(ik_y mh) \quad (104)$$

This simplifies to

$$g^2 = g\frac{\Delta t^2}{h^2}\big(\exp(ik_x h) + \exp(-ik_x h) + \exp(ik_y h) + \exp(-ik_y h) + (2\frac{h^2}{\Delta t^2} - 4)\big) - 1 \quad (105)$$

We can rearrange a bit and also use that the definition of cos appears

$$0 = g^2 - g\frac{\Delta t^2}{h^2}\big(2\cos(k_x h) + 2\cos(k_y h) + (2\frac{h^2}{\Delta t^2} - 4)\big) + 1 \quad (106)$$

Obviously (according to the professor I didn't think this was obvious), we need the half angle formula,

$$\cos(u) = 1 - 2\sin^2(u/2) \quad (107)$$

Using this in our earlier expression

$$0 = g^2 - g\frac{\Delta t^2}{h^2}\big(2 - 4\sin^2(\frac{k_x h}{2}) + 2 - 4\sin^2(\frac{k_y h}{2}) + (2\frac{h^2}{\Delta t^2} - 4)\big) + 1 \quad (108)$$

$$0 = g^2 - g\frac{\Delta t^2}{h^2}\big(-4\sin^2(\frac{k_x h}{2}) - 4\sin^2(\frac{k_y h}{2}) + 2\frac{h^2}{\Delta t^2}\big) + 1 \quad (109)$$

$$0 = g^2 + g\big(4\frac{\Delta t^2}{h^2}\sin^2(\frac{k_x h}{2}) + 4\frac{\Delta t^2}{h^2}\sin^2(\frac{k_y h}{2}) - 2\big) + 1 \quad (110)$$

Now to use the quadratic formula to solve for $g(k)$

$$g = -\frac{4\frac{\Delta t^2}{h^2}\sin^2(\frac{k_x h}{2}) + 4\frac{\Delta t^2}{h^2}\sin^2(\frac{k_y h}{2}) - 2}{2} \pm \frac{\sqrt{(4\frac{\Delta t^2}{h^2}\sin^2(\frac{k_x h}{2}) + 4\frac{\Delta t^2}{h^2}\sin^2(\frac{k_y h}{2}) - 2)^2 - 4}}{2} \quad (111)$$

$$g = 1 - \big(2\frac{\Delta t^2}{h^2}\sin^2(\frac{k_x h}{2}) + 2\frac{\Delta t^2}{h^2}\sin^2(\frac{k_y h}{2})\big) \pm \sqrt{\big(1 - (2\frac{\Delta t^2}{h^2}\sin^2(\frac{k_x h}{2}) + 2\frac{\Delta t^2}{h^2}\sin^2(\frac{k_y h}{2}))^2 - 1} \quad (112)$$

To get the CFL condition, we require that $|g| < 1$. If we let,

$$-\frac{\alpha}{2} = 2\frac{\Delta t^2}{h^2}\sin^2(\frac{k_x h}{2}) + 2\frac{\Delta t^2}{h^2}\sin^2(\frac{k_y h}{2} \quad (113)$$

Then we get

$$g = 1 + \frac{\alpha}{2} \pm \sqrt{(1 + \frac{\alpha}{2})^2 - 1} \quad (114)$$

This is the exact same situation as in part (c), and we know that this requires

$$-4 < \alpha < 0 \quad (115)$$

Now, using our α for this part

$$-4 < -2\bigg(2\frac{\Delta t^2}{h^2}\sin^2(\frac{k_x h}{2}) + 2\frac{\Delta t^2}{h^2}\sin^2(\frac{k_y h}{2})\bigg) < 0 \quad (116)$$

Simplifying

$$\frac{h^2}{\Delta t^2} > \sin^2(\frac{k_x h}{2}) + \sin^2(\frac{k_y h}{2}) > 0 \quad (117)$$

We know that at most,

$$2 = \sin^2(\frac{k_x h}{2}) + \sin^2(\frac{k_y h}{2}) \quad (118)$$

so to make sure we have stability for every k_x,k_y, and h, we get

$$\frac{h^2}{\Delta t^2} > 2 \quad (119)$$

(e) (5 points) Find the modified equation that corresponds to the numerical method in (a) Solve it via Fourier series and comment on the physics of the extra terms. Are they dissipative, dispersive, or something else?

