

# Optimal Upstream Collocation: A Survey of Recent Results

Stephen H. Brill<sup>a</sup>

<sup>a</sup>Department of Mathematics, Boise State University, Boise, Idaho, U.S.A.

We give herein analytical formulas for the Hermite collocation solution of the steady-state convection-diffusion equation with Dirichlet boundary conditions defined on a uniform mesh in one spatial dimension. Analysis is provided which compares the discrete collocation solution to the corresponding exact solution. Extremely accurate collocation solutions are easily attainable in a variety of settings by utilizing the author's "optimal upstream weighting" technique.

## 1. INTRODUCTION AND BACKGROUND

Convection-diffusion (C-D) differential equations (DEs) are used extensively to study physical processes in the sciences and engineering, including the modeling of subsurface contaminant transport. The numerical solution of such equations can, however, be plagued by spurious (and physically unmeaningful) oscillations, particularly when convection is the dominant process. To ameliorate the effect of these oscillations, the technique of "upstream weighting" is often applied to the convective term. This introduces artificial dispersion, often at the expense of "smearing" the sharp solution profile that characterizes the convection-dominated problem. The goal, therefore, is to obtain highly accurate solutions that suffer from neither oscillations nor "smearing."

In this paper, we provide formulas for the Hermite collocation solution of the steady-state convection-diffusion equation

$$-D \frac{d^2 u}{dx^2} + v \frac{du}{dx} = S(x), \quad (1)$$

with boundary conditions

$$\begin{aligned} u(0) &= u_L \\ u(1) &= u_R, \end{aligned} \quad (2)$$

where the diffusion coefficient  $D$  and velocity coefficient  $v$  are both positive. In this paper,  $D$  will always be constant,  $v$  will be piecewise constant and  $S(x) \equiv 0$ . The formulas mentioned above are analyzed so that judicious choices can be made concerning the value to be assigned to free parameters so that highly accurate approximations to the exact solution of our boundary value problem (BVP) can be obtained. The motivation for choosing the Hermite collocation discretization is that collocation provides great flexibility with respect to where to evaluate the various terms in the discretized version of (1).

Although collocation has been widely studied, there had been no work done in which analytical formulas for collocation discretizations of differential equations were derived,

until we turned our attention to this particular problem. In addition to our efforts on the DE (1) (which will be discussed in more detail immediately below), we have studied the one-dimensional self-adjoint constant-coefficient DE [2].

Our initial effort [3] on studying the DE (1) concerned the setting where  $v$  is constant and no upstream weighting was considered. In this case, we proved that oscillations were eliminated but that significant “smearing” occurred in the convection-dominated case. To eliminate the unacceptable smearing, we next studied the effect of including upstream weighting of the convective term [4], which is governed by a parameter  $\zeta$ . An algorithm for how to choose  $\zeta$  in an optimal fashion was derived. Utilizing this algorithm provides extremely accurate collocation solutions, especially when convection dominates.

We then turned our attention to the case where the source/sink term  $S(x)$  is non-zero [5], and derived formulas for the collocation solution for this case. The issue of where to evaluate the source/sink term  $S(x)$  arises naturally. For the case where this term is a linear function, we prove that  $S(x)$  should be evaluated at precisely the same upstream locations as the convective term in (1) in order to obtain extremely accurate collocation solutions.

Our present effort is devoted to the case where the convection coefficient  $v$  is piecewise constant. Although we have not yet determined how to optimally choose the collocation points for this case, we do present here the results we have obtained to date.

The rest of this paper is organized to have one section correspond to each of the versions of (1) described above, with the exception that the  $S(x) \neq 0$  case is not discussed (the interested reader may consult [5]). Description of the Hermite collocation technique of discretization is included when appropriate. We conclude this work with a short section summarizing our results.

## 2. NO UPSTREAM WEIGHTING

As mentioned above, throughout this paper we assume that the source/sink term  $S(x)$  is identically zero. Furthermore, in this section and the next, we assume that  $v$  is constant. Detailed derivations of the results reported in this section may be found in [3]. Upstream weighting will be introduced in the next section.

The differential equation (1) is defined on the interval  $\mathcal{I} = [0, 1]$ . The Dirichlet boundary conditions (2) are included. We partition the interval  $\mathcal{I}$  into  $m$  uniform subintervals, each of length  $h$ , by  $0 = x_0, x_1, x_2, \dots, x_m = 1$ . Note  $x_j = jh$  and  $mh = 1$ .

