*Acta Numerica* (2011), pp. 1–119
doi:10.1017/S096249291100002X
© Cambridge University Press, 2011
Printed in the United Kingdom

# Topics in structure-preserving discretization*

Snorre H. Christiansen
*Centre of Mathematics for Applications and*
*Department of Mathematics, University of Oslo,*
*NO-0316 Oslo, Norway*
*E-mail:* snorrec@math.uio.no

Hans Z. Munthe-Kaas
*Department of Mathematics, University of Bergen,*
*N-5008 Bergen, Norway*
*E-mail:* hans.munthe-kaas@math.uib.no

Brynjulf Owren
*Department of Mathematical Sciences,*
*Norwegian University of Science and Technology,*
*NO-7491 Trondheim, Norway*
*E-mail:* brynjulf.owren@math.ntnu.no

In the last few decades the concepts of structure-preserving discretization, geometric integration and compatible discretizations have emerged as subfields in the numerical approximation of ordinary and partial differential equations. The article discusses certain selected topics within these areas; discretization techniques both in space and time are considered. Lie group integrators are discussed with particular focus on the application to partial differential equations, followed by a discussion of how time integrators can be designed to preserve first integrals in the differential equation using discrete gradients and discrete variational derivatives.

Lie group integrators depend crucially on fast and structure-preserving algorithms for computing matrix exponentials. Preservation of domain symmetries is of particular interest in the application of Lie group integrators to PDEs. The equivariance of linear operators and Fourier transforms on non-commutative groups is used to construct fast structure-preserving algorithms for computing exponentials. The theory of Weyl groups is employed in the construction of high-order spectral element discretizations, based on multivariate Chebyshev polynomials on triangles, simplexes and simplicial complexes.

The theory of mixed finite elements is developed in terms of special inverse systems of complexes of differential forms, where the inclusion of cells corresponds to pullback of forms. The theory covers, for instance, composite

* Colour online available at journals.cambridge.org/anu.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

piecewise polynomial finite elements of variable order over polyhedral grids. Under natural algebraic and metric conditions, interpolators and smoothers are constructed, which commute with the exterior derivative and whose product is uniformly stable in Lebesgue spaces. As a consequence we obtain not only eigenpair approximation for the Hodge–Laplacian in mixed form, but also variants of Sobolev injections and translation estimates adapted to variational discretizations.

## CONTENTS

1 Introduction 2
2 Integration methods based on Lie group techniques 5
3 Schemes which preserve first integrals 21
4 Spatial symmetries, high-order discretizations and fast group-theoretic algorithms 32
5 Finite element systems of differential forms 65
Appendix 108
References 112

## 1. Introduction

The solution of partial differential equations (PDEs) is a core topic of research within pure, applied and computational mathematics. Both measured in the volume of published work[1] and also in terms of its practical influence on application areas such as computational science and engineering, PDEs rank above all other mathematical subjects. Historically, the field of numerical solutions of PDEs has its roots in seminal papers by Richard Courant and his students Kurt Friedrichs and Hans Lewy, in the early twentieth century. Among the nearly 4000 academic descendants of Courant we find a large fraction of the key contributors to the field, such as Friedrichs' student Peter Lax, who received the Abel Prize in 2005 for his ground-breaking contributions to the theory and application of PDEs.

Fundamental properties of PDE discretizations, which have been recognized as crucial ever since the days of the Old Masters, include accuracy, stability, good convergence properties and the existence of efficient computational algorithms. In more recent years various aspects of structure preservation have emerged as important in addition to these fundamental properties.

[1] A count of all mathematical publications from 2001 to 2010, sorted according to the AMS Mathematics Subject Classification, reveals that number **35 PDEs** ranks highest of all primary topics with 46 138 entries; in second place is **62 Statistics**, with 39 176.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

Within the topics of ordinary differential equations and time integration, a systematic study of structure-preserving discretizations was undertaken by Feng Kang in Beijing, starting in the 1980s. During the past decade a systematic study of the preservation of various geometric structures[2] has evolved into a mature branch of numerical analysis, termed the *geometric integration* of differential equations (Hairer, Lubich and Wanner 2006, Leimkuhler and Reich 2004, Sanz-Serna and Calvo 1994). Experience has shown that the preservation of geometric properties can have a crucial influence on the quality of the simulations. In long-term simulations, structure preservation can have a dramatic effect on stability and global error growth. Examples of such structures are symplecticity, volume, symmetry, reversibility and first integrals. In short-term simulations it is frequently seen that discretization schemes designed with structure preservation in mind enjoy small errors per step, and hence become efficient for shorter time simulations too. An important class of problems is that of partial differential equations whose solutions may be subject to blow-up in finite time. We have seen that schemes which are designed to inherit certain symmetries of the continuous problem tend to perform well in capturing such finite-time singularities in the solution.

For spatial PDEs, a parallel investigation of *compatible discretizations* has been undertaken (Arnold, Bochev, Lehoucq, Nicolaides and Shashkov 2006*a*). The equations of mathematical physics (describing fluids, electromagnetic waves or elastic bodies, for instance) have been presented in geometric language, easing the construction of discretizations which preserve important geometric features, such as topology, conservation laws, symmetries and positivity structures. When well-posedness of the continuous PDE depends on the conservation or monotonicity of certain quantities, such as energy, it seems equally important for the stability of numerical schemes that they enjoy similar properties. Rather than approximately satisfying the exact conservation law, as – it could be argued – any consistent scheme would do, it seems preferable, in order to obtain stable methods, to exactly satisfy a discrete conservation law. For PDEs written in terms of grad, curl and div acting on scalar and vector fields, one is led to construct operators acting on certain finite-dimensional spaces of scalar and vector fields, forming a complex which, in spatial domains with trivial topology, is an exact sequence. More generally, various discretizations of the de Rham sequence of differential forms have been introduced, and the successful ones are related to constructs of combinatorial topology such as simplicial cochain complexes (Arnold, Falk and Winther 2010).

[2] A geometric structure is understood as a structural property which can be defined independently of particular coordinate representations of the differential equations.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

This survey paper takes a view of PDEs analogous to looking at the moon through a telescope with very high magnification. In the vast lunar landscape we will focus on a small number of craters with particularly beautiful properties, and leave the rest of the lunar surface out of our main focus.

One goal of the paper has been to tie together recent time integration techniques, in particular Lie group techniques and integral-preserving integration, and combine these with recent developments in structure-preserving spatial discretizations.

The paper consists of four main parts: Sections 2–5. Section 2 covers *integration methods based on Lie group techniques*. We provide an introduction to the theory of Lie group integrators and describe how computational algorithms can be devised, given a group action on a manifold and coordinates on the acting group. Various choices of coordinates are discussed. Another type of Lie group integrator consists of those based on compositions of flows, and we discuss some applications of these methods to the time integration of PDEs. We focus in particular on applications outside the class of exponential integrators for semilinear problems.

In Section 3 we discuss *integral-preserving schemes*. These methods apply to any PDE which has a known first integral, for instance an energy functional. We demonstrate how the method of discrete variational derivatives can be used to devise integral-preserving time integration schemes for PDEs. We consider, in particular, schemes that are linearly implicit, and therefore need the solution of only one linear system per time step. We also discuss methods which apply discrete variational derivatives, or discrete gradients, to conserve an arbitrary number of first integrals.

In Section 4 we cover *spatial symmetries*, *high-order discretizations* and *fast group-theoretic algorithms*. The topic of this section is the exploitation of spatial symmetries. A recurring theme is that of linear differential operators commuting with groups of isometries acting on the domain. We investigate high-order spectral element discretization techniques based on simplicial subdivisions (into triangles, tetrahedra and simplexes in general) using high-order multivariate Chebyshev bases on the triangles. The efficient computation of matrix (and operator) exponentials is crucial for time integrators based on Lie group actions. We survey recent work on high-order discretization techniques, fast Fourier algorithms based on group-theoretic concepts and fundamental results from representation theory.

In Section 5 we discuss *finite element systems for differential forms*. A framework for mixed finite elements is developed, general enough to allow for non-polynomial basis functions and a decomposition of space into non-canonical polytopes, but restrictive enough to yield spaces with local bases, interpolators commuting with the exterior derivative, and exact sequences where desirable. A smoothing technique gives $L^q$-stable commuting quasi-interpolation operators, from which various error estimates can be derived.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

## 2. Integration methods based on Lie group techniques

The use of Lie group techniques for obtaining solutions to differential equations dates back to the Norwegian mathematician Sophus Lie in the second half of the nineteenth century. In Lie's time these were used exclusively as analytical tools; however, more recently it has become increasingly popular to include Lie groups as an ingredient in numerical methods. The use of Lie groups in the numerical approximation of differential equations can be divided into two categories: one in which the aim is to preserve symmetries or invariance of the continuous model, and another in which Lie groups are used as building blocks for a time-stepping procedure. In this section we shall give a brief introduction to the mathematical machinery we use, and we will focus on the second category, that of using Lie groups as a fundamental component when designing numerical time integrators. We shall give a short introduction to the basics of Lie group integrators; for more details consult Iserles, Munthe-Kaas, Nørsett and Zanna (2000) and Hairer *et al.* (2006). The variety of studies related to Lie group integrators is now too large to cover in an exposition of this type. For this reason we shall focus on a selected part of the theory, and mostly consider integrators designed for nonlinear problems. Important classes of schemes that we shall *not* discuss here are methods based on the Magnus expansions and Fer expansions, usually applied to linear differential equations. This work developed in the 1990s, in large part due to Iserles and Nørsett: see, *e.g.*, Iserles and Nørsett (1999). There are several excellent sources for a summary of these methods and their analysis, for instance the *Acta Numerica* article by Iserles *et al.* (2000), the monograph by Hairer *et al.* (2006) and the more recent survey by Blanes, Casas, Oteo and Ros (2009), which also contains many applications of these integrators.

### *2.1. Background and notation*

Let $M$ be some differentiable manifold and let $\mathcal{X}(M)$ be the set of smooth vector fields on $M$. We consider every $X \in \mathcal{X}(M)$ as a differential operator on the set of smooth functions $\mathcal{F}(M)$ on $M$. Thus, in local coordinates $(x_1, \dots, x_m)$ in which $X$ has components $X_1, \dots, X_m$, we write $X = \sum_i X_i(x) \frac{\partial}{\partial x_i}$, so $X[f] = \mathrm{d}f(X)$, $f \in \mathcal{F}(M)$ is the directional derivative of $f$ along the vector field $X$. The flow of a vector field $X \in \mathcal{X}(M)$ is a one-parameter family of maps $\exp(tX) : \mathcal{D}_t \to M$, where $\mathcal{D}_t \subset M$. For any $x \in \mathcal{D}_t$ we have $\exp(tX)x = \gamma(t)$, where

$$\dot{\gamma}(t) = X|_{\gamma(t)}, \quad \gamma(0) = x, \quad t \in (a(x), b(x)),\ a(x) < 0 < b(x).$$

The domain for $\exp(tX)$ is the set $\mathcal{D}_t = \{x \in M : t \in (a(x), b(x))\}$; for further details see, *e.g.*, Warner (1983, 1.48).

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

The chain rule for the derivative of compositions of maps between manifolds is

$$(\psi \circ \phi)' = \psi' \circ \phi', \quad \phi : M \to N,\ \psi : N \to P.$$

From this, we easily obtain the useful formula

$$X[f \circ \phi](x) = \phi'(X|_x)[f](y), \quad \phi : M \to N,\ X \in \mathcal{X}(M),\ f \in \mathcal{F}(N), \quad \text{for } x \in M,\ y = \phi(x). \tag{2.1}$$

The Lie–Jacobi bracket on $\mathcal{X}(M)$ is defined simply as the commutator of vector fields $Z = [X, Y] = XY - YX$. With respect to coordinates $(x_1, \ldots, x_m)$ it has the form

$$Z_i = [X, Y]_i = \sum_{j=1}^{m} X_j \frac{\partial Y_i}{\partial x_j} - Y_j \frac{\partial X_i}{\partial x_j}.$$

This bracket makes $\mathcal{X}(M)$ a Lie algebra; the important properties of the bracket is that it is bilinear, skew-symmetric and satisfies the Jacobi identity

$$[X, [Y, Z]] + [Y, [Z, X]] + [Z, [X, Y]] = 0, \quad \forall X, Y, Z \in \mathcal{X}(M).$$

It is an easy consequence of (2.1) that

$$\phi'([X, Y]) = [\phi'(X), \phi'(Y)], \quad X, Y \in \mathcal{X}(M),\ \phi : M \to N. \tag{2.2}$$

We say that $Y$ is $\phi$-related to $X \in \mathcal{X}(M)$ if $Y|_{\phi(x)} = \phi'(X|_x)$ for each $x \in M$. Note that $Y$ is not generally a vector field on $N$: since $\phi$ does not have to be either injective or surjective, we must treat $Y$ as a pullback section of $\phi^* TN$ over $M$. In the particular case when $\phi$ is a diffeomorphism, there is a unique $Y \in \mathcal{X}(N)$ which is $\phi$-related to $X$; this vector field is then called the pushforward of $X$ with respect to $\phi$:

$$Y = \phi_* X = \phi'(X) \circ \phi^{-1}.$$

A Lie group $G$ is a differentiable manifold which is furnished with a group structure such that the multiplication is a smooth map from $G \times G$ to $G$, and the map $g \mapsto g^{-1}$, $g \in G$ is smooth as well. We briefly review some important concepts related to Lie groups. Every Lie group $G$ has a Lie algebra associated to it, which can be defined as the linear subspace $\mathfrak{g} \subset \mathcal{X}(G)$, of right-invariant vector fields on $G$ equipped with the Lie–Jacobi bracket. By right-invariance we mean invariance under right translation. Consider the diffeomorphism $R_g : G \to G$ defined as $R_g(h) = h \cdot g$. A vector field $X \in \mathcal{X}(G)$ is right-invariant if it is $R_g$-related to itself, *i.e.*, $R_{g*} X = X$. By (2.2) it follows that the Lie–Jacobi bracket between two right-invariant vector fields is again right-invariant. For every $v \in T_{\mathrm{e}} G$, where e is the identity element, there is a unique $X_v \in \mathfrak{g}$ such that $X_v|_g = R'_g(v)$, and for $X \in \mathcal{X}(G)$ there is an element $v_X \in T_{\mathrm{e}} G$ given as $v_X = X|_{\mathrm{e}}$. This shows that $\mathfrak{g}$ is isomorphic to $T_{\mathrm{e}} G$, and it may often be convenient to represent

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

$\mathfrak{g}$ as $T_{\mathrm{e}}G$ rather than as right-invariant vector fields on $G$. We next let the group act locally on a manifold. A left group action on a differentiable manifold $M$ by a Lie group $G$ is a map $G \times M \to M$, which we denote as $y = g \cdot x \in M$ or simply $y = gx$ for $g \in G$ and $x \in M$. The group action must satisfy the conditions

$$\mathrm{e} \cdot x = x, \ \forall x \in M, \quad g \cdot (h \cdot x) = (g \cdot h) \cdot x.$$

The group action is said to be *free* if $g \cdot x = x \Rightarrow g = \mathrm{e}$, so that the only group element which leaves $x \in M$ fixed is the identity element. If it is true that for every pair of points $x \in M, y \in M$ there exists a group element $g$ such that $y = g \cdot x$, the action is called transitive. We usually just need a local version of transitivity, requiring that for every $x \in M$, $G \cdot x$ contains some open neighbourhood of $x$. The *orbit* of the group action containing $x$ is the set $\mathcal{O}_x = \{g \cdot x \,|\, g \in G\}$. It is sometimes useful to restrict the action to an orbit when transitivity is a desired property. Similarly, if one needs the group action to be free, it may be useful to extend the action to multispace $M \times^r M$ ($r$ copies of $M$) or to a suitable jet space whenever $M$ takes the form of a fibred space.

Let $G$ be a Lie group with Lie algebra $\mathfrak{g}$. Assume that $G$ acts locally on the manifold $M$, and set $\Lambda_x(g) = g \cdot x$ for $g \in U \subseteq G$, $x \in M$ where $U$ is some open neighbourhood of the identity element. For every fixed $v \in \mathfrak{g}$ there is a vector field $X_v \in \mathcal{X}(\mathcal{M})$ defined by

$$X_v|_x = \lambda_*(v)|_x = \Lambda'_x(v) = \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \gamma(t) \cdot x,$$

where $\gamma(t)$ is any smooth curve on $G$ such that $\gamma(0) = \mathrm{e}$, $\dot{\gamma}(0) = v$. So $\lambda_*$ is a Lie algebra homomorphism of $\mathfrak{g}$ into its image in $\mathcal{X}(M)$. For Lie group integrators, it is important that the action is locally transitive; in this case it is true that for every $x \in X$

$$\lambda_*(\mathfrak{g})|_x = T_x M.$$

This property means, in particular, that for any smooth vector field $F \in \mathcal{X}(M)$ there exists a map $f : M \to \mathfrak{g}$ such that

$$F|_x = \lambda_*(f(x))|_x, \tag{2.3}$$

a formulation called the *generic presentation of ODEs on manifolds* by Munthe-Kaas (1999). If the action is free then the map $f(x)$ is unique for a given vector field $F$. Suppose further that the Lie algebra $\mathfrak{g}$ is of dimension $d$ and let $e_1, \ldots, e_d$ be a basis for $\mathfrak{g}$. Let $E_i = \lambda_*(e_i)$, $i = 1, \ldots, d$. We call the set $\{E_1, \ldots, E_d\}$ a *frame*,[3] and let $\bar{\mathfrak{g}} = \mathrm{span}\{E_1, \ldots, E_m\}$, that is,

[3] In the literature, a 'frame' is often used as a local object requiring that $E_1|_x, \ldots, E_d|_x$ is a basis for $T_x M$. We do not impose this condition here *per se*, as we find it useful to have a global representation of vector fields on $M$.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

the linear span of the frame fields over $\mathbb{R}$ or $\mathbb{C}$. Now we can represent any smooth vector field $X \in \mathcal{X}(M)$ by means of $d$ functions $f_i : M \to \mathbb{R}$:

$$F|_x = \sum_{i=1}^{d} f_i(x)\, E_i|_x. \tag{2.4}$$

The set of functions $f_i$ is uniquely given only if the action is free. There are two alternative ways to proceed, following either the terminology introduced by Crouch and Grossman (1993) or that of Munthe-Kaas (1995, 1998, 1999). In the former case, one introduces a freeze operator $\mathrm{Fr} : M \times \mathcal{X}(M) \to \bar{\mathfrak{g}}$ relative to the frame. It is defined by

$$X_p|_x := \mathrm{Fr}(p, X)|_x = \sum_i f_i(p) E_i|_x. \tag{2.5}$$

The frozen vector field $X_p$ has the property that it coincides with the unfrozen field $X$ at the point $x = p$, *i.e.*, $X_p|_x = X|_x$. A main assumption underlying many Lie group integrators is that flows of frozen vector fields can be calculated, or in some cases approximated with acceptable computational cost.

### *2.2. Integrators based on coordinates on the Lie group*

In what follows, we shall take $\mathfrak{g}$ to be $T_{\mathrm{e}}G$. Consider a diffeomorphism defined from some open subset of $U \subseteq \mathfrak{g}$ containing 0, *i.e.*, $\Psi : U \to G$ and we require that $\Psi(0) = \mathrm{e}$, $\Psi_0'(v) = v$, $\forall v \in \mathfrak{g}$. For convenience, we shall work with a right-trivialized version of $\Psi_u'$, setting

$$\Psi_u' = R'_{\Psi(u)} \circ \mathrm{d}\Psi_u, \quad \mathrm{d}\Psi_u : \mathfrak{g} \to \mathfrak{g}.$$

Let the group act locally on the manifold $M$; suppose that the action is defined on a subset $G_p \subseteq G$, where $\Psi(U) \subset G_p$ for every $p \in M$. For any $p \in M$ and $u \in U$, set $\lambda_p(u) = \Psi(u) \cdot p$, and let $f : \mathcal{M} \to \mathfrak{g}$ represent the vector field $F \in \mathcal{X}(M)$ through the form (2.3). We define the vector field $\tilde{F}_p$ on $\mathcal{X}(U)$ by

$$\tilde{F}_p\big|_u = \mathrm{d}\Psi_u^{-1} f(\lambda_p(u)). \tag{2.6}$$

A simple calculation (Munthe-Kaas 1999, Owren and Marthinsen 2001) shows that $F$ is $\lambda_p$-related to $\tilde{F}$ such that, for any $u \in U$, we have $F|_{\lambda_p(u)} = \lambda_p'\big|_u(\tilde{F}_p|_u)$. This local relatedness between vector fields $F$ of the form (2.3) whose flows are to be approximated, and vector fields on the algebra of the acting group serves as the key underlying principle of many Lie group integrators. The idea is that near any point $p \in M$ one may represent curves in the form $y(t) = \lambda_p(\sigma(t))$, and if the differential equation for $y(t)$ is given by $F$, then a differential equation for $\sigma(t)$ is that of $\tilde{F}$. Most of the known integrators for ODEs have the property that when the solution belongs to

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

some linear space, the numerical method will provide approximations for the solution belonging to the same linear space. So one may now approximate solutions to the ODE given by $\tilde{F}$, and obtain numerical approximations in the linear space $\mathfrak{g}$. The map $\lambda_p$ will then map these approximations onto $M$ by definition. One may typically choose the initial value in each step to be $p$, so that when solving for $\sigma$ in $\mathfrak{g}$ one sets $\sigma(t_n) = 0$. Suppose that a one-step method map which preserves linear structure is denoted $\Phi_{h,\tilde{F}_p}$, where $h$ is the time step.

**Algorithm 2.1.**

**for** $n = 0, 1, \ldots$ **do**
  $p \leftarrow y_n$
  $\sigma_{n+1} \leftarrow \Phi_{h,\tilde{F}_p}(0)$
  $y_{n+1} \leftarrow \lambda_p(\sigma_{n+1})$
**end for**

We may be more specific and for instance insist that the scheme $\Phi_{h,\tilde{F}}$ to be used in the Lie algebra is an explicit Runge–Kutta method, with weights $b^i$ and coupling coefficients $a_i^j$, $1 \leq j < i \leq s$. In this case we can phrase the scheme as follows for integrating a system in the form (2.3) from $t_0$ to $t_0 + h$, with initial value $y(t_0) = y_0 \in M$.

**Algorithm 2.2. (Runge–Kutta–Munthe-Kaas)**

**for** $i = 1 \to s$ **do**
  $u_i \leftarrow h \sum_{j=1}^{i-1} a_i^j k_j$
  $k_i' \leftarrow f(\Psi(u_i) \cdot y_0)$
  $k_i \leftarrow \mathrm{d}\Psi_{u_i}^{-1}(k_i')$
**end for**
$v \leftarrow h \sum_{j=1}^{s} b^j k_j$
$y_1 = \Psi(v) \cdot y_0$

An obvious benefit of schemes of this form is that they will preserve the manifold structure, a feature which cannot be expected when $M$ is modelled as some embedded submanifold of Euclidean space. If the manifold happens to belong to a level set of one or more first integrals, then the Lie group integrators will automatically preserve these integrals. There are, however, other interesting situations, when the group action is used as a building block for obtaining a more accurate representation of the exact solution than is possible with other methods. The idea behind Lie group integrators can also be seen as a form of preconditioning.

*Coordinate maps.* The computational cost of the Lie group integrators is an important issue, and the freedom one has in choosing the coordinate map $\Psi$ may be used to optimize the computational cost. The generic choice for $\Psi(u)$ is of course the exponential map $\Psi(u) = \exp u$. This choice is called

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

*canonical coordinates of the first kind*. If the Lie group and its Lie algebra are realized as matrices, we have

$$\exp u = \sum_{k=0}^{\infty} \frac{u^k}{k!}.$$

Issues related to computing the matrix exponential have been thoroughly debated in the literature, going back to the seminal paper of Moler and Van Loan (1978), and its follow-up, Moler and van Loan (2003). For dense $n \times n$ matrices one would normally expect a computational complexity of approximately $C\,n^3$, where $C$ will depend on several factors, as the tolerance for the accuracy, the size of the matrix elements and the conditioning of the matrix. The methods presented in these papers, however, do not usually respect the Lie group structure, such that an approximation $\tilde{g} \approx \exp u,\ u \in \mathfrak{g}$ will not belong to the group, *i.e.*, $\tilde{g} \notin G$; this is a crucial issue as far as exact conservation is concerned. In practice, one is left with two alternatives.

(1) Apply a standard method which yields $g = \exp u$ to machine accuracy. For relevant examples with Lie group integrators, the factor $C$ typically lies in the range 20–30 (Owren and Marthinsen 2001).

(2) Apply some approximation which is not exact, but which respects the Lie group structure, *i.e.*, $\tilde{g} \approx \exp u$ with $\tilde{g} \in G$. This approach has been pursued by Celledoni and Iserles (2000, 2001) for approximation by low-rank decomposition, as well as Zanna and Munthe-Kaas (2001/02) and Iserles and Zanna (2005) by means of the generalized polar decomposition. These approaches still have a computational cost of $C\,n^3$ for Lie algebras whose matrix representation yields dense matrices, but the constant $C$ may be smaller than for the general algorithms.

It is interesting to explore other possible choices of analytic matrix functions than the exponential. One then replaces the exponential map with some local diffeomorphism from a neighbourhood of $0 \in \mathfrak{g}$ to $G$, usually required to map $0 \mapsto \mathrm{e}$. But it turns out that the exponential map is the only possible choice of analytic map that works for all Lie groups: this is asserted, for instance, using a result by Kang and Shang (1995).

**Lemma 2.3.** Let $\mathfrak{sl}(d)$ denote the set of all $d \times d$ real matrices with trace equal to zero and let $\mathrm{SL}(d)$ be the set of all $d \times d$ real matrices with determinant equal to one. Then, for any real analytic function $R(z)$ defined in a neighbourhood of $z = 0$ in $\mathbb{C}$ satisfying the conditions: $R(1) = 1$ and $R'(0) = 1$, we have that $R(\mathfrak{sl}(d)) \subseteq \mathrm{SL}(d)$ for some $d \geq 3$ if and only if $R(z) = \exp(z)$.

They prove this result by taking the Lie group $\mathrm{SL}(d)$ of unit determinant $d \times d$ matrices, $d \geq 3$, whose Lie algebra $\mathfrak{sl}(d)$ consists of $d \times d$ matrices with

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

trace zero as an example: see also Hairer *et al.* (2006, p. 102). There are, however, exceptions when certain specific Lie groups are considered. For instance, if a matrix group $G_J$ and its corresponding Lie algebra $\mathfrak{g}_J$ can be characterized as

$$G_J = \{A \in \mathrm{GL}(d) : A^\top J A = J\}, \quad \mathfrak{g}_J = \{a \in \mathfrak{gl}(d) : a^\top J + J a = 0\}, \tag{2.7}$$

for some fixed $d \times d$ matrix $J$, it turns out that every analytic function $R(z)$ satisfying

$$R(z)R(-z) = 1 \tag{2.8}$$

will have the property $a \in \mathfrak{g}_J \Rightarrow R(a) \in G_J$. Such groups include the symplectic group $\mathrm{Sp}(d)$, the orthogonal groups $\mathrm{O}(d), \mathrm{SO}(d)$, and the Lorenz groups $\mathrm{SO}(\ell, d-\ell)$. One of the most popular choices of maps that satisfies (2.8) is the Cayley transformation

$$\Psi_{\mathrm{cay}}(z) = \frac{1+z}{1-z}.$$

This transformation was used by Lewis and Simo (1994), and later by Diele, Lopez and Peluso (1998) for the group $SO(d)$, Lopez and Politi (2001) for general groups of the form (2.7), and Marthinsen and Owren (2001) for linear equations.

If one does not require the map to be realizable as an analytic function of the Lie algebra matrix, there are other choices, for instance the *canonical coordinates of the second kind*. This map requires the use of a basis for $\mathfrak{g}$, say $v_1, \ldots, v_s$ where $s = \dim \mathfrak{g}$. Then the map is defined as

$$\Psi_{\mathrm{ccsk}} : v = \alpha_1 v_1 + \cdots + \alpha_s v_s \mapsto \exp(\alpha_1 v_1) \cdot \exp(\alpha_2 v_2) \cdots \exp(\alpha_s v_s).$$

Here, the choice as well as the ordering of the basis is clearly an issue. In Owren and Marthinsen (2001) this coordinate map was, together with a generic basis of the Lie algebra, known as the Chevalley basis. The main advantage of this map is that when matrices are used, natural choices of bases are frequently realized as sparse matrices whose exponentials may be explicitly known. As an example, one may consider the special linear group $\mathrm{SL}(d)$, which can be realized as the set of $d \times d$ matrices with unit determinant. The Lie algebra is then the set of trace-free $d \times d$ matrices, and the Chevalley basis is obtained by choosing $d-1$ trace-free diagonal matrices, say $\mathrm{e}_{i+1}\mathrm{e}_{i+1}^\top - \mathrm{e}_i \mathrm{e}_i^\top$, $1 \le i \le d-1$ together with all matrices $\mathrm{e}_i \mathrm{e}_j^\top$, $j \neq i$. Also, hybrid variants are possible, where, for instance, the Lie algebra is decomposed into a direct sum of subspaces, say

$$\mathfrak{g} = \mathfrak{g}_1 \oplus \mathfrak{g}_2 \oplus \cdots \oplus \mathfrak{g}_{\bar{s}}, \quad \sum_{j=1}^{\bar{s}} \dim(\mathfrak{g}_j) = s,$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

and a coordinate map can be realized as

$$\Psi_{\text{hyb}} : v = v_1 + \cdots + v_{\bar{s}} \mapsto \exp(v_1) \cdots \exp(v_{\bar{s}}), \quad v_i \in \mathfrak{g}_i.$$

In fact, such coordinate maps derived from the generalized polar decomposition were considered in Krogstad, Munthe-Kaas and Zanna (2009).

*Computing the inverse differential.* In the transformation of the vector field to the Lie algebra, one needs to compute the inverse of the differential $\mathrm{d}\Psi_u^{-1} v$. For the case of the exponential mapping, it was found by Baker (1905) and Hausdorff (1906) that it can be expanded in an infinite series of commutators. Defining the operator $\mathrm{ad}_u(v) = [u, v]$, for any $u$ and $v$ in $\mathfrak{g}$, we have

$$\mathrm{dexp}_u^{-1} v = \sum_{k=0}^{\infty} \frac{B_k}{k!} \mathrm{ad}_u^k(v). \tag{2.9}$$

Here, the constants $B_k$ are the Bernoulli numbers, $B_{2k+1} = 0,\ k \geq 1$, and the first non-zero ones are $B_0 = 1, B_1 = -1/2, B_2 = 1/12, B_4 = -1/720$. We will not dwell on the convergence properties of this series (discussed, for instance, in Varadarajan (1984)), the reason being an observation made by Munthe-Kaas (1999). The series (2.9) may be truncated at any point to yield an approximation in the Lie algebra $\mathfrak{g}$. Since one usually requires that the initial point of each time step in the Lie algebra, $\sigma_0 = 0$, it follows that the term of index $k$ will be of at least order $h^k$ as $h \to 0$. It is therefore sufficient to truncate the series at an index that is consistent with the convergence order of the time-stepper $\Phi_{h,\tilde{F}}$. An extensive analysis of the number of terms that must be kept in such commutator expansions was presented in Munthe-Kaas and Owren (1999), but see also McLachlan (1995).

Considering the coordinate maps $\Psi_{\text{ccsk}}$ and $\Psi_{\text{hyb}}$, their inverse differential maps are studied in Owren and Marthinsen (2001) and Krogstad *et al.* (2009) respectively. In these cases, the maps are computed exactly, but by carefully studying the structural properties of the respective Lie algebras it is possible to find computationally inexpensive algorithms: typically, if the dimension of the Lie algebra is $d$, the cost of computing $\mathrm{d}\Psi_u^{-1}(v)$ is $\mathcal{O}(d^{3/2})$.

*Choosing the group and the group action.* A particularly interesting case for a global group action is the homogeneous space. If the group $G$ is acting transitively on $M$, then the manifold is called a homogeneous space. In this case it is well known (see, *e.g.*, Bryant (1995)) that $M$ is naturally diffeomorphic to $G/G_p$, where $G_p$ is the isotropy or stabilizer subgroup of $G$,

$$G_p = \{g \in G : g \cdot p = p\}.$$

The simplest case is when the action is free, so that $G_p = \mathrm{e}$ for all $p \in M$, meaning $M \cong G$. In this case the function $f$ in (2.3) is unique, and there is no isotropy. In the case that there is a non-trivial isotropy group, the choice

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

of $f$ is not unique, in fact $f$ may be replaced by $f + w$, where $w(x) \in \ker \lambda_*$ for every $x \in M$. It turns out that the choice of $w$ affects the numerical integrator.

We proceed to give some examples frequently seen in applications: see Munthe-Kaas and Zanna (1997). Suppose that the Lie group $G = \mathrm{SO}(d)$ is a group of orthogonal $d \times d$ matrices. Its Lie algebra is $\mathfrak{so}(d)$, the set of $d \times d$ matrices which are real and skew-symmetric. The manifold $M$ is a connected open subset of the $d \times r$ matrices, and we give two examples.

(1) Let the Lie group action be left multiplication $g \cdot m$ where $\cdot$ is now just matrix–matrix multiplication. We may now write the ODE vector field as:

$$F|_y = f(y) \cdot y, \quad f : M \to \mathfrak{g}.$$

   (a) $M$ is the set of vectors in $\mathbb{R}^d$ with unit length, identified with the $(d-1)$-dimensional sphere $S^{d-1}$.

   (b) $M$ is the set of $d \times r$ matrices with orthonormal columns, identified with the Stiefel manifold $V_r(\mathbb{R}^d)$.

(2) $r = d$, and the Lie group action is the conjugation $g \cdot m = gmg^\top$, the right-hand side is just products of matrices. We may choose $M$ to be some isospectral set, such as the set of all matrices with a given prescribed Jordan form; one may also restrict to the set of symmetric matrices with a fixed set of eigenvalues.

Another interesting choice is the affine group, first discussed in the context of Lie group integrators in Munthe-Kaas (1999). One may start with the group $G = \mathrm{GL}(d) \rtimes \mathbb{R}^d$, the semidirect product between the general linear group and $\mathbb{R}^d$. $G$ is the group of all affine linear maps acting on $\mathbb{R}^d$; we have $(A, a) \cdot y = Ay + a$. The group product is given as $(A, a) \cdot (B, b) = (AB, Ab + a)$. The corresponding Lie algebra is $\mathfrak{g} = \mathfrak{gl}(d) \rtimes \mathbb{R}^d$ and the Lie bracket is given as

$$[(A, a), (B, b)] = ([A, B], Ab - Ba).$$

The exponential mapping can be expressed in terms of the matrix exponential

$$\exp(A, b) = (\exp(A), \mathrm{dexp}_A(b)),$$

where

$$\mathrm{dexp}_A(b) = \frac{\exp(A) - I}{A}\, b = \sum_{j=0}^{\infty} \frac{1}{(j+1)!} A^j b.$$

Munthe-Kaas made the observation that this action has a large isotropy group: the dimension of the group is $d(d+1)$ whereas the manifold $M = \mathbb{R}^d$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

only has dimension $d$. Setting $\lambda_p(A,a) = \exp(A,a)\cdot p = \exp(A)p + \mathrm{dexp}_A(a)$, we find

$$\lambda_*(A,a)(y) = Ay + a,$$

such that the isotropy algebra at $y$ consists of all elements in $G$ of the form $(A, -Ay)$. Suppose that the ODE vector field is $F$. Then at one extreme one may choose

$$f(y) = (0, F|_y),$$

in which case one would recover the method $\Phi_{h,F}$ in the algorithm described above. However, as suggested in Munthe-Kaas (1999), one may instead add to this the element $(J, -Jy)$, $J \in \mathrm{GL}(d)$ from $\ker \lambda_*(.)(y)$, to obtain the function

$$f(y) = (J, F|_y - Jy),$$

and a natural choice for $J$ would be the Jacobian of the vector field $F$ evaluated at some point near $y$. One may also replace the group $\mathrm{GL}(d)$ by a subgroup: the group $\mathrm{SO}(d)$ yields the special Euclidean group $\mathrm{SE}(d)$, whereas in applications to PDEs a popular choice is to select some one-parameter subgroup $\exp(tL)$, $L \in \mathfrak{gl}(d)$. The Lie subalgebra elements are now of the form $(\alpha L, a)$ with fixed $L$; thus one may just write $(\alpha, a) \in \mathbb{R}\times\mathbb{R}^d$, and the Lie bracket of this subalgebra is simply

$$[(\alpha,a),(\beta,b)] = [0, L(\alpha b - \beta a)]. \tag{2.10}$$

The isotropy group $G_p$ is now only one-dimensional; its Lie algebra is spanned by the element $(1, -Lp)$.

*Time-symmetric Lie group integrators.* For many problems it is of interest to apply integrators which are symmetric in time or self-adjoint. For one-step methods, say $y_1 = \phi_h(y_0)$, this means that if the integrator is applied backwards in time, *i.e.*, with step size $-h$ and $y_1$ as input, the output $y_0$ results, such that $\phi_h = \phi_{-h}^{-1}$. For Lie group integrators, one may consider as an example the Runge–Kutta–Munthe-Kaas scheme based on the midpoint rule, using the coordinate map $\Psi = \exp$. This scheme is an implicit counterpart to the method given in Algorithm 2.2:

$$u = \frac{h}{2}k, \quad k = f(\exp(u)\cdot y_0), \quad y_1 = \exp(hk)\cdot y_0.$$

Here we have omitted the correction $\mathrm{dexp}_u^{-1}$, since it is not going to alter the order of the method. Following Zanna, Engø and Munthe-Kaas (2001), we define the midpoint approximation $y_{\frac12} = \exp\left(\frac12 hu\right)\cdot y_0$; the method may be written in the form

$$y_1 = \exp\left(\frac{h}{2}f\left(y_{\frac12}\right)\right).y_{\frac12}, \quad y_0 = \exp\left(-\frac{h}{2}f\left(y_{\frac12}\right)\right).y_{\frac12},$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

which is easily seen to be symmetric. Unfortunately, Lie group integrators based on coordinates on the Lie group are not generally symmetric even if the underlying scheme on the Lie algebra is a symmetric method. Zanna *et al.* (2001) considers how a recentering of the coordinate system for the Lie group can be used to obtain symmetric integrators. Rather than letting $\Psi(0) = \mathrm{e}$ be the base point of the coordinate chart, one may choose a $\theta \in \mathfrak{g}$ and define a new base point $q := \Psi(\theta)^{-1} \cdot y_0$. The curve $y(t) \in G$ is now represented locally by a curve $\sigma(t) \in \mathfrak{g}$, as

$$y(t) = \Psi(\sigma(t)) \cdot q = \Psi(\sigma(t)) \cdot \Psi(\theta)^{-1} \cdot y_0.$$

This shift of coordinate chart results in the same differential equation for $\sigma(t)$, but with initial value $\sigma(0) = \theta$ corresponding to $y(0) = y_0$. An implicit Runge–Kutta-type method generalizes Algorithm 2.2 as follows:

$$\left.\begin{aligned} \sigma_i &= \theta + h \textstyle\sum_{j=1}^{s} a_i^j k_j, \\ k_i' &= f(\Psi(\sigma_i) \cdot \Psi(\theta)^{-1} \cdot y_0), \\ k_i &= \mathrm{d}\Psi_{\sigma_i}^{-1}(k_i'), \end{aligned}\right\} \quad i = 1, \ldots, s, \tag{2.11}$$

$$\begin{aligned} v &= \theta + h \textstyle\sum_{i=1}^{s} b^i k_i, \\ y_1 &= \Psi(v) \cdot \Psi^{-1}(\theta) \cdot y_0. \end{aligned}$$

The idea is now to choose the centre $\theta$ depending on the stages $k_i$ and $h$ in such a way that the resulting method is symmetric whenever the underlying Runge–Kutta scheme is symmetric. From Zanna *et al.* (2001) we find that this can be achieved by setting $\theta = \theta(h; k_1, \ldots, k_s)$ and requiring it to satisfy

$$\theta(-h; k_s, \ldots, k_1) = \theta(h; k_1, \ldots, k_s) + h \sum_{i=1}^{s} b^i k_i. \tag{2.12}$$

There is of course no unique way of satisfying (2.12). An obvious choice which has been called the geodesic midpoint is

$$\theta(h, k_1, \ldots, k_s) = -\frac{h}{2} \sum_{i=1}^{s} b^i k_i,$$

but in fact, one may choose more generally

$$\theta(h; k_1, \ldots, k_s) = -h \sum_{i=1}^{s} w^i k_i, \quad \text{where} \quad w^{s+1-i} + w^i = b^i, \quad i = 1, \ldots, s.$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

A concrete example of a symmetric scheme in the format (2.11), using the geodesic midpoint for $\theta$ and the 2-stage Gauss method is

$$\begin{aligned}
\theta &= -\frac{h}{4}(k_1 + k_2), \\
\sigma_1 &= -h\frac{\sqrt{3}}{6}k_2, & \sigma_2 &= h\frac{\sqrt{3}}{6}k_1, \\
k_1' &= f(\Psi(\sigma_1)\cdot\Psi(\theta)^{-1}\cdot y_0), & k_2' &= f(\Psi(\sigma_2)\cdot\Psi(\theta)^{-1}\cdot y_0), \\
k_1 &= \mathrm{d}\Psi_{\sigma_1}^{-1}(k_1'), & k_2 &= \mathrm{d}\Psi_{\sigma_2}^{-1}(k_2'), \\
v &= \frac{h}{4}(k_1 + k_2), & y_1 &= \Psi(v)\cdot\Psi(\theta)^{-1}\cdot y_0.
\end{aligned}$$

McLachlan, Quispel and Tse (2009) recently studied symmetric and symplectic schemes which are linearization-preserving at fixed points, meaning that if $\bar{y}$ is such that $f(\bar{y}) = 0$ in the initial value problem $y' = f(y),\ y(0) = y_0$ then the integrator has the Taylor expansion about $\bar{y}$:

$$y_1 - \bar{y} = \exp(hf'(\bar{y}))(y_0 - \bar{y}) + \mathcal{O}(\|y_0 - \bar{y}\|^2).$$

Symmetric schemes for a class of exponential integrators applied to semilinear problems were proposed by Celledoni, Cohen and Owren (2008).

### *2.3. Lie group integrators based on compositions of flows*

A pioneering article by Crouch and Grossman (1993) suggests applying compositions of flows of frozen vector fields (2.5) to construct integration methods for vector fields which have been expressed in the form (2.4). They propose such generalizations both of multistep methods and of explicit Runge–Kutta methods. The latter takes a step from $t_0$ to $t_0 + h$ as follows.

**Algorithm 2.4. (Crouch–Grossman)**

$Y_1 \leftarrow y_0$
$F_1 \leftarrow \mathrm{Fr}(Y_1, F)$
**for** $r = 2 \to s$ **do**
  $Y_r \leftarrow \exp(ha_r^{r-1}F_{r-1})\cdots\exp(ha_r^1F_1)(y_0)$
  $F_r \leftarrow \mathrm{Fr}(Y_r, F)$
**end for**
$y_1 \leftarrow \exp(hb^sF_s)\cdots\exp(hb^1F_1)(y_0)$

Here, $b^r, a_r^k$ are the weights and the coupling coefficients of the Runge–Kutta method. Explicit third-order with $s = 3$ were presented in Crouch and Grossman (1993), and Owren and Marthinsen (1999) presented a general theory of order conditions and examples of fourth-order methods with four stages. A potential drawback of the Crouch–Grossman methods is that they require the computation of a large number of flows of frozen vector fields in every time step. This led Celledoni, Marthinsen and Owren

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

(2003) to consider a version of these methods where flows of arbitrary linear combinations of frozen vector fields were considered. The methods take the following form.

