Survey of Applied Mathematics Techniques

Brian Wetton [1]

June 21, 2018

[1] www.math.ubc.ca/~wetton, wetton@math.ubc.ca

Lecture 2

Finite Difference Methods

2.1 Motivation

Many problems in Science, Engineering, and Finance involve the solution of differential equations (DE). Often these problems cannot be solved analytically but must be approximated numerically. This approximation must be done to a certain precision that depends on the application. While the differential equations do not exactly describe the real system they model (there is a modelling error) it is important to minimize the errors from the numerical approximation to be able to confirm whether the underlying mathematical model is valid. This is shown graphically in Figure 2.1.

2.2 A First DE Problem

Find $u(x)$, $x \in [0,1]$ with u and its derivatives 1-periodic that satisfies

$$-u'' + u = f(x) \tag{2.1}$$

at every x where $f(x)$ is a given C_2 (periodic) function. This is Problem A from lecture #1.

Here, C^2 (periodic) is the set of functions on the unit interval which have continuous and 1-periodic derivatives up to second order. In general, we write C^n for the set of functions that have continuous derivatives up to order n. These functions have a norm

$$\|u\|_{C^n} = \max_{0 \le j \le n} max_x |u^{(j)}(x)|$$

where $u^{(j)}$ denotes the j'th derivative of u. We will have other norms for functions, so we will label the norm we are using unless it is completely clear.

We will use the following theorem

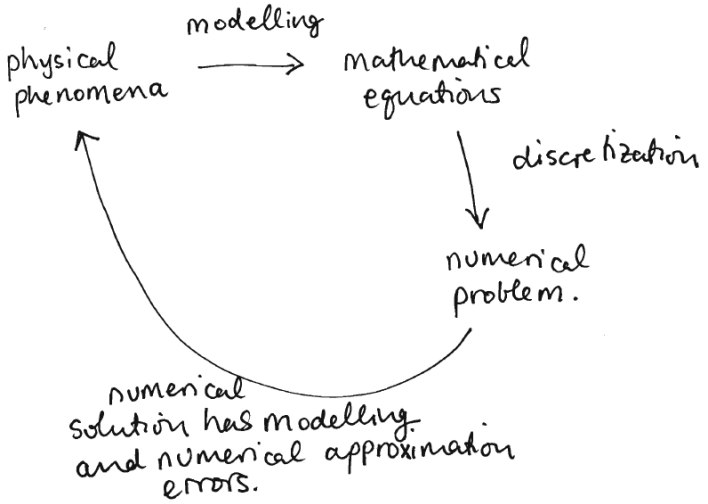

Figure 2.1: Stages in computational modelling. By reducing numerical errors, a clear picture of the model accuracy can be obtained.

Theorem 1 *Problem A (2.1) has a ! (unique) solution with* $u \in C^4$ *with the bound*

$$\|u\|_{C^4} \leq K\|f\|_{C^2}$$

for all given $f \in C^2$ *for some* K *that is independent of* f*.*

The inequality in the Theorem is known as an a-priori bound. For a given f we don't know the solution u but we know there will be a solution, there will only be one solution, and it will have four continuous derivatives with size limited by the derivatives up to second order of f.

2.3 Finite Difference Discretization

2.3.1 Discretization

Determining the solution $u(x)$ of Problem A (2.1) requires finding an infinite number of unknowns (the values of u at every point x in an interval). To proceed computationally, we need to deal with only a finite number of unknowns (discretization). Let's look first at a simple, Finite Difference (FD) discretization. Let U_i, $i = 1, 2, \ldots N$ approximate $u(x)$ at the ends of subintervals with length $h = 1/N$. That is

$$U_i \approx u(ih).$$

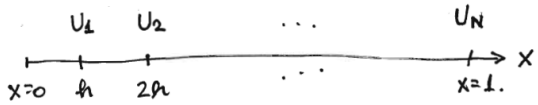

Figure 2.2: Uniform grid in spatial discretization.

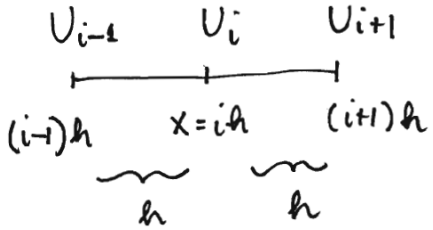

Figure 2.3: Values used in the finite difference approximation of the second derivative.

This is shown in Figure 2.2. Using the periodicity of the problem, we would have

$$U_0 = U_N \text{ and } U_{N+1} = U_1 \tag{2.2}$$

Convention: In these notes, I will always use lower case letters for exact solutions and upper case letters for numerically computed, approximate values.

2.3.2 Approximating Derivatives

Let us assume for the moment that we knew the exact solution, at least at grid points. We can use the values u_{i-1}, u_i and u_{i+1} to approximate $u''(ih)$ with the following formula:

$$D_2 u_i := \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} = u''(ih) + \frac{h^2}{12} u''''(\theta) \tag{2.3}$$

for some $\theta \in [(i-1)h, (i+1)h]$. The geometry of this linear combination of values is shown in Figure 2.3. The geometry and the weights of the linear combination is called the *stencil* of the discrete approximation.

