

## 1 Z-Transform Relation to Fourier Transform

In this problem we study how the grid scaled Z-transform of the discretized solution approaches the Fourier transform of the continuous solution as the space increment  $h \rightarrow 0$ . Recall that if  $u_j = U(jh)$ , the Z-transform  $\hat{u}(\xi)$  is related to the Fourier transform  $\hat{U}(\kappa) = \int_{-\infty}^{\infty} U(x)e^{-ix\kappa} dx$  via

$$h\hat{u}(h\kappa) = \sum_{m=-\infty}^{\infty} \hat{U}\left(\kappa - \frac{2\pi}{h}m\right).$$

Suppose that for  $\kappa \in \mathbb{R}$ ,  $\hat{U}(\kappa)$  satisfies

- (a)  $|\hat{U}(\kappa)| \leq \frac{C_r}{1+|\kappa|^r}$ ,
- (b)  $|\hat{U}(\kappa)| \leq Ce^{-\rho|\kappa|}$ ,

where  $r \geq 2$ ,  $C_r \geq 0$ ,  $C > 0$  and  $\rho > 0$  are constants. Show that for  $|\kappa| \leq \pi/h$ ,  $|h\hat{u}(h\kappa) - \hat{U}(\kappa)|$  is bounded by

- (a)  $\frac{C_r h^r}{4\pi^{r-2}}$
- (b)  $\frac{2Ce^{-\rho\pi/h}}{1-e^{-2\rho\pi/h}}$ .

**Solution:**

- (a) First, note the fact that  $|k \pm \frac{2\pi}{h}m| \geq \frac{2\pi m - \pi}{h}$  and  $\sum_{m=1}^{\infty} \frac{1}{(2m-1)^2} = \frac{\pi^2}{8}$ . Using these facts, we achieve the first upper bound as follows:

$$\begin{aligned} |h\hat{u}(h\kappa) - \hat{U}(\kappa)| &= \left| \sum_{m=-\infty}^{\infty} \hat{U}\left(\kappa - \frac{2\pi}{h}m\right) - \hat{U}(\kappa) \right| \\ &= \left| \sum_{m \neq 0} \hat{U}\left(\kappa - \frac{2\pi}{h}m\right) \right| \\ &\leq \sum_{m \neq 0} \left| \hat{U}\left(\kappa - \frac{2\pi}{h}m\right) \right| \\ &\leq \sum_{m \neq 0} \frac{C_r}{1 + \left| \kappa - \frac{2\pi}{h}m \right|^r} \\ &\leq \sum_{m=1}^{\infty} \frac{2C_r}{1 + \left( \frac{2\pi m - \pi}{h} \right)^r} \\ &= \sum_{m=1}^{\infty} \frac{2C_r h^r}{1 + \pi^r (2m-1)^r} \\ &\leq \frac{2C_r h^r}{\pi^r} \sum_{m=1}^{\infty} \frac{1}{(2m-1)^r} \\ &\leq \frac{2C_r h^r}{\pi^r} \sum_{m=1}^{\infty} \frac{1}{(2m-1)^2} \\ &= \boxed{\frac{C_r h^r}{4\pi^{r-2}}}. \end{aligned}$$

(b) For the second upper bound, we start with the initial simplifications from the previous part as follows:

$$\begin{aligned}
 |h\hat{u}(h\kappa) - \hat{U}(\kappa)| &\leq \sum_{m \neq 0} \left| \hat{U}\left(\kappa - \frac{2\pi}{h}\right) \right| \\
 &\leq \sum_{m \neq 0} C e^{-\rho|\kappa - 2\pi m/h|} \\
 &\leq 2C \sum_{m=1}^{\infty} e^{-\rho(2\pi m - \pi)/h} \\
 &= 2C e^{\rho\pi/h} \sum_{m=1}^{\infty} \left(e^{-2\rho\pi/h}\right)^m \\
 &= \frac{2C e^{\rho\pi/h}}{1 - e^{-2\rho\pi/h}} e^{-2\rho\pi/h} \\
 &= \boxed{\frac{2C e^{-\rho\pi/h}}{1 - e^{-2\rho\pi/h}}}
 \end{aligned}$$

## 2 Circulant Matrix Properties

A circulant matrix is constant along diagonals with entries that "wrap around":

$$A_{jk} = \begin{cases} r_{k-j} & k \geq j, \\ r_{N+k-j} & k < j \end{cases}.$$

For convenience, we will index our matrices starting at zero. Show that  $AU = U\Lambda$ , with

$$U_{jk} = e^{2\pi i j k / N}, \quad \begin{pmatrix} 0 \leq j \leq N-1 \\ 0 \leq k \leq N-1 \end{pmatrix}, \quad \begin{pmatrix} \lambda_0 & & \\ & \ddots & \\ & & \lambda_{N-1} \end{pmatrix}$$

and  $\lambda_k = \sum_{\ell=0}^{N-1} r_\ell e^{2\pi i \ell k / N}$ .

