8 Finite-difference methods for BVPs

In this Lecture, we consider methods for solving BVPs whose (methods') idea is to replace the BVPs with a system of approximating algebraic equations. In the first five sections, we will (mostly) deal with linear BVPs, so that the corresponding system of equations will be linear. The last section will show how nonlinear BVPs can be approached.

8.1 Matrix problem for the discretized solution

Let us begin by considering a linear BVP with Dirichlet boundary conditions:

$$\begin{aligned} &y'' + P(x)y' + Q(x)y = R(x)\,, \\ &y(a) = \alpha, \quad y(b) = \beta\,. \end{aligned} \tag{8.1}$$

As before, we assume that P, Q, and R are twice continuously differentiable, so that y is a four times continuously differentiable function of x. Also, we consider the case where $Q(x) \le 0$ on $[a,\, b]$, so that the BVP (8.1) is guaranteed by Theorem 6.2 to have a unique solution.

Let us replace y'' and y' in (8.1) by their second-order accurate discretizations:

$$y'' \;=\; \frac{1}{h^2}(y_{n+1} - 2y_n + y_{n-1}) + O(h^2)\,, \tag{8.2}$$

$$y' \;=\; \frac{1}{2h}(y_{n+1} - y_{n-1}) + O(h^2)\,. \tag{8.3}$$

Upon omitting the $O(h^2)$-terms and substituting (8.2) and (8.3) into (8.1), we obtain the following system of linear equations:

$$\begin{aligned} &Y_0 = \alpha\,; \\ &(1 + \tfrac{h}{2}P_n)Y_{n+1} - (2 - h^2 Q_n)Y_n + (1 - \tfrac{h}{2}P_n)Y_{n-1} = h^2 R_n\,, \qquad 1 \le n \le N-1\,; \\ &Y_N = \beta\,. \end{aligned} \tag{8.4}$$

In the matrix form, this is

$$A\vec{\mathbf{Y}} \;=\; \vec{\mathbf{r}}\,, \tag{8.5}$$

where

$$A = \begin{pmatrix} -(2-h^2Q_1) & (1+\frac{h}{2}P_1) & 0 & 0 & \cdots & 0 \\ (1-\frac{h}{2}P_2) & -(2-h^2Q_2) & (1+\frac{h}{2}P_2) & 0 & \cdots & 0 \\ 0 & (1-\frac{h}{2}P_3) & -(2-h^2Q_3) & (1+\frac{h}{2}P_3) & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & (1-\frac{h}{2}P_{N-2}) & -(2-h^2Q_{N-2}) & (1+\frac{h}{2}P_{N-2}) \\ 0 & \cdots & 0 & 0 & (1-\frac{h}{2}P_{N-1}) & -(2-h^2Q_{N-1}) \end{pmatrix}, \tag{8.6}$$

$\vec{\mathbf{Y}} = [Y_1,\, Y_2,\, \dots\,,\, Y_{N-1}]^T$, and

$$\vec{\mathbf{r}} = \left[h^2R_1 - \left(1 - \frac{h}{2}P_1\right)\alpha, \quad h^2R_2, \quad h^2R_3, \quad \cdots\,, \quad h^2R_{N-2}, \quad h^2R_{N-1} - \left(1 + \frac{h}{2}P_{N-1}\right)\beta\right]^T; \tag{8.7}$$

the superscript 'T' in (8.7) denotes the transpose.

From Linear Algebra, it is known that the linear system (8.5) has a unique solution if (and only if) the matrix A is nonsingular. Therefore, below we list some results that will allow us to guarantee that under certain conditions, a particular A is nonsingular.

Gerschgorin Circles Theorem, 8.1 Let a_{ij} be entries of an $M \times M$ matrix A and let λ_k, $k = 1, \ldots, M$ be the eigenvalues of A. Then:

(i) Each eigenvalue lies in the union of "row circles" R_i, where

$$R_i = \left\{ z : \; |z - a_{ii}| \le \sum_{j=1,\, j \neq i}^{M} |a_{ij}| \right\} . \tag{8.8}$$

(In words, $\sum_{j=1,\, j\neq i}^{M} |a_{ij}|$ is the sum of all off-diagonal entries in the ith <u>row</u>.)

(ii) Similarly, since A and A^T have the same eigenvalues, then each eigenvalue also lies in the union of "column circles" C_j, where

$$C_j = \left\{ z : \; |z - a_{jj}| \le \sum_{i=1,\, i \neq j}^{M} |a_{ij}| \right\} . \tag{8.9}$$

(In words, $\sum_{i=1,\, i\neq j}^{M} |a_{ij}|$ is the sum of all off-diagonal entries in the jth <u>column</u>.)

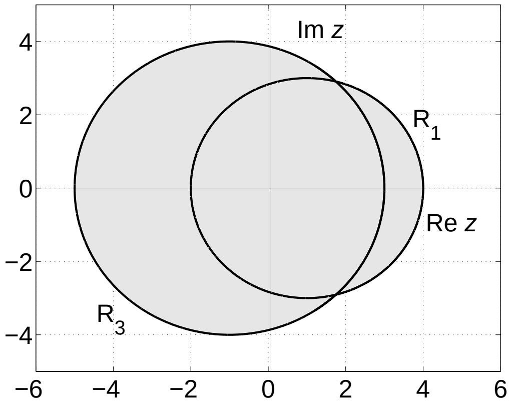
(iii) Let $\bigcup_{i=k}^{i=l} R_i$ be a cluster of $(l - k)$ row circles that is disjoint from all the other row circles. Then it contains exactly $(l - k)$ eigenvalues.
A similar statement holds for column circles.

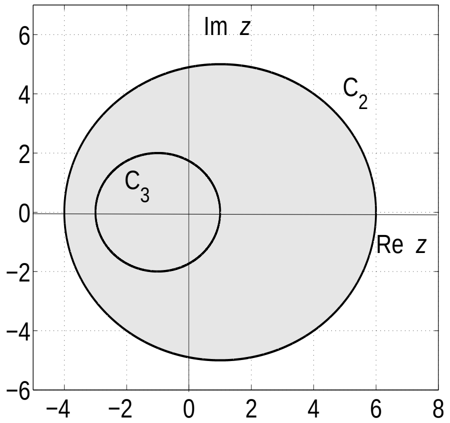
Example Use the Gerschgorin Circles Theorem to estimate eigenvalues of a matrix

$$E = \begin{pmatrix} 1 & 2 & -1 \\ 1 & 1 & 1 \\ 1 & 3 & -1 \end{pmatrix} . \tag{8.10}$$

Solution The circles are listed and sketched below:

$$\begin{aligned} R_1 &= \{z : \; |z - 1| \le |2| + |1| = 3\} & C_1 &= \{z : \; |z - 1| \le |1| + |1| = 2\} \\ R_2 &= \{z : \; |z - 1| \le |1| + |1| = 2\} & C_2 &= \{z : \; |z - 1| \le |2| + |3| = 5\} \\ R_3 &= \{z : \; |z + 1| \le |1| + |3| = 4\} & C_3 &= \{z : \; |z + 1| \le |1| + |-1| = 2\} \end{aligned}$$





(Circles R_2 and C_1 are not sketched because they are concentric with, and lie entirely within, circles R_1 and C_2, respectively.)

Gerschgorin Circles Theorem says that the eigenvalues of E must lie within the intersection of $\bigcup_{i=1}^{3} R_i$ and $\bigcup_{j=1}^{3} C_j$. For example, this gives that $|\mathrm{Re}\,\lambda| \le 4$ for each and any of the eigenvalues.

Before we can apply the Gerschgorin Circles Theorem, we need to introduce some new terminology.

Definition Matrix A is called diagonally dominant if

$$\underline{\text{either}} \quad |a_{ii}| \ge \sum_{j=1,\, j\neq i}^{M} |a_{ij}| \quad \underline{\text{or}} \quad |a_{ii}| \ge \sum_{j=1,\, j\neq i}^{M} |a_{ji}|\,, \qquad 1 \le i \le M\,, \tag{8.11}$$

with the *strict inequality holding for at least one* i.