This can be derived in a number of ways, starting with Taylor Series. Two approaches are described in section 2.4 below. The last term on the right is an error term (we wanted $u''(ih)$ but got that extra term as well). Since u'''' is bounded, we can guarantee answers as accurate as we want by taking $h \to 0$

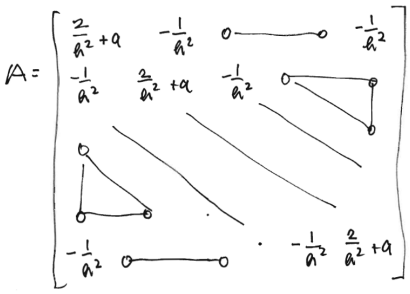

Figure 2.4: Structure of the matrix $\mathcal{A}$.

(that is, by refining the grid). The term h^2 in the error makes this a *second order approximation.*

2.3.3 Discrete Equations

Using the equation that u solves in Problem A (2.1) we can write

$$-D_2 u_i + u_i = f(ih) + \frac{h^2}{12} u''''(\theta) \tag{2.4}$$

The last term above is again an error term. We call it the *truncation error*, the residual when we put the exact solution into a discrete equation. The truncation error goes to zero as $h \to 0$. We say that the discrete equation is *consistent.* By ignoring a small (as $h \to 0$) residual, we could specify our discrete scheme for the approximate values U_i:

$$-D_2 U_i + U_i = F_i \quad \text{for } i = 1, 2, \ldots N \tag{2.5}$$

where $F_i = f(ih)$ are given values. Note that (2.5) is a linear system with N equations for N unknowns. The system can be written in vector form

$$\mathcal{A}\mathbf{U} = \mathbf{F} \tag{2.6}$$

where $\mathcal{A}$ is an $N \times N$ matrix with form shown in Figure 2.4. We will show below that $\mathcal{A}$ is invertible, so the discrete scheme (2.5) has a solution $\mathbf{U}$ for every given right hand side $\mathbf{F}$ for any h.

The matrix $\mathcal{A}$ has mostly zeros. As $N \to \infty$ ($h \to 0$) the fraction of non-zeros decreases. We call such matrices *sparse.* However, $\mathcal{A}^{-1}$ is not sparse. It has no zero entries. This implies that all entries of $\mathbf{F}$ affect solution values at all locations. This is characteristic of elliptic problems.

2.3.4 The Next Step: Test on a Known Problem

We have a scheme for our problem. The very next thing we should do is test out the scheme on as simple a problem as we can that has a known solution. One trick is to *pick* the exact solution u, put it into the DE, and whatever residual there is call it $f(x)$. We can then use the values of f on the grid ($\mathbf{F}$) in the discrete scheme and compute $\mathbf{U}$. We can then compare $\mathbf{U}$ to the exact u at grid points to see how accurate the scheme is at various grid resolutions. For example, we could choose

$$u(x) = e^{\cos x}$$

and find

$$f(x) = (a - \sin^2 x + \cos x)e^{\cos x}$$

where we have taken the periodic interval of Problem A (2.1) to be $[0, 2\pi]$ instead of $[0, 1]$ to make these expressions a little simpler. Choosing good test examples is somewhat of an art. You want them to represent the type of problem you are interested in, but much easier to compute (possibly lower dimensional) with an exact solution you know. In some cases, it is not possible to find an exact solution to a representative problem. In well established fields, there are often benchmark problems with high accuracy solutions you can use to test new schemes. As you are picking test problems, make sure that the solutions you use do not have zero values for the derivatives that compose the truncation error. The scheme will behave anomalously (better than usual) on such problems.

In your test, you should see how errors behave as the grid is refined (by factors of 2 for example, $h = 1/10, 1/20, 1/40, \ldots$). There are three things to look for:

1. As $h \to 0$, do the computed values U tend to the exact u values? That is, do the errors tend to zero? If so, we say that the scheme converges. The computational test is not a proof of convergence, but is strong evidence for it.

2. Is there any odd behaviour in the errors $U - u$? Odd behaviour could be an indication of a program error or it might be just a characteristic of the scheme. Odd error behaviours can be called *numerical artifacts*.

3. If the test case is (roughly) similar to problems you are actually interested in, can you achieve the desired accuracy with an N (h) value that leads to a computation that takes an acceptable length of time? If not, you should spend some time exploring ways to make the computation more efficient or more powerful computer architectures.

Testing the method on the test example above with exact solution gives the results shown in Table 2.1. MATLAB code for this computation is provided. Plots of the errors as functions of x for $N = 40$ and $N = 80$ are shown in Figure 2.5. Note that the error in the computed solution goes down by a factor of approximately 4 when N is doubled (h is halved). This matches with our discussion above, where truncation error goes down by a factor of 4 when N is

N	$E_N = \max_i \lvert U_i - u(ih) \rvert$	E_N/E_{2N}
10	6.71e-2	4.24
20	1.58e-2	4.05
40	3.90e-3	4.01
80	9.73e-3	

Table 2.1: Errors in the finite difference method applied to the example in Section 2.3.4.

