QUEEN'S UNIVERSITY BELFAST

Investigation of Numerical Methods for the Diffusion & Wave Equation for a Square Wave Initial Condition Under Dirichlet Boundary Conditions

Emmet Rice

Introduction

The Diffusion and Wave equations are two of the core partial differential equations in mathematics and have countless applications and relevance across equally many disciplines, such as engineering, finance and practically every science. In most practical circumstances these differential equations are too complex to be solved analytically and require numerical methods to approximate and quantify their behaviour. This report aims to investigate numerical methods for solving the Diffusion equation and the Wave equation for a square wave initial condition when subjected to Dirichlet boundary conditions. MATLAB was chosen as the tool to conduct the numerical analysis in this report. Two approaches are used for this investigation; the first approximates the explicit analytical solution by computationally calculating and summing the Fourier series coefficients from the analytical solution. The other numerical method employed in this investigation is the finite difference method.

Conditionality

Both the Diffusion and Wave models investigated are subject to Dirichlet boundary conditions, which set the dependant parameter "U" at the boundaries of the relevant space itself to always be equal to zero. For the Diffusion equation, the model chosen to be analysed was how the temperature of a one-spatial-dimensional rod evolved from its initial square wave temperature profile over time. For such a system under Dirichlet boundary conditions, it is physically equivalent to having refrigeration mechanisms at either end of the rod which force the temperature to be zero, while allowing heat to flow out of the system. As the bar is one-dimensional, heat can only flow out of the system at the ends of the rod. From a basic understanding of the diffusion equation, the second law of thermodynamics, and empirically that high concentration elements typically flow to areas of low concentration, the temperature of the rod will drop to zero over time. For a rod of length L, and temperature U, the Dirichlet conditions can then be defined by: U(0,t)=0=U(L,t). For this investigation the length L is equal to 2π.

For the Wave equation, one analogous model would be string with no external forces, such as gravity or friction, which is fixed at both ends and held initially as a square wave before being released. In this case, U can be thought of as the vertical displacement of the string.

Examples for two other common boundary conditions are discussed below.

Von-Neumann boundary conditions set the spatial derivative dependant variable U to zero at the spatial limits ($U_x(0,t) = 0 = U_x(L,t)$). For the same Diffusion model outlined above, this is

physically analogous to insulating the rod, preventing heat from flowing out of the system. The entropy of such a closed system will increase, and logically the temperature should then distribute evenly along the bar.
For the Wave equation string model, the string can be imagined attached to a frictionless vertical slider at either ends. A more common model would be wind instrument, open at both ends, producing sound wave harmonics, where U is the displacement of the sound wave.

Periodic Boundary conditions force whatever flows out of one boundary to flow back in to the system from the other side. This physically has the effect of "looping" the bar back on itself, or in higher dimensions forming a ring or toroid shape. This also creates a closed system, and thus temperature should also distribute evenly along the bar. Similar concepts apply for the wave equation.

Analytic Solution by separation of variables

Diffusion Equation

Equation 1, Diffusion Equation

$$\frac{\partial U}{\partial t} = K\frac{\partial^2 U}{\partial x^2} = 0$$

Where K is constant

Aside: When considering temperature, K is the temperature conductivity. K is set as a constant in this report in order to simplify both the analytical and numerical solutions. For Diffusion, K is set as 2. The length of the rod "L" is also set as 2π.

Particular Problem:

$$\frac{\partial U}{\partial t} = 2\frac{\partial^2 U}{\partial x^2} = 0$$

For: $0 \leq x \leq 2\pi \qquad t \geq 0$

Boundary Conditions: $U(0,t) = 0 = U(2\pi,t)$

Initial Condition: $U(x,0) = f(x)$ *where f(x) is a function defining the initial profile of U*

$$f(x) = \begin{cases} 1, & |x-\pi| \leq 1 \\ 0, & otherwise \end{cases}$$

For the separation of variables, seek the solution in the form:

$$U(x,t) = V(x)q(t)$$

The dependent parameter is separated into functions which themselves are only dependant on one variable.

Substituting in to the diffusion equation, and diving across by $U(x,t)$ yields:

$$\frac{1}{Kq}\frac{dq}{dt} - \frac{1}{V}\frac{d^2V}{dx^2} = 0$$

$$\frac{1}{Kq}\frac{dq}{dt} = \frac{1}{V}\frac{d^2V}{dx^2}$$

Now the Partial differential equation is fully separated into functions of only t and x respectively. From the equality above, each can be set equal to an arbitrary constant $-\lambda$ and solved for.

$$Temporal: \frac{1}{Kq}\frac{dq}{dt} = -\lambda \qquad Spatial: \frac{1}{V}\frac{d^2V}{dx^2} = -\lambda$$

Diffusion Spatial Component

Solving for the spatial component, as we can latter apply the Boundary Condition:

$$\frac{d^2V}{dx^2} + V\lambda = 0$$

Using the auxiliary / characteristic equation:

Equation 2, Characteristic Equation

$$am^2 + bm + c = 0$$

Where m terms correspond to the order of the derivative of V, and a = 1, b = 0, c = λ the coefficients

Substituting and solving:

$$(1)m^2 + (0)m + \lambda = 0$$

$$m^2 = \lambda \quad \rightarrow \quad m = \pm\sqrt{-\lambda}$$

This leads to 3 cases depending on the value of λ

Case 1) $\boldsymbol{\lambda < 0}$

Where $w > 0$ is a "Dummy variable" let: $\lambda = -w^2$ $m = \pm\sqrt{-\lambda} = \pm w$

The solution to V must have 2 real distinct roots, and a trial solution of the form:

$$V(x) = Ae^{wx} + Be^{-wx}$$

Where A and B are arbitrary constants

At this point, the boundary conditions must be applied to determine the coefficients A and B, and the validity of the case. For this Project the Dirichlet boundary conditions were applied, but for instance the Von-Neumann conditions could be applied here instead.

$$V(0) = V(2\pi) = 0$$

$$V(0) = Ae^{w(0)} + Be^{-w(0)} = A + B = 0$$

$$\therefore A = -B$$

$$V(2\pi) = B\left(e^{-w(2\pi)} - e^{w(2\pi)}\right) = 0$$

For non-trivial solution, term in brackets must equal zero and $B \neq 0$

Using: $\sinh(x) = \left(\frac{e^{x}-e^{-x}}{2}\right) \rightarrow e^{-x} - e^{x} = -2\sinh(x)$

-2sinh(x) is only 0 when x = 0; Therefore, $V(2\pi) = 0$ only when w = 0

This would require $-w^2 = 0$, thus a contradiction with the initial statement $\lambda < 0$. Only trivial solution exists, and case 1 is invalid.

Case 2) $\boldsymbol{\lambda = 0}$

The solution to V must have 1 real repeated, and a trial solution of the form:

$$V(x) = (A + Bx)e^{mx}$$

Where A and B are arbitrary constants

Apply Boundary Condition:

$$V(0) = (A + 0)(1) = 0$$

$$\therefore A = 0$$

$$V(2\pi) = B(2\pi)e^{m2\pi} = B(2\pi)e^{(0)2\pi} = B(2\pi) = 0$$

$$\therefore B = 0$$

Only trivial solution exists, case 2 is invalid.

Case 3) $\boldsymbol{\lambda > 0}$

Where $w > 0$ is a "Dummy variable" let: $\lambda = w^2$ $m = \pm\sqrt{-\lambda} = \pm\sqrt{-w^2}$

m is 2 distinct complex roots: $m = \alpha \pm i\beta = \pm iw$

Trial solution:

$$V(x) = e^{\alpha}\,(C_1\cos(\beta x) + C_2\sin(\beta x))$$

Where C_1 and C_2 are arbitrary constants

$$V(x) = C_1\cos(wx) + C_2\sin(wx)$$

Apply Dirichlet Boundary Condition:

$$V(0) = C_1\cos(0) + C_2\sin(0) = 0$$

$$V(0) = C_1(1) = 0$$

$$\therefore C_1 = 0$$

$$V(2\pi) = C_2\sin(w2\pi) = 0$$

For non-trivial solution, sine term must equal zero and $C_2 \neq 0$

$$\therefore w2\pi = N\pi$$

Where $N = \pm 1,2,3, \ldots.$

And as $w > 0$:

$$w = \frac{n}{2} = \sqrt{\lambda}$$

Where $n = 1,2,3, \ldots.$

Thus, the spatial component has infinitely many solutions as there are infinitely many valid eigenvalues and eigenfunctions

Equation 3, Diffusion Spatial Component Solution

$$V_n(2\pi) = C_{2n} \sin\left(\frac{nx}{2}\right)$$

Where $n = 1,2,3, \ldots.$

Diffusion Temporal Component

$$\frac{dq}{dt} = -Kq\lambda$$

The derivative of q is itself proportional to q, thus relating q to the exponential function. The above function is a common simple integral which yields:

$$q(t) = Ce^{-K\lambda t}$$

Substituting in for K =2 and $\lambda = \left(\frac{n}{2}\right)$

Equation 4, Diffusion Temporal Component Solution

$$q_n(t) = C_n e^{-\frac{n^2}{2}t} = C_n e^{-nt}$$

Where $n = 1,2,3, \ldots.$

Diffusion Solution

Total solution $U(x,t) = V(x)q(t)$, combining the temporal and spatial components and the arbitrary constants:

$$U(x,t) = a_n e^{-nt} \sin\left(\frac{nx}{2}\right)$$

Where a_n *is an arbituary constant and* $n = 1,2,3, \ldots.$

Utilizing the principle of superposition, where the linear combination of valid solutions to a homogenous linear differential equation, is itself a valid solution:

Equation 5, General Solution for Diffusion Equation with Dirichlet BCs

$$U(x,t) = \sum_{n=1}^{\infty} a_n e^{-nt} \sin\left(\frac{nx}{2}\right)$$

Note: as $t \longrightarrow \infty$, exponential term dominates and $U \longrightarrow 0$, as expected

Where a_n is the combination of arbitrary coefficients.

a_n can be determined from the initial condition:

Where $L = 2\pi$

$$U(x,0) = f(x) = \sum_{n=1}^{\infty} a_n \sin(\frac{n\pi x}{L})(1)$$

Now pre-multiply by $sin(\frac{m\pi x}{L})$ and integrate over $0 \longrightarrow L$: *where m=1,2,3,……..*

$$\int_0^L \sin(\frac{m\pi x}{L}) f(x)dx = \sum_{n=1}^{\infty} B_n \int_0^L \sin\left(\frac{n\pi x}{L}\right) sin\left(\frac{m\pi x}{L}\right) dx = \frac{L}{2}\sum_{n=1}^{\infty} B_n \delta_{mn}$$

Where δ_{mn} is the Kronecker delta function, and is equal to 1 only when m=n (equal to 0 otherwise)

Thus:

$$a_m = \frac{2}{L}\int_0^L \sin(\frac{m\pi x}{L}) f(x)dx$$

$$a_m = \frac{1}{\pi}\int_0^{2\pi} \sin(\frac{nx}{2}) f(x)dx$$

Notably, Equation 5 is equivalent to that of a Fourier series, where functions may be written as the superposition many sinusoidal functions. For continuous functions, as the number of sinusoids "n" approaches the limit infinity, the Fourier approximate solution uniformly converges to the exact solution. The Fourier series can be visualised from the rotation of various circles, where the left hand side of Figure 1 is a visual representation of how the Fourier series approximates functions. The frequency of rotation and the magnitude of the radius for the each circle is related to the coefficient a_n via Euler's formula. Notably, negative frequencies are possible. Intuitively, and as shown in Figure 1, each successive n term (each extra circle) has subsequently less impact on the approximation than the previous term, implying convergence.

Notably for this project, the initial condition is a square wave, which contains two discontinuities at the "steps". This is an issue for the Fourier series and these discontinuities result in the Gibbs Phenomenon. This phenomenon arises from the difficulty of approximating a discrete step with a curve, as the curve "overshoots". As more and more terms are added to the finite summation, the location of the overshoot converges to the location of the discontinuity. As seen in Figure 2. The Fourier series for a discontinuous function converges similarly to that of a continuous function at all points, except at the discontinuity. At the discontinuities, the series converges to the average values of the two points at the discontinuity (Equation 6). Hence the Fourier series for our step function is pointwise convergent, and the Gibbs phenomenon will reduce as n summation terms increases; however, it will never be eliminated entirely.

