Springer Series in Computational Mathematics 41

Jie Shen
Tao Tang
Li-Lian Wang

Spectral Methods

Algorithms, Analysis and Applications

Springer

Springer Series in Computational Mathematics

41

For further volumes:
http://www.springer.com/series/797

Jie Shen · Tao Tang · Li-Lian Wang

Spectral Methods

Algorithms, Analysis and Applications

Springer

Jie Shen
Department of Mathematics
Purdue University
N. University St. 150
West Lafayette, IN 47907-2067
USA
shen@math.purdue.edu

Tao Tang
Department of Mathematics
Hong Kong Baptist University
Waterloo Road 224
Kowloon
Hong Kong SAR
ttang@hkbu.edu.hk

Li-Lian Wang
Division of Mathematical Sciences
School of Physical & Mathematical Sciences
Nanyang Technological University
21 Nanyang Link
637371
Singapore
lilian@ntu.edu.sg

ISSN 0179-3632
ISBN 978-3-540-71040-0 e-ISBN 978-3-540-71041-7
DOI 10.1007/978-3-540-71041-7
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011934044

Mathematics Subject Classification (2010): 65M70, 65M12, 65N15, 65N35, 65N22, 65F05, 35J25, 35J40, 35K15, 42C05

Cover design: deblik, Berlin

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This book is developed from lecture notes of graduate courses taught over the years by the authors at the Pennsylvania State University, Purdue University, Hong Kong Baptist University and Nanyang Technological University of Singapore.

The aim of the book is to provide

- A detailed presentation of basic spectral algorithms
- A systematical presentation of basic convergence theory and error analysis for spectral methods
- Some illustrative applications of spectral methods

For many basic algorithms presented in the book, we provide Matlab codes (which will be made available online) which contain additional programming details beyond the mathematical formulas, so that the readers can easily use or modify these codes to suite their need. We believe that these Matlab codes will help the readers to have a better understanding of these spectral algorithms and provide a useful starting point for developing their own application codes.

There are already quite a few monographs/books on spectral methods. The classical books by Gottlieb and Orszag (1977) and by Canuto et al. (1987)[1] were intended for researchers and advanced graduate students, and they are excellent references for the historical aspects of spectral methods as well as in depth presentations of various techniques and applications in computational fluid dynamics. The book by Boyd (2001) focused on the Fourier and Chebyshev methods with emphasis on implementations and applications. The book by Trefethen (2000) gave an excellent exposition on the spectral-collocation methods through a set of elegant Matlab routines. The books by Deville et al. (2002) and by Karniadakis and Sherwin (2005) concentrated on the spectral-element methods with details on parallel implementations and applications in fluid dynamics, while the more recent book by Hesthaven and Warburton (2008) focused on the discontinuous Galerkin methods with a nodal spectral-element approach. On the other hand, Hesthaven et al. (2007) focused on

[1] An updated and expanded version of Canuto et al. (1987) is recently published. This new version Canuto et al. (2006, 2007) incorporated many new developments made in the last 20 years and provided a more systematical treatment for spectral methods.

the spectral methods for time-dependent problems with a particular emphasis on hyperbolic equations and problems with non-smooth solutions. The book length article by Bernardi and Maday (1997) and their monograph in French Bernardi and Maday (1992a) provided an excellent exposition on the basic approximation theory of spectral methods with a particular emphasis on Stokes equations, while the monograph (Shen and Tang 2006) presented a basic introduction in a lecture note style to the implementation and analysis of spectral methods. The emphasis of the book by Guo (1998b), on the other hand, was on numerical analysis of spectral methods for nonlinear evolution problems. Finally, spectral methods have been playing a very significant role in dealing with stochastic differential equations and uncertainty quantifications, and we refer to the recent books by Le Maître and Knio (2010) and by Xiu (2010) on these emerging topics.

The current book attempts to provide a self-contained presentation for the construction, implementation and analysis of efficient spectral algorithms for some model equations, of elliptic, dispersive and parabolic type, which have wide applications in science and engineering. It strives to provide a systematical approach based on variational formulations for both algorithm development and numerical analysis. Some of the unique features of the current book are

- Our analysis is based on the non-uniformly weighted Sobolev spaces which lead to simplified analysis and more precise estimates, particularly for problems with corner singularities. We also advocate the use of the generalized Jacobi polynomials which are particularly useful for dealing with boundary value problems.
- We develop efficient spectral algorithms and present their error analysis for Volterra integral equations, higher-order differential equations, problems in unbounded domains and in high-dimensional domains. These topics have rarely been covered in detail in the existing books on spectral methods.
- We provide online a set of well structured Matlab codes which can be easily modified and expanded or rewritten in other programming languages.

The Matlab codes as well as corrections/updates to the book will be available at http://www.math.purdue.edu/~shen/STWbook. In case this site becomes unavailable due to unforeseen circumstances in the future, the readers are advised to check the Springer Web site for the updated Web link on the book.

We do not attempt to provide in this book an exhaustive account on the wide range of topics that spectral methods have had impact on. In particular, we do not include some important topics such as spectral methods for hyperbolic equations and spectral-element methods, partly because these topics do not fit well in our uniform framework, and mostly because there are already some excellent books mentioned above on these topics. As such, no attempt is made to provide a comprehensive list of references on the spectral methods. The cited references reflect the topics covered in the book, but inevitably, the authors' bias. While we strive for correctness, it is most likely that errors still exist. We welcome comments, suggestions and corrections.

The book can be used as a textbook for graduate students in both mathematics and other science/engineering. Mathematical analysis and applications are organized

mostly at the end of each chapter and presented in such a way that they can be skipped without affecting the understanding of algorithms in the following chapters. The first four chapters and Sects. 8.1–8.4 provide the basic ingredients on Fourier and polynomial approximations and essential strategies for developing efficient spectral-Galerkin and spectral-collocation algorithms. Section 8.5 deals with sparse spectral methods for high-dimensional problems. The topics in Chaps. 5, 6 and 7 are independent of each other so the readers can choose according to their need. Applications covered in Chap. 9, except for a slight dependence on Sects. 9.4–9.5, are also independent of each other. For the readers' convenience, we provide in the Appendices some essential mathematical concepts, basic iterative algorithms and commonly used time discretization schemes.

The book is also intended as a reference for active practitioners and researchers of spectral methods. The prerequisite for the book includes standard entry-level graduate courses in Numerical Analysis, Functional Analysis and Partial Differential Equations (PDEs). Some knowledge on numerical approximations of PDEs will be helpful in understanding the convergence theory and error analysis but hardly necessary for understanding the numerical algorithms presented in this book.

The authors would like to thank all the people and organizations who have provided support for this endeavor. In particular, the authors acknowledge the general support over the years by NSF and AFOSR of USA, Purdue University; Hong Kong Research Grants Council, the National Natural Science Foundation of China, Hong Kong Baptist University; Singapore Ministry of Education and Nanyang Technological University. We are grateful to Mrs. Thanh-Ha Le Thi of Springer for her support and for tolerating our multiple delays, and to Ms. Xiaodan Zhao of Nanyang Technological University for carefully checking the manuscript. Last but not the least, we would like to thanks our wives and children for their love and support.