The discretization proceeds by introducing a Hermite piecewise cubic interpolating polynomial

$$\hat{u}(x) = \sum_{j=0}^m [u_j f_j(x) + u'_j g_j(x)]. \quad (3)$$

into the differential equation (1), obtaining

$$-D \frac{d^2 \hat{u}}{dx^2} + v \frac{d \hat{u}}{dx} = E(x), \quad (4)$$

where  $E(x)$  is an error function.

The Hermite basis functions, defined for  $\eta \in \left[-\frac{1}{2}, \frac{1}{2}\right]$ , are

$$f_j(x) = \begin{cases} f_L(\eta) = \frac{1}{2}(1+2\eta)^2(1-\eta), & x_{j-1} \leq x = x_j + \left(\eta - \frac{1}{2}\right)h \leq x_j \\ f_R(\eta) = \frac{1}{2}(1-2\eta)^2(1+\eta), & x_j \leq x = x_j + \left(\eta + \frac{1}{2}\right)h \leq x_{j+1} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

and

$$g_j(x) = \begin{cases} g_L(\eta) = \frac{h}{8}(2\eta+1)^2(2\eta-1), & x_{j-1} \leq x = x_j + \left(\eta - \frac{1}{2}\right)h \leq x_j \\ g_R(\eta) = \frac{h}{8}(2\eta-1)^2(2\eta+1), & x_j \leq x = x_j + \left(\eta + \frac{1}{2}\right)h \leq x_{j+1} \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

Note that  $\hat{u}$  in (3) interpolates the values  $u_j = u(x_j)$  and  $u'_j = \frac{du}{dx}(x_j)$ ,  $j = 0, 1, \dots, m$ , because  $f_j(x_k) = \delta_{jk}$ ,  $\frac{df_j}{dx}(x_k) = 0$ ,  $g_j(x_k) = 0$ , and  $\frac{dg_j}{dx}(x_k) = \delta_{jk}$ . Here  $\delta_{jk}$  is the Kronecker symbol.

It is clear that (4) has  $2(m+1)$  coefficients, namely  $u_j$  and  $u'_j$ ,  $j = 0, 1, 2, \dots, m$ . However, the imposition of boundary conditions reduces this number to  $2m$ . To generate the  $2m$  equations necessary to find these undetermined coefficients, we enforce that the error function  $E(x)$  in (4) is identically zero at two distinct ‘‘collocation points’’ in the interior of each of the  $m$  subintervals.

Given certain smoothness conditions, the optimal (in terms of minimizing discretization error) location of the collocation points within each subinterval corresponds to the points of Gaussian quadrature [6]. In this section, we will use these optimal collocation points, which correspond to choosing the collocation points as  $\eta = \pm \frac{1}{\sqrt{12}}$  (see (5) and (6)) in each subinterval  $\left[-\frac{1}{2}, \frac{1}{2}\right]$  (given in local  $\eta$  coordinates). (In the following sections, we will not necessarily collocate all terms of (4) at the Gauss points.)

It is straightforward to see that choosing the collocation points in this manner leads to a matrix equation with the repeated computational molecule

$$\begin{bmatrix} \frac{2\sqrt{3}D}{h^2} - \frac{v}{h} & \frac{(1+\sqrt{3})D}{h} + \frac{v}{2\sqrt{3}} & \frac{-2\sqrt{3}D}{h^2} + \frac{v}{h} & \frac{(-1+\sqrt{3})D}{h} - \frac{v}{2\sqrt{3}} \\ \frac{-2\sqrt{3}D}{h^2} - \frac{v}{h} & \frac{(1-\sqrt{3})D}{h} - \frac{v}{2\sqrt{3}} & \frac{2\sqrt{3}D}{h^2} + \frac{v}{h} & \frac{(-1-\sqrt{3})D}{h} + \frac{v}{2\sqrt{3}} \end{bmatrix} \begin{bmatrix} q_j \\ r_j \\ q_{j+1} \\ r_{j+1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (7)$$

for  $j = 0, 1, 2, \dots, m-1$ . Here  $q_j = u_j$  and  $r_j = u'_j$ ,  $j = 0, 1, 2, \dots, m$ . Note that the matrix equation represented by (7) is a system of  $2m$  equations in  $2(m+1)$  unknowns.