A matrix is called strictly diagonally dominant (SDD) if

$$\underline{\text{either}} \quad |a_{ii}| > \sum_{j=1,\, j\neq i}^{M} |a_{ij}| \quad \underline{\text{or}} \quad |a_{ii}| > \sum_{j=1,\, j\neq i}^{M} |a_{ji}| \qquad \underline{\text{for all}}\ i = 1, \ldots, M. \tag{8.12}$$

In other words, in a SDD matrix, the sums of the off-diagonal elements along *either* every row *or* every colum are less than the corresponding diagonal entries.

Theorem 8.2 If a matrix A is SDD, then it is nonsingular.

Proof If A is SDD, then one of the inequalities (8.12) holds. Suppose the inequality for the rows holds. Comparing that inequality with the r.h.s. of (8.8), we conclude, by the Gerschgorin Circles Theorem, that point $\lambda = 0$ is outside of the union $\bigcup_{i=1}^{M} R_i$ of Gerschgorin circles. Hence it is automatically outside of the intersection of the unions $\bigcup_{i=1}^{M} R_i$ and $\bigcup_{i=1}^{M} C_i$. Therefore, $\lambda = 0$ is not an eigenvalue of A, hence A is nonsingular. ***q.e.d.***

Theorem 8.3 Consider the BVP (8.1). If $Q(x) \le 0$, and if $P(x)$ is bounded on $[a,\, b]$ (i.e. $|P(x)| \le \mathcal{P}$ for some $\mathcal{P}$), then the discrete version (8.4) of the BVP in question has a unique solution, provided that the step size satisfies $h\mathcal{P} \le 2$.

Proof Case (a): $Q(x) < 0$. In this case, matrix A in (8.6) is SDD, provided that $h\mathcal{P} \le 2$. Indeed, A is tridiagonal, with the diagonal elements being

$$a_{ii} = -(2 - h^2 Q_i), \qquad \Rightarrow \qquad |a_{ii}| > 2\,. \tag{8.13}$$

The sum of the absolute values of the off-diagonal elements is

$$|a_{i,i+1}| + |a_{i,i-1}| \;=\; \left|1 + \frac{h}{2}P_i\right| + \left|1 - \frac{h}{2}P_i\right| \;=\; \left(1 + \frac{h}{2}P_i\right) + \left(1 - \frac{h}{2}P_i\right) \;=\; 2\,. \tag{8.14}$$

In removing the absolute value signs in the above equation, we have used the fact that $h\mathcal{P} \le 2$. Now, comparing (8.13) with (8.14), we see that

$$|a_{i,i+1}| + |a_{i,i-1}| < |a_{ii}|\,,$$

which means that A is SDD and hence by Theorem 8.2, (8.4) has a unique solution. ***q.e.d.***

Case (b): $Q(x) \le 0$ requires a more involved proof, which we omit.

Note: We emphasize that Theorem 8.3 gives a bound for the step size,

$$h \cdot \max_{x \in [a,\,b]} |P(x)| \le 2\,, \tag{8.15}$$

which is a sufficient condition for the discretization (8.4) (with $Q(x) \le 0$) to be solvable.

8.2 Thomas algorithm

In the previous section, we have considered the issue of the *possibility* to obtain the unique solution of the linear system (8.4), which (the system) approximates the BVP (8.1). In this section, we will consider the issue of solving (8.4) *in an efficient manner.* The key fact that will allow us to do so is the tridiagonal form of A; that is, A has nonzero entries only on the main diagonal and on the subdiagonals directly above and below it. Let us recall that in order to solve[19] a linear system of the form (8.5) with a *full* $M \times M$ matrix A, one requires $O(M^3)$ operations. However, this would be an inefficient way to solve a linear system with a tridiagonal A. Below we will present an algorithm, which has been discovered independently by several researchers in the 50's, that allows one to solve a linear system with a tridiagonal matrix using only $O(M)$ operations.

A common practical way to numerically solve a linear system of the form

$$A\vec{\mathbf{y}} \;=\; \vec{\mathbf{r}} \tag{8.5}$$

(with *any* matrix A) is via LU decomposition. Namely, one seeks two matrices L (lower triangular) and U (upper triangular) such that [20]

$$LU = A\,. \tag{8.16}$$

Then (8.5) is solved in two steps:

$$\underline{\text{Step 1}}: \quad L\vec{\mathbf{z}} = \vec{\mathbf{r}} \qquad \text{and} \qquad \underline{\text{Step 2}}: \quad U\vec{\mathbf{y}} = \vec{\mathbf{z}}\,. \tag{8.17}$$

[19] by Gaussian elimination or by any other direct (i.e., non-iterative) method

[20] Disclaimer: The theory of LU decomposition is considerably more involved than the simple excerpt from it given here. We will *not* go into further details of that theory in this course.

The linear systems in (8.17) are then solved using forward (for Step 1) and backward (for Step 2) substitutions; details of this will be provided later on.

When A is tridiagonal, both finding the matrices L and U and solving the systems in (8.17) requires only $O(M)$ operations. You will be asked to demonstrate that in a homework problem. Here we will present details of the algorithm itself.

Let

$$A = \begin{pmatrix} b_1 & c_1 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ a_2 & b_2 & c_2 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & a_3 & b_3 & c_3 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & 0 & a_{M-1} & b_{M-1} & c_{M-1} \\ 0 & \cdot & \cdot & \cdot & 0 & 0 & a_M & b_M \end{pmatrix}; \tag{8.18}$$

then we seek L and U in the form:

$$LU = \begin{pmatrix} 1 & 0 & 0 & \cdot & 0 \\ \alpha_2 & 1 & 0 & \cdot & 0 \\ 0 & \alpha_3 & 1 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & 0 & \alpha_M & 1 \end{pmatrix} \begin{pmatrix} \beta_1 & c_1 & 0 & \cdot & 0 \\ 0 & \beta_2 & c_2 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & 0 & \beta_{M-1} & c_{M-1} \\ 0 & \cdot & 0 & 0 & \beta_M \end{pmatrix} \tag{8.19}$$

Multiplying the matrices in (8.19) and comparing the result with (8.18), we obtain:

$$\begin{aligned} &\text{Row 1:} \quad \beta_1 = b_1; \\ &\text{Row 2:} \quad \alpha_2\beta_1 = a_2, \qquad \alpha_2 c_1 + \beta_2 = b_2; \\ &\text{Row } j\text{:} \quad \alpha_j\beta_{j-1} = a_j, \quad \alpha_j c_{j-1} + \beta_j = b_j\,. \end{aligned} \tag{8.20}$$

Equations (8.20) are easily solved for the unknown coefficients α_j and β_j:

$$\begin{aligned} &\beta_1 = b_1, \\ &\alpha_j = a_j/\beta_{j-1}, \quad \beta_j = b_j - \alpha_j c_{j-1}, \quad j = 2, \ldots, M\,. \end{aligned} \tag{8.21}$$

Finally, we show how the systems in (8.17) can be solved. Let

$$\vec{\mathbf{z}} = [z_1,\ z_2,\ \cdots\ ,\ z_M]^T, \qquad \text{etc.}$$

Then the forward substitution in $L\vec{\mathbf{z}} = \vec{\mathbf{r}}$ gives:

$$\begin{aligned} &z_1 = r_1, \\ &z_j = r_j - \alpha_j z_{j-1}, \quad j = 2, \ldots, M\,. \end{aligned} \tag{8.22}$$

The backward substitution in $U\vec{\mathbf{y}} = \vec{\mathbf{z}}$ gives:

$$\begin{aligned} &y_M = z_M/\beta_M, \\ &y_j = (z_j - c_j y_{j+1})/\beta_j, \quad j = M-1, \ldots, 1\,. \end{aligned} \tag{8.23}$$

Thus, $\vec{\mathbf{y}}$ is found in terms of $\vec{\mathbf{r}}$.

The entire procedure (8.21) to (8.23) is fast, as we said earlier, and also requires the storage of only 8 one-dimensional arrays of size $O(M)$ for the coefficients $\{a_j\}$, $\{b_j\}$, $\{c_j\}$, $\{\alpha_j\}$, $\{\beta_j\}$, $\{r_j\}$, $\{z_j\}$, and $\{y_j\}$. Moreover, it is possible to show that when A is SDD, i.e. when

$$|b_j| > |a_j| + |c_j|, \qquad j = 1, \ldots, M, \tag{8.24}$$

then small round-off (or any other) errors do not get amplified by this algorithm; specifically, the numerical error remains small and independent of the size M of the problem[21].