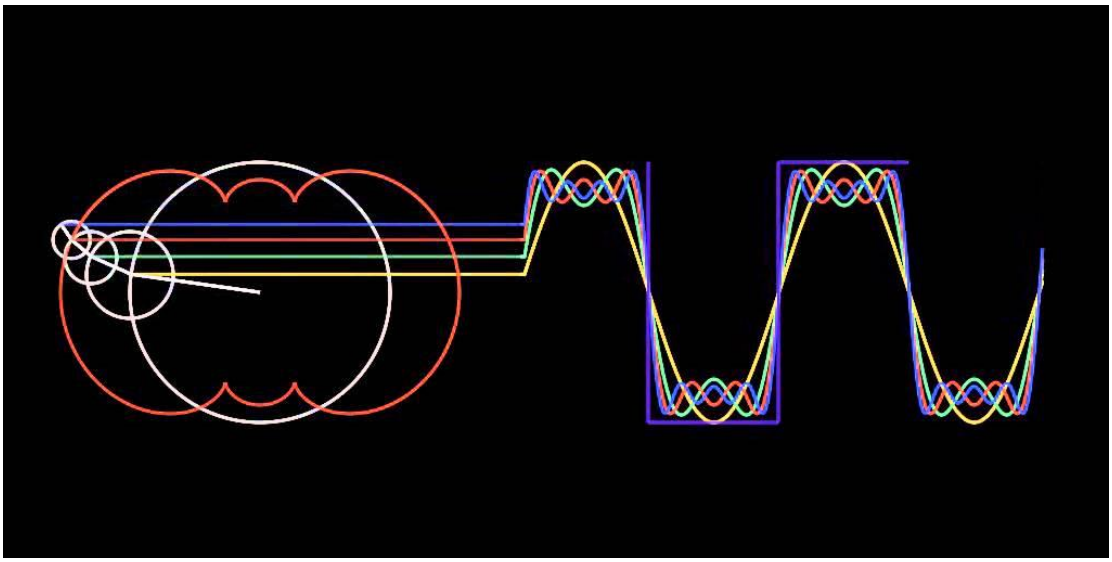


Figure 1. Visual interpretation of how Fourier series approximates a square wave, and how increasing the number of sinusoids better approximates it. Also shows The Gibbs Phenomenon and how the maximum overshoot moves back with increasing terms in the summation. Source: https://www.youtube.com/watch?v=k8FXF1KjzY0

Equation 6, Convergence at Discontinuity for Fourier series

$$\lim_{N\to\infty} S_N f(x_c) = \frac{f(x_c^-) + f(x_c^+)}{2}$$

Where $f(x)$ is an otherwise piecewise continuously differentiable function with a discontinuity at x=c and $S_N f$ is the N^{th} partial Fourier series.

Wave Equation

Equation 7, Wave Equation

$$\frac{\partial^2 U}{\partial t^2} = K\frac{\partial^2 U}{\partial x^2} = 0$$

Where K is constant

Aside: K is typically denoted as c^2, where c is the speed of the wave. For this project wave equation, K is set as 4. The length of the "string" "L" is also set as 2π as before.

Particular Problem:

$$\frac{\partial^2 U}{\partial t^2} = 4\frac{\partial^2 U}{\partial x^2} = 0$$

For: $0 \leq x \leq 2\pi \qquad t \geq 0$

Boundary Conditions: $U(0,t) = 0 = U(2\pi,t)$

Initial Condition: $U(x,0) = f(x)$ *where f(x) is a function defining the initial profile of U*

$$f(x) = \begin{cases} 1, & |x-\pi| \leq 2 \\ 0, & otherwise \end{cases}$$

Note: This is a larger Step function than that of the Diffusion equation.

$$U_t(x,0) = \frac{\partial U(x,0)}{\partial t} = g(x) = \sin\left(\frac{x}{2}\right)$$

As before, for the separation of variables, seek the solution in the form:

$$U(x,t) = V(x)q(t)$$

Applying the same procedure as the diffusion solution yields:

$$Temporal: \frac{1}{Kq}\frac{d^2q}{dt^2} = -\lambda \qquad Spatial: \frac{1}{V}\frac{d^2V}{dx^2} = -\lambda$$

Wave Spatial Component

The Equation is identical to that of the diffusion spatial component thus:

Equation 8, Wave Spatial Component Solution

$$V_n(2\pi) = C_{2n}\sin\left(\frac{nx}{2}\right)$$

Where $n = 1,2,3,\ldots$

Wave Temporal Component

$$\frac{d^2q_n(t)}{dt^2} = -K\lambda_n q_n(t) = -c^2\lambda_n q_n(t) = -4\lambda_n q_n(t)$$

This equation is solved in the same manner as the Diffusion spatial component, as well as knowing $\lambda > 0$ from the Wave spatial component. This is also a well-known standard ordinary differential equation (ODE) with the solution:

$$q_n(t) = D_{1n}\cos\left(\frac{n\pi c}{L}t\right) + D_{2n}\sin\left(\frac{n\pi c}{L}t\right)$$

Where D_{1n} and D_{2n} are arbituary constants and $n = 1,2,3, \ldots$.

Substituting $L = 2\pi$ and c = 2

Equation 9, Wave Temporal Component Solution

$$q_n(t) = D_{1n}\cos(nt) + D_{2n}\sin(nt)$$

Unlike the previous spatial component solution, we do not have apply a temporal boundary condition, and thus cannot solve for the coefficients at this stage.

Wave Solution

As the total solution $U(x,t) = V(x)q(t)$, combining the temporal and spatial components, absorbing the arbitrary constants, and applying the superposition principle:

Equation 10, General Solution for Wave Equation with Dirichlet BCs

$$U(x,t) = \sum_{n=1}^{\infty} (A_n\cos(nt) + B_n\sin(nt))\sin\left(\frac{nx}{2}\right)$$

Where $n = 1,2,3, \ldots$.

The two temporal sinusoidal terms show us that the total solution is the superposition of two traveling waves propagating in opposite direction with equal amplitude and frequency.

Setting t = 0

$$U(x,0) = \sum_{n=1}^{\infty} (A_n(1) + B_n(0))\sin\left(\frac{nx}{2}\right)$$

Which is the same form as the initial condition diffusion equation process to solve for a_n, but notably f(x) is slightly different for the wave equation, with steps at $x = \pi \pm 2$ rather than at $x = \pi \pm 1$ for the diffusion equation square wave.

$$A_n = \frac{1}{\pi}\int_0^{2\pi} \sin(\frac{nx}{2})f(x)dx$$

To solve for B_n the temporal derivative initial condition g(x) must be utilized.

First calculate $U_{n_t}(x,t)$ from our general solution $U(x,t)$ by taking the derivative of the summation terms.

$$U_{n_t}(x,t) = (-nA_n\sin(nt) + nB_n\cos(nt))\sin\left(\frac{nx}{2}\right)$$

Setting t = 0

$$U_{n_t}(x,0) = (nB_n)\sin\left(\frac{nx}{2}\right)$$

$$U_t(x,0) = \sum_{n=1}^{\infty}\left(nB_n \sin\left(\frac{nx}{2}\right)\right)$$

Require $U_t(x,0) = g(x) = \sin\left(\frac{x}{2}\right)$, applying same methodology as used earlier:

$$B_n = \frac{1}{n\pi}\int_0^{2\pi} \sin(\frac{nx}{2})g(x)dx$$

$$B_n = \frac{1}{n\pi}\int_0^{2\pi} \sin(\frac{nx}{2})\sin(\frac{x}{2})dx$$

These analytical solutions as n goes to infinity, must also be summed over infinity; giving infinite solutions. Pragmatically, this leads to Numerical methods being required to solve the Diffusion and wave Equations.

Numerical Methods

Analytical methods allow for continuous infinitesimal change (as limit approaches zero); however, numerical methods cannot, and require planes to be converted to discrete grids of finite differences (Figure 2). It follows then, as the discrete difference approaches zero, the numerical approximation approaches the analytical solution.

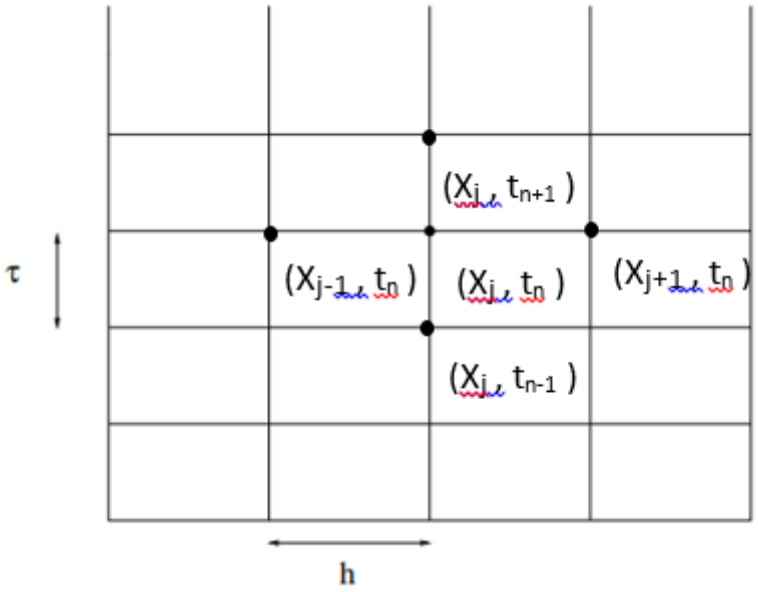


Figure 2. Discrete space for Numerical analysis, where j and n are spatial and temporal steps, h and tau are the discrete special and temporal spacing respectably.

Fourier series Approximation

Now that the parameters have been discretised, the Fourier coefficients may be calculated numerically. For this project the integrals were evaluated using a MATLAB recursive adaptive Lobatto quadrature routine accurate to an error of 10^{-6}. These values were then used in the Equations 5 and 10, looping over all n values, then all space coordinates and finally over all time steps.

Finite Difference Methods

Diffusion Equation

Using the Diffusion equation (Equation 1) and the definition of a derivative, the numerical methods employed in the investigation can be derived.

The derivation from the Diffusion equation is shown below.

$$\frac{\partial U}{\partial t} = K\frac{\partial^2 U}{\partial x^2} = 0$$

For the simplest case where K=1:

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2} = 0$$

From the definition of a derivative:

$$\left(\frac{dU}{dx}\right)_{jn} = \lim_{\Delta x \to 0} \frac{U_{j+1} - U_j}{\Delta x}$$

Numerical methods use these approximate differences for derivatives in order to solve differential equations, and if the initial equation is assumed to be continuously differentiable, the Taylor series may be applied.

Equation 11, Taylor series

$$U(x + \Delta x, t_n) = U(x, t_n) + \Delta x U'(x, t_n) + \frac{\Delta x^2}{2!}U''(x, t_n) + \frac{\Delta x^3}{3!}U^{(3)}(x, t_n) + \cdots$$

It can be seen then that the first order approximation is:

Equation 12, Forward Difference

$$\left(\frac{dU}{dx}\right)_{jn} = \frac{U_{j+1} - U_j}{\mathrm{h}} + \mathcal{O}(\mathrm{h})$$

This is known as the Two Point Forward Difference approximation. $\mathcal{O}(\mathrm{h})$ Is the truncation error term, and tells us that the forward difference is first order accurate in space. Note that if $\lim_{\mathrm{h}\to 0}$ the derivative definition is retrieved, and $(\mathrm{h}) \longrightarrow 0$; therefore, the method is said to be consistent.

For Two Point Forward Difference in space, $\Delta_{+\mathrm{x}}\mathrm{U}_{\mathrm{j}}^{\mathrm{n}} \equiv \mathrm{U}_{\mathrm{j}+1}^{\mathrm{n}} - \mathrm{U}_{\mathrm{j}}^{\mathrm{n}}$

A more accurate approximation is the *Central difference method,* which applies the derivative operator about the spatial points U_{j-1} and U_{j+1} (ie centred around U_j).

Two Point Backward Difference U_{j-1}

$$U(x - \Delta x, t_n) = U(x, t_n) - \Delta x U'(x, t_n) + \frac{\Delta x^2}{2!}U''(x, t_n) - \frac{\Delta x^3}{3!}U^{(3)}(x, t_n) + \cdots$$

Subtracting the backward difference from the forward difference and dividing across by $(2\,\Delta x)$:

$$\frac{U(x + \Delta x, t_n) - U(x - \Delta x, t_n)}{2\Delta x} = U'(x, t_n) + \frac{\Delta x^2}{3!}U^{(3)}(\xi, t_n)$$

Here the error term given by Taylor's theorem tells us it is second order accurate in space, thus:

Equation 13, Three Point Central Difference

$$\left(\frac{dU}{dx}\right)_{jn} = \frac{U_{j+1} - U_{j-1}}{2\mathrm{h}} + \mathcal{O}(h^2)$$

Note: when h<1 (ie, approaches zero), a second order error term is less than a first order, and converges more rapidly.

The operators below can be substituted in as required:

Forward time $\Delta_{+t}U_j^n \equiv U_j^{n+1} - U_j^n$ backward time $\Delta_{-t}U_j^n \equiv U_j^n - U_j^{n-t}$

Centred time $\delta_t U_j^n \equiv U_j^{n+0.5} - U_j^{n-0.5}$

Forward space $\Delta_{+x}U_j^n \equiv U_{j+1}^n - U_j^n$ backward space $\Delta_{-x}U_j^n \equiv U_j^n - U_{j-1}^n$

Centred Space $\delta_x U_j^n \equiv U_{j+0.5}^n - U_{j-0.5}^n$

Second order derivatives:

We can obtain the second order 3 point central difference by applying the central difference method twice which yields:

$$(\frac{\partial^2 U}{\partial x^2})_{jn} = \frac{\delta_x^2 U_j^n}{h^2} = \frac{U_{j+1} - 2U_j + U_{j-1}}{h^2} + \mathcal{O}(h^2)$$

Alternatively, this error term can be seen by adding the forward and backward differences together, rearranging, and dividing across this time by Δx^2 yields:

$$\frac{U(x + \Delta x, t_n) - 2U(x, t_n) + U(x - \Delta x, t_n)}{\Delta x^2} = U''(x, t_n) + \frac{\Delta x^2}{12} U^{(4)}(\xi, t_n)$$

Heat-Diffusion Equation

Thus substituting for The Diffusion Equation:

$$\frac{\partial U}{\partial t} = K \frac{\partial^2 U}{\partial x^2} \Longrightarrow \frac{\Delta_{+t} U_j^n}{\tau} = K \frac{\delta_x^2 U_j^n}{h^2}$$

This is uses Forward time and Centred space, and is hence known as Forward time Centred space Method (FTCS).