It is equally straightforward, although computationally tedious, to show that the solution of (7) is

$$q_j = \frac{u_L(\lambda^m - \lambda^j) + u_R(\lambda^j - 1)}{\lambda^m - 1} \quad (8)$$

and

$$r_j = \frac{\beta m \lambda^j}{\lambda^m - 1} (u_R - u_L), \quad (9)$$

where

$$\lambda = \frac{\beta^2 + 6\beta + 12}{\beta^2 - 6\beta + 12}. \quad (10)$$

The symbol  $\beta$  is the *Péclet number*:

$$\beta = \frac{v}{mD}.$$

It is instructive to compare the collocation solution (8) and (9) of our BVP with its exact solution and its derivative, both evaluated at  $x_j$ :

$$u(x_j) = \frac{u_L(e^{\beta m} - e^{\beta j}) + u_R(e^{\beta j} - 1)}{e^{\beta m} - 1} \quad (11)$$

and

$$u'(x_j) = \frac{\beta m e^{\beta j}}{e^{\beta m} - 1} (u_R - u_L) \quad (12)$$

We see that (8) and (9) are, respectively, extremely similar to (11) and (12). The only difference is that the role of  $e^\beta$  in (11) and (12) is assumed by  $\lambda$  in (8) and (9). For an analysis comparing  $\lambda$  and  $e^\beta$ , the interested reader is referred to [3]. The most interesting result that that is proven in [3] is that the solution collocation solution is oscillation-free. However, for Péclet numbers of even modest size, the collocation solution suffers from significant smearing, as illustrated in Figure 1.

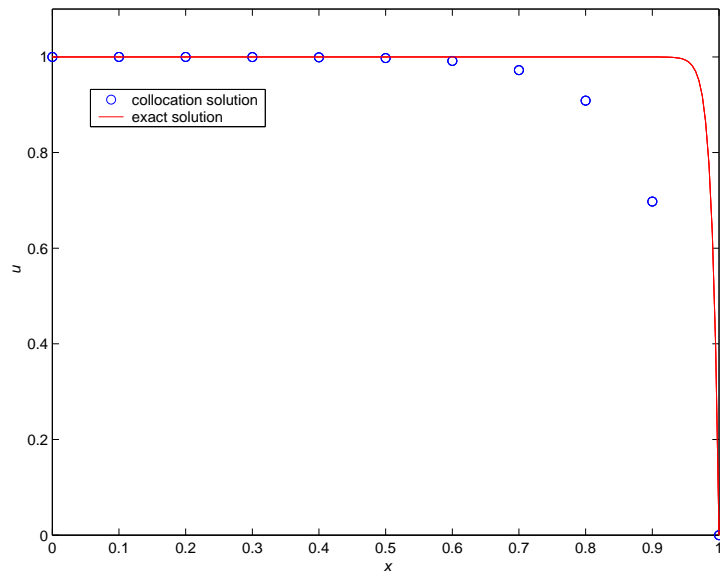


Figure 1. Collocation and exact solutions for  $m = 10$ ,  $\beta = 10$ ,  $u_L = 1$ ,  $u_R = 0$ , with no upstream weighting.

### 3. OPTIMAL UPSTREAM WEIGHTING

In this section we report on results published in [4]. We discuss upstream weighting (originally introduced in [1] for the context of collocation) and its optimal implementation. The advantage of optimal upstream weighting is that we may obtain extremely accurate collocation solutions while still avoiding unwanted oscillations.

When implementing upstreaming, we still enforce  $E(x) = 0$  (see (4)) for each of our  $2m$  equations and we still evaluate  $\frac{d^2\hat{u}}{dx^2}$  in (4) at the Gaussian points  $\eta = \pm\frac{1}{\sqrt{12}}$ . However, we evaluate  $\frac{d\hat{u}}{dx}$  at the points  $\eta = \pm\frac{1}{\sqrt{12}} - \zeta$ , where  $\zeta \geq 0$  controls how much upstreaming occurs. Because the support of each basis function  $f_j$  or  $g_j$  (see (5) and (6)) is the interval  $[-\frac{1}{2}, \frac{1}{2}]$ , it is clear that  $\zeta$  must lie in the interval  $[0, \frac{1}{2} - \frac{1}{\sqrt{12}}]$ .

