Nonlinear Analysis: Modelling and Control, 2003, Vol. 8, No. 1, 33–54

Numerical Analysis of Time-Accurate Solution of Nonlinear Flow Models by Implicit Finite Differences [1]

S.K. Dey

Eastern Illinois University
Charleston, IL 61920, USA
cfskd@eiu.edu

Received: 27.02.2003
Accepted: 13.03.2003

Abstract. Implicit finite differences are often applied to solve flow models. A standard technique to solve these equations is Newton's method. If time step is too large although the difference equation could be computationally stable, Newton's method may fail, and PFI (perturbed functional iterations) may be fruitful. This is discussed in this article applying nonlinear D-mapping analysis.

Keywords: nonlinear flow, finite differences, perturbed functional iterations.

1 Mathematical Preliminaries

Let us consider a stationary iterative process given by:

$$x^{k+1} = Ax^k, \tag{1}$$

where $x = (x_1 x_2 \dots x_I)^T \in D \subseteq \mathbb{R}^I$, R^I – I-dimensional Real space, x^k – value of x a the k^{th} iteration, A: $\mathbb{R}^I \times \mathbb{R}^I \to \mathbb{R}^I$ (A – $I \times I$ square matrix).

If (1) is applied recursively

$$x^{k+1} = Ax^k = A^2 x^{k-1} = \dots = A^{k+1} x^0, \tag{2}$$

where $x^0 = (x_1^0 x_2^0 \dots x_I^0)^T \in \mathbb{R}^I$, ($x_i^0$ – initial value of x_i).

[1]This research was supported by the summer fellowship 2001 from RIACS at NASA-Ames Research Center, Moffett Field, CA 94035

It is well known [1] that if and only if A is a convergent matrix (which means $\lim_{k\to\infty} A^k = \emptyset$, a null matrix)

$$\lim_{k\to\infty} x^k = 0 \tag{3}$$

It is also known [1], that A is a convergent matrix if and only if for some given norm $\|A\| < 1$ which also implies $\rho(A) < 1$, where $\rho(A) = \max_i)|\lambda_i|$, λ_i's are the eigenvalues of A. Let us consider now nonstationary iterative process similar to (1):

$$x^{k+1} = A_k x^k, \tag{4}$$

where $x_k \in D_k$, $A_k : D_k \times D_k \subseteq \mathbb{R}^I \times \mathbb{R}^I \to D_{k+1} \subseteq \mathbb{R}^I$ for $\forall k$. If (4) is applied recursively,

$$x^{k+1} = A_k A_{k-1} x^{k-1} = A_k A_{k-1} \dots A_0 x^0 \tag{5}$$

In order that $\lim_{k\to\infty} x^k = 0$, a necessary and sufficient condition is:

$$\lim_{k\to\infty} A_k A_{k-1} \dots A_0 = 0 \tag{6}$$

It may be noted that if $\forall k$, A_k is a convergent matrix (6) is not necessarily valid.

Definition 1. A square matrix A is super convergent if

$$\rho(A) \leq \alpha < 1, \tag{7}$$

where $\rho(A) =$ the spectral radius of A.

Lemma 1. *A super convergent matrix is also convergent.*

Theorem 1. *If for all $k \geq K$, $\{A_k\}$ is a sequence of super convergent square matrices of the same type and if $\forall k$, A_k's have the same eigenvectors, then* (6) is valid.

Proof. Let $v = (v^1, v^2, \dots, v^I)$ be the eigenvectors of A_k for $\forall k$. Then

$$\begin{aligned} &A_0 v^j = \lambda_{0j} v^j, \quad A_1 A_0 v = \lambda_{0j}\lambda_{1j} v^j, \dots, \\ &A_k A_{k-1} \dots A_o v^j = \lambda_{0j}\lambda_{1j} \dots \lambda_{kj} v^j, \end{aligned} \tag{8}$$

where λ_{mj} is the eigenvalue of A_m corresponding to the eigenvector v^j. From (8)

$$\|(A_k A_{k-1} \dots A_0)v^j\| = |\lambda_{0j}| \cdot |\lambda_{1j}| \cdot \dots \cdot |\lambda_{kj}| \cdot \|v^j\|.$$

Since $\forall k \geq K$, A_k is a superconvergent matrix,

$$|\lambda_{kj}| \leq \alpha < 1 \quad \forall k \geq K.$$

Thus

$$\lim_{k\to\infty} |\lambda_{0j}| \cdot |\lambda_{1j}| \cdot \dots \cdot |\lambda_{kj}| = 0,$$

giving,

$$\lim_{k\to\infty} \|(A_k A_{k-1} \dots A_0)v^j\| = 0.$$

Since for all $j = 1, 2, \dots, I$, v^j is non null,

$$\lim_{k\to\infty} A_k A_{k-1} \dots A_0 = 0.$$

Obviously, the product matrix $A_k A_{k-1} \dots A_0$ is a convergent matrix. □

2 Examples

Example 1. Let

$$A = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.2 \end{bmatrix}, \quad B = \begin{bmatrix} 0.3 & 0 \\ 0 & 0.4 \end{bmatrix}.$$

The eigenvalues of A are $\lambda_1(A) = 0.1$, $\lambda_2(A) = 0.2$, and the eigenvalues of B are $\lambda_1(B) = 0.3$, $\lambda_1 = 0.4$. By definition, both are super convergent matrices. Let us find the eigenvectors of A and B. Let $v^1 = (v_1^1, v_2^1)^T$ be the eigenvector A corresponding to $\lambda_1(A)$. Then

$$\begin{bmatrix} 0.1 & 0 \\ 0 & 0.2 \end{bmatrix} \begin{bmatrix} v_1^1 \\ v_2^1 \end{bmatrix} = 0.1 \begin{bmatrix} v_1^1 \\ v_2^1 \end{bmatrix}$$

giving $v_1^1 = 1$ and $v_2^1 = 0$. Similarly corresponding to $\lambda_1(A)$, if $v^2(v_1^2, v_2^2)^T$ is the eigenvector, $v_1^2 = 0$, $v_2^2 = 1$. Thus the two eigenvectors of A are $(1, 0)^T$

and $(0,1)^T$ corresponding to $\lambda_1(A)$ and $\lambda_2(A)$. It can be easily seen, that they are also the eigenvectors of B corresponding to $\lambda_1(B)$ and $\lambda_2(B)$.

