Geometric numerical integration via auxiliary variables

Boris D. Andrews

Oriel College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2025

Abstract

Geometric numerical integrators are known to exhibit greater accuracy and physical reliability than classical timestepping schemes, in particular over long durations. However, there have long remained difficulties in devising such discretisations that preserve non-quadratic conservation and dissipation laws.

In this thesis we propose a unified framework for the construction of timestepping schemes or arbitrary order for systems of both ordinary (ODE) and partial (PDE) differential equations, that preserve multiple such structures for arbitrary quantities of interest. This jointly employs finite elements (FEs) in time and systematically introduces auxiliary variables (AVs) to transfer the proofs of these structures from the continuous level to the discrete.

We demonstrate the ideas by devising a novel integrator that conserves all known invariants of general conservative ODEs, an energy-conserving and entropy-generating scheme for PDEs within the GENERIC formalism (including the Boltzmann equation), and a FE scheme for the compressible Navier–Stokes (NS) equations that conserves mass, momentum, and energy, and provably possesses non-decreasing entropy. Moreover, we show the approach generalises and unifies several existing ideas in the literature, including Gauss methods, the energy- and helicity-stable integrator of Rebholz for the incompressible NS equations, the energy-stable ODE integrator of Hairer, Cohen & Lubich, the energy- and entropy-stable integrator of Romero for GENERIC ODEs, and the energy-stable PDE integrator of Giesselmann, Karsai & Tscherpel.

When applied to PDEs, we employ the FE method for our schemes' spatial discretisations. While our framework typically imposes no restriction on the FE spaces, we demonstrate how ideas from FE exterior calculus may be used to simplify our schemes, assuming certain compatibility conditions exist on the spaces used. This allows us to derive a novel, mixed velocity–vorticity FE scheme for the incompressible NS equations that is both energy- and (in the 2D case) enstrophy-stable, alongside an energy- and helicity-stable scheme for the equations of incompressible Hall magnetohydrodynamics (MHD) that generalises a scheme of Laakmann, Hu & Farrell.

We derive some general analytic results for our schemes. For certain stable discretisations of a class of advection-diffusion-type PDEs (including the incompressible NS and MHD equations) we are able to show both the existence of discrete solutions, and their uniqueness. For stable discretisations of ODEs, we are able to show both the unique existence of discrete solutions, and their convergence.

Lastly, we extend our framework to the preservation of adiabatic invariants, quantities that exhibit rapid oscillations about a slowly changing value. As an example, we consider the motion of a charged particle in a strong magnetic field. Through a modification of the AVs introduced by our framework, we construct an energy-stable integrator for the motion of such particles that further preserves the adiabatic invariance of the magnetic moment.

In many of these cases, we demonstrate the benefits provided by our structure-preserving approach through numerical simulations. These include the general conservative ODE integrator (using as examples the Kepler problem and the Kovalevskaya top), the conservative Poisson ODE integrator (using as an example the Benjamin–Bona–Mahony equation), the fully stable compressible NS scheme, the energy- and enstrophy-stable integrator for the incompressible NS equations, and the charged particle problem.

Acknowledgements

First and foremost, I would like to express my gratitude for the guidance of my supervisors, Patrick Farrell and Wayne Arter. Four years ago (when I was still young and naïve) I started my DPhil with a brain full of ideas; creative ones, but more often than not misguided to say the least. I had never learnt to communicate my thoughts in an academic setting. I had never learnt to research, to present, to organise my hours, my weeks, my years. You have helped me focus my work and my ideas into something about which I am genuinely excited. Your advice continues to this day to help shape me toward becoming the kind of academic I aspire to be.

I am very grateful to the UK Atomic Energy Authority, in particular to Rob Akers, and to the Engineering and Physical Sciences Research Council for their support throughout my DPhil.

I would now like to extend my gratitude to the rest of the senior and junior faculty at Oxford, both current and past, including Andy Wathen, Charlie Parker, Endre Süli, Georg Maierhofer, Kaibo Hu, José Carrillo, Michael Barnes, Nick Trefethen and Yuji Nakatsukasa; thank you for the fruitful talks and feedback (especially back in the days when they were accompanied by 11 a.m. doughnuts). Beyond Oxford, many inspiring conversations with Aaron Brunk, Ari Stern, Cecilia Pagliantini, Damiano Lombardi, Erwan Faou, Gunnar Hornig, Lee Ricketson, Lorenzo Pareschi, Maria Lukáčová-Medviďová and Tabea Tscherpel have made a deeply meaningful contribution to this thesis. I would especially like to highlight the support of Brendan Keith and Jesse Chan; your continued excitement and enthusiasm about my work have truly gone a long way. A very important shout-out goes to the Firedrake community and Slack, in particular Colin Cotter, Connor Ward, David Ham and Jack Betteridge, without whom this thesis would be but text and equations.

These past four years would have been nothing without the community I have found among my fellow DPhil students. Thank you, Aaron, Alan, Francis, Gonzalo, India, Irina, Jung, Karl, Kars, Lorenzo, Maike, Mingdong, Nathaniel and Nicolas, for making all the long hours so short. In particular (and conveniently last alphabetically), I want to thank Taejun and Umberto; the strength of character it must have taken you to spend (up to) five days a week in my direct company will go down in the annals of Meier Wing history. Academic travel has been one of the great joys of doing a PhD, no less for the friends I have made outside the Oxford community: Alberto, Alice, Astrid, Diana and John. A special thank-you goes to everyone in the

unofficial Mathematical Institute Society for Football Including Tables (MISFITs) and their *relentless* efforts to keep me working hard until 6 p.m. every day.

I'd like now to, at least slowly, begin to conclude, by acknowledging (in a very rough reverse chronological order) all my extracurricular friends: those without whom this thesis may have never come to pass, and especially those without whom I might well have finished in half the time. To quote one of the great philosophers of our age:

"How lucky am I to have something that makes saying goodbye so hard."

Adam, Alex G, Casey, Johny, Kumeren, Maike (again) and Noah, you have offered me warmth, camaraderie and (on occasion) *grade-A* banter. Cyril, Ed and Théana, you have given me a community, a home and a whole lot of cake. For each of you, and all you have brought into my life, I am truly grateful. Thank you, Hannah and Raphael, for your kindness and companionship (in life and in Caffè Nero) without which I may never have written a single chapter. Momo, I am grateful for every second we've spent together. You have filled my life with colour and joy. Thank you.

A warm thank-you now goes to each of my friends from Oriel (directly and indirectly), Ada, Alex C, Angela and Katya, the nickname of the group of whom does not bear repeating here. A second goes to my friends from Worcester (and St Hugh's), Alex S, Anna, Emily, Emma, Jáchym, Jordan, Molly, Teo and Timms, alongside of course The Snakes: Andy R, Chris, Jack, Kieron and Luke. Here's to hoping I can remember the late nights of our next eight years of friendship a little better! A special shout-out goes to Chris, not just a friend to me, but (most Fridays at 5 till 6-ish) a friend to all in the Andrew Wiles Building. To Andy A, Beth, Brad and Olly; thank you for maintaining our friendship so strongly ever since school (even despite 'Boris' never really admitting a cool four-letter nickname).

Lastly, I offer all my love to my family, whose patience with me for the past twenty-six years has been (and continues to be) a far more impressive feat than anything you can find in the coming 200+ pages.

Contents

List of abbreviations

AD . . . . . Advection–diffusion
AV . . . . . Auxiliary variable
BBM . . . . . Benjamin–Bona–Mahony
BC . . . . . Boundary condition
CG . . . . . Continuous Galerkin
CIP . . . . . Continuous interior penalty/penalisation
CMT . . . . . Contraction mapping theorem, a.k.a. Banach's fixed point theorem
CPG . . . . . Continuous Petrov–Galerkin
DAE . . . . . Differential–algebraic equation
DG . . . . . Discontinuous Galerkin
DoF . . . . . Degree of Freedom
ELM . . . . . Edge-localised mode
EM . . . . . Electromagnetic
FE . . . . . Finite element
FEEC . . . . . Finite element exterior calculus
FEM . . . . . Finite element method
FET . . . . . Finite elements in time
GL . . . . . Gauss–Legendre
GJP . . . . . Gradient jump penalty/penalisation
HCT . . . . . Hsieh–Clough–Tocher
KdV . . . . . Korteweg–de Vries
IC . . . . . Initial condition
IM . . . . . Implicit midpoint
IP . . . . . Interior penalty
IPM . . . . . Interior penalty method
IVP . . . . . Initial-value problem

LB–G . . . . . Method of LaBudde & Greenspan [LG74]

LHS . . . . . Left-hand side

LM . . . . . Lagrange multiplier

MEEVC . . . . Mass-, energy-, enstrophy-, vorticity-conserving

MHD . . . . . Magneto-hydrodynamics

MS . . . . . Morgan–Scott

MV–DG . . . . Mean-value (or averaged-vector-field) discrete-gradient method of McLachlan, Quispel & Robidoux [MQR99]

NS . . . . . Navier–Stokes

ODE . . . . . Ordinary differential equation

PDE . . . . . Partial differential equation

QoI . . . . . Quantity of interest

RFP . . . . . Reversed-field pinch

RHS . . . . . Right-hand side

RK . . . . . Runge–Kutta

SP . . . . . Structure-preserving/Structure preservation

SV . . . . . Scott–Vogelius

SVV . . . . . Spectral vanishing viscosity

"It's the job that's never started as takes longest to finish, as my old gaffer used to say."

— Samwise Gamgee [Tol54]

1 Introduction

Contents

Structure-preserving (SP) numerical methods for initial-value problems (IVPs) have proven to be essential tools in the accurate modelling of ordinary (ODE) and partial (PDE) differential equations, particularly in long-time simulations. Alternatively known as geometric numerical integrators, such methods aim not just to deliver approximate solutions, but to replicate, at the discrete level, key geometric structures of the continuous problem, including symmetries, symplecticity, invariants, and dissipation inequalities. In the modern day, the field is mature and widely used; we refer the reader, for instance, to the works of Sanz-Serna & Calvo [SC94], Budd & Piggott [BP03], Hairer *et al.* [HLW06; Hai+06], Christiansen, Munthe-Kaas & Owren [CMO11], Blanes & Casas [BC17] or Iserles & Quispel [IQ18].

Despite extensive development, however, there have long remained difficulties in devising integrators that preserve non-quadratic dissipation laws or invariants. It can well be argued that the focus in the literature has traditionally been primarily on quadratic structures, with schemes that preserve non-quadratic structures being developed generally on a case-by-case basis, in particular in the discretisation of PDEs.

In this thesis we aim to address this issue by presenting a general framework for the construction of SP integrators for systems of both ODEs and PDEs, that preserve arbitrarily many such structures for arbitrary quantities of interest (QoIs); moreover, the approach extends to arbitrary order in time, and in the PDE case to arbitrary spatial discretisations. The approach combines two key ideas: first, the use of finite elements in time (FET) allows us to interact (at the discrete level) with the fundamental theorem of calculus, dictating the change in our QoIs over time intervals; second, the systematic introduction of auxiliary variables (AVs) ensures we can reproduce the evolution law of each considered QoI discretely, by reproducing a discrete form of the proof. Our key contribution lies in the straightforward procedure we propose for the application of these ideas, for general QoIs and general systems.

To demonstrate our framework, we consider a broad class of physically meaningful QoIs for a diverse set of systems. For instance, we construct integrators that preserve all known invariants, including e.g. energies and Casimirs, of general conservative ODE systems. Within the GENERIC formalism, we extend the method to PDEs, deriving schemes that are both energy-conserving and entropy-generating; in particular, this includes certain kinetic-type equations such as the Boltzmann equation. For the compressible Navier–Stokes (NS) equations, we construct finite element (FE) schemes that conserve mass, momentum, and energy, and satisfy a discrete entropy inequality.

The framework further develops and generalise a number of existing schemes in the literature. Within ODE applications, for example, these include the energy-preserving integrators developed by Hairer, Cohen & Lubich [CH11; HL14], and the energy- and entropy-stable GENERIC integrator of Romero [Rom09]. Within PDE applications, we highlight our generalisations of the energy- and helicity-stable incompressible NS integrators of Rebholz [Reb07], and the general energy-stable integrators of Giesselmann, Karsai & Tscherpel [GKT25].

A key strength of the framework is its flexibility in space for PDE discretisations; in particular, as stated above, we typically impose no constraints on the spatial discretisation. When using FE spaces that satisfy certain compatibility conditions deriving from finite element exterior calculus (FEEC), we show that our SP schemes may simplify significantly, allowing for efficient implementation and more elegant

mixed formulations. We demonstrate this with a novel energy-stable velocity–vorticity FE integrator for the incompressible NS equations that preserves the evolution of the enstrophy (a dissipation inequality in 2D), and with an energy- and helicity-stable scheme for incompressible Hall magneto-hydrodynamics (MHD) that generalises the work of Laakmann, Hu & Farrell [LHF23].

By leveraging our schemes' SP properties, we also prove several analytic results. For advection–diffusion (AD) PDEs, such as incompressible NS and MHD, we establish the existence and, under certain stronger conditions, uniqueness of discrete solutions. For ODEs, we prove both the existence of unique solutions, and their convergence.

We explore also the extension of our framework to adiabatic invariants, quantities that are not necessarily conserved, but evolve slowly compared to fast system dynamics. The Lorentz problem considers the motion of a charged particle in a strong magnetic field; using this as an illustration, we demonstrate how a modified idea of the AVs introduced by our framework can lead to an energy-conserving integrator that further preserves the adiabatic invariance of the magnetic moment.

Throughout the thesis, we complement theoretical developments with numerical simulations, demonstrating our integrators' improved accuracy and stability. These include examples ranging from the Kepler problem and Kovalevskaya top (when considering general conservative ODEs) and the Lorentz problem, to the incompressible and compressible NS equations.

The framework presented in this thesis was originally proposed in the preprint [AF25] alongside the example applications used here for the energy- and helicity-stable incompressible NS integrator, and the mass-, momentum-, energy- and entropy-stable integrator for the compressible NS equations. Unless otherwise stated (e.g. in the case of the energy- and helicity-stable integrator the incompressible Hall MHD equations) we believe all remaining schemes introduced in this thesis to be novel in the literature; in particular these include those for general conservative ODE systems, GENERIC PDE systems (including the Boltzmann equation), the Lorentz problem, and the energy- and enstrophy-stable incompressible NS integrator.

1.1 Code availability

The code that was used to generate the numerical results in Figs. 6.1, 6.2 & 10.1 was written in Python using `NumPy` [Har+20], with the former two figures using `PETSc` [Bal+24]. The code for Fig. 6.4 & 6.5 was written in `MATLAB` [The23a] using the `Optimization Toolbox` [The23b].

All remaining numerical simulations were done in `Firedrake` [Ham+23], with the Gauss method in Figs. 7.1 & 7.2 using `Irksome` [FKM21]. Code for reproducing the numerical results of this work can be found at [And25].

1.2 Overview

This thesis is partitioned into three parts. Each includes an introductory chapter; these introductions discuss the relevant motivation and literature for the material discusses in the corresponding part, and an overview of the contents of each contained chapter.

In Part I, we define our general framework for the construction of geometric numerical integrators for IVPs that preserve multiple general conservation and dissipation structures, via FET and the systematic introduction of AVs. We further offer certain preliminary analytic results for our schemes in the case of AD PDEs, demonstrating the existence of solutions and, under certain stronger criteria, their uniqueness. We further discuss the implementation of our schemes, in particular discussing those cases in which the proposed AVs may be pre-computed, and do not require introduction on the computational level. As a running example, we consider the incompressible NS equations, in which both the energy and helicity are conserved in the ideal limit, and energy is dissipated otherwise; we refer to these properties as energy and helicity stability. We apply our framework to derive an energy- and helicity-stable integrator (a high-order-in-time generalisation of that of Rebholz [Reb07]), i.e. one that preserves the conservation and dissipation properties of the energy and helicity discretely. The aforementioned analysis allows to consider the existence and uniqueness of solutions to our discrete scheme, while our notes on the implementation of our schemes indicate how a certain AV (approximating the velocity) may be eliminated on the computational level for more efficient implementation.

In Part II, we demonstrate the framework through application to various conservative and dissipative systems of ODEs and PDEs. For ODEs, we consider Poisson and gradient-descent systems, general conservative systems, and ODEs derived from the GENERIC formalism. In the first case, our work reproduces the schemes of Hairer, Cohen & Lubich [CH11; HL14]; in the last it represents an extension of that of Romero [Rom09] to arbitrary order in time. For the general conservative integrator, however, we believe our scheme to be novel. We establish uniqueness and convergence results that are generally applicable to each of these schemes. For PDEs, we again consider Poisson and gradient-descent systems, PDEs derived from the GENERIC formalism (including a form of the Boltzmann equation), and the compressible NS equations; for the Poisson and gradient-descent systems, our scheme resembles that of Giesselmann, Karsai & Tscherpel [GKT25], however in the latter two cases we again believe our proposed discretisations to be novel. Various numerical examples demonstrate the benefits of the SP schemes, including for the Kepler problem, the Benjamin–Bona–Mahony (BBM) equation (an example Poisson system), and a shockwave formation in the compressible NS equations.

In Part III, we discuss two extensions of the framework: the preservation of certain adiabatic invariants (see Henrard [Hen93] or Arnold, Kozlov & Neishtadt [AKN06]) and connections with FEEC (see the original work of Hiptmair [Hip01] or Arnold, Falk & Winther [AFW06; AFW09; Arn18]). In the former, we construct energy-stable integrators for charged particles in strong magnetic fields that preserve the adiabatic invariance of the magnetic moment; this requires a generalised notion of the AVs introduced by our framework. In the latter, we revisit the energy- and helicity-stable integrators of Part I, using FEEC to eliminate a certain Lagrange multiplier (LM) from the discretisation, simplifying the computational application. We then derive a novel energy-stable integrator for the incompressible NS equations that preserves the evolution of enstrophy, in particular a dissipation inequality in the 2D case; we show that, through FEEC, this may be written in an amenable mixed velocity–vorticity form. We conclude by considering the preservation of energy and helicity stability in the incompressible Hall MHD equations; after extensive manipulation, we are able to show an energy- and helicity-stable scheme proposed by our framework is equivalent to a high-order generalisation of that proposed by Laakmann, Hu & Farrell [LHF23]. Each of the schemes proposed here can be

analysed through the results for general AD systems established in Part I, giving both the existence of solutions and their uniqueness.

Part I

A general framework for geometric numerical integration: conservation laws & dissipation inequalities

„ { *"What one fool can do, another can."* — *Ancient Simian Proverb* } „

— *Silvanus P. Thompson [Tho10]*

2 Introduction

Contents

This part of the thesis introduces our framework for the construction of conservative and accurately dissipative integrators for ODEs and PDEs. Typical approaches for the construction of numerical integrators for PDEs handle the spatial and temporal discretisations separately, the latter typically being done through Runge–Kutta (RK) methods; in contrast, our framework relies on the discretisation of variational problems posed in space-time, albeit over a single time interval, i.e. we do not solve for all times simultaneously. This represents the first of two key ideas in our approach. The second is the systematic introduction of AVs. Through our framework, each structure to be preserved can be connected to a certain associated test function; for each structure, we introduce a specific AV, a certain projection of this associated test function over a space-time domain. These AVs are then coupled back into the original system, creating an SP mixed method.

Incompressible Navier–Stokes, energy, helicity & topology

As a well-studied example of wide interest, we demonstrate our framework through the design of a stable integrator for the NS equations. The first structure we consider

therein is the dissipation (or conservation in the ideal limit) of energy $\frac{1}{2}\|\mathbf{u}\|^2$, where $\mathbf{u}$ is the flow velocity and $\|\cdot\|$ denotes the $\mathbf{L}^2$ norm. This is arguably the most fundamental structure within the NS equations, and crucial to their analysis (see Temam [Tem24] or Girault & Raviart [GR12]); the preservation of energy stability, i.e. the construction of schemes that preserve the dissipation (or conservation) or energy discretely, is therefore essential for the design of well-posed numerical integrators. We use our framework to construct an energy-stable FE scheme for the NS equations, and exploit the preserved energy law in our analysis in Section 3.3 to prove the existence and (under certain conditions) uniqueness of solutions to our discretisation.

The second structure we consider relates to the topology of the incompressible NS equations in 3D. In the ideal case and in the absence of external forces, vortex lines (i.e. streamlines of the vorticity field $\operatorname{curl}\mathbf{u}$) are convected by the flow (see Arnold & Khesin [AK08, Chap. I Cor. 5.11]). As a consequence, the topology of the vortex lines is preserved over time; that is to say, if the initial vortex lines are twisted or knotted, they must remain equivalently twisted or knotted. First observed by Moreau [Mor61] in 1961, the ideal conservation of the fluid helicity $\frac{1}{2}(\mathbf{u}, \operatorname{curl}\mathbf{u})$, where $(\cdot,\cdot)$ denotes the $\mathbf{L}^2$ inner product, is an important (but not equivalent) consequence of this topological persistence property. A result from Arnold [Arn14] in 1974, often called the helicity theorem (see Arnold & Khesin [AK08, Chap. III Th. 4.4]), offers an intuitive topological interpretation of the helicity: for general divergence-free fields in 3D, the helicity is in a certain specific sense a continuous analogue of the linking number, a discrete topological invariant that quantifies the linking of closed curves in 3D (see Cantarella *et al.* [Can+99]). Conservation of the helicity at the discrete level therefore goes some way to the preservation of vortex line topology, ensuring that numerical solutions are unable to untie certain twists in the flow on a global scale. See the review paper of Moffatt & Tsinober [MT92] for a further discussion of the importance of helicity for the dynamics of 3D flows. In our scheme, we are able to preserve both the dissipation of energy and, in the ideal case, the conservation of helicity, for discretisations of arbitrary order in space and time.

2.1 Related literature

Relevant literature relating to specific applications of our framework (i.e. existing stable integrators for previously studied systems) will be reviewed in Parts II & III when these applications are introduced. We restrict our review here to the core techniques applied in our framework and their connection to SP, alongside energy and helicity stable integrators for the incompressible NS equations.

Continuous Petrov–Galerkin & structure preservation

The connections between continuous Petrov–Galerkin (CPG) and SP date back to 1990, when French & Schaeffer [FS90] first observed that CPG time discretisations are SP for many problems. When applied to certain conservative ODE systems, CPG naturally conserves the energy; when applied to gradient descent systems, CPG naturally dissipates the energy. The authors also report that CPG applied to some (but not all) Hamiltonian PDEs is again conservative. In each of these cases our framework recovers the discretisation proposed by French & Schaeffer when applied to the same systems, choosing $\mathcal{I}_n$ (see Step **B** of the framework below) to be the exact integral.

Betsch & Steinmann extended these observations to general Hamiltonian ODEs written in canonical coordinates [BS00b; BS00a] and applied the technique to develop an energy-conserving scheme for elastodynamics [BS01]. This was generalised further by Egger, Habrich & Shashkov [EHS21] to a broad class of conservative or dissipative PDEs written with a skew-symmetric or semidefinite operator acting on the time derivative. Celledoni & Jackaman [CJ21] observed that CPG is energy-conserving for multisymplectic systems. In each of these cases, our framework recovers the proposed discretisations when applied to these problems.

Systematic introduction of auxiliary variables

The idea of using AVs to preserve conservation and dissipation structures dates back to the discrete gradient method of McLachlan, Quispel & Robidoux in 1999 [MQR99]. The authors employ the discrete gradient concept introduced by Gonzalez [Gon96] to derive one-stage energy-conserving discretisations of energy-conserving systems, and dissipative discretisations of gradient descent systems. Under certain conditions,

these discrete gradients identify with the AVs introduced by our framework, in particular when considering the mean-value discrete gradient of Harten, Lax & van Leer [HLL83].

This approach was generalised to higher order in time discretisations by Cohen, Hairer & Lubich [CH11; HL14]. The discrete variational derivative represents an alternative extension of this idea to PDEs [FM10; DO11].

Other methods

Brugnano, Iavernaro & Frasca-Caccia [BI12; BI16; BFI19] have developed line integral methods for conservative ODEs and PDEs. These schemes are closely related to Gauss methods, the framework of Cohen & Hairer [CH11], and CPG schemes with a particular choice of quadrature rule. Of particular relevance to our work is their method for enforcing conservation of invariants other than the energy; they devise a systematic way to perturb the discrete system of Cohen & Hairer in such a way that retains energy conservation and the same order of accuracy, but also conserves other invariants [BI16, Sec. 6.1]. In contrast to our approach, their scheme requires the use of at least as many stages as invariants to be preserved; in particular, the authors devise a 3-stage time discretisation for the Kepler problem that conserves all invariants.

A framework for the construction of SP modifications to explicit RK schemes for certain conservative PDEs was proposed in [EG22]. This was extended to implicit-explicit schemes for systems with a parabolic component in [EG23].

We note in passing the scalar auxiliary variable method of Shen, Xu & Yang [SXY18] which, in contrast to our approach, introduces a single real auxiliary variable involving the energy, rather than a field approximating its gradient.

Energy- and helicity-stable integrators for the incompressible Navier–Stokes equations

As arguably the most fundamental structure in arguably the most fundamental nonlinear equation in the study of numerical PDEs, the literature on the design of energy-stable integrators for the incompressible NS equations is vast. We can, however, trace its origins back to the late 1950s & 1960s.

In 1959, Phillips [Phi59] observed that, when improperly handled, nonlinear advective terms could lead to the breakdown of solutions for numerical integrators. It was shown by Arakawa [Ara66] 7 years later that these instabilities could be avoided by conserving certain quadratic norms on the solution; early finite-difference schemes to do so include those of Harlow & Welch [HW65] and Arakawa, Mesinger & Lamb [MA76; AL77]. In 1970, Piacsek & Williams [PW70] connected the conservation of energy with the preservation of the skew-symmetry of the advective operator on the discrete level. A general framework for the construction of 1-stage, energy-dissipative schemes for the incompressible NS equations was proposed by Simo & Armero in 1994 [SA94]; these correspond to schemes deriving from our framework with $\tilde{F}$ as defined in (3.23) below.

At lowest order in time and after elimination of the AV approximating the velocity (see Section 4.2), our energy- and helicity-stable NS integrator aligns exactly with the 1-stage energy- and helicity-stable scheme proposed by Rebholz [Reb07]; our scheme can therefore be interpreted as a generalisation of this scheme to higher order in time. The uniqueness analysis presented by Rebholz in Subsection 3.1 therein aligns with ours presented in Subsection 3.3.4, albeit with a more careful handling from Rebholz in the continuous setting such that the results are stable under mesh refinement; see also Subsection 4 for a proof of convergence of the scheme of Rebholz.

Zhang *et al.* [Zha+22] proposed an energy- and helicity-stable dual-field discretisation for the incompressible NS equations. Their scheme introduces two variables each for the velocity and vorticity (for a total of four variables) the updates of which are alternated in a staggered fashion, such that each step in the resulting scheme is linear.

Finite element software with support for finite elements in time

While some publicly and commercially available FE software packages support domains of high dimensions (> 3) these implementations are typically not optimised for the use of FET.

The best support in this area comes perhaps in `Firedrake` [Ham+23]. `Fetsome` [La 22] extended the `Irksome` library [FKM21] with some initial support for variational-in-time integrators; work to include more general FET schemes in `Firedrake` is ongoing. The numerical simulations for this thesis rely largely on the use of

`Firedrake`, with the FET aspects implemented using a package made specifically for this thesis.

2.2 Overview

In Chapter 3, we define our general framework for the construction of geometric numerical integrators. As stated above, to fix ideas we consider the incompressible NS equations, constructing FE schemes that preserve both the energy and helicity stability to arbitrary order in space and time. We demonstrate the scheme with a numerical test on an example vortex with varying Reynolds number, comparing our scheme to a classical energy-stable discretisation. While we do not analyse all schemes that may derived from our framework,[1] we present an analysis for SP discretisations of general AD systems, using the preserved structures (in particular the preserved energy stability) to prove certain existence and uniqueness results. Its generality makes it applicable to various other SP discretisations derived from our framework (see Chapter 10).

In Chapter 4, we discuss certain aspects of the implementation of our schemes. Since many commercial FE software options do not support space-time domains, we discuss how we can use tensor-product constructions of our space-time spaces to circumvent these shortcomings. We further discuss how, under certain circumstances (including most SP methods for ODEs), the AVs introduced by our framework can be computed explicitly offline for all solution variables, thus eliminating the need to introduce them in the implementation, and conclude with a discussion about the connections between the SP schemes deriving from our framework and Gauss collocation methods [HLW06, Sec. II.1.3].

[1] Such a complete analysis would be comparable with an analysis of all possible FE methods.

"Marty, you're not thinking fourth dimensionally!"

— Dr. Emmett L. 'Doc' Brown, 1955 version (Christopher A. Lloyd) [Zem90]

3

The general framework: energy- & helicity-stable integrators for the incompressible Navier–Stokes equations

Contents

We present now the general framework. We will show how it can be used to modify a certain given discrete timestepping scheme (e.g. RK/FET) for a transient system to preserve chosen conservation laws or dissipation inequalities. In particular, the ideas will be presented for PDE systems; the extension to ODE systems follows by replacing discrete function spaces with suitable finite-dimensional vector spaces.

Example (Incompressible NS)

To fix ideas, throughout this chapter we will employ the incompressible NS equations as our running example. The equations can be written in strong form as

$$\dot{\mathbf{u}} = \mathbf{u} \times \operatorname{curl} \mathbf{u} - \nabla p + \frac{1}{\mathrm{Re}} \Delta \mathbf{u}, \tag{3.1a}$$

$$0 = \operatorname{div} \mathbf{u}, \tag{3.1b}$$

over a bounded Lipschitz domain $\Omega \subset \mathbb{R}^3$, where $\times$ denotes the cross product, Δ denotes the Laplacian, and throughout this thesis $\dot{*}$ is shorthand for the partial derivative with respect to time t. Here $\mathbf{u} : \mathbb{R}_+ \times \Omega \to \mathbb{R}^3$ is the velocity, $p : \mathbb{R}_+ \times \Omega \to \mathbb{R}$ is the total (or Bernoulli) pressure, and $\mathrm{Re} > 0$ is the Reynolds number; we use a rotational (or Lamb) form for the advective term. We consider periodic boundary conditions (BCs) with the additional constraint on the initial condition (IC)

$$\int_\Omega \mathbf{u}(0) = \mathbf{0}. \tag{3.2}$$

If (3.2) holds then any solution to (3.1) satisfies $\int_\Omega \mathbf{u} = \mathbf{0}$ at all times. This condition is included to ensure certain energy estimates exist on solutions to the scheme; namely, we require that $\|\nabla \mathbf{u}\|$ defines a norm on $\mathbf{u}$ where $\|\cdot\|$ denotes the $L^2(\Omega)$ norm.

Define the energy $Q_1(\mathbf{u}) \coloneqq \frac{1}{2}\|\mathbf{u}\|^2$ and helicity $Q_2(\mathbf{u}) \coloneqq \frac{1}{2}(\mathbf{u}, \operatorname{curl} \mathbf{u})$, where $(\cdot,\cdot)$ denotes the $L^2(\Omega)$ inner product. Under periodic BCs, Q_1 and Q_2 are conserved in solutions of the formal ideal limit $\mathrm{Re} = \infty$, while Q_1 is necessarily dissipated for $\mathrm{Re} < \infty$; we wish to construct a timestepping scheme that preserves these behaviours.

The rest of this chapter proceeds as follows. In Section 3.1, we present our general framework for constructing stable FE integrators, using the incompressible NS equations as a running example. In Section 3.2, we demonstrate the SP properties of our derived stable NS discretisations numerically, through a series of simulations on a vortex at varying Re. In Section 3.3, we present some preliminary analytic results for certain schemes deriving from our framework; in particular, we prove existence

and uniqueness results for stable integrators deriving from our framework for general AD systems, again using the SP incompressible NS discretisation to fix ideas.

3.1 Definition of the framework

A. Definition of semidiscrete form

We define first an abstract semidiscrete formulation of a transient PDE, discretised in space only. This is posed over a general affine space

$$\mathbb{X} := \left\{ u \in C^1(\mathbb{R}_+; \mathbb{U}) : u(0) \text{ satisfies known initial data} \right\}. \tag{3.3}$$

Here, $\mathbb{U}$ denotes an appropriate finite-dimensional spatial function space. The abstract semidiscrete weak problem is then as follows: find $u \in \mathbb{X}$ such that

$$M(u; \dot{u}, v) = F(u; v) \tag{3.4}$$

at all times $t \in \mathbb{R}_+$ and for all $v \in \mathbb{U}$. Here $M : \mathbb{U} \times \mathbb{U} \times \mathbb{U} \to \mathbb{R}$ is possibly nonlinear in u, but linear in $\dot{u}$ and v; this is the significance of the semicolon. Similarly, $F : \mathbb{U} \times \mathbb{U} \to \mathbb{R}$ is possibly nonlinear in u, but linear in the test function v.

Remark 3.1 (Clarifications on the distinction between $\mathbb{U}$ and $\mathbb{X}$)**.** *To avoid confusion, we clarify here a certain distinction: the space $\mathbb{U}$, defined in space only, is a function space, i.e. we require $0 \in \mathbb{U}$; the space $\mathbb{X}$, defined in space and time, is an affine space, i.e. in general $0 \notin \mathbb{X}$ due to the imposed IC on $u(0)$ for $u \in \mathbb{X}$. In particular, since $\mathbb{U}$ must be a function space, our framework is in general only applicable to problems with homogeneous Dirichlet, Neumann or periodic BCs that need not be enforced strongly.*

Example (Incompressible NS)

Working with the NS example, we construct a simple semi-discretisation, which we show can be written in the form (3.4).

Let $\mathbb{V}$, $\mathbb{Q}$ be periodic finite-dimensional function spaces (typically FE spaces) vector-valued for the velocity and scalar-valued for the pressure respectively. To work in the most general case possible, we consider two common alternative semi-discretisations for (3.1). The first formulation is: find

$(\mathbf{u}, p) \in C^1(\mathbb{R}_+; \mathbb{V}) \times C^0(\mathbb{R}_+; \mathbb{Q})$, satisfying known ICs in $\mathbf{u}$ such that

$$(\dot{\mathbf{u}}, \mathbf{v}) = (\mathbf{u} \times \operatorname{curl} \mathbf{u}, \mathbf{v}) + (p, \operatorname{div} \mathbf{v}) - \frac{1}{\mathrm{Re}}(\nabla \mathbf{u}, \nabla \mathbf{v}), \tag{3.5a}$$

$$0 = (\operatorname{div} \mathbf{u}, q), \tag{3.5b}$$

at all times $t \in \mathbb{R}_+$ and for all $(\mathbf{v}, q) \in \mathbb{V} \times \mathbb{Q}$. The second is similar, with variational equations

$$(\dot{\mathbf{u}}, \mathbf{v}) = (\mathbf{u} \times \operatorname{curl} \mathbf{u}, \mathbf{v}) - (\nabla p, \mathbf{v}) - \frac{1}{\mathrm{Re}}(\nabla \mathbf{u}, \nabla \mathbf{v}), \tag{3.6a}$$

$$0 = -(\mathbf{u}, \nabla q). \tag{3.6b}$$

Up to regularity, the formulations (3.5) and (3.6) are equivalent.

In its current form, there are no time derivatives on the pressure term p, implying neither (3.5) nor (3.6) can be written in the form of (3.4).[a] To remedy this, we define a discretely divergence-free subspace $\mathbb{U} \subset \mathbb{V}$, as either

$$\mathbb{U} := \left\{ \mathbf{u} \in \mathbb{V} : (\operatorname{div} \mathbf{u}, q) = 0 \text{ for all } q \in \mathbb{Q} \text{ and } \int_\Omega \mathbf{u} = \mathbf{0} \right\}, \tag{3.7a}$$

or

$$\mathbb{U} := \left\{ \mathbf{u} \in \mathbb{V} : -(\mathbf{u}, \nabla q) = 0 \text{ for all } q \in \mathbb{Q} \text{ and } \int_\Omega \mathbf{u} = \mathbf{0} \right\}. \tag{3.7b}$$

We can then effectively eliminate both p and the mass conservation equation (3.5b) or (3.6b), while further incorporating the condition $\int_\Omega \mathbf{u} = \mathbf{0}$, by posing the semi-discretisation in $\mathbb{U}$; the solution space $\mathbb{X}$ is then defined from $\mathbb{U}$ as in (3.3). The general semi-discretisation, representing either (3.5) or (3.6) depending on the choice (3.7a) or (3.7b) of $\mathbb{U}$, then states: find $\mathbf{u} \in \mathbb{X}$ such that

$$(\dot{\mathbf{u}}, \mathbf{v}) = (\mathbf{u} \times \operatorname{curl} \mathbf{u}, \mathbf{v}) - \frac{1}{\mathrm{Re}}(\nabla \mathbf{u}, \nabla \mathbf{v}) \tag{3.8}$$

at all times $t \in \mathbb{R}_+$ and for all $\mathbf{v} \in \mathbb{U}$. The discrete divergence-free conditions imposed on $\mathbb{U}$ in either (3.5) or (3.6) would then, in implementation, be enforced in (3.8) through LMs, reproducing (3.7a) or (3.7b) respectively, with a further LM enforcing the condition $\int_\Omega \mathbf{u} = \mathbf{0}$ (see (3.32) at the end of this section).

This is in the form of (3.4) with M, F given by

$$M(\mathbf{u}; \dot{\mathbf{u}}, \mathbf{v}) \coloneqq (\dot{\mathbf{u}}, \mathbf{v}), \tag{3.9a}$$

$$F(\mathbf{u}; \mathbf{v}) \coloneqq (\mathbf{u} \times \operatorname{curl} \mathbf{u}, \mathbf{v}) - \frac{1}{\mathrm{Re}}(\nabla \mathbf{u}, \nabla \mathbf{v}). \tag{3.9b}$$

[a] This is because (3.1) represents a differential–algebraic equation (DAE) [AP98].

B. Definition of timestepping scheme

To make this fully discrete, we define a finite-dimensional affine space $\mathbb{X}_n$ over the timestep $T_n = [t_n, t_{n+1}]$. We employ polynomials in time of degree $S \geq 1$:

$$\mathbb{X}_n \coloneqq \{u \in \mathbb{P}_S(T_n; \mathbb{U}) : u(t_n) \text{ satisfies known initial data}\}. \tag{3.10}$$

We further introduce $\mathcal{I}_n[\phi]$, a general linear operator (quadrature rule) that approximates the integral

$$\mathcal{I}_n[\phi] \approx \int_{T_n} \phi. \tag{3.11}$$

The approximation must be sign-preserving, i.e.

$$\phi \geq 0 \implies \mathcal{I}_n[\phi] \geq 0, \tag{3.12a}$$

appropriately scaled in $\Delta t_n \coloneqq t_{n+1} - t_n$, i.e.

$$\mathcal{I}_n[1] = \Delta t_n, \tag{3.12b}$$

and the map $\phi \mapsto \mathcal{I}_n[\phi^2]^{\frac{1}{2}}$ must define a norm on $\mathbb{P}_{S-1}(T_n)$, the space of degree-$(S-1)$ polynomials on T_n. Examples of such linear operators include the exact integral, and any S-stage quadrature rule with positive weights.

The abstract timestepping scheme is then as follows: find $u \in \mathbb{X}_n$ such that

$$\mathcal{I}_n[M(u; \dot{u}, v)] = \mathcal{I}_n[F(u; v)] \tag{3.13}$$

for all $v \in \dot{\mathbb{X}}_n = \mathbb{P}_{S-1}(T_n; \mathbb{U})$.

Example (Incompressible NS)

No specific choice of $\mathcal{I}_n$ is required. For our running example, we might choose $\mathcal{I}_n$ to be a Gauss–Legendre (GL) quadrature rule, yielding a Gauss collocation method.

C. Identification of associated test functions

The properties we wish to preserve (conservation laws or dissipation structures) are associated with particular choices of test functions. For Fréchet-differentiable QoIs $(Q_p : \mathbb{U} \to \mathbb{R})_{p=1}^P$, we assume there exist test functions $(w_p(u))_{p=1}^P$, where each w_p is a functional acting on u, such that the Fréchet derivatives $Q_p'(u; v) = M(u; v, w_p(u))$ for general u, v. Consequently, for u an exact solution of the PDE,

$$Q_p(u(t_{n+1})) - Q_p(u(t_n)) = \int_{T_n} Q_p'(u; \dot{u}) = \int_{T_n} M(u; \dot{u}, w_p(u)) = \int_{T_n} F(u; w_p(u)). \tag{3.14}$$

Note, no constraints are posed here on the space containing $w_p(u)$; it is not generally true that $w_p(u) \in \mathbb{U}$. For each p, the behaviour of Q_p is then encoded in the sign of $F(u; w_p(u))$; in particular for conserved Q_p, $F(u; w_p(u)) = 0$, whereas for dissipated Q_p, $F(u; w_p(u)) \leq 0$.

Example (Incompressible NS)

We consider two QoIs,

$$Q_1(\mathbf{u}) \coloneqq \frac{1}{2}\|\mathbf{u}\|^2, \qquad Q_2(\mathbf{u}) \coloneqq \frac{1}{2}(\mathbf{u}, \operatorname{curl} \mathbf{u}), \tag{3.15}$$

the kinetic energy and the helicity respectively. Consider $\mathbf{u}$ the exact solution of (3.1): for the kinetic energy Q_1,

$$\begin{aligned} Q_1(\mathbf{u}(t_{n+1})) - Q_1(\mathbf{u}(t_n)) & \\ &= \int_{T_n} (\mathbf{u}, \dot{\mathbf{u}}) && \text{(3.16a)} \\ &= \int_{T_n} \left[(\mathbf{u} \times \operatorname{curl} \mathbf{u}, \mathbf{u}) - \frac{1}{\mathrm{Re}} (\nabla \mathbf{u}, \nabla \mathbf{u}) \right] && \text{(3.16b)} \\ &= -\frac{1}{\mathrm{Re}} \int_{T_n} \|\nabla \mathbf{u}\|^2 \leq 0; && \text{(3.16c)} \end{aligned}$$

for the helicity Q_2,

$$\begin{aligned} &Q_2(u(t_{n+1})) - Q_2(u(t_n)) \\ &\qquad = \int_{T_n} (\operatorname{curl} \mathbf{u}, \dot{\mathbf{u}}) && \text{(3.17a)} \\ &\qquad = \int_{T_n} \Big[(\mathbf{u} \times \operatorname{curl} \mathbf{u}, \operatorname{curl} \mathbf{u}) - \frac{1}{\mathrm{Re}} (\nabla \mathbf{u}, \nabla \operatorname{curl} \mathbf{u}) \Big] && \text{(3.17b)} \\ &\qquad = -\frac{1}{\mathrm{Re}} \int_{T_n} (\nabla \mathbf{u}, \nabla \operatorname{curl} \mathbf{u}). && \text{(3.17c)} \end{aligned}$$

In each case we apply integration by parts (IBP) using the periodic BCs in $\mathbf{u}$, while the advection terms vanish due to properties of the cross product. Both (3.16, 3.17) align with (3.14) for respective associated test functions

$$\mathbf{w}_1(\mathbf{u}) \coloneqq \mathbf{u}, \qquad \mathbf{w}_2(\mathbf{u}) \coloneqq \operatorname{curl} \mathbf{u}. \tag{3.18}$$

Both Q_1 and Q_2 are conserved in the ideal limit $\mathrm{Re} = \infty$.

D. Introduction of AVs

Our aim is to replicate the conservation/dissipation properties (3.14) discretely. However, as it stands this cannot be done, as in general $w_p(u) \notin \dot{\mathbb{X}}_n$, so they are not valid choices of discrete test functions. We therefore introduce AVs $(\tilde{w}_p)_{p=1}^P$ into the formulation, computing approximations to the associated test functions $(w_p(u))$ within the discrete test space $\dot{\mathbb{X}}_n$. Namely, for all $p = 1, \dots, P$, $\tilde{w}_p \in \dot{\mathbb{X}}_n$ is defined weakly such that

$$\mathcal{I}_n[M(u; v_p, \tilde{w}_p)] = \int_{T_n} Q_p'(u; v_p) \quad \Big(= \int_{T_n} M(u; v_p, w_p(u)) \Big), \tag{3.19}$$

for all $v_p \in \dot{\mathbb{X}}_n$.

Example (Incompressible NS)

For the NS system, the auxiliary velocity and vorticity $(\tilde{\mathbf{u}}, \boldsymbol{\omega})$ $(= (\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2)) \in \dot{\mathbb{X}}_n^2$ are defined weakly such that

$$\mathcal{I}_n[(\tilde{\mathbf{v}}, \tilde{\mathbf{u}})] = \int_{T_n} (\mathbf{u}, \tilde{\mathbf{v}}), \qquad \mathcal{I}_n[(\boldsymbol{\chi}, \boldsymbol{\omega})] = \int_{T_n} (\operatorname{curl} \mathbf{u}, \boldsymbol{\chi}), \tag{3.20}$$

for all $(\tilde{\mathbf{v}}, \boldsymbol{\chi}) \in \dot{\mathbb{X}}_n^2$. In particular, whereas $\mathbf{u}$ is an approximation of the velocity that is continuous across time intervals of polynomial degree S, $\tilde{\mathbf{u}}$ is another approximation of the velocity with the same spatial discretisation that is discontinuous across time intervals of polynomial degree $S-1$.

Note, in the continuous case, the vorticity $\operatorname{curl} \mathbf{u}$ should satisfy both $\operatorname{div}[\operatorname{curl} \mathbf{u}] = 0$ and $\int_\Omega \operatorname{curl} \mathbf{u} = \mathbf{0}$, by $\operatorname{div} \circ \operatorname{curl} = 0$ and Stokes' theorem respectively. These results are analogous to the restrictions on $\mathbb{U}$ in (3.7); as such, it is appropriate to approximate $\operatorname{curl} \mathbf{u}$ by $\boldsymbol{\omega} \in \dot{\mathbb{X}}_n = \mathbb{P}_{S-1}(T_n; \mathbb{U})$.

Remark 3.2. *In some cases, certain $\tilde{w}_p$ can be computed explicitly, and are therefore not needed in the implementation; this in particular is the case for $\tilde{\mathbf{u}}$ $(= \tilde{\mathbf{w}}_1)$ in the incompressible NS example. We discuss this further in Section 4.*

E. Modification of RHS

We must define $\tilde{F}$, a modification of F in (3.13), so that when the test function is chosen to be an AV we recover the associated conservation or dissipation law.

More specifically, we require the construction of $\tilde{F} : \mathbb{U} \times \mathbb{U}^P \times \mathbb{U} \to \mathbb{R}$ with the following properties:

1. $\tilde{F}(u, (\tilde{w}_p); v)$ is linear in its final argument.

2. $\tilde{F}$ coincides with F when evaluated at the associated test functions: for all $(u, v) \in \mathbb{U} \times \mathbb{U}$,
$$\tilde{F}(u, (w_p(u)); v) = F(u; v), \tag{3.21}$$
when the left-hand side (LHS) is well-defined.

3. $\tilde{F}$ preserves the conservation/dissipation structures of F: for each $q = 1, \dots, P$,
$$\begin{aligned} &\text{if } F(u; w_q(u)) = 0, \text{ then } \tilde{F}(u, (\tilde{w}_p); \tilde{w}_q) = 0; \\ &\text{if } F(u; w_q(u)) \geq 0, \text{ then } \tilde{F}(u, (\tilde{w}_p); \tilde{w}_q) \geq 0; \\ &\text{if } F(u; w_q(u)) \leq 0, \text{ then } \tilde{F}(u, (\tilde{w}_p); \tilde{w}_q) \leq 0. \end{aligned} \tag{3.22}$$

This process is problem-specific, requires some judgement, and is best understood by example.

Example (Incompressible NS)

With the AVs defined to live in $\dot{\mathbb{X}}_n$, the choice $\mathbf{v} = \tilde{\mathbf{u}}\,(= \tilde{\mathbf{w}}_1)$ is now valid in (3.13). We wish to replicate the energy dissipation law (3.16) when this choice is made. By inspection, defining

$$\tilde{F}(\mathbf{u}, \tilde{\mathbf{u}}; \mathbf{v}) \coloneqq (\tilde{\mathbf{u}} \times \operatorname{curl} \mathbf{u}, \mathbf{v}) - \frac{1}{\mathrm{Re}}(\nabla \tilde{\mathbf{u}}, \nabla \mathbf{v}), \tag{3.23}$$

then when testing with $\mathbf{v} = \tilde{\mathbf{u}}$ in (3.23)

$$\tilde{F}(\mathbf{u}, \tilde{\mathbf{u}}; \tilde{\mathbf{u}}) = -\frac{1}{\mathrm{Re}}\|\nabla \tilde{\mathbf{u}}\|^2 \leq 0, \tag{3.24}$$

satisfying (3.22) for $q = 1$. To satisfy (3.22) for $q = 2$, we further modify (3.23) to recover the helicity law (3.17) by defining

$$\tilde{F}(\mathbf{u}, (\tilde{\mathbf{u}}, \boldsymbol{\omega}); \mathbf{v}) \coloneqq (\tilde{\mathbf{u}} \times \boldsymbol{\omega}, \mathbf{v}) - \frac{1}{\mathrm{Re}}(\nabla \tilde{\mathbf{u}}, \nabla \mathbf{v}), \tag{3.25}$$

such that when further testing with $\mathbf{v} = \boldsymbol{\omega}\,(= \tilde{\mathbf{w}}_2)$ in (3.25),

$$\tilde{F}(\mathbf{u}, (\tilde{\mathbf{u}}, \boldsymbol{\omega}); \boldsymbol{\omega}) = -\frac{1}{\mathrm{Re}}(\nabla \tilde{\mathbf{u}}, \nabla \boldsymbol{\omega}). \tag{3.26}$$

In the ideal case $\mathrm{Re} = \infty$, both (3.24, 3.26) evaluate as 0, preserving the conservation structures.

F. Construction of SP scheme

With $\tilde{F}$ defined, the final SP scheme is as follows.

Definition 3.3 (Final discretisation). *Find* $(u, (\tilde{w}_p)) \in \mathbb{X}_n \times \dot{\mathbb{X}}_n^P$ *such that*

$$\mathcal{I}_n[M(u; \dot{u}, v)] = \mathcal{I}_n[\tilde{F}(u, (\tilde{w}_p); v)], \tag{3.27a}$$

$$\mathcal{I}_n[M(u; v_p, \tilde{w}_p)] = \int_{T_n} Q_p'(u; v_p), \tag{3.27b}$$

for all $(v, (v_p)) \in \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n^P$.

Notably, the exact integral on the right-hand side (RHS) of (3.27b) cannot be substituted for $\mathcal{I}_n$ without breaking the SP properties of the scheme, as described in Theorem 3.4 below.

Example (Incompressible NS)

The final energy- and helicity-conserving scheme is as follows: find $(\mathbf{u}, (\tilde{\mathbf{u}}, \boldsymbol{\omega})) \in \mathbb{X}_n \times \dot{\mathbb{X}}_n^2$ such that

$$\mathcal{I}_n[(\dot{\mathbf{u}}, \mathbf{v})] = \mathcal{I}_n\Big[(\tilde{\mathbf{u}} \times \boldsymbol{\omega}, \mathbf{v}) - \frac{1}{\mathrm{Re}}(\nabla\tilde{\mathbf{u}}, \nabla\mathbf{v})\Big], \tag{3.28a}$$

$$\mathcal{I}_n[(\tilde{\mathbf{u}}, \tilde{\mathbf{v}})] = \int_{T_n} (\mathbf{u}, \tilde{\mathbf{v}}), \tag{3.28b}$$

$$\mathcal{I}_n[(\boldsymbol{\omega}, \boldsymbol{\chi})] = \int_{T_n} (\operatorname{curl} \mathbf{u}, \boldsymbol{\chi}), \tag{3.28c}$$

for all $(\mathbf{v}, (\tilde{\mathbf{v}}, \boldsymbol{\chi})) \in \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n^2$. In its current state, the discretisation (3.28) is not amenable to numerical implementation, due to LMs implicitly contained in the definition of $\mathbb{U}$ enforcing the discrete divergence-free conditions; see below for a definition of the scheme (3.28) with these LMs extracted.

Theorem 3.4 (SP of the framework)**.** *Where solutions to* (3.27) *exist, they preserve the sign of the changes to the functionals $Q_q(u)$, $q = 1, \dots, P$, across each timesteps. In particular, if $Q_q(u)$ is conserved by the exact solution, then it is also conserved by the discretisation (up to quadrature errors, solver tolerances, and machine precision).*

Proof. For each QoI Q_q,

$$Q_q(u(t_{n+1})) - Q_q(u(t_n)) = \int_{T_n} Q_q'(u; \dot{u}) = \mathcal{I}_n[M(u; \dot{u}, \tilde{w}_q)] = \mathcal{I}_n[\tilde{F}(u, (\tilde{w}_p); \tilde{w}_q)], \tag{3.29}$$

where the second equality holds by (3.27b), and the final by (3.27a). Thus, if $Q_q(u)$ is conserved by the exact solution, $\tilde{F}(u, (\tilde{w}_p); \tilde{w}_q) = 0$ by (3.22); by the linearity of $\mathcal{I}_n$, $Q_q(u)$ is conserved across timesteps. Otherwise, if $Q_q(u)$ is non-decreasing for the exact solution, $\tilde{F}(u, (\tilde{w}_p); \tilde{w}_q) \geq 0$ by (3.22); by the sign-preserving property of $\mathcal{I}_n$ (3.12a), $Q_q(u)$ is non-decreasing across timesteps. The same argument holds if $Q_q(u)$ is non-increasing. □

Example (Incompressible NS)

For the NS scheme (3.28) we find

$$Q_1(\mathbf{u}(t_{n+1})) - Q_1(\mathbf{u}(t_n)) = -\frac{1}{\mathrm{Re}}\mathcal{I}_n[\|\nabla\tilde{\mathbf{u}}\|^2] \leq 0, \tag{3.30a}$$

$$Q_2(\mathbf{u}(t_{n+1})) - Q_2(\mathbf{u}(t_n)) = -\frac{1}{\mathrm{Re}}\mathcal{I}_n[(\nabla\tilde{\mathbf{u}}, \nabla\boldsymbol{\omega})]. \tag{3.30b}$$

These identities resemble weak forms of (3.16, 3.17).

The framework is summarised in Algorithm 3.5.

Algorithm 3.5 Our proposed framework for constructing SP schemes for IVPs

A. Define the semidiscrete formulation
B. Define the timestepping scheme
C. Identify the associated (spatial) test functions
D. Introduce corresponding AVs
E. Modify the RHS of the weak form
F. Construct the SP scheme

If one begins with a symmetric timestepping scheme, the resulting scheme will inherit its symmetry, a property that is essential for accurately capturing the long-term behaviour of reversible systems [HLW06, Chap. V & XI].

Example (Incompressible NS)

Before demonstrating the scheme (3.28) numerically, we again note that, for numerical implementation, we may remove the LMs contained in $\mathbb{U}$ enforcing the discrete divergence-free and zero-momentum conditions. To do so, let us define spaces $\mathbb{Y}_n, \mathbb{R}_n, \hat{\mathbb{P}}_n$,

$$\mathbb{Y}_n \coloneqq \{\mathbf{u} \in \mathbb{P}_S(T_n;\mathbb{V}) : \mathbf{u}(t_n) \text{ satisfies known initial data in } \mathbb{U}\}, \tag{3.31a}$$

$$\mathbb{S}_n \coloneqq \mathbb{P}_{S-1}(T_n;\mathbb{Q}), \tag{3.31b}$$

$$\hat{\mathbb{P}}_n \coloneqq \mathbb{P}_{S-1}(T_n)^3. \tag{3.31c}$$

With $\mathbb{U}$ defined as in (3.7a), the scheme may be equivalently stated as follows:

find $(\mathbf{u}, (\tilde{\mathbf{u}}, \boldsymbol{\omega}), (p, \theta), (\boldsymbol{\lambda}, \boldsymbol{\alpha})) \in \mathbb{Y}_n \times \dot{\mathbb{Y}}_n^2 \times \mathbb{S}_n^2 \times \hat{\mathbb{P}}_n^2$ such that

$$\mathcal{I}_n[(\dot{\mathbf{u}}, \mathbf{v})] = \mathcal{I}_n\left[(\tilde{\mathbf{u}} \times \boldsymbol{\omega}, \mathbf{v}) - \frac{1}{\mathrm{Re}}(\nabla\tilde{\mathbf{u}}, \nabla\mathbf{v})\right] + \int_{T_n}(p, \operatorname{div}\mathbf{v}) - \boldsymbol{\lambda}\cdot\int_\Omega \mathbf{v}, \tag{3.32a}$$

$$\mathcal{I}_n[(\tilde{\mathbf{u}}, \tilde{\mathbf{v}})] = \int_{T_n}(\mathbf{u}, \tilde{\mathbf{v}}), \tag{3.32b}$$

$$\mathcal{I}_n[(\boldsymbol{\omega}, \boldsymbol{\chi})] = \int_{T_n}(\operatorname{curl}\mathbf{u}, \boldsymbol{\chi}) + \int_{T_n}(\theta, \operatorname{div}\boldsymbol{\chi}) - \boldsymbol{\alpha}\cdot\int_\Omega \boldsymbol{\chi}, \tag{3.32c}$$

$$0 = \int_{T_n}(\operatorname{div}\mathbf{u}, q), \tag{3.32d}$$

$$0 = \int_{T_n}(\operatorname{div}\boldsymbol{\omega}, \eta), \tag{3.32e}$$

$$0 = \int_{T_n}\boldsymbol{\mu}\cdot\int_\Omega \mathbf{u}, \tag{3.32f}$$

$$0 = \int_{T_n}\boldsymbol{\beta}\cdot\int_\Omega \boldsymbol{\omega}, \tag{3.32g}$$

for all $(\mathbf{v}, (\tilde{\mathbf{v}}, \boldsymbol{\chi}), (q, \eta), (\boldsymbol{\mu}, \boldsymbol{\beta})) \in \dot{\mathbb{Y}}_n \times \dot{\mathbb{Y}}_n^2 \times \mathbb{S}_n^2 \times \hat{\mathbb{P}}_n^2$. With $\mathbb{U}$ defined as in (3.7b), this is similar, with each of the 4 $(p, \operatorname{div}\mathbf{v})$-like operators instead taking the form $-(\nabla p, \mathbf{v})$.

3.2 Vortex test

To demonstrate the SP properties of the SP NS scheme (3.28) results, we consider a stationary Hill spherical vortex [Hil94] with swirling motion [Mof69, Sec. 6(b)]. In spherical coordinates (r, θ, φ), define the Stokes stream function

$$\psi(r, \theta, \phi) := \begin{cases} 2\left[\dfrac{J_{\frac{3}{2}}(4\eta r)}{(4r)^{\frac{3}{2}}} - J_{\frac{3}{2}}(\eta)\right](r\sin\theta)^2, & r \le \dfrac{1}{4} \\ 0, & r > \dfrac{1}{4} \end{cases}, \tag{3.33}$$

where J_α denotes the Bessel function of the first kind of order α, and η the first root of $J_{\frac{5}{2}}$, around 5.76. Up to projection onto $\mathbb{U}$, the ICs $\mathbf{u}(0)$ are given by ψ as

$$\mathbf{u}(0) = \frac{\partial_\theta\psi}{r^2\sin\theta}\hat{\mathbf{r}} - \frac{\partial_r\psi}{r\sin\theta}\hat{\boldsymbol{\theta}} + \frac{4\eta\psi}{r\sin\theta}\hat{\boldsymbol{\varphi}} \tag{3.34}$$

where $(\hat{\mathbf{r}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\varphi}})$ are the corresponding spherical unit vectors, and ∂_θ, ∂_r denote partial derivatives with respect to θ, r respectively. This defines the first stationary Hill spherical vortex of radius $\frac{1}{4}$; we consider the domain $\Omega = (-0.5, 0.5)^3$.

In space, we use tetrahedral cells of uniform diameter 2^{-3}; we take $\mathbb{U}$ and $\mathbb{P}$ to form the lowest order Taylor–Hood FE pair (see Ern & Guermond [EG21b, Sec. 54.3]), i.e. continuous Galerkin (CG, a.k.a. Lagrange) elements of degrees 2 and 1 respectively (similarly see Ern & Guermond [EG21a, Sec. 6 & 7]). In time, we take $S = 3$ with $\mathcal{I}_n$ the exact integral, a uniform timestep $\Delta t = 2^{-10}$, and duration $3 \cdot 2^{-6}$; for comparison, we run simulations using the full SP scheme (3.28) alongside one that preserves the structure in the energy only using $\tilde{F}$ as defined in (3.23). We vary $\mathrm{Re} \in 2^{2s}$ over $s \in \{0, \dots, 8\}$.

Fig. 3.1 shows the evolution of the energy Q_1 and helicity Q_2 in the two simulations. From the graph on the lower right, we observe that the energy-preserving scheme has an artificial dissipation in the helicity at all Re, due to the lack of preservation of the helicity-dissipation structure. In all other cases, the dissipations in the energy and helicity decrease in magnitude as Re increases; moreover the energy is universally non-increasing. Fig. 3.1 shows a cross-section of the velocity streamlines at the initial and final times with both schemes, at $\mathrm{Re} = 2^{16}$. When compared with the results from the full SP scheme, one can observe that the artificial helicity dissipation in the energy-preserving scheme causes increased unphysical instability in the vortex.

3.3 Analysis: existence & uniqueness

We detail now some results on the existence and uniqueness properties of certain schemes deriving from the framework. In particular, we restrict our attention to AD systems, with a conserved or dissipated quadratic energy, the behaviour of which is preserved to the discrete level. This includes the stable incompressible NS integrator (3.28) which we will again use as a running example. However, the generality of the analysis means it extends also to the schemes proposed in Sections 10.4 & 10.5; we shall revisit the results of this section in Subsections 10.4.1 & 10.5.1 to show how they can be used to give existence and uniqueness results for the discretisations proposed therein.

Our analysis makes use of the assumption that $\mathbb{U}$ is finite-dimensional, whereby all norms and forms of continuity on $\mathbb{U}$ are equivalent. We make this decision

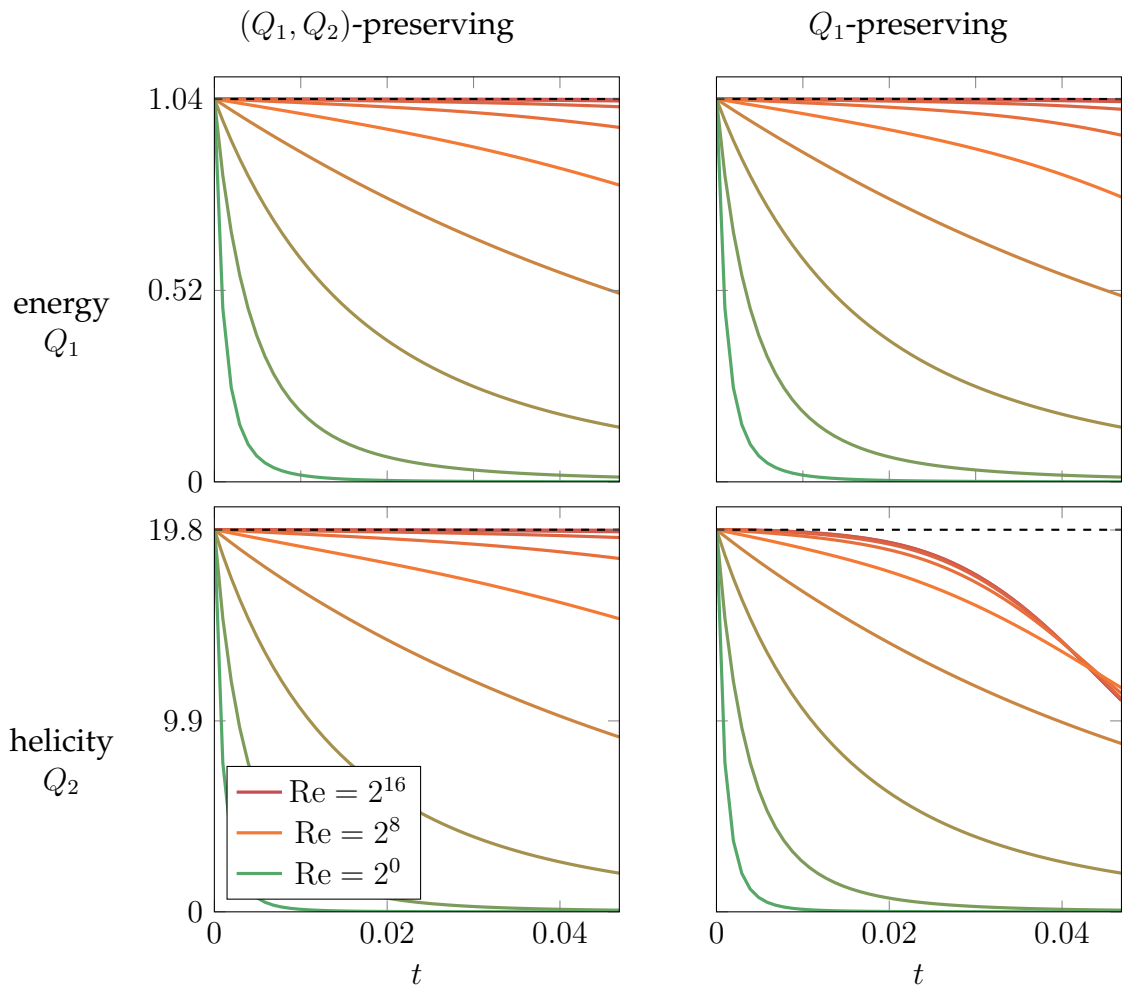
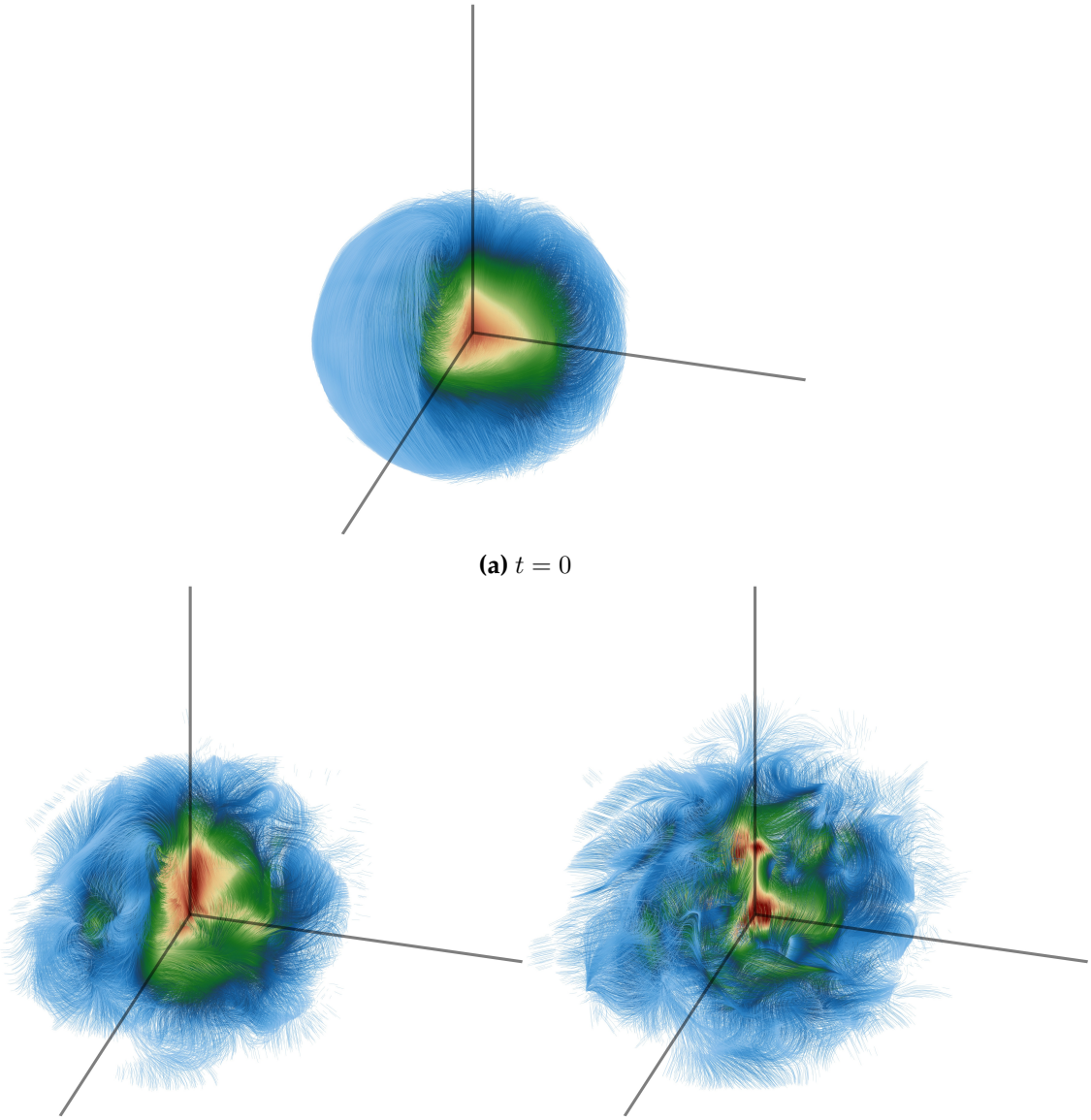


Figure 3.1: Evolution of the energy Q_1 and helicity Q_2 in the (Q_1, Q_2)-preserving scheme (3.28) and the Q_1-preserving scheme derived from (3.23), with varying $\mathrm{Re} = 2^{2s}$ for $s \in \{0, \dots, 8\}$.

as the generality of our framework renders a general infinite-dimensional analysis unsuitable. Under a careful handling of the different norms on $\mathbb{U}$ however, certain aspects of the analysis presented in this section may well be extended to certain infinite-dimensional settings. We specifically make use of theorems that, under stricter regularity conditions, extend to infinite dimensions, in particular the Schauder fixed-point theorem and the contraction mapping theorem (CMT); this is done with the intention that, when investigating a specific problem setting for which the continuous analysis is of importance, certain aspects and results from the discrete analysis could be modified with relative ease.

Again due to our sole consideration of finite-dimensional $\mathbb{U}$, we omit from this section any discussion of convergence. Convergence under refinement of the



(a) $t = 0$

(b) (Q_1, Q_2)-preserving, $t = 3 \cdot 2^{-6}$

(c) Q_1-preserving, $t = 3 \cdot 2^{-6}$

Figure 3.2: Cross-sections of streamlines of the velocity $\mathbf{u}$ for the Hill vortex at times $t \in \{0, 3 \cdot 2^{-6}\}$ in the (Q_1, Q_2)-preserving scheme (3.28) and the Q_1-preserving scheme derived from (3.23) with $\mathrm{Re} = 2^{16}$. Colouring indicates $\|\mathbf{u}\|$.

time discretisation is discussed in Subsection 6.1.1 in the case of ODE systems, for which, with no spatial discretisation, this ceases to be an issue. The same analysis presented therein could well be employed here to show convergence to a solution of an associated semidiscrete problem, discretised in space only; this is of limited use however, as it fails to guarantee convergence under simultaneous refinement of the discretisations in both space and time.

Remark 3.6 (Non-Newton linearisations intended for analysis only)**.** *In the analyses*

to come, we introduce two different linearisations (Definitions 3.19 & 3.14) of the general SP discretisation (3.27). We note here that these are intended as analytic tools only, in particular for their ability to inherit certain energy estimates; we do not necessarily suggest their use as nonlinear solvers as, when convergent, Newton linearisations generally converge more quickly (see Fig. 3.3). In fact, only one of these linearisations, the Picard linearisation (Definition 3.19), we know to converge at all, and only under certain conditions (see Subsection 3.3.4 and Lemma 3.24)

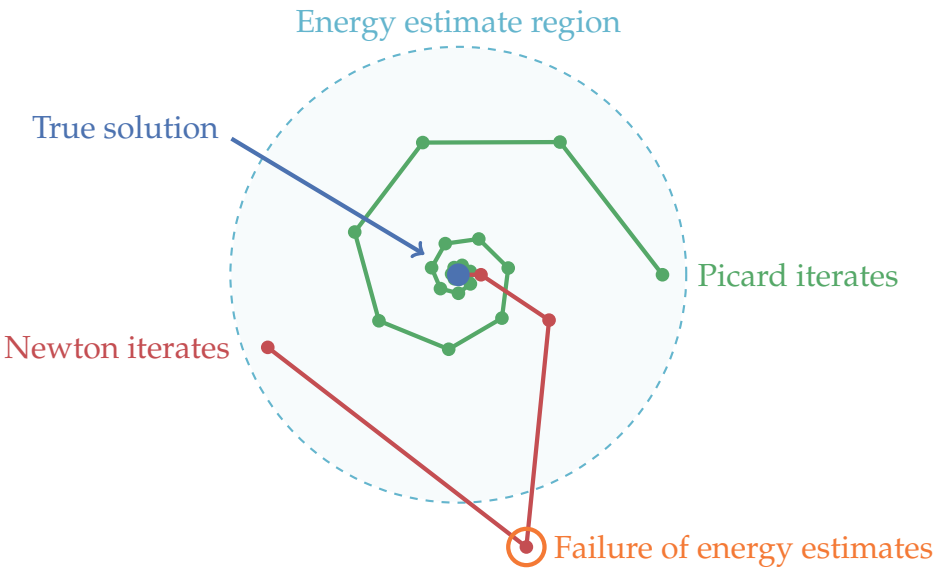


Figure 3.3: Visualisation of the trade-offs in iterate behaviour between a (hypothetical) Newton and Picard linearisation (Definition 3.19) for a (hypothetical) nonlinear problem. Note the Newton iterates temporarily fail the energy estimates, albeit for a faster convergence rate than the Picard iterates in the tail; this typically poses little problem in practice, but a large problem for the analysis.

The rest of this section proceeds as follows. In Subsection 3.3.1, we introduce some preliminary results and notation that will be of general use. In Subsection 3.3.2, we define a general AD system alongside certain technical results. In Subsection 3.3.3, we show that, under relatively loose conditions, solutions to an SP discretisation of such a system exist on arbitrary timesteps Δt_n (Theorem 3.18). In Subsection 3.3.4, we show that, under slightly stricter conditions, these solutions exist uniquely (Theorem 3.26).

Example (Incompressible NS)

Again, for didactic purposes we will be using as a running example the SP discretisation (3.28) of the NS equations.

To simplify the application of the framework to the incompressible NS equations, we assumed no forcing term in our strong form of the equations (3.1). We may retroactively consider a forcing term $\mathbf{f} : \Omega \to \mathbb{R}^3$ by rewriting (3.1a) as

$$\dot{\mathbf{u}} = \mathbf{u} \times \operatorname{curl} \mathbf{u} - \nabla p + \frac{1}{\mathrm{Re}} \Delta \mathbf{u} + \mathbf{f}. \tag{3.35}$$

The application of the framework is then effectively identical, resulting in the slightly modified form of (3.28),

$$\mathcal{I}_n[(\dot{\mathbf{u}}, \mathbf{v})] = \mathcal{I}_n\Big[(\tilde{\mathbf{u}} \times \tilde{\boldsymbol{\omega}}, \mathbf{v}) - \frac{1}{\mathrm{Re}}(\nabla \tilde{\mathbf{u}}, \nabla \mathbf{v}) + (\mathbf{f}, \mathbf{v})\Big], \tag{3.36a}$$

$$\mathcal{I}_n[(\nabla \mathbf{u}, \nabla \mathbf{v})] = \int_{T_n} (\mathbf{u}, \tilde{\mathbf{v}}), \tag{3.36b}$$

$$\mathcal{I}_n[(\tilde{\boldsymbol{\omega}}, \boldsymbol{\chi})] = \int_{T_n} (\operatorname{curl} \mathbf{u}, \boldsymbol{\chi}). \tag{3.36c}$$

For completeness in the analysis, we shall henceforth make this modification, considering instead the more general discretisation (3.36), reducing to (3.28) when $\mathbf{f} = \mathbf{0}$.

3.3.1 Notation & preliminaries

As $\mathbb{U}$ is finite-dimensional, all norms on $\mathbb{U}$ are equivalent; the notation $\|\cdot\|_*$ therefore refers to any fixed norm on $\mathbb{U}$, with a function evaluated on $\mathbb{U}$ said to be continuous if it is continuous with respect to any and all of these norms. We further write $a \lesssim b$ if there exists a constant $C > 0$ dependent only on S and $\mathbb{U}$, in particular independent of Δt_n, such that $a \leq Cb$.

We gather some technical notes and results that will be used in later proofs. Note the following lemmas on $\mathcal{I}_n$, showing respectively that it induces a certain inner product (Lemma 3.7) and that certain norms over T_n are either continuously dependent or, stronger still, equivalent (Lemma 3.8).

