Chapter 2

Geometric numerical integration for ODEs

Here we summarise several key results from the research area of geometric numerical integration. While no new results are given in this chapter, it serves not only as an introduction to structure preserving numerical methods but is the genesis point for this thesis. All results presented in this chapter can be found in [85, 161, 127]. Several ideas here have also been investigated in the masters thesis [104]. As such, we shall not typically present the proofs of results here, instead providing references to where the proofs were originally presented. Geometric numerical integration of ODEs provides powerful tools for the design of temporal discretisations for PDEs. We shall investigate structure preserving temporal discretisations of PDEs from Chapter 4 onwards.

2.1 The continuous problem

Let $\boldsymbol{u}(t) = \boldsymbol{u} \in (C^1([0,T]))^D$ for $t \in [0,T]$ with T some predetermined constant, and where D is a positive integer. Then, we define a general ODE such that

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u} = \boldsymbol{F}(\boldsymbol{u}, t), \tag{2.1}$$

subject to the initial data $\boldsymbol{u}(0) = \boldsymbol{u}_0$ for $\boldsymbol{u}_0 \in \mathbb{R}^D$, and where $\frac{\mathrm{d}}{\mathrm{d}t}$ represents a temporal first derivative. Note that here $\boldsymbol{F}(\boldsymbol{u}, t)$ can include general linear operators acting on $\boldsymbol{u}$ which physically represent either spatial operators, or discretisations of spatial operators. Typically, for an ODE to preserve geometric properties of the continuous level we require

it to be *autonomous*, that is to say that it does not explicitly depend on time, i.e., we seek $\boldsymbol{u}$ such that

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u} = \boldsymbol{f}(\boldsymbol{u}). \tag{2.2}$$

To mark this difference we have changed notation for the right hand side of (2.2) adding additional clarity to the type of operator we are considering. Within this chapter we restrict ourselves to the study of autonomous problems of the form (2.2).
Of course, there is no general structure we expect to preserve for a general autonomous ODE of the form (2.2). For example, when $D = 1$, we could choose $\boldsymbol{f}(u) = u^2$ and have u blow up in finite time, or $\boldsymbol{f}(u) = -u$ and have u decay exponentially. As such, we introduce the notion of a Hamiltonian system.

Definition 2.1.1 (A Hamiltonian system)**.** *Assume that D is* even*, then let $\boldsymbol{u} \in (C^1([0,T]))^D$, $\mathcal{H} : \mathbb{R}^D \to \mathbb{R}$ with $\mathcal{H}(\boldsymbol{u}) \in C^1([0,T])$, and $J \in \mathbb{R}^{D\times D}$ be a* constant *skew-symmetric matrix. A Hamiltonian system is given by seeking $\boldsymbol{u}$ such that*

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u} = J\nabla\mathcal{H}(\boldsymbol{u}), \tag{2.3}$$

subject to the initial condition $\boldsymbol{u}(0) = \boldsymbol{u_0}$. Throughout we assume that J is invertible, which is expected for physical Hamiltonian systems.

In the sequel whenever discussing structure preservation of ODEs we are implicitly restricting our study to Hamiltonian systems. For particular examples of Hamiltonian systems see §3.3.0

Remark 2.1.2 (Skew-symmetric inner product)**.** *By definition, a skew-symmetric matrix $J \in \mathbb{R}^{D\times D}$ satisfies*

$$J^T = -J.$$

This definition implies that for a given vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ we have that

$$\boldsymbol{v}\cdot J\boldsymbol{w} = -\boldsymbol{w}\cdot J\boldsymbol{v},$$

as $\boldsymbol{v}\cdot J\boldsymbol{w} = J^T\boldsymbol{v}\cdot\boldsymbol{w}$. This tells us that the skew-symmetric matrix J induces a skew-symmetric inner product. In fact, this inner product defines a Poisson bracket, see [85], similarly to that we discuss in §4.1 and §8.1.1.

Here we shall focus on two key geometric properties possessed by Hamiltonian systems.

The first of which is the conservation of the Hamiltonian functional $\mathcal{H}(\boldsymbol{u})$, which physically represents energy, over time.

Theorem 2.1.3 (Conservation of the Hamiltonian over time)**.** *The Hamiltonian function (or total energy) $\mathcal{H}(\boldsymbol{u})$, corresponding to the system (2.3), is conserved, i.e.,*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}(\boldsymbol{u}) = 0.$$

Proof. In view of the chain rule we see that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}(\boldsymbol{u}) &= \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}\cdot\nabla\mathcal{H}(\boldsymbol{u}) \\ &= J\nabla\mathcal{H}(\boldsymbol{u})\cdot\nabla\mathcal{H}(\boldsymbol{u}),\end{aligned}$$

after the application of (2.3). Through the skew-symmetric inner product induced by J as discussed in Remark 2.1.2 we observe that the Hamiltonian function is preserved over time.

□

Hamiltonian systems also conserve a symplectic structure, which is characterised by the behaviour of the solution vector $\boldsymbol{u}$ over time. When the vector dimension $D = 2$ the symplectic structure corresponds to area conservation in the solution space (u_1, u_2). For a detailed introduction to the notion of symplectic mappings for Hamiltonian systems where $D = 2$ see [161, §2]. In arbitrary dimension the notion of symplecticity encompasses volume conservation in the solution space, but it is a stronger condition, see [127, §3.5].
There are two main (equivalent) ways of viewing the notion of symplecticity. The first of which is through the flow map of the solution, see for example [94]. The second, which we shall present here, is through differential calculus.

Definition 2.1.4 (Exterior calculus notation and fundamental properties, [127])**.** *Let $\mathbf{d}\boldsymbol{u}$ denote the Cartan exterior derivative of $\boldsymbol{u}$, and additionally let $\wedge$ denote the skew-symmetric wedge product, which is also known as the exterior product. We shall not provide concise definitions of these operators here as they require the introduction of multiple additional concepts, and instead refer the reader to [170, 72] for a gentle introduction. Throughout we shall define the exterior derivative and wedge product through the following identities:*
Let $\mathbf{d}\boldsymbol{a}$, $\mathbf{d}\boldsymbol{b}$ and $\mathbf{d}\boldsymbol{c}$ be Cartan exterior derivatives over a subdomain of $\mathbb{R}^d$ and $\alpha, \beta \in \mathbb{R}$.

Then the wedge product, $\wedge$, respects a skew-symmetry

$$\mathbf{d}\boldsymbol{a} \wedge \mathbf{d}\boldsymbol{b} = -\,\mathbf{d}\boldsymbol{b} \wedge \mathbf{d}\boldsymbol{a}, \tag{2.4}$$

bilinearity

$$\mathbf{d}\boldsymbol{a} \wedge (\alpha\,\mathbf{d}\boldsymbol{b} + \beta\,\mathbf{d}\boldsymbol{c}) = \alpha\,\mathbf{d}\boldsymbol{a} \wedge \mathbf{d}\boldsymbol{b} + \beta\,\mathbf{d}\boldsymbol{a} \wedge \mathbf{d}\boldsymbol{c}, \tag{2.5}$$

and a rule of matrix multiplication

$$\mathbf{d}\boldsymbol{a} \wedge (A\,\mathbf{d}\boldsymbol{b}) = \left(A^T\,\mathbf{d}\boldsymbol{a}\right) \wedge \mathbf{d}\boldsymbol{b}, \tag{2.6}$$

for $A \in \mathbb{R}^{d\times d}$.

Theorem 2.1.5 (Symplectic structure preservation over time)**.** *Let $\boldsymbol{u}$ denote the solution of the Hamiltonian system (2.3), then*

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{d}\boldsymbol{u} \wedge J^{-1}\,\mathbf{d}\boldsymbol{u}\right) = 0, \tag{2.7}$$

i.e, the symplectic structure $\mathbf{d}\boldsymbol{u} \wedge J^{-1}\,\mathbf{d}\boldsymbol{u}$ is preserved over time.

Proof of Theorem 2.1.5. Through application of the product rule, and bilinearity of the wedge product (2.5), we find

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{d}\boldsymbol{u} \wedge J^{-1}\,\mathbf{d}\boldsymbol{u}\right) &= \mathbf{d}\left(\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}\right) \wedge J^{-1}\,\mathbf{d}\boldsymbol{u} + \mathbf{d}\boldsymbol{u} \wedge J^{-1}\,\mathbf{d}\left(\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}\right) \\
&= \mathbf{d}\left(\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}\right) \wedge J^{-1}\,\mathbf{d}\boldsymbol{u} - J^{-1}\,\mathbf{d}\left(\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}\right) \wedge \mathbf{d}\boldsymbol{u} \\
&= 2\,\mathbf{d}\left(\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}\right) \wedge J^{-1}\,\mathbf{d}\boldsymbol{u},
\end{aligned}$$

through (2.4), (2.6) and the skew-symmetry of J^{-1}. Applying the ODE (2.3) we find that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{d}\boldsymbol{u} \wedge J^{-1}\,\mathbf{d}\boldsymbol{u}\right) &= 2J^{-1}\nabla\nabla^{\mathrm{T}}\mathcal{H}\left(\boldsymbol{u}\right)\,\mathbf{d}\boldsymbol{u} \wedge J^{-1}\,\mathbf{d}\boldsymbol{u} \\
&= -2J^{-1}J^{-1}\nabla\nabla^{\mathrm{T}}\mathcal{H}\left(\boldsymbol{u}\right)\,\mathbf{d}\boldsymbol{u} \wedge \mathbf{d}\boldsymbol{u},
\end{aligned}$$

where $\nabla\nabla^{\mathrm{T}}\mathcal{H}\left(\boldsymbol{u}\right)$ is the *Hessian* of $\mathcal{H}\left(\boldsymbol{u}\right)$, through (2.6) and the skew-symmetry of J^{-1}. Note that the square of a skew-symmetric matrix is symmetric. Additionally the Hessian $\nabla\nabla^{\mathrm{T}}\mathcal{H}\left(\boldsymbol{u}\right)$ is symmetric as the partial derivatives of $\mathcal{H}\left(\boldsymbol{u}\right)$ with respect to u_i are smooth

for $i = 1, ..., D$, through Schwarz's theorem. So we can write

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{d}\boldsymbol{u} \wedge J^{-1}\,\mathbf{d}\boldsymbol{u}\right) = -2A\,\mathbf{d}\boldsymbol{u} \wedge \mathbf{d}\boldsymbol{u},$$

where $A = J^{-1}J^{-1}\nabla\nabla^{\mathrm{T}}\mathcal{H}(\boldsymbol{u})$ is symmetric. Through application of (2.4) and (2.6) we observe that

$$A\,\mathbf{d}\boldsymbol{u} \wedge \mathbf{d}\boldsymbol{u} = -A\,\mathbf{d}\boldsymbol{u} \wedge \mathbf{d}\boldsymbol{u},$$

allowing us to conclude the proof.