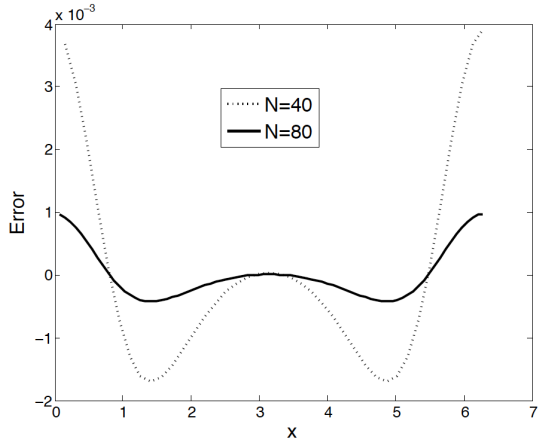

Figure 2.5: Spatial structure of the errors in the finite difference method applied to the example in Section 2.3.4.

doubled. The exact relationship between truncation error and solution error is derived in Section 2.6 where we prove convergence of the method. Note that we have chosen to measure the error in the maximum norm

$$\|\mathbf{U}\|_\infty := \max_i |U_i|$$

in this case. Other norms could have been used and for some types of problems, other norms are more appropriate.

2.4 Derivation of FD formulae

In this section, we prove the result (2.3). We begin with Taylor's Polynomial Approximation Theorem:

Theorem 2 *If $f(x)$ is C_{n+1} in a neighbourhood of a then if x is in that neighbourhood,*

$$f(x) = f(a)+f'(a)(x-a)+\frac{1}{2}f''(a)(x-a)^2+\cdots\frac{1}{n!}f^{(n)}(a)(x-a)^n+\frac{1}{(n+1)!}f^{(n+1)}(\theta)(x-a)^{n+1}$$

for some $\theta \in (a,x)$.

Here, the value of $f(x)$ is approximated by information at $x=a$. The first $n+1$ terms in the expression above are the n'th order Taylor Polynomial approximation of f based at $x=a$ and the last term is a remainder term, the error in the approximation. In case you have never seen the proof of this theorem, I will show the $n=1$ case (linear approximation) in Section 2.4 below. We can use the Theorem to verify the properties of D_2:

$$\begin{aligned}
D_2u_i &= \frac{u_{i-1}-2u_i+u_{i+1}}{h^2} \quad (\text{let } a=ih)\\
&= \frac{1}{h^2}\left(u(a-h)-2u(a)+u(a-h)\right) \quad (\text{use the Theorem})\\
&= \frac{1}{h^2}\left(u(a)-hu'(a)+\frac{h^2}{2}u''(a)-\frac{h^3}{6}u'''(a)+\frac{h^4}{24}u''''(\theta_1)\right.\\
&\quad \left.-2u(a)+u(a)+hu'(a)+\frac{h^2}{2}u''(a)+\frac{h^3}{6}u'''(a)+\frac{h^4}{24}u''''(\theta_2)\right)
\end{aligned}$$

where $\theta_1 \in ((i-1)h, ih)$ and $\theta_2 \in (ih,(i+1)h)$. Continuing with the expression above, we have the desired expression

$$\begin{aligned}
D_2u_i &= u''(a)+\frac{h^2}{12}\left(u''''(\theta_1)+u''''(\theta_2)\right)/2\\
&= u''(a)+\frac{h^2}{12}u''''(\theta)
\end{aligned}$$

for some $\theta \in (\theta_1,\theta_2) \subset ((i-1)h,(i+1)h)$ where in this last step, we have used the intermediate value theorem.

This shows that the D_2 stencil has the desired properties. We could have found the coefficient values in the stencil starting with the same machinery. Taking the Ansatz

$$D_2u_i = \alpha u_{i-1}+\beta u_i+\gamma u_{i+1}$$

and wanting D_2u_i to be $u''(ih)$ with a small truncation error leads to the following requirements for the coefficients when u_{i-1} and u_{i+1} are expanded in Taylor polynomials as above:

$$\begin{aligned}
O(1) \text{ terms:} &\quad \alpha+\beta+\gamma = 0\\
O(h) \text{ terms:} &\quad -h\alpha+h\gamma = 0\\
O(h^2) \text{ terms:} &\quad h^2\alpha/2+h^2\gamma/2 = 1.
\end{aligned}$$

This is a linear system that can be solved for $\alpha = 1/h^2$, $\beta = -2/h^2$, $\gamma = 1/h^2$. Note that the $O(h^3)$ term also cancels with these parameters. It is typical for centred difference approximations of even derivatives that they have order of accuracy one order higher than "expected".

Proof of Theorem 2, $n = 1$ case

We will use Rollé's Theorem from first year calculus:

Theorem 3 *If f is differentiable in (a, b) and continuous in $[a, b]$ and $f(a) = 0$ and $f(b) = 0$ then*

$$f'(\theta) = 0$$

for some $\theta \in (a, b)$.

Consider the $n = 1$ (linear) Taylor Approximation of Theorem 2 at a specific point $x = b$. Then consider the function

$$q(x) = f(x) - L(x) - \frac{f(b) - L(b)}{(b-a)^2}(x-a)^2$$

where $L(x) = f(a) + f'(a)(x - a)$ is the linear approximation. We want to investigate $f(b) - L(b)$, the error of the linear approximation at $x = b$. This is a constant that appears in the $q(x)$ function above. Note that $q(a) = 0$ and $q(b) = 0$ so using Rollé we know that $q'(\theta_1) = 0$ for some $\theta_1 \in (a, b)$. Also, $q'(a) = 0$ so using Rollé again we have $q''(\theta) = 0$ for $\theta \in (a, \theta_1) \subset (a, b)$. We can compute

$$q''(x) = f''(x) - 2\frac{f(b) - f(a)}{(b-a)^2}$$

so $q''(\theta) = 0$ gives

$$f(b) - L(b) = \frac{f''(\theta)}{2}(b-a)^2,$$

the desired result.