We will start with (86),

$$\frac{U_{ij}^{n+1} - 2U_{ij}^{n} + U_{ij}^{n-1}}{\Delta t^2} = \frac{U_{i+1j}^{n} + U_{i-1j}^{n} + U_{ij+1}^{n} + U_{ij-1}^{n} - 4U_{ij}^{n}}{h^2} \tag{120}$$

We will take $U_{ij}^{n} = u(x, y, t)$, thus for the time steps

$$U_{ij}^{n+1} = u(x, y, t + \Delta t) \qquad \text{AND} \qquad U_{ij}^{n-1} = u(x, y, t - \Delta t) \tag{121}$$

For the space steps,

$$U_{i\pm1j}^{n} = u(x \pm h, y, t) \qquad \text{AND} \qquad U_{ij\pm1}^{n} = u(x, y \pm h, t) \tag{122}$$

Inserting all of these, we get

$$\frac{u(x, y, t + \Delta t) - 2u(x, y, t) + u(x, y, t - \Delta t)}{\Delta t^2} = \frac{u(x + h, y, t) + u(x - h, y, t) + u(x, y + h, t) + u(x, y - h, t) - 4u(x, y, t)}{h^2} \tag{123}$$

Next, we will Taylor expand all the terms, leaving off terms of order higher than 4,

$$u(x, y, t \pm \Delta t) = u(x, y, t) \pm \Delta t u_t(x, y, t) + \frac{(\Delta t)^2}{2} u_{tt}(x, y, t) \pm \frac{(\Delta t)^3}{3!} u_{ttt}(x, y, t) + \frac{(\Delta t)^4}{4!} u_{tttt}(x, y, t) \tag{124}$$

$$u(x \pm h, y, t) = u(x, y, t) \pm h u_x(x, y, t) + \frac{h^2}{2} u_{xx}(x, y, t) \pm \frac{h^3}{3!} u_{xxx}(x, y, t) + \frac{h^4}{4!} u_{xxxx}(x, y, t) \tag{125}$$

$$u(x, y \pm h, t) = u(x, y, t) \pm h u_y(x, y, t) + \frac{h^2}{2} u_{yy}(x, y, t) \pm \frac{h^3}{3!} u_{yyy}(x, y, t) + \frac{h^4}{4!} u_{yyyy}(x, y, t) \tag{126}$$

The LHS

$$u(x, y, t + \Delta t) - 2u(x, y, t) + u(x, y, t - \Delta t) = (\Delta t)^2 u_{tt}(x, y, t) + \frac{(\Delta t)^4}{12} u_{tttt}(x, y, t) \tag{127}$$

The RHS

$$u(x + h, y, t) + u(x - h, y, t) + u(x, y + h, t) + u(x, y - h, t) - 4u(x, y, t) = h^2 u_{xx}(x, y, t) + \frac{h^4}{12} u_{xxxx}(x, y, t) + h^2 u_{yy}(x, y, t) + \frac{h^4}{12} u_{yyyy}(x, y, t) \tag{128}$$

All together

$$u_{tt}(x, y, t) + \frac{(\Delta t)^2}{12} u_{tttt}(x, y, t) = u_{xx}(x, y, t) + \frac{h^2}{12} u_{xxxx}(x, y, t) + u_{yy}(x, y, t) + \frac{h^2}{12} u_{yyyy}(x, y, t) \tag{129}$$

At this point, we will drop the function notaion and just use subscripts to denote partial derivatives. We know from the original equation, $u_{tt} = u_{xx} + u_{yy}$, this reduces (129) to

$$u_{tttt} = \frac{h^2}{(\Delta t)^2} \left(u_{xxxx} + u_{yyyy} \right) \tag{130}$$

If we assume a plane like wave solution in space and oscillatory solution in time with frequency ω, we will get

$$u \sim \exp(-i(k_x x + k_y y - \omega(k_x, k_y) t)) \tag{131}$$

This gives that

$$u_x = -i k_x u \qquad \text{AND} \qquad u_y = -i k_y u \qquad \text{AND} \qquad u_t = i\omega u \tag{132}$$

By composing the derivatives, we can find how the higher order derivatives will behave in the same vein as (132). This is essentially transforming to the Fourier domain. Applying this to (133)

$$(i\omega)^4 = \frac{h^2}{(\Delta t)^2}\left((-ik_x)^4 u + (-ik_y)^4\right) \qquad \text{THUS} \qquad \omega^4 = \frac{h^2}{(\Delta t)^2}\left(k_x^4 + k_y^4\right) \tag{133}$$

To determine if this PDE would be dissipative or dispersive, we need to know a few things. If $Im(\omega) > 0$, then we would have a dissipative term, as we could write $\omega = a + bi$, with $b < 0$. Thus

$$\exp(-i(k_x x + k_y y - (a+bi)t)) = \exp(-i(k_x x + k_y y) + iat - bt)) = \exp(-i(k_x x + k_y y - at))\exp(-bt)) \tag{134}$$

The part $\exp(i(k_x x + k_y y - at))$ exhibits normal wave behavior, but it is multiplied by $\exp(-bt)$, which means the amplitude will decay as time goes on. We can determine if the wave is dispersive if the group velocity is a function of the wave numbers. The group velocity is defined as $\frac{d\omega}{dk}$. Here we have two different wave numbers, but we could take the gradient to determine. However, we know from the dispersion relation that we will have a dispersive wave, since $\omega(k_x, k_y)$ does not depend linearly on k_x and k_y. In this case, there is a chance to have both dispersive and dissipative terms. We can only guarantee no dissipation if both k_x and k_y are guaranteed real. If they have imaginary components, it is possible the fourth root would also contain an imaginary part. ˜