□

2.2 Discrete structure preservation

Throughout this section we shall recall numerical algorithms from the literature and investigate how well they preserve discrete equivalents of the structure we investigated on the continuous level.

We partition our temporal interval $[0, T]$ such that $0 := t_0 < t_1 < \cdots < t_N =: T$ where we define the length of each subinterval as $\tau_n := t_{n+1} - t_n$ for $n = 0, ..., N-1$. Here all numerical approximations we consider consist solely of point evaluations at the nodes t_n. We shall denote the point value approximating $\boldsymbol{u}(t_n)$ as $\boldsymbol{u}^n$ throughout, and shall always fix the initial value of our numerical approximation as $\boldsymbol{u}^0 = \boldsymbol{u}(0)$.

As our discrete approximations consist only of a collection of point values we are required to introduce discrete notions of conservation of the Hamiltonian and symplectic structure over time.

Recall in the continuous setting that conservation of the Hamiltonian is written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathcal{H}(\boldsymbol{u})\right) = 0. \tag{2.8}$$

While we no longer have the notation of a time derivative we can instead define conservation of the Hamiltonian through a difference quotient, i.e.,

$$\frac{\mathcal{H}(\boldsymbol{u}^{n+1}) - \mathcal{H}(\boldsymbol{u}^n)}{\tau_n} = 0. \tag{2.9}$$

Note that through integrating the left hand side of (2.8) we observe that

$$\int_{t_n}^{t_{n+1}} \frac{\mathrm{d}}{\mathrm{d}t} \left(\mathcal{H} \left(\boldsymbol{u} \right) \right) \mathrm{d}t = \frac{\mathcal{H} \left(\boldsymbol{u} \left(t_{n+1} \right) \right) - \mathcal{H} \left(\boldsymbol{u} \left(t_n \right) \right)}{\tau_n},$$

through the fundamental theorem of calculus, so (2.9) holds for continuous problems but is weaker.

We define the notion of preservation of the symplectic mapping in the same way as energy conservation, i.e., through difference quotients. We formally define the notion of preservation of the symplectic structure locally as

$$\frac{\mathbf{d}\boldsymbol{u}^{n+1} \wedge J^{-1} \mathbf{d}\boldsymbol{u}^{n+1} - \mathbf{d}\boldsymbol{u}^{n} \wedge J^{-1} \mathbf{d}\boldsymbol{u}^{n}}{\tau_n} = 0, \tag{2.10}$$

for more details on (2.10) see [162] or [127]. Similarly to the discrete notion of conservation of the Hamiltonian, the discrete notion of symplecticity holds for continuous problems but is a weaker notion.

Theorem 2.2.1 (The incompatibility between conserving the Hamiltonian and preserving the symplectic structure in a numerical method, [185])**.** *Let $\boldsymbol{u}^n$ be a discrete approximation to (2.3) for $n = 0, ..., N$, and assume that the Hamiltonian function $\mathcal{H}$ is higher than quadratic in order. If $\boldsymbol{u}^n$ conserves the Hamiltonian in the sense of (2.9) and preserves the symplectic structure (2.10) then $\boldsymbol{u}^n \equiv \boldsymbol{u}\left(t_n\right)$ for $n = 0, ..., N$. That is to say to preserve both the Hamiltonian and the symplectic mapping our numerical approximation* must *be exact.*

Remark 2.2.2 (Choosing which structure to conserve)**.** *As we found in Theorem 2.2.1, in general it is not possible to conserve both the Hamiltonian and the symplectic structure numerically. There has been much debate as to which invariant is more favourable to conserve, see for example [167]. While no conclusions can be drawn on this point which satisfy all in the field of geometric numerical integration, the preservation of the symplectic mapping appears to be favoured in the literature. This may be, in part, due to the following theorem.*

Theorem 2.2.3 (Deviation in the Hamiltonian for symplectic methods over long time, [28])**.** *Let $\boldsymbol{u}^n$ for $n = 0, ..., N$ be a symplectic numerical method, i.e., it satisfies (2.10), then,*

$$\mathcal{H}\left(\boldsymbol{u}^N\right) = \mathcal{H}\left(\boldsymbol{u}^0\right) + \mathcal{O}\left(\tau_{max}^2\right)$$

over an exponentially long time T, where $\tau_{max} := \max_n \tau_n$.

Remark 2.2.4 (Symmetry in Hamiltonian ODEs)**.** *In addition to conservation of the Hamiltonian and the symplectic structure many Hamiltonian ODEs are also symmetric, or time reversible. The symmetry of Hamiltonian ODEs is observed through the* flow map *of the problem, where the flow map describes how the solution changes over time. The notion of symmetry arises from the flow map being* time reversible*, that is to say that assuming the vector $\boldsymbol{u}$ flows forwards in time, if time is then reversed the solution will flow backwards in time along the same trajectory with the same velocity. We shall aim to design numerical methods which are symmetric in time.*
Let

$$\boldsymbol{u}^{n+1} = \Psi_{\tau_n}(\boldsymbol{u}^n) \tag{2.11}$$

define a one step numerical scheme, where Ψ_{τ_n} represents the discrete flow of the numerical approximation from $\boldsymbol{u}^n \to \boldsymbol{u}^{n+1}$, then the method is symmetric if

$$\Psi_{\tau_n}(\boldsymbol{u}^n) = \Psi_{-\tau_n}(\boldsymbol{u}^n)^{-1}. \tag{2.12}$$

We note that while the solution is vectorial the action of the flow map does not change the dimension of the solution. In the discrete setting for a one step method the inverse of the flow map can formally be found by interchanging $\boldsymbol{u}^{n+1}$ and $\boldsymbol{u}^n$ in (2.11). A method being symmetric is equivalent to saying that it is self-adjoined. Physically, a method being symmetric will give rise to symmetries of the solution in phase space when visualised over all time, which we expect to observe physically.

We shall now discuss particular classes of numerical methods and investigate their structure preserving properties.

2.2.1 Runge-Kutta methods

Runge-Kutta methods are an incredibly popular class of temporal discretisation, for both ODEs and as temporal discretisations of PDEs, see [111, 47, 160]. Their popularity is not only due to the ease of their implementation, or the concise formulation of the methods developed by Butcher, see [37]. They have been analysed extensively, and it is possible to construct stable methods of "arbitrary" order, see [39]. For a detailed history of Runge-Kutta methods see [38].

Definition 2.2.5 (Runge-Kutta method)**.** *Let a_{ij} and b_i for $i, j = 1, ..., s$ be real numbers. An s-stage Runge-Kutta method is given by*

$$\begin{aligned} \boldsymbol{u}^{n+1} =& \boldsymbol{u}^n + \tau_n \sum_{i=1}^{s} b_i \boldsymbol{p}^i \\ \boldsymbol{p}^i =& \boldsymbol{f}\left(\boldsymbol{u}^n + \tau_n \sum_{j=1}^{s} a_{ij} \boldsymbol{p}^j\right) \qquad \text{for } i = 1, ..., s. \end{aligned} \tag{2.13}$$

We shall now investigate the geometric properties it is possible to preserve by Runge-Kutta methods.

Theorem 2.2.6 (Condition for symplecticity in Runge-Kutta methods, [160])**.** *Let $\boldsymbol{u}$ describe a Hamiltonian ODE of the form (2.3), i.e., $\boldsymbol{f}(\boldsymbol{u}) = J\nabla\mathcal{H}(\boldsymbol{u})$. Further let a_{ij} and b_i be the coefficients of the Runge-Kutta method (2.13) for $i, j = 1, ..., s$, then if*

$$b_i a_{ij} + b_j a_{ji} - b_i b_j = 0,$$

then the method is symplectic, i.e.,

$$\mathbf{d}\boldsymbol{u}^{n+1} \wedge J^{-1}\, \mathbf{d}\boldsymbol{u}^{n+1} = \mathbf{d}\boldsymbol{u}^n \wedge J^{-1}\, \mathbf{d}\boldsymbol{u}^n.$$

Proof. Taking the differentials of (2.13) and substituting the differential of the numerical scheme in to the left hand side of (2.7) and utilising (2.4), (2.5) and (2.6) we obtain the desired result. The details of this proof can be found in [160].

□

Remark 2.2.7 (Hamiltonian conservation by Runge-Kutta methods)**.** *It was found in [40] that Runge-Kutta methods do not, in general, conserve Hamiltonians of a higher degree than quadratic. However, it is possible to construct a Runge-Kutta method which preserves a particular polynomial Hamiltonian function, see [42]. We shall not press this point here.*

Theorem 2.2.8 (Symmetry condition for Runge-Kutta methods, [179])**.** *Let a_{ij} and b_i for $i, j = 1, ..., s$ determine the Runge-Kutta method (2.13), then the method is symmetric if*

$$a_{s+1-i,s+1-j} + a_{ij} = b_j \qquad \forall i, j = 1, ..., s.$$

Example 2.2.9 (Gauss-Legendre methods: A symplectic family of Runge-Kutta methods)**.** *The Gauss-Legendre family of Runge-Kutta methods are obtained by seeking the*

constants c_i for $i = 1, ..., s$ such that c_i are given by the zeroes of

$$\frac{\mathrm{d}^s}{\mathrm{d}x^s}\left(x^s\left(x-1\right)^s\right).$$

Then the coefficients a_{ij} and b_i for $i, j = 1, ...s$ are determined by

$$a_{ij} = \int_0^{c_i} \mathcal{L}_j(t)\mathrm{d}t, \qquad b_i = \int_0^1 \mathcal{L}_i(t)\mathrm{d}t,$$

where $\mathcal{L}_i(t)$ is the Lagrange polynomial $\mathcal{L}_i(t) = \prod_{j\neq i} \frac{t-c_j}{c_i-c_j}$
Through fixing s we can observe through direct calculation that the condition for a method to be symplectic given in Theorem 2.2.6 is satisfied. For a proof of this for general s see [160].

