

## 10.2.1. Discrete maximum principle.

**Theorem 28.** (*Discrete Maximum Principle*)

(i) If  $L_h u_j \leq 0$  for all  $1 \leq j \leq N-1$  and  $\max(u_0, u_N) \geq 0$ , then  $\max_{0 \leq j \leq N} u_j \leq \max(u_0, u_N)$ .

(ii) If  $L_h u_j \geq 0$  for all  $1 \leq j \leq N-1$ , and  $\min(u_0, u_N) \leq 0$ , then  $\min_{0 \leq j \leq N} u_j \geq \min(u_0, u_N)$ .

*Proof.* We prove (i) by contradiction. Assume that  $\max_{0 \leq j \leq N} u_j = u_k = M$  for some  $1 \leq k \leq N-1$ , where  $M > \max(u_0, u_N)$ . Since  $L_h u_k \leq 0$ ,  $M > 0$ , and  $q(x_k) \geq 0$ , we have

$$\begin{aligned} p_{k-1/2} u_{k-1} + p_{k+1/2} u_{k+1} &\geq [p_{k-1/2} + p(x_{k+1/2} + h^2 q(x_k))] u_k \\ &= [p_{k-1/2} + p_{k+1/2} + h^2 q(x_k)] M \geq [p_{k-1/2} + p_{k+1/2}] M. \end{aligned}$$

But since  $u_{k+1}$  and  $u_{k-1}$  are  $\leq M$ , and  $p_{k-1/2}$  and  $p_{k+1/2}$  are  $> 0$ ,

$$p_{k-1/2} u_{k-1} + p_{k+1/2} u_{k+1} \leq [p_{k-1/2} + p_{k+1/2}] M,$$

with strict inequality holding unless  $u_{k+1} = u_{k-1} = M$ . Hence,

$$p_{k-1/2} u_{k-1} + p_{k+1/2} u_{k+1} = [p_{k-1/2} + p_{k+1/2}] M,$$

and so we must have  $u_{k+1} = u_{k-1} = M$ . We now repeat this argument at all interior points and eventually conclude that  $u_0 = u_N = M$ , which is a contradiction. Thus, we have established (i). In fact, we have established a slightly stronger result, i.e., that if  $\max_{0 \leq j \leq N} u_j \geq 0$  and occurs at an interior point, then  $u_j$  is a constant for all  $j$ . To prove (ii), we let  $z_j = -u_j$ . Then  $L_h z_j = -L_h u_j \leq 0$  if  $L_h u_j \geq 0$  and  $\min u_j = \min(-z_j) = -\max(z_j)$ . Hence,  $\max(z_0, z_N) = -\min(u_0, u_N) \geq 0$ . By part (i),  $\max_{0 \leq j \leq N} z_j \leq \max(z_0, z_N)$ . Hence  $-\min_{0 \leq j \leq N} u_j \leq \max(z_0, z_N)$ , and so  $\min_{0 \leq j \leq N} u_j \geq \min(u_0, u_N)$ .  $\square$

**Corollary 7.** If  $q(x) \equiv 0$ , then the discrete maximum principle holds without the hypotheses  $\max(u_0, u_N) \geq 0$  for part (i) and  $\min(u_0, u_N) \leq 0$  for part (ii).

*Proof.* We see that these hypotheses are only used in one step of the proof to eliminate  $q(x)$  and hence will not be needed if  $q(x) \equiv 0$ .  $\square$

One important consequence of the discrete maximum principle is that it ensures that the discrete system of linear equations always has a unique solution. Since the system  $Au = F$  is a system of  $N-1$  equations in  $N-1$  unknowns, we know that there will exist a unique solution, provided the only solution of the homogeneous problem (i.e., with  $F = 0$ ) is given by  $u = 0$ . But in this case, the linear system is equivalent to the system  $L_h u_j = 0$ ,  $j = 1, \dots, N-1$ , and  $u_0 = u_N = 0$ . Using both parts of the discrete maximum principle, we can conclude that for any  $0 \leq k \leq N$ ,

$$0 = \min(u_0, u_N) \leq \min_{0 \leq j \leq N} u_j \leq u_k \leq \max_{0 \leq j \leq N} u_j \leq \max(u_0, u_N) = 0.$$

Hence  $u_k = 0$  for  $0 \leq k \leq N$ , and so  $u = 0$ .