**Algorithm 2.5. (Commutator-free methods)**

$Y_1 \leftarrow y_0$
$F_1 \leftarrow \mathrm{Fr}(Y_1, F)$
**for** $r = 2 \to s$ **do**
  $Y_r \leftarrow \exp(h \sum_k \alpha^k_{r,J} F_k) \cdots \exp(h \sum_k \alpha^k_{r,1} F_k)(y_0)$
  $F_r \leftarrow \mathrm{Fr}(Y_r, F)$
**end for**
$y_1 \leftarrow \exp(h \sum_k \beta^k_J F_k) \cdots \exp(h \sum_k \beta^k_1 F_k)(y_0)$

Note that when the vector fields commute, the expressions for $Y_r$ and $y_1$ collapse to

$$Y_r = \exp\Big(\sum_k a^k_r F_k\Big)(y_0), \quad y_1 = \exp\Big(\sum_k b^k F_k\Big)(y_0),$$

where

$$a^k_r = \sum_{j=1}^{J} \alpha^k_{r,j}, \quad b^k = \sum_{j=1}^{J} \beta^k_j.$$

These coefficients correspond to the $a^k_r$, $b^k$ in the Runge–Kutta–Munthe-Kaas methods, and they must in particular satisfy the classical order conditions for Runge–Kutta methods. The key idea of the commutator-free methods is to keep the number of exponentials or flows as low as possible in each stage. In Celledoni *et al.* (2003) the following fourth-order method was presented:

$$\begin{aligned}
&Y_1 = y_0, && F_1 = \mathrm{Fr}(Y_1, F),\\
&Y_2 = \exp\big(\tfrac{1}{2} h F_1\big)(y_0), && F_2 = \mathrm{Fr}(Y_2, F),\\
&Y_3 = \exp\big(\tfrac{1}{2} h F_2\big)(y_0), && F_3 = \mathrm{Fr}(Y_3, F),\\
&Y_4 = \exp\big(-\tfrac{1}{2} h F_1 + h F_3\big)(Y_2), && F_4 = \mathrm{Fr}(Y_4, F),\\
&y_1 = \exp\big(\tfrac{h}{12}(-F_1 + 2F_2 + 2F_3 + 3F_4)\big) \exp\big(\tfrac{h}{12}(3F_1 + 2F_2 + 2F_3 - F_4)\big)(y_0).
\end{aligned} \tag{2.13}$$

Note that if the number of exponentials is counted according to the format of Algorithm 2.5, we would have two exponentials in the expression for $Y_4$, but since the rightmost exponential coincides with $Y_2$, we effectively compute only one exponential in each stage $Y_r, r = 2, 3, 4$ and two exponentials in $y_1$, and thus the total number of flow calculations in this method is 4. In comparison, the methods of Crouch and Grossman would require 10 flow calculations. A complete description of the order theory for this type of method was given in Owren (2006). Wensch, Knoth and Galant (2009)

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

extend the idea of re-using flow calculations, as with $Y_2$ in (2.13), to allow for the composition of exponentials to be applied to more general expressions involving $y_0$ and the stage values $Y_i$. For the case of linear differential equations, one can design commutator-free methods based on the Magnus series expansions: see Blanes and Moan (2006).

### *2.4. Applications of Lie group integrators to PDEs*

The Lie group integrators described in the preceding sections can be used as time-steppers in solving partial differential equations. It should come as no surprise that these integrators are intimately linked to a well-known class of methods known as 'exponential integrators', which go back to the early 1960s in pioneering work by Certaine (1960), Nørsett (1969) and Lawson (1967). An excellent account of these methods was presented in a recent *Acta Numerica* article by Hochbruck and Ostermann (2010), and we refer the reader to that article and the references therein for a detailed account of exponential integrators. There are several important papers which study analytical properties of exponential integrators, in particular for parabolic problems. An important issue is to describe the convergence order in the presence of unbounded operators: see, for instance, Hochbruck and Ostermann (2005) and Ostermann, Thalhammer and Wright (2006) and the references therein. In studying such time integrators one possibility is to apply the time-stepper without discretizing in space, thereby interpreting the method as a discrete trajectory in some abstract space of infinite dimension. Another option is to first do a spatial discretization to obtain a finite-dimensional system, and then apply the time-stepping scheme to this finite-dimensional problem.

The most widespread application of exponential integrators is perhaps to the semilinear problem

$$u_t = Lu + N(t, u). \tag{2.14}$$

Here $L$ is a linear operator, and $N(t, u)$ is a nonlinear function or mapping. In many interesting cases, $L$ may be a differential operator of higher order than those appearing in $N(t, u)$; we refer to Minchev (2004) for an extensive treatment of this case. Such a problem lends itself to a formulation by the action of the affine group as described above; then any Lie group integrator can be applied. Cox and Matthews (2002) found an exponential integrator for (2.14) which turned out to be identical to the general commutator-free Lie group integrator in Celledoni *et al.* (2003) applied with the affine action adapted to (2.14), when the Lie algebra $\mathfrak{g}$ is a subalgebra of the affine Lie algebra consisting of elements $(\alpha L, b)$ as described above. Also, Krogstad (2005) derived exponential integrators with excellent properties for semilinear problems. In what follows we shall consider the application of Lie group integrators to some non-standard cases, differing from (2.14).

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

*Celledoni–Kometa methods.* Whereas the most common application of exponential integrators is to problems of the form (2.14), there are other examples in the literature. One is a class of schemes proposed by Celledoni and Kometa (Celledoni 2005, Celledoni and Kometa 2009). They consider convection–diffusion problems of the form

$$\frac{\partial u}{\partial t} + \mathbf{V}(u) \cdot \nabla u = \nu \nabla^2 u + f, \tag{2.15}$$

where the convection term $\mathbf{V}(u) \cdot \nabla u$ is dominating, and $0 < \nu \ll 1$. A semidiscretized version of (2.15) is obtained as

$$y_t - C(y)y = Ay + f, \quad y(0) = y_0. \tag{2.16}$$

Here the nonlinear convection term is assumed to take the form $C(y)y$ for some matrix $C(y)$ and the diffusion operator is similarly replaced by a constant matrix $A$. A typical strategy is now to treat the convection term and the diffusion term separately, using an explicit and possibly high-order approximation for the convection term, and an implicit integrator for the comparably stiff diffusion term. This is the idea behind the IMEX methods, for instance: see Kennedy and Carpenter (2003) for a general theory. Note that the matrix or linear operator $A$ will usually be symmetric and negative definite: therefore implicit schemes can be implemented efficiently in such a splitting. The approach of Celledoni and Kometa (2009) is to use a non-trivial flow calculation for the convection part. A model example of a method in the class they consider for (2.16) is

$$y_1 = \exp(h\, C(y_0))y_0 + h\, A y_1 + h\, f.$$

This scheme can be generalized by adding stages in a Runge–Kutta fashion and by introducing compositions of exponentials in a similar way to that of commutator-free Lie group integrators. This leads to a semi-Lagrangian approach in which linearized versions of the convection part of the problem are solved exactly by the method of characteristics, for example. The general form of the method applied to (2.16) is

$$\left.\begin{aligned} Y_i &= \phi_i y_0 + h \textstyle\sum_{j=1}^{s} a_{i,j} \phi_{i,j} A Y_j, \\ \phi_i &= \exp\big(h \textstyle\sum_k \alpha_{i,J}^k C(Y_k)\big) \cdots \exp\big(h \textstyle\sum_k \alpha_{i,1}^k C(Y_k)\big), \\ \phi_{i,j} &= \phi_i \phi_j^{-1}, \end{aligned}\right\} \quad i = 1, \ldots, s,$$

$$\begin{aligned} y_1 &= \bar{\phi} y_0 + h \textstyle\sum_{i=1}^{s} b_i \bar{\phi}_i A Y_j, \\ \bar{\phi} &= \exp\big(h \textstyle\sum_k \beta_J^k C(Y_k)\big) \cdots \exp\big(h \textstyle\sum_k \beta_1^k C(Y_k)\big), \\ \bar{\phi}_i &= \bar{\phi} \phi_i^{-1}. \end{aligned}$$

Examples of schemes of order two and three are provided in Celledoni and Kometa (2009).

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

*Commutator-free methods.* Wensch *et al.* (2009) introduce the multirate infinitesimal step method for differential equations formulated on Euclidean space of the form

$$y' = f(y) + g(y). \tag{2.17}$$

The method is defined by the authors as follows.

(1) Define the initial points of the stage flows as

$$Z_{ni}(0) = y_n + \sum_j \alpha_{ij}(Y_{nj} - y_n).$$

(2) Compute the stages by solving the following equations exactly:

$$Z'_{ni}(\tau) = \frac{1}{h}\sum_j \gamma_{ij}(Y_{nj} - y_n) + \sum_j \beta_{ij} f(Y_{nj}) + d_i g(Z_{ni}(\tau)),$$
$$Y_{ni} = Z_{ni}(h),$$

and the update stage, $y_{n+1}$, can be interpreted as an additional stage $Y_{n,s+1}$.

It is thus assumed that one can calculate exactly flows of vector fields of the form

$$y' = c + f(y), \tag{2.18}$$

and the explicitness of the scheme is ensured by requiring that $\alpha_{ij} = \beta_{ij} = \gamma_{ij} = 0$ for $j \geq i$. Furthermore, one assumes that the method is balanced, meaning that $d_i = \sum_j \beta_{ij}$ for every $i$. The order analysis can be treated by the theory developed in Owren (2006), and third-order methods are derived under some simplifying assumptions, setting the $\gamma_{ij} = 0$ and demanding that precisely one $\alpha_{ij} \neq 0$ for each $i$. When $g(y) \equiv 0$, the scheme reduces to a standard Runge–Kutta method. In practical situations one may relax the requirement that the flow of (2.18) is solved exactly, but one may instead approximate it with high order of accuracy, for instance by using much smaller time steps than the $h$ which features in the method format above – hence the name multirate methods.

The main application of these methods in Wensch *et al.* (2009) is the 2D Euler equations for atmosphere processes, in which the two main scales are the fast acoustic waves and the slower advection terms:

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot (\rho \mathbf{u}),$$
$$\frac{\partial \rho \mathbf{u}}{\partial t} = -\nabla \cdot (\mathbf{u} \otimes \rho \mathbf{u}) - \frac{\partial p}{\partial \Theta} \nabla \Theta - \mathbf{f}(\rho),$$
$$\frac{\partial \Theta}{\partial t} = -\nabla \cdot (\Theta \mathbf{u}),$$
$$p = \Theta R (p/p_0)^{\kappa}.$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

Here $\mathbf{u} = (u, w)$ denotes the horizontal and vertical velocity, $\rho$ is the density, $\Theta = \theta\rho$ where $\theta$ is the potential temperature, $p$ is the pressure and $R, p_0, \kappa$ are physical parameters. The spatial discretizations used are third-order upwind differences for the advection terms and centred differences for the pressure terms.

## 3. Schemes which preserve first integrals

The development of energy-preserving finite difference schemes go all the way back to the seminal work of Courant, Friedrichs and Lewy (1928) where the celebrated energy method was presented for the first time. Energy is an example of a first integral, a functional which is preserved along exact solutions of the PDE. In the early days of integral-preserving methods, the focus was mostly on stability issues, the preservation of energy, for instance, is an invaluable tool in proving stability of numerical methods for evolutionary problems. More recently some of the interest has shifted to the conservation property itself, considering not only stability issues, but also the ability of the numerical solution to inherit structural properties of the continuous system. In this section we shall discuss some rather general procedures that can be used to construct numerical schemes which exactly preserve one or more first integrals of the partial differential equations. The framework we present includes, but is not limited to, Hamiltonian PDEs. That is, we shall consider systems which can be written in the form

$$u_t = \mathcal{D}\,\frac{\delta \mathcal{H}}{\delta u}, \tag{3.1}$$

where $\mathcal{D}$ is a skew-adjoint operator, which is applied to the *variational derivative* of a first integral $\mathcal{H}$. In general, the operator $\mathcal{D}$ may depend on the solution itself as well as its partial derivatives with respect to spatial variables. In the literature one can also find several examples in which 'dissipative versions' of (3.1) are studied, where the operator $\mathcal{D}$ is replaced by an operator which causes the functional $\mathcal{H}$ to decrease monotonically along exact solutions, the aim then being to design a numerical method which has the same dissipation property. Here we shall consider only the conservative case, and our aim is to derive numerical schemes which, at each time level $t = t_n$, produce a numerical approximation $U^n$ to the exact solution $u(t_n, \cdot)$ of (3.1), satisfying $\mathcal{H}(U^n) = H(U^0)$ for all $n \geq 1$, or $\mathcal{H}_d(\mathbf{u}^n) = \mathcal{H}_d(\mathbf{u}^0)$ where $\mathcal{H}_d$ is a suitable spatially discretized version of $\mathcal{H}$ and $\mathbf{u}^n$ is the corresponding space discretization of $U^n$. We shall also assume that the spatial domain $X$ and boundary conditions are of such a type that (possibly repeated) integration by parts does not cause boundary terms to appear. In the rest of this section we define $M$ to be a fixed connected open subset $M \subset X \times U$, where $X$ is the space of independent variables $x_1, \ldots, x_p$, and

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

$U$ is the space of dependent variables $u_1, \dots, u_q$. The spaces of prolongations $M^{(n)} \subset U^{(n)}$ are then open subsets of the corresponding jet spaces. Following Olver (1993), we define the $\mathbb{C}$-algebra of smooth differential functions $\mathcal{A}$ from $M^{(n)}$ to $\mathbb{C}$. The space $\mathcal{A}^q$ is the set of $q$-plets of differential functions. Usually, the order $n$ of the prolongation is of minor importance to the arguments presented here, and thus we just write $F[u]$ or $F((u^\alpha_J))$ for $F \in \mathcal{A}^q$. We shall make use of operators $\mathcal{D} = \mathcal{D}[u]$, whose adjoint $\mathcal{D}^*[u]$ is defined via

$$\int_X \mathcal{D}P\, Q = \int_X P\mathcal{D}^* Q$$

for all compactly supported differential functions $P$ and $Q$. The operator $\mathcal{D}$ is skew-adjoint if $\mathcal{D}^* = -\mathcal{D}$. We consider the subclass of evolution equations which can be written in the form

$$\frac{\partial u}{\partial t} = F[u] = \mathcal{D}[u]\,\frac{\delta \mathcal{H}}{\delta u}[u], \tag{3.2}$$

where $\mathcal{D}[u]$ is skew-adjoint. The functional $H[u]$ is usually of the form

$$\mathcal{H}[u] = \int \mathcal{G}[u]. \tag{3.3}$$

Equations of this form include the set of Hamiltonian PDEs, but here we do not need to assume that the operator $\mathcal{D}[u]$ defines a Poisson bracket. The general formula for the variational derivative expressed in terms of the Euler operator is included here for later use: $\frac{\delta \mathcal{H}}{\delta u}$ is an $m$-vector depending on $u^\alpha_J$ for $|J| \le \nu'$ where $\nu' \ge \nu$, defined via (Olver 1993, p. 245)

$$\int_\Omega \frac{\delta \mathcal{H}}{\delta u} \cdot \varphi \,\mathrm{d}x = \left.\frac{\partial}{\partial \epsilon}\right|_{\epsilon=0} \mathcal{H}[u + \epsilon\varphi], \tag{3.4}$$

for any sufficiently smooth $m$-vector of functions $\varphi(x)$. One may calculate $\frac{\delta \mathcal{H}}{\delta u}$ by applying the Euler operator to $\mathcal{G}[u]$; the $\alpha$-component is given by

$$\left(\frac{\delta \mathcal{H}}{\delta u}\right)^\alpha = \mathbf{E}_\alpha \mathcal{G}[u], \tag{3.5}$$

where

$$\mathbf{E}_\alpha = \sum_{|J| \le \nu} (-1)^{|J|} D_J \frac{\partial}{\partial u^\alpha_J}, \tag{3.6}$$

so that the sum ranges over all $J$ corresponding to derivatives $u^\alpha_J$ featuring in $\mathcal{G}$. We have used total derivative operators:

$$D_J = D_{j_1} \dots D_{j_k}, \quad D_i = \sum_{\alpha, J} \frac{\partial u^\alpha_J}{\partial x^i} \frac{\partial}{\partial u^\alpha_J}.$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

Another example of a Hamiltonian PDE is a generalized version of the Korteweg–de Vries equation

$$u_t + u_{xxx} + (u^{p-1})_x = 0, \tag{3.7}$$

which can be cast in the form (3.2), with

$$\mathcal{H}[u] = \int \left( \frac{1}{2} u_x^2 - \frac{1}{p} u^p \right) \mathrm{d}x, \quad \mathcal{D} = \frac{\partial}{\partial x}. \tag{3.8}$$

Clearly, the functional $\mathcal{H}[u]$ is conserved along solutions of (3.2), since

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}[u] = \frac{\delta \mathcal{H}}{\delta u}[u]\, u_t = \frac{\delta \mathcal{H}}{\delta u}[u] \mathcal{D}[u] \frac{\delta \mathcal{H}}{\delta u}[u] = 0,$$

because $\mathcal{D}$ is skew-adjoint. The idea behind conservative integration schemes is to construct numerical methods which exactly reproduce this property as the numerical method advances from time $t_n$ to $t_{n+1}$, *i.e.*, assuming that the solution to (3.2) at time $t = t_n$ is approximated by $U^n$, we ask the scheme to satisfy

$$\mathcal{H}[U^n] = \mathcal{H}[U^0], \quad \forall n \geq 1. \tag{3.9}$$

Alternatively, (3.9) may be imposed for some approximation $\mathcal{H}_d$ to $\mathcal{H}$. The methodology we present here has its source in the ODE literature. One of the most prevalent approaches is that of discrete gradients, examples of which were proposed by LaBudde and Greenspan (1974), and treated systematically by Gonzalez (1996) and McLachlan, Quispel and Robidoux (1999). A finite-dimensional counterpart to (3.2) is the system

$$\dot{y} = S(y) \nabla H(y) = f(y), \quad y \in \mathbb{R}^m, \tag{3.10}$$

where $S(y)$ is an antisymmetric matrix. The idea is to introduce a discrete approximation to the gradient, letting $\overline{\nabla} H : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m$ be a continuous map satisfying

$$H(u) - H(v) = \overline{\nabla} H(v, u)^\top (u - v),$$
$$\overline{\nabla} H(u, u) = \nabla H(u).$$

The existence of such discrete gradients is well established in the literature, and their construction is not unique: see, for instance, the monograph by Hairer *et al.* (2006). The averaged vector field (AVF) gradient, for example, is defined by

$$\overline{\nabla}_{\mathrm{AVF}}(v, u) = \int_0^1 \nabla H(\xi u + (1 - \xi) v)\, \mathrm{d}\xi. \tag{3.11}$$

Once a discrete gradient has been found, one immediately obtains an integral-preserving method, simply letting

$$\frac{y^{n+1} - y^n}{\Delta t} = \overline{S}(y^n, y^{n+1})\, \overline{\nabla}(y^n, y^{n+1}).$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

Here $\Delta t$ is the time step, and $\overline{S}(y^n, y^{n+1})$ is some approximation to the matrix $S$ in (3.10); one would normally require that $S(y) = \overline{S}(y, y)$. Discrete gradient methods are usually implicit. The generalization of this approach to PDEs is immediate, and has been developed independently of the ODE literature by Furihata, Matsuo and co-authors (Furihata 1999, Furihata 2001*a*, Furihata 2001*b*, Furihata and Matsuo 2003, Matsuo 2007, Matsuo 2008, Matsuo and Furihata 2001, Matsuo, Sugihara, Furihata and Mori 2000, Matsuo, Sugihara, Furihata and Mori 2002, Yaguchi, Matsuo and Sugihara 2010), and recently presented in its full generality in Dahlby and Owren (2010). One may define a *discrete variational derivative* of the Hamiltonian $\mathcal{H}[u]$ to be a continuous (differential) function of two arguments $u$ and $v$ $\frac{\delta \mathcal{H}}{\delta(v,u)}$ satisfying the relations

$$\mathcal{H}[u] - \mathcal{H}[v] = \int_\Omega \frac{\delta \mathcal{H}}{\delta(v,u)}(u - v)\,\mathrm{d}x, \tag{3.12}$$

$$\frac{\delta \mathcal{H}}{\delta(u,u)} = \frac{\delta \mathcal{H}}{\delta u}. \tag{3.13}$$

To obtain a conservative scheme we simply replace the skew-adjoint operator $\mathcal{D}[u]$ in (3.2) by some approximation $\overline{\mathcal{D}}[v, u]$ satisfying $\overline{\mathcal{D}}[u, u] = \mathcal{D}[u]$, and define the method as

$$\frac{U^{n+1} - U^n}{\Delta t} = \overline{\mathcal{D}}[U^n, U^{n+1}] \frac{\delta \mathcal{H}}{\delta(U^n, U^{n+1})}. \tag{3.14}$$

The AVF scheme can of course also be interpreted as a discrete variational derivative method, where

$$\frac{\delta \mathcal{H}_{\mathrm{AVF}}}{\delta(v,u)} = \int_0^1 \frac{\delta \mathcal{H}}{\delta u}[\xi u + (1-\xi)v]\,\mathrm{d}\xi. \tag{3.15}$$

The fact that (3.15) verifies the condition (3.12) is seen from the elementary identity

$$\mathcal{H}[u] - \mathcal{H}[v] = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}\xi}\mathcal{H}[\xi u + (1-\xi)v]\,\mathrm{d}\xi. \tag{3.16}$$

The derivative under the integral is written

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\xi}\mathcal{H}[\xi u + (1-\xi)v] &= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\bigg|_{\varepsilon=0} \mathcal{H}[v + (\xi+\varepsilon)(u-v)] \\ &= \int_\Omega \frac{\delta \mathcal{H}}{\delta u}[\xi u + (1-\xi)v](u-v)\,\mathrm{d}x.\end{aligned}$$

Now substitute this into (3.16) and interchange the integrals to obtain (3.12).

We illustrate the discrete variational derivative method to the Korteweg–de Vries equaton (3.7)–(3.8) using (3.15). The variational derivative is

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

$\frac{\delta \mathcal{H}}{\delta u} = -u_{xx} - u^{p-1}$ such that

$$\frac{\delta \mathcal{H}_{\mathrm{AVF}}}{\delta(v,u)} = -\int_0^1 (\xi u + (1-\xi)v)_{xx} + (\xi u + (1-\xi)v)^{p-1}\,\mathrm{d}\xi$$
$$= -\frac{u_{xx}+v_{xx}}{2} - \frac{1}{p}\frac{u^p - v^p}{u-v}.$$

Thus, taking $\overline{\mathcal{D}} = \mathcal{D} = \frac{\partial}{\partial x}$, the conservative method for (3.7) reads

$$\frac{U^{n+1}-U^n}{\Delta t} + \frac{U^{n+1}_{xxx} + U^n_{xxx}}{2} + \frac{1}{p}\frac{\partial}{\partial x}\left(\frac{(U^{n+1})^p - (U^n)^p}{U^{n+1}-U^n}\right) = 0.$$

*3.1. Linearly implicit methods for polynomial Hamiltonians*

A major principle behind the conservative schemes is that they preserve the (discretized) first integral exactly. This means that for an implicit scheme, the nonlinear equation to be solved in each time step must be solved exactly or to machine accuracy, and this may sometimes cause the overall method to be expensive compared to other approaches. A possible remedy would be to look for an iteration method having the property that the first integral is preserved in every iteration; another possibility is to design schemes which are linearly implicit by construction, thereby requiring only one linear solve for every time step. In the second of these two approaches, discussed in Dahlby and Owren (2010), one may apply a polarization technique if the Hamiltonian is of polynomial type, meaning that the differential function $\mathcal{G}$ in (3.3) is a multivariate polynomial in the jet space coordinates $u^{(n)}$. The idea is to substitute $\mathcal{G}[u]$ by a function of $k \geq 2$ indeterminates $G[w_1, \ldots, w_k]$, satisfying

$$G[u, \ldots, u] = \mathcal{G}[u] \qquad \text{(consistency)}, \tag{3.17}$$

$$G[w_1, w_2, \ldots, w_k] = G[w_2, \ldots, w_k, w_1] \qquad \text{(cyclicity)}. \tag{3.18}$$

In particular, whenever $\mathcal{G}[u] = \mathcal{G}((u^\alpha_J))$ is a multivariate polynomial of degree $p$, one may construct $G[w_1, \ldots, w_k]$ satisfying (3.17)–(3.18), which is also multi-quadratic as long as $k \geq \lfloor (p+1)/2 \rfloor$. We next replace the skew-adjoint operator $\mathcal{D}[u]$ by some skew-adjoint approximation $D[w_1, \ldots, w_{k-1}]$ of $k-1$ arguments, satisfying

$$D[u, \ldots, u] = \mathcal{D}[u] \qquad \text{(consistency)}, \tag{3.19}$$

$$D[w_1, w_2, \ldots, w_{k-1}] = D[w_2, \ldots, w_{k-1}, w_1] \qquad \text{(cyclicity)}. \tag{3.20}$$

Defining

$$H[w_1, \ldots, w_k] = \int_\Omega G[w_1, \ldots, w_k],$$

we may now define a polarized version of the discrete variational derivative in a similar way to (3.12)–(3.13), but it is now a function of $k+1$ arguments.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

We refer to Dahlby and Owren (2010) for the general definition and give only the particular expression for the AVF case:

$$\frac{\delta H}{\delta(w_1,\dots,w_{k+1})} = \int_0^1 \frac{\delta H}{\delta w_1}[\xi w_{k+1} + (1-\xi)w_1, w_2, \dots, w_k]\,\mathrm{d}\xi. \tag{3.21}$$

Here the variational derivative on the right-hand side, $\frac{\delta H}{\delta w_1}$, is defined as before, considering $H$ as a function of its first argument only, leaving the others fixed. The following method was proposed in Dahlby and Owren (2010):

$$\frac{U^{n+k} - U^n}{k\Delta t} = kD\frac{\delta H}{\delta(U^n,\dots,U^{n+k})}, \quad n \geq 0. \tag{3.22}$$

**Remark 3.1.** Note that the method (3.22) is a multistep method, and thus approximations $U^1,\dots,U^{k-1}$ to the first time steps are required.

**Remark 3.2.** The procedure is linear, in the sense that if the Hamiltonian can be split into a sum of two terms $\mathcal{H}_1 + \mathcal{H}_2$, the method can be applied to each term separately.

We summarize some important properties of the scheme (3.22), the proof of which can be found in Dahlby and Owren (2010).

**Theorem 3.3.**

- The scheme (3.22) is conservative in the sense that
$$H[U^n,\dots,U^{n+k-1}] = H[U^0,\dots,U^{k-1}], \quad \forall n \geq 1,$$
for any polarized Hamiltonian function $H$ satisfying (3.17)–(3.18).
- The polarized AVF scheme (3.22) using (3.21) has formal order of consistency two for any polarized Hamiltonian satisfying (3.17)–(3.18), and skew-symmetric operator $D$ satisfying (3.19)–(3.20).
- If $G[w_1,\dots,w_k]$ is multi-quadratic in all its $k$ arguments, $(w_j^\alpha)_j$, $j = 1,\dots,k$, then the polarized AVF method (3.22) is linearly implicit.

To illustrate the linearly implicit scheme just presented, consider (3.7) with $p = 6$, the mass critical generalized Korteweg–de Vries equation. We then have

$$\mathcal{G}[u] = \frac{1}{2}u_x^2 - \frac{1}{6}u^6.$$

As indicated in Remark 3.2, we can treat the two terms separately. For the second term, one needs to use $k \geq 3$: choosing $k = 3$ leads to the unique multi-quadratic choice $G_2[w,v,u] = G_2(w,v,u) = -\frac{1}{6}u^2v^2w^2$. In the first term we have several options, but for reasons to be explained later, we choose $G_1[w,v,u] = \frac{1}{6}(u_x^2 + v_x^2 + w_x^2)$. As before, we let $\bar{D} := \mathcal{D} = \partial/\partial x$.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

This now leads to the scheme

$$\frac{U^{n+3}-U^n}{3\Delta t}+\frac{U^n_{xxx}+U^{n+3}_{xxx}}{2}+\frac{\partial}{\partial x}\left(\frac{U^n+U^{n+3}}{2}(U^{n+1})^2(U^{n+2})^2\right)=0.$$

Thus the scheme is linear in $U^{n+3}$, and preserves exactly the time-averaged Hamiltonian

$$H[U^n,U^{n+1},U^{n+2}]=\int\frac{1}{6}\big((U^n_x)^2+(U^{n+1}_x)^2+(U^{n+2}_x)^2\big)+\frac{1}{6}\big(U^nU^{n+1}U^{n+2}\big)^2\,\mathrm{d}x.$$

We remark that some care should be taken when choosing the polarization. Since the resulting method is a multistep scheme, it has spurious solutions which might be unstable. Considering as a test case the term $\mathcal{H}[u]=\int u_x^2$, letting $\mathcal{D}$ be a constant operator having the exponentials as eigenfunctions, such as in differentiation operators with constant coefficients, one may choose the one-parameter family of polarizations

$$H[v,u]=\frac{1}{2}\int\left(\theta\frac{u_x^2+v_x^2}{2}+(1-\theta)u_xv_x\right)\mathrm{d}x.$$

Dahlby and Owren (2010) remark that this leads to a stable scheme only if $\theta\geq\frac{1}{2}$.

### *3.2. Preserving multiple first integrals*

The preservation of more than one first integral may often be of interest. For instance, in PDEs which have soliton solutions, a well-known technique for proving stability of the exact solution is based on using the preservation of two specific invariants: see the articles by Benjamin and co-authors (Benjamin 1972, Benjamin, Bona and Mahony 1972). A procedure for preserving multiple first integrals by a numerical integrator for ODEs was described in McLachlan *et al.* (1999). Their method was based on finding a presentation of the ODE vector field $f(y)$ by means of a completely skew-symmetric rank $q+1$ tensor field $S$, writing

$$\dot y=f(y)=S(\cdot,\nabla H_1,\dots,\nabla H_q),\tag{3.23}$$

where $H_1,\dots,H_q$ are $q$ known first integrals. This form effectively encodes the preservation of all the $q$ first integrals, since, for $1\leq i\leq q$,

$$\frac{\mathrm{d}}{\mathrm{d}t}H_i(y)=\nabla H_i(y)\cdot f(y)=S(y)(\nabla H_i(y),\nabla H_1(y),\dots,\nabla H_q(y))=0.$$

Proof of the existence of the formulation (3.23), along with constructive examples, is provided in McLachlan *et al.* (1999). The discrete gradient method is now easily implemented. Given a discrete gradient, $\overline{\nabla}H_i(y^n,y^{n+1})$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

for each of the first integrals, one may, for instance, define the method via

$$\frac{y^{n+1}-y^n}{\Delta t} = \bar{S}(y^n, y^{n+1})(\,\cdot\,, \overline{\nabla} H_1, \ldots, \overline{\nabla} H_q),$$

for a suitable approximation $\bar{S}$ to $S$. Another method for preserving multiple first integrals was presented in Minesaki and Nakamura (2006) for a subclass of Hamiltonian ODEs called Stäckel systems.

In a recent paper by Dahlby, Owren and Yaguchi (2010), the problem of retaining several first integrals simultaneously has been considered. Again we explain this approach for ODEs, or in the case when our PDE has already been discretized in space to yield a system $\dot{y} = f(y)$, $y \in \mathbb{R}^m$, with first integrals $H_1, \ldots, H_q$. Using the Euclidean structure on $\mathbb{R}^m$ one may consider the immersed submanifold $M_c \subset \mathbb{R}^m$, a leaf of the foliation induced by the integrals

$$M = M_c = \{y \in \mathbb{R}^m : H_1(y) = c_1, H_2(y) = c_2, \ldots, H_q(y) = c_q\}.$$

The tangent space to this manifold may be characterized as the orthogonal complement to $\mathrm{span}(\nabla H_1, \ldots, \nabla H_q)$. In order to adapt this setting to the situation with discrete tangents, the *discrete tangent space* was defined relative to two points, sufficiently close to each other in $M \times M$. For all $y$ in a neighbourhood of a point $p \in M_c$, let

$$T_{(p,y)}M = \{\eta \in \mathbb{R}^m : \langle \overline{\nabla} H_j(p,y), \eta \rangle = 0,\ 1 \le j \le q\}$$

for the Euclidean inner product $\langle \cdot, \cdot \rangle$ on $\mathbb{R}^m$. A vector $\eta = \eta_{(p,y)} \in T_{(p,y)}M$ is called a discrete tangent vector. Note that this definition causes $T_{(y,y)}M = T_yM$. The following observation is immediate. Any integrator satisfying

$$y^{n+1} - y^n \in T_{(y^{n+1}, y^n)}M \tag{3.24}$$

preserves all integrals, in the sense that $H_i(y^{n+1}) = H_i(y^n)$, $1 \le i \le q$. The obvious approach to adopt in order to obtain a method satisfying (3.24) is to combine a non-conservative scheme with projection as follows. For simplicity, let $\phi_{\Delta t}$ denote a one-step method of order $p$, $\phi_{\Delta t} : \mathbb{R}^m \to \mathbb{R}^m$. For any $y \in M$, let $y$ also denote its injection into $\mathbb{R}^m$. Then consider

$$u_{n+1} = \phi_{\Delta t}(y_n), \quad y_{n+1} = y_n + \mathcal{P}(y_{n+1}, y_n)(u_{n+1} - y_n), \tag{3.25}$$

where $\mathcal{P}(y_n, y_{n+1})$ is a smooth projection operator onto the discrete tangent space $T_{(y_n, y_{n+1})}M$. An alternative scheme is

$$y_{n+1} = y_n + \Delta t \mathcal{P}(y_n, y_{n+1}) \psi_{\Delta t}(y_{n+1}, y_n), \tag{3.26}$$

where $\psi_{\Delta t}$ is an integrator which can be written in the form

$$y_{n+1} = y_n + \Delta t\, \psi_{\Delta t}(y_{n+1}, y_n).$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

This method is itself assumed to be of order $p$, that is,

$$y(t+\Delta t) - y(t) - h\,\psi_{\Delta t}(y(t+\Delta t), y(t)) = \mathcal{O}(\Delta t^{p+1}). \tag{3.27}$$

The following result was proved in Dahlby *et al.* (2010).

**Theorem 3.4.** If the non-conservative methods $\phi_{\Delta t}, \psi_{\Delta t}$ are of order $p$, then so are the schemes (3.25) and (3.26):

$$y(t+h) - y(t) - \mathcal{P}(y(t), y(t+\Delta t))(u_{n+1} - y(t)) = \mathcal{O}(\Delta t^{p+1}), \tag{3.28}$$

where $u_{n+1} = \phi_\Delta t(y(t))$, and

$$y(t+h) - y(t) - h\mathcal{P}(y(t+h), y(t))\psi_h(y(t+h), y(t)) = \mathcal{O}(h^{p+1}). \tag{3.29}$$

An alternative approach to projection is to derive methods based on a local coordinate representation of the manifold. Based on the assumption that the $q$ first integrals are all independent, we can define a coordinate chart centred at a point $p \in M$ as follows. Suppose $y \in M$ is sufficiently close to $p$ such that all discrete gradients $\overline{\nabla} H_1(p,y), \ldots, \overline{\nabla} H_q(p,y)$ form a linearly independent set of vectors in $\mathbb{R}^m$. Let $T_p(y)$ be a smooth map from the manifold $M$ into the set of orthogonal $m \times m$ matrices $T : M \to O(m)$ such that its last $q$ columns form a basis for the linear span of the discrete gradients. We now define the coordinate map implicitly to be

$$\chi_p : \mathbb{R}^{m-q} \to M : \quad \chi_p(\eta) = y : y - p = T_p(y)\begin{pmatrix} \eta \\ 0_q \end{pmatrix}. \tag{3.30}$$

That these maps induce an atlas on $M$ is proved in Dahlby *et al.* (2010). One may now express the original differential equation in terms of $\eta$:

$$\dot{\eta} = -T^\top(\chi_p \circ \eta) DT_p\big|_{\chi_p \circ \eta}(f(\chi_p \circ \eta))\eta + T^\top(\chi_p \circ \eta) f(\chi_p \circ \eta). \tag{3.31}$$

The method proposed takes one step as follows.

(1) Let $\eta_0 = 0$.

(2) Take a step with any $p$th-order method applied to the ODE (3.31), using $p = y^n$ in (3.30). The result is $\eta_1$.

(3) Compute $y^{n+1} = \chi(\eta_1)$.

The method defined in this way will clearly be of order at least $p$.

Using this approach, one clearly needs to compute $DT_p$ because it is needed explicitly in (3.31), and this map may also be useful in computing the implicitly defined coordinate map. It is shown in Dahlby *et al.* (2010) how $DT_p|_y(v)\eta$ can be computed with a complexity of order $\mathcal{O}(mq^2 + q^3)$.

### *3.3. Discretizing in space*

The adaptation of the approaches just presented to PDEs which have already been discretized in space is of course straightforward: the integral in

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

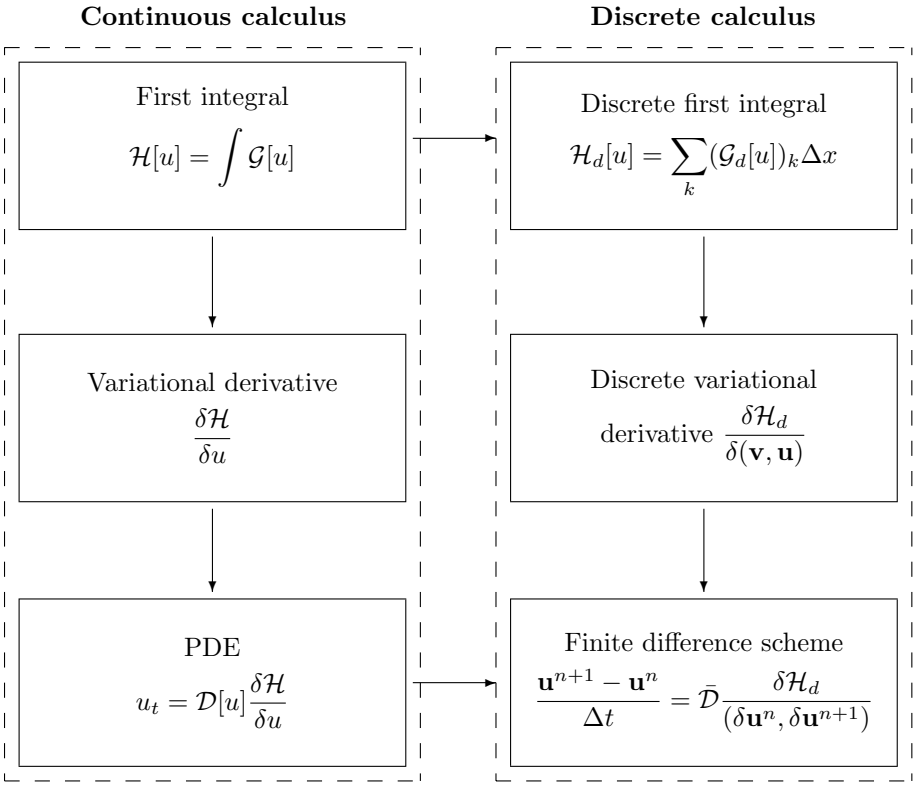



Figure 3.1. The Furihata formalism. This illustration is taken from Furihata (1999).

(3.3) may be replaced by a quadrature rule, resulting in a discrete version, for instance, $\mathcal{H}_d : \mathbb{R}^m \to \mathbb{R}$. The skew-adjoint operator $\mathcal{D}$ must be replaced by a skew-symmetric $m \times m$ matrix, $\mathcal{D}_d$.

We now consider finite difference approximations. The function space to which the solution $u$ belongs is replaced by a finite-dimensional space with functions on a grid indexed by $I_g \subset \mathbb{Z}^d$. We use boldface symbols for these functions. Let there be $N_r$ grid points in the space direction $r$ so that $\mathbf{N} = N_1 \cdots N_d$ is the total number of grid points. We denote by $\mathbf{u}^\alpha$ the approximation to $u^\alpha$ on such a grid, and by $\mathbf{u}$ the vector consisting of $(\mathbf{u}^1, \ldots, \mathbf{u}^m)$. We will replace each derivative $u^\alpha_J$ by a finite difference approximation $\delta_J \mathbf{u}^\alpha$, and replace the integral by a quadrature rule.

We then let

$$\mathcal{H}_d(\mathbf{u}) = \sum_{\mathbf{i} \in I_g} b_{\mathbf{i}} (\mathcal{G}_d((\delta_J \mathbf{u})))_{\mathbf{i}} \, \Delta x. \tag{3.32}$$

Here $\Delta x$ is the volume (length, area) of a grid cell and $\mathbf{b} = (b_{\mathbf{i}})_{\mathbf{i} \in I_g}$ are the

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

weights in the quadrature rule. The discretized $\mathcal{G}_d$ has the same number of arguments as $\mathcal{G}$, and each input argument as well as the output are vectors in $\mathbb{R}^N$. We have here approximated the function $u_J^\alpha$ by a difference approximation $\delta_J \mathbf{u}^\alpha$, where $\delta_J : \mathbb{R}^{\mathbf{N}} \to \mathbb{R}^{\mathbf{N}}$ is a linear map. As in the continuous case, we use square brackets, say $F[\mathbf{u}]$, as shorthand for a list of arguments involving difference operators $F[\mathbf{u}] = F(\mathbf{u}, \delta_{J_1}\mathbf{u}, \dots, \delta_{J_q}\mathbf{u})$. We compute

$$\begin{aligned}\mathcal{H}_d[\mathbf{u}] - \mathcal{H}_d[\mathbf{v}] &= \sum_{\mathbf{i}\in I_g} b_{\mathbf{i}} \sum_{J,\alpha} \int_0^1 \left(\frac{\partial \mathcal{G}_d}{\partial \delta_J \mathbf{u}^\alpha}\right)_{\mathbf{i}} [\xi\mathbf{u} + (1-\xi)\mathbf{v}] \mathrm{d}\xi (\delta_J(\mathbf{u}^\alpha - \mathbf{v}^\alpha))\, \Delta x \\ &= \left\langle \frac{\delta \mathcal{H}_d}{\delta(\mathbf{v},\mathbf{u})}, \mathbf{u} - \mathbf{v} \right\rangle, \end{aligned} \tag{3.33}$$

where

$$\frac{\delta \mathcal{H}_d}{\delta(\mathbf{v},\mathbf{u})} = \sum_{J,\alpha} \delta_J^\top B \left( \int_0^1 \frac{\partial \mathcal{G}_d}{\partial \mathbf{u}_J} [\xi \mathbf{u}^\alpha + (1-\xi)\mathbf{v}^\alpha]\, \mathrm{d}\xi \right),$$

$B$ is the diagonal linear map $B = \mathrm{diag}(b_{\mathbf{i}})$, $\mathbf{i} \in I_g$, and the discrete inner product used in (3.33) is

$$\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{\alpha, \mathbf{i}\in I_g} \mathbf{u}_{\mathbf{i}}^\alpha \mathbf{v}_{\mathbf{i}}^\alpha.$$

Notice the resemblance between the operator acting on $\mathcal{G}_d$ in (3.33) and the continuous Euler operator in (3.6). We make the following assumptions.

(1) The spatially continuous method (3.14) (using (3.15)) is discretized in space, using a skew-symmetric $\mathcal{D}_d$ and a selected set of difference quotients $\delta_J$ for each derivative $\partial_J$.

(2) Considering (3.5) and (3.6), the choice of discretization operators $\delta_J$ used in $\partial \mathcal{G}/\partial u_J^\alpha[u]$ is arbitrary, but the corresponding $D_J$ is replaced by the transpose $\delta_J^\top$.

In this case, using the same $\mathcal{D}_d$, an identical set of difference operators in discretizing $\mathcal{H}$ (3.32), and choosing all the quadrature weights $b_{\mathbf{i}} = 1$, the resulting scheme is the same.

Letting $\mathbf{e}_r$ denote the $r$th canonical unit vector in $\mathbb{R}^d$, we define the most common first-order difference operators

$$\begin{aligned}(\delta_r^+ \mathbf{u})_{\mathbf{i}} &= \frac{\mathbf{u}_{\mathbf{i}+\mathbf{e}_r} - \mathbf{u}_{\mathbf{i}}}{\Delta x_r}, \\ (\delta_r^- \mathbf{u})_{\mathbf{i}} &= \frac{\mathbf{u}_{\mathbf{i}} - \mathbf{u}_{\mathbf{i}-\mathbf{e}_r}}{\Delta x_r}, \\ (\delta_r^{\langle 1\rangle} \mathbf{u})_{\mathbf{i}} &= \frac{\mathbf{u}_{\mathbf{i}+\mathbf{e}_r} - \mathbf{u}_{\mathbf{i}-\mathbf{e}_r}}{2\Delta x_r}. \end{aligned}$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

These difference operators are all commuting, but only the last one is skew-symmetric. However, for the first two we have the useful identities

$$(\delta_r^+)^\top = -\delta_r^-, \quad (\delta_r^-)^\top = -\delta_r^+.$$

Higher-order difference operators $\delta_J$ can generally be defined by taking compositions of these operators: in particular, we shall consider examples in the next section using the second and third derivative approximations

$$\delta_r^{\langle 2\rangle} = \delta_r^+ \circ \delta_r^-, \quad \delta^{\langle 3\rangle} = \delta^{\langle 1\rangle} \circ \delta^{\langle 2\rangle}.$$

We may now introduce numerical approximations $\mathbf{U}^n$ representing the fully discretized system: the scheme is

$$\frac{\mathbf{U}^{n+1} - \mathbf{U}^n}{\Delta t} = \mathcal{D}_d \frac{\delta \mathcal{H}_d}{\delta(\mathbf{U}^n, \mathbf{U}^{n+1})}.$$

The conservative schemes based on polarization are adapted in a straightforward manner, introducing a function $H_d[\mathbf{w_1}, \ldots, \mathbf{w_k}]$ which is consistent and cyclic, and a skew-symmetric map $D_d$ depending on at most $k-1$ arguments. The scheme is then

$$\frac{\mathbf{U}^{n+k} - \mathbf{U}^n}{k\Delta t} = k D_d \frac{\delta H_d}{\delta(\mathbf{U}^n, \ldots, \mathbf{U}^{n+k})}. \tag{3.34}$$

This scheme conserves the function $H_d$ in the sense that

$$H_d[\mathbf{U}^{n+1}, \ldots, \mathbf{U}^{n+k}] = H_d[\mathbf{U}^0, \ldots, \mathbf{U}^{k-1}], \quad n \geq 0.$$

## 4. Spatial symmetries, high-order discretizations and fast group-theoretic algorithms

In this chapter we survey some group-theoretic techniques applied to the discretization and solution of PDEs. Inspired by recent active research in Lie group and exponential time integrators for differential equations, we will in Section 4.1 present algorithms for computing matrix and operator exponentials based on Fourier transforms on finite groups. As a final example, we consider spherically symmetric PDEs, where the discretization preserves the 120 symmetries of the icosahedral group. This motivates the study of spectral element discretizations based on triangular subdivisions.