2.2.2 Collocation methods

Collocation methods are a popular class of methods which can trace their origin back to [86], see [85, §II.1.2]. However, since the popularisation of the work of [84] they are not typically studied independently as they fall within the framework of the Runge-Kutta method. While this may appear to make the independent introduction of this class of methods superfluous we present them here as we utilise them in Chapter 3.

Definition 2.2.10 (Collocation method)**.** *Let c_i for $i = 1, ..., s$ be distinct real numbers, with $c_0 = 0$ and $c_s = 1$. The collocation polynomial $\boldsymbol{w}(t)$ is a degree s polynomial satisfying*

$$\boldsymbol{w}(t_n) = \boldsymbol{u}^n$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{w}\left(t_n + c_i\tau_n\right) = \boldsymbol{f}\left(\boldsymbol{w}\left(t_n + c_i\tau_n\right)\right).$$

The corresponding collocation method is given by

$$\boldsymbol{u}^{n+1} = \boldsymbol{w}\left(t_n + \tau_n\right). \tag{2.14}$$

Note that these methods do not solely consist of point values. In fact, in view of $\boldsymbol{w}$ we obtain a continuous piecewise polynomial solution, similarly to a finite element function.

Theorem 2.2.11 (Collocation methods as Runge-Kutta methods, [181])**.** *The collocation method (2.14) defined by the real constants $c_1, ..., c_s$ is equivalent to the s stage Runge-Kutta*

method (2.13) *with coefficients*

$$a_{ij} = \int_0^{c_i} \mathcal{L}_j(t)\mathrm{d}t, \qquad b_i = \int_0^1 \mathcal{L}_i(t)\mathrm{d}t,$$

where $\mathcal{L}_i(t) = \prod_{j\neq i} \frac{t-c_j}{c_i-c_j}$ *is the Lagrange polynomial.*

Remark 2.2.12 (Hamiltonian conservation of collocation methods)**.** *As collocation methods can be viewed as a subset of Runge-Kutta methods through Theorem 2.2.11 they* cannot *conserve an arbitrary Hamiltonian function over time, see Remark 2.2.7.*

Theorem 2.2.13 (Criteria for symmetric collocation methods)**.** *Let* c_i *for* $i = 1, ..., s$ *describe the collocation method* (2.14)*, then if* $c_i = 1 - c_{s+1-i}$ *then the method is* symmetric*.*

Proof. As no original reference could be found for this proof we shall present it here. Recall from Remark 2.2.4 that a numerical method is symmetric if it is invariant under reversal of time and inversion of the discrete flow map, see (2.12). Inverting the discrete flow map we can rewrite the collocation method in Definition 2.2.10 as

$$\begin{aligned} \boldsymbol{w}(t_{n+1}) &= \boldsymbol{u}^{n+1} \\ \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{w}\left(t_{n+1} + c_i\tau_n\right) &= \boldsymbol{f}\left(\boldsymbol{w}(t_{n+1} + c_i\tau_n)\right) \\ \boldsymbol{u}^n &= \boldsymbol{w}\left(t_{n+1} + \tau_n\right). \end{aligned}$$

Through reversing the flow of time, i.e., mapping $\tau_n \to -\tau_n$, we observe that the collocation method is preserved if and only if $c_i = 1 - c_{s+1-i}$.

□

2.2.3 Discrete gradient methods

Discrete gradient methods, as their name suggests, are numerical methods designed such that they preserve a discrete Hamiltonian, see [137, 139]. Such methods cannot be presented in one class as concisely as Runge-Kutta and collocation methods. We can define a discrete gradient method for a Hamiltonian ODE as follows.

Definition 2.2.14 (Discrete gradient method)**.** *Let* $\boldsymbol{u}$ *be the solution to the Hamiltonian problem* (2.3)*, then a discrete gradient method is of the form*

$$\boldsymbol{u}^{n+1} = \boldsymbol{u}^n + \tau_n J\widetilde{\nabla}\boldsymbol{K}, \tag{2.15}$$

where $\widetilde{\nabla}\boldsymbol{K}$ *denotes a discrete gradient, approximating the continuous function* $\nabla\mathcal{H}(\boldsymbol{u})$*, which satisfies the identity*

$$\widetilde{\nabla}\boldsymbol{K}\cdot\left(\boldsymbol{u}^{n+1}-\boldsymbol{u}^{n}\right)=\mathcal{H}\left(\boldsymbol{u}^{n+1}\right)-\mathcal{H}\left(\boldsymbol{u}^{n}\right). \tag{2.16}$$

Theorem 2.2.15 (Hamiltonian conservation of discrete gradient methods)**.** *Let* $\boldsymbol{u}^{n}$ *for* $n=0,...,N$ *be described by the discrete gradient method* (2.15) *for a Hamiltonian problem* (2.3)*, then the Hamiltonian function is conserved nodally, i.e.,*

$$\mathcal{H}\left(\boldsymbol{u}^{n+1}\right)=\mathcal{H}\left(\boldsymbol{u}^{n}\right).$$

Proof. Through application of (2.16) and (2.15) we see that

$$\begin{aligned}\mathcal{H}\left(\boldsymbol{u}^{n+1}\right)-\mathcal{H}\left(\boldsymbol{u}^{n}\right)&=\widetilde{\nabla}\boldsymbol{K}\cdot\left(\boldsymbol{u}^{n+1}-\boldsymbol{u}^{n}\right)\\&=\tau_n\widetilde{\nabla}\boldsymbol{K}\cdot J\widetilde{\nabla}\boldsymbol{K}\\&=0,\end{aligned}$$

as J is skew-symmetric.

□

Remark 2.2.16 (Preservation of a symplectic structure)**.** *As discrete gradient methods are energy conserving in nature they cannot preserve a discrete symplectic structure by Theorem 2.2.1. For numerical methods which preserve a discrete symplectic map we know that numerical solution remains close to conserving energy, see Theorem 2.2.3. Unfortunately, there are no analytic results stating that the converse is true in the literature.*

There are three main families of discrete gradient methods which are introduced in [81], [89] and [103]. Here we introduce a member of the family discussed in [81] as an illustrative example.

Example 2.2.17 (An example of a discrete gradient method)**.** *Let* $\boldsymbol{u}=(u_1,u_2,...,u_D)$ *describe a Hamiltonian problem of the form* (2.3)*, then a discrete gradient method is given by*

$$u_i^{n+1}=u_i^{n}+\tau_n J\frac{\nabla\mathcal{H}\left(\boldsymbol{u}^{n+1}\right)_i-\nabla\mathcal{H}\left(\boldsymbol{u}^{n}\right)_i}{u_i^{n+1}-u_i^{n}}, \tag{2.17}$$

for $i=1,...,D$*. Clearly, this discrete gradient satisfies* (2.16)*. Additionally,* (2.17) *is symmetric, as interchanging* $\boldsymbol{u}^{n+1}$ *and* $\boldsymbol{u}^{n}$ *then mapping* $\tau_n\to-\tau_n$ *returns the original method.*

2.2.4 Composition methods

Composition methods are *not* a class of method in their own right. A composition method refers to any numerical method in which the flow of the numerical solution over a single step is composed of multiple numerical schemes. Such methods have very useful properties in geometric integration, as well as improved stability, see [101]. It is possible to compose methods which do not preserve geometric properties as a methodology of constructing geometric numerical integrators. Here we focus on composition methods which already possess geometric properties in view of the following remark.

Remark 2.2.18 (Composition of geometric integrators)**.** *Let Ψ_{τ_n} describe the numerical flow of $\boldsymbol{u}^n$ and $\gamma_1, \gamma_2 > 0$. If the numerical flow either conserves a symplectic mapping or the Hamiltonian, then the composition of this flow $\Psi_{\gamma_1\tau_n} \circ \Psi_{\gamma_2\tau_n}$ must also conserve the same symplectic mapping or Hamiltonian due to the local nature of the discrete geometric structure.*

It is possible, through composition, to *increase* the order of a numerical method as has been developed in [184, 136].

Theorem 2.2.19 (Minimum order of a composition method, [85, §II.4])**.** *Consider the composition method*

$$\boldsymbol{u}^{n+1} = \Psi_{\gamma_s\tau_n} \circ \cdots \circ \Psi_{\gamma_1\tau_n} \circ \boldsymbol{u}^n, \tag{2.18}$$

where $\Psi_{\tau_n}(\boldsymbol{u}^n)$ describes the numerical flow of an order p method and $2 < s \in \mathbb{Z}$. If

$$\begin{aligned} \gamma_1 + \cdots + \gamma_s &= 1 \\ \gamma_1^{p+1} + \ldots \gamma_s^{p+1} &= 0, \end{aligned} \tag{2.19}$$

then the composition method (2.18) is at least order $p+1$. Further, if $\Psi_{\tau_n}(\boldsymbol{u}^n)$ is symmetric, and

$$\gamma_{s+1-i} = \gamma_i, \tag{2.20}$$

for $i = 1, ..., s$ then the composition (2.18) is at least order $p+2$.

Remark 2.2.20 (Construction of arbitrarily high degree geometric integrators)**.** *In view of Theorem 2.2.19 and Remark 2.2.18 we can construct arbitrarily high order geometric numerical integrators. As well as preserving either a symplectic mapping or the Hamiltonian function it is also important to preserve symmetry if the method we compose with is*

itself symmetric, to preserve the symmetry we need to satisfy (2.20)*. The minimum number of times we need to compose a method with itself to preserve symmetry and increase the order of accuracy is* 3*, as we have three constraints in* (2.19) *and* (2.20)*, so we can choose* $s = 3$*. The parameters* γ_i *as discussed in Theorem 2.2.19 are then given uniquely by*

$$\gamma_1 = \gamma_3 = \frac{1}{3}2^{\frac{1}{3}} + \frac{1}{6}2^{\frac{2}{3}} + \frac{2}{3}, \qquad \gamma_2 = -\frac{2}{3}2^{\frac{1}{3}} - \frac{1}{3}2^{\frac{2}{3}} - \frac{1}{3}.$$

While we shall not investigate composition methods in the sequel they provide us with a powerful tool for generalising the temporal methods we develop from Chapter 5 onwards. We shall investigate multiple nonlinear PDEs and design fully discrete numerical schemes which conserve physical invariants of the problems. In space our numerical methods can be of arbitrarily high order, however the coupled temporal schemes shall always be $\mathcal{O}(\tau_{max}^2)$. The methodology outlined here allows us to extend the temporal discretisation to be arbitrarily high in order, although this comes at significant computational cost.