2.5 Direct Solution of Sparse Linear Systems

Consider the structure of the nonzero entries of the matrix $\mathcal{A}$ in the discrete problem (2.6) shown in figure 2.4. A matrix such that

$$|A_{ii}| > \sum_{j \neq i} |A_{ij}|$$

for every row i is said to be strictly diagonally dominant. Our matrix $\mathcal{A}$ has this property. It can be shown that Gaussian elimination can be applied to such matrices stably (that is, without significant growth of floating point round-off errors) without pivoting. It can be seen that Gaussian elimination and back substitution can be done for the system (2.6) with a finite number of operations per row independent of the number of rows. The structure of the LU decomposition of $\mathcal{A}$ is shown in figure 2.6. The total operation count to find the solution is $O(N)$, that is the operations are bounded by a constant times N. Thus a direct solver applied to this problem taking into account the sparsity of $\mathcal{A}$ has optional complexity.

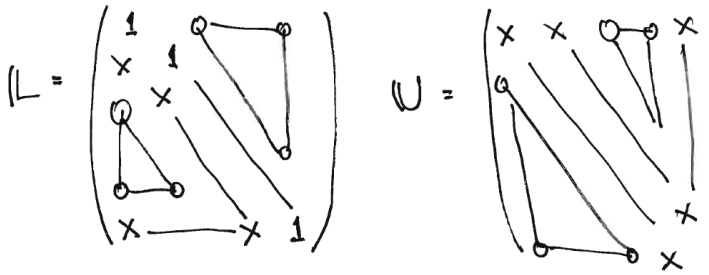

Figure 2.6: Structure of the LU decomposition of the matrix A.

Note: This is close to the ideal example for sparse numerical linear algebra. The matrix A is (almost) narrowly banded, is diagonally dominant, is symmetric and positive definite. We will see that the situation for discretizations in higher dimensional problems is not as ideal. Still, it is possible now to solve 3D problems of *modest* size with direct solvers on basic computers.

2.6 Convergence Proof

We came up with our discrete scheme (2.5) by neglecting a small truncation error in relationships the exact solution at the grid points satisfy (2.4):

$$\begin{aligned} \mathcal{A}\mathbf{U} &= \mathbf{F} \\ \mathcal{A}\mathbf{u} &= \mathbf{F} + \tau \end{aligned}$$

where $\tau_i = h^2 u^{(4)}(\theta_i)/12$ are the truncation errors at every grid point. As discussed above, the truncation errors go to zero as $h \to 0$ (the definition of a consistent scheme). The vector of errors in the computed solutions at grid points is

$$\mathbf{E} = \mathbf{U} - \mathbf{u}.$$

For this linear problem, we can take the difference of the two equations above to obtain

$$\mathcal{A}\mathbf{E} = \tau \quad \text{or} \quad \mathbf{E} = \mathcal{A}^{-1}\tau. \tag{2.7}$$

Note: We haven't shown yet that A is invertible (but our numerical implementation suggests it is for all h) and in practice we would never compute the full matrix $\mathcal{A}^{-1}$. This is just a representation for the theory.

Considering (2.7) we see that with τ small, $\mathbf{E}$ will be small as long as multiplying by $\mathcal{A}^{-1}$ does not increase its size by more than a constant (independent of h). Formally, we want to have the following property:

$$\|\mathcal{A}^{-1}\mathbf{b}\|_\infty \leq C\|\mathbf{b}\|_\infty \tag{2.8}$$

for all $\mathbf{b}$ with C independent of h. Here,

$$\|\mathbf{b}\|_\infty := \max_{i=1,\ldots N} |b_i|$$

is the maximum norm of vectors. For fixed h we can define

$$\|\mathcal{A}^{-1}\|_\infty := \max_{\mathbf{b}\neq\mathbf{0}} \frac{\|\mathcal{A}^{-1}\mathbf{b}\|_\infty}{\|\mathbf{b}\|_\infty}.$$

This is known as an induced matrix norm. The value $\|\mathcal{A}^{-1}\|_\infty$ is the smallest value C for fixed N such that (2.8) holds for all $\mathbf{b}$. The property we are looking for is then that

$$\sup_h \|\mathcal{A}^{-1}\|_\infty$$

is finite. This property defines *maximum norm stability* of the scheme. Showing stability of numerical schemes in general is quite difficult, but easy for this particular scheme. Consider the left hand equation of (2.7) where we look back to (2.5) to see the details of the matrix $\mathcal{A}$:

$$\begin{aligned} -D_2E_i + E_i &= \tau_i \\ -\frac{1}{h^2}E_{i-1} + (\frac{2}{h^2}+1)E_i - \frac{1}{h^2}E_{i+1} &= \tau_i. \end{aligned} \tag{2.9}$$

Suppose that the $\max_i |E_i|$ is attained at an index j and that $E_j > 0$. Thus, $E_{j-1} \leq E_j$ and $E_{j+1} \leq E_j$ and thus

$$2E_j - E_{j+1} - E_{j-1} \geq 0$$

When this used in (2.9) we find that

$$E_j \leq \tau_j \quad \text{or} \quad E_j \leq |\tau_j|.$$

Since $\|\mathbf{E}\|_\infty = E_j$ we have

$$\|\mathbf{E}\|_\infty \leq \|\tau\|_\infty.$$

Since $\mathbf{E} = \mathcal{A}^{-1}\tau$ we have shown the maximum norm stability of the scheme. If $\|\mathbf{E}\|_\infty$ is attained at an index j where $E_j < 0$, a similar argument applies.