In Section 4.2 we introduce novel applications of multivariate non-separable Chebyshev polynomials in the construction of spectral element bases on triangular and simplicial subdomains. These generalized Chebyshev polynomials are intimately connected to kaleidoscopes of mirrors acting on a vector space, *i.e.*, groups of isometries generated by mirrors placed on each face of a triangle (2D), a tetrahedron (3D) or in general simplexes. Mathematically this is described by the theory of root systems and Weyl groups,

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

which we will review. Group-theoretic techniques lead to FFT-based fast computational algorithms. This is well known in the case of commutative groups, but it turns out to be important to understanding such ideas in the setting of non-commutative groups as well.

Lie group integrators enjoy a number of nice geometrical properties, of which we will here focus on their *symmetry* and *equivariance* properties. Fundamental to the theory of differential equations is the equivariance of the solution curves with respect to any diffeomorphism $\phi\colon \mathcal{M} \to \mathcal{M}$ acting on the domain. Let $\phi_* F$ denote the *pushforward* of the vector field $F$, *i.e.*, $(\phi_* F)(z) = T\phi \cdot F(\phi^{-1}(z))$ for all $z \in \mathcal{M}$, where $T\phi$ denotes the tangent map (in coordinates, the Jacobian matrix). Then the two differential equations $y'(t) = F(y(t)), y(0) = y_0$ and $z'(t) = (\phi_* F)(z(t)), z(0) = \phi(y_0)$ have analytical solution curves related by $z(t) = \phi(y(t))$. In particular, if $\phi_* F = F$, we say that $\phi$ is a symmetry of the vector field, and in that case $\phi$ maps solution curves to other solution curves of the same equation.

For numerical integrators it is in general impossible to satisfy equivariance with respect to arbitrary diffeomorphisms, since this would imply an analytically correct solution. (There always exists a local diffeomorphism that straightens the flow to a constant flow in the $x_1$ direction, and this is integrated exactly by any reasonable numerical method.)

The *equivariance group* of a numerical scheme is the largest group of diffeomorphisms under which the numerical solutions transform equivariantly. It is known that the equivariance group of classical Runge–Kutta methods is the group of all affine linear transformations of $\mathbb{R}^n$. Lie group integrators based on exact computation of exponentials have equivariance groups that include the Lie group $G$ on which the method is based; hence, if some elements $g \in G$ are symmetries of the differential equation, then $G$-equivariant Lie group integrators will exactly preserve these symmetries. However, if the exact exponential is replaced with approximations, care must be taken not to destroy $G$-equivariance and symmetry preservation of the numerical scheme. In the case of PDEs, symmetry preservation also depends on symmetry-preserving spatial discretizations.

Equivariance is the foundation of Fourier analysis. Classical Fourier analysis can be defined as the study of linear operators $\mathcal{L}$ which are equivariant with respect to an (abelian) group of translations acting on a domain, *i.e.*, $\mathcal{L} \circ \tau = \tau \circ \mathcal{L}$ for all translations $\tau(t)\colon t \mapsto t + \tau$. The fact that exponential functions are the eigenfunctions of translation operators, $\exp(2\pi \mathrm{i}\lambda(t + \tau)) = \exp(2\pi \mathrm{i}\lambda\tau)\exp(2\pi \mathrm{i}\lambda t)$, and hence also eigenfunctions for $\mathcal{L}$, is the explanation for the omnipresence of Fourier bases in computational science. The more specialized cos and sin bases appear naturally when boundary conditions are taken into account, as symmetrization and skew-symmetrization of the exponentials with respect to the reflection $t \mapsto -t$. Chebyshev polynomials again occur naturally through a change of variables

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

$\cos(t) \mapsto y$. The availability of fast computational algorithms (FFT-based) for all these bases again relies in a crucial manner upon the underlying group structures provided by the equivariance. Products of equivariant operators lead naturally to the convolution product, which is diagonalized by Fourier transforms.

Many important operators are equivariant with respect to larger non-commutative groups of transformations, such as the important Laplacian $\nabla^2$ which is equivariant with respect to every isometry of the domain.[4] The generalized Fourier theory of linear operators equivariant with respect to non-commutative groups of transformations is not as widely known as the abelian theory within computational science. Convolution products also occur naturally in this setting. However, due to non-commutativity, exponential bases are no longer sufficient to diagonalize convolutions. Instead one must also use the higher-dimensional *representations* of the group, which can be understood as homomorphisms of the group into a space of (unitary) matrices. Exponentials are one-dimensional representations of abelian groups. In the natural bases obtained from representation theory, equivariant operators are *block-diagonalized*. This is a starting point for constructing fast algorithms for matrix exponentials, eigenvalue computations and linear equations in cases where the operator is equivariant.

We start our presentation with a discussion of matrices equivariant with respect to a finite non-abelian group acting upon the index set. This is a good example for understanding some basic concepts of non-commutative harmonic analysis, without having to consider more technical analysis questions which occur in the case of infinite groups.

### *4.1. Introduction to discrete symmetries and equivariance*

The topic of this subsection is the application of Fourier analysis on groups to the computation of matrix exponentials. Assuming that the domain is discretized with a symmetry-respecting discretization, we will show that by a change of basis derived from the irreducible representations of the group, the operator is block-diagonalized. This simplifies the computation of matrix exponentials. The basic mathematics here is the *representation theory of finite groups* (James and Liebeck 2001, Lomont 1959, Serre 1977). Applications of this theory in scientific computing have been discussed by a number of authors: see, *e.g.*, Allgower, Böhmer, Georg and Miranda (1992), Allgower, Georg, Miranda and Tausch (1998), Bossavit (1986), Douglas and Mandel (1992) and Georg and Miranda (1992). Our exposition, based on the *group algebra*, is explained in detail in Åhlander and Munthe-Kaas (2005).

[4] The Laplacian can indeed be defined as the unique (up to constant) second-order linear differential operator which commutes with all isometries.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

*$\mathcal{G}$-equivariant matrices.* A *group* is a set $\mathcal{G}$ with a binary operation $g, h \mapsto gh$, inverse $g \mapsto g^{-1}$ and identity element $e$, such that $g(ht) = (gh)t$, $eg = ge = g$ and $gg^{-1} = g^{-1}g = e$ for all $g, h, t \in \mathcal{G}$. We let $|\mathcal{G}|$ denote the number of elements in the group. Let $\mathcal{I}$ denote the set of indexes used to enumerate the nodes in the discretization of a computational domain. We say that a group $\mathcal{G}$ *acts on* a set $\mathcal{I}$ (from the right) if there exists a product $(i, g) \mapsto ig : \mathcal{I} \times \mathcal{G} \to \mathcal{I}$ such that

$$ie = i \quad \text{for all } i \in \mathcal{I}, \tag{4.1}$$

$$i(gh) = (ig)h \quad \text{for all } g, h \in \mathcal{G} \text{ and } i \in \mathcal{I}. \tag{4.2}$$

The map $i \mapsto ig$ is a permutation of the set $\mathcal{I}$, with the inverse permutation being $i \mapsto ig^{-1}$. An action partitions $\mathcal{I}$ into disjoint *orbits*

$$\mathcal{O}_i = \{j \in \mathcal{I} \mid j = ig \text{ for some } g \in \mathcal{G}\}, \quad i \in \mathcal{I}.$$

We let $\mathcal{S} \subset \mathcal{I}$ denote a selection of *orbit representatives*, *i.e.*, one element from each orbit. The action is called *transitive* if $\mathcal{I}$ consists of just a single orbit, $|\mathcal{S}| = 1$. For any $i \in \mathcal{I}$ we let the *isotropy subgroup at* $i$, $\mathcal{G}_i$ be defined as

$$\mathcal{G}_i = \{g \in \mathcal{G} \mid ig = i\}.$$

The action is *free* if $\mathcal{G}_i = \{e\}$ for every $i \in \mathcal{I}$, *i.e.*, there are no fixed points under the action of $\mathcal{G}$.

**Definition 4.1.** A matrix $\boldsymbol{A} \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$, is $\mathcal{G}$-equivariant if

$$\boldsymbol{A}_{i,j} = \boldsymbol{A}_{ig,jg} \quad \text{for all } i, j \in \mathcal{I} \text{ and all } g \in \mathcal{G}. \tag{4.3}$$

The definition is motivated by the result that if $\mathcal{L}$ is a linear differential operator commuting with a group of domain symmetries $\mathcal{G}$, and if we can find a set of discretization nodes $\mathcal{I}$ such that every $g \in \mathcal{G}$ acts on $\mathcal{I}$ as a permutation $i \mapsto ig$, then $\mathcal{L}$ can be discretized as a $\mathcal{G}$-equivariant matrix $A$: see Allgower *et al.* (1998) and Bossavit (1986).

*The group algebra.* We will establish that $\mathcal{G}$-equivariant matrices are associated with (scalar or block) convolutional operators in the *group algebra*.

**Definition 4.2.** The *group algebra* $\mathbb{C}\mathcal{G}$ is the complex vector space $\mathbb{C}^{\mathcal{G}}$ where each $g \in \mathcal{G}$ corresponds to a basis vector $\boldsymbol{g} \in \mathbb{C}\mathcal{G}$. A vector $a \in \mathbb{C}\mathcal{G}$ can be written as

$$a = \sum_{g\in\mathcal{G}} a(g)\boldsymbol{g} \quad \text{where } a(g) \in \mathbb{C}.$$

The convolution product $* : \mathbb{C}\mathcal{G} \times \mathbb{C}\mathcal{G} \to \mathbb{C}\mathcal{G}$ is induced from the product in $\mathcal{G}$ as follows. For basis vectors $\boldsymbol{g}, \boldsymbol{h}$, we set $\boldsymbol{g} * \boldsymbol{h} \equiv \boldsymbol{gh}$, and in general,

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

if $a = \sum_{g\in\mathcal{G}} a(g)\boldsymbol{g}$ and $b = \sum_{h\in\mathcal{G}} b(h)\boldsymbol{h}$, then

$$a * b = \left(\sum_{g\in\mathcal{G}} a(g)\boldsymbol{g}\right) * \left(\sum_{h\in\mathcal{G}} b(h)\boldsymbol{h}\right) = \sum_{g,h\in\mathcal{G}} a(g)b(h)(\boldsymbol{gh}) = \sum_{g\in\mathcal{G}} (a*b)(g)\boldsymbol{g},$$

where

$$(a*b)(g) = \sum_{h\in\mathcal{G}} a(gh^{-1})b(h) = \sum_{h\in\mathcal{G}} a(h)b(h^{-1}g). \tag{4.4}$$

Consider a $\mathcal{G}$-equivariant $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ in the case where $\mathcal{G}$ acts freely and transitively on $\mathcal{I}$. In this case there is only one orbit of size $|\mathcal{G}|$ and hence $\mathcal{I}$ may be identified with $\mathcal{G}$. Corresponding to $\boldsymbol{A}$ there is a unique $A \in \mathbb{C}\mathcal{G}$, given as $A = \sum_{g\in\mathcal{G}} A(g)g$, where $A$ is the first column of $\boldsymbol{A}$, *i.e.*,

$$A(gh^{-1}) = \boldsymbol{A}_{gh^{-1},e} = \boldsymbol{A}_{g,h}. \tag{4.5}$$

Similarly, any vector $\boldsymbol{x} \in \mathbb{C}^n$ corresponds uniquely to $x = \sum_{g\in\mathcal{G}} x(g)g \in \mathbb{C}\mathcal{G}$, where $x(g) = \boldsymbol{x}_g$ for all $g \in \mathcal{G}$. Consider the matrix vector product

$$(\boldsymbol{Ax})_g = \sum_{h\in\mathcal{G}} \boldsymbol{A}_{g,h}\boldsymbol{x}_h = \sum_{h\in\mathcal{G}} A(gh^{-1})x(h) = (A*x)(g).$$

If $\boldsymbol{A}$ and $\boldsymbol{B}$ are two equivariant matrices, then $\boldsymbol{AB}$ is the equivariant matrix, where the first column is given as

$$(\boldsymbol{AB})_{g,e} = \sum_{h\in\mathcal{G}} \boldsymbol{A}_{g,h}\boldsymbol{B}_{h,e} = \sum_{h\in\mathcal{G}} A(gh^{-1})B(h) = (A*B)(g).$$

We have shown that *if $\mathcal{G}$ acts freely and transitively, then the algebra of $\mathcal{G}$-equivariant matrices acting on $\mathbb{C}^n$ is isomorphic to the group algebra $\mathbb{C}\mathcal{G}$ acting on itself by convolutions from the left.*

In the case where $\boldsymbol{A}$ is $\mathcal{G}$-equivariant with respect to a free, but not transitive, action of $\mathcal{G}$ on $\mathcal{I}$, we need a block version of the above theory. Let $\mathbb{C}^{m\times\ell}\mathcal{G} \equiv \mathbb{C}^{m\times\ell}\otimes\mathbb{C}\mathcal{G}$ denote the space of vectors consisting of $|\mathcal{G}|$ matrix blocks, each block of size $m \times \ell$. Thus $A \in \mathbb{C}^{m\times\ell}\mathcal{G}$ can be written as

$$A = \sum_{g\in\mathcal{G}} A(g)\otimes g \quad \text{where } A(g) \in \mathbb{C}^{m\times\ell}. \tag{4.6}$$

The convolution product (4.4) generalizes to a block convolution $* : \mathbb{C}^{m\times\ell}\mathcal{G} \times \mathbb{C}^{\ell\times k}\mathcal{G} \to \mathbb{C}^{m\times k}\mathcal{G}$ given as

$$A * B = \left(\sum_{g\in\mathcal{G}} A(g)\otimes g\right) * \left(\sum_{h\in\mathcal{G}} B(h)\otimes h\right) = \sum_{g\in\mathcal{G}} (A*B)(g)\otimes g,$$

where

$$(A*B)(g) = \sum_{h\in\mathcal{G}} A(gh^{-1})B(h) = \sum_{h\in\mathcal{G}} A(h)B(h^{-1}g), \tag{4.7}$$

and $A(h)B(h^{-1}g)$ denotes a matrix product.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

If the action of $\mathcal{G}$ on $\mathcal{I}$ is free, but not transitive, then $\mathcal{I}$ splits into $m$ orbits, each of size $|\mathcal{G}|$. We let $\mathcal{S}$ denote a selection of one representative from each orbit. We will establish an isomorphism between the algebra of $\mathcal{G}$-equivariant matrices acting on $\mathbb{C}^n$ and the block-convolution algebra $\mathbb{C}^{m\times m}\mathcal{G}$ acting on $\mathbb{C}^m\mathcal{G}$. We define the mappings $\mu : \mathbb{C}^n \to \mathbb{C}^m\mathcal{G}$, $\nu : \mathbb{C}^{n\times n} \to \mathbb{C}^{m\times m}\mathcal{G}$ by

$$\mu(\boldsymbol{y})_i(g) = y_i(g) = \boldsymbol{y}_{ig} \ \forall i \in \mathcal{S},\ g \in \mathcal{G}, \tag{4.8}$$

$$\nu(\boldsymbol{A})_{i,j}(g) = A_{i,j}(g) = \boldsymbol{A}_{ig,j} \ \forall i, j \in \mathcal{S}\ g \in \mathcal{G}. \tag{4.9}$$

In Åhlander and Munthe-Kaas (2005) we show the following.

**Proposition 4.3.** Let $\mathcal{G}$ act freely on $\mathcal{I}$. Then $\mu$ is invertible and $\nu$ is invertible on the subspace of $\mathcal{G}$-equivariant matrices. Furthermore, if $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n\times n}$ are $\mathcal{G}$-equivariant, and $\boldsymbol{y} \in \mathbb{C}^n$, then

$$\mu(\boldsymbol{A}\boldsymbol{y}) = \nu(\boldsymbol{A}) * \mu(\boldsymbol{y}), \tag{4.10}$$

$$\nu(\boldsymbol{A}\boldsymbol{B}) = \nu(\boldsymbol{A}) * \nu(\boldsymbol{B}). \tag{4.11}$$

To complete the connection between $\mathcal{G}$-equivariance and block convolutions, we need to address the general case where the action is not free, and hence some of the orbits in $\mathcal{I}$ have reduced size. One way to treat this case is to duplicate the nodes with non-trivial isotropy subgroups; thus a point $j \in \mathcal{I}$ is considered to be $|\mathcal{G}_j|$ identical points, and the action is extended to a free action on this extended space. Equivariant matrices on the original space is extended by duplicating the matrix entries, and scaled according to the size of the isotropy. We define

$$\mu(\boldsymbol{x})_i(g) = x_i(g) = \boldsymbol{x}_{ig} \ \forall i \in \mathcal{S},\ g \in \mathcal{G}, \tag{4.12}$$

$$\nu(\boldsymbol{A})_{i,j}(g) = A_{i,j}(g) = \frac{1}{|\mathcal{G}_j|}\boldsymbol{A}_{ig,j} \ \forall i, j \in \mathcal{S}\ g \in \mathcal{G}. \tag{4.13}$$

With these definitions it can be shown that (4.10)–(4.11) still hold. It should be noted that $\mu$ and $\nu$ are no longer invertible, and the extended block convolutional operator $\nu(\boldsymbol{A})$ becomes singular. This poses no problems for the computation of exponentials since this is a forward computation. Thus we just exponentiate the block convolutional operator and restrict the result back to the original space. However, for inverse computations such as solving linear systems, the characterization of the image of $\mu$ and $\nu$ as subspaces of $\mathbb{C}^m\mathcal{G}$ and $\mathbb{C}^{m\times m}\mathcal{G}$ is an important issue for finding the correct solution (Åhlander and Munthe-Kaas 2005, Allgower, Georg and Miranda 1993).

*The generalized Fourier transform* (GFT). So far we have argued that a symmetric differential operator becomes a $\mathcal{G}$-equivariant matrix under discretization, which again can be represented as a block convolutional operator. In this subsection we will show how convolutional operators are

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

block-diagonalized by a Fourier transform on $\mathcal{G}$. This is the central part of Frobenius's theory of group representations from 1897–1899. We recommend the monographs by Fässler and Stiefel (1992), James and Liebeck (2001), Lomont (1959) and Serre (1977) as introductions to representation theory with applications.

**Definition 4.4.** A $d$-dimensional group representation is a map $R : \mathcal{G} \to \mathbb{C}^{d\times d}$ such that

$$R(gh) = R(g)R(h) \quad \text{for all } g, h \in \mathcal{G}. \tag{4.14}$$

Generalizing the definition of *Fourier coefficients* we define for any $A \in \mathbb{C}^{m\times k}\mathcal{G}$ and any $d$-dimensional representation $R$ a matrix $\hat{A}(R) \in \mathbb{C}^{m\times k} \otimes \mathbb{C}^{d\times d}$ as

$$\hat{A}(R) = \sum_{g\in\mathcal{G}} A(g) \otimes R(g). \tag{4.15}$$

**Proposition 4.5. (Convolution theorem)** For any $A \in \mathbb{C}^{m\times k}\mathcal{G}$, $B \in \mathbb{C}^{k\times \ell}\mathcal{G}$ and any representation $R$ we have

$$\widehat{(A * B)}(R) = \hat{A}(R)\hat{B}(R). \tag{4.16}$$

*Proof.* The statement follows from

$$\begin{aligned}
\hat{A}(R)\hat{B}(R) &= \left(\sum_{g\in\mathcal{G}} A(g) \otimes R(g)\right)\left(\sum_{h\in\mathcal{G}} B(h) \otimes R(h)\right) \\
&= \sum_{g,h\in\mathcal{G}} A(g)B(h) \otimes R(g)R(h) = \sum_{g,h\in\mathcal{G}} A(g)B(h) \otimes R(gh) \\
&= \sum_{g,h\in\mathcal{G}} A(gh^{-1})B(h) \otimes R(g) = \widehat{(A * B)}(R). \qquad \square
\end{aligned}$$

Let $d_R$ denote the dimension of the representation. For use in practical computations, it is important that $A * B$ can be recovered by knowing $\widehat{(A * B)}(R)$ for a suitable selection of representations, and furthermore that their dimensions $d_R$ are as small as possible. Note that if $R$ is a representation and $X \in \mathbb{C}^{d_R\times d_R}$ is non-singular, then also $\tilde{R}(g) = XR(g)X^{-1}$ is a representation. We say that $R$ and $\tilde{R}$ are equivalent representations. If there exists a similarity transform $\tilde{R}(g) = XR(g)X^{-1}$ such that $\tilde{R}(g)$ has a block diagonal structure, independent of $g \in \mathcal{G}$, then $R$ is called *reducible*, otherwise it is *irreducible*.

**Theorem 4.6. (Frobenius)** For any finite group $\mathcal{G}$ there exists a complete list $\mathcal{R}$ of non-equivalent irreducible representations such that

$$\sum_{R\in\mathcal{R}} d_R^2 = |\mathcal{G}|.$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

Defining the GFT for $a \in \mathbb{C}\mathcal{G}$ as

$$\hat{a}(R) = \sum_{g \in \mathcal{G}} a(g) R(g) \quad \text{for every } R \in \mathcal{R}, \tag{4.17}$$

we may recover $a$ by the inverse GFT (IGFT):

$$a(g) = \frac{1}{|\mathcal{G}|} \sum_{R \in \mathcal{R}} d_R \mathrm{trace}\,(R(g^{-1})\hat{a}(R)). \tag{4.18}$$

For the block transform of $A \in \mathbb{C}^{m \times k}\mathcal{G}$ given in (4.15), the GFT and the IGFT are given componentwise as

$$\hat{A}_{i,j}(R) = \sum_{g \in \mathcal{G}} A_{i,j}(g) R(g) \in \mathbb{C}^{d_R \times d_R}, \tag{4.19}$$

$$A_{i,j}(g) = \frac{1}{|\mathcal{G}|} \sum_{R \in \mathcal{R}} d_R \mathrm{trace}\,(R(g^{-1})\hat{A}_{i,j}(R)). \tag{4.20}$$

Complete lists of irreducible representations for several common groups are to be found in Lomont (1959).

*Applications to the matrix exponential.* We have seen that via the GFT, any $\mathcal{G}$-equivariant matrix is block-diagonalized. Corresponding to an irreducible representation $R$, we obtain a matrix block $\hat{A}(R)$ of size $md_R \times md_R$, where $m$ is the number of orbits in $\mathcal{I}$ and $d_R$ is the size of the representation. Let $W_{\text{direct}}$ denote the computational work, in terms of floating-point operations, for computing the matrix exponential on the original data $A$, and let $W_{\text{fspace}}$ be the cost of doing the same algorithm on the corresponding block diagonal GFT-based data $\hat{A}$. Thus

$$W_{\text{direct}} = c(m|\mathcal{G}|)^3 = cm^3 \Big( \sum_{R \in \mathcal{R}} d_R^2 \Big)^3, \quad W_{\text{fspace}} = cm^3 \sum_{R \in \mathcal{R}} d_R^3$$

and the ratio becomes

$$\mathcal{O}(n^3): \quad W_{\text{direct}}/W_{\text{fspace}} = \Big( \sum_{R \in \mathcal{R}} d_R^2 \Big)^3 / \sum_{R \in \mathcal{R}} d_R^3.$$

Table 4.1 lists this factor for the symmetries of the triangle, the tetrahedron, the 3D cube and the maximally symmetric discretization of a 3D sphere (icosahedral symmetry with reflections).

The cost of computing the GFT is not taken into account in this estimate. There exist fast GFT algorithms of complexity $\mathcal{O}\big(|\mathcal{G}| \log^{\ell}(|\mathcal{G}|)\big)$ for a number of groups, but even if we use a slow transform of complexity $\mathcal{O}\big(|\mathcal{G}^2|\big)$, the total cost of the GFT becomes just $\mathcal{O}\big(m^2|\mathcal{G}|^2\big)$, which is much less than $W_{\text{fspace}}$.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

Table 4.1. Gain in computational complexity for matrix exponential via GFT.

| Domain | $\mathcal{G}$ | $\lvert\mathcal{G}\rvert$ | $\{d_R\}_{R\in\mathcal{R}}$ | $W_{\text{direct}}/W_{\text{fspace}}$ |
|---|---|---|---|---|
| Triangle | $\mathcal{D}_3$ | 6 | $\{1,1,2\}$ | 21.6 |
| Tetrahedron | $\mathcal{S}_4$ | 24 | $\{1,1,2,3,3\}$ | 216 |
| Cube | $\mathcal{S}_4\times\mathcal{C}_2$ | 48 | $\{1,1,1,1,2,2,3,3,3,3\}$ | 864 |
| Icosahedron | $\mathcal{A}_5\times\mathcal{C}_2$ | 120 | $\{1,1,3,3,3,3,4,4,5,5\}$ | 3541 |

**Example 4.7. (Equilateral triangle)** The non-commutative group of smallest order is $\mathcal{D}_3$, the symmetries of an equilateral triangle. There are six linear transformations that map the triangle onto itself: three pure rotations and three rotations combined with reflections. In Figure 4.1(a) we indicate the two generators $\alpha$ (rotation 120° clockwise) and $\beta$ (right–left reflection). These satisfy the algebraic relations $\alpha^3=\beta^2=e$, $\beta\alpha\beta=\alpha^{-1}$, where $e$ denotes the identity transform. The whole group is $\mathcal{D}_3=\{e,\alpha,\alpha^2,\beta,\alpha\beta,\alpha^2\beta\}$.

Given an elliptic operator $\mathcal{L}$ on the triangle such that $\mathcal{L}(u\circ\alpha)=\mathcal{L}(u)\circ\alpha$ and $\mathcal{L}(u\circ\beta)=\mathcal{L}(u)\circ\beta$ for any $u$ satisfying the appropriate boundary conditions on the triangle, let the domain be discretized with a *symmetry-respecting discretization*: see Figure 4.1(b). In this example we consider a finite difference discretization represented by the nodes $\mathcal{I}=\{1,2,\dots,10\}$, such that both $\alpha$ and $\beta$ map nodes to nodes. In finite element discretizations one would use basis functions mapped to other basis functions by the symmetries. We define the action of $\mathcal{D}_3$ on $\mathcal{I}$ as

$$(1,2,3,4,5,6,7,8,9,10)\alpha=(5,6,1,2,3,4,9,7,8,10),$$
$$(1,2,3,4,5,6,7,8,9,10)\beta=(2,1,6,5,4,3,7,9,8,10),$$

and extend to all of $\mathcal{D}_3$ using (4.2). As orbit representatives, we may pick $\mathcal{S}=\{1,7,10\}$. The action of the symmetry group is free on the orbit $\mathcal{O}_1=\{1,2,3,4,5,6\}$, while the points in the orbit $\mathcal{O}_7=\{7,8,9\}$ have isotropy subgroups of size 2, and finally $\mathcal{O}_{10}=\{10\}$ has isotropy of size 6.

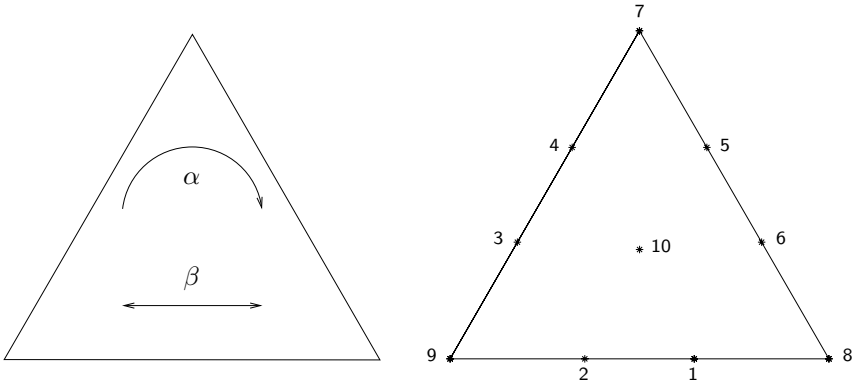
The operator $\mathcal{L}$ is discretized as a matrix $\boldsymbol{A}\in\mathbb{C}^{10\times10}$ satisfying the equivariances $\boldsymbol{A}_{ig,jg}=\boldsymbol{A}_{i,j}$ for $g\in\{\alpha,\beta\}$ and $i,j\in\mathcal{S}$. Thus we have, *e.g.*, $\boldsymbol{A}_{1,6}=\boldsymbol{A}_{3,2}=\boldsymbol{A}_{5,4}=\boldsymbol{A}_{4,5}=\boldsymbol{A}_{2,3}=\boldsymbol{A}_{6,1}$.

$\mathcal{D}_3$ has three irreducible representations given in Table 4.2 (extended to the whole group using (4.14)). To compute $\exp(\boldsymbol{A})$, we find $A=\nu(\boldsymbol{A})\in\mathbb{C}^{3\times3}\mathcal{G}$ from (4.13) and find $\hat{A}=\mathrm{GFT}(A)$ from (4.19). The transformed matrix $\hat{A}$ has three blocks, $\hat{A}(\rho_0),\hat{A}(\rho_1)\in\mathbb{C}^{m\times m}$ and $\hat{A}(\rho_2)\in\mathbb{C}^{m\times m}\otimes\mathbb{C}^{2\times2}\simeq\mathbb{C}^{2m\times2m}$, where $m=3$ is the number of orbits. We exponentiate each of these blocks, and find the components of $\exp(\boldsymbol{A})$ using the inverse GFT (4.20).

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

Table 4.2. A complete list of irreducible representations for $\mathcal{D}_3$.

| | $\alpha$ | $\beta$ |
|---|---|---|
| $\rho_0$ | $1$ | $1$ |
| $\rho_1$ | $1$ | $-1$ |
| $\rho_2$ | $\begin{pmatrix} -1/2 & -\sqrt{3}/2 \\ \sqrt{3}/2 & -1/2 \end{pmatrix}$ | $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ |




(a) (b)

Figure 4.1. Equilateral triangle with a symmetry-preserving set of 10 nodes.

We should remark that in Lie group integrators, it is usually more important to compute $y = \exp(A)\cdot x$ for some vector $x$. In this case, we compute $\hat{y}(\rho_i) = \exp(\hat{A}(\rho_i))\cdot\hat{x}(\rho_i)$, and recover $y$ by inverse GFT. Note that $\hat{x}(\rho_2), \hat{y}(\rho_2) \in \mathbb{C}^m \otimes \mathbb{C}^{2\times 2} \simeq \mathbb{C}^{2m\times 2}$.

**Example 4.8. (Icosahedral symmetry)** As a second example illustrating the general theory, we solve the simple heat equation

$$u_t = \nabla^2 u$$

on the surface of a unit sphere. See Trønnes (2005) for details of the implementation.



The sphere is divided into 20 equilateral triangles, and each triangle is subdivided into a finite difference mesh respecting all the 120 symmetries of the full icosahedral symmetry group (including reflections). To understand this group, it is useful to realize that five tetrahedra can be simultaneously embedded in the icosahedron, so that the 20 triangles correspond to the (in total) 20 corners of these five tetrahedra. From this one sees that the

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

icosahedral rotation group is isomorphic to $A_5$, the group of all 60 even permutations of the five tetrahedra. The 3D reflection matrix $-I$ obviously commutes with any 3D rotation, and hence we realize that the full icosahedral group is isomorphic to the direct product $C_2 \times A_5$, where $C_2 = \{1, -1\}$. The irreducible representations of $A_5$ have dimensions $\{1, 3, 3, 4, 5\}$, and the representations of the full icosahedral group are found by taking tensor products of these with the two one-dimensional representations of $C_2$. The fact that the full icosahedral group is a direct product is also utilized in faster computation of the GFT. This, however, is not of major importance, since the cost of the GFT in any case is much less than the cost of the matrix exponential.

### *4.2. Spatial symmetries and high-order discretizations*

In this subsection we will discuss novel techniques for spatial discretizations based on reflection groups and multivariate Chebyshev polynomials. Given high-order time integration methods, such as Lie group methods, it is desirable to also use high-order discretizations in space. We will see that it is important to employ discretizations respecting spatial symmetries, both for the quality of the discretization error and also for efficiency of linear algebra computations such as matrix exponentials, eigenvalue computations and solution of linear systems.

Spectral methods are important discretization techniques, where special function systems are used to expand the solution on simple (*e.g.*, box-shaped) domains. For PDEs with analytic solutions spectral discretization techniques enjoy exponentially fast convergence (Canuto, Hussaini, Quarteroni and Zang 2006). On more general domains *spectral element* methods are constructed by patching together simple subdomains. Typically the basis functions on the subdomains arise as eigenfunctions for a self-adjoint linear operator, such as the exponential Fourier basis on periodic domains and orthogonal polynomials (Legendre or Chebyshev) on bounded domains with homogeneous boundary conditions. Most of the popular orthogonal polynomial systems used in current spectral (and spectral-element) methods are special cases of Jacobi polynomials, which are the solutions of singular Sturm–Liouville problems on a bounded domain, *i.e.*, they are eigenfunctions of a second-order linear differential operator which is self-adjoint with respect to a weighted inner product, where the weight function is singular on the boundary.

A problem with standard polynomial-based spectral element methods is a lack of flexibility with respect to geometries. On box-shaped domains it is easy to construct polynomial bases by tensor products of univariate polynomials. Box-shaped domains are, however, difficult to patch together in more general geometries, and difficult to match with domain symmetries,

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

*e.g.*, for problems on spherical surfaces. For these reasons, it is desirable to explore high-order function systems based on triangles (2D), tetrahedra (3D) and in general simplexes for higher dimensions.

#### *4.2.1. Root systems and Laplacian eigenfunctions on triangles and simplexes*

For certain triangular or simplicial domains with homogeneous Dirichlet or Neumann boundary conditions, the eigenfunctions of the Laplacian are known explicitly. We will see later that these systems are closely related to reflection groups (crystallographic groups and Coxeter groups) and root systems. In order to motivate a discussion of reflection groups, we start with an important particular example, the eigenfunctions of the Laplacian on an equilateral triangle $\triangle$.

*What is the sound of an equilateral drum?* A detailed discussion of this example is also found in Huybrechs, Iserles and Nørsett (2010). Without loss of generality, we assume that the triangle has corners in the origin $[0;0]$, $\lambda_1 = [1/\sqrt{2}; 1/\sqrt{6}]$ and $\lambda_2 = [0; \sqrt{2}/\sqrt{3}]$, shown as the shaded domain in Figure 4.4(a), labelled $A_2$ (see p. 50). Let $\{s_j\}_{j=1}^3$ denote reflections of $\mathbb{R}^2$ about the edges of the triangle. Let $\tilde{W}$ denote the full group of isometries of $\mathbb{R}^2$ generated by $\{s_j\}_{j=1}^3$. This is an example of a *crystallographic group*,[5] a group of isometries of $\mathbb{R}^d$ where the subgroup of translations form a *lattice* in $\mathbb{R}^d$. From this fact we will derive the Laplacian eigenfunctions on $\triangle$.

Recall that a *lattice* $L$ in $\mathbb{R}^d$ is the $\mathbb{Z}$-linear span of $d$ linearly independent vectors in $\mathbb{R}^d$, $L = \operatorname{span}_{\mathbb{Z}}\{\alpha_1, \ldots, \alpha_d\}$; thus $L$ is an abelian group isomorphic to $\mathbb{Z}^d$. The *reciprocal lattice*

$$L^* = \{\lambda \in \mathbb{R}^d \colon (\lambda, \alpha) \in \mathbb{Z} \text{ for all } \lambda \in \Lambda \text{ and } \alpha \in L\},$$

where $(\cdot, \cdot)$ is the standard inner product on $\mathbb{R}^d$. We have

$$L^* = \operatorname{span}_{\mathbb{Z}}\{\lambda_1, \ldots, \lambda_d\},$$

where $\{\lambda_1, \ldots, \lambda_d\}$ is the dual basis of $\{\alpha_1, \ldots, \alpha_d\}$, *i.e.*, $(\lambda_j, \alpha_k) = \delta_{j,k}$. The reciprocal lattice serves as the index set for the Fourier basis for $L$-periodic functions on $\mathbb{R}^d$. Consider the $L$-periodic domain (a $d$-torus) $\mathbf{T}^d = \mathbb{R}^d/L$ and the space of square-integrable periodic functions $\mathrm{L}^2(\mathbf{T}^d)$. The Fourier basis for $\mathrm{L}^2(\mathbf{T}^d)$ is the $\mathrm{L}^2$-orthogonal family of exponential functions indexed by the dual lattice, $\{\exp(2\pi \mathrm{i}(\lambda, y))\}_{\lambda \in L^*}$.

Back to our equilateral triangle, where the translation lattice of $\tilde{W}$ is the lattice

$$L = \operatorname{span}_{\mathbb{Z}}\{\alpha_1, \alpha_2\} < \mathbb{R}^2$$

generated by the vectors $\alpha_1 = (\sqrt{2}, 0)$ and $\alpha_2 = (-1\sqrt{2}, \sqrt{3}/\sqrt{2})$. The unit

[5] More specifically it is an *affine Weyl group*, to be defined below.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

cell of $L$ can be taken as either the rhombus spanned by $\alpha_1$ and $\alpha_2$ or the hexagon $\hexagon$ indicated in Figure 4.4. The hexagon is the *Voronoi cell* of the origin in the lattice $L$, *i.e.*, its interior consists of the points in $\mathbb{R}^2$ that are closer to the origin than to any other lattice points. As a first step in our construction, we consider the $L$-periodic eigenfunctions of the Laplacian $\nabla^2$. Since $(\lambda_j, \alpha_k) = \delta_{j,k}$, the dual lattice is $L^* = \mathrm{span}_{\mathbb{Z}}\{\lambda_1, \lambda_2\}$ and the periodic eigenfunctions are

$$\nabla_t^2 \mathrm{e}^{2\pi \mathrm{i}(\lambda,t)} = -(2\pi)^2 \|\lambda\|^2 \mathrm{e}^{2\pi \mathrm{i}(\lambda,t)} \quad \text{for all } \lambda \in L^*,\ t \in \mathbb{R}^2/L.$$

We continue to find the Laplacian eigenfunctions on $\triangle$ by folding the exponentials. Let $W < \tilde{W}$ be the linear subgroup which leaves the origin fixed (the symmetries of $\hexagon$):

$$W = \{e, s_1, s_2, s_1 s_2, s_2 s_1, s_1 s_2 s_1\},$$

where $e$ is the identity and $s_i$, $i \in \{1, 2\}$ act on $v \in \mathbb{R}^2$ as $s_i v = v - 2(\alpha_i, v)/(\alpha_i, \alpha_i)$. We define even and odd (cosine- and sine-type) foldings of the exponentials

$$c_\lambda(t) = \frac{1}{|W|} \sum_{w \in W} \mathrm{e}^{2\pi \mathrm{i}(\lambda, wt)} = \frac{1}{|W|} \sum_{w \in W} \mathrm{e}^{2\pi \mathrm{i}(w^\top \lambda, t)}, \tag{4.21}$$

$$s_\lambda(t) = \frac{1}{|W|} \sum_{w \in W} \det(w) \mathrm{e}^{2\pi \mathrm{i}(\lambda, wt)} = \frac{1}{|W|} \sum_{w \in W} \det(w) \mathrm{e}^{2\pi \mathrm{i}(w^\top \lambda, t)}, \tag{4.22}$$

where, in our example, $|W| = 6$. The Laplacian commutes with any isometry, in particular $\nabla^2(f \circ w) = (\nabla^2 f) \circ w$ for $w \in W$. Hence, reflected exponentials and the functions $c_\lambda(t)$ and $s_\lambda(t)$ are also eigenfunctions with the same eigenvalue.

**Lemma 4.9.** The eigenfunctions of $\nabla^2$ on the equilateral triangle $\triangle$ with Dirichlet ($f = 0$) and Neumann boundary conditions ($\nabla f \cdot \mathbf{n} = 0$) are given by

$$\nabla_t^2 c_\lambda(t) = -(2\pi)^2 \|\lambda\|^2 c_\lambda(t),$$
$$\nabla_t^2 s_\lambda(t) = -(2\pi)^2 \|\lambda\|^2 s_\lambda(t),$$

respectively, for $\lambda \in \mathrm{span}_{\mathbb{N}}\{\lambda_1, \lambda_2\}$, the set of all non-negative integer combinations of $\lambda_1 = [1/2; 1/6]$ and $\lambda_2 = [0; 2/3]$.

The set $\mathrm{span}_{\mathbb{N}}\{\lambda_1, \lambda_2\} \subset L^*$ contains exactly one point from each $W$-orbit: see the discussion of Weyl chambers below. An important question is how good (or bad!) these eigenfunctions are as bases for approximating analytic functions on $\triangle$. It is well known from the univariate case that a similar construction, yielding the eigenfunctions $\cos(k\theta)$ and $\sin(k\theta)$ on $[0, \pi]$, does *not* give fast convergence. For an analytic function $f(\theta)$ where $f(0) = f(\pi) = 0$, the even $2\pi$-periodic extension is piecewise smooth, with only $C^0$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

continuity at $\theta \in \{0, \pi\}$. Hence the Fourier cosine series converges only as $\mathcal{O}(k^{-2})$. There are several different ways to achieve spectral convergence $\mathcal{O}(\exp(-ck))$. One possible solution is to approximate $f(\theta)$ in a *frame* (not linearly independent) consisting of both $\{\cos(k\theta)\}_{k\in\mathbb{Z}}$ and $\{\sin(k\theta)\}_{k\in\mathbb{Z}^+}$. Another possibility is to employ a change of variables $x = \cos(\theta)$, yielding (univariate) Chebyshev polynomials of the first and second kind:

$$T_k(x) = \cos(k\theta),$$
$$U_k(x) = \sin((k+1)\theta)/\sin(\theta).$$

These polynomials (in particular the first kind) are ubiquitous in approximation theory and famous for their excellent approximation properties. We will develop the corresponding multivariate theory and see that we have similar possibilities for constructing spectrally converging frames and bases.

*Root systems and affine Weyl groups.* Let $\mathcal{V}$ be a finite-dimensional real Euclidean vector space with standard inner product $(\cdot, \cdot)$. The construction above can be generalized to all those simplexes $\triangle \subset \mathcal{V}$ with the property that the group of isometries $\tilde{W}$ generated by reflecting $\triangle$ about its faces is a crystallography group. All such simplexes are determined by a *root system*, a set of vectors in $\mathcal{V}$ which are perpendicular to the reflection planes of $W$ passing through the origin. We review some basic definitions and results about root systems. For more details we refer to Bump (2004).

**Definition 4.10.** A *root system* in $\mathcal{V}$ is a finite set $\Phi$ of non-zero vectors, called roots, that satisfy the following conditions.

(1) The roots span $\mathcal{V}$.

(2) The only scalar multiples of a root $\alpha \in \Phi$ that belong to $\Phi$ are $\pm\alpha$.

(3) For every root $\alpha \in \Phi$, the set $\Phi$ is invariant under reflection in the hyperplane perpendicular to $\alpha$, *i.e.*, for any two roots $\alpha$ and $\beta$, the set $\Phi$ contains the reflection of $\beta$,

$$s_\alpha(\beta) := \beta - 2\frac{(\alpha, \beta)}{(\alpha, \alpha)}\alpha \in \Phi.$$

(4) (*Crystallographic restriction.*) For any $\alpha, \beta \in \Phi$, we have

$$2\frac{(\alpha, \beta)}{(\alpha, \alpha)} \in \mathbb{Z}.$$

Condition (4) implies that the obtuse angle between two different reflection planes must be either 90°, 120°, 135° or 150°. This is a fundamental fact in crystallography, implying that rotational symmetries of a crystal must be either 2-fold, 3-fold, 4-fold or 6-fold.

The *rank* of the root system is the dimension $d$ of the space $\mathcal{V}$. Any root system contains a subset (not uniquely defined) $\Sigma \subset \Phi$ of so-called *simple*

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

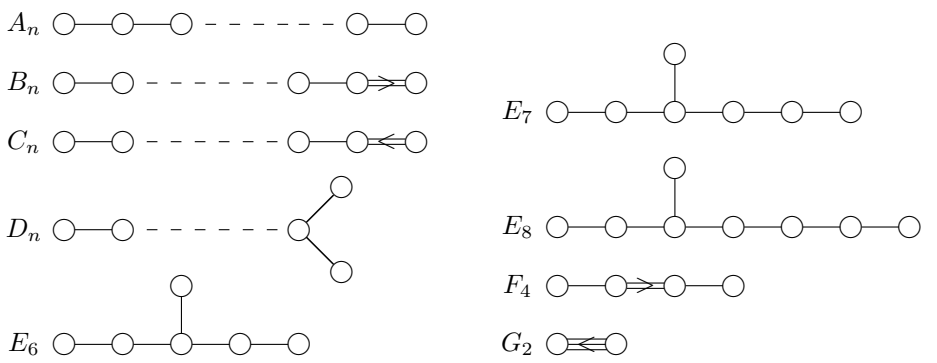



Figure 4.2. Dynkin diagrams for irreducible root systems.