2.3 Conclusion

We gave a brief overview of some of the literature in the area of geometric numerical integration. We introduced the concepts of preservation of symplectic mapping and conservation of energy for Hamiltonian ODEs, and outlined selected numerical schemes which preserve these properties. We also outlined a methodology for the construction of higher order geometric numerical integrators from lower order integrators.

Chapter 3

Finite element methods for ODEs

Here we discuss finite element discretisations for ODEs, both recalling methods from the literature and developing a new finite element method. In the spirit of the previous chapter we shall investigate geometric properties of these methods, similarly to the investigation conducted in [105].

We also investigate the stability and convergence properties here, primarily to gain insight into the properties of our newly developed method.

In this chapter we focus on two main types of ODE, both encompassed by (2.1). The first are *explicit nonlinear functions* of $\boldsymbol{u}$, i.e., $\boldsymbol{F}(\boldsymbol{u}, t) = \boldsymbol{f}(\boldsymbol{u})$. This is a very large class of problems which include Hamiltonian systems, i.e., $\boldsymbol{f}(\boldsymbol{u}) = J\nabla\mathcal{H}(\boldsymbol{u})$ where J is a constant skew-symmetric matrix, and $\mathcal{H}(\boldsymbol{u})$ is a scalar function which physically corresponds to the energy of the system, see Definition 2.1.1. These problems have interesting physical properties, such as conservation of the Hamiltonian function, and preservation of the symplectic structure of the flow map, but due to their skew-symmetric structure they are difficult to analyse. The second class of problems we consider are *linear ODEs with forcing*, i.e., $\boldsymbol{F}(\boldsymbol{u}, t) = \mathfrak{f}(t) - \mathfrak{A}\boldsymbol{u}$, where $\mathfrak{f}(t)$ represents the forcing term and $\mathfrak{A}$ is the sum of a symmetric linear operator A and a skew-symmetric linear operator B, i.e., $\mathfrak{A} = c_1 A + c_2 B$ with c_1 and c_2 known constants. This class of problems is mainly utilised to describe the temporal aspect of evolution PDEs with a combination of symmetric and skew-symmetric operators in space. Here $\mathfrak{A}$ can either represent a continuous spatial operator or an appropriate discretisation of a spatial operator. Note that when required we will divide this case into two sub-cases, when the problem is purely *symmetric*, i.e., when $c_2 = 0$, and when the problem is purely *skew-symmetric*, i.e., when $c_1 = 0$. The symmetric case treats equations such as the heat equation, and the skew-symmetric case

can treat equations such as the wave equation, linear KdV, or even a linear Hamiltonian ODE such as the harmonic oscillator.

3.1 Known methods and their geometric properties

The ideas in this section have been presented in the masters thesis [105], we recall them here for completeness in addition to them informing our study of temporal finite element methods later in this chapter. All results from [105] are referenced appropriately.
Before introducing temporal finite element methods we must first introduce an abundance of notation. Recall from §2.2 that we partitioned our temporal interval $[0,T]$ such that $0 := t_0 < t_1 < \cdots < t_N =: T$. We define a *temporal finite element* as $I_n := (t_n, t_{n+1})$ which possesses an element length $\tau_n := t_{n+1} - t_n$. We shall often write $\max_n \tau_n = \tau_{max}$. Throughout this section we shall write the discrete counterpart to a continuous function through capitalisation, i.e., for a temporal continuous function $\boldsymbol{u} = \boldsymbol{u}(t)$ we write the finite element function approximating it as $\boldsymbol{U} = \boldsymbol{U}(t)$. Note that when there is no ambiguity we shall not explicitly write the dependency of functions. In the sequel, when considering spatial finite element methods, we shall also use capitalisation to refer to *spatial* finite element functions.

Definition 3.1.1 (Temporal finite element spaces)**.** *Let $\mathbb{P}_q(I_n)$ denote the space of polynomials of degree q on an interval $I_n \subset \mathbb{R}$, then the discontinuous finite element space is*

$$\mathbb{V}_q\left([0,T]\right) = \{\boldsymbol{W} : \boldsymbol{W}|_{I_n} \in \left(\mathbb{P}_q(I_n)\right)^D, n = 0, ..., N-1\},$$

further to this the continuous finite element space is defined analogously with global continuity enforced, i.e.,

$$\mathbb{V}_q^C\left([0,T]\right) = \mathbb{V}_q\left([0,T]\right) \cap \left(\mathcal{C}^0\left([0,T]\right)\right)^D.$$

When there is no ambiguity we shall not explicitly write the domain of our finite element spaces. Additionally, in the sequel $\mathbb{V}_q^C\left(I_n\right)$ represents the localisation of the continuous finite element space to a single element. Here, the initial conditions of functions in this space are fixed by the endpoint of functions in $\mathbb{V}_q^C\left(I_{n-1}\right)$, or an appropriate initial condition.

With the definition of the temporal finite element space in mind, we define the mesh function $\tau \in \mathbb{V}_0$ as the piecewise constant finite element function representing the length

of an element, i.e.,

$$\tau|_{I_n} = \tau_n.$$

Remark 3.1.2 (A 1D finite element de Rham complex between spaces,[13])**.** *A de Rham complex, in the general setting, refers to a complex of differential forms acting on a smooth manifold which map all functions in a given space to all functions in a subsequent space through an appropriate exterior derivative. In fact, as outlined in [13], there are discrete counterparts to these continuous mappings in one, two and three dimensional spaces. In the one dimensional case the de Rham mapping is comparatively straightforward. The de Rham complex maps functions in the continuous finite element space* $\mathbb{V}_{q+1}^C$ *to functions in the discontinuous finite element space of one degree lower* $\mathbb{V}_q$*. The exterior derivative which employs this mapping is the standard derivative operator* $\frac{\mathrm{d}}{\mathrm{d}t}$*.*

The first temporal finite element we shall investigate is the continuous, or conforming, Galerkin method. This method has been extensively analysed, see [67, 73].

Definition 3.1.3 (Continuous Galerkin method)**.** *Let* $\mathbb{V}_{q+1}^C$ *and* $\mathbb{V}_q$ *be the vectorial finite dimensional spaces given in Definition 3.1.1. The continuous Galerkin* (cG) *finite element approximation is given by seeking* $\boldsymbol{U} \in \mathbb{V}_{q+1}^C$ *such that*

$$\begin{aligned} \int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \cdot \boldsymbol{V}\,\mathrm{d}t = \int_0^T \boldsymbol{F}\left(\boldsymbol{U}, t\right) \cdot \boldsymbol{V}\,\mathrm{d}t \qquad \forall \boldsymbol{V} \in \mathbb{V}_q, \\ \boldsymbol{U}(0) = \boldsymbol{u}_0. \end{aligned} \tag{3.1}$$

Remark 3.1.4 (Difference between test and trial spaces for the cG method)**.** *For readers with a background in spatial finite element methods continuous Galerkin may initially seem like a misnomer as it utilises a discontinuous test space of one degree lower, so is a Petrov-Galerkin type method. In fact this is the natural continuous finite element formulation due to the lack of symmetry in derivatives, and the degrees of freedom on the left hand side of (3.1) match as* $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}, \boldsymbol{V} \in \mathbb{V}_q$*. This follows from the discrete de Rham complex discussed in Remark 3.1.2.*

Remark 3.1.5 (cG time stepping)**.** *The discontinuous nature of the test functions allow us to rewrite the cG method over a single element, with the initial value of the trial function* $\boldsymbol{U}$ *on the element being the final value of the trial function on the previous element due*

to continuity. Explicitly by choosing the test function

$$\boldsymbol{V}(t) = \begin{cases} \widetilde{\boldsymbol{V}}(t) & \text{for } t \in I_n \\ 0 & \text{else,} \end{cases}$$

where $\widetilde{\boldsymbol{V}}$ *is an arbitrary function in* $\left(\mathbb{P}_q\left(I_n\right)\right)^D$*, over an arbitrary interval, we can write the cG method locally as follows: Let* $\boldsymbol{U}(t_n)$ *be given, then seek* $\boldsymbol{U} \in \mathbb{V}_{q+1}^C\left(I_n\right)$ *such that*

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \cdot \widetilde{\boldsymbol{V}} \mathrm{d}t = \int_{I_n} \boldsymbol{F}\left(\boldsymbol{U}, t\right) \cdot \widetilde{\boldsymbol{V}} \mathrm{d}t \qquad \forall \widetilde{\boldsymbol{V}} \in \left(\mathbb{P}_q\left(I_n\right)\right)^D. \tag{3.2}$$

Theorem 3.1.6 (Preservation of the Hamiltonian function for the cG method, [105, 87])**.** *Let* $\boldsymbol{F}\left(\boldsymbol{U}, t\right) = J\nabla\mathcal{H}\left(\boldsymbol{U}\right)$ *where* J *is constant, i.e., we restrict ourselves to the case of a Hamiltonian system, as described in Definition 2.1.1, then the cG method described in Definition 3.1.3 preserves the Hamiltonian function at the nodes. That is to say that for* $n = 0, ..., N-1$

$$\mathcal{H}\left(\boldsymbol{U}(t_{n+1})\right) = \mathcal{H}\left(\boldsymbol{U}(t_n)\right).$$

Proof. Through the fundamental theorem of calculus we observe that

$$\mathcal{H}\left(\boldsymbol{U}(t_{n+1})\right) - \mathcal{H}\left(\boldsymbol{U}(t_n)\right) = \int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\left(\mathcal{H}\left(\boldsymbol{U}\right)\right) \mathrm{d}t = \int_{I_n} \nabla\mathcal{H}\left(\boldsymbol{U}\right) \cdot \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \mathrm{d}t.$$

Additionally, through choosing $\widetilde{\boldsymbol{V}} = \Pi_{\mathbb{V}_q}\left(\nabla\mathcal{H}\left(\boldsymbol{U}\right)\right)$, where $\Pi_{\mathbb{V}_q}$ is the L_2 projection into $\mathbb{V}_q$, in the local cG method (3.2) we find

$$\begin{aligned} \int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \cdot \nabla\mathcal{H}\left(\boldsymbol{U}\right) \mathrm{d}t &= \int_{I_n} J\nabla\mathcal{H}\left(\boldsymbol{U}\right) \cdot \Pi_{\mathbb{V}_q}\left(\nabla\mathcal{H}\left(\boldsymbol{U}\right)\right) \mathrm{d}t \\ &= \int_{I_n} J\Pi_{\mathbb{V}_q}\left(\nabla\mathcal{H}\left(\boldsymbol{U}\right)\right) \cdot \Pi_{\mathbb{V}_q}\left(\nabla\mathcal{H}\left(\boldsymbol{U}\right)\right) \mathrm{d}t \end{aligned}$$

through the definition of the L_2 projection, and as J is constant. In view of the skew symmetry of J we can write

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \cdot \nabla\mathcal{H}\left(\boldsymbol{U}\right) \mathrm{d}t = 0,$$

allowing us to conclude.

