Department of Mathematics, University of Utah

Analysis of Numerical Methods, II
MATH 6620 – Section 001 – Spring 2024
Homework 5 Solutions
Fourier Series and time-dependent problems

Due Friday, April 5, 2024

Submit your solutions online through Gradescope.

1. (Hyperbolic systems)

a. A linear PDE,

$$\boldsymbol{u}_t + \boldsymbol{A}\boldsymbol{u}_x = 0,$$

governing the vector-valued solution $\boldsymbol{u}(x,t) \in \mathbb{R}^m$, is called *hyperbolic* if $\boldsymbol{A}$ is diagonalizable and $\lambda(\boldsymbol{A}) \subset \mathbb{R}$, with $\lambda(\boldsymbol{A})$ denoting the spectrum of $\boldsymbol{A}$. Assuming this PDE is hyperbolic, and using an equispaced grid in time and space, derive an implementable version of the *upwind scheme* for this problem, where $\boldsymbol{u}_t$ is discretized as $D^+\boldsymbol{u}_j^n$. You may ignore boundary conditions.

b. Derive and state a CFL condition for your upwind scheme.

c. Consider a *nonlinear* PDE,

$$\boldsymbol{u}_t + \boldsymbol{f}(\boldsymbol{u})_x = 0,$$

where $\boldsymbol{f} : \mathbb{R}^m \to \mathbb{R}^m$, and again we ignore boundary conditions. Propose a definition of a *hyperbolic PDE* in this context, and identify an upwind-like numerical scheme on an equidistant time and space mesh. (Again, discretize $\boldsymbol{u}_t \approx D^+\boldsymbol{u}_j^n$.) What is the CFL condition for your scheme?

Solution:

a. By assumption, we have,

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1},$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with real values $\lambda_1, \ldots, \lambda_m$ on the diagonal. We reformulate the PDE through the transformation,

$$\boldsymbol{w} \coloneqq \boldsymbol{V}^{-1}\boldsymbol{u},$$

and so by multiplying the PDE by $\boldsymbol{V}^{-1}$, we obtain,

$$\boldsymbol{w}_t + \boldsymbol{\Lambda}\boldsymbol{w}_x = \boldsymbol{0},$$

which is an uncoupled system of PDEs of the form,

$$(w_i)_t + \lambda_i (w_i)_x = 0, \qquad \boldsymbol{w}(x,t) = (w_1(x,t),\ \ldots,\ w_m(x,t))^T.$$

An upwind scheme for each of these scalar PDEs would be,

$$D^+ w_{i,j}^n + \lambda_i D_{\sigma_i} w_{i,j}^n = 0, \qquad w_{i,j}^n \approx w_i(x_j, t_n),$$

where x_j is an equisdistant grid in space with mesh spacing $h > 0$, and t_n is an equidistant grid in time with mesh spacing $k > 0$. The sign σ_i is determined by the sign of λ_i in the typical upwind manner for scalar, constant wavespeed problems:

$$\sigma_i \coloneqq -\text{sign}(\lambda_j), \implies D_{\sigma_i} = \begin{cases} D_-, & \lambda_i \geq 0 \text{ (rightward moving wave)} \\ D_+, & \lambda_i < 0 \text{ (leftward moving wave)} \end{cases}$$

where it doesn't really matter how σ_i is defined when $\lambda_i = 0$. Then define the size-m vector of differences as,

$$D_{\boldsymbol{\sigma}} \boldsymbol{w}_j^n \coloneqq \begin{pmatrix} D_{\sigma_1} w_{1,j}^n \\ D_{\sigma_2} w_{2,j}^n \\ \vdots \\ D_{\sigma_m} w_{1,m}^n \end{pmatrix}.$$

With this definition, then upwinding is given by,

$$D^+ \boldsymbol{w}_j^n + \boldsymbol{\Lambda} D_{\boldsymbol{\sigma}} \boldsymbol{w}_j^n = 0$$

We translate this to $\boldsymbol{u}$ space by multiplying by $\boldsymbol{V}$ and replacing $\boldsymbol{w}_j^n$ with its $\boldsymbol{u}$-representation:

$$D^+ \boldsymbol{u}_j^n + \boldsymbol{V} \left(\boldsymbol{\Lambda} D_{\boldsymbol{\sigma}} \boldsymbol{V}^{-1} \boldsymbol{u}_j^n \right) = 0,$$

or more explicitly,

$$\boldsymbol{u}_j^{n+1} = \boldsymbol{u}_j^n - k \boldsymbol{V} \left(\boldsymbol{\Lambda} D_{\boldsymbol{\sigma}} \boldsymbol{V}^{-1} \boldsymbol{u}_j^n \right)$$

b. The scheme from the previous part is essentially just an upwind scheme in $\boldsymbol{w}$-space, so the CFL condition is the same as the CFL condition for $\boldsymbol{w}$ space. For every $i \in [m]$, the CFL condition for the scheme evolving $w_{i,j}^n$ is,

$$k \leq \frac{h}{|\lambda_i|}$$

Hence, the CFL restriction for the whole system is the strictest condition for every i:

$$k \leq \min_{i \in [m]} \frac{h}{|\lambda_i|} = \frac{h}{\rho(\boldsymbol{A})},$$

where $\rho(\boldsymbol{A})$ is the spectral radius of $\boldsymbol{A}$:

$$\rho(\boldsymbol{A}) \coloneqq \max_{i \in [m]} |\lambda_i|.$$

c. In the linear case, we have,

$$\boldsymbol{f}(\boldsymbol{u}) = \boldsymbol{A}\boldsymbol{u},$$

and the definition for hyperbolicity was that $\boldsymbol{A}$ is diagonalizable with spectrum $\lambda(\boldsymbol{A})$ satisfying,

$$\lambda(\boldsymbol{A}) \subset \mathbb{R}.$$

Restating this in a way that can generalize to the nonlinear case, the hyperbolicity in the linear case meant that $\boldsymbol{A} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}$ was diagonalizable and satisfies,

$$\lambda\left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}\right) = \lambda(\boldsymbol{A}) \subset \mathbb{R}.$$

This now serves as a potential definition: we might say that our nonlinear PDE is hyperbolic at a state $\boldsymbol{u}$ if the Jacobian matrix $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}(\boldsymbol{u})$ is a diagonalizable matrix, with,

$$\lambda\left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}(\boldsymbol{u})\right) \subset \mathbb{R}.$$