Another approach to existence and uniqueness of the approximate solution and also to stability and error estimates is by using discrete energy estimates. We illustrate this idea by applying it to establish uniqueness of the approximate solution. Our discrete energy estimate

is established using the following summation by parts formula, analogous to integration to parts.

**Lemma 4.**

$$\sum_{j=1}^{N-1} (w_{j+1} - w_j)v_j = w_N v_N - w_1 v_0 - \sum_{j=0}^{N-1} (v_{j+1} - v_j)w_{j+1}.$$

*Proof.* We start from the following identity.

$$w_{j+1}v_{j+1} - w_j v_j = (w_{j+1} - w_j)v_j + (v_{j+1} - v_j)w_{j+1}.$$

Summing this result from  $j = 1$  to  $N - 1$ , we get

$$w_N v_N - w_1 v_1 = \sum_{j=1}^{N-1} [w_{j+1}v_{j+1} - w_j v_j] = \sum_{j=1}^{N-1} [(w_{j+1} - w_j)v_j + (v_{j+1} - v_j)w_{j+1}].$$

Rearranging terms, we get

$$\begin{aligned} \sum_{j=1}^{N-1} (w_{j+1} - w_j)v_j &= w_N v_N - w_1 v_1 - \sum_{j=1}^{N-1} (v_{j+1} - v_j)w_{j+1} = w_N v_N - w_1 v_0 - w_1(v_1 - v_0) \\ &\quad - \sum_{j=1}^{N-1} (v_{j+1} - v_j)w_{j+1} = w_N v_N - w_1 v_0 - \sum_{j=0}^{N-1} (v_{j+1} - v_j)w_{j+1}. \end{aligned}$$

□

Using this lemma, we can establish the following result.

**Theorem 29.** *Let  $L_h$  be defined by (10.1). Then*

$$\begin{aligned} \sum_{j=1}^{N-1} u_j L_h u_j &= -h^{-2} [p_{N-1/2}(u_N - u_{N-1})u_N - p_{1/2}(u_1 - u_0)u_0] \\ &\quad + h^{-2} \sum_{j=0}^{N-1} p_{j+1/2}(u_{j+1} - u_j)^2 + \sum_{j=1}^{N-1} q(x_j)u_j^2. \end{aligned}$$

*Proof.* Let  $w_j = p_{j-1/2}(u_j - u_{j-1})$  and  $v_j = u_j$ . Then

$$u_j L_h u_j = -h^{-2}(w_{j+1} - w_j)v_j + q(x_j)u_j^2.$$

Applying the summation by parts lemma,

$$\begin{aligned} \sum_{j=1}^{N-1} u_j L_h u_j &= -h^{-2} [w_N v_N - w_1 v_0] + h^{-2} \sum_{j=0}^{N-1} (v_{j+1} - v_j)w_{j+1} + \sum_{j=1}^{N-1} q(x_j)u_j^2 \\ &= -h^{-2} [p_{N-1/2}(u_N - u_{N-1})u_N - p_{1/2}(u_1 - u_0)u_0] + h^{-2} \sum_{j=0}^{N-1} p_{j+1/2}(u_{j+1} - u_j)^2 + \sum_{j=1}^{N-1} q(x_j)u_j^2. \end{aligned}$$

□

We note that uniqueness follows directly from this theorem, since if  $f(x_j) = 0$ ,  $j = 0, \dots, N-1$ , and  $u_0 = u_N = 0$ , then

$$h^{-2} \sum_{j=0}^{N-1} p_{j+1/2} (u_{j+1} - u_j)^2 + \sum_{j=1}^{N-1} q(x_j) u_j^2 = 0.$$

Since  $q(x) \geq 0$  and  $p(x) \geq p_* > 0$ , we conclude that  $(u_{j+1} - u_j)^2 = 0$ ,  $j = 0, \dots, N-1$ . Hence  $u_j$  is a constant, and since  $u_0 = 0$ , all the  $u_j$  are zero.

10.2.2. *Stability and error estimates.* A second important consequence of the discrete maximum principle is that we can use it to prove a stability result for our approximation scheme, which can then be used to derive error estimates.

