[1]

CS 714 HOMEWORK 2

Zihao Zheng
Department of Statistics
University of Wisconsin-Madison
Madison, WI, 53705
zihao.zheng@wisc.edu

October 31, 2020

1 Problem A

1.1 (a)

Since v belongs to the span of $w_1, w_2, \cdots, w_n$, then we have the following with coefficient $\alpha_1, \alpha_2, \cdots, \alpha_n$,

$$v = \sum_{i=1}^{n} \alpha_i w_i \tag{1}$$

$$\langle v, w_j \rangle = \langle \sum_{i=1}^{n} \alpha_i w_i, w_j \rangle = \alpha_j ||w_j||^2 \tag{2}$$

Therefore, we can validate the desired equation

$$\sum_{i=1}^{n} \frac{\langle v, w_j \rangle}{||w_j||^2} w_j = \sum_{i=1}^{n} \frac{\alpha_j ||w_j||^2}{||w_j||^2} w_j = v \tag{3}$$

1.2 (b)

1.2.1 (i)

It could converge faster (before N) if the solution lies in the span space of lower dimension (not necessary to be in R^N).

1.2.2 (ii)

I prove by induction.

1. When $n = 1$, The only thing I need to validate is $\langle p_1, p_0 \rangle = 0$. By definition

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

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

2. Inductive step: suppose the statement is true for $n = 1, 2, \cdots, k-1$

[1] Codes can be found at https://github.com/ZihaoZheng-Stat/CS714

3. Consider $n = k$

$$p_k = r_k - \sum_{j=0}^{k-1} \frac{\langle r_j, p_k \rangle_A}{||p_j||_A^2} p_j$$

for any $i, 0 \leq i < k$:

$$\langle p_k, p_i \rangle_A = \langle r_k, r_i \rangle_A - \sum_{j=0}^{k-1} \frac{\langle r_k, p_j \rangle_A}{||p_j||_A^2} \langle p_j, p_i \rangle_A = \langle r_k, p_i \rangle_A - \frac{\langle r_k, p_i \rangle_A}{||p_i||_A^2} ||p_i||_A^2 = 0$$

where the second equality holds by inductive assumption.

1.3 (c)

1.3.1 (i)

Since ϕ_n are orthonormal basis, we have coefficients α_n and β_n such that

$$v = \sum_{i=1}^{N} \alpha_n \phi_n$$
$$w = \sum_{i=1}^{N} \beta_n \phi_n$$

Therefore we have the desired equality by

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

where the last equality holds by $\langle v, \phi_n \rangle = \alpha_n$ and $\langle \phi_n, w \rangle = \beta_n$.

1.3.2 (ii)

This holds simply by definition of matrix A is a symmetric positive definite matrix. Explicitly, if some eigenvalue λ_i is not positive, we can simply find a vector a with only 1 entry of 1 corresponding to that eigenvalue and all the others 0. Then $a' A a = \lambda_i$.

1.3.3 (iii)

Suppose $v = \sum_{i=n}^{N} \alpha_n \phi_n$ and therefore $||v||^2 = \sum_{i=1}^{N} \alpha_n^2$.

$$Av = \sum_{i=n}^{N} \alpha_n \lambda_n \phi_n$$
$$\langle Av, v \rangle = \sum_{i=1}^{N} \lambda_n \alpha_n^2$$

And the last term $\sum_{i=1}^{N} \lambda_n \alpha_n^2$ can be bounded by $\lambda_1^2 \sum_{i=1}^{N} \alpha_n^2$ and $\lambda_N^2 \sum_{i=1}^{N} \alpha_n^2$ by the positivity of λ_i, which is equivalent to the desired inequality.

1.3.4 (iv)

$$||Av||^2 = \sum_{i=1}^{N} \alpha_n^2 \lambda_n^2 \leq \lambda_N^2 \sum_{i=1}^{N} \alpha_n^2 = \lambda_N^2 ||v||^2$$

It yields $||Av|| \leq \lambda_N ||v||$ by taking square root and $\lambda_N > 0$

1.4 (d)

By the definition of $p_{n+1} = r_{n+1} + \beta_n p_n$, the desired equation, $p_{n+1} = (1+\beta_n)p_n - \alpha_n A p_n - \beta_{n-1}p_{n-1}$

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

The last identity holds immediately by definition of p_n and the second $\iff$ is true by definition of r_{n+1}.