*positive roots*. This is a set of $d$ linearly independent roots $\Sigma = \{\alpha_1, \ldots, \alpha_d\}$ such that any root $\beta \in \Phi$ can be written either as a linear combination of $\alpha_j$ with non-negative integer coefficients, or as a linear combination with non-positive integer coefficients. We call $\Sigma$ a *basis* of the root system $\Phi$. A root system is conveniently represented by its *Dynkin diagram*. This is a graph with $d$ nodes corresponding to the simple positive roots. Between two nodes $j$ and $k$, no line is drawn if the angle between $\alpha_j$ and $\alpha_k$ is 90°, a single line is drawn if it is 120°, a double line for 135° and a triple line for 150°. It is only necessary to understand the geometry of *irreducible root systems* when the Dynkin diagram is connected. Disconnected Dynkin diagrams (reducible root systems) are trivially understood in terms of products of irreducible root systems. For irreducible root systems, the roots are either all of the same length, or have just two different lengths. In the latter case a marker $<$ or $>$ on an edge indicates the separation of long and short roots (short $<$ long). Since the work of W. Killing and E. Cartan in the late nineteenth century it has been known that Dynkin diagrams of irreducible root systems must belong to one of four possible infinite cases, $A_n$ $(n > 0)$, $B_n$ $(n > 1)$, $C_n$ $(n > 2)$, $D_n$ $(n > 3)$, or five special cases, $E_6$, $E_7$, $E_8$, $F_4$, $G_2$, shown in Figure 4.2. We say that two root systems are equivalent if they differ only by a scaling or an isometry. Up to equivalence, there corresponds a unique root system to each Dynkin diagram. We will see that the root system and also the Weyl group can be explicitly computed (as real vectors and matrices) using from the Cartan matrix defined below.

A root system $\Phi$ is associated with a dual root system $\Phi^\vee$ defined such that a root $\alpha \in \Phi$ corresponds to a co-root $\alpha^\vee := 2\alpha/(\alpha, \alpha) \in \Phi^\vee$. If all roots have equal lengths then $\Phi^\vee = \Phi$ (up to equivalence), *i.e.*, the root system is *self-dual*. For the cases with two root lengths we have $B_n^\vee = C_n$, $F_4^\vee = F_4$ and $G_2^\vee = G_2$. The *Cartan matrix* is defined as $C_{j,k} = (\alpha_j, \alpha_k^\vee) = 2(\alpha_j, \alpha_k)/(\alpha_k, \alpha_k)$. Its diagonal is $C_{j,j} = 2$. Off-diagonal entries $C_{j,k}$ are found from the number of edges between $\alpha_j$ and $\alpha_k$ and their

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

relative lengths:

$$C_{j,k} = \begin{cases} 0 & \text{if no edge,} \\ -1 & \text{if single edge, or multiple edge where } \|\alpha_j\| < \|\alpha_k\|, \\ -2 & \text{if double edge and } \|\alpha_j\| > \|\alpha_k\|, \\ -3 & \text{if triple edge and } \|\alpha_j\| > \|\alpha_k\|. \end{cases}$$

The ratio of long and short roots is as follows. If $\|\alpha_j\| > \|\alpha_k\|$ then $\|\alpha_j\| = \sqrt{2}\|\alpha_k\|$ in the cases where the division $\nrightarrow$ is a double edge and $\|\alpha_j\| = \sqrt{3}\|\alpha_k\|$ when it is triple $\not\equiv$. It is convenient to choose normalization such that the longest roots have length $\|\alpha_j\| = \sqrt{2}$. This leads to $\alpha_j^\vee = \alpha_j$ for the long roots. For short roots this implies $\alpha_k^\vee = 2\alpha_k$ in the double-line cases $\nrightarrow$ and $\alpha_k^\vee = 3\alpha_k$ in the triple-edge case $\not\equiv$.

Let $D$ be the diagonal matrix $D_{k,k} = \|\alpha_k\|$. Then $S = D^{-1}CD/2$ is a symmetric matrix, and we compute bases for the root system $\Phi$ and the dual root system $\Phi^\vee$ as the columns of the matrices[6] $\Sigma$ and $\Sigma^\vee$ given as

$$\Sigma = RD, \tag{4.23}$$

$$\Sigma^\vee = 2RD^{-1}, \tag{4.24}$$

where $R$ is the Cholesky factorization of $S$, *i.e.*, $R$ is upper-triangular, with positive diagonal elements, such that $R^\top R = D^{-1}CD/2$.

A root system leads to the definition of Weyl groups and affine Weyl groups. Let $\Phi$ be a $d$-dimensional root system with dual root system $\Phi^\vee$. For any root $\alpha \in \Phi$ consider the reflection $s_\alpha \colon \mathcal{V} \to \mathcal{V}$ given by

$$s_\alpha(t) = t - \frac{2(t,\alpha)}{(\alpha,\alpha)}\alpha = t - (t,\alpha^\vee)\alpha.$$

For the dual roots $\alpha^\vee \in \Phi^\vee$ we define translations $\tau_{\alpha^\vee} \colon \mathcal{V} \to \mathcal{V}$ by

$$\tau_{\alpha^\vee}(t) = t + \alpha^\vee.$$

The *Weyl group* of $\Phi$ is the finite group of isometries on $\mathcal{V}$ generated by the reflections $s_\alpha$ for $\alpha \in \Sigma$:

$$W = \left\langle \{s_\alpha\}_{\alpha\in\Sigma} \right\rangle.$$

The *dual root lattice* $L^\vee$ is the lattice spanned by the translations $\tau_{\alpha^\vee}$ for $\alpha^\vee \in \Sigma^\vee$. We identify this with the abelian group of translations on $\mathcal{V}$ generated by the dual roots

$$L^\vee = \left\langle \{\tau_{\alpha^\vee}\}_{\alpha^\vee\in\Sigma^\vee} \right\rangle.$$

The *affine Weyl group* $\tilde{W}$ is the infinite crystallographic symmetry group of $\mathcal{V}$ generated by the reflections $s_\alpha$ for $\alpha \in \Sigma$ and the translations $\tau_{\alpha^\vee}$ for

[6] By abuse of notation we use $\Sigma$ to denote both the basis for the root system and the matrix whose columns form the basis of the root system.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

$\alpha^\vee \in \Sigma^\vee$, thus it is the semidirect product of the Weyl group $W$ with the dual[7] root lattice $L^\vee$

$$\tilde{W} = \left\langle \{s_\alpha\}_{\alpha\in\Sigma}, \{\tau_{\alpha^\vee}\}_{\alpha^\vee\in\Sigma^\vee} \right\rangle = W \rtimes L^\vee.$$

Let $\Lambda$ denote the reciprocal lattice of $L^\vee$, *i.e.*, for all $\lambda \in \Lambda$ and all $\alpha^\vee \in \Phi^\vee$ we have $(\lambda, \alpha^\vee) \in \mathbb{Z}$. The lattice $\Lambda$ is spanned by vectors $\{\lambda_j\}_{j=1}^d$ such that $(\lambda_j, \alpha_k^\vee) = \delta_{j,k}$ for all $\alpha_k^\vee \in \Sigma^\vee$. The vectors $\lambda_j$ are called the *fundamental dominant weights* of the root system $\Phi$ and $\Lambda$ is called the *weights lattice*.

We use the weights $\lambda_j$ and dual roots $\alpha_k^\vee$ to represent the Weyl group in matrix form as follows:

$$w_{j,k} := (\lambda_j, w\alpha_k^\vee) \quad \text{for all } w \in W. \tag{4.25}$$

In particular, the reflections $s_{\alpha_r}$ for $\alpha_r \in \Sigma$ are represented as

$$(s_{\alpha_r})_{j,k} = \delta_{j,k} - C_{r,k}\delta_{r,j} \Rightarrow s_{\alpha_r} = I - e_r e_r^\top C,$$

where $C$ is the Cartan matrix and $\{e_r\}$ is the standard basis on $\mathbb{R}^d$. This shows that the Weyl group can be represented as integer matrices with respect to the basis $\Sigma^\vee$ and the dual basis $\{\lambda_1, \ldots, \lambda_d\}$ for $\mathcal{V}$.

The *positive Weyl chamber* $\mathcal{C}_+$ is defined as the closed conic subset of $\mathcal{V}$ containing the points with non-negative coordinates with respect to the dual basis $\{\lambda_1, \ldots, \lambda_d\}$, in other words

$$\mathcal{C}_+ = \{t \in \mathcal{V} \colon (t, \alpha_j) \geq 0\}.$$

This is a fundamental domain for the Weyl group acting on $\mathcal{V}$. The boundary of $\mathcal{C}_+$ consists of the hyperplanes perpendicular to $\{\alpha_1, \ldots, \alpha_d\}$. The affine Weyl group contains reflection symmetries about affine planes perpendicular to the roots, shifted a half-integer multiple of the length of a co-root away from the origin, *i.e.*, for each $\alpha^\vee \in \Phi^\vee$ and each $k \in Z$ there is an affine plane consisting of the points $P_{k,\alpha^\vee} = \{t \in \mathcal{V} \colon 2(t, \alpha^\vee) = k(\alpha^\vee, \alpha^\vee)\} = \{t \in \mathcal{V} \colon (t, \alpha) = k\}$, and this affine plane is invariant under the affine reflection $\tau_{k\alpha^\vee} \cdot s_\alpha$. A connected closed subset of $\mathcal{V}$ limited by such affine planes is called an *alcove*, and is a fundamental domain for the affine Weyl group $\tilde{W}$.

The situation is particularly simple for irreducible root systems, where the alcoves are always $d$-simplexes. Recall that any root $\alpha \in \Phi$ can be expressed as $\alpha = \sum_{k=1}^d n_k \alpha_k$, where the $n_k = 2(\alpha, \lambda_k)/(\alpha_k, \alpha_k)$ are either all non-negative or all non-positive integers. A root $\tilde{\alpha}$ strictly dominates another root $\alpha$, written $\tilde{\alpha} \succ \alpha$, if $\tilde{n}_k \geq n_k$ for all $k$, with strict inequality for at least one $k$. Irreducible root systems have a unique *dominant root*

[7] Since the Weyl groups of $\Phi$ and of $\Phi^\vee$ are identical, it is no problem to instead define $\tilde{W} = W \rtimes L$ as the semidirect product of the Weyl group with the *primal* root lattice. We have, however, chosen here to follow the most common definition, which leads to a slightly simpler notation for the Fourier analysis.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

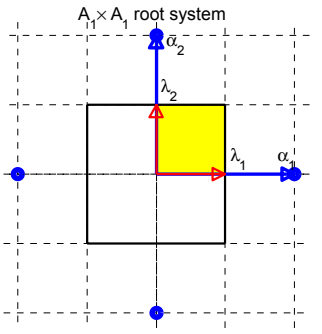



Figure 4.3. Fundamental domains of reducible root system $A_1 \times A_1$.

$\tilde{\alpha} \in \Phi$ such that $\tilde{\alpha} \succ \alpha$ for all $\alpha \neq \tilde{\alpha}$. The dominant root $\tilde{\alpha}$ is the unique long root in the Weyl chamber (possibly on the boundary). Basic geometric properties of affine Weyl groups are summarized by the following lemma.

**Lemma 4.11.**

(i) If $\Phi$ is irreducible with dominant root $\tilde{\alpha}$ then a fundamental domain for $\tilde{W}$ is the simplex $\triangle \subset \mathcal{V}$ given by

$$\triangle = \{t \in \mathcal{V} \colon (t, \tilde{\alpha}) \leq 1 \text{ and } (t, \alpha_j) \geq 0 \text{ for all } \alpha_j \in \Sigma\},$$

where $\triangle$ has corners in the origin and in the points $\lambda_j/(\lambda_j, \tilde{\alpha})$ for $j = 1, \ldots, d$.

(ii) The affine Weyl group is generated by the affine reflections[8] about the boundary faces of the fundamental domain $\triangle$. For irreducible $\Phi$ these are

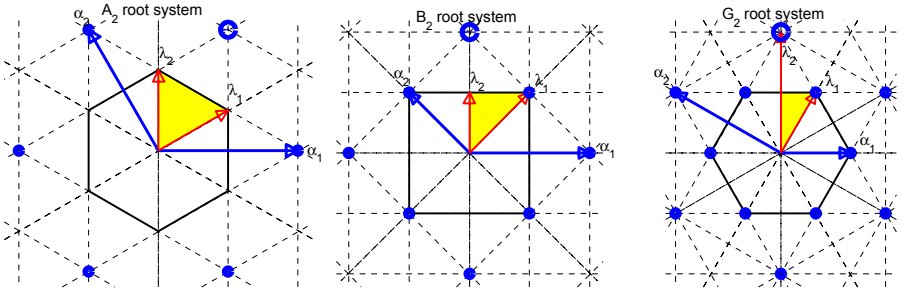
$$\tilde{W} = \langle \{s_{\alpha_j}\}_{j=1}^{d}, \tau_{\tilde{\alpha}} \cdot s_{\tilde{\alpha}} \rangle.$$

(iii) If $\Phi$ is reducible then a fundamental domain for the affine Weyl group is given as the Cartesian product of the fundamental domains for each of its irreducible components.

**Example 4.12.** The simplest rank $d$ root system is the reducible system $A_1 \times \cdots \times A_1$, where the Dynkin diagram consists of $d$ non-connected dots. Figure 4.3 shows $A_1 \times A_1$. The square outlined in black is the fundamental domain of the root lattice, and the small shaded square is the fundamental domain of the affine Weyl group $\tilde{W}$. Reducible root systems can be easily understood as products of irreducible root systems. For instance, the multivariate Chebyshev polynomials corresponding to $A_1 \times \cdots \times A_1$ are the tensor products of $d$ univariate (classical) Chebyshev polynomials on a $d$-cube.

[8] Since $\tilde{W}$ is generated by reflections it is also a special case of a *Coxeter group*.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press




(a) (b) (c)

Figure 4.4. The 2D irreducible root systems.

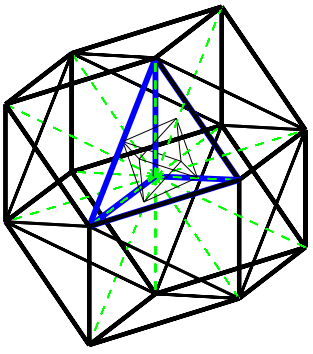


Figure 4.5. Fundamental domains of $A_3$ root system.

**Example 4.13.** The 2D and 3D irreducible root systems are $A_2$, $B_2$, $G_2$, $A_3$, $B_3$ and $C_3$. The Cartan matrix, the matrix representation of generators for the Weyl group, and the size of the Weyl group are shown in Table 4.3.

Figure 4.4 shows the 2D cases with roots $\alpha$ (large dots), dominant root $\tilde{\alpha}$ (larger dot), simple positive roots $(\alpha_1, \alpha_2)$, and their fundamental dominant weights $(\lambda_1, \lambda_2)$. The roots are normalized such that the longest roots have length $\sqrt{2}$, thus for long roots $\alpha^\vee = \alpha$. For short roots we have for $B_2$ that $\alpha^\vee = 2\alpha_2$ and for $G_2$ that $\alpha^\vee = 3\alpha$. The fundamental domain of the dual root lattice (Voronoi region of $L^\vee$) is indicated by ⬡, □ and ⬡, and the fundamental domain for the affine Weyl group is indicated by triangles.

Figure 4.5 shows the $A_3$ case (self-dual), where the fundamental domain (Voronoi region) of the root lattice is a rhombic dodecahedron, a convex polyhedron with 12 rhombic faces. Each of these faces (composed of two triangles) is part of a plane perpendicular to one of the 12 roots, half-way out to the root (roots are not drawn). The fundamental domain of the affine Weyl group is the tetrahedron with an inscribed octahedron. The

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

Table 4.3. The Cartan matrix, the matrix representation of generators for the Weyl group, and the size of the Weyl group.

| | $C$ | $s_1$ | $s_2$ | $s_3$ | $\lvert W\rvert$ |
|---|---|---|---|---|---|
| $A_2$ | $\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ | $\begin{pmatrix} -1 & 1 \\ 0 & 1 \end{pmatrix}$ | $\begin{pmatrix} 1 & 0 \\ 1 & -1 \end{pmatrix}$ | | 6 |
| $B_2$ | $\begin{pmatrix} 2 & -2 \\ -1 & 2 \end{pmatrix}$ | $\begin{pmatrix} -1 & 2 \\ 0 & 1 \end{pmatrix}$ | $\begin{pmatrix} 1 & 0 \\ 1 & -1 \end{pmatrix}$ | | 8 |
| $G_2$ | $\begin{pmatrix} 2 & -1 \\ -3 & 2 \end{pmatrix}$ | $\begin{pmatrix} -1 & 1 \\ 0 & 1 \end{pmatrix}$ | $\begin{pmatrix} 1 & 0 \\ 3 & -1 \end{pmatrix}$ | | 12 |
| $A_3$ | $\begin{pmatrix} 2 & -1 & 0 \\ -2 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$ | $\begin{pmatrix} -1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ | $\begin{pmatrix} 1 & 0 & 0 \\ 1 & -1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$ | $\begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & -1 \end{pmatrix}$ | 24 |
| $B_3$ | $\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -2 \\ 0 & -1 & 2 \end{pmatrix}$ | $\begin{pmatrix} -1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ | $\begin{pmatrix} 1 & 0 & 0 \\ 1 & -1 & 2 \\ 0 & 0 & 1 \end{pmatrix}$ | $\begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & -1 \end{pmatrix}$ | 48 |
| $C_3$ | $\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -2 & 2 \end{pmatrix}$ | $\begin{pmatrix} -1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ | $\begin{pmatrix} 1 & 0 & 0 \\ 1 & -1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$ | $\begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 2 & -1 \end{pmatrix}$ | 48 |

corners of this tetrahedron constitute the fundamental dominant weights $\lambda_1$, $\lambda_2$, $\lambda_3$. The regular octahedron drawn inside the Weyl chamber is important for application of multivariate Chebyshev polynomials, discussed in Section 4.2.2.

*Laplacian eigenfunctions on triangles, tetrahedra and simplexes.* In this subsection we will consider real- or complex-valued functions on $\mathcal{V}$-respecting symmetries of an affine Weyl group. Consider first $\mathrm{L}^2(\mathbf{T})$, the space of complex-valued $\mathrm{L}^2$-integrable periodic functions on the torus $\mathbf{T} = \mathcal{V}/L^\vee$ *i.e.*, functions $f$ such that $f(y + \alpha^\vee) = f(y)$ for all $y \in \mathcal{V}$ and $\alpha^\vee \in L^\vee$. Since the weights lattice $\Lambda$ is reciprocal to $L^\vee$, the Fourier basis for periodic functions is given as $\{\exp(2\pi\mathrm{i}(\lambda, t))\}_{\lambda\in\Lambda}$ and Fourier transforms are defined for $f \in \mathrm{L}^2(\mathbf{T})$:

$$\widehat{f}(\lambda) = \mathcal{F}(f)(\lambda) = \frac{1}{\mathrm{vol}(\mathbf{T})} \int_{\mathbf{T}} f(t)\mathrm{e}^{-2\pi\mathrm{i}(\lambda,t)}\,\mathrm{d}t, \tag{4.26}$$

$$f(t) = \mathcal{F}^{-1}(\widehat{f})(t) = \sum_{\lambda\in\Lambda} \widehat{f}(\lambda)\mathrm{e}^{2\pi\mathrm{i}(\lambda,t)}. \tag{4.27}$$

We are interested in functions which are periodic under translations in $L^\vee$ and which in addition respect the other symmetries in $\tilde{W}$, *e.g.*, functions

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

with odd or even symmetry with respect to the reflections in $\tilde{W}$ (we will see below that there are also other possibilities). Due to the semi-direct product structure $\tilde{W} = W \rtimes L^\vee$ it follows that any $\tilde{w} \in \tilde{W}$ can be written $\tilde{w} = w \cdot \tau_{\alpha^\vee}$, where $w \in W$ and $\alpha^\vee \in L^\vee$. Thus, on the space of periodic functions $\mathrm{L}^2(\mathbf{T})$ the action of $\tilde{W}$ and the finite group $W$ are identical. We define subspaces of symmetric and skew-symmetric periodic functions as follows:

$$\mathrm{L}^2_\vee(\mathbf{T}) = \{f \in \mathrm{L}^2(\mathbf{T})\colon f(wt) = f(t) \text{ for all } w \in W,\, t \in T\},$$
$$\mathrm{L}^2_\wedge(\mathbf{T}) = \{f \in \mathrm{L}^2(\mathbf{T})\colon f(wt) = (-1)^{|w|} f(t) \text{ for all } w \in W,\, t \in T\},$$

where $|w|$ denotes the *length*, defined as $|w| = \ell$, where $w = s_{\alpha_{j_1}} \cdots s_{\alpha_{j_\ell}}$ is written in the shortest possible way as a product of reflections about the simple positive roots $\alpha_j \in \Sigma$. Thus $(-1)^{|w|} = \det(w) = \pm 1$ depending on whether $w$ is a product of an even or odd number of reflections. We define $\mathrm{L}^2$-orthogonal projections $\pi^W_\vee$ and $\pi^W_\wedge$ on these subspaces as

$$\pi^W_\vee f(t) = \frac{1}{|W|} \sum_{w \in W} f(wt), \tag{4.28}$$
$$\pi^W_\wedge f(t) = \frac{1}{|W|} \sum_{w \in W} (-1)^{|w|} f(wt). \tag{4.29}$$

Orthogonal bases for these subspaces are obtained by projecting the exponentials, yielding the cosine- and sine-type basis functions

$$c_\lambda(t) := \pi^W_\vee \exp 2\pi \mathrm{i}(\lambda, t),$$
$$s_\lambda(t) := \pi^W_\wedge \exp 2\pi \mathrm{i}(\lambda, t)$$

as in (4.21)–(4.22). Note that these are not all distinct functions. Since they possess symmetries

$$c_\lambda(t) = c_{w\lambda}(t),$$
$$s_\lambda(t) = (-1)^{|w|} s_{w\lambda}(t),$$

for every $w \in W$, we need only one $\lambda$ from each orbit of $W$. The weights in the Weyl chamber, $\Lambda_+ = \mathcal{C}_+ \cap \Lambda$, form a natural index set of orbit representatives, and we find $\mathrm{L}^2$-orthogonal bases by taking the corresponding $c_\lambda(t)$ and $s_\lambda(t)$. Lemma 4.9 also holds in the following more general case.

**Lemma 4.14.** Let $\phi$ be a rank $d$ root system with weights lattice $\Lambda$. Let $W$ be the Weyl group and let $\triangle$ denote the fundamental domain of the affine Weyl group $\tilde{W} = W \rtimes L^\vee$. The functions $\{c_\lambda(t)\}_{\lambda \in \Lambda_+}$ and $\{s_\lambda(t)\}_{\lambda \in \Lambda_+}$ form two distinct $\mathrm{L}^2$-orthogonal bases for $\mathrm{L}^2(\triangle)$. These basis functions are eigenfunctions of the Laplacian $\nabla^2$ on $\triangle$ satisfying homogeneous Neumann and Dirichlet boundary conditions, as in Lemma 4.9.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

Unfortunately, truncations of these bases do not form excellent approximation spaces for analytic functions on $\triangle$. Approximation of a function $f$, defined on $\triangle$, in terms of $\{c_\lambda(t)\}$ is equivalent to Fourier approximation of the even extension of $f$ in $\mathrm{L}^2_\vee(\mathbf{T})$, and in general we only observe quadratic convergence due to discontinuity of the gradient across the boundary of $\triangle$. A route to spectral convergence is by a change of variables, which turns the trigonometric polynomials $\{c_\lambda(t)\}$ and $\{s_\lambda(t)\}$ into multivariate Chebyshev polynomials of the first and second kind. Before we dive into this topic, we remark on an alternative approach.

*Trigonometric polynomials with generalized symmetries.* In 1D it is known that spectral (exponentially converging) approximations for analytic $f$ defined on $\triangle$ can be obtained by employing the *frame* (*i.e.*, a spanning set of vectors which is not linearly independent) generated by taking both the cosine and sine functions. In the frame we compute an approximating function by solving the following periodic extension problem. Find a smooth extension of $f$ to $\mathrm{L}^2(\mathbf{T})$ and approximate this by the frame on $\triangle$. A detailed discussion of such techniques is found in Huybrechs (2010). One question is whether a frame consisting of only $c_\lambda(t)$ and $s_\lambda(t)$ might also be sufficient to obtain spectral convergence for higher-dimensional domains, or whether we have to include basis functions with other kinds of symmetries.

Consider $\mathrm{L}^2(\mathbf{T},\mathcal{W})$, the space of functions periodic on $\mathbf{T}$ returning values in a Hilbert space $\mathcal{W}$ with inner product $(\cdot,\cdot)_\mathcal{W}$. The inner product on $\mathrm{L}^2(\mathbf{T},\mathcal{W})$ is

$$(f,g) = \int_{\mathbf{T}} (f(t),g(t))_\mathcal{W}\,\mathrm{d}t.$$

Recall that a *representation* of the group $W$ on $\mathcal{W}$ is a map $\rho\colon W \to \mathrm{Gl}(\mathcal{W})$, defining a linear action of $W$ on $\mathcal{W}$, *i.e.*, for $w \in W$, $\rho(w)$ is an invertible linear map on $\mathcal{W}$ such that

$$\rho(w\tilde{w}) = \rho(w)\rho(\tilde{w}).$$

The integer $d_\rho = \dim(\mathcal{W})$ is called the *dimension* of the representation $\rho$. We will (without loss of generality) always assume representations are unitary with respect to $(\cdot,\cdot)_\mathcal{W}$. We say that $f \in \mathrm{L}^2(\mathbf{T},\mathcal{W})$ is $\rho$-*symmetric* if

$$f(wt) = \rho(w)f(t) \quad \text{for all } w \in W,\ t \in \mathbf{T}.$$

The space of such symmetric functions is denoted $\mathrm{L}^2_\rho(\mathbf{T},\mathcal{W})$. As an example, the functions in $\mathrm{L}^2_\vee(\mathbf{T})$ and $\mathrm{L}^2_\wedge(\mathbf{T})$ are $\rho$-symmetric with respect to the trivial representation $\rho_s(w) = 1$ and the alternating representation $\rho_a(w) = (-1)^{|w|}$ acting on $\mathcal{W} = \mathbb{R}$. Since the Weyl group $W$ is constructed as a group of linear transformations on $\mathcal{V}$, we always also have the faithful identity representation $\rho_I(w) = w$, acting on $\mathcal{W} = \mathcal{V}$. Seen through a kaleidoscope,

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

a vector field on $\triangle$ appears to be $\rho_I$-symmetric on $\mathbf{T}$, where the vectors are reflected in the mirrors.

We define a projection $\pi_\rho^W : \mathrm{L}^2(\mathbf{T}, \mathcal{W}) \to \mathrm{L}^2_\rho(\mathbf{T}, \mathcal{W})$ as

$$\pi_\rho^W f(t) = \frac{1}{|W|} \sum_{w \in W} \rho(w^{-1}) f(wt). \tag{4.30}$$

Since $\rho$ is unitary, it can be shown that $\pi_\rho^W$ is an $\mathrm{L}^2$-orthogonal projection.

Note that scalar functions $f : \mathbf{T} \to \mathbb{C}$ can be symmetrized with respect to representations on any vector space $\mathcal{W}$ in a similar manner, *i.e.*,

$$\pi_\rho^W f(t) = \frac{1}{|W|} \sum_{w \in W} \rho(w^{-1}) f(wt),$$

which yields a function $\pi_\rho^W f(t)$ returning values in the space of all matrices over $\mathcal{W}$. Using Theorem 4.6, we can recover $f$ on $\mathbf{T}$ from its symmetrizations on $\triangle$.

**Proposition 4.15.** Let $f$ be a periodic scalar function $f \in \mathrm{L}^2(\mathbf{T})$. Given the symmetrizations $f_\rho = \pi_\rho^W f \in \mathrm{L}^2_\rho(\mathbf{T}, \mathcal{W})$ for all $\rho \in \mathcal{R}$, where $\mathcal{R}$ is a complete list of irreducible representations of $W$, we recover $f$ on all of $\mathbf{T}$ as

$$f(wt) = \sum_{\rho \in \mathcal{R}} d_\rho \mathrm{trace}\,(\rho(w) f_\rho(t)), \quad \text{for all } t \in \triangle \text{ and } w \in W.$$

It should be remarked that for the root system $A_n$ the Weyl group is $W \simeq S_{n+1}$, the symmetric group consisting of all permutations of $n+1$ items. The representation theory of $S_{n+1}$ is well known, and there also exist fast algorithms for computing the corresponding transforms.

Similarly to the symmetrization of $f$ we can symmetrize the Fourier basis with respect to $\rho$ and obtain matrix-valued symmetrized exponentials defined as the $d_\rho \times d_\rho$ matrices

$$\exp_{\rho,\lambda}(t) := \pi_\rho^W \mathrm{e}^{2\pi \mathrm{i}(\lambda, t)}, \tag{4.31}$$

where $d_\rho = \dim(\mathcal{W})$. Restricting to $\lambda \in \Lambda_+$ and $\rho \in \mathcal{R}$, where $\mathcal{R}$ is a complete list of irreducible representations of $W$, the components of $\{\exp_{\rho,\lambda}\}_{\lambda \in \Lambda_+, \rho \in \mathcal{R}}$ form an $\mathrm{L}^2$-orthogonal basis for $\mathrm{L}^2(\mathbf{T})$. Restricted to $\triangle$ they form a tight frame with frame bound $|W|$: see Huybrechs (2010) for a treatment of the classical case $A_1$. These generalized symmetric exponential functions are (almost) equivalent to the irreducible representations of the affine Weyl group $\tilde{W}$ obtained by the technique of small groups (Serre 1977). In Section 4.2.2 we will see that the odd and even symmetrized exponentials $c_\lambda(t)$ and $s_\lambda(t)$ transform to multivariate Chebyshev polynomials of the first and second kind under a particular change of variables. Also, the

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

more general $\rho$-symmetrized exponentials $\exp_{\rho,\lambda}(t)$ are related to multivariate Chebyshev polynomials of 'other kinds'.

Approximation theory on the simplex $\triangle$ using the frame spanned by $\{\exp_{\rho,\lambda}\}$, and the theory of multivariate Chebyshev polynomials of other kinds, is currently under investigation.[9] It is known in the $A_2$ case that the frame consisting only of $c_\lambda(t)$ and $s_\lambda(t)$ is not sufficient to guarantee spectral convergence of arbitrary analytic functions on $\triangle$. By including the last four basis functions arising from the irreducible two-dimensional representation $\rho_I(w) = w$ we do obtain spectral convergence. However, practical use of these techniques in discretizations of PDEs has not yet been completed.

*4.2.2. Multivariate Chebyshev polynomials*

Bivariate Chebyshev polynomials were constructed independently by Koornwinder (1974) and Lidl (1975) by folding exponential functions. Multidimensional generalizations (the $A_2$ family) appeared first in Eier and Lidl (1982). Hoffman and Withers (1988) presented a general folding construction. Characterization of such polynomials as eigenfunctions of differential operators is found in Beerends (1991) and Koornwinder (1974). Applications to the solution of differential equations are found in Munthe-Kaas (2006) and to triangle-based spectral element Clenshaw–Curtis-type quadratures in Ryland and Munthe-Kaas (2011).

Recall that classical Chebyshev polynomials of the first and second kind, $T_k(x)$ and $U_k(x)$, are obtained from $\cos(k\theta)$ and $\sin(k\theta)$ by a change of variable $x = \cos(\theta)$, as

$$T_k(x) = \cos(k\theta),$$
$$U_k(x) = \frac{\sin((k+1)\theta)}{\sin(\theta)}.$$

We want to understand this construction in the context of affine Weyl groups. We recognize $\cos(k\theta)$ and $\sin(k\theta)$ as the symmetrized and skew-symmetrized exponentials. The $\cos(\theta)$ used in the change of variables is the $2\pi$-periodic function, which is symmetric, non-constant and has the longest wavelength (as such, uniquely defined up to constant). In other words $\cos(\theta) = \pi_\vee \exp(\lambda_1\theta)$, where $\lambda_1 = 1$ is the generator of the weights lattice. Any periodic band-limited even function $f$ has a symmetric Fourier series of finite support on the weights lattice, and must hence be a polynomial in the variable $x = \cos(\theta)$. The denominator $\sin(\theta)$ is similarly the odd function of longest possible wavelength. Any periodic band-limited odd function $f$ has a skew-symmetric Fourier series on the weights lattice. Dividing out $\sin(\theta)$ results in a band-limited even function which again must

[9] H. Munthe-Kaas and D. Huybrechs, work in progress.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

map to a polynomial under our change of variables. The denominator, which in our special case is $\sin(\theta)$, we call the *Weyl denominator*. It plays an important role in the representation theory of compact Lie groups as the denominator in *Weyl's* character formula, the jewel in the crown of representation theory. We will detail these constructions below.

*Notation.* As in Section 4.2.1, we let $\Phi$ be a rank $d$ root system on $\mathcal{V} = \mathbb{R}^d$, with Weyl group $W$, co-root lattice $L^\vee$ and affine Weyl group $\tilde{W} = W \rtimes L^\vee$. Let $\mathbf{T} = \mathcal{V}/L^\vee$ be the torus of periodicity and $\Lambda = \mathrm{span}_{\mathbb{Z}}\{\lambda_1, \ldots, \lambda_d\}$ the reciprocal lattice of $L^\vee$. The lattice $\Lambda$ is an abelian group known as the *Pontryagin dual* of the abelian group $\mathbf{T}$. Specifically, each $\lambda \in \Lambda$ corresponds uniquely to an exponential function $\exp(2\pi\mathrm{i}(\lambda, t))$, which is one of the irreducible one-dimensional representations on $\mathbf{T}$. These form an abelian group under multiplication. In the following, it is convenient to write the group $\Lambda$ in multiplicative form as follows. Let $\{e^\lambda\}_{\lambda\in\Lambda}$ denote the elements of the multiplicative group,[10] understood as formal symbols such that for $\lambda, \mu \in \Lambda$ we have $e^\lambda \cdot e^\mu = e^{\lambda+\mu}$.

Let $\mathcal{E} = \mathcal{E}(\mathbb{C})$ denote the free complex vector space over the symbols $e^\lambda$. This consists of all formal sums $a = \sum_{\lambda\in\Lambda} a(\lambda)e^\lambda$ where the coefficients $a(\lambda) \in \mathbb{C}$ and all but a finite number of these are non-zero. This is the complex group algebra over the multiplicative group $\Lambda$, with a commutative product induced from the group product

$$\left(\sum_{\lambda\in\Lambda} a(\lambda)\right)\left(\sum_{\mu\in\Lambda} b(\mu)\right) = \sum_{\nu\in\Lambda}\left(\sum_{\lambda+\mu=\nu} a(\lambda)b(\mu)\right)\cdot e^\nu.$$

We define conjugation $^- : \mathcal{E} \to \mathcal{E}$ as

$$\overline{a} = \sum_{\lambda\in\Lambda} \overline{a(\lambda)}e^{-\lambda}.$$

An element $a \in \mathcal{E}$ is identified with a trigonometric polynomial $f(t) = \mathcal{F}^{-1}(a)(t)$ on the torus $\mathbf{T}$ (*i.e.*, a band-limited periodic function) through the Fourier transforms given in (4.26)–(4.27). Multiplication and conjugation in $\mathcal{E}$ corresponds to pointwise multiplication and complex conjugation of the functions on $\mathbf{T}$.

Let $\mathcal{E}_\vee^W \subset \mathcal{E}$ denote the symmetric subalgebra of those elements that are invariant under the action of the Weyl group $W$ on $\mathcal{E}$. This consists of those $a \in \mathcal{E}$ where $a(\lambda) = a(w\lambda)$ for all $\lambda$ and all $w \in W$. Similarly, $\mathcal{E}_\wedge^W \subset \mathcal{E}$ denotes those $a \in \mathcal{E}$ whose sign alternates under reflections $s_\alpha$, *i.e.*, where the coefficients satisfy $a(\lambda) = (-1)^{|w|}a(w\lambda)$. Here $|w|$ denotes the length

[10] Those who prefer can simply consider $e^\lambda$ to be shorthand for $\exp(2\pi\mathrm{i}(\lambda, t))$, as a function of $t \in \mathbf{T}$.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

of $w$, the length of the shortest factorization of $w$ into a product of simple reflections. Thus $(-1)^{|w|} = \det(w)$. The alternating elements do not form a subalgebra, since the product of two alternating functions is symmetric. The subspaces $\mathcal{E}_\vee^W$ and $\mathcal{E}_\wedge^W$ correspond to $W$-symmetric and $W$-alternating trigonometric polynomials on $\mathbf{T}$.

As a digression we remark that a parallel theory is the representation theory of compact connected Lie groups $G$: see Bump (2004) for an excellent exposition. In this setting $\mathbf{T}$ is the maximal torus of $G$, *i.e.*, a maximal abelian subgroup. Any $g \in G$ is conjugate to a point in $\mathbf{T}$, *i.e.*, there exist a $k \in G$ such that $kgk^{-1} \in \mathbf{T}$. The normalizer $N(\mathbf{T})$ is composed of those elements of $G$ which fix $\mathbf{T}$ under conjugation, $N(\mathbf{T}) = \{k \in G \big| k\mathbf{T}k^{-1} = \mathbf{T}\}$. The Weyl group is $W = N(\mathbf{T})/\mathbf{T}$, which is always finite. It can be shown that the action of $W$ on $T$ through conjugation can be represented in terms of a root system. This implies that class functions on $G$, *i.e.*, functions such that $f(g) = f(kgk^{-1})$ for all $k$, can be represented as $W$-invariant functions on $\mathbf{T}$, and that any function in $\mathcal{E}_\vee^W$ can be interpreted as a class function on $G$. Of particular interest are the class functions which arise as traces of the irreducible representations on $G$. These are called the irreducible characters, and are known explicitly through the celebrated Weyl character formula. We will see that the irreducible characters and multivariate Chebyshev polynomials of the second kind are equivalent: see Beerends (1991).

*Multivariate Chebyshev polynomials.* Define a projection $\pi_\wedge^W : \mathcal{E} \to \mathcal{E}_\wedge^W$ by its action on the coefficients $a \mapsto \pi_\wedge^W a$:

$$(\pi_\wedge^W a)(\lambda) = \frac{1}{|W|} \sum_{w \in W} (-1)^{|w|} a(w\lambda).$$

The projection $\pi_\vee^W : \mathcal{E} \to \mathcal{E}_\vee^W$ is defined similarly, omitting $(-1)^{|w|}$.

The algebra $\mathcal{E}$ is generated by $\{e^{\lambda_j}\}_{j=1}^d \cup \{e^{-\lambda_j}\}_{j=1}^d$ where $\lambda_j$ are the fundamental dominant weights. $\mathcal{E}_\vee^W$ is the subalgebra generated by the symmetric generators $\{z_j\}_{j=1}^d$ defined by

$$z_j = \pi_\vee^W e^{\lambda_j} = \frac{1}{|W|} \sum_{w \in W} e^{w\lambda_j} = \frac{2}{|W|} \sum_{w \in W^+} e^{w\lambda_j},$$

where $W^+$ denotes the even subgroup of $W$ containing those $w$ such that $|w|$ is even. The latter identity follows from $s_{\alpha_j}\lambda_j = \lambda_j$, so it is enough to consider $w$ of even length. The action of $W^+$ on $\lambda_j$ is free and effective.

It can be shown that $\mathcal{E}_\vee^W$ is a unique factorization domain over the generators $\{z_j\}$, *i.e.*, any $a \in \mathcal{E}_\vee^W$ can be expressed uniquely as a polynomial in $\{z_j\}_{j=1}^d$.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

The skew subspace $\mathcal{E}_\wedge^W$ does not form an algebra, but this can be corrected by dividing out the *Weyl denominator*. Define the *Weyl vector* $\rho \in \Lambda$ as

$$\rho = \sum_{j=1}^{d} \lambda_j = \frac{1}{2} \sum_{\alpha \in \Phi^+} \alpha.$$

We define the Weyl denominator $D \in \mathcal{E}_\wedge^W$ as

$$D = \sum_{w \in W} (-1)^{|w|} e^{w\rho}.$$

**Proposition 4.16.** Any $a \in \mathcal{E}_\wedge^W$ is divisible by $D$, *i.e.*, there exists a unique $b \in \mathcal{E}_\vee^W$ such that $a = bD$.

*Proof.* See Bump (2004, Proposition 25.2). □

Any $a \in \mathcal{E}_\vee^W$ can be written as a polynomial in $z_1, \ldots, z_d$, and hence the following polynomials are well-defined.

**Definition 4.17.** For $\lambda \in \Lambda$ we define multivariate Chebyshev polynomials of the first and second kind, $T_\lambda$ and $U_\lambda$, as the unique polynomials that satisfy

$$T_\lambda(z_1, \ldots, z_d) = \pi_\vee^W e^\lambda = \frac{1}{|W|} \sum_{w \in W} e^{w\lambda}, \tag{4.32}$$

$$U_\lambda(z_1, \ldots, z_d) = \frac{|W| \pi_\wedge^W e^{\lambda+\rho}}{D} = \frac{\sum_{w \in W} (-1)^{|w|} e^{w(\rho+\lambda)}}{\sum_{w \in W} (-1)^{|w|} e^{w\rho}}. \tag{4.33}$$

By a slight abuse of notation we will also consider $z_j$ as $W$-invariant functions on $\mathbf{T}$ as

$$z_j(t) = \mathcal{F}^{-1}(z_j)(t) = \frac{2}{|W|} \sum_{w \in W^+} \mathrm{e}^{2\pi \mathrm{i}(\lambda_j, t)}. \tag{4.34}$$

We will study coordinates on $\mathbf{T}$ obtained from these. The functions $z_j(t)$ may be real or complex. If there exists an $w \in W^+$ such that $w\lambda_j = \lambda_j$, then $z_j = \overline{z_j}$ is real. Otherwise there must exist an index $\bar{j} \neq j$ and a $w \in W^+$ such that $w\lambda_j = \lambda_{\bar{j}}$ and we have $\overline{z_j} = z_{\bar{j}}$. In the latter case we can replace these with $d$ real coordinates $x_j = \frac{1}{2}(z_j + z_{\bar{j}})$, $x_{\bar{j}} = \frac{1}{2\mathrm{i}}(z_j - z_{\bar{j}})$.

Using the basis $\{\alpha_j^\vee\}$, we identify $\mathbf{T} \simeq (\mathbb{R}/\mathbb{Z})^d$, and let $t_j \in [0, 1)$ denote standard coordinates on this unit-periodic torus. Let

$$J_{k,l}(t) = \frac{\partial z_k(t)}{\partial \alpha_l^\vee} \equiv (\alpha_l^\vee \cdot \nabla_t) z_j(t) \quad k, l \in \{1, \ldots, d\}$$

be the Jacobian matrix of the map $(t_1, \ldots, t_d) \mapsto (z_1, \ldots, z_d)$.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

**Proposition 4.18.** The Fourier transform of the Jacobian determinant, $\widehat{J} := \mathcal{F}(\det(J)(t))$, is the alternating function

$$\widehat{J} = cD \in \mathcal{E}_{\wedge}^{W},$$

where $D$ is the Weyl denominator and the constant $c = \left(\frac{4\pi \mathrm{i}}{|W|}\right)^d$.