Lemma 3.7 ($\mathcal{I}_n$ defines an inner product)**.** *The map $(\phi, \varphi) \mapsto \mathcal{I}_n(\phi\varphi)$ defines an inner product on $\mathbb{P}_{S-1}(T_n)$.*

Proof. Symmetry and linearity are immediate. For positive-definiteness, recall that the map $\phi \mapsto \mathcal{I}_n[\phi^2]^{\frac{1}{2}}$ defines a norm on $\mathbb{P}_{S-1}(T_n)$. □

Lemma 3.8 (Bounds on norms in time)**.** *For all bounded $v : T_n \to \mathbb{R}$,*

$$\mathcal{I}_n[v] \leq \Delta t_n \sup_{T_n} |v|. \tag{3.37a}$$

For all $v \in \dot{\mathbb{X}}_n$,

$$\Delta t_n \sup_{T_n} \|v\|_*^2 \lesssim \int_{T_n} \|v\|_*^2 \lesssim \mathcal{I}_n[\|v\|_*^2]. \tag{3.37b}$$

Proof. For the former inequality (3.37a),

$$\mathcal{I}_n[v] \leq \mathcal{I}_n[\sup_{T_n} |v|] = \Delta t_n \sup_{T_n} |v|, \tag{3.38}$$

with the inequality holding by (3.12a) and the equality holding by (3.12b).

For the latter inequalities (3.37b) recall Lemma 3.7. Supposing $\|\cdot\|_*$ were a norm induced by an inner product, these results could be found by taking expansions of v in polynomial bases of $\mathbb{P}_{S-1}(T_n)$, constructed to be orthonormal under the inner products of $L^2(T_n)$ and $(\phi, \varphi) \to \mathcal{I}_n[\phi\varphi]$ respectively; off-diagonal terms in the expansion are bounded via Young's inequality. The result then extends trivially to norms $\|\cdot\|_*$ not induced by an inner product by the equivalence of norms on $\mathbb{U}$. □

We lastly introduce an assumption on M (Assumption 3.9) under which we show $\mathcal{I}_n[M]$ defines an inner product (Lemma 3.10).

Assumption 3.9 (M defines an inner product)**.** *$M(u; \dot{u}, v)$ is independent of u; as such, we write $M(\cdot, \cdot) = M(u; \cdot, \cdot)$. Moreover, $M(\cdot, \cdot)$ defines an inner product on $\mathbb{U}$.*

Example (Incompressible NS)

In the case of the incompressible NS discretisation (3.28), $M(\cdot, \cdot)$ is simply the $\mathbf{L}^2$ inner product.

Lemma 3.10 ($\mathcal{I}_n[M]$ defines an inner product)**.** *Under Assumption 3.9, $\mathcal{I}_n[M(\cdot, \cdot)]$ defines an inner product on $\dot{\mathbb{X}}_n$.*

Proof. Symmetry and linearity are immediate. Positive-definiteness holds as a consequence of Assumption 3.9 and (3.37b). □

3.3.2 Advection–diffusion systems

We define now the general AD diffusion system (Assumption 3.11) alongside certain technical lemmas (Lemmas 3.12 & 3.13) that will be of use in the existence (Subsection 3.3.3) and uniqueness (Subsection 3.3.4) analyses to come.

Definition

We use the following assumption to define a general AD system.

Assumption 3.11 (AD conditions)**.** *Assume Assumption 3.9, that M defines an inner product on $\mathbb{U}$. Assume further that the AVs $(\tilde{w}_p)$ may be partitioned into type-A $(\tilde{w}_p^{(A)})$ and a single type-B $\tilde{w}^{(B)}$ such that $\tilde{F}$ is affine in $\tilde{w}^{(B)}$. These AVs are associated with QoIs type-A $(Q_p^{(A)})$ and type-B $Q^{(B)} \geq 0$; we require that $Q^{(B)^{\frac{1}{2}}}$ defines a norm on $\mathbb{U}$. Assume lastly that $\tilde{F}$ can be decomposed as*

$$\tilde{F}(u,(\tilde{w}_p);v) = C\Big(u,(\tilde{w}_p^{(A)});\tilde{w}^{(B)},v\Big) - \kappa D\Big(\tilde{w}^{(B)},v\Big) + G\Big(u,(\tilde{w}_p^{(A)});v\Big), \tag{3.39}$$

where $\kappa \geq 0$ is a diffusivity constant (independent of space $\mathbf{x}$ and time t) and:

- *the advective term $C : \mathbb{U} \times \mathbb{U}^{P-1} \times \mathbb{U} \times \mathbb{U} \to \mathbb{R}$ is linear is its final two arguments, and skew-symmetric with $C(u,(\tilde{w}_p^{(A)});\tilde{w}^{(B)},\tilde{w}^{(B)}) = 0$;*
- *the diffusive term $D : \mathbb{U} \times \mathbb{U} \to \mathbb{R}$ is a bilinear form;*
- *the forcing term $G : \mathbb{U} \times \mathbb{U}^{P-1} \times \mathbb{U} \to \mathbb{R}$ is linear and uniformly continuous in its final argument, i.e. such that the bound*

$$\Big|G\Big(u,(\tilde{w}_p^{(A)});v\Big)\Big| \lesssim \|v\|_* \tag{3.40}$$

holds uniformly for all $u,(\tilde{w}_p^{(A)})$ in $\mathbb{U}$.

Here, type-B abbreviates "bounding"; this quantity will be used to derive energy estimates for the analysis. Type-A abbreviates "additional"; these quantities represent those additional structures preserved by our discretisation. We refer to the case $G = 0$ as force-free. We refer to the case $\kappa > 0$ as diffusive, and the case $\kappa = 0$ as conservative.

Example (Incompressible NS)

For the NS scheme (3.36) the AVs $(\tilde{\mathbf{u}}, \boldsymbol{\omega})$ may be partitioned into the type-B auxiliary velocity $\tilde{\mathbf{u}}$ and type-A auxiliary vorticity $\boldsymbol{\omega}$. Accordingly, the QoIs (Q_1, Q_2) may be partitioned into the type-B energy Q_1 and type-A helicity Q_2; we observe then that $Q_1^{\frac{1}{2}}$ defines the $\mathbf{L}^2$ norm on $\mathbb{U}$.

We may then decompose $\tilde{F}$, with $\kappa = \mathrm{Re}^{-1}$ and C, D, G defined

$$C(\mathbf{u}, \tilde{\boldsymbol{\omega}}; \tilde{\mathbf{u}}, \mathbf{v}) \coloneqq (\tilde{\mathbf{u}} \times \tilde{\boldsymbol{\omega}}, \mathbf{v}), \quad D(\tilde{\mathbf{u}}, \mathbf{v}) \coloneqq (\nabla \tilde{\mathbf{u}}, \nabla \mathbf{v}), \quad G(\mathbf{u}, \boldsymbol{\omega}; \mathbf{v}) \coloneqq (\mathbf{f}, \mathbf{v}). \tag{3.41}$$

The zero-mean constraint $\int_\Omega \mathbf{v} = 0$ imposed on $\mathbb{U}$ (3.7) implies D is an inner product by the Poincaré [Poi90] inequality (see Evans [Eva10, Sec. 5.8]); it was for this precise reason we imposed this restriction on $\mathbb{U}$. As an $\mathbf{L}^2$ inner product against a constant field, G is uniformly continuous in $\mathbf{v}$.

Energy estimates & interpolation of the primal variable

The general AD system (Assumption 3.11) exhibits various structures amenable to analysis. In particular, the diffusivity (at least in the force-free case $G = 0$) in the type-B QoI $Q^{(B)}$ provides a bound for the type-B AV $\tilde{w}^{(B)}$ which we shall make use of in the proofs of both existence (Subsection 3.3.3) and uniqueness (Subsection 3.3.4).

Since energy estimates typically offer bounds on $\tilde{w}^{(B)}$, we show there exists an affine transformation from $\tilde{w}^{(B)}$ to u. To understand this transformation, we define an interpolant $u^* \in \dot{\mathbb{X}}_n$ of $u \in \mathbb{X}_n$ at the GL points within T_n. Let π_S denote the degree-S Legendre polynomial shifted to the interval T_n (see Fig. 3.4) such that the roots of π_S are the GL points with T_n. We may then eliminate the highest-order-in-time component of u as

$$u = u^* + (-1)^S [u(t_n) - u^*(t_n)] \pi_S. \tag{3.42}$$

Lemma 3.12 (Bijection between primal variable and type-B AV)**.** *Assume Assumption 3.11, i.e. we are considering an AD system. There exists a linear bijection between $\tilde{w}^{(B)}$ and u^*, and consequently an affine transformation from $\tilde{w}^{(B)}$ to u via (3.42).*

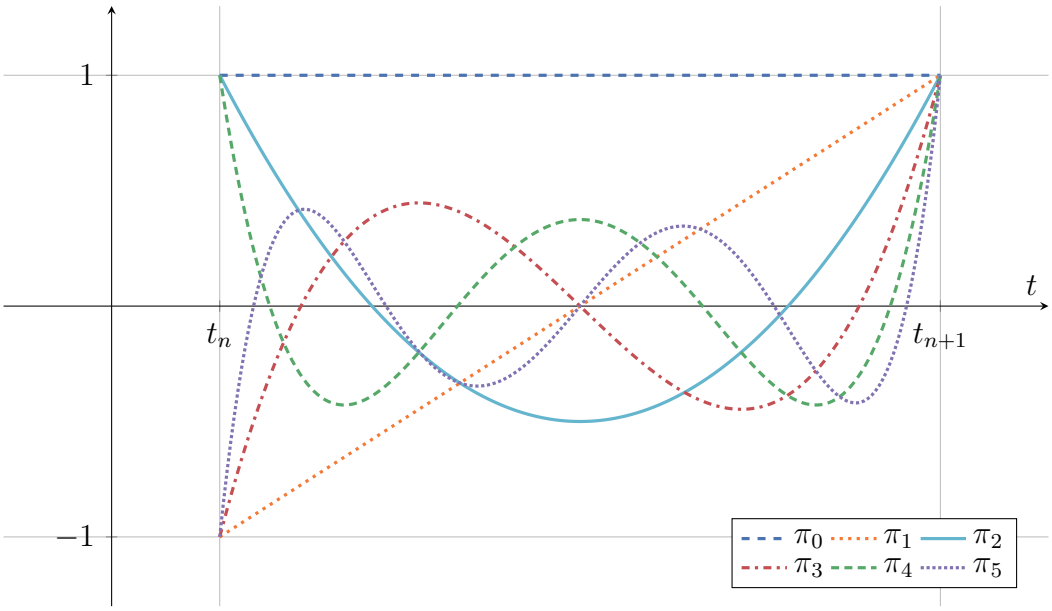


Figure 3.4: The first six Legendre polynomials $\pi_0, \cdots, \pi_5$ shifted to the interval T_n. Note that $\pi_S(t_n) = (-1)^S$.

Proof. The AV $\tilde{w}^{(B)} \in \dot{\mathbb{X}}_n$ is defined to satisfy

$$\mathcal{I}_n\big[M(v^{(B)}, \tilde{w}^{(B)})\big] = \int_{T_n} Q^{(B)\prime}(u; v^{(B)}) \tag{3.43}$$

for all $v^{(B)} \in \dot{\mathbb{X}}_n$. By the orthogonality of the Legendre polynomials, we may rewrite this RHS in the form

$$\mathcal{I}_n\big[M(v^{(B)}, \tilde{w}^{(B)})\big] = \int_{T_n} Q^{(B)\prime}(u^*; v^{(B)}). \tag{3.44}$$

Since $\mathcal{I}_n[M(\cdot,\cdot)]$ and $\int_{T_n} Q^{(B)\prime}(\cdot;\cdot)$ define inner products on $\mathbb{U}$ (by Lemma 3.10 in the case of the former, and by assumption on $Q^{(B)}$ for the latter) the result holds immediately by the Riesz representation theorem. □

Example (Incompressible NS)

For the NS scheme (3.36) $\mathbf{u}^* \in \dot{\mathbb{X}}_n$ is defined to satisfy

$$\mathbf{u} = \mathbf{u}^* + (-1)^S[\mathbf{u}(t_n) - \mathbf{u}^*(t_n)]\pi_S. \tag{3.45}$$

The relation

$$\mathcal{I}_n[(\tilde{\mathbf{v}}, \tilde{\mathbf{u}})] = \int_{T_n} (\mathbf{u}, \tilde{\mathbf{v}}) = \int_{T_n} (\mathbf{u}^*, \tilde{\mathbf{v}}) \tag{3.46}$$

for all $\tilde{\mathbf{v}} \in \dot{\mathbb{X}}_n$ defines a bijection between $\tilde{\mathbf{u}}$ and $\mathbf{u}^*$, and an affine transformation from $\tilde{\mathbf{u}}$ to $\mathbf{u}$.

This construction of u^* is chosen specifically such that the following coercivity result holds.

Lemma 3.13 (Coercivity of AD bilinear form). *Define a bilinear form $A : \dot{\mathbb{X}}_n^2 \to \mathbb{R}$ by*

$$A(\tilde{w}^{(B)}, v) \coloneqq \mathcal{I}_n\Big[M(\dot{u}^* - (-1)^S u^*(t_n)\dot{\pi}_S, v) + \kappa D(\tilde{w}^{(B)}, v)\Big], \tag{3.47}$$

where u^* is a linear transformation of $\tilde{w}^{(B)}$ as in Lemma 3.12 (3.44). Then A is coercive when either $\kappa > 0$ or $S = 1$.

Proof. We first note that, by taking $v^{(B)} = \dot{u}^* - (-1)^S u^*(t_n)\dot{\pi}_S$ in (3.44),

$$\begin{aligned}
&\mathcal{I}_n\Big[M(\dot{u}^* - (-1)^S u^*(t_n)\dot{\pi}_S, \tilde{w}^{(B)})\Big] \\
&\qquad = \int_{T_n} Q^{(B)'}(u^*; \dot{u}^* - (-1)^S u^*(t_n)\dot{\pi}_S) && \text{(3.48a)} \\
&\qquad = \int_{T_n} Q^{(B)'}(u^* - (-1)^S u^*(t_n)\pi_S; \dot{u}^* - (-1)^S u^*(t_n)\dot{\pi}_S) && \text{(3.48b)} \\
&\qquad = \begin{cases} 4Q^{(B)}(u^*(t_{n+1})), & S \text{ odd}, \\ 0, & S \text{ even}, \end{cases} && \text{(3.48c)}
\end{aligned}$$

where in the second equality we use the orthogonality of the Legendre polynomials. Since this is non-negative, the coercivity of A holds immediately when $\kappa > 0$, through the latter, necessarily coercive term $\kappa\mathcal{I}_n[D(\tilde{w}^{(B)}, v)]$.

Otherwise, in the case $S = 1$, we show coercivity by considering the former term $\mathcal{I}_n[M(\dot{u}^* - (-1)^S u^*(t_n)\dot{\pi}_S, v)]$. The above result (3.48) bounds $\mathcal{I}_n[M(\dot{u}^* - (-1)^S u^*(t_n)\dot{\pi}_S, \tilde{w}^{(B)})]$ below by $\|u^*(t_{n+1})\|_*^2$. This is then trivially bounded below by $\sup_{T_n} \|u^*\|_*^2$ as u^* is constant in time when $S = 1$. We may lastly bound this below by $\sup_{T_n} \|\tilde{w}^{(B)}\|_*^2$ through the bijection between u^* and $\tilde{w}^{(B)}$ in Lemma 3.12. Thus, since $\kappa\mathcal{I}_n[D(\tilde{w}^{(B)}, \tilde{w}^{(B)})] \geq 0$, A is coercive when $S = 1$. □

Note, in finite dimensions, A is also trivially continuous by linearity.

Example (Incompressible NS)

For the NS scheme (3.36) $A : \dot{\mathbb{X}}_n^2 \to \mathbb{R}$ is defined

$$A(\tilde{\mathbf{u}}, \mathbf{v}) \coloneqq \mathcal{I}_n\Big[(\mathbf{u}^* - (-1)^S \mathbf{u}^*(t_n), \mathbf{v}) + \frac{1}{\mathrm{Re}}(\nabla\tilde{\mathbf{u}}, \nabla\mathbf{v})\Big], \tag{3.49}$$

for $\mathbf{u}^*$ defined implicitly as a function of $\tilde{\mathbf{u}}$ through (3.46). By Lemma 3.13, this is coercive when either $\mathrm{Re} < \infty$ or $S = 1$.

3.3.3 Existence

We discuss now the existence of solutions to the general SP discretisation (3.27). Our proof strategy relies on the Schauder [Sch30] fixed-point theorem (see Evans [Eva10, Sec. 9.2]). We define a certain Schauder linearisation (Definition 3.14) scaling with a parameter $\gamma \in [0,1]$ from a homogeneous linear problem when $\gamma = 0$ to a full linearisation when $\gamma = 1$, for which fixed points are solutions to the original discretisation (3.27). We apply the Schauder fixed-point theorem to give our final existence result in Theorem 3.18.

Definition of the Schauder linearisation

Definition 3.14 (Schauder linearisation)**.** *Assume Assumption 3.11, i.e. that (3.27) defines an SP discretisation of an AD system. Let $m \in \mathbb{N}$ denote an iteration index, and, on a given timestep T_n, suppose $(u_m, (w_{p,m}^{(A)}), w_m^{(B)}) \in \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n^{P-1} \times \dot{\mathbb{X}}_n$ are given. For $\gamma \in [0,1]$, we define an iteration: find $(u_{m+1}^*, w_{m+1}^{(B)}) \in \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n$ such that*

$$\begin{aligned}\mathcal{I}_n[M(\dot{u}_{m+1}, v)] = \mathcal{I}_n\Big[&\gamma C(u_m, (\tilde{w}_{p,m}^{(A)}); \tilde{w}_m^{(B)}, v) \\ &- \kappa D(\tilde{w}_{m+1}^{(B)}, v) + \gamma G(u_m, (\tilde{w}_{p,m}^{(A)}); v)\Big], \qquad (3.50a)\end{aligned}$$

$$\mathcal{I}_n[M(v^{(B)}, \tilde{w}_{m+1}^{(B)})] = \int_{T_n} (Q^{(B)})'(u_{m+1}; v^{(B)}), \qquad (3.50b)$$

for all $(v, v^{(B)}) \in \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n$, where u_{m+1} is defined in terms of u_{m+1}^* similarly to (3.42) except a weighting by γ,

$$u_{m+1} = u_{m+1}^* + (-1)^S[\gamma u(t_n) - u_{m+1}^*(t_n)]\pi_S; \qquad (3.51)$$

with this, find $(\tilde{w}_{p,m}^{(A)}) \in \dot{\mathbb{X}}_n^{P-1}$ such that

$$\mathcal{I}_n[M(v_p^{(A)}, \tilde{w}_{p,m+1}^{(A)})] = \int_{T_n} (Q_p^{(A)})'(u_{m+1}; v_p^{(A)}) \qquad (3.52)$$

for all $(v_p^{(A)}) \in \dot{\mathbb{X}}_n^{P-1}$.

The Schauder linearisation decouples the updates in the primary variable and type-B AV (3.50) from those in the type-A AVs (3.52).

Example (Incompressible NS)

For the NS discretisation (3.36), the Schauder linearisation takes the form: find $(\mathbf{u}^*_{m+1}, \tilde{\mathbf{u}}_{m+1}) \in \dot{\mathbb{X}}^2_n$ such that

$$\mathcal{I}_n[(\dot{\mathbf{u}}_{m+1}, \mathbf{v})] = \mathcal{I}_n\Big[\gamma(\tilde{\mathbf{u}}_m \times \boldsymbol{\omega}_m, \mathbf{v}) - \frac{1}{\mathrm{Re}}(\nabla\tilde{\mathbf{u}}_{m+1}, \nabla\mathbf{v}) + \gamma(\mathbf{f}, \mathbf{v})\Big], \tag{3.53a}$$

$$\mathcal{I}_n[(\tilde{\mathbf{v}}, \tilde{\mathbf{u}}_{m+1})] = \int_{T_n} (\mathbf{u}_{m+1}, \tilde{\mathbf{v}}), \tag{3.53b}$$

for all $(\mathbf{v}, \tilde{\mathbf{v}}) \in \dot{\mathbb{X}}^2_n$, where $\mathbf{u}_{m+1}$ is defined

$$\mathbf{u}_{m+1} = \mathbf{u}^*_{m+1} + (-1)^S[\gamma\mathbf{u}(t_n) - \mathbf{u}^*_{m+1}(t_n)]\pi_S; \tag{3.54}$$

with this, find $\boldsymbol{\omega}_{m+1} \in \dot{\mathbb{X}}_n$ such that

$$\mathcal{I}_n[(\boldsymbol{\chi}, \boldsymbol{\omega}_{m+1})] = \gamma \int_{T_n} (\operatorname{curl} \mathbf{u}_{m+1}, \boldsymbol{\chi}) \tag{3.55}$$

for all $\boldsymbol{\chi} \in \dot{\mathbb{X}}_n$.

Properties of the Schauder linearisation

We discuss now certain properties held by the Schauder linearisation (3.50, 3.52) in particular its well-posedness and the boundedness of fixed points. Each of these results holds only in either the diffusive case ($\kappa > 0$) or the lowest-order-in-time case ($S = 1$).

Lemma 3.15 (Well-posedness of Schauder linearisation)**.** *The Schauder linearisation (3.50, 3.52) is well-posed with a unique solution for all $\gamma \in [0, 1]$ when either $\kappa > 0$ or $S = 1$.*

Proof. We begin by noting the maps $\tilde{w}^{(B)}_{m+1} \mapsto u^*_{m+1}$ and $\tilde{w}^{(B)}_{m+1} \mapsto u_{m+1}$ (3.50b) are well-defined through a minimal modification of Lemma 3.12. We see then the map $u_{m+1} \mapsto (\tilde{w}^{(A)}_{p,m+1})$ (3.52) is well-defined, since the LHS defines an inner product on $\dot{\mathbb{X}}_n$ by Lemma 3.10. It suffices then to show that $\tilde{w}^{(B)}_{m+1}$ is uniquely determined by (3.50a), i.e. the map $(u_m, (\tilde{w}^{(A)}_{p,m}), \tilde{w}^{(B)}_{p,m}) \mapsto \tilde{w}^{(B)}_{m+1}$ is well-defined.

Let us interpret u^*_{m+1} as an implicit (linear) function of $\tilde{w}^{(B)}_{m+1}$. Define $A : \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n \to \mathbb{R}$ as in (3.47); under the assumptions $\kappa > 0$ or $S = 1$, A is coercive by Lemma 3.13.

Define $B : \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n^{P-1} \times \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n \to \mathbb{R}$, linear in its final argument,

$$B(u_m, (\tilde{w}_{p,m}^{(A)}), \tilde{w}_m^{(B)}; v) = \mathcal{I}_n\Big[G(u_m, (\tilde{w}_{p,m}^{(A)}); v) - (-1)^S M(u(t_n)\dot{\pi}_S, v)\Big]. \tag{3.56}$$

The primal step of the Schauder linearisation (3.50a) is then equivalent to solving the problem: find $\tilde{w}_{m+1}^{(B)} \in \dot{\mathbb{X}}_n$ such that

$$A(\tilde{w}_{m+1}^{(B)}, v) = \gamma\Big\{\mathcal{I}_n[C(u_m, (\tilde{w}_{p,m}^{(A)}); \tilde{w}_m^{(B)}, v)] + B(u_m, (\tilde{w}_{p,m}^{(A)}), \tilde{w}_m^{(B)}; v)\Big\} \tag{3.57}$$

for all $v \in \dot{\mathbb{X}}_n$. This is well-defined by the Lax–Milgram theorem (see Evans [Eva10, Sec. 6.2]). □

The crucial aspect of this result for analytic purposes is that, with u_m and $(\tilde{w}_{p,m}^{(A)})$ defined as implicit functions of $\tilde{w}_m^{(B)}$, the Schauder linearisation may be equivalently interpreted as an iteration on the type-B AVs $\tilde{w}_m^{(B)}$, with the convenient abstract form (3.57).

Example (Incompressible NS)

The incompressible NS Schauder iteration (3.53, 3.55) is well-posed with a unique solution when either $\mathrm{Re} < \infty$ or $S = 1$.

We now show the boundedness of fixed points of the linearisation, uniform in $\gamma \in [0, 1]$.

Lemma 3.16 (Bounded fixed points of Schauder linearisation)**.** *Fixed points $(u, (\tilde{w}_p^{(A)}), \tilde{w}^{(B)})$ of the Schauder linearisation (3.50, 3.52) are bounded in $\tilde{w}^{(B)}$, uniformly over $\gamma \in [0, 1]$, when either $\kappa > 0$ or $S = 1$.*

Proof. Considering $v = \tilde{w}^{(B)}$ in (3.57), by the skew-symmetry of C,

$$A(\tilde{w}^{(B)}, \tilde{w}^{(B)}) = \gamma B(u, (\tilde{w}_p^{(A)}), \tilde{w}^{(B)}; \tilde{w}^{(B)}). \tag{3.58}$$

With the coercivity of the LHS $A(\tilde{w}^{(B)}, \tilde{w}^{(B)})$, it is sufficient to show the RHS is

uniformly bounded by any norm on $\tilde{w}^{(B)}$;

$$
\begin{aligned}
&|\gamma B(u,(\tilde{w}_p^{(A)}),\tilde{w}^{(B)};\tilde{w}^{(B)})| \\
&\qquad \le |B(u,(\tilde{w}_p^{(A)}),\tilde{w}^{(B)};\tilde{w}^{(B)})| && \text{(3.59a)}\\
&\qquad \le \left|\mathcal{I}_n\left[G(u,(\tilde{w}_p^{(A)});\tilde{w}^{(B)}) - (-1)^S M(u(t_n)\dot{\pi}_S,\tilde{w}^{(B)})\right]\right| && \text{(3.59b)}\\
&\qquad \le \Delta t_n \left\{ \sup_{T_n} \left|G(u,(\tilde{w}_p^{(A)});\tilde{w}^{(B)})\right| + \sup_{T_n}\left|M(u(t_n)\dot{\pi}_S,\tilde{w}^{(B)})\right|\right\} && \text{(3.59c)}\\
&\qquad \lesssim (\Delta t_n + 1)\sup_{T_n}\|\tilde{w}^{(B)}\|_*, && \text{(3.59d)}
\end{aligned}
$$

where the second inequality holds by the definition of B (3.56), the third by the bound (3.37a) and triangle inequality, and the last holds by the uniform boundedness of G (3.40). □

Example (Incompressible NS)

Fixed points $(\mathbf{u},\tilde{\mathbf{u}},\boldsymbol{\omega})$ of the NS Schauder linearisation (3.53, 3.55) are bounded in $\tilde{\mathbf{u}}$, uniformly over $\gamma\in[0,1]$, when either $\mathrm{Re}<\infty$ or $S=1$.

Application of Schauder's fixed-point theorem

With the Schauder linearisation established and analysed, we may now proceed to apply Schauder's fixed-point theorem (see Evans [Eva10, Sec. 9.2]) which may be stated in a weak formulation as follows.

Lemma 3.17 (Schauder's fixed-point theorem: weak formulation)**.** *Consider a Hilbert space X. Take a continuous, coercive bilinear form $A:\mathbb{X}^2\to\mathbb{R}$; take also a continuous function $T:X^2\to\mathbb{R}$, linear in its second argument, such that the mapping $x\mapsto T(x;\cdot)$ from $X\to X^*$ is compact. For $\gamma\in[0,1]$, define an iteration on X with iteration index $m\in\mathbb{N}$: for given $x_m\in X$, find $x_{m+1}\in X$ such that*

$$A(x_{m+1},y)=\gamma T(x_m;y) \tag{3.60}$$

for all $y\in X$. If all fixed points of (3.60) for all $\gamma\in[0,1]$ are uniformly bounded in X, then there exists a fixed point when $\gamma=1$, i.e. there exists $x\in X$ such that

$$A(x,y)=T(x;y) \tag{3.61}$$

for all $y\in X$.

When X is finite dimensional, continuity holds as a consequence of linearity, and compactness as a consequence of continuity. In such a case therefore, it is sufficient only that A be coercive, and that T be continuous in its first argument and linear in its second.

With this established, we may apply it to the linearisation (3.50, 3.52) to prove the existence of solutions to the original SP discretisation (3.27).

Theorem 3.18 (Existence of solutions: AD systems)**.** *Assume Assumption 3.11, i.e. we are considering an SP discretisation of an AD system, and that the advective C and forcing G terms are continuous in u, $(\tilde{w}_p^{(A)})$, and that the type-A AVs $(Q_p^{(A)})$ are continuously differentiable. There then exist solutions on arbitrary timesteps Δt_n when either $\kappa > 0$ or $S = 1$.*

Proof. We shall view the Schauder linearisation (3.50, 3.52) as an iteration solely over the type-B AV $\tilde{w}_m^{(B)} \in \dot{\mathbb{X}}_n$, with the primal variable u_m^* and type-A AVs $(\tilde{w}_{p,m}^{(A)})$ defined as implicit functions of $\tilde{w}_m^{(B)}$. We apply the Schauder fixed-point theorem (Lemma 3.17) to the abstract form (3.57).

The coercivity of the LHS A holds from Lemma 3.13 when either $\kappa > 0$ or $S = 1$. By the assumptions of continuity of C and G, the RHS is continuous in u_m, $(\tilde{w}_{p,m}^{(A)})$, $\tilde{w}_m^{(B)}$S. The primal variable u_m is continuously dependent on $\tilde{w}_m^{(B)}$ by the affine transformation (Lemma 3.12) while the type-A AVs are then continuously dependent on u_m, and subsequently $\tilde{w}_m^{(B)}$ by composition, by the assumed continuity of $(Q_p^{(A)})$. When interpreted as an implicit function of $\tilde{w}_m^{(B)}$, the RHS of (3.57) is therefore continuous in $\tilde{w}_m^{(B)}$. The uniform boundedness of the type-B AV $\tilde{w}^{(B)}$ within fixed points of the Schauder linearisation (3.57) then holds from Lemma 3.16. Thus, by the Schauder fixed-point theorem (Lemma 3.17) fixed points to the Schauder iteration (3.50, 3.52) exist when γ, i.e. there exist solutions to the SP discretisation (3.27). □

Example (Incompressible NS)

Solutions to the SP incompressible NS discretisation (3.36) exist for arbitrary timesteps Δt_n, when either $\mathrm{Re} < \infty$ or $S = 1$.

3.3.4 Uniqueness

We discuss now the circumstance under which the solutions to the general SP discretisation (3.27) are unique. Our proof strategy relies instead on the CMT (a.k.a. Banach's fixed point theorem, see Evans [Eva10, Sec. 9.2]). We define a different linearisation, a Picard linearisation (Definition 3.19), for which fixed points are solutions to the original discretisation (3.27). We apply the CMT to give our final uniqueness result in Theorem 3.26.

Picard linearisation

Definition 3.19 (Picard linearisation)**.** *Assume Assumption 3.11, i.e. we are considering an SP discretisation of an AD system. Let $m \in \mathbb{N}$ denote an iteration index, and, on a given timestep T_n, suppose $u_m \in \mathbb{X}_n$ is given. We define an iteration: find $(u_{m+1}, \tilde{w}^{(B)}_{m+1}) \in \mathbb{X}_n \times \dot{\mathbb{X}}_n$ such that*

$$\mathcal{I}_n[M(\dot{u}_{m+1}, v)] = \mathcal{I}_n\Big[\tilde{F}(u_m, (\tilde{w}^{(A)}_{p,m}), \tilde{w}^{(B)}_{m+1}; v)\Big], \tag{3.62a}$$

$$\mathcal{I}_n\Big[M(v^{(B)}, \tilde{w}^{(B)}_{m+1})\Big] = \int_{T_n} Q_p^{(B)\prime}(u_{m+1}; v^{(B)}), \tag{3.62b}$$

for all $(v, v^{(B)}) \in \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n$; with this, find $(\tilde{w}^{(A)}_{p,m+1}) \in \dot{\mathbb{X}}_n^{P-1}$ such that

$$\mathcal{I}_n\Big[M(v_p^{(A)}, \tilde{w}^{(A)}_{p,m+1})\Big] = \int_{T_n} Q_p^{(A)\prime}(u_{m+1}; v_p^{(A)}), \tag{3.63}$$

for all $(v_p^{(A)}) \in \dot{\mathbb{X}}_n^{P-1}$.

Similarly to the Schauder linearisation (3.50, 3.52) the Picard linearisation decouples the updates in the primary variable and type-B AV (3.62) from those in the type-A AVs (3.63), however the difference lies in how this is done.

We recall that the Schauder linearisation depended continuously on an additional parameter $\gamma = [0, 1]$. At $\gamma = 1$, it represented a full linear fixed point iteration for the nonlinear problem, simply handling the nonlinear advective and forcing terms explicitly; transferring gradually to $\gamma = 0$, all inhomogeneous terms (i.e. the advective and forcing terms, as well as the ICs) are gradually eliminated, reducing the problem to a trivial homogeneous one. This had the property that energy estimates held for all fixed points of the Schauder linearisation uniformly for all $\gamma \in [0, 1]$ (Lemma 3.16).

In contrast, the Picard linearisation has no dependence on an additional parameter. Instead, it resembles the Schauder linearisation at $\gamma = 1$—fixed points of the Picard linearisation equate to solutions of the full nonlinear discretisation—however the advective term is handled semi-implicitly, i.e. implicitly in $w_{m+1}^{(B)}$. This does not affect the linearity, however we similarly observe below (Lemma 3.21) that this consturction implies a stronger energy estimate, bounding not just fixed points but all iterates.

Example (Incompressible NS)

For the NS discretisation (3.36), the Picard linearisation takes the form: find $(\mathbf{u}_{m+1}, \tilde{\mathbf{u}}_{m+1}) \in \mathbb{X}_n \times \dot{\mathbb{X}}_n$ such that

$$\mathcal{I}_n[(\dot{\mathbf{u}}_{m+1}, \mathbf{v})] = \mathcal{I}_n\Big[(\tilde{\mathbf{u}}_{m+1} \times \tilde{\boldsymbol{\omega}}_m, \mathbf{v}) - \frac{1}{\mathrm{Re}}(\nabla \tilde{\mathbf{u}}_{m+1}, \nabla \mathbf{v})\Big], \tag{3.64a}$$

$$\mathcal{I}_n[(\tilde{\mathbf{v}}, \tilde{\mathbf{u}}_{m+1})] = \int_{T_n} (\tilde{\mathbf{v}}, \mathbf{u}_{m+1}), \tag{3.64b}$$

for all $(\mathbf{v}, \tilde{\mathbf{v}}) \in \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n$; with this, find $\tilde{\boldsymbol{\omega}}_{m+1} \in \dot{\mathbb{X}}_n$ such that

$$\mathcal{I}_n[(\boldsymbol{\chi}, \tilde{\boldsymbol{\omega}}_{m+1})] = \int_{T_n} (\boldsymbol{\chi}, \operatorname{curl} \mathbf{u}_m) \tag{3.65}$$

for all $\boldsymbol{\chi} \in \dot{\mathbb{X}}_n$.

Properties of the Picard linearisation

We discuss now certain properties held by the Picard linearisation (3.62, 3.63) in particular its well-posedness and the boundedness of iterations. Similarly to the existence analysis above, each of these results holds only in either the diffusive case ($\kappa > 0$) or the lowest-order-in-time case ($S = 1$).

Lemma 3.20 (Well-posedness of Picard linearisation)**.** *The Picard linearisation (3.62, 3.63) is well-posed with a unique solution when either $\kappa > 0$ or $S = 1$.*

Proof. The proof strategy here is very similar to that of Lemma 3.15.

We begin by noting the map $\tilde{w}_{m+1}^{(B)} \mapsto u_{m+1}$ (3.62b) is well-defined through a minimal modification of Lemma 3.12. We see then the map $u_{m+1} \mapsto (\tilde{w}_{p,m+1}^{(A)})$ (3.63)

is well-defined, since the LHS defines an inner product on $\dot{\mathbb{X}}_n$ by Lemma 3.10. It suffices then to show that $\tilde{w}_{m+1}^{(B)}$ is uniquely determined by (3.62a).

Let us interpret u_{m+1} as an implicit (affine) function of $\tilde{w}_{m+1}^{(B)}$. Define $A : \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n \to \mathbb{R}$ as in (3.47); under the assumptions $\kappa > 0$ or $S = 1$, A is coercive by Lemma 3.13. Define $B : \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n^{P-1} \times \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n \to \mathbb{R}$ as in (3.56). The primal step of the Picard linearisation (3.50a) is then equivalent to solving the problem: find $\tilde{w}_{m+1}^{(B)} \in \dot{\mathbb{X}}_n$ such that

$$A(\tilde{w}_{m+1}^{(B)}, v) - \mathcal{I}_n\Big[C(u_m, (\tilde{w}_{p,m+1}^{(A)}); \tilde{w}_m^{(B)}, v)\Big] = B(u_m, (\tilde{w}_{p,m}^{(A)}), \tilde{w}_m^{(B)}; v) \tag{3.66}$$

for all $v \in \dot{\mathbb{X}}_n$. This is well-defined by the Lax–Milgram theorem (see Evans [Eva10, Sec. 6.2]) since the skew-symmetry of C implies its introduction does not affect the coercivity of the LHS. □

Example (Incompressible NS)

The NS Picard linearisation (3.64, 3.65) is well-posed with a unique solution when either $\mathrm{Re} < \infty$ or $S = 1$.

We now show the boundedness of Picard iterations. Notably, this bound differs from that of the Schauder linearisation (Lemma 3.16) in that it holds on all iterations, and not just the fixed points.

Lemma 3.21 (Bounded solutions of Picard linearisation)**.** *Iterates* $(u_{m+1}, (\tilde{w}_{p,m+1}^{(A)}), \tilde{w}_{m+1}^{(B)})$ *of the Picard linearisation (3.62, 3.63) are bounded in* $\tilde{w}_{m+1}^{(B)}$*, when either* $\kappa > 0$ *or* $S = 1$*.*

Proof. The proof here is almost identical to that of Lemma 3.16, by taking $v = \tilde{w}_{m+1}^{(B)}$ in (3.66). □

Example (Incompressible NS)

Iterates $(\mathbf{u}_{m+1}, \tilde{\mathbf{u}}_{m+1}, \boldsymbol{\omega}_{m+1})$ of the NS Picard linearisation (3.64, 3.65) are bounded in $\tilde{\mathbf{u}}_{m+1}$ when either $\mathrm{Re} < \infty$ or $S = 1$.

Application of the contraction mapping theorem

With the Picard linearisation established and analysed, we may now proceed to apply the CMT. We begin by defining certain regularity conditions, under which we may assert the Picard linearisation (3.62, 3.63) is a contraction.

Assumption 3.22 (AD regularity)**.** *Assume Assumption 3.11, i.e. we have an SP discretisation of an AD system. Assume further that $C, G \in \mathrm{Lip}_{\mathrm{loc}}$, i.e. they are locally Lipschitz, and each $Q_p^{(A)} \in \mathrm{Lip}^1_{\mathrm{loc}}$, i.e. it is locally Lipschitz differentiable such that for any compact $K \subset \mathbb{U}$,*

$$|C(u^+, (\tilde{w}_p^{(A)+}); \tilde{w}^{(B)}, v) - C(u^-, (\tilde{w}_p^{(A)-}); \tilde{w}^{(B)}, v)| \lesssim \max\{\|u^+ - u^-\|_*, (\|\tilde{w}_p^{(A)+} - \tilde{w}_p^{(A)-}\|_*)\} \cdot \|\tilde{w}^{(B)}\|_* \|v\|_*, \tag{3.67a}$$

$$|G(u^+, (\tilde{w}_p^{(A)+}); v) - G(u^-, (\tilde{w}_p^{(A)-}); v)| \lesssim \max\{\|u^+ - u^-\|_*, (\|\tilde{w}_p^{(A)+} - \tilde{w}_p^{(A)-}\|_*)\} \cdot \|v\|_*, \tag{3.67b}$$

$$|Q_p^{(A)\prime}(u^+; v) - Q_p^{(A)\prime}(u^+; v)| \lesssim \|u^+ - u^-\|_* \cdot \|v\|_* \quad \forall p, \tag{3.67c}$$

for all $u^\pm \in K$, $(\tilde{w}_p^{(A)\pm} \in K)$ and $\tilde{w}^{(B)}, v \in \mathbb{U}$.

Remark 3.23 (Sufficiency of continuous differentiability)**.** *Noting that, for $r \geq 0$, any $(r+1)$-times continuously differentiable function necessarily lies in $\mathrm{Lip}^r_{\mathrm{loc}}$, it is sufficient to show C, G are continuouosly differentiable and each $Q_p^{(A)}$ is twice continuously differentiable.*

Example (Incompressible NS)

In finite dimensions C (3.41) and Q_2 (3.15) are smooth, therefore Assumption 3.22 holds through Remark 3.23.

Under this assumption, we may show that the Picard linearisation is a contraction, given certain conditions on κ, Δt_n, S. The proof of this result is where the majority of the work is required.

Lemma 3.24 (Picard linearisation is a contraction)**.** *Assume Assumption 3.22, i.e. we have a suitably regular AD problem. Then, given either sufficiently large κ, or in the lowest-order-in-time case $S = 1$ sufficiently small Δt_n, the Picard linearisation (3.63, 3.62) is a contraction.*

Proof. Considering $u_{m+1}^{\pm} \in \mathbb{X}_n$, define $u_{m+1}^{*\pm} \in \dot{\mathbb{X}}_n$, $\tilde{w}_{m+1}^{(B)\pm}$ accordingly by (3.42, 3.62b) respectively; denote also $\delta u_{m+1} := u_{m+1}^{+} - u_{m+1}^{-}$, with δu_{m+1}^{*}, $\delta \tilde{w}_{m+1}^{(B)}$ defined similarly. To show the norms $\sup_{T_n} \|\delta u_{m+1}\|_*$ and $\sup_{T_n} \|\delta \tilde{w}_{m+1}^{(B)}\|_*$ are equivalent, it is sufficient to show $\sup_{T_n} \|\delta u_{m+1}\|_*$ and $\sup_{T_n} \|\delta u_{m+1}^{*}\|_*$ are equivalent, since the norms $\sup_{T_n} \|\delta u_{m+1}^{*}\|_*$ and $\sup_{T_n} \|\delta \tilde{w}_{m+1}^{(B)}\|_*$ are equivalent similarly to Lemma 3.12. By (3.42) we see δu_{m+1}, δu_{m+1}^{*} are related by

$$\delta u_{m+1} = \delta u_{m+1}^{*} - (-1)^S \delta u_{m+1}^{*}(t_n)\pi_S; \tag{3.68a}$$

this relation can be inverted, as

$$\delta u_{m+1}^{*} = \delta u_{m+1} - \frac{1}{\int_{T_n} \pi_S^2}\Big(\int_{T_n} \delta u_{m+1}\pi_S\Big)\pi_S. \tag{3.68b}$$

These linear relations together imply the norms $\sup_{T_n} \|u_{m+1}\|_*$ and $\sup_{T_n} \|\delta u_{m+1}^{*}\|_*$ are equivalent. Together therefore we have the equivalence

$$\sup_{T_n} \|\delta u_{m+1}\|_* \lesssim \sup_{T_n} \|\delta \tilde{w}_{m+1}^{(B)}\|_* \lesssim \sup_{T_n} \|\delta u_{m+1}\|_*. \tag{3.69}$$

It is a simple exercise to confirm each of these scalings is independent of Δt_n.

Before proceeding, we consider the relation briefly the relation between $\delta \tilde{w}_{m+1}^{(B)}$ and δu_{m+1},

$$\mathcal{I}_n\Big[M(v^{(B)}, \delta \tilde{w}_{m+1}^{(B)})\Big] = \int_{T_n} Q^{(B)\prime}(\delta u_{m+1}; v^{(B)}). \tag{3.70}$$

Taking $v^{(B)} = \dot{\delta u}_{m+1}$, we find

$$\mathcal{I}_n\Big[M(\dot{\delta u}_{m+1}, \delta \tilde{w}_{m+1}^{(B)})\Big] = \int_{T_n} Q^{(B)\prime}(\delta u_{m+1}; \dot{\delta u}_{m+1}) = Q^{(B)}(\delta u_{m+1}(t_{n+1})) \geq 0, \tag{3.71}$$

since $\delta u_{m+1}(t_n) = 0$.

Consider now the bound on $\tilde{w}_{m+1}^{(B)}$ from Lemma 3.21 when either $\kappa > 0$ or $S = 1$. It is a simple exercise to confirm this bound is uniform as $\kappa \to \infty$, i.e. for all κ^*, there exists a compact set $K_n \subset \dot{\mathbb{X}}_n$ such that for all $\kappa \geq \kappa^*$, all Picard iterates $(\tilde{w}_m^{(B)})$ lie in K_n. In the case $S = 1$, we may similarly confirm that this bound is uniform as $\Delta t_n \to 0$, i.e. for all Δt_n^*, there exists a compact set $K_n \subset \dot{\mathbb{X}}_n$ such that for all $\Delta t_n \leq \Delta t_n^*$, all Picard iterates $(\tilde{w}_m^{(B)})$ lie in K_n. We restrict our attention then to iterates $\tilde{w}_m^{(B)\pm}$ in a compact set, over which the Lipschitz results (3.67) hold uniformly as we take either $\kappa \to \infty$ or (in the case $S = 1$) $\Delta t_n \to 0$. The affine transformation $\tilde{w}_m^{(B)\pm} \mapsto u_m^{\pm}$ implies the iterates $u_m^{\pm}$ lie similarly in a uniformly compact set.

Now, for each p, $\delta\tilde{w}_p^{(A)} := \tilde{w}_p^{(A)+} - \tilde{w}_p^{(A)-}$ satisfies the relation

$$\mathcal{I}_n\Big[M(v_p^{(A)}, \delta\tilde{w}_{p,m+1}^{(A)})\Big] = \int_{T_n} Q^{(A)\prime}(u_m^+; v_p^{(A)}) - Q^{(A)\prime}(u_m^-; v_p^{(A)}) \tag{3.72}$$

for all $v_p^{(A)} \in \dot{\mathbb{X}}_n$. Taking $v_p^{(A)} = \delta\tilde{w}_{p,m+1}^{(A)}$,

$$\mathcal{I}_n\Big[M(\delta\tilde{w}_{p,m+1}^{(A)}, \delta\tilde{w}_{p,m+1}^{(A)})\Big] = \int_{T_n} Q^{(A)\prime}(u_m^+; \delta\tilde{w}_{p,m+1}^{(A)}) - Q^{(A)\prime}(u_m^-; \delta\tilde{w}_{p,m+1}^{(A)}). \tag{3.73}$$

With $u_m^\pm \in Y_n$, the local Lipschitz differentiability condition on $Q_p^{(A)}$ (3.67c) implies

$$\Big|\mathcal{I}_n\Big[M(\delta\tilde{w}_{p,m+1}^{(A)}, \delta\tilde{w}_{p,m+1}^{(A)})\Big]\Big| \lesssim \int_{T_n} \|\delta u_m\|_* \|\delta\tilde{w}_{p,m+1}^{(A)}\|, \tag{3.74}$$

where on the RHS we use the triangle inequality. We may then bound

$$\begin{aligned}
\Delta t_n \sup_{T_n} \|\delta\tilde{w}_{p,m+1}^{(A)}\|_*^2 &\lesssim \mathcal{I}_n\Big[M(\delta\tilde{w}_{p,m+1}^{(A)}, \delta\tilde{w}_{p,m+1}^{(A)})\Big] && \text{(3.75a)}\\
&\lesssim \int_{T_n} \|\delta u_m\|_* \|\delta\tilde{w}_{p,m+1}^{(A)}\| && \text{(3.75b)}\\
&\lesssim \Delta t_n \sup_{T_n} \|\delta u_m\|_* \sup_{T_n} \|\delta\tilde{w}_{p,m+1}^{(A)}\|_*, && \text{(3.75c)}
\end{aligned}$$

where the first inequality holds by (3.37b). Dividing through by $\Delta t_n \sup_{T_n} \|\delta\tilde{w}_{p,m+1}^{(A)}\|_*$, we bound the difference in the additional AVs,

$$\sup_{T_n} \|\delta\tilde{w}_{p,m+1}^{(A)}\|_* \lesssim \sup_{T_n} \|\delta u_m\|_*. \tag{3.76}$$

By the uniform bounded of $u_m^\pm$, we can assume this scaling constant is independent of either κ or Δt_n.

Lastly, with these bounds, we may consider (3.62a). Taking the difference in (3.62a) between the $*^+$ and $*^-$ iterations, δu_{m+1} must satisfy the relation

$$\mathcal{I}_n\begin{bmatrix} M(\dot{\delta u}_{m+1}, v) \\ \qquad + \kappa D(\delta\tilde{w}_{m+1}^{(B)}, v)\end{bmatrix} = \mathcal{I}_n\begin{bmatrix} C(u_m^+, (\tilde{w}_{p,m}^{(A)+}); \tilde{w}_{m+1}^{(B)+}, v) \\ \qquad - C(u_m^-, (\tilde{w}_{p,m}^{(A)-}); \tilde{w}_{m+1}^{(B)-}, v) \\ \qquad + G(u_m^+, (\tilde{w}_{p,m}^{(A)+}); v) \\ \qquad - G(u_m^-, (\tilde{w}_{p,m}^{(A)-}); v)\end{bmatrix}, \tag{3.77}$$

for all $v \in \dot{\mathbb{X}}_n$. Taking $v = \delta w_{m+1}^{(B)}$ and applying (3.71),

$$\begin{aligned}
&Q^{(B)}(\delta u_{m+1}(t_{n+1})) \\
&\quad + \kappa\mathcal{I}_n\Big[D(\delta\tilde{w}_{m+1}^{(B)}, \delta\tilde{w}_{m+1}^{(B)})\Big]
\end{aligned} = \mathcal{I}_n\begin{bmatrix} C(u_m^+, (\tilde{w}_{p,m}^{(A)+}); \tilde{w}_{m+1}^{(B)+}, \tilde{w}_{m+1}^{(B)}) \\ \qquad - C(u_m^-, (\tilde{w}_{p,m}^{(A)-}); \tilde{w}_{m+1}^{(B)-}, \tilde{w}_{m+1}^{(B)}) \\ \qquad + G(u_m^+, (\tilde{w}_{p,m}^{(A)+}); \tilde{w}_{m+1}^{(B)}) \\ \qquad - G(u_m^-, (\tilde{w}_{p,m}^{(A)-}); \tilde{w}_{m+1}^{(B)})\end{bmatrix}. \tag{3.78}$$

We will use this result to show the iteration is a contraction.

Let us begin by bounding the RHS of (3.78). Note again that $u_m^{\pm}$ are uniformly bounded, and $(\tilde{w}_{p,m}^{(A)\pm})$ are uniformly bounded by (3.76). We may therefore bound the advective term on the RHS via the Lipschitz conditions (3.67a) as

$$\begin{aligned}
&\mathcal{I}_n\begin{bmatrix} C(u_m^+,(\tilde{w}_{p,m}^{(A)+});\tilde{w}_{m+1}^{(B)+},\delta\tilde{w}_{m+1}^{(B)}) \\ \qquad -C(u_m^-,(\tilde{w}_{p,m}^{(A)-});\tilde{w}_{m+1}^{(B)-},\delta\tilde{w}_{m+1}^{(B)})\end{bmatrix} \\
&\qquad\lesssim \mathcal{I}_n\begin{bmatrix} C(u_m^+,(\tilde{w}_{p,m}^{(A)+});\tilde{w}_{m+1}^{(B)+},\delta\tilde{w}_{m+1}^{(B)}) \\ \qquad -C(u_m^-,(\tilde{w}_{p,m}^{(A)-});\tilde{w}_{m+1}^{(B)+},\delta\tilde{w}_{m+1}^{(B)})\end{bmatrix} && \text{(3.79a)}\\
&\qquad\lesssim \Delta t_n \sup_{T_n}\left\{\begin{matrix}|C(u_m^+,(\tilde{w}_{p,m}^{(A)+});\tilde{w}_{m+1}^{(B)+},\delta\tilde{w}_{m+1}^{(B)}) \\ \qquad -C(u_m^-,(\tilde{w}_{p,m}^{(A)-});\tilde{w}_{m+1}^{(B)+},\delta\tilde{w}_{m+1}^{(B)})|\end{matrix}\right\} && \text{(3.79b)}\\
&\qquad\lesssim \Delta t_n \sup_{T_n}\max\{\|\delta u_m\|_*,(\|\delta\tilde{w}_{p,m}^{(A)}\|_*)\}\|\tilde{w}_{m+1}^{(B)+}\|_*\|\delta\tilde{w}_{m+1}^{(B)}\|_* && \text{(3.79c)}\\
&\qquad\lesssim \Delta t_n \sup_{T_n}\|\delta u_m\|_*\|\tilde{w}_{m+1}^{(B)+}\|_*\|\delta\tilde{w}_{m+1}^{(B)}\|_* && \text{(3.79d)}\\
&\qquad\lesssim \Delta t_n \sup_{T_n}\|\delta\tilde{w}_m^{(B)}\|_*\sup_{T_n}\|\delta\tilde{w}_{m+1}^{(B)}\|_*, && \text{(3.79e)}
\end{aligned}$$

where in the first inequality we use the skew-symmetry of C,[1] in the second we use (3.37a), in the third we use the Lipschitz regularity of C (3.67a), in the fourth we use the bound (3.76), and in the final inequality we use (3.69) and the uniform boundedness of $\tilde{w}_{m+1}^{(B)\pm}$; by a similar appeal to (3.67b), the forcing term on the RHS may be bounded as

$$\mathcal{I}_n\begin{bmatrix} G(u_m^+,(\tilde{w}_{p,m}^{(A)+});\delta\tilde{w}_{m+1}^{(B)}) \\ \qquad -G(u_m^-,(\tilde{w}_{p,m}^{(A)-});\delta\tilde{w}_{m+1}^{(B)})\end{bmatrix}\lesssim \Delta t_n \sup_{T_n}\|\delta\tilde{w}_m^{(B)}\|_*\sup_{T_n}\|\delta\tilde{w}_{m+1}^{(B)}\|_*. \tag{3.80}$$

Note, the uniform boundedness as $\kappa\to\infty$ implies the constant contained within $\lesssim$ is independent of κ.[2] The bounds (3.79, 3.80) together in (3.78) then imply

$$\begin{aligned}
&Q^{(B)}(\delta u_{m+1}(t_{n+1}))+\kappa\mathcal{I}_n\Big[D(\delta\tilde{w}_{m+1}^{(B)},\delta\tilde{w}_{m+1}^{(B)})\Big] \\
&\qquad\qquad\lesssim \Delta t_n \sup_{T_n}\|\delta\tilde{w}_m^{(B)}\|_*\sup_{T_n}\|\delta\tilde{w}_{m+1}^{(B)}\|_*. \qquad \text{(3.81)}
\end{aligned}$$

With the RHS of (3.78) now bounded, we proceed to the LHS. In the $\kappa\to\infty$ case, we observe the latter term $\kappa\mathcal{I}_n\Big[D(\delta\tilde{w}_{m+1}^{(B)},\delta\tilde{w}_{m+1}^{(B)})\Big]$ satisfies the bound

$$\kappa\Delta t_n\sup_{T_n}\|\delta\tilde{w}_{m+1}^{(B)}\|_*^2\lesssim\kappa\mathcal{I}_n\Big[D(\delta\tilde{w}_{m+1}^{(B)},\delta\tilde{w}_{m+1}^{(B)})\Big] \tag{3.82}$$

[1] Note the penultimate term $\tilde{w}_{m+1}^{(B)-}$ switching to $\tilde{w}_{m+1}^{(B)+}$.
[2] Independence with respect to Δt_n is included in the definition of $\lesssim$.

by (3.37b). Substituting this bound into (3.81),

$$\kappa\Delta t_n \sup_{T_n} \|\delta\tilde{w}_{m+1}^{(B)}\|_*^2 \lesssim \Delta t_n \sup_{T_n} \|\delta\tilde{w}_m^{(B)}\|_* \sup_{T_n} \|\delta\tilde{w}_{m+1}^{(B)}\|_* \tag{3.83a}$$

$$\sup_{T_n} \|\delta\tilde{w}_{m+1}^{(B)}\|_* \lesssim \frac{1}{\kappa} \sup_{T_n} \|\delta\tilde{w}_m^{(B)}\|_*. \tag{3.83b}$$

As $\kappa \to \infty$, since the constant contained in $\lesssim$ remains fixed, this implies the iteration is ultimately a contraction. In the $\Delta t_n \to 0$ case when $S = 1$, we bound the former term $Q^{(B)}(\delta u_{m+1}(t_{n+1}))$,

$$\sup_{T_n} \|\delta\tilde{w}_{m+1}^{(B)}\|_*^2 \lesssim \sup_{T_n} \|\delta u_{m+1}\|_*^2 \lesssim Q^{(B)}(\delta u_{m+1}(t_{n+1})), \tag{3.84}$$

where in the former inequality we use (3.69), in the latter we note that, for $S = 1$, $\|\delta u_{m+1}\|_*$ attains its maximum value at t_{n+1}. Substituting this bound into (3.81),

$$\sup_{T_n} \|\delta\tilde{w}_{m+1}^{(B)}\|_*^2 \lesssim \Delta t_n \sup_{T_n} \|\delta\tilde{w}_m^{(B)}\|_* \sup_{T_n} \|\delta\tilde{w}_{m+1}^{(B)}\|_* \tag{3.85a}$$

$$\sup_{T_n} \|\delta\tilde{w}_{m+1}^{(B)}\|_* \lesssim \Delta t_n \sup_{T_n} \|\delta\tilde{w}_m^{(B)}\|_*. \tag{3.85b}$$

As $\Delta t_n \to 0$, this bound implies the iteration is ultimately a contraction. Thus, Lemma 3.24 holds. □

Remark 3.25 (Picard linearisation as a solver)**.** *We recall that, as stated in Remark 3.6, we do not introduce the Picard linearisation as a solver for the SP discretisation (3.27) but merely as an analytical tool. If we were to use it however, Lemma 3.24 would offer certain conditions under which the Picard linearisation would be guaranteed to converge (linearly).*

Example (Incompressible NS)

Given either sufficiently small Re, or in the lowest-order-in-time case $S = 1$ sufficiently small Δt_n, the NS Picard linearisation (3.64, 3.65) is a contraction.

With the conditions for the Picard linearisation (3.50, 3.52) to be a contraction established, the proof that unique solutions exist to the original SP discretisation (3.27) is a simple application of the CMT.

Theorem 3.26 (Uniqueness of solutions: AD systems)**.** *Assume Assumption 3.22, i.e. we have a suitably regular AD problem. Then there exists a unique solution to the SP discretisation (3.27) for either sufficiently large κ, or in the lowest-order-in-time case $S = 1$ with sufficiently small Δt_n.*

Proof. This holds an immediate consequence of Lemma 3.24 by the CMT (see [Eva10, Sec. 9.2]). □

Example (Incompressible NS)

The SP discretisation of the NS equations (3.36) is well-posed with a unique solution for either sufficiently small Re, or in the lowest-order-in-time case $S = 1$ with sufficiently small Δt_n.

"[...] it wasn't much good having anything exciting [...] if you couldn't share [it] with somebody."

— Piglet [Mil26]

4 Implementation & computation of auxiliary variables

Contents

The general SP discretisation (3.27) poses a mixed problem in $P+1$ solution and test variables over a space-time domain $\Omega \times T_n$; both of these aspects pose certain computational challenges. The efficiency of numerical solvers typically scales poorly with the number of solution variables, implying each further structure preserved by our framework comes at the expense of the computational speed of the numerical method. Furthermore, a large portion of commercially available FE software is not equipped to solve variational problems over space-time domains. In this chapter, we discuss various approaches that can be taken to mitigate these issues.

For the code that used to generate the numerical results throughout this thesis, see Section 1.1.

The rest of this chapter proceeds as follows. In Section 4.1, we discuss how a polynomial-in-time expansion can be used to reduce the space-time problem (3.27) posed over $\Omega \times T_n$ to a set of S spatial problems posed over the domain Ω, and how careful choice of the basis for this expansion can improve the sparsity of our assembled problem. In Section 4.2, we discuss situations in which one need not solve for certain AVs $(\tilde{w}_p)$ numerically, i.e. when the auxiliary equation (3.27b) can be solved offline for all u (either independent of u, or as an explicit function of it). In Section 4.3, we conclude with a discussion of the special connection held between our SP schemes and Gauss collocation methods [HLW06, Sec. II.1.3], in particular discussing those situations in which the former may reduce to the latter.

4.1 Practical implementation of space-time problems

As a variational problem over the high-dimensional domain $\Omega \times T_n$, it is often not favourable, or possible, to implement (3.27) directly as written.

In this section, we discuss how one can circumvent this issue by writing each of the solution and test functions in terms of a polynomial basis in time, reducing the system from one defined over the space-time domain $\Omega \times T_n$ to one over the domain Ω in space only. We discuss then how, under certain conditions, careful choice of the polynomial basis functions for the expansion in time can lead to increased sparsity in the assembled problem.

4.1.1 Polynomial expansion in time

Let us write each of the solution $(u, (\tilde{w}_p))$ and test $(v, (v_p))$ functions in (3.27) in terms of a polynomial basis in time, and work directly with each component of this expansion. Suppose $(l_s)_{s=1}^S$ forms a basis for $\mathbb{P}_{S-1}(T_n)$; let us write $(\tilde{w}_p)$, v, (v_p), each a function in $\dot{\mathbb{X}}_n$, in terms of (l_s),

$$\tilde{w}_p = \sum_{r=1}^{S} \tilde{w}_{p,r} l_r, \qquad v = \sum_{r=1}^{S} v_r l_r, \qquad v_p = \sum_{r=1}^{S} v_{p,r} l_r. \tag{4.1a}$$

Supposing $(k_s)_{s=0}^S$ form a basis for $\mathbb{P}_S(T_n)$ such that each $k_s(t_n) = 0$ for $s > 0$, we may write u, and consequently $\dot{u}$, as

$$u = u(t_n) k_0 + \sum_{r=1}^{S} u_r k_r, \qquad \dot{u} = u(t_n) \dot{k}_0 + \sum_{r=1}^{S} u_r \dot{k}_r. \tag{4.1b}$$

The scheme (3.27) can then be written in terms of these expansions as follows: find $((u_r), (\tilde{w}_{p,r})) \in \mathbb{U}^S \times \mathbb{U}^{PS}$ such that

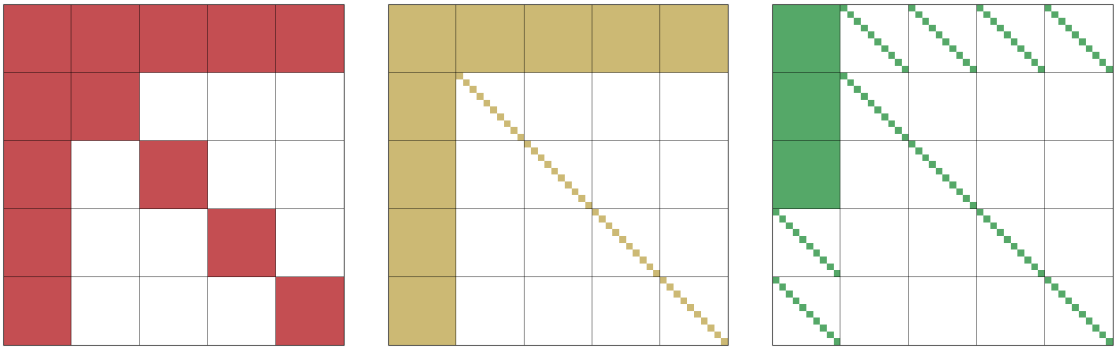
$$\mathcal{I}_n\left[M\left(u; u(t_n)\dot{k}_0 + \sum_{r=1}^{S} u_r \dot{k}_r, v_s l_s\right)\right] = \mathcal{I}_n\left[\tilde{F}\left(u, \left(\sum_{r=1}^{S} \tilde{w}_{p,r} l_r\right); v_s l_s\right)\right], \tag{4.2a}$$

$$\mathcal{I}_n\left[M\left(u; v_{p,s} l_s, \sum_{r=1}^{S} \tilde{w}_{p,r} l_r\right)\right] = \int_{T_n} Q'_p(u; v_{p,s} l_s), \tag{4.2b}$$

for all $((v_s), (v_{p,s})) \in \mathbb{U}^S \times \mathbb{U}^{PS}$, where u can be substituted for its expansion (4.1b). Pre-computing this decomposition then gives a set of $(P+1)S$ discrete problems in $\mathbb{U}$, posed over the spatial domain Ω only.

4.1.2 Sparsity improvements

Consider the formulation (4.2). Without any further work, few terms cancel, leading to a largely dense assembled problem at high order in time, as $S \to \infty$ (see Fig. 4.1a). In certain cases, the sparsity of this problem can be improved through careful choice of basis function (l_s). There are many such cases, and many ways in which this can be done; we consider here two, which will be useful in Section 4.2.



(a) (4.2) **(b)** (4.3) **(c)** (4.4)

Figure 4.1: Graphical illustrations of the sparsity of the problems (4.2–4.4), for a hypothetical example system with $S = 10$ and $P = 4$. Columns align with the solutions functions, ordered $(u_1, \cdots, u_{10}, \tilde{w}_{1,1}, \cdots, \tilde{w}_{1,10}, \cdots, \tilde{w}_{4,1}, \cdots, \tilde{w}_{4,10})$; rows align with the test functions, ordered $(v_1, \cdots, v_{10}, v_{1,1}, \cdots, v_{1,10}, \cdots, v_{4,1}, \cdots, v_{4,10})$. In the final subfigure (Fig. 4.1c) we assume Q_3, Q_4 are quadratic.

4.1.2.1 M independent of u

Let us first suppose M is independent of u, e.g. by Assumption 3.9. By Lemma 3.7, we may choose (l_s) to be orthonormal under the inner product $(\phi, \varphi) \mapsto \mathcal{I}_n(\phi\varphi)$,

e.g. Legendre polynomials when $\mathcal{I}_n$ is the exact integral $\int_{T_n}$; we choose then (k_s) such that each $\dot{k}_s = l_s$, with $k_0 = 1$. In such a case, the system (4.2) simplifies to the following: find $((u_r), (\tilde{w}_{p,r})) \in \mathbb{U}^S \times \mathbb{U}^{PS}$ such that

$$M(u_s, v_s) = \mathcal{I}_n\left[\tilde{F}\left(u, \left(\sum_{r=1}^{S} \tilde{w}_{p,r} l_r\right); v_s l_s\right)\right], \tag{4.3a}$$

$$M(v_{p,s}, \tilde{w}_{p,s}) = \int_{T_n} Q_p'(u; v_{p,s} l_s), \tag{4.3b}$$

for all $((v_s), (v_{p,s})) \in \mathbb{U}^S \times \mathbb{U}^{PS}$. The associated sparsity pattern is illustrated in Fig. 4.1b.

4.1.2.2 $\mathcal{I}_n$ an S-node quadrature rule

Let us now return to the case of general M, and suppose $\mathcal{I}_n$ is an S-node quadrature rule at quadrature points $\{\tau_{n,s}\}_{s=0}^{S-1}$, i.e. our discretisation is an SP modification of an S-stage collocation method. We may choose (l_s) to be the associated Lagrange basis polynomials at $\{\tau_{n,s}\}$. For (k_s), we choose $k_0 = (-1)^S \pi_S$ where π_S is the degree-S Legendre polynomial shifted to the interval T_n (see Fig. 3.4) and k_s for $s > 0$ contructed implicitly such that $\int_{T_n} k_r l_s = \mathcal{I}_n[l_s]\delta_{rs}$. In such a case, the system (4.2) simplifies to the following: find $((u_r), (\tilde{w}_{p,r})) \in \mathbb{U}^S \times \mathbb{U}^{PS}$ such that

$$M(u(\tau_{n,s}); \dot{u}(\tau_{s,n}), v_s) = \tilde{F}(u(\tau_{n,s}), (\tilde{w}_{p,s}); v_s), \tag{4.4a}$$

$$M(u(\tau_{n,s}); v_{p,s}, \tilde{w}_{p,s}) = \frac{1}{\int_{T_n} l_s} \int_{T_n} Q_p'(u; v_{p,s} l_s), \tag{4.4b}$$

for all $((v_s), (v_{p,s})) \in \mathbb{U}^S \times \mathbb{U}^{PS}$. For quadratic invariants Q_p, for which Q_p' defines a bilinear form, the final term on the RHS of (4.4b) reduces to

$$\frac{1}{\int_{T_n} l_s} \int_{T_n} Q_p'(u; v_{p,s} l_s) = Q_p'(u_s; v_{ps}). \tag{4.5}$$

The associated sparsity pattern is illustrated in Fig. 4.1c.

4.2 Elimination of AVs

In certain cases, the AVs $(\tilde{w}_p)$ introduced by our framework need not be introduced into the numerical implementation. They are simply helpful tools in the construction and analysis of the scheme, but do not increase the dimension of the original problem.

4.2.1 ODE systems

The most powerful result arguably comes when we are considering an SP discretisation of an ODE system, such that $\mathbb{U} = \mathbb{R}^n$.

Assume we are considering an SP modification of an S-stage collocation method, such that the system (4.2) can be written in the form (4.4). The auxiliary equation (4.4) defines a linear system in $\tilde{w}_{p,s}$ through M; this can generally be with minimal computational cost, as the operator M is typically largely sparse in the final two arguments, e.g. an ℓ^2 inner product, or one scaled with u. This gives us an explicit equation for $(\tilde{w}_{p,s})$, and consequently $(\tilde{w}_p)$, that can be evaluated offline; rewriting $(\tilde{w}_{p,s})$ or $(\tilde{w}_p)$ in (4.4a) or (3.27a) respectively according to this explicit pre-computation eliminates the AVs from the numerical implementation.

4.2.2 PDE systems & independent associated test functions

Moving beyond ODE systems, the most trivial case to consider comes when $w_p(u)$ is simply independent of u. This is often the case when the QoI is a mass function, and $w_p(u)$ is constant in space and time.

Consider the definition of $\tilde{w}_p$ (3.19):

$$\mathcal{I}_n[M(u; v_p, \tilde{w}_p)] = \int_{T_n} Q'_p(u; v_p) = \int_{T_n} M(u; v_p, w_p(u)), \tag{4.6}$$

for all $v_p \in \mathring{\mathbb{X}}_n$. Let us assume $w_p(u)$ is independent of u with $w_p \in \mathbb{U}$, e.g. constant in space with no relevant zero BCs imposed on $\mathbb{U}$. When $\mathcal{I}_n$ in (4.6) may be equivalently substituted for $\int_{T_n}$, the relation is satisfied exactly when $\tilde{w}_p = w_p$; we may then equivalently substitute the AV $\tilde{w}_p$ for the associated test function w_p, and avoid introducing it into our discretisation.

The most natural case in which $\mathcal{I}_n$ may be substituted for the exact integral $\int_{T_n}$ comes when we choose $\mathcal{I}_n$ to be the exact integral in Step **B** of our framework. We make this choice in Section 7.3 to eliminate certain AVs introduced for mass, momentum and energy conservation (each approximating 1).

Otherwise, $\mathcal{I}_n$ may be substituted for the exact integral $\int_{T_n}$ in (4.6) when the quadrature $\mathcal{I}_n$ is exact on $M(u; v_p, \tilde{w}_p)$. This depends on the order of the quadrature rule $\mathcal{I}_n$ and the order in time of the integrand $M(u; v_p, \tilde{w}_p)$. In the simplest case, where M is independent of u (e.g. under Assumption 3.9), $M(v_p, \tilde{w}_p)$ is of order

$S-1$ in time; any consistent $\geq S$-node quadrature rule $\mathcal{I}_n$ will evaluate the integral on $M(v_p, \tilde{w}_p)$ exactly, and can therefore equivalently be substituted for $\int_{T_n}$.[1] In fact, provided M is polynomial in u, there necessarily exists a sufficiently high order for $\mathcal{I}_n$ above which the quadrature $\mathcal{I}_n[M(u; v_p, \tilde{w}_p)]$ is exact. In any of these cases, the AV $\tilde{w}_p$ may be equivalently substituted for the associated test function $w_p(u)$, and not introduced into the discretisation.

4.2.3 PDE systems & dependent associated test functions

When considering SP discretisations of PDE systems, the less trivial case comes when the associated test function $w_p(u)$ still lies in $\mathbb{U}$, but is dependent on u (e.g. as was the case for the incompressible NS example in Chapter 3). This if often the case when the QoI is a quadratic energy functional when $w_p(u) \propto u$, or for ODEs when the inclusion $w_p(u) \in \mathbb{U} = \mathbb{R}^d$ is trivial.

Let us now assume M defines as an inner product, as in Assumption 3.9. With M independent of u, we may write the system (4.2) in the form (4.3). Noting $Q_p'(u;\cdot) = M(\cdot, w_p(u))$ (by definition) the auxiliary equation (4.3b) can be written in the form

$$M(v_{p,s}, \tilde{w}_{p,s}) = \int_{T_n} M(v_{p,s} l_s, w_p(u)) = M\Big(v_{p,s}, \int_{T_n} w_p(u) l_s\Big), \tag{4.7}$$

for all $v_{p,s} \in \mathbb{U}^S$. We can use the fact that M defines an inner product on $\mathbb{U}$ to write $\tilde{w}_{p,s}$, and consequently $\tilde{w}_p$, explicitly, as

$$\tilde{w}_{p,s} = \int_{T_n} w_p(u) l_s, \qquad \tilde{w}_p = \sum_{r=0}^{S-1} \Big(\int_{T_n} w_p(u) l_r\Big) l_r. \tag{4.8}$$

In such cases, by substituting this identity for $\tilde{w}_{p,s}$ or $\tilde{w}_p$ back into (4.2a) or (3.27a) respectively, we can equivalently apply our SP framework without the additional computational cost of computing the AV.

Example (Incompressible NS)

In the NS scheme (3.28), $\mathbf{w}_1(\mathbf{u}) = \mathbf{u} \in \mathbb{U}$, whereas $\mathbf{w}_2(\mathbf{u}) = \operatorname{curl} \mathbf{u} \notin \mathbb{U}$ necessarily. We can therefore use (4.8) to define $\tilde{\mathbf{u}} = \tilde{\mathbf{w}}_1$ in (3.28a) and eliminate (3.28b) from the mixed formulation. However, we must include $\tilde{\boldsymbol{\omega}} = \tilde{\mathbf{w}}_2$ and

[1] This is analogous to the conservation of linear invariants by consistent RK methods.

(3.28c) if we seek to preserve the helicity.

4.3 Gauss methods

We give special attention here to the relationship between schemes deriving from our framework and Gauss collocation methods [HLW06, Sec. II.1.3]. At least for linear systems, Gauss collocation methods are equivalent to CPG schemes [EG21c, Lem. 70.5] where their SP properties are well-established; it is natural therefore to expect some form of connection between them and the framework presented here. We consider precisely those cases where our proposed schemes are equivalent to collocation methods, in particular Gauss methods.

Theorem 4.1 (Relationship between our proposed schemes and Gauss methods)**.** *Provided at least one (Q_p) has degree ≥ 2, schemes deriving from the framework can be equivalent to a mixed collocation method if and only if all (Q_p) have degree ≤ 2, i.e. only quadratic invariants at most are considered. In such a case, our proposed scheme is equivalent to the S-stage Gauss method.*

Proof. The proposed scheme will be equivalent to a collocation method if and only if $\mathcal{I}_n$ is the corresponding S-point quadrature rule and all exact integrals over T_n can be equivalently substituted for $\mathcal{I}_n$; specifically, from the RHS of (3.27b), we require that for all p,

$$\int_{T_n} Q'_p(u; v_p) = \mathcal{I}_n[Q'_p(u; v_p)]. \tag{4.9}$$

For this to hold, the order in time of the integrand $Q'_p(u; v_p)$ must be no greater than the order of the quadrature rule $\mathcal{I}_n$.

If Q_p is of degree r_p in u, $Q_p(u)$ is of degree $r_p S$ in time by composition; similarly, $Q'_p(u; v_p)$ is of degree $r_p S - 1$ in time. Since at least one $r_p \geq 2$, at least one $r_p S - 1 \geq 2S - 1$. The only S-point quadrature rule with order $2S - 1$ is GL quadrature [Tre20, Thm. 19.1] corresponding to the S-stage Gauss method. Moreover, since we then require $r_p S - 1 \leq 2S - 1$ for all p, it must hold that $r_p \leq 2$, i.e. all (Q_p) have degree ≤ 2. □

Example (Incompressible NS)

As per Theorem 4.1, when one begins with a Gauss collocation method (i.e. $\mathcal{I}_n$ is a GL quadrature rule) the energy- and helicity-conserving NS scheme (3.28) is equivalent to a mixed Gauss collocation method. Further eliminating $\tilde{\mathbf{u}}$ using (4.8), this can be stated as a Gauss method applied to the following semi-discretisation: find $(\mathbf{u}, \boldsymbol{\omega}) \in \mathbb{U} \times \mathbb{U}$ such that

$$(\dot{\mathbf{u}}, \mathbf{v}) = (\mathbf{u} \times \boldsymbol{\omega}, \mathbf{v}) - \frac{1}{\mathrm{Re}}(\nabla \mathbf{u}, \nabla \mathbf{v}), \tag{4.10a}$$

$$(\boldsymbol{\omega}, \boldsymbol{\chi}) = (\operatorname{curl} \mathbf{u}, \boldsymbol{\chi}), \tag{4.10b}$$

at all times $t \in \mathbb{R}_+$ and for all $(\mathbf{v}, \boldsymbol{\chi}) \in \mathbb{U} \times \mathbb{U}$.

