

OPTIMAL UPSTREAM COLLOCATION SOLUTION OF THE ONE-DIMENSIONAL STEADY-STATE CONVECTION-DIFFUSION EQUATION

Stephen H. Brill
Department of Mathematics
Boise State University
Boise, Idaho, U.S.A.
email: `brill@math.boisestate.edu`

Abstract

We give herein analytical formulas for the solution of the Hermite collocation discretization of the unforced steady-state convection-diffusion equation in one spatial dimension and with constant coefficients, defined on a uniform mesh, with Dirichlet boundary conditions. The accuracy of the method is enhanced by employing “upstream weighting” of the convective term in an optimal way, avoiding both “smearing” and unwanted oscillations, particularly for large Péclet numbers. Computational examples illustrate the efficacy of using optimal upstream weighting.

1 Introduction

It is well known that the numerical solution of convection-diffusion differential equations (DEs) is a difficult task when convection is the dominant process. Numerical techniques often give rise to spurious oscillations that are not present in the continuous (i.e., not numerical/discrete) solution of the DE. To ameliorate these physically unmeaningful (and therefore undesirable) oscillations, the technique of upstream weighting is often used (Allen [1], Morton [3], Pinder and Shapiro [4], Shapiro and Pinder [6]). While upstreaming can eliminate the oscillations, it is often at the expense of “smearing” the sharp solution profile of the continuous solution of the DE.

In this work, we study the convection-diffusion equation

$$-D \frac{d^2 u}{dx^2} + v \frac{du}{dx} = 0 \tag{1}$$

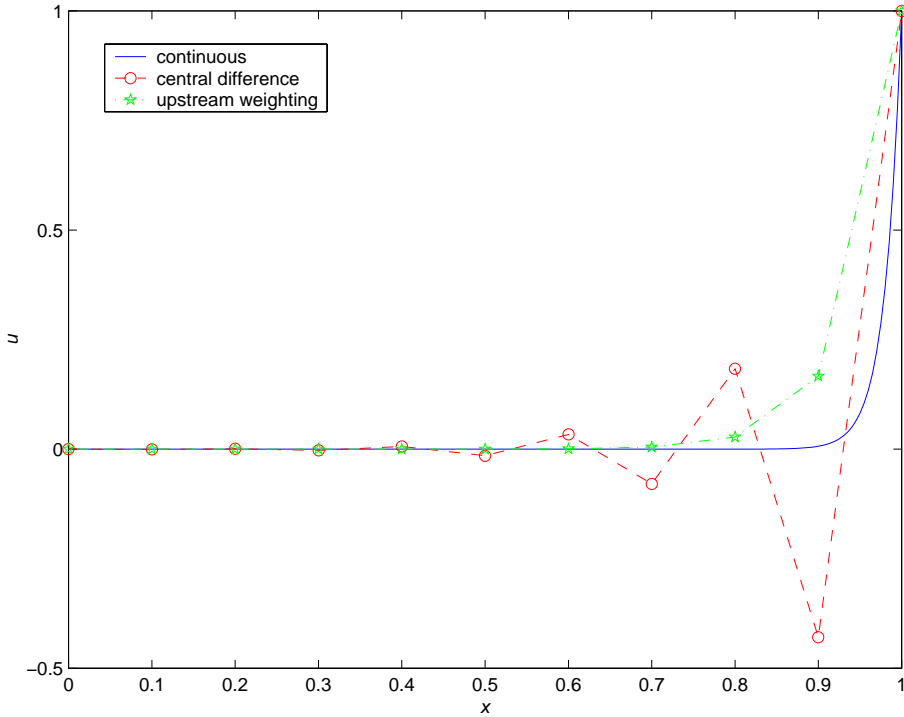


Figure 1: Continuous, central difference, and upstream weighting solutions of (1),(2) with $\beta = 5$ and $m = 10$.

with Dirichlet boundary conditions, defined on the interval $[0, 1]$. The convection coefficient v and diffusion coefficient D are both positive constants. For the purposes of numerical solution, we subdivide the domain $[0, 1]$ into m equal subintervals and thus seek to solve (1) at the nodes $x_j = jh$, $j = 0, 1, 2, \dots, m$, where $h = 1/m$.

The following example, depicted in Figure 1, is discussed in detail in Morton [3] and illustrates the issues that this paper tackles. Suppose along with (1) we have the boundary conditions

$$\begin{aligned} u(0) &= 0 \\ u(1) &= 1. \end{aligned} \tag{2}$$

If we discretize (1) via standard central differences, then we obtain the solution

$$u_j^C = \frac{\lambda_C^j - 1}{\lambda_C^m - 1},$$

$j = 0, 1, 2, \dots, m$, where u_j^C is an approximation to the exact solution of the continuous problem, namely

$$u(x_j) = \frac{e^{\beta j} - 1}{e^{\beta m} - 1}. \tag{3}$$

The lumped parameter $\beta = \frac{hv}{D}$ is known as the Péclet number and

$$\lambda_C = \frac{2 + \beta}{2 - \beta}.$$

It is clear that if $\beta > 2$, then $\lambda_C < -1$. So in this case u_j^C and u_{j+1}^C have opposite signs; i.e. the central difference solution oscillates, which is qualitatively very different from the monotone exact solution (3) (see Figure 1).