To conclude this section, we note that similar algorithms exist for other banded (e.g., pentadiagonal) matrices. The details of those algorithms can be found, e.g., in Sec. 3-2 of W. Ames, "Numerical methods for partial differential equations," 3rd ed. (Academic Press, 1992).

8.3 Error estimates, and higher-order discretization

In this section, we will state, without proof, two theorems about the accuracy of the solutions of systems of discretized equations approximating a given BVP.

Theorem 8.4 Let $\{Y_n\}_{n=1}^{N-1}$ be the solution of the discretized problem (8.4) (or, which is the same, (8.5)–(8.7)) and let $y(x)$ be the exact solution of the original BVP (8.1); then $\epsilon_n = y(x_n) - Y_n$ is the error of the numerical solution. In addition, let $\mathcal{P} = \max_{x\in[a,\,b]} |P(x)|$ and also let $Q(x) \le \mathcal{Q} < 0$ (recall that $Q(x) \le 0$ is required for the BVP to have a unique solution). Then the error satisfies the following estimate:

$$\max|\epsilon_n| \;\le\; \frac{1}{h^2\left(|\mathcal{Q}| + \frac{8}{(b-a)^2}\right)}\left(\frac{1}{12}h^4(M_4 + \mathcal{P}M_3) + 2\rho\right), \tag{8.25}$$

where $M_3 = \max_{x\in[a,\,b]} |y'''|$, $M_4 = \max_{x\in[a,\,b]} |y''''|$, and ρ is the round-off error.

When the round-off error is neglected, estimate (8.25) yields that the discrete approximation (8.4) to the BVP produces the error on the order of $O(h^2)$ (i.e., in other words, is *second-order accurate*). At first sight, this is similar to how the finite-difference approximations (8.2) and (8.3) led to second-order accurate methods for IVPs. In fact, for both the IVPs and BVPs, the expression inside the largest parentheses on the r.h.s. of (8.25) is the local truncation error. However, the interpretation of the $O(1/h^2)$ factor in front of those parentheses differs from the interpretation of the similar factor for IVPs. In the latter case, that factor arose from accumulation of the error over $O(1/h)$ steps. *On the contrary, for the BVPs*, that factor is the so-called *condition number* of the matrix A in (8.5), defined as

$$\text{cond}\,A = ||A|| \cdot ||A^{-1}||, \tag{8.26}$$

where $||A||$ is the norm of matrix A, subordinate to a particular vector norm[22]. In courses on Numerical Analysis, it is shown that the condition number is the factor by which a small error in the vector $\vec{\mathbf{r}}$ is amplified in the corresponding error of the solution $\vec{\mathbf{Y}}$:

$$A\,(\vec{\mathbf{Y}} + \delta\vec{\mathbf{Y}}) = \vec{\mathbf{r}} + \delta\vec{\mathbf{r}} \qquad \Rightarrow \qquad \frac{||\delta\vec{\mathbf{Y}}||}{||\vec{\mathbf{Y}}||} = \text{cond}A \cdot \frac{||\delta\vec{\mathbf{r}}||}{||\vec{\mathbf{r}}||}. \tag{8.27}$$

[21] In line with the earlier disclaimer about the *LU* decomposition, we also note that condition (8.24) is just the simplest of possible conditions that guarantee boundedness of the error. Other conditions exist and are studied in advanced courses in Numerical Analysis.

[22] Definition (8.26) is given here only for completemeness of the presentation; we will not need more precise definitions of the norms involved.

One can show (but we will not do so here) that the condition number of the matrix A in (8.5) is $O(\frac{1}{h^2})$; this fact, along with Eq. (8.27) and the local truncation error $\vec{\delta \mathbf{r}}$ being $O(h^4)$, implies estimate (8.25) for the solution error.

The numerical error of the discrete approximation of the BVP (8.1) can be significantly reduced if instead of the simple central-difference approximation to $y'' = f(x, y)$, as in (8.2), one uses the Numerov's formula (5.18). Specifically, if y' does not enter the BVP, i.e. when the BVP is

$$y'' = f(x, y)\,, \qquad y(a) = \alpha, \quad y(b) = \beta, \tag{8.28}$$

then Numerov's formula leads to the following system of discrete equations:

$$\begin{aligned} &Y_0 = \alpha; \\ &Y_{n+1} - 2Y_n + Y_{n-1} = \tfrac{h^2}{12}(f_{n+1} + 10f_n + f_{n-1}) \qquad 1 \le n \le N-1\,; \\ &Y_N = \beta\,. \end{aligned} \tag{8.29}$$

In particular, when $f(x, y) = -Q(x)y + R(x)$, system (8.29) is linear. Since the local truncation error of Numerov's method is $O(h^6)$, the solution of system (8.29) must be a fourth-order accurate approximation to the exact solution of the BVP (8.28). More precisely, the following theorem holds true:

Theorem 8.5 Let $\{Y_n\}_{n=1}^{N-1}$ be the solution of the discretized problem (8.29), where $f(x, y) = -Q(x)y + R(x)$ (so that this system is linear). Then the error $\epsilon_n = y(x_n) - Y_n$ satisfies the following estimate:

$$\max|\epsilon_n| \le \frac{1}{h^2\left(|Q| + \frac{8}{(b-a)^2}\right)}\left(\frac{1}{240}h^6 M_6 + 2\rho\right)\,, \tag{8.30}$$

where the notations are the same as in Theorem 8.4 and, in addition, $M_6 = \max_{x\in[a,\,b]}|y^{(6)}|$.

8.4 Neumann and mixed boundary conditions

Suppose that at, say, $x = a$, the boundary condition is

$$A_1 y(a) + A_2 y'(a) = \alpha\,, \tag{8.31}$$

where A_1 and A_2 are some constants. (Recall that for $A_1 = 0$, this is the Neumann boundary condition, while for $A_1 \cdot A_2 \neq 0$, this condition is of the mixed type.) Below we present two methods that allow one to construct generalizations of systems (8.4) (simple central-difference discretization) and (8.29) (Numerov's formula) for the case of the boundary condition (8.31) instead of

$$y(a) = \alpha\,.$$

Note that the problem we need to handle for such a generalization is obtaining an approximation to $y'(a)$ in (8.31) that would have the accuracy consistent with the accuracy of the discretization scheme. That is, the accuracy needs to be second-order for a generalization of (8.4) and fourth-order for a generalization of (8.29).

Method 1
This method is efficient only for a generalization of the second-order accurate approximation. Thus, we will look for a second-order accurate finite-difference approximation to $y'(a)$.

Introduce a fictitious point $x_{-1} = a - h$ and let the approximate solution at that point be Y_{-1}. Once one knows Y_{-1}, the approximation sought is

$$y'(a) = \frac{Y_1 - Y_{-1}}{2h} + O(h^2)\,, \tag{8.32}$$

so that the equation approximating the *boundary condition* (8.31) is

$$\underline{x = a}: \qquad A_1 Y_0 + A_2 \frac{Y_1 - Y_{-1}}{2h} = \alpha\,. \tag{8.33}$$

The discrete analog of the ODE itself at $x = a$ is:

$$\left(1 + \frac{h}{2}P_0\right) Y_1 - (2 - h^2 Q_0) Y_0 + \left(1 - \frac{h}{2}P_0\right) Y_{-1} = h^2 R_0\,. \tag{8.34}$$

Equations (8.33) and (8.34) should replace the first equation,

$$Y_0 = \alpha, \tag{8.35}$$

in (8.4), so that the dimension of the resulting system of equations in (8.4) increases from $(N-1)\times(N-1)$ to $(N+1)\times(N+1)$. Note that adding two new equations (8.33) and (8.34) to system (8.4) is consistent with introducing two new unknowns Y_{-1} and Y_0.