Equation 14, FTCS with Truncation

$$\frac{U_j^{n+1} - U_j^n}{\tau} = k \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{h^2} + \mathcal{O}(h^2) + \mathcal{O}(\tau)$$

From the error terms, FTCS method is first order accurate in time, and second order accurate in space. This total error is known as the Truncation Error (T(x,t)) and is found by:

$$T(x,t) = \mathcal{O}(h^2) + \mathcal{O}(h) = \frac{\Delta_{+t} U_j^n}{\tau} - k \frac{\delta_x^2 U_j^n}{h^2}$$

And from Taylor's series:

$$T(x.t) = \frac{1}{2} U_{xx} \tau - \frac{1}{12} U_{xxxx} h^2$$

Returning to FTCS for the diffusion equation, and omitting the truncation error, U_j^{n+1} can be solved explicitly in terms of U^n (a known variable). This fact is why FTCS is known as an *explicit method*. It can be assumed that we know the previous time step, as in general for diffusion equation problems, the purpose of the investigation is to analyse how the system evolves over time from a given initial profile. Rearranging in terms of U_j^{n+1}:

FTCS for Diffusion Equation

$$U_j^{n+1} = U_j^n + \frac{k\tau}{h^2}(U_{j+1}^n - 2U_j^n + U_{j-1}^n)$$

$$U_j^{n+1} = U_j^n + \alpha(U_{j+1}^n - 2U_j^n + U_{j-1}^n)$$

Where $\alpha = \frac{k\tau}{h^2}$

FTCS can be altered for the Dirichlet boundary condition by setting $U_1^{n+1} and\ U_N^{n+1} = 0$ (assuming the spatial grid starts at 1 and not zero, which notably MATLAB arrays indexing start at 1).

Aside; the von Neumann derivative boundary conditions can be forced by setting $U_1^{n+1} = U_2^{n+1} and\ U_N^{n+1} = U_{N-1}^{n+1}$, thus no change of U across the extremity points.
The periodic boundary conditions can be implemented for FTCS by setting $U_1^{n+1} = U_{N-1}^{n+1}$ and $U_N^{n+1} = U_2^{n+1}$.

FTCS Von Neumann Stability Analysis

In order to test that numerical schemes remain stable and do not propagate errors so that the solution does not becomes worse overtime, Von Neumann stability analysis is conducted. From the Lax equivalence theorem, a consistent numerical solution of a partial differential equation is convergent if and only if it is also stable.

Substituting for $U_j^n = G^n e^{idjh}$ into the FTCS scheme, where G(d) is the growth factor.

$$G^{n+1}e^{idjh} = G^n e^{idjh} + \alpha\left(G^n e^{id(j+1)h} - 2G^n e^{idjh} + G^n e^{id(j-1)h} \right)$$
$$G^n e^{idjh}(G) = G^n e^{idjh} + \alpha\left((G^n e^{idjh})e^{idh} - 2G^n e^{idjh} + (G^n e^{idjh})e^{-idh} \right)$$

Removing the common factor

$$G = 1 + \alpha\left(e^{idh} - 2 + e^{-idh}\right)$$

Using trigonometric identities

$$G = 1 + 2\alpha(1 - \cos(dh))$$
$$G = 1 - 4\alpha sin^2\left(\frac{dh}{2}\right)$$

The solution is not amplified, and thus stable, so long as the growth factor $|G(d)| \leq 1\ \forall d$
Sin^2 function maximum value is 1 at $dh = \pi$, therefore worst case $G(d) = 1 - 4\alpha$
Substituting for α and rearranging, then FTCS is conditionally stable so long as:

FTCS Stability Condition

$$\alpha = \frac{k\tau}{h^2} < \frac{1}{2}$$

This makes the scheme very limited and inefficient as Δt must be greatly smaller than Δx.

IMPLICIT SCHEMES:

If we consider backwards time central space (BTCS) for the Diffusion equation:

$$\frac{\partial U}{\partial t} = K\frac{\partial^2 U}{\partial x^2} \Longrightarrow \frac{\Delta_{-t}U_j^n}{\tau} = K\frac{\delta_x^2 U_j^n}{h^2}$$

$$\frac{U_j^n - U_j^{n-1}}{\tau} = k\frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{h^2} + \mathcal{O}(h^2) + \mathcal{O}(\tau)$$

$$U_j^n = U_j^{n-1} + \frac{k\tau}{h^2}(U_{j+1}^n - 2U_j^n + U_{j-1}^n)$$

Here U_j^n is the value sought; however, all values for this future time are also unknown, which U_j^n depends on. Therefore, In order to solve for U_j^n, a series of simultaneous equations must be solved. This is what is known as an *implicit scheme*. It follows then that an implicit scheme uses more computational power to solve per time step than an explicit method.

Upon conduction of Von Neumann stability analysis, that BTCS is unconditionally stable, no matter the value of $\frac{k\tau}{h^2}$, a large advantage over FTCS. For certain problems, τ may have to be impractically small in order to use FTCS, and while an implicit method uses more computational power, larger time steps may result in a faster operation. Note however, that both BTCS and FTCS are of the same accuracy, being only first order accurate in time.

An equation known as the "Theta Method" allows the conversion between the two schemes, where $\theta = 0$ corresponds to the FTCS method, while BTCS is generated when $\theta = 1$, where $0 \leq \theta \leq 1$

Theta Method

$$\frac{\partial U}{\partial t} = K\frac{\partial^2 U}{\partial x^2} \Longrightarrow \frac{\Delta_{+t}U_j^n}{\tau} = K\left(\theta\frac{\delta_x^2 U_j^{n+1}}{h^2} + (1-\theta)\frac{\delta_x^2 U_j^n}{h^2}\right)$$

Crank-Nicholson Scheme

By setting $\theta = \frac{1}{2}$, and expanding out the Theta method yields:

Equation13, Crank-Nicholson Scheme

$$-\frac{\alpha}{2}U_{j-1}^{n+1} + (1+\alpha)U_j^{n+1} - \frac{\alpha}{2}U_{j+1}^{n+1} = \frac{\alpha}{2}U_{j-1}^n + (1+\alpha)U_j^n - \frac{\alpha}{2}U_{j+1}^n$$

Where $\alpha = \frac{k\tau}{h^2}$
Which can also be written as

$$\frac{U_j^{n+1} - U_j^n}{\tau} = \frac{K}{2}\left(\frac{U_{j+1}^{n+1} - 2U_j^{n+1} + U_{j-1}^{n+1}}{h^2} + \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{h^2}\right) + \mathcal{O}(h^2) + \mathcal{O}(\tau^2)$$

Similar to BTCS, the Crank-Nicholson scheme is implicit and unconditionally stable but has the additional benefit of being both second order accurate in both space and time ($\mathcal{O}(h^2) + \mathcal{O}(\tau^2)$). This is due to it being a combination of BTCS and FTCS, which has a similar effect as taking a central difference in time, which we have shown previously is second order accurate; while retaining their spatial accuracy. This can be seen intuitively from Figure 3; note the change of discrete notation. This method being second order accurate in time, also results in the Crank-Nicholson method having second order convergence in time. For these reasons, the Crank-Nicholson scheme is generally better than both FTCS and BTCS. Consequently, this investigation will focus on for solving the diffusion equation with the Crank – Nicholson finite difference scheme.

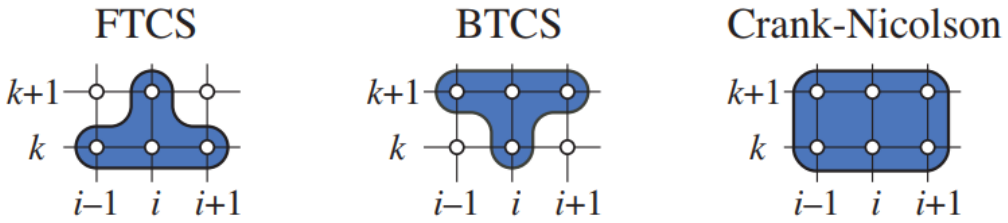


Figure 3, Depicts how the 3 schemes use neighbouring points on the finite grid to calculate the value U_i^{k+1} (where i and k are the temporal and spatial steps respectively). From this diagram, it can be seen intuitively why only the Crank-Nicholson is second order accurate temporally. Source: http://web.cecs.pdx.edu/~gerry/class/ME448/notes/pdf/CN_slides.pdf

Crank-Nicholson Von Neumann Stability Analysis

As before, substituting for $U_j^n = G^n e^{idjh}$ into the Crank-Nicholson scheme, applying laws of indices and factorising yields:

$$\frac{G-1}{\tau} = \frac{K}{2}(G+1)\left(\frac{e^{idh} - 2 + e^{-idh}}{h^2}\right)$$

$$G - 1 = \alpha(G+1)\left(\frac{e^{idh} + e^{-idh}}{2} - 1\right)$$

$$G - 1 = \alpha(G+1)(\cos(dh) - 1)$$

$$G - \alpha(G)(\cos(dh) - 1) = 1 + \alpha(\cos(dh) - 1)$$

$$G = \frac{1 + \alpha(\cos(dh) - 1)}{1 - \alpha(\cos(dh) - 1)}$$

Where $\alpha = \frac{k\tau}{h^2}$ which are all physically non-negative parameters

The solution is not amplified, and thus stable, so long as the growth factor $|G(d)| \leq 1 \; \forall d$
Cos function ranges between ±1, which results in G(d) =±1 respectively.

(For integer multiples of $dh = \pi/2$, $G(d) = \frac{1-\alpha}{1+\alpha}$). As α must be a non-negative value for all reasonably physical systems, the denominator will always be larger than the numerator. Therefore $|G(d)| \leq 1\ \forall d$ and the Crank – Nicholson scheme is unconditionally stable for all values of Δx and Δt.

Solving Implicit Methods

In order to solve implicit methods, the set of simultaneous equations may be placed in matrix format (Figure 4).

$$\begin{pmatrix} b_1 & -c_1 & 0 & . & . & 0 \\ -a_2 & b_2 & -c_2 & . & . & 0 \\ 0 & -a_3 & b_3 & -c_3 & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & . & . & . & -a_N & b_N \end{pmatrix} \begin{pmatrix} u_1^{n+1} \\ u_2^{n+1} \\ u_3^{n+1} \\ . \\ . \\ u_N^{n+1} \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ . \\ . \\ d_N \end{pmatrix}$$

Figure 4, Matrix form for the implicit simultaneous equations. Has the form of a Tridiagonal matrix

For the Crank-Nicholson scheme, from the coefficients where ($\alpha = \frac{k\tau}{h^2}$):

a = α/2 b = (1+α) c= α/2 d=$\frac{\alpha}{2} U_{j-1}^n + (1+\alpha) U_j^n - \frac{\alpha}{2} U_{j+1}^n$

Altered Crank-Nicholson

$$-a_j U_{j-1}^{n+1} + b_j\, U_j^{n+1} - c_j U_{j+1}^{n+1} = d_j$$

As before, due to previous time step, d is completely known. Matrices may be solved via LU factorisation or Gaussian elimination with back substation. For this project a variation of LU decomposition for tridiagonal matrices was used. A fuller explanation of a similar method is provided in the appendix.

Wave Equation Finite Difference Scheme

From the Finite difference analysis overviewed in the diffusion section, it was decided that approximating the second order temporal and spatial derivatives using 3point central differences due to the higher accuracy and convergence rate.

Equation 16, Wave Equation Finite Difference Scheme

$$\frac{U_j^{n+1} - 2U_j^n + U_j^{n-1}}{\tau^2} = c^2\left(+\frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{h^2}\right) + \mathcal{O}(\tau^2, h^2)$$

Where $k = c^2$,c is the speed of the wave and j and n are the spatial and temporal node locations respectively.

This may be rewritten as:

$$U_j^{n+1} = 2(1-\alpha)U_j^n - U_j^{n-1} + \alpha(U_{j-1}^n + U_{j+1}^n)$$

Where $\alpha = c^2\frac{\tau^2}{h^2}$

In this format, it is notable that this is an explicit scheme as the solution at the next time step U_j^{n+1} is only dependant on solutions at previous time steps. Figure 5 gives a visual representation of how this method functions; note that the discrete notation has changed.

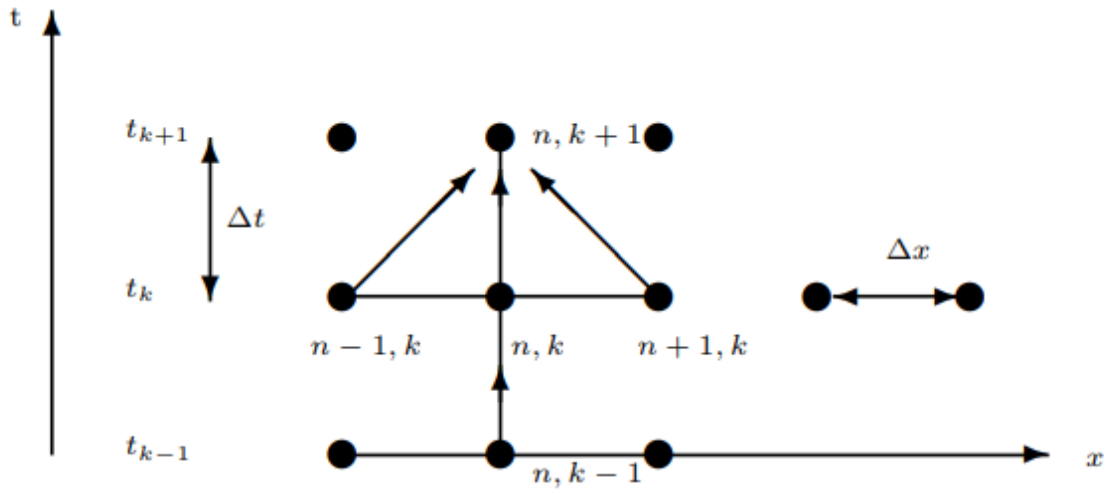


Figure 5, Depicts how the Wave equation numerical scheme uses neighbouring points on the finite grid to calculate the value U_n^{k+1} (where n and k are the temporal and spatial steps respectively). From this diagram, it can be seen intuitively why this is a second order accurate explicit scheme. Source: https://www.math.ubc.ca/~peirce/M257_316_2012_Lecture_8.pdf

Immediately a problem arises with this method. In order to approximate the first-time step beyond the initial condition (n=0), the negative time point U_j^{-1} is required. This may be circumvented as the initial temporal derivative is known, $g(x)$ from the wave equation initial conditions. This allows the FTCS scheme to be used to calculate the first-time step and allow the wave equation model to flow.

