Z. angew. Math. Phys. 57 (2006) 1075–1091
0044-2275/06/061075-17
DOI 10.1007/s00033-005-0048-z

Zeitschrift für angewandte Mathematik und Physik ZAMP

Weighted iterative solutions of linear differential equations and heat conduction of ideal gases in moment theory

I-Shih Liu and Daniel R. Vieira

Abstract. An iterative method is proposed to find a particular solution of a system of linear differential equations, in the form of a fixed-point problem, with no boundary conditions. To circumvent the unboundedness of differential operators, iterative approximation with gradually decreasing weight is used. Conditions for convergence that can easily be checked in numerical iterations are established. Furthermore, for the numerical iterative scheme, uniqueness and stability theorems are proved. These results are applied to heat conduction of ideal gases in moment theory.

Mathematics Subject Classification (2000). 65F10, 65N12, 35K05, 74A25.

Keywords. Iterative approximation, estimated error, numerical stability, uncontrollable boundary value.

1. Introduction

The method of solutions considered in this paper is motivated from boundary value problems in moment equations of ideal gases. Systems of moment equations are regarded as improvements of the classical theories, such as Fourier theory of heat conduction and Navier-Stokes theory of fluid flows. In such systems, besides the usual conservation laws of mass, momentum and energy, there are additional balance equations for higher order moments, which do not have clear physical interpretations and hence their boundary values can not be assigned. Consequently, with insufficient boundary conditions for the problem, the system of moment equations can not be solved uniquely.

Mathematically, the balance equations for higher order moments can be reduced to a system of linear differential equations with non-homogeneous terms, and we proceed to find a solution with no boundary conditions. The method proposed is an iterative approximation with successive application of differential operators in the form of a fixed-point problem.

Well-known fixed-point theorems [2, 12] are usually concerned with bounded operators only. To circumvent the difficulties of unboundedness of differential op-

erators, an iterative approximation with gradually decreasing weight is proposed. Under some conditions, which can easily be checked in numerical iterations, the convergence is proved. Theoretically, the limiting function depends usually on the initial iterate. However, for numerical iterative approximations, it is proved that it converges uniquely to the solution corresponding to the null initial iterate, independent of any particular choice of initial iterates.

Heat conduction of ideal gases in moment theory is given as a numerical example. This problem has been considered in several recent studies [1, 5, 10, 15] of boundary value problems in moment equations with insufficient boundary conditions, by postulating some additional physical assumption or criteria. The present approach does not require any additional physical assumption.

2. Iterative solutions of linear differential equations

We consider a system of non-homogeneous linear differential equations:

$$P(x)\boldsymbol{u}' + Q(x)\boldsymbol{u} = \boldsymbol{r}(x),$$

where $\boldsymbol{u}, \boldsymbol{r} : [a,b] \to I\!R^N$ and $P, Q : [a,b] \to L(I\!R^N)$; $L(I\!R^N)$ is the space of linear map on $I\!R^N$. The prime ($'$) denotes the derivative with respect to x. We are interested in finding a particular solution of the system with no boundary conditions.

The system can conveniently be put into a fixed-point problem of the form:

$$\boldsymbol{u} = G\boldsymbol{u} = \boldsymbol{g} + \boldsymbol{p}\,\boldsymbol{u}', \tag{1}$$

where $\boldsymbol{g} : [a,b] \to I\!R^N$ and $\boldsymbol{p} : [a,b] \to L(I\!R^N)$. The operator G is a differential operator on some function space X with norm denoted by $\|\cdot\|$.

In order to solve the system by iterative approximation, we first note that the simple iterative scheme[1],

$$\boldsymbol{u}_n = G\boldsymbol{u}_{n-1}, \qquad n = 1, 2, 3 \cdots, \tag{2}$$

usually requires that the operator G be contractive for the convergence to a fixed point [12], i.e.,

$$\|G\boldsymbol{u} - G\boldsymbol{v}\| \le C\|\boldsymbol{u} - \boldsymbol{v}\| \quad \text{for} \quad C < 1. \tag{3}$$

It is also known [2, 12] that if the operator G is non-expansive, i.e., for $C = 1$ in (3), the following iteration,

$$\boldsymbol{u}_n = w\,G\boldsymbol{u}_{n-1} + (1-w)\boldsymbol{u}_{n-1} \qquad n = 1, 2, 3, \cdots, \tag{4}$$

with a constant weight $w \in (0,1)$, converges to a fixed point.

[1] This simple iterative scheme is akin to the *Maxwellian iteration* (denominated by Truesdell who also referred to it as *differential iteration* [6]) in the kinetic theory of gases, to obtain approximations of Fourier law of heat conduction and Navier-Stokes law of viscous stress from Grad's moment equations [4]. Usually only few iterations are considered without any assurance of convergence.

3. Weighted iterative approximation

Since differential operators are neither contractive nor non-expansive in general, we propose an iterative approximation with gradually decreasing weight depending on the iteration,

$$\boldsymbol{u}_n = w_n G\boldsymbol{u}_{n-1} + (1-w_n)\boldsymbol{u}_{n-1} \quad \text{for} \quad 0 < w_n \le 1, \quad n = 1, 2, 3, \cdots .$$

It can be rewritten as

$$w_n(G\boldsymbol{u}_{n-1} - \boldsymbol{u}_{n-1}) = \boldsymbol{u}_n - \boldsymbol{u}_{n-1}. \tag{5}$$

We shall take the weights w_n to be monotonically decreasing in n, specifically, $w_n = 1/n^k$ for $k \ge 1$, such that $w_n \to 0$ when $n \to \infty$.

From (5), we can define the *estimated error* of the fixed point of G at the n-th approximation by

$$\mathrm{Err}\,(n) = \|G\boldsymbol{u}_{n-1} - \boldsymbol{u}_{n-1}\|, \tag{6}$$

which estimates the error of the approximate solution as the fixed point of the equation $G\boldsymbol{u} = \boldsymbol{u}$. Accordingly, we have the following theorem for the convergence of the approximation of the fixed-point problem.

Firstly, we need the following lemma for which the proof can be found in textbooks of functional analysis (see e.g., [8, 16]). Let X be the Banach space $C([a,b], I\!R^N)$ with sup-norm and a subspace $\mathcal{D} = C^1([a,b], I\!R^N) \subset X$.