Later on we will see that, rather than dealing with two equations (8.33) and (8.34), it is more practical to solve (8.33) for Y_{-1} and substitute the result in (8.34). Then, instead of two equations (8.33) and (8.34), we will have one equation that needs to replace (8.35):

$$2Y_1 - \left[2 - h^2 Q_0 - 2h\frac{A_1}{A_2}\left(1 - \frac{h}{2}P_0\right)\right] Y_0 = h^2 R_0 + 2h\frac{\alpha}{A_2}\left(1 - \frac{h}{2}P_0\right)\,. \tag{8.36}$$

(We are justified to assume in (8.36) that $A_2 \neq 0$, since otherwise the boundary condition (8.31) becomes Dirichlet.) The resulting system then contains $1 + (N-1) = N$ equations for the N unknowns: Y_0 through Y_{N-1}, and hence can be solved for a unique solution $\{Y_n\}_{n=0}^{N-1}$, unless the coefficient matrix of this system is singular. We will remark on the latter possibility after we describe the other method.

Method 2

This method does *not* use a fictitious point to approximate the boundary condition with the required accuracy. We will first show how this method can be used for the second-order accurate approximation (8.4) and then indicate how it (the method) can be modified for the fourth-order accurate approximation (8.29).

In analogy with Eq. (3.10) of Lecture 3, one can obtain:

$$\frac{y_{n+1} - y_n}{h} = y_n' + \frac{h}{2}y_n'' + O(h^2)\,. \tag{8.37}$$

Then, using the ODE in (8.1) to express y'' as $R - Py' - Qy$, one finds:

$$\begin{aligned} y_n' &= \frac{y_{n+1} - y_n}{h} - \frac{h}{2}(R_n - P_n y_n' - Q_n y_n) + O(h^2) \\ &= \frac{y_{n+1} - y_n}{h} - \frac{h}{2}\left(R_n - P_n\left[\frac{y_{n+1} - y_n}{h} + O(h)\right] - Q_n y_n\right) + O(h^2)\,. \end{aligned} \tag{8.38}$$

Therefore, the boundary condition (8.31) can be approximated by

$$A_1 Y_0 + A_2 \left[\frac{Y_1 - Y_0}{h} - \frac{h}{2}\left(R_0 - P_0 \frac{Y_1 - Y_0}{h} - Q_0 Y_0 \right) \right] = \alpha\,, \qquad (8.39)$$

with the discretization error being $O(h^2)$. To answer some of the Questions for self-assessment, you will need to collect the coefficients of Y_0 and Y_1 in (8.39) to put it in a form similar to that of (8.36). Such a rearranged Eq. (8.39) should then be used along with the $N-1$ equations in (8.4) for $\{Y_n\}_{n=1}^{N-1}$. Thus we have a system of N equations for N unknowns $\{Y_n\}_{n=0}^{N-1}$, which can be solved (again, unless its coefficient matrix is singular).

If one requires a fourth-order accurate approximation to $y'(a)$, one needs to use two more terms in expansion (8.37):

$$\frac{y_{n+1} - y_n}{h} = y_n' + \frac{h}{2} y_n'' + \frac{h^2}{6} y_n''' + \frac{h^3}{24} y_n'''' + O(h^4) \qquad (8.40)$$

and replace the y'' and higher-order derivatives using $y_n'' = f(x_n, y_n) \equiv f_n$, $f_{n+1} = f_n + h\frac{d}{dx} f_n + \ldots = y_n'' + h y_n''' + \ldots$, etc. The result is:

$$y_n' = \frac{y_{n+1} - y_n}{h} - \frac{h}{24}(7f_n + 6f_{n+1} - f_{n+2}) + O(h^4)\,. \qquad (8.41)$$

Equation (8.41) with $n = 0$ can then be used to obtain a fourth-order accurate counterpart of (8.39).

Remark 1: Recall that the coefficient matrix in the discretized BVP (8.4) with Dirichlet boundary conditions is SDD and hence nonsingular, provided that h satisfies the condition of Theorem 8.3. We will now state a requirement that the coefficients A_1, A_2 in the non-Dirichlet boundary condition (8.31) must satisfy in order to guarantee that the corresponding BVP has a unique solution. To this end, we consider the situations arising in Methods 1 and 2 separately.

When Eq. (8.36) replaces the Dirichlet boundary condition (8.35) in Method 1, the resulting coefficient matrix is SDD provided that

$$A_1 A_2 \le 0, \qquad (8.42)$$

in addition to the earlier requirements of $Q(x) < 0$ and $h \cdot \max |P(x)| \le 2$. (You will be asked to verify (8.42) in one of the Questions for self-assessment.) On the other hand, if two individual equations (8.33) and (8.34) are used instead of their combined form (8.36), the corresponding coefficient matrix is *no longer* SDD. This is one advantage of using (8.36) instead of (8.33) and (8.34).

Thus, condition (8.42) is sufficient, but not necessary, to guarantee that the corresponding discretized BVP has a unique solution. That is, even when (8.42) does not hold, one can still attempt to solve the corresponding linear system, since strict diagonal dominance is only a sufficient, but not necessary, condition for A to be nonsingular.

In the case of Method 2, by collecting the coefficients of Y_0 and Y_1 in Eq. (8.39), it is straightforward (although, perhaps, a little tediuos) to show that the coefficient matrix is SDD if (8.42) and the two conditions stated one line below it, are satisfied. You will verify this in a homework problem. The fourth-order accurate analog of (8.39), based on Eq. (8.41), also yields the same conditions for strict diagonal dominance of the coefficient matrix.

Remark 2: In the case of the BVP with Dirichlet boundary conditions, the coefficient matrix is tridiagonal (see Eq. (8.6)), and hence the corresponding linear system can be solved efficiently by the Thomas algorithm. In the case of non-Dirichlet boundary conditions, one can show (and you will be asked to do so) that Method 1 based on Eq. (8.36) yields a tridiagonal system, but the same method using Eqs. (8.33) and (8.34) does not. This is the other advantage of using the single Eq. (8.36) in this method.

Method 2 for a second-order accurate approximation to the BVP gives a tridiagonal matrix. This can be straightforwardly shown by the same rearrangement of terms in Eq. (8.39) that was used above to show that the corresponding matrix is SDD. However, the fourth-order accurate modification of (8.39), based on (8.41), produces a matrix $\tilde{A}$ that is no longer tridiagonal. One can handle this situation in two ways. First, if one is willing to sacrifice one order of accuracy in exchange for the convenience of having a tridiagonal coefficient matrix, then instead of (8.41), one can use

$$y'_n = \frac{y_{n+1} - y_n}{h} - \frac{h}{6}(2f_n + f_{n+1}) + O(h^3)\,. \tag{8.43}$$

Then the modification of (8.39) based on (8.43) does result in a tridiagonal coefficient matrix. Alternatively, one can see that matrix $\tilde{A}$ obtained with the fourth-order accurate formula (8.41) differs from a tridiagonal one only in having its $(1,3)$th entry nonzero (verify). Thus, $\tilde{A}$ is in some sense "very close" to a tridiagonal matrix, and we can hope that this fact could be used to find $\tilde{A}^{-1}$ with only $O(M)$ operations. This can indeed be done by using the algorithm described in the next section.

8.5 Periodic boundary condition; Sherman–Morrison algorithm

In this section, we will consider the case when BVP (8.1) has periodic boundary conditions. We will see that the coefficient matrix arising in this case is *not* tridiagonal, but is, in some sense, close to it. We will then present an algorithm that will allow us to find the inverse of that matrix using only $O(M)$ operations, where $M \times M$ is the dimension of the matrix. The same method can also be used in other situations, including the inversion of matrix $\tilde{A}$ defined in the last paragraph of the preceding section.

We first obtain the analogues of Eqs. (8.5)–(8.7) in the case of the BVP having periodic boundary conditions. Consider the corresponding counterpart of the BVP (8.1):

$$\begin{aligned} &y'' + P(x)y' + Q(x)y = R(x)\,, \\ &y(a) = y(b). \end{aligned} \tag{8.44}$$

The corresponding counterpart of system (8.4) is

$$\begin{aligned} &Y_0 = Y_N\,; \\ &(1 + \tfrac{h}{2}P_n)Y_{n+1} - (2 - h^2Q_n)Y_n + (1 - \tfrac{h}{2}P_n)Y_{n-1} = h^2R_n\,, \qquad 0 \le n \le N\,. \end{aligned} \tag{8.45}$$

Note 1: The index n in (8.45) runs from 0 to N, while in (8.4) in runs from 1 to $N-1$.
Note 2: Since our problem has periodic boundary conditions, it is logical to let

$$Y_{-n} = Y_{N-n} \quad \text{and} \quad Y_{N+n} = Y_n, \qquad 0 \le n \le N\,.$$

In particular,

$$Y_{-1} = Y_{N-1} \quad \text{and} \quad Y_N = Y_0 \tag{8.46}$$

(the last equation here is just the original periodic boundary condition).

