Fourier series and spectral-finite difference methods for the general linear diffusion equation

by

Alet Roux

Submitted in partial fulfilment of the requirements of the degree

Baccalaureus Scientiae (Honores)
Applied Mathematics

in the

Faculty of Natural and Agricultural Science
University of Pretoria
Pretoria

Supervisor
Prof J. M-S. Lubuma

January 2002

Acknowledgement

I wish to thank my supervisor, Prof. J. M-S. Lubuma, for his enthusiasm, kindness and patience while working on this project. His suggestions and words of encouragement are central to the successful completion of this work.

I am grateful to Dr. E. Rapoo, of UNISA, for acting as external examiner for this work and for her valuable suggestions.

I am also indebted to Johan van der Berg, for his help with proofreading, to Leon van der Merwe, for our collaboration on Hilbert-Schmidt theory, and to my family and friends for their support.

Alet Roux

Pretoria, 28 January 2002

Table of Contents

Introduction

Data aequatione quotcunque fluentes quantitae involvente fluxiones invenire et vice versa.
[It is useful to differentiate functions and to solve differential equations.]

- Isaac Newton (Letter to Leibniz, 1676)

This work is an attempt at a comprehensive analysis, from both a theoretical and a numerical point of view, of the general linear diffusion problem

$$\frac{\partial u}{\partial t} - c\frac{\partial^2 u}{\partial x^2} + bu = f \text{ on } (0,2\pi)\times(0,T)$$
$$u(x,0) = u_0(x)$$
$$u(0,t) = u(2\pi,t).$$

There are several methods of solving evolution problems, e.g. Laplace transform methods, semigroup methods, variational methods (cf. Dautray & Lions (2000b)). In this work, we employ the Fourier series or diagonalization method, described by Gustafson (1980: p. xi) as one of "the usual trinities". This method is extensively used in the engineering and applied science, and is often called the method of separation of variables (cf. Creese & Haralick (1978: Section 1D), Greenberg (1978: Section 21.1), Sirovich (1988: Section 7.2)).

Regarding the numerical discretization in the space variable *x*, we use the spectral Galerkin method described by Canuto et al (1988), of which the Fourier series method turns out to be a good motivation. We then couple the spectral method with a standard (cf. Raviart & Thomas (1983: Section 7.5)) and non-standard (cf. Mickens (1994: Chapter 4)) finite difference discretization in the time variable *t*.

The work is organised as follows:

The Hilbert-Schmidt theory of self-adjoint compact linear operators in Hilbert spaces, presented in **Chapter 1**, is a strong motivation to the use of Fourier series and spectral methods. We apply this theory to the Sturm-Liouville problem, as well as to a related boundary value problem with periodic boundary conditions.

In **Chapter 2**, we prove the existence and uniqueness of solutions to the general linear diffusion initial-boundary value problem, using a specific Galerkin method that is related to the Fourier series method.

In view of the theoretical work done in Chapters 1 and 2, it is natural to use the semi-discrete spectral method of Fourier-Galerkin type described in **Chapter 3**. We study the convergence of this method, and present some error estimates.

We then depart temporarily from the general linear diffusion equation. **Chapter 4** starts with some numerical approximations of initial value problems for ordinary differential equations, using the classical θ-method. In addition, we study the concept of elementary stability, and with regard to this property, design an original powerful variant of the θ-method: the non-standard θ-method.

In **Chapter 5**, we return to the general linear diffusion problem. For the semi-discrete approximation, we employ the semi-discrete spectral method developed in Chapter 3. Coupling it with the θ-method discussed in Chapter 4, we obtain a spectral-θ-method, for which we present a stability result as well as error estimates. Finally, we present a new method: the spectral-non-standard θ-method.

> *We have not succeeded in answering all our problems. The answers we have found only serve to raise a whole set of new questions. In some ways, we feel we are as confused as ever, but we believe we are confused on a higher level and about more important things.*
>
> (Posted outside the mathematics reading room, Tromsø University)

Chapter 1. Self-adjoint compact linear operators in Hilbert spaces

Hilbert-Schmidt theory constitutes a solid foundation for and strong motivation to spectral methods. In Section 1.1, we present this theory, essentially following Zeidler (1995: pp. 229-237). In Section 1.2, the Hilbert-Schmidt theory is applied to the Sturm-Liouville problem. A similar boundary value problem with periodic boundary conditions is also considered.

1.1. Hilbert-Schmidt theory

Throughout this section, we consider a complex or real Hilbert space $\mathbf{H} \neq \{0\}$ with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$, and a linear operator $T : \mathbf{H} \to \mathbf{H}$.

We first investigate the properties of self-adjoint operators.

Definition 1.1.1. Self-adjoint operator. The operator T is called self-adjoint (or symmetric) if

$$\langle Tx, y \rangle = \langle x, Ty \rangle \text{ for all } x \in \mathbf{H} \text{ and } y \in \mathbf{H}. \tag{1.1}$$

Remark 1.1.2. If the domain of the operator T is a proper subspace of $\mathbf{H}$, then the concepts of self-adjointness and symmetry for T are distinct (cf. Zeidler (1995: p. 264)).

For a self-adjoint operator T, we have

$$\langle Tx, x \rangle \text{ is real for all } x \in \mathbf{H}. \tag{1.2}$$

Indeed, for any $x \in \mathbf{H}$,

$$\langle Tx, x \rangle = \langle x, Tx \rangle = \overline{\langle Tx, x \rangle}. \tag{1.3}$$

In view of the terminology that we will employ frequently, it is worth stating the following theorem, which is proved by Davis (1963: Theorem 8.9.1):

Theorem 1.1.3. Let $\{w_k\}_{k\in\mathbf{N}}$ be a system of orthonormal vectors in **H**. The following statements are equivalent:

(a) The space spanned by $\{w_k\}_{k\in\mathbf{N}}$ is dense in **H**.

(b) The system $\{w_k\}_{k\in\mathbf{N}}$ is complete, i.e. if $v \in \mathbf{H}$ is such that $\langle v, w_k \rangle = 0$ for all $k \in \mathbf{N}$, then $v = 0$.

(c) The Fourier series $\sum_{k\in\mathbf{N}} \langle v, w_k \rangle w_k$ of any $v \in \mathbf{H}$ converges to v in **H**, i.e.

$$v = \sum_{k\in\mathbf{N}} \langle v, w_k \rangle w_k \ . \tag{1.4}$$

(d) The Parseval identity holds, i.e. for all $v \in \mathbf{H}$,

$$\|v\|^2 = \sum_{k\in\mathbf{N}} |\langle v, w_k \rangle|^2 \ . \tag{1.5}$$

(e) The extended Parseval identity holds, i.e. for all v and w in **H**,

$$\langle v, w \rangle = \sum_{k\in\mathbf{N}} \langle v, w_k \rangle \overline{\langle w, w_k \rangle} . \tag{1.6}$$

Following Schwartz (1979: p. 29), we consider the next definition.

Definition 1.1.4. Hilbert basis. Any orthonormal system $\{w_k\}_{k\in\mathbf{N}}$ that satisfies one of the equivalent statements in Theorem 1.1.3 is called a Hilbert basis of **H**.

The following existence result will be used.

Theorem 1.1.5. Any separable Hilbert space $\mathbf{H} \neq \{0\}$ admits at least one Hilbert basis (cf. Kreyszig (1978: pp. 168 & 171), Schwartz (1979: p. 30)).

The eigenvectors and eigenvalues of self-adjoint operators have some special properties, as depicted in the next result.

Proposition 1.1.6. Suppose that T is self-adjoint. Then

(a) All the eigenvalues of T are real.

(b) Any two eigenvectors of T associated with different eigenvalues are orthogonal.

(c) If $\{w_k\}_{k\in\mathbf{N}}$ is a system of eigenvectors of T that form a Hilbert basis of **H**, then the corresponding system $\{\lambda_k\}_{k\in\mathbf{N}}$ of eigenvalues contains all the eigenvalues of T.

Proof. (a) Consider any eigenvalue λ of T with associated eigenvector x. Then $\lambda\langle x, x\rangle = \langle Tx, x\rangle$, i.e. $\lambda = \dfrac{\langle Tx, x\rangle}{\langle x, x\rangle}$, which is real in view of (1.2).

(b) Suppose that $Tx = \lambda x$ and $Ty = \mu y$ for some eigenvalues $\lambda \neq \mu$. Then

$$\begin{aligned}\langle x, y\rangle &= \frac{1}{\lambda - \mu}\big(\langle \lambda x, y\rangle - \langle x, \mu y\rangle\big)\\ &= \frac{1}{\lambda - \mu}\big(\langle Tx, y\rangle - \langle x, Ty\rangle\big)\\ &= 0\end{aligned}$$

due to the symmetry of T. Hence x and y are orthogonal.

(c) If $\{w_k\}_{k\in\mathbf{N}}$ is a Hilbert basis, then, by Theorem 1.1.3, any $v \in \mathbf{H}$ can be represented in the form (1.4). Suppose, by contradiction, that $Tv = \lambda v$ for some $\lambda \notin \{\lambda_k\}_{k\in\mathbf{N}}$ and $v \neq 0$. Then, by (b), $\langle v, w_k\rangle = 0$ for all $k \in \mathbf{N}$. It follows that $v = 0$ because the system $\{w_k\}_{k\in\mathbf{N}}$ is complete. This contradicts the assumption. Thus $\lambda \in \{\lambda_k\}_{k\in\mathbf{N}}$.

■

The norm of a self-adjoint operator in a Hilbert space can also be expressed in a specific way.

Proposition 1.1.7. If T is self-adjoint, then it is bounded and its norm is given by

$$\|T\| = \sup_{x\in\mathbf{H},\|x\|=1} |\langle Tx, x\rangle|. \tag{1.7}$$

Proof. The claim on boundedness of T is known as the Hellinger-Toeplitz theorem (Kreyszig, 1978: p. 525). Let $(x_k)_{k=1}^{\infty}$ be a sequence in $\mathbf{H}$ such that $\lim_{k\to\infty} x_k = x$ and $\lim_{k\to\infty} Tx_k = y$ in $\mathbf{H}$. For any $z \in \mathbf{H}$, we have

$$\begin{aligned}\langle y, z\rangle &= \lim_{k\to\infty}\langle Tx_k, z\rangle && (\text{Continuity of inner product})\\ &= \lim_{k\to\infty}\langle x_k, Tz\rangle && (\text{Symmetry of } T)\\ &= \langle x, Tz\rangle && (\text{Continuity of inner product})\\ &= \langle Tx, z\rangle && (\text{Symmetry of } T).\end{aligned}$$

Since z is arbitrary, we have $y = Tx$, which shows that the operator T is closed. The closed graph theorem (Kreyszig, 1978: p. 292) permits to conclude that T is bounded.

By the Cauchy-Schwarz inequality and the boundedness of T,

$$\sup_{x\in\mathbf{H},\|x\|=1} |\langle Tx, x\rangle| \leq \|T\|. \tag{1.8}$$

For any $z \in \mathbf{H}$ and a $\lambda > 0$ (to be specified shortly), let $v_+ := \lambda z + \frac{1}{\lambda} Tz$ and $v_- := \lambda z - \frac{1}{\lambda} Tz$. Then $T^2 z = \frac{\lambda}{2}(Tv_+ - Tv_-)$ and $z = \frac{1}{2\lambda}(v_+ + v_-)$, so that

$$\begin{aligned}
\|Tz\|^2 &= \langle T^2 z, z \rangle && \text{(Symmetry of } T\text{)} \\
&= \frac{1}{4}\langle Tv_+ - Tv_-, v_+ + v_- \rangle \\
&= \frac{1}{4}\left| \langle Tv_+, v_+ \rangle - \langle Tv_-, v_- \rangle \right| \\
&\le \frac{1}{4}\left(\left| \langle Tv_+, v_+ \rangle \right| + \left| \langle Tv_-, v_- \rangle \right| \right) && \text{(Triangle inequality)} \\
&= \frac{1}{4}\left(\|v_+\|^2 \left| \left\langle T\left(\frac{1}{\|v_+\|} v_+ \right), \frac{1}{\|v_+\|} v_+ \right\rangle \right| + \|v_-\|^2 \left| \left\langle T\left(\frac{1}{\|v_-\|} v_- \right), \frac{1}{\|v_-\|} v_- \right\rangle \right| \right) \\
&\le \frac{1}{4}\left(\|v_+\|^2 + \|v_-\|^2 \right) \sup_{x \in \mathbf{H}, \|x\|=1} \left| \langle Tx, x \rangle \right| \\
&= \frac{1}{2}\left(\langle \lambda z, \lambda z \rangle + \left\langle \frac{1}{\lambda} Tz, \frac{1}{\lambda} Tz \right\rangle \right) \sup_{x \in \mathbf{H}, \|x\|=1} \left| \langle Tx, x \rangle \right| \\
&= \frac{1}{2}\left(\lambda^2 \|z\|^2 + \frac{1}{\lambda^2} \|Tz\|^2 \right) \sup_{x \in \mathbf{H}, \|x\|=1} \left| \langle Tx, x \rangle \right|.
\end{aligned}$$

Assume that $Tz \neq 0$ (if $Tz = 0$, the result is trivial). If $\lambda := \sqrt{\|Tz\|}$, then

$$\begin{aligned}
\|Tz\|^2 &\le \frac{1}{2}\left(\|Tz\| \|z\|^2 + \|Tz\| \right) \sup_{x \in \mathbf{H}, \|x\|=1} \left| \langle Tx, x \rangle \right| \\
\|Tz\| &\le \frac{1}{2}\left(\|z\|^2 + 1 \right) \sup_{x \in \mathbf{H}, \|x\|=1} \left| \langle Tx, x \rangle \right|,
\end{aligned}$$

so that

$$\|T\| \le \sup_{x \in \mathbf{H}, \|x\|=1} \left| \langle Tx, x \rangle \right|. \tag{1.9}$$

Combining (1.8) and (1.9), we obtain (1.7).

■

We now present some properties of the null space of a self-adjoint operator.

Lemma 1.1.8. Suppose that T is self-adjoint. Then

(a) The null space $\mathbf{N}(T) := \{x \in \mathbf{H} \mid Tx = 0\}$ of T and its orthogonal complement $\mathbf{N}(T)^{\perp}$ are closed subspaces of **H**.

(b) The space $\mathbf{N}(T)^{\perp}$ is invariant with respect to T.

(c) The restricted operator $T\big|_{\mathbf{N}(T)^{\perp}}$ is injective.

Proof. (a) The null space $\mathbf{N}(T) = T^{-1}\{0\}$ is a closed subspace of **H** because T is continuous and $\{0\}$ is a closed set in **H**.

Let $\{y_k\}_{k=1}^{\infty}$ be a sequence in $\mathbf{N}(T)^{\perp}$ that converges to some $y \in \mathbf{H}$. For any $x \in \mathbf{N}(T)$, we have, due to the continuity of the inner product,

$$\langle y, x \rangle = \lim_{k \to \infty} \langle y_k, x \rangle = 0. \tag{1.10}$$

Thus $\langle y, x \rangle = 0$ for all $x \in \mathbf{H}$, so that $y \in \mathbf{N}(T)^{\perp}$. Therefore $\mathbf{N}(T)^{\perp}$ is a closed subspace of **H**.

(b) If $x \in \mathbf{N}(T)^{\perp}$ then, for all $w \in \mathbf{N}(T)$, due to the symmetry of T, we have

$$\langle Tx, w \rangle = \langle x, Tw \rangle = 0\,. \tag{1.11}$$

Hence $Tx \in \mathbf{N}(T)^{\perp}$, and $T(\mathbf{N}(T)^{\perp}) \subseteq \mathbf{N}(T)^{\perp}$.

(c) If $Tx = 0$ and $x \in \mathbf{N}(T)^{\perp}$, then $x \in \mathbf{N}(T) \cap \mathbf{N}(T)^{\perp}$, and $x = 0$. Hence $T\big|_{\mathbf{N}(T)^{\perp}}$ is injective.

■

From now on, we will need a further assumption on T.

Definition 1.1.9. Compact operator. The operator T is called compact if, for every bounded set $\mathbf{M} \subseteq \mathbf{H}$, $T(\mathbf{M})$ is relatively compact, i.e. $\overline{T(\mathbf{M})}$ is compact in **H**.

Remark 1.1.10. A compact linear operator is necessarily bounded. Indeed, suppose that T is compact. Consider the bounded set $\mathbf{S}(0,1) := \{x \in \mathbf{H} \mid \|x\| = 1\}$. By Definition 1.1.9, the set $\overline{T(\mathbf{S}(0,1))}$ is compact in **H** and thus bounded. Hence there exists a $C > 0$ such that $\|Tx\| \le C$ for all $x \in \mathbf{S}(0,1)$. If $y \in \mathbf{H} \setminus \{0\}$ and $x := \dfrac{1}{\|y\|} y$, then $x \in \mathbf{S}(0,1)$. Therefore

$$\left\| T\left(\frac{1}{\|y\|} y \right) \right\| = \|Tx\| \le C, \tag{1.12}$$

which shows that $\|Ty\| \le C\|y\|$.

The following lemma will be instrumental in the proof of the main result, Theorem 1.1.14, below.

Lemma 1.1.11. If T is nonzero, self-adjoint and compact, then there exists an eigenvalue λ and corresponding eigenvector $x \in \mathbf{H}$ such that

$$|\lambda| = \|T\| \text{ and } \|x\| = 1. \tag{1.13}$$

Proof. By Proposition 1.1.7 and the definition of supremum, there exists a sequence $\{v_k\}_{k=1}^{\infty}$ with

$$\|v_k\| = 1 \text{ for all } k \in \mathbf{N}, \tag{1.14}$$

such that

$$\lim_{k\to\infty} |\langle Tv_k, v_k \rangle| = \|T\|. \tag{1.15}$$

The sequence of real numbers (cf. (1.2)) $\{\langle Tv_k, v_k \rangle\}_{k=1}^{\infty}$ is then bounded. By the Bolzano-Weierstrass theorem, there exists a subsequence $\{v_k'\}_{k=1}^{\infty}$ of $\{v_k\}_{k=1}^{\infty}$ and a real number λ that

$$\lim_{k\to\infty} \langle Tv_k', v_k' \rangle = \lambda. \tag{1.16}$$

Given (1.15) and (1.16), we have $|\lambda| = \|T\|$. Thus $\|Tv_k'\| \le |\lambda|$ for all $k \in \mathbf{N}$. Furthermore, we have

$$\begin{aligned} \lim_{k\to\infty} \|Tv_k' - \lambda v_k'\|^2 &= \lim_{k\to\infty} \left(\|Tv_k'\|^2 - 2\lambda \langle Tv_k', v_k' \rangle + \lambda^2 \|v_k'\|^2 \right) \\ &\le \lim_{k\to\infty} \left(\lambda^2 - 2\lambda \langle Tv_k', v_k' \rangle + \lambda^2 \right) \\ &= 0, \end{aligned}$$

which shows that

$$\lim_{k\to\infty} (Tv_k' - \lambda v_k') = 0. \tag{1.17}$$

Since T is compact, it follows from (1.14) that there exists a subsequence $\{v_k''\}_{k=1}^{\infty}$ of $\{v_k'\}_{k=1}^{\infty}$ such that $\{Tv_k''\}_{k=1}^{\infty}$ is convergent. Due to (1.17) and to the fact that $\lambda \neq 0$, the sequence $\{v_k''\}_{k=1}^{\infty}$ also converges to some $x \in \mathbf{H}$, which, in view of (1.14), satisfies $\|x\| = 1$. This implies that $Tx = \lambda x$, and that (1.13) is satisfied.

■

Corollary 1.1.12. Assume that **H** has infinite dimension. If T is nonzero, self-adjoint and compact, then there exist sequences of eigenvalues $(\lambda_k)_{k=1}^{\infty}$ and corresponding eigenvectors $\{w_k\}_{k\in\mathbf{N}}$ such that

$$|\lambda_{m+1}| = \left\|T\big|_{\mathbf{V}_m}\right\| \tag{1.18}$$

where

$$\mathbf{V}_m := \begin{cases} \left\{x \in (\mathbf{N}(T))^{\perp} \middle| \langle x, w_k \rangle = 0 \text{ for } k = 1,2,\mathrm{K}\,,m\right\} & \text{if } m \in \mathbf{N} \\ \mathbf{H} & \text{if } m = 0 \end{cases}, \tag{1.19}$$

$$\mathbf{V}_m \text{ is invariant with respect to } T, \tag{1.20}$$

$$|\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \geq \Lambda \geq |\lambda_{m+1}| \geq \Lambda\ , \tag{1.21}$$

and

$$\{w_k\}_{k\in\mathbf{N}} \text{ is an orthonormal system in } (\mathbf{N}(T))^{\perp}. \tag{1.22}$$

Proof. Two different cases will be considered.

Case 1. The operator *T* is injective, i.e. 0 is not an eigenvalue of *T*

We will prove (1.18)-(1.21) by induction. For m = 0, Lemma 1.1.11 guarantees existence of an eigenvalue λ_1 and associated eigenvector w_1 of T such that (1.18)-(1.21) are satisfied.

Suppose, for $m \in \mathbf{N}$ fixed, that sequences of eigenvalues $(\lambda_k)_{k=1}^{m+1}$ and corresponding eigenvectors $\{w_k\}_{k=1}^{m+1}$ have been found such that (1.18)-(1.21) hold true. Since H has infinite dimension, it follows that $\mathbf{V}_{m+1} \neq \{0\}$. By the injectivity of T, $T\big|_{\mathbf{V}_{m+1}} \neq 0$. Hence Lemma 1.1.11 applies to $T\big|_{\mathbf{V}_{m+1}}$, and there exists an eigenvalue λ_{m+2} and eigenvector w_{m+2} of T with

$$w_{m+2} \in \mathbf{V}_{m+1},\ \|w_{m+2}\| = 1 \tag{1.23}$$

and

$$\begin{aligned} |\lambda_{m+2}| &= \left\|T\big|_{\mathbf{V}_{m+1}}\right\| \\ &\leq \left\|T\big|_{\mathbf{V}_m}\right\| \\ &= |\lambda_{m+1}|. \end{aligned}$$

Therefore, by induction, the sequences $(\lambda_k)_{k=1}^{\infty}$ and $\{w_k\}_{k\in\mathbf{N}}$ satisfying (1.18)-(1.21) can be constructed. Moreover, (1.22) follows from (1.23).

Case 2. The operator *T* is not injective, i.e. 0 is an eigenvalue of *T*

According to Lemma 1.1.8, $T\big|_{(\mathbf{N}(T))^{\perp}}$ satisfies the assumption of Case 1. Hence the results of Case 1 can be applied to $T\big|_{(\mathbf{N}(T))^{\perp}}$ to obtain (1.18)-(1.22).