1.5 (e)

Apply Cayley-Hamilton theorem stated as the following.

For a matrix A and its characteristic polynomial function

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

If one defines an analogous matrix equation, $p(A)$, consisting of the replacement of the scalar variable λ with the matrix A, then this polynomial in the matrix A, $p(A)$ results in the zero matrix.

Note that, the previous defined function $p(A)$ in the theorem provides an equation with A^N (with coefficient exactly 1) and all the other lower-power matrix A^i with certain coefficient c_i. Namely, we will have

$$p(A) = A^N + \sum_{i=1}^{N-1} c_i A^i + c_0 I = 0$$

This immediately verifies that A^N is the linear combination of $I, A, \cdots, A^{N-1}$. The singularity bypasses some pathological cases, when $A^n = 0$.

1.6 (f)

1.6.1 (i)

Starting from $u_{n+1} = u_n + \alpha(f - Au_n)$, we have

$$e_{n+1} = u_{n+1} - u_n = u_n - u + \alpha(f - Au_n) = e_n + \alpha(Au - Au_n) = e_n + \alpha(-Ae_n) = (I - \alpha A)e_n$$

1.6.2 (ii)

By previous result $e_{n+1} = (I - \alpha A)e_n$ and take the norm in both side, we have

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

where ρ is the maximum absolute eigenvalue of matrix $|I - \alpha A|$, which is defined in the desired solution

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

So it follows immediately that $e_n \to 0$ if $\rho < 1$.

1.6.3 (iii)

Notice that $|1 - \alpha\lambda_j|$ is the distance between 1 and $\alpha\lambda_j$ and all $\alpha\lambda_j$ locates as ordered in the line. Therefore, the key observation here is

$$\max_{1\leq j\leq N} |1 - \alpha\lambda_j| = \max(|1 - \alpha\lambda_1|, |1 - \alpha\lambda_N|)$$

Then we discuss the choice of α.

1. When α chosen such that $1-\alpha\lambda_1 \geq 0$ and $-\alpha\lambda_N \geq 0$, then we have $\alpha \leq \frac{1}{\lambda_N}$. aking $\alpha = \frac{1}{\lambda_N}$, we have the following bound
$$\max(|1-\alpha\lambda_1|, |1-\alpha\lambda_N|) = \frac{\lambda_N - \lambda_1}{\lambda_N}$$
2. When α chosen such that $1-\alpha\lambda_1 \geq 0$ and $1-\alpha\lambda_N \leq 0$, then we have $\frac{1}{\lambda_N} \leq \alpha \leq \frac{1}{\lambda_1}$. Taking $\alpha = \frac{2}{\lambda_1+\lambda_N}$, we have the following bound
$$\max(|1-\alpha\lambda_1|, |1-\alpha\lambda_N|) = \frac{\lambda_N - \lambda_1}{\lambda_N + \lambda_1}$$
3. When α chosen such that $1-\alpha\lambda_1 \leq 0$ and $1-\alpha\lambda_N \leq 0$, then we have $\alpha \geq \frac{1}{\lambda_1}$. Taking $\alpha = \frac{1}{\lambda_1}$, we have the following bound
$$\max(|1-\alpha\lambda_1|, |1-\alpha\lambda_N|) = \frac{\lambda_N - \lambda_1}{\lambda_1}$$

Combining all three choices, the ρ is minimized at $\frac{\lambda_N-\lambda_1}{\lambda_N+\lambda_1}$ when $\alpha = \frac{2}{\lambda_1+\lambda_N}$.