Indiana, USA — Jie Shen
Hong Kong, China — Tao Tang
Singapore — Li-Lian Wang

Contents

Symbol List

Common Notation

$\mathbb{C}$	Set of all complex numbers
$\mathbb{R}$	Set of all real numbers
$\mathbb{Z}$	Set of all integers
$\mathbb{N}$	Set of all nonnegative integers
P_N	Set of all real polynomials of degree $\leq N$
i	Complex unit, i.e., $\mathrm{i} = \sqrt{-1}$
δ_{mn}	Kronecker Delta symbol
Γ	Gamma function defined in (A.1)
$\cong$	$z_n \cong w_n$ means that for $w_n \neq 0,\ z_n/w_n \to 1$ as $n \to \infty$
$\sim$	$z_n \sim w_n$ means that for $w_n \neq 0,\ z_n/w_n \to C$ (independent of n) as $n \to \infty$
$\lesssim$	$z_n \lesssim w_n$ means that $z_n \leq C w_n$ with C independent of n

Orthogonal Polynomials/Functions

L_n	Legendre polynomial of degree n defined in (3.168)
T_n	Chebyshev polynomial of degree n defined in (3.207)
$J_n^{\alpha,\beta}$	Jacobi polynomial of degree n with parameter (α,β) defined in (3.110)
$J_n^{k,l}$	generalized Jacobi polynomial of degree n with $k,l \in \mathbb{Z}$ defined in (6.1)
$\mathscr{L}_n$	Laguerre polynomial of degree n defined in (7.4) with $\alpha = 0$
$\widehat{\mathscr{L}_n}$	Laguerre function of degree n defined in (7.16) with $\alpha = 0$
$\mathscr{L}_n^{(\alpha)}$	generalized Laguerre polynomial of degree n with parameter α defined in (7.4)
$\widehat{\mathscr{L}_n^{(\alpha)}}$	generalized Laguerre function of degree n with parameter α defined in (7.16)
H_n	Hermite polynomial of degree n defined in (7.58)
$\widehat{H}_n$	Hermite function of degree n defined in (7.71)

Weight Functions and Weighted Spaces of Functions

ω	A generic non-negative weight function
$\omega^{\alpha,\beta}$	Jacobi weight function: $\omega^{\alpha,\beta}(x)=(1-x)^{\alpha}(1+x)^{\beta}$
ω_{α}	Weight function associated with $\mathscr{L}_n^{(\alpha)}$, i.e., $\omega_{\alpha}(x)=x^{\alpha}e^{-x}$
$\hat{\omega}_{\alpha}$	Weight function associated with $\widehat{\mathscr{L}}_n^{(\alpha)}$, i.e., $\hat{\omega}_{\alpha}(x)=x^{\alpha}$
$L^p(\Omega)$	L^p-space on Ω with $1\le p\le\infty$
$H^r(\Omega)$	Sobolev space on Ω
$H^r_{\omega}(\Omega)$	Weighted Sobolev space on Ω
$B^r_{\alpha,\beta}\big(I^d\big)$	Non-uniformly Jacobi-weighted Sobolev space defined in (3.251) ($d=1$) and in (8.125) with vector-valued $\boldsymbol{\alpha},\boldsymbol{\beta}$
$B^r_{\alpha}\big(\mathbb{R}_+\big)$	Non-uniformly weighted Sobolev space defined in (7.103)
$\hat{B}^r_{\alpha}\big(\mathbb{R}_+\big)$	Non-uniformly weighted Sobolev space defined in (7.110)
$\mathbb{K}^r_{\boldsymbol{\alpha},\boldsymbol{\beta}}(I^d)$	Jacobi-weighted Korobov-type space defined in (8.190)

Inner Products and Norms

$(\cdot,\cdot)_{\omega}$	Inner product of $L^2_{\omega}(\Omega)$
$(\cdot,\cdot)$	Inner product of $L^2(\Omega)$
$\|\cdot\|_{\omega}$	Norm of $L^2_{\omega}(\Omega)$
$\|\cdot\|_{r,\omega}$	Norm of $H^r_{\omega}(\Omega)$
$\|\cdot\|_{r,\omega}$	Semi-norm of $H^r_{\omega}(\Omega)$
$\|\cdot\|$	Norm of $L^2(\Omega)$
$\|\cdot\|_r$	Norm of $H^r(\Omega)$
$\|\cdot\|_r$	Semi-norm of $H^r(\Omega)$
$\|\cdot\|_{\infty}$	Norm of $L^{\infty}(\Omega)$
$\langle\cdot,\cdot\rangle_{N,\omega}$	Discrete inner product associated with a Gauss-type quadrature
$\langle\cdot,\cdot\rangle_N$	$\langle\cdot,\cdot\rangle_N=\langle\cdot,\cdot\rangle_{N,\omega}$ with $\omega\equiv 1$
$\|\cdot\|_{N,\omega}$	Discrete norm associated with $\langle\cdot,\cdot\rangle_{N,\omega}$

One-Dimensional Projection/Interpolation Operators

$\pi_N^{\alpha,\beta}$	$L^2_{\omega^{\alpha,\beta}}$-orthogonal projection operator defined in (3.249)
$\pi^1_{N,\alpha,\beta}$	$H^1_{\omega^{\alpha,\beta}}$-orthogonal projection operator defined in (3.269)
$\pi^{1,0}_{N,\alpha,\beta}$	$H^1_{0,\omega^{\alpha,\beta}}$-orthogonal projection operator defined in (3.290)
$I_N^{\alpha,\beta}$	Jacobi-Gauss-type interpolation operator
π_N, I_N	Operators $\pi_N^{\alpha,\beta}$, $I_N^{\alpha,\beta}$ with $\alpha=\beta=0$
π_N^c, I_N^c	Operators $\pi_N^{\alpha,\beta}, I_N^{\alpha,\beta}$ with $\alpha=\beta=-1/2$
$\Pi_{N,\alpha}$	Orthogonal projection operator in $L^2_{\omega_{\alpha}}(\mathbb{R}_+)$ defined in (7.102)
$\hat{\Pi}_{N,\alpha}$	Orthogonal projection operator in $L^2_{\hat{\omega}_{\alpha}}(\mathbb{R}_+)$ defined in (7.109)
Π_N	Orthogonal projection operator in $L^2_{\omega}(\mathbb{R})$ with $\omega=e^{-x^2}$ defined in (7.125)
$\hat{\Pi}_N$	Orthogonal projection operator defined in (7.128)
I_N^{α}, $\hat{I}_N^{\alpha}$	Laguerre-Gauss-type interpolation operators
I_N^h, $\hat{I}_N^h$	Hermite-Gauss interpolation operators

Chapter 1
Introduction

Numerical methods for partial differential equations can be classified into the *local* and *global* categories. The finite-difference and finite-element methods are based on local arguments, whereas the spectral method is global in character. In practice, finite-element methods are particularly well suited to problems in complex geometries, whereas spectral methods can provide superior accuracy, at the expense of domain flexibility. We emphasize that there are many numerical approaches, such as *hp* finite-elements and spectral-elements, which combine advantages of both the global and local methods. However in this book, we shall restrict our attentions to the *global* spectral methods.