*Proof.* Let $s_\alpha$ be any reflection in $W$. Since $z(t) = z(s_\alpha t)$ we have $J(t) = J(s_\alpha t)s_\alpha$. Hence $\det(J(t)) = -\det(J(s_\alpha t))$ and we conclude that $\widehat{J} \in \mathcal{E}_{\wedge}^{W}$. By Proposition 4.16, $D$ divides $J$. We need to confirm that $J/D$ is constant. We compute

$$\widehat{J_{k,\ell}} = \frac{4\pi \mathrm{i}}{|W|} \sum_{w \in W^+} \langle \alpha_\ell, w\lambda_k \rangle e^{w\lambda_k}.$$

Thus, $\widehat{J_{k,\ell}}$ is supported on $W^+\lambda_k$ and $\widehat{J}$ is supported on the set

$$\sum_{w_1,\ldots,w_d \in W^+} w_j \lambda_j.$$

The highest weight in this set is the Weyl vector $\rho = \sum_{j=1}^{d} \lambda_j$, which is reached if and only if $w_1 = w_2 = \cdots w_d = I$. Since the highest weight is $\rho$ we conclude that $\widehat{J}/D = c$. The constant $c$ is computed as follows:

$$\begin{aligned} c = \widehat{J}(\rho) &= \left( \sum_{\sigma \in S_d} \operatorname{sign}(\sigma) \prod_{j=1}^{d} \widehat{J}_{j,\sigma_j} \right)(\rho) \\ &= \sum_{\sigma \in S_d} \operatorname{sign}(\sigma) \prod_{j=1}^{d} \widehat{J}_{j,\sigma_j}(\lambda_j) = \sum_{\sigma \in S_d} \operatorname{sign}(\sigma) \prod_{j=1}^{d} \frac{4\pi \mathrm{i}}{|W|} \langle \alpha_{\sigma_j}, \lambda_j \rangle \\ &= \left( \frac{4\pi \mathrm{i}}{|W|} \right)^d \sum_{\sigma \in S_d} \operatorname{sign}(\sigma) \prod_{j=1}^{d} \delta_{\sigma_j, j} = \left( \frac{4\pi \mathrm{i}}{|W|} \right)^d. \end{aligned}$$

□

Finally, in this paragraph, we want to remark that (4.33) is exactly the same formula as Weyl's character formula, giving the trace of all the irreducible characters on a semi-simple Lie group (Bump 2004). These characters form an $L^2$-orthogonal basis for the space of class functions on the Lie group. Thus, expansions in terms of multivariate Chebyshev polynomials of the second kind is equivalent to expansions in terms of irreducible characters on a Lie group. Just as the basis given by the irreducible representations block-diagonalize equivariant linear operators on a Lie group, we can also use irreducible characters to obtain block diagonalizations: see James and Liebeck (2001). Thus, our software, which is primarily constructed to deal with spectral element discretizations of PDEs, may also have important applications to computations on Lie groups. This opens up a whole area

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

of possible applications of these approximations, such as in random matrix theory. We will not pursue these topics here.

We will focus below on applications of Chebyshev polynomials of the first kind in spectral element discretizations.

*Analytical properties.* The polynomials $T_\lambda(z)$ of the first kind are by construction W-symmetric in $\lambda$, $T_\lambda = T_{w\lambda}$ for all $w \in W$. Thus the full space of multivariate polynomials is spanned by $\{T_\lambda\}_{\lambda\in\Lambda_+}$, where $\Lambda_+$ are the weights within the positive Weyl chamber. We will first show that this is a family of multivariate orthogonal polynomials.

Let $\triangle \subset \mathbf{T}$ denote the fundamental domain of $\tilde{W}$. The change of variables $t \mapsto z$ in (4.34) has a Jacobian determinant proportional to the Weyl denominator $D$. It is known from representation theory that $D$ is always zero on the boundary of $\triangle$ and non-zero in the interior. Thus the coordinate map $t \mapsto z$ is invertible on $\triangle$, regular in the interior of $\triangle$ and singular on the boundary. Let $\delta = z(\triangle)$. This is a domain which in the $A_2$ case is a *deltoid* or three-cusp Steiner hypocycloid: see Figure 4.6(b). Note that $D \in \mathcal{E}^W_\wedge$ is an alternating function, thus $D\overline{D} \in \mathcal{E}^W_\vee$ is a symmetric function, hence it is a real multivariate polynomial in $z$.

**Proposition 4.19.** The coordinate map $t \mapsto z$ is invertible on $\delta = z(\triangle)$. The absolute value of the Jacobian determinant is

$$|\widehat{J}| = |c|\sqrt{D\widehat{D}},$$

where $D\widehat{D}$ is a real polynomial in $z$. The boundary of $\delta$ is given as an algebraic ideal,

$$D\overline{D} = 0.$$

In the $A_1$ case we have $J = \sin(t)$, hence $D\overline{D} = \sin^2(t) = 1 - z^2$. In the $A_2$ case we have

$$D\overline{D} = \frac{1}{3} + \frac{8}{3}(x^3 - 3xy^2) - (x^2 + y^2)^2 - 2(x^2 + y^2),$$

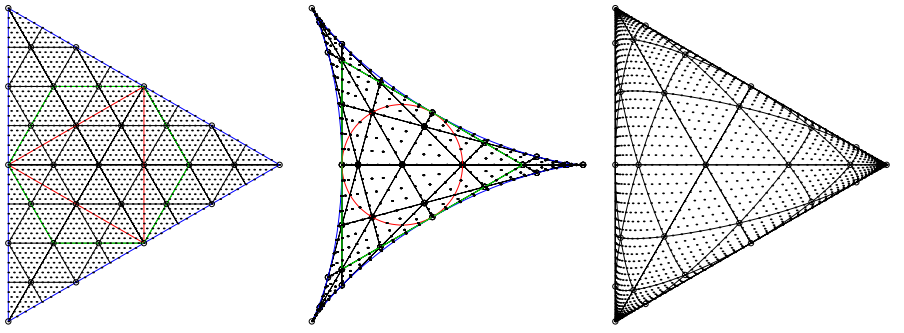
where $x = \frac{1}{2}(z + \overline{z})$ and $y = \frac{1}{2}(z - \overline{z})$.

**Proposition 4.20.** The polynomials $\{T_\lambda\}_{\lambda\in\Lambda_+}$ are orthogonal on $\delta$ with respect to the weighted inner product

$$(f, g) = \int_\delta \overline{f(z)}g(z)\frac{1}{\sqrt{D\overline{D}}}\,\mathrm{d}z.$$

*Proof.* Under the change of variable $z \mapsto t$, we have that $T_\lambda$ maps to $\pi^W_\vee \exp(2\pi\mathrm{i}(\lambda, t))$, and these are $\mathrm{L}^2$-orthogonal on $\triangle$. The weight follows from the formula for the Jacobian determinant. □

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press



(a) (b) (c)

Figure 4.6. The equilateral domain $\Delta$ in (a) maps to the deltoid $\delta$ in (b) under coordinate map $t \mapsto z$, and to a triangle in (c) under a straightening map.

From the convolution product in $\mathcal{E}_\vee^W$ one finds that the $T_\lambda$ satisfy the recurrence relations

$$T_0 = 1, \tag{4.35}$$

$$T_{\lambda_j} = z_j, \tag{4.36}$$

$$T_\lambda = T_{w\lambda} \text{ for } w \in W, \tag{4.37}$$

$$T_{-\lambda} = \overline{T_\lambda}, \tag{4.38}$$

$$T_\lambda T_\mu = \frac{1}{|W|} \sum_{w \in W} T_{\lambda + w\mu}. \tag{4.39}$$

These reduce to classical three-term recurrences for $A_1$ and four-term recurrences for $A_2$: see Munthe-Kaas (2006). The recurrences provide a practical way of computing values of $T_k(z)$ at arbitrary points $z$. For special collocation points we will see below that FFT-based computations are much more efficient.

*Discrete properties.* Pulled back to $t$-coordinates, polynomials in $z$ become band-limited symmetric functions $f \in \mathrm{L}^2_\vee(\mathbf{T})$. Due to the Shannon sampling theorem, any band-limited function $f \in \mathrm{L}^2(\mathbf{T})$ can be exactly reconstructed from sampling on a sufficiently fine lattice $S \subset \mathbf{T}$. Due to periodicity, the reciprocal lattice $S^*$ must be a sublattice of $\Lambda$. The condition for perfect reconstruction is that the only point of the reciprocal lattice $S^*$ within the support of $\widehat{f}$ in $\Lambda$ is the origin. In addition to this condition, we also require the sampling lattice to be $W$-invariant,

$$WS = S.$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

There are several ways to construct such a lattice. A convenient way is to let $S$ be a down-scaled version of the co-root lattice,

$$S = L^\vee/M, \tag{4.40}$$

where $M$ is an integer sufficiently large to ensure the Shannon condition for perfect reconstruction. On the lattice $S$ the sampled function $f_S$ becomes a symmetric function on a finite abelian group. Thus we can employ FFTs to find $\widehat{f} \in \mathcal{E}_\vee^W$. Special group-theoretic versions of symmetrized FFTs can also be used, which are more efficient, but harder to program than the standard FFT: see Munthe-Kaas (1989).

The lattice $S$ maps to collocation points $z(S)$ for the multivariate polynomials, as shown in Figure 4.6 for the $A_2$ case, $M = 12$. The nodal points $z(S) \subset \delta$ form the two-dimensional analogue of Chebyshev extremal points. The straight lines in Figure 4.6(a), which are in the direction of a root, map to straight lines in Figure 4.6(b). Along these lines inside $\delta$, the nodal points distribute like 1D classical Chebyshev collocation points (either as Chebyshev zeros, or as Chebyshev extremal points). Similarly, the multivariate Chebyshev points in higher dimensions contain flat hyperplanes of lower dimension, and under restriction to these one restricts to lower-dimensional Chebyshev nodes. This is an important feature of these polynomials, which simplifies restriction of polynomials to (certain specific) linear subspaces. In particular, we note that the hexagon in Figure 4.6(a) maps to an equilateral triangle embedded nicely inside $\delta$ (Figure 4.6(b)). The fact that the collocation nodes distribute as 1D Chebyshev nodes on the boundary of this triangle allows us to express in closed form conditions for continuity of polynomials patched together along these triangle boundaries.

For high-order polynomial collocation it is extremely important that the Lebesgue number of the collocation points grows slowly in $M$. It is well known that for classical Chebyshev polynomials the Lebesgue number grows at the optimal rate $\mathcal{O}(\log(M))$. From properties of the Dirichlet kernel, one can show a similar result for multivariate Chebyshev polynomials.

**Proposition 4.21.** For the collocation points $z(S)$, where $S = L^\vee/M \subset \mathbb{R}^d$, the Lebesgue number grows as $\mathcal{O}(\log^d(M))$.

*4.2.3. Spectral element methods based on triangles and simplexes*

Triangle-based spectral element methods were first considered in Dubiner (1991), where a basis was constructed by a 'warping' of a 2D tensor product to a triangle. More recent work is Giraldo and Warburton (2005), Hesthaven (1998) and Warburton (2006), based on interpolation of polynomials in good interpolation nodes on a triangle.

Our approach differs from these by basing the construction on the multivariate Chebyshev polynomials, which allows for fast FFT-based computations of all the basic operations on the polynomials. It is well known

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

that a logarithmic growth of the Lebesgue number implies exponentially fast (spectral) convergence of polynomial approximations to analytic functions (Canuto *et al.* 2006). This result, together with the availability of FFT-based algorithms is, just as in the univariate case, among the most important properties of the multivariate Chebyshev polynomials as a practical tool in computational approximation theory and solution of PDEs.

Thus, as a conclusion, the coordinate map $t \mapsto z$ has a dramatic positive effect on the convergence rate of finite approximations using multivariate Chebyshev polynomials, compared to expansions in the symmetric exponentials $\pi_{\wedge}^{W} \exp(2\pi \mathrm{i}(\lambda, t)$. A practical difficulty is, however, the fact that under this change of coordinates the simplex $\triangle$ becomes a significantly more complicated domain with cusps in the corners. This has to be dealt with in an efficient manner if we want to construct practical spectral element methods.

We have been working with three possible solutions to this problem.

- Work with the domain $\triangle \subset \mathbf{T}$. Use the tight frame provided by the $\rho$-symmetrized exponentials for $\rho \in \mathcal{R}$, and solve the periodic extension problem. This approach has not been developed in detail yet.

- Straighten $\delta$ to a nearby simplex by a straightening map (Munthe-Kaas 2006). We have had reasonably good experiences with this approach, but the straightening maps must be singular in the cusps. In numerical experiments with this approach we have seen quite good Lebesgue numbers, and reasonably good convergence rates: see Munthe-Kaas (2006). However, we have not been able to obtain spectral convergence, probably due to the corner singularities of the straightening map.

- Patch together deltoids with overlap, so that the equilateral triangles that are inscribed in the deltoid form a simplicial subdivision of the total space. This approach is working successfully, and will be discussed below.

One important question is whether or not this inscribed triangle is a particular feature for the $A_2$ case, or if the higher-dimensional cases also have a similar structure with a nicely inscribed simplex in $\delta$. This is indeed the case. We can prove this for the $A_n$ family (which is the most important case), and we conjecture that a similar property also holds for the other infinite families of Dynkin diagrams. In Figure 4.7 we illustrate the $\delta$ domain in the $A_3$ case. The 3D deltoid-shaped domain contains an inscribed tetrahedron. This tetrahedron is the image of the regular octahedron inscribed in $\triangle \subset \mathbf{T}$ as seen in Figure 4.5. On the faces of this tetrahedron, the nodal points distribute as the Chebyshev points for the $A_2$ case.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

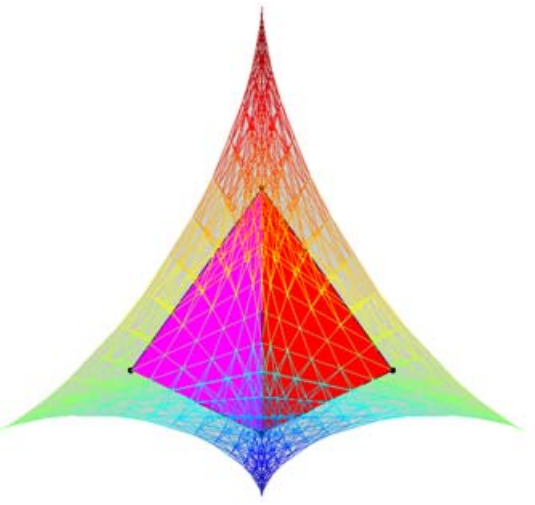


Figure 4.7. The deltoid domain $\delta$ in the $A_3$ case. The inscribed tetrahedron is the image of the regular octahedron inscribed in the Weyl chamber of $\triangle$: see Figure 4.5.

*Numerical experiments.* A library of routines has been implemented for the $A_2$ case, both in MATLAB and in C++. All basic algorithms such as products of functions, derivation and integration is implemented using FFT-based techniques; see Ryland and Munthe-Kaas (2011) for more details. Algorithms exist for exactly integrating the polynomials both over $\delta$ and also over the inscribed triangle, with respect to weight function 1 or the weight function $(D\overline{D})^{-\frac{1}{2}}$.

In Ryland and Munthe-Kaas (2011) we demonstrate the use of these bases in spectrally converging Clenshaw–Curtis-type cubatures based on triangular subdivisions. The software allows for triangles mapped from the equilateral reference triangle both with linear and with non-linear non-singular coordinate maps. The Jacobian of these maps can be supplied analytically, or computed numerically from the sampling values. One problem with this type of quadrature is that, since the deltoid stretches outside its inscribed triangle, one has to evaluate functions at points outside the domain. If this is not possible, one could use the collocation points of the straightened triangle, Figure 4.6(c), for the triangles on the boundary. This, however, reduces the convergence rate.

We are currently exploring spectral element discretizations for PDEs based on similar subdivisions. In the experiments, we have explicitly imposed continuity across triangle boundaries. The numerical experiments confirm the exponential convergence rate for these approximations.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

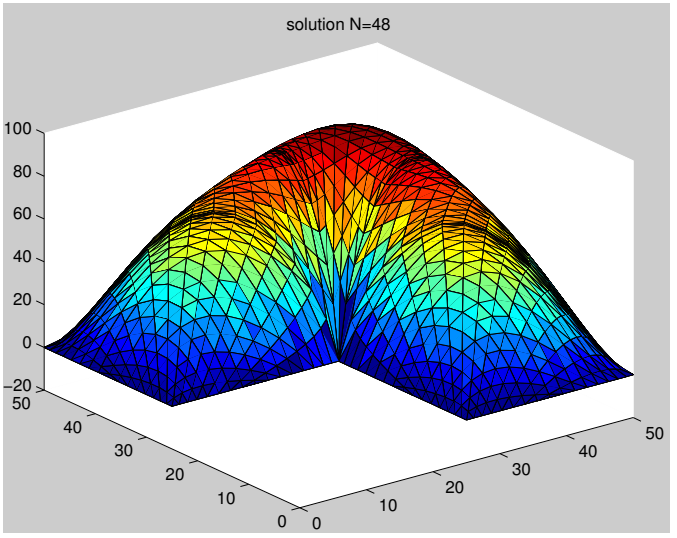



Figure 4.8. The numerical experiment shows spectral element solution of $\nabla^2 u = -1$ on an L-shaped domain composed of three squares, with Dirichlet boundary conditions. Each square is divided into two isosceles triangles, and all these 6 subdomains are patched together by explicitly imposing conditions for continuity across common boundaries of two triangles. Each triangle has a Chebyshev $A_2$ family of Chebyshev polynomials, in this figure using the root lattice down-scaled with the factor 48 as collocation lattice.

## 5. Finite element systems of differential forms

Many partial differential equations (PDEs) can be naturally thought of as expressing that a certain field, say a scalar field or a vector field, is a critical point of a certain functional. Often this functional will be of the form

$$\mathcal{S}(u) = \int_S \mathcal{L}(x, u(x), \nabla u(x), \ldots)\, \mathrm{d}x.$$

Here $S$ is a domain in $\mathbb{R}^d$, $u$ is a section of a vector bundle over $S$ and $\mathcal{L}$ is a Lagrangian involving $x \in S$ as well as values of $u$ and its derivatives. In this context the functional $\mathcal{S}$ is called the action. Criticality of the action can be written as

$$\forall u' \quad \mathrm{D}\mathcal{S}(u)u' = 0.$$

More generally, a PDE written in the form $\mathcal{F}(u) = 0$ can be given a variational formulation:

$$\forall u' \quad \langle \mathcal{F}(u), u' \rangle = 0. \tag{5.1}$$

The operator $\mathcal{F}$ is continuous from a Banach space $X$ to another one $Y$, and $\langle \cdot, \cdot \rangle$ denotes a duality product on $Y \times Y'$.The unknown $u$ is sought in

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

$X$ and the equation is tested with $u'$ in $Y'$. The duality product is most often a generalization of the $\mathrm{L}^2(S)$ scalar product, such as a duality between Sobolev spaces. In the case of least action principles it is of interest to find $X$ such that one can take $Y = X'$ (and $Y' = X$).

This point of view leads quite naturally to discretizations. We construct a trial space $X_n \subseteq X$ and a test space $Y_n' \subseteq Y'$, and solve

$$u \in X_n,\ \forall u' \in Y_n' \quad \langle \mathcal{F}(u), u' \rangle = 0. \tag{5.2}$$

In theory we have sequences indexed by $n \in \mathbb{N}$ and establish convergence properties of the method as $n \to \infty$. This discretization technique is called the Galerkin method when $X_n = Y_n'$ and the Petrov–Galerkin method when $X_n$ and $Y_n'$ are different.

The finite element (FE) method consists in constructing finite-dimensional trial and test spaces adapted to classes of PDEs. Mixed FE spaces have been constructed to behave well for the differential operators grad, curl and div. They provide a versatile tool-box for discretizing PDEs expressed in these terms. Thus Raviart–Thomas div-conforming finite elements (Raviart and Thomas 1977) are popular for PDEs in fluid dynamics and Nédélec's curl-conforming finite elements (Nédélec 1980) have imposed themselves in electromagnetics. For reviews and references see Brezzi and Fortin (1991), Roberts and Thomas (1991), Hiptmair (2002) and Monk (2003).

These spaces fit the definition of a finite element given by Ciarlet (1978). In particular they are equipped with unisolvent degrees of freedom, which determine inter-element continuity and provide interpolation operators: projections $I_n$ onto $X_n$ which are defined at least for smooth fields. In applications one often needs a pair of spaces $X_n^a \times X_n^b$, and these spaces should be compatible in the sense of satisfying a Brezzi inf-sup condition (Brezzi 1974). The spaces are linked by a differential operator $\mathrm{d} : X_n^a \to X_n^b$ (one among grad, curl, div), and the proof of compatibility would follow from a commuting diagram:

$$\begin{array}{ccc} X^a & \xrightarrow{\mathrm{d}} & X^b \\ \downarrow{\scriptstyle I_n^a} & & \downarrow{\scriptstyle I_n^b} \\ X_n^a & \xrightarrow{\mathrm{d}} & X_n^b \end{array}$$

A technical problem is that one would like the interpolators to be defined and bounded on the Banach spaces $X^a$, $X^b$, whereas the natural degrees of freedom are not bounded on these. For instance $X^a$ is usually of the form:

$$X^a = \{u \in \mathrm{L}^2(S) \otimes E^a : \mathrm{d}u \in \mathrm{L}^2(S) \otimes E^b\},$$

for some finite-dimensional fibres $E^a$, $E^b$, but this continuity is usually insufficient for degrees of freedom, such as line integrals, to be well-defined. The convergence of eigenvalue problems requires even more boundedness prop-

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

erties of the interpolation operators, something more akin to boundedness in $\mathrm{L}^2(S)$ (see Christiansen and Winther (2010) for precise statements).

Mixed finite elements have been constructed to contain polynomials of arbitrarily high degree. Applying the Bramble–Hilbert lemma to the interpolators gives high orders of approximation, and under stability conditions this gives a high order of convergence for numerical methods. But here too the intention is hampered by the lack of continuity of interpolators.

As remarked by Bossavit (1988), lowest-order mixed finite elements, when translated into the language of differential forms, correspond to constructs from differential topology called Whitney forms (Weil 1952, Whitney 1957). The above-mentioned differential operators can all be interpreted as instances of the exterior derivative. From this point of view it becomes natural to arrange the spaces in full sequences linked by operators forming commuting diagrams:

$$\begin{array}{ccccccc} X^0 & \xrightarrow{\mathrm{d}} & X^1 & \xrightarrow{\mathrm{d}} & \cdots & \xrightarrow{\mathrm{d}} & X^d \\ \downarrow{\scriptstyle I_n^0} & & \downarrow{\scriptstyle I_n^1} & & \downarrow & & \downarrow{\scriptstyle I_n^d} \\ X_n^0 & \xrightarrow{\mathrm{d}} & X_n^1 & \xrightarrow{\mathrm{d}} & \cdots & \xrightarrow{\mathrm{d}} & X_n^d. \end{array}$$

An important property is that the interpolators should induce isomorphisms on cohomology groups. This is essentially de Rham's theorem, when the top row consists of spaces of smooth differential forms and the bottom row consists of Whitney forms. In particular, on domains homeomorphic to balls, the sequences of FE spaces should be exact. This applies in particular to single elements such as cubes and simplexes.

High-order FE spaces of differential forms naturally generalizing Raviart–Thomas–Nédélec elements were presented by Hiptmair (1999). For a comprehensive review, encompassing Brezzi–Douglas–Marini elements (Brezzi, Douglas and Marini 1985), see Arnold, Falk and Winther (2006*b*). Elements are constructed using the Koszul operator, or equivalently the Poincaré homotopy operator, to ensure local sequence exactness, that is, exactness of the discrete sequence attached to a single element. Mixed FE spaces have thus been defined for simplexes. Tensor product constructions yield spaces on Cartesian products.

Given spaces defined on some mesh, it would sometimes be useful to have spaces constructed on the dual mesh, matching the spaces on the primal mesh in some sense. Since the dual mesh of a simplicial mesh is not simplicial, this motivates the construction of FE spaces on meshes consisting of general polytopes. On these, it seems unlikely that good FE spaces can be constructed with only polynomials: one should at least allow for piecewise polynomials. In some situations, such as convection-dominated fluid flow, stability requires some form of upwinding. This could be achieved by a

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

Petrov–Galerkin method, by including upwinded basis functions in either the trial space $X_n$ or the test space $Y_n'$. For a recent review of this topic see Morton (2010). This provides another motivation for constructing a framework for finite elements that includes non-polynomial functions.

Discretization of PDEs expressed in terms of grad, curl and div on polyhedral meshes has long been pursued in the framework of mimetic finite differences, reviewed in Bochev and Hyman (2006). The convergence of such methods has been analysed in terms of related mixed finite elements (Brezzi, Lipnikov and Shashkov 2005). For connections with finite volume methods see Droniou, Eymard, Gallouët and Herbin (2010). While recent developments tend to exhibit similarities and even equivalences between all these methods, the FE method, as it is understood here, could be singled out by its emphasis on defining fields inside cells, and ensuring continuity properties between them, in such a way that the main discrete differential operators acting on discrete fields are obtained simply by restriction of the continuous ones.

Ciarlet's definition of a finite element does not capture the fact that mixed finite elements behave well with respect to restriction to faces of elements. From the opposite point of view, once this is realized, it becomes natural to prove properties of mixed finite elements by induction on the dimension of the cell.

The goal is then to construct a framework for FE spaces of differential forms on cellular complexes accommodating arbitrary functions. One requires stability of the ansatz spaces under restriction to subcells and under the exterior derivative. For the good properties of standard spaces to be preserved, one must impose additional conditions on the ansatz spaces, essentially surjectivity of the restriction from the cell to the boundary, and sequence exactness under the exterior derivative on each cell. As it turns out, these conditions, which we refer to as compatibility, imply the existence of interpolation operators commuting with the exterior derivative. Degrees of freedom are not part of the definition of compatible finite element systems but are rather deduced, and it becomes more natural to compare various degrees of freedom for a given system. The local properties of surjectivity and sequence exactness also ensure global topological properties by the general methods of algebraic topology.

Interpolation operators still lack desired continuity properties. But combining them with a smoothing technique yields commuting projection operators that are stable in $\mathrm{L}^2(S)$. Stable commuting projections were first proposed in Schöberl (2008). Smoothing was achieved by taking averages over perturbations of the grid. Another construction using cut-off and smoothing by convolution on reference macro-elements was introduced in Christiansen (2007). While commutativity was lost, the lack of it was controlled by an auxiliary parameter. As observed in Arnold *et al.* (2006*b*), for quasi-uniform

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

meshes one can simplify these constructions and use smoothing by convolution on the physical domain. If, in the method of Schöberl (2008), one can say that the nodes of the mesh are shaken independently, smoothing by convolution consists in shaking them in parallel. Much earlier, in Dodziuk and Patodi (1976), convergence for the eigenvalue problem for the Hodge–Laplacian, discretized with Whitney forms, was proved using smoothing by the heat kernel. For scalar functions, smoothing by convolution in the FE method has been used at least as far back as Strang (1972) and Hilbert (1973), but Clément interpolation (Clément 1975) seems to have supplanted it. Christiansen and Winther (2008) introduced a space-dependent smoothing operator, commuting with the exterior derivative, allowing for general shape-regular meshes. These constructions also require a commuting extension operator, extending differential forms outside the physical domain.

This section is organized as follows. In Section 5.2, cellular complexes and the associated framework of finite element systems are introduced. Basic examples are included, as well as some constructions like tensor products. Section 5.2 serves to introduce degrees of freedom and interpolation operators on FE systems. In Section 5.3, we construct smoothers and extensions which commute with the exterior derivative and preserve polynomials locally. When combined with interpolators they yield $\mathrm{L}^q(S)$-stable commuting projections for scale-invariant FE systems. In Section 5.4 we apply these constructions to prove discrete Sobolev injections and a translation estimate.

The framework of FE systems was implicit in Christiansen (2008*a*) and made explicit in Christiansen (2009), but we have improved some of the proofs. Upwinding in this context is new, as well as the discussion of interpolation and degrees of freedom. The construction of smoothers and extensions improves that of Christiansen and Winther (2008) by having the additional property of preserving polynomials locally, up to any given maximal degree. The analysis is also extended from $\mathrm{L}^2$ to $\mathrm{L}^q$ estimates, for all finite $q$. This is used in the proof of the Sobolev injection and translation estimate, which are also new (improving Christiansen and Scheid (2011) and Karlsen and Karper (2010)).

### *5.1. Finite element systems*

*Cellular complexes.* For any natural number $k \geq 1$, let $\mathbb{B}^k$ be the closed unit ball of $\mathbb{R}^k$ and $\mathbb{S}^{k-1}$ its boundary. For instance $\mathbb{S}^0 = \{-1, 1\}$. We also put $\mathbb{B}^0 = \{0\}$.

Let $S$ denote a compact metric space. A $k$-dimensional *cell* in $S$ is a closed subset $T$ of $M$ for which there is a Lipschitz bijection $\mathbb{B}^k \to T$ with a Lipschitz inverse. If a cell $T$ is both $k$- and $l$-dimensional then $k = l$. For $k \geq 1$, we denote by $\partial T$ its boundary, the image of $\mathbb{S}^{k-1}$ by the chosen

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

bi-Lipschitz map. Different such maps give the same boundary. The interior of $T$ is by definition $T \setminus \partial T$ (it is open in $T$ but not necessarily in $S$).

**Definition 5.1.** A *cellular complex* is a pair $(S, \mathcal{T})$ where $S$ is a compact metric space and $\mathcal{T}$ is a finite set of cells in $S$, such that the following conditions hold.

- Distinct cells in $\mathcal{T}$ have disjoint interiors.
- The boundary of any cell in $\mathcal{T}$ is a union of cells in $\mathcal{T}$.
- The union of all cells in $\mathcal{T}$ is $S$.

In this situation we also say that $\mathcal{T}$ is a cellular complex on $S$. We first make the following remarks.

**Proposition 5.2.** The intersection of two cells in $\mathcal{T}$ is a union of cells in $\mathcal{T}$.

*Proof.* Let $T, U$ be two cells in $\mathcal{T}$ and suppose $x \in T \cap U$. Choose a cell $T'$ included in $T$ of minimal dimension such that $x \in T'$. Choose also a cell $U'$ included in $U$ of minimal dimension such that $x \in U'$. Suppose neither of the cells $T'$ and $U'$ are points. Then $x$ belongs to the interiors of both $T'$ and $U'$, so that $T' = U'$. Therefore there exists a cell included in both $T$ and $U$ to which $x$ belongs. This conclusion also trivially holds if $T'$ or $U'$ is a point. □

In fact, if $(S, \mathcal{T})$ is a cellular complex, $S$ can be recovered from $\mathcal{T}$ as follows.

**Proposition 5.3.** Let $(S, \mathcal{T})$ be a cellular complex.

- $S$ is recovered as a set from $\mathcal{T}$ as its union.
- The topology of $S$ is determined by the property that a subset $U$ of $S$ is closed if and only if, for any cell $T$ in $\mathcal{T}$, of dimension $k$, the image of $U \cap T$ under the chosen bi-Lipschitz map is closed in $\mathbb{B}^k$.
- A compatible metric on $S$ can be recovered from metrics $d_T$ on each cell $T$ in $\mathcal{T}$ inherited from $\mathbb{B}^k$, for instance by ($\max \emptyset = +\infty$):

$$d(x, y) = \min\{1,\ \max\{d_T(x, y) : x, y \in T \ \text{ and } \ T \in \mathcal{T}\}\}.$$

A *simplicial complex* is a cellular complex in which the intersection of any two cells is a cell (not just a union of cells) and such that the boundary of any cell is split into subcells in the same way as the boundary of a reference simplex is split into sub-simplexes. The reference simplex $\Delta_k$ of dimension $k$ is

$$\left\{(x_0, \ldots, x_k) \in \mathbb{R}^{k+1} : \sum_i x_i = 1 \ \text{ and } \ \forall i \ \ x_i \geq 0\right\}.$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

Its sub-simplexes are parametrized by subsets $J \subseteq \{0, \cdots, k\}$ as the solution spaces

$$\{x \in \Delta_k : \forall i \notin J \quad x_i = 0\}.$$

For each $k \in \mathbb{N}$ the set of $k$-dimensional subsets of $\mathcal{T}$ is denoted by

$$\mathcal{T}^k = \{T \in \mathcal{T} : \dim T = k\}.$$

The *k-skeleton* of $\mathcal{T}$ is the cellular complex consisting of cells of dimension at most $k$:

$$\mathcal{T}^{(k)} = \mathcal{T}^0 \cup \cdots \cup \mathcal{T}^k.$$

The boundary $\partial T$ of any cell $T$ of $\mathcal{T}$ can be naturally equipped with a cellular complex, namely

$$\{T' \in \mathcal{T} : T' \subseteq T \text{ and } T' \neq T\}.$$

We use the same notation for the boundary of a cell and the cellular complex it carries.

A *refinement* of a cellular complex $\mathcal{T}$ on $S$ is a cellular complex $\mathcal{T}'$ on $S$ such that each element of $\mathcal{T}$ is the union of elements of $\mathcal{T}'$. We will be particularly interested in simplicial refinements of cellular complexes.

A cellular *sub*complex of a cellular complex $\mathcal{T}$ on $S$ is a cellular complex $\mathcal{T}'$ on some closed part $S'$ of $S$ such that $\mathcal{T}' \subseteq \mathcal{T}$. For instance, if $T \in \mathcal{T}$ is a cell, its subcells form a subcomplex of $\mathcal{T}$, which we denote by $\tilde{T}$. We have seen that the boundary of any cell $T \in \mathcal{T}$ can be equipped with a cellular complex which is a subcomplex of $\tilde{T}$.

Fix a cellular complex $(S, \mathcal{T})$. In the following we suppose that for each $T \in \mathcal{T}$ of dimension $\geq 1$, the manifold $T$ has been oriented. The *relative orientation* of two cells $T$ and $T'$ in $\mathcal{T}$, also called the *incidence number*, is denoted $o(T, T')$ and defined as follows. For any edge $e \in \mathcal{T}^1$ its vertices are ordered, from say $\dot{e}$ to $\ddot{e}$. Define $o(e, \dot{e}) = -1$ and $o(e, \ddot{e}) = 1$. As for higher-dimensional cells, fix $k \geq 1$. Given $T \in \mathcal{T}^{k+1}$ and $T' \in \mathcal{T}^k$ such that $T' \subseteq T$, we define $o(T, T') = 1$ if $T'$ is outward-oriented compared with $T$ and $o(T, T') = -1$ if it is inward-oriented. For all $T, T' \in \mathcal{T}$ not covered by these definitions we put $o(T, T') = 0$.

For each $k$, let $\mathcal{C}^k(\mathcal{T})$ denote the set of maps $c : \mathcal{T}^k \to \mathbb{R}$. Such maps associate a real number with each $k$-dimensional cell and are called *k-cochains*. The *coboundary* operator $\delta : \mathcal{C}^k(\mathcal{T}) \to \mathcal{C}^{k+1}(\mathcal{T})$ is defined by

$$(\delta c)_T = \sum_{T' \in \mathcal{T}^k} o(T, T') c_{T'}.$$

The space of $k$-cochains has a canonical basis indexed by $\mathcal{T}^k$. The coboundary operator is the operator whose canonical matrix is the incidence matrix $o$, indexed by $\mathcal{T}^{k+1} \times \mathcal{T}^k$. We remark that the coefficients in the sum can be non-zero only when $T' \in \partial T \cap \mathcal{T}^k$.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

**Lemma 5.4.** We have that $\delta\delta = 0$ as a map $\mathcal{C}^k(\mathcal{T}) \to \mathcal{C}^{k+2}(\mathcal{T})$.

*Proof.* See, *e.g.*, Christiansen (2009, Lemma 3.6). □

In other words the family $\mathcal{C}^\bullet(\mathcal{T})$ is a complex, called the cochain complex and represented by

$$0 \to \mathcal{C}^0(\mathcal{T}) \to \mathcal{C}^1(\mathcal{T}) \to \mathcal{C}^2(\mathcal{T}) \to \cdots .$$

When $S$ is a smooth manifold we denote by $\Omega^k(S)$ the space of smooth differential $k$-forms on $S$. Differential forms can be mapped to cochains as follows. Let $S$ be a manifold and $\mathcal{T}$ a cellular complex on $S$. For each $k$ we denote by $\rho^k : \Omega^k(S) \to \mathcal{C}^k(\mathcal{T})$ the de Rham map, which is defined by

$$\rho^k : u \mapsto \left( \int_T u \right)_{T \in \mathcal{T}^k}.$$

**Proposition 5.5.** For each $k$ the following diagram commutes:

$$\begin{array}{ccc} \Omega^k(S) & \xrightarrow{\mathrm{d}} & \Omega^{k+1}(S) \\ \downarrow{\scriptstyle \rho^k} & & \downarrow{\scriptstyle \rho^{k+1}} \\ \mathcal{C}^k(\mathcal{T}) & \xrightarrow{\delta} & \mathcal{C}^{k+1}(\mathcal{T}) \end{array}$$

*Proof.* This is an application of Stokes' theorem. □

Suppose $\mathcal{T}$ is a cellular complex equipped with an orientation (of the cells) and $\mathcal{T}'$ is a cellular refinement also equipped with an orientation (for instance the same complex but with different orientations). For each cell $T \in \mathcal{T}^k$ and each $T' \in \mathcal{T}'^k$ define $\iota(T, T') = \pm 1$ if $T' \subseteq T$ and they have the same/different orientation, and $\iota(T, T') = 0$ in all other cases.

**Proposition 5.6.** The map $\iota : \mathcal{C}^\bullet(\mathcal{T}') \to \mathcal{C}^\bullet(\mathcal{T})$, defined by

$$(\iota u)_T = \sum_{T' \in \mathcal{T}'} \iota(T, T') u_{T'},$$

is a morphism of complexes, meaning that $\iota$ and $\delta$ commute.

*Proof.* See, *e.g.*, Christiansen (2009, Proposition 3.5). □

We also remark that for a manifold $S$ we have the following.

**Proposition 5.7.** The following diagram commutes:

$$\begin{array}{ccc} & \Omega(S) & \\ {\scriptstyle \rho}\swarrow & & \searrow{\scriptstyle \rho} \\ \mathcal{C}(\mathcal{T}') & \xrightarrow{\iota} & \mathcal{C}(\mathcal{T}) \end{array}$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

*Element systems.* For any cell $T$, we denote by $\Omega^k_{s,q}(T)$ the space of differential $k$-forms on $T$ with $\mathrm{W}^{s,q}(T)$ Sobolev regularity, and put $\Omega^k_q(T) = \Omega^k_{0,q}(T)$. Fix $q \in [1, +\infty[$ and define

$$X^k(T) = \{u \in \Omega^k_q(T) : \mathrm{d}u \in \Omega^k_q(T)\}.$$

When $i : T' \to T$ is an inclusion of cells and $u$ is a smooth enough form on $T$ we denote by $u|_{T'} = i^\star u$ the pullback of $u$ to $T'$. Thus we restrict to the subcell and forget about the action of $u$ on vectors not tangent to it. In the topology (5.1), restrictions to subcells $T'$ of codimension one are well-defined, for instance as elements of $\Omega^k_{-1,q}(T')$. When $T$ is a cell in a given cellular complex $\mathcal{T}$ we may therefore set

$$\hat{X}^k(T) = \{u \in X^k(T) : \forall T' \in \mathcal{T} \quad T' \subseteq T \Rightarrow u|_{T'} \in X^k(T')\}.$$

**Definition 5.8.** Suppose $\mathcal{T}$ is a cellular complex. For each $k \in \mathbb{N}$ and each $T \in \mathcal{T}$ we suppose we are given a space $A^k(T) \subseteq \hat{X}^k(T)$ called a differential $k$-element on $T$. We suppose that the exterior derivative induces maps $\mathrm{d} : A^k(T) \to A^{k+1}(T)$ and that if $i : T' \subseteq T$ is an inclusion of cells, pullback induces a map $i^\star : A^k(T) \to A^k(T')$. Such a family of elements is called an *element system.*

A differential element is said to be finite if it is finite-dimensional. A finite element system is an element system in which all the elements are finite.

**Example 5.9.** The spaces $\hat{X}^\bullet(\bullet)$ themselves define an element system. It is far from finite.

**Example 5.10.** Let $U$ be an open subset of a vector space $V$. We denote by $\mathbb{P}_p(U)$ the space of real polynomials of degree at most $p$ on $U$. For $k \geq 1$ the space of alternating maps $V^k \to \mathbb{R}$ is denoted $\mathbb{A}^k(V)$. The space of differential $k$-forms on $U$, which are polynomial of degree at most $p$, is denoted $\mathbb{PA}^k_p(U)$. We identify

$$\mathbb{PA}^k_p(U) = \mathbb{P}_p(U) \otimes \mathbb{A}^k(V) \quad \text{and} \quad \mathbb{PA}^0_p(U) = \mathbb{P}_p(U).$$

Choose a cellular complex where all cells are flat. Choose a function $\pi : \mathcal{T} \times \mathbb{N} \to \mathbb{N}$ and define

$$A^k(T) = \mathbb{PA}^k_{\pi(T,k)}(T).$$

One gets a finite element system when the following conditions are satisfied:

$$T' \subseteq T \Rightarrow \pi(T', k) \geq \pi(T, k) \quad \text{and} \quad \pi(T, k+1) \geq \pi(T, k) - 1.$$

**Example 5.11.** Denote the Koszul operator on vector spaces by $\kappa$. It is the contraction of differential forms by the identity, considered as a vector field:

$$(\kappa u)_x(\xi_1, \ldots, \xi_k) = u_x(x, \xi_1, \ldots, \xi_k).$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

Alternatively one can use the Poincaré operator associated with the canonical homotopy from the identity to the null-map. Let $\mathcal{T}$ be a simplicial complex. Define, for non-zero $p \in \mathbb{N}$,

$$\Lambda_p^k(T) = \{u \in \mathbb{PA}_p^k(T) : \kappa u \in \mathbb{PA}_p^{k-1}(T)\} = \mathbb{PA}_{p-1}^k(T) + \kappa \mathbb{PA}_{p-1}^{k+1}(T).$$

For fixed $p$ we call this the *trimmed* polynomial finite element system of order $p$. The case $p = 1$ corresponds to constructs in Weil (1952) and Whitney (1957). Arbitrary order elements were introduced in Nédélec (1980) for vector fields in $\mathbb{R}^3$. In Hiptmair (1999) these spaces were extended to differential forms. The correspondence between lowest-order mixed finite elements and Whitney forms was pointed out in Bossavit (1988). See Arnold *et al.* (2006*b*) for a comprehensive review. It was usual to start the indexing at $p = 0$ but, as remarked in the preprint of Christiansen (2007), the advantage of letting the lowest order be $p = 1$ is that the wedge product induces maps:

$$\wedge : \Lambda_{p_0}^{k_0}(T) \times \Lambda_{p_1}^{k_1}(T) \to \Lambda_{p_0+p_1}^{k_0+k_1}(T).$$

See also Arnold *et al.* (2006*b*, p. 34). In words, the wedge product respects the grading in $k$ and the filtering in $p$. This observation was useful in the implementation of a scheme for the Yang–Mills equation (Christiansen and Winther 2006).

The first example of a finite element system yields quite useless Galerkin spaces in general, whereas the second one yields good ones. We shall elaborate on this in what follows, starting by defining what the Galerkin space associated with a finite element system is.

For any subcomplex $\mathcal{T}'$ of $\mathcal{T}$ we define $A^k(\mathcal{T}')$ as follows:

$$A^k(\mathcal{T}') = \left\{ u \in \bigoplus_{T \in \mathcal{T}'} A^k(T) : \forall T, T' \in \mathcal{T} \quad T' \subseteq T \Rightarrow u_T|_{T'} = u_{T'} \right\}.$$

Elements of $A^k(\mathcal{T}')$ may be regarded as differential forms defined piecewise, which are continuous across interfaces between cells, in the sense of equal pullbacks. For a cell $T$ its collection of subcells is the cellular complex $\tilde{T}$. Applied to this case, the above definition gives a space canonically isomorphic to $A^k(T)$. We can identify $A^k(\tilde{T}) = A^k(T)$.

An FE system over a cellular complex is an *inverse system* of complexes: for an inclusion of cells there is a corresponding restriction operator. The space $A^\bullet(\mathcal{T}')$ defined above is an *inverse limit* of this system and is determined by this property up to unique isomorphism. We also point out that this kind of construction, involving glueing of polynomial differential forms on cellular complexes, has been used to study homotopy theory (Griffiths and Morgan 1981).

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

Of particular importance is the application of the above construction to the boundary $\partial T$ of a cell $T$, considered as a cellular complex consisting of all subcells of $T$ except $T$ itself. Considering $\partial T$ as a cellular complex (not only a subset of $T$), we denote the constructed space by $A^k(\partial T)$. If $i : \partial T \to T$ denotes the inclusion map, the pullback by $i$ defines a map $i^\star : A^k(T) \to A^k(\partial T)$, which we denote by $\partial$ and call restriction.

*Conventions.* In the following, the arrows starting or ending at 0 are the only possible ones. Arrows starting at $\mathbb{R}$ are, unless otherwise specified, the maps taking a value to the corresponding constant function. Arrows ending at $\mathbb{R}$ are integration of forms of maximal degree. Other unspecified arrows are instances of the exterior derivative.

Consider now the following conditions on an element system $A$ on a cellular complex $\mathcal{T}$.

- *Extensions.* For each $T \in \mathcal{T}$ and $k \in \mathbb{N}$, restriction $\partial : A^k(T) \to A^k(\partial T)$ is onto.
- *Exactness.* The following sequence is exact for each $T$:

$$0 \to \mathbb{R} \to A^0(T) \to A^1(T) \to \cdots \to A^{\dim T}(T) \to 0. \tag{5.3}$$

The first condition can be written symbolically as $\partial A^k(T) = A^k(\partial T)$.