1.6.4 (iv)

When $c = \frac{2}{c+C}$, we have

$$1 - \alpha\lambda_j \leq 1 - \alpha c = \frac{C-c}{C+c}$$
$$\alpha\lambda_j - 1 \leq \alpha C - 1 = \frac{C-C}{C+c}$$

Therefore $\rho = \max_j(|1-\alpha\lambda)j|) \leq \frac{C-c}{C+c} < 1$

1.7 (g)

1.7.1 (i)

$$r_1 = r_0 - \alpha_0 w_0 = r_0 - \alpha_0 A p_0 = r_0 - \alpha_0 A r_0$$

1.7.2 (ii)

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

Therefore for the desired equality, we need to verify $-\alpha_n\beta_{n-1}Ap_{n-1} = \frac{\alpha_n\beta_{n-1}}{\alpha_{n-1}}(r_n - r_{n-1})$. Also by observing that $r_n - r_{n-1} = \alpha_{n-1}w_{n-1}$, we have

$$\frac{\alpha_n\beta_{n-1}}{\alpha_{n-1}}(r_n - r_{n-1}) = -\frac{\alpha_n\beta_{n-1}}{\alpha_{n-1}}\alpha_{n-1}w_{n-1} = \alpha_n\beta_{n-1}w_{n-1} = -\alpha_n A\beta_{n-1}p_{n-1}$$
where the last equality holds by definition $w_{n-1} = Ap_{n-1}$.

1.7.3 (iii)

For the first identity, we start with the right hand side,

$$r_0 q_0 - \delta_0 q_1 = \frac{1}{\alpha_0} q_0 - \frac{\sqrt{\beta_0}}{\alpha_0} q_1 = \frac{1}{\alpha_0} q_0 - \frac{r_1}{\alpha_0 ||r_0||} = \frac{r_0 - r_1}{\alpha_0 ||r_0||} = \frac{\alpha_0 A r_0}{\alpha_0 ||r_0||} = A q_0$$

For the second identity, we separate the right hand side by three parts,

1. $$\delta_{n-1} q_{n-1} = \frac{1}{\alpha_{n-1}} \frac{||r_n||}{||r_{n-1}||} \frac{r_{n-1}}{||r_{n-1}||}$$

2.
$$\gamma_n q_n = (\frac{1}{\alpha_n} + \frac{\beta_{n-1}}{\alpha_{n-1}})q_n = \frac{1}{\alpha_n}\frac{r_n}{||r_n||} + \frac{1}{\alpha_{n-1}}\frac{||r_n||r_n}{||r_{n-1}||^2}$$

3.
$$-\delta_n q_{n+1} = -\frac{1}{\alpha_n}\frac{r_{n+1}}{||r_n||}$$

Adding these three quantities together and with the help of $r_n - r_{n-1} = -\alpha_{n-1}w_{n-1}$ and $w_n = Ap_n$, the desired result holds.

1.7.4 (iv)

This is the immediate application of the previous property. Explicitly, each column on the left side is just Aq_n. On the right side $Q_nT_n - \delta_{n-1}e_n'$, the first quantity is just $q_0r_0 - \delta_0 q_1$. All the rest, except for the very last column, is $\delta_{n-1}q_{n-1} + \gamma_n q_n - \delta_n q_{n+1}$ and it is true by the previous problem. For the very last column, the right hand side is also true since we add one more term $-\delta_{n-1}e_n'$ to the last column of matrix multiplication Q_nT_n and it is also the immediate application of that in the previous problem.

1.7.5 (v)

Before the argument, we first state and prove a lemma.

$$r_k' r_i = 0, \forall i = 0, 1, \cdots k-1 \tag{4}$$

The prove of this statement is not hard. Because of the direction set is conjugate, we first have $r_k' p_i = 0, \forall i 0, 1, \cdots k-1$ (can be easily verified by induction). Also by definition $p_i = r_i + \beta_i p_{i-1}$, we have $r_i \in span\{p_i, p_{i-1}\}$. Therefore we complete the proof by adding one more verification of $r_k' r_0 = 0$.

After that, we can prove the statement. By the previous problem,

$$Q_n' A Q_n = Q_n'(Q_nT_n - \delta_{n-1}e_n') = Q_n'Q_nT_n - Q_n'\delta_{n-1}e_n' = Q_n'Q_nT_n = T_n$$

Where the second last equality and the last equality holds by definition of q_n and the lemma (actually, $Q_n'Q_n = I$).

2 Problem B

I use `R` with the help of function `approxfun` to implement the linear interpolation. Each time it inputs a vector of x_j and the value $f(x_j)$ and output the linear interpolation function, say $\hat{f}^N$ depending on the grid size parameter N. With a fine enough grid, we can evaluate the uniform norm between $\hat{f}^N$ and the true function f. Figure 1 summarizes examples when $N = 6, 10, 20$.