Lemma 3.1. The linear operator $D : \mathcal{D} \to X$ taking $\boldsymbol{u} \mapsto \boldsymbol{p}\,\boldsymbol{u}'$, where $\boldsymbol{p} \in C^1([a,b], L(I\!R^N))$, $\det \boldsymbol{p} \neq 0$ in $[a,b]$, is closed, i.e., if $x_n \in \mathcal{D}$, $x_n \to x^*$ in X and $Dx_n \to y$ in X, then $x^* \in \mathcal{D}$ and $Dx^* = y$.

Theorem 3.1. Let $G : \mathcal{D} \to X$ be the operator defined by $G\boldsymbol{u} = \boldsymbol{g} + D\boldsymbol{u}$, where the linear differential operator D is defined in the above lemma, and $\boldsymbol{g} \in C^\infty([a,b], I\!R^N)$. Then the weighted iterative approximation,

$$\boldsymbol{u}_n = w_n G\boldsymbol{u}_{n-1} + (1-w_n)\boldsymbol{u}_{n-1}, \qquad w_n = \frac{1}{n^k}, \tag{7}$$

with $\boldsymbol{u}_0 \in C^\infty([a,b], I\!R^N)$, converges to a fixed point $\boldsymbol{u}^*$ if

i) for $k > 1$, $\lim\limits_{n\to\infty} \mathrm{Err}\,(n) = 0$;

ii) for $k = 1$, $\mathrm{Err}\,(n) \le \dfrac{\beta}{n^\alpha}$ when $n > N$, for some constants $\alpha > 0$, $\beta > 0$, and some integer N.

Proof. First, to show that the sequence $\{\boldsymbol{u}_n\}$ converges in X, we shall prove that it is a Cauchy sequence. Let $A_n = \|\boldsymbol{u}_n - \boldsymbol{u}_{n-1}\|$.

For the case $k > 1$, from (5) and (6), the condition $\lim\limits_{n\to\infty} \mathrm{Err}\,(n) = 0$ leads to

$$\lim_{n\to\infty} \frac{A_n}{1/n^k} = \lim_{n\to\infty} \frac{\|\boldsymbol{u}_n - \boldsymbol{u}_{n-1}\|}{w_n} = 0.$$

Since the series $\sum_{n=1}^{\infty} \frac{1}{n^k}$ converges for $k > 1$, by comparison, the series $\sum_{n=1}^{\infty} A_n$ also converges. Therefore, the sequence of its partial sum $S_n = A_1 + A_2 + \cdots + A_n$ is a Cauchy sequence, i.e., $|S_m - S_n| \to 0$ when $m, n \to \infty$.

On the other hand, for $m > n > 0$, we have

$$\begin{aligned}\|\boldsymbol{u}_m - \boldsymbol{u}_n\| &= \|\boldsymbol{u}_m - \boldsymbol{u}_{m-1} + \boldsymbol{u}_{m-1} - \boldsymbol{u}_{m-2} + \cdots + \boldsymbol{u}_{n+1} - \boldsymbol{u}_n\| \\ &\le \|\boldsymbol{u}_m - \boldsymbol{u}_{m-1}\| + \|\boldsymbol{u}_{m-1} - \boldsymbol{u}_{m-2}\| + \cdots + \|\boldsymbol{u}_{n+1} - \boldsymbol{u}_n\| \\ &= A_m + A_{m-1} + \cdots + A_{n+1} = S_m - S_n.\end{aligned}$$

Consequently $\|\boldsymbol{u}_m - \boldsymbol{u}_n\| \to 0$ when $m, n \to \infty$, which proves that $\{\boldsymbol{u}_n\}$ is a Cauchy sequence.

For the case $k = 1$, from (5), (6) and the condition (ii), we have

$$A_n = \|\boldsymbol{u}_n - \boldsymbol{u}_{n-1}\| \le \frac{\beta}{n^{1+\alpha}} \quad \text{for} \quad n > N.$$

Since $\alpha > 0$, the series $\sum_{n=1}^{\infty} \frac{\beta}{n^{1+\alpha}}$ converges and by comparison the series $\sum_{n=1}^{\infty} A_n$ also converges. Using the same arguments, it follows that $\{\boldsymbol{u}_n\}$ is a Cauchy sequence.

To prove that the limit of $\{\boldsymbol{u}_n\}$ is a fixed point of the operator G, let $\boldsymbol{u}_n \to \boldsymbol{u}^* \in X$. Since $\|G\boldsymbol{u}_n - \boldsymbol{u}_n\| \to 0$ in both cases (i) and (ii) for $k \ge 1$, it follows that $G\boldsymbol{u}_n \to \boldsymbol{u}^*$ or $\boldsymbol{g} + D\boldsymbol{u}_n \to \boldsymbol{u}^*$. By the above lemma, the operator D is closed, which implies that $\boldsymbol{u}^* \in \mathcal{D}$ and $D\boldsymbol{u}^* = \boldsymbol{u}^* - \boldsymbol{g}$. Therefore $\boldsymbol{u}^*$ is a fixed point, $G\boldsymbol{u}^* = \boldsymbol{u}^*$. □

3.1. Particular and homogeneous solutions

Theoretically if a boundary value, $\boldsymbol{u}(a)$ or $\boldsymbol{u}(b)$, is given, there exists a unique solution of the differential equation $\boldsymbol{u} = \boldsymbol{g} + D\boldsymbol{u} = G\boldsymbol{u}$. However, in the present method, no boundary value is needed and the fixed point of $\boldsymbol{u} = G\boldsymbol{u}$ obtained from the iterative approximation may depend on the initial iterate $\boldsymbol{u}_0$. Indeed, it is a particular solution of the system of linear differential equations, which may contain a part depending on the homogeneous solution (for $\boldsymbol{g} = 0$). In the following, we shall decompose the approximate solution into two parts, one of which depends only on the function $\boldsymbol{g}$, and the other is a homogeneous solution of the system.

Let the weighted iteration approximation (7) be defined by the iterative operator G_n,

$$G_n\boldsymbol{u} = w_n(\boldsymbol{g} + D\boldsymbol{u}) + (1 - w_n)\boldsymbol{u}, \tag{8}$$

and let H_n be the iterative homogeneous operator defined by

$$H_n\boldsymbol{u} = w_n(D\boldsymbol{u}) + (1 - w_n)\boldsymbol{u}. \tag{9}$$

Note that the operator H_n is linear. Then we have the following result.