**Definition 5.12.** We will say that an element system *admits extensions* if the first property holds, is *locally exact* if the second condition holds and is *compatible* if both hold.

Given a finite element system $A$, we say that its points carry reals if, for each *point* $T \in \mathcal{T}^0$, $A^0(T)$ contains the constant maps $T \to \mathbb{R}$ (so that $A^0(T) = \mathbb{R}^T \approx \mathbb{R}$).

**Proposition 5.13.** If $A$ admits extensions and its points carry reals, then for each cell $T$, $A^{\dim T}(T)$ contains a form with non-zero integral.

*Proof.* By induction on the dimension of the cell, using Stokes' theorem. □

*Notation.* We let $A_0^k(T)$ denote the kernel of $\partial : A^k(T) \to A^k(\partial T)$.

**Proposition 5.14.** We have

$$\dim A^k(\mathcal{T}) \leq \sum_{T \in \mathcal{T}} \dim A_0^k(T),$$

with equality when the finite element system admits extensions.

*Proof.* For a given $m \geq 0$, let $\mathcal{T}^{(m)}$ be the $m$-skeleton of $\mathcal{T}$. We have a sequence:

$$0 \to \bigoplus_{T \in \mathcal{T}^m} A_0^k(T) \to A^k(\mathcal{T}^{(m)}) \to A^k(\mathcal{T}^{(m-1)}) \to 0.$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

The second arrow is bijective onto the kernel of the third. If all cells of dimension $m$ admit extensions, the whole sequence is exact. The proposition follows from applying these remarks for all $m$. □

Combining Propositions 5.13 and 5.14, we get the following.

**Corollary 5.15.** If $A$ admits extensions and its points carry reals, then $A^k(T)$ has dimension at least the number of $k$-dimensional subcells of $T$.

**Proposition 5.16.** When the element system is compatible, the de Rham map $\rho^\bullet : A^\bullet(\mathcal{T}) \to \mathcal{C}^\bullet(\mathcal{T})$ induces isomorphisms in cohomology.

*Proof.* We use induction on the dimension of $\mathcal{T}$. In dimension 0 it is clear. Suppose now that $m \geq 1$, and that we have proved the theorem when $\dim \mathcal{T} < m$. Suppose that $\mathcal{T}$ has dimension $m$.

Note that the de Rham map gives isomorphisms in cohomology on cells,

$$A^\bullet(T) \to \mathcal{C}^\bullet(\tilde{T}),$$

since both complexes are acyclic.

Denote by $\mathcal{U}$ the $(m-1)$-skeleton of $\mathcal{T}$. Consider the diagram:

$$\begin{array}{ccccccccc}
0 & \longrightarrow & A^\bullet(\mathcal{T}) & \longrightarrow & A^\bullet(\mathcal{U}) \bigoplus_{T\in\mathcal{T}^m} A^\bullet(T) & \longrightarrow & \bigoplus_{T\in\mathcal{T}^m} A^\bullet(\partial T) & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \longrightarrow & \mathcal{C}^\bullet(\mathcal{T}) & \longrightarrow & \mathcal{C}^\bullet(\mathcal{U}) \bigoplus_{T\in\mathcal{T}^m} \mathcal{C}^\bullet(\tilde{T}) & \longrightarrow & \bigoplus_{T\in\mathcal{T}^m} \mathcal{C}^\bullet(\partial T) & \longrightarrow & 0
\end{array}$$

The vertical maps are de Rham maps. The second horizontal arrow consists of restricting to the summands whereas the third one consists of restricting and comparing, as in the Mayer–Vietoris sequence.

Both rows are exact sequences of complexes, the diagram commutes, and the last two vertical arrows induce isomorphisms in cohomology by the induction hypothesis. Write the long exact sequences of cohomology groups associated with both rows (*e.g.*, Christiansen (2009, Theorem 3.1)) and connect them with the induced morphisms. Applying the five lemma (*e.g.*, Christiansen (2009, Lemma 3.2)) gives the result for $\mathcal{T}$. □

**Proposition 5.17.** For an element system with extensions the exactness of (5.3) on each $T \in \mathcal{T}$ is equivalent to the combination of the following two conditions.

- For each $T \in \mathcal{T}$, $A^0(T)$ contains the constant functions.
- For each $T \in \mathcal{T}$, the following sequence (with boundary condition) is exact:

$$0 \to A_0^0(T) \to A_0^1(T) \to \cdots \to A^{\dim T}(T) \to \mathbb{R} \to 0. \tag{5.4}$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

*Proof.* When the extension property is satisfied on all cells, both versions (with and without boundary condition) of the cohomological condition guarantee that for each $T$, $A^{\dim T}(T)$ contains a form with non-zero integral, by Proposition 5.13.

When the spaces $A^0(T)$ all contain the constant functions, we may consider the following diagrams, where $T$ is a cell of dimension $m$ and $A^m_-(T)$ denotes the space of forms with zero integral:

$$\begin{array}{ccccccccccc}
0 & \longrightarrow & \mathbb{R} & \longrightarrow & A^0(\partial T) & \longrightarrow & \cdots & \longrightarrow & A^{m-1}(\partial T) & \longrightarrow & \mathbb{R} & \longrightarrow & 0 \\
 & & \uparrow & & \uparrow & & \uparrow & & \uparrow & & \uparrow & & \\
0 & \longrightarrow & \mathbb{R} & \longrightarrow & A^0(T) & \longrightarrow & \cdots & \longrightarrow & A^{m-1}(T) & \longrightarrow & A^m(T) & \longrightarrow & 0 \\
 & & \uparrow & & \uparrow & & \uparrow & & \uparrow & & \uparrow & & \\
 & & 0 & \longrightarrow & A^0_0(T) & \longrightarrow & \cdots & \longrightarrow & A^{m-1}_0(T) & \longrightarrow & A^m_-(T) & \longrightarrow & 0
\end{array} \tag{5.5}$$

The columns, extended by 0, are exact if and only if the extension property holds on $T$ and $A^m(T)$ contains a form of non-zero integral. In this case, if one row is exact, the two other rows are either both exact or both inexact.

We now prove the stated equivalence.

**(i)** If compatibility holds then in (5.5) the extended columns are exact, as well as the first and second row, so also the third. Hence (5.4) is exact.

**(ii)** Suppose now exactness of (5.4) holds for each $T$. Choose $m \geq 1$ and suppose that we have proved exactness of (5.3) for cells of dimension up to $m-1$. Let $T$ be a cell of dimension $m$. In (5.5) the extended columns are exact. Apply Proposition 5.16 to the boundary of $T$ to get exactness of the first row. The third row is exact by hypothesis, and we deduce exactness of the second. The induction, whose initialization is trivial, is complete. □

**Corollary 5.18.** $\hat{X}$ is a compatible element system.

*Tensor products.* Suppose we have two manifolds $M$ and $N$, equipped with cellular complexes $\mathcal{U}$ and $\mathcal{V}$. We suppose we have differential elements $A^k(U)$ for $U \in \mathcal{U}$ and $B^k(V)$ for $V \in \mathcal{V}$, both forming systems as defined above.

Let $\mathcal{U} \times \mathcal{V}$ denote the product cellular complex on $M \times N$, whose cells are all those of the form $U \times V$ for $U \in \mathcal{U}$ and $V \in \mathcal{V}$. Recall that the tensor product of differential forms $u$ on $U$ and $v$ on $V$, is the form on $U \times V$ defined as the wedge product of their pullbacks by the respective canonical projections $p_U : U \times V \to U$ and $p_V : U \times V \to V$. In symbols we can write

$$u \otimes v = (p_U^\star u) \wedge (p_V^\star v).$$

We equip $\mathcal{U} \times \mathcal{V}$ with elements

$$C^\bullet(U \times V) = A^\bullet(U) \otimes B^\bullet(V).$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

Explicitly we put

$$C^k(U \times V) = \bigoplus_l A^l(U) \otimes B^{k-l}(V).$$

This defines an element system $C$, called the tensor product of $A$ and $B$.

**Proposition 5.19.** We have

$$C_0^\bullet(U \times V) = A_0^\bullet(U) \otimes B_0^\bullet(V).$$

*Proof.* See Christiansen (2009, Lemma 3.10). □

**Proposition 5.20.** When $A$ and $B$ admit extensions we have

$$C^\bullet(\mathcal{U} \times \mathcal{V}) = A^\bullet(\mathcal{U}) \otimes B^\bullet(\mathcal{V}).$$

*Proof.* The right-hand side is included in the left-hand side. Moreover, by Proposition 5.14,

$$\dim C(\mathcal{U} \times \mathcal{V}) \leq \textstyle\sum_{U,V} \dim C_0(U \times V).$$

On the other hand Proposition 5.19 gives

$$\begin{aligned}\textstyle\sum_{U,V} \dim C_0(U \times V) &= \textstyle\sum_{U,V} \dim A_0(U) \otimes B_0(V) \\ &= \textstyle\sum_{U,V} \dim A_0(U) \dim B_0(V) \\ &= \textstyle\sum_U \dim A_0(U) \sum_V \dim B_0(V) \\ &= \dim A(\mathcal{U}) \dim B(\mathcal{V}).\end{aligned}$$

This completes the proof. □

**Proposition 5.21.** If $A$ and $B$ admit extensions, so does their tensor product.

*Proof.* Consider cells $U \in \mathcal{U}$ and $V \in \mathcal{V}$. Note that

$$\begin{aligned}(\partial U \times V) \cup (U \times \partial V) &= \partial(U \times V), \\ (\partial U \times V) \cap (U \times \partial V) &= \partial U \times \partial V.\end{aligned}$$

The Mayer–Vietoris principle gives an exact sequence,

$$0 \to C(\partial(U \times V)) \to C(\partial U \times V) \oplus C(U \times \partial V) \to C(\partial U \times \partial V) \to 0,$$

where the second and third mappings are

$$\begin{aligned} w &\mapsto w|_{\partial U \times V} \ \oplus \ w|_{U \times \partial V}, \\ u \oplus v &\mapsto u|_{\partial U \times \partial V} - v|_{\partial U \times \partial V}. \end{aligned}$$

It follows that

$$\dim C(\partial(U \times V)) = \dim C(\partial U \times V) + \dim C(U \times \partial V) - \dim C(\partial U \times \partial V).$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

Applying Proposition 5.20 three times, we get

$$\begin{aligned}\dim C(\partial(U\times V)) &= \dim A(\partial U)\dim B(V) + \dim A(U)\dim B(\partial V)\\ &\quad - \dim A(\partial U)\dim B(\partial V)\\ &= \dim A(U)\dim B(V) - \dim A_0(U)\dim B_0(V)\\ &= \dim C(U\times V) - \dim C_0(U\times V)\\ &= \dim \partial C(U\times V).\end{aligned}$$

Therefore we obtain

$$\partial C(U\times V) = C(\partial(U\times V)),$$

as claimed. □

**Proposition 5.22.** If $A$ and $B$ are locally exact, then so is their tensor product.

*Proof.* This follows from the Kunneth theorem (*e.g.*, Christiansen (2009, Theorem 3.2)). □

*Nesting.* Suppose $\mathcal{T}$ is a cellular complex and that $(\Xi, \preceq)$ is an ordered set. Suppose that, for each parameter $\xi \in \Xi$, an FE system $A[\xi]$ on $\mathcal{T}$ has been chosen. We assume that if $\xi \preceq \xi'$ then $A[\xi]^k(T) \subseteq A[\xi']^k(T)$. Choose now a parameter function $\pi : \mathcal{T} \to \Xi$, which is order-preserving in the sense that if $T' \subseteq T$ then $\pi(T') \preceq \pi(T)$. Define an FE system $A[\pi]$ by

$$A[\pi]^k(T) = \{u \in A[\pi(T)]^k(T) : \forall T' \in \mathcal{T} \quad T' \subseteq T \Rightarrow u|_{T'} \in A[\pi(T')]^k(T')\}.$$

**Proposition 5.23.** If, for each $\xi \in \Xi$, the system $A[\xi]$ is compatible, then the constructed system $A[\pi]$ is too.

*Proof.* Note that if $u \in A[\pi]^k(\partial T)$ then $u \in A[\xi(T)]^k(\partial T)$ so it can be extended to an element $u \in A[\xi(T)]^k(T)$. This element is in $A[\pi]^k(T)$. Thus $A[\pi]^k(T)$ admits extensions.

We also have $A[\pi]^k_0(T) = A[\pi(T)]^k_0(T)$, which gives local exactness thanks to Proposition 5.17. □

**Example 5.24.** For a simplicial complex $\mathcal{T}$, let $A[p]$ denote the trimmed finite element system of order $p$. We model variable orders of approximation by a function $\pi : \mathcal{T} \to \mathbb{N}^*$ such that when $T' \subseteq T$ we have $\pi(T') \leq \pi(T)$. The above construction defines an FE system $A[\pi]$, of the type used for *hp*-methods (Demkowicz, Kurtz, Pardo, Paszyński, Rachowicz and Zdunek 2008). We will check later that the trimmed systems for fixed order $p$ are compatible, and then the above result gives compatibility of the variable-order system defined by $\pi$.

*Locally harmonic forms.* Let $T$ be a cell where, for each $k$, $A^k(T)$ is equipped with a scalar product $a$. Orthogonality with respect to $a$ will be denoted

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

by $\perp$. We say that a $k$-form $u$ on $T$ is *$A$-harmonic* if

$$\mathrm{d}u \perp \mathrm{d}A_0^k(T) \quad \text{and} \quad u \perp \mathrm{d}A_0^{k-1}(T). \tag{5.6}$$

One can, for instance, take $a$ to be the $\mathrm{L}^2$ scalar product on differential forms, associated with some Riemannian metric. Denote by $\mathrm{d}^\star$ the formal adjoint of d with respect to this scalar product. The continuous analogue of the above condition (5.6) is

$$\mathrm{d}^\star \mathrm{d}u = 0 \quad \text{and} \quad \mathrm{d}^\star u = 0. \tag{5.7}$$

From the other point of view, (5.6) is the Galerkin variant of (5.7).

**Proposition 5.25.** Let $A$ be a finite element system where each $A^k(T)$ is equipped with a scalar product $a$. Suppose $T$ is a cell such that (5.4) is exact. Put $m = \dim T$.

- For each $\alpha \in \mathbb{R}$, there is a unique $A$-harmonic $u \in A^m(T)$ such that $\int_T u = \alpha$.
- For $k < m$, any $u \in A^k(\partial T)$ admitting an extension in $A^k(T)$, has a unique $A$-harmonic extension in $A^k(T)$.

Let $A$ be a finite element system on $\mathcal{T}$. Define a finite element system $\mathring{A}$ by

$$\mathring{A}^k(T) = \{u \in A^k(T) : \forall T' \in \mathcal{T} \quad T' \subseteq T \;\Rightarrow\; u|_{T'} \text{ is } A\text{-harmonic}\}.$$

We say that $\mathring{A}$ is the subsystem of locally harmonic forms.

**Proposition 5.26.** If $A$ is a compatible FE system then $\mathring{A}$ is a compatible FE system such that the de Rham map $\rho^k : \mathring{A}^k(\mathcal{T}) \to \mathcal{C}^k(\mathcal{T})$ is an isomorphism.

*Proof.* This was essentially proved in Christiansen (2008*a*). □

This construction generalizes Kuznetsov and Repin (2005), in which div-conforming finite elements are defined on polyhedra in $\mathbb{R}^3$. Since $\mathcal{C}^k(\mathcal{T})$ has a canonical basis, the de Rham map determines a corresponding canonical basis of $\mathring{A}^k(\mathcal{T})$. Its elements can be constructed by recursive harmonic extension.

**Example 5.27.** On simplexes, lowest-order trimmed polynomial differential forms are locally harmonic in the sense of (5.7), with respect to the $\mathrm{L}^2$-product associated with *any* piecewise constant Euclidean metric (Christiansen 2008*a*).

**Example 5.28.** Locally harmonic forms can be used to define finite element spaces on the dual cellular complex of a given simplicial one (Buffa and Christiansen 2007). Choose a simplicial refinement of the dual mesh,

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

for instance the barycentric refinement of the primal mesh. Consider Whitney forms on this refinement as a compatible finite element system on the dual mesh. Then take the subsystem of locally harmonic forms. In space dimension $d$, this provides a space of $k$-forms with the same dimension as the space of $(d-k)$-forms on the primal mesh. Duality in the sense of an inf-sup condition was proved in Buffa and Christiansen (2007) with applications to the preconditioning of integral operators appearing in electromagnetics.

Duality methods are quite common in finite volume settings; for a recent development see Andreianov, Bendahmane and Karlsen (2010).

**Example 5.29.** For a given fine mesh one can agglomerate elements into a coarser cellular mesh. The finite element system on the fine mesh then provides a finite element system on the coarse one, with identical function spaces. If the former system is compatible so is the latter. Associated with the latter, we can consider the subsystem of locally harmonic forms. This procedure can be applied recursively: at each level one can consider the locally harmonic forms of the finer level. This yields a multilevel analysis which can be used for multigrid preconditioning (Pasciak and Vassilevski 2008).

Here is a third application, again involving the dual finite elements.

**Example 5.30.** Recall that the incompressible Euler equation can be written as

$$\dot{u} + \operatorname{div}(u \otimes u) + \operatorname{grad} p = 0 \quad \text{and} \quad \operatorname{div} u = 0.$$

Here, $u$ is a vector field with time derivative $\dot{u}$. One uses div-conforming Raviart–Thomas elements for $u$: let $X_h$ denote this space. One uses a weak formulation with locally harmonic curl-conforming elements on the dual grid as test functions: let $Y_h$ denote this space. The $\mathrm{L}^2$ duality and its extensions to Sobolev spaces are denoted $\langle \cdot, \cdot \rangle$. It is invertible on $X_h \times Y_h$, at least in dimension 2, as proved in Buffa and Christiansen (2007). The semi-discrete problem reads as follows. Find a time-dependent $u_h \in X_h$ such that, for all $v_h \in Y_h$ which are orthogonal to the subspace of $Y_n$ of curl-free elements, we have

$$\langle \dot{u}_h, v_h \rangle + \langle \operatorname{div}(u_h \otimes u_h), v_h \rangle = 0.$$

To see that the second bracket is well-defined, note that $\operatorname{div}(u_h \otimes u_h)$ and $\operatorname{div}\operatorname{div}(u_h \otimes u_h)$ both have $\mathrm{W}^{-1,q}$ Sobolev regularity for all $q < 2$. On the other hand $v_h$ and $\operatorname{curl} v_h$ have $\mathrm{W}^{0,q'}$-regularity, for $q' > 2$.

Analogous spaces, with regularity expressed by a second-order operator, have been used for Regge calculus (Christiansen 2008*b*) and elasticity (Schöberl and Sinwel 2007).

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

In computational fluid dynamics it is often important to include some form of upwinding in the numerical method to obtain stability. A model problem for this situation is the equation

$$\epsilon \Delta u + V \cdot \operatorname{grad} u = f, \tag{5.8}$$

where the field $u \in \mathrm{H}^1(S)$ satisfies homogeneous Dirichlet boundary condition, $f$ is a given forcing term, whereas $V$ is a given flow field, which we take to be divergence-free. One is interested in the asymptotic behaviour for small positive $\epsilon$ (viscosity). Such problems can for instance be solved with Petrov–Galerkin methods. Consider a cellular complex and a large compatible finite element system on it, obtained for instance by refinement. We let one space (say the trial space) consist of locally harmonic forms for the standard $\mathrm{L}^2$-product. For the other space (say the test space) we use locally harmonic forms for a weighted $\mathrm{L}^2$-product.

To motivate our choice of weight we introduce some more notions of differential geometry. Given a 1-form $\alpha$ we define the covariant exterior derivative:

$$\mathrm{d}_\alpha : u \mapsto \mathrm{d}u + \alpha \wedge u,$$

These operators do not form a complex, but we have

$$\mathrm{d}_\alpha \mathrm{d}_\alpha u = (\mathrm{d}\alpha) \wedge u.$$

In gauge theory the term $\mathrm{d}\alpha$ is called curvature. Supposing that $\alpha = \mathrm{d}\beta$ for a function $\beta$, we have

$$\mathrm{d}_\alpha = \exp(-\beta)\, \mathrm{d} \exp(\beta) u.$$

One says that $u \mapsto \exp(\beta)u$ is a gauge transformation.

We suppose the domain $S$ is equipped with a Riemannian metric. It provides in particular an $\mathrm{L}^2$ scalar product on differential forms. We denote by $\mathrm{d}_\alpha^\star$ the formal adjoint of d. When $\alpha = \exp(-\beta)$, we have

$$\mathrm{d}_\alpha^\star u = \exp(\beta) \mathrm{d}^\star \exp(-\beta).$$

A natural generalization of (5.7) is

$$\mathrm{d}_\alpha^\star \mathrm{d} u = 0 \quad \text{and} \quad \mathrm{d}_\alpha^\star u = 0,$$

which can be written as

$$\mathrm{d}^\star \exp(-\beta) \mathrm{d} u = 0 \quad \text{and} \quad \mathrm{d}^\star \exp(-\beta) u = 0. \tag{5.9}$$

**Example 5.31.** To address (5.8) define the 1-form $\alpha$ by

$$\alpha(\xi) = -\epsilon^{-1} V \cdot \xi,$$

and note that (5.8) can be rewritten as

$$\mathrm{d}_\alpha^\star \mathrm{d} u = -\epsilon^{-1} f.$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

Given a cellular complex $\mathcal{T}$ on $S$, with flat cells, and a large compatible FE system $A$, we construct two spaces of locally harmonic forms, distinguished by the choice of scalar product $a$. For one (the trial space) take $a$ to be the $\mathrm{L}^2$ scalar product. For the other (the test space), we choose for each $T$ a constant approximation $\alpha_T$ of the pullback of $\alpha$ to $T$. Let $\beta_T$ be the affine function with zero mean on $T$ such that $\mathrm{d}\beta_T = \alpha_T$. For the trial spaces, use the scalar product defined on a cell $T$ by

$$a(u, v) = \int_T \exp(-\beta_T) u \cdot v \tag{5.10}$$

to define the locally harmonic functions. If $v$ is a constant differential form, $u = \exp(\beta_T)v$ satisfies the equations (5.9).

The canonical basis of the test space will then be upwinded or downwinded (depending on the sign in front of $\beta_T$ in (5.10)), compared with the canonical basis of the trial space.

**Example 5.32.** Similar notions can be used to address the Helmholtz equation:

$$\Delta u + k^2 u = 0. \tag{5.11}$$

We wish to construct a compatible FE system over $\mathbb{C}$, which contains a certain number of plane waves:

$$u_\xi : x \mapsto \exp(\mathrm{i}\xi \cdot x).$$

To contain just one of them we remark that for any (flat) cell $T$, if $\xi_T$ is the tangent component of $\xi$ on $T$ we have, on $T$,

$$\Delta_T u_\xi|_T - \mathrm{i}\xi_T \cdot \operatorname{grad}_T u_\xi|_T = 0.$$

Connecting with the previous example, on a cell $T$ we let $\beta_T$ be the affine real function with zero mean and gradient $\xi_T$ on $T$. Define

$$a(u, v) = \int_T \exp(-\mathrm{i}\beta_T) u \cdot v.$$

Here, extra care must be taken because this bilinear form is not positive definite and will in fact be degenerate at interior resonances of the cell. Away from them, the locally harmonic forms for the infinite-dimensional element system $\hat{X}$ (with $q = 2$) are well behaved, in the sense of satisfying Proposition 5.26, and contain the plane wave $u_\xi$.

More generally, suppose one wants to construct an FE system containing a good approximation to a particular solution of (5.11). To the extent that the solution can be locally approximated by a plane wave, the finite element system will contain a good approximation of it, for a choice of a family of functions $\beta_T$, one for each cell $T$, to be determined (maybe adaptively).

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

*5.2. Interpolators*

*Mirrors and interpolators.* The notion of a mirror system formalizes that of degrees of freedom, with particular emphasis on their geometric location.

**Definition 5.33.** A *mirror system* is a choice, for each $k$ and $T$, of a subspace $\mathcal{Z}^k(T)$ of $\hat{X}^k(T)^\star$, called a $k$-mirror on $T$.

Any $k$-form $u$ in $\hat{X}^k(T)$ then gives a linear form $\langle \cdot, u\rangle$ on $\mathcal{Z}^k(T)$, which we call the mirror image of $u$. For a global $k$-form $u$ the mirror images can be collected into a single object. We define

$$\Phi^k u = \langle \cdot, u|_T\rangle_{T\in\mathcal{T}} \in \mathcal{Z}^k(\mathcal{T})^\star,$$

where, for any subcomplex $\mathcal{T}'$ of $\mathcal{T}$, we define a (global) mirror $\mathcal{Z}^k(\mathcal{T}')$ by

$$\mathcal{Z}^k(\mathcal{T}') = \bigoplus_{T\in\mathcal{T}'} \mathcal{Z}^k(T).$$

We say that a mirror system is *faithful* to an element system $A$ if, for any subcomplex $\mathcal{T}'$, restricting $\Phi$ to $\mathcal{T}'$ determines an isomorphism:

$$\Phi^k(\mathcal{T}') : A^k(\mathcal{T}') \to \mathcal{Z}^k(\mathcal{T}')^\star.$$

**Example 5.34.** The canonical mirror system for the trimmed polynomial FE system of order $p$ is the following, where $\dim T = m$:

$$\mathcal{Z}^k(T) = \{u \mapsto \textstyle\int_T v\wedge u \ : \ v \in \mathbb{PA}^{m-k}_{p-m+k-1}(T)\}.$$

It follows from results in Arnold *et al.* (2006*b*) that it is faithful.

**Proposition 5.35.** When $A$ admits extensions, a given mirror system $\mathcal{Z}$ is faithful if and only if the duality product on $\mathcal{Z}^k(T)\times A_0^k(T)$ is invertible for each $k$ and $T$.

*Proof.* **(i)** Suppose $\mathcal{Z}$ is faithful. By induction on dimension, $\dim \mathcal{Z}^k(T) = \dim A_0^k(T)$. Moreover, $\Phi^k(T)$ induces an injection $A_0^k(T)\to\mathcal{Z}^k(T)^\star$. Thus duality on $\mathcal{Z}^k(T)\times A_0^k(T)$ is invertible.

**(ii)** Suppose duality on $\mathcal{Z}^k(T)\times A_0^k(T)$ is invertible for all $k$ and $T$. Then

$$\dim A^k(\mathcal{T}') = \textstyle\sum_{T\in\mathcal{T}'} \dim A_0^k(T) = \dim \mathcal{Z}^k(\mathcal{T}')^\star.$$

Moreover, $\Phi^k(\mathcal{T}')$ is injective. Indeed, if $\Phi^k u = 0$ then for $T\in\mathcal{T}$, $u|_T$ is proved to be 0 by starting with cells $T$ of dimension $k$, and incrementing cell dimension inductively using $u|_{\partial T} = 0$. □

**Definition 5.36.** For a finite element system $A$, an *interpolator* is a collection of projection operators $I^k(T) : \hat{X}^k(T)\to A^k(T)$, one for each $k\in\mathbb{N}$ and $T\in\mathcal{T}$, which commute with restrictions to subcells.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

One can then denote it simply with $I^\bullet$ and extend it unambiguously to any subcomplex $\mathcal{T}'$ of $\mathcal{T}$. Any faithful mirror system defines an interpolator by

$$\Phi^k I^k u = \Phi^k u.$$

We call this the interpolator associated with the mirror system.

**Proposition 5.37.** The following are equivalent.

- $A$ admits extensions.
- $A$ has a faithful mirror system.
- $A$ can be equipped with an interpolator.

*Proof.* **(i)** Suppose $A$ admits extensions. For each $k$ and $T$, choose a closed supplementary of $A_0^k(T)$ in $\hat{X}^k(T)$ and let $\mathcal{Z}^k(T)$ be its annihilator. The duality product on $\mathcal{Z}^k(T) \times A_0^k(T)$ is then invertible for each $k$ and $T$, so that $\mathcal{Z}$ is faithful to $A$.

**(ii)** As already stated, any faithful mirror system defines an interpolator.

**(iii)** Suppose $A$ has an interpolator. If $u \in A^k(\partial T)$ extend it in $\hat{X}^k(T)$ and interpolate it to get an extension in $A^k(T)$. □

**Example 5.38.** In particular, the trimmed polynomial FE system of order $p$ is compatible. As remarked in Christiansen (2010), it is minimal among compatible finite element systems containing polynomial differential forms of order $p-1$.

For a given element system $A$, admitting extensions, it will be useful to construct extension operators $A^k(\partial T) \to A^k(T)$, *i.e.*, linear left inverses of the restriction. One also remarks that a faithful mirror system determines a particular extension. Namely, to $u \in A^k(\partial T)$ one associates the unique $v \in A^k(T)$ extending $u$ and such that, for all $l \in \mathcal{Z}^k(T)$, $l(v) = 0$.

**Proposition 5.39.** Let $A$ be an element system on $\mathcal{T}$ admitting extensions. Let $\mathcal{M}$ denote the set of mirror systems that are faithful to $A$, let $\mathcal{I}$ be the set of interpolators onto $A$ and let $\mathcal{E}$ be the set of extensions in $A$. The natural map $\mathcal{M} \to \mathcal{I} \times \mathcal{E}$ is bijective.

*Proof.* For a given interpolator $I \in \mathcal{I}$ and $E \in \mathcal{M}$, define a mirror system $z(I, E)$ as follows. For $k \in \mathbb{N}$ and $T \in \mathcal{T}$ consider the map

$$Q_T^k = (\mathrm{id} - E\partial) \circ I : \hat{X}^k(T) \to \hat{X}^k(T).$$

It is a projector with range $A_0^k(T)$. Let $z(I, E)^k(T)$ denote the annihilator of its kernel. In other words $l \in z(I, E)^k(T)$ if and only if

$$\forall u \in \hat{X}^k(T) \quad Iu = EI\partial u \;\Rightarrow\; l(u) = 0.$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

We claim that $z$ inverts the given map $a : \mathcal{M} \to \mathcal{I} \times \mathcal{E}$.

**(i)** Given $(I, E)$, we check that the interpolator and extension defined by $z(I, E)$ are $I$ and $E$.

- Pick $u \in \hat{X}^k(T)$. We have $I(u - Iu) = 0$ and $EI\partial(u - Iu) = 0$, hence $l(u - Iu) = 0$. The interpolator deduced from $z(I, E)$ is thus $I$.
- Pick $u \in A^k(\partial T)$. We have $IEu = EI\partial Eu$, hence for all $l \in z(I, E)^k(T)$ we have $l(u) = 0$. The extension deduced from $z(I, E)$ is thus $E$.

**(ii)** Given a faithful mirror system $\mathcal{Z}$ with associated interpolators $I$ and $E$, we check that $z(I, E) = \mathcal{Z}$.

Pick $l \in \mathcal{Z}^k(T)$. If $u \in \hat{X}^k(T)$ is such that $Iu = EI\partial u$ then $l(u) = l(Iu) = l(EI\partial u) = 0$. Hence $l \in z(I, E)^k(T)$. On the other hand $z(I, E)$ is also faithful to $A$ by Proposition 5.35. The inclusion $\mathcal{Z}^k(T) \subseteq z(I, E)^k(T)$ then implies equality. □

**Example 5.40.** Let $A$ be a finite element system. Equip each $\hat{X}^k(T)$ with a continuous bilinear form $a$ which is non-degenerate on $A_0^k(T)$ (*e.g.*, $\hat{X}^k(T)$ is continuously embedded in a Hilbert space). One can define a mirror system by

$$\mathcal{Z}^k(T) = \{a(\cdot, v) : v \in A_0^k(T)\}.$$

When the FE system admits extensions, the associated interpolator can be interpreted as a recursive $a$-projection, starting from cells of minimal dimension, and continuing by incrementing dimension at each step, projecting with respect to $a$ with given boundary conditions.

*Commuting interpolators.* It is of interest to construct interpolators that commute with the exterior derivative. When $T' \subseteq T$, restriction maps $\hat{X}^k(T) \to \hat{X}^k(T')$, so that a $k$-mirror $\mathcal{Z}^k(T')$ on $T'$ can also be considered as a $k$-mirror on $T$. If the mirror system is faithful it must be in direct sum with $\mathcal{Z}^k(T)$.

**Proposition 5.41.** An interpolator commutes with the exterior derivative if and only if its mirror system satisfies

$$\forall T \in \mathcal{T}\ \forall l \in \mathcal{Z}^k(T) \quad l \circ \mathrm{d} \in \mathcal{Z}^{k-1}(\tilde{T}). \tag{5.12}$$

*Proof.* (*If*) Pick $u \in \hat{X}^{k-1}(T)$. Suppose first that $\Phi^{k-1}u = 0$. Then, for all $l \in \mathcal{Z}^k(\tilde{T})$, $l(\mathrm{d}u) = 0$, hence $\Phi^k(\mathrm{d}u) = 0$. In the general case, since $\Phi^{k-1}(u - \Phi^{k-1}u) = 0$, we deduce $\Phi^k(\mathrm{d}u - \mathrm{d}\Phi^{k-1}u) = 0$, so that $\Phi^k \mathrm{d}u = \mathrm{d}\Phi^{k-1}u$.

(*Only if*) Suppose $l \in \mathcal{Z}^k(T)$ and $l \circ \mathrm{d} \notin \mathcal{Z}^{k-1}(\tilde{T})$. Pick $u \in \hat{X}^{k-1}(T)$ such that $l(\mathrm{d}u) \neq 0$, but for all $l' \in \mathcal{Z}^{k-1}(\tilde{T})$, $l'(u) = 0$. Then $I^k \mathrm{d}u \neq 0$ but $I^{k-1}u = 0$ (so $\mathrm{d}I^{k-1}u = 0$). □

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

**Example 5.42.** The canonical mirror system of trimmed polynomials of order $p$ yields a commuting interpolator.

Suppose that $\mathcal{Z}$ is a mirror system on $\mathcal{T}$ such that (5.12) holds. Then we have a well-defined map $\hat{\mathrm{d}} : l \mapsto l \circ \mathrm{d}$ from $\mathcal{Z}^k(\mathcal{T})$ to $\mathcal{Z}^{k-1}(\mathcal{T})$. Denote by $\delta$ its adjoint, which maps from $\mathcal{Z}^{k-1}(\mathcal{T})^\star$ to $\mathcal{Z}^k(\mathcal{T})^\star$.

**Remark 5.43.** The following diagram commutes:

$$\begin{array}{ccc} \hat{X}^{k-1}(\mathcal{T}) & \xrightarrow{\mathrm{d}} & \hat{X}^k(\mathcal{T}) \\ \downarrow{\scriptstyle \Phi^{k-1}} & & \downarrow{\scriptstyle \Phi^k} \\ \mathcal{Z}^{k-1}(\mathcal{T})^\star & \xrightarrow{\delta} & \mathcal{Z}^k(\mathcal{T})^\star \end{array}$$

*Proof.* Pick $l \in \mathcal{Z}^k(\mathcal{T})$ and $u \in \hat{X}^{k-1}(\mathcal{T})$. We have

$$(\delta\Phi u)(l) = (\Phi u)(\hat{\mathrm{d}} l) = (\Phi u)(l \circ \mathrm{d}) = l(\mathrm{d}u) = (\Phi \mathrm{d} u)(l).$$

This concludes the proof. □

**Proposition 5.44.** Equip each $\hat{X}^k(T)$ with a continuous scalar product $a$. For any compatible finite element system $A$, the following is a faithful mirror system yielding a commuting interpolator. For $k = \dim T$,

$$\mathcal{Z}^k(T) = \{a(\cdot, v) : v \in \mathrm{d}A_0^{k-1}(T)\} + \{\mathbb{R} \textstyle\int \cdot\}.$$

For $k < \dim T$,

$$\mathcal{Z}^k(T) = \{a(\cdot, v) : v \in \mathrm{d}A_0^{k-1}(T)\} + \{a(\mathrm{d}\cdot, v) : v \in \mathrm{d}A_0^k(T)\}.$$

This is the natural generalization, to the adopted setting, of projection-based interpolation, as defined in Demkowicz and Babuška (2003) and Demkowicz and Buffa (2005). When the scalar products $a$ are all the $\mathrm{L}^2$-product on forms, we call it *harmonic interpolation*.

Suppose $\mathcal{U}$ is a cellular complex on $M$ and $\mathcal{V}$ is a cellular complex on $N$ giving rise to the product complex $\mathcal{U} \times \mathcal{V}$ on $M \times N$. If $\mathcal{Z}$ is a mirror system on $\mathcal{U}$ and $\mathcal{Y}$ is a mirror system on $\mathcal{V}$, then $\mathcal{Z} \otimes \mathcal{Y}$ is a mirror system on $\mathcal{U} \times \mathcal{V}$ defined by

$$(\mathcal{Z} \otimes \mathcal{Y})^k(U \times V) = \bigoplus_l \mathcal{Z}^l(U) \otimes \mathcal{Y}(V)^{k-l}.$$

**Proposition 5.45.** The tensor product of two faithful mirror systems is faithful.

*Proof.* We have three FE systems, $A$, $B$ and $A \otimes B$, all admitting extensions. Moreover, by Proposition 5.19,

$$(A \otimes B)_0^k(U \times V) = \bigoplus_l A_0^l(U) \otimes B_0^{k-l}(V).$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

Now apply Proposition 5.35. □

*Extension–projection interpolators.* It will also be of interest to construct extension operators which commute with the exterior derivative. More precisely, for a cell of dimension $m$ a commuting extension operator is a family of operators $E^k : A^k(\partial T) \to A^k(T)$ for $0 \le k \le m-1$, together with a map $E^m : \mathbb{R} \to A^m(T)$ such that the following diagram commutes:

$$\begin{array}{ccccccccccc}
0 & \longrightarrow & \mathbb{R} & \longrightarrow & A^0(\partial T) & \longrightarrow & \cdots & \longrightarrow & A^{m-1}(\partial T) & \longrightarrow & \mathbb{R} & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow{\scriptstyle E^0} & & \downarrow & & \downarrow{\scriptstyle E^{m-1}} & & \downarrow{\scriptstyle E^m} & & \\
0 & \longrightarrow & \mathbb{R} & \longrightarrow & A^0(T) & \longrightarrow & \cdots & \longrightarrow & A^{m-1}(T) & \longrightarrow & A^m(T) & \longrightarrow & 0
\end{array} \tag{5.13}$$

When $A^{m-1}(\partial T)$ has an element with non-zero integral an $E^m$ such that the above diagram commutes, is uniquely determined by $E^{m-1}$ and exists if and only if $E^{m-1}$ sends elements with zero integral to closed forms.

**Proposition 5.46.** Let $A$ be a finite element system where each $A^k(T)$ is equipped with a scalar product $a$. Suppose $T$ is a cell such that (5.4) is exact. Put $m = \dim T$. If $T$ admits extensions, the harmonic extension operators defined by Proposition 5.25 commute in the sense of diagram (5.13).

We suppose that for each cell $T$, extension operators $E : A^\bullet(\partial T) \to A^\bullet(T)$ and projections $P : X^\bullet(T) \to A^\bullet(T)$ have been defined. Then $E$ and $P$ uniquely determine an interpolator $J$ as follows. One constructs $J$ inductively. The initialization on cells of dimension 0 (points) is trivial. Now let $T$ be a cell and suppose that $J$ has been constructed for all cells on its boundary. On $T$ we define, for $u \in \hat{X}^k(T)$ with $k < \dim T$,

$$\begin{aligned} Ju &= EJ\partial u + (\mathrm{id} - E\partial)Pu \\ &= Pu + E(J\partial u - \partial Pu), \end{aligned}$$

and for $k = \dim T$ we simply put

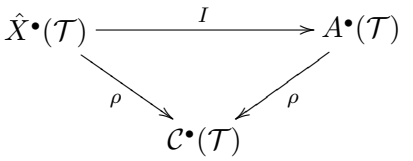
$$Ju = Pu.$$

The only thing to check is that $J$ commutes with restriction from $T$ to cells on the boundary, which is trivial. We call $J$ the associated extension–projection (EP) interpolator.

Let $T$ be cell of dimension $m$. We say that an endomorphism $F$ of $X^m(T)$ preserves integrals if $\int Fu = \int u$ for all $u \in X^m(T)$. An interpolator is said to preserve integrals if it preserves integrals on all cells. For an interpolator $I$, this is equivalent to commutation of the following diagram, involving

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

de Rham maps:




Furthermore, we would like the interpolator to commute with the exterior derivative.

**Proposition 5.47.** Suppose that the projectors $P : X^\bullet(T) \to A^\bullet(T)$ commute with the exterior derivative, that the extensions $E$ commute in the sense of diagram (5.13) and that $P$ preserves integrals. Then the associated EP interpolator commutes with the exterior derivative and preserves integrals.

*Proof.* We use induction on the dimension of cells. Let $T$ be a cell of dimension $m$.

If $u \in \hat{X}^k(T)$ with $k \leq m-2$, the commutation $\mathrm{d}Ju = J\mathrm{d}u$ follows immediately from the commutation of $P$, $E$ and $\partial$, as well as $J$ on $\partial T$ (which follows from the induction hypothesis).

For $u \in \hat{X}^{m-1}(T)$ we have

$$\mathrm{d}Ju = \mathrm{d}Pu + \mathrm{d}E(J\partial u - \partial Pu).$$

But we have

$$\begin{aligned}\int_{\partial T}(J\partial u - \partial Pu) &= \int_{\partial T}\partial u - \int_T \mathrm{d}Pu = \int_{\partial T}\partial u - \int_T P\mathrm{d}u,\\ &= \int_{\partial T}\partial u - \int_T \mathrm{d}u = 0.\end{aligned}$$

The first thing we used is that $J$ preserves integrals on the boundary (which is an induction hypothesis). Hence, by (5.13),

$$\mathrm{d}E(J\partial u - \partial Pu) = 0.$$

Thus

$$\mathrm{d}Ju = \mathrm{d}Pu = P\mathrm{d}u = J\mathrm{d}u.$$

That $J$ preserves integrals on $T$ follows simply from the fact that, for $u \in X^m(T)$, we have $\int Ju = \int Pu = \int u$. □

Compared with the use of mirror systems, the advantage of defining an interpolator from extensions and projections is that approximation properties of the interpolator follow directly from estimates on the extensions and projections. In the following we denote the $\mathrm{L}^q(U)$-norm simply by $\|\cdot\|_U$.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

**Proposition 5.48.** Let $(\mathcal{T}_n)$ be a sequence of cellular complexes, each equipped with a compatible FE system. Suppose that for each $T$, $\tilde{X}^k(T)$ is equipped with a densely defined seminorm $\|\!|\cdot|\!\|_T$. We also suppose that we have functions $\lambda_n : \mathcal{T}_n \to \mathbb{R}_+^*$ and $\tau_n : \mathcal{T}_n \to \mathbb{R}_+^*$ and extensions and projections satisfying

$$\begin{aligned} \|u - P_n u\|_T &\preceq \tau_n(T) \|\!|u|\!\|_T, \\ \|u - P_n u\|_{\partial T} &\preceq \tau_n(T)\lambda_n(T)^{-1} \|\!|u|\!\|_T, \\ \|E_n u\|_T &\preceq \lambda_n(T) \|u\|_{\partial T}. \end{aligned}$$

We suppose that for $T, T' \in \mathcal{T}_n$, if $T' \subseteq T$ then $\lambda_n(T') \simeq \lambda_n(T)$ and $\tau_n(T') \simeq \tau_n(T)$. Then the associated EP interpolator satisfies

$$\|u - I_n u\|_T \preceq \tau_n(T) \sum_{T' \subseteq T} \lambda_n(T)^{\dim T - \dim T'} \|\!|u|\!\|_{T'},$$

where we sum over subcells $T'$ of $T$ in $\mathcal{T}_n$.

We shall express this bound as order optimality.

*Proof.* We use induction. Suppose $T$ is a cell and that $I_n$ is order-optimal on its boundary. For $u \in \hat{X}^k(T)$ with $k < \dim T$, we then have

$$\begin{aligned} \|u - I_n u\|_T &\preceq \|u - P_n u\|_T + \lambda_n(T) \|I_n \partial u - \partial P_n u\|_{\partial T} \\ &\preceq \|u - P_n u\|_T + \lambda_n(T) \|u - P_n u\|_{\partial T} \\ &\qquad + \lambda_n(T) \|u - I_n u\|_{\partial T} \\ &\preceq \tau_n(T) \|\!|u|\!\| + \lambda_n(T) \|u - I_n u\|_{\partial T}. \end{aligned}$$

This completes the proof. □

This proposition was designed with the $p$-version of the finite element method in mind. One can think of $\tau_n(T) = p^{-1}$ and $\lambda_n(T) = p^{-1/q}$. The seminorms involved would correspond to Sobolev spaces, possibly weighted. Thus one would require extension operators whose $\mathrm{L}^q(\partial T) \to \mathrm{L}^q(T)$-norm is of order $p^{-1/q}$. The construction used to prove Proposition 3.3 in the preprint of Christiansen (2007) might be useful here.