It is straightforward to see that collocating in this manner leads to a matrix equation with the repeated computational molecule

$$\begin{bmatrix} M_{11} & M_{12} & -M_{11} & M_{14} \\ M_{21} & M_{22} & -M_{21} & M_{24} \end{bmatrix} \begin{bmatrix} q_j \\ r_j \\ q_{j+1} \\ r_{j+1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad (13)$$

$j = 0, 1, 2, \dots, m-1$ . The entries of the matrix are

$$M_{11} = \frac{2\sqrt{3}D}{h^2} + \frac{v}{h} (6\zeta^2 + 2\sqrt{3}\zeta - 1)$$

$$M_{21} = -\frac{2\sqrt{3}D}{h^2} + \frac{v}{h} (6\zeta^2 - 2\sqrt{3}\zeta - 1)$$

$$M_{12} = \frac{D}{h} (1 + \sqrt{3}) + v \left( \frac{\sqrt{3}}{6} + \zeta + \sqrt{3}\zeta + 3\zeta^2 \right)$$

$$M_{22} = \frac{D}{h} (1 - \sqrt{3}) + v \left( -\frac{\sqrt{3}}{6} + \zeta - \sqrt{3}\zeta + 3\zeta^2 \right)$$

$$M_{14} = \frac{D}{h} (-1 + \sqrt{3}) + v \left( -\frac{\sqrt{3}}{6} - \zeta + \sqrt{3}\zeta + 3\zeta^2 \right)$$

$$M_{24} = -\frac{D}{h} (1 + \sqrt{3}) + v \left( \frac{\sqrt{3}}{6} - \zeta - \sqrt{3}\zeta + 3\zeta^2 \right).$$

It is easy to see that (13) reduces to (7) when  $\zeta = 0$ , i.e., when there is no upstream weighting.

As in the previous section, the collocation solution is governed by the parameter  $\lambda$ . However, now that we have implemented upstream weighting, the value of  $\lambda$  now depends on  $\zeta$ :

$$\lambda = \frac{\beta^2 + 6\beta + 12 + 6\beta\zeta(4 + \beta + \beta\zeta)}{\beta^2 - 6\beta + 12 + 6\beta\zeta(4 - \beta + \beta\zeta)}, \quad (14)$$

which reduces to (10) when  $\zeta = 0$ .

The solution of (13) is given by

$$q_j = \frac{u_L(\lambda^m - \lambda^j) + u_R(\lambda^j - 1)}{\lambda^m - 1} \quad (15)$$

and

$$r_j = \frac{\rho\lambda^j}{\lambda^m - 1}(u_R - u_L), \quad (16)$$

where

$$\rho = \frac{2\beta m(1 + \beta\zeta)}{\beta^2\zeta^2 + 4\beta\zeta + 2}. \quad (17)$$

Note that (15) and (8) are identical (except for the change in the definition of  $\lambda$ ) and that (16) reduces to (9) when  $\zeta = 0$ .

Now that  $\lambda$  is defined as in (14), it is possible for the collocation solution to oscillate. However, this happens for only a very narrow range of parameter values. Specifically, oscillations occur only when  $\beta > 6 + 4\sqrt{3}$  and

$$\frac{1}{2} - \frac{2}{\beta} - \frac{\sqrt{\beta^2 - 12\beta + 24}}{\sqrt{12}\beta} < \zeta \leq \frac{1}{2} - \frac{1}{\sqrt{12}}.$$

This region of the  $\beta$ - $\zeta$  plane may be easily observed. It is the area above the thick dashed curve and below the light dotted line in Figure 2.

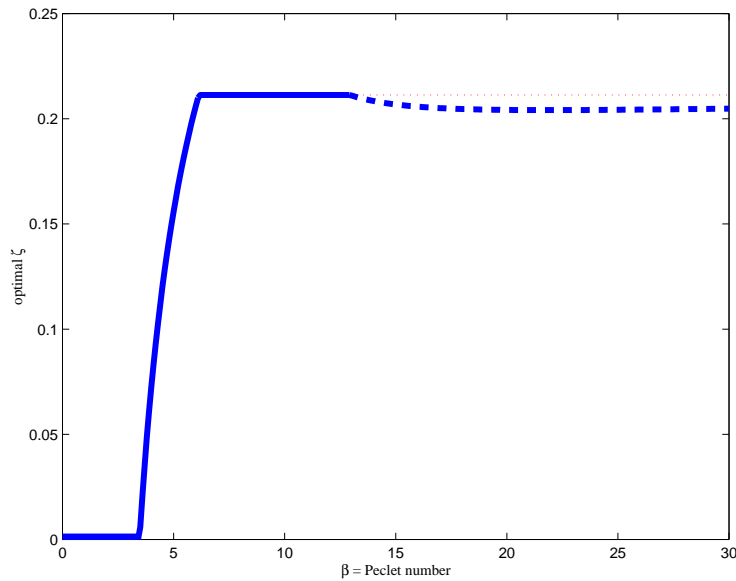


Figure 2. Optimal  $\zeta$  as a function of Péclet number  $\beta$ .