Theorem 3.2. The weighted iterative approximation $\boldsymbol{u}_n = G_n \boldsymbol{u}_{n-1}$ for any initial iterate $\boldsymbol{u}_0$ can be decomposed as

$$\boldsymbol{u}_n = \boldsymbol{u}_n^* + \boldsymbol{u}_n^H, \tag{10}$$

where

$$\boldsymbol{u}_n^* = G_n G_{n-1} \cdots G_3 G_2 \, \boldsymbol{g} \tag{11}$$

and

$$\boldsymbol{u}_n^H = H_n H_{n-1} \cdots H_2 H_1 \, \boldsymbol{u}_0. \tag{12}$$

Proof. Note that $w_1 = 1$ and we have

$$\boldsymbol{u}_1 = G_1 \boldsymbol{u}_0 = w_1(\boldsymbol{g} + D\boldsymbol{u}_0) + (1 - w_1)\boldsymbol{u}_0 = \boldsymbol{g} + H_1 \boldsymbol{u}_0,$$

and

$$\begin{aligned} \boldsymbol{u}_2 &= G_2 \boldsymbol{u}_1 = G_2(\boldsymbol{g} + H_1 \boldsymbol{u}_0) \\ &= w_2(\boldsymbol{g} + D(\boldsymbol{g} + H_1 \boldsymbol{u}_0)) + (1 - w_2)(\boldsymbol{g} + H_1 \boldsymbol{u}_0) = G_2 \boldsymbol{g} + H_2 H_1 \boldsymbol{u}_0. \end{aligned}$$

By mathematical induction, let for $n = m$ the decomposition is valid, that is,

$$\boldsymbol{u}_m = G_m \cdots G_2 \boldsymbol{g} + H_m \cdots H_2 H_1 \boldsymbol{u}_0.$$

Then we have

$$\begin{aligned} \boldsymbol{u}_{m+1} &= G_{m+1}(G_m \cdots G_2 \boldsymbol{g} + H_m \cdots H_2 H_1 \boldsymbol{u}_0) \\ &= w_{m+1}(\boldsymbol{g} + D(G_m \cdots G_2 \boldsymbol{g} + H_m \cdots H_2 H_1 \boldsymbol{u}_0)) \\ &\quad + (1 - w_{m+1})(G_m \cdots G_2 \boldsymbol{g} + H_m \cdots H_2 H_1 \boldsymbol{u}_0) \\ &= w_{m+1}(\boldsymbol{g} + D(G_m \cdots G_2 \boldsymbol{g}) + (1 - w_{m+1})(G_m \cdots G_2 \boldsymbol{g}) \\ &\quad + w_{m+1} D(H_m \cdots H_2 H_1 \boldsymbol{u}_0) + (1 - w_{m+1})(H_m \cdots H_2 H_1 \boldsymbol{u}_0) \\ &= G_{m+1} G_m \cdots G_2 \boldsymbol{g} + H_{m+1} H_m \cdots H_2 H_1 \boldsymbol{u}_0. \end{aligned}$$

Therefore, the decomposition is proved. □

Remark 3.1. Note that the first part $\boldsymbol{u}_n^*$ corresponds to the iterative approximation with initial iterate $\boldsymbol{u}_0 = 0$. If the iteration converges, the first part $\boldsymbol{u}_n^*$ converges to a particular solution, while the second part $\boldsymbol{u}_n^H$ converges to a homogeneous solution of the system. Indeed, if the initial iterate $\boldsymbol{u}_0$ is a homogeneous solution, then analytically $\boldsymbol{u}_n^H$ converges trivially to itself, $\boldsymbol{u}_0$. Nevertheless, since in numerical approximation, the derivatives can not be expressed exactly, whether it would converge to itself or not will be investigated later.

3.2. The estimated error

The estimated error $\mathrm{Err}\,(n)$ defined in (6) clearly depends on the property of the operator G, namely the function $\boldsymbol{g}$ and the linear differential operator D. Although

it is a value that can be checked easily at every iterative step in a numerical scheme, it is inconvenient to analyze theoretically.

In order to characterize the estimated error (6) by the function $\boldsymbol{g}$ and the linear differential operator D, we shall restrict ourselves to the case $\mathbb{R}^N$ for $N = 1$ and $w_n = 1/n$ in the following lemma.

Lemma 3.2. Given the operator $G\boldsymbol{u} = \boldsymbol{g} + D\boldsymbol{u}$, where $\boldsymbol{g} \in C^\infty([a, b], \mathbb{R})$ and D is a linear differential operator, for the weighted iterative approximation, with $\boldsymbol{u}_0 \in C^\infty([a, b], \mathbb{R})$,

$$\boldsymbol{u}_n = w_n G\boldsymbol{u}_{n-1} + (1 - w_n)\boldsymbol{u}_{n-1}, \qquad w_n = \frac{1}{n},$$

the error for the fixed-point approximation can be expressed as

$$G\boldsymbol{u}_n - \boldsymbol{u}_n = \frac{1}{n!}\, P_n(D)\hat{\boldsymbol{g}}\,, \tag{13}$$

where

$$\hat{\boldsymbol{g}} = \boldsymbol{g} + (D\boldsymbol{u}_0 - \boldsymbol{u}_0),$$

and $P_n(D)$ is the polynomial operator defined as

$$P_n(D) = D(D + 1)(D + 2)\cdots(D + n - 1). \tag{14}$$

In other words, we have

$$\mathrm{Err}\,(n + 1) = \frac{1}{n!}\,\|P_n(D)\hat{\boldsymbol{g}}\,\|.$$

Proof. By the existence theorem of system of linear differential equations of first order, let $\hat{\boldsymbol{u}}$ be such that $\hat{\boldsymbol{u}} = \boldsymbol{g} + D\hat{\boldsymbol{u}}$ and $\boldsymbol{v}_n = \boldsymbol{u}_n - \hat{\boldsymbol{u}}$. We have

$$D\boldsymbol{v}_0 = D(\boldsymbol{u}_0 - \hat{\boldsymbol{u}}).$$

Let $\mathrm{Er}\,(n) = G\boldsymbol{u}_n - \boldsymbol{u}_n$, then we have

$$\mathrm{Er}\,(n) = \boldsymbol{g} + D\boldsymbol{u}_n - \boldsymbol{u}_n = \boldsymbol{g} + D\boldsymbol{v}_n + D\hat{\boldsymbol{u}} - \boldsymbol{v}_n - \hat{\boldsymbol{u}},$$