It is well established that RK methods are unable in general to conserve non-quadratic invariants (see Celledoni *et al.* [Cel+09]) and that among all collocation methods this is achieved only by Gauss methods (see Hairer, Lubich & Wanner [HLW06, Sec. IV.2.1]). When the framework is applied to modify Gauss collocation methods, the resulting schemes can be thought of as generalisations of Gauss methods that retain their SP properties for non-quadratic invariants.

Part II

Applications of the framework:
ODEs & PDEs

"This is where the fun begins."

— Anakin Skywalker (Hayden Christensen)
[Luc05]

5 Introduction

Contents

With the general framework established, we turn our attention to certain example applications, within both ODEs and PDEs. In certain cases, including e.g. the energy-stable integrators for Poisson & gradient-descent ODEs (6.11) and PDEs (7.8), these identify with schemes already established in the literature (see the literature review in Section 5.1) whereas in others, including e.g. the conservative integrator for conservative ODEs (6.65) and the mass-, momentum-, energy-conserving and entropy-producing integrator for the compressible NS equations (7.59), we believe the schemes and their SP properties to be novel. For each of the SP ODE integrators, we offer conditions for the unique existence of discrete solutions, and their convergence.

Conservative systems, the Kepler problem, the Kovalevskaya top & the Benjamin–Bona–Mahony equation

Our first example applications concern general conservative systems. Many physical systems evolve under purely conservative dynamics, with time evolution confined to an invariant manifold $\mathcal{M}$, the intersection of level sets of one or more invariants. The geometry of $\mathcal{M}$ fundamentally shapes the behaviour of solutions, in particular

when it is of low dimension; for example, for topologically circular $\mathcal{M}$, typical for maximally superintegrable systems with one fewer invariant than degrees of freedom (DoFs), solutions are necessarily periodic, while for topologically toroidal $\mathcal{M}$ solutions can exhibit quasiperiodicity. Even for relatively high-dimensional $\mathcal{M}$, its structure can be fundamental to the behaviour of solutions. In particular, when the ICs are such that their invariants imply $\mathcal{M}$ is in some way close to a manifold of low dimension, solutions will exhibit perturbed forms of these qualitative dynamics; for example, when $\mathcal{M}$ is close to a one-dimensional set, solutions exhibit a perturbed travelling wave behaviour or form of periodicity, giving rise to the nonlinear stability of solitons, cnoidal waves and breathers. Moreover, the structure of $\mathcal{M}$ can be fundamental to the analysis of the well-posedness of conservative systems, particularly for PDEs.

For ODEs, we construct in Section 6.2 a general-purpose integrator, capable of preserving arbitrarily many invariants. In doing so, we are able to restrict our discrete trajectories to the exact invariant manifold $\mathcal{M}$ when all invariants are known. We see the implications of this on the preservation of dynamics in e.g. the discrete solution for our simulation of the Kovalevskaya top in Subsection 6.2.2, which can clearly be identified as preserving the quasiperiodicity of the exact solution (see Fig. 6.4). For maximally superintegrable systems, in particular periodic ODE systems for which we are able to identify one fewer invariant than DoFs, our method conserves trajectories exactly.[1] The construction of our integrator rewrites the RHS forcing term in terms of an alternating form acting on the invariant gradients (Lemma 6.15) while coupling with a set of AVs that discretely approximate these gradients.

For PDEs, we focus in Section 7.1 on the preservation of a single invariant (typically an energy) with the goal of preserving both the dynamical structure provided by $\mathcal{M}$ and its analytic properties regarding the existence of solutions; this we are able to achieve for general conservative PDE systems (7.8). We see the dynamical benefits in our example application on the BBM equation (Subsection 7.1.1) for which the conservation of energy helps our discrete soliton remain stable over long

[1] We clarify here that the exact preservation of trajectories is not sufficient to imply exact solutions, as a reparametrisation in time of the exact trajectory is still admissible. We see this for instance in our SP discretisation of the Kepler problem in Subsection 6.2.1 (see Fig. 6.1 below).

durations. Our discrete scheme is constructed using an AV approximating a suitable representation of Fréchet derivative of the invariant of interest.

Gradient-descent systems

Ubiquitous throughout physical modelling and engineering, gradient-descent systems' defining structural features are their monotonic dissipations. This is often not just key to the long-term dynamics of solutions, but to their analysis, ensuring the existence of solutions and their convergence as $t \to \infty$.

We construct in Sections 6.1 & 7.1 general-purpose integrators for gradient-descent ODEs and PDEs; these discretisations preserve the dissipation structure discretely, in a way that moreover replicates the dissipation inequality from the continuous level. Similarly to the conservative integrators for ODEs and PDEs, this relies on the introduction of and coupling with an AV approximating either the gradient of the dissipated functional (in the ODE case) or a certain representation of its Fréchet derivative (in the PDE case).

The compressible Navier–Stokes equations, mass, momentum, energy & entropy

While the incompressible NS equations (considered in Chapter 3) provide a robust model for many fluid regimes, particularly those dominated by slow (i.e. low Mach number) flows, they fail to capture essential physical features in settings where density variations, shock formation, or compressional heating are significant. In such regimes, we must consider the compressible NS equations (see Feireisl [Fei04]) bringing with it a host of different conservation and dissipation structures.

With density variations now admissible, the conservation of mass becomes a dynamical law that must be preserved at the discrete level, i.e. no longer enforceable via algebraic constraints as in the incompressible case (3.28) through constraints on $\mathbb{U}$ (3.7a, 3.7b); momentum conservation retains a similar form, while energy balance now becomes a conservation law: kinetic energy (ultimately) converts into internal energy through viscous heating; convergence to equilibrium is driven by the generation (or conservation, in the inviscid Euler case) of entropy, a certain function of the material's thermodynamic quantities related by the medium's constitutive

relation. In essence, mass, momentum and energy conservation ensures boundedness and physical realism, while entropy generation ensures irreversibility and regularity. We refer to a scheme that preserves all these laws as mass-, momentum-, energy-, and entropy-stable.

Compressible effects are relevant for important aspects of plasma dynamics, particularly in the edge region of tokamak plasmas, where steep gradients and heating dominate the plasma state.[2] Owing to the strong anisotropy imposed by the magnetic field, plasma flows parallel and perpendicular to the field lines are often modelled via separate 1D and 2D compressible systems (see Arter [Art95]). Ion and electron species are then typically treated separately, with electron dynamics often significantly simplified.

During edge-localised modes (ELMs), steep pressure gradients drive sound waves and shock fronts. Radio-frequency heating and neutral beam injection can induce localised pressure fluctuations. Confinement loss events such as sawtooth crashes (discussed further in Chapter 8) may also trigger strong compressive fronts. Accurately reproducing these features demands a model that accurately preserves the system's physical structures and laws.

Through our framework, we construct in Chapter 7.3 a mass-, momentum-, energy- and entropy-stable FE integrator for the compressible NS equations (7.59). This requires the introduction of 3 AVs: one approximating the inverse temperature, one for the velocity, and one further AV approximating an (intensive) quantity related to the specific Gibbs free energy. Notably, this guarantees convergence to the correct thermodynamic equilibrium as $t \to \infty$.

Boltzmann, GENERIC, energy & entropy

Fluid models like the compressible NS equations are ineffective in rarefied or low-density conditions, where the mean free path is large and the continuum assumption breaks down. In such regimes, the more fundamental description offered

[2]While we do not, in this thesis, directly propose an SP discretisation for the full compressible MHD equations, such a scheme may be built by combining our SP discretisations of the compressible NS (7.59) and MHD equations (10.82). However, we note that compressible MHD equations are more commonly applied in contemporary literature within astrophysical settings, suggesting that such a discretisation may find broader practical use beyond fusion-specific modelling.

by continuum kinetic models such as the Boltzmann equation (see Cercignani [Cer12]), evolving a particle distribution in phase space, are required.

In this thesis we consider two key thermodynamic structures exhibited by the Boltzmann equation: energy conservation and entropy generation. The former ensures long-term stability, while the latter governs relaxation and irreversibility.

Kinetic effects have a potential to affect all aspects of the operation of plasma in fusion reactors. In the fusion plasma edge and scrape-off layer, low densities, steep gradients, and magnetic field anisotropy can give rise to behaviour necessitating a kinetic treatment. Disruption events (again such as ELMs), pellet injections, and sawtooth crashes can drive the plasma far from equilibrium, leading to anisotropic and multiple-peaked velocity distributions. While kinetic simulations of charged species in 3D often employ gyrokinetic models (which typically further differ from our scheme (7.40) through the use of Fokker–Planck-type collision operators) Boltzmann-like models are still employed, particularly for neutral species. For a summary of the various model equations frequently used in modern tokamak edge codes, see the report from Arter [Art23, Sec. 1.2], in particular Proxyapps 2-6, 2-8 and 3-2 for continuum kinetic models; for further background on kinetic modelling in fusion plasmas, see Stacey [Sta12, Chap. 16].

The Boltzmann equation can be understood as a metriplectic/GENERIC system: a certain class of systems that conserves an energy and generates an entropy (consistent with the first and second laws of thermodynamics).

These formalisms consider ODEs and PDEs governed by two distinct contributions: a Poisson bracket/operator capturing the reversible dynamics (i.e. the transport term in the Boltzmann equation) and a dissipative bracket/operator encoding the irreversible dynamics (i.e. the collisions), the latter driving entropy production and convergence toward equilibrium. While the metriplectic formulation, through Poisson brackets and Hamiltonians, is more favoured in the plasma physics community and literature, we favour the GENERIC formalism, through Poisson operators and energy gradients (see the discussion in Section 5.1).

In either case, respecting this dual conservation–dissipation structure at the discrete level ensures consistency with the system's physical laws, in turn capturing the correct macroscopic behaviour, even far from equilibrium. In Section 6.3 (for ODEs) and Section 7.2 (for PDEs) we construct geometric integrators that respect

preserve these structures for both ODEs (6.81) and PDEs (7.24). These schemes conserve energy exactly, while generating entropy at a rate consistent with the continuous system. This is achieved by introducing two AVs, approximating either the gradients of energy and entropy (in the ODE case) or appropriate Fréchet derivatives (in the PDE case).

We use the Boltzmann equation as a key example application of our reversible–irreversible integrator in Subsection 7.2.1, necessarily preserving both thermodynamic laws even in far-from-equilibrium simulations.

5.1 Related literature

We discuss now relevant literature to the systems and numerical integrators discussed within this part of thesis.

Hamiltonian systems & symplectic integrators

A special kind of Poisson system (as considered in Section 6.1) is the Hamiltonian [Ham34] system, well-studied within geometric numerical integration due to its symplectic structure [HLW06, Chap. VI & VII]. In particular, a famous result of Zhong & Marsden [GM88] asserts that, under certain conditions on the underlying Hamiltonian system,[3] a numerical integrator can only be both symplectic and energy-stable if it solves for the system's trajectory exactly, i.e. it solves the system exactly up to a reparametrisation in time. The impact of this result seems effectively to be that we are offered a choice when discretising a Hamiltonian system, between symplecticity and energy stability; we restrict our attention therefore to the latter.

Regarding GENERIC ODEs, Shang & Öttinger [SÖ20] proposed a splitting, applying a symplectic integrator to the conservative component. They then defined the friction matrix for the dissipative component in terms of the symplectic integrator's modified energy; this somewhat resembles our definition of $\tilde{D}$ as a function of $\widetilde{\nabla H}$ in Section 6.3 (Assumption 6.17).

[3] In particular, the system is required to have no invariants but the energy (or functions thereof).

Energy-stable integrators for Poisson & gradient-descent systems

We begin by revisiting certain literature discussed in Chapter 2. The energy-stable discrete-gradient method of McLachlan, Quispel & Robidoux [MQR99] and in particular its high-order generalisation by Cohen, Hairer & Lubich [CH11; HL14] for Poisson & discrete-gradient ODEs identify exactly with the energy-stable scheme (6.11) when $\mathcal{I}_n$ is an S-stage quadrature rule. When the ODE system (6.3) is written in the form $B(\mathbf{x})^{-1}\dot{\mathbf{x}} = \nabla H(\mathbf{x})$ or the PDE system (7.3) is written in the (variational) form $B^{-1}(u;\dot{u},v) = H'(u;v)$ (for all v) for skew-symmetric/negative semidefinite $B^{-1}(u;\cdot,\cdot)$, the application of our framework with $\mathcal{I}_n = \int_{T_n}$ to construct an energy-stable integrator returns a scheme exactly equivalent to a CPG method, i.e. the introduced AVs do not affect the discretisation. This identifies with the observations of Egger, Habrich & Shashkov [EHS21] on the energy stability properties of CPG.

In his PhD thesis [Jac19, Chap. 4] Jackaman established a framework for the construction of energy-conserving semidiscrete schemes for certain Hamiltonian PDEs, discretised in space only. Similar to (7.8) this introduced an AV approximating the Riesz representation of the Fréchet derivative of the Hamiltonian; he then employs Crank–Nicolson in time, which is conservative for a linearised problem. Jackaman & Pryer [JP21] propose energy-conserving schemes for a certain class of dispersive Hamiltonian PDEs, including the Korteweg–de Vries (KdV) equation, using discontinuous FEs in space. Under a certain handling of the non-conforming terms, this scheme is equivalent to (7.8) at lowest order in time.

In 2023, Brunk *et al.* [Bru+23b] analysed a 1-stage energy-stable integrator for the Cahn–Hilliard [CH58; Cah61] equations, a commonly studied 4^{th}-order PDE. As a gradient-descent system, our integrator (7.8) aligns precisely with theirs at lowest order in time ($S = 1$). Since this initial publication, Brunk *et al.* have extended this scheme to incorporate various additional phenomena such as temperature variations [BLS25] and cross-kinetic coupling [BEH24], and to associated systems such as the Allen–Cahn [BGL24], Cahn–Hilliard–Navier–Stokes [Bru+23a; BS24; BS25; BE25], Cahn–Hilliard–Forchheimer [BF25a] and Cahn–Hilliard–Biot [BF25b] equations. In each case, their proposed SP scheme is reproduced by our framework.

Giesselmann, Karsai & Tscherpel [GKT25] recently announced a very closely related work. They devise energy-conserving and correctly-dissipative FET discretisations for general port-Hamiltonian systems by introducing projections of a representation of the Fréchet derivative of the Hamiltonian onto the discrete set, and explicitly note that this can be understood in terms of an AV. In the absence of the control terms in the port-Hamiltonian formulation, the proposed scheme coincides with (7.8) almost exactly when B and D are independent of u, and M defines an inner product (Assumption 3.9).

We note again the early work of French & Schaeffer [FS90] first establishing the connections between CPG and SP. The authors proposed the introduction of and coupling with an AV for energy conservation in the KdV equation, a Hamiltonian PDE; this AV coincides exactly with that introduced by both our energy-stable method (6.11) and that of Giesselmann, Karsai & Tscherpel [GKT25] at lowest order in time ($S = 1$).

Stable integrators for conservative ODE systems

As discussed in Section 4.3, the most foundational result within conservative integrators for multiple invariants lies in the conservation of quadratic invariants, or preservation of quadratic dissipation laws, by symplectic integrators (see [HLW06, Sec. IV.2]) with RK methods unable in general to preserve non-quadratic structures (see [Cel+09]). While this result is restrictive, it is sufficient in certain circumstances to preserve structures on higher-order QoIs through reparametrisation. For example, introduced by Lax [Lax68] in 1968, the isospectral flow or Lax pair system[4] $\dot{L} = B(L)L - LB(L)$ for symmetric $L(t) \in \mathbb{R}^{d\times d}$ and skew-symmetric $B(L) \in \mathbb{R}^{d\times d}$ conserves the spectrum of L. While the spectrum of L can be characterised by a set of non-quadratic structures, Flaschka [Fla74] observed in 1974 that we may write $L(t) = U(t)L(0)U(t)^\top$, for some orthogonal $U(t) \in \mathbb{R}^{d\times d}$ satisfying $\dot{U} = B(L)U$ with $U(0) = I$ with transpose $U(t)^\top$; in 1997, Calvo, Iserles & Zanna [CIZ97] observed that the conservation of the spectrum of L (i.e. the isospectral property) is then equivalent to the conservation of the orthogonality of U, a quadratic structure easily

[4] See Chu [Chu92] and Calvo, Iserles & Zanna [CIZ97] and citations therein for problems that can be written in this form, including the Toda lattice and a continuous counterpart to the QR algorithm.

conserved by classical numerical integration methods, allowing the conservation of multiple non-quadratic invariants through reparametrisation.

Simpler even than quadratic structures, any consistent RK method applied to an ODE system will preserve all linear structures. Similarly then, this is sufficient under certain circumstances to preserve higher order structures through reparametrisation. For the above Lax pair system for example, Diele, Lopez & Politi [DLP98] suggest the use of the Cayley transform $U = (I - Y)^{-1}(I + Y)$ for skew-symmetric $Y(t) \in \mathbb{R}^{d\times d}$ such that $\dot{Y} = \frac{1}{2}(I - Y)B(L)(I + Y)$ with $Y(0) = 0$; the isospectral property is then equivalent to the conservation of the skew-symmetry of Y, a linear structure.

A number of direct approaches exist for the preservation of conservation laws. One natural such approach is through projection, i.e. the use of an arbitrary numerical integrator, followed by a projection to the target invariant manifold $\mathcal{M}$. Under sufficient regularity conditions, these projection methods should cause no deterioration in the convergence of the method; numerical experiments, however, observe this can damage the long-term behaviour of the solution, in particular due to the loss of any other SP properties the initial integrator may have possessed (see Hairer, Lubich & Wanner [HLW06, Sec. IV.4]). Where there exists a local parametrisation of $\mathcal{M}$, a second approach involves redefining the system in these local coordinates; this in particular is true for systems in Lie groups, for which local parametrisations exist on $\mathcal{M}$ through the associated Lie algebra.[5] In the general case however, such local parametrisations do not exist, or at least are impractical on a computational level. Moreover, neither of these approaches extend to dissipation inequalities.

Conservative integrators for the Kepler problem

The energy- and momentum-stable method of LaBudde & Greenspan (LB–G) [LG74] considered in Subsubsection 6.2.1.1 was first proposed in 1974. This defines a modification to the implicit midpoint (IM) method that preserves the conservation of both linear[6] and angular momentum for general many-body systems with pairwise attractive forces, while further conserving energy.

[5] See [HLW06, Chap. IV.6–IV.8] for a further discussion of geometric numerical integration on Lie groups.

[6] Note, the momentum was not an invariant in our Kepler example in Subsubsection 6.2.1.1 and therefore naturally was not conserved by the LB–G method.

We note in passing that the Kepler system may be transformed to simple harmonic motion, such that each invariant becomes quadratic, by applying an appropriate transformation (Levi–Civita for $d = 2$, Kustaanheimo–Stiefel for $d = 3$). Any quadratic invariant–conserving scheme, such as Gauss methods, will hence conserve each invariant of the transformed problem [MN02; MN04; Koz07]. This approach is restricted to cases where such a transformation can be found.

Energy- and entropy-stable GENERIC numerical integrators

Early considerations of the introduction of dissipative effects into Poisson systems date back to Grmela [Grm84], Kaufman [Kau84] and Morrison [Mor84a; Mor84b] in 1984, with the latter interpreting it is as an entropy dissipation. In the language of traditional Hamiltonian mechanics, formulated through Poisson brackets and Hamiltonians, these systems were originally referred to by Morrison [Mor86] in 1986 as metriplectic, interpreting the entropy as a Casimir of the original Hamiltonian system and the dissipative term as induced by a metric; this terminology remains more commonplace in the plasma physics and Hamiltonian mechanics literature and communities. The title of GENERIC, associated with the more tensorial formulation, was not introduced until 11 years later by Grmela & Öttinger [GÖ97; ÖG97]. However, we favour this terminology throughout this work for the same reason we consider the Poisson system (6.3) not through a Poisson bracket, but through a Poisson matrix: the tensorial formulation of GENERIC explicitly uses gradients of the energy and entropy, thus making the procedure of our framework clearer when we introduce the corresponding AVs.

In 2009, Romero [Rom09] proposed an extension of the discrete gradient method of McLachlan, Quispel & Robidoux [MQR99] to GENERIC ODEs. Discrete gradients for both ∇H and ∇S were introduced, as well as approximations to the Poisson and friction matrices B and D, defined to satisfy the compatibility conditions (6.79) against these discrete gradients; choosing as the discrete gradient the mean-value discrete gradient of Harten, Lax & van Leer [HLL83] yields the lowest-order case ($S = 1$) of our scheme (6.81) when $\mathcal{I}_n$ is the midpoint rule.

In a recent work of Lombardi & Pagliantini [LP24] the authors analyse general Poisson PDEs with a gradient-descent term, i.e. equations of the form (7.19), in particular when M is an inner product independent of u (Assumption 3.9). Similar to

our scheme (7.24) they introduce AVs $\tilde{w}_H \approx w_H(u)$, $\tilde{w}_S \approx w_S(u)$, which they define using discrete gradients, identical to our AVs at lowest order in time ($S = 1$). These AVs replace $w_H(u)$, $w_S(u)$ in the primal equation, preserving some of the PDE's structure. In contrast to our scheme (7.24) however, their focus extends beyond GENERIC systems; they therefore do not modify the operators B, D as we do in (7.24a) through Assumption 7.4, and so do not necessarily preserve the conservation and dissipation structures on the discrete level.

Structure-preserving integrators for the Boltzmann equation

Much of the literature here has generally focused more on the conservation of moments, including the energy, than the generation of entropy, the motivation lying in how efficient algorithms for approximating the collision operator generally rely on spectral methods (see the recent work of Pareschi & Rey [PR22]); such methods are generally unreliable in terms of enforcing bound constraints, in particular positivity, implying the entropy may become ill-defined, and the preservation of its generation nonsensical. This spectral approach traces back to the early work of Bobylev [Bob75] with Pareschi, Perthame & Russo [PP96; PR00a; PR00b] introducing the first Fourier–Galerkin type spectral methods around 20 years later.[7]

When considering further structures, other techniques are preferred, including discrete-velocity methods. See Schneider *et al.* [CRS92; RS94; BPS95] or Buet [Bue96] for an introduction to these ideas, and Bobylev & Vinerean [BV08] for a discussion of invariant preservation.

Structure-preserving integrators for the compressible Navier–Stokes equations

As stated above, the mass-, momentum-, energy- and entropy-stable scheme for the compressible NS equations (7.59) presented in Section 7.3 is, to the best of our knowledge, novel. However, SP methods for the compressible NS equations have been well studied, in particular in the context of finite-volume methods [FLM20]. The concept of entropy-stable methods was introduced and analysed by Tadmor for the barotropic Euler equations [Tad87; Tad03; Tad16]. In the context of discontinuous

[7] See also Mouhot, Pareschi & Filbet [MP06; FMP06] for a discussion of the numerical implementation of these schemes.

Galerkin (DG) methods, AVs mirroring ours for entropy stability were introduced by Parsani *et al.* [Par+16] and Chan [Cha18] preserving the generation of entropy in the semidiscrete case, discretised in space only; see also [Cha20; Cha25]. See Chen & Shu [CS20] for a review on different types of entropy-stable schemes for the compressible NS equations.

The root-density variable present in our scheme (7.59) was employed by Morinishi [Mor10] and Halpern & Waltz [HW18]; see also Nordström [Nor22] where similar forms are used in the context of entropy generation. Without introducing the root-density variable, similar decompositions of the advective term to that of (7.46b) were used be Kennedy & Gruber [KG08], to improve the skew-symmetry of their finite difference discretisation and hence improve its energy conservation properties, and in recent work by Brunk, Jüngel & Lukáčová-Medvid'ová [BJL25], to ensure energy stability in their FE discretisation of a variable-density system related to the NS equations.

The problem of constructing energy-stable FE schemes for compressible flow was considered in the ideal limit by Gawlik & Gay-Balmaz [GG21b] building on a Lagrangian interpretation of the NS equations [Pav+11]. Dissipative terms were later introduced by the same authors [GG21a].

5.2 Overview

In Chapter 6, we begin by demonstrating applications of the framework to certain classes of geometric ODEs. The first of these systems is a simple Poisson or gradient-descent system—both systems have a very similar form, making the application of our framework identical in either case—for which we are able to preserve energy stability, i.e. we are able to use our framework to construct an integrator of arbitrary order that is discretely energy-conserving in the former case and energy-dissipative in the latter. We lead on from here to consider a general conservative ODE system, with potentially multiple invariants; through a general variational form of such equations in terms of a certain alternating form, we are able simultaneously to preserve all known conservation laws. For numerical demonstration, we consider the 2D Kepler problem (a maximally superintegrable system in 4 dimensions with 3 invariants) and the Kovalevskaya top (a superintegrable system in 6 dimensions with

4 invariants); in each case we demonstrate the qualitative improvements offered by the SP approach, as well as a convergence test in the case of the former. We conclude this chapter by considering ODEs deriving from the GENERIC formalism, for which we are able to preserve both the energy conservation and entropy generation structures discretely; as an example application, we consider a classical dissipative thermodynamic system: a C-cylinder engine. We further consider both the existence of unique solutions to and convergence of our SP ODE schemes, establishing general analytic results that we apply to each of these integrators.

In Chapter 7, we proceed by demonstrating applications of the framework within PDEs. Similar to the ODE discussion, we begin by considering a simple Poisson or gradient-descent system, applying our framework in either case to preserve the energy stability; we apply our general scheme to the BBM equation, a Hamiltonian PDE, observing that our energy-stable integrator offers notably greater preservation of the system's dynamics (in particular the long-term stability of an example soliton) over the equivalent Gauss method. We continue then to PDEs deriving from the GENERIC formalism, similarly using our framework to derive a general FE integrator that preserves both energy conservation and entropy generation; our example application here is the Boltzmann equation, arguably the most classical GENERIC PDE, for which we are able to construct an integrator that is necessarily both energy- and entropy-stable. We conclude by considering the compressible NS equations, which conserve mass, momentum, energy and entropy in the Euler/inviscid case, and conserve mass, momentum and energy while necessarily generating entropy otherwise. Through our framework, we are able to construct novel integrators that preserve all of these structures in either case; we demonstrate our scheme with numerical simulations of a shockwave (in the viscous regime) and an adiabatic perturbation (in the Euler regime).

"Euler's method."

— Katherine G. Johnson (Taraji P. Henson)

[...]

"But that's ancient."

— Paul Stafford (James J. 'Jim' Parsons) [Mel16]

6

ODEs: Poisson, gradient-descent, conservative & GENERIC

Contents

This chapter begins the discussion of applications by considering geometric ODE systems, and the discrete preservation of their structures. To use more familiar notation, we shall write the general SP integrator (3.27) for ODEs in the form: find $(\mathbf{x}, (\tilde{\mathbf{w}}_p)) \in \mathbb{X}_n \times \dot{\mathbb{X}}_n^P$ such that

$$\mathcal{I}_n[M(\mathbf{x}; \dot{\mathbf{x}}, \mathbf{y})] = \mathcal{I}_n[\tilde{F}(\mathbf{x}, (\tilde{\mathbf{w}}_p); \mathbf{y})], \qquad \mathcal{I}_n[M(\mathbf{x}; \mathbf{y}_p, \tilde{\mathbf{w}}_p)] = \int_{T_n} \nabla Q_p(\mathbf{x})^\top \mathbf{y}_p, \tag{6.1}$$

for all $(\mathbf{y}, (\mathbf{y}_p)) \in \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n^P$, where $\nabla Q_p(\mathbf{x})^\top$ denotes the tranpose of $\nabla Q_p(\mathbf{x})$, and $\mathbb{X}_n$ is defined as in (3.10) with $\mathbb{U} = \mathbb{R}^d$,

$$\mathbb{X}_n := \Big\{ \mathbf{x} \in \mathbb{P}_S(T_n)^d : \mathbf{x}(t_n) \text{ satisfies known initial data} \Big\}. \tag{6.2}$$

We note throughout this chapter that, as discussed in Subsection 4.2.1, the AVs introduced in the ODE can in general be eliminated on the computational level; they serve primarily as a tool for the construction and analysis of the discretisations.

The rest of this chapter proceeds as follows. In Section 6.1, we begin by considering the application of our framework to a simple Poisson or gradient-descent system (6.3), deriving a discretisation that is energy-stable, i.e. conserves the energy over timesteps in the former case and dissipates it in the latter. We present some further preliminary uniqueness and convergence results for general SP ODE integrators derived from our framework, in particular for our energy-stable integrator for Poisson and gradient-descent systems. In Section 6.2, we extend the idea to general conservative systems with arbitrarily many invariants, constructing novel discretisations that conserve all invariants. We consider as example applications the Kepler problem, and the Kovalevskaya top, and reapply the analytic results of the previous section to discuss again the uniqueness and convergence of solutions. In Section 6.3, we conclude by discussing systems of ODEs deriving from the GENERIC formalism [GÖ97; ÖG97], i.e. a certain class of systems with both a conserved energy and generated entropy. We construct novel integrators that preserve both these structures, applying the results scheme to a model dissipative thermodynamic system, and again discuss the uniqueness and convergence of solutions.

6.1 Poisson & gradient-descent systems: energy stability

As an introductory ODE example, we begin by considering both a general Poisson and gradient-descent system, as the theory and application of our framework (Algorithm 3.5) is similar in either case. These systems exhibit an energy $H(\mathbf{x}) \in \mathbb{R}$ that is either conserved or dissipated (or at least non-increasing) respectively.

Denote the general ODE in $\mathbf{x} : \mathbb{R}_+ \to \mathbb{R}^d$,

$$\dot{\mathbf{x}} = B(\mathbf{x})\nabla H(\mathbf{x}), \tag{6.3}$$

with IC $\mathbf{x}(0) = \mathbf{x}_0$, where $B(\mathbf{x}) \in \mathbb{R}^{d\times d}$ is either skew-symmetric, in the case of a general Poisson system, or negative semidefinite, in the case of a general gradient-descent system. In the Poisson case, this may identify with a Hamiltonian system, with H denoting the Hamiltonian and B encoding the Poisson bracket; in particular

if the system is written in canonical coordinates $\mathbf{x} = [\mathbf{p}, \mathbf{q}]$, B is simply the constant symplectic matrix

$$B(\mathbf{x}) = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}, \tag{6.4}$$

where I is the $d/2$-dimensional identity matrix. By testing (6.3) against ∇H, we see H is either conserved or dissipated in the exact solution. We apply our framework to construct an integrator for (6.3) that preserves this structure.

Application of framework (Algorithm 3.5)

A. Taking $\mathbb{U} \coloneqq \mathbb{R}^d$, we define $\mathbb{X}$ as in (3.3) with $\mathbb{U} = \mathbb{R}^d$,

$$\mathbb{X} \coloneqq \left\{ \mathbf{x} \in C^1(\mathbb{R}_+)^d : \mathbf{x}(0) \text{ satisfies known initial data} \right\}. \tag{6.5}$$

We then arrive at our semidiscrete problem: find $\mathbf{x} \in \mathbb{X}$ such that

$$M(\dot{\mathbf{x}}, \mathbf{y}) = F(\mathbf{x}; \mathbf{y}) \tag{6.6}$$

at all times $t \in \mathbb{R}_+$ and for all $\mathbf{y} \in \mathbb{U} = \mathbb{R}^d$, where M, F are defined

$$M(\dot{\mathbf{x}}, \mathbf{y}) \coloneqq \mathbf{y}^\top \dot{\mathbf{x}}, \qquad F(\mathbf{x}; \mathbf{y}) \coloneqq \mathbf{y}^\top B(\mathbf{x}) \nabla H(\mathbf{x}). \tag{6.7}$$

B. Over the timestep T_n, this is cast into a fully discrete form using our choice of $\mathcal{I}_n$: find $\mathbf{x} \in \mathbb{X}_n$, for $\mathbb{X}_n$ defined as in (6.2), such that

$$\mathcal{I}_n[\mathbf{y}^\top \dot{\mathbf{x}}] = \mathcal{I}_n[\mathbf{y}^\top B(\mathbf{x}) \nabla H(\mathbf{x})], \tag{6.8}$$

for all $\mathbf{y} \in \dot{\mathbb{X}}_n$.

C. Considering the evolution of H, since M is simply the ℓ^2 inner product, the associated test function for the conservation of H is simply ∇H.

D. We accordingly introduce an AV $\widetilde{\nabla H} \in \dot{\mathbb{X}}_n$, approximating $\nabla H(\mathbf{x})$, and defined as in (3.19) such that

$$\mathcal{I}_n[\widetilde{\nabla H}^\top \mathbf{y}_H] = \int_{T_n} \nabla H(\mathbf{x})^\top \mathbf{y}_H, \tag{6.9}$$

for all $\mathbf{y}_H \in \dot{\mathbb{X}}_n$.

E. We introduce $\widetilde{\nabla H}$ into the definition of F as

$$\tilde{F}(\mathbf{x}, \widetilde{\nabla H}; \mathbf{y}) \coloneqq \mathbf{y}^\top B(\mathbf{x}) \widetilde{\nabla H}. \tag{6.10}$$

Clearly this coincides with the definition of F when $\widetilde{\nabla H} = \nabla H$, while ensuring either $\tilde{F} = 0$ (in the Poisson case) or $\tilde{F} \leq 0$ (in the gradient-descent case) when $\mathbf{y} = \widetilde{\nabla H}$.

F. The final SP scheme is then as follows: find $(\mathbf{x}, \widetilde{\nabla H}) \in \mathbb{X}_n \times \dot{\mathbb{X}}_n$ such that

$$\mathcal{I}_n[\mathbf{y}^\top \dot{\mathbf{x}}] = \mathcal{I}_n[\mathbf{y}^\top B(\mathbf{x}) \widetilde{\nabla H}], \qquad \mathcal{I}_n[\widetilde{\nabla H}^\top \mathbf{y}_H] = \int_{T_n} \nabla H(\mathbf{x})^\top \mathbf{y}_H, \tag{6.11}$$

for all $(\mathbf{y}, \mathbf{y}_H) \in \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n$.

Theorem 6.1 (Energy stability of the Poisson & gradient-descent ODE integrator)**.** *The integrator* (6.11) *is energy-stable, with*

$$H(\mathbf{x}(t_{n+1})) - H(\mathbf{x}(t_n)) = \begin{cases} = 0, & \textit{Poisson}, \\ \leq 0, & \textit{gradient-descent}. \end{cases} \tag{6.12}$$

Proof. While this result holds by Theorem 3.4, we confirm the result here for sake of example. By considering respectively $\mathbf{y}_H = \dot{\mathbf{x}}$ and $\mathbf{y} = \widetilde{\nabla H}$ in (6.11),

$$\begin{aligned} H(\mathbf{x}(t_{n+1})) - H(\mathbf{x}(t_n)) &= \int_{T_n} \dot{H} = \int_{T_n} \nabla H^\top \dot{\mathbf{x}} \\ &= \mathcal{I}_n\left[\widetilde{\nabla H}^\top \dot{\mathbf{x}}\right] = \mathcal{I}_n\left[\widetilde{\nabla H}^\top B(\mathbf{x}) \widetilde{\nabla H}\right] \begin{cases} = 0, & \text{Poisson}, \\ \leq 0, & \text{gradient-descent}, \end{cases} \end{aligned} \tag{6.13}$$

with the final result holding by either the skew-symmetry or negative semidefiniteness of $B(\mathbf{x})$, and the sign-preserving property of $\mathcal{I}_n$. □

6.1.1 Analysis: uniqueness & convergence

While in Section 3.3 we detailed some existence and uniqueness results for the incompressible NS integrator (3.28) alongside results for general AD systems (Assumption 3.11) these results are of little use for the analysis of the general energy-stable integrator (6.11). We restrict our attention here instead to general SP discretisations of ODE systems (6.1) derived from our framework (Algorithm 3.5). In such a case, we are able to prove results relating both to the uniqueness of solutions, and convergence on refinement of the timestep Δt_n. We will revisit these results to show uniqueness and convergence results of SP integrators for conservative ODEs and systems deriving from the GENERIC formalism Subsections 6.2.3 & 6.3.2.

Unlike the Picard linearisation (Definition 3.19) used in the uniqueness proof in Subsection 3.3.4, our proof of uniqueness in the ODE case makes use of a certain Picard–Lindelöf linearisation (Definition 6.2) since the preservation of energy estimates on the linearised level is less important in the analysis of ODEs. This definition has the benefit of increased generality. Under relatively loose regularity conditions, we are able to show this linearisation is a contraction on sufficiently small timesteps Δt_n, implying the existence of unique solutions by the CMT.

The rest of this subsection proceeds as follows. In Subsubsection 6.1.1.1, we show that under certain relatively loose regularity conditions on the discretised system, there exist unique solutions to our SP ODE integrators (6.1) on sufficiently small timesteps Δt_n. In Subsubsection 6.1.1.2, we show that under certain conditions, convergence of order S is guaranteed.[1]

6.1.1.1 Uniqueness

We begin by discussing the unique existence of solutions to the general SP ODE integrator (6.1). Similarly to Subsection 3.3.4, our proof strategy relies on the CMT, however differs from that of Subsection 3.3.4 in the choice of linearisation; whereas Subsection 3.3.4 employs a Picard linearisation (Definition 3.19) we employ a so-called Picard–Lindelöf linearisation (Definition 6.2). We derive a certain lemma (Lemma 6.3) from the CMT, showing the existence to certain sufficiently regular nonlinear problems with sufficiently small nonlinear terms. Through a simple corollary (Corollary 6.4) our final uniqueness result holds in Theorem 3.26 for sufficiently small timesteps Δt_n.

Definition 6.2 (Picard–Lindelöf linearisation)**.** *Let $m \in \mathbb{N}$ denote the iteration index. On a given timestep T_n, suppose $\mathbf{x}_m \in \mathbb{X}_n$ is given. Find the AV iterates $(\tilde{\mathbf{w}}_{p,m+1}) \in \dot{\mathbb{X}}_n^P$ such that*

$$\mathcal{I}_n[M(\mathbf{x}_m; \mathbf{y}_p, \tilde{\mathbf{w}}_{p,m+1})] = \int_{T_n} {Q_p}'(\mathbf{x}_m; \mathbf{y}_p), \tag{6.14a}$$

for all $(\mathbf{y}_p) \in \dot{\mathbb{X}}_n^P$. With this, find $\mathbf{x}_{m+1} \in \mathbb{X}_n$ such that

$$\mathcal{I}_n[M(\mathbf{x}_m; \dot{\mathbf{x}}_{m+1}, \mathbf{y})] = \mathcal{I}_n\Big[\tilde{F}(\mathbf{x}_m, (\tilde{\mathbf{w}}_{p,m+1}); \mathbf{y})\Big], \tag{6.14b}$$

[1] Note that, while we are only able to show convergence of order S, numerical experiments appear to indicate an order of $2S$ at time grid points, i.e. at (t_n), assuming one starts with a method of order $2S$ and relevant regularity conditions hold, aligning with known convergence rates for Gauss methods and CPG. This is discussed further in Remark 6.7

for all $\mathbf{y} \in \dot{\mathbb{X}}_n$. *When it is well-defined, the map* $\mathbf{x}_m \mapsto \mathbf{x}_{m+1}$ *is referred to as the Picard–Lindelöf linearisation.*

A simple criterion for the well-definedness of the Picard–Lindelöf linearisation is that $M(\mathbf{x}; \cdot, \cdot)$ is non-singular for all $\mathbf{x} \in \mathbb{R}^d$; in fact, we shall henceforth assume Assumption 3.9, that $M(\mathbf{x}; \cdot, \cdot)$ is both independent of $\mathbf{x}$ and defines an inner product on $\mathbb{R}^d$. Now to apply the CMT to this linearisation, we rely on the following lemma.

Lemma 6.3 (CMT lemma)**.** *For a given map* $T : \mathbb{R}^n \to \mathbb{R}^n$ *and* $\varepsilon > 0$, *consider the nonlinear problem*

$$\mathbf{X} = \varepsilon T(\mathbf{X}). \tag{6.15}$$

Both the following uniqueness results hold.

- *When* $T \in \mathrm{Lip}(\mathbb{R}^n)$ *(i.e. globally Lipschitz) there exists a unique* $\mathbf{X}$ *such that for sufficiently small* ε (6.15) *holds.*
- *With* $T|_{B_s} \in \mathrm{Lip}_{\mathrm{loc}}(B_s)$ *(i.e. locally Lipschitz on an open ball* B_s *around* $\mathbf{0}$ *of radius* s*) over any closed ball* $\overline{B}_r \subset B_s$ *(similarly of radius* $r < s$*) there exists a unique* $\mathbf{X} \in \overline{B}_r$ *such that for sufficiently small* ε (6.15) *holds.*

Proof. Considering the map $\mathbf{X} \mapsto \varepsilon T(\mathbf{X})$, the result in the globally Lipschitz case holds immediately from the CMT for $\varepsilon < \frac{1}{L}$ where $L \geq 0$ is the Lipschitz constant of T.

For the locally Lipschitz case, it suffices to show that for all $\overline{B}_r \subset B_s$ there exists sufficiently small ε such that the map $\mathbf{X} \mapsto \varepsilon T(\mathbf{X})$ maps from $\overline{B}_r$ to $\overline{B}_r$, and that such a map is a contraction. Consider ε bounded above by

$$\varepsilon < \frac{r}{\|T(\mathbf{0})\| + rL_r}, \tag{6.16}$$

where $L_r \geq 0$ is the Lipschitz constant of T on $\overline{B}_r$, and $\|\cdot\|$ denotes the ℓ^2 norm. Checking that the image of T on $\overline{B}_r$ lies in $\overline{B}_r$,

$$\begin{aligned} \|\varepsilon T(\mathbf{X})\| &\leq \varepsilon(\|T(\mathbf{X}) - T(\mathbf{0})\| + \|T(\mathbf{0})\|) \\ &\leq \varepsilon(L_r\|\mathbf{X}\| + \|T(\mathbf{0})\|) \leq \varepsilon(rL_r + \|T(\mathbf{0})\|) < r, \end{aligned} \tag{6.17}$$

where the first inequality holds by the triangle inequality, the second by the Lipschitz bound, the third since $\mathbf{x}_n \in \overline{B}_r$, and the last by the bound on ε (6.16). The rest follows immediately from the CMT, since (6.16) ensures $\varepsilon L_r < 1$. □

To apply Lemma 6.3 to show the existence of solutions to (6.1), we rely on the following simple corollary.

Corollary 6.4 (CMT corollary). *For given maps $T : (\mathbb{R}^n)^{P+1} \to \mathbb{R}^n$ and $(S_p : \mathbb{R}^n \to \mathbb{R}^n)_{p=1}^P$, where $S_p(\mathbf{0}) = \mathbf{0}$ for each p, and $\varepsilon > 0$, consider the nonlinear problem*

$$\mathbf{X} = \varepsilon T\Big(\mathbf{X}, (\tilde{\mathbf{W}}_p)_{p=1}^P\Big), \qquad \tilde{\mathbf{W}}_p = S_p(\mathbf{X}) \quad \forall p. \tag{6.18}$$

Both the following uniqueness results hold:

- *When T, (S_p) are globally Lipschitz, there exists a unique $\mathbf{X}$ such that for sufficiently small ε (6.18) holds.*
- *With T, (S_p) locally Lipschitz on neighbourhoods of $\mathbf{0}$, over a sufficiently small closed ball $\overline{B}_r$ there exists a unique $\mathbf{X} \in \overline{B}_r$ such that for sufficiently small ε (6.18) holds.*

Proof. The nonlinear problem (6.18) may be equivalently written as

$$\mathbf{X} = \varepsilon T\Big(\mathbf{X}, (S_p(\mathbf{X}))_{p=1}^P\Big). \tag{6.19}$$

In the case where T, (S_p) are globally Lipschitz, the map $\mathbf{X} \mapsto T(\mathbf{X}, (S_p(\mathbf{X})))$ is globally Lipschitz by composition; in the case where T, (S_p) are locally Lipschitz, we see the map $\mathbf{X} \mapsto T(\mathbf{X}, (S_p(\mathbf{X})))$ is similarly locally Lipschitz on a sufficiently small neighbourhood of $\mathbf{0}$ by composition, noting each $S_p(\mathbf{0}) = \mathbf{0}$. Corollary 6.4 therefore holds by Lemma 6.3. □

With Corollary 6.4 established, we may lastly show that, under certain regularity conditions, solutions to a general SP ODE discretisation exist uniquely on sufficiently small timesteps Δt_n.

Theorem 6.5 (Uniqueness of solutions: ODEs). *Assume Assumption 3.9. Then both the following results hold:*

- *Assume each QoI Q_p is globally Lipschitz differentiable in $\mathbf{x}$, and $\tilde{F}$ is globally Lipschitz in $\mathbf{x}$, $(\tilde{\mathbf{w}}_p)$. There then exists a unique solution to (6.1) on sufficiently small timesteps Δt_n.*

- *Assume each QoI Q_p is locally Lipschitz differentiable on a neighbourhood of $\mathbf{x}(t_n)$, and $\tilde{F}$ is locally Lipschitz on neighbourhoods of $\mathbf{x}(t_n)$, $\mathbf{w}_p(\mathbf{x}(t_n))$. For sufficiently small $\delta > 0$, there then exists a unique solution to (6.1) satisfying $\sup_{T_n} \|\mathbf{x} - \mathbf{x}(t_n)\| \leq \delta$ on sufficiently small timesteps Δt_n.*

Proof. To make our argument clearer as $\Delta t_n \to 0$, let us write $t = t_n + \tau \Delta t_n$ and reparametrise in $\tau \in [0, 1]$. Define $\mathbf{X}$, $(\tilde{\mathbf{W}}_p)_{p=1}^P$, $\mathbf{Y}$, $(\mathbf{Y}_p)_{p=1}^P$ implicitly as

$$\mathbf{x}(t) = \mathbf{x}(t_n) + \mathbf{X}_n(\tau), \qquad \tilde{\mathbf{w}}_p(t) = \mathbf{w}_p(\mathbf{x}(t_n)) + \tilde{\mathbf{W}}_{p,n}(\tau), \tag{6.20a}$$

$$\mathbf{y}(t) = \mathbf{Y}(\tau), \qquad \mathbf{y}_p(t) = \mathbf{Y}_p(\tau). \tag{6.20b}$$

Letting $\mathcal{J}$ denote the quadrature rule $\mathcal{I}_n$ over the interval $[0, 1]$, after some rearranging the general SP ODE integrator (6.1) can then be written in the form

$$\mathcal{J}[M(\mathbf{X}_n', \mathbf{Y})] = \Delta t_n \mathcal{J}[\tilde{F}(\mathbf{x}(t_n) + \mathbf{X}_n, (\mathbf{w}_p(\mathbf{x}(t_n)) + \tilde{\mathbf{W}}_{p,n}); \mathbf{Y})], \tag{6.21a}$$

$$\mathcal{J}[M(\mathbf{Y}_p, \tilde{\mathbf{W}}_{p,n})] = \int_0^1 [\nabla Q_p(\mathbf{x}(t_n) + \mathbf{X}_n) - \nabla Q_p(\mathbf{x}(t_n))]^\top \mathbf{Y}_p. \tag{6.21b}$$

We may show the existence of unique solutions to (6.21) by Corollary 6.4. Since $\mathcal{J}[M(\cdot, \cdot)]$ defines an inner product similarly to Lemma 3.10, define for each p the map $S_p : \mathbb{P}_{s-1}(0, 1) \to \mathbb{P}_{s-1}(0, 1)$ (where $\mathbb{P}_{s-1}(0, 1)$ denotes the space of degree-$(s-1)$ polynomials on the interval $(0, 1)$) implicitly such that

$$\mathcal{J}[M(\mathbf{Y}_p, S_p(\mathbf{X}_n'))] = \int_0^1 [\nabla Q_p(\mathbf{x}(t_n) + \mathbf{X}_n) - \nabla Q_p(\mathbf{x}(t_n))]^\top \mathbf{Y}_p \tag{6.22a}$$

for all $\mathbf{Y}_p \in \dot{\mathbb{X}}_n$; clearly $S_p(\mathbf{0}) = \mathbf{0}$, while standard arguments alongside the norm equivalences from Lemma 3.8 show S_p is either globally Lipschitz, or locally Lipschitz in a neighbourhood of $\mathbf{0}$ when Q_p is locally Lipschitz differentiable on a neighbourhood of $\mathbf{x}(t_n)$. Define similarly the map $T : (\mathbb{P}_{s-1}(0, 1))^{P+1} \to \mathbb{P}_{s-1}(0, 1)$ implicitly such that

$$\mathcal{J}[M(T(\mathbf{X}_n', (\tilde{\mathbf{W}}_{p,n})), \mathbf{Y})] = \mathcal{J}[\tilde{F}(\mathbf{x}(t_n) + \mathbf{X}_n, (\mathbf{w}_p(\mathbf{x}(t_n)) + \tilde{\mathbf{W}}_{p,n}); \mathbf{Y})] \tag{6.22b}$$

for all $\mathbf{Y} \in \dot{\mathbb{X}}_n$; standard arguments again show T is either globally Lipschitz, or locally Lipschitz in a neighbourhood of $\mathbf{0}$ when $\tilde{F}$ is locally Lipschitz differentiable on a neighbourhood of $\mathbf{x}(t_n), (\mathbf{w}_p(\mathbf{x}(t_n)))$. With T, (S_p) as defined in (6.22), the general SP ODE integrator (6.21) can be written in the form (6.18); since (6.21) is equivalent to (6.1), Theorem 6.5 then holds by Corollary 6.4. □

Remark 6.6 (Sufficiency of continuous differentiability)**.** *Similarly to Remark 3.67, it is sufficient in the latter case to show $\tilde{F}$ is continuously differentiable on a neighbourhood of $\mathbf{x}(t_n)$, $\mathbf{w}_p(\mathbf{x}(t_n))$, and Q_p is twice continuously differentiable on a neighbourhood of $\mathbf{x}(t_n)$.*

Example (Poisson & gradient-descent systems)

The energy-stable integrator for Poisson & gradient-descent systems (6.11) satisfies both the following uniqueness results:

- Assume B and H are globally Lipschitz and Lipschitz differentiable respectively. There then exists a unique solution on sufficiently small timesteps Δt_n.
- Assume B and H are locally Lipschitz and Lipschitz differentiable in neighbourhoods of $\mathbf{x}(t_n)$ respectively. For sufficiently small $\delta > 0$, there then exists a unique solution satisfying $\sup_{T_n} \|\mathbf{x} - \mathbf{x}(t_n)\| \leq \delta$ on sufficiently small timesteps Δt_n.

6.1.1.2 Convergence

We discuss now the convergence of solutions to the general SP ODE integrator (6.1) to an exact solution $\mathbf{X} : \mathbb{R}_+ \to \mathbb{R}^d$. As noted in Section 3.3, the same approach can be used to demonstrate a certain convergence under refinement of the timestep in the PDE case, however this only necessarily implies convergence to a certain semidiscrete solution, discretised in space only; as it stands, the argument here does not demonstrate convergence under simultaneous refinement of the mesh size, and so the results here are only truly directly useful in the ODE case, where no such spatial refinement concerns are present.

Our proof of convergence relies on two technical lemmas: a bound on the approximation error in a certain projection from bounded functions on T_n to $\dot{\mathbb{X}}_n$ (Lemma 6.10) and a resulting error estimate for the AVs $(\tilde{\mathbf{w}}_p)$ (Lemma 6.12). We may then derive an error estimate for $\mathbf{x}$ (i.e. a bound for $\mathbf{x} - \mathbf{X}$) on T_n through the triangle inequality via an intermediate function $u_\dagger$ (6.42) (Theorem 6.13); this is our main result, demonstrating the method is of order S provided sufficient

regularity conditions hold (Assumption 6.11) We conclude with an extension of this bound to $\mathbb{R}_+$ (Corollary 6.14).

Remark 6.7 (Superconvergence). *Theorem 6.13 and Corollary 6.14 imply only a convergence of order S. Numerical experiments however appear to indicate an order of $2S$ at time grid points, i.e. at (t_n), assuming one starts with a method of order $2S$ and relevant regularity conditions hold. (See the convergence rates in Fig. 6.3.) This result would align with the known convergence rates for Gauss methods and CPG, and we expect it to hold here also.*

To discuss the convergence of solutions to our discretisation, it is necessary that these solutions exist uniquely. Sufficient conditions for this unique existence over sufficiently small Δt_n are given by Theorem 6.5 in the case of globally Lipschitz differentiable (Q_p) and globally Lipschitz $\tilde{F}$; in the case where these Lipschitz conditions hold locally, we simply choose the solution $\mathbf{x}$ that minimises $\sup_{T_n} \|\mathbf{x} - \mathbf{x}(t_n)\|$.

Continuing to the proof of convergence, we define first the approximation error $\varepsilon_n(\mathbf{Z})$, quantifying the distance, under the supremum norm, of a certain bounded function $\mathbf{Z} : T_n \to \mathbb{R}^d$ from $\dot{\mathbb{X}}_n$.

Definition 6.8 (Approximation error). *For bounded $\mathbf{Z} : T_n \to \mathbb{R}^d$, define the $\dot{\mathbb{X}}_n$-approximation error $\varepsilon_n(\mathbf{Z})$ of $\mathbf{Z}$,*

$$\varepsilon_n(\mathbf{Z}) \coloneqq \inf_{\mathbf{z} \in \dot{\mathbb{X}}_n} \left\{ \sup_{T_n} \|\mathbf{Z} - \mathbf{z}\| \right\}. \tag{6.23}$$

Since we will be considering global refinement of the timesteps (Δt_n), we define a maximum approximation error $\varepsilon_{\max}(\mathbf{Z}) \coloneqq \max_n \varepsilon_n(\mathbf{Z})$. By taking a truncated Taylor expansion in time of $\mathbf{Z}$, we arrive at the following Jackson-type theorem (see [Jac11] or [Tre20, Chap. 7]), in which similarly to Section 3.3 we write $a \lesssim b$ if there exists a constant $C > 0$ dependent only on S such that $a \leq Cb$.

Lemma 6.9 (Jackson-type theorem). *Suppose $\mathbf{Z} : \mathbb{R}_+ \to \mathbb{R}^d$ has a globally Lipschitz $(S-1)$-times time derivative $\partial_t^{S-1}\mathbf{Z} \in \mathrm{Lip}(\mathbb{R}_+)^d$. Then $\varepsilon_{\max}(\mathbf{Z})$ satisfies the bound $\varepsilon_{\max}(\mathbf{Z}) \lesssim \Delta t_n^S$.*

Under Assumption 3.9, consider the projection under $\mathcal{I}_n[M(\cdot,\cdot)]$ of bounded $\mathbf{Z} : \mathbb{R}_+ \to \mathbb{R}^d$ into $\dot{\mathbb{X}}_n$. We bound the approximation error of this projection by $\varepsilon_n(\mathbf{Z})$.

Lemma 6.10 (General bound on approximation error)**.** *Assuming Assumption 3.9, suppose* $\mathbf{z} \in \dot{\mathbb{X}}_n$, $\mathbf{Z} : \mathbb{R}_+ \to \mathbb{R}^d$, *and* $G \in \dot{\mathbb{X}}_n^*$ *satisfy*

$$\mathcal{I}_n[M(\mathbf{z} - \mathbf{Z}, \mathbf{y})] = G(\mathbf{y}), \tag{6.24}$$

for all $\mathbf{y} \in \dot{\mathbb{X}}_n$. *Then* $\mathbf{z} - \mathbf{Z}$ *satisfies the bound*

$$\sup_{T_n} \|\mathbf{z} - \mathbf{Z}\| \lesssim \varepsilon_n(\mathbf{Z}) + \frac{1}{\Delta t_n} \|G\|_*, \tag{6.25}$$

where $\|G\|_*$ *is defined*

$$\|G\|_* \coloneqq \sup_{\mathbf{y} \in \dot{\mathbb{X}}_n : \sup_{T_n} \|\mathbf{y}\| = 1} |G(\mathbf{y})|, \tag{6.26}$$

the supremum dual norm on $\dot{\mathbb{X}}_n^*$.

Proof. For all $\mathbf{w} \in \dot{\mathbb{X}}_n$,

$$\mathcal{I}_n[M(\mathbf{z} - \mathbf{w}, \mathbf{y})] = \mathcal{I}_n[M(\mathbf{Z} - \mathbf{w}, \mathbf{y})] + G(\mathbf{y}). \tag{6.27}$$

We first bound $\mathbf{z} - \mathbf{w}$ by taking $\mathbf{y} = \mathbf{z} - \mathbf{w}$,

$$\mathcal{I}_n[M(\mathbf{z} - \mathbf{w}, \mathbf{z} - \mathbf{w})] = \mathcal{I}_n[M(\mathbf{Z} - \mathbf{w}, \mathbf{z} - \mathbf{w})] + G(\mathbf{z} - \mathbf{w}) \tag{6.28a}$$

$$\Delta t_n \sup_{T_n} \|\mathbf{z} - \mathbf{w}\|^2 \lesssim \Delta t_n \sup_{T_n} |M(\mathbf{Z} - \mathbf{w}, \mathbf{z} - \mathbf{w})| + |G(\mathbf{z} - \mathbf{w})| \tag{6.28b}$$

$$\lesssim \Delta t_n \sup_{T_n} \|\mathbf{Z} - \mathbf{w}\| \sup_{T_n} \|\mathbf{z} - \mathbf{w}\| + \|G\|_* \sup_{T_n} \|\mathbf{z} - \mathbf{w}\| \tag{6.28c}$$

$$\sup_{T_n} \|\mathbf{z} - \mathbf{w}\| \lesssim \sup_{T_n} \|\mathbf{Z} - \mathbf{w}\| + \frac{1}{\Delta t_n} \|G\|_*, \tag{6.28d}$$

where the first inequality holds by Assumption 3.9 and (3.37), and the second holds by the continuity of M and definition of $\|G\|_*$ (6.26). We then bound $\mathbf{z} - \mathbf{Z}$ through $\mathbf{w}$ using the triangle inequality,

$$\sup_{T_n} \|\mathbf{z} - \mathbf{Z}\| \le \sup_{T_n} \|\mathbf{z} - \mathbf{w}\| + \sup_{T_n} \|\mathbf{Z} - \mathbf{w}\| \lesssim \sup_{T_n} \|\mathbf{Z} - \mathbf{w}\| + \frac{1}{\Delta t_n} \|G\|_*, \tag{6.29}$$

and since $\mathbf{w}$ is arbitrary, we have

$$\sup_{T_n} \|\mathbf{z} - \mathbf{Z}\| \lesssim \varepsilon_n(\mathbf{Z}) + \frac{1}{\Delta t_n} \|G\|_*. \tag{6.30}$$

□

With the projection error (Lemma 6.10) established, we are able to prove our first convergence result. To do so, we define the following regularity assumptions, under which we are able to prove convergence of our scheme. Since we will be considering global refinement of the timesteps (Δt_n), we define a maximum timestep $\Delta t_{\max} := \max_n \Delta t_n$.

Assumption 6.11 (Convergence regularity). *Assume Assumption 3.9, and that $\mathcal{I}_n$ has order at least $2S-1$, i.e.*

$$\left| \int_{T_n} \phi - \mathcal{I}_n[\phi] \right| \lesssim \sup_{T_n} \|\phi^{(2S-1)}\| \Delta t_n^{2S}, \tag{6.31}$$

for all $(2S-1)$-times continuously differentiable $\phi : T_n \to \mathbb{R}$. Assume further that the change $\dot{\mathbf{X}}$ in the exact trajectory $\mathbf{X}$ is uniformly $(2S-2)$-times continuously differentiable, i.e. the r-th derivative $\partial_t^r \mathbf{X}$ is bounded (at least up to a chosen final time) for all $r = 1, \cdots, 2S-2$. Assume further that each QoI Q_p is $2S$-times continuously differentiable, and that (at least) one of the following holds:

- *The RHS $\tilde{F}$ is globally Lipschitz, and that for each p the r-th derivative $\nabla^{\otimes r} Q_p$ is bounded for all $r = 0, \cdots, 2S$.*
- *The RHS $\tilde{F}$ is globally Lipschitz, the QoI Q_p is globally Lipschitz differentiable for each p, and the exact trajectory $\mathbf{X}$ is contained within a compact set.*
- *The RHS $\tilde{F}$ is locally Lipschitz, the exact trajectory $\mathbf{X}$ is contained within a compact set, and for sufficiently small timesteps $\Delta t_{\max}$ all discrete trajectories $\mathbf{x}$ are contained within some compact set.*[2]

Example (Poisson & gradient-descent systems)

For the energy-stable integrator (6.11) the only condition in Assumption 6.11 that simplifies is that of the Lipschitz regularity on $\tilde{F}$: we require that the matrix $B(\mathbf{x})$ is either globally or locally Lipschitz in $\mathbf{x}$ respectively.

We are thus able to prove the following result, bounding the error between the discrete AV $\tilde{\mathbf{w}}_p$ and the associated test function $\mathbf{w}_p(\mathbf{X})$.

[2] In certain cases, these compactness results on the trajectories may be proven through certain conservation and dissipation structures preserved to the discrete level by the discretisation's SP properties.

Lemma 6.12 (Bound on error of AVs). *Assuming Assumption 6.11. Then for each p, $\tilde{\mathbf{w}}_p - \mathbf{w}_p(\mathbf{X})$ satisfies the bound*

$$\sup_{T_n} \|\tilde{\mathbf{w}}_p - \mathbf{w}_p(\mathbf{X})\| \lesssim \sup_{T_n} \|\mathbf{x} - \mathbf{X}\| + \Delta t_n^S. \tag{6.32}$$

Proof. For each p, the discrete AV $\tilde{\mathbf{w}}_p$ is related to its continuous counterpart $\mathbf{w}_p(\mathbf{X})$ by the identity

$$\begin{aligned}
&\mathcal{I}_n[M(\tilde{\mathbf{w}}_p - \mathbf{w}_p(\mathbf{X}), \mathbf{y}_p)] \\
&\qquad = \int_{T_n} \nabla Q_p(\mathbf{x})^\top \mathbf{y}_p - \mathcal{I}_n[\nabla Q_p(\mathbf{X})^\top \mathbf{y}_p] && \text{(6.33a)} \\
&\qquad = \int_{T_n} [\nabla Q_p(\mathbf{x}) - \nabla Q_p(\mathbf{X})]^\top \mathbf{y}_p \\
&\qquad\qquad + \left(\int_{T_n} \nabla Q_p(\mathbf{X})^\top \mathbf{y}_p - \mathcal{I}_n[\nabla Q_p(\mathbf{X})^\top \mathbf{y}_p] \right) =: G_p(\mathbf{y}_p) && \text{(6.33b)}
\end{aligned}$$

for all $\mathbf{y}_p \in \dot{\mathbb{X}}_n$. We may use Lemma 6.10 to bound $\sup_{T_n} \|\tilde{\mathbf{w}}_p - \mathbf{w}_p(\mathbf{X})\|$ by bounding the RHS G_p.

For the former term $\int_{T_n} [\nabla Q_p(\mathbf{x}) - \nabla Q_p(\mathbf{X})]^\top \mathbf{y}_p$ we consider the Lipschitz differentiability condition of Q_p. Since Q_p is at least twice differentiable, it at least locally Lipschitz differentiable; by assumption, Q_p is then either globally Lipschitz differentiable, or there exists a compact set containing both $\mathbf{X}$ and, for sufficiently small $\Delta t_{\max}$, all trajectories $\mathbf{x}$. We may therefore bound $\|\nabla Q_p(\mathbf{x}) - \nabla Q_p(\mathbf{X})\| \lesssim \|\mathbf{x} - \mathbf{X}\|$ over the whole trajectory, implying

$$\begin{aligned}
\left| \int_{T_n} [\nabla Q_p(\mathbf{x}) - \nabla Q_p(\mathbf{X})]^\top \mathbf{y}_p \right| &\leq \sup_{T_n} \|\nabla Q_p(\mathbf{x}) - \nabla Q_p(\mathbf{X})\| \sup_{T_n} \|\mathbf{y}_p\| \Delta t_n \\
&\lesssim \sup_{T_n} \|\mathbf{x} - \mathbf{X}\| \sup_{T_n} \|\mathbf{y}_p\| \Delta t_n
\end{aligned} \tag{6.34}$$

For the latter term $\int_{T_n} \nabla Q_p(\mathbf{X})^\top \mathbf{y}_p - \mathcal{I}_n[\nabla Q_p(\mathbf{X})^\top \mathbf{y}_p]$ we rely on the order of $\mathcal{I}_n$ (6.31),

$$\left| \int_{T_n} \nabla Q_p(\mathbf{X})^\top \mathbf{y}_p - \mathcal{I}_n[\nabla Q_p(\mathbf{X})^\top \mathbf{y}_p] \right| \lesssim \sup_{T_n} \left| \partial_t^{2S-1} [\nabla Q_p(\mathbf{X})^\top \mathbf{y}_p] \right| \Delta t_n^{2S}. \tag{6.35}$$

Each term in the expansion of $\partial_t^{2S-1}[\nabla Q_p(\mathbf{X})^\top \mathbf{y}_p]$ takes the form of an inner product between the following: an r-times spatial derivative derivative $\nabla^{\otimes r} Q_p(\mathbf{X})$ for some $r \leq 2S$; up to $2S-1$ derivatives $\partial_t^r \mathbf{X}$ for certain $1 \leq r \leq 2S-1$; a derivative $\partial_t^r \mathbf{y}_p$ for some $r \leq S-1$ (since for $r \geq S$, $\partial_t^r \mathbf{y}_p = 0$). Each of these terms may be bounded over T_n, in turn allowing us to bound $\sup_{T_n} \left| \partial_t^{2S-1} [\nabla Q_p(\mathbf{X})^\top \mathbf{v}_p] \right|$: we

may bound $\|\nabla^{\otimes r} Q_p(\mathbf{X})\|$ through the $2S$-times continuous differentiability of Q_p, either through the compactly contained trajectory or directly through the explicitly bounded derivatives, depending on the case in question; we may bound each $\|\partial_t^r \mathbf{X}\|$ similarly through the explicitly bounded derivatives; we may bound $\|\partial_t^r \mathbf{y}_p\|$ (in finite dimensions) by some multiple of $\Delta t_n^{-r} \sup_{T_n} \|\mathbf{y}_p\|$, implying a uniform bound of $\Delta t_n^{1-S} \sup_{T_n} \|\mathbf{y}_p\|$. Thus, we may bound

$$\left| \int_{T_n} \nabla Q_p(\mathbf{X})^\top \mathbf{y}_p - \mathcal{I}_n[\nabla Q_p(\mathbf{X})^\top \mathbf{y}_p] \right| \lesssim \sup_{T_n} \|\mathbf{y}_p\| \Delta t_n^{S+1}. \tag{6.36}$$

We may then bound G_p by (6.34, 6.36),

$$\frac{1}{\Delta t_n} \|G_p\|_* \lesssim \sup_{T_n} \|\mathbf{x} - \mathbf{X}\| + \Delta t_n^S, \tag{6.37}$$

implying the bound

$$\begin{aligned} \sup_{T_n} \|\tilde{\mathbf{w}}_p - \mathbf{w}_p(\mathbf{X})\| &\lesssim \varepsilon_n(\mathbf{w}_p(\mathbf{X})) + \sup_{T_n} \|\mathbf{x} - \mathbf{X}\| + \Delta t_n^S \\ &\leq \varepsilon_{\max}(\mathbf{w}_p(\mathbf{X})) + \sup_{T_n} \|\mathbf{x} - \mathbf{X}\| + \Delta t_n^S \end{aligned} \tag{6.38}$$

by Lemma 6.10. To bound $\varepsilon_{\max}(\mathbf{w}_p(\mathbf{X}))$ we consider Lemma 6.9. Each term in the expansion of $\partial_t^{S-1}[\mathbf{w}_p(\mathbf{X})]$ takes the form of an inner product between the following: a derivative $\nabla^{\otimes r} \mathbf{w}_p(\mathbf{X})$ for some $r \leq S-1$; up to $S-1$ derivatives $\partial_t^r \mathbf{X}$ for certain $1 \leq r \leq S-1$. We may see each of these terms is Lipschitz over $\mathbb{R}_+$: the term $\nabla^{\otimes r} \mathbf{w}_p(\mathbf{X})$ is Lipschitz for $r \leq S-1$ as $\mathbf{X}$ is Lipschitz by the assumed continuity of $\dot{\mathbf{X}}$, while $\mathbf{w}_p$ inherits the regularity of ∇Q_p, including either its global $(S-1)$-times Lipschitz differentiability, or its local $(S-1)$-times Lipschitz differentiability with the compactly contained trajectory $\mathbf{X}$; the term $\partial_t^r \mathbf{X}$ is Lipschitz for $1 \leq r \leq S-1$ by the assumed regularity of $\dot{\mathbf{X}}$. Thus, for each n, by (6.38) and Lemma 6.9,

$$\sup_{T_n} \|\tilde{\mathbf{w}}_p - \mathbf{w}_p(\mathbf{X})\| \lesssim \varepsilon_{\max}(\mathbf{w}_p(\mathbf{X})) + \sup_{T_n} \|\mathbf{x} - \mathbf{X}\| + \Delta t_n^S \lesssim \sup_{T_n} \|\mathbf{x} - \mathbf{X}\| + \Delta t_n^S. \tag{6.39}$$

□

Example (Poisson & gradient-descent systems)

Assuming Assumption 6.11 holds, the AV $\widetilde{\nabla H}$ in the integrator (6.11) satisfies

the error estimate

$$\sup_{T_n} \|\widetilde{\nabla H} - \nabla H(\mathbf{X})\| \lesssim \sup_{T_n} \|\mathbf{x} - \mathbf{X}\| + \Delta t_n^S. \tag{6.40}$$

With the error on the AVs established, we arrive at our first convergence results.

Theorem 6.13 (Local convergence)**.** *Assume Assumption 6.11. On sufficiently small timesteps $\Delta t_{\max}$, $\mathbf{x} - \mathbf{X}$ is then bounded by*

$$\sup_{T_n} \|\mathbf{x} - \mathbf{X}\| \lesssim \|\mathbf{x}(t_n) - \mathbf{X}(t_n)\| + \Delta t_n^{S+1}. \tag{6.41}$$

Proof. Intermediate between $\mathbf{x}$ and $\mathbf{X}$, define also $\dot{\mathbf{x}}_\dagger \in \dot{\mathbb{X}}_n$ [3] such that

$$\mathcal{I}_n[M(\dot{\mathbf{x}}_\dagger, \mathbf{y}_\dagger)] = \mathcal{I}_n[\tilde{F}(\mathbf{X}, (\mathbf{w}_p(\mathbf{X})); \mathbf{y}_\dagger)] \tag{6.42}$$

for all $\mathbf{y}_\dagger \in \dot{\mathbb{X}}_n$; the variable $\dot{\mathbf{x}}_\dagger$ necessarily exists by Lemma 3.10. For $t \in T_n$,

$$\mathbf{x}(t) - \mathbf{X}(t) = [\mathbf{x}(t_n) - \mathbf{X}(t_n)] + \int_{t_n}^t [\dot{\mathbf{x}} - \dot{\mathbf{x}}_\dagger] + \int_{t_n}^t [\dot{\mathbf{x}}_\dagger - \dot{\mathbf{X}}] \tag{6.43a}$$

$$\sup_{T_n} \|\mathbf{x} - \mathbf{X}\| \leq \|\mathbf{x}(t_n) - \mathbf{X}(t_n)\| + \int_{T_n} \|\dot{\mathbf{x}} - \dot{\mathbf{x}}_\dagger\| + \int_{T_n} \|\dot{\mathbf{x}}_\dagger - \dot{\mathbf{X}}\|. \tag{6.43b}$$

We shall bound the latter two terms.

For the former term $\int_{T_n} \|\dot{\mathbf{x}} - \dot{\mathbf{x}}_\dagger\|$, note $\dot{\mathbf{x}}$ and $\dot{\mathbf{x}}_\dagger$ are related by the identity

$$\mathcal{I}_n[M(\dot{\mathbf{x}} - \dot{\mathbf{x}}_\dagger, \mathbf{y})] = \mathcal{I}_n[\tilde{F}(\mathbf{x}, (\tilde{\mathbf{w}}_p); \mathbf{y}) - \tilde{F}(\mathbf{X}, (\mathbf{w}_p(\mathbf{X})); \mathbf{y})], \tag{6.44}$$

for all $\mathbf{y} \in \dot{\mathbb{X}}_n$. We seek to bound the RHS term by the Lipschitz condition on $\tilde{F}$. In the case of global Lipschitz regularity, this is immediate. In the case of local Lipschitz regularity, we require $\mathbf{x}$, $(\tilde{\mathbf{w}}_p)$, $\mathbf{X}$, $(\mathbf{w}_p(\mathbf{X}))$ to lie in a uniform compact domain for sufficiently small $\Delta t_{\max}$; this is true for $\mathbf{x}$ and $\mathbf{X}$ by assumption, for $(\mathbf{w}_p(\mathbf{X}))$ by the assumed continuity of $\mathbf{w}_p$ (inherited from the assumed continuity of (∇Q_p)), and for $(\tilde{\mathbf{w}}_p)$ by the bound (6.32). Thus, we may bound

$$|\mathcal{I}_n[M(\dot{\mathbf{x}} - \dot{\mathbf{x}}_\dagger, \mathbf{y})]| = |\mathcal{I}_n[\tilde{F}(\mathbf{x}, (\tilde{\mathbf{w}}_p); \mathbf{y}) - \tilde{F}(\mathbf{X}, (\mathbf{w}_p(\mathbf{X})); \mathbf{y})]| \tag{6.45a}$$

$$\leq \Delta t_n \sup_{T_n} |\tilde{F}(\mathbf{x}, (\tilde{\mathbf{w}}_p); \mathbf{y}) - \tilde{F}(\mathbf{X}, (\mathbf{w}_p(\mathbf{X})); \mathbf{y})| \tag{6.45b}$$

$$\lesssim \Delta t_n \sup_{T_n} \{\max\{\|\mathbf{x} - \mathbf{X}\|, (\|\tilde{\mathbf{w}}_p - \mathbf{w}_p(\mathbf{X})\|)_p\}\|\mathbf{y}\|\} \tag{6.45c}$$

$$\lesssim \Delta t_n \left(\sup_{T_n} \|\mathbf{x} - \mathbf{X}\| + \Delta t_n^S\right) \sup_{T_n} \|\mathbf{y}\|, \tag{6.45d}$$

[3] Without the time derivative, $\mathbf{x}_\dagger$ is defined only up to a constant, however this is irrelevant for our proof as $\mathbf{x}_\dagger$ will appear only through its derivative.

where in the final inequality we use the error estimate (6.32). Taking $\mathbf{y} = \dot{\mathbf{x}} - \dot{\mathbf{x}}_\dagger$,

$$\mathcal{I}_n[\|\dot{\mathbf{x}} - \dot{\mathbf{x}}_\dagger\|^2] \lesssim \Delta t_n (\sup_{T_n} \|\mathbf{x} - \mathbf{X}\| + \Delta t_n^S) \sup_{T_n} \|\dot{\mathbf{x}} - \dot{\mathbf{x}}_\dagger\|. \tag{6.46a}$$

We may bound the LHS below by

$$\sup_{T_n} \|\dot{\mathbf{x}} - \dot{\mathbf{x}}_\dagger\| \int_{T_n} \|\dot{\mathbf{x}} - \dot{\mathbf{x}}_\dagger\| \leq \Delta t_n \sup_{T_n} \|\dot{\mathbf{x}} - \dot{\mathbf{x}}_\dagger\|^2 \lesssim \mathcal{I}_n[\|\dot{\mathbf{x}} - \dot{\mathbf{x}}_\dagger\|^2], \tag{6.47}$$

where the second inequality holds by (3.37b), implying ultimately that

$$\int_{T_n} \|\dot{\mathbf{x}} - \dot{\mathbf{x}}_\dagger\| \lesssim \Delta t_n (\sup_{T_n} \|\mathbf{x} - \mathbf{X}\| + \Delta t_n^S) \tag{6.48}$$

Substituting this bound into (6.43b),

$$\sup_{T_n} \|\mathbf{x} - \mathbf{X}\| \lesssim \|\mathbf{x}(t_n) - \mathbf{X}(t_n)\| + \Delta t_n (\sup_{T_n} \|\mathbf{x} - \mathbf{X}\| + \Delta t_n^S) + \int_{T_n} \|\dot{\mathbf{x}}_\dagger - \dot{\mathbf{X}}\|. \tag{6.49}$$

With a sufficiently small timestep, $\Delta t_n \sup_{T_n} \|\mathbf{x} - \mathbf{X}\| \lesssim \sup_{T_n} \|\mathbf{x} - \mathbf{X}\|$ with the same constant as (6.49), and can therefore be eliminated, giving

$$\sup_{T_n} \|\mathbf{x} - \mathbf{X}\| \lesssim \|\mathbf{x}(t_n) - \mathbf{X}(t_n)\| + \Delta t_n^{S+1} + \int_{T_n} \|\dot{\mathbf{x}}_\dagger - \dot{\mathbf{X}}\|. \tag{6.50}$$

Proceeding to the latter term $\int_{T_n} \|\dot{\mathbf{x}}_\dagger - \dot{\mathbf{X}}\|$,

$$\int_{T_n} \|\dot{\mathbf{x}}_\dagger - \dot{\mathbf{X}}\| \leq \Delta t_n \sup_{T_n} \|\dot{\mathbf{x}}_\dagger - \dot{\mathbf{X}}\|. \tag{6.51}$$

Since $\dot{\mathbf{x}}_\dagger$ and $\dot{\mathbf{X}}$ are related by the identity $\mathcal{I}_n[M(\dot{\mathbf{x}}_\dagger - \dot{\mathbf{X}}, \mathbf{y}_\dagger)] = 0$ for all $\mathbf{y}_\dagger \in \dot{\mathbb{X}}_n$,

$$\int_{T_n} \|\dot{\mathbf{x}}_\dagger - \dot{\mathbf{X}}\| \lesssim \Delta t_n \varepsilon_n(\dot{\mathbf{X}}) \lesssim \|\dot{\mathbf{X}}\|_{\mathrm{Lip}^{S-1}} \Delta t_n^{S+1}, \tag{6.52}$$

where in the first inequality we apply Lemma 6.10, and in the second we apply Lemma 6.9 through the assumed regularity of $\dot{\mathbf{X}}$. Substituting this bound into (6.50) we obtain the desired result (6.41). □

Example (Poisson & gradient-descent systems)

Assuming Assumption 6.11 holds, then for a sufficiently small local timestep Δt_n, the discrete solution $\mathbf{x}$ to (6.11) satisfies the error estimate

$$\sup_{T_n} \|\mathbf{x} - \mathbf{X}\| \lesssim \|\mathbf{x}(t_n) - \mathbf{X}(t_n)\| + \Delta t_n^{S+1}. \tag{6.53}$$

Iterating this bound over multiple timesteps, we obtain the following global convergence result.