We have determined an algorithm for how to choose the value of the upstream parameter  $\zeta$  as a function of the Péclet number  $\beta$  so as to minimize the difference between (15) and (11) (see [4] for details). This algorithm for the optimal choice of the value of  $\zeta$  is given in Table 1 and depicted in Figure 2. (In Table 1, the parameter  $\epsilon$  appears. It is a small positive number, the purpose of which is to ensure that the collocation solution does not oscillate. In all pertinent numerical experiments reported herein, its value is set to  $10^{-6}$ .)

Table 1  
Optimal  $\zeta$  as a function of  $\beta$ .

| $\beta$ interval  | approx $\beta$ interval | optimal $\zeta$   |
|---|-------------------------|---|
| $(0, 2\sqrt{3}]$  | $(0, 3.46410]$          | 0   |
| $[2\sqrt{3}, \sqrt{3} + 2^{-1/2}(3^{3/4} + 3^{5/4})]$     | $[3.46410, 6.13572]$    | $\frac{\sqrt{6\beta^2 - 36} - 6}{6\beta}$   |
| $[\sqrt{3} + 2^{-1/2}(3^{3/4} + 3^{5/4}), 6 + 4\sqrt{3}]$ | $[6.13572, 12.9282]$    | $\frac{1}{2} - \frac{1}{\sqrt{12}}$   |
| $[6 + 4\sqrt{3}, \infty)$                                 | $[12.9282, \infty)$     | $\frac{1}{2} - \frac{2}{\beta} - \frac{\sqrt{\beta^2 - 12\beta + 24}}{\sqrt{12}\beta} - \epsilon$ |

$0 < \epsilon \ll 1.$

We depict the efficacy of optimal upstream weighting in Figure 3. We see that the collocation solution without upstreaming suffers from significant smearing while the optimal upstream collocation solution is visually indistinguishable from the exact solution.

#### 4. PIECEWISE CONSTANT $v$

In this section, we permit the convection coefficient  $v$  to vary while keeping  $S(x) \equiv 0$  and the diffusion coefficient  $D$  constant. We study the case where  $v$  is constant over each subinterval  $[x_j, x_{j+1}]$ ,  $j = 1, 2, \dots, m$ .

Since  $v$  is piecewise-constant over each subinterval, it is logical to describe  $v$  as

$$v = \begin{cases} v_1, & 0 = y_0 \leq x < y_1 \\ v_2, & y_1 < x < y_2 \\ \vdots & \\ v_p, & y_{p-1} < x \leq y_p = 1, \end{cases} \quad (18)$$

where each point in our domain at which  $v$  changes value, i.e.  $y_1, y_2, \dots, y_{p-1}$ , is a mesh point  $x_j$ ,  $j = 1, 2, \dots, m - 1$ . Here  $p$  represents the number of distinct pieces of  $v$ . It is helpful to introduce a mapping from the indices of the set of breakpoints of  $v$  to those of the set of mesh points  $x_j$ :

$$\phi(k) = j,$$

where  $y_k = x_j$ .

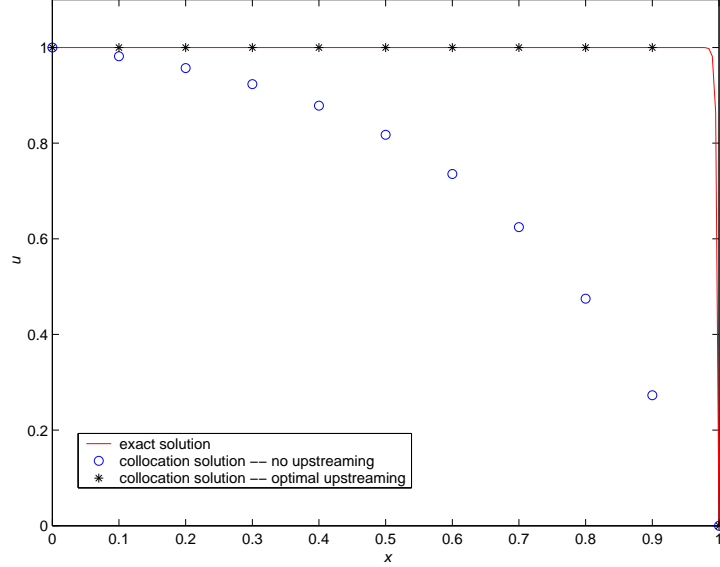


Figure 3. Collocation and exact solutions for  $m = 10$ ,  $\beta = 40$ ,  $u_L = 1$ , and  $u_R = 0$ .