Another way of motivating this definition is to rewrite the PDE by using the chain rule as $\boldsymbol{u}_t + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}\boldsymbol{u}_x = \boldsymbol{0}$, so that the Jacobian matrix is the analog of the wavespeed matrix $\boldsymbol{A}$, and hence the Jacobian should be diagonalizable with real eigenvalues for every relevant $\boldsymbol{u}$. (E.g., every $\boldsymbol{u}$ along a space-time solution trajectory.) Note that this condition seems rather strong, as we require such "real" diagonalizability for a relatively large class of vectors $\boldsymbol{u}$. Nevertheless, this gives us one way to identify a scheme: if the nonlinear system satisfies our definition of hyperbolicity, then we have for every space index j and every time index n,

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}(\boldsymbol{u}_j^n) = \boldsymbol{V}_j^n \boldsymbol{\Lambda}_j^n \left(\boldsymbol{V}_j^n\right)^{-1},$$

where we use sub- and superscripts to emphasize that quantities depend on the space and time index, respectively. (And in particular we must diagonalize $\boldsymbol{f}$ at every time and space point). With this, we may transform the system and perform a scheme that is at least "locally" upwinding:

$$\boldsymbol{w}_j^n \coloneqq \left(\boldsymbol{V}_j^n\right)^{-1} \implies D^+\boldsymbol{w}_j^n + \boldsymbol{\Lambda}_j^n D_{\boldsymbol{\sigma}_j^n}\boldsymbol{w}_j^n = 0,$$

where $\boldsymbol{\Lambda}_j^n$ has diagonal entries $(\lambda_j^n)_1, \ldots, (\lambda_j^n)_m$, and $\boldsymbol{\sigma}_j^n$ is given by,

$$\left(\sigma_j^n\right)_i = -\text{sign}\left(\lambda_j^n\right)_i.$$

Then transforming back into $\boldsymbol{u}$ space and expanding the D^+ operator, we have,

$$\boldsymbol{u}_j^{n+1} = \boldsymbol{u}_j^n - k\boldsymbol{V}_j^n\left(\boldsymbol{\Lambda}_j^n D_{\boldsymbol{\sigma}_j^n}\left(\boldsymbol{V}_j^n\right)^{-1}\boldsymbol{u}_j^n\right)$$

Under the assumption that characteristics travel at constant speed determined by the Jacobian at $\boldsymbol{u}_j^n$, then the CFL condition is given by the assumption that the (j,n)-local characteristics lie within the numerical domain of dependence, i.e.,

$$k \leq \frac{h}{\max_{i\in[m]}(\lambda_j^n)_i} = \frac{h}{\rho\left(\frac{\partial f}{\partial u}(\boldsymbol{u}_j^n)\right)}$$

2. (Hyperbolic systems in two spatial dimensions) Consider a linear system of PDEs of the form,

$$\boldsymbol{u}_t + \boldsymbol{A}\boldsymbol{u}_x + \boldsymbol{B}\boldsymbol{u}_y = 0,$$

for the unknown $\boldsymbol{u}(x,y,t) \in \mathbb{R}^m$, where $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times m}$ are given. Assume this system is *hyperbolic*, meaning that for every $\alpha, \beta \in \mathbb{R}$, then $\alpha\boldsymbol{A}+\beta\boldsymbol{B}$ is diagonalizable with $\lambda(\alpha\boldsymbol{A}+\beta\boldsymbol{B}) \subset \mathbb{R}$. We will use the abbreviation $\boldsymbol{x} = (x,y)^T \in \mathbb{R}^2$.

a. Determine plane wave solutions: let $\boldsymbol{n} \in \mathbb{R}^2$ satisfying $\|\boldsymbol{n}\|_2 = 1$ be a given vector, and assume some initial data $\boldsymbol{u}(x, y, 0) = \boldsymbol{u}_0(\boldsymbol{n} \cdot \boldsymbol{x})$. (Note that $\boldsymbol{u}_0$ is a function of a scalar input.) Ignoring boundary conditions, determine the solution $\boldsymbol{u}(x, y, t)$.

b. Assume further that $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric matrices. Consider discretizing this PDE on an equidistant spatial grid on a square: $h_x > 0$ and $h_y > 0$ are the grid spacing in the x- and y-directions, respectively. Again, assume the discretization $\boldsymbol{u}_t \approx D^+ \boldsymbol{u}^n_{i,j}$, where $\boldsymbol{u}^n_{i,j} \approx u(x_i, y_j, t_n)$. Using plane waves as motivation, derive a CFL condition (a condition on the timestep k) for such a discretization. You may assume that spatial derivatives are approximated using the stencil involving $u^n_{i\pm1,j\pm1}$ and that the numerical domain of dependence is the convex hull of the stencil points (as in the one-dimensional case).

Solution:

a. Given $\boldsymbol{n} = (n_x, n_y)^T$ with $n_x^2 + n_y^2 = 1$, define $r \coloneqq \boldsymbol{n} \cdot \boldsymbol{x} = x n_x + y n_y$. Thus, the initial data is $\boldsymbol{u}(x, y, 0) = \boldsymbol{u}_0(r)$. This suggests we should look for solutions of the form $\boldsymbol{u}(x, y, t) =: \boldsymbol{w}(x n_x + y n_y, t) = \boldsymbol{w}(r, t)$. In order to determine a PDE for $\boldsymbol{w}$, note that,

$$\begin{aligned}
\frac{\partial}{\partial t}\boldsymbol{u}(x, y, t) &= \frac{\partial}{\partial t}\boldsymbol{w}(r, t), \\
\frac{\partial}{\partial x}\boldsymbol{u}(x, y, t) &= \frac{\partial r}{\partial x}\frac{\partial}{\partial r}\boldsymbol{u}(x, y, t) = n_x \frac{\partial}{\partial r}\boldsymbol{u}(x, y, t) \\
\frac{\partial}{\partial y}\boldsymbol{u}(x, y, t) &= \frac{\partial r}{\partial y}\frac{\partial}{\partial r}\boldsymbol{u}(x, y, t) = n_y \frac{\partial}{\partial r}\boldsymbol{u}(x, y, t)
\end{aligned}$$

Therefore,

$$\boldsymbol{u}_t + \boldsymbol{A}\boldsymbol{u}_x + \boldsymbol{B}\boldsymbol{u}_y = 0 \quad \longrightarrow \quad \boldsymbol{w}_t + (n_x \boldsymbol{A} + n_y \boldsymbol{B})\, \boldsymbol{w}_r = 0.$$