Thus by the Theorem 1, AB must be convergent, and we see that

$$AB = \begin{bmatrix} 0.03 & 0 \\ 0 & 0.08 \end{bmatrix}, \quad \lambda_1(AB) = 0.03, \quad \lambda_2(AB) = 0.08$$

showing that AB is convergent.

Example 2. Let

$$A = \begin{bmatrix} 0 & 1 \\ 0.01 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0.01 \\ 1 & 0 \end{bmatrix}.$$

Obviously $B{=}A^T$ (Transpose of A). The eigenvalues of A are given by the equation

$$\begin{vmatrix} -\lambda & 1 \\ 0.01 & -\lambda \end{vmatrix} = 0 \quad \text{or} \quad \lambda^2 = 0.01 \quad \text{giving} \quad \lambda = \pm 0.1.$$

These are also the eigenvalues of B. This may be verified. (Note: If λ is an eigenvalue of A, then it is also an eigenvalue of A^T. Because if λ is an eigenvalue of A, $\det|A - \lambda I| = 0$. Also, $A = (A^T)^T$ and $I = I^T$. Hence, $\det\left|(A^T)^T - \lambda I^T\right| = 0$, giving $\det\left|(A^T - \lambda I)^T\right| = 0$. Since $\det|A| = \det|A^T|$, $\det|A^T - \lambda I| = 0$. Thus λ is an eigenvalue of A^T.)

Let $v^i = (v_1^i, v_2^i)^T$ = the eigenvectors of A corresponding to λ_i $(i = 1, 2)$, the eigenvalues of A. Then by definition,

$$\begin{bmatrix} 0 & 1 \\ 0.01 & 0 \end{bmatrix} \begin{bmatrix} v_1^1 \\ v_2^1 \end{bmatrix} = 0.1 \begin{bmatrix} v_1^1 \\ v_2^1 \end{bmatrix}$$

giving $v_2^1 = 0.1v_1^1$ and $0.01v_1^1 = 0.1v_2^1$. Thus the eigenvector $v^1 = (1, 0.1)^T$ corresponding to the eigenvalue $\lambda_1 = 0.1$. Similarly,

$$\begin{bmatrix} 0 & 1 \\ 0.01 & 0 \end{bmatrix} \begin{bmatrix} v_1^2 \\ v_2^2 \end{bmatrix} = -0.1 \begin{bmatrix} v_1^2 \\ v_2^2 \end{bmatrix},$$

giving $v_1^2 = -0.1v_1^2$ and $0.01v_1^2 = -0.1v_2^2$. Let $v_1^2 = 1$, then $v_2^2 = -0.1$. Thus corresponding to the eigenvalue $\lambda_2 = -0.1$, the eigenvector $v^2 = (1, -0.1)^T$. This shows that the eigenvalues of A are 0.1 and -0.1 and the corresponding

eigenvectors are $(1, 0.1)^T$ and $(1, -0.1)^T$. B has the same eigenvalues. Let us compute its eigenvectors.

$$\begin{bmatrix} 0 & 0.01 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} v_1^1 \\ v_2^1 \end{bmatrix} = 0.1 \begin{bmatrix} v_1^1 \\ v_2^1 \end{bmatrix}$$

giving $0.01v_2^1 = 0.1v_1^1$ and $v_1^1 = 0.1v_2^1$. Then $v_1^1 = 1$, $v_2^1 = 10$. Also

$$\begin{bmatrix} 0 & .01 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} v_1^2 \\ v_2^2 \end{bmatrix} = -0.1 \begin{bmatrix} v_1^2 \\ v_2^2 \end{bmatrix}$$

giving $0.01v_2^2 = -0.1v_1^2$ and $v_1^2 = -0.1v_2^2$. Then $v_1^2 = 1$, $v_2^2 = -10$. Thus, the eigenvectors of A and B are different.

Thus, the eigenvectors of A and B are different. Since $\rho(A) < 1$ and $\rho(B) < 1$, both A and B are convergent. However,

$$AB = \begin{bmatrix} 0 & 1 \\ 0.01 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0.01 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 10^{-4} \end{bmatrix}.$$

Thus, $\rho(AB) = 1$. Hence AB is not a convergent matrix. In general two square matrices of the same type will have different eigenvectors. We need to look into how the product of square matrices of the same type may behave like convergent matrices.

Definition 2. If a sequence of square matrices of the same type $\{A_k\}$ be such that

$$\lim_{k \to \infty} A_k A_{k-1} \dots A_0 = 0 \tag{9}$$

then each A_k is called a *D-matrix*.

Theorem 2. *If for a given norm $\| \cdot \|_q$ (where q is fixed)*

$$\|A_k\|_q \leq \alpha < 1 \tag{10}$$

$\forall k \geq K$, A_k is a D-matrix.

Proof. In the q-norm

$$\begin{aligned} \|A_k A_{k-1} \dots A_0\| &\leq \|A_k\| \cdot \|A_{k-1}\| \cdot \dots \cdot \|A_0\| \\ &= \|A_k\| \cdot \|A_{k-1}\| \cdot \dots \cdot \|A_{K+1}\| \cdot \|A_K\| \|A_{K-1}\| \cdot \dots \cdot \|A_0\| \\ &\leq \alpha^{k-K+1} \cdot (\|A_{K-1}\| \cdot \dots \cdot \|A_0\|) \end{aligned}$$

$\lim_{k\to\infty} \alpha^{k-K+1} = 0$ since $0 < \alpha < 1$. Hence $\lim_{k\to\infty} \|A_k\, A_{k-1} \ldots A_0\|_q = 0$ which proves the theorem. □

Example 3. Let $A_{3n-2} = \text{diag}(\alpha_1, 1, 1)$, $A_{3n-1} = \text{diag}(1, \alpha_2, 1)$, $A_{3n} = \text{diag}(1, 1, \alpha_3)$, $n = 1, 2, 3 \ldots$.

If $\forall k \geq K$, $\max_k |\alpha_k| \leq \alpha < 1$, A_k is a D-matrix. (Note: None of these matrices is a convergent matrix.)

Definition 3. Let $U^k = \left(U_1^k \;\; U_2^k \;\; \ldots \;\; U_I^k\right)^T \in D \subseteq \mathbb{R}^I$. where U_i^k is the value of U_i at some k^{th} iteration. Let $G : D \subseteq \mathbb{R}^I \to D \subseteq \mathbb{R}^I$. If $\forall k \leq K$, $G'(U^k)$, the Jacobian matrix of $G(U^k)$, is a D-matrix, the function $G(U)$ is called a *D-mapping* [1].