**Solution:**

We proceed by showing that  $(AU)_{jk} = (U\Lambda)_{jk}$ . We compute the right hand side first:

$$\begin{aligned} (U\Lambda)_{jk} &= \lambda_k e^{2\pi i j k / N} \\ &= \sum_{\ell=0}^{N-1} r_\ell e^{2\pi i (\ell+j) k / N}. \end{aligned}$$

Now consider the left hand side:

$$\begin{aligned} (AU)_{jk} &= \sum_{\ell=0}^{N-1} A_{j\ell} U_{\ell k} \\ &= \sum_{\ell=0}^{N-1} A_{j\ell} e^{2\pi i \ell k / N} \end{aligned}$$

The value of  $A_{j\ell}$  depends on the relationship between  $j$  and  $\ell$  so we split up the sum into corresponding parts:

$$\begin{aligned} \sum_{\ell=0}^{N-1} A_{j\ell} e^{2\pi i \ell k / N} &= \sum_{\ell=0}^{j-1} A_{j\ell} e^{2\pi i \ell k / N} + \sum_{\ell=j}^{N-1} A_{j\ell} e^{2\pi i \ell k / N} \\ &= \sum_{\ell=0}^{j-1} r_{N+\ell-j} e^{2\pi i \ell k / N} + \sum_{\ell=j}^{N-1} r_{\ell-j} e^{2\pi i \ell k / N} \\ &= \sum_{\ell=N-j}^{N-1} r_\ell e^{2\pi i (\ell+j-N) k / N} + \sum_{\ell=0}^{N-j-1} r_\ell e^{2\pi i (\ell+j) k / N} \\ &= \sum_{\ell=N-j}^{N-1} r_\ell e^{2\pi i (\ell+j) k / N} + \sum_{\ell=0}^{N-j-1} r_\ell e^{2\pi i (\ell+j) k / N} \\ &= \sum_{\ell=0}^{N-1} r_\ell e^{2\pi i (\ell+j) k / N}. \end{aligned}$$

The expressions for  $(AU)_{jk}$  and  $(U\Lambda)_{jk}$  are indeed equal so we conclude that  $AU = U\Lambda$ .

### 3 Crank-Nicolson

Use the Crank-Nicolson scheme to solve the equation

$$u_t = \alpha u_{xx} - \beta u_x, \quad \alpha = \frac{1}{512}, \quad \beta = \frac{33}{32} \quad (\star)$$

on the interval  $[0, 1]$  with periodic boundary conditions and initial conditions

$$u(x, 0) = g(x), \quad g(x) = (\sin \pi x)^{100}.$$

**Solution:**

For the spatial discretizations, we replace the right-hand side of  $(\star)$  by  $\alpha D_x^+ D_x^- u - \beta D_x^0 u$  and then use the trapezoidal rule in time. The trapezoidal rule is the approximation  $(x_{n+1} - x_n)/h \approx (f_n + f_{n+1})/2$ . Applying this method yields the following finite difference:

$$D_t^+ u^n = \frac{1}{2}(\alpha D_x^+ D_x^- u^n + \alpha D_x^+ D_x^- u^{n+1} - \beta D_x^0 u^n - \beta D_x^0 u^{n+1})$$

Consider  $k \in \mathbb{R}, M \in \mathbb{Z}, h = \frac{1}{M}$ , and operators  $A, B$  defined as follows:

$$\begin{aligned} Au_j &= u_{j+1} - u_{j-1} \\ Bu_j &= u_{j+1} - 2u_j + u_{j-1} \end{aligned}$$

We can then rewrite the finite difference as

$$\frac{u^{n+1} - u^n}{k} = \frac{\alpha}{2h^2}(Bu^n + Bu^{n+1}) - \frac{\beta}{4h}(Au^n - Au^{n+1}).$$

We then bring the  $u^{n+1}$  terms to one side to yield an implicit update step

$$\left( I + \frac{\beta k}{4h} A - \frac{\alpha k}{2h^2} B \right) u^{n+1} = \left( I - \frac{\beta k}{4h} A + \frac{\alpha k}{2h^2} B \right) u^n$$

The last part of the scheme construction is accounting for the periodic boundary conditions. We can do this by creating a ghost node  $u_{-1}$  where  $u_{-1} = u_M = u_0$ . The ghost node will then slightly change the update steps for  $u_0$  and  $u_{M-1}$ . Note that the nodes for  $u_0$  and  $u_M$  can be treated as the same. We redefine  $A$  and  $B$  correspondingly to act on  $\mathbb{R}_h^M$  while accounting for the boundary conditions.