Since  $v$  has  $p$  distinct pieces, so does the Péclet number  $\beta$ :

$$\beta = \begin{cases} \beta_1, & 0 = y_0 \leq x < y_1 \\ \beta_2, & y_1 < x < y_2 \\ \vdots & \\ \beta_p, & y_{p-1} < x \leq y_p = 1, \end{cases} \quad (19)$$

In this case, the exact solution of the BVP (1), (2) is

$$u(x) = \begin{cases} u^{(1)}(x), & 0 = y_0 \leq x \leq y_1 \\ u^{(2)}(x), & y_1 \leq x \leq y_2 \\ \vdots & \\ u^{(p)}(x), & y_{p-1} \leq x \leq y_p = 1, \end{cases} \quad (20)$$

where

$$u^{(k)}(x) = \frac{(e^{\beta_k m x} - e^{\beta_k \phi(k-1)})u\left(\frac{\phi(k)}{m}\right) - (e^{\beta_k m x} - e^{\beta_k \phi(k)})u\left(\frac{\phi(k-1)}{m}\right)}{e^{\beta_k \phi(k)} - e^{\beta_k \phi(k-1)}}, \quad (21)$$

$k = 1, 2, \dots, p$ . It is clear that the function  $u(x)$  is continuous on  $\mathcal{I} = [0, 1]$ .

At this point, while (20) and (21) are indisputably correct, we do not yet know the values of  $u\left(\frac{\phi(k)}{m}\right)$ ,  $k = 1, 2, \dots, p-1$ . To uniquely determine these, we enforce the continuity of  $\frac{d}{dx}u(x)$  at the breakpoints  $y_1, y_2, \dots, y_{p-1}$ . (We note that the resulting discontinuities in  $\frac{d^2}{dx^2}u(x)$  are of the same type as those of  $v$ , and thus the DE (1) achieves balance in where it fails to be continuous.) This leads to a tridiagonal system of linear algebraic equations which can be represented by

$$-c_k u\left(\frac{\phi(k-1)}{m}\right) + (b_k + c_k)u\left(\frac{\phi(k)}{m}\right) - b_k u\left(\frac{\phi(k+1)}{m}\right) = 0, \quad (22)$$

$k = 1, 2, \dots, p-1$ . Here

$$c_k = \frac{\beta_k e^{\beta_k \phi(k)}}{e^{\beta_k \phi(k)} - e^{\beta_k \phi(k-1)}} \quad (23)$$

and

$$b_k = \frac{\beta_{k+1} e^{\beta_{k+1} \phi(k)}}{e^{\beta_{k+1} \phi(k+1)} - e^{\beta_{k+1} \phi(k)}} \quad (24)$$

Since  $u\left(\frac{\phi(0)}{m}\right)$  and  $u\left(\frac{\phi(p)}{m}\right)$  are known from the boundary conditions, we may solve the tridiagonal system (22) to obtain the values of  $u(x)$  at the breakpoints  $y_1, y_2, \dots, y_{p-1}$ , thus fully defining the solution function  $u(x)$  in (20).

We will want to compare the exact solution to the collocation solution in this case. For the former, consider (21) at the mesh point  $x_j$ . If  $j = \phi(k)$  or  $j = \phi(k-1)$ , then it is obvious that  $x_j$  satisfies (21). If, however,  $x_{\phi(k-1)} < x_j < x_{\phi(k)}$ , then we see that

$$u^{(k)}(x_j) = \frac{(e^{\beta_k j} - e^{\beta_k \phi(k-1)})u\left(\frac{\phi(k)}{m}\right) - (e^{\beta_k j} - e^{\beta_k \phi(k)})u\left(\frac{\phi(k-1)}{m}\right)}{e^{\beta_k \phi(k)} - e^{\beta_k \phi(k-1)}}. \quad (25)$$

and

$$\frac{du^{(k)}}{dx}(x_j) = \frac{\beta_k m \left( u\left(\frac{\phi(k)}{m}\right) - u\left(\frac{\phi(k-1)}{m}\right) \right)}{e^{\beta_k \phi(k)} - e^{\beta_k \phi(k-1)}} e^{\beta_k j} \quad (26)$$

The solution of the collocation discretization under these conditions corresponds very closely to that of the exact solution (20). Each  $\beta_k$  has a corresponding  $\lambda_k$  (see (14)) and  $\rho_k$  (see (17)). A tridiagonal system (analogous to (22)) must be solved:

$$-c_k^* q_{\phi(k-1)} + (b_k^* + c_k^*) q_{\phi(k)} - b_k^* q_{\phi(k+1)} = 0, \quad (27)$$

$k = 1, 2, \dots, p-1$ , where

$$c_k^* = \frac{\rho_k \lambda_k^{\phi(k)}}{\lambda_k^{\phi(k)} - \lambda_k^{\phi(k-1)}} \quad (28)$$

(compare to (23)) and

$$b_k^* = \frac{\rho_{k+1} \lambda_{k+1}^{\phi(k)}}{\lambda_{k+1}^{\phi(k+1)} - \lambda_{k+1}^{\phi(k)}} \quad (29)$$

(compare to (24)).