Corollary 6.14 (Global convergence). *Assume Assumption 6.11. Then for sufficiently small* $\Delta t_{\max}$

$$\|\mathbf{x}(t) - \mathbf{X}(t)\| \lesssim (t + \Delta t_{\max})\Delta t_{\max}^S \tag{6.54}$$

for all $t \in \mathbb{R}_+$.

Proof. The proof is a standard exercise in induction on (6.41), bounding $\|\mathbf{x}(t_n) - \mathbf{X}(t_n)\|$ in (6.41) applied over T_n by $\sup_{T_n} \|\mathbf{x} - \mathbf{X}\|$ in (6.41) applied over T_{n-1}. One need only assert that each of the bounds holds uniformly over $\mathbb{R}_+$. □

Example (Poisson & gradient-descent systems)

Assuming Assumption 6.11 holds, then for a sufficiently small global timestep $\Delta t_{\max}$, the discrete solution $\mathbf{x}$ to (6.11) satisfies the error estimate

$$\|\mathbf{x}(t) - \mathbf{X}(t)\| \lesssim (t + \Delta t_{\max})\Delta t_{\max}^S \tag{6.55}$$

for all $t \in \mathbb{R}_+$.

6.2 General conservative systems: conservation of arbitrary invariants

We consider now general conservative ODE systems with arbitrarily many invariants of interest.

Let $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^d$ induce the general ODE system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}), \tag{6.56}$$

with IC $\mathbf{x}(0) = \mathbf{x}_0$. Suppose this system is conservative in $P(< d)$ independent invariants $(N_p : \mathbb{R}^d \to \mathbb{R})_{p=1}^P$, such that in the continuous case $\nabla N_p(\mathbf{x})^\top \mathbf{f}(\mathbf{x}) = 0$ for each $p = 1, \dots, P$. Analogous to the conservation of H in Section 6.1, each N_p can then be seen to be conserved over T_n by testing (6.56) with ∇N_p.

We plan to introduce AVs for each ∇N_p, and to use those AVs in the RHS of (6.56). However, as written, (6.56) does not appear to depend on each ∇N_p. Lemma 6.15 demonstrates that $\mathbf{f}$ may be rewritten to make its dependence on (∇N_p) explicit, thereby enabling the introduction of AVs. This result is fully constructive, relying

on the definition of a certain alternating form, i.e. a multilinear map $F : V^n \to \mathbb{R}$ over a vector space V such that $F[v_1, \dots, v_n] = 0$ whenever $v_i = v_j$ for some $i \neq j$. We denote the space of alternating n-forms over V by $\operatorname{Alt}^n V$, and define the alternatisation $\operatorname{Alt} F \in \operatorname{Alt}^n V$ of an n-multilinear map by

$$\operatorname{Alt} F[v_1, \dots, v_n] \coloneqq \sum_{\sigma \in S_n} \operatorname{sgn}_\sigma F[v_{\sigma_1}, \dots, v_{\sigma_n}], \tag{6.57}$$

where S_n denotes the permutation group of degree n, and $\operatorname{sgn}_\sigma \in \{\pm 1\}$ the sign of $\sigma \in S_n$ [Tu10].

Lemma 6.15 (Identification of alternating forms)**.** *For the general conservative system* (6.56) *there exists* $\tilde{F} : \mathbb{R}^d \to \operatorname{Alt}^{P+1} \mathbb{R}^d$ *such that* $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\mathbf{y}^\top \mathbf{f}(\mathbf{x}) = \tilde{F}(\mathbf{x})[\nabla N_1(\mathbf{x}), \dots, \nabla N_P(\mathbf{x}), \mathbf{y}]. \tag{6.58}$$

Proof. We demonstrate the existence of $\tilde{F}$ by construction. Through the independence of (N_p), the gradients (∇N_p) are linearly independent almost everywhere. Consequently, we may define a dual basis $(\mathbf{m}_q : \mathbb{R}^d \to \mathbb{R}^d)_{q=1}^P$ such that almost everywhere $\nabla N_p(\mathbf{x})^\top \mathbf{m}_q(\mathbf{x}) = \delta_{pq}$. For each $\mathbf{x} \in \mathbb{R}^d$, define then the multilinear map $\tilde{G}(\mathbf{x}) : (\mathbb{R}^d)^{P+1} \to \mathbb{R}$,

$$\tilde{G}(\mathbf{x})[\mathbf{n}_1, \dots, \mathbf{n}_P, \mathbf{y}] \coloneqq (\mathbf{n}_1^\top \mathbf{m}_1(\mathbf{x})) \cdots (\mathbf{n}_P^\top \mathbf{m}_P(\mathbf{x}))(\mathbf{y}^\top \mathbf{f}(\mathbf{x})). \tag{6.59}$$

We observe then by the orthogonality property $\nabla N_p(\mathbf{x})^\top \mathbf{m}_q(\mathbf{x}) = \delta_{pq}$ that

$$\tilde{G}(\mathbf{x})[\nabla N_1(\mathbf{x}), \dots, \nabla N_P(\mathbf{x}), \mathbf{y}] \coloneqq \mathbf{y}^\top \mathbf{f}(\mathbf{x}); \tag{6.60}$$

furthermore, by the orthogonality $\nabla N_p(\mathbf{x})^\top \mathbf{f}(\mathbf{x})$ (inherent in the conservation of N_p) this evaluates to 0 under any (non-trivial) permutation of the arguments. Now define $\tilde{F}(\mathbf{x}) \coloneqq \operatorname{Alt} \tilde{G}(\mathbf{x}) \in \operatorname{Alt}^{P+1} \mathbb{R}^d$ to be the alternatisation of $\tilde{G}(\mathbf{x})$ (6.57). This $\tilde{F}(\mathbf{x})$ is alternating for all arguments by construction, and coincides with $\tilde{G}(\mathbf{x})$ when evaluated at $[\nabla N_1(\mathbf{x}), \dots, \nabla N_P(\mathbf{x}), \mathbf{y}]$ for any $\mathbf{y}$ since all but the trivial permutation evaluate to zero in the alternatisation (6.57) Hence (6.58) holds. □

Since the proof here is constructive, one may potentially use it directly when seeking to define such an $\tilde{F}$. In simpler cases however, such an $\tilde{F}$ can often be found simply by inspection, as we will demonstrate for the Kepler problem in Subsection 6.2.1.

With Lemma 6.15 established and $\tilde{F}$ defined, we may apply our framework to construct an integrator for (6.56) that preserves all conservation laws.

Application of framework (Algorithm 3.5)

A. Taking $\mathbb{U} \coloneqq \mathbb{R}^d$, we again define $\mathbb{X}$ as in (6.5). We then arrive at our semidiscrete problem: find $\mathbf{x} \in \mathbb{X}$ such that

$$M(\dot{\mathbf{x}}, \mathbf{y}) = F(\mathbf{x}; \mathbf{y}) \tag{6.61}$$

at all times $t \in \mathbb{R}_+$ and for all $\mathbf{y} \in \mathbb{U} = \mathbb{R}^d$, where M, F are defined

$$M(\dot{\mathbf{x}}, \mathbf{y}) \coloneqq \mathbf{y}^\top \dot{\mathbf{x}}, \qquad F(\mathbf{x}; \mathbf{y}) \coloneqq \mathbf{y}^\top \mathbf{f}(\mathbf{x}). \tag{6.62}$$

B. Over the timestep T_n, this is cast into a fully discrete form using our choice of $\mathcal{I}_n$: find $\mathbf{x} \in \mathbb{X}_n$, for $\mathbb{X}_n$ defined again as in (6.2), such that

$$\mathcal{I}_n[M(\dot{\mathbf{x}}, \mathbf{y})] = \mathcal{I}_n[F(\mathbf{x}, \mathbf{y})], \tag{6.63}$$

for all $\mathbf{y} \in \dot{\mathbb{X}}_n$.

C. Considering the conservation of (N_p), since M is simply the ℓ^2 inner product, the associated test functions for the conservation of (N_p) are (∇N_p).

D. We introduce AVs $(\widetilde{\nabla N}_p) \in \dot{\mathbb{X}}_n^P$, approximating (∇N_p) and defined as in (3.19) such that

$$\mathcal{I}_n[\widetilde{\nabla N}_p^\top \mathbf{y}_p] = \int_{T_n} \nabla N_p(\mathbf{x})^\top \mathbf{y}_p, \tag{6.64}$$

for all $(\mathbf{y}_p)_{p=1}^P \in \dot{\mathbb{X}}_n^P$.

E. We may define $\tilde{F}$ as in Lemma 6.15, such that by (6.58) it coincides with F when each $\widetilde{\nabla N}_p = \nabla N_p$, and by the alternating property of $\tilde{F}(\mathbf{x})$ it preserves each of the conservation structures.

F. The final SP scheme is then as follows: find $(\mathbf{x}, (\widetilde{\nabla N}_p)) \in \mathbb{X}_n \times \dot{\mathbb{X}}_n^P$ such that

$$\mathcal{I}_n[\mathbf{y}^\top \dot{\mathbf{x}}] = \mathcal{I}_n[\tilde{F}(\mathbf{x})[\widetilde{\nabla N}_1, \dots, \widetilde{\nabla N}_P, \mathbf{y}]], \tag{6.65a}$$

$$\mathcal{I}_n[\widetilde{\nabla N}_p^\top \mathbf{y}_p] = \int_{T_n} \nabla N_p(\mathbf{x})^\top \mathbf{y}_p, \tag{6.65b}$$

for all $(\mathbf{y}, (\mathbf{y}_p)) \in \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n^P$.

Theorem 6.16 (Universal stability of the integrator for conservative systems)**.** *The integrator* (6.65) *is universally stable, with* $N_p(\mathbf{x}(t_{n+1})) = N_p(\mathbf{x}(t_n))$ *for all* p*.*

Proof. For each p, by considering respectively $\tilde{\mathbf{y}} = \dot{\mathbf{x}}$ and $\mathbf{y}_p = \widetilde{\nabla N}_p$ in (6.11),

$$\begin{aligned} N_p(\mathbf{x}(t_{n+1})) - N_p(\mathbf{x}(t_n)) = \int_{T_n} \dot{N}_p = \int_{T_n} \nabla N_p(\mathbf{x})^\top \dot{\mathbf{x}} \\ = \mathcal{I}_n\Big[\widetilde{\nabla N}_p^\top \dot{\mathbf{x}}\Big] = \mathcal{I}_n\Big[\tilde{F}(\mathbf{x})[\widetilde{\nabla N}_1, \dots, \widetilde{\nabla N}_P, \widetilde{\nabla N}_p]\Big] = 0, \end{aligned} \tag{6.66}$$

where the final equality holds by the alternating property of $\tilde{F}(\mathbf{x})$. □

6.2.1 The Kepler problem: energy, angular momentum & Runge–Lenz stability

As a numerical demonstration of the scheme (6.65) we discretise the (nondimensionalised) two-body Kepler problem,

$$\dot{\mathbf{x}} = \mathbf{v}, \qquad \dot{\mathbf{v}} = -\frac{1}{\|\mathbf{x}\|^3}\mathbf{x}, \tag{6.67}$$

for $\mathbf{x}, \mathbf{v} : \mathbb{R}_+ \to \mathbb{R}^d$ representing the position and velocity respectively, and $\|\cdot\|$ denoting the ℓ^2 norm. Trajectories of (6.67) preserve the energy H, angular momentum $\mathbf{L}$, and Runge–Lenz vector $\mathbf{A}$, defined

$$H(\mathbf{x},\mathbf{v}) \coloneqq \frac{1}{2}\|\mathbf{v}\|^2 - \frac{1}{\|\mathbf{x}\|}, \quad \mathbf{L}(\mathbf{x},\mathbf{v}) \coloneqq \mathbf{x}\times\mathbf{v}, \quad \mathbf{A}(\mathbf{x},\mathbf{v}) \coloneqq \mathbf{v}\times\mathbf{L}(\mathbf{x},\mathbf{v}) - \frac{1}{\|\mathbf{x}\|}\mathbf{x}, \tag{6.68}$$

where $\times$ denotes the cross product. Roughly speaking, H and $\mathbf{L}$ encode within them the shape of the orbit and the plane to which it is restricted, whereas the orientation of the orbit within that plane is encoded in $\mathbf{A}$ (see Taff [Taf85]). These invariants are not independent, as $\|\mathbf{A}\|^2 = 1 + 2H\|\mathbf{L}\|^2$, while $\mathbf{A}$ and $\mathbf{L}$ are necessarily perpendicular; these 3 invariants thus represent $2d-1$ independent constants of motion (for $d \in \{2,3\}$) the maximum possible number of conserved quantities (i.e. the system is maximally superintegrable).

For our numerical demonstration, we consider the two-dimensional case $d = 2$. In such a case, if H and $\mathbf{A} = (A_1, A_2)$ are conserved, then the scalar angular momentum L will be conserved automatically, since $\|\mathbf{A}\|^2 = 1 + 2HL^2$. We may therefore construct a fully constructive numerical integrator for $d = 2$ using our scheme (6.65) by conserving H, A_1, A_2, i.e. $P = 3$.

To apply (6.65) we must construct some $\tilde{F} : \mathbb{R}^{2\times 2} \to \operatorname{Alt}^4 \mathbb{R}^{2\times 2}$ satisfying the conditions of (6.58), i.e. such that $\forall (\mathbf{x}, \mathbf{v}), (\mathbf{y}, \mathbf{w}) \in \mathbb{R}^{2\times 2}$,

$$\mathbf{y}^\top \mathbf{v} - \mathbf{w}^\top \frac{1}{\|\mathbf{x}\|^3}\mathbf{x} = \tilde{F}\left(\begin{pmatrix}\mathbf{x}\\ \mathbf{v}\end{pmatrix}\right)\left[\begin{pmatrix}\nabla_{\mathbf{x}} H\\ \nabla_{\mathbf{v}} H\end{pmatrix}, \begin{pmatrix}\nabla_{\mathbf{x}} A_1\\ \nabla_{\mathbf{v}} A_1\end{pmatrix}, \begin{pmatrix}\nabla_{\mathbf{x}} A_2\\ \nabla_{\mathbf{v}} A_2\end{pmatrix}, \begin{pmatrix}\mathbf{y}\\ \mathbf{w}\end{pmatrix}\right], \tag{6.69}$$

where $\nabla_{\mathbf{x}}$, $\nabla_{\mathbf{v}}$ denote partial derivatives with respect to $\mathbf{x}$, $\mathbf{v}$ respectively. Instead of using the constructive proof in Lemma 6.15, we may more simply note the space $\operatorname{Alt}^4 \mathbb{R}^4$ merely has dimension 1; any alternating n-form in n dimensions is in fact some multiple of the determinant map on the n-by-n square matrix formed by the n argument vectors, allowing us to vastly reduce the space of potential maps $\tilde{F}$ to consider. Noting the gradients in our QoIs,

$$\nabla_{\mathbf{x}} H = \frac{1}{\|\mathbf{x}\|^3}\mathbf{x}, \qquad \nabla_{\mathbf{x}} \mathbf{A} = \frac{1}{\|\mathbf{x}\|^3}\mathbf{x}^{\otimes 2} - \mathbf{v}^{\otimes 2} + \left(\|\mathbf{v}\|^2 - \frac{1}{\|\mathbf{x}\|}\right) I, \tag{6.70a}$$

$$\nabla_{\mathbf{v}} H = \mathbf{v}, \qquad \nabla_{\mathbf{v}} \mathbf{A} = 2\mathbf{x} \otimes \mathbf{v} - \mathbf{v} \otimes \mathbf{x} - (\mathbf{x} \cdot \mathbf{v}) I, \tag{6.70b}$$

where $\otimes$ denotes the outer product and $\mathbf{x}^{\otimes 2} \coloneqq \mathbf{x} \otimes \mathbf{x}$, we may see by inspection that, for all $(\mathbf{x}, \mathbf{v}), (\mathbf{y}, \mathbf{w}) \in \mathbb{R}^{2\times 2}$,

$$\frac{1}{2L(\mathbf{x}, \mathbf{v}) H(\mathbf{x}, \mathbf{v})} \det \begin{bmatrix} \mathbf{y} & \nabla_{\mathbf{x}} H & \nabla_{\mathbf{x}} \mathbf{A}^\top \\ \mathbf{w} & \nabla_{\mathbf{v}} H & \nabla_{\mathbf{v}} \mathbf{A}^\top \end{bmatrix} = \mathbf{y}^\top \mathbf{v} - \mathbf{w}^\top \frac{1}{\|\mathbf{x}\|^3}\mathbf{x}, \tag{6.71}$$

where $\det : \mathbb{R}^{4\times 4} \to \mathbb{R}$ denotes the determinant map. We may define therefore $\tilde{F} : \mathbb{R}^{2\times 2} \to \operatorname{Alt}^4 \mathbb{R}^{2\times 2}$ as

$$\begin{aligned} &\tilde{F}\left(\begin{pmatrix}\mathbf{x}\\ \mathbf{v}\end{pmatrix}\right)\left[\begin{pmatrix}\widetilde{\nabla_{\mathbf{x}} H}\\ \widetilde{\nabla_{\mathbf{v}} H}\end{pmatrix}, \begin{pmatrix}\widetilde{\nabla_{\mathbf{x}} A_1}\\ \widetilde{\nabla_{\mathbf{v}} A_1}\end{pmatrix}, \begin{pmatrix}\widetilde{\nabla_{\mathbf{x}} A_2}\\ \widetilde{\nabla_{\mathbf{v}} A_2}\end{pmatrix}, \begin{pmatrix}\mathbf{y}\\ \mathbf{w}\end{pmatrix}\right] \\ &\qquad \coloneqq \frac{1}{2L(\mathbf{x}, \mathbf{v}) H(\mathbf{x}, \mathbf{v})} \det \begin{bmatrix} \mathbf{y} & \widetilde{\nabla_{\mathbf{x}} H} & \widetilde{\nabla_{\mathbf{x}} A_1} & \widetilde{\nabla_{\mathbf{x}} A_2} \\ \mathbf{w} & \widetilde{\nabla_{\mathbf{v}} H} & \widetilde{\nabla_{\mathbf{v}} A_1} & \widetilde{\nabla_{\mathbf{v}} A_2} \end{bmatrix}, \end{aligned} \tag{6.72}$$

which we see to be an alternating form by the alternating properties of $\det$, and we see to satisfy (6.58) by (6.71).

Through (6.65) we then arrive at our fully conservative integrator for the two-dimensional Kepler problem: find $((\mathbf{x}, \mathbf{v}), (\widetilde{\nabla_{\mathbf{x}} H}, \widetilde{\nabla_{\mathbf{v}} H}), (\widetilde{\nabla_{\mathbf{x}} \mathbf{A}}, \widetilde{\nabla_{\mathbf{v}} \mathbf{A}})) \in \mathbb{X}_n \times \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n^2$ (with $\mathbb{X}_n$ defined as in (6.2) for $d = 2 \times 2$) such that

$$\mathcal{I}_n\left[\mathbf{y}^\top \dot{\mathbf{x}} + \mathbf{w}^\top \dot{\mathbf{v}}\right] = \mathcal{I}_n\left[\frac{1}{2L(\mathbf{x}, \mathbf{v}) H(\mathbf{x}, \mathbf{v})} \det \begin{bmatrix} \mathbf{y} & \widetilde{\nabla_{\mathbf{x}} H} & \widetilde{\nabla_{\mathbf{x}} \mathbf{A}}^\top \\ \mathbf{w} & \widetilde{\nabla_{\mathbf{v}} H} & \widetilde{\nabla_{\mathbf{v}} \mathbf{A}}^\top \end{bmatrix}\right], \tag{6.73a}$$

$$\mathcal{I}_n\left[\widetilde{\nabla_{\mathbf{x}} H}^\top \mathbf{y}_H + \widetilde{\nabla_{\mathbf{v}} H}^\top \mathbf{w}_H\right] = \int_{T_n} \nabla H(\mathbf{x}, \mathbf{v})^\top \mathbf{y}_H + \nabla H(\mathbf{x}, \mathbf{v})^\top \mathbf{w}_H, \tag{6.73b}$$

$$\mathcal{I}_n\left[\operatorname{tr}\left(\widetilde{\nabla_{\mathbf{x}} \mathbf{A}} Y_A + \widetilde{\nabla_{\mathbf{v}} \mathbf{A}} W_A\right)\right] = \int_{T_n} \operatorname{tr}\left(\nabla_{\mathbf{x}} \mathbf{A}(\mathbf{x}, \mathbf{v}) Y_A + \nabla_{\mathbf{v}} \mathbf{A}(\mathbf{x}, \mathbf{v}) W_A\right), \tag{6.73c}$$

for all $((\mathbf{y}, \mathbf{w}), (\mathbf{y}_H, \mathbf{w}_H), (Y_A, W_A)) \in \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n^2$, where tr denotes the trace.

We test the fully conservative integrator (6.73) numerically on a standard set of ICs (inspired by [HLW06, Sec. I.2.3]) $\mathbf{x}(0) = (0.4, 0)$, $\mathbf{v}(0) = (0, 2)$. In each of these tests, we take $\mathcal{I}_n$ to be an S-stage GL quadrature method, such that the scheme (6.73) is an SP modification of a Gauss method of equal order.

6.2.1.1 Comparison test

To illustrate the qualitative benefits afforded by our fully conservative scheme, Fig. 6.1 simulates our Kepler IVP with timestep $\Delta t_n = 0.1$ and final time $t = 100$, using various classical 1-stage, 2$^{\text{nd}}$-order implicit geometric integrators: IM, the mean-value (or averaged-vector-field) discrete-gradient (MV–DG) method of McLachlan, Quispel & Robidous [MQR99], LB–G [LG74], and our scheme (6.73) at $S = 1$.

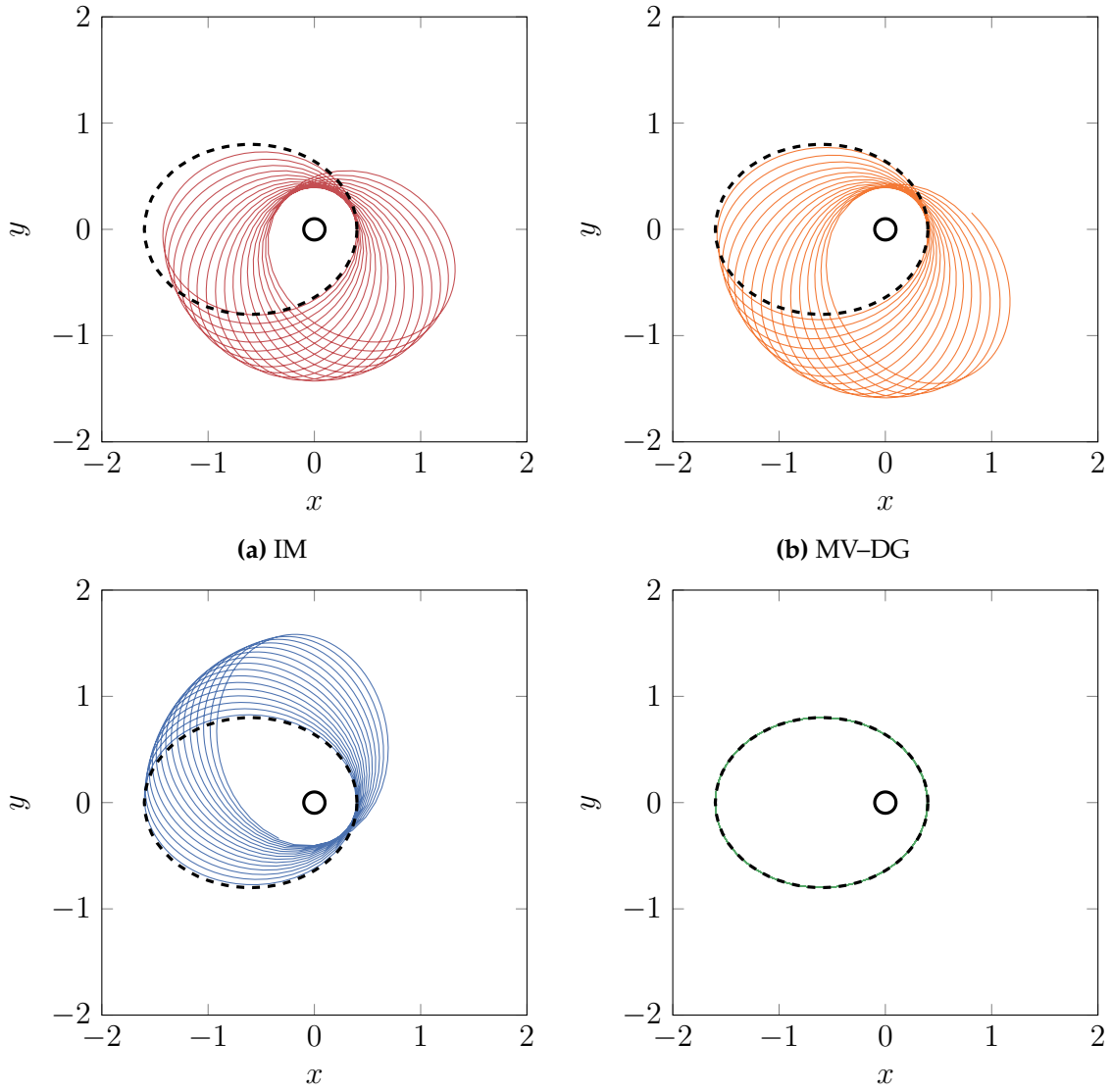
In those cases where they are not conserved, Fig. 6.2 shows the evolution of the invariants H, L, θ up to time $t = 50$, where $\theta \coloneqq \arg \mathbf{A}$.

As a symplectic method, IM conserves the quadratic invariant L (up to quadrature error, solver tolerances and machine precision) but neither H nor θ; it therefore conserves neither the orbit shape nor its orientation, since trajectories in the Kepler problem should be precession-free. IM gives unphysical solutions over this duration with this timestep. The MV–DG scheme conserves H, but neither L nor θ. LB–G conserves H and L by design, and so conserves the orbit shape, but not its orientation θ. In contrast, our scheme (6.73) conserves all three invariants, thereby restricting the discrete solution to the same ellipse traced out by the exact solution.

These results illustrate the potential importance of conserving invariants in Hamiltonian (and non-Hamiltonian) systems: while symplectic methods are likely preferable for e.g. capturing the statistical behaviour of chaotic systems, conservative discretisations may give more physically reasonable results for individual trajectories at coarser timesteps.

6.2.1.2 Convergence test

Fig. 6.3 shows the convergence of (6.73) for our model IVP through the error in the position of the orbital body after the true orbital period (2π for these ICs) at varying timesteps Δt_n and stages S. We observe convergence with rate $2S$ before round-off errors dominate.



(a) IM **(b)** MV–DG

(c) LB–G **(d)** Our proposed scheme

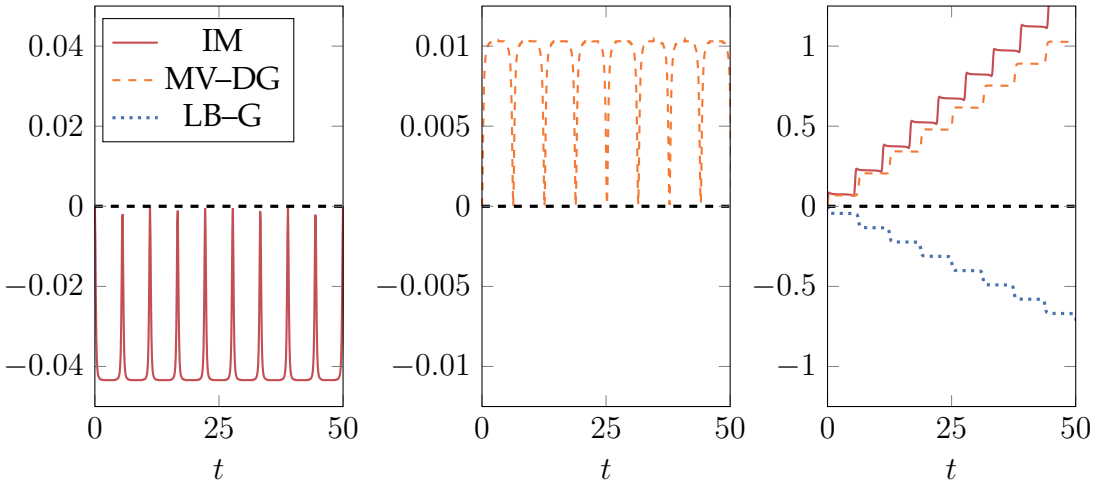
Figure 6.1: Trajectories of the Kepler problem.

6.2.2 **The Kovalevskaya top:** energy, angular momentum, normality & Kovalevskaya invariant stability

As a further example, we consider the (nondimensionalised) Kovalevskaya [Kov89] top,

$$\dot{\mathbf{n}} = \mathbf{n} \times J\mathbf{l}, \qquad \dot{\mathbf{l}} = \mathbf{n} \times \mathbf{e}_1 + \mathbf{l} \times J\mathbf{l}, \tag{6.74}$$

for $\mathbf{n}, \mathbf{l} : \mathbb{R}_+ \to \mathbb{R}^3$ representing the orientation vector (i.e. the z-components of the principal axes) and the angular momentum (i.e. the components of the angular momentum along those principal axes) respectively, $\times$ denoting the cross product,



(a) Energy, $H - H(0)$

(b) Angular momentum, $L - L(0)$

(c) Orbital angle, $\theta - \theta(0)$

Figure 6.2: Error in scalar invariants of the Kepler problem: H, L and θ.

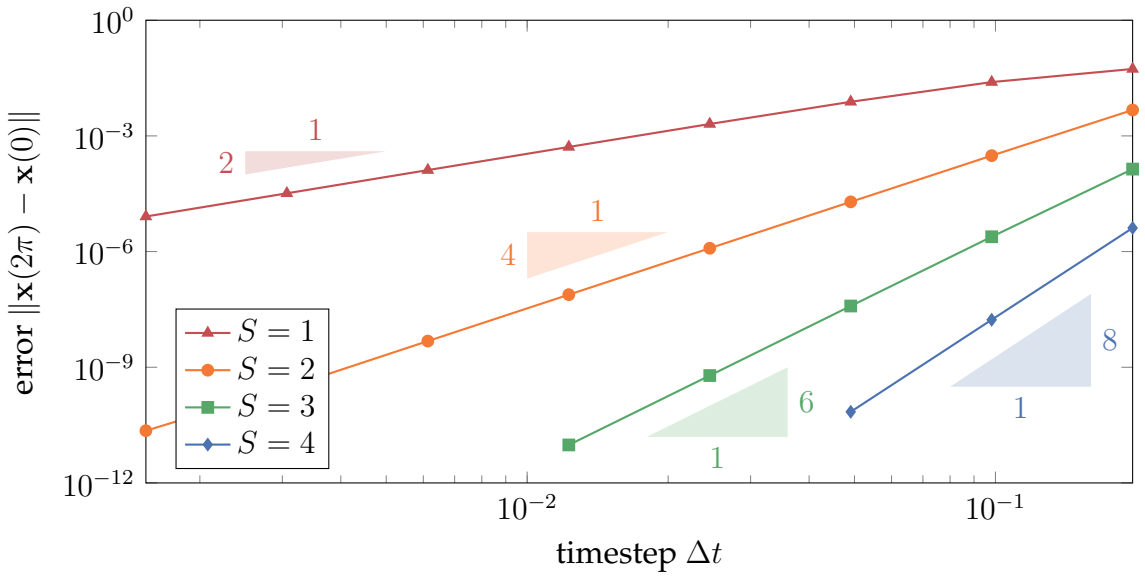


Figure 6.3: Error in the position of the orbital body at $t = 2\pi$ for varying timesteps $\Delta t \in 2\pi \cdot 2^k$, $k \in \{-5, \dots, -12\}$ and stages $S \in \{1, \dots, 4\}$. The convergence curve for $S \in \{3, 4\}$ flattens out at smaller timesteps due to round-off error and solver tolerances. Triangles demonstrate observed convergence rates of $2S$.

$\mathbf{e}_1$ denoting the basis vector $(1, 0, 0)$, and J denoting the matrix

$$J \coloneqq \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}. \tag{6.75}$$

Trajectories of this system have 4 invariants: the energy $H \coloneqq \frac{1}{2}\mathbf{l}^\top J \mathbf{l}$, the (square) norm of the orientation vector $\|\mathbf{n}\|^2$, the angular momentum in the z direction

$L = \mathbf{l} \cdot \mathbf{n}$, and the Kovalevskaya invariant $K = |\xi|^2$ where $\xi = (l_1 + il_2)^2 - 2(n_1 + in_2)$ (i is the imaginary unit). While H, $\|\mathbf{n}\|^2$ and $\mathbf{l} \cdot \mathbf{n}$ are quadratic, K is quartic.

Unlike the Kepler problem (Subsection 6.2.1) it is not immediately clear from inspection how one might define an $\tilde{F} : \mathbb{R}^6 \to \operatorname{Alt}^5 \mathbb{R}^6$ satisfying the conditions of Lemma 6.15 to conserve all 4 of these invariants; we therefore find such an $\tilde{F}$ using a construction similar to that used in the proof of Lemma 6.15. Define the multilinear map $\tilde{G}((\mathbf{n}, \mathbf{l})) : (\mathbb{R}^6)^5 \to \mathbb{R}$,

$$\begin{aligned}\tilde{G}\left(\begin{pmatrix}\mathbf{n}\\ \mathbf{l}\end{pmatrix}\right)\left[\begin{pmatrix}\mathbf{a}_1\\ \mathbf{b}_1\end{pmatrix}, \begin{pmatrix}\mathbf{a}_2\\ \mathbf{b}_2\end{pmatrix}, \begin{pmatrix}\mathbf{a}_3\\ \mathbf{b}_3\end{pmatrix}, \begin{pmatrix}\mathbf{a}_4\\ \mathbf{b}_4\end{pmatrix}, \begin{pmatrix}\mathbf{m}\\ \mathbf{k}\end{pmatrix}\right] \\ := \det[\mathbf{b}_1\ \mathbf{b}_2\ \mathbf{b}_3](\mathbf{n} \cdot \mathbf{a}_4)\left[\mathbf{m}^\top(\mathbf{n} \times J\mathbf{l}) + \mathbf{k}^\top(\mathbf{n} \times \mathbf{e}_1) + \mathbf{k}^\top(\mathbf{l} \times J\mathbf{l})\right]. \quad (6.76a)\end{aligned}$$

Considering the alternatisation $\operatorname{Alt} \tilde{G}((\mathbf{n}, \mathbf{l})) \in \operatorname{Alt}^5 \mathbb{R}^6$, we apply $\operatorname{Alt} \tilde{G}((\mathbf{n}, \mathbf{l}))$ to the gradients of the invariants H, K, L, $\frac{1}{2}\|\mathbf{n}\|^2$ respectively, we see

$$\begin{aligned}\operatorname{Alt} \tilde{G}\left(\begin{pmatrix}\mathbf{n}\\ \mathbf{l}\end{pmatrix}\right)\left[\begin{pmatrix}\mathbf{e}_1\\ J\mathbf{l}\end{pmatrix}, \begin{pmatrix}\nabla_{\mathbf{n}} K\\ \nabla_{\mathbf{l}} K\end{pmatrix}, \begin{pmatrix}\mathbf{l}\\ \mathbf{n}\end{pmatrix}, \begin{pmatrix}\mathbf{n}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{m}\\ \mathbf{k}\end{pmatrix}\right] \\ = 6 \det[J\mathbf{l}\ \nabla_{\mathbf{l}} K\ \mathbf{n}]\|\mathbf{n}\|^2\left[\mathbf{m}^\top(\mathbf{n} \times J\mathbf{l}) + \mathbf{k}^\top(\mathbf{n} \times \mathbf{e}_1) + \mathbf{k}^\top(\mathbf{l} \times J\mathbf{l})\right]. \quad (6.76b)\end{aligned}$$

We therefore define $\tilde{F}((\mathbf{n}, \mathbf{l})) \in \operatorname{Alt}^5 \mathbb{R}^6$,

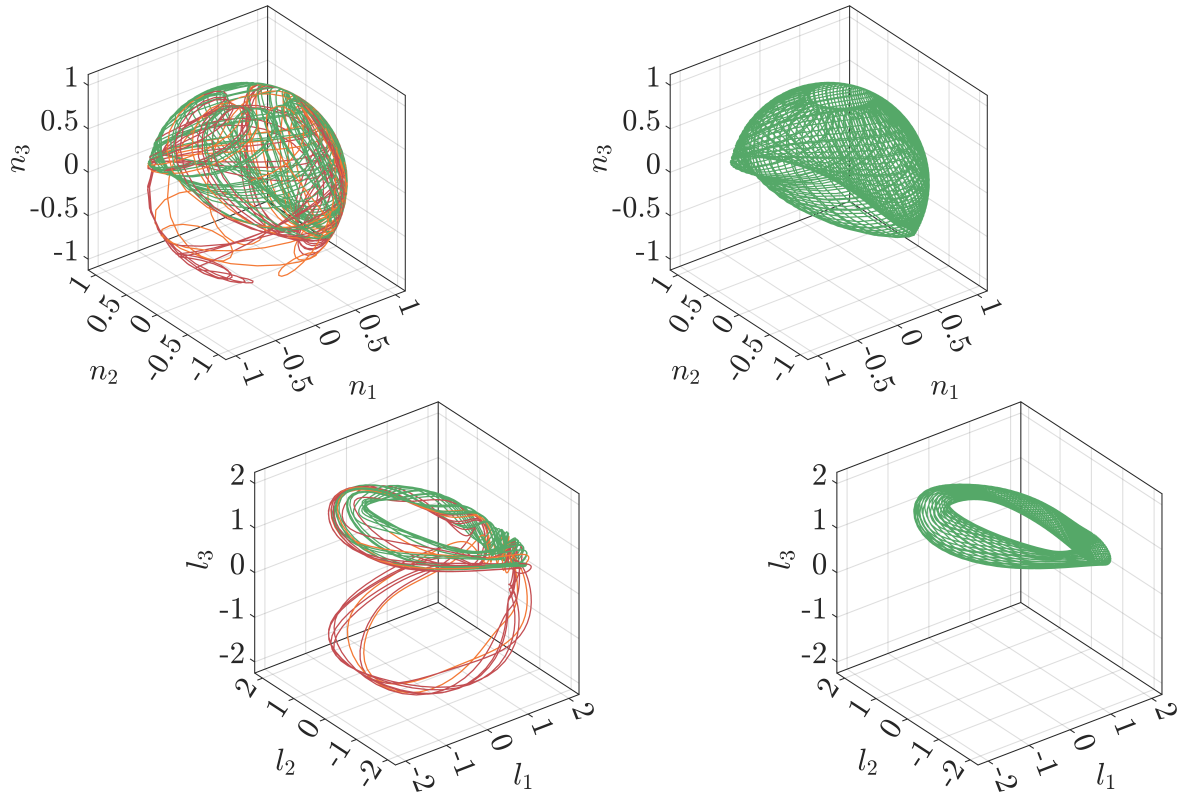
$$\begin{aligned}\tilde{F}\left(\begin{pmatrix}\mathbf{n}\\ \mathbf{l}\end{pmatrix}\right)\left[\begin{pmatrix}\mathbf{a}_1\\ \mathbf{b}_1\end{pmatrix}, \begin{pmatrix}\mathbf{a}_2\\ \mathbf{b}_2\end{pmatrix}, \begin{pmatrix}\mathbf{a}_3\\ \mathbf{b}_3\end{pmatrix}, \begin{pmatrix}\mathbf{a}_4\\ \mathbf{b}_4\end{pmatrix}, \begin{pmatrix}\mathbf{m}\\ \mathbf{k}\end{pmatrix}\right] \\ := \frac{1}{6 \det[J\mathbf{l}\ \nabla_{\mathbf{l}} K\ \mathbf{n}]\|\mathbf{n}\|^2} \operatorname{Alt} \tilde{G}\left(\begin{pmatrix}\mathbf{n}\\ \mathbf{l}\end{pmatrix}\right)\left[\begin{pmatrix}\mathbf{a}_1\\ \mathbf{b}_1\end{pmatrix}, \begin{pmatrix}\mathbf{a}_2\\ \mathbf{b}_2\end{pmatrix}, \begin{pmatrix}\mathbf{a}_3\\ \mathbf{b}_3\end{pmatrix}, \begin{pmatrix}\mathbf{a}_4\\ \mathbf{b}_4\end{pmatrix}, \begin{pmatrix}\mathbf{m}\\ \mathbf{k}\end{pmatrix}\right]. \quad (6.76c)\end{aligned}$$

This satisfies (6.58); we may then use such $\tilde{F}$ in (6.65) to define a fully conservative integrator for the Kovalevskaya top.

6.2.2.1 Numerical test

Fig. 6.4 shows numerical simulations of the Kovalevskaya top with IM and the fully conservative modification of 1-stage CPG (6.65) using $\tilde{F}$ as in (6.76c) with the same ICs $\mathbf{n}(0) = (0.8, 0.6, 0)$, $\mathbf{l}(0) = (2, 0, 0.2)$ and timestep $\Delta t = 0.1$ until final time 300. Fig. 6.5 shows the evolution and drift of the Kovalevskaya invariant K within the IM scheme. In each figure, colouring indicates error in the Kovalevskaya invariant K: green for $|K - K(0)| \leq \frac{1}{2}$, orange for $|K - K(0)| \in (\frac{1}{2}, 1]$, red for $|K - K(0)| > 1$.

All invariants, including K, are conserved by the trajectory of the SP scheme (up to quadrature error, solver tolerances and machine precision). As a quartic invariant,



(a) IM **(b)** Our modification of CPG

Figure 6.4: Trajectories in $\mathbf{n}, \mathbf{l}$ of the Kovalevskaya top, with IM (left) and our proposed modification of CPG (right).

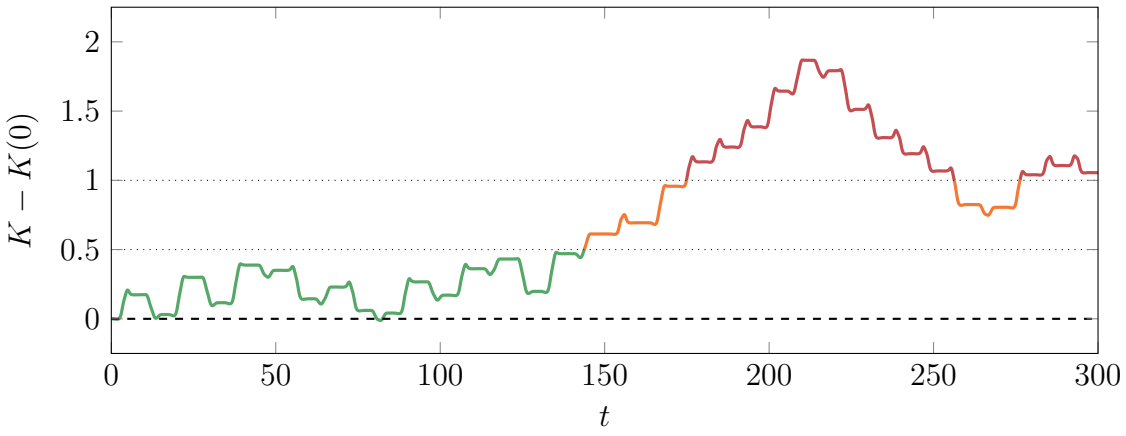


Figure 6.5: Error $K - K(0)$ within the IM simulation of the Kovalevskaya top.

K is not conserved by the IM scheme; we see that the resulting drift in K allows the IM numerical simulation to admit unphysical trajectories after a sufficient duration (approximately 18 rotations of the top for these ICs).

6.2.3 **Analysis:** uniqueness & convergence

We conclude our discussion of the general conservative integrator (6.65) by considering the existence of unique solutions, and their convergence as $\Delta t_{\max} := \max_n \Delta t_n \to 0$ through the results of Subsection 6.1.1.

For uniqueness we refer to Theorem 6.5.

Example (General conservative systems)

The universally stable integrator for conservative systems (6.65) satisfies both the following uniqueness results:

- Assume (N_p) are globally Lipschitz differentiable, and $\tilde{F}$ is globally Lipschitz in $\mathbf{x}$. There then exists a unique solution on sufficiently small timesteps Δt_n.
- Assume (N_p) are locally Lipschitz differentiable, and $\tilde{F}$ is locally Lipschitz in $\mathbf{x}$, each on a neighbourhood of $\mathbf{x}(t_n)$. For sufficiently small $\delta > 0$, there then exists a unique solution satisfying $\sup_{T_n} \|\mathbf{x} - \mathbf{x}(t_n)\| \leq \delta$ on sufficiently small timesteps Δt_n.

The latter result holds for the Kepler integrator (Section 6.2.1) provided both the energy H and the angular momentum L are non-zero,[a] and for the Kovalevskaya integrator (Section 6.2.2) provided $\det[J\mathbf{l}\ \nabla_{\mathbf{l}} K\ \mathbf{n}]|_{t=t_n}$ is non-zero.

[a] We require also that $\mathbf{x}(t_n) \neq \mathbf{0}$, however this is necessary for the continuous system even to be well-defined. Similarly, a non-zero angular momentum is sufficient for the existence of exact solutions for all time $t > 0$.

For convergence we refer to Theorem 6.13 and Corollary 6.14.

Example (General conservative systems)

Assuming Assumption 6.11 holds (in particular such that $\tilde{F}$ is either globally of locally Lipschitz in $\mathbf{x}$) then the discrete solution $\mathbf{x}$ to (6.65) satisfies the following error estimates:

- On each n, for a sufficiently small local timestep Δt_n,

$$\sup_{T_n} \|\mathbf{x} - \mathbf{X}\| \lesssim \|\mathbf{x}(t_n) - \mathbf{X}(t_n)\| + \Delta t_n^{S+1}. \tag{6.77a}$$

- For a sufficiently small global timestep $\Delta t_{\max}$,

$$\|\mathbf{x}(t) - \mathbf{X}(t)\| \lesssim (t + \Delta t_{\max})\Delta t_{\max}^{S} \tag{6.77b}$$

  for all $t \in \mathbb{R}_+$.

The final condition in Assumption 6.11 holds for both the Kepler (Section 6.2.1) and Kovalevskaya (Section 6.2.2) integrators, provided again in the former case that neither H nor L is zero, and in the latter that $\det[J\mathbf{l}\ \nabla_{\mathbf{l}} K\ \mathbf{n}]$ remains non-zero along all discrete trajectories $\mathbf{x}$ as $\Delta t_{\max} \to 0$.

6.3 GENERIC formalism: energy & entropy stability

We consider now systems of ODEs deriving from the GENERIC formalism [GÖ97; ÖG97] which we refer to as GENERIC ODEs. A GENERIC ODE can be most simply interpreted as a combination of a Poisson and gradient-descent ODE (6.3),

$$\dot{\mathbf{x}} = B(\mathbf{x})\nabla E(\mathbf{x}) + D(\mathbf{x})\nabla S(\mathbf{x}), \tag{6.78}$$

where $B(\mathbf{x}), D(\mathbf{x}) \in \mathbb{R}^{d\times d}$ are skew-symmetric and positive semidefinite respectively (the Poisson and friction matrices) and $E(\mathbf{x}), S(\mathbf{x}) \in \mathbb{R}$ are conserved and generated respectively (the energy and entropy). For these conservation and dissipation structures to hold, the additional orthogonality constraints

$$\nabla S(\mathbf{x})^\top B(\mathbf{x}) = \mathbf{0}, \qquad \nabla E(\mathbf{x})^\top D(\mathbf{x}) = \mathbf{0} \tag{6.79}$$

are imposed. In the construction of our SP scheme, we rely on the following assumption.

Assumption 6.17 (Characterisation of GENERIC matrix compatibility)**.** *Assume the existence of $\tilde{B}, \tilde{D} : (\mathbb{R}^d)^2 \to \mathbb{R}^{d\times d}$ such that the following hold:*

1. *$\tilde{B}, \tilde{D}$ coincide with B, D: for all $\mathbf{x} \in \mathbb{R}^d$,*

$$\tilde{B}(\mathbf{x}, \nabla S(\mathbf{x})) = B(\mathbf{x}), \qquad \tilde{D}(\mathbf{x}, \nabla E(\mathbf{x})) = D(\mathbf{x}). \tag{6.80a}$$

2. *$\tilde{B}, \tilde{D}$ are skew-symmetric and positive semidefinite respectively for all arguments.*

3. *$\tilde{B}, \tilde{D}$ preserve the compatibility conditions (6.79) for all arguments: for all $\mathbf{x}, \mathbf{y}_E, \mathbf{y}_S \in \mathbb{R}^d$,*

$$\mathbf{y}_S^\top \tilde{B}(\mathbf{x}, \mathbf{y}_S) = \mathbf{0}, \qquad \mathbf{y}_E^\top \tilde{D}(\mathbf{x}, \mathbf{y}_E) = \mathbf{0}. \tag{6.80b}$$

With Assumption 6.17 established, we may apply our framework to construct an integrator for (6.78) that preserves both the conservation law in E and dissipation inequality in S. The argument being very similar to that in Sections 6.1 & 6.2, we omit the details here and state only the final scheme: find $(\mathbf{x}, (\widetilde{\nabla E}, \widetilde{\nabla S})) \in \mathbb{X}_n \times (\dot{\mathbb{X}}_n^2)$ such that

$$\mathcal{I}_n[\mathbf{y}^\top \dot{\mathbf{x}}] = \mathcal{I}_n[\mathbf{y}^\top \tilde{B}(\mathbf{x}, \widetilde{\nabla S})\widetilde{\nabla E} + \mathbf{y}^\top \tilde{D}(\mathbf{x}, \widetilde{\nabla E})\widetilde{\nabla S}], \tag{6.81a}$$

$$\mathcal{I}_n[\widetilde{\nabla E}^\top \mathbf{y}_E] = \int_{T_n} \nabla E(\mathbf{x})^\top \mathbf{y}_E, \tag{6.81b}$$

$$\mathcal{I}_n[\widetilde{\nabla S}^\top \mathbf{y}_S] = \int_{T_n} \nabla S(\mathbf{x})^\top \mathbf{y}_S, \tag{6.81c}$$

for all $(\mathbf{y}, (\mathbf{y}_E, \mathbf{y}_S)) \in \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n^2$.

Theorem 6.18 (Energy & entropy stability of the GENERIC ODE integrator)**.** *The integrator (6.65) is energy and entropy stable, with $E(\mathbf{x}(t_{n+1})) = E(\mathbf{x}(t_n))$ and $S(\mathbf{x}(t_{n+1})) \geq S(\mathbf{x}(t_n))$.*

Proof. By considering respectively $\tilde{\mathbf{y}}_E = \tilde{\mathbf{y}}_S = \dot{\mathbf{x}}$ and $\mathbf{y}_E = \widetilde{\nabla E}$, $\mathbf{y}_S = \widetilde{\nabla S}$ in (6.81),

$$\begin{aligned}
&E(\mathbf{x}(t_{n+1})) - E(\mathbf{x}(t_n)) && S(\mathbf{x}(t_{n+1})) - S(\mathbf{x}(t_n)) \\
&= \int_{T_n} \dot{E} && = \int_{T_n} \dot{S} && (6.82a)\\
&= \int_{T_n} \nabla E(\mathbf{x})^\top \dot{\mathbf{x}} && = \int_{T_n} \nabla S(\mathbf{x})^\top \dot{\mathbf{x}} && (6.82b)\\
&= \mathcal{I}_n\left[\widetilde{\nabla E}^\top \dot{\mathbf{x}}\right] && = \mathcal{I}_n\left[\widetilde{\nabla S}^\top \dot{\mathbf{x}}\right] && (6.82c)\\
&= \mathcal{I}_n\left[\begin{array}{l}\widetilde{\nabla E}^\top \tilde{B}(\mathbf{x}, \widetilde{\nabla S})\widetilde{\nabla E} \\ \qquad + \widetilde{\nabla E}^\top \tilde{D}(\mathbf{x}, \widetilde{\nabla E})\widetilde{\nabla S}\end{array}\right] && = \mathcal{I}_n\left[\begin{array}{l}\widetilde{\nabla S}^\top \tilde{B}(\mathbf{x}, \widetilde{\nabla S})\widetilde{\nabla E} \\ \qquad + \widetilde{\nabla S}^\top \tilde{D}(\mathbf{x}, \widetilde{\nabla E})\widetilde{\nabla S}\end{array}\right] && (6.82d)\\
&= 0, && \geq 0, && (6.82e)
\end{aligned}$$

where the final equality and inequality hold by Assumption 6.17. □

6.3.1 A simple thermodynamic engine

Inspired by the classical thermodynamic systems considered by e.g. Ottinger [Ött05, Ex. 3] or Gay-Balmaz & Yoshimura [GY17, Sec. 3.1], we consider as an example a (nondimensionalised) model for an idealised, unpowered, C-cylinder thermodynamic engine with thermal dissipation,

$$\dot{\theta} = \omega, \qquad \dot{S}_c = \frac{1}{T_c}[T_0 - T_c], \tag{6.83a}$$

$$\dot{\omega} = \sum_{c=1}^{C} P_c \sin\left(\theta - \frac{2\pi c}{C}\right), \qquad \dot{S}_0 = \frac{1}{T_0}\sum_{c=1}^{C}[T_c - T_0]. \tag{6.83b}$$

Here, θ represents the engine phase and ω its rate of change. The thermodynamic variables $(S_c)_{c=1}^C$, $(P_c)_{c=1}^C$ and $(T_c)_{c=1}^C$ represent the entropies, pressures and temperatures respectively within each piston, satisfying the fluid's equation of state at volumes $(V_c \coloneqq V_p - \cos(\theta - \frac{2\pi c}{C}))_c$ for a constant, uniform reference volume $V_p > 1$; similarly, the thermodynamic variables S_0 and T_0 represents the entropy and temperature respectively of the surrounding environment, the latter of which being constant.

Example (Ideal fluid)

For an ideal fluid, (P_c), (T_c) can be related to (S_c), $(V_c \coloneqq V_p - \cos(\theta - \frac{2\pi c}{C}))$ by

$$P_c(S_c, V_c) = \exp\left(\frac{S_c}{C_V}\right) V_c^{-\gamma}, \qquad T_c(S_c, V_c) = P_c(S_c, V_c) V_c, \tag{6.84}$$

where C_V is the nondimensionalised heat capacity at constant volume and $\gamma = 1 + \frac{1}{C_V}$ is the adiabatic index ($\frac{3}{2}$ and $\frac{5}{3}$ respectively for a monatomic gas).

Define a state variable $\mathbf{x} \coloneqq (\theta, \omega, (S_c), S_0)$ accordingly, with total energy E and entropy S,

$$E(\mathbf{x}) \coloneqq \frac{1}{2}\omega^2 + \sum_c U_c + U_0, \qquad S(\mathbf{x}) \coloneqq \sum_c S_c + S_0, \tag{6.85}$$

where $(U_c)_{c=1}^C$ denote the internal energies within each piston and U_0 the energy dissipated to the surrounding environment, again satisfying the fluid's equation of state.

Example (Ideal fluid)

For an ideal fluid, (U_c) can be related to (S_c), (V_c) and U_0 to S_0 by

$$U_c(S_c, V_c) = C_V T_c(S_c, V_c), \qquad U_0(S_0) = T_0 S_0. \tag{6.86}$$

Observing the gradients in E, S,

$$\nabla E(\mathbf{x}) \coloneqq \begin{pmatrix} -\sum_c P_c \sin\left(\theta - \frac{2\pi c}{C}\right) \\ \omega \\ (T_c)_c \\ T_0 \end{pmatrix}, \qquad \nabla S(\mathbf{x}) \coloneqq \begin{pmatrix} 0 \\ 0 \\ \mathbf{1} \\ 1 \end{pmatrix}, \tag{6.87}$$

where the former identity may be derived by the fundamental thermodynamic relations

$$dU_c = T_c dS_c - P_c dV_c = T_c dS_c - P_c \sin\left(\theta - \frac{2\pi c}{C}\right) d\theta, \qquad dU_0 = T_0 dS_0. \tag{6.88}$$

The system (6.83) can then be written under the GENERIC formalism (6.78) for

$$B(\mathbf{x}) \coloneqq \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}, \quad D(\mathbf{x}) \coloneqq \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \frac{T_0}{T_1} & \cdots & 0 & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{T_0}{T_C} & -1 \\ 0 & 0 & -1 & \cdots & -1 & \sum_c \frac{T_C}{T_0} \end{pmatrix}. \tag{6.89}$$

For B, D as defined in the thermodynamic engine model (6.89), $\tilde{B}$ may simply be defined as B to satisfy the conditions of Assumption 6.17 (as it is constant) while $\tilde{D}$ may be defined

$$D(\mathbf{x}, \nabla E) \coloneqq \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \frac{\partial_{S_0} E}{\partial_{S_1} E} & \cdots & 0 & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{\partial_{S_0} E}{\partial_{S_C} E} & -1 \\ 0 & 0 & -1 & \cdots & -1 & \sum_c \frac{\partial_{S_c} E}{\partial_{S_0} E} \end{pmatrix}, \tag{6.90}$$

where ∂_{S_c} denotes the partial derivative with respect to S_c. Through (6.81) we then construct the following energy- and entropy-stable integrator for (6.83): find

$((\theta, \omega, (S_c), S_0), (\tilde{P}, \tilde{\omega}, (\tilde{T}_c))) \in \mathbb{X}_n^{C+3} \times \dot{\mathbb{X}}_n^{C+2}$ (with $\dot{\mathbb{X}}_n$ defined as in (6.2) for $d = 1$) such that

$$\mathcal{I}_n[\dot{\theta}\eta] = \mathcal{I}_n[\tilde{\omega}\eta] \tag{6.91a}$$

$$\mathcal{I}_n[\dot{\omega}\psi] = -\mathcal{I}_n[\tilde{P}\psi] \tag{6.91b}$$

$$\mathcal{I}_n[\dot{S}_c R_c] = \mathcal{I}_n\left[\frac{1}{\tilde{T}_c}(T_0 - \tilde{T}_c)R_c\right] \tag{6.91c}$$

$$\mathcal{I}_n[\dot{S}_0 R_0] = \mathcal{I}_n\left[\frac{1}{T_0}\sum_c (\tilde{T}_c - T_0)R_0\right] \tag{6.91d}$$

$$\mathcal{I}_n[\tilde{P}\tilde{Q}] = -\int_{T_n} \sum_c P_c(S_c, V_c) \sin\left(\theta - \frac{2\pi c}{C}\right)\tilde{Q} \tag{6.91e}$$

$$\mathcal{I}_n[\tilde{\omega}\tilde{\psi}] = \int_{T_n} \omega\tilde{\psi} \tag{6.91f}$$

$$\mathcal{I}_n[\tilde{T}_c\tilde{V}_c] = \int_{T_n} T_c(S_c, V_c)\tilde{V}_c \tag{6.91g}$$

for all $((\eta, \psi, (R_c), R_0), (\tilde{Q}, \tilde{\psi}, (\tilde{V}_c))) \in \dot{\mathbb{X}}_n^{C+3} \times \dot{\mathbb{X}}_n^{C+2}$, where again each $V_c := V_p - \cos(\theta - \frac{2\pi c}{C})$. Specifically, $\tilde{P}$, $\tilde{\omega}$, $(\tilde{T}_c)$ are the AVs for energy conservation, approximating $-\sum_c P_c \sin\left(\theta - \frac{2\pi c}{C}\right)$, ω, $T_c(S_c, V_c)$ respectively.[4]

One may observe in fact that the scheme (6.91) may be simplified, as the RHSs of (6.91a, 6.91b) identify with the LHSs of (6.91f, 6.91e) respectively: find $((\theta, \omega, (S_c), S_0), (\tilde{T}_c)) \in \mathbb{X}_n^{C+3} \times \dot{\mathbb{X}}_n^{C}$ such that

$$\mathcal{I}_n[\dot{\theta}\eta] = \int_{T_n} \omega\eta \tag{6.92a}$$

$$\mathcal{I}_n[\dot{\omega}\psi] = \int_{T_n} \sum_c P_c(S_c, V_c) \sin\left(\theta - \frac{2\pi c}{C}\right)\tilde{Q} \tag{6.92b}$$

$$\mathcal{I}_n[\dot{S}_c R_c] = \mathcal{I}_n\left[\frac{1}{\tilde{T}_c}(T_0 - \tilde{T}_c)R_c\right] \tag{6.92c}$$

$$\mathcal{I}_n[\dot{S}_0 R_0] = \mathcal{I}_n\left[\frac{1}{T_0}\sum_c (\tilde{T}_c - T_0)R_0\right] \tag{6.92d}$$

$$\mathcal{I}_n[\tilde{T}_c\tilde{V}_c] = \int_{T_n} T_c(S_c, V_c)\tilde{V}_c \tag{6.92e}$$

for all $((\eta, \psi, (R_c), R_0), (\tilde{V}_c)) \in \dot{\mathbb{X}}_n^{C+3} \times \dot{\mathbb{X}}_n^{C}$.

Remark 6.19 (Choice of parametrisation). *As with any thermodynamic system, there exists here a choice of parametrisation of the thermodynamic variables, identical on the continuous level, but implying different properties on the discrete. In particular, choosing*

[4] We do not require an AV approximating T_0 as it constant, nor do we need those approximating the gradients in S for the same reason.

a parametrisation in which a certain QoI is linear will necessarily preserve the behaviour of that QoI under any consistent timestepping scheme. To represent classical schemes in a fair light, we therefore consider a parametrisation in the internal (S_c) and external S_0 entropies, under which any consistent timestepping scheme will preserve the generation of entropy; within our framework, this has the equivalent implication that the associated test function for entropy ∇S is constant, thus already lies in $\mathbb{X}_n$ automatically, and necessitates the introduction of no AV.

A consequence of this fair comparison is that the scheme here appears somewhat simpler than that proposed in (6.81), with only a minor subset of the AVs being required.[5]

Similar eliminations could be performed for the AV approximating ∇E had we chosen to parametrise in the internal (U_c) and external U_0 energies. Alternatively, parametrising in the internal (T_c) and external T_0 temperatures would have created a richer discretisation better reflecting (6.81) with AVs introduced approximating both ∇E and ∇S; classical timestepping schemes however would fail under such a parametrisation on both energy and entropy stability, making this comparison unfair.

6.3.2 **Analysis:** uniqueness & convergence

We conclude our discussion of the integrator for GENERIC ODEs (6.81) by considering the existence of unique solutions, and their convergence as $\Delta t_{\max} := \max_n \Delta t_n \to 0$ through the results of Subsection 6.1.1.

For uniqueness we refer to Theorem 6.5.

Example (GENERIC ODEs)

The universally stable integrator for conservative systems (6.65) satisfies both the following uniqueness results:

- Assume E, S are globally Lipschitz differentiable, and $\tilde{B}$, $\tilde{D}$ are globally Lipschitz. There then exists a unique solution on sufficiently small timesteps Δt_n.
- Assume E, S are locally Lipschitz differentiable on a neighbourhood

[5] In fact, as per the argument in Section 4.2.1, we can eliminate each AV on the computational level regardless, by writing it as an explicit function of the primal variables θ, ω, (S_c), S_0.

of $\mathbf{x}(t_n)$, and $\tilde{B}$, $\tilde{D}$ are locally Lipschitz on a neighbourhood of $\mathbf{x}(t_n)$, $\nabla E(\mathbf{x}(t_n))$, $\nabla S(\mathbf{x}(t_n))$. For sufficiently small $\delta > 0$, there then exists a unique solution satisfying $\sup_{T_n} \|\mathbf{x} - \mathbf{x}(t_n)\| \leq \delta$ on sufficiently small timesteps Δt_n.

In particular, assuming the constitutive relations determining the pressure P and temperature T as functions of the entropy S and volume V are locally Lipschitz, the latter result holds for our integrator for the thermodynamic engine (6.91).

For convergence we refer to Theorem 6.13 and Corollary 6.14.

Example (GENERIC ODEs)

Assuming Assumption 6.11 holds (in particular such that $\tilde{B}$, $\tilde{D}$ are either globally of locally Lipschitz in $\mathbf{x}$, $\widetilde{\nabla E}$, $\widetilde{\nabla S}$) then the discrete solution $\mathbf{x}$ to (6.65) satisfies the following error estimates:

- On each n, for a sufficiently small local timestep Δt_n,

$$\sup_{T_n} \|\mathbf{x} - \mathbf{X}\| \lesssim \|\mathbf{x}(t_n) - \mathbf{X}(t_n)\| + \Delta t_n^{S+1}. \tag{6.93a}$$

- For a sufficiently small global timestep $\Delta t_{\max}$,

$$\|\mathbf{x}(t) - \mathbf{X}(t)\| \lesssim (t + \Delta t_{\max}) \Delta t_{\max}^{S} \tag{6.93b}$$

for all $t \in \mathbb{R}_+$.

The final condition in Assumption 6.11 holds in particular for our thermodynamic engine integrator (6.91), provided the exact solution is uniformly $(2S-1)$-times continuously differentiable, and that the constitutive relations determining P, T in terms of S, V are also $(2S-1)$-times continuously differentiable (albeit not necessarily uniformly).

'Listen, do you want the job done right, or do you want it done fast?"

— Homer J. Simpson (Daniel L. 'Dan' Castellaneta) [FP00]

7

PDEs: Poisson, gradient-descent, GENERIC & compressible Navier–Stokes

Contents

This chapter continues the discussion of applications by considering geometric PDE systems. In contrast to Chapter 6, these schemes will be presented without analysis; none of the PDEs considered in this chapter fall under the class of AD systems analysed in 3.3 while, as discussed earlier, the results in Subsection 6.1.1 for the analysis of SP discretisations for ODE systems do not give meaningful results in PDE settings.

The rest of this chapter proceeds as follows. In Section 7.1, we begin similarly to Section 6.1 by considering the application of our framework to a simple Poisson or gradient-descent system (7.1), deriving a discretisation that is energy-stable, i.e. conserves the energy over timesteps in the former case and dissipates it in the

latter. We consider as an example system the BBM equations. In Section 6.3, similarly to Section 7.2 we consider systems of PDEs deriving from the GENERIC formalism [GÖ97; ÖG97]. As a canonical dissipative thermodynamic PDE, we consider the Boltzmann equation, deriving novel integrators that simultaneously preserve both the conservation of energy and generation of entropy. In Section 7.3, we conclude by considering the compressible NS equations, deriving novel integrators that are mass-, momentum-, energy- and entropy-stable, i.e. conserve mass, momentum, energy and entropy in the inviscid Euler case, and conserve mass, momentum and energy and necessarily generate entropy in the viscous case.

7.1 Poisson & gradient-descent systems: energy stability

As an introductory PDE example, we begin similarly to Section 6.1 by considering either a general Poisson or gradient-descent system, with a single QoI, the energy $H(u) \in \mathbb{R}$, either conserved or dissipated respectively.

Unlike the ODE system (6.3) this is most conveniently stated for our purposes in a variational form over some space U: find $u \in C^1(\mathbb{R}_+; U)$ satisfying known initial data, such that

$$M(u; \dot{u}, v) = B(u; w_H(u), v) \tag{7.1}$$

at all times $t \in \mathbb{R}_+$ and for all $v \in U$, where the operators $M, B : U \times U \times U \to \mathbb{R}$ are linear in their final two arguments, with the latter furthermore either skew-symmetric, in the case of a general Poisson system, or negative semidefinite, in the case of a general gradient-descent system. The functional $w_H : U \to U$ is such that $M(u; \cdot, w_H(u)) = H'(u; \cdot)$, the Fréchet derivative of H, i.e. it is the associated test function for H as defined in Step C of our framework. It is straightforward then to see the behaviour of H over T_n by considering $v = w_H(u)$:

$$\begin{aligned} H(u(t_{n+1})) - H(u(t_n)) &= \int_{T_n} \dot{H} = \int_{T_n} H'(u; \dot{u}) = \int_{T_n} M(u; \dot{u}, w_H(u)) \\ &= \int_{T_n} B(u; w_H(u), w_H(u)) \begin{cases} = 0, & \text{Poisson}, \\ \leq 0, & \text{gradient-descent}, \end{cases} \end{aligned} \tag{7.2}$$

with the final result holding by either the skew-symmetry of negative semidefiniteness of $B(u; \cdot, \cdot)$ respectively.

Remark 7.1 (Poisson & gradient-descent PDEs without M)**.** *The Poisson or gradient-descent PDE (7.1) can be stated without the choice of M by considering test functions in the dual space U^* of U: find $u \in C^1(\mathbb{R}_+; U)$ satisfying known initial data, such that*

$$L[\dot{u}] = B^*(u; H'(u), L) \tag{7.3}$$

at all times $t \in \mathbb{R}_+$ and for all $L \in U^*$, where $B : U \times U \times U \to \mathbb{R}$ and $B^* : U \times U^* \times U^* \to \mathbb{R}$ are related by

$$B^*(u; M(u; \cdot, w), M(u; \cdot, v)) = B(u; w, v). \tag{7.4}$$

However, as our framework requires the test space and solution space to be identical, this more general, abstract form is less useful for our purposes.

We may now apply our framework to construct an energy-stable integrator for the Poisson or gradient-descent PDE (7.1).

Application of framework (Algorithm 3.5)

A. Taking U to be a finite-dimensional function space $\mathbb{U}$, the system (7.1) defines our semidiscrete form.[1]

B. Over the timestep T_n, this is cast into a fully discrete form using our choice of $\mathcal{I}_n$: find $\mathbf{x} \in \mathbb{X}_n$ (for $\mathbb{X}_n$ defined as in (3.10)) such that

$$\mathcal{I}_n[M(u; \dot{u}, v)] = \mathcal{I}_n[B(u; w_H(u), v)], \tag{7.5}$$

for all $v \in \dot{\mathbb{X}}_n$.

C. As established earlier, the associated test function for the evolution of H is $w_H(u)$ satisfying $H'(u; \cdot) = M(u; \cdot, w_H(u))$.

D. We introduce an AV $\tilde{w}_H \in \dot{\mathbb{X}}_n$, approximating $w_H(u)$, and defined as in (3.19) such that

$$\mathcal{I}_n[M(u; v_H, \tilde{w}_H)] = \int_{T_n} H'(u; v) \quad \left(= \int_{T_n} M(u; v_H, w_H(u))\right), \tag{7.6}$$

for all $v_H \in \dot{\mathbb{X}}_n$.

[1] In the language of Section 3.1, $F : \mathbb{U} \times \mathbb{U} \to \mathbb{R}$ is defined $F(u; v) \coloneqq B(u; w_H(u), v)$

E. To introduce $\tilde{w}_H$ into the RHS of (7.5) in a way to preserve the behaviour of H, we propose the modified discrete form,[2]

$$\mathcal{I}_n[M(u;\dot{u},v)] = \mathcal{I}_n[B(u;\tilde{w}_H,v)]. \tag{7.7}$$

F. The final SP scheme is then as follows: find $(u,\tilde{w}_H) \in \mathbb{X}_n \times \dot{\mathbb{X}}_n$ such that

$$\mathcal{I}_n[M(u;\dot{u},v)] = \mathcal{I}_n[B(u;\tilde{w}_H,v)], \tag{7.8a}$$

$$\mathcal{I}_n[M(u;v_H,\tilde{w}_H)] = \int_{T_n} H'(u;v_H), \tag{7.8b}$$

for all $(v,v_H) \in \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n$.

Theorem 7.2 (Energy stability of the Poisson & gradient-descent ODE integrator)**.** *The integrator* (7.8) *is energy-stable, with*

$$H(\mathbf{x}(t_{n+1})) - H(\mathbf{x}(t_n)) \begin{cases} = 0, & \textit{Poisson}, \\ \leq 0, & \textit{gradient-descent}. \end{cases} \tag{7.9}$$

Proof. By considering respectively $v = \dot{u}$ and $v_H = \tilde{w}_H$ in (6.11),

$$\begin{aligned} H(\mathbf{x}(t_{n+1})) - H(\mathbf{x}(t_n)) = \int_{T_n} \dot{H} = \int_{T_n} H'(u;\dot{u}) = \mathcal{I}_n[M(u;\dot{u},\tilde{w}_H)] \\ = \mathcal{I}_n[B(u;\tilde{w}_H,\tilde{w}_H)] \begin{cases} = 0, & \text{Poisson}, \\ \leq 0, & \text{gradient-descent}, \end{cases} \end{aligned} \tag{7.10}$$

with the final result holding by either the skew-symmetry of negative semidefiniteness of $B(u;\cdot,\cdot)$ respectively, and the sign-preserving property of $\mathcal{I}_n$. □

7.1.1 The Benjamin–Bona–Mahony equation

As a motivating example, consider the BBM equation [BBM97] in $u : \mathbb{R}_+ \times \Omega \to \mathbb{R}$ over an interval $\Omega \subset \mathbb{R}$,

$$\dot{u} - \partial_x^2 \dot{u} = -\partial_x u - u\partial_x u, \tag{7.11}$$

where ∂_x denotes the partial derivative with respect to the spatial coordinate x. We shall assume periodic BCs. The BBM equation is a Poisson[3] system with energy

$$H(u) \coloneqq \int_\Omega \frac{1}{2}u^2 + \frac{1}{6}u^3. \tag{7.12}$$

[2] In the language of Section 3.1, $\tilde{F} : \mathbb{U} \times \mathbb{U} \times \mathbb{U} \to \mathbb{R}$ is defined $\tilde{F}(u,\tilde{w}_H;v) \coloneqq B(u;\tilde{w}_H,v)$.

[3] Hamiltonian, in fact.

To apply the scheme (7.8) to construct an energy-stable integrator for (7.11), we must first show (7.11) may be written in the form (7.1).

A typical semidiscrete variational form of (7.11) can be found by testing in L^2 against some test function v: find $u \in U$ satisfying known ICs such that

$$(\dot{u} - \partial_x^2 \dot{u}, v) = -(\partial_x u + u\partial_x u, v) \tag{7.13a}$$

$$(\dot{u}, v) + (\partial_x \dot{u}, \partial_x v) = -(\partial_x [u + \frac{1}{2}u^2], v) \tag{7.13b}$$

$$(\dot{u}, v)_{H^1} = (u + \frac{1}{2}u^2, \partial_x v) \tag{7.13c}$$

at all times $t \in \mathbb{R}_+$ and for all $v \in U$, implying M is simply the H^1 inner product. The associated test function $w_H(u)$ for H must then satisfy

$$(\delta u, w_H(u))_{H^1} = H'(u; \delta u) \tag{7.14a}$$

$$(\delta u, w_H(u) - \partial_x^2 w_H(u)) = (u + \frac{1}{2}u^2, \delta u), \tag{7.14b}$$

implying $w_H(u)$ is defined in strong form implicitly to satisfy

$$w_H(u) - \partial_x^2 w_H(u) = u + \frac{1}{2}u^2. \tag{7.15}$$

We may write our semidiscrete variational form (7.13c) in terms of $w_H(u)$ therefore as,

$$(\dot{u}, v)_{H^1} = (w_H(u) - \partial_x^2 w_H(u), \partial_x v) \tag{7.16a}$$

$$= \frac{1}{2}[(w_H(u), \partial_x v) + (\partial_x w_H(u), \partial_x^2 v) - (\partial_x w_H(u), v) - (\partial_x^2 w_H(u), \partial_x v)], \tag{7.16b}$$

$$= \frac{1}{2}[(w_H(u), \partial_x v)_{H^1} - (\partial_x w_H(u), v)_{H^1}]. \tag{7.16c}$$

Defining skew-symmetric $B(w, v) \coloneqq \frac{1}{2}[(w, \partial_x v)_{H^1} - (\partial_x w, v)_{H^1}]$ therefore, this aligns with the variational form for general Poisson PDEs (7.1).