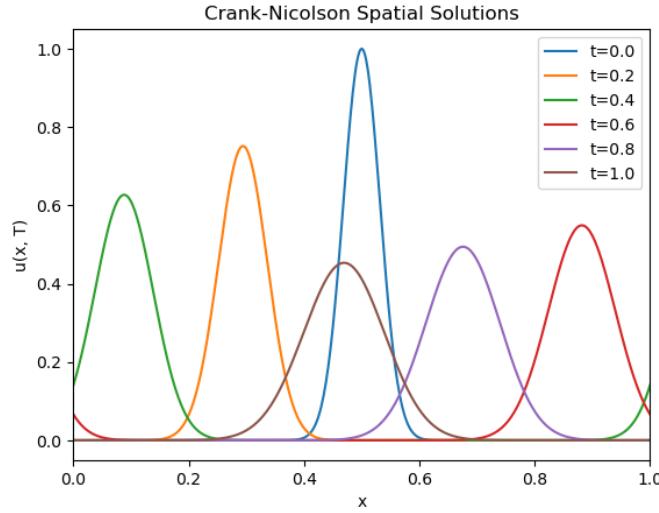
$$\begin{aligned} Au_0 &= u_1 - u_{M-1} \\ Au_{M-1} &= u_0 - u_{M-2} \\ Bu_0 &= u_1 - 2u_0 + u_{M-1} \\ Bu_{M-1} &= u_0 - 2u_{M-1} + u_{M-2} \end{aligned}$$

Shown below are the matrix representations of the operators  $A, B \in \mathbb{R}_h^{M \times M}$ .

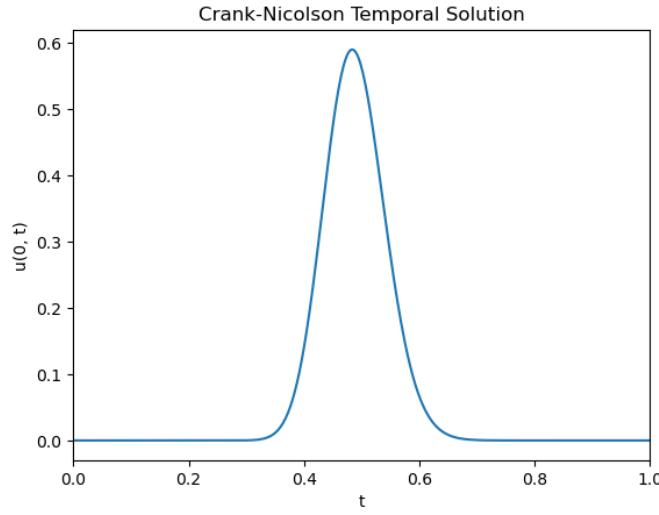
$$A = \begin{pmatrix} 0 & 1 & \cdots & 0 & -1 \\ -1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & \cdots & -1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} -2 & 1 & \cdots & 0 & 1 \\ 1 & -2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -2 & 1 \\ 1 & 0 & \cdots & 1 & -2 \end{pmatrix}$$

We discuss results and analysis of the Crank-Nicolson scheme below.

- (a) We first solve the PDE for various time steps on the entire interval  $[0, 1]$ . These solutions are plotted below. Observe that the solution is both translational and dissipative.



- (b) We can gain further insights by looking at the solution trajectory over time for  $u(0, t)$  where  $0 \leq t \leq 1$ . This is again plotted below. The translational effect is seen by the increasing and then decreasing value at  $x = 0$ .



- (c) The amplification factor of the scheme is found via the  $Z$ -transform of the update step. We compute this below.

$$u^{n+1} = \underbrace{\left( I + \frac{\beta k}{4h} A - \frac{\alpha k}{2h^2} B \right)^{-1}}_{\mathcal{B}} \left( I - \frac{\beta k}{4h} A + \frac{\alpha k}{2h^2} B \right) u^n$$

$$\hat{u}^{n+1}(\xi) = \underbrace{\frac{1 - \frac{\beta k}{4h} G_A(\xi) + \frac{\alpha k}{2h^2} G_B(\xi)}{1 + \frac{\beta k}{4h} G_A(\xi) - \frac{\alpha k}{2h^2} G_B(\xi)}}_{G(\xi)} \hat{u}^n(\xi)$$