In view of Eq. (8.46), the equations in system (8.45) with $n = 0$ and $n = N-1$ can be written as follows:

$$\begin{aligned} &(1+\tfrac{h}{2}P_0)Y_1 - (2-h^2Q_0)Y_0 + (1-\tfrac{h}{2}P_0)Y_{N-1} = h^2R_0\,, \\ &(1+\tfrac{h}{2}P_{N-1})Y_0 - (2-h^2Q_{N-1})Y_{N-1} + (1-\tfrac{h}{2}P_{N-1})Y_{N-2} = h^2R_{N-1}\,. \end{aligned} \tag{8.47}$$

With this result, system (8.45) can be written in the matrix form (8.5), where now

$$\vec{\mathbf{Y}} = [Y_0,\, Y_1,\, \ldots\,, Y_{N-1}]^T\,, \tag{8.48}$$

$$A = \begin{pmatrix} -(2-h^2Q_0) & (1+\frac{h}{2}P_0) & 0 & 0 & \cdots & \boxed{(1-\frac{h}{2}P_0)} \\ (1-\frac{h}{2}P_2) & -(2-h^2Q_2) & (1+\frac{h}{2}P_2) & 0 & \cdots & 0 \\ 0 & (1-\frac{h}{2}P_3) & -(2-h^2Q_3) & (1+\frac{h}{2}P_3) & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & (1-\frac{h}{2}P_{N-2}) & -(2-h^2Q_{N-2}) & (1+\frac{h}{2}P_{N-2}) \\ \boxed{(1+\frac{h}{2}P_{N-1})} & \cdots & 0 & 0 & (1-\frac{h}{2}P_{N-1}) & -(2-h^2Q_{N-1}) \end{pmatrix} \tag{8.49}$$

and

$$\vec{\mathbf{r}} = \left[h^2R_0,\quad h^2R_1,\quad h^2R_2,\quad \cdots,\quad h^2R_{N-2},\quad h^2R_{N-1}\right]^T\,. \tag{8.50}$$

Matrix A in Eq. (8.49) differes from its counterpart in Eq. (8.6) in two respects: (i) its dimension is $N \times N$ rather than $(N-1)\times(N-1)$ and (ii) it is not tridiagonal due to the terms in its upper-right and lower-left corners. Such matrices are called *circulant.*

Thus, to obtain the solution of the BVP (8.44), we will need to solve the linear system (8.5 with the non-tridiagonal matrix A. We will now show how this problem can be reduced to the solution of a system with a tridiagonal matrix. To this end, we first make a preliminary observation. Let $\vec{\mathbf{w}}$ be a vector whose only nonzero entry is its ith entry and equals w_i. Let $\vec{\mathbf{z}}$ be a vector whose only nonzero entry is its jth entry and equals z_j. Then $C = \vec{\mathbf{w}}\vec{\mathbf{z}}^T$ is an $N \times N$ matrix whose only nonzero entry is $C_{ij} = w_iz_j$ (verify). Similarly, if $\vec{\mathbf{w}} = [w_1,\, 0,\, 0,\, \ldots\,, w_N]^T$ and $\vec{\mathbf{z}} = [z_1,\, 0,\, 0,\, \ldots\,, z_N]^T$, then

$$C = \vec{\mathbf{w}}\,\vec{\mathbf{z}}^T = \begin{pmatrix} w_1z_1 & 0 & \cdots & 0 & w_1z_N \\ 0 & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & 0 \\ w_Nz_1 & 0 & \cdots & 0 & w_Nz_N \end{pmatrix}. \tag{8.51}$$

Therefore, the circulant matrix A in Eq. (8.49) can be represented as:

$$A = A_{\text{tridiag}} + \vec{\mathbf{w}}\,\vec{\mathbf{z}}^T\,, \tag{8.52}$$

where A_{tridiag} is some tridiagonal matrix and $\vec{\mathbf{w}}$ and $\vec{\mathbf{z}}$ are some properly chosen vectors. Note that while the choice of $\vec{\mathbf{w}}$ and $\vec{\mathbf{z}}$ allows much freedom, the form of A_{tridiag} is unique for a given circulant matrix A, once $\vec{\mathbf{w}}$ and $\vec{\mathbf{z}}$ have been chosen. In a homework problem, you will be asked to make a choice of $\vec{\mathbf{w}}$ and $\vec{\mathbf{z}}$ and consequently come up with the expression for A_{tridiag}, given that A is as in Eq. (8.49).

Linear systems (8.5) with the coefficient matrix A given by (8.52) can be time-efficiently — i.e., in $O(M)$ operations — solved by the so-called Sherman–Morrison algorithm. This algorithm can be found in most textbooks on Numerical Analysis, or online.

8.6 Nonlinear BVPs

The analysis presented in this section is carried out with three restrictions. First, we consider only BVPs with Dirichlet boundary conditions. Generalizations for boundary conditions of the form (8.31) can be done straightforwardly along the lines of the previous section.

Second, we only consider the BVPs

$$y'' = f(x, y), \qquad y(a) = \alpha, \quad y(b) = \beta\,, \tag{8.28}$$

which does not involve y'. Although the methods described below can also be adopted to a more general BVP

$$y'' = f(x, y, y'), \qquad y(a) = \alpha, \quad y(b) = \beta\,, \tag{7.1}$$

the analysis of *convergence* of those methods to a solution of the BVP (7.1) is significantly more complex than the one presented here. Thus, the methods that we will develop in this section can be applied to (7.1) *without a guarantee* that one will obtain a solution of that BVP.

Third, we will focus our attention on the second-order accurate discretization of the BVPs. The analysis for the fourth-order accurate discretization scheme is essentially the same, and produces similar results.

Consider the BVP (8.28). The counterpart of system (8.4) for this BVP has the same form as (8.5), i.e.:

$$A\vec{\mathbf{Y}} \;=\; \vec{\mathbf{r}}\,, \tag{8.5}$$

where now

$$A = \begin{pmatrix} -2 & 1 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 1 & -2 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 1 & -2 & 1 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & 0 & 1 & -2 & 1 \\ 0 & \cdot & \cdot & \cdot & 0 & 0 & 1 & -2 \end{pmatrix}; \qquad \vec{\mathbf{r}} = \begin{pmatrix} h^2 f(x_1, Y_1) - \alpha \\ h^2 f(x_2, Y_2) \\ \cdots \\ h^2 f(x_{N-2}, Y_{N-2}) \\ h^2 f(x_{N-1}, Y_{N-1}) - \beta \end{pmatrix}. \tag{8.53}$$

Equations (8.5) and (8.53) constitute a system of nonlinear algebraic equations. Below we consider three methods of *iterative* solution of such a system. Of these methods, Method 1 is an analogue of the fixed-point iteration method for solving a single linear equation[23]

$$\mathcal{A} \cdot y = r(y), \tag{8.54}$$

Method 2 contains its modifications, and Method 3 is the analog of the Newton–Raphson method for (8.54).

Method 1 (Picard iterations)

The fixed-point iteration scheme, also called the Picard iteration scheme, for the single nonlinear equation (8.54) is simply

$$y^{(k+1)} = \frac{1}{\mathcal{A}} r(y^{(k)}), \tag{8.55}$$

where $y^{(k)}$ denotes the kth iteration of the solution of (8.54). To start the iteration scheme (8.55), one, of course, needs an *initial guess* $y^{(0)}$.

[23] The constant $\mathcal{A}$ in (8.54) could have been absorbed into the function $r(y)$, but we kept it so as to mimic the notations of (8.5).