which gives

$$\mathrm{Er}\,(n) = (D - 1)\boldsymbol{v}_n.$$

On the other hand, we have

$$\begin{aligned}
\boldsymbol{u}_n &= w_n G\boldsymbol{u}_{n-1} + (1 - w_n)\boldsymbol{u}_{n-1} \\
&= w_n(\boldsymbol{g} + D\boldsymbol{u}_{n-1}) + (1 - w_n)\boldsymbol{u}_{n-1} \\
&= w_n(\boldsymbol{g} + D\hat{\boldsymbol{u}} + D\boldsymbol{v}_{n-1}) + (1 - w_n)(\hat{\boldsymbol{u}} + \boldsymbol{v}_{n-1}) \\
&= \hat{\boldsymbol{u}} + \boldsymbol{v}_n,
\end{aligned}$$

which leads to

$$\boldsymbol{v}_n = w_n D\boldsymbol{v}_{n-1} + (1 - w_n)\boldsymbol{v}_{n-1} = (1 - w_n + w_n D)\boldsymbol{v}_{n-1}.$$

Therefore, with $w_n = 1/n$, we obtain iteratively

$$\begin{aligned}\mathrm{Er}\,(n) &= (D-1)\boldsymbol{v}_n = (D-1)(1-w_n+w_nD)\boldsymbol{v}_{n-1} = \cdots \\ &= (D-1)(1-w_n+w_nD)(1-w_{n-1}+w_{n-1}D)\cdots(1-w_1+w_1D)\boldsymbol{v}_0 \\ &= \frac{1}{n!}(D-1)\big((D+n-1)(D+n-2)\cdots(D+1)D\big)\boldsymbol{v}_0.\end{aligned}$$

Since D is a linear operator, by the use of (14), we can write the above relation in the form,

$$\begin{aligned}\mathrm{Er}\,(n) &= \frac{1}{n!}P_n(D)(D-1)\boldsymbol{v}_0 = \frac{1}{n!}P_n(D)(D\boldsymbol{v}_0-\boldsymbol{v}_0) \\ &= \frac{1}{n!}P_n(D)(D\boldsymbol{u}_0-\boldsymbol{u}_0+\boldsymbol{g}) = \frac{1}{n!}P_n(D)\hat{\boldsymbol{g}}\,,\end{aligned}$$

which proves the lemma. □

3.3. Properties of Lagrange polynomial

Before giving an example, we shall first consider some properties of the polynomial $P_n(x)$ defined by (14). Recall that a polynomial of degree $\le n$, which takes some prescribed values at $n+1$ different points is called a Lagrange polynomial (see e.g., [3]). One can easily prove the following lemmas.

Lemma 3.3. The polynomial

$$L_n(x) = \frac{1}{n!}\,P_n(x) = \frac{1}{n!}\,\prod_{k=1}^{n}\big(x+(k-1)\big) \tag{15}$$

is the Lagrange polynomial which takes the values

$$L_n(0) = L_n(-1) = \cdots = L_n(1-n) = 0 \quad \text{and} \quad L_n(1) = 1.$$

Lemma 3.4. For any $0 \ge x \ge (1-n)$, it follows that

$$\lim_{n\to\infty} L_n(x) = 0.$$

Example. Consider the following first order linear differential equation,

$$u = Gu = e^{-x} + u'.$$

For the simple scheme (2), i.e., with weight $w_n = 1$ and $u_0 = 0$, we have

$$u_n = \begin{cases} e^{-x} & \text{if } n \text{ is odd,} \\ 0 & \text{if } n \text{ is even.}\end{cases}$$

Therefore it does not converge. However with $w_n = 1/n$, the approximation (7) gives

$$u_1 = e^{-x}, \qquad u_n = \frac{1}{2}e^{-x} \quad \text{for} \quad n \ge 2,$$

which converges trivially and, in fact, is a particular solution of the equation $u = e^{-x} + u'$.

In this case, we have $g(x) = e^{-x}$ and $\hat{g}(x) = e^{-x} + (u_0'(x) - u_0(x))$. It follows that

$$\frac{1}{n!}\,P_n(D)\hat{g} = \frac{1}{n!}\,P_n(D)e^{-x} + \frac{1}{n!}\,P_n(D)(u_0'(x) - u_0(x)).$$

The first term on the right-hand side leads to

$$\frac{1}{n!}\,P_n(D)e^{-x} = \frac{1}{n!}\sum_{k=1}^{n}(-1)^k a_k\, e^{-x} = \frac{1}{n!}\,P_n(-1)e^{-x} = L_n(-1)e^{-x} = 0,$$

by Lemma 3.3, where a_k is the coefficient of x^k of the polynomial $P_n(x)$. Furthermore, for the initial iterate u_0 given by any polynomial, the second term also goes to zero for $n \to \infty$. Therefore, by Theorem 3.1 the convergence is guaranteed. Moreover, if $u_0(x)$ is a homogeneous solution of the differential equation, i.e. $u_0 = u_0'$, the second term is identically zero, and therefore, the approximation also converges.

4. Numerical scheme

In most cases, unlike the example in the previous section, the conditions (i) or (ii) of Theorem 3.1 can hardly be verified analytically. However, such conditions can easily be verified in the numerical iterative approximations.

Let $a = x_0 < x_1 < \cdots < x_{m-1} < x_m = b$ be a division of the interval $[a, b]$ into evenly spaced subintervals of length h and denote $\boldsymbol{u}(x_j) = \boldsymbol{u}_j$. The derivative will be approximated with a central difference scheme,

$$\boldsymbol{u}'(x_j) = \frac{1}{2h}(S_+ - S_-)\boldsymbol{u}_j,$$

where S_+ and S_- are the shift operators defined by

$$S_+\boldsymbol{u}_j = \boldsymbol{u}_{j+1}, \qquad S_-\boldsymbol{u}_j = \boldsymbol{u}_{j-1}.$$

Consider the fixed-point problem

$$\boldsymbol{u} = G\boldsymbol{u} = \boldsymbol{g} + \boldsymbol{p}\,\boldsymbol{u}',$$

with $\boldsymbol{g}, \boldsymbol{p} \in C^\infty([a, b], I\!R)$ and $\boldsymbol{p} \neq 0$ in $[a, b]$. The numerical scheme of the weighted iterative approximation can be written as

$$\begin{aligned}\boldsymbol{u}_j^n &= w_n\Big(\boldsymbol{g}_j + \frac{\boldsymbol{p}_j}{2h}(S_+ - S_-)\boldsymbol{u}_j^{n-1}\Big) + (1 - w_n)\boldsymbol{u}_j^{n-1} \\ &= G_n\boldsymbol{u}_j^{n-1} = w_n\boldsymbol{g}_j + H_n\boldsymbol{u}_j^{n-1},\end{aligned} \tag{16}$$

where G_n and the homogeneous operator

$$H_n = 1 - w_n + w_n\frac{\boldsymbol{p}_j}{2h}(S_+ - S_-) \tag{17}$$

are the iterative operators defined in (8) and (9).