□

Remark 3.1.7 ("Exact" conservation of the Hamiltonian)**.** *While the cG method exactly*

preserves the Hamiltonian function this is not necessarily true for an implementation of the method. To implement the method we are required to make a quadrature approximation, and if this quadrature approximation is not *exact, then the Hamiltonian will not be exactly preserved. In practice, we will only be able to preserve the Hamiltonian exactly if our Hamiltonian function is polynomial and the quadrature method is chosen to be of high enough order. When dealing with non-polynomial Hamiltonian functions it is also possible to take an "overkill" approach, i.e., we can take a very high order quadrature approximation so the quadrature error is of the same order of magnitude as machine precision.*

Remark 3.1.8 (A comparison between the cG method and discrete gradient methods)**.** *As we discussed in §2.2.3, a discrete gradient method (2.15) satisfies the identity*

$$\widetilde{\nabla}\boldsymbol{K}\cdot\left(\boldsymbol{u}^{n+1}-\boldsymbol{u}^{n}\right)=\mathcal{H}\left(\boldsymbol{u}^{n+1}\right)-\mathcal{H}\left(\boldsymbol{u}^{n}\right),$$

for some discrete gradient $\widetilde{\nabla}\boldsymbol{K}$. The cG method (3.2), while not strictly falling within this framework, satisfies the similar identity in the continuous framework

$$\int_{I_n}\nabla\mathcal{H}\left(\boldsymbol{U}\right)\cdot\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\mathrm{d}t=\mathcal{H}\left(\boldsymbol{U}(t_{n+1})\right)-\mathcal{H}\left(\boldsymbol{U}(t_n)\right),$$

where $\boldsymbol{U}$ is the cG solution, which is crucial in the proof of Theorem 3.1.6.

Theorem 3.1.9 (cG as a Runge-Kutta method, [105])**.** *Let $\boldsymbol{F}\left(\boldsymbol{U},t\right)=\boldsymbol{f}\left(\boldsymbol{U}\right)$ be an explicit function of $\boldsymbol{U}$, then under a $q+1$ point quadrature choice the cG scheme, (3.1), agrees with a Runge-Kutta method, see Definition 2.2.5, at the nodes. Furthermore this Runge-Kutta method is given by*

$$\begin{aligned}b_i&=\int_0^1\mathcal{L}_i(t)\mathrm{d}t\\ a_{ij}&=\int_0^{c_i}\mathcal{L}_j(t)\mathrm{d}t\end{aligned}\tag{3.3}$$

for $i,j=1,...,q+1$, where $c_i=\sum_{j=1}^{q+1}a_{ij}$ are our quadrature points and $\mathcal{L}_i=\prod_{j\neq i}\frac{t-c_j}{c_i-c_j}$ is a Lagrange polynomial.

Proof. Under a $q+1$ point quadrature we can write the cG method, (3.1), as

$$\begin{aligned}\sum_{n=0}^{N-1}\sum_{iq=1}^{q+1}c_{iq}\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}(t_{n,iq})\cdot\boldsymbol{V}(t_{n,iq})&=\sum_{n=0}^{N-1}\sum_{iq=1}^{q+1}c_{iq}\boldsymbol{f}\left(\boldsymbol{U}(t_{n,iq})\right)\cdot\boldsymbol{V}(t_{n,iq})\\ \boldsymbol{U}(0)&=\boldsymbol{u}_0.\end{aligned}\tag{3.4}$$

As $\boldsymbol{V} \in \mathbb{V}_q$ we can choose

$$\boldsymbol{V} = \begin{cases} 1 & t = t_{n,iq} \\ 0 & \text{all other quadrature points,} \end{cases}$$

for an arbitrary quadrature point iq in an arbitrary interval I_n. Under this choice of test function we can write (3.4) as

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}(t_{n,iq}) = \boldsymbol{f}\left(\boldsymbol{U}(t_{n,iq})\right). \tag{3.5}$$

As (3.5) is for an arbitrary quadrature point on an arbitrary interval we can conclude that $\boldsymbol{U}$ is a collocation polynomial, see Definition 2.2.10. This implies that at the nodes our cG formulation agrees with a collocation method defined by the quadrature points. This collocation method is equivalent to the Runge-Kutta method (3.3) by Theorem 2.2.11. □

Remark 3.1.10 (Symplectic implementation of cG, [105])**.** *Applying particular quadrature choices (which introduce leading order errors) we obtain a scheme equivalent on the nodes to families of symplectic methods. For example choosing the $q+1$ point Gauss quadrature, see [85, P34], the cG method (3.1) agrees with the Gauss Runge-Kutta family on the nodes. These methods are known to be symplectic, as they satisfy Theorem 2.2.6.*

We now shift our focus to a discontinuous Galerkin method often employed for temporal ODEs, which has also been extensively studied in the literature, see [113, 66, 68]. Before we can introduce this method we first must define additional notation.

Definition 3.1.11 (Discontinuous finite element notation)**.** *Due to the discontinuous nature of the finite element space finite element functions are permitted to be multi-valued at the nodes of the elements. With this in mind we write*

$$\boldsymbol{U}_n^+ := \boldsymbol{U}(t_n^+) := \lim_{t \searrow t_n} \boldsymbol{U}(t), \qquad \boldsymbol{U}_n^- := \boldsymbol{U}(t_n^-) := \lim_{t \nearrow t_n} \boldsymbol{U}(t),$$

to describe the values of the function on the right and left of a discontinuity respectively. Additionally we define the jump of a function at the node t_n to be

$$[\![\boldsymbol{U}_n]\!] = \boldsymbol{U}_n^- - \boldsymbol{U}_n^+ \tag{3.6}$$

and the average as

$$\{\boldsymbol{U}_n\} = \frac{1}{2}\left(\boldsymbol{U}_n^- + \boldsymbol{U}_n^+\right). \tag{3.7}$$

Definition 3.1.12 (Upwind discontinuous Galerkin method)**.** *Let $\mathbb{V}_q$ be the space of piecewise polynomial functions of degree q as described in Definition 3.1.1. The upwind discontinuous Galerkin* (dG) *approximation is given by seeking $\boldsymbol{U} \in \mathbb{V}_q$ such that*

$$\begin{aligned} \sum_{n=0}^{N-1} \int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \cdot \boldsymbol{V} \mathrm{d}t = \sum_{n=0}^{N-1} \int_{I_n} \boldsymbol{f}\left(\boldsymbol{U}, \boldsymbol{V}\right) \mathrm{d}t + [\![\boldsymbol{U}_n]\!] \cdot \boldsymbol{V}_n^+ \qquad \forall \boldsymbol{V} \in \mathbb{V}_q \\ \boldsymbol{U}^-(0) = \boldsymbol{u}_0. \end{aligned} \tag{3.8}$$

Remark 3.1.13 (dG time stepping)**.** *Similarly to the cG method, the upwind dG method can be implemented in a time-stepping fashion utilising the discontinuous nature of the test function. To be concise, we can reduce the upwind scheme to a single element by choosing our test function such that*

$$\boldsymbol{V}(t) = \begin{cases} \widetilde{\boldsymbol{V}}(t) & \text{for } t \in I_n \\ 0 & \text{else,} \end{cases}$$

where $\widetilde{V}$ is an arbitrary function in $\left(\mathbb{P}_q\left(I_n\right)\right)^D$ allowing us to write

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \cdot \widetilde{\boldsymbol{V}} \mathrm{d}t = \int_{I_n} \boldsymbol{f}\left(\boldsymbol{U}, \widetilde{\boldsymbol{V}}\right) \mathrm{d}t + [\![\boldsymbol{U}_n]\!] \cdot \widetilde{\boldsymbol{V}}_n^+ \qquad \forall \widetilde{\boldsymbol{V}} \in \left(\mathbb{P}_q\left(I_n\right)\right)^D. \tag{3.9}$$

Note that we cannot typically write dG methods in a time stepping fashion. The choice of flux needs to be chosen carefully such that on an arbitrary element it only depends on the value of the solution on the previous element. Many standard flux choices for spatial finite element methods require information from both neighbouring elements and the methods need to be solved globally. These global methods are not practical temporal discretisations.

Remark 3.1.14 (Dissipation of the upwind dG method)**.** *In [68] the upwind dG method* (3.9) *is viewed as a* discrete dynamical system *for problems with a linear right hand side. It is found that the associated dynamical system is dissipative in the sense that the numerical solution over time can not increase. In addition, we will discover in the sequel, if the right hand side is chosen to describe a purely skew-symmetric problem then the upwind dG method is dissipative in the same sense that the solution over time does not increase, see Theorem 3.1.22.*

Due to the dissipative nature of (3.9) it does not *fall within the framework of geometric numerical integration when applied to Hamiltonian systems.*

Error bounds for both the cG and upwind dG methods have been fully developed in the literature for linear problems, these bounds can be described as follows.

Theorem 3.1.15 (A priori error bound for the cG method (3.1), [67]). *Let $\boldsymbol{U}$ be the solution of the cG method (3.1), where $\boldsymbol{u} \in C^{q+2}([0,T])$ denotes the corresponding exact solution given in (2.1). Under the assumption that $\boldsymbol{F}(\boldsymbol{U},t)$ is linear we have for $q = 0, 1$ that*

$$\sup_{[0,T]} |\boldsymbol{U} - \boldsymbol{u}| \leq C(T)\tau_{max}^{q+2} \sup_{[0,T]} \left| \frac{\mathrm{d}^{q+2}}{\mathrm{d}t^{q+2}} \boldsymbol{u} \right|,$$

where the constant $C(T)$ depends exponentially on T.