Each time of choice N, we can evaluate the error. In order to find the smallest N that bound the error by 0.01, we use binary search with the lower bound 10 and upper bound 200. After only a few iterations, we find the smallest $N = 100$ where the error $e_{100} = 0.0096, e_{99} = 0.0101$.

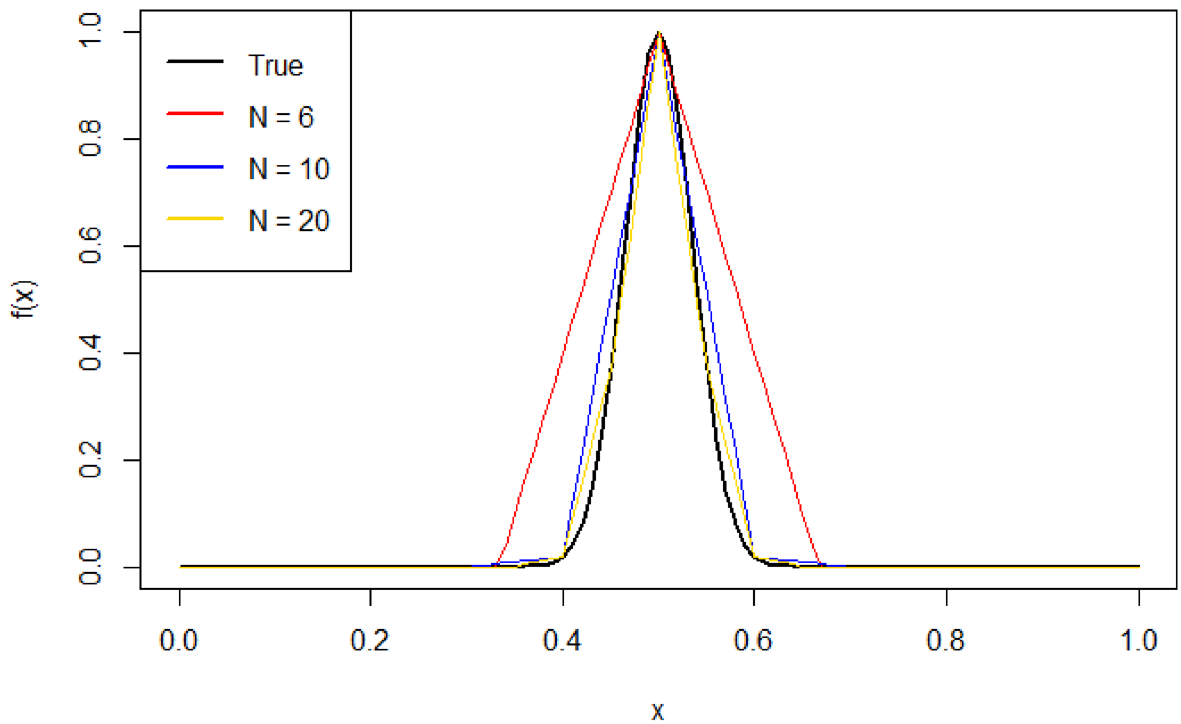

Figure 1: Figure of Problem B with some simple example demonstrations.

3 Problem C

3.1 (a)

For a specified Δ_x and Δ_t, we could use the finite difference approximation to approximate the second order over time and the Laplacian on a fixed time point t. Denote $U^n_{j_1,j_2}$ be the finite approximation on the grid $x = j_1\Delta_x, y = j_2\delta_x, t_n = n\Delta_t$ where $j_1, j_2 \in \{1, 2, \cdots, \frac{1}{\Delta_x}\}$ and n also has a finite upper boundary. Then we have the following approximation.