Theorem. *Let an iterative process be given as*

$$U^{k+1} = G(U^k), \tag{11}$$

where $G : D \subseteq \mathbb{R}^I \to D \subseteq \mathbb{R}^I$. Let $U^* \in D$, be the fixed point of G. Then the iterative scheme (11) converges to U^* if and only if $G'(U^k)$ is a D-matrix (which implies that G is a D-mapping on D).

Proof.

$$U^{k+1} - U^* = G(U^k) - G(U^*).$$

Let $\epsilon^k = U^k - U^*$. Then, $G(U^*) = G(U^* - \epsilon^k)$. Thus

$$\epsilon^{k+1} = G'(\xi^k) \cdot \epsilon^k, \tag{12}$$

where

$$\xi^k = \left(\xi_1^k \; \xi_2^k \; \ldots \; \xi_I^k\right)^T \in D \text{ and}$$
$$\xi_i^k = U_i^k + \theta_i \epsilon_i^k,\ 0 < \theta_i < 1 \text{ for } \forall i.$$

Applying (12) recursively, we get

$$\epsilon^{k+1} = G'(\xi^k) \cdot G'(\xi^{k-1}) \cdot \ldots \cdot G'(\xi^0) \cdot \epsilon^0,$$

where $\epsilon^0 = U^0 - U^*$. From (9) (since $G'(\xi^k)$ is a D-matrix for $\forall k \geq K$)

$$\lim_{k\to\infty} \epsilon^k = 0$$

if and only if $G'(\xi^k)$ is a D-matrix. □

3 Analysis of D-Mapping Compatibility of Nonlinear Solvers

Newton's method is widely applied to solve nonlinear systems. Thus analysis of D-mapping compatibility of Newton's method will be done first. Consider the nonlinear system

$$F(U) = 0, \tag{13}$$

where $U = (U_1 \;\; U_2 \;\ldots\; U_I)^T \in \mathbb{R}^I$. In the element form (13) is:

$$f_i(U_1, U_2, \ldots, U_I) = 0, \quad i = 1, 2, \ldots, I.$$

Newton's iterative technique to solve (13) is:

$$U^{k+1} = U^k - F'(U^k)^{-1} F(U^k), \tag{14}$$

where

$$F'(U^k) = \begin{bmatrix} \frac{\partial f_1}{\partial U_1} & \frac{\partial f_1}{\partial U_2} & \cdots & \frac{\partial f_1}{\partial U_I} \\ \frac{\partial f_2}{\partial U_1} & \frac{\partial f_2}{\partial U_2} & \cdots & \frac{\partial f_2}{\partial U_I} \\ \vdots & & & \\ \frac{\partial f_I}{\partial U_1} & \frac{\partial f_I}{\partial U_2} & \cdots & \frac{\partial f_I}{\partial U_I} \end{bmatrix} \quad \text{(evaluated at } U^k\text{)}.$$

Let $U = U^*$ – the root $\in \mathbb{R}^I$. Then

$$U^{k+1} - U^* = U^k - U^* - F'(U^k)^{-1}\big(F(U^k) - F(U^*)\big), \tag{15}$$

since $F(U^*) = 0$.

Let $\epsilon^k = U^k - U^*$.

In the element form,

$$\begin{aligned} &f_i\big(U_1^k - \epsilon_1^k, U_2^k - \epsilon_2^k, \ldots, U_I^k - \epsilon_I^k\big) \\ &\quad = f_i(U^k) - \sum_{j=1}^{I} \frac{\partial f_i(U^k)}{\partial U_j}\epsilon_j^k + \frac{1}{2!}\left(\sum_{j=1}^{I} \frac{\partial^2 f_i}{\partial U_j^2}\big(\epsilon_j^k\big)^2\right) \\ &\quad + \left(\frac{\partial^2 f_i}{\partial U_1 \partial U_2}\epsilon_1^k\epsilon_2^k + \frac{\partial^2 f_i}{\partial U_1 \partial U_3}\epsilon_1^k\epsilon_3^k + \ldots + \frac{\partial^2 f_i}{\partial U_1 \partial U_I}\epsilon_1^k\epsilon_I^k\right) \\ &\quad + \left(\frac{\partial^2 f_i}{\partial U_2 \partial U_3}\epsilon_2^k\epsilon_3^k + \ldots + \frac{\partial^2 f_i}{\partial U_2 \partial U_I}\epsilon_2^k\epsilon_I^k\right) + \ldots + \frac{\partial^2 f_i}{\partial U_{I-1} \partial U_I}\epsilon_{I-1}^k\epsilon_I^k. \end{aligned} \tag{16}$$

All partial derivatives in the third term (second partial derivatives) are evaluated at $\xi_i^k = U_i^k + \theta_i \epsilon_i^k$, $0 < \theta_i < 1$ for $\forall i$.

Thus

$$F(U^*) = F(U^k - \epsilon^k) = F(U^k) - F'(U^k)\epsilon^k + A_0(\epsilon^k)^2 + A_1\epsilon_1^k(\epsilon^k) + A_2\epsilon_2^k(\epsilon^k) + \cdots + A_{I-1}\epsilon_{I-1}^k(\epsilon^k), \tag{17}$$

where

$$(\epsilon^k)^2 = \left((\epsilon_1^k)^2 \;\; (\epsilon_2^k)^2 \;\; \ldots \;\; (\epsilon_I^k)^2\right)^T,$$
$$\epsilon_1^k(\epsilon^k) = \left((\epsilon_1^k)^2 \;\; \epsilon_1^k\epsilon_2^k \;\; \ldots \;\; \epsilon_1^k\epsilon_I^k\right)^T.$$