We may therefore use our general SP integrator (7.8) for Poisson PDEs to derive the following energy-conserving scheme: find $(u, \tilde{w}_H) \in \mathbb{X}_n \times \dot{\mathbb{X}}_n$ such that

$$\mathcal{I}_n[(\dot{u}, v)_{H^1}] = \frac{1}{2}\mathcal{I}_n[(\tilde{w}_H, \partial_x v)_{H^1} - (\partial_x \tilde{w}_H, v)_{H^1}], \tag{7.17a}$$

$$\mathcal{I}_n[(v_H, \tilde{w}_H)_{H^1}] = \int_{T_n} (u + \frac{1}{2}u^2, v_H), \tag{7.17b}$$

for all $(v, v_H) \in \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n$. Again, taking $(v, v_H) = (\tilde{w}_H, \dot{u})$ confirms that (7.17) conserves H.

7.1.1.1 Soliton test

To numerically verify and motivate these conservation properties, we consider the domain $\Omega = (-50, 50)$. Up to projection, the following ICs form a soliton[4] of speed $\frac{1+\sqrt{5}}{2}$:

$$u(0) = \frac{3\sqrt{5}-3}{2} \operatorname{sech}\left(\frac{\sqrt{5}-1}{4}x\right)^2, \tag{7.18}$$

where sech is the hyperbolic secant function. Over an interval mesh of uniform mesh width 2, we take $\mathbb{U}$ to be the (degree-3) Hermite space (see Ern & Guermond [EG21a, Chap. 5]); in time, we take a uniform timestep $\Delta t_n = 1$. Under these conditions, we compare the results from a 2-stage (symplectic) Gauss method as applied to (7.13c) with that of the scheme (7.17) with $\mathcal{I}_n$ the exact integral[5] and $S = 2$.

Fig. 7.1 shows the evolution of the energy $H(u)$ under each scheme. Artificial dissipation in the energy under the Gauss method causes the value to decrease from its initial value of around 11.1 to around 6.2 at the final time $t = 2 \cdot 10^4$.

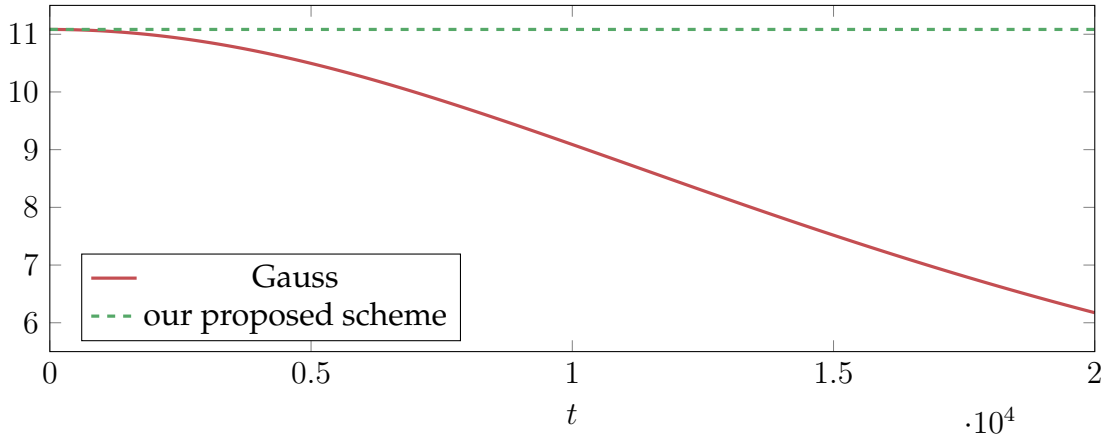
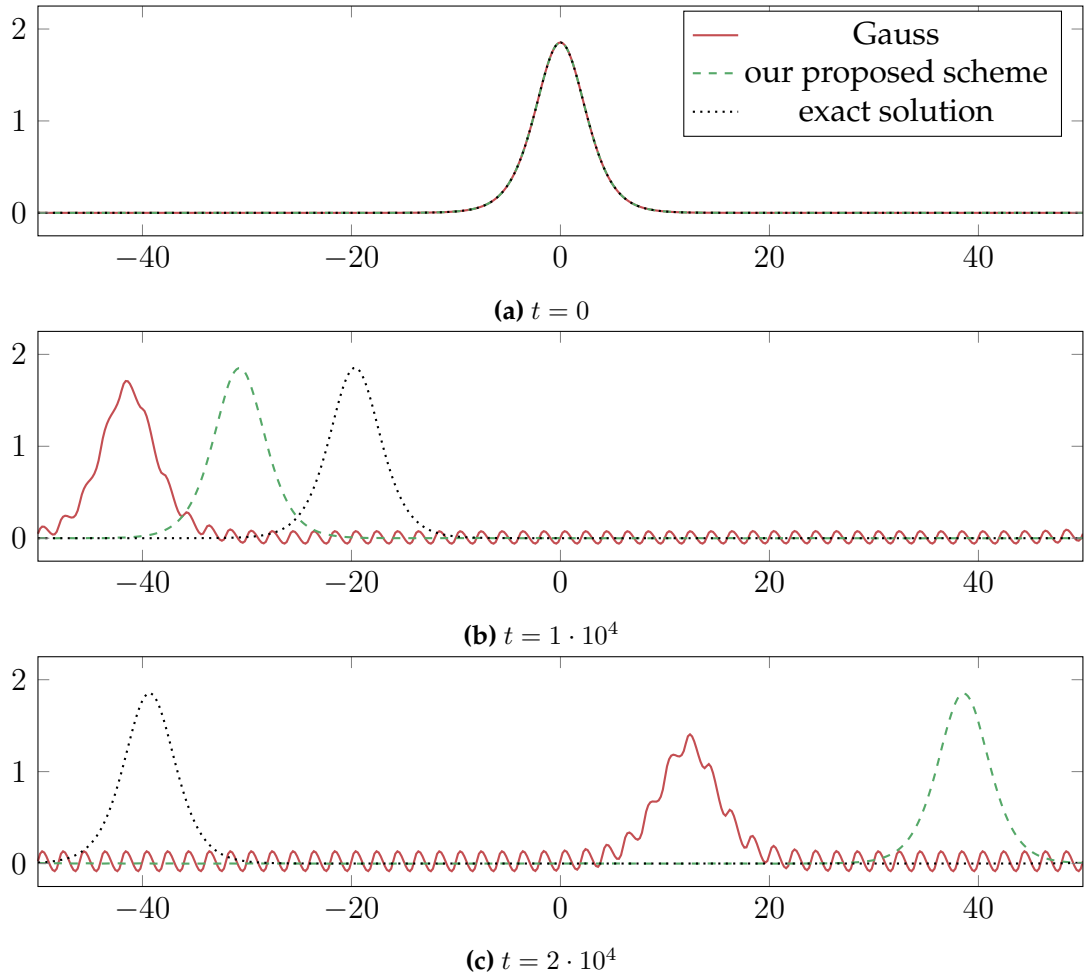


Figure 7.1: Evolution of the energy $H(u)$ when solving the BBM equations with a Gauss method and our proposed scheme.

Fig. 7.2 shows u under each scheme at various times along the simulation. The dissipation in $H(u)$ under the Gauss method correlates with a reduction in the amplitude of u, causing the speed of the discrete soliton to decrease. At $t = 2 \times 10^4$, the discrete soliton in the Gauss simulation has speed approximately 1.45; compare with the exact value of approximately 1.618, and that of the numerical solution from our proposed scheme of approximately 1.617.

[4]The appropriate notion of a nonlinear wave under periodic BCs is not a soliton, but a cnoidal wave (see Ablowitz & Segur [AS81, Sec. 2.3]). The value the ICs at the boundary $x = \pm 50$ however are approximately 2×10^{-13}, implying this distinction on this domain is negligible, especially after



(a) $t = 0$

(b) $t = 1 \cdot 10^4$

(c) $t = 2 \cdot 10^4$

Figure 7.2: Plots of $u(x)$ in the BBM simulations using a Gauss method and our proposed scheme at times $t \in \{0, 1 \cdot 10^4, 2 \cdot 10^4\}$. The exact solution is included for comparison.

Of note is the conservation of the H^1 norm $\|u\|_{H^1}$, a further invariant of the BBM equation. Fig. 7.1 shows the evolution of $\|u\|_{H^1}$ under our proposed scheme. While the construction of the scheme (7.17) does not guarantee the discrete conservation of $\|u\|_{H^1}$, we find numerically that $\|u(t_n)\|_{H^1}$ oscillates within the small interval $(15.9660, 15.9667)$ over the simulation duration; this is reminiscent of the approximate conservation of energy exhibited by symplectic integrators (see Fig. 6.2a or e.g. [HLW06, Chap. I, Fig 4.1]). The proof of this property remains an open problem.

projection into the discrete space $\mathbb{U}$.

[5] We are able to compute this exactly, as all terms in the discretisation (7.17) are polynomial.

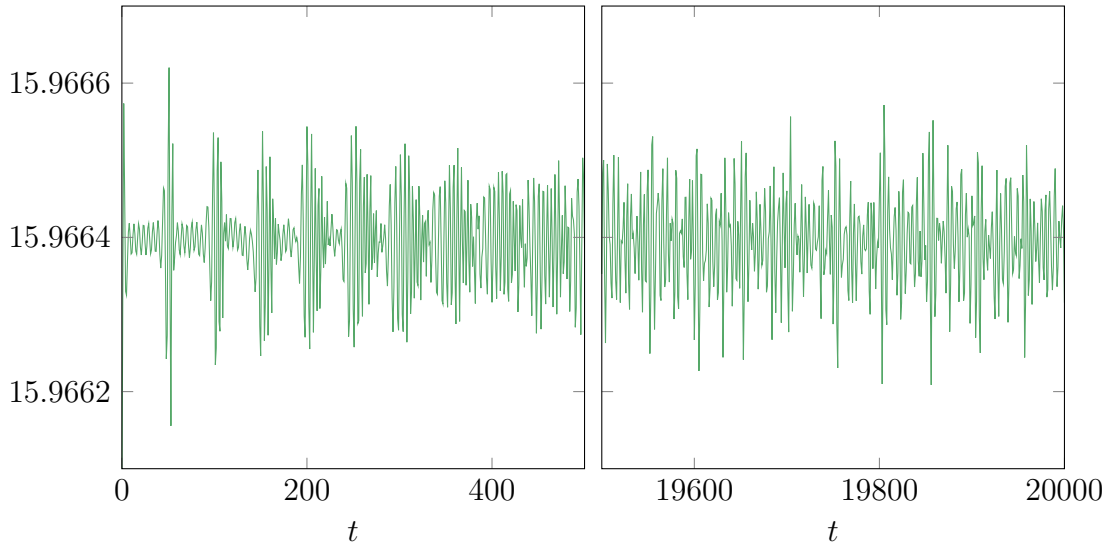


Figure 7.3: Evolution of the H^1 norm $\|u\|_{H^1}$ when solving the BBM equations with our proposed scheme.

7.2 GENERIC systems: energy & entropy stability

We consider now systems of PDEs deriving from the GENERIC formalism [GÖ97; ÖG97] which we refer to the as GENERIC PDEs. A combination of (7.1) and (6.78), these are most conveniently stated in a variational form over some space U: find $u \in C^1(\mathbb{R}_+; U)$ satisfying known initial data, such that

$$M(u; \dot{u}, v) = B(u; w_H(u), v) + D(u; w_S(u), v) \tag{7.19}$$

at all times $t \in \mathbb{R}_+$ and for all $v \in U$, where the operators $M, B, D : U \times U \times U \to \mathbb{R}$ are each linear in their final two arguments, with B (the Poisson operator) furthermore skew-symmetric and D (the friction operator) furthermore positive semidefinite. For our purposes, this system (7.19) has two relevant QoIs: $E, S : U \to \mathbb{R}$, with $E(u)$ (the energy) conserved and $S(u)$ (the entropy) non-decreasing; similarly to (7.1) the functionals $w_E, w_S : U \to U$ are such that $M(u; \cdot, w_E(u)) = E'(u; \cdot)$, $M(u; \cdot, w_S(u)) = S'(u; \cdot)$, i.e. they are the associated test functions for E, S. For these conservation and dissipation structures to hold, the GENERIC formalism imposes the additional orthogonality constraints

$$B(u; \cdot, w_S(u)) = 0, \qquad D(u; \cdot, w_E(u)) = 0. \tag{7.20}$$

are required. It is straightforward then to see the conservation and generation structures by considering $v = w_E(u), w_S(u)$ respectively.

Remark 7.3 (GENERIC PDEs without M)**.** *Similarly to Remark 7.1, the GENERIC PDE (7.19) can be stated without the choice of M by considering test functions in the dual space U^* of U: find $u \in C^1(\mathbb{R}_+; U)$ satisfying known initial data, such that*

$$L[\dot{u}] = B^*(u; H'(u), L) + D(u; S'(u), L) \tag{7.21}$$

at all times $t \in \mathbb{R}_+$ and for all $L \in U^*$, where $B : U \times U \times U \to \mathbb{R}$ and $B^* : U \times U^* \times U^* \to \mathbb{R}$ are related by

$$B^*(u; M(u; \cdot, w), M(u; \cdot, v)) = B(u; w, v), \tag{7.22a}$$

and $D : U \times U \times U \to \mathbb{R}$ and $D^* : U \times U^* \times U^* \to \mathbb{R}$ are related by

$$D^*(u; M(u; \cdot, w), M(u; \cdot, v)) = D(u; w, v). \tag{7.22b}$$

Similar to Assumption 6.17, in the construction of our SP scheme we rely on the following assumption.

Assumption 7.4 (Characterisation of GENERIC operator compatibility)**.** *Assume the existence of $\tilde{B}, \tilde{D} : \mathbb{U} \times \mathbb{U} \times \mathbb{U} \times \mathbb{U} \to \mathbb{R}$, linear in their final two arguments, such that the following hold:*

1. *$\tilde{B}, \tilde{D}$ coincide with B, D: for all u,*

$$\tilde{B}(u, w_S(u); \cdot, \cdot) = B(u; \cdot, \cdot), \qquad \tilde{D}(u, w_E(u); \cdot, \cdot) = D(u; \cdot, \cdot). \tag{7.23a}$$

2. *$\tilde{B}, \tilde{D}$ are skew-symmetric and positive semidefinite respectively in their final two arguments.*

3. *$\tilde{B}, \tilde{D}$ preserve the compatibility conditions (7.20) for all arguments: for all $u, \tilde{w}_E, \tilde{w}_S \in U$,*

$$\tilde{B}(u, \tilde{w}_S; \cdot, \tilde{w}_S) = 0, \qquad \tilde{D}(u, \tilde{w}_E; \cdot, \tilde{w}_E) = 0. \tag{7.23b}$$

With Assumption 7.4 established, we may apply our framework to construct an integrator for (7.19) that preserves both the conservation law in E and dissipation inequality in S. Both the stages in the application of the framework and the resultant scheme are largely similar to a combination of the SP integrators for GENERIC ODEs (6.81) and Poisson & gradient-descent PDEs (7.8), again relying on the introduction

of chosen LHS operator $M : \mathbb{U}^3 \to \mathbb{R}$. For brevity therefore, we state only the final scheme: find $(u, (\tilde{w}_E, \tilde{w}_S)) \in \mathbb{X}_n \times \dot{\mathbb{X}}_n^2$ (for $\mathbb{X}_n$ defined as in (3.10)) such that for all $(v, (v_E, v_S)) \in \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n^2$,

$$\mathcal{I}_n[M(u; \dot{u}, v)] = \mathcal{I}_n[\tilde{B}(u, \tilde{w}_S; \tilde{w}_E, v) + \tilde{D}(u, \tilde{w}_E; \tilde{w}_S, v)], \tag{7.24a}$$

$$\mathcal{I}_n[M(u; v_E, \tilde{w}_E)] = \int_{T_n} E'(u; v_E), \tag{7.24b}$$

$$\mathcal{I}_n[M(u; v_S, \tilde{w}_S)] = \int_{T_n} S'(u; v_S). \tag{7.24c}$$

By (7.23a) we see (7.24a) identifies with the original weak formulation (7.19) when $(\tilde{w}_E, \tilde{w}_S) = (w_E(u), w_S(u))$.

Theorem 7.5 (Energy & entropy stability of the GENERIC ODE integrator)**.** *The integrator* (6.65) *is energy and entropy stable, with* $E(u(t_{n+1})) = E(u(t_n))$ *and* $S(u(t_{n+1})) \geq S(u(t_n))$.

Proof. By considering respectively $v_E = v_S = \dot{u}$ and $v_E = \tilde{w}_E$, $v_S = \tilde{w}_S$ in (7.24),

$$\begin{aligned}
&E(u(t_{n+1})) - E(u(t_n)) && S(u(t_{n+1})) - S(u(t_n)) \\
&= \int_{T_n} \dot{E} && = \int_{T_n} \dot{S} && \text{(7.25a)} \\
&= \int_{T_n} E'(u; \dot{u}) && = \int_{T_n} S'(u; \dot{u}) && \text{(7.25b)} \\
&= \mathcal{I}_n[M(u; \dot{u}, \tilde{w}_E)] && = \mathcal{I}_n[M(u; \dot{u}, \tilde{w}_S)] && \text{(7.25c)} \\
&= \mathcal{I}_n \begin{bmatrix} \tilde{B}(u, \tilde{w}_S; \tilde{w}_E, \tilde{w}_E) \\ \qquad + \tilde{D}(u, \tilde{w}_E; \tilde{w}_S, \tilde{w}_E) \end{bmatrix} && = \mathcal{I}_n \begin{bmatrix} \tilde{B}(u, \tilde{w}_S; \tilde{w}_E, \tilde{w}_S) \\ \qquad + \tilde{D}(u, \tilde{w}_E; \tilde{w}_S, \tilde{w}_S) \end{bmatrix} && \text{(7.25d)} \\
&= 0, && \geq 0, && \text{(7.25e)}
\end{aligned}$$

where the final equality and inequality hold by Assumption 7.4. □

7.2.1 The Boltzmann equation

Inspired by [Ött05, Chap. 7] we consider as a key example application the (nondimensionalised) Boltzmann equation in d dimensions,

$$\dot{f} = -\mathbf{v} \cdot \nabla_{\mathbf{x}} f + \nabla_{\mathbf{x}} \phi \cdot \nabla_{\mathbf{v}} f + \frac{1}{\mathrm{Kn}} C(f). \tag{7.26}$$

Here, $f(\mathbf{x}, \mathbf{v}, t) \in \mathbb{R}$ represents the particle density function in the position and velocity $\mathbf{x}, \mathbf{v} \in \mathbb{R}^d$, $\nabla_{\mathbf{x}}$ and $\nabla_{\mathbf{v}}$ denote the partial derivatives with respect to $\mathbf{x}$ and

$\mathbf{v}$, $\phi(\mathbf{x}) \in \mathbb{R}$ represents a potential energy density, and Kn is the Knudsen number. The term C denotes the Boltzmann collision operator, defined by

$$C(f) \coloneqq \int_{\mathbf{v}^*,\mathbf{n}\in S^{d-1}} \beta(\mathbf{n}, \|\mathbf{v}-\mathbf{v}^*\|)(f^\dagger f^{*\dagger} - f f^*)\,\mathrm{d}\mathbf{v}^*\mathrm{d}\mathbf{n}, \tag{7.27}$$

where $\beta(\mathbf{n}, \|\mathbf{v}-\mathbf{v}^*\|) \geq 0$ is the collision kernel, $\|\cdot\|$ denotes the ℓ^2 norm, $S^{d-1} \subset \mathbb{R}^n$ is the unit $(d-1)$-sphere, and f^*, $f^\dagger$, $f^{*\dagger}$ are shorthand for

$$f^* = f|_{\mathbf{v}=\mathbf{v}^*}, \qquad f^\dagger = f|_{\mathbf{v}=\mathbf{v}^\dagger}, \qquad f^{*\dagger} = f|_{\mathbf{v}=\mathbf{v}^{*\dagger}}, \tag{7.28a}$$

where in turn $\mathbf{v}^\dagger, \mathbf{v}^{*\dagger}$ are the unique post/pre-collision velocities satisfying

$$\mathbf{v}+\mathbf{v}^* = \mathbf{v}^\dagger + \mathbf{v}^{*\dagger}, \quad \frac{1}{2}\|\mathbf{v}\|^2 + \frac{1}{2}\|\mathbf{v}^*\|^2 = \frac{1}{2}\|\mathbf{v}^\dagger\|^2 + \frac{1}{2}\|\mathbf{v}^{*\dagger}\|^2, \quad \frac{\mathbf{v}-\mathbf{v}^\dagger}{\|\mathbf{v}-\mathbf{v}^\dagger\|} = \mathbf{n}. \tag{7.28b}$$

We assume periodic BCs in $\mathbf{x}$, and a vanishing asymptotic BC in $\mathbf{v}$ of $f \to 0$ as $\|\mathbf{v}\| \to \infty$. The Boltzmann equation (7.26) has a conserved energy E and generated entropy S,

$$E \coloneqq \int_{\mathbf{x},\mathbf{v}} \left(\frac{1}{2}\|\mathbf{v}\|^2 + \phi\right) f, \qquad S \coloneqq \int_{\mathbf{x},\mathbf{v}} (1 - \log f) f. \tag{7.29}$$

To apply the scheme (7.24) to construct an energy- and entropy-stable integrator for (7.26), we must first show (7.26) fits within the GENERIC formalism, i.e. may be written in the form (7.19).

To first handle the asymptotic BCs in $\mathbf{v}$, let us parametrise f as

$$f(t;\mathbf{x},\mathbf{v}) = f_0(\mathbf{v}) \exp(u(t;\mathbf{x},\mathbf{v})) \tag{7.30}$$

for $u \in U$. The function $f_0 > 0$ characterises the asymptotic behaviour in $\mathbf{v}$, with $f_0 \to 0$ and $u = o[\log f_0]$ as $\|\mathbf{v}\| \to \infty$. Note then that $\dot{f} = f\dot{u}$.

We now cast (7.26) into a variational form. By testing against some $v \in U$ and after some classical manipulation of the collision term, we arrive at the following: find $u \in C^1(\mathbb{R}_+; U)$ satisfying known initial data, such that

$$\int_{\mathbf{x},\mathbf{v}} f\dot{u}v = \int_{\mathbf{x},\mathbf{v}} (\nabla_{\mathbf{x}}\phi \cdot \nabla_{\mathbf{v}} v - \mathbf{v}\cdot\nabla_{\mathbf{x}} v) f + \frac{1}{4\mathrm{Kn}} \int_{\mathbf{x},\mathbf{v},\mathbf{v}^*,\mathbf{n}} \beta(f^\dagger f^{*\dagger} - f f^*)(v + v^* - v^\dagger - v^{*\dagger}) \tag{7.31}$$

at all times $t \in \mathbb{R}_+$ and for all $v \in U$, where v^*, $v^\dagger$, $v^{*\dagger}$ are defined analogously to f^*, $f^\dagger$, $f^{*\dagger}$ (7.28a). This induces the choice of the LHS operator $M : U^3 \to \mathbb{R}$,

$$M(u; w, v) \coloneqq \int_{\mathbf{x},\mathbf{v}} fwv = \int_{\mathbf{x},\mathbf{v}} f_0 \exp(u) wv. \tag{7.32}$$

Now, consider the energy E and entropy S (7.29) as functions in u. These QoIs have Fréchet derivatives

$$E'(u; \delta u) = \int_{\mathbf{x},\mathbf{v}} \left(\frac{1}{2}\|\mathbf{v}\|^2 + \phi\right) f \delta u, \qquad S'(u; \delta u) = -\int_{\mathbf{x},\mathbf{v}} f \log f \delta u, \tag{7.33}$$

where again f is defined in terms of u by (7.30). Seeking $w_E(u)$, $w_S(u)$ such that $M(u; \cdot, w_E(u)) = E'(u; \cdot)$, $M(u; \cdot, w_S(u)) = S'(u; \cdot)$, the solution is immediate:

$$w_E(u) = \frac{1}{2}\|\mathbf{v}\|^2 + \phi, \qquad w_S(u) = -\log f. \tag{7.34}$$

Define then the Poisson B and friction D operators,[6]

$$B(u; \tilde{w}_E, v) \coloneqq \int_{\mathbf{x},\mathbf{v}} (\nabla_{\mathbf{x}} \tilde{w}_E \cdot \nabla_{\mathbf{v}} v - \nabla_{\mathbf{v}} \tilde{w}_E \cdot \nabla_{\mathbf{x}} v) f, \tag{7.35a}$$

$$D(u; \tilde{w}_S, v) \coloneqq \frac{1}{4\mathrm{Kn}} \int_{\mathbf{x},\mathbf{v},\mathbf{v}^*,\mathbf{n}} \beta(\exp(-\tilde{w}_S^\dagger - \tilde{w}_S^{*\dagger}) - \exp(-\tilde{w}_S - \tilde{w}_S^*)) (v + v^* - v^\dagger - v^{*\dagger}), \tag{7.35b}$$

where $\tilde{w}_S^*$, $\tilde{w}_S^\dagger$, $\tilde{w}_S^{*\dagger}$ are again defined analogously to f^*, $f^\dagger$, $f^{*\dagger}$ (7.28a) and v^*, $v^\dagger$, $v^{*\dagger}$. The skew-symmetry of B is immediate, while the positive-definiteness of D relies on the observation that $(e^{-x} - e^{-y})(y - x) \geq 0$. The GENERIC compatibility condition $B(u; \cdot, \tilde{w}_S) = 0$ holds immediately, while $D(u; \cdot, \tilde{w}_E) = 0$ can be seen from the conservation of energy condition $\frac{1}{2}\|\mathbf{v}\|^2 + \frac{1}{2}\|\mathbf{v}^*\|^2 = \frac{1}{2}\|\mathbf{v}^\dagger\|^2 + \frac{1}{2}\|\mathbf{v}^{*\dagger}\|^2$ (7.28b). With M, B, D as defined, the Boltzmann equation is a GENERIC PDE of the form (7.19); we can thus apply (7.24) to preserve the energy and entropy stability.

We must define $\tilde{B}, \tilde{D} : \mathbb{U}^4 \to \mathbb{R}$ satisfying Assumption 7.4. Take $\tilde{w}_H, \tilde{w}_S$ to be approximations to $w_H(u), w_S(u)$. The incorporation of $\tilde{w}_S$ into B is simple, with $\tilde{B}$ defined

$$\tilde{B}(u, \tilde{w}_S; \tilde{w}_E, v) \coloneqq \int_{\mathbf{x},\mathbf{v}} (\nabla_{\mathbf{x}} \tilde{w}_E \cdot \nabla_{\mathbf{v}} v - \nabla_{\mathbf{v}} \tilde{w}_E \cdot \nabla_{\mathbf{x}} v) \exp(-\tilde{w}_S). \tag{7.36}$$

Considering $\tilde{w}_S = w_S(u) = -\log f$, we see this identifies with B as $\exp(-\tilde{w}_S) = \exp(\log f) = f$; we see $\tilde{B}$ evaluates to 0 for $v = \tilde{w}_S$ by the substitution $\exp(-\tilde{w}_S)\nabla\tilde{w}_S =$

[6] Technically speaking, this friction operator D is not linear in $\tilde{w}_S$, a requirement that was imposed on D in our definition. However, this has no effect on the SP properties of our discretisation.

$-\nabla[\exp(-\tilde{w}_S)]$ and IBP in $\mathbf{x}$, $\mathbf{v}$ noting the periodic BCs and assuming $\tilde{w}_S \to \infty$ as $\|\mathbf{v}\| \to \infty$ such that $\exp(-\tilde{w}_S) \to 0$.

The incorporation of $\tilde{w}_E$ into D is somewhat more involved. Let $\Sigma \subset \mathbb{R}^3 \times (\mathbb{R}^3)^4 \times S^{d-1}$ denote the ($3d$-dimensional) manifold of tuples $(\mathbf{x}, (\mathbf{v}, \mathbf{v}^*, \mathbf{v}^\dagger, \mathbf{v}^{*\dagger}), \mathbf{n})$ satisfying the relations (7.28b), which we endow with the metric induced from $\Sigma \subset \mathbb{R}^3 \times (\mathbb{R}^3)^4 \times S^{d-1}$. The friction operator D can then be written as an integral over Σ,

$$D(u; \tilde{w}_S, v) \coloneqq \frac{1}{4\mathrm{Kn}} \int_\Sigma \beta(\exp(-\tilde{w}_S^\dagger - \tilde{w}_S^{*\dagger}) - \exp(-\tilde{w}_S - \tilde{w}_S^*)) \\ (v + v^* - v^\dagger - v^{*\dagger}). \quad (7.37)$$

We may similarly define an auxiliary ($3d$-dimensional) manifold $\tilde{\Sigma} \subset \mathbb{R}^3 \times (\mathbb{R}^3)^4 \times S^{d-1}$ of tuples $(\mathbf{x}, (\mathbf{v}, \mathbf{v}^*, \mathbf{v}^{\tilde{\dagger}}, \mathbf{v}^{*\tilde{\dagger}}), \mathbf{n})$ satisfying the auxiliary relations

$$\mathbf{v} + \mathbf{v}^* = \mathbf{v}^{\tilde{\dagger}} + \mathbf{v}^{*\tilde{\dagger}}, \quad \tilde{w}_H|_{\mathbf{v}=\mathbf{v}} + \tilde{w}_H|_{\mathbf{v}=\mathbf{v}^*} = \tilde{w}_H|_{\mathbf{v}=\mathbf{v}^{\tilde{\dagger}}} + \tilde{w}_H|_{\mathbf{v}=\mathbf{v}^{*\tilde{\dagger}}}, \quad \frac{\mathbf{v} - \mathbf{v}^{\tilde{\dagger}}}{\|\mathbf{v} - \mathbf{v}^{\tilde{\dagger}}\|} = \mathbf{n}, \quad (7.38a)$$

again with the metric induced from $\Sigma \subset \mathbb{R}^3 \times (\mathbb{R}^3)^4 \times S^{d-1}$. Similarly to $f^*, f^\dagger, f^{*\dagger}$ (7.28a) we take $\psi^*, \psi^{\tilde{\dagger}}, \psi^{*\tilde{\dagger}}$, for an arbitrary function ψ in $\mathbf{v}$, as shorthand for

$$\psi^* \coloneqq \psi|_{\mathbf{v}=\mathbf{v}^*}, \qquad \psi^{\tilde{\dagger}} \coloneqq \psi|_{\mathbf{v}=\mathbf{v}^{\tilde{\dagger}}}, \qquad \psi^{*\tilde{\dagger}} \coloneqq \psi|_{\mathbf{v}=\mathbf{v}^{*\tilde{\dagger}}}. \quad (7.38b)$$

We may then introduce $\tilde{w}_H$ implicitly into the definition of $\tilde{D}$ implicitly through $\tilde{\Sigma}$:

$$\tilde{D}(u, \tilde{w}_E; \tilde{w}_S, v) \coloneqq \frac{1}{4\mathrm{Kn}} \int_{\tilde{\Sigma}} \beta(\exp(-\tilde{w}_S^{\tilde{\dagger}} - \tilde{w}_S^{*\tilde{\dagger}}) - \exp(-\tilde{w}_S - \tilde{w}_S^*)) \\ (v + v^* - v^{\tilde{\dagger}} - v^{*\tilde{\dagger}}). \quad (7.39)$$

Considering $\tilde{w}_E = w_E(u) = \frac{1}{2}\|\mathbf{v}\|^2 + \phi(\mathbf{x})$, we see this identifies with D as the conditions on $\mathbf{v}^*, \mathbf{v}^{\tilde{\dagger}}, \mathbf{v}^{*\tilde{\dagger}}$ (7.38a) align with those on $\mathbf{v}^*, \mathbf{v}^\dagger, \mathbf{v}^{*\dagger}$ (7.28b); we see $\tilde{D}$ evaluates to 0 for $v = \tilde{w}_E$ as the conditions on $\mathbf{v}^*, \mathbf{v}^{\tilde{\dagger}}, \mathbf{v}^{*\tilde{\dagger}}$ (7.38a) imply $\tilde{w}_E + \tilde{w}_E^* - \tilde{w}_E^{\tilde{\dagger}} - \tilde{w}_E^{*\tilde{\dagger}} = 0$ on $\tilde{\Sigma}$.

We finally derive the following SP scheme for the Boltzmann scheme (7.26): find $(u, (\tilde{w}_E, \tilde{w}_S)) \in \mathbb{X}_n \times \dot{\mathbb{X}}_n^2$ such that for all $(v, (v_H, v_S)) \in \dot{\mathbb{X}}_n$,

$$\mathcal{I}_n\Big[\int_{\mathbf{x},\mathbf{v}} f\dot{u}v\Big] = \mathcal{I}_n\Big[\int_{\mathbf{x},\mathbf{v}} (\nabla_{\mathbf{x}}\tilde{w}_E \cdot \nabla_{\mathbf{v}} v - \nabla_{\mathbf{v}}\tilde{w}_E \cdot \nabla_{\mathbf{x}} v)\tilde{f} \\ + \frac{1}{4\mathrm{Kn}} \int_{\tilde{\Sigma}} \beta(\tilde{f}^{\tilde{\dagger}}\tilde{f}^{*\tilde{\dagger}} - \tilde{f}\tilde{f}^*)(v + v^* - v^{\tilde{\dagger}} - v^{*\tilde{\dagger}})\Big], \quad (7.40a)$$

$$\mathcal{I}_n\Big[\int_{\mathbf{x},\mathbf{v}} f\tilde{w}_E v_H\Big] = \int_{T_n}\int_{\mathbf{x},\mathbf{v}} f\Big(\frac{1}{2}\|\mathbf{v}\|^2 + \phi(\mathbf{x})\Big)v_H, \quad (7.40b)$$

$$\mathcal{I}_n\Big[\int_{\mathbf{x},\mathbf{v}} f\tilde{w}_S v_S\Big] = -\int_{T_n}\int_{\mathbf{x},\mathbf{v}} f\log f v_S, \quad (7.40c)$$

where again f is defined as in (7.30), the auxiliary density function $\tilde{f}$ is shorthand for $\tilde{f} \coloneqq \exp(-\tilde{w}_S)$, the functions $\tilde{f}^*, \tilde{f}^{\dagger}, \tilde{f}^{*\dagger}$ and v^*, $v^{\dagger}$, $v^{*\dagger}$ are defined via (7.38b) for $\mathbf{v}^*$, $\mathbf{v}^{\dagger}$, $\mathbf{v}^{*\dagger}$ defined as in (7.38a), and $\tilde{\Sigma} \subset \mathbb{R}^3 \times (\mathbb{R}^3)^4 \times S^{d-1}$ is the auxiliary ($3d$-dimensional) manifold $(\mathbf{x}, (\mathbf{v}, \mathbf{v}^*, \mathbf{v}^{\dagger}, \mathbf{v}^{*\dagger}), \mathbf{n})$ satisfying (7.38a).

The conservation of H and non-dissipation of S can then be shown by testing with $(v, v_H) = (\tilde{w}_E, \dot{u})$ and $(v, v_S) = (\tilde{w}_S, \dot{u})$ respectively.

7.3 The compressible Navier–Stokes equations: mass, momentum, energy & entropy stability

We now consider SP schemes for the compressible NS equations. We seek a scheme that will conserve the mass, momentum and energy, and preserve the behaviour of the entropy; specifically, we would like the entropy to be conserved in the ideal limit, and non-decreasing otherwise. The non-dissipation of entropy is a crucial aspect in the analysis and behaviour of solutions to the compressible NS equations [Fei+21] with quantitative implications on the regularity of solutions and qualitative implications on the dissipation rate; it is therefore essential we preserve this.

The compressible NS equations can be written in the following nondimensionalised form over a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$:

$$\dot{\rho} = -\operatorname{div}[\rho \mathbf{u}], \tag{7.41a}$$

$$\rho \dot{\mathbf{u}} = -\rho \mathbf{u} \cdot \nabla \mathbf{u} - \nabla p + \operatorname{div}\left[\frac{2}{\mathrm{Re}} \rho \tau[\mathbf{u}]\right], \tag{7.41b}$$

$$\dot{\varepsilon} = -\operatorname{div}[\varepsilon \mathbf{u}] - p \operatorname{div} \mathbf{u} + \frac{2}{\mathrm{Re}} \rho \upsilon[\mathbf{u}, \mathbf{u}] + \operatorname{div}\left[\frac{1}{\mathrm{RePr}} \rho \nabla \theta\right]. \tag{7.41c}$$

Here, ρ, p, $\mathbf{u}$, ε and θ are the density, pressure, velocity, internal energy density and temperature respectively, $\mathrm{Re} > 0$ and $\mathrm{Pr} > 0$ are the Reynolds and Prandtl numbers (potentially functions of ρ and ε) the deviatoric strain $\tau : \mathbb{R}^d \to \mathbb{R}^d_{\mathrm{sym}}$ is defined

$$\tau[\mathbf{u}] \coloneqq \frac{1}{2} \nabla \mathbf{u} + \frac{1}{2} \nabla \mathbf{u}^{\top} - \frac{1}{3} (\operatorname{div} \mathbf{u}) I, \tag{7.42a}$$

trace-free when $d = 3$, and the positive semidefinite bilinear form $\upsilon : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is defined

$$\upsilon[\mathbf{u}, \mathbf{v}] \coloneqq \left(\frac{1}{2} \nabla \mathbf{u} + \frac{1}{2} \nabla \mathbf{u}^{\top}\right) : \left(\frac{1}{2} \nabla \mathbf{v} + \frac{1}{2} \nabla \mathbf{v}^{\top}\right) - \frac{1}{3} (\operatorname{div} \mathbf{u})(\operatorname{div} \mathbf{v}) = \tau[\mathbf{u}] : \nabla \mathbf{v}. \tag{7.42b}$$

We assume the Stokes hypothesis, that the bulk viscosity is zero [Sto45]. This is for brevity only and is not necessary; the ideas we present in this section readily extend to more complex stress tensors. For simplicity, we assume periodic BCs.

The system (7.41) is completed by constitutive relations relating two of ρ, θ, p, ε to the others.

Example (Ideal fluid)

The constitutive relations for a nondimensionalised ideal fluid can be written as

$$p = \rho\theta, \qquad \varepsilon = C_V p, \tag{7.43}$$

where C_V is the nondimensionalised heat capacity at constant volume ($\frac{3}{2}$ for a monatomic gas).

It will be convenient for our purposes to further define the inverse temperature $\beta := \theta^{-1}$. We also define the specific entropy s, corresponding to the total entropy $\int_\Omega \rho s$, satisfying the (intensive) fundamental thermodynamic relation

$$\beta \mathrm{d}\varepsilon = \mathrm{d}[\rho s] - g\mathrm{d}\rho. \tag{7.44a}$$

Here, $g = s - (\varepsilon + p)\beta/\rho$ is the negation of the specific free energy, or Gibbs free energy per unit mass, per unit temperature. Taking differentials gives the second thermodynamic relation,

$$\rho \mathrm{d}g + \varepsilon \mathrm{d}\beta + \mathrm{d}[p\beta] = 0. \tag{7.44b}$$

Example (Ideal fluid)

For an ideal gas, s and g evaluate as

$$s = \log\left(\frac{\theta^{C_V}}{\rho}\right), \qquad g = s - (C_V + 1). \tag{7.45}$$

We now apply the framework to construct a FE discretisation for (7.41) with the desired SP properties.

Application of framework (Algorithm 3.5)

A. To define the semidiscrete form, we must first choose a convenient parametrisation. Many options are available here, such as primitive or conservative variables. We shall choose $\sigma = \rho^{\frac{1}{2}}$, $\boldsymbol{\mu} = \rho^{\frac{1}{2}}\mathbf{u}$, and $\zeta = \log(\varepsilon)$. This parametrisation is chosen with some hindsight. The choice of $\boldsymbol{\mu}$ ensures the energy is independent of the density, limiting the number of AVs that must be introduced; the choice of σ balances this in a way that later simplifies the conservation of momentum; the choice of ζ ensures the internal energy remains positive. Writing (7.41) in terms of σ, $\boldsymbol{\mu}$, ζ yields

$$(\dot{\rho} =) \quad 2\sigma\dot{\sigma} = -\operatorname{div}[\rho\mathbf{u}], \tag{7.46a}$$

$$\sigma\dot{\boldsymbol{\mu}} = -\frac{1}{2}(\rho\mathbf{u}\cdot\nabla\mathbf{u} + \operatorname{div}[\rho\mathbf{u}^{\otimes 2}]) - \nabla p + \operatorname{div}\left[\frac{2}{\mathrm{Re}}\rho\tau[\mathbf{u}]\right], \tag{7.46b}$$

$$(\dot{\varepsilon} =) \quad \varepsilon\dot{\zeta} = -\operatorname{div}[\varepsilon\mathbf{u}] - p\operatorname{div}\mathbf{u} + \frac{2}{\mathrm{Re}}\rho\upsilon[\mathbf{u},\mathbf{u}] + \operatorname{div}\left[\frac{1}{\mathrm{RePr}}\rho\nabla\theta\right], \tag{7.46c}$$

where $\rho = \sigma^2$, $\mathbf{u} = \sigma^{-1}\boldsymbol{\mu}$, $\varepsilon = \exp(\zeta)$, $\mathbf{u}^{\otimes 2}$ denotes the outer product $\mathbf{u}\otimes\mathbf{u}$, and it is assumed that known constitutive relations determine p, θ as functions of ρ, ε. For some continuous,[7] spatially periodic FE space $\mathbb{V} \subset C(\overline{\Omega})$, we define the mixed FE space $\mathbb{U} \coloneqq \mathbb{V}^{1+d+1}$; we use the same space for each variable both for simplicity, and as it will help in ensuring momentum conservation. We may then define a semidiscrete variational problem: find $(\sigma, \boldsymbol{\mu}, \zeta) \in \mathbb{U}$, for $\mathbb{U}$ defined as in (3.3), such that

$$M((\sigma,\zeta);(\dot{\sigma},\dot{\boldsymbol{\mu}},\dot{\zeta}),(v_\rho,\mathbf{v}_m,v_\varepsilon)) = F((\sigma,\boldsymbol{\mu},\zeta);(v_\rho,\mathbf{v}_m,v_\varepsilon)) \tag{7.47}$$

at all times $t \in \mathbb{R}_+$ and for all $(v_\rho, \mathbf{v}_m, v_\varepsilon) \in \mathbb{U}$, where M, F are defined

$$M \coloneqq \int_\Omega 2\sigma\dot{\sigma}v_\rho + \int_\Omega \sigma\dot{\boldsymbol{\mu}}\cdot\mathbf{v}_m + \int_\Omega \varepsilon\dot{\zeta}v_\varepsilon, \tag{7.48a}$$

$$\begin{aligned} F \coloneqq &\int_\Omega \rho\mathbf{u}\cdot\nabla v_\rho \\ &+ \int_\Omega \frac{1}{2}\rho\mathbf{u}\cdot(\nabla\mathbf{v}_m\cdot\mathbf{u} - \nabla\mathbf{u}\cdot\mathbf{v}_m) + p\operatorname{div}\mathbf{v}_m - \frac{2}{\mathrm{Re}}\rho\upsilon[\mathbf{u},\mathbf{v}_m] \\ &+ \int_\Omega (\varepsilon\mathbf{u}\cdot\nabla v_\varepsilon + pv_\varepsilon\operatorname{div}\mathbf{u}) + \frac{2}{\mathrm{Re}}\rho\upsilon[\mathbf{u},\mathbf{u}]v_\varepsilon - \frac{1}{\mathrm{RePr}}\rho\nabla\theta\cdot\nabla v_\varepsilon. \end{aligned} \tag{7.48b}$$

[7] Discontinuous spaces $\mathbb{V} \not\subset C(\overline{\Omega})$ are often preferred for discretisation. This necessitates the introduction of facet and penalty terms to handle the non-conformity; such an extension is possible, but omitted here for brevity.

Example (Ideal fluid)

Writing the equations of state for an ideal fluid (7.43) in terms of $\rho = \sigma^2$ and $\varepsilon = \exp(\zeta)$ yields

$$p = \frac{\varepsilon}{C_V}, \qquad \theta = \frac{p}{\rho}. \tag{7.49}$$

B. Two of the AVs proposed by our framework, in particular for mass, momentum and energy conservation, will be uniform and constant, i.e. 1; with regard to the argument in Subsection 4.2.2 therefore, we choose $\mathcal{I}_n$ to be the exact integral $\int_{T_n}$, eliminating the need to introduce corresponding AVs into our discretisation. Over the timestep T_n therefore, we cast (7.47) into a fully discrete form (i.e. a CPG discretisation): find $(\sigma, \boldsymbol{\mu}, \zeta) \in \mathbb{X}_n$ such that

$$\int_{T_n} M((\sigma,\zeta);(\dot\sigma,\dot{\boldsymbol{\mu}},\dot\zeta),(v_\rho,\mathbf{v}_m,v_\varepsilon)) = \int_{T_n} F((\sigma,\boldsymbol{\mu},\varepsilon);(v_\rho,\mathbf{v}_m,v_\varepsilon)), \tag{7.50}$$

for all $(v_\rho, \mathbf{v}_m, v_\varepsilon) \in \dot{\mathbb{X}}_n$, with $\mathbb{X}_n$ defined as in (3.10). For simplicity, we assume henceforth that a sufficiently small timestep and fine mesh are chosen so that σ remains positive, implying the constitutive relations remain well-defined.

C. Including each component of the momentum, we have $3 + d$ QoIs,

$$Q_1 \coloneqq \int_\Omega \sigma^2, \qquad \mathbf{Q}_2 \coloneqq \int_\Omega \sigma\boldsymbol{\mu}, \qquad Q_3 \coloneqq \int_\Omega \frac{1}{2}\|\boldsymbol{\mu}\|^2 + \varepsilon, \qquad Q_4 \coloneqq \int_\Omega \rho s, \tag{7.51}$$

the mass, momentum, energy and entropy respectively, where $\|\cdot\|$ denotes the $\mathbf{L}^2(\Omega)$ norm, s is a function of $\rho = \sigma^2$ and $\varepsilon = \exp(\zeta)$. By evaluating the Fréchet derivatives, we identify these with the respective associated test functions

$$(1,\mathbf{0},0), \qquad (\frac{1}{2}u_i,\mathbf{e}_i,0) \text{ for each } i, \qquad (0,\mathbf{u},1), \qquad (g,\mathbf{0},\beta), \tag{7.52}$$

where $\mathbf{e}_i$ denotes the i-th basis vector, (u_i) denote the components of $\mathbf{u} = u_i \sum_i \mathbf{e}_i$, and again $\beta = \theta^{-1}$ is the inverse temperature. The associated test functions for Q_4 are found from (7.44a).

D. We introduce AVs for each of the associated test functions in (7.52) according to (3.19), where they are required according to the argument in Subsection 4.2.2, i.e. when they are not equal to 1 or 0. Those remaining associated test functions

include two for $\mathbf{u}$, and one each for g, β. Furthermore, we can see that the variational relations (3.19) satisfied by each of the AVs for $\mathbf{u}$ are identical, so these two AVs are identical. This leaves three AVs, $(\tilde{g}, \tilde{\mathbf{u}}, \tilde{\beta}) \in \dot{\mathbb{X}}_n$ satisfying

$$\int_{T_n} M((\sigma,\zeta);(v_g,\mathbf{v}_u,v_\beta),(\tilde{g},\tilde{\mathbf{u}},\tilde{\beta})) = \int_{T_n} M((\sigma,\zeta);(v_g,\mathbf{v}_u,v_\beta),(g,\mathbf{u},\beta)), \tag{7.53}$$

for all $(v_g, \mathbf{v}_u, v_\beta) \in \dot{\mathbb{X}}_n$, where again $\beta = \theta^{-1}$, g are functions of $\rho = \sigma^2$, $\varepsilon = \exp(\zeta)$, and $\mathbf{u} = \sigma^{-1}\boldsymbol{\mu}$. Like β, we assume that $\tilde{\beta} > 0$.

Example (Ideal fluid)

Recall (7.45). The negative specific free energy per unit temperature g can be defined for an ideal fluid in terms of $\rho = \sigma^2$, $\varepsilon = \exp(\zeta)$,

$$g = \log\left(\frac{\varepsilon^{C_V}}{\rho^{C_V+1}}\right) - (C_V \log C_V + C_V + 1). \tag{7.54}$$

E. We now introduce $\tilde{g}$, $\tilde{\beta}$ into F. The primal variables $\rho = \sigma^2$, $p(\sigma,\zeta)$, $\varepsilon = \exp(\zeta)$ were defined to be functions of σ, ζ, with p determined by the fluid's constitutive relations; in contrast, let $\tilde{\rho}(\tilde{\beta},\tilde{g})$, $\tilde{p}(\tilde{\beta},\tilde{g})$, $\tilde{\varepsilon}(\tilde{\beta},\tilde{g})$ denote an auxiliary density, pressure, energy density determined by the fluid's constitutive relations as functions of the auxiliary inverse temperature $\tilde{\beta}$ and auxiliary negative specific free energy per unit temperature $\tilde{g}$. Crucially, in this sense $\tilde{\rho}$ differs from $\rho = \sigma^2$, $\tilde{p}$ from $p(\sigma,\zeta)$, and $\tilde{\varepsilon}$ from $\varepsilon = \exp(\zeta)$. By inspection, we define $\tilde{F}((\sigma,\boldsymbol{\mu},\zeta),(\tilde{g},\tilde{\mathbf{u}},\tilde{\beta});(v_\rho,\mathbf{v}_m,v_u))$ to be

$$\begin{aligned}\tilde{F} \coloneqq &\int_\Omega \tilde{\rho}\tilde{\mathbf{u}}\cdot\nabla v_\rho \\ &+ \int_\Omega \frac{1}{2}\tilde{\rho}\tilde{\mathbf{u}}\cdot(\nabla\mathbf{v}_m\cdot\tilde{\mathbf{u}} - \nabla\tilde{\mathbf{u}}\cdot\mathbf{v}_m) + \tilde{p}\,\mathrm{div}\,\mathbf{v}_m - \frac{2}{\mathrm{Re}}\rho\upsilon[\tilde{\mathbf{u}},\mathbf{v}_m] \\ &+ \int_\Omega (\tilde{\varepsilon}\tilde{\mathbf{u}}\cdot\nabla v_\varepsilon - \tilde{p}v_\varepsilon\,\mathrm{div}\,\tilde{\mathbf{u}}) + \frac{2}{\mathrm{Re}}\rho\upsilon[\tilde{\mathbf{u}},\tilde{\mathbf{u}}]v_\varepsilon + \frac{1}{\mathrm{RePr}}\rho\theta^2\nabla\tilde{\beta}\cdot\nabla v_\varepsilon.\end{aligned} \tag{7.55}$$

Substituting $(v_\rho, \mathbf{v}_m, v_\varepsilon)$ for each set of AVs for each QoI,

$$\tilde{F}(\ldots;(1,\mathbf{0},0)) = 0, \tag{7.56a}$$

$$\tilde{F}(\ldots;(\frac{1}{2}\tilde{\mathbf{u}},I,0)) = 0, \tag{7.56b}$$

$$\tilde{F}(\ldots;(0,\tilde{\mathbf{u}},1)) = 0, \tag{7.56c}$$

$$\tilde{F}(\ldots;(\tilde{g},\mathbf{0},\tilde{\beta})) = \frac{1}{\mathrm{Re}}\int_\Omega \rho\tilde{\beta}\upsilon[\tilde{\mathbf{u}},\tilde{\mathbf{u}}] + \frac{1}{\mathrm{Pr}}\rho\theta^2\|\nabla\tilde{\beta}\|^2 \geq 0. \tag{7.56d}$$

These identities are immediate by evaluation of the LHS. The evaluation of (7.56d) includes the integral

$$\int_\Omega \tilde{\mathbf{u}} \cdot \left(\tilde{\rho} \nabla \tilde{g} + \tilde{\varepsilon} \nabla \tilde{\beta} + \nabla[\tilde{p}\tilde{\beta}] \right). \tag{7.57}$$

Any set of intensive thermodynamic quantities satisfying a valid constitutive law must satisfy the thermodynamic relation (7.44b). As $\tilde{\rho}$, $\tilde{g}$, $\tilde{\varepsilon}$, $\tilde{\beta}$, $\tilde{p}$ are constructed to satisfy such a law, we see $\tilde{\rho}\nabla\tilde{g} + \tilde{\varepsilon}\nabla\tilde{\beta} + \nabla[\tilde{p}\tilde{\beta}] = \mathbf{0}$ everywhere by construction, and (7.57) must evaluate to zero.

Example (Ideal fluid)

The auxiliary $\tilde{\rho}$, $\tilde{p}$, $\tilde{\varepsilon}$ can be written in $\tilde{g}$, $\tilde{\beta}$ as

$$\tilde{\rho} = \tilde{\beta}^{-C_V} \exp(-\tilde{g} - (C_V + 1)), \qquad \tilde{p} = \frac{\tilde{\rho}}{\tilde{\beta}}, \qquad \tilde{\varepsilon} = C_V \tilde{p}. \tag{7.58}$$

F. The final SP scheme is as follows: find $((\sigma, \boldsymbol{\mu}, \zeta), (\tilde{g}, \tilde{\mathbf{u}}, \tilde{\beta})) \in \mathbb{X}_n \times \dot{\mathbb{X}}_n$ such that

$$\int_{T_n} M((\sigma,\zeta);(\dot{\sigma},\dot{\boldsymbol{\mu}},\dot{\zeta}),(v_\rho,\mathbf{v}_m,v_\varepsilon)) = \int_{T_n} \tilde{F}((\sigma,\boldsymbol{\mu},\zeta),(\tilde{g},\tilde{\mathbf{u}},\tilde{\beta});(v_\rho,\mathbf{v}_m,v_\varepsilon)), \tag{7.59a}$$

$$\int_{T_n} M((\sigma,\zeta);(v_g,\mathbf{v}_u,v_\beta),(\tilde{g},\tilde{\mathbf{u}},\tilde{\beta})) = \int_{T_n} M((\sigma,\zeta);(v_g,\mathbf{v}_u,v_\beta),(g,\mathbf{u},\beta)), \tag{7.59b}$$

for all $((v_\rho, \mathbf{v}_m, v_\varepsilon), (v_g, \mathbf{v}_u, v_\beta)) \in \dot{\mathbb{X}}_n \times \dot{\mathbb{X}}_n$.

Theorem 7.6 (Mass, momentum, energy and entropy stability of the compressible NS integrator)**.** *The integrator* (7.59) *is mass, momentum, energy and entropy stable, with each of Q_1, $\mathbf{Q}_2$, Q_3 conserved across timesteps T_n, and Q_4 generated at a rate*

$$Q_4|_{t=t_{n+1}} - Q_4|_{t=t_n} = \frac{1}{\mathrm{Re}} \int_{T_n} \int_\Omega \rho\tilde{\beta} v[\tilde{\mathbf{u}}, \tilde{\mathbf{u}}] + \frac{1}{\mathrm{Pr}} \rho\theta^2 \|\nabla\tilde{\beta}\|^2 \geq 0. \tag{7.60}$$

Proof. Each of these results holds from the results (7.56) by testing in (7.59) respectively against

$$v_\rho = 1, \tag{7.61a}$$

$$(v_\rho, \mathbf{v}_m, \mathbf{v}_u) = (\frac{1}{2}\tilde{u}_i, \mathbf{e}_i, \dot{\sigma}\mathbf{e}_i) \text{ for each } i, \tag{7.61b}$$

$$(\mathbf{v}_m, v_\varepsilon, \mathbf{v}_u) = (\tilde{\mathbf{u}}, 1, \dot{\boldsymbol{\mu}}), \tag{7.61c}$$

$$(v_\rho, v_\varepsilon, v_g, v_\beta) = (\tilde{g}, \tilde{\beta}, \dot{\sigma}, \dot{\zeta}), \tag{7.61d}$$

where $(\tilde{u}_i)$ denotes the components of $\tilde{\mathbf{u}}$. □

To numerically verify the stability results, we run two tests for an ideal gas with $C_V = 2.5$ (typical for air at room temperature) over a unit square domain $\Omega = (0,1)^2$. We take $\mathbb{V}$ to be the lowest order, degree-1 CG (or Lagrange) space (see Ern & Guermond [EG21a, Sec. 6 & 7]) comparing the scheme (7.59) at $S = 1$ (i.e. at lowest order in time) with an IM discretisation of (7.46).

7.3.1 Shockwave test

We consider first a supersonic perturbation in the velocity field, with $\mathrm{Pr} = 0.71$ (typical for air) and $\mathrm{Re} = 2^7$. ICs are

$$\sigma(0) = 1, \tag{7.62a}$$

$$\boldsymbol{\mu}(0) = 2^3 \exp(\cos(2\pi x) + \cos(2\pi(y - 0.5)) - 2)\mathbf{e}_1, \tag{7.62b}$$

$$\zeta(0) = 0, \tag{7.62c}$$

up to projection. The FE space $\mathbb{V}$ is defined over a grid of square cells of uniform width 2^{-8}; we take a uniform timestep $\Delta t = 2^{-11}$.

Fig. 7.4 shows plots of the velocity, density, temperature, and specific entropy at various times in the SP scheme.[8] The shockwave is clearly visible at the final time. We use continuous approximations to all variables, causing oscillations in ρ and s; this could potentially be improved with a non-conforming DG spatial discretisation (see Ern & Guermond [EG21a, Sec. 6 & 7]).

Fig. 7.5 shows the error in the mass Q_1, momentum $\mathbf{Q}_2$, and energy Q_3 for each simulation. Each is conserved (up to quadrature error, solver tolerances and machine precision) for the scheme (7.59) while only the mass is conserved for the IM scheme.[9] The error in the energy increases exponentially in the IM scheme from the point of formation of the shockwave, rising from a value of around 4.046 to around 4.059.

[8] The results from the IM scheme exhibit no visible visual difference.

[9] One can verify that any order of Gauss method applied to (7.46) will be mass-conserving.

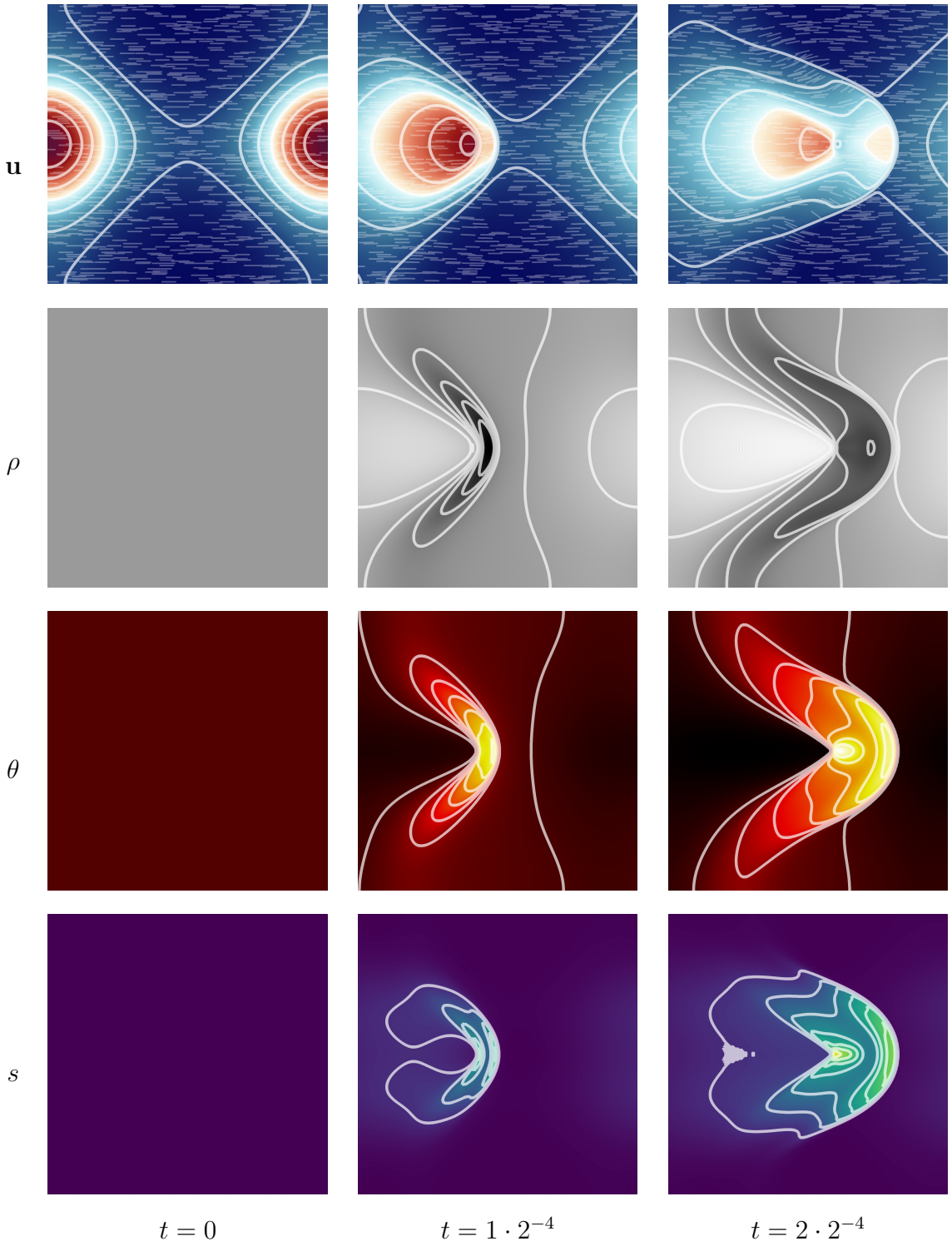
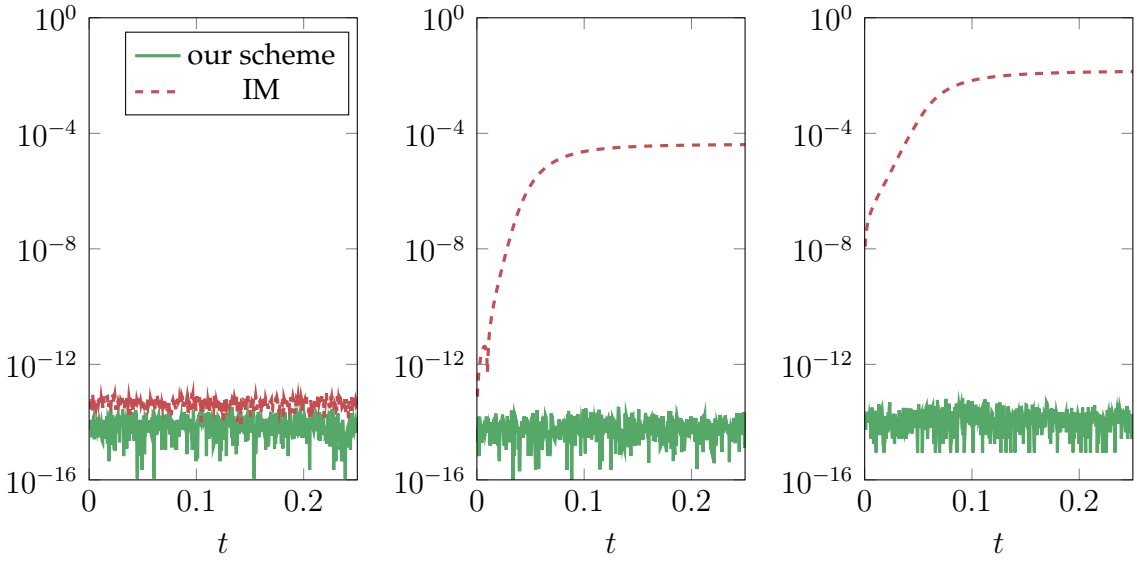


Figure 7.4: Contours of the velocity magnitude $\|\mathbf{u}\|$, density ρ, temperature θ, and specific entropy s at times $t \in \{0, 1 \cdot 2^{-4}, 2 \cdot 2^{-4}\}$ in the SP simulation of the supersonic test (Subsection 7.3.1).



(a) Mass, $|Q_1 - Q_1(0)|$ **(b)** Momentum, $\|\mathbf{Q}_2 - \mathbf{Q}_2(0)\|$ **(c)** Energy, $|Q_3 - Q_3(0)|$

Figure 7.5: Errors in different invariants over time within the supersonic test (Subsection 7.3.1) for IM and our proposed scheme.

7.3.2 Euler test

To better illustrate the preservation of the entropy structure, we consider an adiabatic (uniform s) perturbation in the state functions σ, ζ, with $\mathrm{Re} = \infty$, i.e. discarding viscous and thermally dissipative terms. We take the ICs to be

$$\sigma(0) = \exp\left(\frac{1}{2}\sin(2\pi x)\sin(2\pi y)\right), \tag{7.63a}$$

$$\boldsymbol{\mu}(0) = \mathbf{0}, \tag{7.63b}$$

$$\zeta(0) = \left(1 + \frac{1}{C_V}\right)\sin(2\pi x)\sin(2\pi y), \tag{7.63c}$$

again up to projection. The FE space $\mathbb{V}$ is defined over a uniform grid of triangular cells of width 2^{-5}; we take a uniform timestep $\Delta t = 2^{-7}$.

With $\mathrm{Re} = \infty$, entropy Q_4 should be conserved both in an exact solution, and in the scheme (7.59). Fig. 7.6 shows the error in the entropy for each simulation. The lines terminate when the nonlinear solver fails to converge, potentially due to a solution to the scheme no longer existing; we observe that the SP scheme fails after 515 timesteps, whereas the IM scheme fails after 392. Our scheme (7.59) conserves entropy throughout (up to quadrature error, solver tolerances, and machine precision) whereas IM does not, introducing spurious (unphysical) entropy decrease.

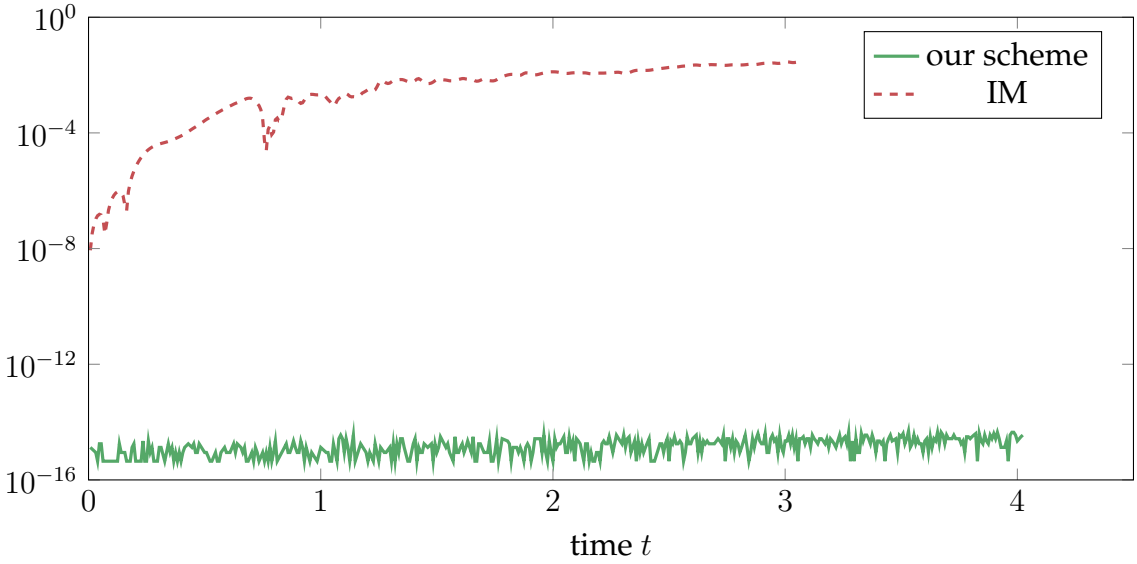


Figure 7.6: Error in the entropy $|Q_4 - Q_4(0)|$ over time within the inviscid test (Subsection 7.3.2) for IM and our proposed scheme.

Part III

Extensions of the framework: adiabatic invariants & FEEC

"A Globetrotter always saves the good algebra for the final minutes."

— Ethan 'Bubblegum' Tate (Phillip 'Phil' LaMarr) [Kee01]

8 Introduction

Contents

With various simple example applications discussed in Part II, we now turn our attention to two extensions that go beyond the framework as presented in Part I.

Adiabatic invariants & the Lorentz problem

The first of these extensions lies in adiabatic invariants [Hen93; AKN06], quantities that are neither conserved nor dissipated, but exhibit rapid, bounded oscillations about a certain more slowly changing value. Since these adiabatic invariants remain within a bounded interval, this property can have many of the same implications for the dynamic behaviour of solutions as typical conservation structures. However, they fall outside the remit of the framework presented in Chapter 3.

Chapter 9 concerns how we may extend our framework to the preservation of adiabatic invariants. In place of considering the preservation of general adiabatic invariants, we consider the Lorentz problem, modelling the motion of a charged particle in a strong, non-uniform magnetic field.

Within such systems, the magnetic moment μ is adiabatically invariant. In combination with the conservation of energy ε, this implies particle trajectories are

confined to regions where the magnetic field strength does not exceed a threshold value determined by ε and μ. This behaviour is critical for charged particle dynamics; as a key astrophysical example, it is in part the adiabatic invariance of μ that enforces a planetary magnetosphere (e.g. the Earth's Van Allen belts) keeping particles away from the planetary surface where the magnetic field is stronger. This bounding behaviour is enforced by an effective magnetic mirror force, acting in the direction opposing the gradient of the magnetic field strength; devices that exploit the magnetic field bound induced by the adiabatic invariance of μ are hence referred to as magnetic mirrors. Such devices offered one of the earliest options for magnetic confinement fusion.

Through a different notion of associated test functions in Step **C**, we are able to preserve the adiabatic invariance of μ discretely. Crucially, this property holds independently of the timestep Δt_n.

The trajectories of charged particles in strong magnetic fields can be seen as fast, low-amplitude cyclotron oscillations around a guiding centre $\boldsymbol{\xi}$. This centre $\boldsymbol{\xi}$ moves parallel to the magnetic field lines, with slight slow perpendicular drifts due to magnetic field gradients and curvature. Modern toroidal magnetic confinement fusion devices, such as tokamaks and stellarators, aim to utilise this behaviour by confining particles to closed, nested magnetic flux surfaces, presenting an effective alternative to the adiabatic invariance of μ for magnetic confinement.

We do not, in Chapter 9, preserve these drifts in $\boldsymbol{\xi}$. It is our hope however that the ideas we develop for the asymptotic preservation of μ may be extended further to preserve these guiding centre drifts in the future, just as our general framework allows us equivalently to preserve both conservation and dissipation structures. In particular, the generalised notion of associated test function we use to preserve the behaviour of μ can readily be extended to $\boldsymbol{\xi}$; it is not yet clear, however, where to re-introduce the associated AVs in Step **E** to preserve these drifts (see Remark 9.3).

FEEC, simplification & reparametrisation

Our second extension relates to FEEC [Hip01; AFW06; AFW09; Arn18; Hu25]. FEEC arguably represents a generally different style of thinking within SP to that presented in this thesis, as those structures that FEEC has historically considered have generally been geometric and topological properties from exterior calculus (e.g. complex

exactness and cohomologies). These structures are generally local/pointwise, such as exact divergence-free conditions, and not global, such as those we consider here. However, these ideas do not exist independently; Stern, Zampa & McLachlan [MS20; SZ23] for example have established connections between FEEC and symplecticity (or more specifically multisymplecticity) while Hu *et al.* [HLX21; LHF23; He+25] have used FE de Rham complexes from FEEC to derive energy- and helicity-stable integrators in MHD.

Chapter 10 concerns how FEEC may be applied to our SP discretisations, in particular in creating equivalent but more computationally practical schemes. In particular, after re-assessing in Section 10.3 the energy- and helicity-stable incompressible NS integrator (3.28) of Chapter 3, we consider two new sets of structures for PDE systems.

Enstrophy & stabilisation on under-resolved meshes.

While we studied the incompressible NS equations in Chapter 3 from the perspective of energy and helicity stability, they exhibit further structures, especially in 2D. In particular, we highlight the enstrophy $Q_3(\mathbf{u}) \coloneqq \frac{1}{2}\|\nabla\mathbf{u}\|^2$, where $\|\cdot\|$ denotes the $\mathbf{L}^2$ norm. Just as the energy $Q_1(\mathbf{u}) \coloneqq \frac{1}{2}\|\mathbf{u}\|^2$ is dissipated in the incompressible NS equations as $\dot{Q}_1 = \frac{1}{\mathrm{Re}}\|\nabla\mathbf{u}\|^2$, the enstrophy is dissipated as $\dot{Q}_3 = \frac{1}{\mathrm{Re}}\|\Delta\mathbf{u}\|^2$ when $d = 2$ and assuming appropriate BCs, where Δ denotes the Laplacian.

In Section 10.4, we derive energy- and enstrophy-stable integrators for the incompressible NS equations, i.e. schemes that necessarily dissipate energy and preserve the evolution, or dissipation when $d = 2$, of enstrophy. Through FEEC, we are able to reparametrise our scheme into an approachable velocity–vorticity formulation.

Typical $\mathbf{H}^1$-conforming methods for the incompressible NS equations that are solely energy-stable struggle on under-resolved meshes, i.e. when $h^2\mathrm{Re} \gg 1$ for a given mesh size h. Namely, they exhibit instability: large, spurious oscillations in the velocity field.

While many explanations can be offered for this phenomenon (see Moura *et al.* [Mou+22]), we can attribute it to the lack of a meaningful $\mathbf{H}^1$ energy estimate holding on the discrete solution. To clarify this reasoning, consider the energy

estimates provided by energy stability alone in the continuous case:

$$\sup_{t\geq 0}\|\mathbf{u}\|^2 \leq \|\mathbf{u}(0)\|^2, \qquad \int_0^\infty \|\nabla\mathbf{u}\|^2 \leq \frac{\mathrm{Re}}{2}\|\mathbf{u}(0)\|^2. \tag{8.1a}$$

Ostensibly, we have both an $\mathbf{L}^2$ and $\mathbf{H}^1$ bound on our solution, that would both be inherited by an energy-stable discretisation. However, the issue is as follows: in finite dimensions, one may recall that all norms are equivalent; an $\mathbf{L}^2$ bound implies an $\mathbf{H}^1$ bound, up to some constant of proportionality. For the $\mathbf{H}^1$ bound provided by energy stability to have any impact in the discrete setting, it must be stronger than what this norm equivalence already provides. Crucially, however, the $\mathbf{H}^1$ bound scales with Re; as $\mathrm{Re}\to\infty$, this $\mathbf{H}^1$ bound becomes weaker, and ultimately vacuous. In particular, when $h^2\mathrm{Re}\gg 1$, the $\mathbf{H}^1$ bound no longer constrains the discrete solution beyond what is already implied by the $\mathbf{L}^2$ bound. The $\mathbf{H}^1$ bound then provided by energy stability has no impact on the regularity of discrete solutions, tending $\mathbf{H}^1$-conforming schemes towards instability on coarse meshes and at high Re. This argument is even clearer in the inviscid Euler case $\mathrm{Re}=\infty$, where no $\mathbf{H}^1$ bound is available at all.

What is required then is a Re-robust $\mathbf{H}^1$ bound, i.e. one that is independent of Re. In two dimensions, and again under suitable BCs, such a bound arises from enstrophy dissipation. In the continuous case, we have

$$\sup_{t\geq 0}\|\nabla\mathbf{u}\|^2 \leq \|\nabla\mathbf{u}(0)\|^2, \qquad \int_0^\infty \|\Delta\mathbf{u}\|^2 \leq \frac{\mathrm{Re}}{2}\|\nabla\mathbf{u}(0)\|^2. \tag{8.1b}$$

In particular, this provides a Re-robust $\mathbf{H}^1$ bound (alongside an $\mathbf{H}^2$ bound that scales with Re). In constrast to the energy-based $\mathbf{H}^1$ estimate, that which is provided by enstrophy stability remains meaningful as $\mathrm{Re}\to\infty$, even in the inviscid Euler limit $\mathrm{Re}=\infty$. This motivates the use of enstrophy-stable schemes, as they are expected to offer improved stability over solely energy-stable ones on under-resolved 2D meshes. We demonstrate this improved stability in the inviscid Euler case $\mathrm{Re}=\infty$ in Subsection 10.4.4.