4.1. Stability and uniqueness of numerical solutions

We shall consider the case that the coefficient function $\boldsymbol{p}(x)$ is constant for $x \in [a,b]$ and the weight $w_n = 1/n$. We shall prove that the homogeneous part of (10) tends to zero for any initial iterate $\boldsymbol{u}_0$ in the numerical scheme.

Theorem 4.1. For the homogeneous differential equation with constant coefficient, $\boldsymbol{u}(x) = \boldsymbol{p}\,\boldsymbol{u}'(x)$, $x \in [a,b]$, with $\boldsymbol{p}$ a non-zero constant, the numerical iterative approximation with $w_n = 1/n$,

$$\boldsymbol{u}_j^n = H_n \boldsymbol{u}_j^{n-1}, \qquad H_n = 1 - w_n + w_n \frac{\boldsymbol{p}}{2h}(S_+ - S_-), \tag{18}$$

implies that

$$\boldsymbol{u}^H(x_j) = \lim_{n\to\infty} \boldsymbol{u}_n(x_j) = 0$$

for any initial iterate $\boldsymbol{u}_0 \in C^\infty([a,b], I\!R)$.

Proof. We shall apply Neumann finite Fourier analysis [13, 14] to the numerical scheme (18). Let the *discrete Fourier transform* of the grid function $\boldsymbol{u}$ be defined by ($i = \sqrt{-1}$)

$$\hat{\boldsymbol{u}}(\xi) = \sum_j \boldsymbol{u}_j e^{ij\xi}, \quad 0 \le \xi < \pi.$$

It leads to

$$\widehat{S_+\boldsymbol{u}} = e^{-i\xi}\hat{\boldsymbol{u}}, \qquad \widehat{S_-\boldsymbol{u}} = e^{i\xi}\hat{\boldsymbol{u}},$$

and from (18) it follows, with the identity $\sin\xi = (e^{i\xi} - e^{-i\xi})/2i$, that

$$\begin{aligned}\hat{\boldsymbol{u}}_n &= \Big(1 - w_n + w_n \frac{\boldsymbol{p}}{2h}(e^{-i\xi} - e^{i\xi})\Big)\hat{\boldsymbol{u}}_{n-1} \\ &= \big(1 - w_n - i\, w_n \frac{\boldsymbol{p}}{h}\sin\xi\big)\hat{\boldsymbol{u}}_{n-1}.\end{aligned}$$

It multiplied by its complex conjugate gives

$$\begin{aligned}|\hat{\boldsymbol{u}}_n|^2 &= \big((1-w_n)^2 + w_n^2\lambda\big)|\hat{\boldsymbol{u}}_{n-1}|^2 \\ &= \frac{1}{n^2}\big((n-1)^2 + \lambda\big)|\hat{\boldsymbol{u}}_{n-1}|^2,\end{aligned}$$

where we have used $w_n = 1/n$ and

$$\lambda = \big(\frac{\boldsymbol{p}}{h}\sin\xi\big)^2 \ge 0.$$

Applying this relation iteratively, we obtain

$$|\hat{\boldsymbol{u}}_n|^2 = \Big(\prod_{k=1}^{n} \frac{(k-1)^2 + \lambda}{k^2}\Big)|\hat{\boldsymbol{u}}_0|^2 = R_n(\lambda)|\hat{\boldsymbol{u}}_0|^2,$$

where

$$R_n(\lambda) = \prod_{k=1}^{n} \frac{(k-1)^2 + \lambda}{k^2}.$$

In the following lemma, we shall prove that $\lim_{n\to\infty} R_n(\lambda) = 0$ for any $\lambda \geq 0$. Therefore, $\lim_{n\to\infty} \hat{\boldsymbol{u}}_n = 0$, which implies $\lim_{n\to\infty} \boldsymbol{u}_n = 0$ and the theorem is proved. □

Lemma 4.1. For any $\lambda \geq 0$, $\lim_{n\to\infty} R_n(\lambda) = 0$.

Proof. Let m be a positive integer and $\lambda < m$. Then for $k \geq m+1$, We have

$$(k-1)^2 + \lambda \leq (k-1)^2 + m = (k^2 - k) - (k - m - 1) \leq k(k-1).$$

On the other hand, for $m < n$, we have

$$R_n(\lambda) = \prod_{k=1}^{n} \frac{(k-1)^2 + \lambda}{k^2} = \Big(\prod_{k=1}^{m} \frac{(k-1)^2 + \lambda}{k^2}\Big)\Big(\prod_{k=m+1}^{n} \frac{(k-1)^2 + \lambda}{k^2}\Big).$$

Since the first part is a product of finite terms, it is bounded and let its value be denoted by M. Therefore, by the above estimate, we have

$$\begin{aligned} R_n(\lambda) = \prod_{k=1}^{n} \frac{(k-1)^2 + \lambda}{k^2} &= M \prod_{k=m+1}^{n} \frac{(k-1)^2 + \lambda}{k^2} \leq M \prod_{k=m+1}^{n} \frac{k(k-1)}{k^2} \\ &\leq M \frac{m}{m+1} \frac{m+1}{m+2} \cdots \frac{n-2}{n-1} \frac{n-1}{n} = \frac{Mm}{n}, \end{aligned}$$

which tends to zero as $n \to \infty$. □

Remark 4.1. If $\boldsymbol{u}_0(x)$ is a non-trivial homogeneous solution of $\boldsymbol{u}(x) = \boldsymbol{p}\,\boldsymbol{u}'(x)$, since the derivative is approximated by the numerical scheme, it is equivalent to small perturbations of the exact derivatives in the iterative approximation. The fact that any such perturbations will lead to the trivial null solution, can be stated as instability of failing to remain close to the initial homogeneous solution $\boldsymbol{u}_0(x)$.

Corollary 4.1. (Stability). For the numerical scheme (18), the only numerically stable homogeneous solution of $\boldsymbol{u} = \boldsymbol{p}\,\boldsymbol{u}'$ with constant $\boldsymbol{p} \neq 0$ is the trivial solution $\boldsymbol{u}^H = 0$.