Spectral methods, in the context of numerical schemes for differential equations, belong to the family of weighted residual methods (WRMs), which are traditionally regarded as the foundation of many numerical methods such as finite element, spectral, finite volume, boundary element (cf. Finlayson (1972)). WRMs represent a particular group of approximation techniques, in which the residuals (or errors) are minimized in a certain way and thereby leading to specific methods including Galerkin, Petrov-Galerkin, collocation and tau formulations.

The objective of this introductory chapter is to formulate spectral methods in a general way by using the notion of residual. Several important tools, such as *discrete transform* and *spectral differentiation*, will be introduced. These are basic ingredients for developing efficient spectral algorithms.

1.1 Weighted Residual Methods

Prior to introducing spectral methods, we first give a brief introduction to the WRM. Consider the general problem:

$$\partial_t u(x,t) - \mathscr{L}u(x,t) = \mathscr{N}(u)(x,t), \quad t > 0,\ x \in \Omega, \tag{1.1}$$

J. Shen et al., *Spectral Methods: Algorithms, Analysis and Applications*,
Springer Series in Computational Mathematics 41, DOI 10.1007/978-3-540-71041-7_1,

where $\mathscr{L}$ is a leading spatial derivative operator, and $\mathscr{N}$ is a lower-order linear or nonlinear operator involving only spatial derivatives. Here, Ω denotes a bounded domain of $\mathbb{R}^d$, $d = 1, 2$ or 3. Equation (1.1) is to be supplemented with an initial condition and suitable boundary conditions.

We shall only consider the WRM for the spatial discretization, and assume that the time derivative is discretized with a suitable time-stepping scheme. Among various time-stepping methods (cf. Appendix D), semi-implicit schemes or linearly-implicit schemes, in which the principal linear operators are treated *implicitly* to reduce the associated stability constraint, while the nonlinear terms are treated explicitly to avoid the expensive process of solving nonlinear equations at each time step, are most frequently used in the context of spectral methods.

Let τ be the time step size, and $u^k(\cdot)$ be an approximation of $u(\cdot, k\tau)$. As an example, we consider the Crank-Nicolson leap-frog scheme for (1.1):

$$\frac{u^{n+1} - u^{n-1}}{2\tau} - \mathscr{L}\Big(\frac{u^{n+1} + u^{n-1}}{2}\Big) = \mathscr{N}(u^n), \quad n \geq 1. \tag{1.2}$$

We can rewrite (1.2) as

$$\mathbf{L}u(x) := \alpha u(x) - \mathscr{L}u(x) = f(x), \quad x \in \Omega, \tag{1.3}$$

where, with a slight abuse of notation, $u = \frac{u^{n+1}+u^{n-1}}{2}$, $\alpha = \tau^{-1}$ and $f = \alpha u^{n-1} + \mathscr{N}(u^n)$. Hence, at each time step, we need to solve a steady-state problem of the form (1.3).

At this point, it is important to emphasize that the construction of efficient numerical solvers for some important equations in the form of (1.3), such as Poisson-type equations and advection-diffusion equations, is an essential step in solving general nonlinear PDEs. With this in mind, a particular emphasis of this book is to design and analyze efficient spectral algorithms for equations of the form (1.3) where $\mathscr{L}$ is a *linear elliptic* operator.

The starting point of the WRM is to approximate the solution u of (1.3) by a finite sum

$$u(x) \approx u_N(x) = \sum_{k=0}^{N} a_k \phi_k(x), \tag{1.4}$$

where $\{\phi_k\}$ are the *trial* (*or basis*) *functions*, and the expansion coefficients $\{a_k\}$ are to be determined. Substituting u_N for u in (1.3) leads to the *residual*:

$$\mathbf{R}_N(x) = \mathbf{L}u_N(x) - f(x) \neq 0, \quad x \in \Omega. \tag{1.5}$$

The notion of the WRM is to force the residual to zero by requiring

$$(\mathbf{R}_N, \psi_j)_\omega := \int_\Omega \mathbf{R}_N(x)\psi_j(x)\omega(x)dx = 0, \quad 0 \leq j \leq N, \tag{1.6}$$

where $\{\psi_j\}$ are the *test functions*, and ω is a positive weight function; or

$$\langle \mathbf{R}_N, \psi_j \rangle_{N,\omega} := \sum_{k=0}^{N} \mathbf{R}_N(x_k)\psi_j(x_k)\omega_k = 0, \quad 0 \le j \le N, \tag{1.7}$$

where $\{x_k\}_{k=0}^N$ are a set of preselected collocation points, and $\{\omega_k\}_{k=0}^N$ are the weights of a numerical quadrature formula.

The choice of trial/test functions is one of the main features that distinguishes spectral methods from finite-element and finite-difference methods. In the latter two methods, the trial/test functions are local in character with finite regularities. In contrast, spectral methods employ globally smooth functions as trial/test functions. The most commonly used trial/test functions are trigonometric functions or orthogonal polynomials (typically, the eigenfunctions of singular Sturm-Liouville problems), which include

- $\phi_k(x) = e^{ikx}$ (Fourier spectral method)
- $\phi_k(x) = T_k(x)$ (Chebyshev spectral method)
- $\phi_k(x) = L_k(x)$ (Legendre spectral method)
- $\phi_k(x) = \mathscr{L}_k(x)$ (Laguerre spectral method)
- $\phi_k(x) = H_k(x)$ (Hermite spectral method)

Here, $T_k, L_k, \mathscr{L}_k$ and H_k are the Chebyshev, Legendre, Laguerre and Hermite polynomials of degree k, respectively.

The choice of test functions distinguishes the following formulations:

- *Galerkin.* The test functions are the same as the trial ones (i.e., $\phi_k = \psi_k$ in (1.6) or (1.7)), assuming the boundary conditions are periodic or homogeneous.
- *Petrov-Galerkin.* The test functions are different from the trial ones.
- *Collocation.* The test functions $\{\psi_k\}$ in (1.7) are the Lagrange basis polynomials such that $\psi_k(x_j) = \delta_{jk}$, where $\{x_j\}$ are preassigned collocation points. Hence, the residual is forced to zero at $\{x_j\}$, i.e., $\mathbf{R}_N(x_j) = 0$.