Recall that $\tau_i = \frac{h^2}{12}u^{(4)}(\theta_i)$ for some $\theta_i \in ((i-1)h, (i+1)h)$ so

$$\|\tau\|_\infty \leq \frac{\|u\|_{C_4}}{12}h^2.$$

using Theorem 1 we have

$$\|\tau\|_\infty \leq \frac{K\|f\|_{C_2}}{12}h^2.$$

Using the stability result we derived,

$$\|\mathbf{E}\|_\infty \leq \frac{K\|f\|_{C_2}}{12}h^2. \tag{2.10}$$

This proves the convergence of the scheme since as $h \to 0$, $\|\mathbf{E}\|_\infty \to 0$. With $\|\mathbf{E}\|_\infty < Ch^2$ we say that the convergence is *second order*.

The analysis leading to the convergence result above is an example of the Lax Equivalence Theorem for linear problems, often stated informally as

```
Consistency + Stability = Convergence
```

Note that in (2.10) that if $\|f\|_{C_2}$ is large (*i.e.* f is highly oscillatory) then h must be quite small to make the errors small. This makes sense: to resolve oscillatory behaviour, a fine grid is needed.

2.7 von Neumann Analysis

Consider again our discretization of Problem A (2.1),

$$\mathcal{A}\mathbf{U} = \mathbf{F}, \quad \mathcal{A} = -D_2 + aI.$$

It is possible to show the stability of the scheme in other norms. Here, we will consider the l_2 norm, also known as the energy norm or the (scaled) Euclidean norm

$$\|\mathbf{b}\|_2 := \sqrt{h \sum_{i=1}^{N} |b_i|^2}. \tag{2.11}$$

Note that the scaling by h is such that if $b_i = C$ for all i, then $\|\mathbf{b}\|_2 = C$ for every N (h) since $N = 1/h$. Also, this makes the norm analogous to the continuous energy norm for the space of functions L_2:

$$\|f\|_{L_2} = \sqrt{\int_0^1 |f(x)|^2 dx}.$$

The discrete l_2 norm (2.11) can be seen as an approximation (trapezoidal rule) of the continuum norm.

2.7.1 Discrete Fourier vectors

To proceed, we introduce the complex valued Discrete Fourier (DF) vectors $\mathbf{f}_\alpha$:

$$f_{\alpha,j} = e^{2\pi i \alpha j h}.$$

To clarify the labelling, $\mathbf{f}_\alpha$ is a vector with N complex components for every $\alpha = 0, \ldots N-1$. The N components are indexed by j above and the i is the pure imaginary unit. The set $\{\mathbf{f}_\alpha\}$ is a basis of $\mathbf{C}_N$, orthonormal in the inner product that corresponds to the l_2 norm:

$$\begin{aligned} (\mathbf{a}, \mathbf{b}) &:= h \sum_{j=1}^{N} a_j b_j^* \\ \text{so } (\mathbf{a}, \mathbf{a}) &= \|\mathbf{a}\|_2^2 \end{aligned}$$

Since the DF vectors are a basis, we can write any vector as a linear combination of these vectors, that is

$$\mathbf{a} = \hat{a}_0\mathbf{f}_0 + \hat{a}_1\mathbf{f}_1 + \ldots + \hat{a}_{N-1}\mathbf{f}_{N-1}$$

for ! coefficients $\hat{\mathbf{a}}$, the scaled DF transform of $\mathbf{a}$. We are pursuing some theoretical properties here, but there are practical situations where $\hat{\mathbf{a}}$ is desired and it can be computed efficiently using the Fast Fourier Transform algorithm. Since $\{\mathbf{f}_\alpha\}$ is an orthonormal basis,

$$\|\mathbf{a}\|_2 = \|\hat{\mathbf{a}}\|_2. \tag{2.12}$$

2.7.2 Application to discretizations

It can be shown that the DF vectors are always the complete set of eigenvectors of any linear, constant coefficient, periodic, finite difference discretization on a uniform grid. As an example, this will be shown explicitly for our discretization of Problem A (2.1).

$$\begin{aligned} D_2 f_{\alpha,j} &= \frac{1}{h^2}(f_{\alpha,j-1} - 2f_{\alpha,j} + f_{\alpha,j+1}) \\ &= \frac{1}{h^2}e^{2\pi i\alpha jh}\left(e^{-2\pi i\alpha h} - 2 + e^{2\pi i\alpha h}\right) \\ &= \frac{2}{h^2}(\cos(2\pi\alpha h) - 1)f_{\alpha,j}. \end{aligned}$$