Now, let $\vec{\mathbf{Y}}^{(k)}$ denote the kth iteration of the solution of the matrix nonlinear equation (8.5), (8.53). Then the corresponding Picard iteration scheme is

$$A\vec{\mathbf{Y}}^{(k+1)} = \vec{\mathbf{r}}\left(\vec{x}, \vec{\mathbf{Y}}^{(k)}\right), \qquad k = 0, 1, 2, \ldots . \tag{8.56}$$

Importantly, unlike the original nonlinear Eqs. (8.5) and (8.53), Eq. (8.56) is a *linear* system. Indeed, at the $(k+1)$th iteration, the unknown vector $\vec{\mathbf{Y}}^{(k+1)}$ enters linearly, while the nonlinear r.h.s. contains $\vec{\mathbf{Y}}^{(k)}$ that has been determined at the previous iteration.

Let us invetsigate the rate of convergence of the iteration scheme (8.56). For its one-variable counterpart, Eq. (8.54), the convergence condition of the fixed-point iterations (8.55) is well-known to be

$$\left| \mathcal{A}^{-1} \cdot \frac{dr(y)}{dy} \right| < 1 \tag{8.57}$$

for all y sufficiently close to the fixed point. The condition we will establish for (8.56) will be a direct analog of (8.57).

Before we proceed, let us remark that the analysis of convergence of Picard's iteration scheme (8.56) can proceed in two slightly different ways. In one way, one can transform (8.56) into an iteration scheme for $\vec{\varepsilon}^{(k)} = \vec{\mathbf{Y}}^{(k)} - \vec{\mathbf{Y}}^{(k-1)}$. Then the conditions that the iterations converge is expressed by

$$\left\|\vec{\varepsilon}^{(k)}\right\| < \left\|\vec{\varepsilon}^{(k-1)}\right\| \qquad \text{for all sufficiently large } k, \tag{8.58}$$

where, as in Lecture 4 (see (4.1)), $|| \ldots ||$ denotes the ∞-norm:

$$\|\vec{\varepsilon}\| \equiv \|\vec{\varepsilon}\|_\infty = \max_n |\varepsilon_n| \,.$$

Indeed, this condition implies that $\lim_{k\to\infty} \left\|\vec{\varepsilon}^{(k)}\right\| = 0$. This, in its turn, implies that the sequence of iterations $\left\{\vec{\mathbf{Y}}^{(k)}\right\}$ tends to a limit, which then, according to (8.56), must be a solution of (8.5).

We will, however, proceed in a slightly different way. Namely, we will assume that our starting guess, $\vec{\mathbf{Y}}^{(0)}$, is sufficiently close to the exact solution, $\vec{\mathbf{Y}}$, of (8.5). Then, as long as the iterations converge, $\vec{\mathbf{Y}}^{(k)}$ will stay close to $\vec{\mathbf{Y}}$ for all k, and one can write

$$\vec{\mathbf{Y}}^{(k)} = \vec{\mathbf{Y}} + \vec{\varepsilon}^{(k)}, \qquad \text{where } \left\|\vec{\varepsilon}^{(k)}\right\| \ll 1. \tag{8.59}$$

The condition that iterations (8.56) converge has exactly the same form, (8.58), as before. However, its interpretation is slightly different (although equivalent): Now the fact that $\lim_{k\to\infty} \left\|\vec{\varepsilon}^{(k)}\right\| = 0$ implies that $\lim_{k\to\infty} \vec{\mathbf{Y}}^{(k)} = \vec{\mathbf{Y}}$, i.e. that the iterative solutions converge to the exact solution of (8.5).

Both methods of convergence analysis described above can be shown to yield the same conclusions. We chose to follow the second method, based on (8.59), because it can be more naturally related to the linearization of the nonlinear equation at hand (see below). Linearization results in replacing the analysis of the original nonlinear equation (8.5) by the analysis of a system of linear equations, which can be carried out using well-developed methods of Linear Algebra.

Thus, we begin the convergence analysis of the iteration scheme (8.56) by substituting there expression (8.59) and linearizing the right-hand side using the first two terms of the Taylor expansion near $\vec{\mathbf{Y}}$:

$$\cancel{A\vec{\mathbf{Y}}} + A\vec{\varepsilon}^{(k+1)} = \cancel{\vec{\mathbf{r}}(\vec{x}, \vec{\mathbf{Y}})} + \frac{\partial \vec{\mathbf{r}}}{\partial \vec{\mathbf{Y}}}\, \vec{\varepsilon}^{(k)} + O\big(\left\|\vec{\varepsilon}^{(k)}\right\|^2 \big). \tag{8.60}$$

The first terms on both sides of the above equation cancel out by virtue of the exact equation (8.5). Then, upon discarding quadratically small terms and using the definition of $\vec{\mathbf{r}}$ from (8.53), one obtains a linear system

$$A\vec{\varepsilon}^{(k+1)} = \frac{\partial \vec{\mathbf{r}}}{\partial \vec{\mathbf{Y}}}\,\vec{\varepsilon}^{(k)}, \qquad \text{where} \quad \frac{\partial \vec{\mathbf{r}}}{\partial \vec{\mathbf{Y}}} = h^2\,\mathrm{diag}\left(\frac{\partial f(x_1,Y_1)}{\partial Y_1},\ \ldots\ ,\frac{\partial f(x_{N-1},Y_{N-1})}{\partial Y_{N-1}}\right). \tag{8.61}$$

We will use this equation to relate the norms of $\vec{\varepsilon}^{(k+1)}$ and $\vec{\varepsilon}^{(k)}$ and thereby establish a sufficient condition for convergence of iterations (8.56).

Let us assume that

$$\max_{1\le n\le N-1}\left|\frac{\partial f(x_n,Y_n)}{\partial Y_n}\right| = L\,. \tag{8.62}$$

Multiplying both sides of (8.56) by A^{-1} and taking their norm, one obtains:

$$\left|\left|\vec{\varepsilon}^{(k+1)}\right|\right| \le ||A^{-1}||\cdot h^2 L\cdot \left|\left|\vec{\varepsilon}^{(k)}\right|\right|\,, \tag{8.63}$$

where we have also used a known fact from Linear Algebra, stating that for any matrix A and vector $\vec{\mathbf{z}}$,

$$||A\vec{\mathbf{z}}|| \le ||A||\cdot||\vec{\mathbf{z}}||$$

(actually, the latter inequality follows directly from the definition of a matrix norm). For completeness, let us mention that

$$||A||_\infty = \max_{1\le i\le M}\sum_{j=1}^{M}|a_{ij}|\,. \tag{8.64}$$

Inequality (8.63) shows that

$$\left|\left|\vec{\varepsilon}^{(k)}\right|\right| \le \left(h^2L\,||A^{-1}||\right)^k \left|\left|\vec{\varepsilon}^{(0)}\right|\right|, \tag{8.65}$$

which implies that Picard iterations converge when

$$h^2L\,||A^{-1}|| < 1\,. \tag{8.66}$$

As promised earlier, this condition is analogous to (8.57).

It now remains to find $||A^{-1}||$. Since matrix A shown in Eq. (8.53) arises in a great many applications, the explicit form of its inverse has been calculated for any size $N=(b-a)/h$. The derivation of A^{-1} can be found on photocopied pages posted on the course website; the corresponding result for $||A^{-1}||$, obtained with the use of (8.64), is:

$$||A^{-1}|| = \frac{(b-a)^2}{8h^2}\,. \tag{8.67}$$

Substituting (8.67) into (8.66), we finally obtain that *for Picard iterations to converge, it is sufficient (but not necessary)* that

$$\frac{(b-a)^2}{8}L\,<\,1\,. \tag{8.68}$$

Thus, whether the Picard iterations converge to the discretized solution of the BVP (8.28) depends not only on the function $f(x,y)$ but also on the length of the interval $[a,b]$.