■

Proposition 1.1.13. Assume that **H** is separable and has infinite dimension, and T is nonzero, self-adjoint and compact. Then

(a) If the set of eigenvalues $(\lambda_k)_{k=1}^{\infty}$ of T in Corollary 1.1.12 is countable, i.e. it has a one-to-one correspondence with **N**, then $\lim_{k\to\infty}\lambda_k = 0$.

(b) The associated eigenvectors $\{w_k\}_{k\in\mathbf{N}}$ of T in Corollary 1.1.12 may be chosen and augmented so that the new system, still denoted by $\{w_k\}_{k\in\mathbf{N}}$, form a Hilbert basis of **H**.

Proof. (a) Suppose, by contradiction, that there exists a constant $C > 0$ such that $|\lambda_k| \geq C$ for all $k \in \mathbf{N}$. Then the sequence $\left(\frac{1}{\lambda_k} w_k\right)_{k=1}^{\infty}$ is bounded. Furthermore, by the compactness of T, the sequence $(w_k)_{k=1}^{\infty} = \left(T\left(\frac{1}{\lambda_k} w_k\right)\right)_{k=1}^{\infty}$ contains a convergent subsequence. This is impossible since, by (1.22), for $m \neq n$,

$$\|w_m - w_n\|^2 = \|w_m\|^2 + \|w_n\|^2 = 2, \tag{1.24}$$

and therefore $(w_k)_{k=1}^{\infty}$ cannot even contain a Cauchy subsequence. Thus $\lim_{k\to\infty}\lambda_k = 0$.

(b) For any $x \in (\mathbf{N}(T))^{\perp}$, and $m \in \mathbf{N}$, let

$$v_m := x - \sum_{k=1}^{m} \langle x, w_k \rangle w_k . \tag{1.25}$$

Fix $x \in \mathbf{V}_m$ (cf. (1.19)). It follows from (1.18) and Remark 1.1.10 that

$$\|Tx\| \leq \left\|T\big|_{\mathbf{V}_m}\right\| \|x\| = |\lambda_{m+1}| \|x\| . \tag{1.26}$$

Since $v_m \in \mathbf{V}_m$, we have, in view of (1.22) and (1.25), that

$$\|v_m\|^2 = \|x\|^2 - \sum_{k=1}^{m} |\langle x, w_k \rangle|^2 . \tag{1.27}$$

Therefore $\|v_m\| \leq \|x\|$. Using (a) and (1.26), we then obtain

$$\begin{aligned} \lim_{m\to\infty} \|Tv_m\| &\leq \lim_{m\to\infty} (|\lambda_{m+1}| \|v_m\|) \\ &\leq \|x\| \lim_{m\to\infty} |\lambda_{m+1}| \\ &= 0. \end{aligned}$$

Fix any $x \in (\mathbf{N}(T))^{\perp}$. Then

$$\lim_{m\to\infty}\left(Tx-\sum_{k=1}^{m}\langle x,w_k\rangle\lambda_k w_k\right)=\lim_{m\to\infty}\left(Tx-\sum_{k=1}^{m}\langle x,w_k\rangle Tw_k\right)$$
$$=\lim_{m\to\infty}Tv_m$$
$$=0.$$

It follows that $Tx=\sum_{k\in\mathbf{N}}\langle x,w_k\rangle\lambda_k w_k$ for all $x\in(\mathbf{N}(T))^{\perp}$.

The Bessel inequality

$$\sum_{k=1}^{m}\left|\langle x,w_k\rangle\right|^2\le\|x\|^2 \text{ for } m\in\mathbf{N} \tag{1.28}$$

follows directly from (1.27). Therefore the Fourier series $\sum_{k\in\mathbf{N}}\langle x,w_k\rangle w_k$ is convergent for all $x\in(\mathbf{N}(T))^{\perp}$. Hence there exists some $y\in\mathbf{H}$ such that $y=\sum_{k\in\mathbf{N}}\langle x,w_k\rangle w_k$. Actually, $y\in(\mathbf{N}(T))^{\perp}$ because $\sum_{k=1}^{m}\langle x,w_k\rangle w_k\in(\mathbf{N}(T))^{\perp}$ and $(\mathbf{N}(T))^{\perp}$ is closed. Then $Tx = Ty$, and, by the injectivity of $T\big|_{(\mathbf{N}(T))^{\perp}}$, it follows that $x = y$. Therefore

$$x=\sum_{k\in\mathbf{N}}\langle x,w_k\rangle w_k\,. \tag{1.29}$$

In view of Theorem 1.1.3,

$$\{w_k\}_{k\in\mathbf{N}} \text{ is a Hilbert basis of } (\mathbf{N}(T))^{\perp}. \tag{1.30}$$

If, on the one hand, $\mathbf{N}(T)=\{0\}$, then $\mathbf{H}=(\mathbf{N}(T))^{\perp}$, and $\{w_k\}_{k\in\mathbf{N}}$ is a Hilbert basis of $\mathbf{H}$. Suppose now, on the other hand, that $\mathbf{N}(T)\neq\{0\}$. Theorem 1.1.5 guarantees the existence of a Hilbert basis $\{u_k\}_{k\in\mathbf{N}}$ of $\mathbf{N}(T)$, with $Tu_k=0$ for all $k\in\mathbf{N}$. For each $x\in\mathbf{H}$, there exists a unique decomposition

$$x = y + z \text{ with } y\in\mathbf{N}(T) \text{ and } z\in(\mathbf{N}(T))^{\perp}. \tag{1.31}$$

Using (1.30), it follows that

$$x=y+z$$
$$=\sum_{k\in\mathbf{N}}\langle y,u_k\rangle u_k+\sum_{k\in\mathbf{N}}\langle z,w_k\rangle w_k$$
$$=\sum_{k\in\mathbf{N}}\langle x,u_k\rangle u_k+\sum_{k\in\mathbf{N}}\langle x,w_k\rangle w_k.$$

Consequently $\{u_k\}_{k\in\mathbf{N}}\cup\{w_k\}_{k\in\mathbf{N}}$ represents a Hilbert basis of $\mathbf{H}$.

■

A further property of the eigenvalues $(\lambda_k)_{k=1}^{\infty}$ of T is that each nonzero eigenvalue has finite multiplicity. Indeed, in view of Proposition 1.1.13(b) and Proposition 1.1.6(c), any eigenvalue $\lambda \neq 0$ of T must be equal to some λ_m. In view of Proposition 1.1.13(a), there exist integers M and N ($M \leq N$) such that

$$\lambda_M = \lambda_{M+1} = \Lambda = \lambda_m = \Lambda = \lambda_{N-1} = \lambda_N \tag{1.32}$$

and $\lambda_j \neq \lambda_m$ for $j < M$ and $j > N$. If $x \neq 0$ is any eigenvector associated with λ, then, due to Proposition 1.1.6(b),

$$\begin{aligned} x &= \frac{1}{\lambda} Tx \\ &= \frac{1}{\lambda} \sum_{k=M}^{N} \lambda_k \langle x, w_k \rangle w_k \\ &= \sum_{k=M}^{N} \langle x, w_k \rangle w_k . \end{aligned}$$

Hence $\{w_k\}_{k=M}^{N}$ form a (Hamel) basis (cf. Kreyszig (1978: pp. 54-55)) for the space $\{x \in \mathbf{H} | Tx = \lambda x\}$, and thus

$$\lambda \text{ has multiplicity } N - M + 1. \tag{1.33}$$

With the material collected so far, we are in a position to state the main result, or the Hilbert-Schmidt theory for self-adjoint compact operators in Hilbert spaces.

Theorem 1.1.14. Hilbert-Schmidt theory. Assume that **H** is separable. Suppose that T is self-adjoint and compact. Then

(a) All the eigenvalues of T are real.

(b) The eigenvalues of T form a sequence $(\lambda_k)_{k=1}^{\infty}$ such that $(|\lambda_k|)_{k=1}^{\infty}$ is decreasing. If the set of eigenvalues is countable, then $\lim_{k \to \infty} \lambda_k = 0$.

(c) The multiplicity of each nonzero eigenvalue is finite.

(d) With each eigenvalue λ_k can be associated an eigenvector w_k such that the collection $\{w_k\}_{k \in \mathbf{N}}$ is a Hilbert basis of **H**.

(e) Each eigenvalue λ_k can be expressed in terms of the Rayleigh quotient, i.e.

$$|\lambda_1| = \sup_{x \in \mathbf{H} \setminus \{0\}} \frac{|\langle Tx, x \rangle|}{\langle x, x \rangle} \text{ and } |\lambda_k| = \sup_{\substack{x \in \mathbf{H} \setminus \{0\} \\ x \perp \operatorname{span}\{w_1, w_2, \ldots, w_{k-1}\}}} \frac{|\langle Tx, x \rangle|}{\langle x, x \rangle}. \tag{1.34}$$

(f) Any two eigenvectors of T corresponding to different eigenvalues are orthogonal.

Proof. We give the proof for a Hilbert space **H** with infinite dimension. The proof of the finite dimensional case entails an obvious simplification (see e.g. Lubuma (1994)).

(a) This is a direct consequence of Proposition 1.1.6(a).

(b) This is a direct consequence of (1.21) and Proposition 1.1.13(a).

(c) See (1.33).

(d) This is a direct consequence of Proposition 1.1.13(b).

(e) After careful comparison of (1.7) and (1.19), (1.34) follows.

(f) This is a direct consequence of Proposition 1.1.6(b).

■

1.2. Sturm-Liouville problem

The Sturm-Liouville problem provides a direct application of Hilbert-Schmidt theory. It is also central to the application of spectral methods. In this section, we investigate some qualitative and quantitative properties of this problem. We will also determine the properties of the solutions to the associated eigenvalue problem.

Given the real-valued functions

$$p \in \mathbf{C}^1[0,2\pi],\ q \in \mathbf{C}^0[0,2\pi] \text{ and } f \in \mathbf{L}^2(0,2\pi), \tag{1.35}$$

we consider the ordinary differential equation

$$-\frac{d}{dx}\left(p\frac{du}{dx}\right)+qu=f \text{ on } (0,2\pi) \tag{1.36}$$

with general boundary conditions

$$\left.\begin{aligned}\alpha_1 u(0)+\beta_1\frac{du}{dx}(0)+\chi_1 u(2\pi)+\delta_1\frac{du}{dx}(2\pi)=0\\ \alpha_2 u(2\pi)+\beta_2\frac{du}{dx}(2\pi)+\chi_2 u(0)+\delta_2\frac{du}{dx}(0)=0.\end{aligned}\right\} \tag{1.37}$$

The analysis of the boundary value problem (1.36)-(1.37) is done by e.g. Dieudonné (1980: Section XIV.8). In this study, we will focus on two particular cases of the boundary conditions (1.37), as described below.

Firstly, we consider the boundary conditions

$$\alpha_1 u(0)+\beta_1\frac{du}{dx}(0)=0 \tag{1.38}$$

$$\alpha_2 u(2\pi)+\beta_2\frac{du}{dx}(2\pi)=0\,, \tag{1.39}$$

(i.e. $\chi_1=\delta_1=\chi_2=\delta_2=0$ in (1.37)), which, together with (1.36), form a Sturm-Liouville problem.

Definition 1.2.1. Regular Sturm-Liouville problem. (cf. Canuto et al (1988: p. 282)) The Sturm-Liouville problem (1.36), (1.38)-(1.39) is called regular if

$$p(x)>0 \text{ and } q(x)\geq 0 \text{ for all } x\in[0,2\pi] \tag{1.40}$$

and

$$|\alpha_1|+|\beta_1|>0 \text{ and } |\alpha_2|+|\beta_2|>0\,. \tag{1.41}$$

Remark 1.2.2. Dautray & Lions (2000a: Section 2.7) give a detailed discussion of the regular Sturm-Liouville problem (1.36), (1.38)-(1.39) with $p = 1$.

Remark 1.2.3. The problem (1.36), (1.38)-(1.39) together with (1.40), where $p(0)=0$ or $p(2\pi)=0$, is a singular Sturm-Liouville problem. For such a problem, the boundary conditions (1.38)-(1.39) must be suitably changed due to the singularity of the solution at 0 or 2π. Typically, if $p(0)=0$, we have the boundary conditions (1.39) together with

$$\lim_{x\to 0^+} p(x)\frac{du}{dx}(x)=0, \tag{1.42}$$

or, if $p(2\pi)=0$, we have (1.38) with

$$\lim_{x\to 2\pi^-} p(x)\frac{du}{dx}(x)=0. \tag{1.43}$$

The second type of boundary value problem that we consider is the following: We couple the differential equation (1.36) with periodic boundary condition

$$u(0)=u(2\pi) \tag{1.44}$$

(i.e. $\alpha_1=\chi_2=1$, $\alpha_2=\chi_1=-1$ and $\beta_1=\delta_1=\beta_2=\delta_2=0$ in (1.37)).

To solve these problems, we use the Lax-Milgram Lemma. Firstly, we consider two definitions, and then present the Lax-Milgram Lemma in a form suitable to our needs.

Definition 1.2.4. Antilinear form. Let **V** be a complex vector space. A form $l:\mathbf{V}\to\mathbf{C}$ is called antilinear if

$$l(\alpha v)=\overline{\alpha}l(v) \text{ for each scalar } \alpha \text{ and } v\in\mathbf{V}. \tag{1.45}$$

Definition 1.2.5. Hermitian form, sesquilinear form. (cf. Kreyszig (1978: pp. 191 & 195)) Let **V** be a complex vector space. A form $a:\mathbf{V}\times\mathbf{V}\to\mathbf{C}$ is called sesquilinear if it is linear in its first argument and antilinear in its second argument, i.e. for scalars α and ß and all $v\in\mathbf{V}$, $w\in\mathbf{V}$ and $x\in\mathbf{V}$, we have

$$a(\alpha v+\beta w,x)=\alpha a(w,x)+\beta a(w,x) \text{ and } a(v,\alpha w+\beta x)=\overline{\alpha}a(v,w)+\overline{\beta}a(v,x). \tag{1.46}$$

The sesquilinear form a is called hermitian if it satisfies the additional condition

$$a(v,w)=\overline{a(w,v)} \text{ for } v\in\mathbf{V},\ w\in\mathbf{V}. \tag{1.47}$$

Theorem 1.2.6. Lax-Milgram Lemma. Let **V** be a complex Banach space with norm $\|\cdot\|$, and let l be a bounded antilinear form on **V**. Suppose that $a:\mathbf{V}\times\mathbf{V}\to\mathbf{C}$ is a hermitian form. We assume that a is

- Continuous, i.e. there exists a constant $K\geq 0$ such that

$$|a(v,w)|\leq K\|v\|\|w\| \text{ for all } v\in\mathbf{V} \text{ and } w\in\mathbf{V}, \tag{1.48}$$

- **V**-elliptic or **V**-coercive in the sense that there exists an $\alpha > 0$ such that

$$a(v,v)\geq\alpha\|v\|^2 \text{ for all } v\in\mathbf{V}. \tag{1.49}$$

Then the variational problem to find u such that

$$\left.\begin{array}{c} u\in\mathbf{V} \\ a(u,v)=l(v) \text{ for all } v\in\mathbf{V} \end{array}\right\} \tag{1.50}$$

is well-posed, i.e. there exists a unique solution $u\in\mathbf{V}$ that depends continuously upon the data l, i.e.

$$\|u\|\leq C\|l\|. \tag{1.51}$$

for some constant $C > 0$.

Proof. With (1.49), the hermitian form a clearly defines an inner product on **V**. Moreover, the norm $\|v\|_a := \sqrt{a(v,v)}$ induced by a is equivalent to $\|\cdot\|$ since, by (1.48) and (1.49), we have

$$\sqrt{\alpha}\|v\|\leq\|v\|_a\leq\sqrt{K}\|v\|. \tag{1.52}$$

The space **V**, being complete with respect to $\|\cdot\|$, is also complete with respect to $\|\cdot\|_a$. Hence **V** is a Hilbert space with respect to a.

On the other hand, $\bar{l}:\mathbf{V}\to\mathbf{C}$ defined by $\bar{l}(v):=\overline{l(v)}$ is a bounded linear form on **V**. By Riesz' Representation Theorem, there exists a unique $u\in\mathbf{V}$ such that

$$a(v,u)=\overline{l(v)} \text{ for all } v\in\mathbf{V} \tag{1.53}$$

and

$$\|u\|=\|\bar{l}\|=\|l\|. \tag{1.54}$$

But

$$l(v)=\overline{a(v,u)}=a(u,v). \tag{1.55}$$

Therefore (1.50) and (1.51) are satisfied.

■

For the regular Sturm-Liouville problem (1.36), (1.38)-(1.41), we consider the following space of test and trial functions:

$$\mathbf{V} := \begin{cases} \mathbf{H}_0^1(0,2\pi) := \left\{ v \in \mathbf{H}^1(0,2\pi) \middle| v(0) = 0, v(2\pi) = 0 \right\} & \text{if } \beta_1 = 0, \beta_2 = 0 \\ \left\{ v \in \mathbf{H}^1(0,2\pi) \middle| v(0) = 0 \right\} & \text{if } \beta_1 \neq 0, \beta_2 = 0 \\ \left\{ v \in \mathbf{H}^1(0,2\pi) \middle| v(2\pi) = 0 \right\} & \text{if } \beta_1 = 0, \beta_2 \neq 0 \\ \mathbf{H}^1(0,2\pi) & \text{if } \beta_1 \neq 0, \beta_2 \neq 0 \end{cases}. \tag{1.56}$$

In (1.56), for $m \in \mathbf{N}$,

$$\mathbf{H}^m(0,2\pi) := \left\{ v \in \mathbf{L}^2(0,2\pi) \middle| \frac{d^k v}{dx^k} \in \mathbf{L}^2(0,2\pi) \text{ for } k = 0,1,\mathrm{K}\ ,m \right\} \tag{1.57}$$

is the Sobolev space of order m equipped with the norm and inner product

$$\|v\|_m = \sqrt{\sum_{k=0}^{m} \left\| \frac{d^k v}{dx^k} \right\|_0^2} \text{ and } \langle v, w \rangle_m = \sum_{k=0}^{m} \left\langle \frac{d^k v}{dx^k}, \frac{d^k w}{dx^k} \right\rangle_0, \tag{1.58}$$

with

$$\|v\|_0 := \sqrt{\int_0^{2\pi} |v(x)|^2 dx} \text{ and } \langle v, w \rangle_0 := \int_0^{2\pi} v(x) \overline{w(x)} dx \tag{1.59}$$

the norm and inner product of $\mathbf{L}^2(0,2\pi)$.

For the periodic problem (1.36) & (1.44), we assume in addition to (1.35) that p is periodic, i.e.

$$p \in \mathbf{C}_{\text{per}}^0[0,2\pi] := \left\{ v \in \mathbf{C}^0[0,2\pi] \middle| v(0) = v(2\pi) \right\}. \tag{1.60}$$

In this case, the space of test and trial functions is

$$\mathbf{H}_{\text{per}}^1(0,2\pi) := \left\{ v \in \mathbf{H}^1(0,2\pi) \middle| v(0) = v(2\pi) \right\}. \tag{1.61}$$

Let us assume that a classical solution $u \in \mathbf{C}^2[0,2\pi]$ of the regular Sturm-Liouville problem (1.36), (1.38)-(1.41) exists. Let $v \in \mathbf{V}$ be an arbitrary test function. Multiplication by $\bar{v}$ of (1.36) and integration by parts yield, using (1.38)-(1.39),

$$a(u,v) = \langle f, v \rangle_0, \tag{1.62}$$

with the sesquilinear form a defined by

$$a(v,w) := \left\langle p \frac{dv}{dx}, \frac{dw}{dx} \right\rangle_0 + \langle qv, w \rangle_0 - \gamma_1 p(0) v(0) \overline{w(0)} + \gamma_2 p(2\pi) v(2\pi) \overline{w(2\pi)}, \tag{1.63}$$

where

$$\gamma_k := \begin{cases} \dfrac{\alpha_k}{\beta_k} & \text{if } \beta_k \neq 0 \\ 0 & \text{if } \beta_k = 0 \end{cases} \text{ for } k = 1, 2. \tag{1.64}$$

The sesquilinear form a is clearly hermitian.

Definition 1.2.7. Sturm-Liouville variational problem. The problem of finding a function u such that

$$\left.\begin{array}{c} u \in \mathbf{V} \\ a(u,v) = \langle f, v \rangle_0 \text{ for all } v \in \mathbf{V}, \end{array}\right\} \tag{1.65}$$

is called the variational formulation of the Sturm-Liouville problem (1.36), (1.38)-(1.41).

We will use the following result (cf. Dautray & Lions (2000a: pp. 42-44)) in the proof of Theorem 1.2.9 below:

Lemma 1.2.8. For every $\varepsilon > 0$, there exists a constant $C \equiv C_\varepsilon > 0$ such that

$$\|v\|_\infty^2 \le \varepsilon \left\| \frac{dv}{dx} \right\|_0^2 + C\|v\|_0^2 \text{ for all } v \in \mathbf{H}^1(0,2\pi), \tag{1.66}$$

where $\|\cdot\|_\infty$ denotes the supremum norm.

Theorem 1.2.9. There exists an $M \ge 0$ such that if

$$p(x) \ge M \text{ and } q(x) \ge M \text{ for all } x \in [0,2\pi], \tag{1.67}$$

then the Sturm-Liouville variational problem (1.65) is well-posed.

Proof. By the Cauchy-Schwarz inequality, the antilinear form $l(v) := \langle f, v \rangle_0$ and the sesquilinear form a are both clearly continuous on $\mathbf{H}^1(0,2\pi)$. Moreover, we infer from (1.40) that there exists an $M_1 > 0$ such that

$$\inf_{x \in [0,2\pi]} p(x) \ge M_1. \tag{1.68}$$

Regarding the **V**-ellipticity of a, we shall now distinguish between two cases.

Case 1: $\beta_1 = 0$ and $\beta_2 = 0$ in (1.56), and thus $\gamma_1 = 0$ and $\gamma_2 = 0$ in (1.63)

In this case, $\mathbf{V} = \mathbf{H}_0^1(0,2\pi)$. We may take $M = 0$ in (1.67). For $v \in \mathbf{H}_0^1(0,2\pi)$, we have

$$\begin{aligned} a(v,v) &= \left\langle p\frac{dv}{dx}, \frac{dv}{dx} \right\rangle_0 + \langle qv, v \rangle_0 \\ &\ge \left\langle p\frac{dv}{dx}, \frac{dv}{dx} \right\rangle_0 \\ &\ge M_1 \left\| \frac{dv}{dx} \right\|_0^2 \\ &\ge K\|v\|_1^2 \end{aligned}$$

because of the Poincaré-Friedrichs inequality. This proves the **V**-ellipticity of a in this case.