We may eliminate these oscillations via upstream weighting. For the case of finite difference discretization, this is most easily accomplished by replacing the central difference approximation to $\frac{du}{dx}$ in (1) with a one-sided backward difference approximation. The resulting solution is now

$$u_j^U = \frac{\mu^j - 1}{\mu^m - 1},$$

where $\mu = 1 + \beta$. While this last solution is monotone, we have lost the sharp solution profile that is present in (3) (see Figure 1). This example clearly illustrates the limitations of the finite difference discretization to solve (1). We are thus motivated to investigate other methods of discretization.

Collocation discretization of the transient convection-diffusion equation has been studied in a variety of papers. Allen [1], using Taylor series analysis, explains why Hermite collocation can eliminate the unwanted oscillations provided upstream weighting is utilized but does not address simultaneously eliminating the “smearing” effect. In Pinder and Shapiro [4] and Shapiro and Pinder [6] the authors consider replacing the traditional cubic Hermite basis with quartic polynomials, successfully eliminating the oscillations but not the residual smearing. They also provide a Fourier analysis for their method. In a previous paper (Brill [2]), the present author derived analytical solutions for the Hermite collocation discretization of (1) without considering the effects of upstream weighting. Thus the present work may be viewed as an extension of the previous one.

This paper is organized as follows. We first describe Hermite collocation in the context of the DE (1), both with and without upstream weighting. We then provide the analytical solution of the matrix equation that arises from the discretization. Subsequently, we provide an analysis which compares the discrete collocation solution to the continuous solution. In particular, we will discuss how to select the upstream parameter ζ as a function of the Péclet number β in such a way as to eliminate spurious oscillations and minimize the error between the continuous and discrete solutions. We then provide several computational examples which illustrate the theory. A short section summarizing our results concludes the paper.

2 Hermite Collocation

The differential equation (1) is defined on the interval $[0, 1]$ with Dirichlet boundary conditions included. We partition the interval $[0, 1]$ into m uniform subintervals as described above, using the same notation.

The discretization proceeds by introducing a piecewise cubic Hermite interpolating polynomial

$$\hat{u}(x) = \sum_{j=0}^m [u_j f_j(x) + u'_j g_j(x)] \quad (4)$$

into the ODE (1), obtaining

$$-D \frac{d^2 \hat{u}}{dx^2} + v \frac{d\hat{u}}{dx} = E(x), \quad (5)$$

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

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

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

and

$$g_j(x) = \begin{cases} \frac{h}{8} (2\eta + 1)^2 (2\eta - 1), & x_{j-1} \leq x = x_j + \left(\eta - \frac{1}{2}\right) h \leq x_j \\ \frac{h}{8} (2\eta - 1)^2 (2\eta + 1), & x_j \leq x = x_j + \left(\eta + \frac{1}{2}\right) h \leq x_{j+1} \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

Note that \hat{u} in (4) 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 δ_{jk} is the Kronecker symbol.

It is clear that (5) 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, the traditional choice is to enforce that the error function $E(x)$ in (5) 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 local discretization error) location of the collocation points within each subinterval corresponds to the points of Gaussian quadrature (Prenter [5]), which in turn corresponds to choosing the collocation points as $\eta = \pm \frac{1}{\sqrt{12}}$ (see (6) and (7)) in each subinterval $[-\frac{1}{2}, \frac{1}{2}]$ (given in local η coordinates). In our work, this choice will correspond to an absence of upstream weighting. However, large Péclet numbers violate the smoothness conditions stipulated in Prenter [5]; thus the Gaussian points are not, in general, optimal for our DE. As we shall see, use of the Gaussian points (i.e. no upstream weighting) is optimal only for a very small range of values of β .

Upstream weighting is implemented in the following manner, which was introduced by Allen [1]. As we mentioned above, we have $2m$ equations in $2m$ unknowns, where the $2m$ equations are traditionally generated by forcing $E(x) = 0$ in (5) at two collocation points in each of the m subintervals $[x_j, x_{j+1}]$, $j = 0, 1, 2, \dots, m-1$. When implementing upstreaming, we still enforce $E(x) = 0$ for each of our $2m$

equations and we still evaluate $\frac{d^2\widehat{u}}{dx^2}$ in (5) at the Gaussian points $\eta = \pm\frac{1}{\sqrt{12}}$. However, we evaluate $\frac{d\widehat{u}}{dx}$ at the points $\eta = \pm\frac{1}{\sqrt{12}} - \zeta$, where $\zeta > 0$ controls how much upstreaming occurs. Because the support of each basis function f_j or g_j (see (6) and (7)) is the interval $[-\frac{1}{2}, \frac{1}{2}]$, it is clear that ζ must lie in the interval $[0, \frac{1}{2} - \frac{1}{\sqrt{12}}]$.