Method 2 (modified Picard iterations)

The idea of the modified Picard iterations method can be explained using the example of the single equation (8.54), where we will set $\mathcal{A} = 1$ for convenience and without loss of generality. Suppose the simple fixed-point iteration scheme (8.55) does not converge because at the fixed point $\bar{y}$, $dr(\bar{y})/dy \approx \kappa > 1$,[24] so that the convergence condition (8.57) is violated. Then, instead of iterating (8.54), let us iterate

$$y - \kappa y = r(y) - \kappa y \qquad \Rightarrow \qquad (1-\kappa) y^{(k+1)} = r(y^{(k)}) - \kappa y^{(k)} \qquad \Rightarrow \qquad y^{(k+1)} = \frac{r(y^{(k)}) - \kappa y^{(k)}}{1-\kappa}\,. \tag{8.69}$$

Note that the *exact* equation which we start with in (8.69) is equivalent to Eq. (8.54) (with $\mathcal{A} = 1$), but the iteration equation in (8.69) is *different* from (8.55). If our guess κ at the true value of the derivative $dr(\bar{y})/dy$ is "sufficiently close", then the derivative of the r.h.s. of (8.69) is less than 1, and hence, by (8.57), the iteration scheme (8.69) converges, in contrast to (8.54), which diverges. In other words, by subtracting from both sides of the equation a linear term whose slope closely matches the slope of the nonlinear term at the solution $\bar{y}$, one drastically reduces the magnitude of the slope of the right-hand side of the iterative scheme, thereby making it converge.

Let us return to the iterative solution of Eq. (8.28), where now, instead of iterating, as in Picard's method, the equation

$$\left(y^{(k+1)}\right)'' = f\left(x, y^{(k)}\right), \tag{8.70}$$

we will iterate the equation

$$\left(y^{(k+1)}\right)'' - c\, y^{(k+1)} = f\left(x, y^{(k)}\right) - c\, y^{(k)} \tag{8.71}$$

with some constant c. The corresponding linearized equation in vector form is easily obtained from (8.61):

$$(A - h^2 c\, I)\vec{\varepsilon}^{(k+1)} = h^2\, \mathrm{diag}\left(\frac{\partial f(x_1, Y_1)}{\partial Y_1} - c,\ \ldots,\ \frac{\partial f(x_{N-1}, Y_{N-1})}{\partial Y_{N-1}} - c\right) \vec{\varepsilon}^{(k)}, \tag{8.72}$$

where I is the identity matrix of the same size as A. We will now address the question of how one should choose the constant c so as to ensure convergence of (8.72) and hence of the modified scheme (8.71).

The main difference between the multi-component equation (8.72) and the single-component equation (8.69) is that no single value of c could simultaneously match all of the values $\partial f(x_n, Y_n)/\partial Y_n$, which, in general, are distributed in some interval

$$L^- \le \frac{\partial f(x_n, Y_n)}{\partial Y_n} \le L^+\,, \qquad n = 1, \ldots, N-1\,. \tag{8.73}$$

It may be intuitive to suppose that the optimal choice for c may be at the midpoint of that interval, i.e.,

$$c_{\mathrm{opt}} = L^{\mathrm{av}} = \frac{1}{2}\big(L^- + L^+\big). \tag{8.74}$$

Below we will show that this is indeed the case.

Specifically, let us only consider the case where $\partial f/\partial y > 0$, when a unique solution of BVP (8.28) is guaranteed to exist by Theorem 6.1 of Lecture 6. Then in (8.73), both

$$L^{\pm} > 0, \tag{8.75}$$

[24] Here '$\approx$' is used instead of '$=$' because $dr(\bar{y})/dy$ is usually not known exactly.

and hence $L^{\rm av} > 0$. Next, by following the steps of the derivation of (8.63) one obtains from (8.72):

$$\left\|\vec{\varepsilon}^{(k+1)}\right\| \;\le\; \left(||(A-h^2cI)^{-1}|| \cdot h^2 \max_{L^- \le \ell \le L^+} |\ell - c| \right) \cdot \left\|\vec{\varepsilon}^{(k)}\right\| \,. \tag{8.76}$$

Here ℓ stands for any of $\partial f(x_n, Y_n)/\partial Y_n$, which satisfy (8.73). Note that the maximum above is taken with respect to ℓ while c is assumed to be fixed. Then:

$$L^- - c \le \ell - c \le L^+ - c \quad \Rightarrow \quad \max_{L^- \le \ell \le L^+} |\ell - c| = \max\{|c - L^-|,\, |L^+ - c|\}. \tag{8.77}$$

Our immediate goal is to determine for what c the coefficient multiplying $\left\|\vec{\varepsilon}^{(k)}\right\|$ in (8.76) is the smallest: this will yield the fastest convergence of the modified Picard iterations (8.71). The entire analysis of this question is a little teduious, since one will have to consider separately the cases where $c \in [L^-,\, L^+]$ and $c \not\in [L^-,\, L^+]$. Since we have announced that the answer, (8.74), corresponds to the former case, we will present the details only for it. Details for the case $c \not\in [L^-,\, L^+]$ are similar, but will not yield an optimal value of c, and so we will omit them.

Thus, we are looking to determine

$$K \equiv \min_{L^- \le c \le L^+} \left(h^2\, ||(A - h^2cI)^{-1}||\, \max\{|c-L^-|,\, |L^+ - c|\} \right) \,, \tag{8.78}$$

for which we will first need to find the norm $||(A-h^2cI)^{-1}||$. Since A is a symmetric matrix (see (8.53)), so are matrices $(A-h^2cI)$ and $(A-h^2cI)^{-1}$. In graduate-level courses on Linear Algebra it is shown that the norm of a real symmetric matrix B equals the modulus of its largest eigenvalue. Since (an eigenvalue of B^{-1}) = 1/(an eigenvalue of B), then

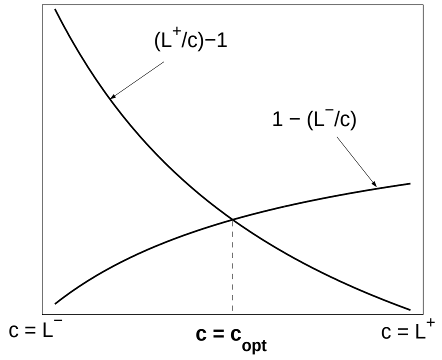
$$||(A-h^2cI)^{-1}|| \;=\; \frac{1}{\min|\text{eigenvalue of }\; (A - h^2cI)\,|} \,. \tag{8.79a}$$

To find the lower bound for the latter eigenvalue, we use the result of Problem 2 of HW 8 (which is based on the Gerschgorin Circles Theorem of Section 8.1). Thus,

$$||(A-h^2cI)^{-1}|| \;=\; \frac{1}{h^2c} \tag{8.79b}$$

(recall that $c > 0$ because we assumed in (8.75) that $L^\pm > 0$ and also since $c \in [L^-,\, L^+]$). Combining (8.78) and (8.79b), we will now determine

$$\begin{aligned} K \;&=\; \min_{L^- \le c \le L^+} \left(\max\left\{ \left|\frac{c-L^-}{c}\right|,\, \left|\frac{L^+-c}{c}\right| \right\} \right) \\ &=\; \min_{L^- \le c \le L^+} \left(\max\left\{ 1 - \frac{L^-}{c},\, \frac{L^+}{c} - 1 \right\} \right) \,. \end{aligned} \tag{8.80}$$



The corresponding optimal c is found as shown in the figure on the left:

$$1-\frac{L^-}{c_{\rm opt}}=\frac{L^+}{c_{\rm opt}}-1$$

$$\Rightarrow \quad c_{\rm opt}=\frac{1}{2}\big(L^-+L^+\big),$$

which is (8.74).

Substituting (8.74) into (8.80) and then using the result in (8.76), one finally obtains:

$$||\bar{\varepsilon}^{(k+1)}|| \le \left(\frac{L^+-L^-}{L^++L^-}\right)||\bar{\varepsilon}^{(k)}||\,. \tag{8.81}$$

Since according to our assumption (8.75) $L^{\pm}>0$, then the factor $(L^+-L^-)/(L^++L^-)<1$. Thus, the modified Picard scheme (8.71), (8.74) converges.

The issues of using the modified Picard iterations on BVPs where conditions (8.75) are not met, or using scheme (8.71) with a non-optimal constant c, are considered in homework problems. Let us only emphasize that the modified Picard iterations *can* sometimes converge even when (8.75) and/or (8.74) do not hold.