In general,

$$\epsilon_j^k(\epsilon^k) = \left(\epsilon_j^k\epsilon_1^k \;\; \epsilon_j^k\epsilon_2^k \;\; \ldots \;\; \epsilon_j^k\epsilon_{j-1}^k, \; (\epsilon_j^k)^2, \; \epsilon_j^k\epsilon_{j+1}^k \;\; \ldots \;\; \epsilon_j^k\epsilon_I^k\right)^T, \quad j = 1, 2, \ldots I, \tag{18a}$$

$$A_0 = \frac{1}{2!}\begin{bmatrix} \frac{\partial^2 f_1}{\partial U_1^2} & \frac{\partial^2 f_1}{\partial U_2^2} & \cdots & \frac{\partial^2 f_1}{\partial U_I^2} \\ \frac{\partial^2 f_2}{\partial U_1^2} & \frac{\partial^2 f_2}{\partial U_2^2} & \cdots & \frac{\partial^2 f_2}{\partial U_I^2} \\ \vdots & & & \\ \frac{\partial^2 f_I}{\partial U_1^2} & \frac{\partial^2 f_I}{\partial U_2^2} & \cdots & \frac{\partial^2 f_I}{\partial U_I^2} \end{bmatrix}, \tag{18b}$$

$$A_1 = \begin{bmatrix} 0 & \frac{\partial^2 f_1}{\partial U_1 \partial U_2} & \frac{\partial^2 f_1}{\partial U_1 \partial U_3} & \cdots & \frac{\partial^2 f_1}{\partial U_1 \partial U_I} \\ 0 & \frac{\partial^2 f_2}{\partial U_1 \partial U_2} & \frac{\partial^2 f_2}{\partial U_1 \partial U_3} & \cdots & \frac{\partial^2 f_2}{\partial U_1 \partial U_I} \\ \vdots & & & & \\ 0 & \frac{\partial^2 f_I}{\partial U_1 \partial U_2} & \frac{\partial^2 f_I}{\partial U_1 \partial U_3} & \cdots & \frac{\partial^2 f_I}{\partial U_1 \partial U_I} \end{bmatrix}, \tag{18c}$$

$$A_2 = \begin{bmatrix} 0 & 0 & \frac{\partial^2 f_1}{\partial U_2 \partial U_3} & \frac{\partial^2 f_1}{\partial U_2 \partial U_4} & \cdots & \frac{\partial^2 f_1}{\partial U_2 \partial U_I} \\ 0 & 0 & \frac{\partial^2 f_2}{\partial U_2 \partial U_3} & \frac{\partial^2 f_2}{\partial U_2 \partial U_4} & \cdots & \frac{\partial^2 f_2}{\partial U_2 \partial U_I} \\ \vdots & & & & & \\ 0 & 0 & \frac{\partial^2 f_I}{\partial U_2 \partial U_3} & \frac{\partial^2 f_I}{\partial U_2 \partial U_4} & \cdots & \frac{\partial^2 f_I}{\partial U_2 \partial U_I} \end{bmatrix}, \tag{18d}$$

$$A_{I-1} = \underbrace{\begin{bmatrix} 0 & 0 & \dots & 0 & \frac{\partial^2 f_1}{\partial U_{I-1}\partial U_I} \\ 0 & 0 & \dots & 0 & \frac{\partial^2 f_2}{\partial U_{I-1}\partial U_I} \\ \vdots & & & & \\ 0 & 0 & \dots & 0 & \frac{\partial^2 f_I}{\partial U_{I-1}\partial U_I} \end{bmatrix}}_{\text{1st (I-1) columns}}. \tag{18e}$$

Substituting (16) in (15) we get,

$$\epsilon^{k+1} = \epsilon^k - F'(U^k)^{-1}\{F'(U^k)\epsilon^k - \sigma\}, \tag{19}$$

where

$$\sigma = A_0(\epsilon^k)^2 + A_1\epsilon_1^k(\epsilon^k) + \cdots + A_{I-1}\epsilon_{I-1}^k(\epsilon^k). \tag{20}$$

From (19) and (20), for some norm

$$\|\epsilon^{k+1}\| \leq I \cdot \|F'(U^k)^{-1}\| \cdot \|A\|_{\max} \cdot \|\epsilon^k\|_{\max}^2, \tag{21}$$

where

$$\|A\|_{\max} = \max_j \|A_j\|, \quad j = 0, 1, 2, \dots, I-1,$$
$$\|\epsilon^k\|_{\max} = \max_j \|\epsilon_j^k\|, \quad j = 1, 2, \dots, I.$$

The inequality (21) is often referred to as the quadratic rate of convergence of Newton's method. If initial guess is sufficiently close to the root or in other words, if

$$\|\epsilon^0\| = \|U^0 - U^*\| < 1, \tag{22}$$

(21) generates a faster rate of convergence, but the algorithm is not compatible with the analysis of D-mapping.

For application of PFI (perturbed Functional Iterations) [2] (13) is expressed as

$$U = G(U), \tag{23}$$

$U \in D \subseteq \mathbb{R}^I, \quad U = U^*$ – root.

In the element form (23) is

$$U_i = G_i(U_1, U_2, \ldots U_I), \quad i = 1, 2, \ldots, I. \tag{24}$$

Also, $U^* = (U_1^* \, U_2^* \ldots U_I^*)^T \in D \subseteq \mathbb{R}^I$. Algorithm of PFl consists of three steps:

Step 1. Compute intermediary iterates

$$\begin{aligned} &\hat{U}_i = G_i(U_1^k, U_2^k \ldots U_I^k), \quad i = 1, 2, \ldots I, \\ &U_i^k \text{ – value of } U_i \text{ at the } k^{\text{th}} \text{ iteration.} \end{aligned} \tag{25}$$

Step 2. Compute perturbation parameters

$$\begin{aligned} &\omega_i^k = \frac{G_i(\hat{U}_1, \hat{U}_2, \ldots \hat{U}_I) - \hat{U}_i}{1 - \frac{\partial G_i(\hat{U})}{\partial U_i}}, \quad i = 1, 2, \ldots, I, \\ &\frac{\partial G_i(\hat{U})}{\partial U_i} \text{ – value of } \frac{\partial G_i}{\partial U_i} \text{ at } \left(\hat{U}_1, U_2^k, \ldots, U_I^k\right). \end{aligned} \tag{26}$$

Step 3. Compute the new $(k+1)^{\text{th}}$ iterates

$$U_i^{k+1} = \omega_i^k + \hat{U}_i, \quad i = 1, 2, \ldots, I. \tag{27}$$

Iterations stop, if at some k

$$\max_i |\omega_i^k| < \epsilon, \tag{28}$$

where ϵ is positive and arbitrarily small. It may be proved that (28) is both necessary and sufficient for convergence of PFl.

Let us study the analysis of convergence of PFl.

$$U^{k+1} - U^* = \omega^k + \hat{U} - U^*, \quad \omega^k = (\omega_1^k \;\; \omega_2^k \;\; \ldots \;\; \omega_I^k)^T. \tag{29}$$

From (26),

$$\omega^k = (I - \partial G)^{-1}\left(G(\hat{U}) - \hat{U}\right), \tag{30}$$

where

$$I - \partial G = \text{diag}\left(1 - \frac{\partial G_1}{\partial U_1}, 1 - \frac{\partial G_2}{\partial U_2}, \ldots, 1 - \frac{\partial G_I}{\partial U_I}\right)$$

and all partial derivatives are evaluated at $(\hat{U}_1, \hat{U}_2, \ldots, \hat{U}_I)$.