Since by assumption $n_x \boldsymbol{A} + n_y \boldsymbol{B}$ is diagonalizable with real spectrum,

$$n_x \boldsymbol{A} + n_y \boldsymbol{B} = \boldsymbol{V}_{\boldsymbol{n}} \boldsymbol{\Lambda}_{\boldsymbol{n}} \boldsymbol{V}_{\boldsymbol{n}}^{-1}, \qquad \boldsymbol{\Lambda}_{\boldsymbol{n}} = \operatorname{diag}(\lambda_{\boldsymbol{n},1}, \ldots, \lambda_{\boldsymbol{n},m}), \qquad \lambda_{\boldsymbol{n},q} \in \mathbb{R} \;\; \forall q \in [m],$$

where we have used $\boldsymbol{n}$ subscripts to emphasize quantities that depend on $\boldsymbol{n}$. Then the PDE for $\boldsymbol{w}$ is a system of linear hyperbolic PDEs in a scalar space dimension r. The solution is given by linearly transforming $\boldsymbol{w}$ into the variable $\boldsymbol{W} \coloneqq \boldsymbol{V}_{\boldsymbol{n}}^{-1} \boldsymbol{w}$ to decouple the system:

$$\boldsymbol{W}_t + \boldsymbol{\Lambda}_{\boldsymbol{n}} \boldsymbol{W}_r = 0, \qquad \boldsymbol{W}(r, 0) = \boldsymbol{V}_{\boldsymbol{n}}^{-1} \boldsymbol{w}(r, 0) = \boldsymbol{V}_{\boldsymbol{n}}^{-1} \boldsymbol{u}_0(r).$$

Hence, the exact solution for $\boldsymbol{W}$ and hence $\boldsymbol{u}$ is given by,

$$\begin{aligned}
\boldsymbol{W}(r, t) = \boldsymbol{W}(r - \boldsymbol{\lambda}_{\boldsymbol{n}} t, 0) &\coloneqq \begin{pmatrix} W_1(r - \lambda_{\boldsymbol{n},1} t, 0) \\ \vdots \\ W_m(r - \lambda_{\boldsymbol{n},m} t, 0) \end{pmatrix} \\
\boldsymbol{u}(x, y, t) &= \boldsymbol{V}_{\boldsymbol{n}} \boldsymbol{W}(x n_x + y n_y - \boldsymbol{\lambda}_{\boldsymbol{n}} t, 0).
\end{aligned}$$

Through this argument, we have constructed a particular solution $\boldsymbol{u}$ to the PDE (that satisifes the initial data). Note that more general solutions, e.g., $u = u(r, s, t)$ with nontrivial s dependence, with $s = -n_y x + n_x y$, cannot exist due to well-posedness and linearity of the PDE. If such alternative solutions did exist, then the difference between the s-dependent and s-independent solution above would satisfy the same PDE but with zero initial data, meaning that the solution difference would be the zero function.

b. Note that for fixed $\boldsymbol{n}$, the solution to the above system for $\boldsymbol{u}$ (or $\boldsymbol{W}$) involves waves traveling in $\mathbb{R}^2$ in the direction $\boldsymbol{n}$ at speeds,

$$\lambda_{\boldsymbol{n},1}, \ldots \lambda_{\boldsymbol{n},m}.$$

For the one-directional dimension $\boldsymbol{n}$, we consider the origin-centered numerical domain of dependence to be the convex hull of neighboring points:

$$\text{numerical domain of dependence} = \text{conv}\left(h_x(1,0)^T,\ h_x(-1,0)^T,\ h_y(0,1)^T,\ h_y(0,-1)^T\right),$$

which is an origin-centered rotated quadrilateral. The distance from the origin to the boundary of this set along the direction $\boldsymbol{n}$ is given by,

$$h_r \coloneqq \frac{1}{\frac{|n_x|}{h_x} + \frac{|n_y|}{h_y}} = \frac{h_x h_y}{h_y|n_x| + h_x|n_y|}$$

Hence, in the one-dimensional direction $\boldsymbol{n}$, this problem has the following CFL restriction:

$$k \leq \min_{q \in [m]} \frac{h_r}{|\lambda_{\boldsymbol{n},q}|} = \frac{h_r}{\rho\left(n_x \boldsymbol{A} + n_y \boldsymbol{B}\right)}.$$

While this is the CFL restriction considering waves along a direction $\boldsymbol{n}$, the more practical restriction is an $\boldsymbol{n}$-independent condition. I.e., we should seek a condition without *a priori* knowledge of $\boldsymbol{n}$. Then formally this would be the condition,

$$k \leq \sup_{\|\boldsymbol{n}\|_2 = 1} \frac{h_r}{\rho\left(n_x \boldsymbol{A} + n_y \boldsymbol{B}\right)} \tag{1}$$

In order to derive a computable condition from this, we exploit the further assumptions that $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric. We claim that under this extra assumption,

$$\rho\left(n_x \boldsymbol{A} + n_y \boldsymbol{B}\right) \leq |n_x|\rho(\boldsymbol{A}) + |n_y|\rho(\boldsymbol{B}). \tag{2}$$

To prove this, note that since $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric, then their spectra are characterized by the Rayleigh quotient. E.g., for $\boldsymbol{A}$:

$$\lambda_{\min}(\boldsymbol{A}) = \min_{\boldsymbol{x} \neq \boldsymbol{0}} R_{\boldsymbol{A}}(\boldsymbol{x}), \qquad \lambda_{\max}(\boldsymbol{A}) = \max_{\boldsymbol{x} \neq \boldsymbol{0}} R_{\boldsymbol{A}}(\boldsymbol{x}), \qquad \rho(\boldsymbol{A}) = \max_{\boldsymbol{x} \neq \boldsymbol{0}} \left|R_{\boldsymbol{A}}(\boldsymbol{x})\right|,$$

with,

$$R_{\boldsymbol{A}}(\boldsymbol{x}) \coloneqq \frac{\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}}.$$

Note in particular that $\boldsymbol{R}_{\boldsymbol{A}}(\cdot)$ is linear in $\boldsymbol{A}$. Using this characterization, then,

$$\begin{aligned}
\rho(n_x \boldsymbol{A} + n_y \boldsymbol{B}) &= \max_{\boldsymbol{x} \neq \boldsymbol{0}} \left|R_{n_x \boldsymbol{A} + n_y \boldsymbol{B}}(\boldsymbol{x})\right| \\
&= \max_{\boldsymbol{x} \neq \boldsymbol{0}} \left|n_x R_{\boldsymbol{A}}(\boldsymbol{x}) + n_y R_{\boldsymbol{B}}(\boldsymbol{x})\right| \\
&\leq |n_x| \max_{\boldsymbol{x} \neq \boldsymbol{0}} \left|R_{\boldsymbol{A}}(\boldsymbol{x})\right| + |n_y| \max_{\boldsymbol{x} \neq \boldsymbol{0}} \left|R_{\boldsymbol{B}}(\boldsymbol{x})\right| \\
&= |n_x|\rho(\boldsymbol{A}) + |n_y|\rho(\boldsymbol{B}),
\end{aligned}$$

as claimed in (2). To see how (2) assists us with coming up with a more transparent CFL condition, note that we have just shown that,