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

$$\begin{bmatrix} M_{11} & M_{12} & -M_{11} & M_{14} \\ M_{21} & M_{22} & -M_{21} & M_{24} \end{bmatrix} \begin{bmatrix} q_j \\ r_j \\ q_{j+1} \\ r_{j+1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad (8)$$

$j = 0, 1, 2, \dots, m-1$. Here $q_j = u_j$ and $r_j = u'_j$, $j = 0, 1, 2, \dots, m$. Note that the matrix equation represented by (8) is a system of $2m$ equations in $2(m+1)$ unknowns. The entries of the matrix are

$$\begin{aligned} M_{11} &= \frac{2\sqrt{3}D}{h^2} + \frac{v}{h} (6\zeta^2 + 2\sqrt{3}\zeta - 1) \\ M_{21} &= -\frac{2\sqrt{3}D}{h^2} + \frac{v}{h} (6\zeta^2 - 2\sqrt{3}\zeta - 1) \\ M_{12} &= \frac{D}{h}(1 + \sqrt{3}) + v \left(\frac{\sqrt{3}}{6} + \zeta + \sqrt{3}\zeta + 3\zeta^2 \right) \\ M_{22} &= \frac{D}{h}(1 - \sqrt{3}) + v \left(-\frac{\sqrt{3}}{6} + \zeta - \sqrt{3}\zeta + 3\zeta^2 \right) \\ M_{14} &= \frac{D}{h}(-1 + \sqrt{3}) + v \left(-\frac{\sqrt{3}}{6} - \zeta + \sqrt{3}\zeta + 3\zeta^2 \right) \\ M_{24} &= -\frac{D}{h}(1 + \sqrt{3}) + v \left(\frac{\sqrt{3}}{6} - \zeta - \sqrt{3}\zeta + 3\zeta^2 \right) \end{aligned}$$

which reduce to those given in Brill [2] for the case of $\zeta = 0$.

3 Analytical Solution of Upstream Collocation

We begin by deriving an equivalent but simpler way of expressing (8). Let

$$k_1 = [6h(D + hv\zeta) + \sqrt{3}h^2v(1 - 6\zeta^2)]D^{-2},$$

$$k_2 = [6h(D + hv\zeta) - \sqrt{3}h^2v(1 - 6\zeta^2)]D^{-2},$$

and

$$k_3 = \frac{h}{2D}.$$

Now, add k_1 times the first row of (8) to k_2 times the second row of (8). Also, separately, add the first row of (8) to the second row of (8) and multiply this sum by k_3 . We obtain

$$M \begin{bmatrix} q_j \\ r_j \\ q_{j+1} \\ r_{j+1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad (9)$$

where

$$M = \begin{bmatrix} 0 & \lambda_{\text{num}} & 0 & -\lambda_{\text{den}} \\ -\beta m(1 - 6\zeta^2) & 1 + \beta\zeta + 3\beta\zeta^2 & \beta m(1 - 6\zeta^2) & -1 - \beta\zeta + 3\beta\zeta^2 \end{bmatrix}.$$

Here λ_{num} and λ_{den} are, respectively, the numerator and denominator of (12).

We now give the solution of the matrix equation (9), which is, of course, also the solution of (8). The proof of this result is completely straightforward, though computationally tedious.

Theorem 3.1 *The general solution of (8) and (9) is*

$$q_j = c_1 + c_2 \lambda^j \quad (10)$$

$$r_j = \rho c_2 \lambda^j, \quad (11)$$

where c_1 and c_2 are constants determined by boundary conditions,

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

and

$$\rho = \frac{2\beta m(1 + \beta\zeta)}{\beta^2\zeta^2 + 4\beta\zeta + 2}. \quad (13)$$

We note that this result reduces, for the case $\zeta = 0$, to the analogous formulas in Brill [2].

Since we are given the Dirichlet boundary conditions $u(0) = q_0 = b_0$ and $u(1) = q_m = b_1$ we conclude from (10) with $j = 0, m$, that

$$c_2 = \frac{b_1 - b_0}{\lambda^m - 1}$$

and

$$c_1 = b_0 - \frac{b_1 - b_0}{\lambda^m - 1}.$$

Thus

$$q_j = b_0 + (b_1 - b_0) \frac{\lambda^j - 1}{\lambda^m - 1}. \quad (14)$$

The solution of the corresponding continuous problem is

$$u(x_j) = b_0 + (b_1 - b_0) \frac{e^{\beta j} - 1}{e^{\beta m} - 1}. \quad (15)$$

4 Oscillations

That oscillations are undesirable is illustrated in the following example. Suppose the solution u of (1) represents the concentration of a contaminant in water, with $u = 0$ representing pristine water and $u = 1$ representing pure contaminant. Then oscillations (like those in Figure 1) may produce values of u less than 0 and/or greater than 1, which are physically absurd.