Here  $G_A(\xi)$  and  $G_B(\xi)$  are the  $Z$ -transforms of  $A$  and  $B$ , and  $G(\xi)$  is the amplification factor.

$$G_A(\xi) = e^{i\xi} - e^{-i\xi} = 2i \sin \xi$$

$$G_B(\xi) = e^{i\xi} - 2 + e^{-i\xi} = 2 \cos \xi - 2 = -4 \sin^2 \frac{\xi}{2}$$

Putting these terms together, we can explicitly write the amplification factor.

$$G(\xi) = \frac{1 - \frac{2\alpha k}{h^2} \sin^2 \frac{\xi}{2} - \frac{\beta k}{2h} i \sin \xi}{1 + \frac{2\alpha k}{h^2} \sin^2 \frac{\xi}{2} + \frac{\beta k}{2h} i \sin \xi}$$

Now define variables  $a = \frac{2\alpha k}{h^2} \sin^2 \frac{\xi}{2}$  and  $b = \frac{\beta k}{2h} \sin \xi$ . Note that  $a, b \in \mathbb{R}$  and  $a \geq 0$ . We then plug the expressions for  $G_A(\xi)$  and  $G_B(\xi)$  into the earlier expression to upper bound the amplification factor.

$$\begin{aligned} |G(\xi)| &= |\hat{u}^{n+1}(\xi)| \\ &= \frac{|1 - a - bi|}{|1 + a + bi|} \\ |1 - a - bi| &= |1 - a + bi| \\ &= |1 + a + bi| \\ |G(\xi)| &\leq \frac{|1 + a + bi|}{|1 + a + bi|} \\ &= 1 \end{aligned}$$

Furthermore,  $G(0) = 1$ , so the upper bound is achievable on the finite interval and  $\|\mathcal{B}\|_{\ell_h^2} = \|G\|_\infty = 1$ . Since 1 is an upper bound for  $\|\mathcal{B}\|_{\ell_h^2}$ , and in fact an equality, we conclude that the scheme is **unconditionally stable**.

(d) Consider the equation  $v(x, t) = u(x + \beta t, t)$ . Consider the partial  $v_x$ .

$$\begin{aligned} v_t(x, t) &= \beta u_x(x + \beta t, t) + u_t(x + \beta t, t) \\ &= \beta v_x(x, t) + u_t(x + \beta t, t) \\ &= \beta v_x(x, t) + \alpha u_{xx}(x + \beta t, t) - \beta u_x(x + \beta t, t) \\ &= \cancel{\beta v_x(x, t)} + \alpha v_{xx}(x, t) - \cancel{\beta u_x(x, t)} \\ &= \alpha v_{xx}(x, t) \end{aligned}$$

This yields the following PDE:  $v_t = \alpha v_{xx}, \quad v(x, 0) = g(x)$ . The definition of  $v(x, t)$  is just a shift of the  $x$  coordinate by a scalar multiple of  $t$  so it will also have periodic boundary conditions.

(e) We now derive an exact formula for  $v(x, t)$ . First, we expand  $g(x)$  in terms of its Fourier coefficients.

$$\begin{aligned} g(x) &= (\sin \pi x)^{100} \\ &= \left( \frac{1}{2i} \right)^{100} (e^{\pi ix} - e^{-\pi ix})^{100} \\ &= \frac{1}{2^{100}} \sum_{k=0}^{100} \binom{100}{k} (-1)^{100-k} e^{2\pi ikx} e^{-100\pi ix} \\ &= \frac{1}{2^{100}} \sum_{k=-50}^{50} \binom{100}{k+50} (-1)^{k+50} e^{2\pi ikx} \end{aligned}$$

Thus, the Fourier coefficients of  $g(x)$  are  $c_k = \binom{100}{k+50} (-1)^{k+50} 2^{-100}$  for  $-50 \leq k \leq 50$ . Observe that the PDE is separable so basis elements of the solution space will be of the form  $v_k(x, t) = T_k(t)X_k(t)$ . By observation of the initial condition and the periodicity of  $v(x, t)$ , we can write the general form as

$$v(x, t) = \sum_{k=-50}^{50} c_k T_k(t) e^{2\pi i k x}$$

where each  $c_k$  is the coefficient of a basis element. We apply the PDE to an arbitrary  $k$ th term to calculate  $T_k$  since each term is in the solution space.

$$\begin{aligned} c_k T'_k(t) e^{2\pi i k t} &= c_k \alpha T_k(t) (-4\pi^2 k^2) e^{2\pi i k t} \\ T'_k(t) &= (-4\alpha\pi^2 k^2) T_k(t) \end{aligned}$$