$$\frac{h_r}{|n_x|\rho(\boldsymbol{A}) + |n_y|\rho(\boldsymbol{B})} \leq \frac{h_r}{|\rho(n_x\boldsymbol{A} + n_y\boldsymbol{B})|},$$

and hence is we can assure that k is smaller than the left-hand side above for all $\boldsymbol{n}$, then it will satisfy our CFL condition (1). This is enough to create a servicable coarse bound, since,

$$\begin{aligned}\frac{1}{h_r}\left(|n_x|\rho(\boldsymbol{A}) + |n_y|\rho(\boldsymbol{B})\right) &= \left(\frac{|n_x|}{h_x} + \frac{|n_y|}{h_y}\right)\left(|n_x|\rho(\boldsymbol{A}) + |n_y|\rho(\boldsymbol{B})\right) \\ &\leq \left(\frac{1}{h_x} + \frac{1}{h_y}\right)\left(\rho(\boldsymbol{A}) + \rho(\boldsymbol{B})\right),\end{aligned}$$

so that a sufficient CFL condition reads,

$$k \leq \frac{h_x h_y}{(h_x + h_y)\left(\rho(\boldsymbol{A}) + \rho(\boldsymbol{B})\right)},$$

where this bound ensures (1). (One may of course create a tighter bound by optimizing over all $\boldsymbol{n}$, but this produces only a slightly better bound.) Note that this bound has all the right general behavior for $h_x = h_y$ and/or $\rho(\boldsymbol{A}) = \rho(\boldsymbol{B})$.

3. (Finite differences for non-smooth problems) Consider Burgers' equation:

$$u_t + f(u)_x = 0, \qquad f(u) = \frac{u^2}{2}, \qquad (x,t) \in [-\pi,\pi] \times (0,T].$$

where we will take $T = 1.0$. Supplement this PDE with the boundary conditions,

$$u(\pm\pi, t) = u(\pm\pi, 0),$$

where the initial condition function $u(\cdot, 0)$ will be specified below. Note that for smooth u, then $f(u)_x = uu_x$. Based on this observation, and using an equidistant grid in both space and time, we will consider two schemes for this PDE:

$$\begin{aligned}\text{Scheme } A: &\quad D^+u_j^n + D_0 f(u_j^n) = 0, \\ \text{Scheme } B: &\quad D^+u_j^n + u_j^n D_0 u_j^n = 0.\end{aligned}$$

Numerically test these schemes for solving the PDE up to time $t = T$ with the following three initial data:

$$u(x,0) = u_1(x) = -\sin(x). \tag{3a}$$

$$u(x,0) = u_2(x) = \begin{cases} 1, & x \leq 0 \\ 0, & x > 0 \end{cases} \tag{3b}$$

$$u(x,0) = u_3(x) = \begin{cases} 0, & x \leq 0 \\ 1, & x > 0 \end{cases} \tag{3c}$$

Based on the experiments above, which scheme would you prefer to use for each example, and in general? (Feel free to try other schemes as well, e.g., upwind versions of Schemes A and B as identified in problem 1c.)

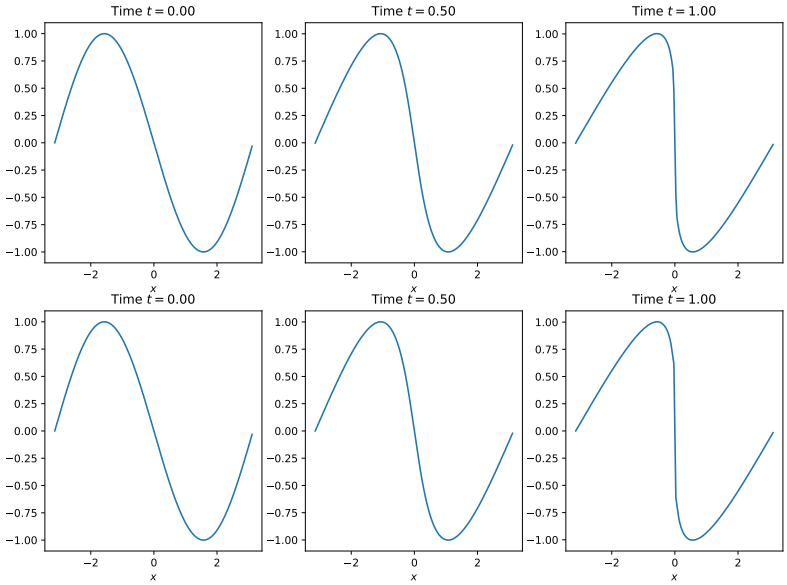

Figure 1: Problem 3: Scheme A (top) and Scheme B (bottom) for the initial condition (3a) to Burgers' equation. The left, center, and right columns correspond to times $t = 0, 0.5$, and 1.0.

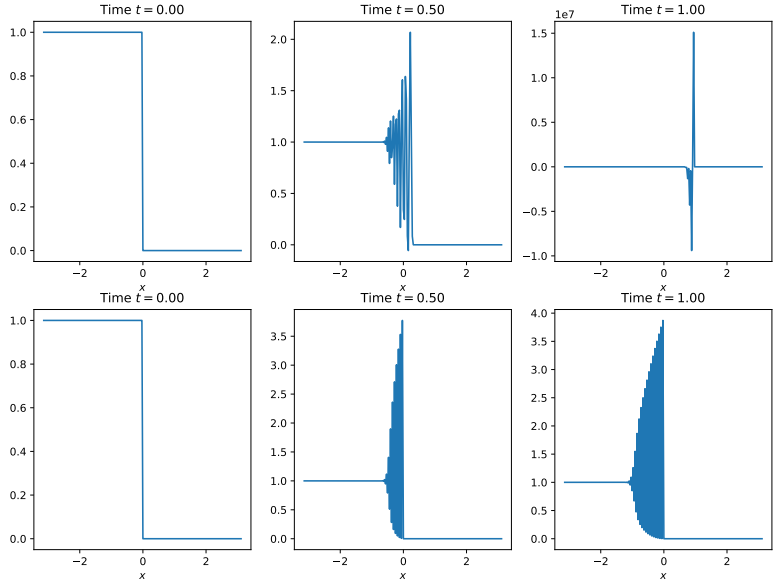

Figure 2: Problem 3: Scheme A (top) and Scheme B (bottom) for the initial condition (3b) to Burgers' equation. The left, center, and right columns correspond to times $t = 0, 0.5$, and 1.0.