Remark 1.1. *In the literature, the term of pseudo-spectral method is often used to describe any spectral method where some operations involve a collocation approach or a numerical quadrature which produces aliasing errors* (cf. Gottlieb and Orszag (1977)). *In this sense, almost all practical spectral methods are pseudo-spectral. In this book, we shall not classify a method as pseudo-spectral or spectral. Instead, it will be classified as Galerkin type or collocation type.*

Remark 1.2. *The so-called* tau *method is a particular class of Petrov-Galerkin method. While the* tau *method offers some advantages in certain situations, for most problems, it is usually better to use a well-designed Galerkin or Petrov-Galerkin method. So in this book, we shall not touch on this topic, and refer to* El-Daou and Ortiz (1998), Canuto et al. (2006) *and the references therein for a thorough discussion of this approach.*

In the forthcoming sections, we shall demonstrate how to construct spectral methods for solving differential equations by examining several spectral schemes based on Galerkin, Petrov-Galerkin and collocation formulations in a general manner. We shall revisit these illustrative examples in a more rigorous fashion in the main body of the book.

1.2 Spectral-Collocation Method

To fix the idea, we consider the following linear problem:

$$\begin{aligned} &\mathbf{L}u(x) = -u''(x) + p(x)u'(x) + q(x)u(x) = f(x), \quad x \in (-1,1), \\ &B_{\pm}u(\pm 1) = g_{\pm}, \end{aligned} \tag{1.8}$$

where $B_{\pm}$ are linear operators corresponding to Dirichlet, Neumann or Robin boundary conditions (see Sect. 4.1), and the data p, q, f and $g_{\pm}$ are given such that the above problem is well-posed.

As mentioned earlier, the collocation method forces the residual to vanish pointwisely at a set of preassigned points. More precisely, let $\{x_j\}_{j=0}^{N}$ (with $x_0 = -1$ and $x_N = 1$) be a set of Gauss-Lobatto points (see Chap. 3), and let P_N be the set of all real algebraic polynomials of degree $\leq N$. The spectral-collocation method for (1.8) amounts to finding $u_N \in P_N$ such that (a) the residual $\mathbf{R}_N(x) = \mathbf{L}u_N(x) - f(x)$ equals to zero at the interior collocation points, namely,

$$\mathbf{R}_N(x_k) = \mathbf{L}u_N(x_k) - f(x_k) = 0, \quad 1 \leq k \leq N-1, \tag{1.9}$$

(b) u_N satisfies exactly the boundary conditions, i.e.,

$$B_{-}u_N(x_0) = g_{-}, \quad B_{+}u_N(x_N) = g_{+}. \tag{1.10}$$

The spectral-collocation method is usually implemented in the physical space by seeking approximate solution in the form

$$u_N(x) = \sum_{j=0}^{N} u_N(x_j) h_j(x), \tag{1.11}$$

where $\{h_j\}$ are the Lagrange basis polynomials (also referred to as *nodal* basis functions), i.e., $h_j \in P_N$ and $h_j(x_k) = \delta_{kj}$. Hence, inserting (1.11) into (1.9)-(1.10) leads to the linear system

$$\begin{aligned} &\sum_{j=0}^{N} \big[\mathbf{L}h_j(x_k)\big] u_N(x_j) = f(x_k), \quad 1 \leq k \leq N-1, \\ &\sum_{j=0}^{N} \big[B_{-}h_j(x_0)\big] u_N(x_j) = g_{-}, \quad \sum_{j=0}^{N} \big[B_{+}h_j(x_N)\big] u_N(x_j) = g_{+}. \end{aligned} \tag{1.12}$$

The above system contains $N+1$ equations and $N+1$ unknowns, so we can rewrite it in a matrix form. To fix the idea, we consider (1.8) with Dirichlet boundary conditions: $u(\pm 1) = g_{\pm}$. In this case, setting $u_N(x_0) = g_{-}$ and $u_N(x_N) = g_{+}$ in the first equation of (1.12), we find that the system (1.12) reduces to

$$\sum_{j=1}^{N-1} \big[\mathbf{L}h_j(x_k)\big]u_N(x_j) = f(x_k) - \big\{\big[\mathbf{L}h_0(x_k)\big]g_- + \big[\mathbf{L}h_N(x_k)\big]g_+\big\}, \tag{1.13}$$

for $1 \le k \le N-1$. Differentiating (1.11) m times leads to

$$u_N^{(m)}(x_k) = \sum_{j=0}^{N} d_{kj}^{(m)} u_N(x_j) \ \text{ where } \ d_{kj}^{(m)} = h_j^{(m)}(x_k). \tag{1.14}$$

The matrix $D^{(m)} = \big(d_{kj}^{(m)}\big)_{k,j=0,\dots,N}$ is called the differentiation matrix of order m relative to $\{x_j\}_{j=0}^N$. If we denote by $\mathbf{u}^{(m)}$ the vector whose components are the values of $u_N^{(m)}$ at the collocation points, it follows from (1.14) that

$$\mathbf{u}^{(m)} = D^{(m)}\mathbf{u}^{(0)}, \quad m \ge 1. \tag{1.15}$$

Hence, we have

$$\mathbf{L}h_j(x_k) = -d_{kj}^{(2)} + p(x_k)d_{kj}^{(1)} + q(x_k)\delta_{kj}. \tag{1.16}$$

Denote by $\mathbf{f}$ the vector with $N-1$ components given by the right-hand side of (1.13). Setting

$$\begin{aligned} &\widetilde{D}_m = \big(d_{kj}^{(m)}\big)_{k,j=1,\dots,N-1}, \quad m = 1,2, \\ &P = \operatorname{diag}\big(p(x_1),\dots,p(x_{N-1})\big), \quad Q = \operatorname{diag}\big(q(x_1),\dots,q(x_{N-1})\big), \end{aligned} \tag{1.17}$$

the system (1.13) reduces to

$$\big(-\widetilde{D}_2 + P\widetilde{D}_1 + Q\big)\mathbf{u}^{(0)} = \mathbf{f}. \tag{1.18}$$

Observe that the collocation method is easy to implement, once the differentiation matrices are precomputed. Moreover, it is very convenient for solving problems with variable coefficients and/or nonlinear problems, since we work in the physical space and derivatives can be evaluated by (1.14) directly. As a result, the collocation method has been extensively used in practice. However, three important issues should be considered in the implementation and analysis of a collocation method:

- The coefficient matrix of the collocation system is always full with a condition number behaving like $O(N^{2m})$ (m is the order of the differential equation).
- The choice of collocation points is crucial in terms of stability, accuracy and ease of dealing with boundary conditions. In general, they are chosen as nodes (typically, zeros of orthogonal polynomials) of Gauss-type quadrature formulas.
- The aforementioned collocation scheme is formulated in a *strong* form. In terms of error analysis, it is more convenient to reformulate it as a (but not always equivalent) *weak* form, see Sect. 1.3.3 and Chap. 4.