To simplify the presentation, consider the special case when  $p(x) = 1$  and  $q(x) = 0$ , so we are considering the differential equation  $-u''(x) = f(x)$ . We can then establish the following stability result for our approximation scheme.

**Theorem 30.** *Let  $v$  be a function defined on the mesh points  $x_j$ ,  $j = 0, \dots, N$ . Then*

$$\max_{0 \leq j \leq N} |v_j| \leq \max(|v_0|, |v_N|) + \frac{(b-a)^2}{2} \max_{1 \leq j \leq N-1} |L_h v_j|.$$

*Proof.* Let  $w_j = (x_j - a)^2/2$ ,  $j = 0, 1, \dots, N$ . Then,  $0 \leq w_j \leq (b-a)^2/2$ . Furthermore,

$$L_h w_j = \frac{1}{2h^2} [-(x_j - h - a)^2 + 2(x_j - a)^2 - (x_j + h - a)^2] = -1.$$

Next, define mesh functions  $\phi_j^+$  and  $\phi_j^-$  by  $\phi_j^\pm = \pm v_j + M w_j$ , where  $M = \max_{1 \leq j \leq N-1} |L_h v|$ . Then for  $1 \leq j \leq N-1$ ,

$$L_h \phi_j^\pm = \pm L_h v_j + M L_h w_j = \pm L_h v_j - M \leq 0.$$

Hence, by the discrete maximum principle, for  $1 \leq j \leq N$ ,

$$\begin{aligned} \phi_j^\pm &\leq \max(\phi_0^\pm, \phi_N^\pm) = \max(\pm v_0 + M w_0, \pm v_N + M w_N) \\ &\leq \max(\pm v_0, \pm v_N) + M \max(w_0, w_N) = \max(\pm v_0, \pm v_N) + M(b-a)^2/2. \end{aligned}$$

Since  $M w_j \geq 0$ ,

$$\pm v_j = \phi_j^\pm - M w_j \leq \phi_j^\pm.$$

Hence,

$$\pm v_j \leq \max(\pm v_0, \pm v_N) + M(b-a)^2/2,$$

and so

$$\max_{0 \leq j \leq N} |v_j| \leq \max(|v_0|, |v_N|) + \frac{(b-a)^2}{2} \max_{1 \leq j \leq N-1} |L_h v_j|.$$

□

Using this stability result together with the estimate for the local truncation error of the method, we easily derive an error estimate for our finite difference scheme. First, we establish the following result.

**Theorem 31.** *Suppose  $u$  is the exact solution of the boundary value problem and  $u_j$  the approximation to  $u$  at  $x_j$  given by the finite difference scheme. Then*

$$\max_{0 \leq j \leq N} |u(x_j) - u_j| \leq (b-a)^2/2 \max_{1 \leq j \leq N-1} |L_h u(x_j) - Lu(x_j)|.$$

*Proof.* Set  $v_j = u(x_j) - u_j$ . Then  $v_0 = v_N = 0$  and

$$\begin{aligned} L_h v_j &= L_h u(x_j) - L_h u_j = L_h u(x_j) - Lu(x_j) + Lu(x_j) - L_h u_j \\ &= L_h u(x_j) - Lu(x_j) + f(x_j) - f(x_j) = L_h u(x_j) - Lu(x_j). \end{aligned}$$

Hence,

$$\max_{1 \leq j \leq N-1} |L_h v_j| = \max_{1 \leq j \leq N-1} |L_h u(x_j) - Lu(x_j)|,$$

and the result follows directly from our previous theorem.  $\square$

Noting that the quantity  $|L_h u(x_j) - Lu(x_j)|$  is just the local truncation error of the method, we immediately get the following corollary.

**Corollary 8.** *If  $u \in C^4[a, b]$ , then*

$$\max_j |u(x_j) - u_j| \leq \frac{(b-a)^2}{24} h^2 \max_{[a,b]} |u^{(4)}(x)|.$$

**Remark:** The quantity  $Lu - L_h u$  is called the consistency error in the approximation of  $Lu$  by  $L_h u$ . The statement that  $\max |Lu - L_h u| \rightarrow 0$  as  $h \rightarrow 0$  says the approximation is consistent.

