MA8404 Numerical solution of time dependent differential equations

Elena Celledoni
Department of Mathematical Sciences, NTNU
7491 Trondheim, Norway. email: elenac@math.ntnu.no

26. august 2013

Innhold

Kapittel 1

Lie group methods

1.1 Background material

This is section a summary of the results of the first and second chapter in [Ol] which are particularly relevant for the introduction of Lie group methods.

1.1.1 Manifolds

Definition 1.1.1. *[Ol] An m-dimensional manifold $\mathcal{M}$ is a topological space covered by a collection of open subsets $W_\alpha \subset \mathcal{M}$ (coordinate charts) and maps $\mathcal{X}_\alpha : W_\alpha \to V_\alpha \subset \mathbf{R}^m$ one-to-one and onto, where V_α is an open, connected subset of $\mathbf{R}^m$. $(W_\alpha, \mathcal{X}_\alpha)$ define coordinates on $\mathcal{M}$.*
$\mathcal{M}$ is a smooth manifold if the maps $\mathcal{X}_{\alpha\beta} = \mathcal{X}_\beta \circ \mathcal{X}_\alpha^{-1}$, are smooth where they are defined, i.e. on $\mathcal{X}_\alpha(W_\alpha \cap W_\beta)$ to $\mathcal{X}_\beta(W_\alpha \cap W_\beta)$.

Example 1.1.2. *$\mathbf{R}^m$ is a m-dimensional manifold covered with a single chart.*

Example 1.1.3. *The unit sphere $\mathbf{S}^{m-1} := \{\mathbf{x} \in \mathbf{R}^m \mid \sum_{i=1}^m x_i^2 = 1\}$ is a $m-1$-dimensional manifold covered with two charts obtained by omitting the north and south poles respectively. The coordinate maps are obtained considering the stereographic projection from the north and south pole respectively.*

Given two smooth manifolds $\mathcal{M}$ and $\mathcal{N}$ we say that $F : \mathcal{M} \to \mathcal{N}$ is a *smooth map* if it is smooth in local coordinates. Introducing local coordinates on the manifolds we get $x \in \mathcal{M}$, $x = (x_1, \ldots, x_m)$, and $y \in \mathcal{N}$, $y = (y_1, \ldots, y_n)$. Assume $y = F(x)$ and $y_i = F_i(x)$ $i = 1, \ldots, n$; if F_i is smooth as a map from an open subset of $\mathbf{R}^m$ to $\mathbf{R}$, then F is smooth also as map between $\mathcal{M}$ and $\mathcal{N}$.

The *rank of a map $F : \mathcal{M} \to \mathcal{N}$* is the rank of the Jacobian of F, the map is said to be *regular* if its rank is constant.

A subset $\mathcal{N} \subset \mathcal{M}$ of a manifold which is a manifold in its own right is a submanifold.

Definition 1.1.4. Submanifolds *An immersed submanifold $\mathcal{N}$ of a manifold $\mathcal{M}$ is a subset $\mathcal{N} \subset \mathcal{M}$ and a map F smooth and one-to-one $F : \tilde{\mathcal{N}} \to \mathcal{N} \subset \mathcal{M}$ with F everywhere of maximal rank and $\tilde{\mathcal{N}}$ an n-dimensional manifold.*

Example 1.1.5. *$\mathcal{M} = \mathbf{R}^3$ consider the parametrized curve $\phi(t) = (\cos(t), \sin(t), t)$ (a circular elix), ϕ is one-to-one and $\dot{\phi} = (-\sin(t), \cos(t), 1)$ is never 0 so the maximal rank condition is satisfied and the elix is an immersed submanifold of $\mathbf{R}^3$.*

1.1.2 Vector fields

A tangent vector to a manifold $\mathcal{M}$ at a point is the tangent to a smooth curve passing through the point: given $x \in \mathcal{M}$ and $\phi(t) \in \mathcal{M}$ the curve such that $\phi(0) = x$ then

$$\mathbf{v}|_x := \frac{d}{dt}\phi(t)\Big|_{t=0}.$$

Definition 1.1.6. *The tangent space to a m-dimensional manifold $\mathcal{M}$ at the point x is the vector space of dimension m formed by the collection of the tangent vectors at x and is denoted by $T_x\mathcal{M}$.*

$$\mathrm{T}_x\mathcal{M} := \{v = \frac{d}{dt}\phi(t)\Big|_{t=0}, \text{ s.t. } \phi(t) \in \mathcal{M}, \ \forall t, \ \phi(0) = x\}.$$

The definition of the tangent space of a sub-manifold of $\mathbf{R}^n$ is also given in [HLW] chapter IV.5.