1.3 Spectral Methods of Galerkin Type

The collocation method described in the previous section is implemented in the physical space. In this section, we shall describe Galerkin-type spectral methods in the frequency space, and present the basic principles of the spectral-Galerkin method, spectral-Petrov-Galerkin method, and spectral-Galerkin method with numerical integration.

1.3.1 Galerkin Method

Without loss of generality, we consider (1.8) with $g_{\pm} = 0$. The non-homogeneous boundary conditions can be easily handled by considering $v = u - \tilde{u}$, where $\tilde{u}$ is a "simple" function satisfying the non-homogeneous boundary conditions (cf. Chap. 4).

Define the finite-dimensional approximation space:

$$X_N = \left\{ \phi \in P_N \ : \ B_{\pm}\phi(\pm 1) = 0 \right\} \ \Rightarrow \ \dim(X_N) = N - 1.$$

Let $\{\phi_k\}_{k=0}^{N-2}$ be a set of basis functions of X_N. We expand the approximate solution as

$$u_N(x) = \sum_{k=0}^{N-2} \hat{u}_k \phi_k(x) \in X_N. \tag{1.19}$$

Then, the expansion coefficients $\{\hat{u}_k\}_{k=0}^{N-2}$ can be determined by the residual equation (1.6) with $\{\psi_j = \phi_j\}$:

$$\int_{-1}^{1} \left(\mathbf{L}u_N(x) - f(x) \right) \phi_j(x) \omega(x) dx = 0, \quad 0 \le j \le N-2, \tag{1.20}$$

which is equivalent to

$$\begin{cases} \text{Find } u_N \in X_N \text{ such that} \\ \left(\mathbf{L}u_N, v_N \right)_{\omega} = \left(f, v_N \right)_{\omega}, \quad \forall v_N \in X_N. \end{cases} \tag{1.21}$$

Here, $(\cdot,\cdot)_{\omega}$ is the inner product of $L^2_{\omega}(-1,1)$ (cf. Appendix B).

The linear system of the above scheme is obtained by substituting (1.19) into (1.20). More precisely, setting

$$\begin{aligned} &\mathbf{u} = \left(\hat{u}_0, \hat{u}_1, \ldots, \hat{u}_{N-2} \right)^T; \quad f_j = (f, \phi_j)_{\omega}, \quad \mathbf{f} = \left(f_0, f_1, \ldots, f_{N-2} \right)^T; \\ &s_{jk} = \left(\mathbf{L}\phi_k, \phi_j \right)_{\omega}, \quad S = (s_{jk})_{j,k=0,\ldots,N-2}, \end{aligned}$$

the system (1.20) reduces to

$$S\mathbf{u} = \mathbf{f}. \tag{1.22}$$

Therefore, it is crucial to choose basis functions $\{\phi_j\}$ such that:

- The right-hand side $(f,\phi_j)_\omega$ can be computed efficiently.
- The linear system (1.22) can be solved efficiently.

The key idea is to use *compact combinations* of orthogonal polynomials or orthogonal functions to construct basis functions. To demonstrate the basic principle, we consider the Legendre spectral approximation (i.e., $\omega \equiv 1$ in (1.20)-(1.22)). Let $L_k(x)$ be the Legendre polynomial of degree k, and set

$$\phi_k(x) = L_k(x) + \alpha_k L_{k+1}(x) + \beta_k L_{k+2}(x), \quad k \geq 0, \tag{1.23}$$

where the constants α_k and β_k are uniquely determined by the boundary conditions: $B_\pm \phi_k(\pm 1) = 0$ (cf. Sect. 4.1). We shall refer to such basis functions as *modal* basis functions. Therefore, we have

$$X_N = \text{span}\{\phi_0, \phi_1, \ldots, \phi_{N-2}\}. \tag{1.24}$$

Using the properties of Legendre polynomials (cf. Sect. 3.3), one verifies easily that, *if $p(x)$ and $q(x)$ are constants*, the coefficient matrix S is *sparse* so the linear system (1.22) can be solved efficiently. However, for more general $p(x)$ and $q(x)$, the coefficient matrix S is full and one needs to resort to an iterative method (cf. Sect. 4.4).

In the above, we just considered the Legendre case. In fact, the construction of such a basis is also feasible for the Chebyshev, Laguerre and Hermite cases (see Chaps. 4–7). The notion of using compact combinations of orthogonal polynomials/functions to develop efficient spectral solvers will be repeatedly emphasized in this book.

We now consider the evaluation of $(f,\phi_j)_\omega$. In general, this term can not be computed exactly and is usually approximated by $(I_N f,\phi_j)_\omega$, where I_N is an interpolation operator upon P_N relative to the Gauss-Lobatto points. Thus, we can write

$$(I_N f)(x) = \sum_{k=0}^{N} \tilde{f}_k \varphi_k(x), \tag{1.25}$$

where $\{\varphi_k\}$ is an orthonormal polynomial basis of P_N (orthogonal with respect to ω, i.e., $(\varphi_k,\varphi_j)_\omega = \delta_{jk}$). Thanks to the orthogonality, the *discrete transforms* between the physical values $\{f(x_j)\}_{j=0}^N$ and the expansion coefficients $\{\tilde{f}_k\}_{k=0}^N$ can be computed efficiently. In particular, the computational complexity of the Fourier and Chebyshev discrete transforms can be reduced to $O(N\log_2 N)$ by using the fast Fourier transform (FFT). An approach for implementing discrete transforms relative to general orthogonal polynomials is given in Sect. 3.1.5.

It is important to point out that in solving time-dependent nonlinear problems, f usually contains nonlinear terms involving derivatives of the numerical solution u_N at previous time steps (cf. (1.3)). Hence, numerical differentiations in the frequency space and/or in the physical space are required. Differentiation techniques relative to general orthogonal polynomials are addressed in Sects. 3.1.6 and 3.1.7.

1.3.2 Petrov-Galerkin Method

As pointed out in Sect. 1.1, the use of different test and trial functions distinguishes the Petrov-Galerkin method from the Galerkin method. Thanks to this flexibility, the Petrov-Galerkin method can be very useful for some non-self-adjoint problems such as odd-order equations.