For the second order over t, we use the simplest 3-points formula:

$$u_{tt} = \frac{U^{n-1}_{j_1,j_2} - 2U^n_{j_1,j_2} + U^{n+1}_{j_1,j_2}}{\Delta_t^2} \tag{5}$$

For the Laplacian, we use the 5-point approximation:

$$\Delta u = \frac{U^n_{j_1-1,j_2} + U^n_{j_1+1,j_2} + U^n_{j_1,j_2-1} + U^n_{j_1,j_2+1} - 4U^n_{j_1,j_2}}{\Delta_x^2} \tag{6}$$

Therefore, the 2D wave equation could be formulated as the following:

$$\frac{U^{n-1}_{j_1,j_2} - 2U^n_{j_1,j_2} + U^{n+1}_{j_1,j_2}}{\Delta_t^2} = \frac{U^n_{j_1-1,j_2} + U^n_{j_1+1,j_2} + U^n_{j_1,j_2-1} + U^n_{j_1,j_2+1} - 4U^n_{j_1,j_2}}{\Delta_x^2} \tag{7}$$

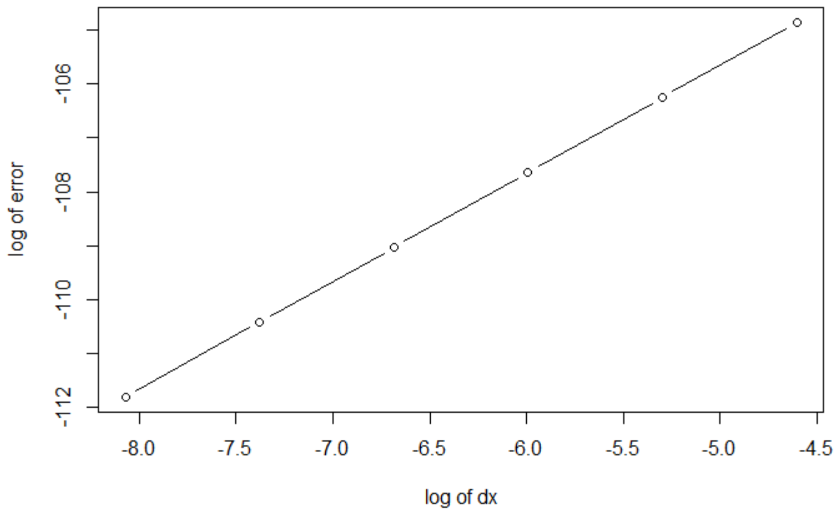

Figure 2: log-log plot between Δ_x and max-norm error

with the following boundary condition:

$$U^0_{j_1,j_2} = 0 \tag{8}$$
$$U^1_{j_1,j_2} = f(\Delta_x j_1) f(\Delta_x j_2) \Delta_t \tag{9}$$

Note that, formula (7) can be also modeled as the following, a two-step approach (in other words, update $U^{n+1}_{j_1,j_2}$ based on $U^n_{j_1,j_2}, U^{n-1}_{j_1,j_2}$), in update $U^n_{j_1,j_2}$:

$$U^{n+1}_{j_1,j_2} = \frac{\Delta_t^2}{\Delta_x^2}\left(U^n_{j_1-1,j_2} + U^n_{j_1+1,j_2} + U^n_{j_1,j_2-1} + U^n_{j_1,j_2+1} - 4U^n_{j_1,j_2}\right) \tag{10}$$
$$- U^{n-1}_{j_1,j_2} + 2U^n_{j_1,j_2} \tag{11}$$

In the simulation study, we try $\Delta_x = \frac{1}{N}, \frac{1}{2N}, \frac{1}{4N}, \frac{1}{8N}, \frac{1}{16N}, \frac{1}{32N}, \frac{1}{64N}$ and the correspondence $\Delta_t = \frac{1}{2}\Delta_x$ (this would be argued later that this choice would be sufficient to be stable). N is the lower bound in problem B ($N = 100$ in this study).

In order to evaluate the error, we choose the finest grid as the "true answer" and evaluate the max-norm over the grid $\Delta_x j_1, \Delta_x j_2$ and the time t (we run the update from $t = 0$ until $t = 1$). Figure 2 shows the log-log plot and it confirms that the slope is approximately 2.