Now,

$$G(\hat{U}) - \hat{U} = G(\hat{U}) - G(U^*) + U^* - \hat{U}, \tag{31}$$

since $U^* = G(U^*)$.

Let $\hat{\epsilon} = \hat{U} - U^*$.

$G(U^*) = G(\hat{U} - \hat{\epsilon}) = G(\hat{U}) - G'(\hat{U}) \cdot \hat{\epsilon}$ (Neglecting terms of the order $\|\hat{\epsilon}\|^2$). Substituting in (31)

$$G(\hat{U}) - \hat{U} = \big(G'(\hat{U}) - I\big)\hat{\epsilon}. \tag{32}$$

From (29), (30) and (32)

$$\begin{aligned} \epsilon^{k+1} &= (I - \partial G)^{-1}\big(G'(\hat{U}) - I\big)\hat{\epsilon} + \hat{\epsilon} \\ &= \Big\{I - (I - \partial G)^{-1}\big(I - G'(\hat{U})\big)\Big\}\hat{\epsilon}' \end{aligned} \tag{33}$$

where $\epsilon^{k+1} = U^{k+1} - U^*$. Now,

$$\begin{aligned} \hat{\epsilon} &= \hat{U} - U^* = G(U^k) - G(U^*) = G(U^k) - G(U^k - \epsilon^k) \\ &= G'(U^k) \cdot \epsilon^k. \end{aligned} \tag{34}$$

Thus, from (33) and (34)

$$\epsilon^{k+1} = A_k \cdot B_k \cdot \epsilon^k, \tag{35}$$

where

$$A_k = I - (I - \partial G)^{-1}\big(I - G'(\hat{U})\big), \tag{36a}$$

$$B_k = G'(U^k). \tag{36b}$$

Obviously

$$B_k = \begin{bmatrix} \frac{\partial G_1}{\partial U_1} & \frac{\partial G_1}{\partial U_2} & \cdots & \frac{\partial G_1}{\partial U_I} \\ \frac{\partial G_2}{\partial U_1} & \frac{\partial G_2}{\partial U_2} & \cdots & \frac{\partial G_2}{\partial U_I} \\ \vdots & & & \\ \frac{\partial G_I}{\partial U_1} & \frac{\partial G_I}{\partial U_2} & \cdots & \frac{\partial G_I}{\partial U_I} \end{bmatrix}.$$

Considering the $\infty-$norm, (with $G_{ij} = \frac{\partial G_i}{\partial U_j}$), if B_k is diagonally dominant, $\|B_k\|_\infty = \max\limits_i \sum\limits_{j=1}^{I} |G_{ij}(U^k)| < 1$. This implies,

$$\forall i, \quad \sum_{\substack{j=1 \\ j\neq i}}^{I} \left|G_{ij}(U^k)\right| < 1 - \left|G_{ii}(U^k)\right| \tag{37}$$

obviously $\left|G_{ii}(U^k)\right| < 1$).

Since $1 = |x + 1 - x| \leq |x| + |1 - x|$ or $1 - |x| \leq |1 - x|$, thus from (37)

$$\forall i, \quad \sum_{\substack{j=1 \\ j\geq i}}^{I} \left|G_{ij}(U^k)\right| < \left|1 - G_{ii}(U^k)\right|. \tag{38}$$

Now

$$A_k = \begin{bmatrix} 0 & \frac{G_{12}(\hat{U})}{1-G_{11}(\hat{U})} & \cdots & \frac{G_{1I}(\hat{U})}{1-G_{11}(\hat{U})} \\ \frac{G_{21}(\hat{U})}{1-G_{22}(\hat{U})} & 0 & \frac{G_{23}(\hat{U})}{1-G_{22}(\hat{U})} & \cdots & \frac{G_{2I}(\hat{U})}{1-G_{22}(\hat{U})} \\ \vdots & & & & \\ \frac{G_{I1}(\hat{U})}{1-G_{II}(\hat{U})} & \frac{G_{I2}(\hat{U})}{1-G_{II}(\hat{U})} & \cdots & 0 \end{bmatrix}. \tag{39}$$

If (38) is valid,

$$\|A_k\|_\infty = \max_i \sum_{\substack{j=1 \\ j\neq i}}^{I} \frac{\left|G_{ij}(\hat{U})\right|}{\left|1 - G_{ii}(\hat{U})\right|} < 1. \tag{40}$$

Then from (35)

$$\|\epsilon^{k+1}\|_\infty \leq \|M_k\|_\infty \cdot \|\epsilon^k\|_\infty, \tag{41}$$

where

$$\|M_k\|_\infty = \|A_k\|_\infty \cdot \|B_k\|_\infty \leq \alpha < 1. \tag{42}$$

From (41) it is evident that if (37) is valid, or in other words, if B_k is diagonally dominant, the algorithm of PF1 is compatible with the properties of D-mapping.

Since in general, nonlinear systems may not be diagonally dominant, PFl has recently been modified so that it may be fruitful when diagonal dominance is not displayed by the nonlinear system.

The new algorithm is: MPFl (modified perturbed functional iterations) which requires the following steps:

Define,

$$\begin{aligned} &H(U) = (I - \Gamma)U + \Gamma G(U), \\ &\Gamma = \text{diag}(\gamma_1, \gamma_2, \dots \gamma_I), \\ &\hat{U} = H(U^k), \\ &\omega^k = \frac{1}{\alpha}(I - \partial H)^{-1}\big(H(\hat{U}) - \hat{U}\big), \\ &\alpha - \text{a scalar}, \\ &U^{k+1} = \omega^k + \hat{U}. \end{aligned} \tag{43a}$$

By applying the analysis of convergence, it has been found that [3], it is preferable to choose

$$\alpha = \frac{1}{I - 1} \tag{43b}$$

and $\Gamma = \beta^{-1}(I - \partial G)^{-1}$, where $\beta = I/(I - 1)$.

These transformations render PFl more effective to solve a vast number of nonlinear systems quite accurately, which may not display diagonal dominance in their linearized versions. The convergence of PFl is global and thus it is independent of the initial guess.