As an illustrative example, we consider the following third-order equation:

$$\begin{aligned} &\mathbf{L}u(x) := u'''(x) + u(x) = f(x), \quad x \in (-1,1), \\ &u(\pm 1) = u'(1) = 0. \end{aligned} \tag{1.26}$$

As with the Galerkin case, we enforce the boundary conditions on the approximate solution. So we set

$$X_N = \left\{ \phi \in P_N : \phi(\pm 1) = \phi'(1) = 0 \right\} \;\Rightarrow\; \dim(X_N) = N-2.$$

Assuming that $\{\phi_k\}_{k=0}^{N-3}$ is a basis of X_N, we expand the approximate solution as

$$u_N(x) = \sum_{k=0}^{N-3} \hat{u}_k \phi_k(x) \in X_N.$$

The expansion coefficients $\{\hat{u}_k\}_{k=0}^{N-3}$ are determined by the residual equation (1.6) (with $\omega = 1$):

$$\int_{-1}^{1} \big(\mathbf{L}u_N(x) - f(x)\big)\, \psi_j(x) dx = 0, \quad 0 \le j \le N-3. \tag{1.27}$$

Since the leading third-order operator is not self-adjoint, it is natural to use a Petrov-Galerkin method with the test function space:

$$X_N^* = \left\{ \psi \in P_N : \psi(\pm 1) = \psi'(-1) = 0 \right\} \;\Rightarrow\; \dim(X_N^*) = N-2.$$

Assume that $\{\psi_k\}_{k=0}^{N-3}$ is a basis of X_N^*. Then, (1.27) is equivalent to the variational formulation:

$$\begin{cases} \text{Find } u_N \in X_N \text{ such that} \\ \big(\mathbf{L}u_N, v_N\big) = \big(f, v_N\big), \quad \forall v_N \in X_N^*, \end{cases} \tag{1.28}$$

where $(\cdot,\cdot)$ is the inner product of the usual L^2-space.

The theoretical aspects of the above scheme will be examined in Chap. 6. We now consider its implementation. Setting

$$\begin{aligned} &\mathbf{u} = \big(\hat{u}_0, \hat{u}_1, \ldots, \hat{u}_{N-3}\big)^T; \quad f_j = (f, \psi_j), \quad \mathbf{f} = \big(f_0, f_1, \ldots, f_{N-3}\big)^T; \\ &s_{jk} = (\phi_k', \psi_j''), \quad S = \big(s_{jk}\big)_{j,k=0,\ldots,N-3}; \\ &m_{jk} = (\phi_k, \psi_j), \quad M = \big(m_{jk}\big)_{j,k=0,\ldots,N-3}, \end{aligned}$$

the linear system (1.28) becomes

$$\big(S+M\big)\mathbf{u}=\mathbf{f}. \tag{1.29}$$

As described in the previous section, we wish to construct basis functions for X_N and X_N^*, so that the linear system (1.29) can be inverted efficiently. Once again, this goal can be achieved by using compact combinations of orthogonal polynomials. It can be checked that for $0 \le k \le N-3$,

$$\begin{aligned}\phi_k &= L_k - \frac{2k+3}{2k+5}L_{k+1} - L_{k+2} + \frac{2k+3}{2k+5}L_{k+3} \in X_N;\\ \psi_k &= L_k + \frac{2k+3}{2k+5}L_{k+1} - L_{k+2} - \frac{2k+3}{2k+5}L_{k+3} \in X_N^*,\end{aligned} \tag{1.30}$$

where L_n is the Legendre polynomial of degree n (cf. Sect. 3.3). Hence, $\{\phi_k\}_{k=0}^{N-3}$ (resp. $\{\psi_j\}_{j=0}^{N-3}$) forms a basis of X_N (resp. X_N^*). Moreover, using the properties of the Legendre polynomials, one verifies easily that the matrix M is seven-diagonal, i.e., $m_{jk}=0$ for all $|j-k|>3$. More importantly, the matrix S is diagonal.

1.3.3 Galerkin Method with Numerical Integration

We considered previously Galerkin-type methods in the frequency space, which are well suited for linear problems with constant (or polynomial) coefficients. However, their implementations are not convenient for problems with general variable coefficients. On the other hand, the collocation method is easy to implement, but it can not always be reformulated as a suitable variational formulation (most convenient for error analysis). A combination of these two approaches leads to the so-called *Galerkin method with numerical integration*, or sometimes called the *collocation method in the weak form.*

The key idea of this approach is to *replace the continuous inner products in the Galerkin formulation by the discrete ones*. As an example, we consider again (1.8) with $g_\pm = 0$. The spectral-Galerkin method with numerical integration is

$$\begin{cases}\text{Find } u_N \in X_N := \{\phi \in P_N : B_\pm\phi(\pm1)=0\} \text{ such that}\\ a_N(u_N,v_N) := \langle Lu_N, v_N\rangle_N = \langle f, v_N\rangle_N, \quad \forall v_N \in X_N,\end{cases} \tag{1.31}$$

where the discrete inner product is defined by

$$\langle u,v\rangle_N = \sum_{j=0}^{N} u(x_j)v(x_j)\omega_j,$$

with $\{x_j,\omega_j\}_{j=0}^N$ being the set of Legendre-Gauss-Lobatto quadrature nodes and weights (cf. Theorem 3.29).

For problems with variable coefficients, the above method is easier to implement, thanks to the discrete inner product, than the spectral-Galerkin method (1.21). It is also more convenient for error analysis, thanks to the weak formulation, than the spectral-collocation method (1.12).

We note that in the particular case of homogeneous Dirichlet boundary conditions, i.e., $B_{\pm}u(\pm 1) = u(\pm 1) = 0$, by taking $v_N = h_j$, $1 \le j \le N-1$ in (1.31) and using the exactness of Legendre-Gauss-Lobatto quadrature, i.e.,

$$\langle u, v\rangle_N = (u, v), \quad \forall u \cdot v \in P_{2N-1}, \tag{1.32}$$

we find that the formulation (1.31) is equivalent to the collocation formulation (1.12). However, this is not true for general boundary conditions (see Chap. 4).

1.4 Fundamental Tools for Error Analysis

In the previous sections, we briefly described several families of spatial discretization schemes using the notion of weighted residual methods. In this section, we present some fundamental apparatuses for stability and convergence analysis of numerical schemes based on weak (or variational) formulations.

We consider the linear boundary value problem (1.3):

$$\mathbf{L}u = f, \quad \text{in } \Omega; \quad Bu = 0, \quad \text{on } \partial\Omega, \tag{1.33}$$

where $\mathbf{L}$ and B are linear operators, and f is a given function on Ω.

As shown before, the starting point is to reformulate (1.33) in a *weak formulation*:

$$\begin{cases} \text{Find } u \in X \text{ such that} \\ a(u, v) = F(v), \quad \forall v \in Y, \end{cases} \tag{1.34}$$

where X is the space of trial functions, Y is the space of test functions, and F is a linear functional on Y. The expression $a(u, v)$ defines a bilinear form on $X \times Y$. It is conventional to assume that X and Y are Hilbert spaces. We refer to Appendix B for basic functional analysis settings.

Now, we consider what conditions should be placed on (1.34) to guarantee its well-posedness in the sense that:

- *Existence-uniqueness:* There exists exactly one solution of the problem.
- *Stability:* The solution must be stable which means that it depends on the data continuously. In other words, a small change of the given data produces a small change of the solution correspondingly.