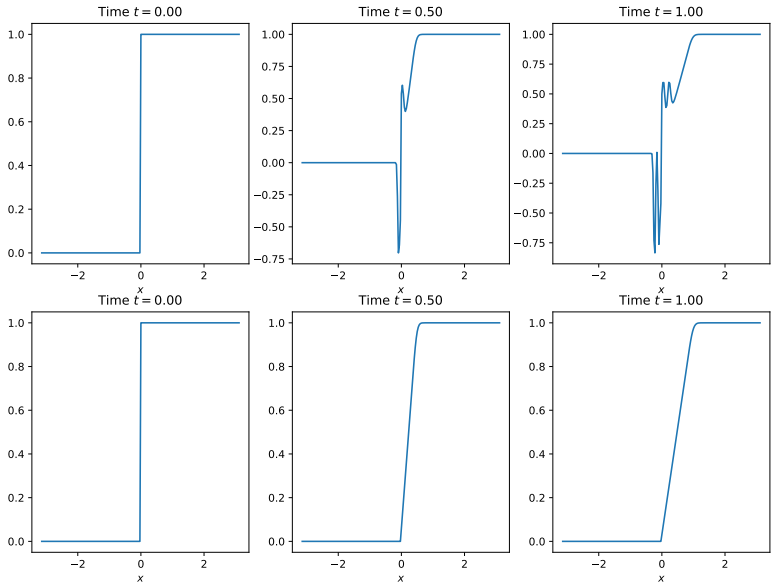

Figure 3: Problem 3: Scheme A (top) and Scheme B (bottom) for the initial condition (3c) to Burgers' equation. The left, center, and right columns correspond to times $t = 0, 0.5$, and 1.0.

Solution: Using the schemes described, we simulate the solution up to tim $T = 1$ using $M = 200$ equispaced points in space, and $N = 200$ timesteps. (I.e., $h = 2\pi/M$ and $k = 1/N$.) The numerical results for initial condition (3a) are presented in figure 1, for (3b) in figure 2, and for (3c) in figure 3.

The first major observation is that for initial condition (3a), both schemes perform comparably. This is perhaps not surprising since both solutions appear to be relatively smooth functions and these two schemes are essentially identical so long as $f(u)$ is smooth. However, for both discontinuous initial conditions (3b) and (3c), it's not quite clear which scheme is better. Both schemes produce seemingly unphysical oscillations for initial condition (3b), and both produce a somewhat reasonable-looking solution for initial condition (3c). Thus, these schemes produce significantly different behavior for non-smooth initial data, and it's not clear which one is better.

In fact, neither of these schemes is bulletproof for these examples: For initial condition (3c), it turns out that Scheme B produces a numerical solution that is at least a reasonable approximation to the "true" solution. In contrast, Scheme A produces something that is stable, but "very" incorrect. However, when we look at initial condition (3b), both schemes produce rather bad oscillations, and Scheme A is (eventually) unstable. However, Scheme A for this case actually produces a solution that has some of the correct characteristics; in particular, the real solution to this problem is a moving discontinuity, and Scheme A produces an essentially correct movement of the discontinuity, even if it's accompanied by very incorrect oscillations and eventual instability. While Scheme B appears stable for this initial condition, the discontinuity does not move in the correct way, and for example at $t = 0.5$, the Scheme B solution is worse than the Scheme A solution. (Although again both are quite bad.) This highlights a major point: these two initial conditions are seemingly the same experiment, but in fact

the correctness of each scheme varies considerably between these cases. The major point here is that the methods we've developed are not quite suitable (at least without modification) to tackle nonlinear problems with non-smooth solutions.

4. (Fourier approximation)

a. Prove that if $u \in H^s_p$ for some integer $s \geq 0$, then

$$\|u - P_N u\|_{L^2} \leq N^{-s} \|u\|_{H^s_p},$$

where the space H^s_p is defined on slide `D13-S10` and P_N is defined in slide `D13-S07`.

b. Confirm this behavior by numerically computing $\|u_j - P_N u_j\|_{L^2}$ as a function of N for each $j = 0, 1, 2, 3$. The functions u_j, $j \geq 0$, are defined as,

$$u_0(x) \coloneqq \begin{cases} 1, & |x - \pi| < \frac{\pi}{2} \\ -1, & else \end{cases} \qquad u_j(x) \coloneqq c_j + \int_0^x u_{j-1}(y)\,\mathrm{d}y \quad (j \geq 1),$$

where c_j is chosen so that u_j is a mean-0 function. Based on your numerical results what type of regularity (s) does u_j seem to have?

Solution:

a. If $s = 0$, the desired statement is,

$$\|u - P_N u\|_{L^2} \leq \|u\|_{L^2},$$

which is true since $(I - P_N)$ is an orthogonal projector. Therefore, suppose $s \geq 1$. We have,

$$u \in H^s_p \implies u^{(j)} \in L^2,\ 0 \leq j \leq s, \quad u^{(\ell)}(0) = u^{(\ell)}(2\pi),\ 0 \leq \ell \leq s - 1.$$

Hence, for $j \leq s$, we define expansion coefficients for any $k \in \mathbb{Z}$,

$$\widehat{u}^{(j)}_k \coloneqq \left\langle u^{(j)}(x), \phi_k(x) \right\rangle = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} u^{(j)}(x) e^{-ikx}\,\mathrm{d}x.$$

These coefficients satisfy Parseval's relation,

$$\left\| u^{(j)} | \right|^2_{L^2} = \sum_{k \in \mathbb{Z}} \left| \widehat{u}^{(j)}_k \right|^2.$$

Through integration by parts, we find for any $1 \leq j \leq s$:

$$\begin{aligned} \widehat{u}^{(j)}_k &= \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} u^{(j)}(x) e^{-ikx}\,\mathrm{d}x \\ &\overset{\text{(IbP)}}{=} -\frac{ik}{\sqrt{2\pi}} e^{-ikx} u^{(j-1)}(x)\Big|_0^{2\pi} + \frac{ik}{\sqrt{2\pi}} \int_0^{2\pi} u^{(j-1)}(x) e^{-ikx}\,\mathrm{d}x \\ &\overset{(*)}{=} \frac{ik}{\sqrt{2\pi}} \int_0^{2\pi} u^{(j-1)}(x) e^{-ikx}\,\mathrm{d}x \\ &= ik \widehat{u}^{(j-1)}_k. \end{aligned}$$

where $(*)$ uses $u^{(j-1)}(0) = u^{(j-1)}(2\pi)$, which is true since $0 \leq j - 1 \leq s - 1$. Hence, by finite induction, for any $k \neq 0$:

$$\widehat{u}_k = \widehat{u}^{(0)}_k = \frac{\widehat{u}^{(s)}_k}{(ik)^s}. \tag{4}$$