Unfortunately, it is known that the FTCS method is less than ideal; however, the $U_j^n - U_j^{n-1}$ terms in Equation 16 also appear in the temporal form of the more accurate three-point central difference and can be substituted as $2\Delta t g(x)$. While the truncation error for this approximation is now equivalent to that used for the further times steps, this initial error is propagated and compounded throughout the further approximations. This should theoretically affect then accuracy of the entire method.

Wave Equation Scheme Von Neumann Stability Analysis

As before, substituting for $U_j^n = G^n e^{idjh}$ into the Wave equation scheme:

$$G^{n+1}e^{idjh} = 2(1-\alpha)G^n e^{idjh} - G^{n-1}e^{idjh} + \alpha(G^n e^{id(j-1)h} + G^n e^{id(j+1)h})$$

Removing the common factor

$$G = 2(1-\alpha) - G^{-1} + \alpha(e^{-idjh} + e^{idh})$$

Using trigonometric identities

$$G = 2(1-\alpha) + 2\alpha(\cos(dh) - G^{-1}$$
$$G = 2 - 2\alpha + 2\alpha(\cos(dh) - G^{-1}$$
$$G = 2(1 + \alpha(\cos(dh) - 1) - G^{-1}$$
$$G = 2\left(1 - 2\alpha sin^2\left(\frac{dh}{2}\right)\right) - G^{-1}$$
$$G = 2\rho - G^{-1}$$

Where $\rho = 1 - 2\alpha sin^2\left(\frac{dh}{2}\right)$

Multiple across by the Growth factor G and rearranging gives the following quadratic equation:

$$G^2 - 2\rho G + 1 = 0$$

With roots G_1 and G_2 given by:

$$G_{1,2} = \rho \pm \sqrt{\rho^2 - 1}$$

If $\rho^2 > 1$ then the magnitude of one of the growth factors is greater than 1, which means there is an instability.

If $\rho^2 \leq 1$, then $|\rho| \leq 1$
Substituting back for ρ

$$-1 \leq 1 - 2\alpha sin^2\left(\frac{dh}{2}\right) \leq 1$$
$$-2 \leq 1 - 2\alpha sin^2\left(\frac{dh}{2}\right) \leq 0$$

The range of $Sin^2(z)$ and as α must be non-negative, ensure the right-hand side condition is always met. The left-hand side is met if and only if:

$$\alpha sin^2\left(\frac{dh}{2}\right) \leq 1$$

Considering the maximum of $Sin^2(z) = 1$ then:

$$c^2\frac{\tau^2}{h^2} = \alpha \leq 1$$

CFL Condition for stability of Wave equation scheme.

$$c \leq \frac{h}{\tau}$$

This condition for stability is known as the Courant-Friedrichs-Lewy (CFL) condition for stability.

TAYLOR SERIES

Notably, as the Taylor series, which the finite difference methodology is based on, approximates solutions by fitting polynomial curves of higher and higher orders. Intuitively this leads to the same Gibbs phenomenon issue for discrete steps as the Fourier series. This can also be understood as viewing the real Taylor series and Fourier series as particular cases of a complex Taylor Series.

$$f(x) = C_0 + C_1 z + C_2 z^2 + C_3 z^3 + \cdots$$

To obtain the real Taylor series, set $z = x$ and the coefficients $c_n = \frac{f^{(n)}(0)}{n!}$. Substitute for $z = e^{i\theta}$ and $c_n = \frac{1}{2\pi}\int_0^{2\pi} e^{-ni\theta} g(\theta) d\theta$ to obtain the Fourier series. These two forms for the coefficients c_n are related by the Cauchy integral formula. The forms of these coefficients discussed above are analogous to those used in this project.

Results and Analysis

For ease of reference, time is assumed to be measured in seconds.

Fourier Analysis

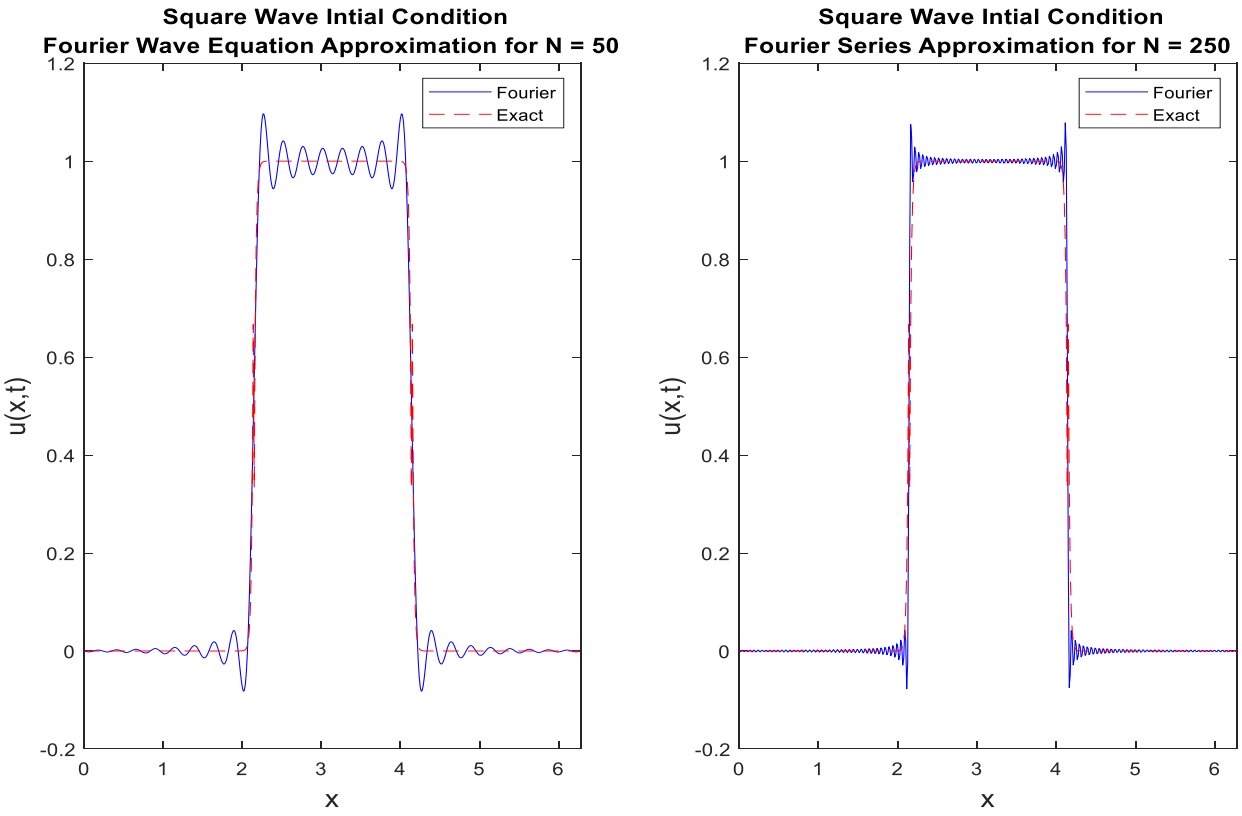


Figure 6(a) and 6(b), (left, right) depict the Fourier series approximation comparison to the exact square wave initial condition for N=50 and N=250 series summation respectively. Both have clearly visible Gibbs phenomenon; however, the higher summation model is visibly a better approximation due to the Gibbs phenomenon more localized to the discrete step as predicted. (Modelled Using Fourier heat equation model and Crank Nicholson "hard coded" initial condition, arbitrary choice as the wave equation models would be the same for the initial profile. $\tau = 3.125e-05$; $h = 1.25e-02$*)*

The Fourier series approximations in Figure 6(a) and (b) show clearly the Gibbs phenomena, which is more pronounced in the vicinity of the discontinuity. As predicted, the phenomena "moves back" to the discontinuity region as more terms in the summation are used, providing a more accurate approximation. A quantitative measurement of this was conducted using the reduced chi squared equation.

Equation 16, Reduced Chi Squared

$$\chi^2 = \frac{\sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}}{n - m - 1}$$

Where n is the number of observations, E is the expected data, O is the observed (modelled) data, and m is the number of fitting parameters.

For measuring goodness of fit, the smaller the Chi squared value, the better the observed data fits to the expected model. Care must be taken when using Chi squared value as the number of fitting parameters, or number of degrees of freedom, affects the measurement value and if errors are estimated incorrectly it is common to misinterpret and derive wrong conclusions from the resultant value. Reduced chi squared value which is less than 1 can signify an "over fit" model or an over estimation of the errors. For this investigation, the chi squared value was only used as a rudimentary aid to quantify the goodness of fit and significant weighting should not be given to its value. This is compounded further by the decision to systemically set the number of fitting parameters in the equation to 1 throughout the investigations. This was chosen as in general the true value of m was considered negligible relative to the large size of the data sets used, which dominates the denominator. The chi squared value for N=50 and N=250 models was 4.09e-06 and 9.63e-06 respectively. As these values are extremely small, and from the aforementioned points, more weight was given to visual analysis of Figures 6(a) and (b), where using a larger number of summation terms N provides a visually more appropriate approximation; as expected.

In order to determine an appropriate value of N, a tolerance difference of 0.005 between approximations (separated by N + 10) for the initial square wave was chosen as sufficient. In order to account for the Gibbs phenomena, the average absolute difference of all points was used rather than an individual point; as the tolerance was quickly met for low number of summation terms when considering the individual value at pi. For this tolerance level, N ~230 summation terms were required. This convergence to within this desired tolerance implies empirically that the Fourier model used is pointwise convergent for this initial condition.

Diffusion Equation

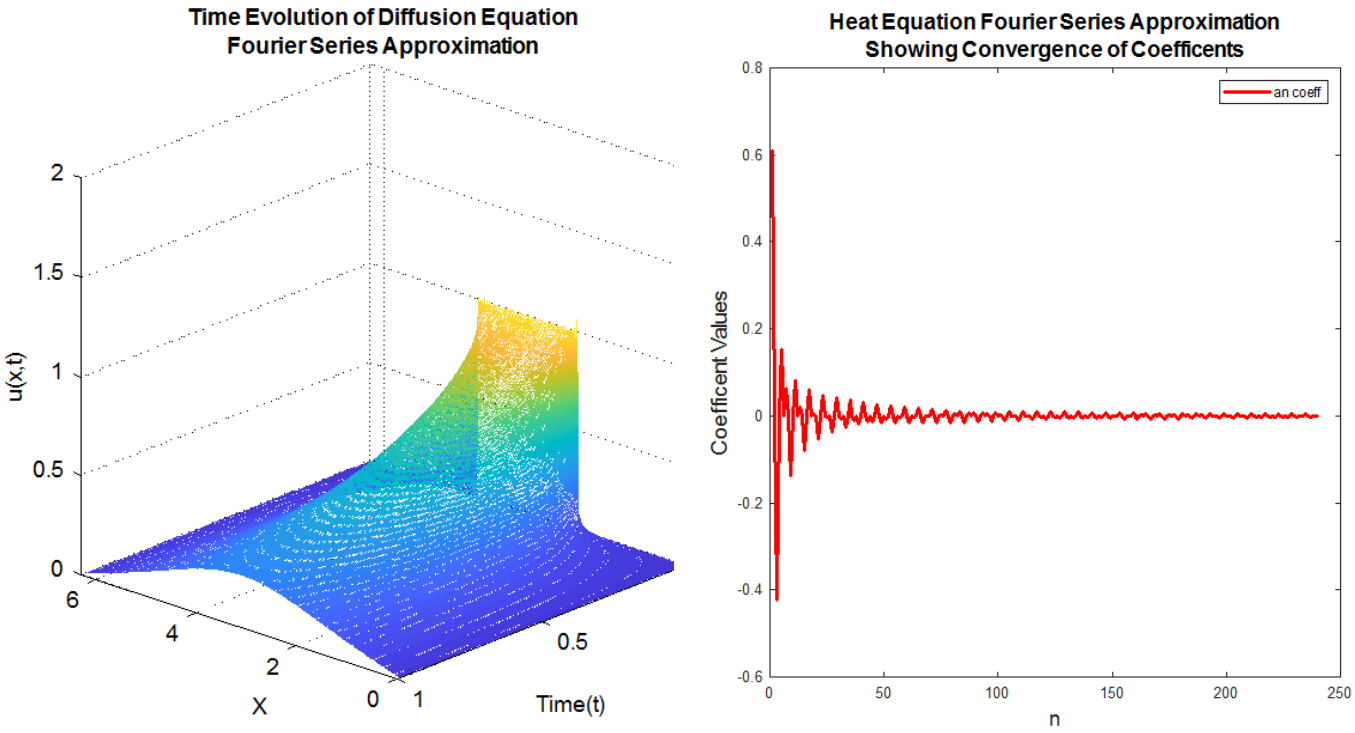
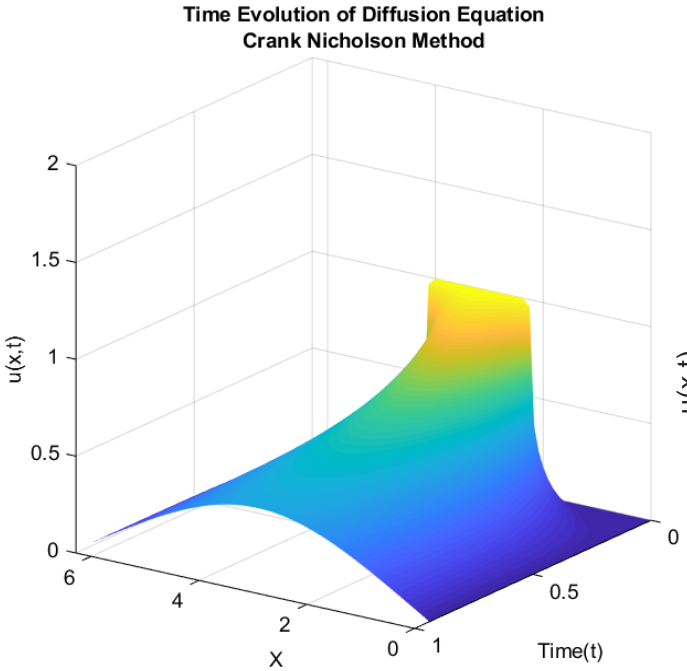