Thus

$$D_2\mathbf{f}_\alpha = \frac{2}{h^2}(\cos(2\pi\alpha h) - 1)\mathbf{f}_\alpha$$

and

$$\mathcal{A}\mathbf{f}_\alpha = \lambda_\alpha\mathbf{f}_\alpha$$

with $\lambda_\alpha = \frac{2}{h^2}(1 - \cos(2\pi\alpha h)) + 1$. The set of the eigenvalues $\{\lambda_\alpha\}$ of $\mathcal{A}$ corresponding to the DF vectors is called the *symbol* of $\mathcal{A}$. Consider again our discretization:

$$\mathcal{A}\mathbf{U} = \mathbf{F}.$$

We can write it in terms of the the DF components

$$\hat{U}_\alpha = \frac{1}{\lambda_\alpha}\hat{F}_\alpha$$

where we have gained considerable insight from diagonalizing the problem. The eigenvalues λ_α are all positive and ≥ 1 so

$$\begin{aligned} & |\hat{U}_\alpha| \leq |\hat{F}_\alpha| \quad \text{for every } \alpha \\ \Rightarrow \quad & \|\hat{\mathbf{U}}\|_2 \leq \|\hat{\mathbf{F}}\|_2 \\ \Rightarrow \quad & \|\mathbf{U}\|_{\mathbf{2}} \leq \|\mathbf{F}\|_{\mathbf{2}} \end{aligned}$$

using (2.12). This shows the l_2 norm stability of the scheme. Applying the same result to $\mathcal{A}\mathbf{E} = \tau$ shows the l_2 convergence of the scheme,

$$\|\mathbf{E}\|_2 \leq Ch^2$$

using the bounds on the truncation error τ from the previous section. Note that a (sub-optimal) maximum norm convergence result can be derived from this, since

$$\begin{aligned} & h|E_i|^2 \leq h\sum_{j=1}^{N}|E_j|^2 := \|\mathbf{E}\|_2^2 \leq (Ch^2)^2 \text{ for each } i \\ \Rightarrow \quad & |E_i|^2 \leq C^2h^3 \text{ for each } i \\ \Rightarrow \quad & \|\mathbf{E}\|_\infty \leq Ch^{3/2} \end{aligned}$$

So from von Neumann analysis we can show that scheme does converge in maximum norm, but at the non-optimal rate of 3/2. We know from our numerical test that second order accuracy is maximum norm is observed and confirmed that in our first, maximum norm, stability analysis.

Remark: Sometimes there is a gap between what can be proved and the actual behaviour of the method.

2.8 Implementing Boundary Conditions

Consider Problem A (2.1) in the interval [0,1], but with local boundary conditions specified at the ends $x = 0$ and $x = 1$ rather than periodic conditions. Possible conditions for this problem at $x = 0$ are

$$u(0) = a, \quad a \text{ given (Dirichlet)} \tag{2.13}$$

$$u'(0) = a \quad \text{(Neumann)} \tag{2.14}$$

$$u'(0) - \alpha u(0) = a, \quad \alpha > 0 \text{ given (Robin).} \tag{2.15}$$

Similar conditions can be given at $x = 1$. For (2.15) at $x = 1$, $\alpha < 0$ for physically stable models. In this setting with a uniform grid with spacing $h = 1/N$ we will in general have $N + 1$ discrete unknowns $U_0, U_1, \ldots U_N$ at $x = 0, h, \ldots 1$. If we have Dirichlet conditions at both interval ends, we can apply $U_0 = a$ directly, and the same for U_N. The value of U_0 only appears in the stencil at grid point 1:

$$D_2U_1 = \frac{U_0 - 2U_1 + U_2}{h^2} = \frac{-2U_1 + U_2}{h^2} + \frac{a}{h^2}.$$

In the implementation, the first two terms of the right expression above become part of the matrix $\mathcal{A}$ and the last term contributes to the right hand side vector. In this case, the end point values have been eliminated and the system to be solved is of size $N - 1$.

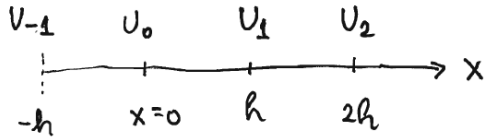

Figure 2.7: Grid points near the $x = 0$ boundary.

First derivative approximations

To proceed to the other conditions (Neumann and Robin) we need to discuss finite difference approximations of the first derivative.

$$D_+U_j := \frac{U_{j+1} - U_j}{h} = u'(jh) + \frac{h}{2}u''(jh) + \dots \text{ (forward differencing, first order)}$$

$$D_-U_j := \frac{U_j - U_{j-1}}{h} = u'(jh) - \frac{h}{2}u''(jh) + \dots \text{ (backward differencing, first order)}$$

$$D_1U_j := \frac{U_{j+1} - U_{j-1}}{2h} = u'(jh) + \frac{h^2}{6}u'''(jh) + \dots \text{ (centred differencing, second order)}$$

$$\tilde{D}_+U_j := \frac{-\frac{3}{2}U_j + 2U_{j+1} - \frac{1}{2}U_{j+2}}{h} = u'(jh) + \frac{h^2}{3}u'''(jh) + \dots \text{ (second order forward differencing)}$$

Notice that although both D_1 and $\tilde{D}_+$ are second order accurate, $\tilde{D}_+$ has a larger error constant. Note also that $D_1 = \frac{1}{2}(D_+ + D_-)$ and that this combination cancels the first order error terms.