The first fundamental result concerning the existence-uniqueness and stability is known as the Lax-Milgram lemma (see Theorem B.1) related to the abstract problem (1.34) with $X = Y$, i.e.,

$$\begin{cases} \text{Find } u \in X \text{ such that} \\ a(u,v) = F(v), \quad \forall v \in X. \end{cases} \tag{1.35}$$

More precisely, if the bilinear form $a(\cdot,\cdot) : X \times X \to \mathbb{R}$ satisfies

- Continuity:
$$\exists C > 0 \quad \text{such that} \quad |a(u,v)| \le C\|u\|_X \|v\|_X, \tag{1.36}$$
- Coercivity:
$$\exists \alpha > 0 \quad \text{such that} \quad a(u,u) \ge \alpha \|u\|_X^2, \tag{1.37}$$

then for any $F \in X'$ (the dual space of X as defined in Appendix B), the problem (1.35) admits a unique solution $u \in X$, satisfying

$$\|u\|_X \le \frac{1}{\alpha}\|F\|_{X'}. \tag{1.38}$$

Remark 1.3. *The constant*

$$\alpha = \inf_{0 \ne u \in X} \frac{|a(u,u)|}{\|u\|_X^2} \tag{1.39}$$

is referred to as the ellipticity constant of (1.35)*.*

The above result can only be applied to the problem (1.34) with $Y = X$. We now present a generalization of the Lax-Milgram lemma for the case $X \ne Y$ (see, e.g., Babuška and Aziz (1972)).

Theorem 1.1. *Let X and Y be two real Hilbert spaces, equipped with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$, respectively. Assume that $a(\cdot,\cdot) : X \times Y \to \mathbb{R}$ is a bilinear form and $F(\cdot) : Y \to \mathbb{R}$ is a linear continuous functional, i.e., $F \in Y'$ (the dual space of Y) satisfying*

$$\|F\|_{Y'} = \sup_{0 \ne v \in Y} \frac{|F(v)|}{\|v\|_Y} < \infty. \tag{1.40}$$

Further, assume that $a(\cdot,\cdot)$ satisfies

- *Continuity*:
$$\exists C > 0 \quad \textit{such that} \quad |a(u,v)| \le C\|u\|_X \|v\|_Y, \tag{1.41}$$
- *Inf-sup condition*:
$$\exists \beta > 0 \quad \textit{such that} \quad \sup_{0 \ne v \in Y} \frac{|a(u,v)|}{\|u\|_X \|v\|_Y} \ge \beta, \quad \forall 0 \ne u \in X, \tag{1.42}$$
- *"Transposed" inf-sup condition*:
$$\sup_{0 \ne u \in X} |a(u,v)| > 0, \quad \forall 0 \ne v \in Y. \tag{1.43}$$

Then, for any $F \in Y'$, the problem (1.34) *admits a unique solution $u \in X$, which satisfies*

$$\|u\|_X \le \frac{1}{\beta}\|F\|_{Y'}. \tag{1.44}$$

Remark 1.4. *The condition* (1.42) *is also known as the Babuška-Brezzi inf-sup condition* (cf. Babuška (1973), Brezzi (1974))*, and the real number*

$$\beta = \inf_{0\neq u\in X}\sup_{0\neq v\in Y}\frac{|a(u,v)|}{\|u\|_X\|v\|_Y} \tag{1.45}$$

is called the inf-sup constant.

Remark 1.5. *Theorem* 1.1 *with $X = Y$ is not equivalent to the Lax-Milgram lemma. In fact, one can verify readily the relation between the ellipticity and inf-sup constants: $\alpha \le \beta$. Indeed, by* (1.37)*,*

$$\alpha\|u\|_X \le \frac{|a(u,u)|}{\|u\|_X} \le \sup_{0\neq v\in X}\frac{|a(u,v)|}{\|v\|_X}, \quad \forall 0 \neq u \in X,$$

which implies

$$\alpha \le \inf_{0\neq u\in X}\sup_{0\neq v\in X}\frac{|a(u,v)|}{\|u\|_X\|v\|_X} = \beta.$$

This means that one can have $\alpha = 0$ but $\beta > 0$. In other words, the bilinear form is not coercive, but satisfies the inf-sup condition.

We review below the fundamental theory on convergence analysis of numerical approximations to (1.34).

We first consider the case $X = Y$. Assume that $X_N \subseteq X$ and

$$\forall v \in X, \quad \inf_{v_N\in X_N}\|v - v_N\|_X \to 0 \quad \text{as} \quad N \to \infty. \tag{1.46}$$

The Galerkin approximation to (1.35) is

$$\begin{cases} \text{Find } u_N \in X_N \text{ such that} \\ a(u_N, v_N) = F(v_N), \quad \forall v_N \in X_N. \end{cases} \tag{1.47}$$

The stability and convergence of this scheme can be established by using the following lemma (cf. Céa (1964)):

Theorem 1.2. (Céa Lemma). *Under the assumptions of the Lax-Milgram lemma* (*see Theorem* B.1)*, the problem* (1.47) *admits a unique solution $u_N \in X_N$ such that*

$$\|u_N\|_X \le \frac{1}{\alpha}\|F\|_{X'}. \tag{1.48}$$

Moreover, if u is the solution of (1.35)*, we have*

$$\|u - u_N\|_X \le \frac{C}{\alpha}\inf_{v_N\in X_N}\|u - v_N\|_X. \tag{1.49}$$

Here, the constants C and α are given in (1.36) *and* (1.37)*, respectively.*

Proof. Since X_N is a subspace of X, applying the Lax-Milgram lemma to (1.47) leads to the existence-uniqueness of u_N and the stability result (1.48). Now, taking $v = v_N$ in (1.35), and subtracting (1.47) from the resulting equation, we obtain the error equation

$$a(u - u_N, v_N) = 0, \quad \forall v_N \in X_N, \tag{1.50}$$

which, together with (1.36)-(1.37), implies

$$\begin{aligned} \alpha \|u - u_N\|_X^2 &\le a(u - u_N, u - u_N) = a(u - u_N, u - v_N) \\ &\le C \|u - u_N\|_X \|u - v_N\|_X, \quad \forall v_N \in X_N, \end{aligned}$$

from which (1.49) follows. □

Remark 1.6. *If, in addition, the bilinear form is symmetric, i.e., $a(u,v) = a(v,u)$, the Galerkin method is referred to as the Ritz method. In this case, the constant in the upper bound of* (1.49) *can be improved to $\sqrt{C\alpha^{-1}}$.*

Remark 1.7. *In performing error analysis of spectral methods, we usually take v_N in* (1.49) *to be a suitable orthogonal projection of u upon X_N, denoted by $\pi_N u$, which leads to*

$$\|u - u_N\|_X \le \frac{C}{\alpha} \|u - \pi_N u\|_X. \tag{1.51}$$

Hence, the error estimate follows from the approximation result on $\|u - \pi_N u\|_X$, which takes a typical form:

$$\|u - \pi_N u\|_X \le c N^{-\sigma(m)} \|u\|_{H^m}, \tag{1.52}$$

where c is a generic positive constant independent of N and any function, $\sigma(m) > 0$ is the so-called order of convergence in terms of the regularity index m, and H^m is a suitable Sobolev space with a norm involving derivatives of u up to m-th order. The establishment of such approximation results for each family of orthogonal polynomials/functions will be another emphasis of this book.