Figure 7a) and 7(b), (left, right) depict the time evolution of Fourier series approximation for the Diffusion equation and the values of the coefficient "an" as the summation term n increases respectively. Figure 7(a) aligns with the physical interpretation of heat flowing out of a bar as the temperature approaches uniform zero over time. Figure 7(b) further supports the convergence of the Fourier series model as each successive "a_n" term are of lower magnitude than the previous. The oscillatory nature and sign change is hypothesised to be related to change in rotational orientation and frequency. ($\tau = 3.125e-05$; $h = 1.25e-02$)

The U(x,t) function profile in Figures 7(a) follows the shape to be expected for the physical temperature system discussed previously, as the temperature evolves over time from the square wave initial condition and spreads uniformly along the bar and dissipates out at the edges. Notably, as the profile becomes uniform and similar to that of a continuous Gaussian, the Gibbs phenomena present at the initial condition is smoothed out as well. Figures 7(b) provides further evidence of the Fourier series being pointwise convergent as the value of a_n is inversely proportional to n as theorised.

Crank-Nicholson



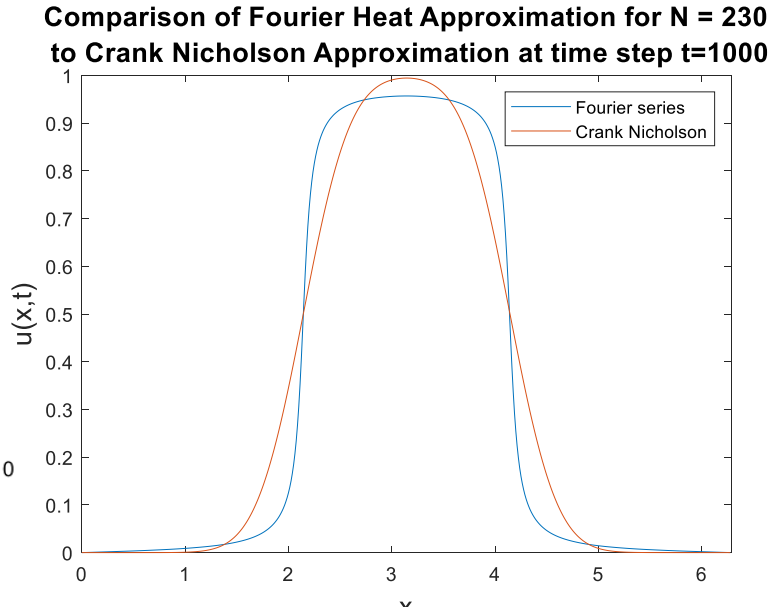


Figure 8(a) and 8(b), (left, right) depict the time evolution of The Crank Nicholson Scheme and the comparison with the Fourier approximation at a specific time respectively. Figure 8(a) aligns with the physical interpretation of heat flowing out of a bar as the temperature approaches uniform zero over time, just as the Fourier series previously. Figure 8(b) aids in visualising how the 2 models differ. As seen in figures 6 , due to the Fourier approximation starting with the Gibbs phenomena (where as the crank scheme initiates from the hard codes square wave), the Fourier series has more unphysical temperature in the system and so takes more time to dissipate. $(\tau = 3.125e-05\ ;\ h = 1.25e-02)$

Figure 8(a) profile is very similar to that of the Fourier series approximation, but notably as the Crank Nicholson scheme starts from a "hard coded" initial condition, there is no Gibbs phenomena at all. This is due to the scheme never having to model a discrete step as it immediately begins to model as a "Gaussian-esque" curve. Conversely, due to the Fourier approximation starting with the Gibbs phenomena (where as the crank scheme initiates from the hard codes square wave), the Fourier series has more unphysical temperature in the system and so takes more time to dissipate.

Convergence of crank Nicholson scheme:

In order to assess the convergence of the Crank Nicholson scheme, the temporally and spatially characteristics must be conducted separately. The tolerance desired for convergence due to complications discussed below was later adjusted to 0.1. Spatial convergence testing was conducted first by choosing an arbitrary discrete temporal step "deltat" which would remain fixed throughout the spatial testing. The Crank Nicholson scheme was then conducted for ever decreasing spatial steps, "deltax", which halted in size each iteration. This process continued until the maximum difference between each model for identical spatial coordinates (which lined up across all time steps) fell within the desired tolerance. This approach to look at multiple shared spatial coordinates

were used in account of the Gibbs phenomena which would be prevalent for the wave equation convergence. Once the maximum spatial step which fell within tolerance was found, this was then used to find the maximum temporal step using a similar procedure. In order for multiple x coordinates to align, as the spatial steps were halved and, the maximum x points were doubled and subtracted by 1, as there is one more maximum points than step. One source of potential error is the additional fail safe in the code to use the floor or ceiling functions to force the number of spatial steps to be an integer, though due to the setup this should not theoretically be necessary at this stage; however it can vary the total length of the bar.

Interestingly, while iterating large discrepancies appeared in the model close to where the initial discontinuities were (Figure 9). I hypothesis that these are somehow related to the Gibbs phenomena; however, these are more likely artefacts produced by erroneous code. Due to time constraints this behaviour was not investigated further and the convergence for deltax was accepted at this limit. Interestingly, even though the Crank Nicholson scheme is theoretically unconditionally stable, this suggests otherwise. Notably these artefacts from my preliminary research did not arise for the temporal convergence even though the same detax was used. This further suggests a codding error. Further investigations using alternative initial conditions should be conducted.

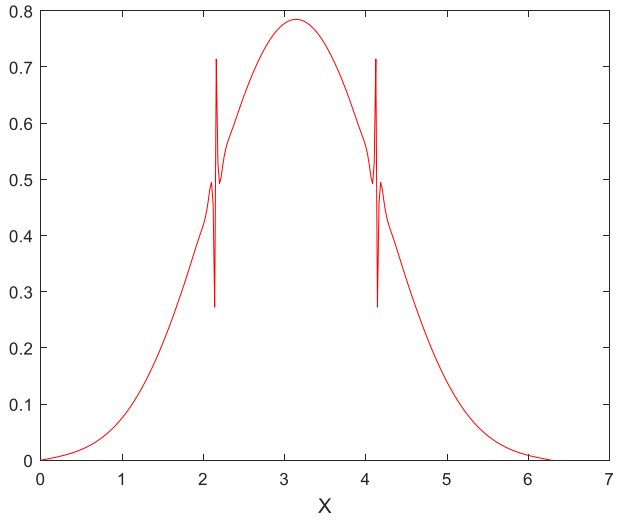


Figure 9 is a snapshot of the time evolution of The Crank Nicholson Scheme during the convergence testing programme. The Graph headers are missing as the image was extracted during the running programme. The Gibb like artefacts are seen around the x value of pi plus and minus one. This relates to the position of the initial discontinuities. $(\tau = (1.00e-02\ ;\ h = 7.85e-02\ ; approimate\ time\ step = \ 20 \pm 15)$

Step Size Required Crank Nicholson Convergence

The maximum spatial and temporal step size for the Crank Nicholson scheme for within a tolerance difference of 0.1 due to the artefact phenomena was 7.853982e-02 and 1.562500e-04 respectively. Therefore, empirically the crank Nicholson scheme is convergent in time, and most likely convergent in space. The artefact phenomena prevents the concrete conclusion that the Crank Nicholson scheme is empirically spatially convergent.

Truncation Error Analysis

To test the order of accuracy of this method, for comparison with that expected analytically, it was assumed that the step values given by the convergence procedure provided an exact solution. Then

by working in reverse to the convergence procedure and increasing the step size away from that of the exact solution, and taking the difference in the generated u(x,t) values, the temporal and spatial order of error were approximated. Unfortunately, due to time constraints only the spatial order error testing was completed for the Diffusion Equation.

Similar to the convergence testing, the number of spatial steps needed for each model must decrease in the same ratio as the spatial step size increases, while the number of steps remained an integer value. The coding structure for this truncation analysis was developed originally for the more limiting Finite difference method for the Wave Equation, as the CFL condition must be satisfied for that case. Initially, step sizes which when integer multiplied resulted in all corresponding spatial and temporal coordinates aligning, which was confirmed by using the modulus function so that there was no remainder. However, this conditionality greatly limited the number of appropriate step sizes and was not a large enough data set.

Instead, spatial values which were closest to the arbitrary point pi at a specific time step were chosen to calculate the tolerance. This decision adds further inaccuracy to the model, which decreased the quality of the resultant data but allowed more data to be extracted.

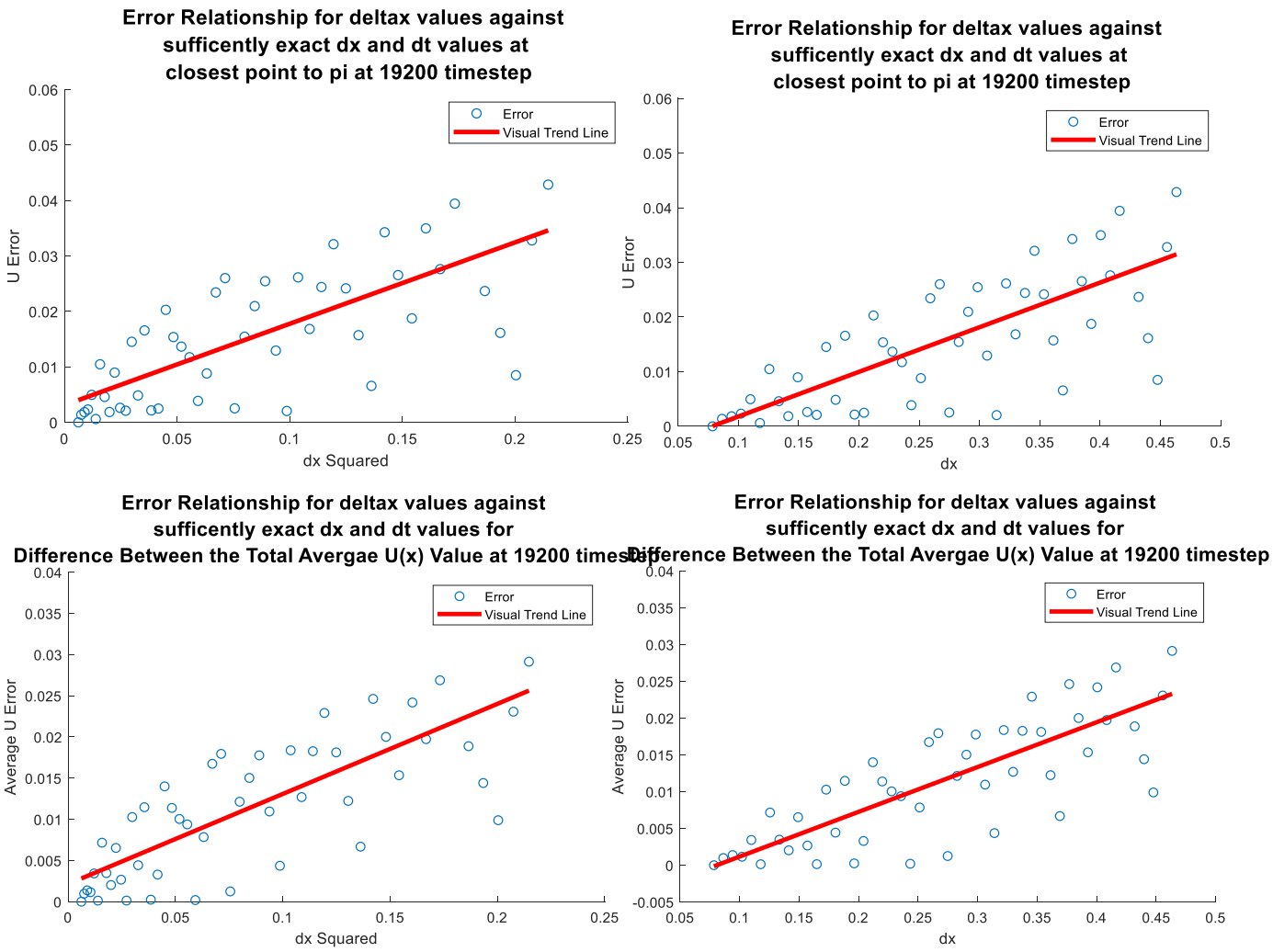


Figure 10(a), 10 (b), 10(c) and 10 (d), (top left, top right, bottom left, bottom right) Depict the Crank Nicholson schemes error relationships as titled. Trend lines are for visual aid and for calculating rudimentary Chi squared values: 1.427e-02, 1.303e-02, 9.938e-03, and 8.998e-03 respectively. $(\tau\ exact\ = 3.1251.562500e-04\ ;\ h\ exact\ = 7.853982e-02)$

The expected $\mathcal{O}(h^2)$ relationship investigated in Figures 10a had a shape reminiscent of square root relationship, and so $\mathcal{O}(h)$ was plotted in Figures 10b. Visually this appeared to be a better fit, and was supported by the relatively lower Chi Squared value calculated from the trend line. As before, this Chi Squared value was only used as a qualitative "bench mark" for the trend. In order to reduce the discrepancy caused by taking the value closest to pi, the error between the average values of U(x) was calculated and plotted. Of note, the averaging process involves further uncertainty in the data due to the principle of propagation of uncertainty. All Figures support the basic principle that increasing the step size reduces the accuracy of the model. Visually figures 10b and 10d, and their smaller chi squared values, suggest that the order of accuracy of the model is more likely $\mathcal{O}(h)$. I suggest more analysis is required. Potential causes of this seeming discrepancy from the analytical expectation include computational errors associated with solving the simultaneous equations, manual coding errors, as well as those arising from the point selection and averaging method discussed above.