Theorem 3.1.16 (A priori error bound for the upwind dG method (3.8), [113]). *Let $\boldsymbol{U}$ be the solution of the upwind dG method (3.8), where $\boldsymbol{u} \in C^{q+1}([0,T])$ represents the corresponding exact solution given in (2.1). Under the assumption that $\boldsymbol{F}(\boldsymbol{U},t)$ is linear we have for $q = 0, 1$ that*

$$\sup_{[0,T]} |\boldsymbol{U} - \boldsymbol{u}| \leq C(T)\tau_{max}^{q+1} \sup_{[0,T]} \left| \frac{\mathrm{d}^{q+1}}{\mathrm{d}t^{q+1}} \boldsymbol{u} \right|,$$

where the constant $C(T)$ depends exponentially on T.

Note that while for the cG method (3.1) we seek a polynomial solution of degree $q+1$, for the upwind dG method (3.8) we seek a polynomial solution of degree q, ergo both schemes converge at the same rate.
While these error bounds are well established, we selectively develop equivalent bounds in the sequel as they inform the analysis of the new finite element method we propose in §3.2.

3.1.1 Stability

Here we investigate the stability properties of the temporal finite element methods presented in this section. Note that while similar results are often cited in the literature they are rarely presented in full. We consider the stability of symmetric linear ODEs, skew-symmetric linear ODEs, and Hamiltonian ODEs independently.

3.1.1.1 Symmetric linear ODEs

In this subsection we restrict ourselves to *symmetric* linear ODEs, that is we choose the right hand side of the ODE (2.1) such that $\boldsymbol{F}(\boldsymbol{u}, t) = \mathfrak{f} - A\boldsymbol{u}$, where $\mathfrak{f} = \mathfrak{f}(t)$ and A is a symmetric linear operator. The operator A represents a symmetric spatial operator, which can be either continuous in space or discrete. An example of an appropriate continuous spatial operator could be the second spatial derivative, and a discrete operator could be a symmetric finite element discretisation. Physically $\mathfrak{f}$ represents the forcing of the system. Our symmetric linear ODE can be written in the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u} + A\boldsymbol{u} = \mathfrak{f}.$$

We further assume that this operator is symmetric and consequently induces a norm, in addition to this at various points in the sequel we will assume that the operator A induces a bilinear form which is coercive, i.e., for $\boldsymbol{w} \in \mathbb{R}^D$

$$A\boldsymbol{w} \cdot \boldsymbol{w} \geq C_A \left|\boldsymbol{w}\right|_A^2 \tag{3.10}$$

and/or continuous, i.e., for $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^D$

$$\left|A\boldsymbol{w}_1 \cdot \boldsymbol{w}_2\right| \leq c_A \left|\boldsymbol{w}_1\right|_A \left|\boldsymbol{w}_2\right|_A, \tag{3.11}$$

where $\left|\cdot\right|_A$ represents some appropriate semi-norm depending on A. For example, A could represent, or approximate, the second order spatial Laplacian

$$A = -\triangle.$$

In this case the semi-norm $\left|\cdot\right|_A$ is in fact the spatial H^1 norm, see for example [33]. If A is a matrix then the coercivity and continuity conditions correspond to A being nondegenerate and bounded.

Theorem 3.1.17 (Stability of the cG method for symmetric problems)**.** *Consider the cG method (3.1) applied to a symmetric linear ODE, i.e., $\boldsymbol{F}(\boldsymbol{U}, t) = \mathfrak{f} - A\boldsymbol{U}$ where $\mathfrak{f}$ is bounded and A is a symmetric operator. The solution to the cG method satisfies*

$$\left\|\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\right\|_{L_2([0,T])}^2 + A\boldsymbol{U}(T) \cdot \boldsymbol{U}(T) \leq A\boldsymbol{U}(0) \cdot \boldsymbol{U}(0) + \|\mathfrak{f}\|_{L_2([0,T])}^2.$$

Remark 3.1.18 (Stability in multiple norms)**.** *Under the assumption that A induces a coercive and continuous form, see (3.10) and (3.11), we can obtain the following stability bound*

$$\left\|\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\right\|^2_{L_2([0,T])} \leq c_A \left|\boldsymbol{U}(0)\right|^2_A + \|\mathfrak{f}\|^2_{L_2([0,T])}\,.$$

Additionally we have stability in the norm induced by A, i.e.,

$$|\boldsymbol{U}(T)|^2_A \leq |\boldsymbol{U}(0)|^2_A + \|\mathfrak{f}\|^2_{L_2([0,T])}\,.$$

Proof of Theorem 3.1.17. For a symmetric linear problem we may write the cG method as follows: seek $\boldsymbol{U} \in \mathbb{V}^C_{q+1}$ such that

$$\begin{aligned}\int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\cdot\boldsymbol{V}\,\mathrm{d}t + \int_0^T A\boldsymbol{U}\cdot\boldsymbol{V}\,\mathrm{d}t = \int_0^T \mathfrak{f}\cdot\boldsymbol{V}\,\mathrm{d}t \qquad \forall \boldsymbol{V}\in\mathbb{V}_q\\ \boldsymbol{U}(0) = \boldsymbol{u}_0.\end{aligned}$$

Choosing $\boldsymbol{V} = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}$ we have

$$\int_0^T \left|\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\right|^2 + A\boldsymbol{U}\cdot\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\,\mathrm{d}t = \int_0^T \mathfrak{f}\cdot\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\,\mathrm{d}t,$$

which, after using the symmetry of A, can be written as

$$\begin{aligned}\int_0^T \left|\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\right|^2 + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left(A\boldsymbol{U}\cdot\boldsymbol{U}\right)\mathrm{d}t &= \int_0^T \mathfrak{f}\cdot\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\,\mathrm{d}t\\ &\leq \|\mathfrak{f}\|_{L_2([0,T])}\left\|\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\right\|_{L_2([0,T])}\\ &\leq \frac{1}{4\epsilon}\|\mathfrak{f}\|^2_{L_2([0,T])} + \epsilon\left\|\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\right\|^2_{L_2([0,T])},\end{aligned}$$

through Hölder's inequality and Cauchy's inequality with ϵ. Choosing $\epsilon = \frac{1}{2}$ and applying the fundamental theorem of calculus

$$\left\|\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\right\|^2_{L_2([0,T])} + A\boldsymbol{U}(T)\cdot\boldsymbol{U}(T) \leq A\boldsymbol{U}(0)\cdot\boldsymbol{U}(0) + \|\mathfrak{f}\|^2_{L_2([0,T])}\,,$$

as required.

□

Theorem 3.1.19 (Stability of the upwind dG method for symmetric problems)**.** *Let* $\boldsymbol{U}$ *be the solution of the upwind dG method* (3.8)*. Further assume that* $f(\boldsymbol{U}) = \mathfrak{f} - A\boldsymbol{U}$*, where* $\mathfrak{f} := \mathfrak{f}(t)$ *is bounded and* A *is symmetric. Additionally assume that* A *is a coercive operator, i.e., for all* $\boldsymbol{W} \in \mathbb{V}_q$

$$\|\boldsymbol{W}\|^2_{L_2(I_n)} \leq C_A \int_{I_n} A\boldsymbol{W} \cdot \boldsymbol{W} \mathrm{d}t,$$

where we have assumed that $|\boldsymbol{W}| \leq C |\boldsymbol{W}|_A$*, then*

$$\left|\boldsymbol{U}^-_{n+1}\right|^2 + \int_0^T A\boldsymbol{U} \cdot \boldsymbol{U} \mathrm{d}t \leq |\boldsymbol{u}_0|^2 + \frac{C_A}{2} \|\mathfrak{f}\|^2_{L_2((0,T))},$$

where $\boldsymbol{u}_0$ *denotes the initial data.*

Before presenting the proof of Theorem 3.1.19 we first require the following result.

Lemma 3.1.20 (A useful dG inequality)**.** *Let* $\boldsymbol{W} \in \mathbb{V}_q$*, then*

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{W} \cdot \boldsymbol{W} \mathrm{d}t + [\![\boldsymbol{W}_n]\!] \cdot \boldsymbol{W}^+_n \geq \frac{1}{2}\left|\boldsymbol{W}^-_{n+1}\right|^2 - \frac{1}{2}\left|\boldsymbol{W}^-_n\right|^2.$$

Proof. Applying the definition of the jump and the fundamental theorem of calculus we find

$$\begin{aligned}\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{W} \cdot \boldsymbol{W} \mathrm{d}t + [\![\boldsymbol{W}_n]\!] \cdot \boldsymbol{W}^+_n &= \int_{I_n} \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|\boldsymbol{W}|^2 \mathrm{d}t + \left(\boldsymbol{W}^+_n - \boldsymbol{W}^-_n\right) \cdot \boldsymbol{W}^+_n \\ &= \frac{1}{2}\left(\left|\boldsymbol{W}^-_{n+1}\right|^2 - \left|\boldsymbol{W}^+_n\right|^2\right) + \left|\boldsymbol{W}^+_n\right|^2 - \boldsymbol{W}^-_n \cdot \boldsymbol{W}^+_n.\end{aligned}$$

Through Cauchy's inequality we have that

$$-\left|\boldsymbol{W}^-_n \cdot \boldsymbol{W}^+_n\right| \geq -\frac{1}{2}\left|\boldsymbol{W}^-_n\right|^2 - \frac{1}{2}\left|\boldsymbol{W}^+_n\right|^2,$$

which tells us

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{W} \cdot \boldsymbol{W} \mathrm{d}t + [\![\boldsymbol{W}_n]\!] \cdot \boldsymbol{W}^+_n \geq \frac{1}{2}\left|\boldsymbol{W}^-_{n+1}\right|^2 - \frac{1}{2}\left|\boldsymbol{W}^-_n\right|^2,$$

as required.