4 Time Accurate Numerical Solution of Nonlinear Flow Models

A time dependent flow model may be expressed as

$$\frac{\partial u}{\partial t} = F(u, u_x, u_{xx}) \tag{44}$$

subject to a given set of initial-boundary conditions. A consistent implicit finite difference approximation of (44) may be expressed as

$$U^{n+1} = U^n + G(U^{n+1}), \tag{45}$$

when $U^n = (U_1^n \ U_2^n \ \dots \ U_I^n)^T$

$U_i^n = U(x_i, t_n)$ = the net function corresponding to u_i^n.

x_i are the grid points, $i = 1, 2, \dots I$, t_n = value of t at the n time frame.

In general, implicit finite difference methods have much better stability properties, which means that if Δt is chosen to be large, often much larger than Δx (mesh size), the error $\|U^n - u^n\|$ remains bounded which implies that with an appropriate iterative scheme, solution of (45) is guaranteed. Yet in several cases Netwon's method failed. The question is why?

Let us replace U^{n+1} by U and express (45) as

$$U = U^n + G(U). \tag{46}$$

This is a nonlinear system and could be solved by iterative scheme

$$U^{k+1} = U^n + G(U^k). \tag{47}$$

If $U = U^{*,n+1}$ – the solution at t_{n+1},

$$U^{*n+1} = U^n + G(U^{*,n+1}). \tag{48}$$

Let $\epsilon^{k+1} = U^{k+1} - U^{*,n+1}$, then

$$\epsilon^{k+1} = G'(U^k)\epsilon^k. \tag{49}$$

If $G'(U^k)$ is a D-matrix (or could be transformed into a D-matrix by appropriate transformations), $\lim_{k\to\infty} \epsilon^k = 0$, and the numerical method must converge, and the convergence is global, not local which means a large time step Δt, should work. Newton's method does not require this condition to be satisfied.

Iterative solution for (46) always requires an initial guess. The initial guess for U at $t = t_{n+1}$ is U^n (which is U at t_n). If $\Delta t = t_{n+1} - t_n$, is large, $\|U^{*,n+1} - U^n\|$ could be significantly large which should cause any numerical nonlinear system solver which is only locally convergent to fail. This also could cause Newton's method to fail. PFI is globally convergent. However, it is dependent of the properties of D-mapping. In many cases modified perturbed functional iterations (MPFI) was used and results seem to be quite encouraging.

5 Applications

Application 1. Let us first consider Burgers' Equation

$$u_t + uu_x = \nu u_{xx}. \tag{50}$$

We have considered several sets of initial-boundary conditions. Let $0 \le x \le 1$.

$$\text{Case 1} \quad \text{If } 0 \le x \le 0.02,\ u(x,0) = 1, \text{ else } u(x,0) = -0.5. \tag{51a}$$

$$\begin{aligned}\text{Case 2} \quad &\text{If } 0 \le x \le 0.02,\ u(x,0) = 1, \\ &\text{If } 0.02 < x \le 0.15,\ u(x,0) = 0.5, \text{ else } u(x,0) = -0.5.\end{aligned} \tag{51b}$$

$$\begin{aligned}\text{Case 3} \quad &\text{If } 0 \le x \le 0.02,\ u(x,0) = 1, \\ &\text{If } 0.02 < x \le 0.15,\ u(x,0) = 0.5, \\ &\text{If } 0.15 < x \le 0.25,\ u(x,0) = 0.13, \text{ else } u(x,0) = -0.5.\end{aligned} \tag{51c}$$

$$\begin{aligned}\text{Case 4} \quad &\text{If } 0 \le x \le 0.25,\ u(x,0) = \sin \pi x, \\ &\text{If } 0.25 \le x \le 0.75,\ u(x,0) = 0.5 - x, \\ &\text{If } 0.75 < x \le 1,\ U(x,0) = \sin\big(\pi(x-1)\big).\end{aligned} \tag{51d}$$

The implicit finite difference equation for the Burgers' Equation (50) has been obtained by approximating the differential equation by the algorithm of flux-vector splitting, developed by Steger and Warming [4]. The finite difference representation of (50) is:

$$\begin{aligned}U^{n+1} =& U_j^n - a\Big\{U_{j+1}^{n+1}\big(U_{j+1}^{n+1} - |U_{j+1}^{n+1}|\big) + 2U_j^{n+1}\|U_j^{n+1}\| \\ &- U_{j-1}^{n+1}\big(U_{j-1}^{n+1} + |U_{j-1}^{n+1}|\big)\Big\} + b\big(U_{j-1}^{n+1} - 2U_j^{n+1} + U_{j+1}^{n+1}\big),\end{aligned} \tag{52}$$

where $a = \Delta t/(2\Delta x)$, $b = \nu\Delta t/(\Delta x^2)$ Δt– time step, Δx – mesh size. $j = 1, 2, \ldots J$. This equation may be expressed as

$$\begin{aligned}U_j^{n+1} &= U_j^n + G\big(U_{j-1}^{n+1},\ U_j^{n+1},\ U_{j+1}^{n+1}\big), \\ \frac{\partial G_j}{\partial U_j} &= -2aU_j^{n+1} - 2b, \quad \text{if} \quad U_j^{n+1} > 0, \\ &= 2aU_j^{n+1} - 2b, \quad \text{if} \quad U_j^{n+1} < 0.\end{aligned} \tag{53}$$

The Fig. 1 corresponds to Case 1. Here with $\nu = 10^{-8}$, $\Delta x = 0.01$, $\Delta t = 10\Delta x$, the one dimensional motion of the shock is clearly displayed. There

exists no overshooting or undershooting. The shock is very slightly smeared. In general it is quite crisp.

The Fig. 2, corresponds to Case 2. Here with $\nu = 10^{-7}$, $\Delta x = 0.01$, $\Delta t = 3.5\Delta x$, the merging of two shocks is depicted. With no over and undershooting, the shock is relatively crisp.

The Fig. 3, corresponds to Case 3. Here triple shocks merge with each other with $\nu = 10^{-7}$, $\Delta x = 0.01$, $\Delta t = 3.5\Delta x$.

The Fig. 4 corresponds to Case 4. The jump discontinuities occur at $x = 0.25$ and $x = 0.75$ respectively.

Here $\Delta x = 0.001$, $\Delta t = 3.5\Delta x$, $\nu = 10^{-8}$. In the Fig. 5, dissipation of u can be noticed with no over/under shooting of computational results.