With respect to (10) and (11), it is clear that our collocation solution will oscillate if and only if $\lambda < 0$ in (12). Since $\beta \in (0, \infty)$ and $\zeta \in [0, \frac{1}{2} - \frac{1}{\sqrt{12}}]$, it is clear that λ is negative if and only if its denominator is negative, which in turn leads to

$$\frac{1}{2} - \frac{2}{\beta} - \frac{\sqrt{\beta^2 - 12\beta + 24}}{\sqrt{12}\beta} < \zeta < \frac{1}{2} - \frac{2}{\beta} + \frac{\sqrt{\beta^2 - 12\beta + 24}}{\sqrt{12}\beta} \quad (16)$$

which, for positive β , is defined only if $0 < \beta \leq 6 - \sqrt{12}$ or $\beta \geq 6 + \sqrt{12}$. However, if $0 < \beta \leq 6 - \sqrt{12}$, then $\frac{1}{2} - \frac{2}{\beta} + \frac{\sqrt{\beta^2 - 12\beta + 24}}{\sqrt{12}\beta} < 0$, which implies $\zeta < 0$, which is not allowed. On the other hand, if $\beta \geq 6 + \sqrt{12}$, then $\frac{1}{2} - \frac{2}{\beta} + \frac{\sqrt{\beta^2 - 12\beta + 24}}{\sqrt{12}\beta} > \frac{1}{2} - \frac{1}{\sqrt{12}}$. Thus the condition (16) which leads to negative λ may be replaced by the more precise condition

$$\frac{1}{2} - \frac{2}{\beta} - \frac{\sqrt{\beta^2 - 12\beta + 24}}{\sqrt{12}\beta} < \zeta \leq \frac{1}{2} - \frac{1}{\sqrt{12}}, \quad (17)$$

which can occur only when $\beta > 6 + 4\sqrt{3}$. That is,

Theorem 4.1 *The upstream collocation solution (10) (11) of (1) will exhibit oscillations if and only if $\beta > 6 + 4\sqrt{3}$ and (17) holds.*

This situation is depicted in the small shaded region in the upper part of Figure 2.

5 Optimal Upstream Weighting

Since we are armed with analytical formulas for the solution of both the collocation discretization of (1) and the solution to the corresponding continuous problem, we investigate whether, for a given Péclet number β , we can find a corresponding value of ζ that will minimize the difference between the discrete collocation solution of (1) and its corresponding continuous solution.

We first require two lemmas.

Lemma 5.1 *Let $\lambda \geq 0$. Then $\lambda > 1$.*

Proof: Since we assume that $\lambda \geq 0$, we are restricting ourselves to the domain that is outlined in Figure 2, where we have used the convention that the domain includes the portions of its boundary depicted by solid lines but does not include yet gets arbitrarily close to those parts of the boundary depicted by dashed lines.

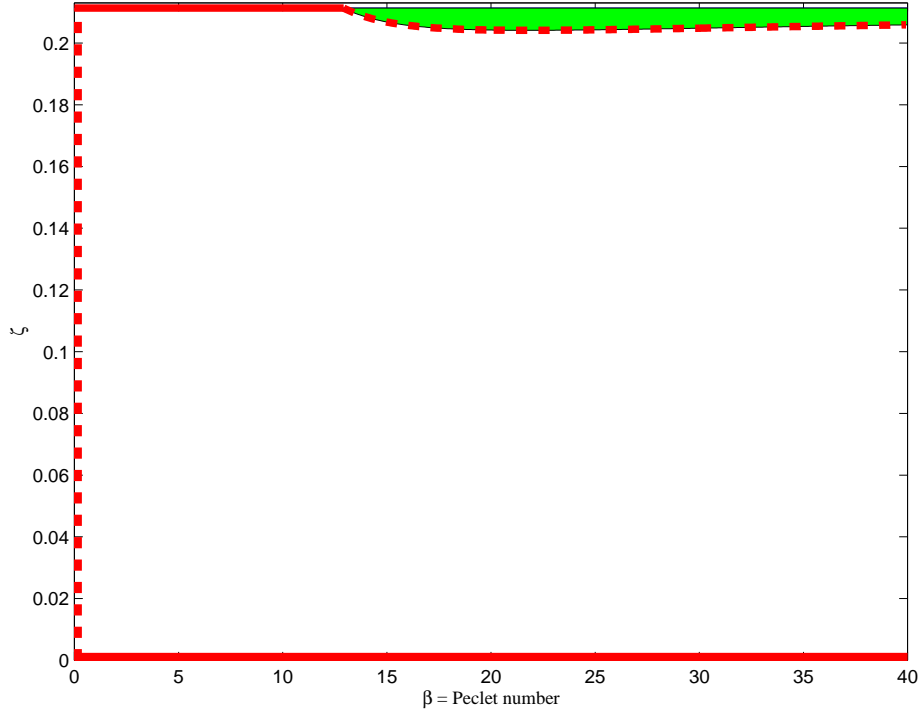


Figure 2: Shaded region indicates where $\lambda < 0$. Outlined region indicates where $\lambda > 0$.