This result, together with Theorem 3.2, implies the following uniqueness theorem for the above numerical iterative approximation irrespective of the initial iterate function[2].

Corollary 4.2. (Uniqueness). For the fixed-point problem $\boldsymbol{u} = \boldsymbol{g} + \boldsymbol{p}\,\boldsymbol{u}'$, with constant $\boldsymbol{p} \neq 0$ sufficiently small, if the numerical iterative scheme (16) converges, it converges uniquely independent of any particular choice of initial iterate $\boldsymbol{u}_0$. The (unique) limit $\boldsymbol{u}^* = \lim_{n\to\infty} \boldsymbol{u}_n^*$ from (11) corresponds to the starting iterate $\boldsymbol{u}_0 = 0$.

[2] In Sect. §23 of [6] concerning the iterative approximation based on *infinite differentiation*, it is stated that "there is one particular solution which is especially important, and that for a certain [fairly broad] class of initial iterates and for small enough [coefficient] the method of differential iteration converges to that particular solution". This particular solution is referred to as the *asymptotic* solution (see the final remarks in Sect. §27 of [6]).

4.2. Remarks on variable coefficient function

We have proved that $R_n(\lambda)$ is bounded for $\lambda \geq 0$ and tends to zero as n tends to infinity. However, one can see that for large values of λ the maximum value of $R_n(\lambda)$ increases rapidly at the beginning stage of the iterations when $n < m$. Therefore, although $R_n(\lambda)$ will eventually tends to zero as $n \to \infty$, its value may become too large too soon, resulting in great errors, and spoil the subsequent iterative calculations. Nevertheless, in our numerical simulations, for some moderate values of λ, the errors grow within reasonable bound and converge to zero eventually so that Theorem 4.1 remains valid.

For the case of variable coefficient function $\boldsymbol{p}(x)$, such that $\boldsymbol{p}(x) \neq 0$, if $|\boldsymbol{p}(x)|$ is small enough, the above analysis can be regarded as a *local* stability consideration, which is a necessary condition for the global stability analysis under quite general conditions (for more information, see [13]). In the following numerical applications, we shall verify that Theorem 4.1 remains valid, at least numerically, for the case of variable coefficient function $\boldsymbol{p}(x)$.

5. Heat conduction of ideal gases in moment theory

For ideal gases, the moment theory can be obtained either from Grad distribution for the Boltzmann equation (see e.g., [4, 7]) or from the theory of *extended thermodynamics* [9, 11]. We shall consider the example, treated in [10] with the 14-moment theory – one-dimensional stationary heat conduction between two coaxial cylindrical walls.

Specifically, we consider the boundary value problem of heat conduction in a monatomic gas at rest between two coaxial cylinders, with radii r_0 and r_1. In this case, we have the following basic equations [10],

$$\begin{cases} \dfrac{dq}{dr} + \dfrac{1}{r}\, q = 0, \\ q = -\Big(\dfrac{5}{2}\,\tau R p + \dfrac{14}{5}\,\tau^2 R\,\dfrac{q}{r}\Big)\dfrac{d\theta}{dr} - \dfrac{1}{6}\,\tau\dfrac{d\boldsymbol{u}}{dr}, \quad \text{for} \quad r_0 \leq r \leq r_1, \\ \boldsymbol{u} = -28\tau R q \dfrac{d\theta}{dr}, \end{cases} \tag{19}$$

and the boundary conditions,

$$\theta(r_1) = \theta_1, \qquad q(r_0) = q_0. \tag{20}$$

The equation $(19)_1$ is the conservation of energy, while the relation $(19)_2$ can be regarded as a generalization of Fourier law of heat conduction, with the thermal conductivity $\kappa = (5/2)R\tau p$. The relation $(19)_3$ is the additional balance equation for the non-equilibrium fourth order moment $\boldsymbol{u}$ (denoted by Δ in [10]). The pressure p is constant, θ is the temperature and τ is the relaxation time of the gas. The gas constant is denoted by R (the Boltzmann constant divided by the mass

of the gas molecule).

It is well-known [10, 15] that the boundary conditions (20), which are well-posed in the classical Fourier theory, consisting of equations $(19)_{1,2}$ by neglecting second order terms in τ when $\tau \to 0$, are not sufficient for the uniqueness of the above problem in moment theory, and indeed, the additional boundary value, say, $\boldsymbol{u}(r_0)$, is needed. Such boundary data are referred to as *uncontrollable* parameters, because from the physical viewpoint, their assignment will lead to undesirable results as pointed out in [10].

We introduce the following dimensionless quantities,

$$\tilde{r} = \frac{r}{r_0}, \qquad \tilde{\theta} = \frac{\theta}{\theta_1}, \qquad \tilde{q} = \frac{q}{p\sqrt{R\theta_1}}, \qquad \tilde{\boldsymbol{u}} = \frac{\boldsymbol{u}}{pR\theta_1},$$

and the Knudsen number that characterizes the classical theory as the limit when $\mathtt{Kn} \to 0$,

$$\mathtt{Kn} = \frac{\tau}{r_0}\sqrt{R\theta_1}.$$

In terms of dimensionless variables (neglecting the tilde for simplicity), the last two equations of (19) becomes

$$\begin{aligned} q &= -\mathtt{Kn}\Big(\frac{5}{2} + \frac{14}{5}\,\mathtt{Kn}\,\frac{q}{r}\Big)\frac{d\theta}{dr} - \frac{1}{6}\,\mathtt{Kn}\,\frac{d\boldsymbol{u}}{dr}, \\ \boldsymbol{u} &= -28\,\mathtt{Kn}\,q\,\frac{d\theta}{dr}, \end{aligned} \tag{21}$$

for $1 < r < r_1/r_0$, while $(19)_1$ can be solved explicitly with the boundary condition $(20)_2$,

$$q = \frac{c}{r}, \qquad q(1) = c = \frac{q_0}{p\sqrt{R\theta_1}},$$

where the constant c depends on the assigned boundary values and the physical data of the gas. Eliminating q and $d\theta/dr$ from (21), we obtain

$$\boldsymbol{u}(r) = \boldsymbol{g}(r) + \boldsymbol{p}(r)\,\frac{d\boldsymbol{u}}{dr}, \tag{22}$$

where

$$\boldsymbol{g}(r) = \frac{56}{5}\,\frac{\dfrac{c^2}{r^2}}{1 + \dfrac{28}{25}\,\mathtt{Kn}\,\dfrac{c}{r^2}}, \qquad \boldsymbol{p}(r) = \frac{28}{15}\,\mathtt{Kn}\,\frac{\dfrac{c}{r}}{1 + \dfrac{28}{25}\,\mathtt{Kn}\,\dfrac{c}{r^2}}.$$

The first order differential equation (22) takes the form of the fixed-point problem (1), for which no boundary value of $\boldsymbol{u}$ is given, because, in the kinetic theory, it is the higher moments that we know nothing about physically.