### *5.3. Quasi-interpolators*

Let $S$ be a domain in $\mathbb{R}^d$. Let $(\mathcal{T}_n)$ be a sequence of cellular complexes on $S$, each equipped with a compatible FE system $A[n]$. We define $X_n^k = A[n]^k(\mathcal{T}_n)$.

**Definition 5.49.** Consider a sequence of maps $Q_n^k : \mathrm{L}^q(S) \to X_n^k$.

- We say that they are *stable* if, for $u \in \mathrm{L}^q(S)$,

$$\|Q_n^k u\|_{\mathrm{L}^q(S)} \preceq \|u\|_{\mathrm{L}^q(S)}.$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

- We say that they are *order-optimal* if, for $u \in \mathrm{W}^{\ell,q}(S)$,

$$\|u - Q_n^k u\|_{\mathrm{L}^q(S)} \preceq \tau_n^\ell \|u\|_{\mathrm{W}^{\ell,q}(S)},$$

  where $\tau_n^\ell$ is the order of best approximation on $X_n^k$ in the $\mathrm{W}^{\ell,q}(S) \to \mathrm{L}^q(S)$-norm.

- We say that they are *quasi-projections* if, for some $\alpha < 1$, we have for all $u \in X_n^k$

$$\|u - Q_n^k u\|_{\mathrm{L}^q(S)} \leq \alpha \|u\|_{\mathrm{L}^q(S)}.$$

- We say that they *commute* if they commute with the exterior derivative.

We shall construct stable quasi-interpolators of the form

$$Q = IRE,$$

where $I$ is an interpolator, $R$ is a regularization (smoother) approximating the identity and $E$ is an extension operator (usually $R$ requires values outside $S$).

The following technique will be referred to as *scaling*.

**Lemma 5.50.** For a cell $T$ of diameter $h_T$ and barycentre $b_T$, consider the scaling map $\sigma_T : x \to h_T x + b_T$ and let $\hat{T}$ be the pre-image of $T$ by $\sigma_T$, called the reference cell. Let $m$ be the dimension of $T$ ($0 \leq m \leq d$). Let $u$ be a $k$-multilinear form on $T$ and let $\hat{u} = \sigma_T^\star u$ be the pullback of $u$ to $\hat{T}$. Then we have

$$\|u\|_{\mathrm{L}^q(T)} = h_T^{-k+m/q} \|\hat{u}\|_{\mathrm{L}^q(\hat{T})}.$$

We adopt the $h$-setting, in that the differential elements $A[n]^k(T)$, when pulled back to reference domains $\hat{T}$, belong to compact families: see Arnold *et al.* (2006*b*, Remark p. 64). This hypothesis excludes, for instance, methods where the polynomial degree is unbounded as $n \to \infty$, and normally requires the cells to be shape-regular.

**Proposition 5.51.** Suppose the finite elements $A[n]^k(T)$ contain polynomials of degree $\ell - 1$. A sequence of commuting interpolators $I_n$ can be constructed to satisfy

$$\|u - I_n u\|_{\mathrm{L}^q(T)} \preceq \sum_{T' \subseteq T} h_T^{\ell + (\dim T - \dim T')/q} \|\nabla^\ell u\|_{\mathrm{L}^q(T')}, \tag{5.14}$$

where we sum over subcells $T'$ of $T$ in $\mathcal{T}_n$.

We shall express this bound as order optimality.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

*Proof.* Choose commuting interpolators $I_n$, which when pulled back to the reference cell satisfy

$$\|u - \hat{I}_n u\|_{\mathrm{L}^q(\hat{T})} \preceq \sum_{T' \subseteq T} \|\nabla^\ell u\|_{\mathrm{L}^q(\hat{T}')}.$$

Estimate (5.14) follows by scaling. Such interpolators can be constructed from mirror systems (*e.g.*, harmonic interpolation), or from extensions and projections as in Proposition 5.48. □

*Regularizer.* We consider regularizing operators constructed as follows. We require a function $\psi$ supported in the unit ball $\mathbb{B}^d$ and a function

$$\Phi : \begin{cases} \mathbb{B}^d \times \mathbb{R}^d \to \mathbb{R}^d, \\ (y, x) \mapsto \Phi_y(x). \end{cases}$$

For any given $y$ we denote by $\Phi_y^\star$ the pullback by $\Phi_y : \mathbb{R}^d \to \mathbb{R}^d$. That is, for any $k$-multilinear form $u$,

$$(\Phi_y^\star u)[x](\xi_1, \ldots, \xi_k) = u[\Phi_y(x)](\mathrm{D}_x\Phi_y(x)\xi_1, \ldots, \mathrm{D}_x\Phi_y(x)\xi_k).$$

We define

$$Ru = \int_{\mathbb{B}^d} \psi(y)\Phi_y^\star u \,\mathrm{d}y. \tag{5.15}$$

For the function $\psi$ we choose one which is smooth, rotationally invariant, non-negative, with support in the unit ball and with integral 1. Later on, we will require convolution by $\psi$ to preserve polynomials up to a certain degree (see (5.24)), but until then this hypothesis will be irrelevant. In what follows, when we integrate with respect to the variable $y$ it will always be on the unit ball $\mathbb{B}^d$, so we omit this from expressions such as (5.15).

Given such a $\psi$ we are interested in how properties of $\Phi$ reflect upon $R$. For this purpose we call $R$ defined by (5.15) the regularizer associated with $\Phi$. To emphasize its dependence on $\Phi$ we sometimes denote the regularizer by $R[\Phi]$. Formally, $R[\Phi]$ will commute with the exterior derivative, since pullbacks do. The minimal regularity we assume of $\Phi$ is to have continuous second derivatives. We also suppose that, for given $x$, $y \mapsto \Phi_y(x)$ is a diffeomorphism from $\mathbb{B}^d$ to its range. This is enough for the commutation to hold. In what follows we shall be a bit more careful about the regularizing effects of $R[\Phi]$ for given $\Phi$.

Let $\mathbb{B}(x, \delta)$ denote the ball with centre $x$ and radius $\delta$. For any subset $U$ of $\mathbb{R}^d$, its $\delta$-neighbourhood is defined by

$$\mathcal{V}^\delta(U) = \cup\{\mathbb{B}(x, \delta) : x \in U\}.$$

As a first result we state the following.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

**Proposition 5.52.** For any $\delta > 0$ and $C > 0$, there exists $C' > 0$ such that, for all $\Phi$ satisfying, at a point $x$, for all $y \in \mathbb{B}^d$,

$$|\Phi_y(x) - x| \leq \delta,$$

and

$$\|\mathrm{D}_y\Phi_y(x)^{-1}\|, \|\mathrm{D}_x\Phi_y(x)\| \leq C,$$

the associated regularizer satisfies

$$\|(R[\Phi]u)(x)\| \leq C'\|u\|_{\mathrm{L}^1(\mathcal{V}^\delta(x))}. \tag{5.16}$$

*Proof.* With $x$ fixed, the Jacobian of the map

$$\mathbb{B}^d \ni y \mapsto \Phi_y(x) \in \mathcal{V}^\delta(x)$$

has an inverse bounded by $C$. □

We are interested in estimates on derivatives of $R[\Phi]u$ when $u$ is a $k$-form. We have

$$(\nabla R[\Phi]u)(x) = \int \psi(y)\mathrm{D}_x(\Phi_y^\star u)[x]\,\mathrm{d}y, \tag{5.17}$$

where we can substitute the expression

$$\begin{aligned}
&\mathrm{D}_x\Phi_y^\star u[x](\xi_0,\dots,\xi_k) \\
&\quad = \mathrm{D}u[\Phi_y(x)](\mathrm{D}_x\Phi_y(x)\xi_0,\dots,\mathrm{D}_x\Phi_y(x)\xi_k) \\
&\qquad + \textstyle\sum_{i=1}^k u[\Phi_y(x)](\mathrm{D}_x\Phi_y(x)\xi_1,\dots,\mathrm{D}^2_{xx}\Phi_y(x)(\xi_0,\xi_i),\dots,\mathrm{D}_x\Phi_y(x)\xi_k).
\end{aligned} \tag{5.18}$$

The purpose of the following lemma is to get an integral expression for $\nabla R[\Phi]u$ not involving any derivatives of $u$. Essentially, in (5.17), derivatives acting on $u$ are transferred to other terms under the integral sign, using integration by parts. Without the integral sign, this corresponds to identifying the derivatives of $u$ as a 'total' divergence, up to expressions involving no derivatives of $u$.

**Lemma 5.53.** We have

$$\begin{aligned}
&\mathrm{D}_x\Phi_y^\star u[x](\xi_0,\dots,\xi_k) \\
&\quad = \textstyle\sum_{i=1}^k u[\Phi_y(x)](\mathrm{D}_x\Phi_y(x)\xi_1,\dots,\mathrm{D}^2_{xx}\Phi_y(x)(\xi_0,\xi_i),\dots,\mathrm{D}_x\Phi_y(x)\xi_k) \\
&\qquad - \textstyle\sum_{i=1}^k u[\Phi_y(x)](\mathrm{D}_x\Phi_y(x)\xi_1,\dots,\mathrm{D}^2_{yx}\Phi_y(x)(\Xi_y(x)\xi_0,\xi_i),\dots,\mathrm{D}_x\Phi_y(x)\xi_k) \\
&\qquad + \mathrm{D}_y\Phi_y^\star u[x](\Xi_y(x)\xi_0,\xi_1,\dots,\xi_k),
\end{aligned}$$

with $\Xi_y(x)$ defined by

$$\Xi_y(x)\xi = \mathrm{D}_y\Phi_y(x)^{-1}\mathrm{D}_x\Phi_y(x)\xi.$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

We also have

$$\begin{aligned}\psi(y)\mathrm{D}_y\Phi_y^\star u[x](\Xi_y(x)\xi_0,\xi_1,\dots,\xi_k)\\ = \mathrm{div}_y\left(\psi(y)\Xi_y(x)\xi_0\,\Phi_y^\star u[x](\xi_1,\dots,\xi_k)\right)\\ - \mathrm{div}_y\left(\psi(y)\Xi_y(x)^\star\right)\xi_0\,\Phi_y^\star u[x](\xi_1,\dots,\xi_k).\end{aligned}$$

The following proposition shows that the regularizer maps forms of $\mathrm{L}^1$-regularity to continuously differentiable forms, while being careful about the operator norm.

**Proposition 5.54.** For any $\delta > 0$ and $C > 0$, there exists $C' > 0$ such that, for all $\Phi$ satisfying, at a point $x$, for all $y \in \mathbb{B}^d$,

$$|\Phi_y(x) - x| \le \delta,$$

and

$$\left.\begin{array}{c}\|\mathrm{D}_y\Phi_y(x)^{-1}\|,\ \|\mathrm{D}_x\Phi_y(x)\|,\ \|\mathrm{D}^2_{xx}\Phi_y(x)\|\\ \|\mathrm{D}^2_{yx}\Phi_y(x)\|,\ \|\mathrm{D}_y(\mathrm{D}_y\Phi_y(x)^{-1})\|\end{array}\right\} \le C,$$

the associated regularizer satisfies

$$\|(\nabla R[\Phi]u)(x)\| \le C'\|u\|_{\mathrm{L}^1(\mathcal{V}^\delta(x))}. \tag{5.19}$$

*Proof.* The preceding lemma gives an expression for $(\nabla R[\Phi]u)(x)$, from which the claim follows. □

In the estimate (5.19) one has one order more of differentiation on the left-hand side. In the following proposition we consider an equal amount of differentiation on both sides.

**Proposition 5.55.** Pick an integer $\ell \ge 0$, and $\delta > 0$. For any $C > 0$, there exists $C' > 0$ such that, for all $\Phi$ satisfying, at a point $x$, for all $y \in \mathbb{B}^d$,

$$|\Phi_y(x) - x| \le \delta, \tag{5.20}$$

$$\|\mathrm{D}_y\Phi_y(x)^{-1}\| \le C, \tag{5.21}$$

and

$$\|\mathrm{D}_x\Phi_y(x)\|,\ \|\mathrm{D}^2_{xx}\Phi_y(x)\|,\dots,\|\mathrm{D}^{\ell+1}_{x\dots x}\Phi_y(x)\| \le C, \tag{5.22}$$

the associated regularizer satisfies

$$\|(\nabla^\ell R[\Phi]u)(x)\| \le C'\|u\|_{\mathrm{W}^{\ell,1}(\mathcal{V}^\delta(x))}.$$

*Proof.* For $\ell = 0$ one uses the change of variable formula in the definition (5.15), the Jacobian being taken care of by estimate (5.21). The case $\ell = 1$ follows from expression (5.18). Differentiating this expression several times gives the claimed result for general $\ell$ (an expression for the differential of order $\ell$ of a pullback will be given later). □

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

From now on we suppose $\Phi$ has the form

$$\Phi_y(x) = x + \phi(x)y, \tag{5.23}$$

for some function $\phi : S \to \mathbb{R}$. Most of the discussion would work with matrix-valued maps $\phi : S \to \mathbb{R}^{d\times d}$, which could be important for anisotropic meshes. But for simplicity we do not consider anisotropic meshes here, and thus scalar $\phi$ are sufficient. We also assume that $\psi$ satisfies

$$\int \psi(y) f(y)\,\mathrm{d}y = f(0), \tag{5.24}$$

for all polynomials $f$ of degree at most $p + d$, for some integer $p \geq 0$.

The purpose of these hypotheses is to make the regularizer preserve polynomials of degree up to $p$. To see this, first note that property (5.24) guarantees that convolution by $\psi$ preserves polynomials of degree $p+d$. We state the following.

**Proposition 5.56.** Suppose $|\phi(x)| \leq \delta$ for some $\delta > 0$.

If $u$ is also a polynomial of degree at most $p$ on $\mathcal{V}^\delta(x)$, then $\nabla^\ell Ru(x) = \nabla^\ell u(x)$ for all $\ell$.

*Proof.* We have

$$\mathrm{D}_x\Phi_y(x)\xi = \xi + (\mathrm{D}\phi(x)\xi)y.$$

Suppose $u$ is a $k$-form which is a polynomial of degree at most $p$. Recall that

$$(\Phi_y^\star u)[x](\xi_1, \ldots, \xi_k) = u[\Phi_y(x)](\mathrm{D}_x\Phi_y(x)\xi_1, \ldots, \mathrm{D}_x\Phi_y(x)\xi_k).$$

As a function of $y$ this is a polynomial of degree at most $p + k \leq p + d$. Its value at $y = 0$ is

$$u[x](\xi_1, \ldots, \xi_k).$$

This gives the case $\ell = 0$ of the proposition. For $\ell = 1$ one uses expression (5.18). Greater $\ell$ are obtained by further differentiation of this expression. □

To see to what extent Propositions 5.52, 5.54 and 5.55 can be applied, we remark that

$$\begin{aligned}
\Phi_y(x) - x &= \phi(x)y,\\
\mathrm{D}_y\Phi_y(x)\xi &= \phi(x)\xi,\\
\mathrm{D}_y(\mathrm{D}_y\Phi_y(x)^{-1}) &= 0,\\
\mathrm{D}_x\Phi_y(x)\xi &= (\mathrm{D}\phi(x)\xi)y,\\
\mathrm{D}^2_{yx}\Phi_y(x)(\xi,\xi') &= (\mathrm{D}\phi(x)\xi)\xi',\\
\mathrm{D}^\ell_{x\ldots x}\Phi_y(x)(\xi_1,\ldots,\xi_\ell) &= \mathrm{D}^\ell\phi(x)(\xi_1,\ldots,\xi_\ell)\,y.
\end{aligned}$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

**Proposition 5.57.** Pick $\ell \leq p+1$. For any $\delta > 0$ and any $C > 0$, there exists $C' > 0$ such that, for all $\phi$ satisfying, at some point $x$,

$$|\phi(x)| \leq \delta,$$
$$|\phi(x)^{-1}| \leq C,$$

and

$$\|\mathrm{D}_x\phi(x)\|,\ \|\mathrm{D}^2_{xx}\phi(x)\|, \ldots, \|\mathrm{D}^{\ell+1}_{x\ldots x}\phi(x)\| \leq C,$$

the associated regularizer satisfies

$$\|(\nabla^\ell Ru)(x)\| \leq C'\|\nabla^\ell u\|_{\mathrm{L}^1(\mathcal{V}^\delta(x))}.$$

*Proof.* By the Deny–Lions lemma there exists $C > 0$ such that, for all $u \in \mathrm{W}^{\ell,1}(\mathcal{V}^\delta(x))$,

$$\inf_{f\in\mathbb{P}^{\ell-1}} \|u-f\|_{\mathrm{W}^{\ell,1}(\mathcal{V}^\delta(x))} \preceq \|\nabla^\ell u\|_{\mathrm{L}^1(\mathcal{V}^\delta(x))}.$$

The regularizer $R$ preserves the space $\mathbb{P}^{\ell-1}$ of polynomials of degree up to $\ell-1$. Combining this with Proposition 5.55 gives the claimed result. □

The regularizer is adapted to the mesh $\mathcal{T}_n$ as follows. We choose $\phi_n$ such that, for all $x \in T \in \mathcal{T}_n$,

$$\phi_n(x) \simeq h_T, \tag{5.25}$$
$$\|\mathrm{D}^{1+r}\phi_n(x)\| \preceq h_T^{-r}, \text{ for } 0 \leq r \leq \ell. \tag{5.26}$$

We introduce a parameter $\epsilon > 0$ and consider the regularizations $R^\epsilon_n = R[\Phi]$ associated with the maps

$$\Phi_y(x) = x + \epsilon\phi_n(x)y.$$

We define

$$\mathcal{V}^\epsilon_n(T) = \cup\{\mathbb{B}(x, \epsilon\phi_n(x)) : x \in T\}.$$

We choose $\epsilon$ fixed but small.

**Proposition 5.58.** Fix $\epsilon > 0$. For any $T \in \mathcal{T}_n$ we have an estimate

$$h_T^{1+(d-\dim T)/q}\|\nabla R^\epsilon_n u\|_{\mathrm{L}^q(T)} \preceq \|u\|_{\mathrm{L}^q(\mathcal{V}^\epsilon_n(T))}, \tag{5.27}$$
$$h_T^{(d-\dim T)/q}\|\nabla^\ell R^\epsilon_n u\|_{\mathrm{L}^q(T)} \preceq \|\nabla^\ell u\|_{\mathrm{L}^q(\mathcal{V}^\epsilon_n(T))}. \tag{5.28}$$

For $T$ of maximal dimension we also have

$$\|u - R^\epsilon_n u\|_{\mathrm{L}^q(T)} \preceq h_T^\ell\|\nabla^\ell u\|_{\mathrm{L}^q(\mathcal{V}^\epsilon_n(T))}. \tag{5.29}$$

*Proof.* Pick $T \in \mathcal{T}_n$. Its diameter is $h_T$ and its barycentre $b_T$. Consider the scaling map $\sigma_T : x \to h_T x + b_T$ and let $\hat{T}$ be the pre-image of $T$ by $\sigma_T$, called the reference cell.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

Note that, quite generally, the regularization $R$ transforms as follows under pullback by a diffeomorphism $\sigma$:

$$R[\Phi] = (\sigma^\star)^{-1} R[\sigma^{-1} \circ \Phi_\bullet \circ \sigma]\sigma^\star,$$

where

$$\sigma^{-1} \circ \Phi_\bullet \circ \sigma : (y, x) \mapsto \sigma^{-1}(\Phi_y(\sigma(x))).$$

In our case the operator $R_n^\epsilon[\sigma_T^{-1} \circ \Phi_\bullet \circ \sigma_T]$ is regularizing on the reference cell $\hat{T}$ and we denote it by $\hat{R}_n^\epsilon$. We have

$$(\sigma_T^{-1} \circ \Phi_y \circ \sigma_T)(x) = x + \epsilon h_T^{-1} \phi_n(h_T x + b_T) y.$$

The conditions (5.25), (5.26) put us in a position to conclude from Propositions 5.54 and 5.55 that

$$\|\nabla \hat{R}_n^\epsilon u\|_{\mathrm{L}^\infty(\hat{T})} \preceq \|u\|_{\mathrm{L}^1(\sigma_T^{-1}\mathcal{V}_n^\epsilon(T))},$$
$$\|\nabla^\ell \hat{R}_n^\epsilon u\|_{\mathrm{L}^\infty(\hat{T})} \preceq \|\nabla^\ell u\|_{\mathrm{L}^1(\sigma_T^{-1}\mathcal{V}_n^\epsilon(T))}.$$

From this we deduce

$$\|\nabla \hat{R}_n^\epsilon u\|_{\mathrm{L}^q(\hat{T})} \preceq \|u\|_{\mathrm{L}^q(\sigma_T^{-1}\mathcal{V}_n^\epsilon(T))},$$
$$\|\nabla^\ell \hat{R}_n^\epsilon u\|_{\mathrm{L}^q(\hat{T})} \preceq \|\nabla^\ell u\|_{\mathrm{L}^q(\sigma_T^{-1}\mathcal{V}_n^\epsilon(T))}.$$

Then the estimates (5.27) and (5.28) follow from scaling.

Let $T \in \mathcal{T}_n$ have dimension $d$. From Proposition 5.52, we get

$$\|\hat{R}u\|_{\mathrm{L}^q(\hat{T})} \preceq \|u\|_{\mathrm{L}^q(\sigma_T^{-1}\mathcal{V}_n^\epsilon(T))}.$$

Preservation of polynomials and the Deny–Lions lemma then gives

$$\|u - \hat{R}u\|_{\mathrm{L}^q(\hat{T})} \preceq \|\nabla^\ell u\|_{\mathrm{L}^q(\sigma_T^{-1}\mathcal{V}_n^\epsilon(T))}.$$

Scaling then gives (5.29). □

For a cell $T \in \mathcal{T}$ we denote by $\mathcal{M}_n(T)$ the macro-element surrounding $T$ in $\mathcal{T}_n$, that is, the union of the cells $T' \in \mathcal{T}_n$ touching $T$:

$$\mathcal{M}_n(T) = \cup\{T' \in \mathcal{T}_n : T' \cap T \neq \emptyset\}.$$

Choose $\epsilon$ so small that, for all $n$ and all $T \in \mathcal{T}_n$,

$$\mathcal{V}_n^\epsilon(T) \cap S \subseteq \mathcal{M}_n(T).$$

**Proposition 5.59.** For $u$ defined on $\mathcal{V}_n^\epsilon(S)$ we have estimates

$$\|I_n R_n^\epsilon u\|_{\mathrm{L}^q(S)} \preceq \|u\|_{\mathrm{L}^q(\mathcal{V}_n^\epsilon(S)}$$

and, for $\ell \leq p + 1$,

$$\|u - I_n R_n^\epsilon u\|_{\mathrm{L}^q(S)} \preceq h_n^\ell \|u\|_{\mathrm{L}^q(\mathcal{V}_n^\epsilon(S)}.$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

*Extension.* We shall define extension operators which extend differential forms on $S$ to some neighbourhood of $S$, preserve polynomials up to a certain degree $p$, commute with the exterior derivative and are continuous in $\mathrm{W}^{\ell,q}$-norms for $\ell \leq p+1$.

For this purpose we will use maps $\Phi_s$ depending on a parameter $s$, defined outside $S$ with values in $S$, and pull back by these maps. By taking judiciously chosen linear combinations of such pullbacks we meet the requirements of continuity and polynomial preservation.

First we derive some formulas for the derivative of order $\ell$ of the pullback of a differential form. Antisymmetry in the variables will not be important for these considerations, so we consider multilinear rather than differential forms.

Consider then a smooth map $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ and $u$, a $k$-multilinear form on $\mathbb{R}^d$, with $k \geq 1$. We want to give an expression for

$$(\nabla^\ell \Phi^\star u)[x](\xi_1, \ldots, \xi_{k+\ell}),$$

as a linear combination of terms of the form

$$(\nabla^r u)[\Phi(x)]\big(\mathrm{D}^{m_1}\Phi(x)(\zeta_1, \ldots, \zeta_{m_1}), \mathrm{D}^{m_2}\Phi(x)(\zeta_{m_1+1}, \ldots, \zeta_{m_1+m_2}), \ldots\big), \tag{5.30}$$

where $m_1 + \cdots + m_{k+r} = k + \ell$ and

$$(\zeta_1, \ldots, \zeta_{k+\ell}) = (\xi_{\sigma(1)}, \ldots, \xi_{\sigma(k+\ell)}),$$

for any permutation $\sigma$ of the indexes $\{1, \ldots, k+\ell\}$.

To handle the combinatorics behind this problem, we let $\mathcal{D}$ be the set of pairs $(\sigma, m)$ where $m = (m_1, \ldots, m_{v(m)})$ is a multi-index of valence $v(m) \geq 1$ and weight $w(m) = m_1 + \cdots + m_{v(p)}$, such that $m_i \geq 1$ for each $1 \leq i \leq v(m)$, and $\sigma$ is a permutation of the indexes $(1, \ldots, w(m))$. We define the partial weights

$$|m|_0 = 0 \text{ and } |m|_i = m_1 + \cdots + m_i \text{ for } 1 \leq i \leq v(m).$$

We define $(\sigma, m)^\circ[\Phi, x](\xi_1, \ldots, \xi_{w(m)})$ to be the $v(m)$-multivector:

$$\otimes_{i=1}^{v(m)} \mathrm{D}^{m_i}\Phi(x)(\zeta_{|m|_{i-1}+1}, \ldots, \zeta_{|m|_i}),$$

with

$$(\zeta_1, \ldots, \zeta_{w(m)}) = (\xi_{\sigma(1)}, \ldots, \xi_{\sigma(w(m))}).$$

Thus

$$(\sigma, m)^\circ[\Phi, x] \in \mathrm{L}(\otimes^{w(m)}\mathbb{R}^d, \otimes^{v(m)}\mathbb{R}^d). \tag{5.31}$$

Let $\overline{\mathcal{D}}$ be the free group generated by $\mathcal{D}$. We define some operations in $\overline{\mathcal{D}}$, which correspond to differentiating (5.31) with respect to $x$.

For a given $(\sigma, m)$ we define, for $j = 1, \ldots, v(m)$, $\partial_j(\sigma, m)$ to consist of

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

the multi-index $(m_1, \dots, m_j + 1, \dots, m_{v(m)})$ and the permutation

$$1 \le i \le w(m) \mapsto \begin{cases} \sigma(i) & \text{if } \sigma(i) \le |m|_j, \\ \sigma(i) + 1 & \text{if } \sigma(i) > |m|_j, \end{cases}$$
$$w(m) + 1 \mapsto |m|_j + 1.$$

Defining

$$\mathrm{D}(\sigma, m) = \sum_{j=1}^{v(m)} \partial_j(\sigma, m),$$

we have

$$\mathrm{D}((\sigma, m)^\circ[\Phi, x]) = (\mathrm{D}(\sigma, m))^\circ[\Phi, x].$$

On the left, D is ordinary differentiation with respect to $x$, and on the right, D is an operation in $\overline{\mathcal{D}}$.

If $u$ is a $k$-multilinear form and $\zeta = \zeta_1 \otimes \ldots \otimes \zeta_k \in \otimes^k \mathbb{R}^d$, we define the contraction

$$u : \zeta = u(\zeta_1, \dots, \zeta_k).$$

Contraction is bilinear. Given a $k$-multilinear form $u$, we want to differentiate, with respect to $x$, expressions of the form

$$(\nabla^r u)[\Phi(x)] : (\sigma, m)^\circ[\Phi, x],$$

where $v(m) = k+r$. This corresponds to (5.30). Since contraction is bilinear and we know how to differentiate the right-hand side, it remains to differentiate the left-hand side. For this purpose we introduce one more operation in $\overline{\mathcal{D}}$. Define $e$ such that $e(\sigma, m)$ consists of the multi-index $(m_1, \dots, m_{v(m)}, 1)$ and the permutation

$$1 \le i \le w(m) \mapsto \sigma(i),$$
$$w(m) + 1 \mapsto w(m) + 1.$$

Then we have

$$\nabla\big((\nabla^r u)[\Phi(x)] : (\sigma, m)^\circ[\Phi, x]\big) \tag{5.32}$$
$$= (\nabla^{r+1} u)[\Phi(x)] : (e(\sigma, m))^\circ[\Phi, x] + (\nabla^r u)[\Phi(x)] : (\mathrm{D}(\sigma, m))^\circ[\Phi, x].$$

In the free group $\overline{\mathcal{D}}$ we now define $\Gamma^\ell_r[k]$ for $\ell \ge 0$ and $0 \le r \le \ell$ recursively. We initialize by

$$\Gamma^0_0[k] = (\mathrm{id}, (1, \dots, 1)),$$

with the identity permutation and $k$ terms in the multi-index. Then we define for $\ell \ge 0$

$$\Gamma^{\ell+1}_0[k] = e\Gamma^\ell_0[k],$$
$$\Gamma^{\ell+1}_i[k] = e\Gamma^\ell_i[k] + \mathrm{D}\Gamma^\ell_{i-1}[k], \text{ for } 1 \le i \le \ell,$$
$$\Gamma^{\ell+1}_{\ell+1}[k] = \mathrm{D}\Gamma^\ell_\ell[k].$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

We check that

$$\Gamma_0^\ell[k] = \Gamma_0^0[k+\ell],$$
$$\Gamma_\ell^\ell[k] = \mathrm{D}^\ell \Gamma_0^0[k].$$

**Proposition 5.60.** We have, for a given $k$-form $u$,

$$\nabla^\ell(\Phi^\star u)[x] = \sum_{i=0}^{\ell} (\nabla^{\ell-i} u)[\Phi(x)] : \Gamma_i^\ell[k]^\circ[\Phi, x].$$

The expression $\Gamma_i^\ell[k] \in \overline{\mathcal{D}}$ is a sum of terms with valence $k+\ell-i$ and weight $k+\ell$.

*Proof.* We use induction on $\ell$, using (5.32). □

For $x$ outside $S$, let $\delta(x)$ denote the distance from $x$ to $S$.

Given a polynomial degree $p$, we will construct an extension operator $E$ as a linear combination of pullbacks:

$$E = \sum_{s\in I} \psi_s \Phi_s^\star, \text{ with } \Phi_s : \mathcal{V}^\epsilon(S) \setminus S \to S,$$

subject to the following conditions.

- The index set $I$ is a finite subset of the interval $[2,3]$ and the coefficients $(\psi_s)_{s\in I}$ are chosen such that, for any polynomial $f$ of degree at most $p+d$,
$$\sum_{s\in I} \psi_s f(s) = f(0).$$
- There is a function $\phi : \mathcal{V}^\epsilon(S) \setminus S \to \mathbb{R}^d$ such that, for all $s$ and $x$,
$$\Phi_s(x) = x + s\phi(x), \tag{5.33}$$
and, moreover,
$$\|\phi(x)\| \simeq \delta(x),$$
$$\|\mathrm{D}^{1+r}\phi(x)\| \preceq \delta(x)^{-r}, \quad \text{for } 0 \le r \le p.$$
- Finally, for $s \in [2,3]$ we should have
$$\|\mathrm{D}\Phi_s(x)^{-1}\| \preceq 1,$$
and the $\Phi_s$ should determine diffeomorphisms from $\mathcal{V}^\epsilon(S) \setminus S$ for some $\epsilon > 0$, to an interior neighbourhood of $\partial S$.

In order to show that the above list of conditions can be met, we need some results of a general nature that are given in the Appendix.

**Proposition 5.61.** The conditions listed above can be met.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

*Proof.* Let $I$ consist of $p+d+1$ points in the interval $[2,3]$. Numbers $\psi_s$ are then determined by solving a linear system with a Vandermonde matrix.

If $\partial S$ had been smooth, the orthogonal projection $\wp$ onto $\partial S$ would be well-defined and smooth on a neighbourhood. Then we could have taken $\phi(x) = \wp(x) - x$. In the following we modify this construction to allow for Lipschitz boundaries.

Choose a smooth vector field $\nu$ according to Proposition A.3, pointing outwards on $\partial S$, so that for some $\epsilon' > 0$ the following map is a Lipschitz isomorphism onto its open range:

$$g : \begin{cases} \partial S \times ]-\epsilon', \epsilon'[ & \to \mathbb{R}^d, \\ (z,t) & \mapsto z + t\nu(z). \end{cases}$$

Define $f$ on $\mathcal{V}^\epsilon(S) \setminus S$ by

$$f(g(z,t)) = z - g(z,t) = -t\nu(z).$$

The problem with $f$, to serve as $\phi$ in (5.33), is its lack of regularity.

From Theorem 2 in Stein (1970, p. 171) we get a regularized distance function $\tilde{\delta}$ defined outside $S$, such that

$$\tilde{\delta}(x) \simeq \delta(x),$$

and for all $r \geq 1$

$$\|\mathrm{D}^r \tilde{\delta}(x)\| \preceq \delta(x)^{1-r}.$$

We regularize $f$ by a variant of (5.15), (5.23). We put

$$\phi(x) = \int \psi(y) f(x - \epsilon\tilde{\delta}(x)y)\,\mathrm{d}y,$$

where the parameter $\epsilon$ is chosen to satisfy $\epsilon\tilde{\delta}(x) \leq 1/2\delta(x)$, so that $f$ is evaluated far enough from $S$. For an illustration we refer to Figure 5.1.

The conditions are then met. □

In the following we choose an integer $\ell \leq p$.

**Proposition 5.62.** We have

$$\nabla^\ell Eu = \nabla^\ell u[\Phi_2(x)] \tag{5.34a}$$

$$+ \sum_s \psi_s (\nabla^\ell u[\Phi_s(x)] - \nabla^\ell u[\Phi_2(x)]) : \otimes^{k+\ell}(\mathrm{id} + s\mathrm{D}\phi(x)) \tag{5.34b}$$

$$+ \sum_s \psi_s \int_0^1 \nabla^\ell u[t\Phi_s(x) + (1-t)\Phi_2(x)] : \tag{5.34c}$$

$$\sum_{i=1}^{\ell} \otimes^i (s-2)\phi(x) \otimes \Gamma_i^\ell[k]^\circ[\Phi_s, x] \frac{(1-t)^{i-1}}{(i-1)!}\,\mathrm{d}t. \tag{5.34d}$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

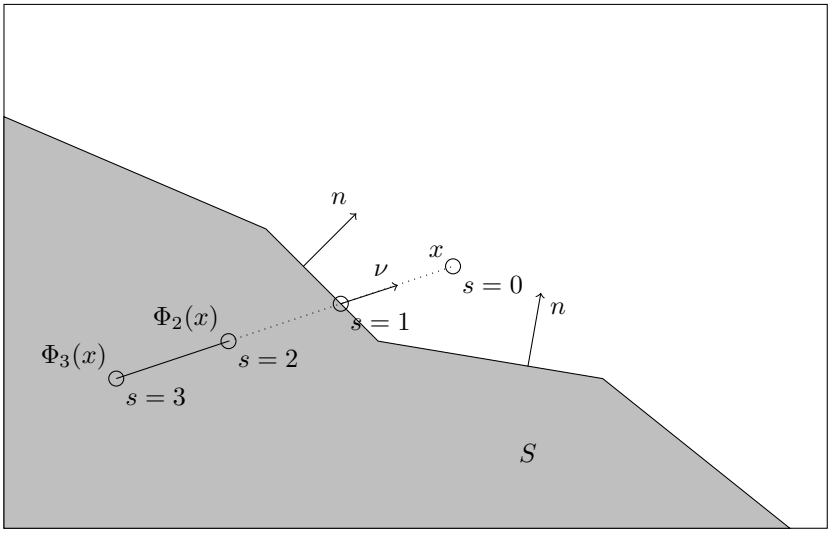



Figure 5.1. Definition of $\Phi_s(x)$, without smoothing.

*Proof.* We have

$$\nabla^\ell(\Phi_s^\star u)[x] = \sum_{i=0}^{\ell} (\nabla^{\ell-i} u)[\Phi_s(x)] : \Gamma_i^\ell[k]^\circ[\Phi_s, x]. \tag{5.35}$$

Concerning $\Gamma_i^\ell[k]^\circ[\Phi_s, x]$, note that

$$\begin{aligned} \mathrm{D}\Phi_s(x) &= \mathrm{id} + s\mathrm{D}\phi(x), \\ \mathrm{D}^r\Phi_s(x) &= s\mathrm{D}^r\phi(x) \text{ for } r \geq 2. \end{aligned}$$

Therefore $\Gamma_i^\ell[k]^\circ[\Phi_s, x]$ is a polynomial in $s$ of degree at most $k + \ell - i$. For $i = 0$ the value in $s = 0$ is

$$\Gamma_0^\ell[k]^\circ[\Phi_s, x]|_{s=0} = \otimes^{k+\ell}(\mathrm{id} + s\mathrm{D}\phi(x))|_{s=0} = \otimes^{k+\ell}\mathrm{id}.$$

For $i \geq 1$ we have a sum of terms that are products where at least one derivative of order at least two appears, so the value at $s = 0$ is 0.

In (5.35) the $i = 0$ term gives rise to

$$\begin{aligned} \sum_s \psi_s \nabla^\ell u[\Phi_s(x)] : \Gamma_0^\ell[k]^\circ[\Phi_s, x] &= \nabla^\ell u[\Phi_2(x)] \\ &+ \sum_s \psi_s (\nabla^\ell u[\Phi_s(x)] - \nabla^\ell u[\Phi_2(x)]) : \otimes^{k+\ell}(\mathrm{id} + s\mathrm{D}\phi(x)). \end{aligned}$$

This corresponds to (5.34a,b).

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

For $i \geq 1$, Taylor's formula with integral remainder gives

$$(\nabla^{\ell-i}u)[\Phi_s(x)] = \sum_{j=0}^{i-1} \nabla^{\ell-i+j}u[\Phi_2(x)] : \otimes^j (s-2)\phi(x)\ \frac{1}{j!} \tag{5.36}$$
$$+ \int_0^1 \nabla^\ell u[t\Phi_s(x) + (1-t)\Phi_2(x)] : \otimes^i (s-2)\phi(x)\ \frac{(1-t)^{i-1}}{(i-1)!}\,\mathrm{d}t.$$

When this expression is contracted with $\Gamma_i^\ell[k]^\circ[\Phi_s, x]$, the sum (5.36) consists of polynomials in $s$ with degree $j + k + \ell - i \leq \ell + d$, with value 0 at $s = 0$. Therefore

$$\sum_s \psi_s (\nabla^{\ell-i}u)[\Phi_s(x)] : \Gamma_i^\ell[k]^\circ[\Phi_s, x]$$
$$= \sum_s \psi_s \int_0^1 \nabla^\ell u[t\Phi_s(x) + (1-t)\Phi_2(x)] :$$
$$\otimes^i (s-2)\phi(x) \otimes \Gamma_i^\ell[k]^\circ[\Phi_s, x] \frac{(1-t)^{i-1}}{(i-1)!}\,\mathrm{d}t.$$

Summing over $1 \leq i \leq \ell$, this corresponds to (5.34c,d). □

From this formula several conclusions can be drawn.

**Proposition 5.63.**

- If $u$ is of class $\mathcal{C}^\ell(S)$, $Eu$ is of class $\mathcal{C}^\ell(\mathcal{V}^\epsilon(S))$.
- If $u$ is a polynomial of degree at most $p$, $Eu$ is too. More precisely, if $T$ is a cell touching $\mathcal{S}$, and $u$ is a polynomial on its macro-element, of degree at most $p$, then $Eu$ is a polynomial of degree at most $p$ on $\mathcal{V}^\epsilon(T)$.
- $E$ is bounded $\mathrm{W}^{\ell,q}(S) \to \mathrm{W}^{\ell,q}(\mathcal{V}^\epsilon(S))$. For cells $T \in \mathcal{T}_n$ touching $\partial S$,
$$\|\nabla^\ell Eu\|_{\mathrm{L}^q(\mathcal{V}_n^\epsilon(T))} \preceq \|\nabla^\ell u\|_{\mathrm{L}^q(\mathcal{M}_n(T))}.$$

*Proof.* For $(\sigma, m) \in \mathcal{D}$,

$$\|(\sigma, m)^\circ[\Phi_s, x]\| \preceq \delta(x)^{v(m)-w(m)},$$

from which it follows that

$$\|\Gamma_i^\ell[k]^\circ[\Phi_s, x]\| \preceq \delta(x)^{-i}.$$

This gives

$$\| \otimes^i \phi(x) \otimes \Gamma_i^\ell[k]^\circ[\Phi_s, x]\| \preceq 1. \tag{5.37}$$

We now check continuity properties on $\partial S$. Suppose $x_0 \in \partial S$ and $x \to x_0$. Then we have

$$\|\nabla^\ell u[\Phi_s(x)] - \nabla^\ell u[\Phi_2(x)]\| \to 0,$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

so that (5.34b) converges to 0. If the integrals (5.34c) had been evaluated at $\nabla^\ell u[\Phi_2(x)]$, their sum over $s \in I$ would be 0. An argument similar to the above, combined with (5.37), then shows that (5.34c,d) also converges to 0. We are left with the term on (5.34a), which converges to $\nabla^\ell u(x_0)$.

Boundedness properties of $E$ follow from (5.37) and the assumption that the $\Phi_s$ determine diffeomorphisms with uniformly bounded Jacobian determinants. □

*Quasi-interpolator.* Putting together the pieces, we get the following.

**Theorem 5.64.** For any $\epsilon > 0$, the operators $Q_n^\epsilon = I_n R_n^\epsilon E$ satisfy local estimates, for $T \in \mathcal{T}_n^d$,

$$\|Q_n^\epsilon u\|_{\mathrm{L}^q(T)} \preceq \|u\|_{\mathrm{L}^q(\mathcal{M}_n(T))},$$
$$\|u - Q_n^\epsilon u\|_{\mathrm{L}^q(T)} \preceq h_T^\ell \|\nabla^\ell u\|_{\mathrm{L}^q(\mathcal{M}_n(T))},$$

as the corresponding global ones:

$$\|Q_n^\epsilon u\|_{\mathrm{L}^q(S)} \preceq \|u\|_{\mathrm{L}^q(S)},$$
$$\|u - Q_n^\epsilon u\|_{\mathrm{L}^q(S)} \preceq h^\ell \|\nabla^\ell u\|_{\mathrm{L}^q(S)}.$$

Moreover, for any $\epsilon'$, choosing $\epsilon$ small enough will yield, for $u \in X_n^k$,

$$\|u - Q_n^\epsilon u\|_{\mathrm{L}^q(T)} \leq \epsilon' \|u\|_{\mathrm{L}^q(\mathcal{M}_n(T))},$$
$$\|u - Q_n^\epsilon u\|_{\mathrm{L}^q(S)} \preceq \epsilon' \|u\|_{\mathrm{L}^q(S)}.$$

Finally, $Q_n^\epsilon$ commutes with the exterior derivative (when it is in $\mathrm{L}^q(S)$).

When $\epsilon$ is chosen so small that $\|(\mathrm{id} - Q_n^\epsilon)|_{X_n^k}\|_{\mathrm{L}^q(S)\to\mathrm{L}^q(S)} \leq 1/2$, $Q_n^\epsilon|_{X_n^k}$ is invertible with norm less than 2. We define operators:

$$P_n = (Q_n^\epsilon|_{X_n^k})^{-1} Q_n^\epsilon.$$

**Proposition 5.65.** The operators $P_n$ are $\mathrm{L}^q(S)$-stable projections onto $X_n^k$ which commute with the exterior derivative.

The case $q = 2$ leads to eigenvalue convergence for the operator $\mathrm{d}^\star\mathrm{d}$ discretized by the Galerkin method on $X_n^k$ and therefore for the Hodge–Laplacian in mixed form. For a discussion of eigenvalue convergence we refer to Boffi (2010), Arnold *et al.* (2010) and Christiansen and Winther (2010).

### *5.4. Sobolev injection and translation estimate*

In this subsection we prove a Sobolev injection theorem generalizing the one we introduced in Christiansen and Scheid (2011). The proof technique is slightly different, and we generalize to differential forms in all dimensions. We also prove a translation estimate of the type introduced in Karlsen and

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

Karper (2010, Theorem A.1). Compared with that paper we get an optimal bound and a generalization to all known mixed finite elements in the $h$-version.

Given a cellular complex $\mathcal{T}$, we define the broken $\mathrm{H}^1$-seminorm $\|\!|\cdot|\!\|$ by

$$\|\!|u|\!\|^2 = \sum_{T\in\mathcal{T}^d} \|\nabla u\|^2_{\mathrm{L}^2(T)} + \sum_{T\in\mathcal{T}^{d-1}} h_T^{-1}\|[u]_T\|^2_{\mathrm{L}^2(T)}. \tag{5.38}$$

Given a simplex $T \in \mathcal{T}^{d-1}$, $[u]_T$ denotes the jump of $u$ across $T$. The scaling factor in front of the jump terms is chosen so that the two terms that are summed scale in the same way: see Lemma 5.50.

We assume we have a sequence of cellular complexes $(\mathcal{T}_n)$ and that each $\mathcal{T}_n$ is equipped with a compatible FE system $A[n]$. We also assume that they are of the type discussed in Section 5.3 so that those constructions apply. We define $X_n^k = A[n]^k(\mathcal{T}_n)$. The broken seminorm (5.38) will be defined relative to one of the $\mathcal{T}_n$ and, since the particular $\mathcal{T}_n$ should be clear from the context, we omit dependence on $\mathcal{T}_n$ from our notation.

