Error in FD methods for BVPs

We will continue our exploration of the 2^{nd} order finite difference method for solving the 1D Poisson equation. Our focus will be on defining and characterizing the *convergence* behavior of our solution method.

Again, remember that the analysis done here extends straightforwardly to the more general BVP

$$\alpha \frac{d^2u}{dx^2} + \beta \frac{du}{dx} + \gamma u = f, \quad x \in [a, b] \quad (1)$$

1 *Convergence of the 2^{nd} order finite difference method*

Motivated by the cliffhanger ending of the last lecture, we continue our question of whether we can characterize how $\hat{\boldsymbol{u}}$ converges to $\bar{\boldsymbol{u}}$. Before proceeding with this investigation, let us finally add some precision to this notion of *convergence*. By asking about the convergence behavior of our finite difference solution, we are asking the following mathematical question:

$$\lim_{n \to \infty} ||\bar{\boldsymbol{u}} - \hat{\boldsymbol{u}}||_2 \quad (2)$$

In words, when we consider convergence we are asking the question of how the numerical solution approaches (or does not approach) the exact solution as the number of discretization points grows infinitely large (we could equivalently consider the limit as $\Delta x \to 0$).

The aim of this section will be to characterize the convergence behavior of our 2^{nd}-order finite difference method. To facilitate this, let us define the *error vector* $\boldsymbol{e} := \bar{\boldsymbol{u}} - \hat{\boldsymbol{u}}$, and write the matrix system from the last lecture more succinctly as $\boldsymbol{A}\hat{\boldsymbol{u}} = \boldsymbol{f}$. Now note that

$$\begin{aligned} \boldsymbol{A}\boldsymbol{e} &= \boldsymbol{A}(\bar{\boldsymbol{u}} - \hat{\boldsymbol{u}}) \\ &= \boldsymbol{A}\bar{\boldsymbol{u}} - \boldsymbol{f} \end{aligned} \quad (3)$$

We used the fact that $\boldsymbol{A}\hat{\boldsymbol{u}} = \boldsymbol{f}$.

We can use (3) to arrive at the desired expression for $||\boldsymbol{e}||_2 = ||\bar{\boldsymbol{u}} - \hat{\boldsymbol{u}}||_2$. Pre-multiplying (3) by $\boldsymbol{A}^{-1}$ and taking the grid function 2-norm of the result gives

Recall that we defined the grid function 2-norm in the last lecture.

$$\begin{aligned} ||\boldsymbol{e}||_2 &= \left|\left| \boldsymbol{A}^{-1}\left[\boldsymbol{A}\bar{\boldsymbol{u}} - \boldsymbol{f}\right] \right|\right|_2 \\ &\leq \left|\left| \boldsymbol{A}^{-1} \right|\right|_2 ||\boldsymbol{A}\bar{\boldsymbol{u}} - \boldsymbol{f}||_2 \end{aligned} \quad (4)$$

The inequality arises from the definition of the induced matrix 2-norm. We do not have time to discuss the details of this beautiful quantity, but you are encouraged to explore the matrix norm on Wikipedia if you are interested!

The beautiful error bound (4) demonstrates that there are two sources that contribute to the error:

a) The error associated with using the *exact* solution at the interpolation points, $\bar{\boldsymbol{u}}$, in the finite difference method. This source of error is called the *truncation error*.

b) The bound associated with the 2-norm of the matrix $\boldsymbol{A}^{-1}$.

We will characterize each of the two sources of error in turn.

Take some time to convince yourself that the error associated with b) is indeed nonzero. Why is this the case? The reason is that the finite difference operator $\boldsymbol{A}$ approximates—but is not equal to—d^2/dx^2. If $\boldsymbol{A}$ exactly represented the second derivative operator, then there would be zero error in using $\bar{\boldsymbol{u}}$ in b).