2.8.1 Implementing Neumann conditions

Consider now implementing the Neumann condition (2.14). There are several approaches. Since implementing boundary conditions is often a source of confusion, I will go through some of the options in detail. In what follows, refer to Figure 2.7 for the numbering of the unknowns.

U_0 **equation for Neumann condition:** In this scenario, U_0 remains an unknown and we use an approximation of the Neumann condition for its corresponding discrete equation. Using first order differencing

$$D_+U_0 := \frac{U_1 - U_0}{h} = a$$

leads to approximate values that are only first order accurate at *all* grid points. We can easily maintain global second order accuracy by using

$$\tilde{D}_+U_0 = a \tag{2.16}$$

instead.

U_0 eliminated using one sided differencing: We notice again that U_0 only appears in the U_1 equation and so we can use (2.16) to eliminate it:

$$\begin{aligned} D_2U_1 &:= \frac{U_0 - 2U_1 + U_2}{h^2} \\ &= \frac{\frac{2}{3}(2U_1 - \frac{1}{2}U_2 - ha) - 2U_1 + U_2}{h^2} \\ &= \frac{2(U_2 - U_1)}{3h^2} - \frac{2a}{3h} \end{aligned} \tag{2.17}$$

However, (2.17) can cause some confusion since the truncation error in (2.16) is second order but (2.17) is only first order accurate. Analytically, (2.17) is the right way to view the system since the corresponding $(N+1) \times (N+1)$ matrices are max-norm stable.

Introduce the ghost point U_{-1}: You can also introduce the ghost point U_{-1} as shown in Figure 2.7. Consider U_{-1} to approximate the solution extended from the interior to $x = -h$ using a Taylor polynomial of high enough order. The Neumann condition $u'(0) = a$ can then be approximated by

$$D_1U := \frac{U_1 - U_{-1}}{2h} = a.$$

This equation and U_{-1} can be added to the system or U_{-1} can be eliminated using this condition as done above. This approximation has a smaller truncation error constant than the second order one-sided approach and so is preferred.

Robin conditions can be implemented in a similar manner.

2.8.2 Implementing boundary conditions for staggered grid discretizations

We can consider approximate values at subinterval centres rather than subinterval ends. For second order methods, this is equivalent to considering unknowns that are the integral average of the unknown function over the subinterval (the basis of finite volume methods). Values on this grid near the $x = 0$ boundary are shown in Figure 2.8. Here, a ghost value is needed even for a Dirichlet condition, which is then approximated to second order using linear interpolation (averaging):

$$\frac{U_{1/2} + U_{-1/2}}{2} = a \quad \text{approximates } u(0) = a$$

and short centred differencing is used for a Neumann condition

$$\frac{U_{1/2} - U_{-1/2}}{h} = a \quad \text{approximates } u'(0) = a.$$

Note that this approximation has a dominant truncation error term of $\frac{h^2}{24}u'''(0)$. This is the most accurate second order way to approximate Neumann conditions

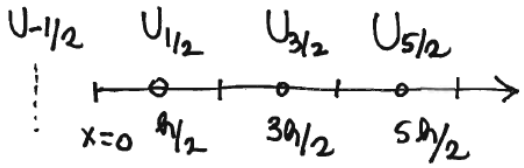

Figure 2.8: Staggered grid points near the $x = 0$ boundary.

in this finite difference framework. As discussed above, these equations can be incorporated into a matrix of the discretization $\mathcal{A}$, or the ghost value $U_{-1/2}$ can be eliminated.

2.9 Asymptotic Error Analysis

Our convergence proof for the discretization of the periodic problem showed that

$$\|\mathbf{E}\|_\infty \leq Ch^2$$

where $\mathbf{E} = \mathbf{U} - \mathbf{u}$. Computationally, we saw that in fact

$$E_i = c(ih)h^2 + o(h^2)$$

with an underlying smooth function $c(x)$ independent of h. Above, $o(h^2)$ (notice the lower case o) is a quantity smaller than any constant times h^2 as $h \to 0$. For smooth data f this result is not hard to show and the remainder term $o(h^2)$ is $O(h^4)$. Let us see what $c(x)$ would have to be to have the property

$$U_i = u(ih) + c(ih)h^2 + \ldots \tag{2.18}$$

This is known as an asymptotic error expansion. Plug this into the discrete equations $-D_2U_i + aU_i = F_i$ and use the remainder terms for the approximation for D_2 we derived in Section 2.4:

$$-u'' - \frac{h^2}{12}u'''' + O(h^4) - h^2c'' + au + ah^2c = f \tag{2.19}$$

where these terms are all evaluated at $x = ih$. If (2.19) is to hold for all h, then the coefficients of powers of h must match:

$$O(1): \qquad -u'' + au = f \quad \text{with } u \text{ periodic} \tag{2.20}$$

$$O(h^2): \qquad -c'' + cu = \frac{1}{12}u'''' \quad \text{with } c \text{ periodic.} \tag{2.21}$$

Equation (2.20) is satisfied by the exact solution. We don't have to actually solve (2.21) but we do know that this problem has a smooth solution $c(x)$. Note

that $c(x)h^2$ is the dominant error term (2.18) and from (2.21) we see that $c(x)$ is the response of the system to a RHS that is the truncation error. This makes sense.