Typically, if u is sufficiently smooth, the estimate (1.52) *is valid for every m. However, for a finite-element method, the order of convergence is restricted by the order of local basis functions. The explicit dependence of the estimates of* (1.52) *type on the regularity index m will also be explored in this book.*

Observe that the bilinear form and the functional F in the discrete problem (1.47) are the same as those in the continuous problem (1.35). However, it is often convenient to use suitable approximate bilinear forms and/or functionals (see, for example, (1.31)). Hence, it is necessary to consider the following approximation to (1.35):

$$\begin{cases} \text{Find } u_N \in X_N \text{ such that} \\ a_N(u_N, v_N) = F_N(v_N), \quad \forall v_N \in X_N, \end{cases} \tag{1.53}$$

where X_N still satisfies (1.46), and $a_N(\cdot,\cdot)$ and $F_N(\cdot)$ are suitable approximations to $a(\cdot,\cdot)$ and $F(\cdot)$, respectively. In general, although X_N is a subspace of X, the

properties of the discrete bilinear form can not carry over from those of the continuous one. Hence, they have to be derived separately.

The result below, known as the first *Strang lemma* (see, e.g., Strang and Fix (1973), Ciarlet (1978)), is a generalization of Theorem 1.2.

Theorem 1.3. (First Strang lemma). *Under the assumptions of the Lax-Milgram lemma, suppose further that the discrete forms $F_N(\cdot)$ and $a_N(\cdot,\cdot)$ satisfy the same properties in the subspace $X_N \subset X$, and $\exists\, \alpha_* > 0$, independent of N, such that*

$$a_N(v,v) \geq \alpha_* \|v\|_X^2, \quad \forall v \in X_N. \tag{1.54}$$

Then, the problem (1.53) admits a unique solution $u_N \in X_N$, satisfying

$$\|u_N\|_X \leq \frac{1}{\alpha_*} \sup_{0 \neq v_N \in X_N} \frac{|F_N(v_N)|}{\|v_N\|_X}. \tag{1.55}$$

Moreover, if u is the solution of (1.35), we have

$$\begin{aligned} \|u - u_N\|_X \leq \inf_{w_N \in X_N} \Bigg\{ & \left(1 + \frac{C}{\alpha_*}\right) \|u - w_N\|_X \\ & + \frac{1}{\alpha_*} \sup_{0 \neq v_N \in X_N} \frac{|a(w_N, v_N) - a_N(w_N, v_N)|}{\|v_N\|_X} \Bigg\} \\ & + \frac{1}{\alpha_*} \sup_{0 \neq v_N \in X_N} \frac{|F(v_N) - F_N(v_N)|}{\|v_N\|_X}. \end{aligned} \tag{1.56}$$

Here, the constant C is given in (1.36).

Proof. The existence-uniqueness and stability of (1.55) follow from the Lax-Milgram lemma. The proof of (1.56) is slightly different from that of (1.49). For any $w_N \in X_N$, let $e_N = u_N - w_N$. Using (1.54), (1.35) and (1.53) leads to

$$\begin{aligned} \alpha^* \|e_N\|_X^2 \leq a_N(e_N, e_N) &= a(u - w_N, e_N) + a(w_N, e_N) \\ &\quad - a_N(w_N, e_N) + F_N(e_N) - F(e_N). \end{aligned}$$

Since the result is trivial for $e_N = 0$, we derive from (1.36) that for $e_N \neq 0$,

$$\begin{aligned} \alpha^* \|e_N\|_X &\leq C\|u - w_N\|_X + \frac{|a(w_N, e_N) - a_N(w_N, e_N)|}{\|e_N\|_X} \\ &\quad + \frac{|F(e_N) - F_N(e_N)|}{\|e_N\|_X} \\ &\leq C\|u - w_N\|_X + \sup_{0 \neq v_N \in X_N} \frac{|a(w_N, v_N) - a_N(w_N, v_N)|}{\|v_N\|_X} \\ &\quad + \sup_{0 \neq v_N \in X_N} \frac{|F(v_N) - F_N(v_N)|}{\|v_N\|_X}, \end{aligned}$$

which, together with the triangle inequality, yields

$$\|u-u_N\|_X \le \|u-w_N\|_X + \|e_N\|_X.$$

Finally, taking the infimum over $w_N \in X_N$ leads to the desired result. □

The previous discussions were restricted to approximations of the abstract problem (1.35) based on Galerkin-type formulations. Similar analysis can be done for the Petrov-Galerkin approximation of (1.34) by using Theorem 1.1. Indeed, let $X_N \subseteq X$ and $Y_N \subseteq Y$. Consider the approximation to (1.34):

$$\begin{cases} \text{Find } u_N \in X_N \text{ such that} \\ a(u_N, v_N) = F(v_N), \quad \forall v_N \in Y_N. \end{cases} \tag{1.57}$$

Unlike the coercivity property, the inf-sup property can not carry over from the whole space to the subspace. Indeed, the infimum in (1.39) will not decrease if it is taken on a subspace, whereas the supremum in the inf-sup constant (1.45), in general, becomes smaller on a subspace. Consequently, we have to prove

- Discrete inf-sup condition:

$$\exists \beta_* > 0 \quad \text{such that} \quad \sup_{0\neq v_N \in Y_N} \frac{|a(u_N, v_N)|}{\|u_N\|_X \|v_N\|_Y} \ge \beta_*, \quad \forall 0 \neq u_N \in X_N, \tag{1.58}$$

- Discrete "transposed" inf-sup condition:

$$\sup_{0\neq u_N \in X_N} |a(u_N, v_N)| > 0, \quad \forall 0 \neq v_N \in Y_N. \tag{1.59}$$

The following result, which is another generalization of Theorem 1.2, can be found in Babuška and Aziz (1972).

Theorem 1.4. *Under the assumptions of Theorem 1.1, assume further that (1.58) and (1.59) hold. Then the discrete problem (1.57) admits a unique solution $u_N \in X_N$, satisfying*

$$\|u_N\|_X \le \frac{1}{\beta_*} \|F\|_{Y'}. \tag{1.60}$$

Moreover, if u is the solution of (1.34), we have

$$\|u-u_N\|_X \le \left(1 + \frac{C}{\beta_*}\right) \inf_{v_N \in X_N} \|u - v_N\|_X, \tag{1.61}$$

where the constant C is given in (1.41).

Remark 1.8. *If we consider the following approximation to (1.34):*

$$\begin{cases} \textit{Find } u_N \in X_N \textit{ such that} \\ a_N(u_N, v_N) = F_N(v_N), \quad \forall v_N \in Y_N, \end{cases} \tag{1.62}$$

then a result similar to Theorem 1.3 can be derived, provided that (1.58) and (1.59) hold in the subspaces X_N and Y_N.