The tangent bundle

$$T\mathcal{M} = \cup_{x\in\mathcal{M}} T_x\mathcal{M}$$

is the collection of all tangent spaces, it can be given the structure of a manifold of dimension $2m$.

The tangent bundle to the circle can be identified with the cartesian product of the circle with $\mathbf{R}$, $TS^1 \simeq S^1 \times \mathbf{R}$. The tangent bundle to the sphere TS^2 can NOT be identified with the cartesian product of the sphere and $\mathbf{R}^2$.

A *vector field* on $\mathcal{M}$ is a section of the tangent bundle of $\mathcal{M}$, i.e. is a smoothly varying assignment of tangent vectors: $\mathbf{v} : \mathcal{M} \to T\mathcal{M}$ such that $\mathbf{v}(x) = \mathbf{v}|_x \in T_x\mathcal{M}$. In local coordinates

$$\mathbf{v}(x) = \sum_{i=1}^{m} \xi^i(x)\frac{\partial}{\partial x^i},$$

$\xi^i(x)$ are smooth functions and $\frac{\partial}{\partial x^i}$ denote a basis of the tangent space $T_x\mathcal{M}$.

A curve $\phi : \mathbf{R} \to \mathcal{M}$ is an *integral curve* of the vector field $\mathbf{v}$ if, when $\phi(t) = x$, the tangent to the curve at t coincides with the vector field at x, i.e. $\dot{\phi}(t) = \mathbf{v}(x)$. This means that in local coordinates

$$\frac{dx^i}{dt} = \xi^i(x), \quad x^i = \phi_i(t).$$

Example 1.1.7. *We consider a vector field on $\mathbf{R}^2$,*

$$\mathbf{v}(x, y) = y\partial_x - x\partial_y,$$

$$\mathbf{v}(x, y) = \begin{pmatrix} \xi^1(x, y) \\ \xi^2(x, y) \end{pmatrix} = \begin{pmatrix} y \\ -x \end{pmatrix},$$

and to find the integral curve one has to solve

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= -x, \end{aligned}$$

obtaining

$$\begin{aligned} x(t) &= \cos(t)x_0 + \sin(t)y_0, \\ y(t) &= -\sin(t)x_0 + \cos(t)y_0. \end{aligned}$$

If $\phi(t)$ is a maximal integral curve of the vector field we denote it with

$$\phi(t) = \exp(t\mathbf{v})x_0, \quad x_0 = \phi(0),$$

$\exp(t\mathbf{v})x_0$ is the *flow* generated by the vector field $\mathbf{v}$, while $\mathbf{v}$ is called the *infinitesimal generator* of the flow. This notation is justified by some fundamental properties of the flow resembling known properties of the exponential mapping:

$$\exp(t\mathbf{v})\exp(s\mathbf{v})x_0 = \exp((t+s)\mathbf{v})x_0, \quad \exp(0\mathbf{v})x_0 = x_0,$$

$$\exp(t\mathbf{v})^{-1}x_0 = \exp(-t\mathbf{v})x_0, \quad \frac{d}{dt}\exp(t\mathbf{v})x_0 = \mathbf{v}|_{\exp(t\mathbf{v})x_0}$$

and also

$$\mathbf{v}|_{x_0} = \frac{d}{dt}\exp(t\mathbf{v})x_0\Big|_{t=0}, \quad \forall x_0 \in \mathcal{M},$$

i.e. given the flow starting from x_0 we can retrieve the vector field at x_0 by differentiating the flow with respect to t and then setting $t = 0$. The flow of a vector field can be expanded as

$$\exp(t\mathbf{v})x_0 = x_0 + t\ \mathbf{v}|_{x_0} + \mathcal{O}(t^2).$$

If x_0 is such that $\mathbf{v}|_{x_0} = 0$ we say that x_0 is a *singularity or equilibrium point* of the vector field, and this implies

$$\exp(t\mathbf{v})x_0 = x_0, \quad \forall t.$$

Points that are not equilibrium points are called *regular*.

Vector fields can operate on functions as *derivations*.