MHD, topology, helicity & perturbation analysis

MHD equations govern the dynamics of electrically conducting fluids, particularly plasmas. Various systems of equations are studied throughout MHD, each exhibiting similar and important structures; the preservation of these structures have proven generally critical for accurate computation.

In Section 10.5, we consider how our framework may be used to preserve these structures, in particular within the incompressible Hall MHD equations (10.68), and how FEEC may be used to simplify the structure and application of the resulting discretisations.

Arguably most crucial among these structures is the dissipation of an energy functional, either $\frac{1}{2}\|\mathbf{B}\|^2$ (where $\mathbf{B}$ is the magnetic field) or this value featuring an additional hydrodynamic contribution, e.g. $\frac{1}{2}\|\mathbf{u}\|^2$ (where $\mathbf{u}$ is the fluid velocity field). Brackbill & Barnes [BB80] observed in 1980 that pointwise violation of the magnetic Gauss law $\operatorname{div}\mathbf{B} = 0$ can imply instability in the energy, making it of equal importance; see also Dai & Woodward [DW98]. We circumvent this issue in our discretisation by working with a formulation in the magnetic potential $\mathbf{A}$, with $\mathbf{B} = \operatorname{curl}\mathbf{A}$ defined implicitly such that $\operatorname{div}\mathbf{B} = 0$ is enforced naturally by the complex condition $\operatorname{div}\circ\operatorname{curl} = 0$. We are then able to use FEEC to reparametrise our discretisation in the more traditional magnetic field $\mathbf{B}$; since this reparametrisation is equivalent, it necessarily inherits the pointwise $\operatorname{div}\mathbf{B} = 0$ property. Energy stability is then enforced through the introduction of AVs approximating the current $\operatorname{curl}\mathbf{B}$ and flow velocity $\mathbf{u}$.

Beyond energy stability, an important aspect of MHD systems in the ideal setting is Alfvén's theorem [Alf43], or the frozen-in flux theorem, on the preservation of the topology of magnetic fields (see Choudhuri [Cho98, Chap. 15]). Similarly to the advection of vortex lines for the ideal incompressible NS equations as discussed in Chapter 2, this states that the magnetic fields in ideal MHD are convected by the flow; the ideal conservation of the magnetic helicity $\frac{1}{2}(\mathbf{A}, \mathbf{B})$ then serves a similar role to the ideal conservation of the fluid helicity in the NS equations. Violation of Alfvén's theorem in the non-ideal, resistive case is referred to as magnetic reconnection (again, see Choudhuri [Cho98, Sec. 15.2]). Through reconnection, the distribution of helicity across length scales satisfies an inverse cascade (see Frisch *et al.* [Fri+75]) with small-scale structures merging to form larger, more coherent ones that are stable over long time periods; as such, the magnetic helicity generally dissipates at a much slower rate than the energy, which satisfies a typical forward cascade. Through Arnold's inequality (see Arnold & Khesin [AK08, p. 122]) the helicity serves as a lower bound for the energy. These two results together have a profound impact, as the energy is prevented from decaying to zero at a rate it otherwise would by

the lower bound of the slowly decaying helicity; in essence, the magnetic field is prevented from decaying through large knotted structures that persist over long time periods, even in the non-ideal case. On the discrete level, failure to accurately preserve the evolution of the magnetic helicity allows the numerical solution to violate this topological barrier, with solutions decaying at an unphysically high rate. Through the introduction of an AV approximating the magnetic field $\mathbf{B}$, we are able to preserve the evolution of the magnetic helicity discretely. Further discussion on the helicity and the roles of knottedness in plasma physics can be found in the works of Arnold & Khesin [AK08], Berger & Field [BF84], Moffatt & Tsinober [Mof81; MT92; Mof14], and Smiet [Smi17].

For the Hall MHD equations, a further ideal invariant is found in the hybrid helicity $\frac{1}{2}(\mathbf{A} + a\mathbf{u}, \mathbf{B} + b\,\mathrm{curl}\,\mathbf{u})$ where certain conditions hold on a, b. This can be viewed as a linear combination of the magnetic helicity, fluid helicity as considered in Chapter 3, and cross helicity $\frac{1}{2}(\mathbf{u}, \mathbf{B})$ (see Mininni, Gomez & Mahajan [MGM03]). Similarly to the magnetic and fluid helicities, in the case $a = b$ the hybrid helicity quantifies the knottedness of streamlines in $\mathbf{B} + b\,\mathrm{curl}\,\mathbf{u}$, showing a global topological conservation of the knottedness of these lines in the ideal case. Moreover, similar to the Arnold inequality, the hybrid helicity serves as a lower bound for both the energy and enstrophy $\frac{1}{2}\|\mathrm{curl}\,\mathbf{u}\|^2$. In the non-Hall case, the cross helicity takes the place of the hybrid helicity as the third QoI (see e.g [PB09]). We are able to preserve these final structures through the introduction of an AV approximating the vorticity $\mathrm{curl}\,\mathbf{u}$.

In magnetic confinement fusion, reversed-field pinch (RFP) devices rely heavily on plasma currents to generate and sustain their magnetic fields. Unlike tokamaks, where externally imposed magnetic fields play a dominant role in shaping the configuration, RFP field structures emerge largely through self-organisation governed by the MHD equations. This makes accurate numerical discretisation of the MHD system particularly important for RFP design. In particular, preserving magnetic helicity is crucial, as many RFP concepts rely on Taylor's theory, which posits that plasmas relax to minimum energy states consistent with helicity conservation.

A significant portion of fusion reactor design and plasma modelling focuses on stability analysis, particularly the behaviour of small perturbations, such as kink, ballooning, tearing, and Alfvénic modes. Among these, instabilities driven by incompressible dynamics, such as internal kink modes which have been linked to

sawtooth relaxations, are typically those that require the least free energy. Accordingly, linearised MHD, particularly in its incompressible form, is widely used in the design and analysis of magnetic confinement fusion devices, with `NIMROD` [Sov+03] and `JOREK` [HC07; Hoe+21] representing two of the more widely used codebases (see also the comparative review article of Artola *et al.* [Art+21]).

Due to the symmetry of the (non-Hall) MHD equations in the sign of $\mathbf{B}$, the dominant nonlinear terms guiding the behaviour of small perturbations are not quadratic, but cubic, leading to long-lived but sudden, violent transitions. As such, while linearised MHD may capture the initial stability of perturbations well, accurate transient simulations require a proper handling of the nonlinearities, further motivating the need for stable SP discretisations for the nonlinear MHD equations.

For further reading on the use of MHD in fusion plasma modelling, we refer the reader to formal analysis of Goedbloed, Keppens & Poedts [GKP19]. With relation to the above, we highlight Chapters 16 (on axisymmetric equilibria and background magnetic fields in tokamaks), 18 (on linearised MHD in similar axisymmetric settings), and in particular 19 (on the full incompressible MHD equations including helicity preservation).

8.1 Related literature

Asymptotic-preserving integrators for the Lorentz problem

One of the most commonly applied integrators for the charged particle problem (9.1) is a modified Störmer–Verlet method, referred to as the Boris method [Bor70]. Under a general magnetic field, this can exhibit a drift or random walk in the energy [HL18]. In [HL20] the authors propose a further modification of the Störmer–Verlet method based on a Lagrangian interpretation of the system, which they show to conserve both the energy and magnetic moment as adiabatic invariants over long timescales; this however relies on the use of timesteps much smaller than the oscillation period.

On longer timesteps, much of the research into asymptotic-preserving schemes for charged particles has focused on the problem of capturing guiding centre drifts, typically through the introduction of a fictitious force [BF85; VB95], fictitious velocity [Coh+07] or both [GCW10]. As a Hamiltonian system, exactly energy-conserving

schemes for charged particles in magnetic fields on arbitrary timesteps are well-established [MQR99; CH11]. In [RC20] the authors propose a modification of the scheme of Brackbill, Forslund and Vu [BF85; VB95] with exact energy conservation, using an adaptive timestepping scheme to transition between regions of high and low gyration radius.

Typically, one of the fictitious forces introduced to preserve the asymptotic behaviour in such schemes resembles the fictitious mirror force $-\mu \mathbf{b}_{\|} \cdot \nabla B$. This is observed numerically to improve the preservation of the adiabatic invariance of μ [RC20, Fig. 10] however these results are typically hard to prove or quantify in comparison to our schemes (9.11, 9.15). In particular, the mirror force implicitly appears in our discrete solutions as a consequence of the adiabatic invariance of μ, with no fictitious terms needed in the formulation.

Energy- & enstrophy-stable integrators for the incompressible Navier–Stokes equations

In 2017, Palha & Gerritsma [PG17] proposed a dual-field velocity–vorticity discretisation for the 2D incompressible NS equations. This dual-field concept resembles that used in the energy- and helicity-stable discretisation proposed by Zhang *et al.* [Zha+22], as discussed in Section 2.1. While this scheme is not an instance of our framework—the discrete vorticity ω therein is not a projection of a function of the discrete velocity $\mathbf{u}$, but evolves according to its own coupled equation—we note that the spaces occupied by ω, $\mathbf{u}$, and the pressure p are required in their work to satisfy the same FE complex relations as those identified in Subsection 10.4.2. The authors also observe the stabilisation properties of their scheme on under-resolved meshes, using as a numerical demonstration the roll-up of a shear layer with $\mathrm{Re} = \infty$.

In 3D, Hanot [Han23] recently proposed a 1-stage, mixed velocity–vorticity incompressible NS integrator. Again, this requires the same FE complex relations between the velocity, vorticity and pressure spaces as specified in Subsection 10.4.2, however the proposed discretisation defines the vorticity $\boldsymbol{\omega}$ via a projection on the velocity $\mathbf{u}$; consequently, this scheme more closely resembles one that might derive from our framework than that as mentioned above proposed by Palha & Gerritsma [PG17]. The discretisation differs from ours (10.41) in the definition of

$\boldsymbol{\omega}$; while we define $\boldsymbol{\omega}$ to be a projection of $\operatorname{curl}\mathbf{u}$ under the $\mathbf{H}(\operatorname{curl})$ inner product,[1] Hanot defines this projection in $\mathbf{L}^2$. While the author's goal was the numerical simulation of the equations in 3D, the reduction to 2D is immediate. In the 2D case, Zhang *et al.* [Zha+24] analysed the SP properties of this scheme, including its enstrophy stability. Specifically, they observe it to dissipate an auxiliary enstrophy $\tilde{Q}_3(\omega) \coloneqq \frac{1}{2}\|\omega\|^2$ defined on the AV ω; this differs from the enstrophy stability we show for our 2D integrator (10.44), which holds on the primal variable $\mathbf{u}$. Adopting the terminology of Zhang *et al.* [Zha+24], we refer to this discretisation as the mass-, energy-, enstrophy-, vorticity-conserving (MEEVC) scheme. It is not clear in what ways these primal (in our scheme) and auxiliary (in the MEEVC scheme) forms of enstrophy stability might be equivalent. It is further unclear if this scheme can in some way be derived from our framework; the similarities are striking, however we believe this not to be the case, as the enstrophy stability in the MEEVC scheme holds in this alternative auxiliary form only. A numerical comparison of our scheme in 2D to the MEEVC scheme is shown in Subsection 10.4.4.

In the context of metriplectic systems, the recent preprint of Lombardi & Pagliantini [LP24] proposed a 1-stage, mixed stream function–vorticity discretisation for the 2D incompressible NS equations, that is both energy- and enstrophy-stable. This can be shown to be equivalent to the MEEVC scheme of Hanot [Han23] and Zhang *et al.* [Zha+24], while avoiding the FE complex criteria. The relation between the MEEVC scheme and that proposed by Lombardi & Pagliantini [LP24] is equivalent to the relation between our S-stage stream function–vorticity schemes (10.30, 10.34) and their velocity-vorticity counterparts (10.41, 10.44).

An alternative approach to enstrophy stability comes in vorticity formulations of the NS equations. Here, one begins with a vorticity parametrisation, and discretises the vorticity equation

$$\dot{\boldsymbol{\omega}} = \boldsymbol{\omega} \cdot \nabla \mathbf{u} - \mathbf{u} \cdot \nabla \boldsymbol{\omega} + \frac{1}{\mathrm{Re}} \Delta \boldsymbol{\omega}. \tag{8.2}$$

As the square norm of $\boldsymbol{\omega}$, it is then simple to preserve the dissipation of enstrophy assuming an appropriate handling of the nonlinear advective term (see the schemes of Charnyi *et al.* [Cha+17]), however the issue lies then, not in preserving enstrophy stability, but in preserving energy stability.

[1] More specifically, our projection is under the bilinear form $(\operatorname{curl}\cdot, \operatorname{curl}\cdot)$, which defines an inner product after restriction to a certain space of discretely divergence-free functions.

Stabilisation for the incompressible Navier–Stokes equations on under-resolved meshes

Typical approaches to stabilisation have not focused on enstrophy stability, the most common approach for $\mathbf{H}^1$-conforming schemes being the introduction of an artificial viscosity. Spectral vanishing viscosity (SVV) methods, one such example originally proposed by Maday & Tadmor [MT89] in 1989, allow the viscous term to act only on the high-order modes; the motivation for this approach lies in how traditional schemes accumulate energy in the finer length scales, which are in turn associated with the higher-order polynomials. With appropriate tuning, Tadmor [Tad90; Tad93] showed this to yield entropy solutions in the convergent limit, however the selective damping of high-order modes is known to reduce the effective resolution of the discrete solution; the approximation power of the highest order polynomials is lost.

Continuous interior penalty (CIP)/gradient jump penalisation (GJP) methods, proposed by Douglas & Dupont [DD76] in 1976, aim to tackle the fine-scale oscillations by penalising the jump on the solution gradient across facets. An analysis of these methods for AD systems including the NS equations at high Re was done by Burman, Hansbo & Fernández [BH04; BFH06; BF07]. At moderate polynomial orders e.g. $p \approx 3$ (where there are few high-order terms to penalise) CIP/GJP methods are known to outperform SVV methods (see Moura *et al.* [Mou+22]). We note, however, that the introduction of artificial dissipation terms renders these schemes to be irreversible in the Euler case $\mathrm{Re} = \infty$, i.e. they introduce energy dissipation when it should not exist.

In contrast, DG methods (see Cockburn, Karniadakis & Shu [CKS00]) offer intrinsic upwind stabilisation, which is biased towards these finer scales, as discussed by Moura *et al.* [MSP15; Mou+17]. The stability properties of DG methods are similar to those of CIP/GJP methods, however it can be argued they introduce a significant overhead, with more DoFs for the same order on the same mesh.

Finite element Stokes complexes & their implementation

It was noted in Subsection 10.4.3 that conforming schemes for (10.41, 10.44) require the knowledge and implementation of FE Stokes complexes.

In 2D, such FE complexes are well-known. For example, the Scott–Vogelius (SV) [SV85a; SV85b] complex (as considered in Subsection 10.4.3) takes $\mathbb{U}$ to be the Morgan–Scott (MS) [MS75] space. Outside the workaround presented in Subsection 10.4.3 inspired by Ainsworth & Parker [AP24a] the MS element is not implemented in general purpose FE software. Over Alfeld-split [Alf84] (or barycentrically refined) meshes, the MS space becomes the Hsieh–Clough–Tocher (HCT) [CT65] space; the HCT space, on the other hand, is relatively well-supported, e.g. in `GetFem++` (see Renard & Poulios [RP20]), `FreeFEM++` (see Hecht [Hec12]), `libMesh` (see Stogner & Carey [SC07]), and `Firedrake` through `FIAT` (see Brubeck & Kirby [BK25]).

FE Stokes complexes in 3D are less common. The first FE Stokes complexes on arbitrary tetrahedral meshes were proposed by Neilan [Nei15], further generalised by Chen & Huang [CH24] to de Rham complexes of arbitrary smoothness; these spaces required elements of very order, namely polynomials degrees 9, 8, 7 and 6 respectively along the complex. Over split meshes, this degree can be reduced. We refer the reader to the works of Fu, Guzmán & Neilan [FGN20] and Hu, Zhang & Zhang [HZZ22] for such complexes on Alfeld-split meshes, and Guzman, Lischke & Neilan [GLN22] on Worsey–Farin splits. As far as we are aware, none of these spaces are yet supported in any publicly available FE software.

In the consideration of biharmonic-like equations, Ainsworth & Parker [AP24a; AP24b] propose reparametrisation systems by which one may effectively use H^2-conforming scalar FEs while only requiring FEs with at most H^1 conformity to be implemented. The schemes (10.56, 10.60) are very similar and heavily inspired by their construction.

Energy- & helicity-stable integrators in MHD

After the application of FEEC, our final energy- and helicity-scheme for the incompressible Hall MHD equations is equivalent at lowest order in time ($S = 1$) to that proposed by Laakmann, Hu & Farrell [LHF23]. Our intention is not to claim the scheme is novel, but to show how it may be derived through our framework; for these reason, we do not demonstrate any numerical simulations, instead referring the reader to those done by the authors above.

The scheme as presented by Laakmann, Hu & Farrell builds upon the earlier work of Hu, Lee & Xu [HLX21] in the non-Hall case; our framework as applied to this form of the MHD equations similarly derives their energy- and helicity-scheme. Equivalently, we re-derive the scheme of He *et al.* [He+25] when applying our framework to the magneto-frictional equations.

8.2 Overview

In Chapter 9, we consider an application of our framework to the Lorentz problem, modelling the motion of a charged particle in a strong, non-uniform magnetic field. We show that, with a certain different notion of associated test functions (see 3.1), we are able to preserve the adiabatic invariance [Hen93; AKN06] of the magnetic moment, i.e. ensure its oscillations remains bounded against a certain value.

In Chapter 10, we consider how we may use results from FEEC to simplify our SP discretisations. In particular, we show how the satisfaction of certain compatibility conditions between the FE spaces allows us to both eliminate certain LM-like terms from our discretisations, and reparametrise our discretisations in terms of spaces of lower regularity. To apply these ideas, we revisit the energy- and helicity-stable integrator (3.28) of Chapter 3; we consider as further examples an energy- and (in the 2D case) enstrophy-stable integrator for the incompressible NS equations, and an energy- and helicity-stable integrators for the incompressible Hall MHD equations.

"Sir, could we take a break for a while? It appears my intelligence circuits have melted."

— Kryten 2X4B-523P (Robert Llewellyn) [GN91]

9

The Lorentz problem: magnetic moment stability & adiabatic invariants

Contents

We consider in this chapter an SP scheme for the Lorentz problem, governing the motion of a charged particle in a magnetic field. In particular, we consider strong, non-uniform, stationary magnetic fields, within which particles oscillate over length scales much smaller than the characteristic length scale of the magnetic field. We seek an SP integrator that will both conserve the energy, and preserve the adiabatic invariance of the magnetic moment. Combined, these structures have important impacts on the long-term dynamics of solutions. These adiabatic invariants fall outside the scope of our framework as stated in Algorithm 3.5. We find, however, that we are able to preserve the adiabatic invariance structure through a generalised notion of the associated test function presented in Step **C**.

Consider then the (nondimensionalised) Lorentz problem,

$$\dot{\mathbf{x}} = \mathbf{v}, \qquad \dot{\mathbf{v}} = \frac{1}{\rho}\mathbf{v} \times \mathbf{B}(\mathbf{x}), \tag{9.1}$$

where, $\mathbf{x}(t), \mathbf{v}(t) \in \mathbb{R}^3$ represent the particle's position and velocity respectively, $\mathbf{B}(\mathbf{x}) \in \mathbb{R}^3$ represents the magnetic field, $\times$ denotes the cross product, and $\rho \ll 1$ is

the dimensionless gyroradius, representing the ratio of a characteristic gyroradius to the characteristic length scale of variations in the magnetic field. Solutions to (9.1) have characteristic trajectories, with high-frequency ($\mathcal{O}[\rho^{-1}]$) oscillations, compounding into medium-frequency ($\mathcal{O}[1]$) motion parallel to the field lines B, and low-frequency ($\mathcal{O}[\rho]$) perpendicular drifts (see Fig. 9.1).

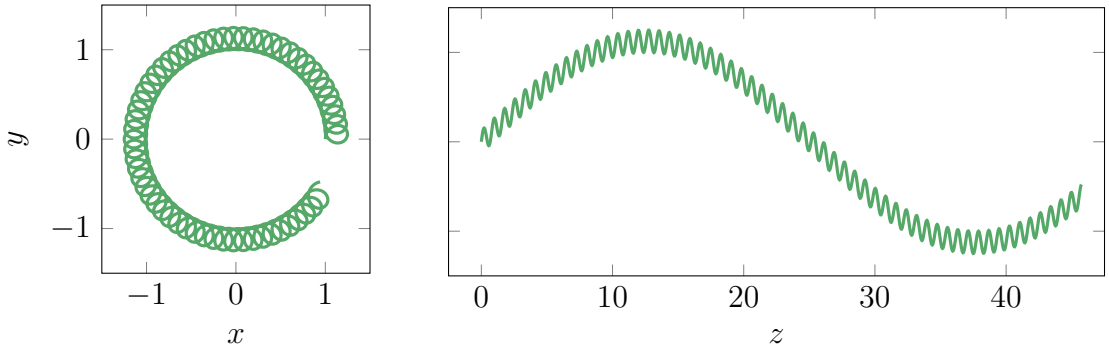


Figure 9.1: Trajectories of a charged particle moving in a magnetic field according to the Lorentz system (9.1) with $\mathbf{B} = (0, 0, x^2 + y^2)$, $\rho = 2^{-3}$ and ICs $\mathbf{x}(0) = (1, 0, 0)$, $\mathbf{v}(0) = (0, 1, 1)$. Both the high-frequency ($\mathcal{O}[\rho^{-1}]$) oscillations and low-frequency ($\mathcal{O}[\rho]$) drifts (in this case grad-B drift manifesting as a further circular motion) are visible in either figure.

Remark 9.1 (Negible electric field and constant magnetic field)**.** *In the Lorentz model* (9.1) *we both neglect to include an electric field, and consequently assume the magnetic field to be constant. This is done simply to ensure the adiabatic invariance of μ on the continuous level. As desired, one may simply assume a varying magnetic field directly in our SP scheme, or introduce an electric field through e.g. a splitting, using our SP integrator for the magnetic field component.*

For ease of notation, define the magnetic field strength $B(\mathbf{x}) \coloneqq \|\mathbf{B}(\mathbf{x})\|$, where $\|\cdot\|$ denotes the ℓ^2 norm, and the normalised magnetic field $\mathbf{b}_\parallel(\mathbf{x}) \coloneqq \frac{1}{B(\mathbf{x})}\mathbf{B}(\mathbf{x})$. Of unit length and perpendicular to $\mathbf{b}_\parallel$, define $\mathbf{b}_\perp(\mathbf{x}) \in \mathbb{R}^3$ implicitly through the expansion $\mathbf{v} = v_\parallel \mathbf{b}_\parallel + v_\perp \mathbf{b}_\perp$; this in turn defines the parallel and perpendicular velocities $v_\parallel(\mathbf{x}), v_\perp(\mathbf{x}) \in \mathbb{R}$ respectively. The basis $(\mathbf{b}_\parallel, \mathbf{b}_\perp, \mathbf{b}_*)$ for $\mathbb{R}^3$ is completed by $\mathbf{b}_*(\mathbf{x}) \coloneqq \mathbf{b}_\parallel(\mathbf{x}) \times \mathbf{b}_\perp(\mathbf{x})$.

As state above, we consider two QoIs: the energy $\varepsilon(\mathbf{v}) \coloneqq \frac{1}{2}\|\mathbf{v}\|^2$, and magnetic moment $\mu \coloneqq \mu(\mathbf{x}, \mathbf{v}) \coloneqq \frac{1}{2B(\mathbf{x})} v_\perp^2$. The conservation of energy ε is trivial to prove, whereas the adiabatic invariance of the magnetic moment μ is more involved; simply evaluating $\dot{\mu}$ we find it to be $\mathcal{O}[1]$, implying very little. The typical approach for

proving the adiabatic invariance of μ is through gyro-averaging i.e. by averaging $\dot{\mu}$ over the rapid oscillation and confirming the result to be negligible. We refer the reader to Northrop [Nor63] for a classical text on adiabatic invariants and gyroaveraging methods for charged particles (in particular Section 3.A on the magnetic moment) or more recently Hazeltine & Meiss [HM03, Sec. 2.4]. Regardless, this proof is difficult to preserve on discretisation; we therefore seek an alternative.

Let us define a μ-correction term $\Delta\mu(\mathbf{x}, \mathbf{v}) \in \mathbb{R}$,

$$\Delta\mu \coloneqq \frac{1}{B^3}\left[\frac{1}{4}v_\parallel^2 v_\perp \mathbf{b}_\perp \otimes \mathbf{b}_* + \frac{1}{4}v_\parallel^2 v_\perp \mathbf{b}_* \otimes \mathbf{b}_\perp + \frac{1}{2}v_\perp^3 \mathbf{b}_\parallel \otimes \mathbf{b}_* + v_\parallel^2 v_\perp \mathbf{b}_* \otimes \mathbf{b}_\parallel\right] : \nabla\mathbf{B}, \tag{9.2}$$

where $\otimes$ denotes the outer product, such that e.g. $\mathbf{b}_\perp \otimes \mathbf{b}_* : \nabla\mathbf{B} = (\mathbf{b}_* \cdot \nabla\mathbf{B}) \cdot \mathbf{b}_\perp$ with $(\mathbf{b}_* \cdot \nabla\mathbf{B})$ denoting the convective derivative in the direction of $\mathbf{b}_*$. We may observe by careful direct calculation that

$$(\mathbf{v} \times \mathbf{B}) \cdot \nabla_{\mathbf{v}}[\mu + \rho\Delta\mu] + \rho\mathbf{v} \cdot \nabla_{\mathbf{x}}\mu = 0, \tag{9.3}$$

where $\nabla_{\mathbf{x}}$ denotes the partial derivatives with respect to $\mathbf{x}$. Denoting by ∂_t the partial derivative with respect to t, we may then evaluating the change in $\mu + \rho\Delta\mu$ over the timestep T_n,

$$\begin{aligned}
&[\mu + \rho\Delta\mu]|_{t=t_{n+1}} - [\mu + \rho\Delta\mu]|_{t=t_n} \\
&\qquad = \int_{T_n} \partial_t[\mu + \rho\Delta\mu] && \text{(9.4a)}\\
&\qquad = \int_{T_n} \nabla_{\mathbf{x}}[\mu + \rho\Delta\mu] \cdot \dot{\mathbf{x}} + \nabla_{\mathbf{v}}[\mu + \rho\Delta\mu] \cdot \dot{\mathbf{v}} && \text{(9.4b)}\\
&\qquad = \int_{T_n} \nabla_{\mathbf{x}}\mu \cdot \dot{\mathbf{x}} + \nabla_{\mathbf{v}}[\mu + \rho\Delta\mu] \cdot \dot{\mathbf{v}} + \mathcal{O}[\rho] && \text{(9.4c)}\\
&\qquad = \int_{T_n} \nabla_{\mathbf{x}}\mu \cdot \mathbf{v} + \frac{1}{\rho}\nabla_{\mathbf{v}}[\mu + \rho\Delta\mu] \cdot (\mathbf{v} \times \mathbf{B}) + \mathcal{O}[\rho] && \text{(9.4d)}\\
&\qquad = \int_{T_n} \mathcal{O}[\rho] && \text{(9.4e)}\\
&\qquad = \mathcal{O}[\rho\Delta t_n]. && \text{(9.4f)}
\end{aligned}$$

Thus, we may bound the change in μ over T_N as $\mu|_{t=t_{n+1}} - \mu|_{t=t_{n+1}} = \mathcal{O}[\rho](1 + \mathcal{O}[\Delta t_n])$. Plotting $\mu + \rho\Delta\mu$ for an example trajectory, we see the introduction of the correction $\rho\Delta\mu$ balances the high-frequency oscillations in μ (see Fig. 9.2 below) demonstrating a smoother time derivative. The proof (9.4) offers an alternative option for conserving the adiabatic invariance of μ. In particular, the fourth variational equality (9.4d)

requires the Lorentz system $(\dot{\mathbf{x}}, \dot{\mathbf{v}}) = (\mathbf{v}, \frac{1}{\rho}(\mathbf{v} \times \mathbf{B}))$ to hold when tested against test functions $(\nabla_{\mathbf{x}}\mu, \nabla_{\mathbf{v}}[\mu + \rho\Delta\mu])$. This informs our general strategy for the preservation of the adiabatic invariant structure: interpret this gradient tuple $(\nabla_{\mathbf{x}}\mu, \nabla_{\mathbf{v}}[\mu + \rho\Delta\mu])$ as the associated test function for the preservation of the adiabatic invariance of μ in Step **C**; define AVs approximating $(\nabla_{\mathbf{x}}\mu, \nabla_{\mathbf{v}}[\mu + \rho\Delta\mu])$ in Step **D**; introduce these AVs into the RHS of the variational form of (9.1) in Step **E** such that it evaluates to 0 when testing against these AVs, as in (9.3).

Application of framework (Algorithm 3.5)

A. We define $\mathbb{X}$ as in (6.5) for $d = 3$,

$$\mathbb{X} \coloneqq \left\{\mathbf{x} \in C^1(\mathbb{R}_+)^3 : \mathbf{x}(0) \text{ satisfies known initial data}\right\}. \tag{9.5}$$

We then arrive at our semidiscrete formulation for (9.1): find $(\mathbf{x}, \mathbf{v}) \in \mathbb{X}^2$ such that

$$\mathbf{y} \cdot \dot{\mathbf{x}} = \mathbf{y} \cdot \mathbf{v}, \qquad \mathbf{w} \cdot \dot{\mathbf{v}} = \frac{1}{\rho}\mathbf{w} \cdot (\mathbf{v} \times \mathbf{B}). \tag{9.6}$$

at all times $t \in \mathbb{R}_+$ and for all $(\mathbf{y}, \mathbf{w}) \in \mathbb{U}^2$.

B. Over the timestep T_n, this is cast into a fully discrete form using our choice of $\mathcal{I}_n$, over $\mathbb{X}_n$ defined as in (6.2),

$$\mathbb{X}_n \coloneqq \left\{\mathbf{x} \in \mathbb{P}_S(T_n)^3 : \mathbf{x}(t_n) \text{ satisfies known initial data}\right\}. \tag{9.7}$$

find $(\mathbf{x}, \mathbf{v}) \in \mathbb{X}_n^2$ such that

$$\mathcal{I}_n[\mathbf{y} \cdot \dot{\mathbf{x}}] = \mathcal{I}_n[\mathbf{y} \cdot \mathbf{v}], \qquad \mathcal{I}_n[\mathbf{w} \cdot \dot{\mathbf{v}}] = \frac{1}{\rho}\mathcal{I}_n[\mathbf{w} \cdot (\mathbf{v} \times \mathbf{B})], \tag{9.8}$$

for all $(\mathbf{y}, \mathbf{w}) \in \dot{\mathbb{X}}_n^2$.

C. Following the argument above regarding the proof of adiabatic invariance of the magnetic moment μ, its associated test functions are $\boldsymbol{\alpha}(\mathbf{x}, \mathbf{v}) \coloneqq \nabla_{\mathbf{x}}\mu$ and $\boldsymbol{\beta}(\mathbf{x}, \mathbf{v}) \coloneqq \nabla_{\mathbf{v}}[\mu + \rho\Delta\mu]$. For the conservation of energy ε, the associated test functions are simply $\mathbf{0}$ (which can be ignored) and $\mathbf{v}$.

D. We introduce AVs $(\tilde{\mathbf{v}}, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}) \in \dot{\mathbb{X}}_n^3$, approximating $(\mathbf{v}, \boldsymbol{\alpha}(\mathbf{x}, \mathbf{v}), \boldsymbol{\beta}(\mathbf{x}, \mathbf{v}))$, and defined as in (3.19) such that

$$\mathcal{I}_n[\tilde{\mathbf{v}} \cdot \tilde{\mathbf{w}}] = \int_{T_n} \mathbf{v} \cdot \tilde{\mathbf{w}}, \tag{9.9a}$$

$$\mathcal{I}_n[\tilde{\boldsymbol{\alpha}} \cdot \tilde{\boldsymbol{\zeta}}] = \int_{T_n} \nabla_{\mathbf{x}}\mu \cdot \tilde{\boldsymbol{\zeta}}, \tag{9.9b}$$

$$\mathcal{I}_n[\tilde{\boldsymbol{\beta}} \cdot \tilde{\boldsymbol{\eta}}] = \int_{T_n} \nabla_{\mathbf{v}}[\mu + \rho\Delta\mu] \cdot \tilde{\boldsymbol{\eta}}, \tag{9.9c}$$

for all $(\tilde{\mathbf{w}}, \tilde{\boldsymbol{\zeta}}, \tilde{\boldsymbol{\eta}}) \in \dot{\mathbb{X}}^3_n$.

E. We must now introduce $(\tilde{\mathbf{v}}, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})$ into the RHS of (9.8) such that it evaluates to 0 when considering $(\mathbf{y}, \mathbf{w})$ as either $(\mathbf{0}, \tilde{\mathbf{v}})$ or $(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})$. It was shown in Lemma 6.15 (when constructing universally stable integrators for general conservative ODE systems in Section 6.2) that this can assuredly always be done, even when no immediate solution presents itself. Indeed, through a construction as in the proof of Lemma 6.15 we arrive at the following modified form of (9.8): find $(\mathbf{x}, \mathbf{v}) \in \mathbb{X}^2_n$ such that

$$\mathcal{I}_n[\mathbf{y} \cdot \dot{\mathbf{x}}] = \mathcal{I}_n\left[\frac{1}{\|\tilde{\boldsymbol{\alpha}}\|^2}(\tilde{\boldsymbol{\alpha}} \times \mathbf{y}) \cdot (\tilde{\boldsymbol{\alpha}} \times \tilde{\mathbf{v}}) - \frac{1}{\rho\|\tilde{\boldsymbol{\alpha}}\|^2}(\tilde{\boldsymbol{\beta}} \cdot (\tilde{\mathbf{v}} \times \mathbf{B}))(\tilde{\boldsymbol{\alpha}} \cdot \mathbf{y})\right], \tag{9.10a}$$

$$\mathcal{I}_n[\mathbf{w} \cdot \dot{\mathbf{v}}] = \frac{1}{\rho}\mathcal{I}_n[\mathbf{w} \cdot (\tilde{\mathbf{v}} \times \mathbf{B})], \tag{9.10b}$$

for all $(\mathbf{y}, \mathbf{w}) \in \dot{\mathbb{X}}^2_n$. We see first that this RHS coincides with that of (9.8) when $(\tilde{\mathbf{v}}, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}) = (\mathbf{v}, \boldsymbol{\alpha}(\mathbf{x}, \mathbf{v}), \boldsymbol{\beta}(\mathbf{x}, \mathbf{v}))$ by noting the classical vector calculus identity $\|\mathbf{X}\|^2\mathbf{Y} \cdot \mathbf{Z} = (\mathbf{X} \times \mathbf{Y}) \cdot (\mathbf{X} \times \mathbf{Z}) + (\mathbf{X} \cdot \mathbf{Y})(\mathbf{X} \cdot \mathbf{Z})$ alongside the relation (9.3) between $\boldsymbol{\alpha}(\mathbf{x}, \mathbf{v})$ and $\boldsymbol{\beta}(\mathbf{x}, \mathbf{v})$. We see then that it preserves the conservation of energy ε and the adiabatic invariance of the magnetic moment μ by considering the test function $(\mathbf{y}, \mathbf{w})$ to be $(\mathbf{0}, \tilde{\mathbf{v}})$ and $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ respectively, cancelling like terms, and using the orthogonality of the cross product.

F. The final SP integrator is then as follows: find $((\mathbf{x}, \mathbf{v}), (\tilde{\mathbf{v}}, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})) \in \mathbb{X}^2_n \times \dot{\mathbb{X}}^3_n$ such that

$$\mathcal{I}_n[\mathbf{y} \cdot \dot{\mathbf{x}}] = \mathcal{I}_n\left[\frac{1}{\|\tilde{\boldsymbol{\alpha}}\|^2}(\tilde{\boldsymbol{\alpha}} \times \mathbf{y}) \cdot (\tilde{\boldsymbol{\alpha}} \times \tilde{\mathbf{v}}) - \frac{1}{\rho\|\tilde{\boldsymbol{\alpha}}\|^2}(\tilde{\boldsymbol{\beta}} \cdot (\tilde{\mathbf{v}} \times \mathbf{B}))(\tilde{\boldsymbol{\alpha}} \cdot \mathbf{y})\right], \tag{9.11a}$$

$$\mathcal{I}_n[\mathbf{w} \cdot \dot{\mathbf{v}}] = \frac{1}{\rho}\mathcal{I}_n[\mathbf{w} \cdot (\tilde{\mathbf{v}} \times \mathbf{B})], \tag{9.11b}$$

$$\mathcal{I}_n[\tilde{\mathbf{v}} \cdot \tilde{\mathbf{w}}] = \int_{T_n} \mathbf{v} \cdot \tilde{\mathbf{w}}, \tag{9.11c}$$

$$\mathcal{I}_n[\tilde{\boldsymbol{\alpha}} \cdot \tilde{\boldsymbol{\zeta}}] = \int_{T_n} \nabla_{\mathbf{x}}\mu \cdot \tilde{\boldsymbol{\zeta}}, \tag{9.11d}$$

$$\mathcal{I}_n[\tilde{\boldsymbol{\beta}} \cdot \tilde{\boldsymbol{\eta}}] = \int_{T_n} \nabla_{\mathbf{v}}[\mu + \rho\Delta\mu] \cdot \tilde{\boldsymbol{\eta}}. \tag{9.11e}$$

for all $((\mathbf{y}, \mathbf{w}), (\tilde{\mathbf{w}}, \tilde{\boldsymbol{\zeta}}, \tilde{\boldsymbol{\eta}})) \in \dot{\mathbb{X}}^2_n \times \dot{\mathbb{X}}^3_n$.

Theorem 9.2 (Energy stability of the Poisson & gradient-descent ODE integrator)**.** *The integrator (9.11) conserves ε exactly, and $\mu + \rho\Delta\mu$ up to order $\mathcal{O}[\rho\Delta t_n]$,*

$$\varepsilon|_{t_{n+1}} - \varepsilon|_{t_n} = 0, \qquad [\mu + \rho\Delta\mu]|_{t_{n+1}} - [\mu + \rho\Delta\mu]|_{t_n} = \mathcal{O}[\rho\Delta t_n]. \tag{9.12}$$

The latter result quantifies the preserved adiabatic invariance of μ.

Proof. By considering respectively $\tilde{\mathbf{w}} = \dot{\mathbf{v}}$ in (9.11c) and $\mathbf{w} = \tilde{\mathbf{v}}$ in (9.11b),

$$\varepsilon|_{t_{n+1}} - \varepsilon|_{t_n} = \int_{T_n} \dot{\varepsilon} = \int_{T_n} \mathbf{v} \cdot \dot{\mathbf{v}} = \mathcal{I}_n[\tilde{\mathbf{v}} \cdot \dot{\mathbf{v}}] = \frac{1}{\rho}\mathcal{I}_n[\tilde{\mathbf{v}} \cdot (\tilde{\mathbf{v}} \times \mathbf{B})] = 0, \tag{9.13}$$

with the final equality holding by the orthogonality of the cross product. By considering respectively $(\tilde{\boldsymbol{\zeta}}, \tilde{\boldsymbol{\eta}}) = (\dot{\mathbf{x}}, \dot{\mathbf{v}})$ in (9.11d, 9.11e) and $(\mathbf{y}, \mathbf{w}) = (\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})$ in (9.11a, 9.11b),

$$\begin{aligned}
&[\mu + \rho\Delta\mu]|_{t_{n+1}} - [\mu + \rho\Delta\mu]|_{t_n} && \text{(9.14a)}\\
&\quad = \int_{T_n} \partial_t[\mu + \rho\Delta\mu] && \text{(9.14b)}\\
&\quad = \int_{T_n} \nabla_{\mathbf{x}}[\mu + \rho\Delta\mu] \cdot \dot{\mathbf{x}} + \nabla_{\mathbf{v}}[\mu + \rho\Delta\mu] \cdot \dot{\mathbf{v}} && \text{(9.14c)}\\
&\quad = \int_{T_n} \nabla_{\mathbf{x}}\mu \cdot \dot{\mathbf{x}} + \nabla_{\mathbf{v}}[\mu + \rho\Delta\mu] \cdot \dot{\mathbf{v}} + \mathcal{O}[\rho] && \text{(9.14d)}\\
&\quad = \mathcal{I}_n[\tilde{\boldsymbol{\alpha}} \cdot \dot{\mathbf{x}} + \tilde{\boldsymbol{\beta}} \cdot \dot{\mathbf{v}}] + \mathcal{O}[\rho\Delta t_n] && \text{(9.14e)}\\
&\quad = \mathcal{I}_n\left[\frac{1}{\|\tilde{\boldsymbol{\alpha}}\|^2}(\tilde{\boldsymbol{\alpha}} \times \tilde{\boldsymbol{\alpha}}) \cdot (\tilde{\boldsymbol{\alpha}} \times \tilde{\mathbf{v}}) - \frac{1}{\rho\|\tilde{\boldsymbol{\alpha}}\|^2}(\tilde{\boldsymbol{\beta}} \cdot (\tilde{\mathbf{v}} \times \mathbf{B}))(\tilde{\boldsymbol{\alpha}} \cdot \tilde{\boldsymbol{\alpha}})\right.\\
&\qquad\qquad \left. + \frac{1}{\rho}\tilde{\boldsymbol{\beta}} \cdot (\tilde{\mathbf{v}} \times \mathbf{B})\right] + \mathcal{O}[\rho\Delta t_n] && \text{(9.14f)}\\
&\quad = \mathcal{O}[\rho\Delta t_n], && \text{(9.14g)}
\end{aligned}$$

where again we use the orthogonality of the cross product. As intended, this mimics the earlier proof (9.4) of the adiabatic invariance of μ. □

While this scheme is effective for preserving the conservation and adiabatic invariance structures, we find the reciprocal $\frac{1}{\|\boldsymbol{\alpha}\|^2}$ terms can lead to some numerical instability. For this reason, we propose the following slightly modified form: find $((\mathbf{x}, \mathbf{v}), (\tilde{\mathbf{v}}, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})) \in \mathbb{X}_n^2 \times \dot{\mathbb{X}}_n^3$ such that

$$\begin{aligned}
\mathcal{I}_n[\|\nabla_{\mathbf{x}}\mu\|\dot{\mathbf{x}} \cdot \mathbf{y}] &= \mathcal{I}_n\left[\|\nabla_{\mathbf{x}}\mu\|(\tilde{\boldsymbol{\alpha}} \times \tilde{\mathbf{y}}) \cdot (\tilde{\boldsymbol{\alpha}} \times \mathbf{v}) - \frac{1}{\rho}(\tilde{\boldsymbol{\beta}} \cdot (\tilde{\mathbf{v}} \times \mathbf{B}))(\tilde{\boldsymbol{\alpha}} \cdot \mathbf{y})\right], && \text{(9.15a)}\\
\mathcal{I}_n[\dot{\mathbf{v}} \cdot \mathbf{w}] &= \frac{1}{\rho}\mathcal{I}_n[\|\tilde{\boldsymbol{\alpha}}\|^2 \mathbf{w} \cdot (\tilde{\mathbf{v}} \times \mathbf{B})], && \text{(9.15b)}\\
\mathcal{I}_n[\tilde{\mathbf{v}} \cdot \tilde{\mathbf{w}}] &= \int_{T_n} \mathbf{v} \cdot \tilde{\mathbf{w}}, && \text{(9.15c)}\\
\mathcal{I}_n[\|\nabla_{\mathbf{x}}\mu\|\tilde{\boldsymbol{\alpha}} \cdot \tilde{\boldsymbol{\zeta}}] &= \int_{T_n} \nabla_{\mathbf{x}}\mu \cdot \tilde{\boldsymbol{\zeta}}, && \text{(9.15d)}\\
\mathcal{I}_n[\tilde{\boldsymbol{\beta}} \cdot \tilde{\boldsymbol{\eta}}] &= \int_{T_n} \nabla_{\mathbf{v}}[\mu + \rho\Delta\mu] \cdot \tilde{\boldsymbol{\eta}}. && \text{(9.15e)}
\end{aligned}$$

for all $((\mathbf{y},\mathbf{w}),(\tilde{\mathbf{w}},\tilde{\boldsymbol{\zeta}},\tilde{\boldsymbol{\eta}})) \in \dot{\mathbb{X}}^2_n \times \dot{\mathbb{X}}^3_n$. Here, the AV $\tilde{\boldsymbol{\alpha}}$ is instead a discrete approximation to the normalised $\frac{1}{\nabla_{\mathbf{x}}\mu}\nabla_{\mathbf{x}}\mu$. The proofs of energy and magnetic moment stability are identical.

As discussed in Subsection 4.2.1, the AVs $(\tilde{\mathbf{v}},\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}})$ in (9.11) can be eliminated on the computational level. In particular, when $\mathcal{I}_n$ is the midpoint rule with $S=1$, we derive the following MV–DG-style method: find $(\mathbf{x}_{n+1},\mathbf{v}_{n+1}) \in \mathbb{R}^3$ such that

$$\frac{1}{\Delta t_n}\|\tilde{\boldsymbol{\alpha}}_n\|^2(\mathbf{x}_{n+1}-\mathbf{x}_n) = (\tilde{\boldsymbol{\alpha}}_n \times \mathbf{v}_{n+\frac{1}{2}}) \times \tilde{\boldsymbol{\alpha}}_n - \frac{1}{\rho}(\tilde{\boldsymbol{\beta}}_n \cdot (\mathbf{v}_{n+\frac{1}{2}} \times \mathbf{B}(\mathbf{x}_{n+\frac{1}{2}})))\,\tilde{\boldsymbol{\alpha}}_n, \tag{9.16a}$$

$$\frac{1}{\Delta t_n}(\mathbf{v}_{n+1}-\mathbf{v}_n) = \frac{1}{\rho}(\mathbf{v}_{n+\frac{1}{2}} \times \mathbf{B}(\mathbf{x}_{n+\frac{1}{2}})), \tag{9.16b}$$

where $\mathbf{x}_{n+\frac{1}{2}} \coloneqq \frac{1}{2}(\mathbf{x}_n+\mathbf{x}_{n+1})$, $\mathbf{v}_{n+\frac{1}{2}} \coloneqq \frac{1}{2}(\mathbf{v}_n+\mathbf{v}_{n+1})$, and $\tilde{\boldsymbol{\alpha}}_n$, $\tilde{\boldsymbol{\beta}}_n$ are defined

$$\tilde{\boldsymbol{\alpha}}_n \coloneqq \int_0^1 \nabla_{\mathbf{x}}\mu((1-\tau)\mathbf{x}_n+\tau_{n+1}\mathbf{x}_{n+1},(1-\tau)\mathbf{v}_n+\tau_{n+1}\mathbf{v}_{n+1})\,\mathrm{d}\tau, \tag{9.17}$$

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_n \coloneqq \int_0^1 &\nabla_{\mathbf{v}}\mu((1-\tau)\mathbf{x}_n+\tau_{n+1}\mathbf{x}_{n+1},(1-\tau)\mathbf{v}_n+\tau_{n+1}\mathbf{v}_{n+1})\\ &+\rho\nabla_{\mathbf{v}}\Delta\mu((1-\tau)\mathbf{x}_n+\tau_{n+1}\mathbf{x}_{n+1},(1-\tau)\mathbf{v}_n+\tau_{n+1}\mathbf{v}_{n+1})\,\mathrm{d}\tau;\end{aligned} \tag{9.18}$$

we may interpret these as adiabatic discrete gradients.

Remark 9.3 (Preservation of guiding centre drifts)**.** *A further QoI for* (9.1) *is the guiding centre* $\boldsymbol{\xi}(\mathbf{x},\mathbf{v}) \in \mathbb{R}^3$*,*

$$\boldsymbol{\xi} \coloneqq \mathbf{x} + \rho\frac{1}{B^2}\mathbf{v}\times\mathbf{B}, \tag{9.19}$$

evolving, up to oscillations, according to transport a combination of ($\mathcal{O}[1]$) *transport parallel to the field lines, and a slow* ($\mathcal{O}[\rho]$) *drift term*

$$\dot{\boldsymbol{\xi}} \approx v_\parallel \mathbf{b}_\parallel + \rho\left[\underbrace{\frac{v_\perp^2}{2B^3}\mathbf{B}\times\nabla B}_{\text{grad-B drift}} + \underbrace{\frac{v_\parallel^2}{B^2}\mathbf{B}\times(\mathbf{B}\cdot\nabla\mathbf{b}_\parallel)}_{\text{curvature drift}} - \underbrace{\frac{v_\perp^2}{2B^4}(\mathbf{B}\cdot\operatorname{curl}\mathbf{B})\mathbf{B}}_{\text{polarisation drift}}\right]. \tag{9.20a}$$

We can formalise this structure through the definition of a certain guiding centre correction $\Delta\boldsymbol{\xi}(\mathbf{x},\mathbf{v}) \in \mathbb{R}^3$ *such that*

$$\dot{\boldsymbol{\xi}} + \rho^2\dot{\Delta\boldsymbol{\xi}} = v_\parallel \mathbf{b}_\parallel + \rho\left[\underbrace{\frac{v_\perp^2}{2B^3}\mathbf{B}\times\nabla B}_{\text{grad-B drift}} + \underbrace{\frac{v_\parallel^2}{B^2}\mathbf{B}\times(\mathbf{B}\cdot\nabla\mathbf{b}_\parallel)}_{\text{curvature drift}} - \underbrace{\frac{v_\perp^2}{2B^4}(\mathbf{B}\cdot\operatorname{curl}\mathbf{B})\mathbf{B}}_{\text{polarisation drift}}\right] + \mathcal{O}[\rho^2]. \tag{9.20b}$$

A similar argument to the above on the adiabatic invariance of μ may then be applied, to preserve the guiding centre drifts, motivating in Step **D** *the introduction of AVs approximating $\nabla_{\mathbf{x}}\boldsymbol{\xi}$ and $\nabla_{\mathbf{v}}[\boldsymbol{\xi} + \rho^2\Delta\boldsymbol{\xi}]$; it remains unclear in Step* **E**, *however, where these AVs should be introduced in the RHS of* (9.8) *to preserve this structure. In its current state, the scheme* (9.15) *does not necessarily preserve guiding centre drifts; artificial drifts can be seen, for example, in Fig. 9.3 below, with the discrete trajectories under our scheme* (9.15) *drifting away from the z axis.*

9.1 Magnetic mirror test

Inspired by [RC20, Fig. 1] we test the integrator (9.15) using a magnetic mirror induced by two circular currents loops of radius r, oriented normal to the z-axis and centred at $z = \pm L$,

$$\mathbf{B}(\mathbf{x}) = \frac{[r^2 + L^2]^{\frac{3}{2}}}{2} \sum_{\pm} \frac{1}{[r^2 + (z \pm L)^2]^{\frac{3}{2}}} \left[\frac{3}{2} \cdot \frac{z \pm L}{r^2 + (z \pm L)^2} \begin{pmatrix} x \\ y \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right]. \tag{9.21}$$

We take $r = 2^2$, $L = 2^3$, $\rho = 2^{-5}$, and ICs $\mathbf{x} = (0, \rho, 0)$, $\mathbf{v} = (1, 0, 2.1)$. Under these conditions, the particle has insufficient energy to pass the magnetic mirror induced by the current loop at $z = 2^3$ due to the bounds on $B(\mathbf{x})$ over the particle imposed by the conservation of ε and adiabatic invariance of μ; in the exact trajecory, the particle should be reflected back to the plane $z = 0$.

Using a timestep $\Delta t = 2^{-4}$ (i.e. on the order of $\mathcal{O}[\rho]$) we compare our scheme (9.15) with $\mathcal{I}_n = \int_{T_n}$ to a simple IM method. Fig. 9.2 shows the evolution of the magnetic moment μ and the corrected magnetic moment $\mu + \rho\Delta\mu$ in either case; both schemes conserve energy up to solver tolerances. We see that, for our scheme (9.15) μ is restricted to the interval $[0.5000, 0.5004]$, whereas with IM μ drops from its initial value of 0.5 to a minimum value of around 0.12.

Fig. 9.3 shows the trajectories in either case. The particle in our scheme 9.15 is fully reflected by the mirror, whereas under IM the particle breaks through the mirror, due to the lack of preservation of the adiabatic invariance of μ; this occurs approximately when μ attains its lowest value, at around $t \approx 3.9$. Under these ICs in fact, among timesteps $\Delta t \in 2^{\mathbb{Z}}$ we require a timestep as low as $\Delta t = 2^{-9}$ (i.e. on the order of $\mathcal{O}[\rho^2]$) before we observe successful mirroring in the IM scheme.

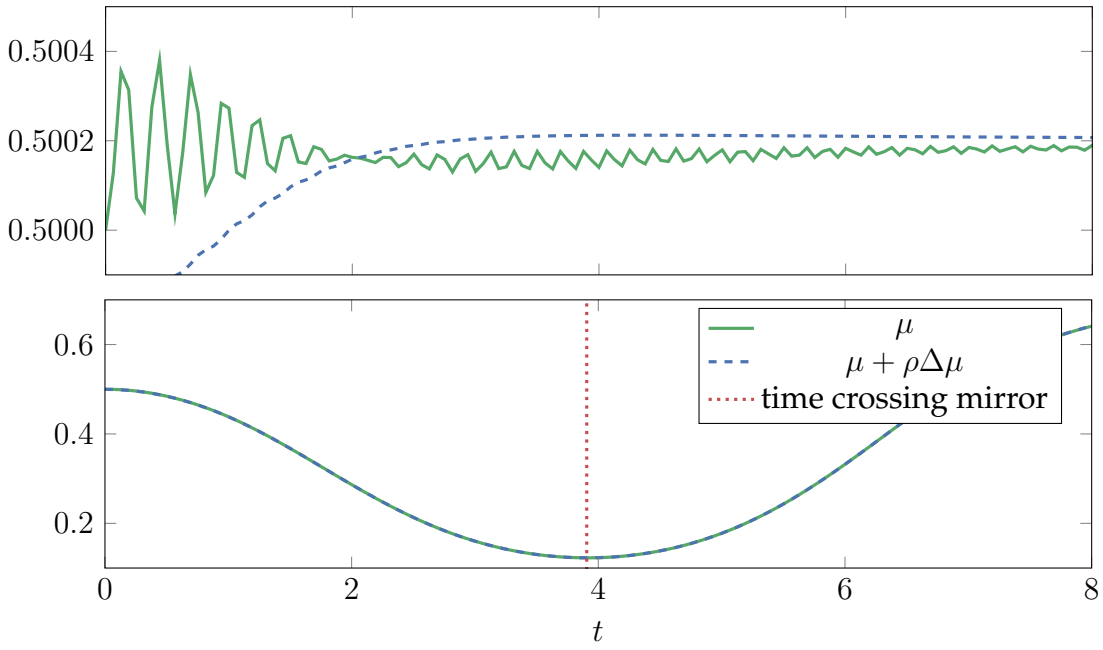


Figure 9.2: Magnetic moment μ and corrected magnetic moment $\mu + \rho\Delta\mu$ for the magnetic mirror test, using our scheme (9.15) (above) and IM (below). Note the differing scales on each y axis.

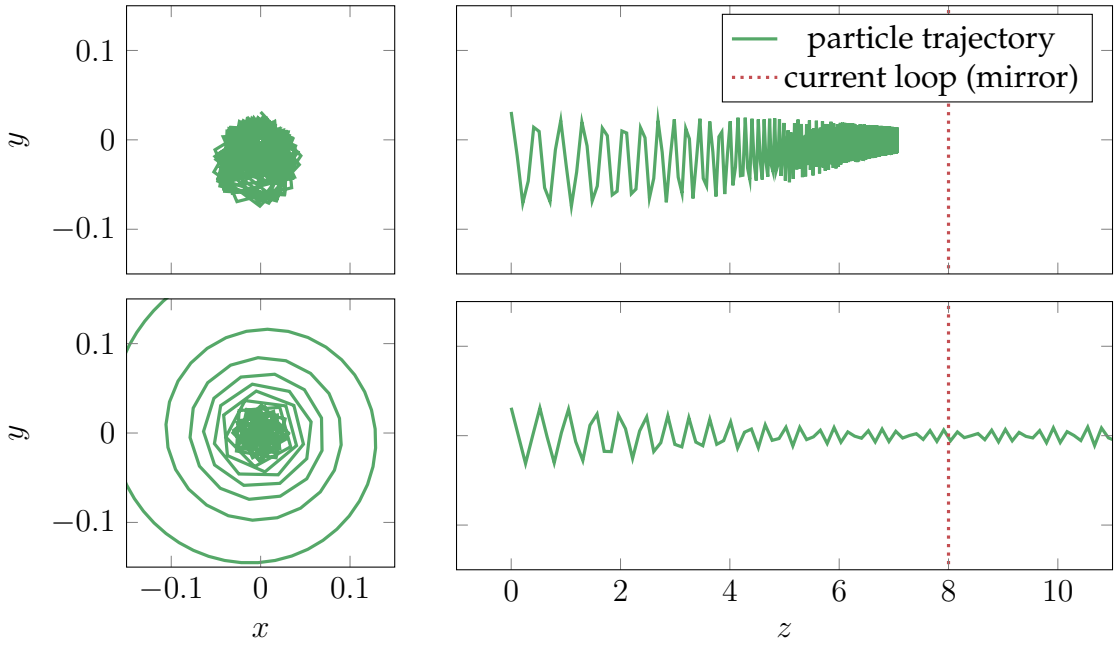


Figure 9.3: Numerical trajectories for the magnetic mirror test, using our scheme (9.15) (above) and IM (below).

"Hey, every triangle is a love triangle when you love triangles."

— James Acaster [Aca18]

10

Simplification of discretisations through FEEC: incompressible Navier–Stokes & MHD

Contents

We consider now two ways in which FEEC [Hip01; AFW06; AFW09; Arn18] may be applied to simplify schemes deriving from our framework in Chapter 3. We shall apply this to the energy- and helicity-stable integrator (3.28) presented in Chapter 3, alongside a novel energy- and enstrophy-stable integrator for the

incompressible NS equations, and an energy- and helicity-stable integrators for the incompressible Hall MHD equations. Both of the novel integrators represent AD systems (Assumption 3.11) allowing us to apply the existence and uniqueness results from Section 3.3.

The rest of this chapter proceeds as follows. In Section 10.1, we begin by overviewing the notation from exterior calculus that will be used throughout this chapter. In Section 10.2, we highlight the two ways in which we shall use FEEC to simplify schemes deriving from our framework, the first involving the elimination of LMs induced in the definition of AVs, and the second involving the reparametrisation of our schemes along FE complexes. In Section 10.3, we revisit the energy- and helicity-stable integrator (3.28) of Chapter 3. We show that, when $\mathbb{U}$ is defined as in (3.7b) and the spaces $\mathbb{Q}$, $\mathbb{V}$ satisfy a certain compatibility condition, the LM required to enforce the divergence-free condition on the auxiliary vorticity $\tilde{\boldsymbol{\omega}}$ (denoted θ in (3.32)) can be eliminated.

In Section 10.4, we again consider the incompressible NS equations, which, under appropriate BCs, dissipate both the energy $\frac{1}{2}\|\mathbf{u}\|^2$ and, in the 2D case, the enstrophy $\frac{1}{2}\|\operatorname{curl}\mathbf{u}\|^2$; we apply our framework to construct FE integrators that preserve both these structures. Applying the analytic results of Section 3.3, we prove similar existence and uniqueness results for this discretisation. These integrators initially use a stream function–vorticity parametrisation; provided the FE spaces in which these functions are defined derive from certain FE complexes, we are then able to reparametrise this into a more traditional velocity-vorticity formulation. To illustrate the stability properties offered by enstrophy stability, we consider a numerical test in the Euler case $\mathrm{Re} = \infty$.

In Section 10.5, we consider the incompressible Hall MHD equations, which conserve an energy, magnetic helicity, and hybrid helicity in the ideal limit, with the first of these dissipated in the nonideal case. We apply our framework to construct FE integrators that preserve each of these structures, for which we again show the existence and uniqueness results of Section 3.3 hold. These integrators use an electromagnetic (EM) potential parametrisation; similarly to Section 10.4, provided the FE spaces in which these functions are defined derive from certain FE complexes, we are able to reparametrise this into a more traditional EM field formulation. Moreover, similarly to Section 10.3, this FE complex compatibility

condition allows us to eliminate a certain LM enforcing the discrete divergence-free criteria on the introduced AVs.[1]

10.1 Notation & preliminaries

We begin by introducing various relevant notation and preliminaries from exterior calculus.

10.1.1 Primal complexes

We consider general Hilbert complexes

$$\cdots \longrightarrow V_{r-1} \xrightarrow{d_{r-1}} V_r \xrightarrow{d_r} V_{r+1} \xrightarrow{d_{r+1}} V_{r+2} \longrightarrow \cdots , \tag{10.1}$$

along sequences $(V_r, (\cdot,\cdot)_{V_r})$ of Hilbert spaces.[2] We shall denote the nullspace of (bounded) $d_r : V_r \to V_{r+1}$ by $\mathcal{N}d_r \subseteq V_r$, and its range by $\mathcal{R}d_r \subseteq V_{r+1}$.

We make broad use of the de Rham complex, which we write for simplicity with familiar vector proxies. For a general operator d, define the Hilbert space $H(d) \coloneqq \{u \in L^2 : du \in L^2\}$, using boldface when it is vector-valued. In 3D the Hilbert de Rham complex then takes the form

$$H^1 \xrightarrow{\text{grad}} \mathbf{H}(\text{curl}) \xrightarrow{\text{curl}} \mathbf{H}(\text{div}) \xrightarrow{\text{div}} L^2 . \tag{10.2a}$$

In 2D we write it as

$$H^1 \xrightarrow{\text{curl}} \mathbf{H}(\text{div}) \xrightarrow{\text{div}} L^2 . \tag{10.2b}$$

The 2D curl, mapping from scalars to vectors, is defined $\operatorname{curl}\phi \coloneqq (\partial_{x_2}\phi, -\partial_{x_1}\phi)$.

In Section 10.4, when considering the preservation of enstrophy in the incompressible NS equations, we require Stokes complexes in 3D and 2D (see Chen &

[1] As noted in Section 8.1, the resulting scheme identifies with that proposed by Laakmann, Hu & Farrell [LHF23]. The authors' analysis proceeds similarly to ours, with theirs offering a more careful handling of the function spaces that extends to the non-discrete setting, and ours extending to higher order in time.

[2] These are typically closed subspaces of the Hilbert space $H(d_r) \coloneqq \{v_r \in L^2 : d_r v_r \in L^2\}$

Huang [CH24]) i.e. certain de Rham complexes (10.2) with enhanced regularity. While different forms for the Stokes complex exist in 3D,[3] we consider the form[4]

$$H^1 \xrightarrow{\text{grad}} \mathbf{H}(\text{curl}^2) \xrightarrow{\text{curl}} \mathbf{H}^1 \xrightarrow{\text{div}} L^2 \,. \tag{10.3a}$$

In 2D, the spaces in the Stokes complex are generally more familiar:

$$H^2 \xrightarrow{\text{curl}} \mathbf{H}^1 \xrightarrow{\text{div}} L^2 \,. \tag{10.3b}$$

Considering a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, denote the boundary by $\partial\Omega$ with outward-facing unit normal $\mathbf{n}$ on $\partial\Omega$. We define then the following Hilbert subspaces with natural zero Dirichlet BCs:

$$H^1_0 \coloneqq \{u \in H^1 : u = 0 \text{ on } \partial\Omega\}, \tag{10.4a}$$

$$\mathbf{H}_0(\text{div}) \coloneqq \{\mathbf{u} \in \mathbf{H}(\text{div}) : \mathbf{u} \cdot \mathbf{n} = 0 \text{ on } \partial\Omega\}, \tag{10.4b}$$

$$\mathbf{H}_0(\text{curl}) \coloneqq \{\mathbf{u} \in \mathbf{H}(\text{curl}) : \mathbf{u} \times \mathbf{n} = 0 \text{ on } \partial\Omega\}. \tag{10.4c}$$

In the final case, $\times$ denotes the cross product. In 3D, this induces a de Rham complex (10.2) with zero BCs

$$H^1_0 \xrightarrow{\text{grad}} \mathbf{H}_0(\text{curl}) \xrightarrow{\text{curl}} \mathbf{H}_0(\text{div}) \xrightarrow{\text{div}} L^2 \,, \tag{10.5a}$$

while in 2D this induces the complex

$$H^1_0 \xrightarrow{\text{curl}} \mathbf{H}_0(\text{div}) \xrightarrow{\text{div}} L^2 \,. \tag{10.5b}$$

Exactness is especially important to us, required in particular for the conditions of Lemma 10.4 below. In Sections 10.3 & 10.5, the periodic BCs under consideration cause exactness to fail, due to the presence of constant harmonic forms. To remedy this, let the subscript $U_\#$ denote the restriction of a general space U to those functions $u \in U$ with zero mean $\int_\Omega u = 0$. When Ω is a rectangular domain with full periodic

[3] The term *Stokes complex* is more usually used to refer to the complex $\mathbf{H}^2 \to \mathbf{H}^1(\text{curl}) \to \mathbf{H}^1 \to L^2$. However, we neither consider nor use this complex here.

[4] Generally this space $\mathbf{H}(\text{curl}^2)$ is denoted in the literature by $\mathbf{H}(\text{grad curl})$, the two spaces being equivalent on sufficiently regular domains. We choose this notation however to better align with our discrete problem in Section 10.4.

BCs (i.e. topologically equivalent to a d-torus) we may take zero means on each space to ensure exactness; this is equivalent to removing the harmonic forms. In 3D for example, whereas the typical de Rham (10.2) and Stokes (10.3) complexes are not exact on periodic domains, the following is:

$$H^1_{\#} \xrightarrow{\text{grad}} \mathbf{H}_{\#}(\text{curl}) \xrightarrow{\text{curl}} \mathbf{H}_{\#}(\text{div}) \xrightarrow{\text{div}} L^2_{\#} \,. \tag{10.6}$$

In Section 10.4, we consider the intersection of the Stokes complexes (of enhanced regularity) (10.3), and the standard de Rham complexes (or reduced regularity) with zero Dirichlet BCs (10.5). In 3D we consider the complex

$$H^1_0 \xrightarrow{\text{grad}} \mathbf{H}(\text{curl}^2) \cap \mathbf{H}_0(\text{curl}) \xrightarrow{\text{curl}} \mathbf{H}^1 \cap \mathbf{H}_0(\text{div}) \xrightarrow{\text{div}} L^2 \,, \tag{10.7a}$$

while in 2D we consider the complex

$$H^2 \cap H^1_0 \xrightarrow{\text{curl}} \mathbf{H}^1 \cap \mathbf{H}_0(\text{div}) \xrightarrow{\text{div}} L^2 \,. \tag{10.7b}$$

Lemma 10.1 (Exactness of the Stokes complexes with zero BCs)**.** *When substituting the final space L^2 for its zero-averaged counterpart $L^2_{\#}$, the Stokes complexes with zero BCs* (10.7) *are exact over contractible domains.*

Proof. Considering the 3D case (10.7a) we prove exactness at each space, with similar arguments holding in the 2D case (10.7b):

1. For $v \in \mathcal{N}\,\text{grad} \subset H^1_0$, v is constant, implying $v = 0$ due to the BCs on H^1_0.
2. For $\mathbf{v} \in \mathcal{N}\,\text{curl} \subset \mathbf{H}(\text{curl}^2) \cap \mathbf{H}_0(\text{curl})$, the exactness of (10.5a) immediately implies the existence of $\phi \in H^1_0$ such that $\mathbf{v} = \nabla\phi$.
3. For $\mathbf{v} \in \mathcal{N}\,\text{div} \subset \mathbf{H}^1 \cap \mathbf{H}_0(\text{div})$, the exactness of (10.5a) similarly implies the existence of $\boldsymbol{\phi} \in \mathbf{H}_0(\text{curl})$ such that $\mathbf{v} = \text{curl}\,\boldsymbol{\phi}$; since $\mathbf{v} \in \mathbf{H}^1$ and $\mathbf{v} = \text{curl}\,\boldsymbol{\phi}$, we see we have sufficiently regularity $\boldsymbol{\phi} \in \mathbf{H}(\text{grad curl}) = \mathbf{H}(\text{curl}^2)$.
4. For $v \in L^2_{\#}$, we see by solving the associated Neumann problem that there exists (over sufficiently regular domains) $\Delta^{-1}v \in H^2$ such that $\Delta[\Delta^{-1}v] = v$ and $\nabla\Delta^{-1}v \cdot \mathbf{n} = 0$; defining $\boldsymbol{\phi} \coloneqq \nabla\Delta^{-1}v$ gives $\boldsymbol{\phi} \in \mathbf{H}^1 \cap \mathbf{H}_0(\text{div})$ satisfying $\text{div}\,\boldsymbol{\phi} = v$.

□

10.1.2 Dual complexes

Denote by $d_r^* : V_{r+1} \to V_r$ the Hilbert adjoint of $d_r : V_r \to V_{r+1}$, defined for $\phi_{r+1} \in V_{r+1}$ such that $(d_r^* \phi_{r+1}, \theta_r)_{V_r} = (\phi_{r+1}, d_r \theta_r)_{V_{r+1}}$ for all $\theta_r \in V_r$. The associated Hilbert dual complex of (10.1) is then

$$\cdots \longleftarrow V_{r-1} \xleftarrow{d_{r-1}^*} V_r \xleftarrow{d_r^*} V_{r+1} \xleftarrow{d_{r+1}^*} V_{r+2} \longleftarrow \cdots . \tag{10.8}$$

Our interaction with these adjoint operators is generally restricted to considering their kernels $\mathcal{N} d_r^*$, e.g. in Lemma 10.3 & 10.4 below.

There are many similar ideas of dual complexes in the literature. To fix ideas with the Hilbert dual (10.8), consider the 3D de Rham complex with zero BCs (10.5a); under our definition, this has a dual

$$\begin{aligned} H_0^1 &\xleftarrow{-(\mathrm{id} - \operatorname{div}\operatorname{grad})^{-1} \operatorname{div}} \mathbf{H}_0(\mathrm{curl}) \xleftarrow{(\mathrm{id} + \operatorname{curl}^2)^{-1} \operatorname{curl}} \cdots \\ \cdots &\xleftarrow{(\mathrm{id} + \operatorname{curl}^2)^{-1} \operatorname{curl}} \mathbf{H}_0(\mathrm{div}) \xleftarrow{-(\mathrm{id} - \operatorname{grad}\operatorname{div})^{-1} \operatorname{grad}} L^2 , \end{aligned} \tag{10.9}$$

where id is the identity map, and these inverse maps $(\mathrm{id} \pm *)^{-1}$ are well-defined from the Dirichlet BCs inherited from the Hilbert spaces (10.4). While the inclusion of these inverse maps in the dual operators may seem unusual, we see immediately that the dual kernels $\mathcal{N} d_r^*$ still take the familiar forms,

$$\mathcal{N} \operatorname{grad}^* = \{\mathbf{v} \in \mathbf{H}_0(\mathrm{curl}) : -(\mathbf{v}, \operatorname{grad} \phi) = 0 \text{ for all } \phi \in H_0^1\}, \tag{10.10a}$$

$$\mathcal{N} \operatorname{curl}^* = \{\mathbf{v} \in \mathbf{H}_0(\mathrm{div}) : (\mathbf{v}, \operatorname{curl} \boldsymbol{\phi}) = 0 \text{ for all } \boldsymbol{\phi} \in \mathbf{H}_0(\mathrm{curl})\}, \tag{10.10b}$$

$$\mathcal{N} \operatorname{div}^* = \{v \in L^2 : -(v, \operatorname{div} \boldsymbol{\phi}) = 0 \text{ for all } \boldsymbol{\phi} \in \mathbf{H}_0(\mathrm{div})\}, \tag{10.10c}$$

with these variational definitions holding similarly on discrete subcomplexes.

10.2 Outline of techniques from FEEC

We highlight here the two systems by which we may apply FEEC to simplify the schemes deriving from our framework (Algorithm 3.5) without affecting the underlying scheme.

Method 1: Elimination of LMs The first idea we present involves the elimination of LMs. Typically, when we consider DAEs in our framework, we enforce the algebraic structures through a restriction on the solution space $\mathbb{U}$; in practical implementation, these restrictions then manifest as LMs.

Example (Incompressible NS)

In the energy- and helicity-stable incompressible NS integrator (3.28) presented in Chapter 3, we enforced the divergence-free condition as a discrete condition on $\mathbb{U}$ (3.7). For practical implementation, this divergence-free restriction is implemented via a LM, analogous to the pressure p (3.32).

Each AV we introduce therein is then a projection into this same restricted $\mathbb{U}$ space, and induces its own LM.

Example (Incompressible NS)

The AV $\boldsymbol{\omega}$ represents the vorticity. As it lies in $\mathbb{U}$, it requires its own LM. This is denoted by θ in (3.32).

FEEC can alleviate the need to introduce certain auxiliary LMs. When the spaces in the definition of $\mathbb{U}$ satisfy specific complex compatibility conditions, these auxiliary LMs can simply be removed, with no effect on the discrete solution. To fix this idea, we introduce the following two lemmas.

Lemma 10.2 (Elimination of LMs: primal)**.** *Suppose $\phi_{r+1} \in V_{r+1}$ is defined, for some $\varphi_r \in V_r$, by the projection*

$$(\phi_{r+1}, \theta_{r+1})_{V_{r+1}} = (d_r\varphi_r, \theta_{r+1})_{V_{r+1}}, \tag{10.11}$$

for all $\theta_{r+1} \in V_{r+1}$. Then we may equivalently seek $\phi_{r+1} \in \mathcal{N}d_{r+1} \subset V_{r+1}$ (i.e. the nullspace of d_{r+1}) such that (10.11) holds for all $\theta_{r+1} \in \mathcal{N}d_{r+1}$.

Proof. It suffices to show that, for ϕ_{r+1} as defined by (10.11), $\phi_{r+1} \in \mathcal{N}d_{r+1}$, i.e. $d_{r+1}\phi_{r+1} = 0$. As $d_r\varphi_r \in V_r$ automatically, the projection is trivial, i.e. $\phi_{r+1} = d_r\varphi_r$. The result then holds immediately by the complex property. □

Lemma 10.2 has an important analogue in the dual complex, Lemma 10.3.

Lemma 10.3 (Elimination of LMs: dual)**.** *Let V_{r+1} be continuously embedded in some larger space $\hat{V}_{r+1} \hookleftarrow V_{r+1}$. Suppose $\phi_r \in V_r$ is defined, for some $\hat{\varphi}_{r+1} \in \hat{V}_{r+1}$, by the projection*

$$(\phi_r, \theta_r)_{V_r} = (\hat{\varphi}_{r+1}, d_r\theta_r)_{\hat{V}_{r+1}}, \tag{10.12}$$

for all $\theta_r \in V_r$. Then we may equivalently seek $\phi_r \in \mathcal{N}d^*_{r-1} \subset V_r$ (i.e. the nullspace of the dual of d_{r-1}) such that (10.12) holds for all $\theta_r \in \mathcal{N}d^*_{r-1}$.

Proof. It suffices to show that $\phi_r \in \mathcal{N}d^*_{r-1}$, i.e. $d^*_{r-1}\phi_r = 0$. Equivalently we may show that for all $\vartheta_{r-1} \in V_{r-1}$,

$$(\phi_r, d_{r-1}\vartheta_{r-1})_{V_r} = 0. \tag{10.13}$$

Considering $\theta_r = d_{r-1}\vartheta_{r-1}$ in (10.12) this holds immediately by the complex property. □

Notably, this dual result is stronger, in so far as it holds for all $\hat{\varphi}_{r+1} \in \hat{V}_{r+1} \hookleftarrow V_{r+1}$, whereas the primal result only necessarily holds for $\varphi_r \in V_r$.

Example (Incompressible NS)

Under certain conditions, we are able to appeal to Lemma 10.3 to show the LM θ may be eliminated from our discrete scheme, without affecting the solution. This is discussed in detail in Section 10.3.

Method 2: Reparametrisation along complexes The second idea we present involves the reparametrisation of a PDE discretisation in terms of certain function's derivative. In particular, we highlight that we seek to use FEEC to reparametrise our discretisations exactly, i.e. such that, despite using different variables, the FE scheme and its discrete solutions remain necessarily unchanged.

In certain applications of our framework, we arrive at a discretisation in which a certain quantity appears solely in terms of a certain derivative, e.g. its curl. Often, this derivative may in fact be a more physically meaningful quantity, e.g. the magnetic field instead of the magnetic potential. We may consider then those situations where our discretisation may be equivalently rewritten in terms of this derivative,

to be defined in more traditional variables. This is of further interest from the perspective of computational complexity, as the spaces later in a FE complex are typically of lower regularity, leading to discrete problems that are generally easier to solve and precondition.[5] FEEC gives sufficient conditions where we may make this reparametrisation, detailed in the following lemma.