Case 2: $\beta_1 \neq 0$ or $\beta_2 \neq 0$ in (1.56), and thus $\gamma_1 \neq 0$ or $\gamma_2 \neq 0$ in (1.63)

Let $v \in \mathbf{V}$. From (1.63), we have

$$\begin{aligned} a(v,v) &\geq \left\langle p\frac{dv}{dx}, \frac{dv}{dx} \right\rangle_0 + \langle qv, v \rangle_0 - \left| \gamma_2 p(2\pi)|v(2\pi)|^2 - \gamma_1 p(0)|v(0)|^2 \right| \\ &\geq \inf_{x\in[0,2\pi]} \min\{p(x), q(x)\} \left(\left\| \frac{dv}{dx} \right\|_0^2 + \|v\|_0^2 \right) - \left| \gamma_2 p(2\pi)|v(2\pi)|^2 - \gamma_1 p(0)|v(0)|^2 \right| \\ &\geq \inf_{x\in[0,2\pi]} \min\{p(x), q(x)\} \left(\left\| \frac{dv}{dx} \right\|_0^2 + \|v\|_0^2 \right) - (|\gamma_1| + |\gamma_2|) \|p\|_\infty \|v\|_\infty^2. \end{aligned}$$

Using Lemma 1.2.8 with $\varepsilon := \dfrac{\beta}{(|\gamma_1| + |\gamma_2|)\|p\|_\infty}$ for some $\beta \in (0,1)$, we obtain

$$\begin{aligned} a(v,v) &\geq \inf_{x\in[0,2\pi]} \min\{p(x), q(x)\} \left(\left\| \frac{dv}{dx} \right\|_0^2 + \|v\|_0^2 \right) - (|\gamma_1| + |\gamma_2|) \|p\|_\infty \left(\varepsilon \left\| \frac{dv}{dx} \right\|_0^2 + C\|v\|_0^2 \right) \\ &= \left(\inf_{x\in[0,2\pi]} \min\{p(x), q(x)\} - C(|\gamma_1| + |\gamma_2|)\|p\|_\infty \right) \|v\|_0^2 \\ &\quad + \left(\inf_{x\in[0,2\pi]} \min\{p(x), q(x)\} - \varepsilon(|\gamma_1| + |\gamma_2|)\|p\|_\infty \right) \left\| \frac{dv}{dx} \right\|_0^2 \\ &= \left(\inf_{x\in[0,2\pi]} \min\{p(x), q(x)\} - \beta\frac{C}{\varepsilon} \right) \|v\|_0^2 + \left(\inf_{x\in[0,2\pi]} \min\{p(x), q(x)\} - \beta \right) \left\| \frac{dv}{dx} \right\|_0^2 \\ &> \left(\inf_{x\in[0,2\pi]} \min\{p(x), q(x)\} - \frac{C}{\varepsilon} \right) \|v\|_0^2 + \left(\inf_{x\in[0,2\pi]} \min\{p(x), q(x)\} - 1 \right) \left\| \frac{dv}{dx} \right\|_0^2 \qquad (\beta < 1). \end{aligned}$$

On taking $M > \max\left\{\dfrac{C}{\varepsilon}, 1\right\}$ in (1.67), it follows that a is $\mathbf{H}^1(0,2\pi)$-elliptic.

By Theorem 1.2.6, there exists a unique $u \in \mathbf{H}^1(0,2\pi)$ that satisfies (1.65) and depends continuously upon f.

■

We now present a regularity result.

Theorem 1.2.10. The solution u of the Sturm-Liouville variational problem (1.65) has the regularity $u \in \mathbf{H}^2(0,2\pi)$.

Proof. Suppose that u is the solution of (1.65). Fix any $v \in \mathbf{D}(0,2\pi)$, where $\mathbf{D}(0,2\pi)$, the space of test functions, consists of infinitely differentiable functions with compact support in $(0,2\pi)$. In view of (1.63) and of the definition of the distributional derivative, it follows that

$$-\frac{d}{dx}\left(p\frac{du}{dx}\right)+qu=f \text{ in } \mathbf{D}'(0,2\pi), \tag{1.69}$$

where $\mathbf{D}'(0,2\pi)$, the space of distributions on $(0,2\pi)$, consists of linear continuous functionals on $\mathbf{D}(0,2\pi)$ equipped with the canonical topology of L. Schwartz.

As $f\in\mathbf{L}^2(0,2\pi)$, it follows from (1.69) that $-\frac{d}{dx}\left(p\frac{du}{dx}\right)+qu\in\mathbf{L}^2(0,2\pi)$. With $qu\in\mathbf{L}^2(0,2\pi)$, we have $\frac{d}{dx}\left(p\frac{du}{dx}\right)\in\mathbf{L}^2(0,2\pi)$. Since $\frac{d}{dx}\left(p\frac{du}{dx}\right)=\frac{dp}{dx}\frac{du}{dx}+p\frac{d^2u}{d^2x}$, where $\frac{dp}{dx}\frac{du}{dx}\in\mathbf{L}^2(0,2\pi)$ and p is bounded from below, we have $u\in\mathbf{H}^2(0,2\pi)$.

■

Using the same method as with the Sturm-Liouville problem (1.36), (1.38)-(1.39), we will now solve the periodic boundary value problem (1.36) & (1.44).

If we define the continuous hermitian form

$$a(v,w):=\left\langle p\frac{dv}{dx},\frac{dw}{dx}\right\rangle_0+\langle qv,w\rangle_0, \tag{1.70}$$

then the variational formulation of the boundary value problem (1.36) & (1.44) is to find a function u such that

$$\left.\begin{aligned} u&\in\mathbf{H}^1_{\text{per}}(0,2\pi)\\ a(u,v)&=\langle f,v\rangle_0 \text{ for all } v\in\mathbf{H}^1_{\text{per}}(0,2\pi).\end{aligned}\right\} \tag{1.71}$$

We still need an assumption of the form (1.67). In this case, it will be sufficient for M to be any positive value. Indeed, for any $M > 0$,

$$\begin{aligned} a(v,v)&\geq M\left(\left\|\frac{dv}{dx}\right\|_0^2+\|v\|_0^2\right)\\ &=M\|v\|_1^2,\end{aligned}$$

which shows that a is $\mathbf{H}^1(0,2\pi)$-elliptic. By Theorem 1.2.6, we have proved the following result:

Theorem 1.2.11. Under the condition (1.67), with M an arbitrary positive number, the variational problem (1.71) is well-posed.

Remark 1.2.12. The proofs of Theorem 1.2.9 and Theorem 1.2.11 motivate that we assume throughout this work that the condition (1.67) holds.

In a similar fashion to Theorem 1.2.10, we obtain the following:

Theorem 1.2.13. The solution u of the variational problem (1.71) has the regularity $u \in \mathbf{H}^2(0,2\pi)$.

Remark 1.2.14. The regularity $u \in \mathbf{H}^2(0,2\pi)$ in Theorem 1.2.10 and Theorem 1.2.13 permits us to give an interpretation of the variational problems (1.65) and (1.71).

If we replace the function f in (1.65) by the expression given in (1.69), we obtain, for all $v \in \mathbf{V}$,

$$a(u,v) = \left\langle -\frac{d}{dx}\left(p\frac{du}{dx}\right) + qu, v \right\rangle_0 \tag{1.72}$$

or (cf. (1.63))

$$\left.\begin{aligned} \left\langle p\frac{du}{dx}, \frac{dv}{dx}\right\rangle_0 + \langle qu, v\rangle_0 - \gamma_1 p(0)u(0)\overline{v(0)} \\ + \gamma_2 p(2\pi)u(2\pi)\overline{v(2\pi)} = \left\langle -\frac{d}{dx}\left(p\frac{du}{dx}\right) + qu, v\right\rangle_0 . \end{aligned}\right\} \tag{1.73}$$

But integration by parts yields

$$\left.\begin{aligned} \left\langle -\frac{d}{dx}\left(p\frac{du}{dx}\right) + qu, v\right\rangle_0 = \left\langle p\frac{du}{dx}, \frac{dv}{dx}\right\rangle_0 + \langle qu, v\rangle_0 - p(0)\frac{du}{dx}(0)\overline{v(0)} \\ + p(2\pi)\frac{du}{dx}(2\pi)\overline{v(2\pi)}. \end{aligned}\right\} \tag{1.74}$$

Comparison of (1.73) and (1.74) shows, in view of (1.64), that

$$p(0)\frac{du}{dx}(0) = \gamma_1 p(0)u(0) \text{ and } p(2\pi)\frac{du}{dx}(2\pi) = \gamma_2 p(2\pi)u(2\pi), \tag{1.75}$$

i.e. the boundary conditions (1.38)-(1.39) are satisfied. Thus $u \in \mathbf{V}$ solves (1.36), (1.38)-(1.39) in the sense of distributions.

A similar interpretation may be obtained for the variational problem (1.71) in connection with the periodic problem (1.36) & (1.44).

Theorem 1.2.15. The eigenvalue problem

$$\left.\begin{aligned} -\frac{d}{dx}\left(p\frac{du}{dx}\right) + qu = \lambda u \text{ on } (0,2\pi) \\ u(0) = u(2\pi) \end{aligned}\right\} \tag{1.76}$$

has eigenvalues $(\lambda_k)_{k=1}^{\infty}$ and eigenfunctions $\{w_k\}_{k\in\mathbf{N}}$ that satisfy the following:

(a) The eigenvalues $(\lambda_k)_{k=1}^{\infty}$ form a positive increasing unbounded sequence of real numbers, i.e.

$$0 < \lambda_1 \le \lambda_2 \le \lambda_3 \le \Lambda \ . \tag{1.77}$$

(b) The eigenfunctions $\{w_k\}_{k\in\mathbf{N}}$, where $w_k \in \mathbf{H}^1_{\text{per}}(0,2\pi)$, form a Hilbert basis of $\mathbf{L}^2(0,2\pi)$.

(c) There holds the identity

$$a(w_k, v) = \lambda_k \langle w_k, v\rangle_0 \text{ for } k \in \mathbf{N} \text{ and } v \in \mathbf{H}^1_{\text{per}}(0,2\pi). \tag{1.78}$$

Consequently, the eigenfunctions $\left\{\frac{1}{\sqrt{\lambda_k}} w_k\right\}_{k\in\mathbf{N}}$ form a Hilbert basis of $\mathbf{H}^1_{\text{per}}(0,2\pi)$ with respect to the inner product *a*.

Proof. Define the operator $K : \mathbf{L}^2(0,2\pi) \to \mathbf{H}^1_{\text{per}}(0,2\pi) \subset \mathbf{H}^1(0,2\pi)$ that associates with each *f* the corresponding solution *u* to the problem (1.36) & (1.44). *K* is linear and, by (1.51), bounded. By the Rellich Theorem, the embedding $J : \mathbf{H}^1(0,2\pi) \to \mathbf{L}^2(0,2\pi)$ is compact. If $T := J \circ K$, then *T* is compact as well.

(a) For any $v \in \mathbf{L}^2(0,2\pi)$ and $w \in \mathbf{L}^2(0,2\pi)$,

$$\begin{aligned} \langle v, Tw\rangle_0 &= a(Tv, Tw) && (\text{Definition of } T) \\ &= \overline{a(Tw, Tv)} && (a \text{ is hermitian}) \\ &= \overline{\langle w, Tv\rangle_0} && (\text{Definition of } T) \\ &= \langle Tv, w\rangle_0. \end{aligned}$$

Therefore *T* is self-adjoint. By Theorem 1.1.14(b), the eigenvalues $(\mu_k)_{k=1}^{\infty}$ of *T* form a decreasing sequence of real numbers with limit 0.

For each eigenvalue μ and associated eigenvector $v \in \mathbf{L}^2(0,2\pi)$ of *T*, we have

$$\begin{aligned} \mu \|v\|_0^2 &= \langle v, Tv\rangle_0 \\ &= a(Tv, Tv) && (\text{Definition of } T) \\ \mu &= \frac{a(Tv, Tv)}{\|v\|_0^2} \\ &\ge \frac{\alpha \|Tv\|_1^2}{\|v\|_0^2} && (\mathbf{H}^1(0,2\pi)\text{- coercivity of } a) \\ &> 0 && (\text{Injectivity of } T). \end{aligned}$$

Hence $(\mu_k)_{k=1}^{\infty}$ is a positive sequence.

If $\lambda_k := \frac{1}{\mu_k}$ and w_k is the eigenvector with $\|w_k\|_0 = 1$ associated with each μ_k, then $(\lambda_k)_{k=1}^{\infty}$ is an increasing unbounded sequence, and (1.76) becomes, for any $k \in \mathbf{N}$,

$$w_k = \lambda_k T w_k \text{ where } w_k \in \mathbf{H}^1_{\text{per}}(0,2\pi). \tag{1.79}$$

(b)-(c) By Theorem 1.1.14(d), $\{w_k\}_{k\in\mathbf{N}}$ is a Hilbert basis of $\mathbf{L}^2(0,2\pi)$. Moreover, for all $v\in\mathbf{H}^1_{\text{per}}(0,2\pi)$, using (1.79), we have

$$\begin{aligned} a(w_k,v) &= a(\lambda_k T w_k, v) \\ &= \lambda_k a(Tw_k, v) \\ &= \lambda_k \langle w_k, v\rangle_0, \end{aligned}$$

which proves (1.78). If $v_k := \frac{1}{\sqrt{\lambda_k}} w_k$, then

$$\begin{aligned} a(v_m, v_n) &= \frac{1}{\sqrt{\lambda_m \lambda_n}} a(w_m, w_n) \\ &= \sqrt{\frac{\lambda_m}{\lambda_n}} \langle w_m, w_n\rangle. \end{aligned}$$

Therefore $\{v_k\}_{k\in\mathbf{N}}$ is a Hilbert basis of $\mathbf{H}^1_{\text{per}}(0,2\pi)$ with respect to the inner product a.

■

Chapter 2. General linear diffusion problem

In this chapter, we consider existence and uniqueness of solutions to the general linear diffusion initial-boundary value problem

$$\frac{\partial u}{\partial t} - c\frac{\partial^2 u}{\partial x^2} + bu = f \text{ on } (0,2\pi)\times(0,T) \tag{2.1}$$

$$u(x,0) = u_0(x) \text{ for } x \in (0,2\pi) \tag{2.2}$$

$$u(0,t) = u(2\pi,t) \text{ for } t \in (0,T). \tag{2.3}$$

Here the constants $b > 0$ and $c > 0$, the finite limit time $T > 0$ and the functions $u_0 : (0,2\pi) \to \mathbf{R}$ and $f : (0,2\pi)\times(0,T) \to \mathbf{R}$ are given, whereas the function $u : (0,2\pi)\times(0,T) \to \mathbf{R}$ is unknown.

There are several methods of solving evolution problems (see e.g. Dautray & Lions (2000b: pp. 509-523)). In view of the numerical approach followed in this work, we shall use a specific Galerkin method that is related to the Fourier series or diagonalization method in Hilbert spaces, as developed by Raviart & Thomas (1983: pp. 155-161).

The following notation and function spaces will be used frequently. We systematically separate the variables x and t for a given function $w : [0,2\pi]\times[0,T] \to \mathbf{R}$ by associating with w, for $t \in [0,T]$, the function $w(t) : [0,2\pi] \to \mathbf{R}$ such that $w(t)(x) = w(x,t)$.

Given a Hilbert space $\mathbf{V}$ and an integer $m \geq 0$, we consider on the one hand the space

$$\mathbf{C}^m([0,T],\mathbf{V}) := \left\{ v : [0,T] \to \mathbf{V} \,\middle|\, t \;\alpha\; \frac{d^k v}{dt^k} \text{ is continuous for } k = 0,1,\mathrm{K}\ m \right\}, \tag{2.4}$$

which is a Banach space under the norm

$$\|v\|_{\mathbf{C}^m([0,T],\mathbf{V})} := \max_{0\leq k\leq m}\left(\sup_{t\in[0,T]} \left\| \frac{d^k u(t)}{dt^k} \right\|_{\mathbf{V}} \right). \tag{2.5}$$

On the other hand, we consider the Hilbert space

$$\mathbf{L}^2((0,T),\mathbf{V}) := \left\{ \text{class of } v : (0,T) \to \mathbf{V} \,\middle|\, \int_0^T \|v(t)\|_{\mathbf{V}}^2 \, dt < \infty \right\} \tag{2.6}$$

equipped with the norm and inner product

$$\|v\|_{\mathbf{L}^2((0,T),\mathbf{V})} := \left(\int_0^T \|v(t)\|_{\mathbf{V}}^2 \, dt \right)^{\frac{1}{2}} \text{ and } \langle v, w \rangle_{\mathbf{L}^2((0,T),\mathbf{V})} := \int_0^T \langle v(t), w(t) \rangle_{\mathbf{V}} \, dt \,. \tag{2.7}$$

We assume once and for all that the data u_0 and f satisfy

$$u_0 \in \mathbf{L}^2(0,2\pi) \text{ and } f \in \mathbf{L}^2((0,2\pi)\times(0,T)) = \mathbf{L}^2\big((0,T),\mathbf{L}^2(0,2\pi)\big). \tag{2.8}$$

In order to obtain a variational formulation of (2.1)-(2.3), we assume, as usual, that this problem admits a smooth enough solution u, e.g.

$$u \in \mathbf{C}^1\left([0,T], \mathbf{H}^1_{\text{per}}(0,2\pi) \cap \mathbf{H}^2(0,2\pi)\right). \tag{2.9}$$

Fix a test function $\psi \in \mathbf{L}^2\left((0,T), \mathbf{H}^1_{\text{per}}(0,2\pi)\right)$. Multiplication of (2.1) by $\overline{\psi}$ and integration over $(0,2\pi)\times(0,T)$ lead to

$$\left.\begin{aligned} \int_0^T\int_0^{2\pi} \frac{\partial u}{\partial t}(x,t)\overline{\psi(x,t)}\,dx\,dt - c\int_0^T\int_0^{2\pi} \frac{\partial^2 u}{\partial x^2}(x,t)\overline{\psi(x,t)}\,dx\,dt + b\int_0^T\int_0^{2\pi} u(x,t)\overline{\psi(x,t)}\,dx\,dt \\ = \int_0^T\int_0^{2\pi} f(x,t)\overline{\psi(x,t)}\,dx\,dt. \end{aligned}\right\} \tag{2.10}$$

Integration by parts with respect to x in the second term on the left hand side yields, because of the boundary condition (2.3),

$$\left.\begin{aligned} \int_0^T\int_0^{2\pi} \frac{\partial u}{\partial t}(x,t)\overline{\psi(x,t)}\,dx\,dt + c\int_0^T\int_0^{2\pi} \frac{\partial u}{\partial x}(x,t)\overline{\frac{\partial \psi}{\partial x}(x,t)}\,dx\,dt + b\int_0^T\int_0^{2\pi} u(x,t)\overline{\psi(x,t)}\,dx\,dt \\ = \int_0^T\int_0^{2\pi} f(x,t)\overline{\psi(x,t)}\,dx\,dt. \end{aligned}\right\} \tag{2.11}$$

Recognising that, due to the smoothness of u,

$$\frac{d}{dt}\langle u(t), v\rangle_0 = \left\langle \frac{du(t)}{dt}, v\right\rangle_0, \tag{2.12}$$

and taking $\psi(x,t) = \gamma(t)v(x)$, with $\gamma \in \mathbf{D}(0,T)$ and $v \in \mathbf{H}^1_{\text{per}}(0,2\pi)$, we have

$$\frac{d}{dt}\langle u(t), v\rangle_0 + c\left\langle \frac{du(t)}{dx}, \frac{dv}{dx}\right\rangle_0 + b\langle u(t), v\rangle_0 = \langle f(t), v\rangle_0 \text{ in } \mathbf{D}'(0,T). \tag{2.13}$$

If the sesquilinear form $a : \mathbf{H}^1(0,2\pi)\times\mathbf{H}^1(0,2\pi) \to \mathbf{C}$ is defined as

$$a(v,w) := c\left\langle \frac{dv}{dx}, \frac{dw}{dx}\right\rangle_0 + b\langle v, w\rangle_0 \text{ for all } v \in \mathbf{H}^1(0,2\pi) \text{ and } w \in \mathbf{H}^1(0,2\pi), \tag{2.14}$$

then it is continuous, i.e.

$$\left.\begin{aligned} |a(v,w)| \le \|a\|\|v\|_1\|w\|_1 \text{ where } \|a\| := \max\{b,c\} \text{ for all } v \in \mathbf{H}^1(0,2\pi) \\ \text{and } w \in \mathbf{H}^1(0,2\pi), \end{aligned}\right\} \tag{2.15}$$

and $\mathbf{H}^1(0,2\pi)$-elliptic or $\mathbf{H}^1(0,2\pi)$-coercive in the sense that

$$a(v,v) \ge \alpha\|v\|_1^2 \text{ for all } v \in \mathbf{H}^1(0,2\pi), \text{ where } \alpha := \min\{b,c\}. \tag{2.16}$$

The form a is also clearly hermitian.

The equation (2.13) leads to the variational formulation of the general linear diffusion problem:

Definition 2.1. General linear diffusion variational problem. The problem of finding a real-valued function

$$u \in \mathbf{L}^2\left((0,T),\mathbf{H}^1_{\text{per}}(0,2\pi)\right) \cap \mathbf{C}^0\left([0,T],\mathbf{L}^2(0,2\pi)\right) \tag{2.17}$$

such that, for almost all t in $(0,T)$,

$$\frac{d}{dt}\langle u(t),v\rangle_0 + a(u(t),v) = \langle f(t),v\rangle_0 \text{ for all } v \in \mathbf{H}^1_{\text{per}}(0,2\pi) \tag{2.18}$$

$$u(0) = u_0\,, \tag{2.19}$$

with the derivative $\frac{d}{dt}$ in the sense of distributions, is called the variational formulation of the general linear diffusion problem (2.1)-(2.3).

Remark 2.2. Strictly speaking, the test function $\psi \in \mathbf{L}^2\left((0,T),\mathbf{H}^1_{\text{per}}(0,2\pi)\right)$ is taken such that

$$\frac{\partial \psi}{\partial t} \in \mathbf{L}^2\left((0,T),\mathbf{L}^2(0,2\pi)\right) \text{ and } \psi(T) = 0\,, \tag{2.20}$$

where $\frac{\partial \psi}{\partial t}$ is the derivative in the sense of distributions. Integration by parts in the variable t in the first integral in (2.11) shows that, for any $\psi \in \mathbf{L}^2\left((0,T),\mathbf{H}^1_{\text{per}}(0,2\pi)\right)$ satisfying (2.20), we have

$$-\int_0^T \left\langle u(t), \frac{d\psi}{dt}(t)\right\rangle_0 dt + \int_0^T a(u(t),\psi(t))dt = \int_0^T \langle f(t),\psi(t)\rangle_0\, dt + \langle u_0,\psi(0)\rangle_0\,. \tag{2.21}$$

The equation (2.21) is an alternative variational formulation of the general linear diffusion problem. This type of variational formulation is studied in, e.g. Lions (1961: p. 44).