□

Proof of Theorem 3.1.19. Choosing $\boldsymbol{V} = \boldsymbol{U}$ in the localised dG method (3.9) we have

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\cdot\boldsymbol{U} + A\boldsymbol{U}\cdot\boldsymbol{U}\mathrm{d}t + [\![\boldsymbol{U}_n]\!]\cdot\boldsymbol{U}_n^+ = \int_{I_n} \mathfrak{f}\cdot\boldsymbol{U}\mathrm{d}t.$$

Applying Lemma 3.1.20 we observe that

$$\left|\boldsymbol{U}_{n+1}^-\right|^2 - \left|\boldsymbol{U}_n^-\right|^2 + \int_{I_n} A\boldsymbol{U}\cdot\boldsymbol{U}\mathrm{d}t \leq \int_{I_n} \mathfrak{f}\cdot\boldsymbol{U}\mathrm{d}t.$$

Through Hölder's and Cauchy's inequality with ϵ we see

$$\begin{aligned}\left|\boldsymbol{U}_{n+1}^-\right|^2 - \left|\boldsymbol{U}_n^-\right|^2 + \int_{I_n} A\boldsymbol{U}\cdot\boldsymbol{U}\mathrm{d}t \leq& \frac{1}{4\epsilon}\left\|\mathfrak{f}\right\|_{L_2(I_n)}^2 + \epsilon\left\|\boldsymbol{U}\right\|_{L_2(I_n)}^2 \\ \leq& \frac{1}{4\epsilon}\left\|\mathfrak{f}\right\|_{L_2(I_n)}^2 + \epsilon C_A\int_{I_n} A\boldsymbol{U}\cdot\boldsymbol{U}\mathrm{d}t,\end{aligned}$$

after applying coercivity of A. Choosing $\epsilon = \frac{1}{2C_A}$ we find

$$\left|\boldsymbol{U}_{n+1}^-\right|^2 - \left|\boldsymbol{U}_n^-\right|^2 + \frac{1}{2}\int_{I_n} A\boldsymbol{U}\cdot\boldsymbol{U}\mathrm{d}t \leq \frac{C_A}{2}\left\|\mathfrak{f}\right\|_{L_2(I_n)}^2.$$

As this bound holds for arbitrary n we can iterate back in time to $n = 0$ finding

$$\left|\boldsymbol{U}_{n+1}^-\right|^2 + \frac{1}{2}\int_0^T A\boldsymbol{U}\cdot\boldsymbol{U}\mathrm{d}t \leq \left|\boldsymbol{U}_0^-\right|^2 + \frac{C_A}{2}\left\|\mathfrak{f}\right\|_{L_2([0,T])}^2,$$

noting that $\boldsymbol{U}_0^-$ is, by definition, the initial condition we can conclude.

□

3.1.1.2 Skew-symmetric linear ODEs

Here we restrict the right hand side function of the general ODE (2.1) to $\boldsymbol{F}(\boldsymbol{U}, t) = -B\boldsymbol{U}$, where B is a linear skew-symmetric operator. Here the operator B can represent either: a continuous spatial operator, a discretised spatial operator, or a constant operator. Possible PDEs which can be represented when B is a spatial operator include the Airy equation and linear KdV. We have assumed that there is no forcing term, as forcing destroys a lot of the structure of these problems.

Theorem 3.1.21 (Stability of the cG method for skew-symmetric problems)**.** *Consider the cG method (3.1) applied to a skew-symmetric linear ODE, i.e., $\boldsymbol{F}(\boldsymbol{U}, t) = -B\boldsymbol{U}$ where*

B is a skew-symmetric operator. The solution to the cG method satisfies

$$|\boldsymbol{U}(T)| = |\boldsymbol{U}(0)|.$$

Proof. The following argument follows directly from the "conservative" nature of the ODE, and the fact that the discretisation mimics this conservative structure. Choosing $\boldsymbol{V} = \Pi_{\mathbb{V}_q}(\boldsymbol{U})$, where $\Pi_{\mathbb{V}_q}$ represents the L_2 projection into $\mathbb{V}_q$, in (3.1) for skew-symmetric linear ODEs we find

$$\begin{aligned} 0 &= \int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \cdot \Pi_{\mathbb{V}_q}(\boldsymbol{U}) + B\boldsymbol{U} \cdot \Pi_{\mathbb{V}_q}(\boldsymbol{U})\mathrm{d}t \\ &= \int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \cdot \boldsymbol{U} + B\Pi_{\mathbb{V}_q}(\boldsymbol{U}) \cdot \Pi_{\mathbb{V}_q}(\boldsymbol{U})\mathrm{d}t, \end{aligned}$$

through the definition of the L_2 projection into $\mathbb{V}_q$ as $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \in \mathbb{V}_q$ by Remark 3.1.2. Through skew-symmetry of B, i.e., as $B\Pi_{\mathbb{V}_q}(\boldsymbol{U}) \cdot \Pi_{\mathbb{V}_q}(\boldsymbol{U}) = 0$ we find that

$$\begin{aligned} 0 &= \int_0^T \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|\boldsymbol{U}|^2 \,\mathrm{d}t \\ &= |\boldsymbol{U}(T)|^2 - |\boldsymbol{U}(0)|^2, \end{aligned}$$

through the fundamental theorem of calculus as required.

□

Theorem 3.1.22 (Stability of the upwind dG method for skew-symmetric problems)**.** *Consider the upwind dG method (3.8) applied to the skew-symmetric linear ODE, i.e., $\boldsymbol{F}(\boldsymbol{U}, t) = -B\boldsymbol{U}$ where B is a skew-symmetric operator. The solution to the upwind dG method satisfies*

$$\left|\boldsymbol{U}_{n+1}^-\right|^2 \leq \left|\boldsymbol{U}_n^-\right|^2,$$

i.e., the numerical solution is stable over time. Note that for the true solution we have that

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}|\boldsymbol{u}|^2 \,\mathrm{d}t = 0,$$

so this bound allows for artificial diffusion, but does not ensure it as we have not explicitly excluded the possibility that $\left|\boldsymbol{U}_{n+1}^-\right|^2 = \left|\boldsymbol{U}_n^-\right|^2$.

Proof. Choosing $\boldsymbol{V} = \boldsymbol{U}$ in the localised upwind dG method (3.9) we have that

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \cdot \boldsymbol{U}\mathrm{d}t + [\![\boldsymbol{U}_n]\!] \cdot \boldsymbol{U}_n^+ + \int_{I_n} B\boldsymbol{U} \cdot \boldsymbol{U}\mathrm{d}t = 0.$$

Applying Lemma 3.1.20 and the skew-symmetry of B we find

$$\frac{1}{2}\left|\boldsymbol{U}_{n+1}^{-}\right| \leq \frac{1}{2}\left|\boldsymbol{U}_{n}^{-}\right|,$$

as required.

□

3.1.1.3 Hamiltonian systems

In the continuous setting the natural notion of non-linear stability arises from conservation of the Hamiltonian, i.e., $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}(\boldsymbol{u}) = 0$. For the cG method (3.1) applied to Hamiltonian systems a discrete notion of non-linear stability follows directly from Theorem 3.1.6 which tells us that

$$\mathcal{H}(\boldsymbol{U}_{n+1}) = \mathcal{H}(\boldsymbol{U}_{n}),$$

where $\boldsymbol{U}$ is the solution of the cG method (3.1) applied to a Hamiltonian system.
For the upwind dG method (3.8) stability does not follow as immediately in general, as the Hamiltonian is not conserved at the nodes. However, the diffuse nature of the method, see Remark 3.1.14, leads to stability. Note that the diffuse nature of the method has only been proven in the case that $J\nabla\mathcal{H}(\boldsymbol{u})$ is linear. To prove the diffuse nature of nonlinear problems we can follow the methodology outlined in §3.1.1.2.

3.1.2 Convergence

Here we develop a new a priori error bound for the cG method (3.1) for symmetric linear problems. Similar order bounds can be developed for skew-symmetric linear problems, however, for brevity we shall not discuss this case here. As presented in Theorem 3.1.15, a priori error bounds already exist for this method which utilise duality arguments. Here we employ a simpler approach. While we are proving this a priori bound for the cG method we are primarily introducing the techniques so they can be applied to the new temporal finite element method we introduce in §3.2. Throughout we shall assume that the exact solution $\boldsymbol{u}$ is sufficiently smooth, to be concise we will require that $\boldsymbol{u} \in C^{q+2}([0,T])$.

Definition 3.1.23 (Interpolation operator for the cG method)**.** *Let* $\boldsymbol{w} \in \left(\mathcal{C}^0([0,T]) \cap L_2([0,T])\right)^D$*, then we define the interpolation operator* $\mathcal{I}(\boldsymbol{w}) \in \mathbb{V}_{q+1}^{C}(I_n)$

locally such that

$$\int_{I_n} \left(\mathcal{I}\left(\boldsymbol{w}\right) - \boldsymbol{w}\right) \cdot \boldsymbol{\phi} \mathrm{d}t = 0 \qquad \forall \boldsymbol{\phi} \in \left(\mathbb{P}_{q-1}\left(I_n\right)\right)^D, \tag{3.12}$$

for $n = 0, ..., N-1$, *and* $\mathcal{I}\left(\boldsymbol{w}(t_n)\right) = \boldsymbol{w}(t_n)$ *for* $n = 0, .., N$. *Note that this interpolation is uniquely defined as* (3.12) *locally fixes* q *degrees of freedom and nodal conditions fix the remaining degree of freedom on each element. When* $q = 0$ *our interpolation operator is uniquely defined through being exact at the nodes.*

Lemma 3.1.24 (Convergence of the cG interpolation operator)**.** *Let* $\boldsymbol{w} \in \left(\mathcal{C}^{q+1}([0,T]) \cap L_2([0,T])\right)^D$, *further let the interpolation operator* $\mathcal{I}$ *be as discussed in Definition 3.1.23, then*

$$\left\|\boldsymbol{w} - \mathcal{I}\left(\boldsymbol{w}\right)\right\|_{L_2(I_n)} \leq C\tau_n^{q+2} \left\| \frac{\mathrm{d}^{q+2}}{\mathrm{d}^{q+2}t} \boldsymbol{w} \right\|_{L_2(I_n)}.$$

Proof. Lemma 3.1.24 follows from the interpolation operator given in Definition 3.1.23 being exact for functions in $\mathbb{V}_{q+1}^C\left([0,T]\right)$.