If we take the ansatz  $T_k(t) = e^{\lambda_k t}$ , we see that  $\lambda_k = -4\alpha\pi^2 k^2$ . Thus, the exact solution for  $v(x, t)$  expressed in terms of the Fourier coefficients  $c_k$  is

$$v(x, t) = \sum_{k=-50}^{50} c_k e^{-4\alpha\pi^2 k^2 t} e^{2\pi i k x}$$

From our discussion earlier, we know  $u(x, t) = v(x - \beta t, t)$ , so we can recover the solution to original PDE.

$$u(x, t) = \sum_{k=-50}^{50} c_k e^{-(4\alpha\pi^2 k^2 + 2\beta\pi k)t} e^{2\pi i k x}$$

While we could compute  $c_k$  explicitly with the exact formula found earlier, this can be computationally challenging due to floating point errors. Instead, we demonstrate an alternative approximate method. First, sample  $g(x)$  on a uniform grid with  $M = 128$  grid-points and apply the DFT. The DFT yields coefficients

$$\tilde{g}_k = \sum_{j=0}^{M-1} g_j e^{-2\pi i k x_j}$$

where  $x_j = j/M$  are the uniform sample positions. The coefficients  $\tilde{g}_j$  relate to  $c_k$  by the following inverse transform:

$$g_j = \frac{1}{M} \sum_{k=0}^{M-1} \tilde{g}_k e^{2\pi i k x_j}.$$

However, we need coefficients between  $-50$  and  $50$ . This can be done by shifting the indices cyclically. Here we assume  $M$  is even. Then,

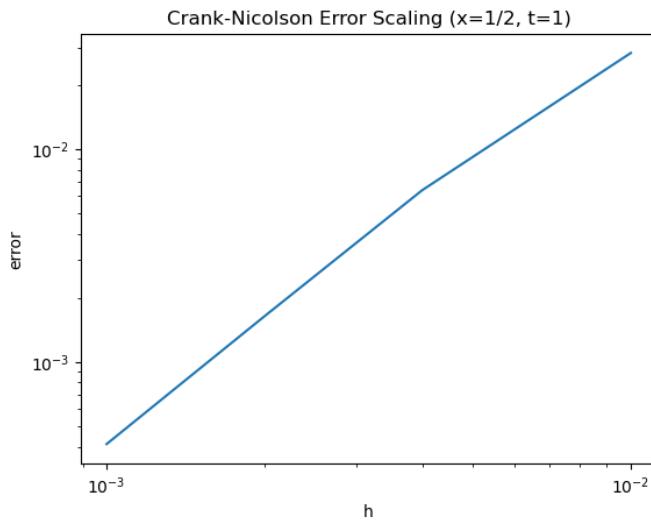
$$\begin{aligned} g_j &= \frac{1}{M} \sum_{k=0}^{M-1} \tilde{g}_k e^{2\pi i k j / M} \\ &= \frac{1}{M} \sum_{k=0}^{M-1} \tilde{g}_k e^{2\pi i (k-M) j / M} \\ &= \frac{1}{M} \sum_{k=-M/2}^{-1} \tilde{g}_{k+M} e^{2\pi i k j / M} + \frac{1}{M} \sum_{k=0}^{M/2-1} \tilde{g}_k e^{2\pi i k j / M} \end{aligned}$$

Finally, we can match terms to get approximations for the Fourier coefficients.

$$c_k = \begin{cases} \tilde{g}_{k+M}/M, & -M/2 \leq k \leq -1 \\ \tilde{g}_k/M, & 0 \leq k \leq M/2 - 1 \end{cases}$$

Note that we only need  $-50 \leq k \leq 50$  so  $M = 128 > 50 \times 2$  is sufficiently large. We also compute the exact solutions for  $c_k$  as a baseline by using the combinatorial formula described earlier. With the given selection for  $M$ , the corresponding relative error is less than  $10^{-15}$ .

- (f) We can now compute  $u(x, t)$  directly by using the exact form in part (e) and the DFT approximations of  $c_k$ . As an example,  $u(1/2, 1) = 0.41023088876098807$ .
- (g) We can now use our solution for  $u(1/2, 1)$  in part (f) to compute the error of the Crank-Nicolson method and show its order of accuracy. Let the error be  $e_h = |u_h(1/2, 1) - u(1/2, 1)|$  where  $u_h(x, t)$  is the numerical solution from Crank-Nicolson with the refinement path  $k = h$ . The truncation error of the method is  $O(k^2 + h^2)$  so we expect  $O(h^2)$  error.



The line plot above has a slope of  $\approx 2$ , thus verifying the expected second order error convergence.