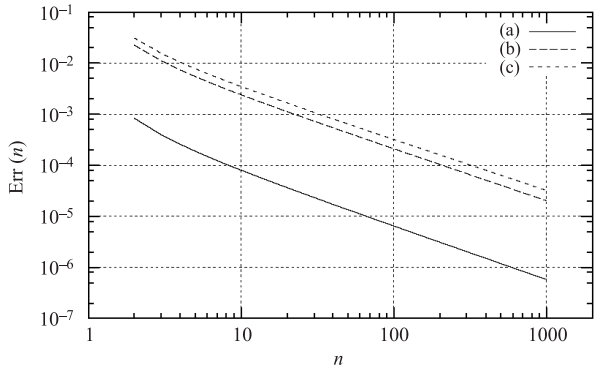

Figure 1. The plot of estimated error Err (n) in log-log scale for different initial iterates: (a) $\boldsymbol{u}_0(r) = 0$, (b) $\boldsymbol{u}_0(r) = \sin \pi(r-1)$, (c) $\boldsymbol{u}_0(r) = \exp(r-1)$.

5.1. Numerical iterative solutions

There are previous attempts [1, 5, 10, 15] to determine the uncontrollable parameter by some additional physical assumptions or criteria, mostly with unsatisfactory results [5, 15] or cumbersome numerical procedures [10]. Unlike those attempts, the present iterative solution does not require any additional assumption. At each iteration, $\boldsymbol{u}_n(r)$ at the interior mesh points are calculated from the iterative scheme (16), and by continuity, extrapolation formulas with interior mesh points are used to determine the value of $\boldsymbol{u}_n(r)$ at the end points. The mesh points are evenly spaced with length h.

For the numerical calculations the following data are used:

$$\mathtt{Kn} = 0.1, \qquad c = 0.1, \qquad 1 \le r \le 3, \qquad h = \frac{1}{100}, \qquad w_n = \frac{1}{n}.$$

Three cases of initial iterates are considered:

$$\text{(a)} \quad \boldsymbol{u}_0(r) = 0, \quad \text{(b)} \quad \boldsymbol{u}_0(r) = \sin \pi(r-1), \quad \text{(c)} \quad \boldsymbol{u}_0(r) = \exp(r-1).$$

Fig. 1 shows the estimated error Err(n) for the three cases of initial iterates. Note that for $n > 10$, the curves in the log-log scale plot are almost parallel straight lines with negative slope, i.e., they can be represented by the equation

$$\log(\text{Err}\,(n)) = \log \beta - \alpha \log n$$

for some $\alpha > 0$ and $\beta > 0$, and it follows that

$$\text{Err}\,(n) = \frac{\beta}{n^\alpha} \quad \text{for} \quad n > 10.$$

Therefore, the condition (ii) of Theorem 3.1 for the convergence of the iterative approximation is satisfied.

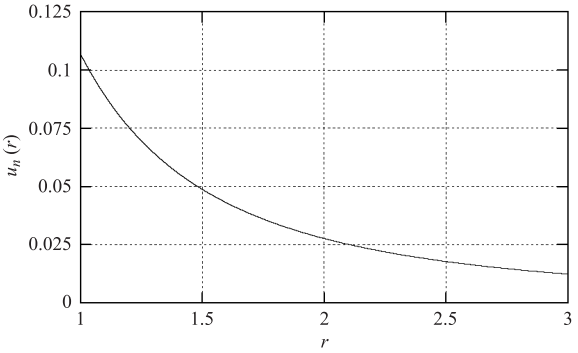

Figure 2. The approximate solutions $\boldsymbol{u}_n(r)$ at $n = 1000$ for all three different cases (a), (b), and (c). Their curves are practically identical.

On the other hand, for all three different cases of initial iterates, the numerical results show that the approximate solutions $\boldsymbol{u}_n(r)$ for large values of n are identical to within an insignificant error. The profile of $\boldsymbol{u}_n(r)$ at $n = 1000$ is plotted in Fig. 2. This confirms our observation (in Sect. 4.2) of uniqueness that Corollary 4.2 might be valid for the case of small enough variable coefficient function $\boldsymbol{p}(r)$. In this example, we have $|\boldsymbol{p}(r)| < 2h$ for $1 \le r \le 3$ and $h = 0.01$.

In order to further assure the uniqueness of iterative solution independent of some particular choice of initial iterates, the homogeneous part of the iterative solutions are also calculated for the case (b) and (c). The results are plotted in Fig. 3. It can be seen that the homogeneous parts at $n = 1000$ for both cases are nearly zero, i.e., $\|\boldsymbol{u}_n^H\| < 6{\times}10^{-5}$ insignificantly compared with the initial iterates. Therefore, in this example the validity of Corollary 4.1 is also justified.

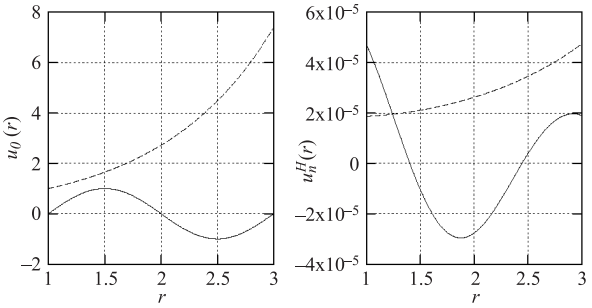

Figure 3. The initial iterates $\boldsymbol{u}_0(r)$ and the corresponding iterations $\boldsymbol{u}_n^H(r)$ at $n = 1000$ for the cases: (b) solid lines, (c) dashed lines.

5.2. Other observations

Physically the null initial iterate is the state that the non-equilibrium higher order moments vanish, corresponding to the knudsen number $\mathtt{Kn} \to 0$, which reduces the problem to the classical theory. On the other hand, since moment theories are regarded as improvements of the classical theory for moderate Knudsen numbers, it is essential for the iterative approximation to obtain the (unique) solution starting from the classical theory. In [10] an iterative minimization scheme was proposed, also starting with the null initial iterate. Their results are consistent with ours.