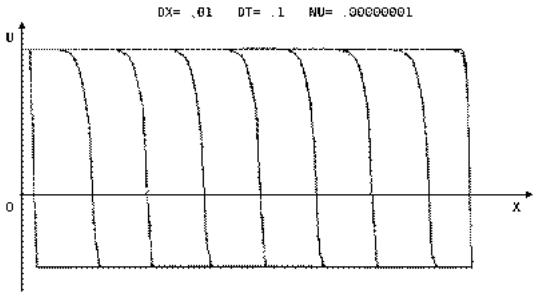

Fig. 1. One dimensional motion of a shock.

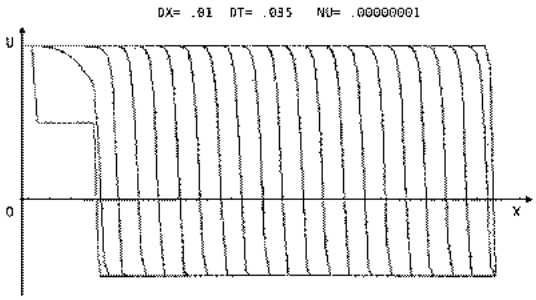

Fig. 2. Double shocks merging with each other.

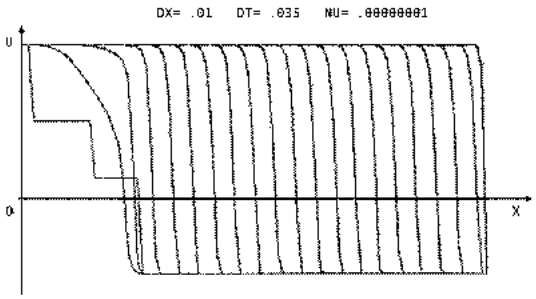

Fig. 3. Triple shocks merging with each other.

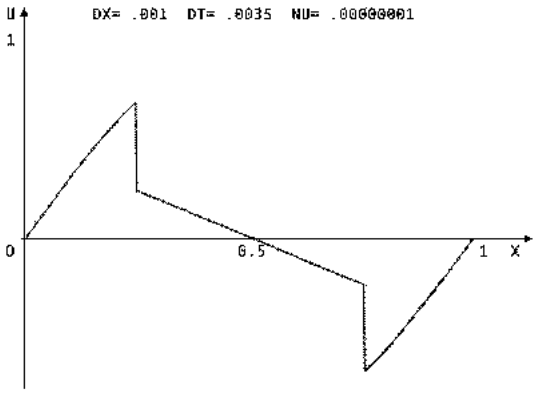

Fig. 4. Jump discontinuities.

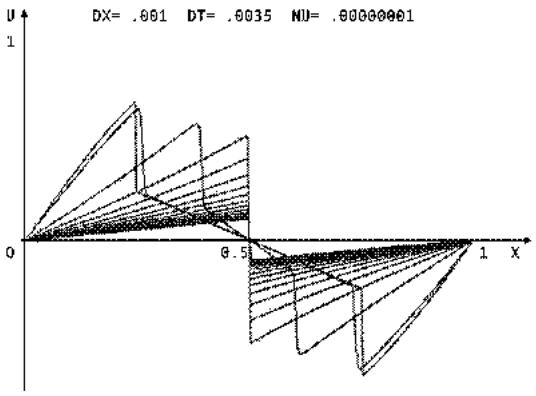

Fig. 5. Dissipation of u as time progresses.

Application 2.

$$u_t + uu_x + uu_y = \nu\Delta^2 u. \tag{54}$$

An exact solution is $u = \dfrac{\nu(x+y)}{2(1+\nu t)}$. A simple implicit finite difference analog for (54) is:

$$\begin{aligned} U_{ij}^{n+1} = U_{ij}^{n} &- a_1 U_{i,j}^{n+1}\left(U_{i+1,j}^{n+1} - U_{i-1,j}^{n+1}\right) \\ &- a_2 U_{i,j}^{n+1}\left(U_{i,j+1}^{n+1} - U_{i,j-1}^{n+1}\right) \\ &+ b_1\left(U_{i-1,j}^{n+1} - 2U_{i,j}^{n+1} + U_{i+1,j}^{n+1}\right) \\ &+ b_2\left(U_{i,j-1}^{n+1} - 2U_{i,j}^{n+1} + U_{i,j+1}^{n+1}\right) \qquad (55) \\ a_1 = \Delta t/(2\Delta x),&\ \ a_2 = \Delta t/(2\Delta y), \\ b_1 = \nu\Delta t/(\Delta x^2),&\ \ b_2 = \nu\Delta t/(\Delta y^2), \\ i = 1, 2, \ldots I,&\ \ j = 1, 2, \ldots J. \end{aligned}$$

In the code mesh sizes have been chosen as $\Delta x = \Delta y = 0.01$, $0 \le x \le 1$ and $0 \le y \le 1$.

At each time step, the largest error in computations, defined by:

$$E_{\max} = \max_{i,j} \left| u_{i,j}^{n} - U_{i,j}^{n} \right|,$$

where $u_{i,j}^{n} = u(x_i, y_j, t_n)$ = the exact solution has been computed.

With $\nu = 10^{-2}$, and $\Delta t = 10^5 \Delta x = 1000$, $E_{\max} = 8.5475 \times 10^{-3}$ and remains the same after $n = 600$ time steps.

With $\nu = 10^{-3}$, and $\Delta t = 10^5 \Delta x = 1000$, $E_{\max} = 8.516792 \times 10^{-4}$ and remains the same after $n = 200$ time steps.

With $\nu = 10^{-7}$, and $\Delta t = 10^5 \Delta x = 1000$, $E_{\max} = 6.5608 \times 10^{-9}$ and remains the same after $n = 3000$ time steps.

A mathematical analysis is now under investigation to understand this pattern of results.

6 Discussions

PFI is indeed a quite powerful method to solve nonlinear differential equations numerically. The recent modifications of the method [3] given in (43a) and

(43b) added more to its computational strength. When the initial-boundary conditions are continuous, $G'(U)$, the Frechet derivative of $G(U)$ exists and PFI does display its properties of global convergence. However, if continuity of initial-boundary conditions is violated and these discontinuities propagate with time, PFI could fail if the time step is large. That is why, whereas $\Delta t = 10\Delta x$ for the Case 1, $\Delta t = 3.5\Delta x$ for other cases in the Application 1. For the Application 2, there are no discontinuities and solutions are always smooth. That is why a large $\Delta t = 1000$ worked quite well signifying the global convergence of PFI.