□

Theorem 3.1.25 (A discrete error for the cG method for symmetric linear ODEs)**.** *Let* $\boldsymbol{U}$ *be the solution to the cG method* (3.1) *with right hand side* $\boldsymbol{F}\left(\boldsymbol{U}, t\right) = \mathfrak{f}\left(t\right) - A\boldsymbol{U}$, *additionally let* $\mathcal{I}\left(\boldsymbol{u}\right)$ *be the interpolation of the exact solution of the ODE* $\boldsymbol{u}$ *as described in Definition 3.1.23. The cG solution satisfies the bound*

$$\frac{1}{2}\left|\boldsymbol{U}_n - \mathcal{I}\left(\boldsymbol{u}\right)_n\right|^2 + \left\|\left|\Pi_{\mathbb{V}_q}\left(\boldsymbol{U} - \mathcal{I}\left(\boldsymbol{u}\right)\right)\right|_A\right\|_{L_2([0,t_n])}^2 \leq C\left\|\left|\mathcal{I}\left(\boldsymbol{u}\right) - \boldsymbol{u}\right|_A\right\|_{L_2([0,t_n])}^2,$$

with $\Pi_{\mathbb{V}_q}$ *denoting the* L_2 *projection into the finite element space* $\mathbb{V}_q$. *Note that here the constant* C *depends on the coercivity and continuity constants given in* (3.10) *and* (3.11) *respectively.*

Proof of Theorem 3.1.25. We begin by splitting the error of the cG solution $\boldsymbol{U}$ with respect to the exact solution $\boldsymbol{u}$, by adding and subtracting the cG interpolant described in Definition 3.1.23 applied to the exact solution $\mathcal{I}\left(\boldsymbol{u}\right)$, i.e.,

$$\boldsymbol{U} - \boldsymbol{u} = \left(\boldsymbol{U} - \mathcal{I}\left(\boldsymbol{u}\right)\right) + \left(\mathcal{I}\left(\boldsymbol{u}\right) - \boldsymbol{u}\right) =: \boldsymbol{\theta} + \boldsymbol{\rho}.$$

Adding and subtracting the aforementioned interpolant appropriately to the local cG method (3.2) we observe that

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\theta}\cdot\boldsymbol{V} + A\boldsymbol{\theta}\cdot\boldsymbol{V} + \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{I}(\boldsymbol{u})\cdot\boldsymbol{V} + A\mathcal{I}(\boldsymbol{u})\cdot\boldsymbol{V}\,\mathrm{d}t = \int_{I_n} \mathfrak{f}\cdot\boldsymbol{V}\,\mathrm{d}t \qquad \forall \boldsymbol{V}\in\mathbb{V}_q. \tag{3.13}$$

Additionally, we can write the general linear ODE (2.1) with $\boldsymbol{F}(\boldsymbol{U},t) = \mathfrak{f}(t) - A\boldsymbol{U}$ variationally as

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u} + A\boldsymbol{u}\cdot\boldsymbol{V}\,\mathrm{d}t = \int_{I_n} \mathfrak{f}\cdot\boldsymbol{V}\,\mathrm{d}t \qquad \forall \boldsymbol{V}\in\mathbb{V}_q. \tag{3.14}$$

Eliminating $\mathfrak{f}$ from (3.13) and (3.14) we can write

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\theta}\cdot\boldsymbol{V} + A\boldsymbol{\theta}\cdot\boldsymbol{V}\,\mathrm{d}t = -\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\rho}\cdot\boldsymbol{V} + A\boldsymbol{\rho}\cdot\boldsymbol{V}\,\mathrm{d}t. \tag{3.15}$$

Through integration by parts we have that

$$\begin{aligned}\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\rho}\cdot\boldsymbol{V}\,\mathrm{d}t &= -\int_{I_n} \boldsymbol{\rho}\cdot\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{V}\,\mathrm{d}t + \boldsymbol{\rho}_{n+1}\cdot\boldsymbol{V}_{n+1} - \boldsymbol{\rho}_n\cdot\boldsymbol{V}_n \\ &= 0,\end{aligned} \tag{3.16}$$

after applying the definition of the interpolant, as $\boldsymbol{\rho} := \mathcal{I}(\boldsymbol{u}) - \boldsymbol{u}$. Note that in the case $q = 0$ we have that $\int_{I_n}\boldsymbol{\rho}\cdot\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{V}\mathrm{d}t$ is trivially zero. In view of (3.16), we can choose $\boldsymbol{V} = \Pi_{\mathbb{V}_q}(\boldsymbol{\theta})$ in (3.15) allowing us to write

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\theta}\cdot\boldsymbol{\theta} + A\Pi_{\mathbb{V}_q}(\boldsymbol{\theta})\cdot\Pi_{\mathbb{V}_q}(\boldsymbol{\theta})\,\mathrm{d}t = -\int_{I_n} \boldsymbol{\rho}\cdot\Pi_{\mathbb{V}_q}(\boldsymbol{\theta})\,\mathrm{d}t,$$

through the definition of the L_2 projection, as $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\theta}\in\mathbb{V}_q$. Through Hölder's inequality, and then continuity of A we have

$$\int_{I_n} \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left(|\boldsymbol{\theta}|^2\right) + A\Pi_{\mathbb{V}_q}(\boldsymbol{\theta})\cdot\Pi_{\mathbb{V}_q}(\boldsymbol{\theta})\,\mathrm{d}t \leq C_A\left\||\boldsymbol{\rho}|_A\right\|_{L_2(I_n)}\left\|\left|\Pi_{\mathbb{V}_q}(\boldsymbol{\theta})\right|_A\right\|_{L_2(I_n)}$$

where C_A is given by (3.11). Further, through coercivity (3.10) and Cauchy's inequality with ϵ

$$\frac{1}{2}|\boldsymbol{\theta}_{n+1}|^2 + c_A\left\|\left|\Pi_{\mathbb{V}_q}(\boldsymbol{\theta})\right|_A\right\|^2_{L_2(I_n)} \leq \frac{1}{2}|\boldsymbol{\theta}_n|^2 + \frac{C_A^2}{4\epsilon}\left\||\boldsymbol{\rho}|_A\right\|^2_{L_2(I_n)} + \epsilon\left\|\left|\Pi_{\mathbb{V}_q}(\boldsymbol{\theta})\right|_A\right\|^2_{L_2(I_n)}.$$

Choosing $\epsilon = \frac{c_A}{2}$ and iterating back to the initial point in time

$$\frac{1}{2}\left|\boldsymbol{\theta}_{n+1}\right|^2 + \frac{c_A}{2}\left\|\left|\Pi_{\mathbb{V}_q}\left(\boldsymbol{\theta}\right)\right|_A\right\|^2_{L_2([0,t_{n+1}])} \leq \frac{1}{2}\left|\boldsymbol{\theta}_0\right|^2 + \frac{C_A}{2c_A}\left\|\left|\boldsymbol{\rho}\right|_A\right\|^2_{L_2([0,t_{n+1}])}.$$

Observing that $\boldsymbol{\theta}_0 = \boldsymbol{\rho}_0 = 0$ allows us to conclude.

□

Corollary 3.1.26 (Convergence of the cG method for symmetric linear ODEs)**.** *Let $\boldsymbol{U}$ be the solution to the cG method (3.1) with the symmetric right hand side $\boldsymbol{F}\left(\boldsymbol{U},t\right) = \mathfrak{f}\left(t\right) - A\boldsymbol{U}$ where A is a symmetric linear operator. Additionally, let $\boldsymbol{u} \in C^{q+2}([0,T])$ be the exact solution of the ODE (2.1), then*

$$\left|\boldsymbol{U}_n - \boldsymbol{u}_n\right|^2 + \left\|\left|\Pi_{\mathbb{V}_q}\left(\boldsymbol{U}-\boldsymbol{u}\right)\right|_A\right\|^2_{L_2([0,t_n])} \leq C\left\|\tau^{q+2}\left|\frac{\mathrm{d}^{q+2}}{\mathrm{d}t^{q+2}}\boldsymbol{u}\right|_A\right\|^2_{L_2([0,t_n])}.$$

Proof. Recall that through the introduction of the interpolant $\mathcal{I}\left(\boldsymbol{u}\right)$ described in Definition 3.1.23 we write

$$\boldsymbol{U} - \boldsymbol{u} = \left(\boldsymbol{U} - \mathcal{I}\left(\boldsymbol{u}\right)\right) + \left(\mathcal{I}\left(\boldsymbol{u}\right) - \boldsymbol{u}\right) =: \boldsymbol{\theta} + \boldsymbol{\rho}.$$

Through the triangle inequality, and Cauchy's inequality, we can write

$$\begin{aligned}\left|\boldsymbol{U}_n - \boldsymbol{u}_n\right|^2 \leq& \frac{3}{2}\left|\boldsymbol{\theta}_n\right|^2 + \frac{3}{2}\left|\boldsymbol{\rho}_n\right|^2\\ \leq& \frac{3}{2}\left|\boldsymbol{\theta}_n\right|^2,\end{aligned} \tag{3.17}$$

through Definition 3.1.23. Following a similar argument we see that

$$\begin{aligned}\left\|\left|\Pi_{\mathbb{V}_q}\left(\boldsymbol{U}-\boldsymbol{u}\right)\right|_A\right\|^2_{L_2([0,t_n])} \leq& \frac{3}{2}\left\|\left|\Pi_{\mathbb{V}_q}\left(\boldsymbol{\theta}\right)\right|_A\right\|^2_{L_2([0,t_n])} + \frac{3}{2}\left\|\left|\Pi_{\mathbb{V}_q}\left(\boldsymbol{\rho}\right)\right|_A\right\|^2_{L_2([0,t_n])}\\ \leq& \frac{3}{2}\left\|\left|\Pi_{\mathbb{V}_q}\left(\boldsymbol{\theta}\right)\right|_A\right\|^2_{L_2([0,t_n])} + \frac{3}{2}\left\|\left|\boldsymbol{\rho}\right|_A\right\|^2_{L_2([0,t_n])},\end{aligned} \tag{3.18}$$

through the stability of the L_2 projector. Combining (3.17) and (3.18), and applying Theorem 3.1.25 we find that

$$\left|\boldsymbol{U}_n - \boldsymbol{u}_n\right|^2 + \left\|\left|\Pi_{\mathbb{V}_q}\left(\boldsymbol{U}-\boldsymbol{u}\right)\right|_A\right\|^2_{L_2([0,t_n])} \leq C\left\|\left|\boldsymbol{\rho}\right|_A\right\|_{L_2([0,t_n])}.$$

We may conclude through the application of Lemma 3.1.24.

□