All of this may seem like formal so far, but now consider

$$\tilde{\mathbf{E}} = \mathbf{U} - (\mathbf{u} + h^2\mathbf{c}).$$

We have

$$\mathcal{A}\tilde{\mathbf{E}} = O(h^4)$$

using (2.20) and (2.21). Using the stability result for $\mathcal{A}$ from Section 2.6 we have

$$\tilde{\mathbf{E}} = O(h^4).$$

This shows that (2.18) is accurate to fourth order, the desired result.

There are several consequences of the result

$$\mathbf{U} = \mathbf{u} + h^2\mathbf{c} + O(h^4)$$

with $c(x)$ smooth, independent of h:

Richardson Extrapolation: This justifies Richardson extrapolation

$$\tilde{\mathbf{U}}_h := \frac{4}{3}\mathbf{U}_{h/2} - \frac{1}{3}\mathbf{U}_h = \mathbf{u} + O(h^4)$$

where in the expression above $\mathbf{U}_h$ is a coarse grid computation and $\mathbf{U}_{h/2}$ are values from a fine grid (refined by a factor of 2) computation taken at coarse grid points. Note that this is not an efficient way to make a fourth order method. In addition, it is not reliable since not all schemes have regular errors like this one.

Discrete Smoothness: This result also shows that derivative approximations converge with full order. Consider our basic estimate

$$\mathbf{U} = \mathbf{u} + O(h^2).$$

If we were interested in values of the first derivative of the solution, we would compute

$$D_1\mathbf{U} = D_1\mathbf{u} + O(h) = \mathbf{u}' + O(h)$$

where the order of h is lost because we divide the $O(h^2)$ by h when we apply D_1 and we have not taken into account the structure of the $O(h^2)$ term. However, if we know the asymptotic error result applies

$$\mathbf{U} = \mathbf{u} + h^2\mathbf{c} + O(h^4)$$

with $c(x)$ smooth then we see that

$$D_1\mathbf{U} = D_1\mathbf{u} + h^2 D_1\mathbf{c} + O(h^3) = \mathbf{u}' + h^2(\mathbf{c}' + \frac{1}{6}\mathbf{u}''') + O(h^3) = \mathbf{u}' + O(h^2)$$

and we see convergence order preserved for derivatives. In Finite Element Method (FEM) literature, this "unexpected" increase in convergence order is sometimes called "superconvergence" (although this term also has other meanings).

Discrete Embedding: An asymptotic error expansion can also overcome deficiencies in the stability analysis using weak norms. Consider the conversion of the l_2 estimate to the maximum norm estimate considered in Section 2.7. Following that argument with

$$\tilde{\mathbf{E}} := \mathbf{U} - (\mathbf{u} + h^2\mathbf{c}) = O(h^4)$$

gives

$$\begin{aligned} \|\tilde{\mathbf{E}}\|_\infty &\leq Ch^{7/2} \\ \|\mathbf{E} - h^2\mathbf{c}\|_\infty &\leq Ch^{7/2} \\ \|\mathbf{E}\|_\infty &= O(h^2) \end{aligned}$$

where in the last step we have used the triangle inequality.

The existence of regular asymptotic error behaviour (and so all the results above) relies on the problem having smooth solutions and being computed on a regular (structured) grid. This shows one of the advantages of using structured meshes. In addition, discretizations on structured meshes are more efficiently implemented, especially on specific computational architectures like Graphical Processing Units (GPUs). However, structured meshes restrict local adaptivity and do not apply in a straightforward way to general problem geometries in higher dimensions.

2.10 Lecture #2 Problems

Problem 1 *Derive a fourth order accurate approximation of the second derivative at a grid point x_i using values $F_{i+2}, F_{i+1}, F_i, F_{i-1}, F_{i-2}$.*

Problem 2 *Apply the approximation of the previous question to Problem A. Use the test problem to verify that the method converges with fourth order accuracy. Show the l_2-stability (energy stability) of the method using von Neumann analysis.*

Problem 3 *Consider the wide approximation above applied to Problem A with homogeneous Dirichlet conditions. An additional numerical boundary condition must be given at each end (for U_{-1} and U_{N+1}) because of the wide stencil. Experiment numerically with different conditions until you find some that give fourth order convergence. As an optional question, show the asymptotic error analysis of your scheme.*

Problem 4 *Suppose the $N \times N$ matrix $\mathcal{A}$ is strictly diagonally dominant,* i.e. *all rows satisfy*

$$|A_{ii}| > \sum_{j \neq i} |A_{ij}|.$$

(a) *Show that $\mathcal{A}$ is invertible.*

(b) *If $\mathcal{A}$ is real and symmetric (so all the eigenvalues are real) and $A_{ii} > 0$ for all i show that all eigenvalues of $\mathbf{A}$ are positive.*

(c) *Show that Gaussian Elimination can be performed on $\mathcal{A}$ without pivoting.*

Problem 5 *Consider the boundary value problem for $u(x)$, $x \in [0, 1]$:*

$$-(a(x)u')' + c(x)u = f \quad \text{with } u(0) = 0 \text{ and } u(1) = 0.$$

with $a(x) > 0$ and $c(x) > 0$ and smooth for all x.

(a) *Show that this problem has unique solutions.*

(b) *Write down the conditions that the Green's function for this problem must satisfy.*

(c) *Show that the Green's function with the properties above exists.*