First note that $\lambda = 1$ only when $\beta = 0$ (this follows directly from (12)). We now compute

$$\frac{\partial \lambda}{\partial \beta} = \frac{12(12 + 18\beta^2\zeta^2 + 24\beta\zeta - \beta^2)}{[\beta^2 - 6\beta + 12 + 6\beta\zeta(4 - \beta + \beta\zeta)]^2}$$

and note that $\frac{\partial \lambda}{\partial \beta} = 1 > 0$ at $\beta = 0$. Since λ is a continuous function of β and ζ on our domain, we may therefore conclude that $\lambda > 1$ at all points in our domain. Q.E.D.

Lemma 5.2 *Let $\lambda > 1$ and define*

$$\phi(\alpha) = \frac{\alpha\lambda^\alpha}{\lambda^\alpha - 1}.$$

Then ϕ is an increasing function for $\alpha \geq 1$.

Proof: Note

$$\frac{d\phi}{d\alpha} = \frac{\lambda^\alpha}{(\lambda^\alpha - 1)^2} (\lambda^\alpha - 1 - \alpha \log \lambda)$$

can equal zero only if

$$\lambda^\alpha = 1 + \alpha \log \lambda. \tag{18}$$

Letting $p = \alpha \log \lambda$ and taking the logarithm of both sides of (18) yields

$$p = \log(1 + p). \quad (19)$$

Note that our definition of p requires that p be positive. But the only solution of (19) is $p = 0$. Thus $\frac{d\phi}{d\alpha}$ can never equal 0. Furthermore, evaluation of $\frac{d\phi}{d\alpha}$ at any $\lambda > 1$ and any $\alpha \geq 1$ yields a positive result. Since $\frac{d\phi}{d\alpha}$ is a continuous function of λ and α , we may conclude that $\frac{d\phi}{d\alpha}$ is always positive. Q.E.D.

Throughout the rest of this section, we will assume that we are in the domain discussed in Lemma 5.1 and thus we have $\lambda > 1$.

The (discrete) collocation and continuous solutions of (1) with Dirichlet boundary conditions are given in (14) and (15), respectively. We thus wish to minimize the maximum value of $|E_j|$ with respect to ζ , where

$$E_j = \frac{e^{\beta j} - 1}{e^{\beta m} - 1} - \frac{\lambda^j - 1}{\lambda^m - 1}. \quad (20)$$

Note here that $j \in \{0, 1, 2, \dots, m\}$.

We first note that $|E_j|$ is certainly minimized at both $j = 0$ and $j = m$, since $E_j = 0$ at these values of j . Thus there exists a number $\bar{j} \in \{1, 2, \dots, m-1\}$ such that $|E_j|$ is maximized at \bar{j} . We now find the the value of ζ that minimizes $|E_{\bar{j}}|$.

By the chain rule for differentiation, we have

$$\frac{\partial E_{\bar{j}}}{\partial \zeta} = \frac{\partial E_{\bar{j}}}{\partial \lambda} \frac{\partial \lambda}{\partial \zeta}.$$

Now,

$$\frac{\partial E_{\bar{j}}}{\partial \lambda} = \frac{m\lambda^m(\lambda^{\bar{j}} - 1) - \bar{j}\lambda^{\bar{j}}(\lambda^m - 1)}{\lambda(\lambda^m - 1)^2}$$

which can equal zero only if

$$\frac{\bar{j}\lambda^{\bar{j}}}{\lambda^{\bar{j}} - 1} = \frac{m\lambda^m}{\lambda^m - 1}. \quad (21)$$

However, since $\bar{j} \in \{1, 2, \dots, m-1\}$, Lemma 5.2 says that (21) cannot occur. Thus $\frac{\partial E_{\bar{j}}}{\partial \lambda}$ cannot equal zero and so $\frac{\partial E_{\bar{j}}}{\partial \zeta} = 0$ only if $\frac{\partial \lambda}{\partial \zeta} = 0$.

We now compute

$$\frac{\partial \lambda}{\partial \zeta} = \frac{-12\beta^2(12 + 6\beta^2\zeta^2 + 12\beta\zeta - \beta^2)}{[\beta^2 - 6\beta + 12 + 6\beta\zeta(4 - \beta + \beta\zeta)]^2}$$

which equals zero only if

$$\zeta = \frac{\sqrt{6\beta^2 - 36} - 6}{6\beta}. \quad (22)$$

To ensure that $\zeta \in [0, \frac{1}{2} - \frac{1}{\sqrt{12}}]$, we must restrict β in (22) to the interval $[2\sqrt{3}, \sqrt{3} + 2^{-1/2}(3^{3/4} + 3^{5/4})] \approx [3.46410, 6.13572]$.