3.2 (b)

Considering the ODE $y''(t) = \lambda_y$, we would like to solve the following approximation equation:

$$\frac{y^{n+1} - 2y^n + y^{n-1}}{\Delta_t^2} = \lambda y^n$$

By denoting ρ and let $\lambda \Delta_t^2 = k$, we are trying to solve the following equation:

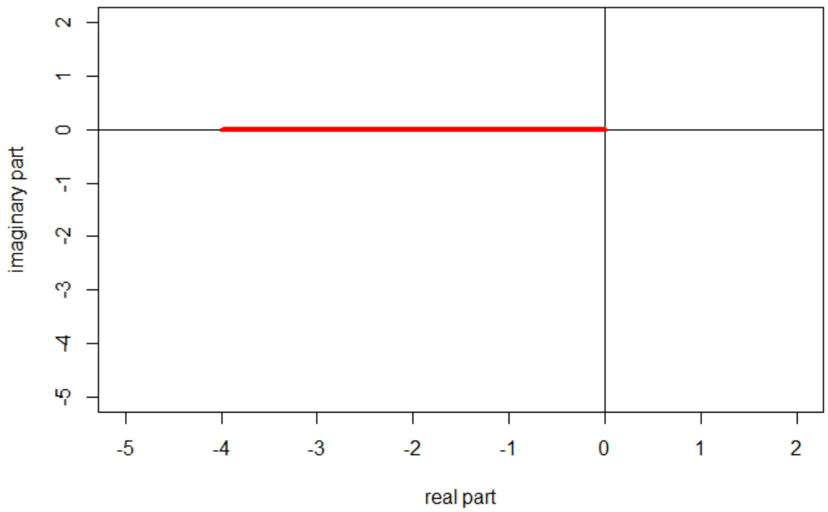

Figure 3: feasible set for $\lambda\Delta_t^2$

$$\rho - 2 + \frac{1}{\rho} = k$$
$$\rho^2 - (k+2)\rho + 1 = 0$$

In other words, we would like to find conditions for $k = \lambda\Delta_t^2$ to make sure $|\rho| \leq 1$.

This problem is not obvious since both ρ and k need to be considered in the complex plane. Therefore, I denote $k = a + bi$ and solve the complex equation and check whether or not the solution, stated as the following, has mod smaller than or equal to 1.

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

Then we could conclude the condition for a, b to make $|\rho| \leq 1$ (see simulation codes for detail): $b = 0, -4 \leq a \leq 0$ (shown as Figure 3).

3.3 (c)

From the previous problem, we known that $-4 \leq \lambda\Delta_t^2 \leq 0$, however λ is related to Δ_x. Therefore, we apply the following method of lines (12):

$$U''_{j_1,j_2}(t) = \frac{1}{\Delta_x^2}(U_{j_1-1,j_2}(t) + U_{j_1+1,j_2}(t) + U_{j_1,j_2-1}(t) + U_{j_1,j_2+1}(t) - 4U_{j_1,j_2}(t)) \tag{12}$$

By vectorizing $U_{j_1,j_2}(t)$ to get a $\frac{1}{\Delta_x^2}$ length vector, we could have the matrix form of the linear system, as the following:

$$U'' = MU$$
$$M = \frac{1}{\Delta_x^2}(A \otimes I_y + I_x \otimes A)$$

where $I_x = I_y$ are $\frac{1}{\Delta_x}$ by $\frac{1}{\Delta_x}$ matrix and A is the matrix that we are really familiar with (with diagonal -2, upper and lower diagonal 1).

Notice that the eigenvalue of M is bounded by $\frac{-8}{\Delta_x^2}$ (two times eigenvalue bound of A, which is $\frac{-4}{\Delta_x^2}$), combining with the previous result $\lambda\Delta_t^2 \geq -4$, we have the CFL condition:

$$\frac{\Delta_t^2}{\Delta_x^2} \leq \frac{1}{2}$$

3.4 (d)

Let's denote:

$$\begin{aligned}
U_{j_1,j_2}^{n} &= \exp(ik_1j_1\Delta_x + ik_2j_2\Delta_x) \\
U_{j_1,j_2}^{n+1} &= g\exp(ik_1j_1\Delta_x + ik_2j_2\Delta_x) \\
U_{j_1,j_2}^{n-1} &= \frac{1}{g}\exp(ik_1j_1\Delta_x + ik_2j_2\Delta_x)
\end{aligned}$$