From Fig. 1, it can be seen that the estimated error is about 10^{-4} at $n = 10$ (for null initial iterate), which means that the numerical solution at the 10-th iteration is practically the same as that in Fig. 2. Therefore, although we have run the iterations up to very large n for the sake of convergence, the process of iterations can be terminated by checking the estimated error to within an acceptable tolerance for practical purpose. It usually can be achieved with very few iterations.

In the example, we have taken $w_n = 1/n$. However, from numerical viewpoint, if instead, we have taken $w_n = 1/n^k$ for some k slightly greater then 1, say $k = 1.001$, there are no any significant changes in our numerical simulations. In other words, the condition (i) of Theorem 3.1 for convergence and Corollary 4.2 for uniqueness can also be verified.

In Sect. 4, the analysis is based on the numerical scheme with central difference scheme for the derivative. Numerical schemes with forward or backward difference schemes have not been analyzed. However, in our numerical simulations, both these schemes have been tested and the results are similar.

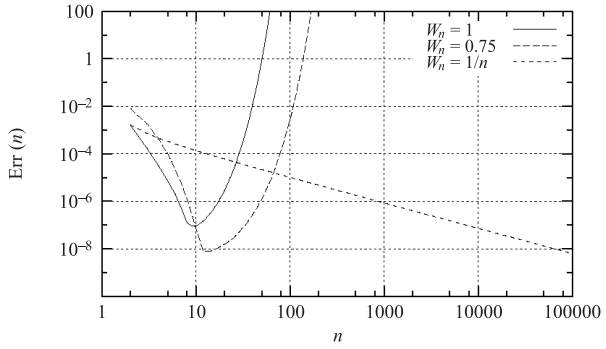

Figure 4. Comparison of estimated errors Err (n) vs. n in log-log scale for the schemes with constant weights $w_n = 1$, $w_n = 0.75$ as well as with decreasing weight $w_n = 1/n$.

6. Remarks on approximate solutions and convergence

In order to emphasize the differences between the iterative approximation with decreasing weight and the simpler schemes (2) and (4) with constant weights, the respective estimated errors are plotted in Fig. 4 for the numerical data: $\mathtt{Kn} = 0.2$, $c = 0.1$, $h = 0.01$, $\boldsymbol{u}_0(r) = 0$. Note that for the approximations with constant weights, the errors blow up for $n > 10$ showing the eventual divergence of the approximations, while for the approximation with decreasing weight the convergence is evident.

On the other hand, for the cases of constant weights, the errors are very small after few iterations. Indeed, from Fig. 4, one can see that the same estimated error (about 10^{-7}) for both cases of constant weights, $w_n = 1$ and $w_n = 0.75$, at $n = 10$ can only be achieved roughly at $n = 9,000$ for the scheme with decreasing weight. As noted in the previous section, the iterative scheme with $w_n = 1/n$ converges but obviously with a much slower rate.

It is amazing that even through the iterative approximations (2) and (4) do not converge, they do give rather good approximate solutions at the first few iterations. From practical point of view, this may be regarded as a justification to obtain numerical solutions by terminating the approximations to within an acceptable tolerance of error, without further assurance concerning the convergence of the approximations.

If we recall the remark in the footnote 1 about the similarity between the simple iterative scheme (2) and the Maxwellian iterations in the kinetic theory, which usually proceed with few iterations to obtain more general physical laws, this observation may seem quite interesting.

References

[1] Brini, F. and Ruggeri, T., Entropy principle for the moment systems of degree α associated to the Boltzmann equation. Critical derivatives and non controllable boundary data, *Continuum Mech. Thermodyn.* **14** (2002), 165–189.

[2] Browder, F. E. and Petryshyn, V., The solution by iteration of nonlinear funcional equations in Banach spaces, *Bull. Amer. Math. Soc.* **72** (1966), 571–575.

[3] Burden, R. L. and Faires, J. D., *Numerical Analysis*, sixth edition, Brooks-Cole Publishing Co. Pacific Grove, California 1997.

[4] Grad, H., On the kinetic theory of rarefied gases, *Commun. Pure Appl. Math.* **2** (1949), 331–407.

[5] Grmela, M., Karlin, I. V. and Zmievski, V. B., Boundary layer variational principles: A case study, *Phys. Review* E, **66**, 011201 (2002).

[6] Ikenberry, E. and Trusdell, C., On the pressures and the flux of energy in a gas according to Maxwell's kinetic theory, I & II *J. Rational Mech. Anal.* **5** (1956), 1–54, 55–128.

[7] Kogan, M. N., *Rarefied Gas Dynamics*, Plenum Press, New York 1969.

[8] Kreyszig, E., *Introductory Functional Analysis with Applications*, Wiley, New York 1978.

[9] Liu, I-Shih and Müller, I., Extended thermodynamics of classical and degenerate gases, *Arch. Rational Mech. Anal.* **83** (1983), 285–332.

[10] Liu, I-Shih, Rincon, M. A. and Müller, I., Iterative approximation of stationary heat conduction in extended thermodynamics, *Continuum Mech. Thermodyn.* **14** (2002), 483–493.
[11] Müller, I. and Ruggeri, T., *Rational Extended Thermodynamics*, 2nd edition, Springer, New York 1998.
[12] Ortega, J. M. and Reinboldt, W. C., *Iterative Solutions of Non-Linear Equations in Several Variables*, Academic Press, New York, London 1970.
[13] Richtmyer, R. D. and Morton, K. W., *Difference Methods for Initial Value Problems*, John Wiley (Interscience), New York 1967.
[14] Sod, G. A., *Numerical Methods in Fluid Dynamics*, Cambridge University Press, Cambridge-New York 1985.
[15] Struchtrup, H. and Weiss, W., Maximum of the local entropy production becomes minimal in stationary processes, *Phy. Rev. Lett.* **80** (1998), 5048–5051.
[16] Yosida, K., *Functional Analysis*, Springer, New York 1968.

I-Shih Liu and Daniel R. Vieira
Instituto de Matemática
Universidade Federal do Rio de Janeiro
21945-970, Rio de Janeiro
Brasil
e-mail: liu@im.ufrj.br

(Received: January 5, 2006)

Published Online First: August 23, 2006