Method 3 (Newton–Raphson method)
Although this method can be described for a general BVP of the form (7.1), we will only do so for a particular BVP, encountered in one of the homework problems for Lecture 7. Namely, consider the BVP

$$y''=\frac{y^2}{2+x}, \qquad y(0)=1, \quad y(2)=1. \tag{8.82}$$

Let us begin by writing down the system of second-order accurate discrete equations for this BVP:

$$Y_{n+1}-2Y_n+Y_{n-1}=\frac{h^2\,Y_n^2}{2+x_n}, \qquad Y_0=1, \quad Y_N=1\,. \tag{8.83}$$

Let $Y_n^{(0)}$ be the initial guess for the solution of (8.83) at x_n. Similarly to (8.59), we relate it to the *exact* solution $\{Y_n\}$ of (8.83):

$$Y_n^{(0)}=Y_n+\epsilon_n^{(0)}\,, \qquad |\epsilon_n|\ll 1\,. \tag{8.84}$$

Substituting (8.84) into (8.83) and using the fact that the exact solution $\{Y_n\}$ satisfies (8.83), we obtain an equivalent system:

$$\left[\epsilon_{n+1}^{(0)}-2\epsilon_n^{(0)}+\epsilon_{n-1}^{(0)}-\frac{h^2\cdot 2Y_n^{(0)}\epsilon_n^{(0)}}{2+x_n}\right]+\frac{h^2\left(\epsilon_n^{(0)}\right)^2}{2+x_n}=\left(Y_{n+1}^{(0)}-2Y_n^{(0)}+Y_{n-1}^{(0)}\right)-\frac{h^2\left(Y_n^{(0)}\right)^2}{2+x_n}. \tag{8.85}$$

If we now assume that $\epsilon_n^{(0)}$ are sufficiently small for all n, then we neglect the last term on the l.h.s. of (8.85), and obtain a *linear* system for $\epsilon_n^{(0)}$:

$$\epsilon_{n+1}^{(0)}-\left(2+\frac{2h^2Y_n^{(0)}}{2+x_n}\right)\epsilon_n^{(0)}+\epsilon_{n-1}^{(0)}=\left(Y_{n+1}^{(0)}-2Y_n^{(0)}+Y_{n-1}^{(0)}\right)-\frac{h^2\left(Y_n^{(0)}\right)^2}{2+x_n}\,. \tag{8.86}$$

If our initial guess satisfies the boundary conditions of the BVP (8.82), then we also have

$$\epsilon_0^{(0)} = 0, \qquad \epsilon_N^{(0)} = 0. \tag{8.87}$$

System (8.86), (8.87) can be solved time-efficiently by the Thomas algorithm. Thus we obtain $\epsilon_n^{(0)}$. According to (8.84), this gives the next iteration for our solution $\{Y_n\}$:

$$Y_n \approx Y_n^{(1)} = Y_n^{(0)} - \epsilon_n^{(0)}. \tag{8.88}$$

We then substitute

$$Y_n^{(1)} = Y_n + \epsilon_n^{(1)} \tag{8.89}$$

into (8.83), obtain a system analogous to (8.85), and then solve it for $\epsilon_n^{(1)}$ in exactly the same way we have solved (8.85) for $\epsilon_n^{(0)}$. Repeating these steps, we stop when the process converges, i.e. $\left\|\vec{\mathbf{Y}}^{(k+1)} - \vec{\mathbf{Y}}^{(k)}\right\|$ becomes less than a given tolerance.

Using iterative methods in more complicated situations

First, let us note that iterative methods, which we have described above for ODEs, are equally applicable to partial differential equations (PDEs). This distinguishes iterative methods from the shoting method considered in Lecture 7: as we stressed at the end of that Lecture, the shooting method *cannot* be extended to solving BVPs for PDEs. Thus, iterative methods remain *the only* group of method for solving nonlinear BVPs for PDEs. The ideas of these methods are the same as we have described above for ODEs.

Second, the fixed-point iteration methods (i.e., counterparts of the Picard and modified Picard described above) are not used (or are rarely used) in commercial software. The main reason is that they are considerably slower than the Newton–Raphson method and its variations and also than so-called Krylov subspace methods, studied in advanced courses on Numerical Linear Algebra. We will only mention that the most famous of those Krylov subspace methods is the Conjugate Gradient method (CGM) for solving symmetric positive (or negative) definite linear systems. An accessible introduction to this method can be found in many textbooks and also in an online paper by J.R. Shewchuk, "An Introduction to the Conjugate Gradient Method Without the Agonizing Pain".[25] An extension of the CGM to nonlinear BVPs whose linearization yields symmetric matrices is known as the Newton–CGM or, more generally, as a class of Newton–Krylov methods.

Another reason why Krylov subspace methods are used much more widely than fixed-point iterative methods is that convergence conditions of the latter methods are typically restricted by a condition analogous to (8.75). Some of Krylov subspace methods either do not have restrictions, or are less sensitive to them. Also, the Newton–Raphson method does not have any restrictions similar to (8.75). Yet, this method also has its own issues, and a number of books are written about application of the Newton–Raphson method to systems of nonlinear equations.

8.7 Questions for self-assessment

1. Verify that (8.4) follows from (8.1)–(8.3).

[25] Some pain, however, is to be expected.

2. Explain why the form of the first and last terms of $\vec{\mathbf{r}}$ in (8.7) is different from the form of the other terms.
3. What does the Gerschgorin Circles Theorem allow one to do?
4. What is the difference between a diagonally dominant and strictly diagonally dominant matrices?
5. Make sure you can follow the derivations in the Example about the Gerschgorin Circles Theorem.
6. Make sure you can follow the proof of Theorem 8.2.
7. What can you say about a solution of a linear system (8.5) where A is SDD?
8. Under what condition(s) on h is the discretized BVP (8.4) guaranteed to have a unique solution?
9. Verify (8.20).
10. Verify (8.21) through (8.23).
11. Explain why the local truncation error of the discretized BVP is multiplied by a factor $O(\frac{1}{h^2})$ to give the error of the solution.
12. Describe the idea of Method 1 for handling non-Dirichlet boundary conditions.
13. Derive (8.36).
14. Describe the idea of Method 2 for handling non-Dirichlet boundary conditions.
15. Convince yourself (and be prepared to convince the instructor) that *if* condition (8.42) and the two conditions stated one line below it, hold, then the coefficient matrix in Method 1 based on Eq. (8.36) is SDD.
    Hint: Write the condition that you want to prove and *then* assume that (8.42) and the other two conditions hold.
16. Suppose that you need to solve a BVP with a mixed-type boundary condition and such that (8.42) does *not* hold. Will you attempt to solve such a BVP? What should your expectations be?
17. Suppose that the BVP has a non-Dirichlet boundary condition of the form (8.31) at $x = b$ (the right end point of the interval) rather than at $x = a$. What will the analog of condition (8.42) be in this case?
    Hint 1: Remember that this is a QSA and not an exercise requiring calculations.
    Hint 2: To better visualize the situation, suppose $[a, b] = [-1, 1]$. What mathematical operation will take the left end point of this interval into the right end point? How will this operation transform the terms on the l.h.s. of (8.31)?
18. Convince yourself (and be prepared to convince the instructor) that the statements in Remark 2 in Sec. 8.4 about the coefficient matrices for the second-order methods being tridiagonal, are correct.
19. Describe the idea of the Picard iterations.

20. Derive (8.63) as explained in the text.
21. What is the condition for the Picard iterations to converge? Is this a necessary or sufficient condition? In other words, if that condition does not hold, should one still attempt to solve the BVP?
22. Describe the idea behind the modified Picard iterations.
23. How is the finite-difference implementation of (8.71) different from that of (8.56)?
24. Make sure you can obtain (8.72).
25. Make sure you can obtain (8.76) and (8.77).
26. Explain how the result of Problem 2 of HW 8 leads to (8.79b).
27. Derive (8.74) from the explanation found after (8.80).
28. Obtain (8.81) as explained in the text.
29. Are (8.75) and (8.74) necessary or sufficient conditions for convergence of the modified Picard iterations (8.71)?
30. Make sure you can see where (8.85) comes from.
31. Write down a linear system satisfied by $\epsilon_n^{(1)}$ defined in (8.89).
32. Describe the idea behind the Newton–Raphson method.
33. Can one use iterative methods when solving nonlinear BVPs for PDEs?