A derivation is a linear operator defined on an algebra[1] $\mathcal{A}$, $D : \mathcal{A} \to \mathcal{A}$ satisfying the Leibniz rule, $D(ab) = D(a)b + aD(b)$, for all $a, b \in \mathcal{A}$, where ab is the product of a and b in $\mathcal{A}$.

Given $f : \mathcal{M} \to \mathbf{R}$ the result of applying $\mathbf{v}$ as a derivation on f is a new function $\mathbf{v}(f)$ such that

$$\mathbf{v}(f(x)) = \sum_{i=1}^{m} \xi^i(x)\frac{\partial f}{\partial x^i} = \frac{d}{dt}f(\exp(t\mathbf{v})x)\Big|_{t=0},$$

$\mathbf{v}(f)$ determines the infinitesimal change of f along the flow of $\mathbf{v}$. It is easy to verify that $\mathbf{v}$ acts as a derivation

1. $\mathbf{v}(\lambda f + \mu g) = \lambda\mathbf{v}(f) + \mu\mathbf{v}(g)$,
2. $\mathbf{v}(fg) = f\mathbf{v}(g) + g\mathbf{v}(f)$.

The *Lie series* expansion of $f : \mathcal{M} \to \mathbf{R}$ is an expansion of f evaluated along the flow of $\mathbf{v}$

$$f(\exp(t\mathbf{v})x) = f(x) + t\mathbf{v}(f(x)) + \frac{1}{2}t^2\mathbf{v}(\mathbf{v}(f(x))) + \ldots,$$

converging for t sufficiently near 0. This is a way of reconstructing f along the flow of $\mathbf{v}$ given $\mathbf{v}$.

[1] An algebra is a vector space equipped with a multiplication operation, "$\cdot$" : $\mathcal{A}\times\mathcal{A} \to \mathcal{A}$. This operation is distributive with respect to the addition of the vector space and is compatible with the product by scalars in an appropriate sense.

The *Lie bracket* of vector fields is an operation on the set of vector fields, given two vector fields $\mathbf{v}$ and $\mathbf{w}$, $[\mathbf{v}, \mathbf{w}]$ is also a vector field. Such vector field is identified by the way it is acting on smooth functions, i.e. for all smooth $f : \mathcal{M} \to \mathbf{R}$,

$$[\mathbf{v}, \mathbf{w}](f) = \mathbf{v}(\mathbf{w}(f)) - \mathbf{w}(\mathbf{v}(f)).$$

In coordinates, assuming

$$\mathbf{v} = \sum_{i=1}^{m} \xi^i \frac{\partial}{\partial x^i}, \quad \mathbf{w} = \sum_{i=1}^{m} \eta^i \frac{\partial}{\partial x^i},$$

we obtain

$$\begin{aligned} [\mathbf{v}, \mathbf{w}] &= \sum_{i=1}^{m} \xi^i \sum_{j=1}^{m} \frac{\partial \eta^j}{\partial x^i} \frac{\partial}{\partial x^j} - \sum_{i=1}^{m} \eta^i \sum_{j=1}^{m} \frac{\partial \xi^j}{\partial x^i} \frac{\partial}{\partial x^j} \\ &= \sum_{j=1}^{m} \left(\sum_{i=1}^{m} \xi^i \frac{\partial \eta^j}{\partial x^i} - \sum_{i=1}^{m} \eta^i \frac{\partial \xi^j}{\partial x^i} \right) \frac{\partial}{\partial x^j}. \end{aligned}$$

One can verify that the following important properties hold for the Lie bracket of vector fields:

1. *bilinearity*: $[\lambda_1 \mathbf{v}_1 + \lambda_2 \mathbf{v}_2, \mathbf{w}] = \lambda_1 [\mathbf{v}_1, \mathbf{w}] + \lambda_2 [\mathbf{v}_2, \mathbf{w}]$.

2. *skew-symmetry*: $[\mathbf{v}, \mathbf{w}] = -[\mathbf{w}, \mathbf{v}]$.

3. *Jacobi identity*: $[\mathbf{v}, [\mathbf{w}, \mathbf{u}]] + [\mathbf{u}, [\mathbf{v}, \mathbf{w}]] + [\mathbf{w}, [\mathbf{u}, \mathbf{v}]] = 0$.