At values of β outside this interval, $|E_{\bar{j}}|$ is minimized by taking ζ to be either zero or $\frac{1}{2} - \frac{1}{\sqrt{12}}$, i.e., the minimum and maximum permissible values of ζ . However,

if we want to eliminate any chance of oscillations, we must ensure that $\zeta < \frac{1}{2} - \frac{2}{\beta} - \frac{\sqrt{\beta^2 - 12\beta + 24}}{\sqrt{12}\beta}$. Thus, when $\beta > 6 + 4\sqrt{3}$, we select ζ to be

$$\zeta = \frac{1}{2} - \frac{2}{\beta} - \frac{\sqrt{\beta^2 - 12\beta + 24}}{\sqrt{12}\beta} - \epsilon, \quad (23)$$

where ϵ is a small positive number large enough to ensure that the computed value of the right side of (23) is indeed less than $\frac{1}{2} - \frac{2}{\beta} - \frac{\sqrt{\beta^2 - 12\beta + 24}}{\sqrt{12}\beta}$.

We thus have an algorithm for choosing ζ optimally:

Theorem 5.3 *Let ϵ be a small positive number. Assuming that oscillations in the collocation solution are unacceptable, the value of ζ (as a function of β) that minimizes the maximum difference between the discrete collocation solution of (1) and the corresponding continuous solution (both with given Dirichlet boundary conditions) is given in Table 1 (and depicted in Figure 3).*

Table 1: Optimal ζ as a function of β .

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

6 Numerical Experiments

In this section we give the results of numerical experiments that mirror the theoretical conclusions given in Theorem 5.3, Table 1, and Figure 3. Results are reported for the case $m = 10$, but qualitatively similar results hold for all values of m we tested. We used

$$\beta = 0.1, 1, 2, 3, \dots, 12, 13, 15, 20, 50, 100, 200, 300, 500, 1000. \quad (24)$$

and

$$\zeta = 0, 0.005, 0.01, \dots, 0.205, 0.21, 0.211, 0.2113, 0.21132, 0.211324, \frac{1}{2} - \frac{1}{\sqrt{12}}. \quad (25)$$

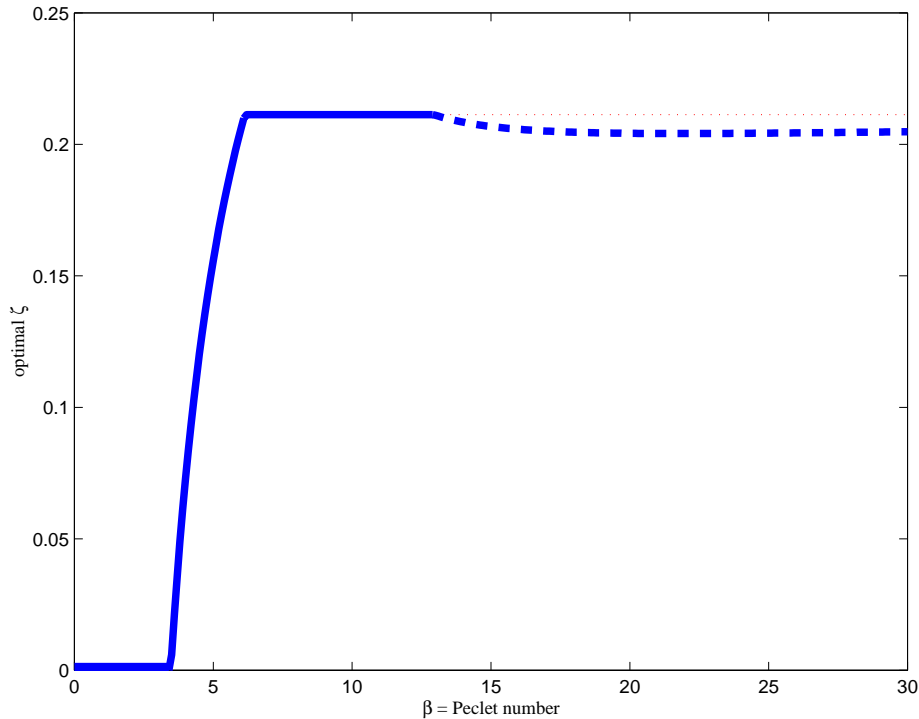


Figure 3: Optimal ζ as a function of Péclet number β

For each value of β , we used all given values of ζ which would not result in an oscillatory solution. For each pair (β, ζ) , we computed the values $|E_j|$, $j = 1, 2, \dots, m - 1$ (see (20)), and then selected the largest one. The results are depicted in Figure 4, where for each value of β (from among the values in (24)) we have plotted the value of ζ (from among the values in (25)) that yields the smallest maximum value of $|E_j|$. These data points are then superimposed over the theoretical optimal ζ curve from Figure 3. We plainly see excellent agreement between the numerical results and the theoretical predictions.