The solution to the tridiagonal system (27) gives values for  $q_{\phi(k)}$ ,  $k = 1, 2, \dots, p-1$ . From these, we can compute the rest of the  $q_j$ 's. For each  $q_j$  yet to be determined, locate  $j$  (i.e., find  $k$ ) such that  $\phi(k-1) < j < \phi(k)$ . Then

$$q_j = \frac{(\lambda_k^j - \lambda_k^{\phi(k-1)})q_{\phi(k)} - (\lambda_k^j - \lambda_k^{\phi(k)})q_{\phi(k-1)}}{\lambda_k^{\phi(k)} - \lambda_k^{\phi(k-1)}} \quad (30)$$

(compare to (25)) and

$$r_j = \frac{\rho_k(q_{\phi(k)} - q_{\phi(k-1)})}{\lambda_k^{\phi(k)} - \lambda_k^{\phi(k-1)}} \lambda_k^j$$

(compare to (26)).

Now that we have the formulas (25) and (30), we want to select the upstream parameters  $\zeta_k$ ,  $k = 1, 2, \dots, p$ , in such a way as to have (25) and (30) approximately equal. This is achieved by (approximately) equating the two tridiagonal systems (22) and (27), i.e., by enforcing

$$c_k \approx c_k^*$$

(see (23) and (28)) and

$$b_k \approx b_k^*$$

(see (24) and (29)), which, respectively, lead to

$$f(\gamma_k) \approx \frac{1 - \lambda_k^{-Q_k}}{1 - e^{-\beta_k Q_k}} \quad (31)$$

and

$$f(\gamma_{k+1}) \approx \frac{\lambda_{k+1}^{Q_{k+1}} - 1}{e^{\beta_{k+1} Q_{k+1}} - 1}, \quad (32)$$

where

$$f(x) \equiv \frac{2(1+x)}{x^2 + 4x + 2} \quad (33)$$

(see (17)),

$$\gamma_k = \beta_k \zeta_k,$$

and

$$Q_k = \phi(k) - \phi(k-1),$$

$$k = 1, 2, \dots, p-1.$$

If  $\beta$  is small, we know from (14) and [3] that  $\lambda \approx e^\beta$  is achieved by choosing  $\zeta \approx 0$ . Note that if  $\zeta \approx 0$ , then  $\gamma \approx 0$  and that  $f(0) \approx 0$ . Thus, if  $\beta$  is small, then (31) and (32) are satisfied by choosing  $\zeta \approx 0$ .

On the other hand, if  $\beta$  is large, then (31) and (32) reduce to

$$f(\gamma_k) \approx 1 - \lambda_k^{-Q_k} \quad (34)$$

and

$$f(\gamma_{k+1}) \approx 0. \quad (35)$$

Examination of (33) and (14) reveals that (34) is satisfied by choosing  $\zeta \approx 0$ .

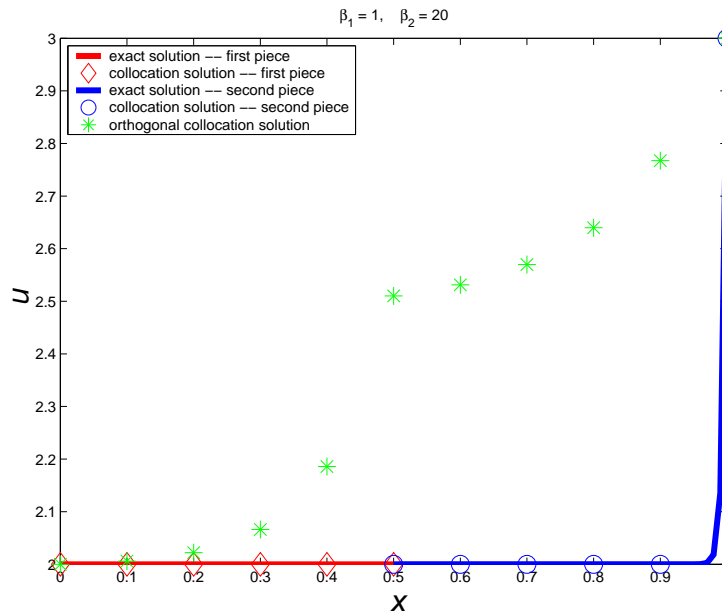


Figure 4.  $\zeta_1 = 0$ ,  $\zeta_2 = 0.204210$ .