Let us study how numerical solution by PFI could be affected by errors introduced in the initial-boundary conditions. Let at $t = t_1$, these errors be introduced. Then

$$U^1 = (U_0 + \alpha) + G(U^1) + (b + \beta), \tag{56}$$

where U_0 – initial value of U at $t = t_0$, U^1 – value of U at $t = t_1$, b – boundary conditions at $t = t_1$, α – small error in U_0, β – small error in b. Let $\mu = \alpha + \beta$, then (56) may be expressed as:

$$U = U_0 + G(U) + b + \mu. \tag{57}$$

Let U^* = the solution for U at $t = t_1$. Then

$$U^* = U_0 + G(U^*) + b. \tag{58}$$

The equation (57) is solved by iterations as follows:

$$U^{k+1} = U_0 + G(U^k) + b + \mu. \tag{59}$$

Let $\delta_k = U^k - U^*$.

Let us assume that $G'(U^k)$ is a D-matrix $\forall k \geq K$ and $\|G'(U^k)\| \leq \alpha < 1$ for some norm. Subtracting (58) from (59) and linearizing at U^*,

$$\delta_{k+1} = G'(U^*) \cdot \delta_k + \mu \tag{60}$$

Let $\theta = G'(U^*)$. Then $\|\theta\| \leq \alpha < 1$ and θ is also a convergent matrix. Then applying (60) recursively,

$$\delta_{k+1} = \theta^{k+1}\delta_0 + (I + \theta + \theta^2 + \cdots + \theta^k)\mu. \tag{61}$$

If for some $k = K$,

$$\|\delta_{K+1}\| < \epsilon, \quad \epsilon \text{ is positive and arbitrarily small,} \tag{62}$$

it implies convergence.

Equation (62) is true if

$$\|\theta\|^{k+1} \cdot \|\delta_0\| + \left(\frac{1 - \|\theta\|^{k+1}}{1 - \|\theta\|}\right) \|\mu\| < \epsilon \tag{63}$$

Since $\|\theta\| \leq \alpha < 1$ and K is large, for convergence of PFl, $\|\mu\|$ has to stay bounded. Also, from (61)

$$\lim_{k\to\infty} \delta_{k+1} = \left\{ \lim_{k\to\infty} \left(I + \theta + \theta^2 + \cdots + \theta^k\right)\right\}\mu = (I - \theta)^{-1}\mu, \tag{64}$$

giving $\mu = (I - \theta) \lim_{k\to\infty} \delta_{k+1}$. Thus whenever PFl converges, it implies that errors from initial/boundary conditions are bounded. (Theoretically they are damped out.) Thus, theoretically convergence of PFl is a sufficient condition that the errors from initial/boundary conditions are damped out.

Let us consider Newton's method and see how Newton's iterations may be affected by small changes in the initial-boundary conditions.

The finite difference equations (52) (or (55)), may be set up as:

$$F(U) = \epsilon \tag{65}$$

at a given time step t_{n+1}, where ϵ consists of the perturbations given to initial-boundary conditions. If U^* is the root, $F(U^*) = \epsilon$. Let U^0 be the initial guess for U^*. Then

$$\epsilon = F(U^*) = F(U^0 + \delta^0), \tag{66}$$

where $\delta^0 = U^* - U^0$. Linearizing (66) we see,

$$\delta^0 = F'(U^0)^{-1}\left(\epsilon - F(U^0)\right). \tag{67}$$

Then U^1, the first Newton-iterate is given by:

$$U^1 = U^0 + F'(U^0)^{-1}\left(\epsilon - F(U^0)\right) \tag{68}$$

Thus, at each iteration level, an error $F'(U^k)^{-1}\epsilon$, $(k = 0, 1, 2, \dots)$ is added to the standard Newton's iterate $U^{k+1} = U^k - F'(U^k)^{-1}F(U^k)$. This will cause enhancement of error caused by perturbing initial-boundary conditions. The method will converge to U^*, if and only if

$$\lim_{k\to\infty} F'(U^k)^{-1} = 0 \tag{69}$$

The convergence properties of PFI as we have noticed earlier, keep the errors from initial-boundary conditions bounded at each iteration.

7 Conclusion

For the past two decades PFI has been applied successfully to solve nonlinear differential equations. Because of its properties of global convergence, several boundary value problems represented by integro-differential equations were also solved by PFI [5], where most Newton-type methods involving direct or indirect inversion of matrices failed.

PFI linearizes the model only along the diagonal thus it maintains most information regarding nonlinearities. This is why when it works, the solutions are highly accurate.

References

1. Dey S.K. "D-Mapping Analysis for Numerical Solution of Nonlinear Systems with Applications to Differential Equations", *International Journal of Modelling and Simulation*, **16**(1), 1996

2. Dey S.K. "Numerical solution of Euler's Equation by Perturbed Functionals", *Lectures in Applied Mathematics*, **26**, American Mathematical Society 1985

3. Dey S.K. "Modified Perturbed Functional Iterations (MPFI) for Nonlinear Equations", *International Conference on Math. Modelling*, Singh B., Gupta U.S., Srivastava G.S., Gulati T.R., Katiyar V.K. (Eds.), Tata McGraw Hill, New Delhi, 2001

4. Steger J.L., Warming R.F. "Flux Vector Splitting of the Inviscid Gasdynamics Equations with Application to Finite Difference Methods", *Journal of Comp. Physics*, **40**, 1981

5. Dey S.K. "Finite Difference Solution of Boundary-Layer Equations", *Num. Heat Transfer,* **3**, 1980

6. Ames W.F. *Numerical Methods for Partial Differential Equations,* 4th ed., Academic Press, New York, 1994

7. Anderson D.A., Tannehill J.C., Pletcher R.H. *Computational Fluid Mechanics and Heat Transfer,* Taylor and Francis, 1984

8. Beam R.M., Warming R.F. "An Implicit Finite-Difference Algorithm for Hyperbolic Systems in Conservation Law Form", *J. Comp. Physics,* **22**, 1976

9. Bugers J.M. "A Mathematical Model Illustrating the Theory of Turbulence", *Adv. Applied Mech.,* **1**, 1948

10. Lax P.D., Wendroff B. "Difference Schemes for Hyperbolic Equaitons and Higher Order Accuracy", *Comm. Pure Appl. Math.,* **17**, 1964