The variational problem (2.17)-(2.19) is closely related to the following eigenvalue problem of the type studied in Section 1.2: Find

$$w \in \mathbf{H}^1_{\text{per}}(0,2\pi) \tag{2.22}$$

and $\lambda \in \mathbf{C}$ such that

$$-c\frac{d^2w}{dx^2} + bw = \lambda w \text{ on } (0,2\pi) \tag{2.23}$$

or, equivalently,

$$a(w,v) = \lambda\langle w,v\rangle_0 \text{ for all } v \in \mathbf{H}^1_{\text{per}}(0,2\pi). \tag{2.24}$$

Regarding this eigenvalue problem, we have the following result:

Theorem 2.3.

(a) The eigenvalues in (2.23) and (2.24) are given by the positive unbounded sequence

$$\lambda_k = ck^2 + b \text{ for } k \in \mathbf{Z}. \tag{2.25}$$

(b) Associated with each λ_k is the eigenfunction

$$w_k(x) = \frac{1}{\sqrt{2\pi}} e^{ikx} \text{ for } k \in \mathbf{Z}. \tag{2.26}$$

(c) The system $\{w_k\}_{k\in\mathbf{Z}}$ is a Hilbert basis of $\mathbf{L}^2(0,2\pi)$, and the system $\left\{\frac{1}{\sqrt{\lambda_k}} w_k\right\}_{k\in\mathbf{Z}}$ is a Hilbert basis of $\mathbf{H}^1_{\text{per}}(0,2\pi)$ with respect to the inner product a.

Proof. All we have to show is the explicit form of λ_k and w_k. By the properties of the hermitian form a (cf. (2.15) and (2.16)), Theorem 1.2.15 guarantees that the eigenvalues form an increasing unbounded sequence of positive real numbers with corresponding eigenfunctions forming a Hilbert basis of $\mathbf{L}^2(0,2\pi)$ and $\mathbf{H}^1_{\text{per}}(0,2\pi)$.

The solution to the ordinary differential equation (2.23) is of the form

$$w(x) = Ae^{i\sqrt{\frac{\lambda-b}{c}}x} + Be^{-i\sqrt{\frac{\lambda-b}{c}}x}. \tag{2.27}$$

In view of the periodicity condition (2.22), we must have

$$1 = e^{i2\pi\sqrt{\frac{\lambda-b}{c}}x} \text{ and } 1 = e^{-i2\pi\sqrt{\frac{\lambda-b}{c}}x}. \tag{2.28}$$

This implies that, for some $k \in \mathbf{Z}$,

$$\sqrt{\frac{\lambda_k - b}{c}} = k, \tag{2.29}$$

and therefore λ_k is given by (2.25).

For $k \in \mathbf{Z}$, let us define w_k by (2.26). Clearly, w_k satisfies (2.27). Moreover, $\{w_k\}_{k\in\mathbf{Z}}$ is a Hilbert basis of $\mathbf{L}^2(0,2\pi)$.

■

Theorem 2.4. The variational problem (2.17)-(2.19) has a unique solution u. With the eigenvalues $\{\lambda_k\}_{k\in\mathbf{Z}}$ and eigenfunctions $\{w_k\}_{k\in\mathbf{Z}}$ defined in Theorem 2.3, u has, for $t \in [0,T]$, the Fourier series representation

$$u(t) = \sum_{k\in\mathbf{Z}} \left(e^{-\lambda_k t} \langle u_0, w_k \rangle_0 + \int_0^t e^{-\lambda_k (t-s)} \langle f(s), w_k \rangle_0 \, ds \right) w_k \,. \tag{2.30}$$

Proof. The proof will be done in two parts.

Part 1: Existence of solution.

We proceed by the Galerkin method, the four main steps of which are listed below.

1. Approximation of (2.17)-(2.19) in a finite dimensional space

For each $m \in \mathbf{N}$, let

$$\mathbf{S}_m := \operatorname{span}\{w_k\}_{k=-m}^{m}. \tag{2.31}$$

Following Temam (1979: p. 44), the expanding sequence $(\mathbf{S}_m)_{m=1}^{\infty}$ is a Galerkin approximation to the space $\mathbf{H}^1_{\text{per}}(0,2\pi)$ due to the fact that

$$\overline{\bigcup_{m=1}^{\infty}} \mathbf{S}_m \text{ is dense in } \mathbf{H}^1_{\text{per}}(0,2\pi). \tag{2.32}$$

Consider the finite dimensional analogue of the variational problem (2.17)-(2.19): For $m \in \mathbf{N}$, find

$$u_m \in \mathbf{L}^2((0,T),\mathbf{S}_m) \cap \mathbf{C}^0([0,T],\mathbf{L}^2(0,2\pi)) \tag{2.33}$$

such that

$$\frac{d}{dt}\langle u_m(t),v\rangle_0 + a(u_m(t),v) = \langle f(t),v\rangle_0 \text{ for all } v \in \mathbf{S}_m \text{ and } t \in (0,T) \tag{2.34}$$

$$u_m(0) = \sum_{k=-m}^{m} \langle u_0, w_k\rangle_0 w_k \,. \tag{2.35}$$

The condition (2.33) shows that any solution of (2.33)-(2.35) admits the representation

$$u_m(t) := \sum_{k=-m}^{m} \alpha_k(t) w_k \text{ for } t \in (0,T), \tag{2.36}$$

with unknown coefficients $\{\alpha_k\}_{k=-m}^{m}$. For $k = -m,-m+1,\mathrm{K}\,,m$, take $v = w_k$ in (2.34). Using (2.24) and (2.35), it follows that solving (2.34)-(2.35) is equivalent to solving, for $k = -m,-m+1,\mathrm{K}\,,m$, the initial value problem

$$\left.\begin{aligned} \frac{d}{dt}\alpha_k(t) + \lambda_k \alpha_k(t) &= \langle f(t), w_k\rangle_0 \text{ for } t \in (0,T) \\ \alpha_k(0) &= \langle u_0, w_k\rangle_0. \end{aligned}\right\} \tag{2.37}$$

Easy calculation shows that the unique solution of the initial value problem (2.37) is

$$\alpha_k(t) = e^{-\lambda_k t}\langle u_0, w_k\rangle_0 + \int_0^t e^{-\lambda_k (t-s)} \langle f(s), w_k\rangle_0 \, ds \,. \tag{2.38}$$

Therefore the problem (2.33)-(2.35) has a unique solution given by

$$u_m(t) = \sum_{k=-m}^{m} \left(e^{-\lambda_k t} \langle u_0, w_k \rangle_0 + \int_0^t e^{-\lambda_k (t-s)} \langle f(s), w_k \rangle_0 \, ds \right) w_k \,. \tag{2.39}$$

2. *A priori* estimates

In this specific case of a Galerkin method, which is related to the Fourier series or diagonalization method, the *a priori* estimates step amounts to proving that $(u_m)_{m=1}^{\infty}$ is a Cauchy sequence in both $\mathbf{C}^0([0,T],\mathbf{L}^2(0,2\pi))$ and $\mathbf{L}^2((0,T),\mathbf{H}^1_{\text{per}}(0,2\pi))$ (For the general procedure involving *a priori* estimates, refer to Dautray & Lions (2000b: pp. 514-515)).

Since Parseval's identity gives

$$\begin{aligned} \sum_{k\in\mathbf{Z}} \left| \langle u_0, w_k \rangle_0 \right|^2 &= \|u_0\|_0^2 \\ &< \infty \end{aligned}$$

and

$$\begin{aligned} \sum_{k\in\mathbf{Z}} \int_0^T \left| \langle f(s), w_k \rangle_0 \right|^2 ds &= \int_0^T \sum_{k\in\mathbf{Z}} \left| \langle f(s), w_k \rangle_0 \right|^2 ds \\ &= \|f\|^2_{\mathbf{L}^2((0,T),\mathbf{L}^2(0,2\pi))} \\ &< \infty, \end{aligned}$$

it follows that

$$\lim_{\substack{m\to\infty \\ p\to\infty \\ p>m}} \sum_{|k|=m+1}^{p} \left| \langle u_0, w_k \rangle_0 \right|^2 = 0 \text{ and } \lim_{\substack{m\to\infty \\ p\to\infty \\ p>m}} \sum_{|k|=m+1}^{p} \int_0^T \left| \langle f(s), w_k \rangle_0 \right|^2 ds = 0 \,. \tag{2.40}$$

For any $m \in \mathbf{N}$ and $p \in \mathbf{N}$ with $p > m$, we have

$$\begin{aligned} &\left\| u_p(t) - u_m(t) \right\|_0 \\ &= \sqrt{\sum_{|k|=m+1}^{p} \left| e^{-\lambda_k t} \langle u_0, w_k \rangle_0 + \int_0^t e^{-\lambda_k (t-s)} \langle f(s), w_k \rangle_0 ds \right|^2} && \begin{pmatrix} \text{Form of } u_m \\ \text{Parseval identity} \end{pmatrix} \\ &\le \sqrt{\sum_{|k|=m+1}^{p} e^{-2\lambda_k t} \left| \langle u_0, w_k \rangle_0 \right|^2} + \sqrt{\sum_{|k|=m+1}^{p} \left| \int_0^t e^{-\lambda_k (t-s)} \langle f(s), w_k \rangle_0 ds \right|^2} && (\text{Minkowski inequality}) \\ &\le \sqrt{\sum_{|k|=m+1}^{p} e^{-2\lambda_k t} \left| \langle u_0, w_k \rangle_0 \right|^2} + \sqrt{\sum_{|k|=m+1}^{p} \left(\int_0^t e^{-2\lambda_k (t-s)} ds \right) \left(\int_0^t \left| \langle f(s), w_k \rangle_0 \right|^2 ds \right)} && (\text{Cauchy - Schwarz inequality}) \\ &= \sqrt{\sum_{|k|=m+1}^{p} e^{-2\lambda_k t} \left| \langle u_0, w_k \rangle_0 \right|^2} + \sqrt{\sum_{|k|=m+1}^{p} \frac{1 - e^{-2\lambda_k}}{2\lambda_k} \left(\int_0^t \left| \langle f(s), w_k \rangle_0 \right|^2 ds \right)}. \end{aligned}$$

It then follows that

$$\begin{aligned}
&\left\|u_p-u_m\right\|_{\mathbf{C}^0\left([0,T],\mathbf{L}^2(0,2\pi)\right)} \\
&=\sup_{t\in[0,T]}\left\|u_p(t)-u_m(t)\right\|_0 && \left(\text{Definition of } \|\cdot\|_{\mathbf{C}^0\left([0,T],\mathbf{L}^2(0,2\pi)\right)}\right) \\
&\le\sqrt{\sum_{|k|=m+1}^{p}\left|\left\langle u_0,w_k\right\rangle_0\right|^2}+\sqrt{\frac{1}{2b}\sum_{|k|=m+1}^{p}\left(\int_0^T\left|\left\langle f(s),w_k\right\rangle_0\right|^2 ds\right)} && \left(\text{Definition of } \{\lambda_k\}_{k\in\mathbf{Z}}\right).
\end{aligned}$$

In view of (2.40), $\lim\limits_{\substack{m\to\infty\\ p\to\infty\\ p>m}}\left\|u_p-u_m\right\|_{\mathbf{C}^0\left([0,T],\mathbf{L}^2(0,2\pi)\right)}=0$, which shows that $\left(u_m\right)_{m=1}^{\infty}$ is a Cauchy sequence in $\mathbf{C}^0\left([0,T],\mathbf{L}^2(0,2\pi)\right)$.

For any $m\in\mathbf{N}$ and $p\in\mathbf{N}$ with $p>m$, and a the hermitian form defined in (2.14), we have

$$\begin{aligned}
&\left\|u_p(t)-u_m(t)\right\|_1^2 \\
&\le\frac{1}{\alpha}a\left(u_p(t)-u_m(t),u_p(t)-u_m(t)\right) && \left(\mathbf{H}^1(0,2\pi)\text{-coercivity of } a\right) \\
&=\frac{1}{\alpha}\sum_{|k|=m+1}^{p}\sum_{|l|=m+1}^{p}a\left(\alpha_k(t)w_k,\alpha_l(t)w_l\right) && \left(\text{Form of } u_m \text{ and } u_p\right) \\
&=\frac{1}{\alpha}\sum_{|k|=m+1}^{p}\sum_{|l|=m+1}^{p}\alpha_k(t)\overline{\alpha_l(t)}\sqrt{\lambda_k\lambda_l}\,a\left(\frac{1}{\sqrt{\lambda_k}}w_k,\frac{1}{\sqrt{\lambda_l}}w_l\right) \\
&=\frac{1}{\alpha}\sum_{|k|=m+1}^{p}\lambda_k\left|\alpha_k(t)\right|^2 && \left(\text{Orthonormality of } \left\{\frac{1}{\sqrt{\lambda_k}}w_k\right\}_{k\in\mathbf{Z}}\right) \\
&=\frac{1}{\alpha}\sum_{|k|=m+1}^{p}\lambda_k\left|e^{-\lambda_k t}\left\langle u_0,w_k\right\rangle_0+\int_0^t e^{-\lambda_k(t-s)}\left\langle f(s),w_k\right\rangle_0 ds\right|^2 && \left(\text{Form of } \alpha_k\right) \\
&\le\frac{2}{\alpha}\sum_{|k|=m+1}^{p}\lambda_k\left(e^{-2\lambda_k t}\left|\left\langle u_0,w_k\right\rangle_0\right|^2+\left|\int_0^t e^{-\lambda_k(t-s)}\left\langle f(s),w_k\right\rangle_0 ds\right|^2\right) && \left((x+y)^2\le 2\left(x^2+y^2\right)\right).
\end{aligned}$$

Thus

$$\begin{aligned}
\left\|u_p-u_m\right\|^2_{\mathbf{L}^2\left((0,T),\mathbf{H}^1_{\text{per}}(0,2\pi)\right)}&=\int_0^T\left\|u_p(t)-u_m(t)\right\|_1^2 dt \\
&\le\frac{2}{\alpha}\int_0^T\sum_{|k|=m+1}^{p}\lambda_k\left(e^{-2\lambda_k t}\left|\left\langle u_0,w_k\right\rangle_0\right|^2+\left|\int_0^t e^{-\lambda_k(t-s)}\left\langle f(s),w_k\right\rangle_0 ds\right|^2\right)dt \\
&=\frac{2}{\alpha}\sum_{|k|=m+1}^{p}\int_0^T\lambda_k\left(e^{-2\lambda_k t}\left|\left\langle u_0,w_k\right\rangle_0\right|^2+\left|\int_0^t e^{-\lambda_k(t-s)}\left\langle f(s),w_k\right\rangle_0 ds\right|^2\right)dt \\
&=\frac{2}{\alpha}\sum_{|k|=m+1}^{p}\left(\lambda_k\left|\left\langle u_0,w_k\right\rangle_0\right|^2\int_0^T e^{-2\lambda_k t}dt+\lambda_k\int_0^T\left|\int_0^t e^{-\lambda_k(t-s)}\left\langle f(s),w_k\right\rangle_0 ds\right|^2 dt\right)
\end{aligned}$$

$$
\begin{aligned}
&\left\|u_p-u_m\right\|^2_{\mathbf{L}^2\left((0,T),\mathbf{H}^1_{\mathrm{per}}(0,2\pi)\right)}\\
&\le\frac{2}{\alpha}\sum_{|k|=m+1}^{p}\left(\lambda_k\left|\left\langle u_0,w_k\right\rangle_0\right|^2\int_0^T e^{-2\lambda_k t}dt+\lambda_k\int_0^T\left(\int_0^t e^{-2\lambda_k(t-s)}ds\right)\left(\int_0^t\left|\left\langle f(s),w_k\right\rangle_0\right|^2 ds\right)dt\right)\quad\begin{pmatrix}\text{Cauchy-Schwarz}\\ \text{inequality}\end{pmatrix}\\
&=\frac{1}{\alpha}\sum_{|k|=m+1}^{p}\left(\left|\left\langle u_0,w_k\right\rangle_0\right|^2\left(1-e^{-2\lambda_k T}\right)+\int_0^T e^{-2\lambda_k t}\left(1-e^{2\lambda_k t}\right)\left(\int_0^t\left|\left\langle f(s),w_k\right\rangle_0\right|^2 ds\right)dt\right)\\
&\le\frac{1}{\alpha}\sum_{|k|=m+1}^{p}\left(\left|\left\langle u_0,w_k\right\rangle_0\right|^2+T\int_0^T\left|\left\langle f(s),w_k\right\rangle_0\right|^2 ds\right),
\end{aligned}
$$

which, by (2.40), shows that $\lim\limits_{\substack{m\to\infty\\ p\to\infty\\ p>m}}\left\|u_p-u_m\right\|_{\mathbf{L}^2\left((0,T),\mathbf{H}^1_{\mathrm{per}}(0,2\pi)\right)}=0$, and $\left(u_m\right)_{m=1}^{\infty}$ is a Cauchy sequence in $\mathbf{L}^2\left((0,T),\mathbf{H}^1_{\mathrm{per}}(0,2\pi)\right)$ as well.

3. Passage to limit *u*

In the general situation, a weak compactness argument is used in this step (see e.g. Dautray & Lions (2000b: pp. 515-516)). This step is much simpler in our case, because of the specific method under consideration.

Since both $\mathbf{C}^0\left([0,T],\mathbf{L}^2(0,2\pi)\right)$ and $\mathbf{L}^2\left((0,T),\mathbf{H}^1_{\mathrm{per}}(0,2\pi)\right)$ are complete and continuously embedded in $\mathbf{L}^2\left((0,T)\times(0,2\pi)\right)$, there exists a unique $u\in\mathbf{L}^2\left((0,T),\mathbf{H}^1_{\mathrm{per}}(0,2\pi)\right)\cap\mathbf{C}^0\left([0,T],\mathbf{L}^2(0,2\pi)\right)$ such that the Cauchy sequence $\left(u_m\right)_{m=1}^{\infty}$ converges to u in both $\mathbf{C}^0\left([0,T],\mathbf{L}^2(0,2\pi)\right)$ and $\mathbf{L}^2\left((0,T),\mathbf{H}^1_{\mathrm{per}}(0,2\pi)\right)$. Note that, in view of (2.39), we have (2.30).

4. Confirmation that the limit *u* is a solution to (2.17)-(2.19)

Fix $\psi\in\mathbf{D}(0,T)$ and $n\in\mathbf{N}$. For any $m\ge n$, multiplication of (2.34) by $\overline{\psi}$ yields, by the distributional definition of derivatives,

$$
-\int_0^T\left\langle u_m(t),v\right\rangle_0\overline{\frac{d\psi}{dt}(t)}dt+\int_0^T a(u_m(t),v)\overline{\psi(t)}dt=\int_0^T\left\langle f(t),v\right\rangle_0\overline{\psi(t)}dt\ \text{ for all }\ v\in\mathbf{S}_n. \qquad (2.41)
$$

Since, in (2.41), $\left(u_m\right)_{m=1}^{\infty}$ converges to u in the sense of the space of vector distributions $\mathbf{D}'\left((0,T),\mathbf{H}^1_{\mathrm{per}}(0,2\pi)\right)\cap\mathbf{D}'\left((0,T),\mathbf{L}^2(0,2\pi)\right)$, it follows that

$$
-\left\langle\left\langle u(t),v\right\rangle_0,\frac{d\psi}{dt}\right\rangle_{0,(0,T)}+\left\langle a(u(t),v),\psi\right\rangle_{0,(0,T)}=\left\langle\left\langle f(t),v\right\rangle_0,\psi\right\rangle_{0,(0,T)}\ \text{ for all }\ v\in\mathbf{S}_n. \qquad (2.42)
$$

By (2.32), the relation (2.42) holds for all $v\in\mathbf{H}^1_{\mathrm{per}}(0,2\pi)$. Hence the function u satisfies (2.18) on $(0,T)$ in the sense of distributions. Moreover, since $\left(u_m\right)_{m=1}^{\infty}$ converges to u in $\mathbf{C}^0\left([0,T],\mathbf{L}^2(0,2\pi)\right)$, it follows from the pointwise convergence in $\mathbf{L}^2(0,2\pi)$ that

$$\begin{aligned} u(0) &= \lim_{m\to\infty} u_m(0) \\ &= \lim_{m\to\infty} \sum_{k=-m}^{m} \langle u_0, w_k \rangle_0 w_k \\ &= \sum_{k\in\mathbf{Z}} \langle u_0, w_k \rangle_0 w_k \\ &= u_0. \end{aligned}$$

Therefore u satisfies (2.19).

Part 2: Uniqueness of solution

The solution u is unique as a result of the representation (2.30).