The truncation error

Before analyzing the truncation error, let us stop and make sure we understand what we need from this source of error to arrive at a convergent finite difference method. Let us assume for now that $||\boldsymbol{A}^{-1}||_2 < c$ (we will show that this is true later in this lecture). Thus, by virtue of (4), if we want the size of the error $||\boldsymbol{e}||_2$ to vanish as $\Delta x \to 0$, we *must* have that the truncation error goes to zero as $\Delta x \to 0$. We will show that this is indeed the case for our finite difference method.

Recall from our observation a) above that the truncation error arises from applying the finite difference approximation to the exact solution evaluated at the interpolation points. Defining by $\boldsymbol{\tau}$ the truncation error vector, $\boldsymbol{\tau} = \boldsymbol{A}\bar{\boldsymbol{u}} - \boldsymbol{f}$, we have that the j^{th} entry of $\boldsymbol{\tau}$ is

$$\tau_j = \frac{1}{\Delta x^2}\left(u(x_{j-1}) - 2u(x_j) + u(x_{j+1})\right) - f(x_j) \tag{5}$$

To arrive at a meaningful variant of this expression, we will Taylor expand $u(x_{j-1})$ and $u(x_{j+1})$ about x_j. Doing this yields an approximation for τ_j that is valid in the limit of small Δx:

For example,

$$u(x_{j+1}) = u(x_j) + \Delta x \frac{du}{dx} + \frac{\Delta x^2}{2}\frac{d^2u}{dx^2} + \frac{\Delta x^3}{6}\frac{d^3u}{dx^3} + O(\Delta x^4)$$

$$\tau_j = \frac{\Delta x^2}{12}\frac{d^4u}{dx^4} + O(\Delta x^4) \tag{6}$$

This analysis demonstrates that the truncation error indeed vanishes to zero as Δx vanishes. But we can say more: the truncation error decays at a *rate* of Δx^2 as $\Delta x \to 0$.

The importance of the qualities that i) the truncation error vanishes and ii) the rate at which it vanishes motivate the following definition:

Definition: consistency

A finite difference method that is written in the form $\boldsymbol{Au} = \boldsymbol{f}$ is called *consistent* if it has a truncation error that goes to zero as $\Delta x \to 0$. If, moreover, the truncation error vanishes at a rate of Δx^r, the method is said to be consistent with order r.

The error contribution from $\boldsymbol{A}^{-1}$

Our aim will be to demonstrate that $||\boldsymbol{A}^{-1}||_2$ remains bounded in the limit as $\Delta x \to 0$. If this is true, then $||\boldsymbol{A}^{-1}||_2 < c$ for some constant c *irrespective* of the value of Δx. This fact will gave us a useful way to bound (4) from above.

How do we demonstrate this nonsingular nature of $\boldsymbol{A}$? Our approach will be to construct the eigendecomposition of $\boldsymbol{A}$, and use this to arrive at a bound for the norm of its inverse. That is, we seek a set

of matrices $\boldsymbol{V}$ and $\boldsymbol{\Lambda}$ such that

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1} \tag{7}$$

Let us pull a rabbit out of a hat: the i^{th} element of the j^{th} eigenvector of $\boldsymbol{A}$, $\boldsymbol{v}_j$, is given by

$$\begin{aligned} (v_j)_i &= \sin(j\pi(i\Delta x)), \\ j &= 1, \ldots, n-2; \;\; i = 1, \ldots, n-2 \end{aligned} \tag{8}$$

This is not as out of nowhere as it may first appear. Recall that $v_j(x) = \sin(j\pi(x-a)/(b-a))$ is the eigenfunction associated with the operator d^2/dx^2. That is, we can think of d^2/dx^2 as an infinite-dimensional matrix. Defining $L := d^2/dx^2$, notice that $Lv_j = -(j\pi/(b-a))^2 v_j = \lambda_j v_j$. So the eigenvectors of $\boldsymbol{A}$ mimic the eigenvectors of the operator it approximates!