Lemma 10.4 (Reparametrisation along complexes)**.** *Suppose the complex is exact at V_r and V_{r+1}. The operator d_r then defines an isomorphism from $\mathcal{N}d^*_{r-1} \subset V_r$ (the nullspace of the dual of d_{r-1}) to $\mathcal{N}d_{r+1} \subset V_{r+1}$ (the nullspace of d_{r+1}).*

Proof. To show d_r is an isomorphism, we must show surjectivity and injectivity.

For surjectivity, noting $\mathcal{N}d_{r+1} = \mathcal{R}d_r = d_r[V_r]$ by exactness, it suffices to show $d_r[\mathcal{N}d^*_{r-1}] = d_r[V_r]$. By the Hodge decomposition (see Arnold [Arn18, Sec. 4.2]) V_r may be decomposed as $V_r = \mathcal{N}d_r \oplus \mathcal{N}d^*_{r-1}$; evaluating then $d_r[V_r]$,

$$d_r[V_r] = d_r[\mathcal{N}d_r \oplus \mathcal{N}d^*_{r-1}] = d_r[\mathcal{N}d^*_{r-1}]. \tag{10.14}$$

For injectivity, it suffices to show that if $\phi_r \in \mathcal{N}d^*_{r-1}$, satisfies $d_r\phi_r = 0$, then $\phi_r = 0$. For $\phi_r \in \mathcal{N}d^*_{r-1} = \mathcal{R}d^*_r$, there exists $\varphi_{r+1} \in V_{r+1}$ such that $\phi_r = d^*_r\varphi_{r+1}$, i.e. for all $\theta_r \in V_r$

$$(\phi_r, \theta_r) = (\varphi_{r+1}, d_r\theta_r). \tag{10.15a}$$

In particular, considering $\theta_r = \phi_r$,

$$\|\phi_r\|^2 = (\phi_r, \phi_r) = (\varphi_{r+1}, d_r\phi_r). \tag{10.15b}$$

When $d_r\phi_r = 0$ we must have $\phi_r = 0$. □

Under such conditions, if we have in our discretisation a quantity $\phi_r \in \mathcal{N}d^*_{r-1}$ that appears only through its derivative $d_r\phi_r$, then we may equivalently parametrise and implement our scheme in $\varphi_{r+1} = d_r\phi_r$ with $\varphi_{r+1} \in \mathcal{N}d_{r+1}$.

[5] Consider for example hybridisation, which in its typical form can only be performed on $\mathbf{H}(\mathrm{div})$-conforming spaces.

10.3 Energy- & helicity-stable integrators for the incompressible Navier–Stokes equations (revisited): elimination of the Lagrange multiplier

We revisit the energy- and helicity-stable integrator (3.28) proposed in Chapter 3 for the incompressible NS equations, again with periodic BCs. For simplicity, we shall assume $\mathcal{I}_n$ to be an S-node GL quadrature rule, such that, as observed in Section 4.3, this is equivalent to an S-stage Gauss collocation method applied to the following semi-discretisation: find $(\mathbf{u}, \boldsymbol{\omega}) \in \mathbb{U}^2$ (for discretely divergence-free $\mathbb{U}$ defined as in either (3.7a) or (3.7b)) such that

$$(\dot{\mathbf{u}}, \mathbf{v}) = (\mathbf{u} \times \boldsymbol{\omega}, \mathbf{v}) - \frac{1}{\mathrm{Re}}(\operatorname{curl} \mathbf{u}, \operatorname{curl} \mathbf{v}), \tag{10.16a}$$

$$(\boldsymbol{\omega}, \boldsymbol{\chi}) = (\operatorname{curl} \mathbf{u}, \boldsymbol{\chi}), \tag{10.16b}$$

for all $(\mathbf{v}, \boldsymbol{\chi}) \in \mathbb{U}^2$. Note, the elimination of the auxiliary velocity $\tilde{\mathbf{u}}$ in the case of $\mathcal{I}_n$ being a GL quadrature rule is simple, as $\tilde{\mathbf{u}}$ simply becomes the interpolant of $\mathbf{u}$ at the GL points in T_n, and is therefore effectively equivalent. We elect also to rewrite the dissipative term $(\nabla \mathbf{u}, \nabla \mathbf{v})$ in the form $(\operatorname{curl} \mathbf{u}, \operatorname{curl} \mathbf{v})$, equivalent in the continuous setting for exactly divergence-free functions $\mathbf{u}$, as it better aligns with the regularity assumptions when applying FEEC in the following subsection.[6]

10.3.1 Application of FEEC

Expanding out the LMs in (10.16) yields a 4-field discretisation (excluding the LMs enforcing the zero-mean conditions $\int_\Omega \mathbf{u} = \mathbf{0}$ and $\int_\Omega \boldsymbol{\omega} = \mathbf{0}$). Through the use of FE spaces compatible with FEEC, we may eliminate one of these LMs, the one enforcing the discrete divergence-free condition on $\boldsymbol{\omega}$.

Let us first apply IBP to rewrite (10.16b) in the form

$$(\boldsymbol{\omega}, \boldsymbol{\chi}) = (\mathbf{u}, \operatorname{curl} \boldsymbol{\chi}). \tag{10.17}$$

We take then $\mathbb{U}$ to be defined as in (3.7b),

$$\mathbb{U} \coloneqq \left\{ \mathbf{u} \in \mathbb{V} : -(\mathbf{u}, \nabla q) = 0 \text{ for all } q \in \mathbb{Q} \text{ and } \int_\Omega \mathbf{u} = \mathbf{0} \right\}, \tag{10.18}$$

[6] It may readily be confirmed that, under the assumptions in the following section (and again restricting our attention to functions of zero mean), the $\mathbf{H}(\operatorname{curl})$ seminorm $\mathbf{u} \mapsto \|\operatorname{curl} \mathbf{u}\|$ defines a norm on $\mathbb{U}$. Accordingly, this modification does not affect the existence and uniqueness results of Section 3.3.

and suppose the spaces $\mathbb{V}$, $\mathbb{Q}$ exist as part of a subcomplex of the periodic de Rham complex (10.6):

$$\begin{array}{ccccccc} H^1 & \xrightarrow{\text{grad}} & \mathbf{H}_{\#}(\text{curl}) & \xrightarrow{\text{curl}} & \mathbf{H}_{\#}(\text{div}) & \xrightarrow{\text{div}} & L^2 \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ \mathbb{Q} & \xrightarrow{\text{grad}} & \mathbb{V} & \xrightarrow{\text{curl}} & \overline{\mathbb{V}} & \xrightarrow{\text{div}} & \overline{\mathbb{Q}} \end{array} . \tag{10.19}$$

Exactness is ensured at the vector-valued spaces $\mathbb{V}$ and $\overline{\mathbb{V}}$ by eliminating the harmonic forms.[7] The space $\mathbb{U}$ may be identified as the nullspace $\mathcal{N}\,\text{grad}^* \subset \mathbb{V}$ of the dual operator $\text{grad}^* : \mathbb{V} \to \mathbb{Q}$. Lemma 10.3 implies then that the projection (10.17) defining the auxiliary vorticity $\boldsymbol{\omega} \in \mathbb{U}$ may equivalently be written as a projection in the larger space $\mathbb{V}$. Thus, the integrator (10.16) may be written equivalently as follows: find $(\mathbf{u}, \boldsymbol{\omega}) \in \mathbb{U} \times \mathbb{V}$ such that (10.16) holds for all $(\mathbf{v}, \boldsymbol{\chi}) \in \mathbb{U} \times \mathbb{V}$. Expanding the LMs in $\mathbb{U}$ gives the final equivalent semi-discretisation: find $(\mathbf{u}, p, \boldsymbol{\omega}) \in \mathbb{V} \times \mathbb{Q} \times \mathbb{V}$ such that

$$(\dot{\mathbf{u}}, \mathbf{v}) = (\mathbf{u} \times \boldsymbol{\omega}, \mathbf{v}) - (\nabla p, \mathbf{v}) - \frac{1}{\text{Re}}(\text{curl}\,\mathbf{u}, \text{curl}\,\mathbf{v}), \tag{10.20a}$$

$$0 = (\mathbf{u}, \nabla q), \tag{10.20b}$$

$$(\boldsymbol{\omega}, \boldsymbol{\chi}) = (\text{curl}\,\mathbf{u}, \boldsymbol{\chi}), \tag{10.20c}$$

for all $(\mathbf{v}, q, \boldsymbol{\chi}) \in \mathbb{V} \times \mathbb{Q} \times \mathbb{V}$. Discretising in time with a Gauss method, energy Q_1 and helicity Q_2 stability may be shown directly in (10.20) by considering $(\mathbf{v}, q) = (\mathbf{u}, p)$ and $(\mathbf{v}, \boldsymbol{\chi}) = (\boldsymbol{\omega}, \dot{\mathbf{u}} + \nabla p)$ respectively.

10.4 Energy- & enstrophy-stable integrators for the incompressible Navier–Stokes equations: stream function–to–velocity reparametrisation

We now reconsider the incompressible NS equations (3.1) through the perspective of enstrophy stability, in place of helicity stability. We intend to construct a stable integrator that preserves the behaviour of the energy and enstrophy, i.e. dissipation and ideal conservation of energy and, in the 2D case, enstrophy. Each subsection herein will first consider the 3D case, for which the differential operators are generally

[7] We do not eliminate the harmonic forms at $\mathbb{Q}$ and $\overline{\mathbb{Q}}$, as exactness there is not required.

more familiar and we are able to at least preserve the equation governing the evolution of enstrophy even if it is not (necessarily) dissipative; we shall consider then the 2D case, for which the enstrophy evolution equation becomes a stronger dissipation result.

3D In place of periodic BCs, we shall assume the no-flux and no-tangential-stress BCs

$$\mathbf{u} \cdot \mathbf{n} = 0, \qquad \operatorname{curl} \mathbf{u} \times \mathbf{n} = \mathbf{0}. \tag{10.21a}$$

We have then two QoIs,

$$Q_1(\mathbf{u}) \coloneqq \frac{1}{2}\|\mathbf{u}\|^2, \qquad Q_3(\mathbf{u}) \coloneqq \frac{1}{2}\|\operatorname{curl} \mathbf{u}\|^2, \tag{10.21b}$$

the energy (as defined in (3.15)) and enstrophy respectively. For an exact solution $\mathbf{u}$ of the incompressible NS equations (3.1), Q_1 and Q_3 satisfy

$$\dot{Q}_1 = -\frac{1}{\mathrm{Re}}\|\operatorname{curl} \mathbf{u}\|^2 (\leq 0), \qquad \dot{Q}_3 = -\int_\Omega \mathbf{u} \cdot (\operatorname{curl} \mathbf{u} \cdot \nabla \operatorname{curl} \mathbf{u}) - \frac{1}{\mathrm{Re}}\|\operatorname{curl}^2 \mathbf{u}\|^2. \tag{10.21c}$$

Remark 10.5 (Different forms of the no-tangential-stress BC)**.** *The more common (and physically meaningful) form for the no-tangential-stress BC in the NS equations is* $\mathbf{n} \cdot (\nabla \mathbf{u} + \nabla \mathbf{u}^\top) \cdot \mathbf{t} = 0$*, for any tangential vector* $\mathbf{t}$ *on* $\partial\Omega$ *such that* $\mathbf{n} \cdot \mathbf{t} = 0$*. However, it may readily be shown that this is equivalent to the condition* $\operatorname{curl} \mathbf{u} \times \mathbf{n} = \mathbf{0}$ *under the no-flux condition* $\mathbf{u} \cdot \mathbf{n} = 0$*, which will be a more convenient form for our purposes.*

2D In 2D, the BCs (10.21a) become

$$\mathbf{u} \cdot \mathbf{n} = 0, \qquad \operatorname{rot} \mathbf{u} = 0, \tag{10.22a}$$

where $\operatorname{rot} \mathbf{u} \coloneqq \partial_{x_1} u_2 - \partial_{x_2} u_1$ for $\mathbf{u} = (u_1, u_2)$. The energy Q_1 and enstrophy Q_3 are then defined

$$Q_1(\mathbf{u}) \coloneqq \frac{1}{2}\|\mathbf{u}\|^2, \qquad Q_3(\mathbf{u}) \coloneqq \frac{1}{2}\|\operatorname{rot} \mathbf{u}\|^2. \tag{10.22b}$$

These evolve, for an exact solution $\mathbf{u}$, according to

$$\dot{Q}_1 = -\frac{1}{\mathrm{Re}}\|\operatorname{rot} \mathbf{u}\|^2 (\leq 0), \qquad \dot{Q}_3 = -\frac{1}{\mathrm{Re}}\|\operatorname{rot} \operatorname{curl} \mathbf{u}\|^2 (\leq 0). \tag{10.22c}$$

In particular, both of these quantities are dissipated, and conserved in the ideal case.

Proceeding with our framework similarly to the energy- and helicity-stable FE integrator in Chapter 3 would motivate in Step D the introduction of AVs approximating the velocity $\mathbf{u}$ and the curl of the vorticity $\operatorname{curl}^2 \mathbf{u}$ (or, in 2D, $\operatorname{curl}\operatorname{rot}\mathbf{u}$). This poses an issue in Step E; when substituting in the AVs on the RHS of (3.8), there is no clear place for the AV approximating $\operatorname{curl}^2 \mathbf{u}$.

To circumvent this issue, we first pass to a stream function $\boldsymbol{\psi}$ formulation of (3.1). This has the consequence that the introduced AVs for energy and enstrophy stability respectively in Step D will instead approximate the stream function $\boldsymbol{\psi}$ and vorticity $\operatorname{curl}^2 \boldsymbol{\psi}$ (or, in 2D, $\Delta\psi$) respectively; these will have clear places in the RHS of (3.8) in Step E.

Remark 10.6 (Assumption of contractible domain)**.** *To transfer to the stream function formulation, we shall assume a contractible domain Ω. Taking care to handle the domain cohomology appropriately, there exist stream function formulations of the incompressible NS equations on non-contractible domains through the inclusion of appropriate harmonic functions, however we assume contractibility here for simplicity, and defer the analysis of topological non-trivial domains to future work.*

3D In 3D, with a contractible domain Ω, the Hodge (or in this case Helmholtz) decomposition (see Arnold [Arn18, Sec. 4.2]) indicates that, up to regularity, any divergence-free function with zero normal component on the boundary may be written as the curl of a divergence-free function with zero tangential component on the boundary. In particular, we shall write $\mathbf{u} = \operatorname{curl}\boldsymbol{\psi}$, writing the BCs (10.21a) as

$$\boldsymbol{\psi} \times \mathbf{n} = \mathbf{0}, \qquad \operatorname{curl}^2 \boldsymbol{\psi} \times \mathbf{n} = \mathbf{0}. \tag{10.23}$$

Energy Q_1 and enstrophy Q_3 are defined similarly to (10.21b) on $\boldsymbol{\psi}$ as

$$Q_1(\boldsymbol{\psi}) \coloneqq \frac{1}{2}\|\operatorname{curl}\boldsymbol{\psi}\|^2, \qquad Q_3(\boldsymbol{\psi}) \coloneqq \frac{1}{2}\|\operatorname{curl}^2\boldsymbol{\psi}\|^2. \tag{10.24}$$

Taking the curl of (3.1a) to eliminate the pressure, we have the incompressible NS equations in stream function formulation,

$$\operatorname{curl}^2 \dot{\boldsymbol{\psi}} = \operatorname{curl}[\operatorname{curl}\boldsymbol{\psi} \times \operatorname{curl}^2 \boldsymbol{\psi}] - \frac{1}{\mathrm{Re}}\operatorname{curl}^4 \boldsymbol{\psi}, \tag{10.25a}$$

$$0 = \operatorname{div}\boldsymbol{\psi}. \tag{10.25b}$$

We may now apply our framework to (10.25) to construct the desired energy- and enstrophy-stable integrator.

Application of framework (Algorithm 3.5)

A. Let vector-valued $\mathbb{V}$, satisfying $\boldsymbol{\phi}\times\mathbf{n}=\mathbf{0}$ on $\partial\Omega$ for all $\boldsymbol{\phi}\in\mathbb{V}$, and scalar-valued $\mathbb{Q}$, satisfying $\eta=0$ on $\partial\Omega$ for all $\eta\in\mathbb{Q}$, be suitable finite-dimensional function spaces. Similar to in Chapter 3, define a $\mathbb{Q}$-discretely divergence-free subspace $\mathbb{U}\subset\mathbb{V}$ as in either (3.7a) or (10.37). We then arrive at our semidiscrete form: find $\boldsymbol{\psi}\in\mathbb{U}$ such that

$$(\operatorname{curl}\dot{\boldsymbol{\psi}},\operatorname{curl}\boldsymbol{\phi})=(\operatorname{curl}\boldsymbol{\psi}\times\operatorname{curl}^2\boldsymbol{\psi},\operatorname{curl}\boldsymbol{\phi})-\frac{1}{\mathrm{Re}}(\operatorname{curl}^2\boldsymbol{\psi},\operatorname{curl}^2\boldsymbol{\phi}),\tag{10.26}$$

at all times $t\in\mathbb{R}_+$ and for all $\boldsymbol{\phi}\in\mathbb{U}$.

B. This is fully discretised in time over $\mathbb{X}_n$ defined as in (3.10) with $\mathcal{I}_n$. As in Section 10.3, let us again assume for simplicity that $\mathcal{I}_n$ is simply an S-node GL quadrature rule, such that the fully discrete system is equivalent to an S-stage Gauss collocation method applied to (10.26).

C. As stated above, the associated test functions can be identified as the stream function $\boldsymbol{\psi}$ and vorticity $-\Delta\boldsymbol{\psi}$ respectively.

D. Nominally, our framework indicates we must introduce AVs $(\tilde{\boldsymbol{\psi}},\boldsymbol{\omega})\in(\dot{\mathbb{X}}_n)^2$, projections of $(\boldsymbol{\psi},-\Delta\boldsymbol{\psi})$ into $\dot{\mathbb{X}}_n$ under the $\mathbf{H}(\operatorname{curl})$ seminorm:

$$\mathcal{I}_n[(\operatorname{curl}\tilde{\boldsymbol{\psi}},\operatorname{curl}\tilde{\boldsymbol{\phi}})]=\int_{T_n}(\operatorname{curl}\boldsymbol{\psi},\operatorname{curl}\tilde{\boldsymbol{\phi}}),\tag{10.27a}$$

$$\mathcal{I}_n[(\operatorname{curl}\boldsymbol{\omega},\operatorname{curl}\boldsymbol{\chi})]=\int_{T_n}(\operatorname{curl}^2\boldsymbol{\psi},\operatorname{curl}\boldsymbol{\chi}^2),\tag{10.27b}$$

for all $(\tilde{\boldsymbol{\phi}},\boldsymbol{\chi})\in(\dot{\mathbb{X}}_n)^2$. With $\mathcal{I}_n$ an S-node GL quadrature rule however, this is equivalent at each of the GL points in T_n to $\tilde{\boldsymbol{\psi}}=\boldsymbol{\psi}$ and $\boldsymbol{\omega}\in\mathbb{U}$ defined such that

$$(\operatorname{curl}\boldsymbol{\omega},\operatorname{curl}\boldsymbol{\chi})=(\operatorname{curl}^2\boldsymbol{\psi},\operatorname{curl}^2\boldsymbol{\chi})\tag{10.28}$$

for all $\boldsymbol{\chi}\in\mathbb{U}$. Note, similarly to the construction of the helicity-stable integrator in Section 3.1, the vorticity $\operatorname{curl}^2\boldsymbol{\psi}$ should, in the continuous case, satisfy $\operatorname{div}[\operatorname{curl}^2\boldsymbol{\psi}]=0$, $\int_\Omega\operatorname{curl}^2\boldsymbol{\psi}=\mathbf{0}$ and $\operatorname{curl}^2\boldsymbol{\psi}\times\mathbf{n}=\mathbf{0}$ on the boundary $\partial\Omega$; these results are analogous to the restrictions on $\mathbb{U}$, and as such it is appropriate to approximate $\operatorname{curl}^2\boldsymbol{\psi}$ by $\boldsymbol{\omega}\in\mathbb{U}$.

E. We now introduce the AVs $\tilde{\boldsymbol{\psi}}$, $\boldsymbol{\omega}$ into the RHS of (10.26) as

$$(\operatorname{curl}\dot{\boldsymbol{\psi}},\operatorname{curl}\boldsymbol{\phi})=(\operatorname{curl}\tilde{\boldsymbol{\psi}}\times\boldsymbol{\omega},\operatorname{curl}\boldsymbol{\phi})-\frac{1}{\mathrm{Re}}(\operatorname{curl}\boldsymbol{\omega},\operatorname{curl}\boldsymbol{\phi}),\tag{10.29}$$

where $\tilde{\boldsymbol{\psi}}$ may be substituted for $\boldsymbol{\psi}$ at the GL points.

F. The final SP scheme is then any S-stage Gauss collocation method applied to the following semidiscrete system: find $(\boldsymbol{\psi}, \boldsymbol{\omega}) \in \mathbb{U}^2$ such that

$$(\operatorname{curl} \dot{\boldsymbol{\psi}}, \operatorname{curl} \boldsymbol{\phi}) = (\operatorname{curl} \boldsymbol{\psi} \times \boldsymbol{\omega}, \operatorname{curl} \boldsymbol{\phi}) - \frac{1}{\mathrm{Re}}(\operatorname{curl} \boldsymbol{\omega}, \operatorname{curl} \boldsymbol{\phi}), \tag{10.30a}$$

$$(\operatorname{curl} \boldsymbol{\omega}, \operatorname{curl} \boldsymbol{\chi}) = (\operatorname{curl}^2 \boldsymbol{\psi}, \operatorname{curl}^2 \boldsymbol{\chi}), \tag{10.30b}$$

for all $(\boldsymbol{\phi}, \boldsymbol{\chi}) \in \mathbb{U}^2$.

Theorem 10.7 (Energy & enstrophy stability of the incompressible NS integrator)**.** *When integrating in time using a Gauss method, the incompressible NS integrator* (10.30) *is energy- and enstrophy-stable, with discrete analogues of the following results holding across each timestep T_n:*

$$\dot{Q}_1 = -\frac{1}{\mathrm{Re}} \| \operatorname{curl}^2 \boldsymbol{\psi} \|^2 \leq 0, \tag{10.31a}$$

$$\dot{Q}_3 = -\int_\Omega \operatorname{curl} \boldsymbol{\psi} \cdot (\boldsymbol{\omega} \cdot \nabla \boldsymbol{\omega}) + \frac{1}{\mathrm{Re}} \| \operatorname{curl} \boldsymbol{\omega} \|^2. \tag{10.31b}$$

Proof. The former energy stability result (10.31a) holds by considering $(\boldsymbol{\phi}, \boldsymbol{\chi}) = (\boldsymbol{\psi}, \boldsymbol{\psi})$ while the latter enstrophy stability result (10.31b) holds by considering $(\boldsymbol{\phi}, \boldsymbol{\chi}) = (\boldsymbol{\omega}, \dot{\boldsymbol{\psi}})$. □

2D In 2D, with a contractible domain, the Hodge decomposition similarly indicates that any sufficiently regular divergence-free function with zero normal component on the boundary may be written as the curl of a function that is zero on the boundary. We therefore write $\mathbf{u} = \operatorname{curl} \psi$, with BCs (10.22a) taking the form

$$\psi = 0, \qquad \Delta \psi = 0. \tag{10.32}$$

The analogous stream function formulation to (10.25) in 2D is then

$$\Delta \dot{\psi} = \operatorname{rot}[\Delta \psi \nabla \psi] - \frac{1}{\mathrm{Re}} \Delta^2 \psi. \tag{10.33}$$

Applying our framework (with $\mathcal{I}_n$ a GL quadrature) gives a scheme equivalent to a Gauss method applied to the following energy- and enstrophy-stable semi-discretisation: find $(\psi, \omega) \in \mathbb{U}^2$ such that

$$(\nabla \dot{\psi}, \nabla \phi) = -(\omega \nabla \psi, \operatorname{curl} \phi) - \frac{1}{\mathrm{Re}} (\nabla \omega, \nabla \phi), \tag{10.34a}$$

$$(\nabla \omega, \nabla \chi) = (\Delta \psi, \Delta \chi), \tag{10.34b}$$

for all $(\phi, \chi) \in \mathbb{U}^2$, where $\mathbb{U}$ is a finite-dimensional function space such that $\phi = 0$ on $\partial\Omega$ for all $\phi \in \mathbb{U}$, and the 2D scalar-to-vector curl is defined $\operatorname{curl}\phi := (\partial_{x_2}\phi, -\partial_{x_1}\phi)$. The stability results for (10.34) hold identically to those of Theorem 10.7, with energy Q_1 and enstrophy Q_3 defined on ψ as

$$Q_1(\psi) := \frac{1}{2}\|\nabla\psi\|^2, \qquad Q_3(\psi) := \frac{1}{2}\|\Delta\psi\|^2; \tag{10.35}$$

in the 2D case however, both of these quantities are conserved in the ideal limit $\mathrm{Re} = \infty$, and dissipated otherwise.

10.4.1 Analysis: existence & uniqueness

Before proceeding to the application of FEEC to simplify our energy- and enstrophy-stable integrators (10.30, 10.34) we discuss certain preliminary existence and uniqueness results, using the results for AD systems from Section 3.3.

We can see this to be a compatible SP discretisation of an AD system (Assumption 3.11) by interpreting the energy Q_1 as the dissipated type-B QoI, and the enstrophy Q_3 as the sole additional type-A QoI; this requires a slight but equivalent rewriting of the dissipative term of (10.30) as: find $(\boldsymbol{\psi}, \boldsymbol{\omega}) \in \mathbb{U}^2$ such that

$$(\operatorname{curl}\dot{\boldsymbol{\psi}}, \operatorname{curl}\boldsymbol{\phi}) = (\operatorname{curl}\boldsymbol{\psi} \times \boldsymbol{\omega}, \operatorname{curl}\boldsymbol{\phi}) - \frac{1}{\mathrm{Re}}(\operatorname{curl}^2\boldsymbol{\psi}, \operatorname{curl}^2\boldsymbol{\phi}), \tag{10.36a}$$

$$(\operatorname{curl}\boldsymbol{\omega}, \operatorname{curl}\boldsymbol{\chi}) = (\operatorname{curl}^2\boldsymbol{\psi}, \operatorname{curl}^2\boldsymbol{\chi}), \tag{10.36b}$$

for all $(\boldsymbol{\phi}, \boldsymbol{\chi}) \in \mathbb{U}^2$; the modification to the 2D equations (10.34) is similar. Since all operators in this scheme are smooth, most required regularity results (i.e. those in Assumption 3.11, Theorem 3.18 and Assumption 3.22) hold immediately. The only result that requires some new analysis is that the dissipative term defines an inner product on $\mathbb{U}$; equivalently, we require that a norm is defined by either $\boldsymbol{\psi} \mapsto \|\operatorname{curl}^2\boldsymbol{\psi}\|$ in the 3D case, or $\psi \mapsto \|\Delta\psi\|$ in 2D.

For the former, 3D case, this can be shown by assuming some compatibility between $\mathbb{V}$ and $\mathbb{Q}$. Namely, we require $\mathbb{U}$ to be defined similarly to (10.18),

$$\mathbb{U} := \{\mathbf{u} \in \mathbb{V} : -(\mathbf{u}, \nabla q) = 0 \text{ for all } q \in \mathbb{Q}\}, \tag{10.37}$$

and that $\mathbb{V}$, $\mathbb{Q}$ exist as part of a subcomplex of the Stokes complex (10.7a), exact over contractible domains (Lemma 10.1):

$$\begin{array}{ccccccc} H_0^1 & \xrightarrow{\text{grad}} & \mathbf{H}(\text{curl}^2) \cap \mathbf{H}_0(\text{curl}) & \xrightarrow{\text{curl}} & \mathbf{H}^1 \cap \mathbf{H}_0(\text{div}) & \xrightarrow{\text{div}} & L^2 \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ \mathbb{Q} & \xrightarrow{\text{grad}} & \mathbb{V} & \xrightarrow{\text{curl}} & \overline{\mathbb{V}} & \xrightarrow{\text{div}} & \overline{\mathbb{Q}} \end{array} . \tag{10.38}$$

In such a case, define the divergence-free subspace $\overline{\mathbb{U}} \subset \overline{\mathbb{V}}$ similarly to (3.7a),

$$\overline{\mathbb{U}} \coloneqq \{\mathbf{u} \in \overline{\mathbb{V}} : (\operatorname{div} \mathbf{u}, q) = 0 \text{ for all } q \in \overline{\mathbb{Q}}\}. \tag{10.39}$$

Observing that $\overline{\mathbb{U}} = \mathcal{N} \operatorname{div} \subset \overline{\mathbb{V}}$, while $\mathbb{U} = \mathcal{N} \operatorname{grad}^* \subset \mathbb{V}$, Lemma 10.4 implies $\operatorname{curl} : \mathbb{U} \to \overline{\mathbb{U}}$ defines an isomorphism. Writing $\mathbf{u} = \operatorname{curl} \boldsymbol{\psi}$, it is then sufficient to show that $\mathbf{u} \mapsto \|\operatorname{curl} \mathbf{u}\|$ defines a norm on $\overline{\mathbb{U}}$, a result that holds over sufficiently regular domains by the (generalised) Gaffney inequality (see He, Hu & Xu [HHX19]).

The latter, 2D case is immediate. If $\Delta\psi = 0$ on Ω, with the Dirichlet BC $\psi = 0$ on $\partial\Omega$, solving the associated Laplace equation asserts that $\psi = 0$ on Ω.

For existence we refer to Theorem 3.18.

Example (Incompressible NS: energy- and enstrophy-stable case)

If $d = 3$, assume that $\mathbb{V}$, $\mathbb{Q}$ form part of a discrete Stokes complex (10.38) and that we define $\mathbb{U}$ as in (10.37). Then solutions to our proposed energy- and enstrophy-stable integrators for the NS equations (10.30, 10.34) exist on arbitrary timesteps Δt_n in either the viscous ($\mathrm{Re} < \infty$) or lowest-order-in-time ($S = 1$) case.

For uniqueness we refer to Theorem 3.26.

Example (Incompressible NS: energy- and enstrophy-stable case)

If $d = 3$, assume again that $\mathbb{V}$, $\mathbb{Q}$ form part of a discrete Stokes complex (10.38) and that we define $\mathbb{U}$ as in (10.37). Then the integrators (10.30, 10.34) are well-posed with a unique solution for either sufficiently small Re, or in the

lowest-order-in-time case ($S = 1$) with sufficiently small Δt_n.

10.4.2 Application of FEEC

One immediate observation about the schemes (10.30, 10.34) is that all terms except the vorticity feature only through their gradient. In particular in the case of the stream function $\boldsymbol{\psi}$, this gradient is its curl, i.e. the velocity $\mathbf{u} = \operatorname{curl} \boldsymbol{\psi}$; this represents a far more typical field over which to pose the NS equations. We may therefore consider those circumstances under which we may equivalently reparametrise (10.30, 10.34) in the velocity $\mathbf{u}$ (and vorticity $\boldsymbol{\omega}$). To do so, we require that there exists a FE parametrisation of $\mathbf{u} = \operatorname{curl} \boldsymbol{\psi}$ (or $\operatorname{curl} \psi$ in the 2D case).

3D In the 3D case, we require the same FE compatibility conditions between $\mathbb{V}$, $\mathbb{Q}$ as proposed in the analysis above, namely that we define $\mathbb{U}$ as in (10.37) and suppose $\mathbb{V}$, $\mathbb{Q}$ come from a discrete Stokes complex (10.38). As shown in the analysis, by Lemma 10.4, $\operatorname{curl} : \mathbb{U} \to \overline{\mathbb{U}}$ defines an isomorphism, where $\overline{\mathbb{U}}$ is defined as in (10.39).

In the case where $\mathbb{V}$, $\mathbb{Q}$ come from such a complex, and $\mathbb{U}$ is defined as in (10.37), the semi-discretisation (10.30) may then be equivalently written as follows: find $(\mathbf{u}, \boldsymbol{\omega}) \in \overline{\mathbb{U}} \times \mathbb{U}$ such that

$$(\dot{\mathbf{u}}, \mathbf{v}) = (\mathbf{u} \times \boldsymbol{\omega}, \mathbf{v}) - \frac{1}{\mathrm{Re}}(\operatorname{curl} \boldsymbol{\omega}, \mathbf{v}), \tag{10.40a}$$

$$(\operatorname{curl} \boldsymbol{\omega}, \operatorname{curl} \boldsymbol{\chi}) = (\operatorname{curl} \mathbf{u}, \operatorname{curl}^2 \boldsymbol{\chi}), \tag{10.40b}$$

for all $(\mathbf{v}, \boldsymbol{\chi}) \in \overline{\mathbb{U}} \times \mathbb{U}$. Energy and enstrophy stability may be similarly shown by testing against $(\mathbf{v}, \boldsymbol{\chi}) = (\mathbf{u}, \boldsymbol{\psi})$ and $(\mathbf{v}, \boldsymbol{\chi}) = (\operatorname{curl} \boldsymbol{\omega}, \dot{\boldsymbol{\psi}})$ respectively, where $\boldsymbol{\psi} \in \mathbb{U}$ is the stream function as above such that $\operatorname{curl} \boldsymbol{\psi} = \mathbf{u}$. Similar to (3.32), we may define the scheme (10.40) in a more familiar way amenable to implementation by extracting the LMs contained in $\overline{\mathbb{U}}$, $\mathbb{U}$: find $(\mathbf{u}, p, \boldsymbol{\omega}, \theta) \in \overline{\mathbb{V}} \times \overline{\mathbb{Q}} \times \mathbb{V} \times \mathbb{Q}$ such that

$$(\dot{\mathbf{u}}, \mathbf{v}) = (\mathbf{u} \times \boldsymbol{\omega}, \mathbf{v}) + (p, \operatorname{div} \mathbf{v}) - \frac{1}{\mathrm{Re}}(\operatorname{curl} \boldsymbol{\omega}, \mathbf{v}), \tag{10.41a}$$

$$0 = (\operatorname{div} \mathbf{u}, q), \tag{10.41b}$$

$$(\operatorname{curl} \boldsymbol{\omega}, \operatorname{curl} \boldsymbol{\chi}) = (\operatorname{curl} \mathbf{u}, \operatorname{curl}^2 \boldsymbol{\chi}) + (\nabla \theta, \boldsymbol{\chi}), \tag{10.41c}$$

$$0 = (\boldsymbol{\omega}, \nabla \eta), \tag{10.41d}$$

for all $(\mathbf{v}, q, \boldsymbol{\chi}, \eta) \in \overline{\mathbb{V}} \times \overline{\mathbb{Q}} \times \mathbb{V} \times \mathbb{Q}$. Energy and enstrophy stability can be seen by testing against $(\mathbf{v}, q, \boldsymbol{\chi}) = (\mathbf{u}, p, \boldsymbol{\psi})$ and $(\mathbf{v}, \boldsymbol{\chi}) = (\operatorname{curl} \boldsymbol{\omega}, \dot{\boldsymbol{\psi}})$ respectively.

2D In 2D, suppose $\mathbb{U}$ comes from a discrete Stokes complex (10.7b),

$$\begin{array}{ccccc} H^2 \cap H^1_0 & \xrightarrow{\operatorname{curl}} & \mathbf{H}^1 \cap \mathbf{H}_0(\operatorname{div}) & \xrightarrow{\operatorname{div}} & L^2 \\ \downarrow & & \downarrow & & \downarrow \\ \mathbb{U} & \xrightarrow{\operatorname{curl}} & \overline{\mathbb{V}} & \xrightarrow{\operatorname{div}} & \overline{\mathbb{Q}} \end{array}, \tag{10.42}$$

similarly exact over contractible domains (Lemma 10.1). Defining the $\overline{\mathbb{Q}}$-discretely divergence-free subspace $\overline{\mathbb{U}} \subset \overline{\mathbb{V}}$ as in (10.39), Lemma 10.4 similarly implies $\operatorname{curl} : \mathbb{U} \to \overline{\mathbb{U}}$ defines an isomorphism. In such a case, the semi-discretisation (10.34) may be equivalently written as follows: find $(\mathbf{u}, \omega) \in \overline{\mathbb{U}} \times \mathbb{U}$ such that

$$(\dot{\mathbf{u}}, \mathbf{v}) = (\omega \mathbf{k} \times \mathbf{u}, \mathbf{v}) - \frac{1}{\mathrm{Re}} (\operatorname{curl} \omega, \mathbf{v}), \tag{10.43a}$$

$$(\nabla \omega, \nabla \chi) = -(\operatorname{rot} \mathbf{u}, \Delta \chi), \tag{10.43b}$$

for all $(\mathbf{v}, \chi) \in \overline{\mathbb{U}} \times \mathbb{U}$, and $\mathbf{k}$ represents the unit normal to the plane, oriented such that $\operatorname{curl} \phi = \mathbf{k} \times \nabla \phi$. Energy and enstrophy stability may be shown by testing against $(\mathbf{v}, \chi) = (\mathbf{u}, \psi)$ and $(\mathbf{v}, \chi) = (\operatorname{curl} \omega, \dot{\psi})$ respectively. We may then extract the LM contained in $\overline{\mathbb{U}}$ with the following: find $(\mathbf{u}, p, \omega) \in \overline{\mathbb{V}} \times \overline{\mathbb{Q}} \times \mathbb{U}$ such that

$$(\dot{\mathbf{u}}, \mathbf{v}) = (\omega \mathbf{k} \times \mathbf{u}, \mathbf{v}) + (p, \operatorname{div} \mathbf{v}) - \frac{1}{\mathrm{Re}} (\operatorname{curl} \omega, \mathbf{v}), \tag{10.44a}$$

$$0 = (\operatorname{div} \mathbf{u}, q), \tag{10.44b}$$

$$(\nabla \omega, \nabla \chi) = -(\operatorname{rot} \mathbf{u}, \Delta \chi), \tag{10.44c}$$

for all $(\mathbf{v}, q, \chi) \in \overline{\mathbb{V}} \times \overline{\mathbb{Q}} \times \mathbb{U}$. Energy and enstrophy stability may be shown by testing against $(\mathbf{v}, q, \chi) = (\mathbf{u}, p, \psi)$ and $(\mathbf{v}, \chi) = (\operatorname{curl} \omega, \dot{\psi})$ respectively.

Remark 10.8 (Application of analysis after reparametrisation)**.** *Note, provided all the discrete Sokes complex criteria (10.38, 10.42) hold, this reparametrisation is equivalent, i.e. the velocity–vorticity schemes (10.41, 10.44) are equivalent to the original stream function–vorticity schemes (10.30, 10.34). Consequently, both the existence and uniqueness results from Subsection 10.4.1 hold after reparametrisation.*

Again, in either case (10.41, 10.44) we require a Gauss method for the time integration to ensure energy and enstrophy stability.

10.4.3 Usage without implementation of discrete Stokes complexes: interior penalty methods & a conforming workaround

As established, the integrators (10.41, 10.44) are necessarily both energy- and enstrophy-stable, satisfying our SP requirements. However, for numerical implementation they require discrete de Rham complexes of enhanced regularity, i.e. they require discrete Stokes complexes (10.38, 10.42). While some such discrete Stokes complexes do exist in the FEEC literature (see the literature review in Section 8.1) they are uncommon, and implementations remain even rarer.

In this subsection therefore, we discuss two practical options for the implementation of these schemes (10.41, 10.44): non-conforming interior penalty methods (IPMs) that rely on (reduced regularity) discrete de Rham complexes (10.5) only, and a technique for implementing equivalent schemes without requiring the implementation of either of the high-regularity spaces $\mathbb{V}$ or $\mathbb{Q}$.

10.4.3.1 Interior penalty methods with reduced regularity

Our first approach is the use of IPMs for non-conforming discretisations.

3D Beginning with the 3D case, suppose we do not have computational access to a discrete Stokes complex (10.38), but we do to a discrete de Rham complex (10.5a) of reduced regularity,

$$\begin{array}{ccccccc} H^1_0 & \xrightarrow{\text{grad}} & \mathbf{H}_0(\text{curl}) & \xrightarrow{\text{curl}} & \mathbf{H}_0(\text{div}) & \xrightarrow{\text{div}} & L^2 \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ \mathbb{Q} & \xrightarrow{\text{grad}} & \mathbb{V} & \xrightarrow{\text{curl}} & \overline{\mathbb{V}} & \xrightarrow{\text{div}} & \overline{\mathbb{Q}} \end{array} . \tag{10.45}$$

We must first introduce some notation. Letting $\mathcal{K}$ denote a mesh, i.e. a set of cells $K \in \mathcal{K}$, over Ω, let $\mathcal{F}$ denote the set of its facets $F \in \mathcal{F}$. On $\mathcal{F}$, let $\mathbf{v}_\pm$ indicate the value of some vector-field $\mathbf{v}$, discontinuous across $\mathcal{F}$, on either side of the facet; let $\mathbf{n}_\pm$ similarly denote the outward-pointing normals in either cell (such that $\mathbf{n}_+ = -\mathbf{n}_-$). Define the jumps $[\![\mathbf{v} \times \mathbf{n}]\!] \coloneqq \mathbf{v}_+ \times \mathbf{n}_+ + \mathbf{v}_- \times \mathbf{n}_-$ and $[\![\mathbf{v}]\!]^* \coloneqq \mathbf{v}_+ - \mathbf{v}_-$. Define the mean $\{\!\{\text{curl}\,\mathbf{v}\}\!\} \coloneqq \frac{1}{2}(\text{curl}\,\mathbf{v}_+ + \text{curl}\,\mathbf{v}_-)$. Lastly, let $h_F \coloneqq \frac{1}{2}(h_+ + h_-)$ denote the mean mesh size on a facet $F \in \mathcal{F}$, where $h_\pm$ denote the size of the cells on either side of F.

With this notation, define $\mathcal{C}^*[\cdot,\cdot]$, a broken $\mathbf{H}(\mathrm{curl})$ bilinear form on $\overline{\mathbb{V}}$,

$$\begin{aligned}\mathcal{C}[\mathbf{u},\mathbf{v}] \coloneqq &\sum_{K\in\mathcal{K}}\int_K \operatorname{curl}\mathbf{u}\cdot\operatorname{curl}\mathbf{v}\\ &+\sum_{F\in\mathcal{F}}\int_F\Big[\frac{\sigma}{h_F}[\![\mathbf{u}]\!]^*\cdot[\![\mathbf{v}]\!]^* - [\![\mathbf{u}\times\mathbf{n}]\!]\cdot\{\!\{\operatorname{curl}\mathbf{v}\}\!\} - \{\!\{\operatorname{curl}\mathbf{u}\}\!\}\cdot[\![\mathbf{v}\times\mathbf{n}]\!]\Big],\end{aligned} \tag{10.46}$$

where $\sigma > 0$ is a sufficiently large interior penalty (IP) parameter (see Ern & Guermond [EG21a, Chap. 18] and [EG21b, Chap. 38]). We propose the following IP semi-discretisation: find $(\mathbf{u},p,\boldsymbol{\omega},\theta)\in\overline{\mathbb{V}}\times\overline{\mathbb{Q}}\times\mathbb{V}\times\mathbb{Q}$ such that

$$(\dot{\mathbf{u}},\mathbf{v}) = (\mathbf{u}\times\boldsymbol{\omega},\mathbf{v}) + (p,\operatorname{div}\mathbf{v}) - \frac{1}{\mathrm{Re}}(\operatorname{curl}\boldsymbol{\omega},\mathbf{v}), \tag{10.47a}$$

$$0 = (\operatorname{div}\mathbf{u},q), \tag{10.47b}$$

$$(\operatorname{curl}\boldsymbol{\omega},\operatorname{curl}\boldsymbol{\chi}) = \mathcal{C}^*[\mathbf{u},\operatorname{curl}\boldsymbol{\chi}] + (\nabla\theta,\boldsymbol{\chi}), \tag{10.47c}$$

$$0 = (\boldsymbol{\omega},\nabla\eta), \tag{10.47d}$$

for all $(\mathbf{v},q,\boldsymbol{\chi},\eta)\in\overline{\mathbb{V}}\times\overline{\mathbb{Q}}\times\mathbb{V}\times\mathbb{Q}$. Defining the broken enstrophy Q_3^*,

$$Q_3^*(\mathbf{u}) \coloneqq \mathcal{C}^*[\mathbf{u},\mathbf{u}] \geq 0, \tag{10.48}$$

we see by similar arguments that, assuming the de Rham complex criteria (10.45) holds, (10.47) satisfies the semidiscrete structures

$$\dot{Q}_1 = -\frac{1}{\mathrm{Re}}\mathcal{C}^*[\mathbf{u},\mathbf{u}] \leq 0, \qquad \dot{Q}_3^* = -\int_\Omega \mathbf{u}\cdot(\boldsymbol{\omega}\cdot\nabla\boldsymbol{\omega}) - \frac{1}{\mathrm{Re}}\|\operatorname{curl}\boldsymbol{\omega}\|^2, \tag{10.49}$$

with the fully discrete analogues holding in time when using a Gauss method for the time discretisation.

2D In the 2D case, we suppose again we have access only to discrete de Rham complex (10.2a) of reduced regularity,

$$\begin{array}{ccccc} H_0^1 & \xrightarrow{\operatorname{curl}} & \mathbf{H}_0(\operatorname{div}) & \xrightarrow{\operatorname{div}} & L^2 \\ \downarrow & & \downarrow & & \downarrow \\ \mathbb{U} & \xrightarrow{\operatorname{curl}} & \overline{\mathbb{V}} & \xrightarrow{\operatorname{div}} & \overline{\mathbb{Q}} \end{array}. \tag{10.50}$$

We must first extend the notation from above. Letting $\mathbf{t}$ denote the unit clockwise-facing tangential vector on $\partial\Omega$, define the jump $[\![\mathbf{v}\cdot\mathbf{t}]\!] \coloneqq \mathbf{v}_+\cdot\mathbf{t}_+ + \mathbf{v}_-\cdot\mathbf{t}_-$ over $\mathcal{F}$. Define the mean $\{\!\{\operatorname{rot}\mathbf{v}\}\!\} \coloneqq \frac{1}{2}(\operatorname{rot}\mathbf{v}_+ + \operatorname{rot}\mathbf{v}_-)$ similarly.

We then redefine $\mathcal{C}^*[\cdot,\cdot]$ in the 2D case to be a broken $\mathbf{H}(\mathrm{rot})$ bilinear form on $\overline{\mathbb{V}}$,

$$\begin{aligned}\mathcal{C}^*[\mathbf{u},\mathbf{v}] \coloneqq &\sum_{K\in\mathcal{K}}\int_K \mathrm{rot}\,\mathbf{u}\,\mathrm{rot}\,\mathbf{v}\\ &+\sum_{F\in\mathcal{F}}\int_F\Big[\frac{\sigma}{h_F}[\![\mathbf{u}]\!]^*\cdot[\![\mathbf{v}]\!]^* - [\![\mathbf{u}\cdot\mathbf{t}]\!]\{\!\{\mathrm{rot}\,\mathbf{v}\}\!\} - \{\!\{\mathrm{rot}\,\mathbf{u}\}\!\}[\![\mathbf{v}\cdot\mathbf{t}]\!]\Big].\end{aligned} \tag{10.51}$$

We propose then the following IP semi-discretisation: find $(\mathbf{u},p,\omega)\in\overline{\mathbb{V}}\times\overline{\mathbb{Q}}\times\mathbb{U}$ such that

$$(\dot{\mathbf{u}},\mathbf{v}) = (\omega\mathbf{k}\times\mathbf{u},\mathbf{v}) + (p,\mathrm{div}\,\mathbf{v}) - \frac{1}{\mathrm{Re}}(\mathrm{curl}\,\omega,\mathbf{v}), \tag{10.52a}$$

$$0 = (\mathrm{div}\,\mathbf{u},q), \tag{10.52b}$$

$$(\nabla\omega,\nabla\chi) = \mathcal{C}^*[\mathbf{u},\mathrm{curl}\,\chi], \tag{10.52c}$$

for all $(\mathbf{v},q,\chi)\in\overline{\mathbb{V}}\times\overline{\mathbb{Q}}\times\mathbb{U}$. We see then by similar arguments that, assuming the de Rham complex criteria (10.50) holds, (10.52) satisfies the semidiscrete structures (10.49) with the broken enstrophy defined as in (10.48); the fully discrete analogues then hold in time when using a Gauss method for the time discretisation.

10.4.3.2 Workaround for implementation with enhanced regularity

Looking in the 3D case (10.41) (with the argument in 2D (10.44) being similar), assuming one could compute the inverse $\mathrm{curl}^{-1}:\overline{\mathbb{U}}\to\mathbb{U}$ of $\mathrm{curl}:\mathbb{U}\to\overline{\mathbb{U}}$ exactly, the scheme (10.40) could be written as one entirely in $\overline{\mathbb{U}}$: find $(\mathbf{u},\boldsymbol{\alpha})\in\overline{\mathbb{U}}^2$ such that

$$(\dot{\mathbf{u}},\mathbf{v}) = (\mathbf{u}\times\mathrm{curl}^{-1}\boldsymbol{\alpha},\mathbf{v}) - \frac{1}{\mathrm{Re}}(\boldsymbol{\alpha},\mathbf{v}), \tag{10.53a}$$

$$(\boldsymbol{\alpha},\boldsymbol{\beta}) = (\mathrm{curl}\,\mathbf{u},\mathrm{curl}\,\boldsymbol{\beta}), \tag{10.53b}$$

for all $(\mathbf{v},\boldsymbol{\beta})\in\overline{\mathbb{U}}^2$. This is simply a re-writing of $(\mathrm{curl}\,\boldsymbol{\omega},\mathrm{curl}\,\boldsymbol{\chi})\mapsto(\boldsymbol{\alpha},\boldsymbol{\beta})$. Introducing LMs for numerical implementation would then give a problem over $\overline{\mathbb{V}}$ and $\overline{\mathbb{Q}}$ only.

While full 3D FE Stokes complexes (10.38) are often not implemented numerically, spaces $\overline{\mathbb{V}}$ and $\overline{\mathbb{Q}}$ (i.e. such that $\mathrm{div}\,\overline{\mathbb{V}}=\overline{\mathbb{Q}}$) often are, in particular the SV [SV85a; SV85b] pair: $\overline{\mathbb{V}}=[\mathbb{CG}_{p+1}]^3$, the CG (or Lagrange) space of degree $p+1$, and $\overline{\mathbb{Q}}=\mathbb{DG}_p$, the DG space of one lower degree p over a simplicial mesh (see Ern & Guermond [EG21a, Sec. 6 & 7]).[8] This offers immediate appeal: if, for any $\boldsymbol{\alpha}\in\overline{\mathbb{U}}$, one were able to find some discretely divergence-free $\boldsymbol{\omega}$ such that $\mathrm{curl}\,\boldsymbol{\omega}=\boldsymbol{\alpha}$, and this map

[8] Note, to ensure $\mathrm{div}\,\overline{\mathbb{V}}=\overline{\mathbb{Q}}$ and inf-sup stability, we require certain conditions on the mesh and the order p (see Farrell, Mitchell & Scott [FMS24] for a recent review of these conditions).

could be computed efficiently and without the use of any high-regularity spaces, then we could use this map in (10.53) to implement a conforming energy- and enstrophy-stable scheme, even in cases when spaces of more refined regularity than $\mathbf{H}^1$ are unavailable.

3D To demonstrate that this can often be done in 3D, suppose that the FE software in question does feature a discrete 3D de Rham complex (as in (10.45)) of standard reduced regularity,

$$\begin{array}{ccccccc} H^1_0 & \xrightarrow{\text{grad}} & \mathbf{H}_0(\text{curl}) & \xrightarrow{\text{curl}} & \mathbf{H}_0(\text{div}) & \xrightarrow{\text{div}} & L^2 \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ \hat{\mathbb{Q}} & \xrightarrow{\text{grad}} & \hat{\mathbb{V}} & \xrightarrow{\text{curl}} & \hat{\overline{\mathbb{V}}} & \xrightarrow{\text{div}} & \hat{\overline{\mathbb{Q}}} \end{array} . \tag{10.54}$$

Define then $\hat{\mathbb{U}}$ and $\hat{\overline{\mathbb{U}}}$ similarly to (10.37) and (10.39) respectively. By Lemma 10.4, $\text{curl} : \hat{\mathbb{U}} \to \hat{\overline{\mathbb{U}}}$ defines an isomorphism, which can furthermore be inverted numerically as follows: for $\boldsymbol{\alpha} \in \hat{\overline{\mathbb{U}}}$, find $\boldsymbol{\omega} \in \hat{\mathbb{U}}$ such that

$$(\text{curl}\,\boldsymbol{\omega}, \text{curl}\,\boldsymbol{\chi}) = (\boldsymbol{\alpha}, \text{curl}\,\boldsymbol{\chi}) \tag{10.55}$$

for all $\boldsymbol{\chi} \in \hat{\mathbb{U}}$; it is straightforward to confirm this is well-posed, and that $\text{curl}\,\boldsymbol{\omega} = \boldsymbol{\alpha}$. In fact, if we further assume the inclusion $\overline{\mathbb{V}} \subset \hat{\overline{\mathbb{V}}}$, we see $\overline{\mathbb{U}} \subset \hat{\overline{\mathbb{U}}}$,[9] implying that simply by restricting our attention to $\boldsymbol{\alpha} \in \overline{\mathbb{U}} \subset \hat{\overline{\mathbb{U}}}$ the above variational map (10.55) defines a right-inverse of curl on $\overline{\mathbb{U}}$ with a discretely divergence-free image, as required.

Combining (10.53, 10.55) then and eliminating LMs, our final scheme is as follows: find $((\mathbf{u}, \boldsymbol{\alpha}), (p, r), \boldsymbol{\omega}, \theta) \in \overline{\mathbb{V}}^2 \times \overline{\mathbb{Q}}^2 \times \hat{\mathbb{V}} \times \hat{\mathbb{Q}}$ such that

$$(\dot{\mathbf{u}}, \mathbf{v}) = (\mathbf{u} \times \boldsymbol{\omega}, \mathbf{v}) + (p, \text{div}\,\mathbf{v}) - \frac{1}{\text{Re}}(\boldsymbol{\alpha}, \mathbf{v}), \tag{10.56a}$$

$$0 = (\text{div}\,\mathbf{u}, q), \tag{10.56b}$$

$$(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\text{curl}\,\mathbf{u}, \text{curl}\,\boldsymbol{\beta}) + (r, \text{div}\,\boldsymbol{\beta}), \tag{10.56c}$$

$$0 = (\text{div}\,\boldsymbol{\alpha}, s), \tag{10.56d}$$

$$(\text{curl}\,\boldsymbol{\omega}, \text{curl}\,\boldsymbol{\chi}) = (\boldsymbol{\alpha}, \text{curl}\,\boldsymbol{\chi}) - (\nabla\theta, \boldsymbol{\chi}), \tag{10.56e}$$

$$0 = -(\boldsymbol{\omega}, \nabla\eta) \tag{10.56f}$$

[9] We may see this by noting $\hat{\overline{\mathbb{U}}}$ is the exactly divergence-free subspace of $\hat{\mathbb{V}}$, and $\overline{\mathbb{U}}$ the exactly divergence-free subspace of $\overline{\mathbb{V}}$.

for all $((\mathbf{v},\boldsymbol{\beta}),(q,s),\boldsymbol{\chi},\eta) \in \overline{\mathbb{V}}^2 \times \overline{\mathbb{Q}}^2 \times \hat{\mathbb{V}} \times \hat{\mathbb{Q}}$. One may then see energy stability by taking $(\mathbf{v},\boldsymbol{\beta},q) = (\mathbf{u},\mathbf{u},p-\frac{1}{\mathrm{Re}}r)$ and enstrophy stability by taking $(\mathbf{v},\boldsymbol{\beta},s) = (\boldsymbol{\alpha},\dot{\mathbf{u}},p)$.

An example set of compatible FE spaces for this scheme are

$$\hat{\mathbb{Q}} = \mathbb{CG}_{p+2}, \qquad \hat{\mathbb{V}} = \mathbb{Ned}^{\mathrm{curl}}_{\mathrm{p+2}}, \qquad \overline{\mathbb{V}} = [\mathbb{CG}_{p+1}]^3, \qquad \overline{\mathbb{Q}} = \mathbb{DG}_p, \tag{10.57}$$

with $\mathbb{Ned}^{\mathrm{curl}}_{\mathrm{p+2}}$ being the degree-$(p+2)$ curl-conforming Nédélec [Néd86] element of the first kind.

The new scheme (10.56) is equivalent to solving the original scheme (10.41) over the discrete Stokes complex (10.38)

$$\begin{array}{ccccccc}
H^1_0 & \xrightarrow{\mathrm{grad}} & \mathbf{H}(\mathrm{curl}^2)\cap\mathbf{H}_0(\mathrm{curl}) & \xrightarrow{\mathrm{curl}} & \mathbf{H}^1\cap\mathbf{H}_0(\mathrm{div}) & \xrightarrow{\mathrm{div}} & L^2 \\
\downarrow & & \downarrow & & \downarrow & & \downarrow \\
\hat{\mathbb{Q}} & \xrightarrow{\mathrm{grad}} & \mathrm{grad}\,\hat{\mathbb{Q}}\oplus\mathrm{curl}^{-1}\,\overline{\mathbb{U}} & \xrightarrow{\mathrm{curl}} & \overline{\mathbb{V}} & \xrightarrow{\mathrm{div}} & \overline{\mathbb{Q}}
\end{array}, \tag{10.58}$$

where curl^{-1} maps to the discretely divergence-free space $\hat{\mathbb{U}}$. While we are implicitly using the finite-dimensional function space $\mathrm{grad}\,\hat{\mathbb{Q}}\oplus\mathrm{curl}^{-1}\,\overline{\mathbb{U}}$, we have no guarantee it is a FE space in the traditional sense, even if $\overline{\mathbb{V}}$ and $\overline{\mathbb{Q}}$ are FE spaces, as we can not guarantee it has a local basis. Regardless, this observation ensures that both the existence and uniqueness results from Section 10.4.1 hold for this scheme (10.56).

2D The same ideas in the construction of the conforming workaround (10.56) may be applied also to the 2D scheme (10.44) albeit with fewer introduced AVs. We require computational access to a discrete 2D de Rham complex (as in (10.45)) again of standard reduced regularity,

$$\begin{array}{ccccc}
H^1_0 & \xrightarrow{\mathrm{curl}} & \mathbf{H}_0(\mathrm{div}) & \xrightarrow{\mathrm{div}} & L^2 \\
\downarrow & & \downarrow & & \downarrow \\
\hat{\mathbb{U}} & \xrightarrow{\mathrm{curl}} & \hat{\overline{\mathbb{V}}} & \xrightarrow{\mathrm{div}} & \overline{\mathbb{Q}}
\end{array}, \tag{10.59}$$

such that $\overline{\mathbb{V}} \subset \hat{\overline{\mathbb{V}}}$. Our scheme is then as follows: find $((\mathbf{u}, \boldsymbol{\alpha}), (p, r), \omega) \in \overline{\mathbb{V}}^2 \times \overline{\mathbb{Q}}^2 \times \hat{\mathbb{U}}$ such that

$$(\dot{\mathbf{u}}, \mathbf{v}) = (\omega \mathbf{k} \times \mathbf{u}, \mathbf{v}) + (p, \operatorname{div} \mathbf{v}) - \frac{1}{\mathrm{Re}}(\boldsymbol{\alpha}, \mathbf{v}), \tag{10.60a}$$

$$0 = (\operatorname{div} \mathbf{u}, q), \tag{10.60b}$$

$$(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\operatorname{rot} \mathbf{u}, \operatorname{rot} \boldsymbol{\beta}) + (r, \operatorname{div} \boldsymbol{\beta}), \tag{10.60c}$$

$$0 = (\operatorname{div} \boldsymbol{\alpha}, s), \tag{10.60d}$$

$$(\operatorname{curl} \omega, \operatorname{curl} \chi) = (\boldsymbol{\alpha}, \operatorname{curl} \chi), \tag{10.60e}$$

for all $((\mathbf{v}, \boldsymbol{\beta}), (q, s), \chi) \in \overline{\mathbb{V}}^2 \times \overline{\mathbb{Q}}^2 \times \hat{\mathbb{U}}$. It is again a simple exercise to confirm that (10.60e) ensures $\operatorname{curl} w = \boldsymbol{\alpha}$ exactly.

Natural FE spaces here are

$$\hat{\mathbb{U}} = \mathbb{C}\mathbb{G}_{p+2}, \qquad \overline{\mathbb{V}} = [\mathbb{C}\mathbb{G}_{p+1}]^2, \qquad \overline{\mathbb{Q}} = \mathbb{D}\mathbb{G}_p, \tag{10.61}$$

with similar degree p and mesh structure constraints for the SV pair $(\overline{\mathbb{V}}, \overline{\mathbb{Q}}) = ([\mathbb{C}\mathbb{G}_{p+1}]^2, \mathbb{D}\mathbb{G}_p)$.

The new scheme (10.60) is equivalent to solving the original scheme (10.44) implicitly over the discrete Stokes complex

$$\begin{array}{ccccc} H^2 \cap H_0^1 & \xrightarrow{\operatorname{curl}} & \mathbf{H}^1 \cap \mathbf{H}_0(\operatorname{div}) & \xrightarrow{\operatorname{div}} & L^2 \\ \downarrow & & \downarrow & & \downarrow \\ \operatorname{curl}^{-1} \overline{\mathbb{U}} & \xrightarrow{\operatorname{curl}} & \overline{\mathbb{V}} & \xrightarrow{\operatorname{div}} & \overline{\mathbb{Q}} \end{array} . \tag{10.62}$$

For the spaces proposed above (10.61), the space $\operatorname{curl}^{-1} \overline{\mathbb{U}} \subset H^2 \cap H_0^1$ is the MS [MS75] space.

10.4.4 2D vortex test

To demonstrate and motivate our scheme's SP properties numerically, we conclude by considering the numerical behaviour of a vortex in a 2D box $\Omega = (0, 1)^2$ at $\mathrm{Re} = \infty$ under our enstrophy-stable scheme.

To set up the ICs, define the Weierstrass elliptic function $\wp : \mathbb{C} \to \mathbb{C}$ over the complex plane $\mathbb{C}$ (see Fig. 10.1a)

$$\wp(z) \coloneqq \frac{1}{z^2} + \sum_{\substack{(m,n) \in \\ \mathbb{Z}^2 \setminus \{(0,0)\}}} \frac{1}{(z - 2m - 2ni)^2} - \frac{1}{(2m + 2ni)^2}; \tag{10.63}$$

this is analytic and doubly periodic with period 2 in both the real and imaginary axes. Up to projection onto $\mathbb{U}$, the ICs $\psi(0)$ are defined for $\mathbf{x} = (x, y) \in \Omega$ as proportional to

$$\psi(0) \propto \Re\{\log[\wp(x + iy) - \wp(x_0 + iy_0)] - \log[\wp(x + iy) - \wp(x_0 - iy_0)]\}, \tag{10.64}$$

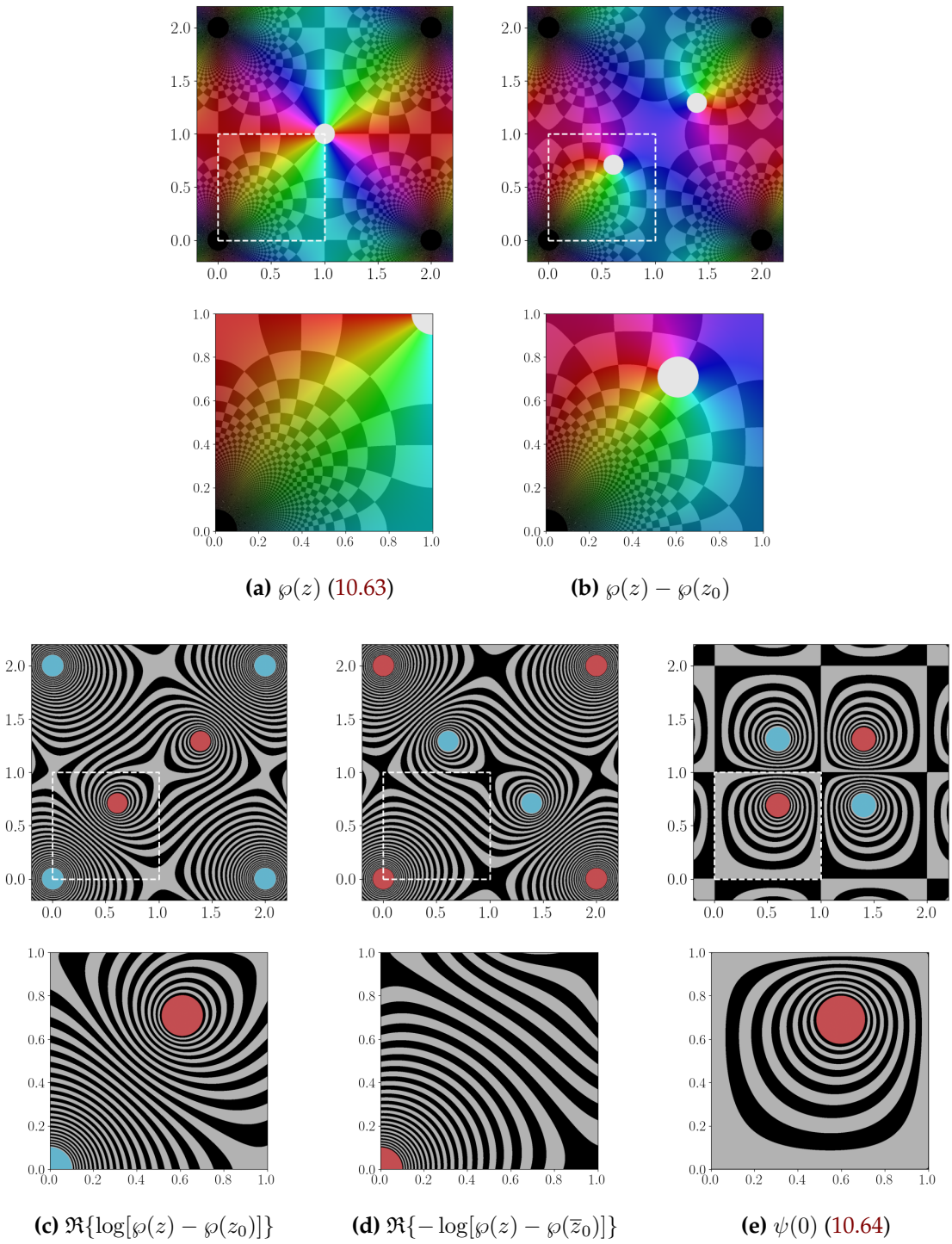
where $\Re : \mathbb{C} \to \mathbb{R}$ denotes the real component (see Fig. 10.1e) and $x_0, y_0 \in (0, 1)$. The initial velocity $\mathbf{u}(0) \in \overline{\mathbb{U}}$ is then defined from $\psi(0)$ by $\mathbf{u}(0) = \operatorname{curl} \psi(0)$, with the constant of proportionality chosen such that $Q_1(\mathbf{u}) = 1$.[10] These ICs are designed to model a vortex initially at $\mathbf{x}_0 = (x_0, y_0) \in \Omega$ while preserving each of the BCs; the no-flux BC $\mathbf{u} \cdot \mathbf{n} = 0$ holds by the symmetries of $\wp$, while the construction of $\psi(0)$ as the real part of an analytic function ensures $\operatorname{rot} \mathbf{u} = -\Delta\psi = 0$ outside the vortex (x_0, y_0), in particular enforcing the BC $\operatorname{rot} \mathbf{u} = 0$ on $\partial\Omega$.

Remark 10.9 (Motivation for construction of 2D vortex)**.** *Fig. 10.1 offers physical intuition for this definition of $\psi(0)$ (10.64). In each subfigure, the upper plot shows a broader view of the function, with $\Omega = [0, 1]^2$ marked by a dashed white square; the lower plot focuses on Ω. We write $z = x + iy$, $z_0 = x_0 + iy_0$:*

- *Figs. 10.1a & 10.1b visualise argument with hue, and modulus by shade with darker values at higher moduli; roots are indicated in white, poles in black, and a general checkerboard illustrates the conformality. In Fig. 10.1b, the subtraction of $\wp(z_0)$ moves the root to z_0.*
- *Figs. 10.1c, 10.1d & 10.1e visualise real functions through contour stripes, with negative poles indicated in red, and positive in cyan; each of these functions is harmonic outside the poles, and can be viewed as an irrotational stream function with clockwise (red) and counterclockwise (cyan) vortices. In Fig. 10.1c, the $\log$ map turns the poles and roots of $\wp(z) - \wp(z_0)$ into vortices. In Fig. 10.1d, we consider a similar function, constructed to have opposing vortices along the grid $(2\mathbb{Z})^2$, but differing poles outside $(2\mathbb{Z})^2$. In Fig. 10.1e, we see that summing these functions returns a stream function satisfying the no-flux BCs, and with a solitary vortex within Ω at z_0.*

In the continuous case, the incompressible NS equations under ICs given by (10.64) have very specific dynamics. In the Euler/inviscid case $\mathrm{Re} = \infty$, the vortex

[10] We note that, in the continuous case, these ICs (10.64) do not have a well-defined enstrophy, i.e. the velocity $\mathbf{u}(0) = \operatorname{curl} \psi(0)$ is not $\mathbf{H}(\operatorname{curl})$-conforming. After projection into $\mathbb{U}$ however, this ceases to be an issue.



(a) $\wp(z)$ (10.63) **(b)** $\wp(z) - \wp(z_0)$

(c) $\Re\{\log[\wp(z) - \wp(z_0)]\}$ **(d)** $\Re\{-\log[\wp(z) - \wp(\overline{z}_0)]\}$ **(e)** $\psi(0)$ (10.64)

Figure 10.1: Step-by-step illustration of the construction of (10.64) with $z_0 = 0.6 + 0.7i$, i.e. $\mathbf{x}_0 = (0.6, 0.7)$.

will retain its shape, orbiting clockwise around the perimeter of the domain Ω. To test the discrete replication of this behaviour, we consider in our numerical tests $\mathrm{Re} = \infty$, guaranteeing the mesh is under-resolved.

Assuming the mesh to be triangular, let $\mathbb{CG}_{p+2}$ and $\mathbb{DG}_p$ denote the continuous and DG spaces of degree $p+2$ and p respectively over this mesh (see Ern & Guermond [EG21a, Sec. 6 & 7]). Defining the degree-$(p+1)$ Brezzi–Douglas–Marini [BDM85] space $\mathbb{BDM}_p$,

$$\mathbb{BDM}_p \coloneqq [\mathbb{DG}_p]^2 \cap \mathbf{H}(\mathrm{div}), \tag{10.65}$$

we note that, for $p \geq 0$, this induces a discrete de Rham complex (10.50),

$$\mathbb{CG}_{p+2} \xrightarrow{\quad\mathrm{curl}\quad} \mathbb{BDM}^{\mathrm{d}}_{p+1} \xrightarrow{\quad\mathrm{div}\quad} \mathbb{DG}_p \,. \tag{10.66}$$

We take therefore $(\mathbb{U}, \overline{\mathbb{V}}, \overline{\mathbb{Q}}) = (\mathbb{CG}_3, \mathbb{BDM}_2, \mathbb{DG}_1)$ (i.e. with $p=1$) observing that this complex is non-conforming ($\mathbb{BDM}_p \not\subset \mathbf{H}(\mathrm{curl})$) and so will necessitate the use of our non-conforming scheme (10.52). Taking the mesh $\mathcal{K}$ to be triangular of uniform width 2^{-5}, we compare 3 different 1-stage integrators: a classical energy-stable IPM with no auxiliary velocity[11] with (uniform) timestep $\Delta t_n = 2^{-10}$;[12] the comparable MEEVC scheme of Hanot [Han23] and Zhang *et al.* [Zha+24] (see our discussion in Section 8.1) with timestep $\Delta t_n = 2^{-8}$; our non-conforming scheme (10.52) with IP parameter $\sigma = 2^5$ and timestep $\Delta t_n = 2^{-8}$. We ensure the vortex is not initially aligned with the mesh by taking $\mathbf{x}_0 = (3-\sqrt{5})(\frac{1}{2}, 1)$ and scale the ICs (10.64) such that the initial energy $Q_1(\mathbf{u}(0)) = 1$.

Fig. 10.2 shows plots of the stream function ψ (i.e. $\psi \in \mathbb{U}$ such that $\mathrm{curl}\,\psi = \mathbf{u}$) at various (exponentially increasing) times in each of the schemes. The vortex in the numerical results from the classical integrator dissipates after relatively few iterations, despite the shorter timestep; the results from the auxiliary enstrophy–stable MEEVC scheme and our enstrophy-stable discretisation (10.52) remain stable until the final time $t = 2^4$, after performing between 2 and 3 circuits of the domain Ω.

[11] In particular, we consider an IM discretisation of the following semi-discretisation: find $(\mathbf{u}, p) \in \overline{\mathbb{V}} \times \overline{\mathbb{Q}}$ such that

$$(\dot{\mathbf{u}}, \mathbf{v}) = \sum_{K\in\mathcal{K}} \int_K (\mathrm{rot}\,\mathbf{u})\mathbf{k}\cdot(\mathbf{u}\times\mathbf{v}) - \sum_{F\in\mathcal{F}} \int_F [\![\mathbf{u}\cdot\mathbf{t}]\!]\mathbf{k}\cdot(\{\!\{\mathbf{u}\}\!\}\times\{\!\{\mathbf{v}\}\!\}) + (p, \mathrm{div}\,\mathbf{v}), \tag{10.67a}$$

$$0 = (\mathrm{div}\,\mathbf{u}, q) \tag{10.67b}$$

for all $(\mathbf{v}, q) \in \overline{\mathbb{V}} \times \overline{\mathbb{Q}}$, where $\{\!\{\mathbf{u}\}\!\} \coloneqq \frac{1}{2}(\mathbf{u}_+ + \mathbf{u}_-)$.

[12] We take the timestep $\Delta t_n = 2^{-10}$ for the classical integrator to be shorter than that of the MEEVC scheme and our discretisation $\Delta t_n = 2^{-8}$, as the discretisation under the classical integrator failed to converge on the longer timestep.