■

Remark 2.5. The solution u in (2.30) is real-valued, even though it is expressed as a series of complex functions. Indeed, we have, for $(x,t)\in(0,2\pi)\times(0,T)$,

$$\begin{aligned} \overline{u(x,t)} &= \sum_{k\in\mathbf{Z}} \left(e^{-\lambda_k t} \overline{\langle u_0, w_k \rangle_0} + \int_0^t e^{-\lambda_k (t-s)} \overline{\langle f(s), w_k \rangle_0} ds \right) \overline{w_k(x)} \qquad (\text{Form of } u) \\ &= \sum_{k\in\mathbf{Z}} \left(e^{-\lambda_k t} \langle u_0, w_{-k} \rangle_0 + \int_0^t e^{-\lambda_k (t-s)} \langle f(s), w_{-k} \rangle_0 ds \right) w_{-k}(x) \quad \left(\begin{array}{l} f, u_0 \text{ real-valued} \\ \overline{w_k} = w_{-k} \end{array} \right) \\ &= \sum_{k\in\mathbf{Z}} \left(e^{-\lambda_k t} \langle u_0, w_k \rangle_0 + \int_0^t e^{-\lambda_k (t-s)} \langle f(s), w_k \rangle_0 ds \right) w_k \\ &= u(x,t). \end{aligned}$$

Theorem 2.6. The solution u of (2.17)-(2.19) satisfies the inequality

$$\|u(t)\|_0 \le e^{-bt} \|u_0\|_0 + \int_0^t e^{-b(t-s)} \|f(s)\|_0 ds \text{ for } t\in[0,T], \tag{2.43}$$

from which it follows that the solution u depends continuously upon the data, in the sense that

$$\|u(t)\|_0 \le \left(1 + \sqrt{\frac{e^{2T}-1}{2b}} \right) \left(\|u_0\|_0 + \|f\|_{\mathbf{L}^2((0,T),\mathbf{L}^2(0,2\pi))} \right) \text{ for } t\in[0,T]. \tag{2.44}$$

Proof. By (2.30), we have

$$\begin{aligned} \|u(t)\|_0 &= \left\| \sum_{k\in\mathbf{Z}} \left(e^{-\lambda_k t} \langle u_0, w_k \rangle_0 + \int_0^t e^{-\lambda_k (t-s)} \langle f(s), w_k \rangle_0 ds \right) w_k \right\|_0 \\ &\le \left\| \sum_{k\in\mathbf{Z}} e^{-\lambda_k t} \langle u_0, w_k \rangle_0 w_k \right\|_0 + \left\| \sum_{k\in\mathbf{Z}} \left(\int_0^t e^{-\lambda_k (t-s)} \langle f(s), w_k \rangle_0 ds \right) w_k \right\|_0 \quad (\text{Triangle inequality}) \\ &\le \left\| \sum_{k\in\mathbf{Z}} e^{-\lambda_k t} \langle u_0, w_k \rangle_0 w_k \right\|_0 + \int_0^t \left\| e^{-\lambda_k (t-s)} \langle f(s), w_k \rangle_0 w_k \right\|_0 ds \end{aligned}$$

$$\begin{aligned}
\|u(t)\|_0 &\le \sqrt{\sum_{k\in\mathbf{Z}} e^{-2\lambda_k t}\left|\langle u_0, w_k\rangle_0\right|^2} + \int_0^t \sqrt{\sum_{k\in\mathbf{Z}} e^{-2\lambda_k (t-s)}\left|\langle f(s), w_k\rangle_0\right|^2}\,ds \quad (\text{Parseval identity}) \\
&\le e^{-bt}\sqrt{\sum_{k\in\mathbf{Z}}\left|\langle u_0, w_k\rangle_0\right|^2} + \int_0^t e^{-b(t-s)}\sqrt{\sum_{k\in\mathbf{Z}}\left|\langle f(s), w_k\rangle_0\right|^2}\,ds \qquad (\text{Definition of } \{\lambda_k\}_{k\in\mathbf{Z}}) \\
&= e^{-bt}\|u_0\|_0 + \int_0^t e^{-b(t-s)}\|f(s)\|_0\,ds \qquad (\text{Parseval identity}).
\end{aligned}$$

It then follows that

$$\begin{aligned}
\|u(t)\|_0 &\le \|u_0\|_0 + \int_0^T e^{bs}\|f(s)\|_0\,ds \\
&\le \|u_0\|_0 + \sqrt{\int_0^T e^{2bs}ds}\,\sqrt{\int_0^T \|f(s)\|_0^2\,ds} \qquad \begin{pmatrix}\text{Cauchy - Schwarz}\\ \text{inequality}\end{pmatrix} \\
&= \|u_0\|_0 + \sqrt{\frac{e^{2T}-1}{2b}}\,\|f\|_{\mathbf{L}^2((0,T),\mathbf{L}^2(0,2\pi))} \\
&\le \left(1+\sqrt{\frac{e^{2T}-1}{2b}}\right)\left(\|u_0\|_0 + \|f\|_{\mathbf{L}^2((0,T),\mathbf{L}^2(0,2\pi))}\right).
\end{aligned}$$

■

Remark 2.7. It is possible to prove the inequality of continuous dependence (2.44) with a constant not dependent on *T*. If in (2.34), we set $v = u_m(t)$ and use the relation

$$\left\langle \frac{du_m}{dt}(t), u_m(t)\right\rangle_0 = \frac{1}{2}\frac{d}{dt}\|u_m(t)\|_0^2, \tag{2.45}$$

we obtain

$$\begin{aligned}
\frac{1}{2}\|u_m(t)\|_0^2 - \frac{1}{2}\left\|\sum_{k=-m}^{m}\langle u_0, w_k\rangle_0 w_k\right\|_0^2 &= \frac{1}{2}\int_0^t \frac{d}{ds}\|u_m(s)\|_0^2\,ds \\
&= \int_0^t \langle f(s), u_m(s)\rangle_0\,ds - \int_0^t a(u(s), u_m(s))ds.
\end{aligned}$$

This means that u_m satisfies the relation

$$\frac{1}{2}\|u_m(t)\|_0^2 + \int_0^t a(u_m(s), u_m(s))ds = \frac{1}{2}\left\|\sum_{k=-m}^{m}\langle u_0, w_k\rangle_0 w_k\right\|_0^2 + \int_0^t \langle f(s), u_m(s)\rangle_0\,ds\,. \tag{2.46}$$

Taking the limit as $m\to\infty$, we have

$$\frac{1}{2}\|u(t)\|_0^2 + \int_0^t a(u(s), u(s))ds = \frac{1}{2}\|u_0\|_0^2 + \int_0^t \langle f(s), u(s)\rangle_0\,ds\,. \tag{2.47}$$

The relation (2.47) is called the energy equality, since the quantity

$$E(t) := \frac{1}{2}\|u(t)\|_0^2 + \int_0^t a(u(s),u(s))ds \tag{2.48}$$

represents the energy of the system (see Dautray & Lions (2000b: p. 520)).

We then have

$$\begin{aligned}
\frac{1}{2}\|u(t)\|_0^2 + \frac{\alpha}{2}\int_0^t \|u(s)\|_1^2 ds &\le \frac{1}{2}\|u(t)\|_0^2 + \alpha\int_0^t \|u(s)\|_1^2 ds \\
&\le \frac{1}{2}\|u(t)\|_0^2 + \int_0^t a(u(s),u(s))ds && \left(\mathbf{H}^1(0,2\pi)\text{-coercivity of } a\right) \\
&= \frac{1}{2}\|u_0\|_0^2 + \int_0^t \langle f(s),u(s)\rangle_0 ds && (\text{Energy equality}) \\
&\le \frac{1}{2}\|u_0\|_0^2 + \int_0^t \|f(s)\|_0 \|u(s)\|_1 ds && (\text{Cauchy - Schwarz inequality}) \\
&\le \frac{1}{2}\|u_0\|_0^2 + \frac{1}{2\alpha}\int_0^t \|f(s)\|_0^2 ds + \frac{\alpha}{2}\int_0^t \|u(s)\|_1^2 ds && (\text{Young inequality}) \\
\|u(t)\|_0^2 &\le \|u_0\|_0^2 + \frac{1}{\alpha}\int_0^T \|f(s)\|_0^2 ds \\
\|u(t)\|_0^2 &\le \|u_0\|_0^2 + \frac{1}{\alpha}\|f\|^2_{\mathbf{L}^2((0,T),\mathbf{L}^2(0,2\pi))}.
\end{aligned}$$

A further consequence of the energy equality (2.47) is the continuous dependence of u in the norm of $\mathbf{L}^2\left((0,T),\mathbf{H}^1_{\text{per}}(0,2\pi)\right)$ with respect to u_0 and f, given by

$$\|u\|^2_{\mathbf{L}^2((0,T),\mathbf{H}^1_{\text{per}}(0,2\pi))} \le \frac{1}{\alpha}\|u_0\|_0^2 + \frac{1}{\alpha^2}\|f\|^2_{\mathbf{L}^2((0,T),\mathbf{L}^2(0,2\pi))}. \tag{2.49}$$

Indeed,

$$\begin{aligned}
\alpha\int_0^T \|u(t)\|_1^2 dt &\le \frac{1}{2}\|u(T)\|_0^2 + \int_0^T a(u(t),u(t))dt && \left(\mathrm{H}^1(0,2\pi)\text{-coercivity of } a\right) \\
&= \frac{1}{2}\|u_0\|_0^2 + \int_0^T \langle f(t),u(t)\rangle_0 dt && (\text{Energy equality}) \\
&\le \frac{1}{2}\|u_0\|_0^2 + \int_0^T \|f(t)\|_0 \|u(t)\|_1 dt && (\text{Cauchy - Schwarz inequality}) \\
&\le \frac{1}{2}\|u_0\|_0^2 + \frac{1}{2\alpha}\int_0^T \|f(t)\|_0^2 dt + \frac{\alpha}{2}\int_0^T \|u(t)\|_1^2 dt && (\text{Young inequality}) \\
\alpha\int_0^T \|u(t)\|_1^2 dt &\le \|u_0\|_0^2 + \frac{1}{\alpha}\int_0^T \|f(t)\|_0^2 dt \\
\|u\|^2_{\mathbf{L}^2((0,T),\mathbf{H}^1_{\text{per}}(0,2\pi))} &\le \frac{1}{\alpha}\|u_0\|_0^2 + \frac{1}{\alpha^2}\|f\|^2_{\mathbf{L}^2((0,T),\mathbf{L}^2(0,2\pi))}.
\end{aligned}$$

We now present a regularity result:

Theorem 2.8. If the initial data u_0 is of class $\mathbf{H}^1_{\text{per}}(0,2\pi)$, then the solution u of the problem (2.17)-(2.19) has the regularity

$$\frac{\partial u}{\partial t} \in \mathbf{L}^2\big((0,T),\mathbf{L}^2(0,2\pi)\big) \tag{2.50}$$

and

$$u(t) \in \mathbf{H}^2(0,2\pi) \text{ for almost all } t \in (0,T). \tag{2.51}$$

Proof. For any $m \in \mathbf{N}$, and $k = -m,-m+1,\mathrm{K}\,,m$, taking $v = w_k$ in (2.34) and multiplication by $\overline{\frac{d}{dt}\langle u_m(t), w_k\rangle}$ results in

$$\left|\frac{d}{dt}\langle u_m(t), w_k\rangle_0\right|^2 + a\left(u_m(t), \frac{d}{dt}\langle u_m(t), w_k\rangle_0 w_k\right) = \left\langle f(t), \frac{d}{dt}\langle u_m(t), w_k\rangle_0 w_k\right\rangle_0 . \tag{2.52}$$

Recognising that $a\left(u_m(t), \frac{du_m}{dt}(t)\right) = \frac{1}{2}\frac{d}{dt}a(u_m(t),u_m(t))$, summation of (2.52) over $-m,-m+1,\mathrm{K}\,,m$, integration over $[0,T]$ and rearrangement of terms yield, by the Parseval identity,

$$\begin{aligned}
\int_0^T \left\|\frac{du_m}{dt}(t)\right\|_0^2 dt &= \frac{1}{2}a(u_m(0),u_m(0)) - \frac{1}{2}a(u_m(T),u_m(T)) + \int_0^T \left\langle f(t), \frac{du_m}{dt}(t)\right\rangle_0 dt \\
&\le \frac{1}{2}a(u_m(0),u_m(0)) + \int_0^T \left\langle f(t), \frac{du_m}{dt}(t)\right\rangle_0 dt \\
&\le \frac{1}{2}a(u_m(0),u_m(0)) + \int_0^T \|f(t)\|_0 \left\|\frac{du_m}{dt}(t)\right\|_0 dt && \begin{pmatrix}\text{Cauchy - Schwarz}\\ \text{inequality}\end{pmatrix} \\
&\le \frac{1}{2}a(u_m(0),u_m(0)) + \frac{1}{2}\int_0^T \|f(t)\|_0^2 dt + \frac{1}{2}\int_0^T \left\|\frac{du_m}{dt}(t)\right\|_0^2 dt && (\text{Young inequality}) \\
\int_0^T \left\|\frac{du_m}{dt}(t)\right\|_0^2 dt &\le a(u_m(0),u_m(0)) + \|f\|^2_{\mathbf{L}^2((0,T),\mathbf{L}^2(0,2\pi))} \\
&\le \|a\|\|u_0\|_1^2 + \|f\|^2_{\mathbf{L}^2((0,T),\mathbf{L}^2(0,2\pi))} && (\text{Boundedness of } a) \\
&< \infty.
\end{aligned}$$

Hence $\frac{\partial u_m}{\partial t} \in \mathbf{L}^2\big((0,T),\mathbf{L}^2(0,2\pi)\big)$ for each $m \in \mathbf{N}$, and, taking the limit as $m \to \infty$, it follows that $\frac{\partial u}{\partial t} \in \mathbf{L}^2\big((0,T),\mathbf{L}^2(0,2\pi)\big)$.

Since u satisfies (2.1) in the sense of distributions (see Remark 2.9 below), with $f(t) \in \mathbf{L}^2(0,2\pi)$ and $\frac{\partial u}{\partial t}(t) \in \mathbf{L}^2(0,2\pi)$ for almost all $t \in (0,T)$, it follows that $-c\frac{\partial^2 u}{\partial x^2}(t) + bu(t) \in \mathbf{L}^2(0,2\pi)$ for

almost all $t\in(0,T)$. Then, since $u(t)\in \mathbf{L}^2(0,2\pi)$, $\dfrac{\partial^2 u}{\partial x^2}\in \mathbf{L}^2(0,2\pi)$, and therefore $u(t)\in \mathbf{H}^2(0,2\pi)$, for almost all $t\in(0,T)$.

■

Remark 2.9. The regularity (2.50)-(2.51) in Theorem 2.8 permits us to give an interpretation of the variational problem (2.17)-(2.19).

Suppose that u is the solution of the problem (2.17)-(2.19). Fix $\psi\in\mathbf{D}(0,T)$ and $v\in\mathbf{D}(0,2\pi)$. Substitution of v in (2.18), multiplication by $\overline{\psi}$ and integration over $(0,2\pi)\times(0,T)$ leads to

$$\left.\begin{aligned} -\int_0^T\int_0^{2\pi} u(x,t)\overline{v(x)\frac{d\psi}{dt}(t)}\,dx\,dt - c\int_0^T\int_0^{2\pi} u(x,t)\overline{\frac{d^2v}{dx^2}(x)\psi(t)}\,dx\,dt + b\int_0^T\int_0^{2\pi} u(x,t)\overline{v(x)\psi(t)}\,dx\,dt \\ = \int_0^T\int_0^{2\pi} f(x,t)\overline{v(x)\psi(t)}\,dx\,dt. \end{aligned}\right\} \quad (2.53)$$

If

$$\phi(x,t) := v(x)\psi(t), \quad (2.54)$$

then $\phi\in\mathbf{D}((0,2\pi)\times(0,T))$ and (2.53) becomes

$$\left.\begin{aligned} -\int_0^T\int_0^{2\pi} u(x,t)\overline{\frac{\partial\phi}{\partial t}(x,t)}\,dx\,dt - c\int_0^T\int_0^{2\pi} u(x,t)\overline{\frac{\partial^2\phi}{\partial x^2}(x,t)}\,dx\,dt + b\int_0^T\int_0^{2\pi} u(x,t)\overline{\phi(x,t)}\,dx\,dt \\ = \int_0^T\int_0^{2\pi} f(x,t)\overline{\phi(x,t)}\,dx\,dt. \end{aligned}\right\}. \quad (2.55)$$

As the space consisting of finite linear combinations of functions ϕ of the form (2.54) is dense in $\mathbf{D}((0,2\pi)\times(0,T))$, (2.55) is equivalent to

$$\left\langle \frac{\partial u}{\partial t} - c\frac{\partial^2 u}{\partial x^2} + bu, \phi \right\rangle_{\mathbf{D}'\times\mathbf{D}} = \langle f, \phi\rangle_{\mathbf{D}'\times\mathbf{D}} \text{ for all } \phi\in\mathbf{D}((0,2\pi)\times(0,T)). \quad (2.56)$$

Therefore any solution of the variational problem (2.17)-(2.19) is also a solution, in the sense of distributions, of the initial-boundary value problem (2.1)-(2.3).

Chapter 3. A semi-discrete spectral method for the general linear diffusion problem

The work in this chapter depends on the theory done in Chapter 2 (especially in Theorem 2.4). We therefore continue with the notational conventions used in the previous chapter. For convenience, we will repeat the most important formulae here. We will also assume that u denotes the solution of (2.1)-(2.3), or

$$\left.\begin{aligned} \frac{\partial u}{\partial t} - c\frac{\partial^2 u}{\partial x^2} + bu &= f \text{ on } (0,2\pi)\times(0,T) \\ u(x,0) &= u_0(x) \\ u(0,t) &= u(2\pi,t), \end{aligned}\right\} \tag{3.1}$$

in the sense of Theorem 2.4.

In this chapter, we investigate a semi-discrete (in the x variable) approximation of the general linear diffusion problem (2.1)-(2.3) (or (3.1)) or (2.17)-(2.19). Given the analysis done in Chapter 2, it is natural to use a spectral method of Fourier-Galerkin type in the sense described below.

First requirement of the Fourier-Galerkin spectral method

With each positive integer m, we associate the finite dimensional subspace $\mathbf{S}_m$ of $\mathbf{H}^1_{\text{per}}(0,2\pi)$ defined in (2.31) by

$$\mathbf{S}_m := \text{span}\{w_k\}_{k=-m}^{m}, \text{ where } w_k(x) = \frac{1}{\sqrt{2\pi}} e^{ikx}. \tag{3.2}$$

Second requirement of the Fourier-Galerkin spectral method

For each positive integer m, we approximate the solution u of (2.1)-(2.3) or (3.1) by u_m, which is represented in the form

$$u_m(t) = \sum_{k=-m}^{m} \alpha_k(t) w_k, \tag{3.3}$$

with unknown coefficients $\{\alpha_k\}_{k=-m}^{m}$.

Third requirement of the Fourier-Galerkin spectral method

In general, u_m will not satisfy

$$\frac{\partial u}{\partial t} - c\frac{\partial^2 u}{\partial x^2} + bu = f \text{ and } u(0) = u_0, \tag{3.4}$$

i.e. the residuals

$$Ru_m := \frac{\partial u_m}{\partial t} - c\frac{\partial^2 u_m}{\partial x^2} + bu_m - f \text{ and } Iu_m := u_m(0) - u_0 \tag{3.5}$$

will not vanish everywhere. Consequently, the third requirement of the Fourier-Galerkin spectral method is to minimize the residuals Ru_m and Iu_m by demanding that

$$\left.\begin{aligned} \langle Ru_m(t), v\rangle_0 = 0 \text{ for all } v \in \mathbf{S}_m \text{ and } t \in (0,T) \\ \langle Iu_m, v\rangle_0 = 0 \text{ for all } v \in \mathbf{S}_m \end{aligned}\right\} \tag{3.6}$$

or, equivalently, for $k = -m, -m+1, \mathrm{K}, m$, that (2.37) be satisfied:

$$\left.\begin{aligned} \frac{d}{dt}\alpha_k(t) + \lambda_k\alpha_k(t) = \langle f(t), w_k\rangle_0 \text{ for } t \in (0,T) \text{ where } \lambda_k = ck^2 + b \\ \alpha_k(0) = \langle u_0, w_k\rangle_0. \end{aligned}\right\} \tag{3.7}$$

Theorem 3.1. The Fourier-Galerkin spectral approximation (3.3) & (3.7) is the m^{th} partial sum of the Fourier series of u. Consequently $(u_m)_{m=1}^{\infty}$ converges to u in $\mathbf{L}^2\left((0,T), \mathbf{H}^1_{\text{per}}(0,2\pi)\right)$ and $\mathbf{C}^0\left([0,T], \mathbf{L}^2(0,2\pi)\right)$.

Proof. This is a consequence of (2.30) and (2.38), where the Fourier series of u and the solution to (3.1) and (3.7) are given respectively. The convergence of the sequence $(u_m)_{m=1}^{\infty}$ to u in the stated topologies was established in Step 3 (p. 32) of the proof of Theorem 2.4.

■

Remark 3.2. An alternative proof of the convergence of the Fourier-Galerkin spectral approximation (3.3) & (3.7) can be found using the classical framework of Lax-Richtmyer, i.e. consistency and stability (cf. Dautray & Lions (2000c: p. 37)).

The approximation (3.3) & (3.7) is consistent in the sense that there exist, for $m \in \mathbf{N}$, projection operators $P_m : \mathbf{L}^2(0,2\pi) \to \mathbf{S}_m$ such that

$$\lim_{m\to\infty} \|v - P_m v\|_0 = 0 \text{ for all } v \in \mathbf{L}^2(0,2\pi). \tag{3.8}$$

Indeed, taking $P_m v := \sum_{k=-m}^{m} \langle v, w_k\rangle w_k$, Theorem 1.1.3 immediately leads to (3.8).

In addition, the approximation is stable in the sense that

$$\|u_m(t)\|_0^2 \le \|u_0\|_0^2 + \frac{1}{\alpha}\|f\|^2_{\mathbf{L}^2((0,T),\mathbf{L}^2(0,2\pi))} \text{ where } \alpha = \min\{b,c\}. \tag{3.9}$$

This follows directly from (2.46), since

$$
\begin{aligned}
&\frac{1}{2}\|u_m(t)\|_0^2+\frac{\alpha}{2}\int_0^t\|u_m(s)\|_1^2\,ds\\
&\le\frac{1}{2}\|u_m(t)\|_0^2+\alpha\int_0^t\|u_m(s)\|_1^2\,ds\\
&\le\frac{1}{2}\sum_{k=-m}^{m}\left|\langle u_0,w_k\rangle_0\right|^2+\int_0^t a(u_m(s),u_m(s))ds && \begin{pmatrix}\mathbf{H}^1(0,2\pi)\text{-coercivity of } a\\ \text{Parseval identity}\end{pmatrix}\\
&=\frac{1}{2}\sum_{k=-m}^{m}\left|\langle u_0,w_k\rangle_0\right|^2+\int_0^t\langle f(s),u_m(s)\rangle_0\,ds && (\text{Energy equality})\\
&\le\frac{1}{2}\sum_{k=-m}^{m}\left|\langle u_0,w_k\rangle_0\right|^2+\int_0^t\|f(s)\|_0\|u_m(s)\|_1\,ds && \begin{pmatrix}\text{Cauchy - Schwarz}\\ \text{inequality}\end{pmatrix}\\
&\le\frac{1}{2}\sum_{k=-m}^{m}\left|\langle u_0,w_k\rangle_0\right|^2+\frac{1}{2\alpha}\int_0^t\|f(s)\|_0^2\,ds+\frac{\alpha}{2}\int_0^t\|u_m(s)\|_1^2\,ds && (\text{Young inequality})\\
\|u_m(t)\|_0^2&\le\sum_{k=-m}^{m}\left|\langle u_0,w_k\rangle_0\right|^2+\frac{1}{\alpha}\int_0^T\|f(s)\|_0^2\,ds\\
&\le\|u_0\|_0^2+\frac{1}{\alpha}\|f\|^2_{\mathbf{L}^2((0,T),\mathbf{L}^2(0,2\pi))} && (\text{Bessel inequality}).
\end{aligned}
$$

In addition to the above convergence results, we have the following exponential rate of convergence:

Theorem 3.3. For $m \in \mathbf{N}$, there holds the pointwise error estimate

$$\|u(t)-u_m(t)\|_0 \le \frac{1}{m}\left\|\frac{du}{dx}(t)\right\|_0 \quad \text{for } t\in(0,T). \tag{3.10}$$

Furthermore, if u satisfies the regularity condition

$$u(t)\in \mathbf{H}^p(0,2\pi) \text{ for } t\in(0,T) \tag{3.11}$$

for some integer $p > 1$, then

$$\|u(t)-u_m(t)\|_0 \le \frac{1}{m^p}\left\|\frac{d^p u}{dx^p}(t)\right\|_0 \quad \text{for } t\in(0,T) \tag{3.12}$$

$$\|u(t)-u_m(t)\|_1 \le \frac{1}{m^{p-1}}\left\|\frac{d^{p-1} u}{dx^{p-1}}(t)\right\|_1 \quad \text{for } t\in(0,T). \tag{3.13}$$

Proof. If u satisfies (3.11) for some $p \in \mathbf{N}$, then

$$
\begin{aligned}
\|u(t)-u_m(t)\|_0^2 &= \sum_{|k|>m}\left|\langle u(t),w_k\rangle_0\right|^2 && (\text{Parseval identity})\\
&= \sum_{|k|>m}\frac{1}{k^{2p}}\left|k^p\langle u(t),w_k\rangle_0\right|^2\\
&\le \frac{1}{m^{2p}}\sum_{|k|>m}\left|k^p\langle u(t),w_k\rangle_0\right|^2
\end{aligned}
$$

$$\begin{aligned}\|u(t)-u_m(t)\|_0^2 &\le \frac{1}{m^{2p}}\sum_{k\in\mathbf{Z}}\left|k^p\langle u(t),w_k\rangle_0\right|^2\\ &= \frac{1}{m^{2p}}\left\|\frac{d^p u}{dx^p}(t)\right\|_0^2 \qquad \text{(Parseval identity)}.\end{aligned}$$

Since

$$\|u(t)-u_m(t)\|_1^2 = \|u(t)-u_m(t)\|_0^2 + \left\|\frac{du}{dx}(t)-\frac{du_m}{dx}(t)\right\|_0^2 \text{ with } \frac{du}{dx}(t)\in\mathbf{H}^{p-1}(0,2\pi), \qquad (3.14)$$

the estimate (3.13) follows from (3.12).