Let us verify that these are indeed eigenvectors of $\boldsymbol{A}$ by direct computation. The k^{th} entry of the vector $\boldsymbol{A}\boldsymbol{v}_j$ is

$$\begin{aligned} (Av_j)_k &= \frac{1}{\Delta x^2}\left[(v_j)_{k-1} - 2(v_j)_k + (v_j)_{k+1}\right] \\ &= \frac{1}{\Delta x^2}\left[\sin(j\pi(k-1)\Delta x) - 2\sin(j\pi k\Delta x) + \sin(j\pi(k+1)\Delta x)\right] \\ &= \frac{1}{\Delta x^2}\left[\sin(j\pi k\Delta x)\cos(j\pi\Delta x) - 2\sin(j\pi k\Delta x) + \right. \\ &\qquad \left. \sin(j\pi k\Delta x)\cos(j\pi\Delta x)\right] \\ &= \frac{2(\cos(j\pi\Delta x) - 1)}{\Delta x^2}\sin(j\pi k\Delta x) \\ &= \lambda_j (v_j)_k \end{aligned} \tag{9}$$

In going from the second to third line, we wrote $(k-1)\Delta x$ and $(k+1)\Delta x$ as $k\Delta x - \Delta x$ and $k\Delta x + \Delta x$, respectively, and used standard trigonometric identities.

So $\boldsymbol{v}_j$ is indeed an eigenvector of $\boldsymbol{A}$, and it has an associated eigenvalue

$$\lambda_j = (2/\Delta x^2)(\cos(j\pi\Delta x) - 1) \tag{10}$$

The eigenvectors enjoy an additional property:

$$\boldsymbol{v}_j^T \boldsymbol{v}_i = \begin{cases} 0 & i \neq j \\ n-1 & i = j \end{cases} \tag{11}$$

This orthogonality property is also a feature of the eigenfunctions of $L = d^2/dx^2$. One can show that $\int_a^b \sin(j\pi(x-a)/(b-a))\sin(i\pi(x-a)/(b-a))dx = 0$ if $i \neq j$.

Thus, the expression for $\boldsymbol{V}^{-1}$ is straightforward: $\boldsymbol{V}^{-1} = 1/(n-1)\boldsymbol{V}^T$. In fact, we could have anticipated this property of the eigenvectors of $\boldsymbol{A}$. For any real symmetric matrix, its eigendecomposition leads to an eigenvector matrix that is *unitary*, *i.e.*, having an inverse equal to its transpose (to within a constant scaling, which is not particularly important because eigenvectors are only unique up to a constant anyway).

We now have what we need to bound $||\boldsymbol{A}^{-1}||_2$. First, observe that

$$\begin{aligned} \boldsymbol{A}^{-1} &= \left(\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T\right)^{-1} \\ &= \left(\boldsymbol{V}^T\right)^{-1}\boldsymbol{\Lambda}^{-1}\boldsymbol{V}^{-1} \\ &= \boldsymbol{V}\boldsymbol{\Lambda}^{-1}\boldsymbol{V}^T \end{aligned} \tag{12}$$

and since $\boldsymbol{\Lambda}$ is a diagonal matrix, its inverse is also diagonal with entries $(\Lambda^{-1})_{ii} = \lambda_i^{-1}$. Thus, like $\boldsymbol{A}$, $\boldsymbol{A}^{-1}$ also possesses unitary eigenvectors. Its norm is therefore given by

$$\left|\left|\boldsymbol{A}^{-1}\right|\right|_2 = \left|\left|\boldsymbol{\Lambda}^{-1}\right|\right|_2 = \max_{i=1,\ldots,n} |\lambda_j^{-1}| \tag{13}$$

That is, the norm of $\boldsymbol{A}^{-1}$ is given by the eigenvalue of $\boldsymbol{A}$ of smallest modulus. From the definition of the eigenvalues of $\boldsymbol{A}$ provided in (10), this smallest eigenvalue is $\lambda_1 = (2/\Delta x^2)(\cos(\pi\Delta x) - 1)$. Is this bounded as $\Delta x \to 0$? We may consider a Taylor series about $\Delta x = 0$:

$$\begin{aligned}\lambda_1 &= \left(\frac{2}{\Delta x^2}\right)(\cos(\pi\Delta x) - 1) \\ &= \left(\frac{2}{\Delta x^2}\right)\left(-\frac{1}{2}\pi^2\Delta x^2 + \frac{1}{24}\pi^4\Delta x^4 + O\left(\Delta x^6\right)\right) \\ &\approx -\pi^2\end{aligned} \tag{14}$$

This expansion is valid for sufficiently small Δx.