Finally, for any $N \geq 1$:

$$\begin{aligned}
\|u - P_N u\|_{L^2}^2 = \left\| \sum_{|k|>N} \widehat{u}_k \phi_k \right\|_{L^2}^2 &= \sum_{|k|>N} |\widehat{u}_k|^2 \\
&\overset{(4)}{=} \sum_{|k|>N} \frac{1}{k^{2s}} |\widehat{u}_k^{(s)}|^2 \\
&\leq N^{-2s} \sum_{|k|>N} |\widehat{u}_k^{(s)}|^2 \\
&\leq N^{-2s} \sum_{k\in\mathbb{Z}} |\widehat{u}_k^{(s)}|^2 \\
&= N^{-2s} \left\| u^{(s)} \right\|_{L^2}^2 \\
&\leq N^{-2s} \sum_{j=0}^{s} \left\| u^{(j)} \right\|_{L^2}^2 = N^{-2s} \|u\|_{H_p^s}^2
\end{aligned}$$

which proves the result.

b. One can very well proceed to compute the requested norms by numerical computation (involving some approximation). Some options include computing projection coefficients approximately using dense quadrature, or even using simple $(2N+1)$-point interpolation since question 5 below establishes that interpolation will produce essentially the same asymptotic error. However, here is a way to compute the exact L^2 error norms: First we observe the following: if f is a given function, and $s \geq 1$, then

$$f \in L^2, \;\; f(0) = f(2\pi), \;\; \text{and} f' \in H_p^{s-1} \quad \Longrightarrow f \in H_p^s,$$

which follows directly from the definition of H_p^s. Coming to the current problem, we see directly that $u_0 \in L^2 = H_p^0$. Furthermore, for $j \geq 1$:

$$u_j(2\pi) = c_j + \int_0^{2\pi} u_{j-1}(y)\, \mathrm{d}y \overset{(*)}{=} c_j = u_j(0),$$

where $(*)$ uses that u_{j-1} is a mean-0 function. Furthermore, $u_j(x)$ by definition is a continuous function on the closed interval $[0, 2\pi]$. Therefore, it has a finite maximum inside this interval, and hence is an L^2 function. Hence, by finite induction, $u_j \in H_p^j$. Hence, the computations from part (a) allow us to conclude for $k \in \mathbb{Z}\backslash\{0\}$:

$$u_j \in H_p^j \text{ and } u_j^{(j)}(x) = u_0(x) \;\; \Longrightarrow \;\; \widehat{u}_{j,k} = \frac{\widehat{u}_{j,0}}{(ik)^j},$$

where $\widehat{u}_{j,k}$ is the ϕ_k-Fourier expansion coefficient for u_j. Therefore, we need only compute the Fourier coefficients for u_0 in order to determine the exact coefficients for all the remaining functions. (Since u_j is mean-0 for every j, then $\widehat{u}_{j,0} = 0$.) We compute the

coefficients for u_0, $k \neq 0$, explicitly:

$$\begin{aligned}
\langle u_0, \phi_k \rangle &= \frac{1}{\sqrt{2\pi}} \left[\int_{\pi/2}^{3\pi/2} e^{-ikx} \, \mathrm{d}x - \int_{-\pi/2}^{\pi/2} e^{-ikx} \, \mathrm{d}x \right] \\
&= -\frac{1}{ik\sqrt{2\pi}} \left[e^{-ikx} \big|_{\pi/2}^{3\pi/2} - e^{-ikx} \big|_{-\pi/2}^{\pi/2} \right] \\
&= -\frac{4}{\sqrt{2\pi}} \frac{\sin(k\pi/2)}{k} \\
&= \begin{cases} \frac{-4(-1)^{(k-1)/2}}{k\sqrt{2\pi}}, & k \text{ odd} \\ 0, & k \text{ even} \end{cases}
\end{aligned}$$

Then this gives us the result:

$$\widehat{u}_{j,k} = \begin{cases} \frac{-4(-1)^{(k-1)/2}}{i^j k^{j+1} \sqrt{2\pi}}, & k \text{ odd} \\ 0, & k \text{ even} \end{cases}$$

which holds even for $k = 0$. In order to compute the error norms, we will use the fact that,

$$\|u - P_N u\|_{L^2}^2 = \|u\|_{L^2}^2 - \sum_{|k| \leq N} |\widehat{u}_k|^2 ,$$

which is the Pythagorean theorem $\|u\|^2 = \|P_N u\|^2 + \|(I - P_N)u\|^2$. Note that this furnishes an exact formula for the L^2 error if we know the L^2 norms of u_j along with a finite number of expansions coefficients (which we just computed above). In order to compute the exact L^2 norms, instead of taking antiderivatives to compute formulas for u_j, we utilize Parseval's identity:

$$\|u_j\|_{L^2}^2 = \sum_{k \in \mathbb{Z}} |\widehat{u}_{j,k}|^2 = 2 \sum_{\ell=1}^{\infty} \frac{16}{2\pi(2\ell - 1)^{2j+2}} = \frac{16}{\pi} \sum_{\ell=1}^{\infty} \frac{1}{(2\ell - 1)^{2j+2}}$$

In order to compute this infinite series, note that

$$\zeta(s) \coloneqq \sum_{n=1}^{\infty} \frac{1}{n^s},$$

is the Riemann Zeta function, and in particular for positive $s > 1$:

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \sum_{\ell=1}^{\infty} \frac{1}{(2\ell)^s} + \sum_{\ell=1}^{\infty} \frac{1}{(2\ell - 1)^s} = \frac{1}{2^s} \zeta(s) + \sum_{\ell=1}^{\infty} \frac{1}{(2\ell - 1)^s},$$

so that,

$$\sum_{\ell=1}^{\infty} \frac{1}{(2\ell - 1)^s} = \left[1 - \frac{1}{2^s} \right] \zeta(s).$$

Hence, the L^2 norms we wish to compute are,

$$\|u_j\|_{L^2}^2 = \frac{16}{\pi} \sum_{\ell=1}^{\infty} \frac{1}{(2\ell - 1)^{2j+2}} = \frac{16}{\pi} \left[1 - \frac{1}{2^{2j+2}} \right] \zeta(2j + 2).$$

The Riemann Zeta function in particular has the following known analytical values for even integers:

$$\zeta(2j+2) = \frac{(2\pi)^{2j+2}\,|B_{2j+2}|}{2(2j+2)!},$$

where B_n are the Bernoulli numbers, with the following relevant values:

$$|B_n| = \begin{cases} \frac{1}{6}, & n = 2 \\ \frac{1}{30}, & n = 4 \\ \frac{1}{42}, & n = 6 \\ \frac{1}{30}, & n = 8. \end{cases}$$

Putting everything together, we conclude:

$$\|u_j\|_{L^2}^2 = \frac{16(2\pi)^{2j+1}}{(2j+2)!}\left[1 - \frac{1}{2^{2j+2}}\right]|B_{2j+2}|.$$