■

Remark 3.4. If $u_0\in\mathbf{H}^1_{\text{per}}(0,2\pi)$, we know that $u(t)\in\mathbf{H}^2(0,2\pi)$, i.e. $p = 2$ (cf. Theorem 2.8).

Remark 3.5. The estimates in Theorem 3.3 are sharper than those obtained by Thomée (1997: Chapter 1), essentially because in this case $u_m(t)$ is the best approximation of $u(t)$ in $\mathbf{S}_m$ with respect to the inner product $\langle\cdot,\cdot\rangle_0$ (cf. Reddy (1998: pp. 192-193), Canuto et al (1988: p. 277)).

For example, Thomée (1997: Theorem 1.3) states that

$$\|u(t)-u_m(t)\|_0 \le \|u_0-u_m(0)\|_0 + \frac{C}{m}\left(\|u_0\|_1 + \int_0^t\left\|\frac{du}{ds}(s)\right\|_1 ds\right) \qquad (3.15)$$

under the assumptions

$$u(t)\in\mathbf{H}^2(0,2\pi) \text{ and } \frac{du}{dt}(t)\in\mathbf{H}^1(0,2\pi). \qquad (3.16)$$

Our rate of convergence $\|u(t)-u_m(t)\|_0 = \mathrm{O}\left(\frac{1}{m}\right)$ in (3.10) is valid without this assumption. Even when (3.16) is met, our estimate is sharper. Indeed, using the seminorm on $\mathbf{H}^1(0,2\pi)$ defined by

$$|v|_1 := \left\|\frac{dv}{dx}\right\|_0 \text{ for } v\in\mathbf{H}^1(0,2\pi), \qquad (3.17)$$

we have

$$\begin{aligned}\|u(t)-u_m(t)\|_0 &\le \frac{1}{m}\left\|\frac{du}{dx}(t)\right\|_0\\ &= \frac{1}{m}|u(t)|_1\end{aligned}$$

$$\begin{aligned}\left\|u(t)-u_m(t)\right\|_0 &\le \frac{1}{m}\left|u_0+\int_0^t \frac{du}{dt}(s)ds\right|_1 \\ &\le \frac{1}{m}\left(\left\|\frac{du_0}{dx}\right\|_0+\int_0^t\left\|\frac{\partial^2 u}{\partial x\partial t}(s)\right\|_0 ds\right).\end{aligned}$$

Chapter 4. Finite difference methods for a first order initial value problem

This chapter is devoted to some numerical approximations of initial value problems for ordinary differential equations. In Section 4.1, we define, in a general framework, the concepts of consistency, zero-stability and convergence of finite difference methods that correspond to linear one-step methods. These concepts are applied to the classical θ-method in Section 4.2. In Section 4.3, we design an original scheme, the non-standard θ-method, which has the advantage of preserving some qualitative properties of the solution of the initial value problem.

4.1. Generalities

We consider the initial value problem for the first order ordinary differential equation

$$\left.\begin{aligned} Dy \equiv \frac{dy}{dt} &= g(t,y) \text{ in } (0,T) \\ y(0) &= \eta. \end{aligned}\right\} \tag{4.1}$$

We assume once and for all that the problem (4.1) is well-posed, i.e. it has a unique solution that depends continuously upon the data. Typically, this is true when the function $(t,z) \to g(t,z)$ is continuous on $(0,T)\times(-\infty,\infty)$ and satisfies the Lipschitz condition

$$|g(t,z_1)-g(t,z_2)| \le L|z_1-z_2| \text{ for all } t \in (0,T),\ z_1 \in (-\infty,\infty) \text{ and } z_2 \in (-\infty,\infty) \tag{4.2}$$

for some constant *L* (cf. Lambert (1991: p. 6), Burden & Faires (1997: pp. 256-257)). Consequently, we assume that (4.2) is true. Whenever necessary, we assume more regularity on *g* so that the solution *y* is smoother.

For the numerical approximation of (4.1), we fix an integer *N*, and define the step-size

$$\Delta t := \frac{T}{N}. \tag{4.3}$$

We then replace the continuous interval $[0,T]$ by the mesh of equidistant points

$$\{t_n | t_n := n\Delta t \text{ for } n = 0,1,\text{K},N\}. \tag{4.4}$$

For $n = 0,1,\text{K},N$, y_n is used to denote an approximation of $y(t_n)$, whereas g_n denotes $g(t_n,y_n)$. The finite difference method entails solving in y_n the equation

$$D_{\Delta t} y_n = F_{\Delta t}(y_n, g_n), \tag{4.5}$$

where $D_{\Delta t}$ is a difference operator such that $D_{\Delta t} y_n$ approximates $Dy(t_n)$, and $F_{\Delta t}(y_n, g_n)$ approximates $g(t_n, y(t_n))$ in some way.

The finite difference method that we consider in this work is a particular case of the linear one-step methods investigated by Lambert (1973: Chapters 2 & 3, 1991: Chapter 3). Consequently the three concepts that form an integral part of contemporary analysis of numerical methods are defined below in accordance to Lambert.

Definition 4.1.1.

(a) **Consistency.** (Lambert, 1991: p. 28) The finite difference scheme (4.5), viewed as a linear one-step method, is called consistent, provided that for all well-posed initial value problems (4.1) with exact solution *y*, the truncation error $D_{\Delta t} y(t) - F_{\Delta t}(y(t), g(t, y(t)))$ satisfies

$$\lim_{\Delta t \to \infty} \left(D_{\Delta t} y(t) - F_{\Delta t}(y(t), g(t, y(t))) \right) = 0 \text{ for any } t \in [0, T]. \tag{4.6}$$

(b) **Zero-stability.** (Lambert, 1991: p. 32) The finite difference scheme (4.5) is called zero-stable if there exist constants $K > 0$ and $h_0 > 0$ such that, for all $\Delta t \in (0, h_0]$, the relation $|y_n - \tilde{y}_n| \leq K\varepsilon$ holds whenever $|\delta_n - \tilde{\delta}_n| \leq \varepsilon$ for a given accuracy $\varepsilon > 0$ and any two perturbations $(\delta_n)_{n=0}^N$ and $(\tilde{\delta}_n)_{n=0}^N$ of the data in (4.1) with corresponding perturbed solutions $(y_n)_{n=0}^N$ and $(\tilde{y}_n)_{n=0}^N$.

(c) **Convergence.** (Lambert, 1973: p. 22) The finite difference scheme (4.5), viewed as a linear one-step method, is called convergent if, for all well-posed initial value problems (4.1) with exact solution *y*, we have

$$\lim_{\substack{\Delta t \to 0 \\ n\Delta t = t}} |y_n - y(t)| = 0 \text{ for any } t \in [0, T] \tag{4.7}$$

for all solutions $(y_n)_{n=0}^N$ of (4.5) satisfying $y_0 = \eta(\Delta t)$, where $\lim_{\Delta t \to 0} \eta(\Delta t) = \eta$.

In this conceptual framework, we have the following equivalence theorem due to Dahlquist (cited by Lambert (1973: p. 33, 1991: p. 36)):

Theorem 4.1.2. The necessary and sufficient conditions of the method (4.5) to be convergent are that it be both consistent and zero-stable.

4.2. The θ-method

The θ-method for approximating (4.1) reads

$$\left.\begin{aligned} \frac{y_{n+1}-y_n}{\Delta t} &= \theta g_{n+1}+(1-\theta)g_n \text{ for } n=0,1,\mathrm{K}\,,N-1 \\ y_0 &= \eta, \end{aligned}\right\} \tag{4.8}$$

where $\theta\in[0,1]$ is a given parameter. In the notation of (4.5), we have

$$D_{\Delta t}y_n=\frac{y_{n+1}-y_n}{\Delta t} \text{ and } F_{\Delta t}(y_n,g_n)=\theta g_{n+1}+(1-\theta)g_n\,. \tag{4.9}$$

As mentioned earlier, the θ-method (4.8) is a linear one-step method, the structure of a linear *k*-step method being (Lambert, 1973: p. 4)

$$\frac{1}{\Delta t}\sum_{j=0}^{k}\alpha_j y_{n+j}=\sum_{j=0}^{k}\beta_j g_{n+j} \text{ for } n=0,1,\mathrm{K}\,,N-1\,. \tag{4.10}$$

For some specific values of θ, the θ-method is known under the names indicated below:

Value of θ	**Name of method**
$\theta=0$	Forward Euler method
$\theta=\frac{1}{2}$	Trapezoidal rule, Crank-Nicolson method
$\theta=1$	Backward Euler method

We now present the convergence of the θ-method.

Theorem 4.2.1. The θ-method (4.8) is consistent, zero-stable and thus convergent.

Proof. To prove consistency, we consider any well-posed initial value problem (4.1) with continuously differentiable solution *y* on $[0,T]$. The truncation error of the θ-method (4.8) at a fixed $t\in[0,T]$ is

$$\tau_{\Delta t}(t):=\frac{y(t+\Delta t)-y(t)}{\Delta t}-\left(\theta g(t+\Delta t,y(t+\Delta t))+(1-\theta)g(t,y(t))\right). \tag{4.11}$$

We therefore have

$$\begin{aligned}\lim_{\Delta t\to 0}\tau_{\Delta t}(t)&=\lim_{\Delta t\to 0}\left(\frac{y(t+\Delta t)-y(t)}{\Delta t}-\left(\theta g(t+\Delta t,y(t+\Delta t))+(1-\theta)g(t,y(t))\right)\right)\\ &=\lim_{\Delta t\to 0}\left(\frac{y(t+\Delta t)-y(t)}{\Delta t}\right)-\lim_{\Delta t\to 0}\left(\theta g(t+\Delta t,y(t+\Delta t))\right)+(1-\theta)g(t,y(t))\end{aligned}$$

$$\lim_{\Delta t \to 0} \tau_{\Delta t}(t) = \frac{dy}{dt}(t) - g(t, y(t))$$
$$= 0.$$

By Definition 4.1.1(a), this means that the θ-method is consistent.

For zero-stability, it is difficult to proceed directly by using Definition 4.1.1(b). However, it is known that the linear multi-step method (4.10) is zero-stable if and only if every root of the first characteristic polynomial

$$\rho(z) := \sum_{j=0}^{k} \alpha_j z^j \tag{4.12}$$

has modulus less than or equal to 1, and those with modulus 1 are simple (cf. Lambert (1991: p. 35)). Since the θ-method (4.8) is a linear one-step method for which the only root, $z = 1$, of the corresponding polynomial $\rho(z)$ is simple, we conclude that the θ-method is zero-stable.

Using Theorem 4.1.2, (4.7) follows.

■

Remark 4.2.2. Some clarification is necessary about the convergence proved in Theorem 4.2.1. In view of Definition 4.1.1(b), the zero-stability of the θ-method simply means that this method is insensitive to perturbations whenever the step-size Δt approaches 0. In other words, zero-stability controls the manner in which errors accumulate, but only in the limit as $\Delta t \to 0$. However, in practice the limit of Δt is never reached. What one obtains are different discrete solutions corresponding to different non-zero values of Δt.

In this regard, there exist several examples of linear multi-step methods that are convergent, i.e. consistent and zero-stable, but for which a value h_0 of the step-size may be found such that, for $\Delta t > h_0$, the error of the method increases as Δt increases, whereas for $\Delta t < h_0$ it decreases. Let us illustrate this fact for the forward Euler method (i.e. $\theta = 0$ in (4.8)). For this method, it can be shown (cf. Burden & Faires (1997: pp. 264-266)) that there holds the error bound

$$|y(t_n) - y_n| \le \frac{M\Delta t}{2L}\left(e^{Lt_n} - 1\right) \text{ for } n = 0,1,\text{K}\,,N\,, \tag{4.13}$$

where L is the Lipschitz constant in (4.2) and y is supposed to be of class $\mathbf{C}^2[0,T]$ such that

$$\left|\frac{d^2 y}{dt^2}(t)\right| \le M < \infty \text{ for } t \in [0,T] \tag{4.14}$$

for some constant M. Furthermore, if one considers the perturbed Euler method

$$\begin{aligned} \tilde{y}_{n+1} &= \tilde{y}_n + \Delta t g(t_n, \tilde{y}_n) + \delta_{n+1} \text{ for } n = 0,1,\text{K}\,,N-1 \\ \tilde{y}_0 &= \eta + \delta_0, \end{aligned} \tag{4.15}$$

where $|\delta_n| \le \delta$ for $n = 0,1,\mathrm{K}\,,N$, then Burden & Faires show that

$$|y(t_n) - \tilde{y}_n| \le \frac{2}{L}\left(\frac{M\Delta t}{2} + \frac{\delta}{\Delta t}\right)\left(e^{Lt_n} - 1\right) + |\delta_0| e^{Lt_n} \text{ for } n = 0,1,\mathrm{K}\,,N. \tag{4.16}$$

The error bound (4.16) is no longer linear in Δt. In fact, since

$$\lim_{\Delta t \to 0}\left(\frac{M\Delta t}{2} + \frac{\delta}{\Delta t}\right) = \infty, \tag{4.17}$$

the error would be expected to become large for sufficiently small values of Δt. Let

$$E(\Delta t) := \frac{M\Delta t}{2} + \frac{\delta}{\Delta t}. \tag{4.18}$$

Then

$$\frac{dE}{d\Delta t}(\Delta t) = \frac{M}{2} - \frac{\delta}{(\Delta t)^2}, \tag{4.19}$$

and if $\Delta t < \sqrt{\frac{2\delta}{M}}$, then $\frac{dE}{d\Delta t}(\Delta t) < 0$ and E is decreasing. Similarly, if $\Delta t > \sqrt{\frac{2\delta}{M}}$, then $\frac{dE}{d\Delta t}(\Delta t) > 0$ and E is increasing. The minimal value of E occurs when

$$\Delta t = \sqrt{\frac{2\delta}{M}}. \tag{4.20}$$

Decreasing Δt beyond this value will tend to increase the total error in the approximation.

The above comments motivate the need for another stability theory that applies when Δt takes a fixed non-zero value. For a fixed step-size $\Delta t > 0$, stability of the θ-method means that the propagation of error is insignificant as $n \to \infty$. Following Raviart & Thomas (1983: Chapter 7), we assume, as usual, that the derivative $\frac{\partial g}{\partial y}$ is constant, and consider the model and test problem

$$\left.\begin{aligned} \frac{dy}{dt} &= -\lambda y \text{ in } (0,\infty) \\ y(0) &= \eta, \end{aligned}\right\} \tag{4.21}$$

where, for convenience, we have

$$-\lambda \equiv \frac{\partial g}{\partial y} < 0. \tag{4.22}$$

The solution of (4.21) is

$$y(t) = \eta e^{-\lambda t}. \tag{4.23}$$

The θ-method (4.8) applied to (4.21) yields

$$y_{n+1} = \frac{1-(1-\theta)\lambda\Delta t}{1+\theta\lambda\Delta t} y_n \text{ for } n \in \mathbf{N}$$
$$y_0 = \eta,$$

and therefore

$$y_n = \left(\frac{1-(1-\theta)\lambda\Delta t}{1+\theta\lambda\Delta t}\right)^n \eta \text{ for } n \in \mathbf{N}. \tag{4.24}$$

The exact solution (4.23) and the discrete solution (4.24) are identical at the initial time $t = 0$. In view of (4.23) and (4.24), the propagation of error will be insignificant as $n \to \infty$ if and only if

$$\left|\frac{1-(1-\theta)\lambda\Delta t}{1+\theta\lambda\Delta t}\right| \le 1, \tag{4.25}$$

which is the condition of stability for the θ-method for a fixed Δt. When

$$\left|\frac{1-(1-\theta)\lambda\Delta t}{1+\theta\lambda\Delta t}\right| < 1, \tag{4.26}$$

we say that the θ-method is absolutely stable for the fixed step-size Δt to express the fact that $(y_n)_{n=1}^{\infty}$ in (4.24) goes to 0 as $n \to \infty$, like $y(t)$ in (4.23) as $t \to \infty$.

Remark 4.2.3. It is easy to see that if $\theta \ge \frac{1}{2}$, then the θ-method is stable for any $\Delta t > 0$. According to standard terminology, the θ-method is then said to be unconditionally stable. If $\theta < \frac{1}{2}$, then the θ-method is stable only for Δt satisfying

$$\lambda\Delta t \le \frac{2}{1-2\theta}. \tag{4.27}$$

In this case, the θ-method is called conditionally stable.

Remark 4.2.4. Setting $h := -\lambda\Delta t$, the number

$$x = \frac{1+(1-\theta)h}{1-\theta h} \tag{4.28}$$

is the only root of the stability polynomial $\pi(r,h)$ (defined by Lambert (1973: p. 65)) for the θ-method (4.8) viewed as a linear one-step method (4.10). It is in terms of (4.28) and (4.26) that Lambert (1991: p. 70) defines the concept of absolute stability.

4.3. The non-standard θ-method

In this section, we will develop an original powerful variant of the θ-method. We assume that the initial value problem for the first order differential equation (4.1) is autonomous, i.e.

$$\left.\begin{aligned} Dy \equiv \frac{dy}{dt} &= g(y) \text{ in } (0,T) \\ y(0) &= \eta, \end{aligned}\right\} \tag{4.29}$$

the other conditions on *g* being satisfied.

Non-standard finite difference schemes for (4.29) were introduced by Mickens (1994: Chapter 4) as powerful numerical methods that preserve significant properties of exact solutions of the involved differential equations. Schemes were empirically developed using a collection of rules set by Mickens.

Anguelov & Lubuma (2000, 2001a, 2001b) provide some mathematical justifications for the success of these empirical procedures. In particular, non-standard finite difference schemes can be defined as follows by using two of Mickens' rules:

Definition 4.3.1. The scheme (4.5) is called a non-standard finite difference method if at least one of the following conditions is met:

(a) In the first order discrete derivative $D_{\Delta t} y_n$ that occurs in (4.5), the traditional denominator Δt is replaced by a positive function ϕ such that

$$\phi(\Delta t) = \Delta t + O\left((\Delta t)^2\right) \text{ as } \Delta t \to 0 . \tag{4.30}$$

(b) Non-linear terms in $g(y)$ are approximated in a non-local way, i.e. by a suitable function of several points of the mesh (e.g. $(y(t_n))^2 \approx y_n y_{n+1}$, $(y(t_n))^3 \approx y_n^2 y_{n+1}$).

The power of the non-standard finite difference method over the standard method is expressed in the next definition:

Definition 4.3.2. P-stability. (Anguelov & Lubuma (2000, 2001a)) Assume that the solutions of (4.29) satisfy some property **P**. The numerical scheme (4.5) is called (qualitatively) stable with respect to the property **P** (or **P**-stable) if, for every value of $\Delta t > 0$, the set of solutions of (4.5) satisfy property **P**.

Our aim is to construct a non-standard θ-method that is stable with respect to the properties of fixed-points of (4.29). Our construction is based on Definition 4.3.1(a) (often referred to as renormalization of the denominator). In other words, we are, in view of (4.8), looking for discrete schemes of the form

$$\frac{y_{n+1}-y_n}{\phi(\Delta t)} = \theta g_{n+1} + (1-\theta) g_n \text{ for } n = 0,1,\mathrm{K}\,, N-1\,, \tag{4.31}$$

or, equivalently,

$$y_{n+1} = y_n + \phi(\Delta t)\left(\theta g_{n+1} + (1-\theta) g_n\right) \equiv G(\Delta t, y_n, y_{n+1}) \text{ for } n = 0,1,\mathrm{K}\,, N-1 \tag{4.32}$$

with suitable initial values.

Definition 4.3.3. Fixed-point of differential equation. Any constant $\tilde{y}$ such that $g(\tilde{y}) = 0$, is called a fixed-point of the differential equation (4.29).

We consider fixed-points with regard to the behaviour of the other solutions of the differential equation around them. Generally, fixed-points which attract other solutions in some neighbourhood when $t \to \infty$, are called stable, those that repulse other solutions are called unstable. We restrict our study to hyperbolic fixed-points $\tilde{y}$, i.e. fixed-points $\tilde{y}$ satisfying the relation

$$J \equiv \frac{dg}{dy}(\tilde{y}) \neq 0\,. \tag{4.33}$$

The reason for this restriction is the Hartman & Grobman theorem (Stuart & Humphries, 1998: pp. 156 & 164), which guarantees that the asymptotic behaviour of solutions of (4.29) with initial data near $\tilde{y}$ may be reduced to the behaviour of solutions of the linear equation

$$\frac{d\varepsilon}{dt} = J\varepsilon\,. \tag{4.34}$$

The equation (4.34) may formally be obtained as follows. Consider the perturbed trajectory $y(t) := \tilde{y} + \varepsilon(t)$. We have

$$\begin{aligned} \frac{d\varepsilon}{dt} &= \frac{dy}{dt} \\ &= g(\tilde{y} + \varepsilon) \\ &= g(\tilde{y}) + \varepsilon \frac{dg}{dy}(\tilde{y}) + O(\varepsilon^2), \end{aligned}$$

using a Taylor expansion of g around $\tilde{y}$. Since we are only interested in small values of ε, we retain only the linear part of this differential equation, and (4.34) follows.