Wave Equation

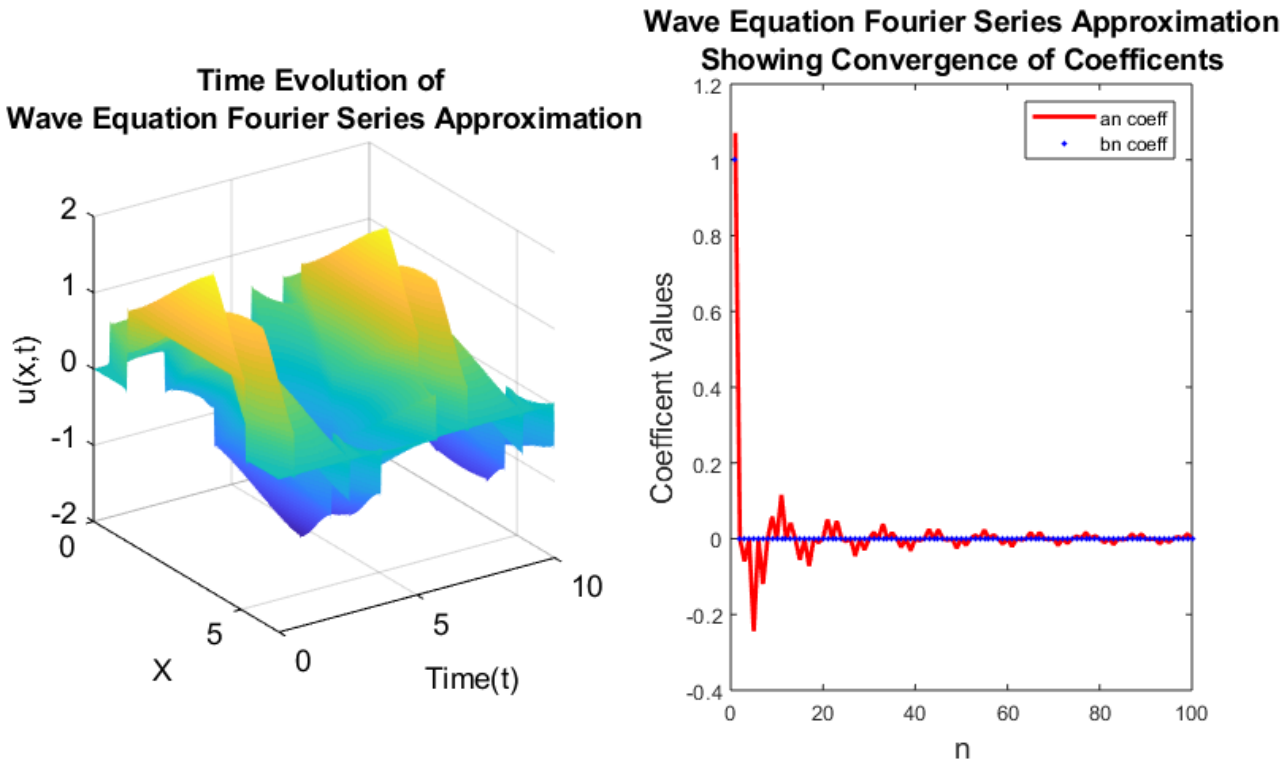


Figure 11(a) and 11(b), (left, right) depict the time evolution of Fourier series approximation for the wave equation and the values of the coefficient "an" and "bn" as the summation term n increases respectively. Figure 11a aligns with the physical interpretation of a trapped standing wave reflecting causing positive and negative superposition. Figure 11b Further supports the convergence of the Fourier series model as each successive "a_n" and "bn" terms are of lower magnitude than the previous. The oscillatory nature and sign change of "an" is hypothesised to be related to change in rotational orientation and frequency. ($\tau = 0.005$; $h = 0.0125$, 2000 time steps)

As the initial condition was a square wave, the two fundamental waves are also square waves. This results in the jagged wave structure as multiple discontinuities are propagating. . Consequently, the Gibbs phenomena has a much larger impact than for the diffusion equation as they are not smoothed out (Figure 12). These miniature peaks meant the "Peak finder" MATLAB function could not be used, meaning there was no efficient way in order to determine the time period (and consequently it's stability) of the wave compared to that theoretically predictable.

Similar to the diffusion equation investigation, the decay of the Fourier coefficients supports the Fourier series being pointwise convergent.

There is an issue with the model, Figure 11a shows while the two fundamental waves predicted analytically, travel in opposite directions at the same speed and reflect off the boundaries producing positive and negative interference, their summation is incorrect. The two waves should not combine into a sharp peak but instead recombine into a larger square wave, as seen in the finite difference method results (Figure 13). This is due to coding error that was not discovered until late into the investigation process and was not corrected due to time constraints.

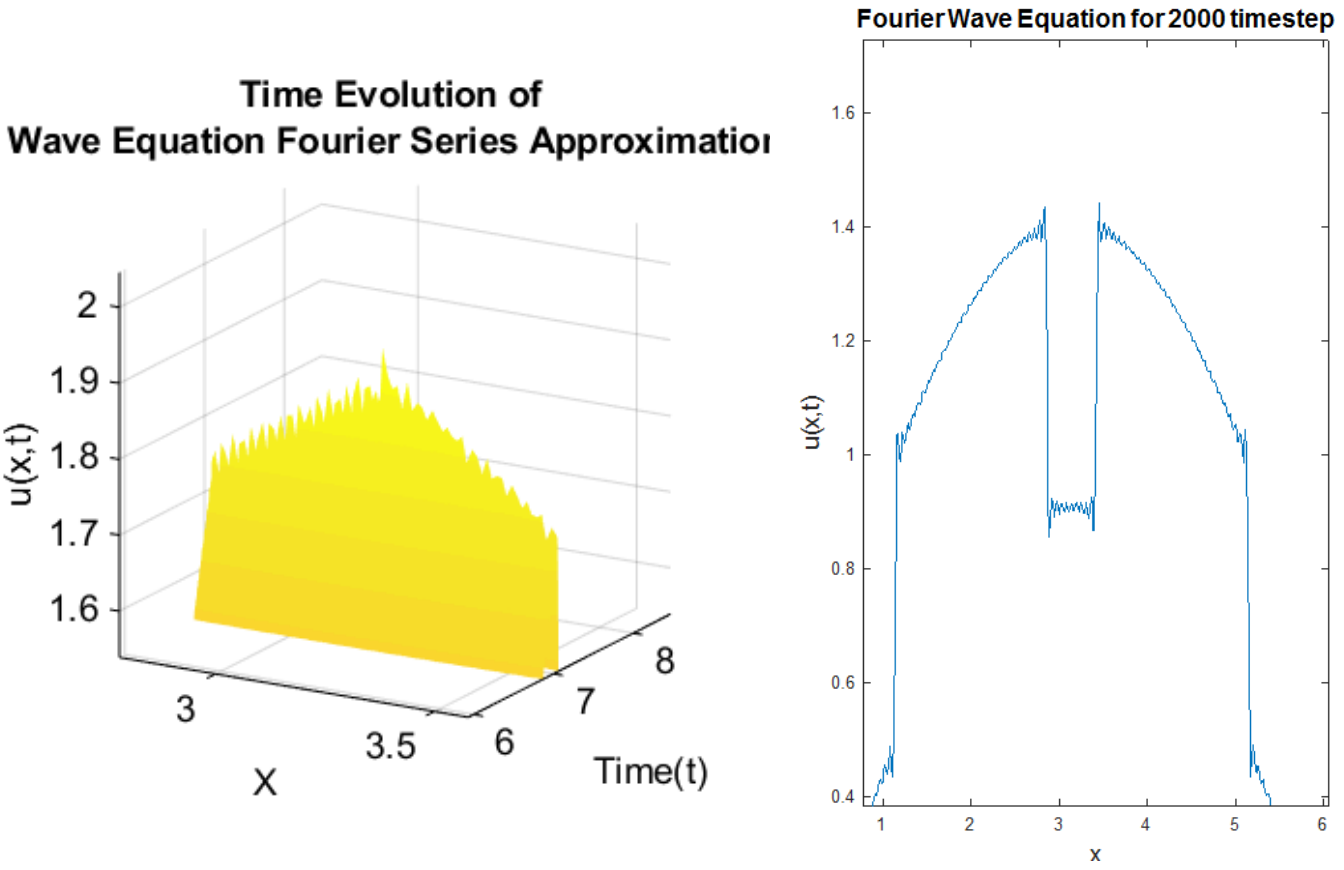


Figure 12(a) and 12(b), (left, right) are sections of the time evolution of Fourier series approximation for the wave equation model used in figures 11(a). ($\tau = 0.01$; $h = 0.0125$, 2000 time steps)

Wave Equation Finite Difference Scheme

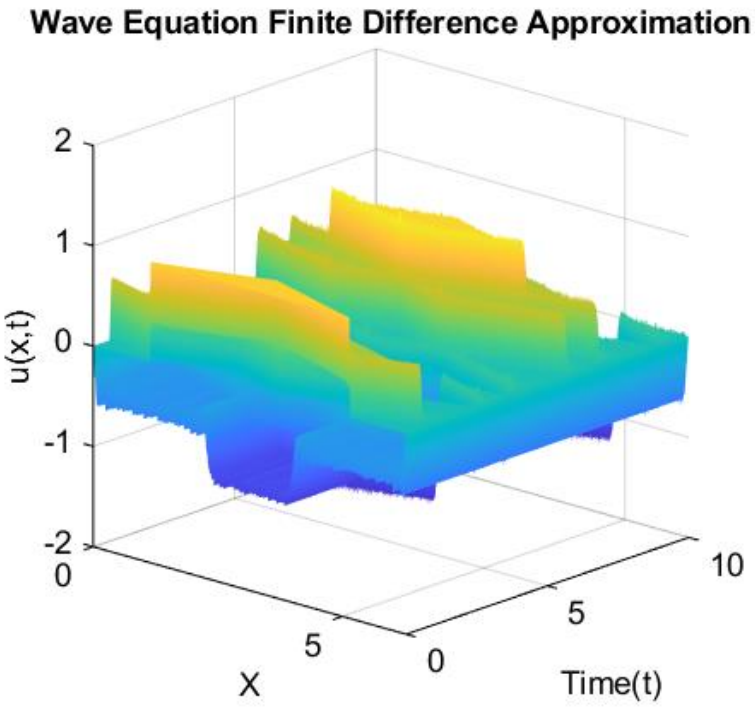


Figure 13 depicts the time evolution of the Finite Difference approximation for the wave equation. The profile aligns with the physical interpretation of a trapped standing wave reflecting causing positive and negative superposition ($\tau = 0.005$; $h = 0.0125$, 2000 time steps)

The finite difference method for the wave equation produced the physically correct model for the super position of two square waves, periodically reproducing the initial condition. Due to the severity of the Fourier modelling error, the two numerical methods could not be compared effectively. This was unfortunate as the Gibbs phenomena affected this scheme in a similar manner as the Fourier approximation, as predicted. As the Gibbs phenomena was present in both approximations, and had the Fourier approximation been coded correctly, this would have provided a more ideal comparison between the two approximations than the Diffusion Equation.

Convergence of the Wave Equation Finite Difference Scheme

The same principles to test for convergence the Crank Nicholson scheme were used, again the average difference of U(x,t) values was used to account for the large fluctuations at various points arising from the Gibbs Phenomena. However, the limitation of the CFL condition also required the temporal step size to be adjusted during the spatial convergence testing due to the reduction in the spatial step size. The absence of the "artefact phenomena" allowed for a lower spatial and temporal tolerance of 0.05 to be chosen.