Hence, to numerically test L^2 convergence of $P_N u_j$ as a function of N for $j = 0, 1, 2, 3$, we plot the (square root of the) following explicit errors:

$$\begin{aligned}\|u_j - P_N u_j\|_{L^2}^2 &= \frac{16(2\pi)^{2j+1}}{(2j+2)!}\left[1 - \frac{1}{2^{2j+2}}\right]|B_{2j+2}| - \sum_{|k|\le N} |\widehat{u}_{j,k}|^2 \\ &= \frac{16(2\pi)^{2j+1}}{(2j+2)!}\left[1 - \frac{1}{2^{2j+2}}\right]|B_{2j+2}| - \frac{16}{\pi}\sum_{\ell=1}^{\lceil N/2\rceil}\frac{1}{(2\ell-1)^{2j+2}} \qquad (5)\end{aligned}$$

We demonstrate convergence behavior using this formula in Figure 4. We see, as expected, that as j increases, the error decreases more quickly for increasing N. In particular, from the figure we also observe that the error appears to decay like,

$$\|u_j - P_N u_j\|_{L^2} \sim N^{-(j+1/2)},$$

so that at least empirically this suggests that $u_j \in H_p^{j+1/2}$. (It's actually $H_p^{j+1/2-\epsilon}$ for every $\epsilon > 0$.)

5. (Fourier interpolation)

a. For any $N \ge 1, k \in \mathbb{Z}$, prove that $I_N \phi_k = \phi_\ell$, where ℓ satisfying $|\ell| \le N$ is the modular restriction of k to $[-N, N]$:

$$\ell = \ell(k) \coloneqq k - (2N+1)j \in [-N, N], \qquad\qquad j \in \mathbb{Z}.$$

An equivalent definition: $\ell(k) = -N + [(k+N) \pmod{2N+1}]$. The operator I_N is defined on `D14-S10` as the $M = (2N+1)$-point Fourier interpolation operator, and ϕ_k is defined on slide `D13-S03(b)`.

b. If $u \in H_p^s$, prove that,

$$\|u - I_N u\|_{L^2} \lesssim N^{-s}\|u\|_{H_p^s},$$

where $a \lesssim b$ means that $a \le Cb$ for some constant C independent of N and u.

Solution:

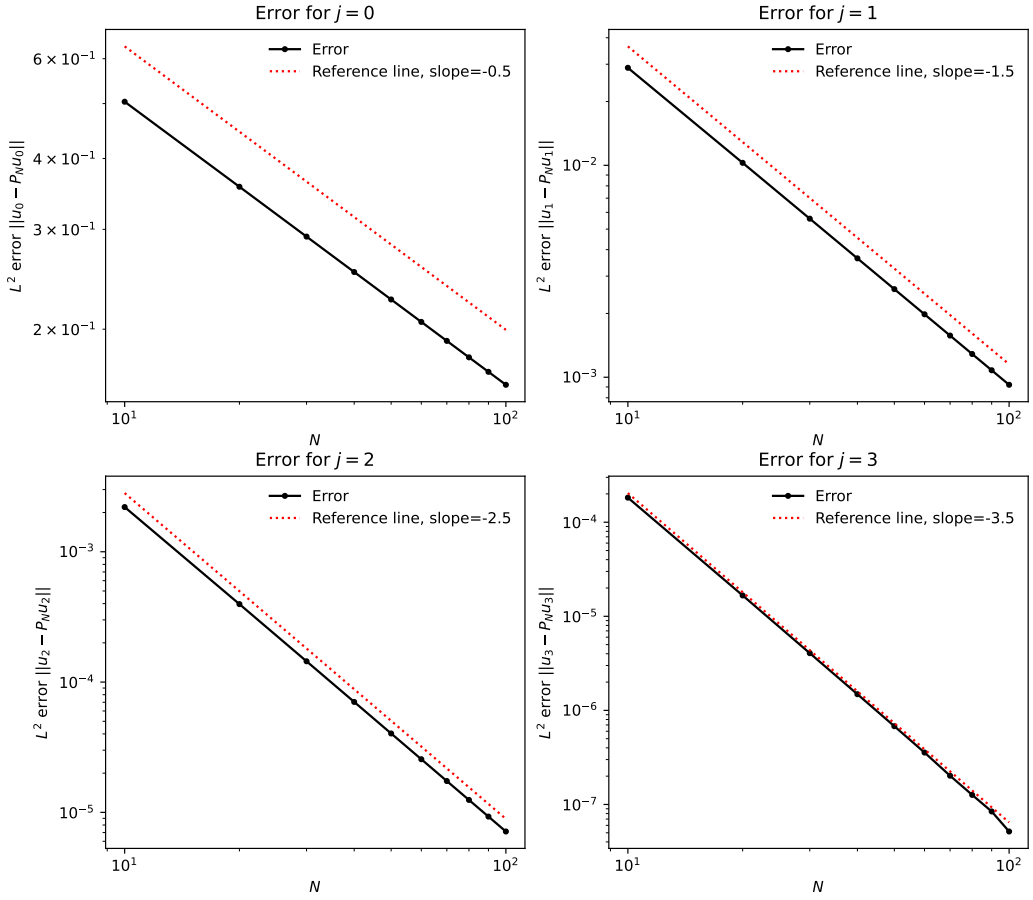

Figure 4: Problem 4: Fourier projection errors as a function of N for u_j, $j = 0, 1, 2, 3$. The errors are computed through the formula (5). Reference lines of given slopes are also plotted. One may also compute an approximate error using a truncated Parseval's sum or an interpolatory approximation to achieve essentially the same results.

a. If $|k| \leq N$, then $k = \ell$ and ϕ_k is in the range of I_N. Since I_N is a projector, then $I_N\phi_k = \phi_k = \phi_\ell$. For $|k| > N$, then there is some $j \in \mathbb{Z}$ such that

$$\ell = k - (2N+1)j$$

with $|\ell| \leq N$. With $x_m = 2\pi m/(2N+1)$, $m \in [2N+1]$ the interpolatory grid, note that,

$$\begin{aligned}\phi_k(x_m) = \frac{1}{\sqrt{2\pi}} e^{ikx_m} = \frac{1}{\sqrt{2\pi}} e^{i\ell x_m} e^{-i(2N+1)jx_m} &= \frac{1}{\sqrt{2\pi}} e^{i\ell x_j} e^{-i(2N+1)j2\pi m/(2N+1)} \\ &= \frac{1}{\sqrt{2\pi}} e^{i\ell x_j} e^{-i2\pi jm} = \frac{1}{\sqrt{2\pi}} e^{i\ell x_j} = \phi_\ell(x_m).\end{aligned}$$