Definition 4.3.4. Linear stability of fixed-point of differential equation. A hyperbolic fixed-point $\tilde{y}$ of (4.29) is called linearly stable provided that the solution ε of (4.34) corresponding to any small enough initial data satisfies $\lim_{t\to\infty}\varepsilon(t)=0$. Otherwise, the fixed-point is called linearly unstable.

Remark 4.3.5. The linear stability of a hyperbolic fixed-point $\tilde{y}$ of (4.29) is equivalent to having $J < 0$ in (4.33). Likewise, a fixed-point $\tilde{y}$ is linearly unstable if and only if $J > 0$ in (4.33).

We now turn to the discrete analogue of our discussion on fixed-points.

Definition 4.3.6. Fixed-point of discrete scheme. A constant $\tilde{y}$ is called a fixed-point of the difference scheme (4.31) or (4.32) if $y_n = \tilde{y}$ is a fixed-point of the mapping G in (4.32).

Remark 4.3.7. In view of the definition of G, $\tilde{y}$ is a fixed-point of the discrete scheme (4.31) if and only if $\tilde{y}$ is a fixed-point of the differential equation in (4.29).

Let $\tilde{y}$ be a fixed-point of the discrete scheme (4.31). Consider, from (4.31), the discrete trajectory $y_n := \tilde{y} + \varepsilon_n$ and the discrete perturbation equation

$$\frac{\varepsilon_{n+1}-\varepsilon_n}{\phi(\Delta t)} = \theta g(\tilde{y}+\varepsilon_{n+1}) + (1-\theta) g(\tilde{y}+\varepsilon_n) \text{ for } n = 0,1,\mathrm{K}\,,N-1\,. \tag{4.35}$$

Taylor expansion of $g(\tilde{y}+\varepsilon_{n+1})$ and $g(\tilde{y}+\varepsilon_n)$ around $\tilde{y}$ yields, on retaining only the linear part in ε_n, the equation

$$\frac{\varepsilon_{n+1}-\varepsilon_n}{\phi(\Delta t)} = \theta J\varepsilon_{n+1} + (1-\theta) J\varepsilon_n \text{ for } n = 0,1,\mathrm{K}\,,N-1\,, \tag{4.36}$$

or equivalently

$$\varepsilon_{n+1} = J_{\Delta t}\varepsilon_n \text{ with } J_{\Delta t} = \frac{1+\phi(\Delta t)(1-\theta)J}{1-\phi(\Delta t)\theta J} \text{ for } n = 0,1,\mathrm{K}\,,N-1\,, \tag{4.37}$$

which is the discrete analogue of the error equation (4.34).

Definition 4.3.8. Linear stability of fixed-point of discrete scheme. Assume that $\tilde{y}$ is a hyperbolic fixed-point of (4.29). As a fixed-point of (4.31), this $\tilde{y}$ is called linearly stable provided that the solution $(\varepsilon_n)_{n\geq 0}$ of (4.37) corresponding to any small enough initial data satisfies $\lim_{n\to\infty}\varepsilon_n = 0$. Otherwise, the fixed-point is called linearly unstable.

Remark 4.3.9. The linear stability of a hyperbolic fixed-point $\tilde{y}$ of (4.31) is equivalent to having $|J_{\Delta t}| < 1$ in (4.37). Likewise, a fixed-point $\tilde{y}$ is linearly unstable if and only if $|J_{\Delta t}| \geq 1$ in (4.37).

We now formalize the stability under consideration, as a special case of the **P**-stability in Definition 4.3.2:

Definition 4.3.10. Elementary stability. The finite difference scheme (4.5) is called elementary stable if, for any value of the step-size Δt, its only fixed-points $\tilde{y}$ are those of the differential equation (4.29), the linear stability properties of each $\tilde{y}$ being the same for both the differential equation and the discrete scheme.

Theorem 4.3.11. The standard θ-method (4.8) is not elementary stable.

Proof. It is sufficient to consider the forward Euler method. i.e. $\theta = 0$ in (4.8). Let $\tilde{y}$ be a fixed-point of the differential equation (4.29) and the finite difference scheme (4.8). The discrete error equation (4.37) becomes

$$\varepsilon_{n+1} = (1 + J\Delta t)\varepsilon_n \text{ for } n = 0,1,\text{K}\,,N-1, \tag{4.38}$$

so that

$$\varepsilon_n = (1 + J\Delta t)^n \varepsilon_0 \text{ for } n = 0,1,\text{K}\,,N. \tag{4.39}$$

Suppose that the fixed-point $\tilde{y}$ is linearly stable for the differential equation (4.29) (i.e. $J < 0$). However, $(\varepsilon_n)_{n=1}^{\infty}$ diverges for $\Delta t \geq \frac{2}{-J}$, which means that $\tilde{y}$ is linearly unstable for the finite difference scheme (4.8).

■

Remark 4.3.12. The only exception to the result in Theorem 4.3.11 is the Crank-Nicolson method, i.e. $\theta = \frac{1}{2}$, which is elementary stable (cf. Anguelov & Lubuma (2001a)).

In contrast to the negative result in Theorem 4.3.11, we have the following new elementary stable non-standard scheme that extends the results of Anguelov & Lubuma (2001a) and Mickens (1994: Section 4.2) regarding the forward and backward Euler methods:

Theorem 4.3.13. Let ϕ be a real-valued function on **R** that satisfies (4.30) and the additional condition

$$0 < \phi(x) < 1 \text{ for } x > 0 \tag{4.40}$$

(e.g. $\phi(x) = 1 - e^{-x}$). Assume that the differential equation (4.29) has a finite number of fixed-points, all hyperbolic. Set

$$q := \max\left\{ \left| \frac{dg}{dy}(\tilde{y}) \right| \middle| g(\tilde{y}) = 0 \right\}. \tag{4.41}$$

Then the non-standard θ-method

$$\left.\begin{aligned} \frac{y_{n+1} - y_n}{\frac{\phi(q\Delta t)}{q}} &= \theta g_{n+1} + (1-\theta) g_n \text{ for } n = 0,1,\mathrm{K}\,, N-1 \\ y_0 &= \eta, \end{aligned}\right\} \tag{4.42}$$

or

$$D_{\Delta t} y_n = \frac{y_{n+1} - y_n}{\frac{\phi(q\Delta t)}{q}} \text{ and } F_{\Delta t}(y_n, g_n) = \theta g_{n+1} + (1-\theta) g_n \tag{4.43}$$

in the notation of (4.5), is elementary stable.

Proof. By Remark 4.3.7, $\tilde{y}$ is a fixed-point of the differential equation (4.29) if and only if $\tilde{y}$ is a fixed-point of (4.42).

Suppose that $\tilde{y}$ is a hyperbolic fixed-point of (4.29). The discrete error equation (4.37) reads

$$\varepsilon_{n+1} = \left(\frac{1 + \phi(q\Delta t)(1-\theta)\frac{J}{q}}{1 - \phi(q\Delta t)\theta\frac{J}{q}} \right) \varepsilon_n \text{ for } n = 0,1,\mathrm{K}\,, N-1, \tag{4.44}$$

so that

$$\varepsilon_n = \left(\frac{1 + \phi(q\Delta t)(1-\theta)\frac{J}{q}}{1 - \phi(q\Delta t)\theta\frac{J}{q}} \right)^n \varepsilon_0 \text{ for } n = 0,1,\mathrm{K}\,, N. \tag{4.45}$$

If $\tilde{y}$ is linearly unstable for the differential equation (i.e. $J > 0$), then

$$\left| \frac{1 + \phi(q\Delta t)(1-\theta)\frac{J}{q}}{1 - \phi(q\Delta t)\theta\frac{J}{q}} \right| = \frac{1 + \phi(q\Delta t)(1-\theta)\frac{J}{q}}{1 - \phi(q\Delta t)\theta\frac{J}{q}} \geq 1, \tag{4.46}$$

which, in view of (4.45), implies that $(\varepsilon_n)_{n\geq 1}$ diverges. Thus $\tilde{y}$ is also a linearly unstable fixed-point of the non-standard scheme (4.42). If, on the other hand, $\tilde{y}$ is linearly stable for the differential equation (i.e. $J < 0$), then

$$\left|\frac{1+\phi(q\Delta t)(1-\theta)\frac{J}{q}}{1-\phi(q\Delta t)\theta\frac{J}{q}}\right| = \left|\frac{1-\phi(q\Delta t)(1-\theta)\frac{|J|}{q}}{1+\phi(q\Delta t)\theta\frac{|J|}{q}}\right|$$

$$= \frac{1-\phi(q\Delta t)(1-\theta)\frac{|J|}{q}}{1+\phi(q\Delta t)\theta\frac{|J|}{q}}$$

$$< 1$$

because of (4.40) and (4.41). Consequently, in view of (4.45), $(\varepsilon_n)_{n\geq 1}$ converges to 0. Therefore $\tilde{y}$ is also linearly stable for the non-standard scheme (4.42).

■

The non-standard θ-method preserves the zero-stability, consistency and convergence properties of the standard θ-method.

Theorem 4.3.14. The non-standard θ-method (4.42) is consistent, zero-stable and thus convergent.

Proof. This is a consequence of Theorem 6 in Anguelov & Lubuma (2001a), observing that the function $F_{\Delta t}$ in (4.43) satisfies the Lipschitz condition

$$\sup_{n\in\mathbf{N}}|F_{\Delta t}(y_n, g(y_n)) - F_{\Delta t}(z_n, g(z_n))| \leq L\sup_{n\in\mathbf{N}}|y_n - z_n| \tag{4.47}$$

for any two bounded sequences $(y_n)_{n=1}^{\infty}$ and $(z_n)_{n=1}^{\infty}$, and L the constant in (4.2). Indeed,

$$\begin{aligned}
&|F_{\Delta t}(y_n, g(y_n)) - F_{\Delta t}(z_n, g(z_n))| \\
&= |\theta(g(y_{n+1}) - g(z_{n+1})) + (1-\theta)(g(y_n) - g(z_n))| \\
&\leq \theta|g(y_{n+1}) - g(z_{n+1})| + (1-\theta)|g(y_n) - g(z_n)| && \text{(Triangle inequality)} \\
&\leq L(\theta|y_{n+1} - z_{n+1}| + (1-\theta)|y_n - z_n|) && \text{(Lipschitz condition on } g) \\
&\leq L\sup_{n\in\mathbf{N}}|y_n - z_n|.
\end{aligned}$$

■

Remark 4.3.15. The non-standard θ-method is absolutely stable for all values of $\theta \in [0,1]$ and $\Delta t > 0$. Indeed, the non-standard analogue of (4.24) is given by

$$y_n = \left(\frac{1-\phi(\lambda\Delta t)(1-\theta)}{1+\phi(\lambda\Delta t)\theta} \right)^n \eta . \tag{4.48}$$

Using (4.40), it follows that

$$\left| \frac{1-\phi(\lambda\Delta t)(1-\theta)}{1+\phi(\lambda\Delta t)\theta} \right| < 1 . \tag{4.49}$$

At the end of this chapter, we have the following comparative table, which clearly shows the power of the non-standard θ-method (4.42) over the standard θ-method (4.8):

Property	**Standard θ-method**	**Non-standard θ-method**
Consistency	Yes	Yes
Zero-stability	Yes	Yes
Convergence	Yes	Yes
Absolute stability for all Δt	No	Yes
Elementary stability for all θ	No	Yes

Chapter 5. Full discretization of the general linear diffusion problem

In this chapter, we consider a fully discrete (i.e. in both the *x* and *t* variables) approximation of the general linear diffusion problem (2.1)-(2.3) or (2.17)-(2.19):

$$\left.\begin{aligned} \frac{\partial u}{\partial t} - c\frac{\partial^2 u}{\partial x^2} + bu &= f \text{ on } (0,2\pi)\times(0,T) \\ u(x,0) &= u_0(x) \\ u(0,t) &= u(2\pi,t). \end{aligned}\right\} \tag{5.1}$$

The Fourier-Galerkin spectral method (3.3) & (3.7) is used for the spatial approximation. For the discretization in time, we will consider, in turn, the θ-method (as discussed in Section 4.2) and the non-standard θ-method (Section 4.3).

We return to the notation used in Chapters 2 and 3. In particular, we will make use of the eigenvalues $\{\lambda_k\}_{k\in\mathbf{Z}}$ and eigenfunctions $\{w_k\}_{k\in\mathbf{Z}}$ of the eigenvalue problem (2.23), as given by (2.25) and (2.26):

$$\lambda_k = ck^2 + b \text{ and } w_k(x) = \frac{1}{\sqrt{2\pi}} e^{ikx}. \tag{5.2}$$

As previously, we will also assume that *u* denotes the solution of (5.1) in the sense of Theorem 2.4.

In addition, for a given number of time steps *N* and step-size Δt defined by (4.3), we replace, as before, the interval $[0,T]$ by the mesh (4.4). We adopt the notation

$$g_{k,n} := \langle f(t_n), w_k \rangle_0 \text{ and } f_{m,n} := \sum_{k=-m}^{m} \langle f(t_n), w_k \rangle_0 w_k, \tag{5.3}$$

and denote by $\alpha_{k,n}$ the finite difference approximation to $\alpha_k(t_n)$. Moreover, we set

$$u_{m,n} := \sum_{k=-m}^{m} \alpha_{k,n} w_k. \tag{5.4}$$

5.1. Spectral-θ-method

For fixed $m \in \mathbf{N}$ and $k = -m, -m+1, \mathrm{K}, m$, the sequence $(\alpha_{k,n})_{n=0}^{N}$ is determined by applying the θ-method (4.8) to the initial value problem (2.37) given by

$$\left.\begin{aligned} \frac{d}{dt}\alpha_k(t) + \lambda_k \alpha_k(t) &= \langle f(t), w_k \rangle \text{ for } t \in (0,T) \\ \alpha_k(0) &= \langle u_0, w_k \rangle. \end{aligned}\right\} \tag{5.5}$$

In other words, $(\alpha_{k,n})_{n=0}^{N}$ solves the difference equation

$$\left.\begin{aligned} \frac{\alpha_{k,n+1} - \alpha_{k,n}}{\Delta t} + \lambda_k\left(\theta\alpha_{k,n+1} + (1-\theta)\alpha_{k,n}\right) &= \theta g_{k,n+1} + (1-\theta) g_{k,n} \text{ for } n = 0,1,\mathrm{K}, N-1 \\ \alpha_{k,0} &= \langle u_0, w_k \rangle_0. \end{aligned}\right\} \tag{5.6}$$

Multiplying (5.6) by w_k, summing over $k = -m, -m+1, \mathrm{K}, m$, taking the inner product with any $v \in \mathbf{S}_m$ (with $\mathbf{S}_m$ defined in (2.31)) and using the identity (2.24) in successive order lead to the spectral-θ-method

$$\left.\begin{aligned} \left\langle \frac{u_{m,n+1} - u_{m,n}}{\Delta t}, v \right\rangle_0 + a\left(\theta u_{m,n+1} + (1-\theta)u_{m,n}, v\right) &= \left\langle \theta f_{m,n+1} + (1-\theta) f_{m,n}, v \right\rangle_0 \\ &\text{for } n = 0,1,\mathrm{K}, N-1 \text{ and } v \in \mathbf{S}_m \\ u_{m,0} &= \sum_{k=-m}^{m} \langle u_0, w_k \rangle_0 w_k. \end{aligned}\right\} \tag{5.7}$$

Remark 5.1.1. The simple structure of (5.5) is due to the orthonormality of the basis $\{w_k\}_{k=-m}^{m}$ of $\mathbf{S}_m$. In general, for an arbitrary basis $\{w_k\}_{k=-m}^{m}$, the analogue of (5.5) is obtained as follows. Define the vectors

$$\underline{\alpha}(t) := \begin{bmatrix} \alpha_{-m}(t) \\ \alpha_{-m+1}(t) \\ \mathrm{M} \\ \alpha_m(t) \end{bmatrix}, \quad \underline{\beta} := \begin{bmatrix} \langle u_0, w_{-m} \rangle_0 \\ \langle u_0, w_{-m+1} \rangle_0 \\ \mathrm{M} \\ \langle u_0, w_m \rangle_0 \end{bmatrix} \text{ and } \underline{\chi}(t) := \begin{bmatrix} \langle f(t), w_{-m} \rangle_0 \\ \langle f(t), w_{-m+1} \rangle_0 \\ \mathrm{M} \\ \langle f(t), w_m \rangle_0 \end{bmatrix}, \tag{5.8}$$

and the mass matrix $\mathbf{M} = (m_{ij}) := \left(\langle w_j, w_i \rangle_0\right)$ as well as the stiffness matrix $\mathbf{R} = (r_{ij}) := \left(a(w_j, w_i)\right)$. Then the general form of (5.5) is the first-order initial-value problem

$$\left.\begin{aligned} \mathbf{M}\frac{d\underline{\alpha}}{dt}(t) + \mathbf{R}\underline{\alpha}(t) &= \underline{\chi}(t) \text{ for } t \in (0,T) \\ \underline{\alpha}(0) &= \underline{\beta}. \end{aligned}\right\} \tag{5.9}$$

The θ-method for the system (5.9) is

$$\frac{1}{\Delta t}\mathbf{M}(\underline{\alpha}_{n+1}-\underline{\alpha}_n)+\mathbf{R}(\theta\underline{\alpha}_{n+1}+(1-\theta)\underline{\alpha}_n)=\theta\underline{\chi}(t_{n+1})+(1-\theta)\underline{\chi}(t_n)\text{ for } n=0,1,\mathrm{K}\,,N-1$$
$$\underline{\alpha}_0=\underline{\beta}.$$

Despite the simple structure of (5.5), it is a stiff system, as is the general system (5.9). The stiffness of (5.9) means the following (Dautray & Lions, 2000c: p. 68): Some of the characteristic values of the system (5.9), i.e. eigenvalues λ satisfying

$$\lambda\mathbf{M}\underline{\alpha}(t)=\mathbf{R}\underline{\alpha}(t), \tag{5.10}$$

have small or bounded modulus as $m\to\infty$, i.e. the dimension of the system tends to ∞ (which is necessary to obtain convergence in space), while others have very large modulus, which tends to ∞ with *m*.

Theorem 5.1.2. For $\frac{1}{2}\le\theta\le 1$, the spectral-θ-method (5.7) is stable in the sense of Lax-Richtmyer, i.e.

$$\left\|u_{m,n}\right\|_0\le K\left\|u_{m,0}\right\|_0 \text{ for } n\Delta t\le T \tag{5.11}$$

for some constant $K\ge 0$.

Proof. We use $\mathbf{diag}(a_k)_{k=-m}^{m}$ to denote the diagonal matrix with entries $(a_k)_{k=-m}^{m}$, and define

$$\underline{\alpha}_{m,n}:=\begin{bmatrix}\alpha_{-m,n}\\ \alpha_{-m+1,n}\\ \mathrm{M}\\ \alpha_{m,n}\end{bmatrix}. \tag{5.12}$$

Then (5.7) (with $f = 0$) is equivalent to

$$\underline{\alpha}_{m,n+1}-\underline{\alpha}_{m,n}=\mathbf{diag}(-\Delta t\lambda_k)_{k=-m}^{m}\left(\theta\underline{\alpha}_{m,n+1}+(1-\theta)\underline{\alpha}_{m,n}\right)$$
$$\mathbf{diag}(1+\Delta t\theta\lambda_k)_{k=-m}^{m}\underline{\alpha}_{m,n+1}=\mathbf{diag}(1-\Delta t(1-\theta)\lambda_k)_{k=-m}^{m}\underline{\alpha}_{m,n}$$
$$\underline{\alpha}_{m,n+1}=\mathbf{diag}\left(\frac{1-\Delta t(1-\theta)\lambda_k}{1+\Delta t\theta\lambda_k}\right)_{k=-m}^{m}\underline{\alpha}_{m,n}.$$

Using $\|\cdot\|_{\mathrm{E}}$ to denote the Euclidean norm, we have

$$\left\|\underline{\alpha}_{m,n+1}\right\|_{\mathrm{E}}=\left\|\mathbf{diag}\left(\frac{1-\Delta t(1-\theta)\lambda_k}{1+\Delta t\theta\lambda_k}\right)_{k=-m}^{m}\underline{\alpha}_{m,n}\right\|_2\le\left\|\underline{\alpha}_{m,n}\right\|_{\mathrm{E}} \tag{5.13}$$

since $\frac{1-\Delta t(1-\theta)\lambda_k}{1+\Delta t\theta\lambda_k}\le 1$ for $\frac{1}{2}\le\theta\le 1$ (cf. Remark 4.2.3). Hence $\left\|\underline{\alpha}_{m,n}\right\|_{\mathrm{E}}\le\left\|\underline{\alpha}_{m,0}\right\|_{\mathrm{E}}$. Thus

$$\begin{aligned}\left\|u_{m,n}\right\|_0^2 &= \sum_{k=-m}^{m}\left|\alpha_{k,n}\right|^2 \quad (\text{Parseval identity})\\ &\le \sum_{k=-m}^{m}\left|\alpha_{k,0}\right|^2\\ &= \left\|u_{m,0}\right\|_0^2 \qquad (\text{Parseval identity}).\end{aligned}$$

■

Remark 5.1.3. We know (cf. Theorem 2.4) that the problem (5.1) is well-posed (see Dautray & Lions (2000c: p. 36)). Therefore, in view of the Lax equivalence theorem (Dautray & Lions, 2000c: p. 37), the spectral-θ-method would be convergent for $\frac{1}{2}\le\theta\le 1$, in the sense that

$$\lim_{\substack{m\to\infty\\ \Delta t\to 0\\ n\Delta t=t}}\left\|u(t)-u_{m,n}\right\|_0 = 0 \text{ for all } t\in(0,T) \tag{5.14}$$

under the assumption that the method is consistent.