Since each piece of the exact solution is monotone, we wish to avoid oscillatory collocation solutions. We thus choose  $\zeta_{k+1}$  in (35) to be as large as possible while avoiding the small region of the  $\beta$ - $\zeta$  plane where oscillatory solutions exist (see Figure 2 and Table 1).

We now have an algorithm for how to choose each  $\zeta_k$  in the case where the corresponding  $\beta_k$  is large or small, the efficacy of each we illustrate (for the case  $p = 2$ ) in Figures 4 and 5. In these figures, the Péclet numbers  $\beta$  and corresponding upstream parameters  $\zeta$  are given. In each figure, we show the exact (continuous) solution to the given BVP, the upstream collocation solution with the amount of upstreaming given by our algorithm, and the collocation solution with no upstreaming (orthogonal collocation). In each case, our algorithm gives collocation results which are visually indistinguishable from the exact solution, while the collocation method without upstreaming gives poor results indeed.

## 5. SUMMARY AND CONCLUSIONS

Our recent work concerning analytical formulas for the solution of the Hermite collocation discretization of the one-dimensional convection-diffusion equation has been summarized in this paper. Because this work is intended to be a survey, the exposition here is lacking in detail, which the interested reader may find in [3], [4], and [5]. The unifying concept of these works is that if one applies “optimal upstream weighting” to convection-dominated problems, then one obtains numerical solutions that do not suffer from the “smearing” effect, are non-oscillatory, and are visually indistinguishable from the corresponding exact solution.

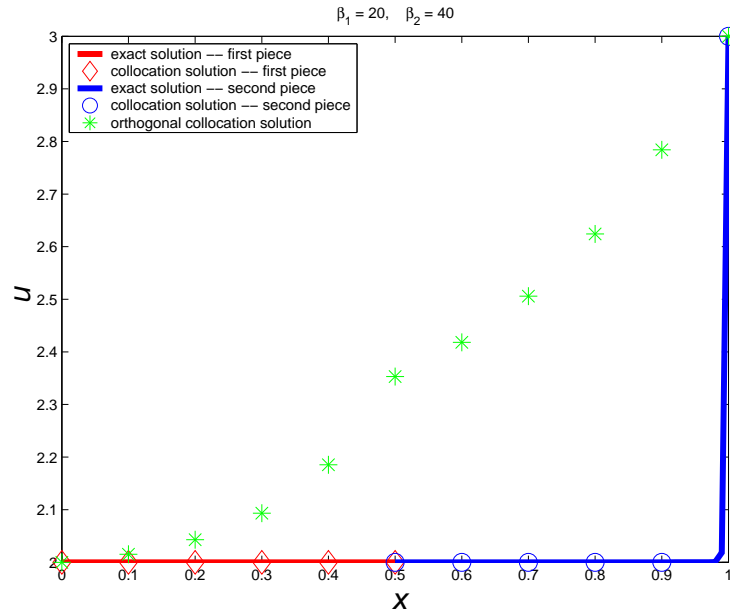


Figure 5.  $\zeta_1 = 0, \zeta_2 = 0.205902$ .

## REFERENCES

1. M.B. Allen, How Upstream Collocation Works, *Int. J. Num. Meth. Eng.*, 19 (1983) 1753–1763.
2. S.H. Brill, Analytical Solution of Hermite Collocation Discretization of Self-Adjoint Ordinary Differential Equations, *International Journal of Differential Equations and Applications*, 6 (2002) 1–18.
3. S.H. Brill, Analytical Solution of Hermite Collocation Discretization of the Steady-State Convection-Diffusion Equation, *International Journal of Differential Equations and Applications*, 4 (2002) 141–155.
4. S.H. Brill, Optimal Upstream Collocation Solution of the One-Dimensional Steady-State Convection-Diffusion Equation, *International Journal of Applied Mathematics*, 10 (2002) 197–215.
5. S.H. Brill, Analytical Upstream Collocation Solution of a Forced One-Dimensional Steady-State Convection-Diffusion Equation, *International Journal of Differential Equations and Applications*, 7 (2003) 69–99.
6. P.M. Prenter, *Splines and Variational Methods*, John Wiley & Sons, New York, 1975.