For spatial convergence, the spatial step size required was 3.125e-03, with a temporal step size of 6.25e-04 required to satisfy the CFL condition. The average difference in U(x,t) was 2.233077e-02 which was within the desired tolerance.

The temporal step size required for stable spatial convergence was also sufficient to satisfy the temporal convergence. The average difference in U(x,t) was 2.019187e-02, which was within the desired tolerance.

Therefore, the Wave Equation Finite Difference scheme is empirically both convergent for space and time.

Truncation Error Analysis

As stated previously, the CFL condition complicated the error testing. The same methodology was applied as before; assuming the convergent step sizes to provide an exact solution. Temporal error testing was carried out for the wave equation, but similar reasoning as that for choosing a point closest to the arbitrary spatial point "pi" was required to find an approximately identical time point. The arbitrary point chosen for analysis around was 6 seconds, again at the spatial point pi. This double approximation meant that the temporal error testing produced less reliable data than the spatial testing.

The error between the average values of U(x) were also calculated to reduce the error from choosing the point closest to pi. More importantly for the wave equation, this was also used to mitigate for the Gibbs phenomena. As when analysing the difference at one specific point, if the Gibbs Phenomena shifts at all due to the changing step size, this can produce relatively extremely large in the U(x,t) value at that point.

Wave Finite Difference Spatial Truncation Error Analysis Results

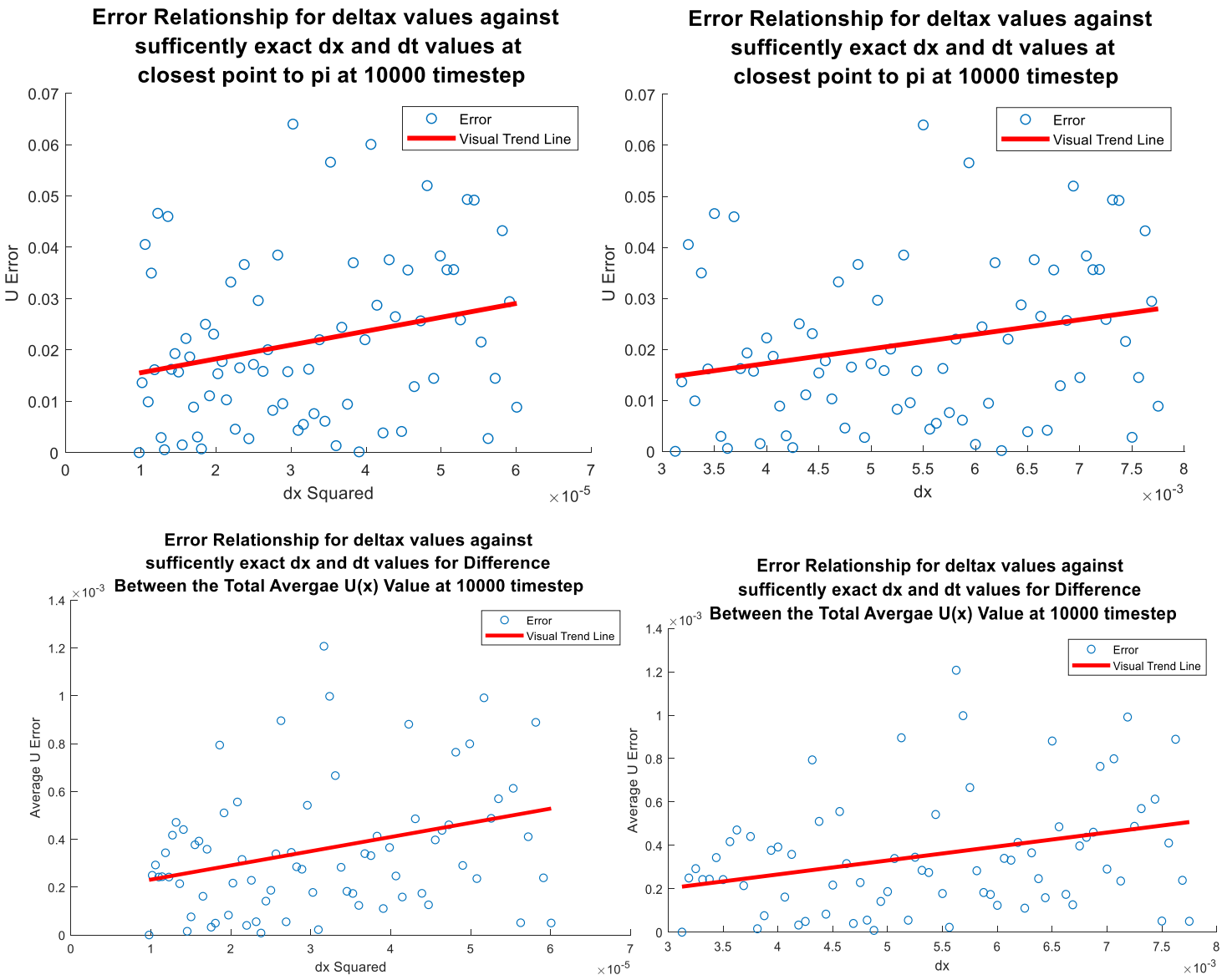


Figure 14(a), 14 (b), 14(c) and 14 (d), (top left, top right, bottom left, bottom right) Depict the Wave Equation Finite Difference schemes spatial error relationships as titled. Trend lines are for visual aid and for calculating rudimentary Chi squared values: 1.335e-02, 1.328e-02, 2.539e-04, and 2.522e-04 respectively. Note that the Average difference values are much less (y axis scale e-03) than those from pi; yet the profile remains very similar. ($\tau\ exact\ = 6.250000e\text{-}04$; $h\ exact\ = 3.125e\text{-}03$)

As expected, the error taken from the single pi point are of orders of magnitude larger than that taken from the average value. All relationships (Figure 14) show very similar spread and trends, as well as extremely similar chi squared values between dx and dx squared plots. Due to this large spread, the error profile appears more similar to that expected from a random distribution than a trend relation. Yet, especially noticeable from Figure 14(d), there does appear to be a slight positive correlation. While the Chi squared values for the dx fit are systematically lower than for dx squared

fit, the difference is extremely small. This fact, along with the profile of each plot being visually similar, it cannot be reasonably concluded what the empirical spatial error relationship is; however, the basic principle that errors increases with step size seems to hold. Potential causes of this seeming discrepancy from the analytical expectation include, manual coding errors, errors arising from the point selection and averaging method discussed above, the large spread resulting from the Gibbs function, and the error propagation of the initial approximation used to start the model flowing. Further tests of the scheme with a continuous initial condition would be beneficial.

Wave Finite Difference Temporal Truncation Error Analysis Results

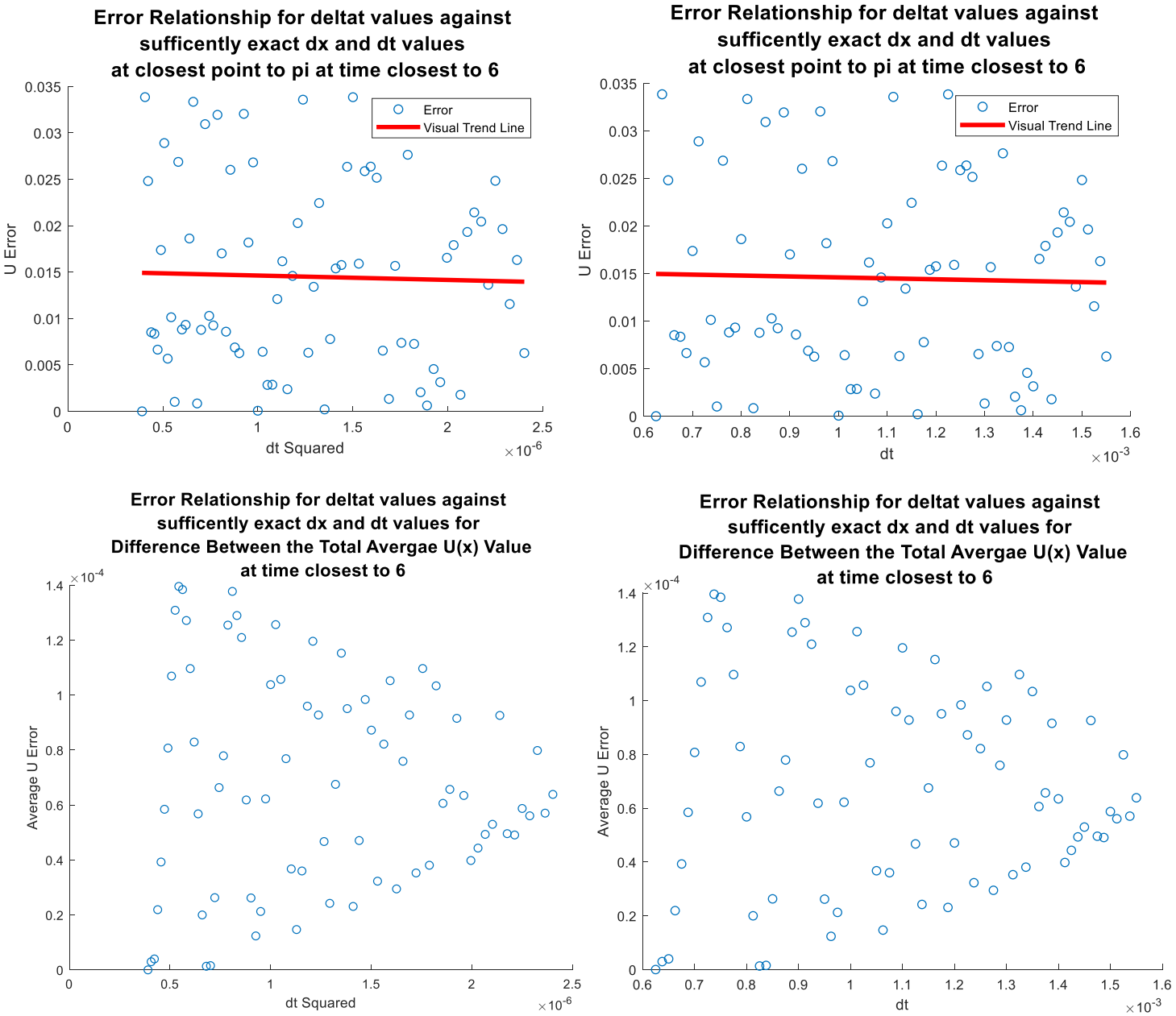


Figure 15(a), 15 (b), 15(c) and 15(d), (top left, top right, bottom left, bottom right) Depict the Wave Equation Finite Difference schemes temporal error relationships as titled. Note that the Average difference values are much less (y axis scale e-03) than those from only around pi. ($\tau\ exact$ = 6.250000e-04; $h\ exact$ =3.125e-03)

The temporal error distribution around pi and the 6 second time point (Figures 15(a,b)) are even more similar to that expected from a random distribution than the spatial error figures. This is likely due to the compounding of the uncertainty associated with not having exactly corresponding spatial and temporal points for the error calculation, along with the previously mentioned error sources such as the Gibbs phenomena. The slight negative gradient of the fitted trend line is hypothesised to be due to this seeming stochasticism. The Average difference relationships conical profiles imply that the basic expectation that the error increases as temporal step size increase holds. The data however is not conclusive enough to suggest the empirical order of temporal truncation error. Further investigation using a continuous initial condition is suggested.

Technique to Reduce Gibbs Phenomena

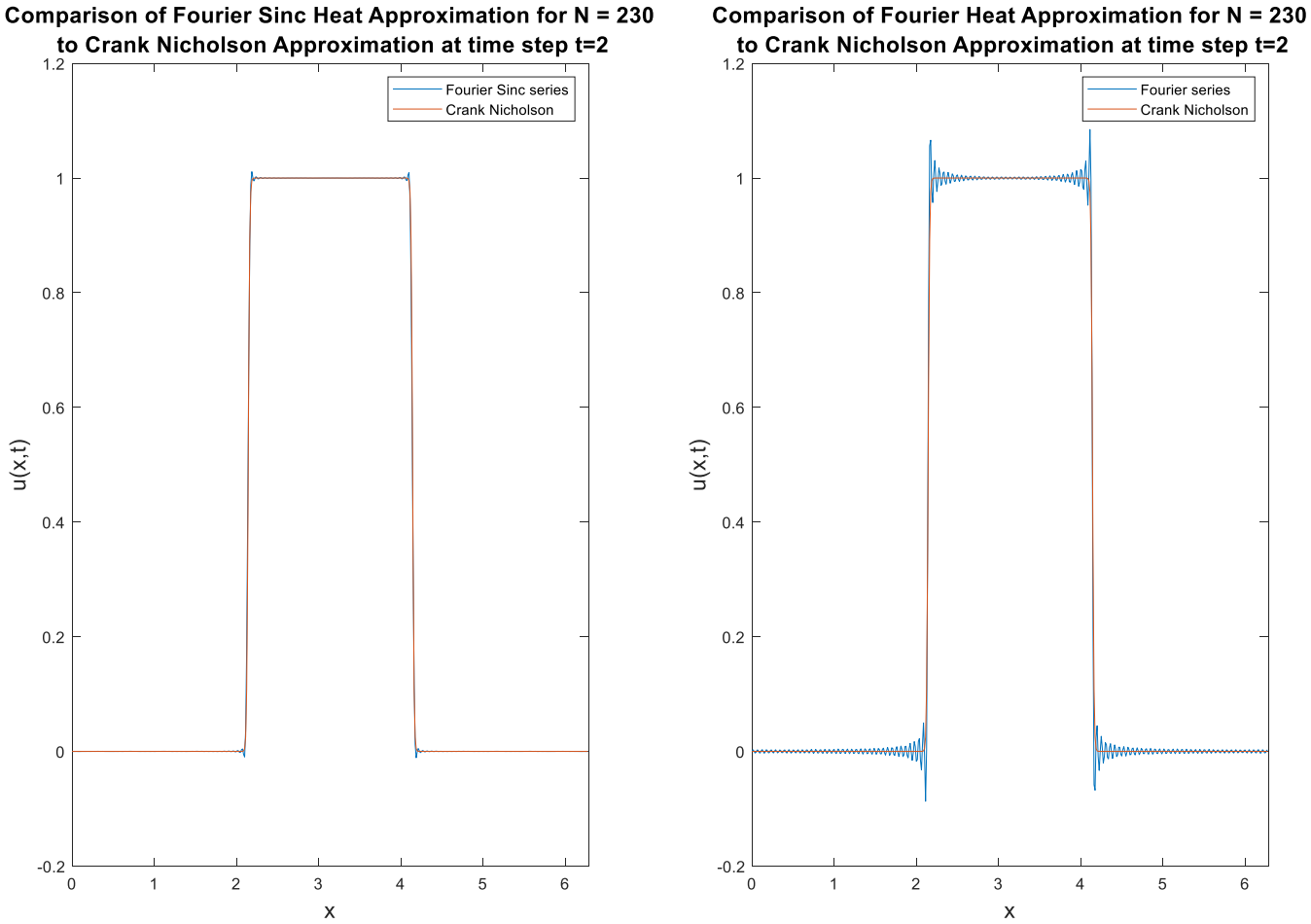


Figure 16(a) and 16(b), (left, right) depict the Fourier series approximation comparison to the exact square wave initial condition with and without using the sinc function correction. Using the sigma approximation method (ie applying the sinc function) greatly reduces the Gibbs Phenomena. ($\tau = 3.125e-05$; $h = 1.25e-02$)