Therefore, since I_N interpolates ϕ_k based only on its values at x_m and since these values equal those of ϕ_ℓ that in the range I_N, then $I_N\phi_k = \phi_\ell$. One can make this more formal: We have

$$I_N\phi_\ell = \phi_\ell = \sum_{|q|\leq N} b_q \phi_q I_N \phi_k \qquad\qquad = \sum_{|q|\leq N} c_q \phi_q$$

where we know that the coefficients $\boldsymbol{b}$ are such that $b_q = 1$ with $q = \ell$ and $b_q = 0$ otherwise, but also that $\boldsymbol{b}, \boldsymbol{c} \in \mathbb{C}^{2N+1}$ are given by solutions to the quadrature/interpolation problem:

$$\begin{aligned}\boldsymbol{c} &= \widetilde{\boldsymbol{V}}^* \phi_k(\boldsymbol{x}), \\ \boldsymbol{b} &= \widetilde{\boldsymbol{V}}^* \phi_\ell(\boldsymbol{x}),\end{aligned}$$

where $\widetilde{\boldsymbol{V}}$ is the scaled Fourier Vandermonde-like (DFT) matrix on slides `D14-S04(b)`, and $\boldsymbol{x} = (x_m)_{m\in[2N+1]}$. Since we have already shown that $\phi_k(\boldsymbol{x}) = \phi_\ell(\boldsymbol{x})$, then $\boldsymbol{c} = \boldsymbol{b}$, implying $I_N\phi_k = I_N\phi_\ell = \phi_\ell$.

b. Let $u = \sum_{k\in\mathbb{Z}} \widehat{u}_k \phi_k$. Since the interpolation operator is linear, we have,

$$I_N u = \sum_{k\in\mathbb{Z}} \widehat{u}_k I_N \phi_k = \sum_{k\in\mathbb{Z}} \widehat{u}_k \phi_{\ell(k)} = \sum_{|\ell|\leq N} \left(\sum_{j\in\mathbb{Z}} \widehat{u}_{\ell+j(2N+1)} \right) \phi_\ell$$

Therefore,

$$I_N u = \sum_{|k|\leq N} (\widehat{u}_k + a_k)\phi_k, \qquad\qquad a_k = \sum_{j\in\mathbb{Z}\setminus\{0\}} \widehat{u}_{k+j(2N+1)}$$

Then the error between u and its interpolatory approximation, using the Pythagorean theorem, is,

$$\begin{aligned}\|u - I_N u\|_{L^2}^2 &\leq \|u - P_N u\|_{L^2}^2 + \|P_N u - I_N u\|_{L^2}^2 \\ &\overset{(*)}{\leq} N^{-2s} \|u\|_{H_p^s}^2 + \left\| \sum_{|k|\leq N} (\widehat{u}_k - (\widehat{u}_k - a_k))\phi_k \right\|^2 \\ &= N^{-2s} \|u\|_{H_p^s}^2 + \sum_{|k|\leq N} |a_k|^2\end{aligned}$$

where $(*)$ uses the result from problem 3(a). We have shown in problem 3(a) that so long as $u \in H^s_p$, then

$$\widehat{u}_k = \frac{\widehat{u}^{(s)}_k}{(ik)^s},$$

where $\widehat{u}^{(s)}_k$ are the Fourier expansion coefficients for $u^{(s)} \in L^2$. Using this formula for $\widehat{u}_k$, we can bound the a_k coefficients; for example, if $k \leq 0$:

$$\begin{aligned}
|a_k|^2 = \left| \sum_{j \in \mathbb{Z}\setminus\{0\}} \widehat{u}_{k+j(2N+1)} \right|^2 &= \left| \sum_{j \in \mathbb{Z}\setminus\{0\}} \frac{\widehat{u}^{(s)}_{k+j(2N+1)}}{(k+j(2N+1))^s} \right|^2 \\
&\overset{(*)}{\leq} \left(\sum_{j \in \mathbb{Z}\setminus\{0\}} \left| \widehat{u}^{(s)}_{k+j(2N+1)} \right|^2 \right) \left(\sum_{j \in \mathbb{Z}\setminus\{0\}} \frac{1}{|k+j(2N+1)|^{2s}} \right) \\
&\overset{k\leq 0}{\leq} \left(\sum_{j \in \mathbb{Z}\setminus\{0\}} \left| \widehat{u}^{(s)}_{k+j(2N+1)} \right|^2 \right) \left(2 \sum_{j \in \mathbb{N}} \frac{1}{(k+j(2N+1))^{2s}} \right) \\
&\overset{k\geq -N}{\leq} \left(\sum_{j \in \mathbb{Z}\setminus\{0\}} \left| \widehat{u}^{(s)}_{k+j(2N+1)} \right|^2 \right) \left(2 \sum_{j \in \mathbb{N}} \frac{1}{(jN)^{2s}} \right) \\
&= \left(\sum_{j \in \mathbb{Z}\setminus\{0\}} \left| \widehat{u}^{(s)}_{k+j(2N+1)} \right|^2 \right) \left(2N^{-2s} \sum_{j \in \mathbb{N}} \frac{1}{j^{2s}} \right) \\
&= 2N^{-2s} \zeta(2s) \left(\sum_{j \in \mathbb{Z}\setminus\{0\}} \left| \widehat{u}^{(s)}_{k+j(2N+1)} \right|^2 \right),
\end{aligned}$$

where $(*)$ is the Cauchy-Schwarz inequality (for sequences in $\ell^2(\mathbb{N})$), and $\zeta(\cdot)$ is the Riemann zeta function, which for real inputs $s > 1$ is defined as,

$$\zeta(s) \coloneqq \sum_{n \in \mathbb{N}} \frac{1}{n^s}.$$

Note in particular that $\zeta(2s)$ is a finite number for any real $s \geq 1$. An analogous computation (resulting in the same bound) can be accomplished if $k > 0$. Therefore we have,

$$\begin{aligned}
\|u - I_N u\|^2_{L^2} &\leq N^{-2s} \|u\|^2_{H^s_p} + \sum_{|k| \leq N} |a_k|^2 \\
&\leq N^{-2s} \|u\|^2_{H^s_p} + 2\zeta(2s) N^{-2s} \sum_{|k| \leq N} \sum_{j \in \mathbb{Z}\setminus\{0\}} \left| \widehat{u}^{(s)}_{k+j(2N+1)} \right|^2 \\
&= N^{-2s} \|u\|^2_{H^s_p} + 2\zeta(2s) N^{-2s} \sum_{|k| \leq N} \|u^{(s)}\|_{L^2} \\
&\leq N^{-2s} \|u\|^2_{H^s_p} + 2\zeta(2s) N^{-2s} \|u^{(s)}\|_{L^2} \\
&\leq N^{-2s} \|u\|^2_{H^s_p} + 2\zeta(2s) N^{-2s} \|u\|^2_{H^s_p} \\
&= (1 + 2\zeta(2s)) N^{-2s} \|u\|^2_{H^s_p},
\end{aligned}$$

which proves the result.