Then the equation (10) is equivalent of the following:

$$\begin{aligned}
g\exp(ik_1j_1\Delta_x + ik_2j_2\Delta_x) = \frac{\Delta_t^2}{\Delta_x^2}\{&\exp((ik_1(j_1-1) + ik_2j_2)\Delta_x) \\
&+ \exp((ik_1(j_1+1) + ik_2j_2)\Delta_x) \\
&+ \exp((ik_1j_1 + ik_2(j_2-1))\Delta_x) \\
&+ \exp((ik_1j_1 + ik_2(j_2+1))\Delta_x) \\
&- 4\exp((ik_1j_1 + ik_2j_2)\Delta_x)\} \\
&- \frac{1}{g}\exp((ik_1j_1 - +ik_2j_2)\Delta_x) + 2\exp((ik_1j_1 - +ik_2j_2)\Delta_x)
\end{aligned}$$

which is then equivalent to:

$$g = \frac{\Delta_t^2}{\Delta_x^2}(2\cos k_1\Delta_x + 2\cos k_2\Delta_x - 4) - \frac{1}{g} + 2$$

Note that $2\cos k_1\Delta_x + 2\cos k_2\Delta_x - 4$ is bounded by 0 and -8. In order to make $|g| \leq 1$, the CFL condition for $\frac{\Delta_t^2}{\Delta_x^2}$ we need to assure is:

$$\frac{\Delta_t^2}{\Delta_x^2} \leq \frac{1}{2}$$

And this agrees with what we find in previous question.

3.5 (e)

Consider $v(x, y, t)$ for the modified equation as the following:

$$\begin{aligned}
&\frac{v(x,y,t+\Delta_t) - 2v(x,y,t) + v(x,y,t-\Delta_t)}{\Delta_t^2} \\
=& \frac{v(x-\Delta_x,y,t) + v(x+\Delta_x,y,t) + v(x,y+\Delta_x,t) + +v(x,y-\Delta_x,t) - 4v(x,y,t)}{\Delta_x^2}
\end{aligned}$$

Using Taylor expansion, we could get (ignore the higher order tail with order higher than Δ_t^2, Δ_x^2):

$$v_{tt} + \frac{1}{12}\Delta_t^2 v_{tttt} = \Delta v + \frac{1}{12}\Delta_x^2(v_{xxxx} + v_{yyyy}) \tag{13}$$

Compared to the original equation, we have extra term of order $O(\Delta_t^2)$ and $O(\Delta_x^2)$.
If using Fourier series, we will have the following dispersive system:

$$\frac{e^{-iw\Delta_t} + e^{iw\Delta_t} - 2}{\Delta_t^2} = \frac{e^{ik_1\Delta_x} + e^{ik_2\Delta_x} + e^{-ik_1\Delta_x} + e^{-ik_2\Delta_x} - 4}{\Delta_x^2} \tag{14}$$

which is equivalent to:

$$\cos(w\Delta_t) = 1 + \frac{\Delta_t^2}{\Delta_x^2}(\cos(k_1\Delta_x) + \cos(k_2\Delta_x) - 2)$$
$$w = \frac{1}{\Delta_t}\arccos(1 + \frac{\Delta_t^2}{\Delta_x^2}(\cos(k_1\Delta_x) + \cos(k_2\Delta_x) - 2))$$

4 Problem D

Consider the function $y(t)$ the linear ODE with two derivatives (I suppose in general form):

$$b(t) + a_0(t)y + a_1(t)y' + a_2(t)y'' = 0 \tag{15}$$

Using finite difference approximation, we get the following with $y_j = y(t_j)$:

$$b(t_j) + a_0(t_j)y_j + a_1(t_j)\frac{y_{j+1} - y_{j-1}}{h} + a_2(t_j)\frac{y_{j+1} - 2y_j + y_{j-1}}{h^2} = 0$$

which is equivalent to a linear system:

$$Y_{t+1} = AY_t + b$$

where each entry of b is known by $b(t_j)$ and the matrix A is similar to what we see in the class, with only diagonal and immediate upper/lower diagonal terms.

Then the Lax equivalence theorem states that if the system is consistent, and lax-stable ($|A^n| \geq C_T$), then the ODE with finite difference approximation is convergent. The proof exactly mirrors the general case stated in Theorem 9.2.