Since our approximate problem has a unique solution, there will be a constant  $C_h$  depending on  $h$  such for any any mesh function  $v$ ,

$$\max_{0 \leq j \leq N} |v_j| \leq C_h \left[ \max(|v_0|, |v_N|) + \max_{1 \leq j \leq N-1} |L_h v_j| \right].$$

If there exists a constant  $C$  (the stability constant) such that  $C_h \leq C$  for all  $0 < h \leq h_0$ , (i.e., the estimate holds with a constant  $C$  independent of  $h$ ), then we say the approximation scheme is stable.

Our theorem showed the the error  $\max |u(x_i) - u_i|$  was bounded by the stability constant times the maximum of the consistency error. So we have the result that stability + consistency implies convergence, i.e.,  $\lim_{h \rightarrow 0} \max |u(x_i) - u_i| = 0$ . In this case, we imposed the boundary conditions exactly, so there was no consistency error due to approximation of the boundary conditions. However, in other problems, such as those described in the next section, this will not be the case.

10.2.3. *Other boundary conditions.* So far, we have only considered boundary conditions of the form  $u(a) = g_a$  (Dirichlet boundary conditions). Another important type of boundary condition is  $p(a)u'(a) = g_a$  (Neumann boundary condition). For such a boundary condition,  $u_0$  becomes an unknown, and we need to impose an additional equation. The simplest approach is to use a finite difference approximation to  $u'(a)$  using only boundary and interior points. Using the Taylor series expansion, we have

$$u(a+h) = u(a) + hu'(a) + h^2u''(\xi_a)/2.$$

Hence,

$$u'(a) = [u(a+h) - u(a)]/h - hu''(\xi_a)/2.$$

Discarding the last term, we could then approximate the boundary condition  $p(a)u'(a) = g_a$  by the equation  $p(a)[u_1 - u_0]/h = g_a$ . In order to preserve the symmetry of the matrix, we could use instead the approximate boundary condition  $p_{1/2}[u_1 - u_0]/h = g_a$ . An analogous approximation can be used if we replace the boundary condition  $u(b) = g_b$  by the boundary condition  $p(b)u'(b) = g_b$ . In this approach, we use only an  $O(h)$  approximation to the derivative and have approximated  $p(a)$  by  $p(a+h/2)$ , again a first order approximation. One must then check the overall effect on the error because of use of these low order approximations to the boundary conditions.

Another approach, which gives an  $O(h^2)$  approximation to the boundary conditions, is to introduce a fictitious value  $u_{-1}$  at  $x = a - h$  and assume the differential equation also holds at the point  $x = a$ . Thus, the unknowns are now  $u_{-1}, u_0, \dots, u_{N-1}$  (assuming we keep the boundary condition  $u(b) = g_b$ ). The discrete equations at  $x = a + h$  and  $x = a$  are:

$$-p_{1/2}u_0 + a_1u_1 - p_{3/2}u_2 = h^2f(x_1), \quad -p_{-1/2}u_{-1} + a_0u_0 - p_{1/2}u_1 = h^2f(x_0).$$

We then approximate the boundary condition to produce an equation that can be used to eliminate  $u_{-1}$ . We note from our previous Taylor expansions that

$$[v(x+\epsilon) + v(x-\epsilon)]/2 = v(x) + O(\epsilon^2).$$

Choosing  $v = pu'(x)$ ,  $x = a$ , and  $\epsilon = h/2$ , we see that

$$\begin{aligned} g_a = p(a)u'(a) &= [p(a+h/2)u'(a+h/2) + p(a-h/2)u'(a-h/2)]/2 + O(h^2) \\ &= p(a+h/2)\frac{u(a+h) - u(a)}{2h} + p(a-h/2)\frac{u(a) - u(a-h)}{2h} + O(h^2). \end{aligned}$$

Hence, we take as the additional equation:

$$p_{1/2}u_1 - p_{-1/2}u_{-1} + [p_{-1/2} - p_{1/2}]u_0 = g_a.$$

A similar procedure can be employed at the point  $x = b$ .

10.2.4. *Nonlinear problems.* If we return to the more general problem,

$$u'' = f(x, u, u'), \quad u(a) = g_a, \quad u(b) = g_b,$$

we can employ many of the same ideas to produce a discrete system. However, in this case, we will end up with a nonlinear systems of equations, so we will need to use the methods of the previous section, and the problem becomes much more complicated.