Finally, we examine how well optimal upstream collocation removes the “smearing” effect initially illustrated in Figure 1. (We noted in Brill [2] that, for fairly modest Péclet numbers, there was significant smearing in the collocation solution of (1) without upstreaming.) The results are depicted in Figure 5, where for Péclet numbers 2, 5, 10, and 40, we plotted the continuous, optimal upstream collocation, and non-upstream collocation solutions for $m = 10$ and the boundary conditions $u(0) = 1$, $u(1) = 0$. When applicable, we used $\epsilon = 10^{-6}$ (see Theorem 5.3). For very modest Péclet numbers (e.g., $\beta = 2$), we see that all three solutions are visually indistinguishable (in fact, since $\zeta = 0$ is optimal for this case, the two collocation solutions are identical). For $\beta = 5$, we see some smearing of both collocation solutions, but about only half as much in the optimal case as compared to the $\zeta = 0$ case. For the Péclet numbers 10 and 40, we see significant smearing of the col-

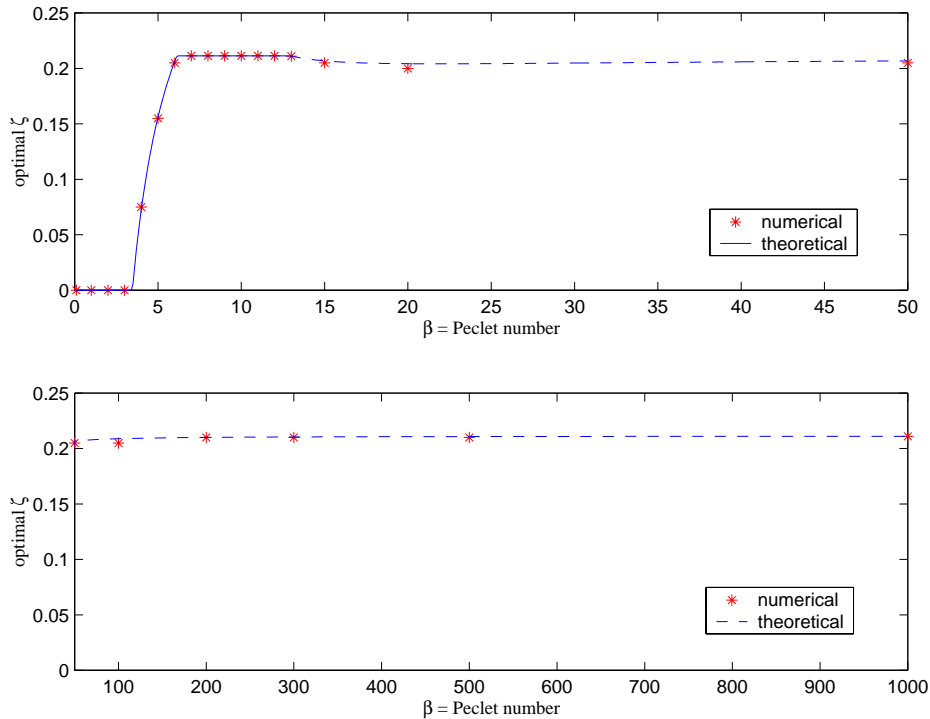


Figure 4: Theoretical and numerical values of optimal ζ .

location solution without upstreaming, but the continuous and optimal upstream collocation solutions are visually indistinguishable.

In Figure 6 we examine how profoundly optimal upstream collocation outperforms collocation without upstreaming. We define improvement as

$$\text{improvement} = \frac{\text{maximum error without upstreaming}}{\text{maximum error with optimal upstreaming}}$$

and compute this ratio for β between 0.5 and 10,000. (ϵ in Theorem 5.3 is 10^{-6} here). Although the graph in Figure 6 is for $m = 10$, this curve is visually indistinguishable when compared to those curves for $m = 100$ or $m = 1000$. We see that we obtain significant improvement for all values of the Péclet number other than those for which the optimal value of ζ is 0. In particular, for $\beta \geq 14$, we get improvement on the order of half a million.

7 Summary and Conclusions

In this paper, we give formulas for the analytical solution of the Hermite collocation discretization of the one-dimensional constant-coefficient unforced convection-diffusion equation, defined on a uniform mesh with Dirichlet boundary conditions. Upstream weighting is employed in the evaluation of the derivative of the convective