In the $\mathrm{L}^2(S)$ case define

$$\begin{aligned} X^k &= \{u \in \mathrm{L}^2(S) : \mathrm{d}u \in \mathrm{L}^2(S)\},\\ W^k &= \{u \in X^k : \mathrm{d}u = 0\},\\ V^k &= \{u \in X^k : \forall w \in W^k \quad \textstyle\int u\cdot w = 0\}. \end{aligned}$$

Define

$$\begin{aligned} W_n^k &= \{u \in X_n^k : \mathrm{d}u = 0\},\\ V_n^k &= \{u \in X_n^k : \forall w \in W_n^k \quad \textstyle\int u\cdot w = 0\}. \end{aligned}$$

Denote by $\mathfrak{H}$ the $\mathrm{L}^2$-orthogonal projection onto $\overline{V}^k$, the completion of $V^k$ in $\mathrm{L}^2(S)$. This operator realizes a Hodge decomposition of $u$ in the form $u = (u - \mathfrak{H}u) + \mathfrak{H}u$. The following well-known trick is also useful in the proof of eigenvalue convergence.

**Proposition 5.66.** We have, for $u \in V_n^k$,

$$\|u - \mathfrak{H}u\|_{\mathrm{L}^2(S)} \le \|\mathfrak{H}u - P_n\mathfrak{H}u\|_{\mathrm{L}^2(S)}.$$

*Proof.* We have

$$u - P_n\mathfrak{H}u = P_n(u - \mathfrak{H}u) \in W_n^k.$$

Now write

$$\mathfrak{H}u - P_n\mathfrak{H}u = (u - P_n\mathfrak{H}u) - (u - \mathfrak{H}u),$$

and note that the two terms on the left-hand side are orthogonal. □

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

**Proposition 5.67.** For $u \in X_n^k$,

$$\|\!|u|\!\| \simeq \|\nabla R_n^\epsilon E u\|_{\mathrm{L}^2(S)}, \tag{5.39}$$

$$\|u - R_n^\epsilon E u\|_{\mathrm{L}^2(S)} \preceq h \|\!|u|\!\|. \tag{5.40}$$

*Proof.* The parameter $\epsilon$ is chosen so small that, for all $u \in X_n^k$,

$$\|\!|u - R_n^\epsilon E u|\!\| \leq 1/2 \|\!|u|\!\|.$$

Then we have

$$\|\!|u|\!\| \leq 2 \|\!|R_n^\epsilon E u|\!\|,$$
$$\|\!|R_n^\epsilon E u|\!\| \leq 3/2 \|\!|u|\!\|.$$

Since $R_n^\epsilon E u$ is smooth, we have

$$\|\!|R_n^\epsilon E u|\!\| = \|\nabla R_n^\epsilon E u\|_{\mathrm{L}^2(S)}.$$

This gives (5.39).

The other estimate is proved locally by scaling from a reference macro-element. □

**Proposition 5.68.** For $u \in \mathrm{H}^1(S)$,

$$\|\!|P_n u|\!\| \preceq \|\nabla u\|_{\mathrm{L}^2(S)}.$$

*Proof.* By restriction to reference simplexes and scaling,

$$\|\!|Q_n u|\!\| \preceq \|\nabla u\|_{\mathrm{L}^2(S)}.$$

Choosing $\epsilon$ small enough, we also have, for $u \in X_n^k$,

$$\|\!|u - Q_n u|\!\| \leq 1/2 \|\!|u|\!\|.$$

Combining the two, we get the proposition. □

**Proposition 5.69.** Suppose $S$ is convex and that the meshes are quasi-uniform. Then, for all $u \in V_n^k$,

$$\|\!|u|\!\| \preceq \|\mathrm{d}u\|_{\mathrm{L}^2(S)}. \tag{5.41}$$

*Proof.* We have, for $u \in V_n^k$,

$$\begin{aligned}
\|\!|u|\!\| &= \|\!|P_n u|\!\| \\
&\leq \|\!|P_n(u - \mathfrak{H} u)|\!\| + \|\!|P_n \mathfrak{H} u|\!\| \\
&\leq h^{-1} \|P_n(u - \mathfrak{H} u)\|_{\mathrm{L}^2} + \|\nabla \mathfrak{H} u\|_{\mathrm{L}^2} \\
&\preceq h^{-1} \|u - \mathfrak{H} u\|_{\mathrm{L}^2} + \|\nabla \mathfrak{H} u\|_{\mathrm{L}^2} \\
&\preceq \|\nabla \mathfrak{H} u\|_{\mathrm{L}^2}.
\end{aligned}$$

From this (5.41) follows. □

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

**Proposition 5.70.** Let $q$ be the relevant Sobolev exponent, so that $\mathrm{H}^1(S)$ is contained in $\mathrm{L}^q(S)$. For all $u \in X_n^k$,

$$\|u\|_{\mathrm{L}^q(S)} \preceq |\!|\!| u |\!|\!| + \|u\|_{\mathrm{L}^2(S)}.$$

*Proof.* We have

$$\begin{aligned}\|u\|_{\mathrm{L}^q(S)} &\preceq \|R_n^\epsilon E u\|_{\mathrm{L}^q(S)} \\ &\preceq \|\nabla R_n^\epsilon E u\|_{\mathrm{L}^2(S)} + \|R_n^\epsilon E u\|_{\mathrm{L}^2(S)} \\ &\preceq |\!|\!| u |\!|\!| + \|u\|_{\mathrm{L}^2(S)},\end{aligned}$$

as claimed. □

Denote by $\tau_y$ the translation by the vector $y$, so that when $u$ is defined in $x - y$ we have

$$(\tau_y u)(x) = u(x - y).$$

**Proposition 5.71.** For all $u \in X_n^k$,

$$\|u - \tau_y E u\|_{\mathrm{L}^2(S)} \preceq (|y| + h^{1/2}|y|^{1/2}) |\!|\!| u |\!|\!|.$$

*Proof.* On a reference simplex $\hat{T}$ we can write for $u \in X_n^k$ pulled back:

$$\|\hat{u} - \tau_{\hat{y}} \hat{u}\|^2_{\mathrm{L}^2(\hat{T})} \preceq |\hat{y}|^2 \sum_{T' \in \mathcal{M}_n(\hat{T})^d} \|\nabla \hat{u}\|^2_{\mathrm{L}^2(T')} + |\hat{y}| \sum_{T' \in \mathcal{M}_n(\hat{T})^{d-1}} \|[\hat{u}]_{T'}\|^2_{\mathrm{L}^2(T')}.$$

Scaling back to $T$ of size $h$, we get, with $y = h\hat{y}$,

$$\|u - \tau_y u\|^2_{\mathrm{L}^2(T)} \preceq |y|^2 \sum_{T' \in \mathcal{M}_n(T)^d} \|\nabla u\|^2_{\mathrm{L}^2(T')} + |y| \sum_{T' \in \mathcal{M}_n(T)^{d-1}} \|[u]_{T'}\|^2_{\mathrm{L}^2(T')},$$

so that

$$\|u - \tau_y u\|^2_{\mathrm{L}^2(T)} \preceq (|y|^2 + h|y|) |\!|\!| u |\!|\!|^2_{\mathcal{M}_n(T)}.$$

This estimate comes with a restriction of the type $|y| \leq h/C$, ensuring that one does not translate $T$ out of its associated macro-element. For $|y| \geq h/C$ we can write

$$u - \tau_y E u = (u - R_n^\epsilon E u) + (R_n^\epsilon E u - \tau_y E R_n^\epsilon E u) + \tau_y E (R_n^\epsilon E u - u).$$

From this we deduce

$$\begin{aligned}\|u - \tau_y E u\|_{\mathrm{L}^2} &\preceq \|u - R_n^\epsilon E u\|_{\mathrm{L}^2} + \|R_n^\epsilon E u - \tau_y E R_n^\epsilon E u\|_{\mathrm{L}^2} \\ &\preceq h |\!|\!| u |\!|\!| + |y| \|\nabla R_n^\epsilon E u\|_{\mathrm{L}^2} \\ &\preceq |y| \, |\!|\!| u |\!|\!|.\end{aligned}$$

This concludes the proof. □

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

## Acknowledgements

*First author.* On finite elements, I have benefited from the insights of, in particular, Annalisa Buffa, Jean-Claude Nédélec and Ragnar Winther. On algebraic topology I am grateful for the help of John Rognes and Bjørn Jahren, particularly with Propositions 5.16 and 5.21. I am also grateful to Martin Costabel for a helpful introduction to Stein's construction of universal extension operators. Example 5.30 is from a discussion with Thomas Dubos (shallow water on the sphere) in June 2009. Trygve Karper's remarks on upwinding were also helpful.

*Second author.* I would, in particular, like to thank Brett Ryland for his important contribution to the development of the theory of generalized Chebyshev polynomials and also for doing a major part of the computer implementations. Furthermore, I thank Morten Nome for his efforts in realizing the spectral element implementations. I am also grateful to Daan Huybrechs for important contributions to the understanding of generalized symmetries and Chebyshev polynomials of 'other kinds'. Finally, thanks to Krister Åhlander for his contributions to developing both the theory and the computer codes for numerical treatment of equivariance.

*Third author.* Morten Dahlby and Takaharu Yaguchi have contributed substantially to the section on integral-preserving methods. Elena Celledoni has proofread parts of this paper and made numerous suggestions for improvement.

This work, conducted as part of the award 'Numerical Analysis and Simulations of Geometric Wave Equations', made under the European Heads of Research Councils and European Science Foundation EURYI (European Young Investigator) Awards scheme, was supported by funds from the Participating Organizations of EURYI and the EC Sixth Framework Program. The work is also supported by the Norwegian Research Council through the project 'Structure Preserving Algorithms for Differential Equations: Applications, Computation and Education' (SpadeACE).

## Appendix

**Lemma A.1.** In a Banach space $\mathbb{E}$, let $U$ be an open set and let $f : U \to \mathbb{E}$ be a contraction mapping, *i.e.*, for some $\delta < 1$,

$$\|f(x) - f(y)\| \leq \delta \|x - y\|. \tag{A.1}$$

Then the map $g : U \to \mathbb{E}$, $x \mapsto x + f(x)$ has an open range $V$ and determines a Lipschitz bijection $U \to V$, with a Lipschitz inverse.

*Proof.* Suppose that $g(x_0) = y_0$. For $\|y - y_0\| \leq \epsilon$ find the solution $x$ of $g(x) = y$ as a fixed point of the map $z \mapsto y - f(z)$. More precisely, construct

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

a sequence starting at $x_0$ and defined by $x_{n+1} = y - f(x_n)$. Then, as long as it is defined (in $U$), we have

$$\begin{aligned}\|x_{n+1} - x_n\| &\leq \delta \|x_n - x_{n-1}\| \\ &\leq \delta^{n-1}\epsilon.\end{aligned}$$

If $\epsilon$ is chosen so small that the closed ball with centre $x_0$ and radius $(1-\delta)^{-1}\epsilon$ is included in $U$, the sequence is defined for all $n$, and converges to a limit $x \in U$ solving $g(x) = y$.

It follows that $g$ is an open mapping. In particular, the range $V$ is open. Moreover,

$$\|g(x) - g(y)\| \geq (1 - \delta)\|x - y\|.$$

This gives injectivity, so that $g : U \to V$ is bijective. The inverse is Lipschitz with constant no worse than $(1 - \delta)^{-1}$. □

**Lemma A.2.** In some Euclidean space $\mathbb{E}$, let $S$ be a bounded domain whose boundary $\partial S$ is locally the graph of a Lipschitz function. Let $n$ be the outward-pointing normal on $\partial S$. Suppose $m$ is a unit vector, that $x_0 \in \partial S$, and that for $x$ in a neighbourhood of $x_0$ in $\partial S$ we have $n(x) \cdot m \geq \epsilon$, for some $\epsilon > 0$. Then there is a neighbourhood of $x_0$ in $\partial S$ which is a Lipschitz graph above the plane orthogonal to $m$.

*Proof.* We know that for a certain outward-pointing unit vector $m_0$, a neighbourhood $\mathcal{U}_0$ of $x_0$ is a Lipschitz graph above an open ball $B_0$ in $m_0^{\perp}$. Choose $\theta \in [0, \pi/2[$ such that, for $x \in \mathcal{U}_0$, $n(x) \cdot m_0 \geq \cos\theta$, and moreover $\epsilon \geq \cos\theta$. Let $f : B_0 \to \mathbb{R}$ be the function such that $\mathcal{U}_1$ is the range of

$$\begin{cases} B_0 & \to \mathbb{E}, \\ y & \mapsto y + f(y)m_0. \end{cases}$$

Since, for $y \in B_0$,

$$n(y + f(y)m_0) = (m_0 - \operatorname{grad} f(y))/(1 + |\operatorname{grad} f(y)|^2)^{1/2},$$

we get $|\operatorname{grad} f(x)| \leq \tan\theta$.

Choose $y, y' \in B_0$ and put $x = y + f(y)m_0$ and $x' = y' + f(y')m_0$. For $s, s' \in \mathbb{R}$ we have

$$\begin{aligned}&|(x + sm_0) - (x' + s'm_0)|^2 \\ &\quad = |y - y'|^2 + (f(y) - f(y'))^2 + 2(f(y) - f(y'))(s - s') + (s - s')^2.\end{aligned}$$

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

Then note that, for $M > 0$,

$$\begin{aligned} 2(f(y) - f(y'))(s - s') &\le (1 + M^{-2})(f(y) - f(y'))^2 \\ &\quad + (1 + M^{-2})^{-1}(s - s')^2 \\ &\le (f(y) - f(y'))^2 + M^{-2}\tan^2\theta |y - y'|^2 \\ &\quad + (1 + M^{-2})^{-1}(s - s')^2. \end{aligned}$$

In particular, with $M = \tan\theta$ we get

$$|(x + sm_0) - (x' + s'm_0)|^2 \ge (1 - (1 + \tan^{-2}\theta)^{-1})(s - s')^2,$$

which simplifies to

$$|(x + sm_0) - (x' + s'm_0)| \ge \cos\theta |s - s'|.$$

Define the function

$$g_0 : \begin{cases} \mathcal{U}_0 \times \mathbb{R} & \to \mathbb{E}, \\ x & \mapsto y + sm_0. \end{cases} \tag{A.2}$$

Its range is $B_0 + \mathbb{R}m_0$, and it is bijective onto it.

Consider now a unit vector $m_1$ such that $n(x) \cdot m_1 \ge \cos\theta$ for $x \in \mathcal{U}_0$. Define $g_1$ as in (A.2), replacing $m_0$ by $m_1$. We shall show that $g_1$ is open and bi-Lipschitz, when $|m_1 - m_0| < \cos\theta$. Define $f_1$ on $B_0 + \mathbb{R}m_0$, by $f_1(x) = g_1 \circ g_0^{-1} - x$. With the preceding notation, we have

$$f_1(x + sm_0) - f_1(x' + s'm_0) = (s - s')(m_1 - m_0),$$

and hence

$$|f_1(x + sm_0) - f_1(x' + s'm_0)| \le |m_1 - m_0| / \cos\theta |(x + sm_0) - (x' + s'm_0)|.$$

We then apply Lemma A.1 and deduce that $g_1$ is open and bi-Lipschitz. It follows that there is a ball $B_1$ in $m_1^\perp$ above which a neighbourhood $\mathcal{U}_1 \subseteq \mathcal{U}_0$ of $x_0$ is a graph. We may repeat the above considerations to construct a sequence of such vectors $m_1, m_2, \ldots, m_k$ reaching $m$ in a finite number of steps. □

**Proposition A.3.** In some Euclidean space $\mathbb{E}$, let $S$ be a bounded domain whose boundary $\partial S$ is locally the graph of a Lipschitz function. Then there exists a smooth vector field $\nu$ on $\mathbb{E}$, of unit length and outward-pointing on $\partial S$, such that, for some $\epsilon > 0$, the map

$$\begin{cases} \partial S \times ]-\epsilon, \epsilon[ & \to \mathbb{E}, \\ (x, s) & \mapsto x + s\nu(x) \end{cases} \tag{A.3}$$

has open range and determines a Lipschitz isomorphism onto it.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

*Proof.* Cover $\partial S$ with a finite number of orthogonal cylinders $C_i$ directed along a unit vector $n_i$ pointing out of $S$, and with a base $U_i$, such that above $U_i$, $\partial S$ is the graph of a Lipschitz function. Let $n$ denote the outward-pointing normal on $\partial S$. For some $\theta \in [0, \pi/2[$, we have for all $i$ and all $x \in \partial S \cap C_i$

$$n_i \cdot n(x) \geq \cos(\theta). \tag{A.4}$$

Choose smooth functions $\alpha_i$ on $\mathbb{E}$ whose restrictions to $\partial S$ form a partition of unity. Define

$$\tilde{\nu}(x) = \sum_i \alpha_i(x) n_i,$$

and normalize by putting

$$\nu(x) = \tilde{\nu}(x)/|\nu(x)|.$$

We have $\nu(x) \cdot n(x) \geq \cos(\theta)$, for all $x \in \partial S$.

Denote by $g$ the function

$$g : \begin{cases} \partial S \times \mathbb{R} & \to \mathbb{E}, \\ (x, s) & \mapsto x + s\nu(x). \end{cases}$$

Pick $x_0 \in \partial S$ and put $m_0 = \nu(x_0)$. By Lemma A.2, $\partial S$ is locally a Lipschitz graph above the plane orthogonal to $m_0$. Let $\mathcal{U}(x_0)$ be the corresponding neighbourhood of $x_0$ in $\partial S$. Denote by $g_0$ the function

$$g_0 : \begin{cases} \partial S \cap \mathcal{U}(x_0) \times \mathbb{R} & \to \mathbb{E}, \\ (x, s) & \mapsto x + s m_0. \end{cases}$$

It is a Lipschitz isomorphism onto its range, which is open in $\mathbb{E}$.

Define $f$ by $f(x) = g \circ g_0^{-1}(x) - x$. We have

$$f(x + s m_0) = s(\nu(x) - m_0),$$

so that

$$f(x + s m_0) - f(y + t m_0) = (s - t)(\nu(x) - m_0) + t(\nu(x) - \nu(y)).$$

It follows that, for a small enough $\epsilon$ and possibly reducing $\mathcal{U}(x_0)$, $f$ is short on $g_0(\mathcal{U}(x_0) \times ]-\epsilon, \epsilon[)$. By Lemma A.1 it follows that $g \circ g_0^{-1}$ restricted to $g_0(\mathcal{U}(x_0) \times ]-\epsilon, \epsilon[)$, for some small enough $\epsilon$, called $\epsilon(x_0)$, has open range and determines a Lipschitz isomorphism onto it.

Hence $g$ restricted to $\mathcal{U}(x_0) \times ]-\epsilon(x_0), \epsilon(x_0)[$ has open range and determines a Lipschitz isomorphism onto it. In particular, $g$ is an open mapping.

The open subsets $\mathcal{U}(x_0)$ associated with each $x_0 \in \partial S$ cover $\partial S$. Choose a finite subset $\mathcal{F}$ of $\partial S$ such that the sets $\mathcal{U}(x)$ for $x \in \mathcal{F}$ cover $\partial S$. Pick $\mu > 0$ such that if $|x - x'| \leq \mu$ they belong to a common $\mathcal{U}(x)$ for $x \in \mathcal{F}$. In (A.3), choose $\epsilon$ smaller than each $\epsilon(x)$ for $x \in \mathcal{F}$, and also smaller than

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

$\mu/3$. Now pick two points $x, x' \in \partial S$, and $s, s'$ in $]-\epsilon, \epsilon[$. If $|x - x'| \leq \mu$, they belong to a common $\mathcal{U}(x)$, $x \in \mathcal{F}$. If, on the other hand, $|x - x'| \geq \mu$, we have

$$\begin{aligned} |g(x,s) - g(x',s')| &\geq |x - x'| - |s| - |s'|, \\ &\geq \mu/3. \end{aligned}$$

Based on these two cases we may conclude that, for some global $m$,

$$|g(x,s) - g(x',s')| \geq m(|x - x'|^2 + |s - s'|^2).$$

The lemma follows. □

## REFERENCES[11]

K. Åhlander and H. Munthe-Kaas (2005), 'Applications of the generalized Fourier transform in numerical linear algebra', *BIT* **45**, 819–850.

E. L. Allgower, K. Böhmer, K. Georg and R. Miranda (1992), 'Exploiting symmetry in boundary element methods', *SIAM J. Numer. Anal.* **29**, 534–552.

E. L. Allgower, K. Georg and R. Miranda (1993), Exploiting permutation symmetry with fixed points in linear equations. In *Lectures in Applied Mathematics* (E. L. Allgower, K. Georg and R. Miranda, eds), Vol. 29, AMS, pp. 23–36.

E. L. Allgower, K. Georg, R. Miranda and J. Tausch (1998), 'Numerical exploitation of equivariance', *Z. Angew. Math. Mech.* **78**, 185–201.

B. Andreianov, M. Bendahmane and K. H. Karlsen (2010), 'Discrete duality finite volume schemes for doubly nonlinear degenerate hyperbolic–parabolic equations', *J. Hyperbolic Diff. Equations* **7**, 1–67.

D. N. Arnold, P. B. Bochev, R. B. Lehoucq, R. A. Nicolaides and M. Shashkov, eds (2006*a*), *Compatible Spatial Discretizations*, Vol. 142 of *The IMA Volumes in Mathematics and its Applications*, Springer.

D. N. Arnold, R. S. Falk and R. Winther (2006*b*), Finite element exterior calculus, homological techniques, and applications. In *Acta Numerica*, Vol. 15, Cambridge University Press, pp. 1–155.

D. N. Arnold, R. S. Falk and R. Winther (2010), 'Finite element exterior calculus: From Hodge theory to numerical stability', *Bull. Amer. Math. Soc.* (NS) **47**, 281–354.

H. F. Baker (1905), 'Alternants and continuous groups', *Proc. London Math. Soc.* **3**, 24–47.

R. J. Beerends (1991), 'Chebyshev polynomials in several variables and the radial part of the Laplace–Beltrami operator', *Trans. Amer. Math. Soc.* **328**, 779–814.

T. B. Benjamin (1972), 'The stability of solitary waves', *Proc. Roy. Soc. London Ser. A* **328**, 153–183.

[11] The URLs cited in this work were correct at the time of going to press, but the publisher and the authors make no undertaking that the citations remain live or are accurate or appropriate.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

T. B. Benjamin, J. L. Bona and J. J. Mahony (1972), 'Model equations for long waves in nonlinear dispersive systems', *Philos. Trans. Roy. Soc. London Ser. A* **272**, 47–78.
S. Blanes and P. Moan (2006), 'Fourth- and sixth-order commutator-free Magnus integrators for linear and non-linear dynamical systems', *Appl. Numer. Math.* **56**, 1519–1537.
S. Blanes, F. Casas, J. A. Oteo and J. Ros (2009), 'The Magnus expansion and some of its applications', *Phys. Rep.* **470**, 151–238.
P. B. Bochev and J. M. Hyman (2006), Principles of mimetic discretizations of differential operators. In *Compatible Spatial Discretizations*, Vol. 142 of *The IMA Volumes in Mathematics and its Applications*, Springer, pp. 89–119.
D. Boffi (2010), Finite element approximation of eigenvalue problems. In *Acta Numerica*, Vol. 19, Cambridge University Press, pp. 1–120.
A. Bossavit (1986), 'Symmetry, groups, and boundary value problems: A progressive introduction to noncommutative harmonic analysis of partial differential equations in domains with geometrical symmetry', *Comput. Methods Appl. Mech. Engrg* **56**, 167–215.
A. Bossavit (1988), Mixed finite elements and the complex of Whitney forms. In *The Mathematics of Finite Elements and Applications VI* (Uxbridge 1987), Academic Press, pp. 137–144.
F. Brezzi (1974), 'On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers', *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge* **8**, 129–151.
F. Brezzi and M. Fortin (1991), *Mixed and Hybrid Finite Element Methods*, Vol. 15 of *Springer Series in Computational Mathematics*, Springer.
F. Brezzi, J. Douglas, Jr. and L. D. Marini (1985), 'Two families of mixed finite elements for second order elliptic problems', *Numer. Math.* **47**, 217–235.
F. Brezzi, K. Lipnikov and M. Shashkov (2005), 'Convergence of the mimetic finite difference method for diffusion problems on polyhedral meshes', *SIAM J. Numer. Anal.* **43**, 1872–1896.
R. L. Bryant (1995), An introduction to Lie groups and symplectic geometry. In *Geometry and Quantum Field Theory* (D. S. Freed and K. K. Uhlenbeck, eds), Vol. 1 of *IAS/Park City Mathematics Series*, AMS.
A. Buffa and S. H. Christiansen (2007), 'A dual finite element complex on the barycentric refinement', *Math. Comp.* **76**, 1743–1769.
D. Bump (2004), *Lie Groups*, Springer.
C. Canuto, M. Hussaini, A. Quarteroni and T. Zang (2006), *Spectral Methods: Fundamentals in Single Domains*, Scientific Computation series, Springer.
E. Celledoni (2005), Eulerian and semi-Lagrangian schemes based on commutator-free exponential integrators. In *Group Theory and Numerical Analysis*, Vol. 39 of *CRM Proc. Lecture Notes*, AMS, pp. 77–90.
E. Celledoni and A. Iserles (2000), 'Approximating the exponential from a Lie algebra to a Lie group', *Math. Comp.* **69**, 1457–1480.
E. Celledoni and A. Iserles (2001), 'Methods for the approximation of the matrix exponential in a Lie-algebraic setting', *IMA J. Numer. Anal.* **21**, 463–488.
E. Celledoni and B. K. Kometa (2009), 'Semi-Lagrangian Runge–Kutta exponential integrators for convection dominated problems', *J. Sci. Comput.* **41**, 139–164.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

E. Celledoni, D. Cohen and B. Owren (2008), 'Symmetric exponential integrators with an application to the cubic Schrödinger equation', *Found. Comput. Math.* **8**, 303–317.

E. Celledoni, A. Marthinsen and B. Owren (2003), 'Commutator-free Lie group methods', *Future Generation Computer Systems* **19**, 341–352.

J. Certaine (1960), The solution of ordinary differential equations with large time constants. In *Mathematical Methods for Digital Computers*, Wiley, pp. 128–132.

S. H. Christiansen (2007), 'Stability of Hodge decompositions in finite element spaces of differential forms in arbitrary dimension', *Numer. Math.* **107**, 87–106.

S. H. Christiansen (2008*a*), 'A construction of spaces of compatible differential forms on cellular complexes', *Math. Models Methods Appl. Sci.* **18**, 739–757.

S. H. Christiansen (2008*b*), On the linearization of Regge calculus. E-print, Department of Mathematics, University of Oslo.

S. H. Christiansen (2009), Foundations of finite element methods for wave equations of Maxwell type. In *Applied Wave Mathematics*, Springer, pp. 335–393.

S. H. Christiansen (2010), 'Éléments finis mixtes minimaux sur les polyèdres', *CR Math. Acad. Sci. Paris* **348**, 217–221.

S. H. Christiansen and C. Scheid (2011), 'Convergence of a constrained finite element discretization of the Maxwell Klein Gordon equation', *ESAIM: Math. Model. Numer. Anal.* **45**, 739–760.

S. H. Christiansen and R. Winther (2006), 'On constraint preservation in numerical simulations of Yang–Mills equations', *SIAM J. Sci. Comput.* **28**, 75–101.

S. H. Christiansen and R. Winther (2008), 'Smoothed projections in finite element exterior calculus', *Math. Comp.* **77**, 813–829.

S. H. Christiansen and R. Winther (2010), On variational eigenvalue approximation of semidefinite operators. Preprint: arXiv.org/abs/1005.2059.

P. G. Ciarlet (1978), *The Finite Element Method for Elliptic Problems*, Vol. 4 of *Studies in Mathematics and its Applications*, North-Holland.

P. Clément (1975), 'Approximation by finite element functions using local regularization', *RAIRO Analyse Numérique* **9**, 77–84.

R. Courant, K. Friedrichs and H. Lewy (1928), 'Über die partiellen Differenzengleichungen der mathematischen Physik', *Math. Ann.* **100**, 32–74.

S. M. Cox and P. C. Matthews (2002), 'Exponential time differencing for stiff systems', *J. Comput. Phys.* **176**, 430–455.

P. E. Crouch and R. Grossman (1993), 'Numerical integration of ordinary differential equations on manifolds', *J. Nonlinear Sci.* **3**, 1–33.

M. Dahlby and B. Owren (2010), A general framework for deriving integral preserving numerical methods for PDEs. Technical report 8/2010, Norwegian University of Science and Technology. arXiv.org/abs/1009.3151.

M. Dahlby, B. Owren and T. Yaguchi (2010), Preserving multiple first integrals by discrete gradients. Technical report 11/2010, Norwegian University of Science and Technology. arXiv.org/abs/1011.0478.

L. Demkowicz and I. Babuška (2003), '$p$ interpolation error estimates for edge finite elements of variable order in two dimensions', *SIAM J. Numer. Anal.* **41**, 1195–1208.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

L. Demkowicz and A. Buffa (2005), '$H^1$, $H$(curl) and $H$(div)-conforming projection-based interpolation in three dimensions: Quasi-optimal $p$-interpolation estimates', *Comput. Methods Appl. Mech. Engrg* **194**, 267–296.

L. Demkowicz, J. Kurtz, D. Pardo, M. Paszyński, W. Rachowicz and A. Zdunek (2008), *Computing with hp-adaptive Finite Elements*, Vol. 2, *Frontiers: Three Dimensional Elliptic and Maxwell Problems with Applications*, Applied Mathematics and Nonlinear Science Series, Chapman & Hall/CRC.

F. Diele, L. Lopez and R. Peluso (1998), 'The Cayley transform in the numerical solution of unitary differential systems', *Adv. Comput. Math.* **8**, 317–334.

J. Dodziuk and V. K. Patodi (1976), 'Riemannian structures and triangulations of manifolds', *J. Indian Math. Soc.* (NS) **40**, 1–52.

C. C. Douglas and J. Mandel (1992), 'Abstract theory for the domain reduction method', *Computing* **48**, 73–96.

J. Droniou, R. Eymard, T. Gallouët and R. Herbin (2010), 'A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods', *Math. Models Methods Appl. Sci.* **20**, 265–295.

M. Dubiner (1991), 'Spectral methods on triangles and other domains', *J. Sci. Comput.* **6**, 345–390.

R. Eier and R. Lidl (1982), 'A class of orthogonal polynomials in $k$ variables', *Math. Ann.* **260**, 93–99.

A. F. Fässler and E. Stiefel (1992), *Group Theoretical Methods and their Applications*, Birkhäuser.

D. Furihata (1999), 'Finite difference schemes for $\partial u/\partial t = (\partial/\partial x)^{\alpha}\delta G/\delta u$ that inherit energy conservation or dissipation property', *J. Comput. Phys.* **156**, 181–205.

D. Furihata (2001*a*), 'Finite-difference schemes for nonlinear wave equation that inherit energy conservation property', *J. Comput. Appl. Math.* **134**, 37–57.

D. Furihata (2001*b*), 'A stable and conservative finite difference scheme for the Cahn–Hilliard equation', *Numer. Math.* **87**, 675–699.

D. Furihata and T. Matsuo (2003), 'A stable, convergent, conservative and linear finite difference scheme for the Cahn–Hilliard equation', *Japan J. Indust. Appl. Math.* **20**, 65–85.

K. Georg and R. Miranda (1992), Exploiting symmetry in solving linear equations. In *Bifurcation and Symmetry* (E. L. Allgower, K. Böhmer and M. Golubisky, eds), Vol. 104 of *International Series of Numerical Mathematics*, Birkhäuser, pp. 157–168.

F. X. Giraldo and T. Warburton (2005), 'A nodal triangle-based spectral element method for the shallow water equations on the sphere', *J. Comput. Phys.* **207**, 129–150.

O. Gonzalez (1996), 'Time integration and discrete Hamiltonian systems', *J. Nonlinear Sci.* **6**, 449–467.

P. A. Griffiths and J. W. Morgan (1981), *Rational Homotopy Theory and Differential Forms*, Vol. 16 of *Progress in Mathematics*, Birkhäuser.

E. Hairer, C. Lubich and G. Wanner (2006), *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, second edition, Vol. 31 of *Springer Series in Computational Mathematics*, Springer.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

F. Hausdorff (1906), 'Die symbolische Exponential Formel in der Gruppentheorie', *Leipziger Ber.* **58**, 19–48.
J. S. Hesthaven (1998), 'From electrostatics to almost optimal nodal sets for polynomial interpolation in a simplex', *SIAM J. Numer. Anal.* **35**, 655–676.
S. Hilbert (1973), 'A mollifier useful for approximations in Sobolev spaces and some applications to approximating solutions of differential equations', *Math. Comp.* **27**, 81–89.
R. Hiptmair (1999), 'Canonical construction of finite elements', *Math. Comp.* **68**, 1325–1346.
R. Hiptmair (2002), Finite elements in computational electromagnetism. In *Acta Numerica*, Vol. 11, Cambridge University Press, pp. 237–339.
M. Hochbruck and A. Ostermann (2005), 'Explicit exponential Runge–Kutta methods for semilinear parabolic problems', *SIAM J. Numer. Anal.* **43**, 1069–1090.
M. Hochbruck and A. Ostermann (2010), Exponential integrators. In *Acta Numerica*, Vol. 19, Cambridge University Press, pp. 209–286.
M. E. Hoffman and W. D. Withers (1988), 'Generalized Chebyshev polynomials associated with affine Weyl groups', *Trans. Amer. Math. Soc.* **308**, 91–104.
D. Huybrechs (2010), 'On the Fourier extension of non-periodic functions', *SIAM J. Numer. Anal.* **47**, 4326–4355.
D. Huybrechs, A. Iserles and S. Nørsett (2010), 'From high oscillation to rapid approximation V: The equilateral triangle', *IMA J. Numer. Anal.* doi:10.1093/imanum/drq010.
A. Iserles and S. P. Nørsett (1999), 'On the solution of linear differential equations in Lie groups', *Philos. Trans. Roy. Soc. London Ser. A* **357**, 983–1019.
A. Iserles and A. Zanna (2005), 'Efficient computation of the matrix exponential by generalized polar decompositions', *SIAM J. Numer. Anal.* **42**, 2218–2256.
A. Iserles, H. Munthe-Kaas, S. P. Nørsett and A. Zanna (2000), Lie-group methods. In *Acta Numerica*, Vol. 9, Cambridge University Press, pp. 215–365.
G. James and M. Liebeck (2001), *Representations and Characters of Groups*, second edition, Cambridge University Press.
F. Kang and Z.-J. Shang (1995), 'Volume-preserving algorithms for source-free dynamical systems', *Numer. Math.* **71**, 451–463.
K. H. Karlsen and T. K. Karper (2010), 'Convergence of a mixed method for a semi-stationary compressible Stokes system', *Math. Comp.* doi:10.1090/S0025-5718-2010-02446-9.
C. A. Kennedy and M. H. Carpenter (2003), 'Additive Runge–Kutta schemes for convection–diffusion–reaction equations', *Appl. Numer. Math.* **44**, 139–181.
T. Koornwinder (1974), 'Orthogonal polynomials in two variables which are eigenfunctions of two algebraically independent partial differential operators I–IV', *Indag. Math.* **36**, 48–66 and 357–381.
S. Krogstad (2005), 'Generalized integrating factor methods for stiff PDEs', *J. Comput. Phys.* **203**, 72–88.
S. Krogstad, H. Munthe-Kaas and A. Zanna (2009), 'Generalized polar coordinates on Lie groups and numerical integrators', *Numer. Math.* **114**, 161–187.
Y. Kuznetsov and S. Repin (2005), 'Convergence analysis and error estimates for mixed finite element method on distorted meshes', *J. Numer. Math.* **13**, 33–51.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

R. A. LaBudde and D. Greenspan (1974), 'Discrete mechanics: A general treatment', *J. Comput. Phys.* **15**, 134–167.
J. D. Lawson (1967), 'Generalized Runge–Kutta processes for stable systems with large Lipschitz constants', *SIAM J. Numer. Anal.* **4**, 372–380.
B. Leimkuhler and S. Reich (2004), *Simulating Hamiltonian Dynamics*, Vol. 14 of *Cambridge Monographs on Applied and Computational Mathematics*, Cambridge University Press.
D. Lewis and J. C. Simo (1994), 'Conserving algorithms for the dynamics of Hamiltonian systems on Lie groups', *J. Nonlinear Sci.* **4**, 253–299.
R. Lidl (1975), 'Tchebyscheffpolynome in mehreren Variablen', *J. Reine Angew. Math.* **273**, 178–198.
J. S. Lomont (1959), *Applications of Finite Groups*, Academic Press.
L. Lopez and T. Politi (2001), 'Applications of the Cayley approach in the numerical solution of matrix differential systems on quadratic groups', *Appl. Numer. Math.* **36**, 35–55.
A. Marthinsen and B. Owren (2001), 'Quadrature methods based on the Cayley transform', *Appl. Numer. Math.* **39**, 403–413.
T. Matsuo (2007), 'New conservative schemes with discrete variational derivatives for nonlinear wave equations', *J. Comput. Appl. Math.* **203**, 32–56.
T. Matsuo (2008), 'Dissipative/conservative Galerkin method using discrete partial derivatives for nonlinear evolution equations', *J. Comput. Appl. Math.* **218**, 506–521.
T. Matsuo and D. Furihata (2001), 'Dissipative or conservative finite-difference schemes for complex-valued nonlinear partial differential equations', *J. Comput. Phys.* **171**, 425–447.
T. Matsuo, M. Sugihara, D. Furihata and M. Mori (2000), 'Linearly implicit finite difference schemes derived by the discrete variational method', *Sūrikaisekikenkyūsho Kōkyūroku* **1145**, 121–129.
T. Matsuo, M. Sugihara, D. Furihata and M. Mori (2002), 'Spatially accurate dissipative or conservative finite difference schemes derived by the discrete variational method', *Japan J. Indust. Appl. Math.* **19**, 311–330.
R. I. McLachlan (1995), 'On the numerical integration of ordinary differential equations by symmetric composition methods', *SIAM J. Sci. Comput.* **16**, 151–168.
R. I. McLachlan, G. R. W. Quispel and N. Robidoux (1999), 'Geometric integration using discrete gradients', *Philos. Trans. Roy. Soc. London Ser. A* **357**, 1021–1045.
R. I. McLachlan, G. R. W. Quispel and P. S. P. Tse (2009), 'Linearization-preserving self-adjoint and symplectic integrators', *BIT* **49**, 177–197.
B. V. Minchev (2004), *Exponential Integrators for Semilinear Problems*, University of Bergen. PhD thesis, University of Bergen, Norway.
Y. Minesaki and Y. Nakamura (2006), 'New numerical integrator for the Stäckel system conserving the same number of constants of motion as the degree of freedom', *J. Phys. A* **39**, 9453–9476.
C. Moler and C. Van Loan (1978), 'Nineteen dubious ways to compute the exponential of a matrix', *SIAM Review* **20**, 801–836.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

C. B. Moler and C. F. van Loan (2003), 'Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later', *SIAM Review* **45**, 3–49.
P. Monk (2003), *Finite Element Methods for Maxwell's Equations*, Numerical Mathematics and Scientific Computation, Oxford University Press.
K. W. Morton (2010), 'The convection–diffusion Petrov–Galerkin story', *IMA J. Numer. Anal.* **30**, 231–240.
H. Munthe-Kaas (1989), Symmetric FFTs: A general approach. In Topics in Linear Algebra for Vector and Parallel Computers, PhD thesis, NTNU, Trondheim, Norway. Available at: hans.munthe-kaas.no.
H. Munthe-Kaas (1999), 'High order Runge–Kutta methods on manifolds', *Appl. Numer. Math.* **29**, 115–127.
H. Munthe-Kaas (2006), 'On group Fourier analysis and symmetry preserving discretizations of PDEs', *J. Phys. A* **39**, 5563.
H. Z. Munthe-Kaas (1995), 'Lie–Butcher theory for Runge–Kutta methods', *BIT* **35**, 572–587.
H. Z. Munthe-Kaas (1998), 'Runge–Kutta methods on Lie groups', *BIT* **38**, 92–111.
H. Z. Munthe-Kaas and B. Owren (1999), 'Computations in a free Lie algebra', *Philos. Trans. Roy. Soc. London Ser. A* **357**, 957–981.
H. Z. Munthe-Kaas and A. Zanna (1997), Numerical integration of differential equations on homogeneous manifolds. In *Foundations of Computational Mathematics* (F. Cucker and M. Shub, eds), Springer, pp. 305–315.
J.-C. Nédélec (1980), 'Mixed finite elements in $\mathbf{R}^3$', *Numer. Math.* **35**, 315–341.
S. P. Nørsett (1969), An $A$-stable modification of the Adams–Bashforth methods. In *Conf. Numerical Solution of Differential Equations* (Dundee 1969), Springer, pp. 214–219.
P. J. Olver (1993), *Applications of Lie Groups to Differential Equations*, second edition, Vol. 107 of *Graduate Texts in Mathematics*, Springer.
A. Ostermann, M. Thalhammer and W. M. Wright (2006), 'A class of explicit exponential general linear methods', *BIT* **46**, 409–431.
B. Owren (2006), 'Order conditions for commutator-free Lie group methods', *J. Phys. A* **39**, 5585–5599.
B. Owren and A. Marthinsen (1999), 'Runge–Kutta methods adapted to manifolds and based on rigid frames', *BIT* **39**, 116–142.
B. Owren and A. Marthinsen (2001), 'Integration methods based on canonical coordinates of the second kind', *Numer. Math.* **87**, 763–790.
J. E. Pasciak and P. S. Vassilevski (2008), 'Exact de Rham sequences of spaces defined on macro-elements in two and three spatial dimensions', *SIAM J. Sci. Comput.* **30**, 2427–2446.
P.-A. Raviart and J. M. Thomas (1977), A mixed finite element method for second order elliptic problems. In *Mathematical Aspects of Finite Element Methods*, Vol. 606 of *Lecture Notes in Mathematics*, Springer, pp. 292–315.
J. E. Roberts and J.-M. Thomas (1991), Mixed and hybrid methods. In *Handbook of Numerical Analysis*, Vol. II, North-Holland, pp. 523–639.
B. N. Ryland and H. Munthe-Kaas (2011), On multivariate Chebyshev polynomials and spectral approximations on triangles. In *Spectral and High Order Methods for Partial Differential Equations*, Vol. 76 of *Lecture Notes in Computational Science and Engineering*, Springer, pp. 19–41.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press

J. M. Sanz-Serna and M. P. Calvo (1994), *Numerical Hamiltonian Problems*, Vol. 7 of *Applied Mathematics and Mathematical Computation*, Chapman & Hall.
J. Schöberl (2008), '*A posteriori* error estimates for Maxwell equations', *Math. Comp.* **77**, 633–649.
J. Schöberl and A. Sinwel (2007), Tangential-displacement and normal-normal-stress continuous mixed finite elements for elasticity. RICAM report.
J. P. Serre (1977), *Linear Representations of Finite Groups*, Springer.
E. M. Stein (1970), *Singular Integrals and Differentiability Properties of Functions*, Vol. 30 of *Princeton Mathematical Series*, Princeton University Press.
G. Strang (1972), 'Approximation in the finite element method', *Numer. Math.* **19**, 81–98.
A. Trønnes (2005), Symmetries and generalized Fourier transforms applied to computing the matrix exponential. Master's thesis, University of Bergen, Norway.
V. S. Varadarajan (1984), *Lie Groups, Lie Algebras, and Their Representations*, Vol. 102 of *Graduate Texts in Mathematics*, Springer.
T. Warburton (2006), 'An explicit construction of interpolation nodes on the simplex', *J. Engng Math.* **56**, 247–262.
F. W. Warner (1983), *Foundations of Differentiable Manifolds and Lie Groups*, Vol. 94 of *Graduate Texts in Mathematics*, Springer.
A. Weil (1952), 'Sur les théorèmes de de Rham', *Comment. Math. Helv.* **26**, 119–145.
J. Wensch, O. Knoth and A. Galant (2009), 'Multirate infinitesimal step methods for atmospheric flow simulation', *BIT* **49**, 449–473.
H. Whitney (1957), *Geometric Integration Theory*, Princeton University Press.
T. Yaguchi, T. Matsuo and M. Sugihara (2010), 'Conservative numerical schemes for the Ostrovsky equation', *J. Comput. Appl. Math.* **234**, 1036–1048.
A. Zanna and H. Z. Munthe-Kaas (2001/02), 'Generalized polar decompositions for the approximation of the matrix exponential', *SIAM J. Matrix Anal. Appl.* **23**, 840–862.
A. Zanna, K. Engø and H. Z. Munthe-Kaas (2001), 'Adjoint and selfadjoint Lie-group methods', *BIT* **41**, 395–421.

https://doi.org/10.1017/S096249291100002X Published online by Cambridge University Press