A common method used to greatly reduced, but not entirely eliminate, the Gibbs Phenomena is by using the Sigma approximation, a modified version of the Fourier approximation. This approximation makes use of the normalised Sine cardinal function which is used in information and signal processing. Due to time constraints, no further research was conducted into this methodology; however, as seen in Figure 16, it is extremely promising and warrants further investigation.

Equation 17, Sigma Approximation for Equation 5

$$U(x,t) = \sum_{n=1}^{m-1} sinc\left(\frac{n}{m}\right) a_n e^{-nt} \sin\left(\frac{nx}{2}\right)$$

Where all symbols have their previous meanings, and m is the maximum number of terms in the summation, and sinc is the sine cardinal function.

Conclusion

Numerical methods have undeniable benefits for diffusion and wave problems, and allow for solutions which would take significantly more time to calculate fully analytically, if at all possible. The methods used in this report, with the exception of the Fourier wave equation scheme most likely due to coding error, produced physically meaningful approximate solutions which concur with the expected result for the Dirichlet boundary condition. Modelling square waves proved cumbersome due to the Gibbs Phenomena, which hard large repercussions for later analysis.

All schemes were shown to be convergent both spatial and temporally as expected theoretically, with the exception of the spatial convergence for the Crank Nicolson scheme due to the artefact anomalies. These anomalies also draw concerns for the unconditional stability of the Crank-Nicholson method, but as they are hypothesized to be related to the Gibbs function, and more likely due to human coding errors, further analysis should be conducted and is expected to affirm the full spatial convergence and unconditional stability of the scheme.

Analysis as to whether finite difference methods or Fourier approximations are more appropriate could not be conducted due to the effect of the Gibbs function asymmetrical impact for the diffusion equation, and the invalid wave equation Fourier approximation.

No conclusive supporting evidence for the true empirical order for the finite difference methods truncation error was produced. Further investigations into using smoother initial conditions are required, as well as further research into the Sigma approximation.

APPENDIX

In order to solve implicit methods, the set of simultaneous equations may be placed in matrix format (Figure (17)).

$$\begin{pmatrix} b_1 & -c_1 & 0 & . & . & 0 \\ -a_2 & b_2 & -c_2 & . & . & 0 \\ 0 & -a_3 & b_3 & -c_3 & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & . & . & . & -a_N & b_N \end{pmatrix} \begin{pmatrix} u_1^{n+1} \\ u_2^{n+1} \\ u_3^{n+1} \\ . \\ . \\ u_N^{n+1} \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ . \\ . \\ d_N \end{pmatrix}$$

Figure 17, Matrix form for the implicit simultaneous equations. Has the form of a Tridiagonal matrix

For the Crank-Nicholson scheme, from the coefficients where ($\alpha = \frac{k\tau}{h^2}$):

$a = \alpha/2$ $\quad b = (1+\alpha)$ $\quad c = \alpha/2$ $\quad d = \frac{\alpha}{2} U_{j-1}^n + (1+\alpha) U_j^n - \frac{\alpha}{2} U_{j+1}^n$

Altered Crank-Nicholson

$$-a_j U_{j-1}^{n+1} + b_j U_j^{n+1} - c_j U_{j+1}^{n+1} = d_j$$

As before, due to previous time step, d is completely known. Matrices may be solved via LU factorisation or Gaussian elimination with back substation, however since the matrix is tridiagonal, the Thomas / Tridiagonal matrix algorithm can be used in order to reduce the matrix to its upper triangular form. This extremely beneficial as the order of operations for the Thomas algorithm is O(n), compared to $O(n^3)$ for Gaussian elimination.
When the matrix has been reduced to the upper triangular form, the equations are in the form:

Equation 18

$$U_j^{n+1} - e_j U_{j+1}^{n+1} = f_j$$

Where the new coefficients given by the Thomas algorithm are:

$$e_j = \frac{c_j}{b_j - a_j e_{j-1}} \qquad f_j = \frac{d_j + a_j f_{j-1}}{b_j - a_j e_{j-1}}$$

Giving the upper triangular matrix:

$$\begin{pmatrix} 1 & -e_1 & 0 & . & . & 0 \\ 0 & 1 & -e_2 & . & . & 0 \\ 0 & 0 & 1 & -e_3 & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & . & . & . & 0 & 1 \end{pmatrix} \begin{pmatrix} u_1^{n+1} \\ u_2^{n+1} \\ u_3^{n+1} \\ . \\ . \\ u_N^{n+1} \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ . \\ . \\ f_N \end{pmatrix}$$

Figure *18, Upper triangular matrix form for the implicit simultaneous equations*

It is clear from Figure 18 that $U_N^{n+1} = f_N$. Knowing this value, it can be substituted in to the 2 variable equation above to retrieve U_{N-1}^{n+1}, thus the standard method of back substitution is conducted all the way until U_1^{n+1}. This can be coded as for loop from j=N-1 → 1 over:

$$U(j) = f(j) + e(j)U(j+1)$$

In order to implement the Crank-Nicholson method, it should be noted that a(1) and c(N) must be set to zero as they do not exist in the matrix. Similarly, the corresponding terms should be omitted for d(1) and d(N) calculations, as U_{1-1} and U_{N+1} do not exist (assuming the numerical tool used array starts at 1,not zero, alter as necessary). This should force the Dirichlet boundary condition.

The implementation of the Von Neumann condition, like the FTCS method, requires $U_1^{n+1} = U_2^{n+1} and\ U_N^{n+1} = U_{N-1}^{n+1}$; However, array elements b_1 and b_N must be changed as well:

$$b_1 = 1 + \frac{\alpha}{2} \qquad \text{and} \qquad b_N = 1 + \frac{\alpha}{2}$$

Crank-Nicholson with periodic boundary conditions

To apply periodic boundary conditions to the Crank-Nicholson method, the matrix must be altered to what is known as a cyclic tridiagonal matrix. (Figure (19))

$$\begin{pmatrix} b & -c & 0 & . & . & -a \\ -a & b & -c & . & . & 0 \\ 0 & -a & b & -c & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ -c & . & . & . & -a & b \end{pmatrix} \begin{pmatrix} u_1^{n+1} \\ u_2^{n+1} \\ u_3^{n+1} \\ . \\ . \\ u_N^{n+1} \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ . \\ . \\ d_N \end{pmatrix}$$

Figure 19, Cyclic Tridiagonal Matrix

The Thomas algorithm can only be applied to Tridiagonal matrices (figure (19)) ; however, by applying the Sherman-Morrison formula, the periodic matrix above can be solved by two consecutive Thomas algorithms.

Sherman-Morrison formula

$$(A + xy^T)^{-1} = A^{-1} - \frac{A^{-1}xy^TA^{-1}}{1 + y^TA^{-1}x}$$

Where A is an invertible tridiagonal square matrix, n x n
x, y are n x 1 vectors

Setting $(A + xy^T)$ equal to Cyclic tridiagonal matrix, such that:

Equation 19

$$(A + xy^T)u = d$$

And letting:

$$x = [-b\ 0\ 0 \dots 0 - c]^T \qquad \text{and} \qquad y = \left[1\ 0\ 0 \dots 0\ \frac{a}{b}\right]^T$$

Matrix A is found below:

$$\begin{pmatrix} b & -c & 0 & . & . & -a \\ -a & b & -c & . & . & 0 \\ 0 & -a & b & -c & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ -c & . & . & . & -a & b \end{pmatrix} = A + \begin{pmatrix} -b \\ 0 \\ 0 \\ . \\ . \\ -c \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & . & . & \frac{a}{b} \end{pmatrix}$$

$$\begin{pmatrix} b & -c & 0 & . & . & -a \\ -a & b & -c & . & . & 0 \\ 0 & -a & b & -c & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ -c & . & . & . & -a & b \end{pmatrix} = A + \begin{pmatrix} -b & 0 & . & . & . & -a \\ 0 & 0 & . & . & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ -c & . & . & . & . & \frac{-ac}{b} \end{pmatrix}$$

$$\therefore A = \begin{pmatrix} 2b & -c & 0 & . & . & 0 \\ -a & b & -c & . & . & 0 \\ 0 & -a & b & -c & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & . & . & . & -a & b + \frac{ac}{b} \end{pmatrix}$$

This gives a Tridiagonal matrix for which the Thomas algorithm can be applied, however the variable being sought is U. Returning to Equation 7 and using Sherman-Morison formula, as well as the property of linear algebra that $A^{-1}A = I$ (where I is the indemnity matrix):

$$(A + xy^T)^{-1}(A + xy^T)U = d\left(A^{-1} - \frac{A^{-1}xy^TA^{-1}}{1 + y^TA^{-1}x}\right)$$

$$\Longrightarrow U = d\left(A^{-1} - \frac{A^{-1}xy^TA^{-1}}{1 + y^TA^{-1}x}\right)$$

Where U is a matrix

Substituting now for pre-transformation variables w and z, where A is the tridiagonal matrix found earlier.

$$d = Aw \qquad \text{and} \qquad x = Az$$

$$U = Aw\left(A^{-1} - \frac{A^{-1}Azy^TA^{-1}}{1 + y^TA^{-1}Az}\right) = Aw\left(A^{-1} - \frac{zy^TA^{-1}}{1 + y^Tz}\right) = w - \frac{zy^Tw}{1 + y^Tz}$$

Equation 19

$$U = w - \left(\frac{y^Tw}{1 + y^Tz}\right)z$$

The solution for U_j^{n+1} given by Equation 19 can thus be solved in terms of $w_j^{n+1}$and z_j^{n+1}. The Equations for these variables above, as A is tridiagonal and both d and x are known, take the same form as figure 3, and therefore may both be solved for using two separate Thomas algorithms. As stated earlier, in this investigation, the conducting medium's conductivity (K) is assumed to be uniform and not a function of x, which means that $\alpha = \frac{k\tau}{h^2}$ is a constant. The elements of the matrix (a, b, c) are all functions of α, and if α is a constant then the Thomas algorithm to calculate z can be conducted outside of the computational loop, which reduces operation time. The implementation of A is the same as for the standard Crank-Nicholson scheme, but requires b(1) = 2(1+ α) and b(N) = (1++ α)+ $\frac{\alpha}{1+\alpha}$ which is obvious from the matrix itself.

Bibliography

Peter Duffy , *An Introduction to Finite Difference Methods for Advection Problems,* Dep. of Maths Physics, UCD

Gerald Recktenwald, *Crank Nicolson Solution to the Heat Equation*, Dep of Mechanical Engineering, Portland State University

Gerald W. Recktenwald, 2011, *Finite-Difference Approximations to the Heat Equation*

Strikwerda, John C. (1989). Finite Difference Schemes and Partial Differential Equations (1st ed.). Chapman & Hall. pp. 26, 222. ISBN 0-534-09984-X*.*

http://texas.math.ttu.edu/~gilliam/fall03/m4354_f03/heat_N_web/heat_ex_homo_neum.pdf

http://mathworld.wolfram.com/LanczosSigmaFactor.html

http://mathworld.wolfram.com/SincFunction.html

https://www.math.ubc.ca/~peirce/M257_316_2012_Lecture_8.pdf

http://www-users.math.umn.edu/~olver/num_/lnp.pdf

http://hplgit.github.io/INF5620/doc/pub/main_trunc-2up.pdf

http://mathfaculty.fullerton.edu/mathews/n2003/differentiation/numericaldiffproof.pdf

http://web.math.ucsb.edu/~grigoryan/124B/lecs/lec18.pdf

https://web.stanford.edu/class/math220a/handouts/waveequation2.pdf

https://www.math.ubc.ca/~peirce/HeatProblems.pdf