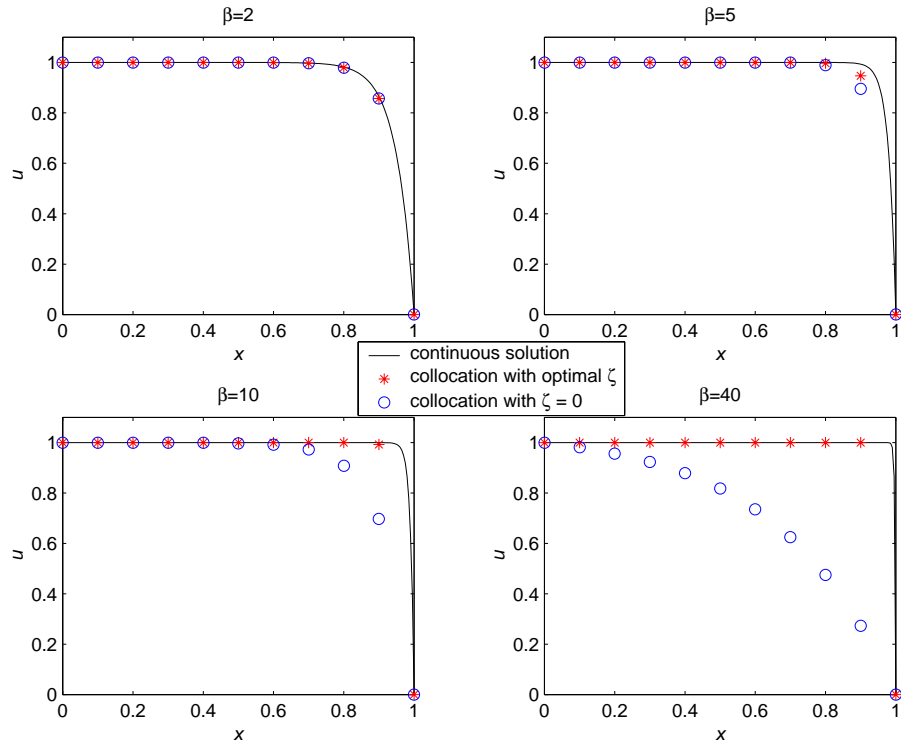


Figure 5: Comparison of continuous solution, optimal upstream collocation solution, and collocation solution with no upstreaming ($\zeta = 0$), for $\beta = 2, 5, 10, 40$.

term. The upstream parameter ζ may be chosen in an optimal manner to both eliminate physically absurd oscillations and to capture the sharp solution profile that exists in the exact solution of the corresponding continuous problem. Numerical experiments conform to and illustrate the theory derived herein.

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 the Steady-State Convection-Diffusion Equation, *International Journal of Differential Equations and Applications*, **4** (2002), 141–155.
- [3] K. W. Morton, *Numerical Solution of Convection-Diffusion Problems*, Chapman & Hall, London (1996).
- [4] G. F. Pinder and A. Shapiro, A New Collocation Method for the Solution of the Convection-Dominated Transport Equation, *Water Resources Research*, **15** (1979), 1177–1182.

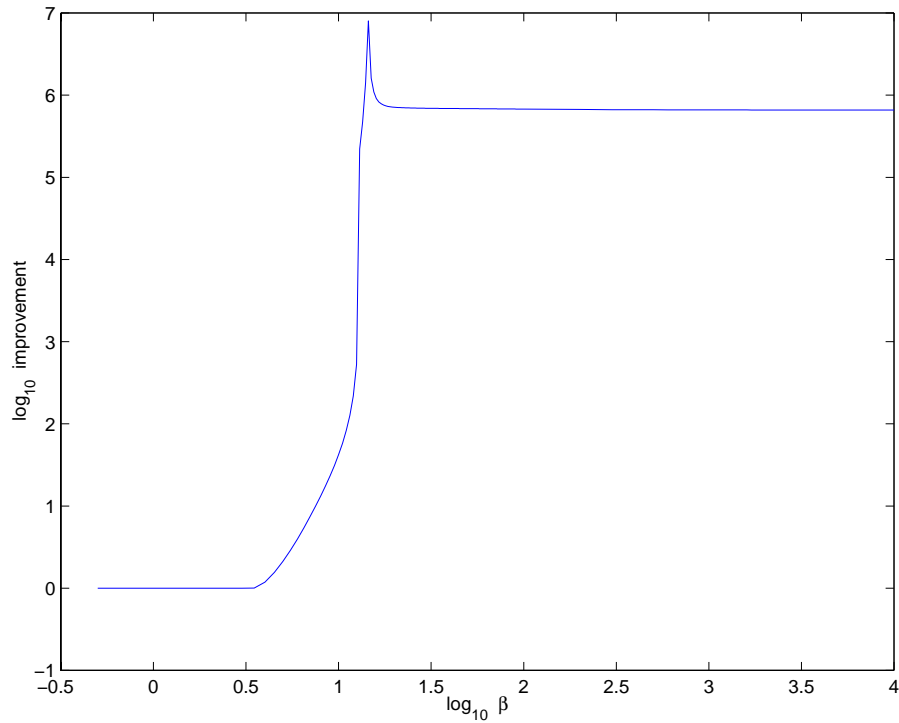


Figure 6: Improvement obtained by using optimal upstream collocation compared to using collocation with no upstreaming. β here varies from 0.5 to 10,000.

- [5] P. M. Prenter, *Splines and Variational Methods*, John Wiley & Sons, New York (1975).
- [6] A. Shapiro and G. F. Pinder, Analysis of an Upstream Weighted Collocation Approximation to the Transport Equation, *J. Comp. Phys.*, **39** (1981), 46–71.