So we can indeed say that $||\boldsymbol{A}^{-1}||_2$ is bounded by a constant (which must be larger than $\sim 1/\pi^2$). This fact is crucial to establishing the convergence of our finite difference method. If $||\boldsymbol{A}^{-1}||_2$ were not bounded, our finite difference method would not converge. Because of its importance, we assign a definition related to the boundedness of $\boldsymbol{A}^{-1}$:

Definition: stability

A finite difference method that is written in the form $\boldsymbol{Au} = \boldsymbol{f}$ is called *stable* if $||\boldsymbol{A}^{-1}||_2 < c$ for some constant c that is independent of Δx. Stability is a *necessary* condition for the finite difference method to converge.

We are defining stability in terms of finite difference methods, but there are analogous definitions for other numerical methods.

One important note before we move onto the truncation error of our finite difference method: the eigenvectors $\boldsymbol{v}_j$ and eigenvalues λ_j only belong to the matrix $\boldsymbol{A}$ associated with the truncated linear system from the last lecture. The fact that this truncated matrix yields such a cleanly expressible set of eigenvectors and eigenvalues makes it invaluable for studying the convergence of our second order finite difference method.

Convergence

Now that we have the definitions of stability and consistency, we can recast our definition of convergence in terms of these important concepts:

Definition: convergence (take 2)

A finite difference method that is written in the form $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f}$ is called *convergent* if it is *stable* and *consistent*. If, moreover, the method is consistent with order r, it is called an r^{th} order method (this is because the error decays at the same rate as the truncation error by virtue of (4)).

Why have we spent so much time defining these ideas of stability and convergence? The answer is because these concepts apply to general finite difference methods, and are not restricted to the 2^{nd} order method we derived in Lecture 3. Convince yourself of this by going back through this derivation and verifying that we did not make any assumptions except that the method could be cast as $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f}$!

2 *A practical convergence test*

We now have analytical arguments that tell us that our method converges and, beyond this, that it is a second order method. Do we see this in practice?

We already know from the last lecture that method indeed converges as $\Delta x \to 0$. But does it converge at the expected rate. We reproduce the convergence plot from the last lecture below, now superposed with an additional curve of Δx^2. Indeed, our method is of order 2.

Brain teaser: why is the rate of error decay different for the largest two values of Δx?

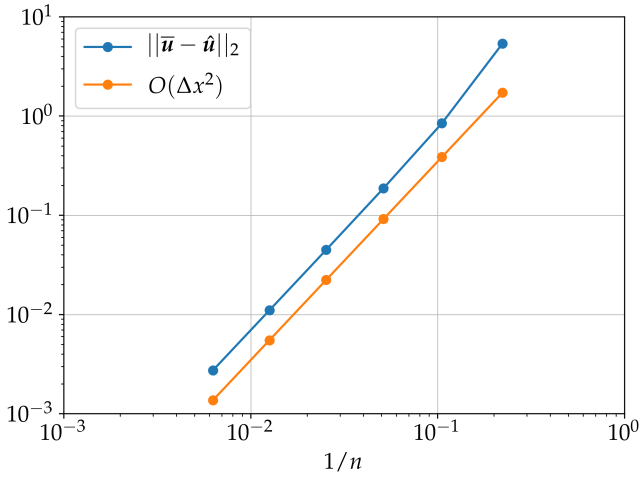

Figure 1: Error in the numerical solution of the 1D Poisson equation as a function of $1/n$, along with the expected Δx^2 scaling.