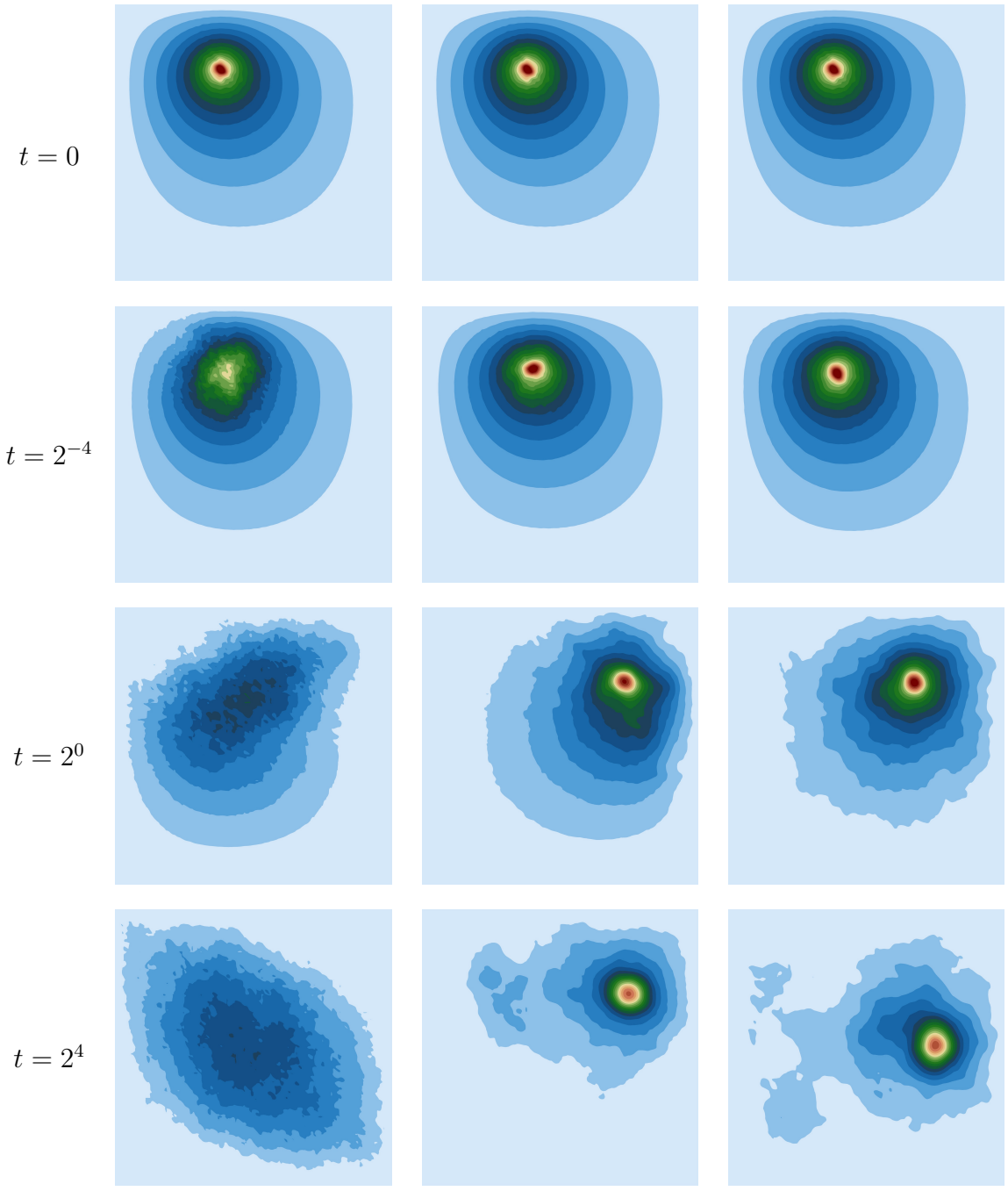
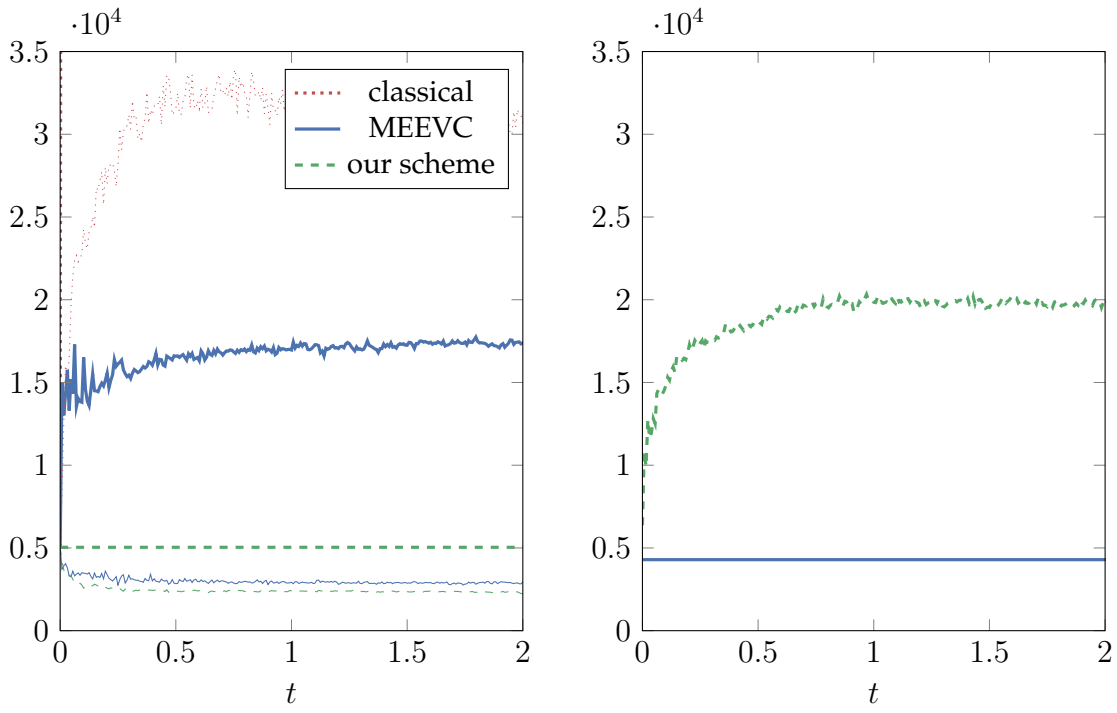


Figure 10.2: Contours, from red at $\psi = -1$ to light blue at $\psi = 0$, in the stream function ψ at times $t \in \{0, 2^{-4}, 2^{0}, 2^{4}\}$ for 3 different 1-stage integrators in the 2D vortex test (Subsection 10.4.4).

Fig. 10.3 shows the evolution of the enstrophy within each of the simulations, up to time $t = 2$; we need not plot the energies for each simulation, as each scheme is energy-stable. Since we are using a non-conforming discretisation, there are various notions of enstrophy available. Fig. 10.3a illustrates both the broken enstrophy $Q_3^*(\mathbf{u})$

(10.48) with $\sigma = 2^5$ (thick upper line) and an internal enstrophy $\frac{1}{2}\sum_{K\in\mathcal{K}}\int_K(\operatorname{rot}\mathbf{u})^2$, the component of the broken enstrophy on the cell interiors only (thin lower line); the upper line for the broken enstrophy in the classical scheme is not visible on the figure, as it reaches and fluctuates around a value of approximately 5×10^5. The MEEVC scheme is constructed not to conserve, not the broken enstrophy, but an auxiliary enstrophy $\tilde{Q}_3(\omega) \coloneqq \frac{1}{2}\|\omega\|^2$ evaluated on the auxiliary vorticity.[13] Fig. 10.3b illustrates the evolution of the auxiliary enstrophy $\tilde{Q}_3(\omega)$ for the MEEVC scheme and our scheme (10.52) only; this is ill-defined for the classical scheme, as ω is not specified.



(a) Broken and internal enstrophy

(b) Auxiliary enstrophy

Figure 10.3: Evolution in different forms of the enstrophy for 3 different 1-stage integrators in the 2D vortex test (Subsection 10.4.4). Fig. 10.3a shows the broken $Q_3^*(\mathbf{u})$ (10.48) (thick upper line) and internal $\frac{1}{2}\sum_{K\in\mathcal{K}}\|\operatorname{curl}\mathbf{u}\|^2$ (thin lower line) enstrophies; Fig. 10.3b shows the auxiliary enstrophy $\tilde{Q}_3(\omega) \coloneqq \frac{1}{2}\|\omega\|^2$.

[13] In the FET interpretation, this requires ω to be interpreted as a projection into $\mathbb{X}_n$, continuous in time, as opposed to $\dot{\mathbb{X}}_n$, discontinuous in time. Both projections are equivalent when $\mathcal{I}_n$ is an S-node GL rule, i.e. one discretises in time using an S-stage Gauss method.

10.5 Energy- & helicity-stable integrators in MHD: elimination of Lagrange mutipliers & electromagnetic potential–to–field reparametrisation

We now consider the incompressible Hall MHD equations. This system may be written in the following nondimensionalised potential form:

$$\dot{\mathbf{u}} = \mathbf{u} \times \operatorname{curl} \mathbf{u} - \nabla p + \frac{2}{\beta} \operatorname{curl}^2 \mathbf{A} \times \operatorname{curl} \mathbf{A} - \frac{1}{\mathrm{Re}} \operatorname{curl}^2 \mathbf{u}, \tag{10.68a}$$

$$0 = \operatorname{div} \mathbf{u}, \tag{10.68b}$$

$$\dot{\mathbf{A}} = \mathbf{u} \times \operatorname{curl} \mathbf{A} - \nabla \varphi - \mathrm{R_H} \operatorname{curl}^2 \mathbf{A} \times \operatorname{curl} \mathbf{A} - \frac{1}{\mathrm{Re_m}} \operatorname{curl}^2 \mathbf{A}, \tag{10.68c}$$

$$0 = \operatorname{div} \mathbf{A}. \tag{10.68d}$$

Here (as ever) $\mathbf{u}$ and p denote the velocity and total pressure respectively, and $\mathrm{Re} > 0$ is the (fluid) Reynolds number; the new variables $\mathbf{A}$ and φ denote the magnetic and electric potential, and $\mathrm{Re_m}, \beta, \mathrm{R_H} > 0$ are the magnetic Reynolds number, plasma beta, and Hall coefficients. Similarly to Chapter 3, we consider periodic BCs with the additional constraints on the ICs

$$\int_\Omega \mathbf{u}(0) = \mathbf{0}, \qquad \int_\Omega \mathbf{A}(0) = \mathbf{0}. \tag{10.69}$$

Define the energy Q_1, magnetic helicity Q_2, and modified fluid/cross helicity Q_3 respectively,

$$Q_1(\mathbf{u}, \mathbf{A}) \coloneqq \frac{1}{2}\left[\|\mathbf{u}\|^2 + \frac{2}{\beta}\|\operatorname{curl} \mathbf{A}\|^2\right], \tag{10.70a}$$

$$Q_2(\mathbf{u}, \mathbf{A}) \coloneqq \frac{1}{2}(\mathbf{A}, \operatorname{curl} \mathbf{A}), \tag{10.70b}$$

$$Q_3(\mathbf{u}, \mathbf{A}) \coloneqq \frac{1}{2}\left[(\mathbf{u}, \operatorname{curl} \mathbf{u}) + \frac{4}{\beta \mathrm{R_H}}(\mathbf{u}, \operatorname{curl} \mathbf{A})\right]. \tag{10.70c}$$

Note for a, b satisfying $4ab = \beta \mathrm{R_H}(a + b)$, the hybrid helicity [MGM03] can be written as a combination of these QoIs as

$$\frac{1}{2}(\mathbf{A} + a\mathbf{u}, \operatorname{curl}[\mathbf{A} + b\mathbf{u}]) = Q_2 + abQ_3. \tag{10.71}$$

Preserving the behaviour of Q_2 and Q_3 is therefore sufficient to preserve the behaviour of the hybrid helicity. Under periodic BCs, Q_1, Q_2, Q_3 are each conserved in solutions of the formal ideal limit $\mathrm{Re} = \mathrm{Re_m} = \infty$, with Q_1 necessarily dissipated

for $\mathrm{Re} < \infty$; we wish to construct an energy- and helicity-stable timestepping scheme for the incompressible Hall MHD equations (10.68) i.e. one that preserves these behaviours.

The application of our framework (Algorithm 3.5) is largely similar to that in Section 3.1 & 10.4; we therefore omit the details for brevity, and move directly to the final energy- and helicity-stable integrator. Assuming for simplicity we take $\mathcal{I}_n$ to be an S-node quadrature rule, our SP scheme is equivalent to an S-stage Gauss collocation method applied to the following semi-discretisation: for a discretely divergence-free space $\mathbb{U}$, find $(\mathbf{u}, \mathbf{A}, \mathbf{j}, \mathbf{H}, \boldsymbol{\omega}) \in \mathbb{U}^5$ such that

$$(\dot{\mathbf{u}}, \mathbf{v}) = (\mathbf{u} \times \boldsymbol{\omega}, \mathbf{v}) + \frac{2}{\beta}(\mathbf{j} \times \mathbf{H}, \mathbf{v}) - \frac{1}{\mathrm{Re}}(\operatorname{curl} \mathbf{u}, \operatorname{curl} \mathbf{v}), \tag{10.72a}$$

$$(\dot{\mathbf{A}}, \mathbf{D}) = (\mathbf{u} \times \mathbf{H}, \mathbf{D}) - \mathrm{R_H}(\mathbf{j} \times \mathbf{H}, \mathbf{D}) - \frac{1}{\mathrm{Re_m}}(\operatorname{curl} \mathbf{A}, \operatorname{curl} \mathbf{D}), \tag{10.72b}$$

$$(\mathbf{j}, \mathbf{k}) = (\operatorname{curl} \mathbf{A}, \operatorname{curl} \mathbf{k}), \tag{10.72c}$$

$$(\mathbf{H}, \mathbf{G}) = (\operatorname{curl} \mathbf{A}, \mathbf{G}), \tag{10.72d}$$

$$(\boldsymbol{\omega}, \boldsymbol{\chi}) = (\operatorname{curl} \mathbf{u}, \boldsymbol{\chi}), \tag{10.72e}$$

for all $(\mathbf{v}, \mathbf{D}, \mathbf{k}, \mathbf{G}, \boldsymbol{\chi}) \in \mathbb{U}^5$. The AV $\mathbf{H}$ approximates the magnetic field $\operatorname{curl} \mathbf{A}$,[14] while $\mathbf{j}$ approximates the current $\operatorname{curl}^2 \mathbf{A}$; the AV $\boldsymbol{\omega}$ similarly approximates the vorticity $\operatorname{curl} \mathbf{u}$.

Theorem 10.10 (Energy & helicity stability of the incompressible Hall MHD integrator)**.** *When integrating in time using a Gauss method, the MHD integrator* (10.72) *is energy- and helicity-stable, with discrete analogues of the following results holding across each timestep T_n:*

$$\dot{Q}_1 = -\frac{1}{\mathrm{Re}}\|\operatorname{curl} \tilde{\mathbf{u}}\|^2 - \frac{2}{\beta \mathrm{Re_m}}\|\mathbf{j}\|^2 \leq 0, \tag{10.73a}$$

$$\dot{Q}_2 = -\frac{1}{\mathrm{Re_m}}(\mathbf{j}, \mathbf{H}), \tag{10.73b}$$

$$\dot{Q}_3 = -\frac{1}{\mathrm{Re}}(\operatorname{curl} \tilde{\mathbf{u}}, \operatorname{curl} \boldsymbol{\omega}) - \frac{2}{\beta \mathrm{R_H Re}}(\operatorname{curl} \tilde{\mathbf{u}}, \operatorname{curl} \mathbf{H}) - \frac{2}{\beta \mathrm{R_H Re_m}}(\mathbf{j}, \boldsymbol{\omega}), \tag{10.73c}$$

[14] We avoid calling this AV $\mathbf{B}$, as a variable which we call $\mathbf{B}$ will be introduced in the following subsection.

Proof. Each of these results holds by testing respectively against

$$(\mathbf{v}, \mathbf{D}, \mathbf{k}) = (\mathbf{u}, \frac{2}{\beta}\mathbf{j}, \frac{2}{\beta}\dot{\mathbf{A}}), \tag{10.74a}$$

$$(\mathbf{D}, \mathbf{G}) = (\mathbf{H}, \dot{\mathbf{A}}), \tag{10.74b}$$

$$(\mathbf{v}, \mathbf{D}, \mathbf{G}, \boldsymbol{\chi}) = (\boldsymbol{\omega} + \frac{2}{\beta \mathrm{R_H}}\mathbf{H}, \frac{2}{\beta \mathrm{R_H}}\boldsymbol{\omega}, \frac{2}{\beta \mathrm{R_H}}\dot{\mathbf{u}}, \dot{\mathbf{u}} + \frac{2}{\beta \mathrm{R_H}}\dot{\mathbf{A}}). \tag{10.74c}$$

□

10.5.1 Analysis: existence & uniqueness

Similarly to our analysis of the energy- and enstrophy-stable integrators (10.41, 10.44) in Subsection 10.4.1, we discuss certain preliminary existence and uniqueness results for the energy- and helicity-stable integrator (10.72) using the results for AD systems derived in Section 3.3, before proceeding to the application of FEEC.

We can see this to be a compatible SP discretisation of an AD system (Assumption 3.11) by interpreting the energy Q_1 as the dissipated type-B QoI, and the helicities Q_2, Q_3 as the additional type-A QoIs. Since all operators in this scheme are smooth, most required regularity results hold immediately, with the only result requiring some analysis being that the dissipative term defines an inner product on $\mathbb{U}$; equivalently, it is sufficient to show that the map $(\mathbf{u}, \mathbf{A}) \mapsto (\| \operatorname{curl} \mathbf{u}\|^2 + \| \operatorname{curl} \mathbf{A}\|^2)^{\frac{1}{2}}$ defines a norm on $\mathbb{U}^2$. Similarly to Subsection 10.4.1, this can be shown under compatibility conditions from FEEC. In particular, we assume the following: take $\mathbb{U}$ to be defined as in (10.18),

$$\mathbb{U} \coloneqq \left\{ \mathbf{u} \in \mathbb{V} : -(\mathbf{u}, \nabla q) = 0 \text{ for all } q \in \mathbb{Q} \text{ and } \int_\Omega \mathbf{u} = \mathbf{0} \right\}, \tag{10.75}$$

and assume $\mathbb{V}$, $\mathbb{Q}$ come form a discrete periodic de Rham complex as in (10.19)

$$\begin{array}{ccccccc} H^1 & \xrightarrow{\text{grad}} & \mathbf{H}_\#(\text{curl}) & \xrightarrow{\text{curl}} & \mathbf{H}_\#(\text{div}) & \xrightarrow{\text{div}} & L^2 \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ \mathbb{Q} & \xrightarrow{\text{grad}} & \mathbb{V} & \xrightarrow{\text{curl}} & \overline{\mathbb{V}} & \xrightarrow{\text{div}} & \overline{\mathbb{Q}} \end{array}. \tag{10.76}$$

Similarly to Subsection 10.3.1, we ensure exactness at $\mathbb{V}$ and $\overline{\mathbb{V}}$ by eliminating the harmonic forms. In such a case, defining the $\overline{\mathbb{Q}}$-discretely divergence-free subspace $\overline{\mathbb{U}} \subset \overline{\mathbb{V}}$ as in (3.7a),

$$\overline{\mathbb{U}} \coloneqq \left\{ \mathbf{u} \in \overline{\mathbb{V}} : (\operatorname{div} \mathbf{u}, q) = 0 \text{ for all } q \in \overline{\mathbb{Q}} \text{ and } \int_\Omega \mathbf{u} = \mathbf{0} \right\}, \tag{10.77}$$

we see as in Subsection 10.4.1 that $\operatorname{curl} : \mathbb{U} \to \overline{\mathbb{U}}$ defines an isomorphism by Lemma 10.4, implying immediately that $(\mathbf{u}, \mathbf{A}) \mapsto (\|\operatorname{curl} \mathbf{u}\|^2 + \|\operatorname{curl} \mathbf{A}\|^2)^{\frac{1}{2}}$ defines a norm on $\mathbb{U}^2$.

For existence we refer again to Theorem 3.18.

Example (Incompressible Hall MHD)

Assuming that $\mathbb{V}$, $\mathbb{Q}$ form part of a discrete de Rham complex (10.76) and that we define $\mathbb{U}$ as in (10.75), solutions to our proposed energy- and helicity-stable integrator for the incompressible Hall MHD equations (10.72) exist on arbitrary timesteps Δt_n in either the viscous ($\mathrm{Re}, \mathrm{Re_m} < \infty$) or lowest-order-in-time ($S = 1$) case.

For uniqueness we refer to Theorem 3.26.

Example (Incompressible Hall MHD)

Assuming that $\mathbb{V}$, $\mathbb{Q}$ form part of a discrete de Rham complex (10.76) and that we define $\mathbb{U}$ as in (10.75), the integrators (10.30, 10.34) are well-posed with a unique solution for either sufficiently small Re and $\mathrm{Re_m}$, or in the lowest-order-in-time case ($S = 1$) with sufficiently small Δt_n.

10.5.2 Application of FEEC

We now consider the ways in which FEEC may be used to simplify the scheme (10.72). This involves both the elimination of LMs and an equivalent reparametrisation in the more approachable EM field $\mathbf{B} = \operatorname{curl} \mathbf{A}$, $\mathbf{E} = -\dot{\mathbf{A}} - \nabla\varphi$.

10.5.2.1 Elimination of Lagrange multipliers

The first observation is similar to that on the energy- and helicity-stable integrator in Section 10.3.1.

Expanding out the LMs in (10.72) would yield a 10-field discretisation, with 5 LMs. Through the use of FE spaces compatible with FEEC, we may eliminate 3 of these LMs: those enforcing the discrete divergence-free conditions on $\mathbf{H}$, $\mathbf{j}$,

$\boldsymbol{\omega}$. This is very similar in practice to the elimination of LMs in the energy- and helicity-stable NS integrators in Section 10.3.

Let us first apply IBP to write (10.72c–10.72e) in the form

$$(\mathbf{j}, \mathbf{k}) = (\operatorname{curl} \mathbf{A}, \operatorname{curl} \mathbf{k}), \tag{10.78a}$$

$$(\mathbf{H}, \mathbf{G}) = (\mathbf{A}, \operatorname{curl} \mathbf{G}), \tag{10.78b}$$

$$(\boldsymbol{\omega}, \boldsymbol{\chi}) = (\mathbf{u}, \operatorname{curl} \boldsymbol{\chi}). \tag{10.78c}$$

We suppose then (aligning with the requirements in the analysis above) that the spaces $\mathbb{V}, \mathbb{Q}$ exist as part of a discrete periodic de Rham complex (10.76), and that the $\mathbb{Q}$-discretely divergence-free subspace $\mathbb{U} \subset \mathbb{V}$ is defined as in (10.75). Identifying the space $\mathbb{U}$ may be identified as the nullspace $\mathcal{N} \operatorname{grad}^* \subset \mathbb{V}$ of the dual operator $\operatorname{grad}^* : \mathbb{V} \to \mathbb{Q}$. Lemma 10.3 then implies that the projection (10.17) defining the auxiliary current, magnetic field, and vorticity $\mathbf{j}, \mathbf{H}, \boldsymbol{\omega} \in \mathbb{U}$ may equivalently be written as projections in the larger space $\mathbb{V}$. The integrator (10.72) may be written equivalently as: find $((\mathbf{u}, \mathbf{A}), (\mathbf{j}, \mathbf{H}, \boldsymbol{\omega})) \in \mathbb{U}^2 \times \mathbb{V}^3$ such that (10.72) holds for all $((\mathbf{v}, \mathbf{D}), (\mathbf{k}, \mathbf{G}, \boldsymbol{\chi})) \in \mathbb{U}^2 \times \mathbb{V}^3$. Expanding the LMs in $\mathbb{U}$ gives then the equivalent, partially simplified, 7-field semi-discretisation: find $((\mathbf{u}, \mathbf{A}, \mathbf{j}, \mathbf{H}, \boldsymbol{\omega}), (p, \varphi)) \in \mathbb{V}^5 \times \mathbb{Q}^2$ such that

$$(\dot{\mathbf{u}}, \mathbf{v}) = (\mathbf{u} \times \boldsymbol{\omega}, \mathbf{v}) - (\nabla p, \mathbf{v}) + \frac{2}{\beta}(\mathbf{j} \times \mathbf{H}, \mathbf{v}) - \frac{1}{\mathrm{Re}}(\operatorname{curl} \mathbf{u}, \operatorname{curl} \mathbf{v}), \tag{10.79a}$$

$$(\dot{\mathbf{A}}, \mathbf{D}) = (\mathbf{u} \times \mathbf{H}, \mathbf{D}) - (\nabla \varphi, \mathbf{D}) - \mathrm{R_H}(\mathbf{j} \times \mathbf{H}, \mathbf{D}) - \frac{1}{\mathrm{Re_m}}(\mathbf{j}, \mathbf{D}), \tag{10.79b}$$

$$(\mathbf{j}, \mathbf{k}) = (\operatorname{curl} \mathbf{A}, \operatorname{curl} \mathbf{k}), \tag{10.79c}$$

$$(\mathbf{H}, \mathbf{G}) = (\operatorname{curl} \mathbf{A}, \mathbf{G}), \tag{10.79d}$$

$$(\boldsymbol{\omega}, \boldsymbol{\chi}) = (\operatorname{curl} \mathbf{u}, \boldsymbol{\chi}), \tag{10.79e}$$

$$0 = (\mathbf{u}, \nabla q), \tag{10.79f}$$

$$0 = (\mathbf{A}, \nabla \phi), \tag{10.79g}$$

for all $((\mathbf{v}, \mathbf{D}, \mathbf{k}, \mathbf{G}, \boldsymbol{\chi}), (q, \phi)) \in \mathbb{V}^5 \times \mathbb{Q}^2$.

10.5.2.2 Electromagnetic potential–to–field reparametrisation

The second observation is similar to that on the energy- and enstrophy-stable integrator in Section 10.4.2.

One immediate observation about the scheme (10.79) is that, except in (10.79g), the EM potentials $\mathbf{A}$, φ only feature in the forms of their derivatives $\mathbf{B} = \operatorname{curl}\mathbf{A}$, $\mathbf{E} = -\dot{\mathbf{A}} - \nabla\varphi$, the magnetic and electric field respectively. These fields represent far more typical fields over which to pose the MHD equations. We may therefore consider those circumstances under which we may equivalently reparametrise (10.79) in the EM fields $\mathbf{B}$, $\mathbf{E}$.

Similarly to Section 10.4, with $\mathbb{V}$, $\mathbb{Q}$ coming from the above discrete de Rham complex (10.76), Lemma 10.4 implies $\operatorname{curl} : \mathbb{U} \to \overline{\mathbb{U}}$ defines an isomorphism, with the divergence-free subspace $\overline{\mathbb{U}} \subset \overline{\mathbb{V}}$ defined as in (10.77). We may then introduce $\mathbf{B} = \operatorname{curl}\mathbf{A} \in \mathbb{U}$ as an additional variable through the projection

$$(\dot{\mathbf{B}}, \mathbf{C}) = (\operatorname{curl}\dot{\mathbf{A}}, \mathbf{C}) \tag{10.80a}$$

for all $\mathbf{C} \in \overline{\mathbb{U}}$, and the IC $\mathbf{B} = \operatorname{curl}\mathbf{A}$ at $t = 0$; one may see this is sufficient to ensure $\dot{\mathbf{B}} = \operatorname{curl}\dot{\mathbf{A}}$ exactly by taking $\mathbf{C} = \dot{\mathbf{B}} - \operatorname{curl}\dot{\mathbf{A}}$. By the complex property $\operatorname{curl}\circ\operatorname{grad} = 0$, this may be written identically as

$$(\dot{\mathbf{B}}, \mathbf{C}) = (\operatorname{curl}[\dot{\mathbf{A}} + \nabla\varphi], \mathbf{C}). \tag{10.80b}$$

Lemma 10.2 then implies we may equivalently define this as the projection onto $\overline{\mathbb{V}}$ in place of $\overline{\mathbb{U}}$ (provided the IC on $\mathbf{B}$ lies in $\overline{\mathbb{U}}$) yielding the 8-field semi-discretisation: find $((\mathbf{u}, \mathbf{A}, \mathbf{j}, \mathbf{H}, \boldsymbol{\omega}), \mathbf{B}, (p, \varphi)) \in \mathbb{V}^5 \times \overline{\mathbb{V}} \times \mathbb{Q}^2$ such that

$$(\dot{\mathbf{u}}, \mathbf{v}) = (\mathbf{u} \times \boldsymbol{\omega}, \mathbf{v}) - (\nabla p, \mathbf{v}) + \frac{2}{\beta}(\mathbf{j} \times \mathbf{H}, \mathbf{v}) - \frac{1}{\mathrm{Re}}(\operatorname{curl}\mathbf{u}, \operatorname{curl}\mathbf{v}), \tag{10.81a}$$

$$0 = (\mathbf{u} \times \mathbf{H}, \mathbf{D}) - (\dot{\mathbf{A}} + \nabla\varphi, \mathbf{D}) - \mathrm{R}_{\mathrm{H}}(\mathbf{j} \times \mathbf{H}, \mathbf{D}) - \frac{1}{\mathrm{Re}_{\mathrm{m}}}(\mathbf{j}, \mathbf{D}), \tag{10.81b}$$

$$(\mathbf{j}, \mathbf{k}) = (\mathbf{B}, \operatorname{curl}\mathbf{k}), \tag{10.81c}$$

$$(\mathbf{H}, \mathbf{G}) = (\mathbf{B}, \mathbf{G}), \tag{10.81d}$$

$$(\boldsymbol{\omega}, \boldsymbol{\chi}) = (\operatorname{curl}\mathbf{u}, \boldsymbol{\chi}), \tag{10.81e}$$

$$(\dot{\mathbf{B}}, \mathbf{C}) = (\operatorname{curl}[\dot{\mathbf{A}} + \nabla\varphi], \mathbf{C}), \tag{10.81f}$$

$$0 = (\mathbf{u}, \nabla q), \tag{10.81g}$$

$$0 = (\mathbf{A}, \nabla\phi), \tag{10.81h}$$

for all $((\mathbf{v}, \mathbf{D}, \mathbf{k}, \mathbf{G}, \boldsymbol{\chi}), \mathbf{C}, (q, \phi)) \times \mathbb{V}^5 \times \overline{\mathbb{V}} \times \mathbb{Q}^2$.

Since $\dot{\mathbf{A}}$ spans $\mathbb{U} = \mathcal{N}\operatorname{grad}^*$ and $\nabla\varphi$ spans $\nabla\mathbb{Q} = \mathcal{R}\operatorname{grad}$, by the Hodge decomposition $-\dot{\mathbf{A}} - \nabla\varphi$ spans $\mathcal{N}\operatorname{grad}^* \oplus \mathcal{R}\operatorname{grad} = \mathbb{V}$. We may therefore reparametrise our

semi-discretisation (10.81) through a variable $\mathbf{E} = -\dot{\mathbf{A}} - \nabla\varphi \in \mathbb{V}$ in place of both $\mathbf{A}$ and φ: find $((\mathbf{u}, \mathbf{E}, \mathbf{j}, \mathbf{H}, \boldsymbol{\omega}), \mathbf{B}, p) \in \mathbb{V}^5 \times \overline{\mathbb{V}} \times \mathbb{Q}$ such that

$$(\dot{\mathbf{u}}, \mathbf{v}) = (\mathbf{u} \times \boldsymbol{\omega}, \mathbf{v}) - (\nabla p, \mathbf{v}) + \frac{2}{\beta}(\mathbf{j} \times \mathbf{H}, \mathbf{v}) - \frac{1}{\mathrm{Re}}(\operatorname{curl} \mathbf{u}, \operatorname{curl} \mathbf{v}), \tag{10.82a}$$

$$0 = (\mathbf{u} \times \mathbf{H}, \mathbf{D}) + (\mathbf{E}, \mathbf{D}) - \mathrm{R_H}(\mathbf{j} \times \mathbf{H}, \mathbf{D}) - \frac{1}{\mathrm{Re_m}}(\mathbf{j}, \mathbf{D}), \tag{10.82b}$$

$$(\mathbf{j}, \mathbf{k}) = (\mathbf{B}, \operatorname{curl} \mathbf{k}), \tag{10.82c}$$

$$(\mathbf{H}, \mathbf{G}) = (\mathbf{B}, \mathbf{G}), \tag{10.82d}$$

$$(\boldsymbol{\omega}, \boldsymbol{\chi}) = (\operatorname{curl} \mathbf{u}, \boldsymbol{\chi}), \tag{10.82e}$$

$$(\dot{\mathbf{B}}, \mathbf{C}) = -(\operatorname{curl} \mathbf{E}, \mathbf{C}), \tag{10.82f}$$

$$0 = (\mathbf{u}, \nabla q) \tag{10.82g}$$

for all $((\mathbf{v}, \mathbf{D}, \mathbf{k}, \mathbf{G}, \boldsymbol{\chi}), \mathbf{C}, q) \in \mathbb{V}^5 \times \overline{\mathbb{V}} \times \mathbb{Q}$. This is our final semi-discretisation, which, as noted in Section 8.1, aligns precisely with that proposed by Laakmann, Hu & Farrell [LHF23] at lowest order in time $S = 1$.

Using the EM field reparametrisation (10.82) ICs must be posed on $\mathbf{B} \in \overline{\mathbb{U}}$, i.e. must be exactly divergence-free in $\overline{\mathbb{V}}$. Each of the structures in Theorem 10.10 can then be seen (again, when using a Gauss method for the time discretisation) by testing respectively against

$$(\mathbf{v}, \mathbf{D}, \mathbf{k}, q) = (\mathbf{u}, \frac{2}{\beta}\mathbf{j}, \frac{2}{\beta}\mathbf{E}, p), \tag{10.83a}$$

$$(\mathbf{D}, \mathbf{G}) = (\mathbf{H}, \mathbf{E}), \tag{10.83b}$$

$$(\mathbf{v}, \mathbf{D}, \mathbf{G}, \boldsymbol{\chi}) = (\boldsymbol{\omega} + \frac{2}{\beta \mathrm{R_H}}\mathbf{H}, \frac{2}{\beta \mathrm{R_H}}\boldsymbol{\omega}, \frac{2}{\beta \mathrm{R_H}}\dot{\mathbf{u}} + \frac{2}{\beta \mathrm{R_H}}\nabla p, \dot{\mathbf{u}} + \nabla p - \frac{2}{\beta \mathrm{R_H}}\mathbf{E}), \tag{10.83c}$$

noting (10.82f) implies $\dot{\mathbf{B}} = -\operatorname{curl} \mathbf{E}$ holds exactly.

Remark 10.11 (Application of analysis after reparametrisation)**.** *Similarly to Remark 10.8, provided all the FE complex criteria hold, this reparametrisation* (10.82) *is exactly equivalent to the original scheme* (10.72)*, and both the existence and uniqueness results from Subsection 10.5.1 still necessarily hold.*

References

[AS81] M. J. Ablowitz and H. Segur. *Solitons and the Inverse Scattering Transform*. SIAM, Jan. 1981.

[Aca18] J. W. Acaster. *"Recognise". James Acaster: Repertoire, episode 1*. Distributor: Netflix. Mar. 2018.

[AP24a] M. Ainsworth and C. Parker. "Computing H^2-conforming finite element aproximations without having to implement C^1-elements". In: *SIAM Journal on Scientific Computing* 46.4 (Aug. 2024), A2398–A2420.

[AP24b] M. Ainsworth and C. Parker. "Two and three dimensional H^2-conforming finite element approximations without C^1-elements". In: *Computer Methods in Applied Mechanics and Engineering* 431 (Nov. 2024), p. 117267.

[Alf84] P. Alfeld. "A trivariate Clough–Tocher scheme for tetrahedral data". In: *Computer Aided Geometric Design* 1.2 (Nov. 1984), pp. 169–181.

[Alf43] H. Alfvén. "On the existence of electromagnetic-hydrodynamic waves". In: *Arkiv for matematik, astronomi och fysik* 29B.2 (1943), pp. 1–7.

[And25] B. D. Andrews. *Software used in 'Geometric numerical integration via auxiliary variables'*. GitHub pre-release. July 2025. URL: https://github.com/BorisAndrews/thesis_code/releases/tag/v1.0.

[AF25] B. D. Andrews and P. E. Farrell. *Enforcing conservation laws and dissipation inequalities numerically via auxiliary variables*. arXiv manuscript. Apr. 2025.

[Ara66] A. Arakawa. "Computational design for long-term numerical integration of the equations of fluid motion: Two-dimensional incompressible flow. Part I". In: *Journal of Computational Physics* 1.1 (1966), pp. 119–143.

[AL77] A. Arakawa and V. R. Lamb. *Computational design of the basic dynamical processes of the UCLA general circulation model*. Tech. rep. Jan. 1977.

[Arn18] D. N. Arnold. *Finite Element Exterior Calculus*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, 2018.

[AFW06] D. N. Arnold, R. S. Falk, and R. Winther. "Finite Element Exterior Calculus, Homological Techniques, and Applications". In: *Acta Numerica* 15 (May 2006), pp. 1–155.

[AFW09] D. N. Arnold, R. S. Falk, and R. Winther. "Finite Element Exterior Calculus: From Hodge Theory to Numerical Stability". In: *Bulletin of the American Mathematical Society* 47 (June 2009).

[Arn14] V. I. Arnold. "The asymptotic Hopf invariant and its applications". In: *Vladimir I. Arnold - Collected works: Hydrodynamics, bifurcation theory, and algebraic geometry 1965-1972*. Berlin, Heidelberg: Springer, 2014, pp. 357–375.

[AK08] V. I. Arnold and B. A. Khesin. *Topological Methods in Hydrodynamics*. Springer Science & Business Media, Jan. 2008.

[AKN06] V. I. Arnold, V. V. Kozlov, and A. I. Neishtadt. *Mathematical Aspects of Classical and Celestial Mechanics*. 3rd ed. Vol. 3. Encyclopaedia of Mathematical Sciences. Springer, 2006.

[Art95] W. Arter. "Numerical simulation of magnetic fusion plasmas". In: *Reports on Progress in Physics* 58.1 (1995).

[Art23] W. Arter. *Equations for EXCALIBUR/NEPTUNE Proxyapps*. Tech. rep. CD/EXCALIBUR-FMS/0021-1.32-M1.2.1. UKAEA, Oct. 2023. URL: https://github.com/ExCALIBUR-NEPTUNE/Documents/blob/main/reports/ukaea_reports/CD-EXCALIBUR-FMS0021-1.31-M1.2.1.pdf.

[Art+21] F. J. Artola et al. "3D simulations of vertical displacement events in tokamaks: A benchmark of M3D-C1, NIMROD, and JOREK". In: *Physics of Plasmas* 28.5 (May 2021), p. 052511.

[AP98] U. M. Ascher and L. R. Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. Philadelphia, PA, United States: SIAM, Jan. 1998.

[Bal+24] S. Balay et al. *PETSc/TAO users manual*. ANL-21/39 - Revision 3.21. 2024.

[BBM97] T. B. Benjamin, J. L. Bona, and J. J. Mahony. "Model equations for long waves in nonlinear dispersive systems". In: *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 272.1220 (Jan. 1997), pp. 47–78.

[BF84] M. A. Berger and G. B. Field. "The topological properties of magnetic helicity". In: *Journal of Fluid Mechanics* 147 (Oct. 1984), pp. 133–148.

[BS00a] P. Betsch and P. Steinmann. "Conservation properties of a time FE method—part I: time-stepping schemes for N-body problems". In: *International Journal for Numerical Methods in Engineering* 49.5 (2000), pp. 599–638.

[BS00b] P. Betsch and P. Steinmann. "Inherently energy conserving time finite elements for classical mechanics". In: *Journal of Computational Physics* 160.1 (May 2000), pp. 88–116.

[BS01] P. Betsch and P. Steinmann. "Conservation properties of a time FE method—part II: time-stepping schemes for non-linear elastodynamics". In: *International Journal for Numerical Methods in Engineering* 50.8 (2001), pp. 1931–1955.

[BC17] S. Blanes and F. Casas. *A Concise Introduction to Geometric Numerical Integration*. Boca Raton, FL, United States: CRC Press, Nov. 2017.

[Bob75] A. V. Bobylev. "Exact solutions of the Boltzmann equation". In: *Akademiia Nauk SSSR Doklady* 225 (Dec. 1975), pp. 1296–1299.

[BPS95] A. V. Bobylev, A. Palczewski, and J. Schneider. "On approximation of the Boltzmann equation by discrete velocity models". In: *Comptes rendus de l'Académie des sciences. Série I, Mathématique* 320.5 (1995), pp. 639–644.

[BV08] A. V. Bobylev and M. C. Vinerean. "Construction of discrete kinetic models with given invariants". In: *Journal of Statistical Physics* 132.1 (July 2008), pp. 153–170.

[Bor70] J. P. Boris. "Relativistic plasma simulation-optimization of a hybrid code". In: *Proc. Fourth Conf. Num. Sim. Plasmas*. 1970, pp. 3–67.

[BB80] J. U Brackbill and D. C Barnes. "The effect of nonzero $\nabla \cdot \mathbf{B}$ on the numerical solution of the magnetohydrodynamic equations". In: *Journal of Computational Physics* 35.3 (May 1980), pp. 426–430.

[BF85] J. U. Brackbill and D. W. Forslund. "Simulation of low-frequency, electromagnetic phenomena in plasmas". In: *Multiple time scales*. Ed. by J. U. Brackbill and B. I. Cohen. Academic Press, Jan. 1985, pp. 271–310.

[BDM85] F. Brezzi, J. Douglas, and L. D. Marini. "Two families of mixed finite elements for second order elliptic problems". In: *Numerische Mathematik* 47 (1985), pp. 217–235.

[BK25] P. D. Brubeck and R. C. Kirby. *FIAT: enabling classical and modern macroelements*. arXiv manuscript. Jan. 2025.

[BFI19] L. Brugnano, G. Frasca-Caccia, and F. Iavernaro. "Line integral solution of Hamiltonian PDEs". In: *Mathematics* 7.3 (Mar. 2019), p. 275.

[BI12] L. Brugnano and F. Iavernaro. "Line integral methods which preserve all invariants of conservative problems". In: *Journal of Computational and Applied Mathematics* 236.16 (Oct. 2012), pp. 3905–3919.

[BI16] L. Brugnano and F. Iavernaro. *Line Integral Methods for Conservative Problems*. Boca Raton, FL, United States: CRC Press, Mar. 2016.

[BEH24] A. Brunk, H. Egger, and O. Habrich. "A second-order structure-preserving discretization for the Cahn-Hilliard/Allen-Cahn system with cross-kinetic coupling". In: *Applied Numerical Mathematics* 206 (Dec. 2024), pp. 12–28.

[BE25] A. Brunk and M. F. P. ten Eikelder. *A simple, fully-discrete, unconditionally energy-stable method for the two-phase Navier-Stokes Cahn-Hilliard model with arbitrary density ratios*. arXiv manuscript. Apr. 2025.

[BF25a] A. Brunk and M. Fritz. *Analysis and structure-preserving approximation of a Cahn-Hilliard-Forchheimer system with solution-dependent mass and volume source*. Apr. 2025.

[BF25b] A. Brunk and M. Fritz. "Structure-preserving approximation of the Cahn-Hilliard-Biot system". In: *Numerical Methods for Partial Differential Equations* 41.1 (Nov. 2025).

[BGL24] A. Brunk, J. Giesselmann, and M. Lukáčová-Medviďová. *Robust a posteriori error control for the Allen–Cahn equation with variable mobility*. arXiv manuscript. Mar. 2024.

[BJL25] A. Brunk, A. Jüngel, and M. Lukáčová-Medviďová. *A structure-preserving numerical method for quasi-incompressible Navier-Stokes-Maxwell-Stefan systems*. Apr. 2025.

[BLS25] A. Brunk, M. Lukáčová-Medviďová, and D. Schumann. *Structure-preserving approximation of the non-isothermal Cahn-Hilliard system*. arXiv manuscript. June 2025.

[BS24] A. Brunk and D. Schumann. "Nonisothermal Cahn–Hilliard Navier–Stokes system". In: *Proceedings in Applied Mathematics and Mechanics* 24.2 (2024).

[BS25] A. Brunk and D. Schumann. "Structure-preserving approximation for the non-isothermal Cahn–Hilliard–Navier–Stokes system". In: *Numerical Mathematics and Advanced Applications ENUMATH 2023*. Ed. by A. Sequeira et al. Vol. 1. Cham: Springer Nature Switzerland, Apr. 2025, pp. 188–197.

[Bru+23a] A. Brunk et al. "A second-order fully-balanced structure-preserving variational discretization scheme for the Cahn–Hilliard–Navier–Stokes system". In: *Mathematical Models and Methods in Applied Sciences* 33.12 (Nov. 2023), pp. 2587–2627.

[Bru+23b] A. Brunk et al. “Stability and discretization error analysis for the Cahn–Hilliard system via relative energy estimates”. In: *ESAIM: Mathematical Modelling and Numerical Analysis* 57.3 (May 2023), pp. 1297–1322.

[BP03] C. J. Budd and M. D. Piggott. “Geometric integration and its applications”. In: *Handbook of numerical analysis*. Vol. XI. Amsterdam: North-Holland Publishing Co., 2003, pp. 35–139.

[Bue96] C. Buet. “A discrete-velocity scheme for the Boltzmann operator of rarefied gas dynamics”. In: *Transport Theory and Statistical Physics* 25.1 (Jan. 1996), pp. 33–60.

[BF07] E. Burman and M. A. Fernández. “Continuous interior penalty finite element method for the time-dependent Navier–Stokes equations: Space discretization and convergence”. In: *Numerische Mathematik* 107.1 (July 2007), pp. 39–77.

[BFH06] E. Burman, M. A. Fernández, and P. Hansbo. “Continuous interior penalty finite element method for Oseen’s equations”. In: *SIAM Journal on Numerical Analysis* 44.3 (Jan. 2006), pp. 1248–1274.

[BH04] E. Burman and P. Hansbo. “Edge stabilization for Galerkin approximations of convection–diffusion–reaction problems”. In: *Computer Methods in Applied Mechanics and Engineering*. Recent Advances in Stabilized and Multiscale Finite Element Methods 193.15 (Apr. 2004), pp. 1437–1453.

[Cah61] J. W. Cahn. “On spinodal decomposition”. In: *Acta Metallurgica* 9.9 (Sept. 1961), pp. 795–801.

[CH58] J. W. Cahn and J. E. Hilliard. “Free energy of a non-uniform system I: Interfacial free energy”. In: *Journal of Chemistry and Physics* 28 (1958), pp. 258–267.

[CIZ97] M. Calvo, A. Iserles, and A. Zanna. “Numerical solution of isospectral flows”. In: *Mathematics of Computation* 66.220 (1997), pp. 1461–1486.

[Can+99] J. Cantarella et al. “Influence of geometry and topology on helicity”. In: *Geophysical Monograph Series* 111 (Jan. 1999), pp. 17–24.

[CJ21] E. Celledoni and J. I. Jackaman. “Discrete conservation laws for finite element discretisations of multisymplectic PDEs”. In: *Journal of Computational Physics* 444 (Nov. 2021), p. 110520.

[Cel+09] E. Celledoni et al. “Energy-preserving Runge–Kutta methods”. In: *ESAIM: Mathematical Modelling and Numerical Analysis* 43.4 (2009), pp. 645–649.

[Cer12] C. Cercignani. *The Boltzmann Equation and its Applications*. Springer New York, Oct. 2012.

[Cha20] J Chan. “Entropy stable reduced order modeling of nonlinear conservation laws”. In: *Journal of Computational Physics* 423 (Dec. 2020), p. 109789.

[Cha18] J. Chan. “On discretely entropy conservative and entropy stable discontinuous Galerkin methods”. In: *Journal of Computational Physics* 362 (June 2018), pp. 346–374.

[Cha25] J. Chan. *An artificial viscosity approach to high order entropy stable discontinuous Galerkin methods*. arXiv manuscript. Jan. 2025.

[Cha+17] S. Charnyi et al. “On conservation laws of Navier–Stokes Galerkin discretizations”. In: *Journal of Computational Physics* 337 (May 2017), pp. 289–308.

[CH24] L. Chen and X. Huang. “Finite element de Rham and Stokes complexes in three dimensions”. In: *Mathematics of Computation* 93.345 (Jan. 2024), pp. 55–110.

[CS20] T. Chen and C.-W. Shu. "Review of entropy stable discontinuous Galerkin methods for systems of conservation laws on unstructured simplex meshes". In: *CSIAM Transactions on Applied Mathematics* 1.1 (2020), pp. 1–52.

[Cho98] A. R. Choudhuri. *The Physics of Fluids and Plasmas: An Introduction for Astrophysicists*. Cambridge University Press, Nov. 1998.

[CMO11] S. H. Christiansen, H. Z. Munthe-Kaas, and B. Owren. "Topics in structure-preserving discretization". In: *Acta Numerica* 20 (May 2011), pp. 1–119.

[Chu92] M. T. Chu. "Matrix differential equations: A continuous realization process for linear algebra problems". In: *Nonlinear Analysis* 18.12 (1992), pp. 1125–1146.

[CT65] R. W. Clough and J. L. Tocher. "Finite element stiffness matricess for analysis of plate bending". In: *Proceedings of the First Conference on Matrix Methods in Structural Mechanics* (1965), pp. 515–546.

[CKS00] B. Cockburn, G. E. Karniadakis, and C.-W. Shu. "The development of discontinuous Galerkin methods". In: *Discontinuous Galerkin Methods*. Berlin: Springer, 2000, pp. 3–50.

[CH11] D. Cohen and E. Hairer. "Linear energy-preserving integrators for Poisson systems". In: *BIT Numerical Mathematics* 51.1 (Mar. 2011), pp. 91–101.

[Coh+07] R. H. Cohen et al. "Large-timestep mover for particle simulations of arbitrarily magnetized species". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*. Proceedings of the 16th international symposium on heavy ion inertial fusion 577.1 (July 2007), pp. 52–57.

[CRS92] F. Coquel, F. Rogier, and J. Schneider. "A deterministic method for solving the homogeneous Boltzmann equation". In: *La Recherche Aerospatiale (English Edition)* 3 (Jan. 1992), pp. 1–10.

[DO11] M. Dahlby and B. Owren. "A general framework for deriving integral preserving numerical methods for PDEs". In: *SIAM Journal on Scientific Computing* 33.5 (Jan. 2011). Publisher: Society for Industrial and Applied Mathematics, pp. 2318–2340.

[DW98] W. Dai and P. R. Woodward. "On the divergence-free condition and conservation laws in numerical simulations for supersonic magnetohydrodynamical flows". In: *The Astrophysical Journal* 494.1 (Feb. 1998), p. 317.

[DLP98] F. Diele, L. Lopez, and T. Politi. "One step semi-explicit methods based on the Cayley transform for solving isospectral flows". In: *Journal of Computational and Applied Mathematics* 89.2 (Mar. 1998), pp. 219–223.

[DD76] J. Douglas Jr. and T. Dupont. "Interior Penalty Procedures for Elliptic and Parabolic Galerkin Methods". In: *Computing Methods in Applied Sciences*. Vol. 58. Lecture Notes in Physics. Berlin: Springer Verlag, 1976, pp. 207–216.

[EHS21] H. Egger, O. Habrich, and V. Shashkov. "On the energy stable approximation of Hamiltonian and gradient systems". In: *Computational Methods in Applied Mathematics* 21.2 (Apr. 2021), pp. 335–349.

[EG21a] A. Ern and J.-L. Guermond. *Finite Elements I: Approximation and Interpolation*. Vol. 72. Texts in Applied Mathematics. Cham, Switzerland: Springer International Publishing, 2021.

[EG21b] A. Ern and J.-L. Guermond. *Finite Elements II: Galerkin Approximation, Elliptic and Mixed PDEs*. Vol. 73. Texts in Applied Mathematics. Cham, Switzerland: Springer International Publishing, 2021.

[EG21c] A. Ern and J.-L. Guermond. *Finite Elements III: First-Order and Time-Dependent PDEs*. Vol. 74. Texts in Applied Mathematics. Cham, Switzerland: Springer International Publishing, 2021.

[EG22] A. Ern and J.-L. Guermond. "Invariant-domain-preserving high-order time stepping: I. Explicit Runge–Kutta schemes". In: *SIAM Journal on Scientific Computing* 44.5 (Oct. 2022). Publisher: Society for Industrial and Applied Mathematics, A3366–A3392.

[EG23] A. Ern and J.-L. Guermond. "Invariant-domain preserving high-order time stepping: II. IMEX schemes". In: *SIAM Journal on Scientific Computing* 45.5 (Oct. 2023). Publisher: Society for Industrial and Applied Mathematics, A2511–A2538.

[Eva10] L. C. Evans. *Partial Differential Equations*. American Mathematical Soc., 2010.

[FKM21] P. E. Farrell, R. C. Kirby, and J. Marchena-Menéndez. "Irksome: Automating Runge–Kutta time-stepping for finite element methods". In: *ACM Transactions on Mathematical Software* 47.4 (Sept. 2021), 30:1–30:26.

[FMS24] P. E. Farrell, L. Mitchell, and L. R. Scott. "Two conjectures on the Stokes complex in three dimensions on Freudenthal meshes". In: *SIAM Journal on Scientific Computing* 46.2 (Apr. 2024), A629–A644.

[Fei04] E. Feireisl. *Dynamics of Viscous Compressible Fluids*. OUP Oxford, 2004.

[FLM20] E. Feireisl, M. Lukáčová-Medviďová, and H. Mizerová. "A finite volume scheme for the Euler system inspired by the two velocities approach". In: *Numerische Mathematik* 144.1 (Jan. 2020), pp. 89–132.

[Fei+21] E. Feireisl et al. *Numerical Analysis of Compressible Fluid Flows*. Cham, Switzerland: Springer International Publishing, 2021.

[FMP06] F. Filbet, C. Mouhot, and L. Pareschi. "Solving the Boltzmann equation in $N \log_2 N$". In: *SIAM Journal on Scientific Computing* 28.3 (Jan. 2006), pp. 1029–1053.

[Fla74] H. Flaschka. "The Toda lattice. II. Existence of integrals". In: *Physical Review B* 9.4 (Feb. 1974), pp. 1924–1925.

[FS90] D. A. French and J. W. Schaeffer. "Continuous finite element methods which preserve energy properties for nonlinear problems". In: *Applied Mathematics and Computation* 39.3 (Oct. 1990), pp. 271–295.

[FP00] J. Frink and D. Payne. *"Insane Clown Poppy". The Simpsons, season 12, episode 3*. Director: R. Anderson. Distributor: Fox Broadcasting Company, LLC. Nov. 2000.

[Fri+75] U. Frisch et al. "Possibility of an inverse cascade of magnetic helicity in magnetohydrodynamic turbulence". In: *Journal of Fluid Mechanics* 68.4 (Apr. 1975), pp. 769–778.

[FGN20] G. Fu, J. Guzmán, and M. Neilan. "Exact smooth piecewise polynomial sequences on Alfeld splits". In: *Mathematics of Computation* 89.323 (May 2020), pp. 1059–1091.

[FM10] D. Furihata and T. Matsuo. *Discrete Variational Derivative Method: A Structure-Preserving Numerical Method for Partial Differential Equations*. Boca Raton, FL, United States: CRC Press, Dec. 2010.

[GG21a] E. S. Gawlik and F. Gay–Balmaz. "A structure-preserving finite element method for compressible ideal and resistive magnetohydrodynamics". In: *Journal of Plasma Physics* 87.5 (Oct. 2021), p. 835870501.

[GG21b] E. S. Gawlik and F. Gay-Balmaz. "A variational finite element discretization of compressible flow". In: *Foundations of Computational Mathematics* 21.4 (Aug. 2021), pp. 961–1001.

[GY17] F. Gay-Balmaz and H. Yoshimura. "A Lagrangian variational formulation for nonequilibrium thermodynamics. Part I: Discrete systems". In: *Journal of Geometry and Physics* 111 (Jan. 2017), pp. 169–193.

[GM88] Z. Ge and J. E. Marsden. "Lie–Poisson Hamilton–Jacobi theory and Lie–Poisson integrators". In: *Physics Letters A* 133.3 (Nov. 1988), pp. 134–139.

[GCW10] T. C. Genoni, R. E. Clark, and D. R. Welch. "A fast implicit algorithm for highly magnetized charged particle motion". In: *The Open Plasma Physics Journal* 3.1 (Apr. 2010).

[GKT25] J. Giesselmann, A. Karsai, and T. Tscherpel. "Energy-consistent Petrov–Galerkin time discretization of port-Hamiltonian systems". In: *The SMAI Journal of Computational Mathematics* 11 (2025), pp. 335–367.

[GR12] V. Girault and P.-A. Raviart. *Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms*. Springer Science & Business Media, Dec. 2012.

[GKP19] H. Goedbloed, R. Keppens, and S. Poedts. *Magnetohydrodynamics of Laboratory and Astrophysical Plasmas*. Cambridge: Cambridge University Press, 2019.

[Gon96] O. Gonzalez. "Time integration and discrete Hamiltonian systems". In: *Journal of Nonlinear Science* 6.5 (Sept. 1996), pp. 449–467.

[GN91] R. Grant and D. R. Naylor. *"Justice". Red Dwarf, series 4, episode 3.* Director: E. Bye. Distributor: BBC2. Feb. 1991.

[Grm84] M. Grmela. "Particle and bracket formulations of kinetic equations". In: *Contemporary Mathematics* 28 (1984), pp. 125–132.

[GÖ97] M. Grmela and H. C. Öttinger. "Dynamics and thermodynamics of complex fluids. I. Development of a general formalism". In: *Physical Review E* 56.6 (Dec. 1997), pp. 6620–6632.

[GLN22] J. Guzmán, A. Lischke, and M. Neilan. "Exact sequences on Worsey–Farin splits". In: *Mathematics of Computation* 91.338 (Nov. 2022), pp. 2571–2608.

[HL14] E. Hairer and C. Lubich. "Energy-diminishing integration of gradient systems". In: *IMA Journal of Numerical Analysis* 34.2 (Apr. 2014), pp. 452–461.

[HL18] E. Hairer and C. Lubich. "Energy behaviour of the Boris method for charged-particle dynamics". In: *BIT Numerical Mathematics* 58.4 (Dec. 2018), pp. 969–979.

[HL20] E. Hairer and C. Lubich. "Long-term analysis of a variational integrator for charged-particle dynamics in a strong magnetic field". In: *Numerische Mathematik* 144.3 (Mar. 2020), pp. 699–728.

[HLW06] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. Heidelberg, Germany: Springer Science & Business Media, May 2006.

[Hai+06] E. Hairer et al. "Geometric Numerical Integration". In: *Oberwolfach Reports* 3.1 (Dec. 2006), pp. 805–882.

[HW18] F. D. Halpern and R. E. Waltz. "Anti-symmetric plasma moment equations with conservative discrete counterparts". In: *Physics of Plasmas* 25.6 (June 2018), p. 060703.

[Ham+23] D. A. Ham et al. *Firedrake user manual*. Tech. rep. May 2023.

[Ham34] W. R. Hamilton. "XV. On a general method in dynamics; by which the study of the motions of all free systems of attracting or repelling points is reduced to the search and differentiation of one central relation, or characteristic function". In: *Philosophical Transactions of the Royal Society* 124 (1834), pp. 247–308.

[Han23] M.-L. Hanot. "An arbitrary order and pointwise divergence-free finite element scheme for the incompressible 3D Navier–Stokes equations". In: *SIAM Journal on Numerical Analysis* 61.2 (2023), pp. 784–811.

[HW65] F. J. Harlow and J. E. Welch. "Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface". In: *Physics of Fluids* 8.12 (1965), p. 2182.

[Har+20] C. R. Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (2020), pp. 357–362.

[HLL83] A. Harten, P. D. Lax, and B. van Leer. "On upstream differencing and Godunov-type schemes for hyperbolic conservation laws". In: *SIAM Review* 25.1 (1983), pp. 35–61.

[HM03] R. D. Hazeltine and J. D. Meiss. *Plasma Confinement*. Courier Corporation, Jan. 2003.

[HHX19] J. He, K. Hu, and J. Xu. "Generalized Gaffney inequality and discrete compactness for discrete differential forms". In: *Numerische Mathematik* 143.4 (Dec. 2019), pp. 781–795.

[He+25] M. He et al. *Topology-preserving discretization for the magneto-frictional equations arising in the Parker conjecture*. arXiv manuscript. Jan. 2025.

[Hec12] F. Hecht. "New development in freefem++". In: *Journal of Numerical Mathematics* 20.3-4 (Dec. 2012), pp. 251–266.

[Hen93] J. Henrard. "The adiabatic invariant in classical mechanics". In: *Dynamics Reported: Expositions in Dynamical Systems*. Ed. by C. K. R. T. Jones, U. Kirchgraber, and H. O. Walther. Springer, 1993, pp. 117–235.

[Hil94] M. J. M. Hill. "VI. On a spherical vortex". In: *Philosophical Transactions of the Royal Society of London. (A.)* 185 (1894), pp. 213–245.

[Hip01] R. Hiptmair. "Higher order Whitney forms". In: *Progress in Electromagnetics Research* 32 (2001), pp. 271–299.

[Hoe+21] M. Hoelzl et al. "The JOREK non-linear extended MHD code and applications to large-scale instabilities and their control in magnetically confined fusion plasmas". In: *Nuclear Fusion* 61.6 (May 2021), p. 065001.

[Hu25] K. Hu. *Many facets of cohomology: Differential complexes and structure-aware formulations*. arXiv manuscript. Apr. 2025.

[HLX21] K. Hu, Y.-J. Lee, and J. Xu. "Helicity-conservative finite element discretization for incompressible MHD systems". In: *Journal of Computational Physics* 436 (Mar. 2021), p. 110284.

[HZZ22] K. Hu, Q. Zhang, and Z. Zhang. "A family of finite element Stokes complexes in three dimensions". In: *SIAM Journal on Numerical Analysis* 60.1 (Feb. 2022), pp. 222–243.

[HC07] G. T. A. Huysmans and O. Czarny. "MHD stability in X-point geometry: simulation of ELMs". In: *Nuclear Fusion* 47.7 (June 2007), p. 659.

[IQ18] A. Iserles and G. R. W. Quispel. "Why Geometric Numerical Integration?" In: *Discrete Mechanics, Geometric Integration and Lie–Butcher Series*. Ed. by K. Ebrahimi-Fard and M. Barbero Liñán. Springer Proceedings in Mathematics & Statistics. Cham: Springer International Publishing, 2018, pp. 1–28.

[Jac19] J. I. Jackaman. "Finite element methods as geometric structure preserving algorithms". PhD. University of Reading, Jan. 2019.

[JP21] J. I. Jackaman and T. Pryer. "Conservative Galerkin methods for dispersive Hamiltonian problems". In: *Calcolo* 58.3 (2021).

[Jac11] D. Jackson. "Über die Genauigkeit der Annäherung stetiger Funktionen durch ganze rationale Funktionen gegebenen Grades und trigonometrische Summen gegebener Ordnung". PhD. Göttingen, 1911.

[Kau84] A. N. Kaufman. "Dissipative Hamiltonian systems: A unifying principle". In: *Physics Letters A* 100.8 (Feb. 1984), pp. 419–422.

[Kee01] K. Keeler. *"Time Keeps on Slippin"'. Futurama, season 3, episode 14*. Director: C. Louden. Distributor: Fox Broadcasting Company, LLC. May 2001.

[KG08] C. A. Kennedy and A. Gruber. "Reduced aliasing formulations of the convective terms within the Navier–Stokes equations for a compressible fluid". In: *Journal of Computational Physics* 227.3 (Jan. 2008), pp. 1676–1700.

[Kov89] S. Kovalevskaya. "Sur le problème de la rotation d'un corps solide autour d'un point fixe". In: *Acta Mathematica* 12 (1889), pp. 177–232.

[Koz07] R. Kozlov. "Conservative discretizations of the Kepler motion". In: *Journal of Physics A: Mathematical and Theoretical* 40.17 (Apr. 2007), p. 4529.

[La 22] G. La Scala. *Variational Time-Steppers that are Finite Element in Time for Firedrake and Irksome*. MEng Project Report. Imperial College London, June 2022, p. 63. (Visited on 12/02/2023).

[LHF23] F. Laakmann, K. Hu, and P. E. Farrell. "Structure-preserving and helicity-conserving finite element approximations and preconditioning for the Hall MHD equations". In: *Journal of Computational Physics* 492 (Nov. 2023), p. 112410.

[LG74] R. A. LaBudde and D. Greenspan. "Discrete mechanics—a general treatment". In: *Journal of Computational Physics* 15.2 (June 1974), pp. 134–167.

[Lax68] P. D. Lax. "Integrals of nonlinear equations of evolution and solitary waves". In: *Communications on Pure and Applied Mathematics* 21.5 (1968), pp. 467–490.

[LP24] D. Lombardi and C. Pagliantini. *Conformal variational discretisation of infinite dimensional Hamiltonian systems with gradient flow dissipation*. arXiv manuscript. Dec. 2024.

[Luc05] G. W. Lucas Jr. *Star Wars: Episode III – Revenge of the Sith*. United States. Distributor: 20th Century Fox. May 2005.

[MT89] Y. Maday and E. Tadmor. "Analysis of the spectral vanishing viscosity method for periodic conservation laws". In: *SIAM Journal on Numerical Analysis* 26.4 (Aug. 1989), pp. 854–870.

[MQR99] R. I. McLachlan, G. R. W. Quispel, and N. Robidoux. "Geometric integration using discrete gradients". In: *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 357.1754 (Apr. 1999). Publisher: Royal Society, pp. 1021–1045.

[MS20] R. I. McLachlan and A. Stern. "Multisymplecticity of hybridizable discontinuous Galerkin methods". In: *Foundations of Computational Mathematics* 20.1 (Feb. 2020), pp. 35–69.

[Mel16] T. Melfi. *Hidden Figures*. United States, Distributor: 20th Century Fox. Dec. 2016.

[MA76] F. Mesinger and A. Arakawa. *Numerical methods used in atmospheric models*. Report. Global Atmospheric Research Programme (GARP), 1976.

[Mil26] A. A. Milne. *Winnie-the-Pooh*. en. London, United Kingdom: Metheun, 1926.

[MN02] Y. Minesaki and Y. Nakamura. "A new discretization of the Kepler motion which conserves the Runge–Lenz vector". In: *Physics Letters A* 306.2 (Dec. 2002), pp. 127–133.

[MN04] Y. Minesaki and Y. Nakamura. "A new conservative numerical integration algorithm for the three-dimensional Kepler motion based on the Kustaanheimo–Stiefel regularization theory". In: *Physics Letters A* 324.4 (Apr. 2004), pp. 282–292.

[MGM03] P. D. Mininni, D. O. Gómez, and S. M. Mahajan. "Dynamo action in magnetohydrodynamics and Hall-magnetohydrodynamics". In: *The Astrophysical Journal* 587.1 (Apr. 2003), p. 472.

[Mof69] H. K. Moffatt. "The degree of knottedness of tangled vortex lines". In: *Journal of Fluid Mechanics* 35.1 (Jan. 1969), pp. 117–129.

[Mof81] H. K. Moffatt. "Some developments in the theory of turbulence". In: *Journal of Fluid Mechanics* 106 (May 1981), pp. 27–47.

[Mof14] H. K. Moffatt. "Helicity and singular structures in fluid dynamics". In: *Proceedings of the National Academy of Sciences* 111.10 (Mar. 2014), pp. 3663–3670.

[MT92] H. K. Moffatt and A. Tsinober. "Helicity in laminar and turbulent flow". In: *Annual Review of Fluid Mechanics* 24.1 (1992), pp. 281–312.

[Mor61] J. J. Moreau. "Constantes d'un Îlot Tourbillonnaire en Fluide Parfait Barotrope". In: *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences* 252 (1961), pp. 2810–2812.

[MS75] J. Morgan and R. Scott. "A nodal basis for C^1 piecewise polynomials of degree $n \geq 5$". In: *Mathematics of Computation* 29.131 (1975), pp. 736–740.

[Mor10] Y. Morinishi. "Skew-symmetric form of convective terms and fully conservative finite difference schemes for variable density low-Mach number flows". In: *Journal of Computational Physics* 229.2 (Jan. 2010), pp. 276–300.

[Mor84a] P. J. Morrison. "Bracket formulation for irreversible classical fields". In: *Physics Letters A* 100.8 (Feb. 1984), pp. 423–427.

[Mor84b] P. J. Morrison. *Some observations regarding brackets and dissipation*. Tech. rep. PAM-228. University of California at Berkeley, 1984.

[Mor86] P. J. Morrison. "A paradigm for joined Hamiltonian and dissipative systems". In: *Physica D: Nonlinear Phenomena* 18.1 (1986), pp. 410–419.

[MP06] C. Mouhot and L. Pareschi. "Fast algorithms for computing the Boltzmann collision operator". In: *Mathematics of Computation* 75.256 (2006), pp. 1833–1852.

[MSP15] R. C. Moura, S. J. Sherwin, and J. Peiró. "Linear dispersion–diffusion analysis and its application to under-resolved turbulence simulations using discontinuous Galerkin spectral/hp methods". In: *Journal of Computational Physics* 298 (Oct. 2015), pp. 695–710.

[Mou+17] R. C. Moura et al. “On the eddy-resolving capability of high-order discontinuous Galerkin approaches to implicit LES / under-resolved DNS of Euler turbulence”. In: *Journal of Computational Physics* 330 (Feb. 2017), pp. 615–623.

[Mou+22] R. C. Moura et al. “Gradient jump penalty stabilisation of spectral/hp element discretisation for under-resolved turbulence simulations”. In: *Computer Methods in Applied Mechanics and Engineering* 388 (Jan. 2022).

[Néd86] J.-C. Nédélec. “A new family of mixed finite elements in $\mathbb{R}^3$”. In: *Numerische Mathematik* 50.1 (1986), pp. 57–81.

[Nei15] M. Neilan. “Discrete and conforming smooth de Rham complexes in three dimensions”. In: *Mathematics of Computation* 84.295 (Sept. 2015), pp. 2059–2081.

[Nor22] J. Nordström. “A skew-symmetric energy and entropy stable formulation of the compressible Euler equations”. In: *Journal of Computational Physics* 470 (Dec. 2022), p. 111573.

[Nor63] T. G. Northrop. *The Adiabatic Motion of Charged Particles*. New York (State): Interscience Publishers, 1963.

[Ött05] H. C. Öttinger. *Beyond Equilibrium Thermodynamics*. John Wiley & Sons, May 2005.

[ÖG97] H. C. Öttinger and M. Grmela. “Dynamics and thermodynamics of complex fluids. II. Illustrations of a general formalism”. In: *Physical Review E* 56.6 (Dec. 1997), pp. 6633–6655.

[PG17] A. Palha and M. Gerritsma. “A mass, energy, enstrophy and vorticity conserving (MEEVC) mimetic spectral element discretization for the 2D incompressible Navier–Stokes equations”. In: *Journal of Computational Physics* 328 (Jan. 2017), pp. 200–220.

[PP96] L. Pareschi and B. Perthame. “A Fourier spectral method for homogeneous boltzmann equations”. In: *Transport Theory and Statistical Physics* 25.3-5 (Apr. 1996), pp. 369–382.

[PR22] L. Pareschi and T. Rey. “Moment preserving Fourier–Galerkin spectral methods and application to the Boltzmann equation”. In: *SIAM Journal on Numerical Analysis* 60.6 (Dec. 2022), pp. 3216–3240.

[PR00a] L. Pareschi and G. Russo. “Numerical solution of the Boltzmann equation I: Spectrally accurate approximation of the collision operator”. In: *SIAM Journal on Numerical Analysis* 37.4 (Jan. 2000), pp. 1217–1245.

[PR00b] L. Pareschi and G. Russo. “On the stability of spectral methods for the homogeneous Boltzmann equation”. In: *Transport Theory and Statistical Physics* 29.3-5 (Apr. 2000), pp. 431–447.

[Par+16] M. Parsani et al. “Entropy stable staggered grid discontinuous spectral collocation methods of any order for the compressible Navier–Stokes equations”. In: *SIAM Journal on Scientific Computing* 38.5 (Jan. 2016), A3129–A3162.

[Pav+11] D. Pavlov et al. “Structure-preserving discretization of incompressible fluids”. In: *Physica D: Nonlinear Phenomena* 240.6 (Mar. 2011), pp. 443–458.

[PB09] J. C. Perez and S Boldyrev. “Role of cross-helicity in magnetohydrodynamic turbulence”. In: *Physical Review Letters* 102.2 (Jan. 2009), p. 025003.

[Phi59] N. A. Phillips. “An example of non-linear computational instability”. In: *The Atmosphere and the Sea in Motion* 501 (1959), p. 504.

[PW70] S. A. Piacsek and G. P. Williams. "Conservation properties of convection difference schemes". In: *Journal of Computational Physics* 6.3 (Dec. 1970), pp. 392–405.

[Poi90] H. Poincaré. "Sur les equations aux dérivées partielles de la physique mathématique". In: *American Journal of Mathematics* 12.3 (1890).

[Reb07] L. G. Rebholz. "An energy- and helicity-conserving finite element scheme for the Navier–Stokes equations". In: *SIAM Journal on Numerical Analysis* 45.4 (Jan. 2007), pp. 1622–1638.

[RP20] Y. Renard and K. Poulios. "GetFEM: Automated FE modeling of multiphysics problems based on a generic weak form language". In: *ACM Trans. Math. Softw.* 47.1 (Dec. 2020), 4:1–4:31.

[RC20] L. F. Ricketson and L. Chacón. "An energy-conserving and asymptotic-preserving charged-particle orbit implicit time integrator for arbitrary electromagnetic fields". In: *Journal of Computational Physics* 418 (Oct. 2020), p. 109639.

[RS94] F. Rogier and J. Schneider. "A direct method for solving the Boltzmann equation". In: *Transport Theory and Statistical Physics* 23.1-3 (Jan. 1994), pp. 313–338.

[Rom09] I. Romero. "Thermodynamically consistent time-stepping algorithms for non-linear thermomechanical systems". In: *International Journal for Numerical Methods in Engineering* 79.6 (Mar. 2009), pp. 706–732.

[SC94] J. M. Sanz-Serna and M. P. Calvo. *Numerical Hamiltonian Problems*. London, United Kingdom: CRC Press, May 1994.

[Sch30] J. Schauder. "Der fixpunktsatz in funktionalraümen". In: *Studia Mathematica* 2.1 (1930), pp. 171–180.

[SV85a] L. R. Scott and M. Vogelius. "Conforming finite element methods for incompressible and nearly incompressible continua". In: *Large Scale Computations in Fluid Mechanics*. Vol. 22. Providence: American Mathematical Society, 1985, pp. 221–244.

[SV85b] L. R. Scott and M. Vogelius. "Norm estimates for a maximal right inverse of the divergence operator in spaces of piecewise polynomials". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 19.1 (1985), pp. 111–143.

[SÖ20] X. Shang and H. C. Öttinger. "Structure-preserving integrators for dissipative systems based on reversible–irreversible splitting". In: *Proceedings of the Royal Society A* 476.2234 (Feb. 2020).

[SXY18] J. Shen, J. Xu, and J. Yang. "The scalar auxiliary variable (SAV) approach for gradient flows". In: *Journal of Computational Physics* 353 (Jan. 2018), pp. 407–416.

[SA94] J. C. Simo and F. Armero. "Unconditional stability and long-term behavior of transient algorithms for the incompressible Navier–Stokes and Euler equations". In: *Computer Methods in Applied Mechanics and Engineering* 111.1 (Jan. 1994), pp. 111–154.

[Smi17] C. B. Smiet. "Knots in plasma". PhD. Leiden University, 2017.

[Sov+03] C. R. Sovinec et al. "NIMROD: A computational laboratory for studying nonlinear fusion magnetohydrodynamics". In: *Physics of Plasmas* 10.5 (May 2003), pp. 1727–1732.

[Sta12] W. M. Stacey. *Fusion Plasma Physics*. John Wiley & Sons, Nov. 2012.

[SZ23] A. Stern and E. Zampa. *Multisymplecticity in finite element exterior calculus*. arXiv manuscript. Dec. 2023.

[SC07] R. H. Stogner and G. F. Carey. "C1 macroelements in adaptive finite element methods". In: *International Journal for Numerical Methods in Engineering* 70.9 (2007), pp. 1076–1095.

[Sto45] G. G. Stokes. "On the theories of internal friction of fluids in motion". In: *Transactions of the Cambridge Philosophical Society* 8 (1845), pp. 287–305.

[Tad87] E. Tadmor. "The numerical viscosity of entropy stable schemes for systems of conservation laws. I". In: *Mathematics of Computation* 49.179 (1987), pp. 91–103.

[Tad90] E. Tadmor. "Shock capturing by the spectral viscosity method". In: *Computer Methods in Applied Mechanics and Engineering* 80.1 (June 1990), pp. 197–208.

[Tad93] E. Tadmor. "Total variation and error estimates for spectral viscosity approximations". In: *Mathematics of Computation* 60.201 (1993), pp. 245–256.

[Tad03] E. Tadmor. "Entropy stability theory for difference approximations of nonlinear conservation laws and related time-dependent problems". In: *Acta Numerica* 12 (May 2003), pp. 451–512.

[Tad16] E. Tadmor. "Perfect derivatives, conservative differences and entropy stable computation of hyperbolic conservation laws". In: *Discrete and Continuous Dynamical Systems* 36.8 (2016), pp. 4579–4598.

[Taf85] L. G. Taff. *Celestial Mechanics*. New York City, NY, United States: Wiley, May 1985.

[Tem24] R. Temam. *Navier–Stokes Equations: Theory and Numerical Analysis*. American Mathematical Society, May 2024.

[The23a] The MathWorks Inc. *MATLAB version: 24.1 (R2024a)*. Natick, MA, United States, 2023.

[The23b] The MathWorks Inc. *Optimization Toolbox version: 24.1 (R2024a)*. Natick, MA, United States, 2023.

[Tho10] S. P. Thompson. *Calculus Made Easy*. en. Macmillan, 1910.

[Tol54] J. R. R. Tolkien. *The Two Towers*. Vol. 2. The Lord of the Rings. London: George Allen & Unwin, 1954.

[Tre20] L. N. Trefethen. *Approximation Theory and Approximation Practice*. Philadelphia, PA, United States: SIAM, 2020.

[Tu10] L. W. Tu. *An Introduction to Manifolds*. Springer Science & Business Media, Oct. 2010.

[VB95] H. X. Vu and J. U. Brackbill. "Accurate numerical solution of charged particle motion in a magnetic field". In: *Journal of Computational Physics* 116.2 (Feb. 1995), pp. 384–387.

[Zem90] R. L. Zemeckis. *Back to the Future Part III*. United States. Distributor: Universal Pictures. May 1990.

[Zha+22] Y. Zhang et al. "A mass-, kinetic energy- and helicity-conserving mimetic dual-field discretization for three-dimensional incompressible Navier–Stokes equations, part I: Periodic domains". In: *Journal of Computational Physics* 451 (Feb. 2022), p. 110868.

[Zha+24] Y. Zhang et al. "A MEEVC discretization for two-dimensional incompressible Navier-Stokes equations with general boundary conditions". In: *Journal of Computational Physics* 510 (Aug. 2024), p. 113080.