We now present a result on the order of convergence, following, to some extent, Raviart & Thomas (1983: pp. 177-178):

Theorem 5.1.4. For $m\in\mathbf{N}$, the solution $\{u_{m,n}\}_{n=0}^{N}$ in $\mathbf{S}_m$ of the scheme (5.7) satisfies the following estimates:

(a) If $\frac{1}{2}<\theta\le 1$ and $u\in\mathbf{C}^1([0,T],\mathbf{H}^1_{\text{per}}(0,2\pi))\cap\mathbf{C}^2([0,T],\mathbf{L}^2(0,2\pi))$, there exists for any $h_0>0$ a constant $C>0$ such that for all $\Delta t\le h_0$, we have

$$\left\|u_{m,n}-u(t_n)\right\|_0 \le \frac{1}{m}\left\|\frac{du}{dx}(t_n)\right\|_0 + C\left(\int_0^{t_n}\left\|\frac{du_m}{dt}(s)-\frac{du}{dt}(s)\right\|_0 ds + \Delta t\int_0^{t_n}\left\|\frac{d^2u}{dt^2}(s)\right\|_0 ds\right). \tag{5.15}$$

(b) If $\theta=\frac{1}{2}$ and $u\in\mathbf{C}^1([0,T],\mathbf{H}^1_{\text{per}}(0,2\pi))\cap\mathbf{C}^3([0,T],\mathbf{L}^2(0,2\pi))$, then

$$\left\|u_{m,n}-u(t_n)\right\|_0 \le \frac{1}{m}\left\|\frac{du}{dx}(t_n)\right\|_0 + \int_0^{t_n}\left\|\frac{du_m}{dt}(s)-\frac{du}{dt}(s)\right\|_0 ds + C(\Delta t)^2\int_0^{t_n}\left\|\frac{d^3u}{dt^3}(s)\right\|_0 ds, \tag{5.16}$$

where C is independent of m, Δt and u.

(c) If $0 \le \theta < \frac{1}{2}$ and $u \in \mathbf{C}^1\left([0,T],\mathbf{H}^1_{\text{per}}(0,2\pi)\right) \cap \mathbf{C}^2\left([0,T],\mathbf{L}^2(0,2\pi)\right)$, we have, under the stability condition $\lambda_m \Delta t \le \frac{2}{1-2\theta}$ (cf. (4.27)), the formula

$$\left\| u_{m,n} - u(t_n) \right\|_0 \le \frac{1}{m} \left\| \frac{du}{dx}(t_n) \right\|_0 + \int_0^{t_n} \left\| \frac{du_m}{dt}(s) - \frac{du}{dt}(s) \right\|_0 ds + C\Delta t \int_0^{t_n} \left\| \frac{d^2 u}{dt^2}(s) \right\|_0 ds, \qquad (5.17)$$

where C is independent of m, Δt and u.

Proof. The proof is given in full by Raviart & Thomas (1983: p. 178). We shall restrict ourselves to a straightforward proof for the case $\theta = 1$, following Thomée (1997: Theorem 1.5). Consider

$$e_{m,n} := u_{m,n} - u_m(t_n), \qquad (5.18)$$

the error between the fully discrete approximation and the semi-discrete spectral approximation at $t = t_n$. Adding and subtracting terms in (5.7) and rearrangement yield, for all $v \in \mathbf{S}_m$,

$$\left\langle \frac{e_{m,n+1} - e_{m,n}}{\Delta t}, v \right\rangle_0 + a(e_{m,n+1}, v) = \left\langle f_{m,n+1}, v \right\rangle_0 - \left\langle \frac{u_m(t_{n+1}) - u_m(t_n)}{\Delta t}, v \right\rangle_0 - a(u_m(t_{n+1}), v). \qquad (5.19)$$

Using (2.18), (5.19) becomes

$$\left\langle \frac{e_{m,n+1} - e_{m,n}}{\Delta t}, v \right\rangle_0 + a(e_{m,n+1}, v) = \left\langle \frac{du}{dt}(t_{n+1}) - \frac{u_m(t_{n+1}) - u_m(t_n)}{\Delta t}, v \right\rangle_0 \text{ for all } v \in \mathbf{S}_m. \qquad (5.20)$$

Substituting $v = e_{m,n+1}$ into (5.20) and observing that $a(e_{m,n+1}, e_{m,n+1}) \ge 0$, we obtain

$$\begin{aligned}
\left\langle \frac{e_{m,n+1} - e_{m,n}}{\Delta t}, e_{m,n+1} \right\rangle_0 &\le \left\| \frac{u_m(t_{n+1}) - u_m(t_n)}{\Delta t} - \frac{du}{dt}(t_{n+1}) \right\|_0 \left\| e_{m,n+1} \right\|_0 && \begin{pmatrix} \text{Cauchy - Schwarz} \\ \text{inequality} \end{pmatrix} \\
\frac{1}{\Delta t} \left\| e_{m,n+1} \right\|_0^2 &\le \left\| \frac{u_m(t_{n+1}) - u_m(t_n)}{\Delta t} - \frac{du}{dt}(t_{n+1}) \right\|_0 \left\| e_{m,n+1} \right\|_0 + \frac{1}{\Delta t} \left\langle e_{m,n}, e_{m,n+1} \right\rangle_0 \\
\left\| e_{m,n+1} \right\|_0^2 &\le \left(\Delta t \left\| \frac{u_m(t_{n+1}) - u_m(t_n)}{\Delta t} - \frac{du}{dt}(t_{n+1}) \right\|_0 + \left\| e_{m,n} \right\|_0 \right) \left\| e_{m,n+1} \right\|_0 && \begin{pmatrix} \text{Cauchy - Schwarz} \\ \text{inequality} \end{pmatrix},
\end{aligned}$$

so that

$$\left\| e_{m,n+1} \right\|_0 \le \Delta t \left\| \frac{u_m(t_{n+1}) - u_m(t_n)}{\Delta t} - \frac{du}{dt}(t_{n+1}) \right\|_0 + \left\| e_{m,n} \right\|_0. \qquad (5.21)$$

Repeated application of (5.21) leads to

$$\begin{aligned}
\left\| e_{m,n} \right\|_0 &\le \Delta t \sum_{k=0}^{n-1} \left\| \frac{u_m(t_{k+1}) - u_m(t_k)}{\Delta t} - \frac{du}{dt}(t_{k+1}) \right\|_0 \\
&\le \Delta t \sum_{k=0}^{n-1} \left\| \frac{u_m(t_{k+1}) - u_m(t_k)}{\Delta t} - \frac{u(t_{k+1}) - u(t_k)}{\Delta t} \right\|_0 \\
&\quad + \Delta t \sum_{k=0}^{n-1} \left\| \frac{u(t_{k+1}) - u(t_k)}{\Delta t} - \frac{du}{dt}(t_{k+1}) \right\|_0
\end{aligned} \Bigg\} \text{ (Triangle inequality)}$$

$$\begin{aligned}\left\|e_{m,n}\right\|_0 &\le \sum_{k=0}^{n-1}\left\|\int_{t_k}^{t_{k+1}}\left(\frac{du}{dt}(s)-\frac{du_m}{dt}(s)\right)ds\right\|_0+\sum_{k=0}^{n-1}\left\|\int_{t_k}^{t_{k+1}}(s-t_k)\frac{d^2u}{dt^2}(s)ds\right\|_0 \quad \begin{pmatrix}\text{Taylor formula with integral}\\ \text{form of the remainder}\end{pmatrix}\\ &\le \sum_{k=0}^{n-1}\int_{t_k}^{t_{k+1}}\left\|\frac{du}{dt}(s)-\frac{du_m}{dt}(s)\right\|_0 ds+\Delta t\sum_{k=0}^{n-1}\int_{t_k}^{t_{k+1}}\left\|\frac{d^2u}{dt^2}(s)\right\|_0 ds\\ &= \int_0^{t_n}\left\|\frac{du}{dt}(s)-\frac{du_m}{dt}(s)\right\|_0 ds+\Delta t\int_0^{t_n}\left\|\frac{d^2u}{dt^2}(s)\right\|_0 ds.\end{aligned}$$

Then we have

$$\begin{aligned}\left\|u_{m,n}-u(t_n)\right\|_0 &\le \left\|u_m(t_n)-u(t_n)\right\|_0+\left\|e_{m,n}\right\|_0 \qquad (\text{Triangle inequality})\\ &\le \frac{1}{m}\left\|\frac{du}{dx}(t_n)\right\|_0+\int_0^{t_n}\left\|\frac{du_m}{dt}(s)-\frac{du}{dt}(s)\right\|_0 ds+\Delta t\int_0^{t_n}\left\|\frac{d^2u}{dt^2}(s)\right\|_0 ds,\end{aligned}$$

due to (3.10) and our calculations here.

■

The error estimates can be improved upon by using any regularity properties of *u*.

Corollary 5.1.5. Suppose that

$$u(t)\in \mathbf{H}^2(0,2\pi) \text{ and } \frac{du}{dt}(t)\in \mathbf{H}^1(0,2\pi) \tag{5.22}$$

(cf. (2.50)-(2.51)). For $m \in \mathbf{N}$, the solution $\{u_{m,n}\}_{n=0}^N$ in $\mathbf{S}_m$ of the scheme (5.7) satisfies the following estimates:

(a) If $\frac{1}{2}<\theta\le 1$ and $u\in \mathbf{C}^1\left([0,T],\mathbf{H}^1_{\text{per}}(0,2\pi)\right)\cap \mathbf{C}^2\left([0,T],\mathbf{L}^2(0,2\pi)\right)$, there exists for any $h_0>0$ a constant $C > 0$ such that for all $\Delta t \le h_0$, we have

$$\left\|u_{m,n}-u(t_n)\right\|_0 \le \frac{1}{m^2}\left\|\frac{d^2u}{dx^2}(t)\right\|_0+C\left(\frac{1}{m}\int_0^{t_n}\left\|\frac{\partial^2 u}{\partial x\partial t}(s)\right\|_0 ds+\Delta t\int_0^{t_n}\left\|\frac{d^2u}{dt^2}(s)\right\|_0 ds\right). \tag{5.23}$$

(b) If $\theta=\frac{1}{2}$ and $u\in \mathbf{C}^1\left([0,T],\mathbf{H}^1_{\text{per}}(0,2\pi)\right)\cap \mathbf{C}^3\left([0,T],\mathbf{L}^2(0,2\pi)\right)$, then

$$\left\|u_{m,n}-u(t_n)\right\|_0 \le \frac{1}{m^2}\left\|\frac{d^2u}{dx^2}(t)\right\|_0+\frac{1}{m}\int_0^{t_n}\left\|\frac{\partial^2 u}{\partial x\partial t}(s)\right\|_0 ds+C(\Delta t)^2\int_0^{t_n}\left\|\frac{d^3u}{dt^3}(s)\right\|_0 ds, \tag{5.24}$$

where *C* is independent of *m*, Δt and *u*.

(c) If $0 \le \theta < \frac{1}{2}$ and $u \in \mathbf{C}^1\left([0,T],\mathbf{H}^1_{\text{per}}(0,2\pi)\right) \cap \mathbf{C}^2\left([0,T],\mathbf{L}^2(0,2\pi)\right)$, we have, under the stability condition $\lambda_m \Delta t \le \frac{2}{1-2\theta}$ (cf. (4.27)), the formula

$$\left\| u_{m,n} - u(t_n) \right\|_0 \le \frac{1}{m^2} \left\| \frac{d^2 u}{dx^2}(t) \right\|_0 + \frac{1}{m} \int_0^{t_n} \left\| \frac{\partial^2 u}{\partial x \partial t}(s) \right\|_0 ds + C\Delta t \int_0^{t_n} \left\| \frac{d^2 u}{dt^2}(s) \right\|_0 ds , \qquad (5.25)$$

where C is independent of m, Δt and u.

Proof. These results follow easily by using (3.12).

■

5.2. Spectral-non-standard θ-method

The choice of the spectral-θ-method in Section 5.1, with θ lying in the interval $\left[\frac{1}{2},1\right]$, was motivated by what we studied in Section 4.2. In that section, we proved that the θ-method, applied to an initial value problem for a first order differential equation, is stable for any $\Delta t > 0$ if and only if $\theta \in \left[\frac{1}{2},1\right]$ (cf. Remark 4.2.3).

However, in Section 4.3, we considered the non-standard θ-method, which is elementary stable (Theorem 4.3.13) as well as absolutely stable (Remark 4.3.15) for any $\Delta t > 0$ and $\theta \in [0,1]$. It is therefore natural to wonder whether the use of the non-standard θ-method could provide fully discrete spectral methods with better stability properties than the method we studied in Section 5.1.

In this section, we describe some possible spectral non-standard θ-methods, the full analysis of which will be done in further study. The procedure to design non-standard schemes for initial value-boundary value problems such as (5.1) is due to Mickens (1994: Chapter 7) and formalized by Anguelov & Lubuma (2001a).

The stationary case of the problem (5.1) is the boundary value problem

$$\left.\begin{aligned} -c\frac{d^2u}{dx^2} + bu = f \text{ on } (0,2\pi) \\ u(0) = u(2\pi). \end{aligned}\right\} \tag{5.26}$$

According to the analysis done in Section 1.2 and Chapter 4, a discrete solution to (5.26) is obtained by the Fourier-Galerkin spectral method as

$$u_m(x) = \sum_{k=-m}^{m} \alpha_k w_k(x), \tag{5.27}$$

or, equivalently,

$$\left.\begin{aligned} u_m \in \mathbf{S}_m \\ a(u_m, v) = \langle f, v \rangle_0 \text{ for all } v \in \mathbf{S}_m. \end{aligned}\right\} \tag{5.28}$$

The space-independent case of (5.1) is the initial value problem

$$\left.\begin{aligned} \frac{du}{dt} + bu = f \text{ on } (0,T) \\ u(0) = u_0 \equiv \eta. \end{aligned}\right\} \tag{5.29}$$

For the decay problem

$$\left.\begin{array}{c} \dfrac{du}{dt}+bu=0 \text{ on } (0,T) \\ u(0)=\eta, \end{array}\right\} \tag{5.30}$$

the non-standard θ-method developed in Theorem 4.3.13 reads

$$\left.\begin{array}{c} \dfrac{u_{n+1}-u_n}{\dfrac{\phi(b\Delta t)}{b}}+\theta b u_{n+1}+(1-\theta)bu_n=0 \text{ for } n=0,1,\mathrm{K}\,,N-1 \\ u_0=\eta, \end{array}\right\} \tag{5.31}$$

where ϕ satisfies the conditions (4.30), (4.40). Notice that his scheme is elementary stable, even if $b < 0$. Another elementary stable non-standard scheme that could be considered for (5.30) is

$$\left.\begin{array}{c} \dfrac{u_{n+1}-u_n}{\dfrac{1-e^{-b\Delta t}}{b}}+bu_n=0 \text{ for } n=0,1,\mathrm{K}\,,N-1 \\ u_0=\eta. \end{array}\right\} \tag{5.32}$$

The scheme (5.32) is exact (cf. Mickens (1994: p. 71)) in the sense that $u(t_n)=u_n$ for u the solution of (5.30).

In view of (5.31) and (5.32), the following non-standard schemes may be considered for (5.29):

$$\left.\begin{array}{c} \dfrac{u_{n+1}-u_n}{\dfrac{\phi(b\Delta t)}{b}}+\theta b u_{n+1}+(1-\theta)bu_n=\theta f_{n+1}+(1-\theta)f_n \text{ for } n=0,1,\mathrm{K}\,,N-1 \\ u_0=\eta \end{array}\right\} \tag{5.33}$$

$$\left.\begin{array}{c} \dfrac{u_{n+1}-u_n}{\dfrac{1-e^{-b\Delta t}}{b}}+bu_n=f_n \text{ for } n=0,1,\mathrm{K}\,,N-1 \\ u_0=\eta. \end{array}\right\} \tag{5.34}$$

We may now combine (5.28) and (5.33) to obtain the following spectral-non-standard θ-method for (5.1):

$$\left.\begin{array}{c} \left\langle \dfrac{u_{m,n+1}-u_{m,n}}{\dfrac{\phi(b\Delta t)}{b}}, v \right\rangle_0 + a\left(\theta u_{m,n+1}+(1-\theta)u_{m,n}, v\right)=\left\langle \theta f_{m,n+1}+(1-\theta)f_{m,n}, v\right\rangle_0 \\ \text{for } n=0,1,\mathrm{K}\,,N-1 \text{ and } v\in \mathbf{S}_m \\ u_{m,0}=\sum_{k=-m}^{m}\langle u_0, w_k\rangle_0 w_k. \end{array}\right\} \tag{5.35}$$

Another spectral-non-standard scheme results from (5.28) and (5.34), and reads as follows:

$$\left.\begin{aligned}\left\langle \frac{u_{m,n+1}-u_{m,n}}{\frac{1-e^{-b\Delta t}}{b}}, v \right\rangle_0 + a\left(bu_{m,n}, v\right) &= \left\langle f_{m,n}, v\right\rangle_0 \\ &\text{for } n = 0,1,\mathrm{K}\ ,N-1 \text{ and } v \in \mathbf{S}_m \\ u_{m,0} &= \sum_{k=-m}^{m} \left\langle u_0, w_k \right\rangle_0 w_k .\end{aligned}\right\} \tag{5.36}$$

Notice that the non-standard schemes (5.35) and (5.36) can be defined as such even when $f = f(u)$ in (5.1).

For $m \in \mathbf{N}$ fixed, applying the non-standard θ-method (4.42) (with ϕ satisfying (4.30) and (4.40), and q defined by (4.41)) to (5.5) yields

$$\left.\begin{aligned}\frac{\alpha_{k,n+1}-\alpha_{k,n}}{\frac{\phi(\lambda_m \Delta t)}{\lambda_m}} + \lambda_k\left(\theta\alpha_{k,n+1} + (1-\theta)\alpha_{k,n}\right) &= \theta g_{k,n+1} + (1-\theta) g_{k,n} \text{ for } n = 1,2,\mathrm{K}\ , N \\ \alpha_{k,0} &= \left\langle u_0, w_k \right\rangle_0 .\end{aligned}\right\} \tag{5.37}$$

If, in (5.1), we have $f = 0$, then the method (5.37) is elementary stable (by Theorem 4.3.13). From (5.37) it follows (similar to the derivation of (5.7) from (5.6)) that

$$\left.\begin{aligned}\left\langle \frac{u_{m,n+1}-u_{m,n}}{\frac{\phi(\lambda_m \Delta t)}{\lambda_m}}, v \right\rangle_0 + a\left(\theta u_{m,n+1} + (1-\theta)u_{m,n}, v\right) &= \left\langle \theta f_{m,n+1} + (1-\theta) f_{m,n}, v\right\rangle_0 \\ &\text{for } n = 0,1,\mathrm{K}\ ,N-1 \text{ and } v \in \mathbf{S}_m \\ u_{m,0} &= \sum_{k=-m}^{m} \left\langle u_0, w_k \right\rangle_0 w_k .\end{aligned}\right\} \tag{5.38}$$

The spectral-non-standard θ-methods (5.35), (5.36) and (5.38) obtained for (5.1) are all elementary stable in the limit space-independent case (5.30). Further qualitative properties (e.g. convergence) of these schemes, as well as the study of some numerical experiments, form an integral part of our ongoing research.

References

ANGUELOV, R. & LUBUMA, J.M-S. 2000. On the non-standard finite difference method (Keynote address at the annual congress of the South African Mathematical Society, Pretoria, South Africa, 16-18 October 2000). *Notices of the South African Mathematical Society*. 31 (3): pp. 143-152.

ANGUELOV, R. & LUBUMA, J.M-S. 2001a. Contributions to the mathematics of the non-standard finite difference method and applications. *Numerical methods for partial differential equations*. 17 (5): pp. 518-543.

ANGUELOV, R. & LUBUMA, J.M-S. 2001b. Nonstandard finite difference method by nonlocal approximation. *Technical report UPWT 2001/5*. Pretoria: University of Pretoria.

BURDEN, R.L. & FAIRES, J.D. 1997. *Numerical analysis, 6th edition.* Pacific Grove: Brooks/Cole.

CANUTO, C., HUSSAINI, M.Y., QUARTERONI, A. & ZANG, T.A. 1988. *Spectral methods in fluid dynamics*. New York: Springer-Verlag. (Springer series in computational physics.)

CREESE, T.M. & HARALICK, R.M. 1978. *Differential equations for engineers*. New York: McGraw-Hill.

DAUTRAY, R. & LIONS, J-L. 2000a. *Mathematical analysis and numerical methods for science and technology. Volume 3: Spectral theory and applications.* Berlin: Springer-Verlag.

DAUTRAY, R. & LIONS, J-L. 2000b. *Mathematical analysis and numerical methods for science and technology. Volume 5: Evolution problems I.* Berlin: Springer-Verlag.

DAUTRAY, R. & LIONS, J-L. 2000c. *Mathematical analysis and numerical methods for science and technology. Volume 6: Evolution problems II.* Berlin: Springer-Verlag.

DAVIS, P.J. 1963. *Interpolation and approximation*. Massachusetts: Blaisdell. (Introductions to higher mathematics.)

DIEUDONNÉ, J. 1980. *Calcul infinitésimal, deuxième édition revue et corrigée.* Paris: Hermann. (Collection méthodes.)

GREENBERG, M.D. 1978. *Foundations of applied mathematics.* Eaglewood Cliffs: Prentice-Hall.

GUSTAFSON, K.E. 1980. *Introduction to partial differential equations and Hilbert space methods.* New York: Wiley.

KREYSZIG, E. 1978. *Introductory functional analysis with applications*. New York: Wiley.

LAMBERT, J.D. 1973. *Computational methods in ordinary differential equations.* London: Wiley. (Introductory mathematics for scientists and engineers.)

LAMBERT, J.D. 1991. *Numerical methods in ordinary differential systems*. Chichester: Wiley.

LIONS, J.L. 1961. *Equations differentielles operationnelles et problèmes aux limites.* Berlin: Springer-Verlag. (De grundlehren der mathematischen wissenshaften, band 111.)

LUBUMA, J.M-S. 1994. *Analyse numerique elementaire I: Cours du 1er cycle en sciences et sciences appliquées*. Kinshasa: University of Kinshasa.

MICKENS, R.E. 1994. *Nonstandard finite difference models of differential equations*. Singapore: World Scientific.

RAVIART, P.A. & THOMAS, J.M. 1983. *Introduction á l'analyse numérique des équations aux dérivées partielles*. Paris: Masson.

REDDY, B.D. 1998. *Introductory functional analysis with applications to boundary value problems and finite elements*. New York: Springer. (Texts in applied mathematics, no. 27.)

SIROVICH, L. 1988. *Introduction to applied mathematics*. New York: Springer-Verlag. (Texts in applied mathematics, no. 1.)

SCHWARTZ, L. 1979. *Analyse Hilbertienne*. Paris: Hermann.

STUART, A.M. & HUMPHRIES, A.R. 1998. *Dynamical systems and numerical analysis.* New York: Cambridge. (Cambridge monographs on applied and computational mathematics.)

TEMAM, R. 1979. *Navier-Stokes equations, revised edition.* Amsterdam: North-Holland. (Studies in mathematics and its applications, vol. 2.)

THOMÉE, V. 1997. *Galerkin finite element methods for parabolic problems.* Berlin: Springer. (Springer series in computational mathematics, no. 25.)

ZEIDLER, E. 1995. *Applied functional analysis: applications to mathematical physics.* New York: Springer-Verlag. (Applied mathematical sciences, no. 108.)