The *derivative map* or *differential* of a given map $F : \mathcal{M} \to \mathcal{N}$ is a map

$$dF : T\mathcal{M} \to T\mathcal{N}, \quad s.t. \quad dF|_x : T_x\mathcal{M} \to T_{F(x)}\mathcal{N}$$

which is such that for any curve $\phi(t)$ such that $\phi(t)|_{t=0} = x$ and correspondingly $F(\phi(t))|_{t=0} = F(x)$, with tangent vectors $\mathbf{v}_x := \frac{d}{dt}\phi(t)\big|_{t=0}$ and $\mathbf{w}_{F(x)} := \frac{d}{dt}F(\phi(t))\big|_{t=0}$ we have

$$dF|_x (\mathbf{v}_x) = \mathbf{w}_{F(x)}.$$

The differential is a linear map and in coordinates it is represented by the Jacobian of F. Only when F is one-to-one dF maps vector fields to vector fields, [Ol].

Assume F is one-to-one and $\mathbf{v}$ a vector field on $\mathcal{M}$ and $d\,F(\mathbf{v})$ a vector field on $\mathcal{N}$, one can prove that

$$F(\exp(t\mathbf{v})x) = \exp(td\,F(\mathbf{v}))F(x). \tag{1.1}$$

If $\mathbf{w}$ is also a vector field on $\mathcal{M}$ than one can also prove that the Lie bracket of vector fields is invariant under $d\,F$, i.e.

$$d\,F([\mathbf{v}, \mathbf{w}]) = [d\,F(\mathbf{v}), d\,F(\mathbf{w})].$$

1.1.3 Lie groups

A Lie group is a manifold G equipped with a smooth product operation "$\cdot$" which gives to G a group structure, e.g. there exists an identity element $e \in G$ and for any $g \in G$ its inverse g^{-1} is in G.

Here follows a list of examples of Lie groups,

- $(\mathbf{R}, +)$
- $\mathrm{GL}(n) = \{A \in M^{n\times n} | \det(A) \neq 0\}$ with the product between $n \times n$ matrices as group product,
- $\mathrm{SL}(n) = \{A \in M^{n\times n} | \det(A) = 1\}$ with the product between $n \times n$ matrices as group product,
- $\mathrm{SO}(n) = \{A \in M^{n\times n} | \det(A) = 1,\ A^T A = I\}$ with the product between $n \times n$ matrices as group product,
- $\mathrm{SP}(2r) = \{A \in M^{2r\times 2r} |\, A^T J A = J\}$ with the product between $2r \times 2r$ matrices as group product,

here $M^{n\times n}$ is the set of $n \times n$ real matrices.

1.1.4 Transformation groups

A transformation group acting on a smooth manifold $\mathcal{M}$ is a Lie group G and a smooth map $\Lambda : \mathrm{G} \times \mathcal{M} \to \mathcal{M}$ such that

- $\Lambda(e, x) = x$ for all $x \in \mathcal{M}$.
- $\Lambda(g, \Lambda(h, x)) = \Lambda(g \cdot h, x)$ for all $x \in \mathcal{M}$ and $g, h \in \mathrm{G}$.

Λ is called a Lie group action. We say that the Lie group action is *global* when $\Lambda(g, x)$ is defined for all $x \in \mathcal{M}$ and $g \in \mathrm{G}$ and local if it is defined on an open subset $\mathcal{V} \subset \mathrm{G} \times \mathcal{M}$ such that $\{e\} \times \mathcal{M} \subset \mathcal{V}$.

Some examples:

- $\mathrm{GL}(n, \mathbf{R})$ (or any of its subgroups) acting on $\mathbf{R}^n$ by matrix-vector multiplication.
- Any Lie group can act on itself by the group multiplication.

The set

$$\mathcal{O}_x = \{m \in \mathcal{M} \,|\, m = \Lambda(g, x),\ g \in \mathrm{G}\}$$

is called *orbit* of the Lie group action.

Example 1.1.8. *Consider the group* $\mathrm{O}(2)$ *acting on* $\mathbf{R}^2$ *the orbits are circles around the origin of* $\mathbf{R}^2$*. Analogously for* $\mathrm{O}(n)$ *acting on* $\mathbf{R}^n$ *the orbits are spheres:*

$$\{\, x \in \mathrm{R}^n \,|\, \|x\| = C \,\}$$

with C *a constant.*

A lie group action is said to be *transitive* when there is only one orbit, $\mathcal{O}_x = \mathcal{M}$, i.e.

$$\forall y \in \mathcal{M}\, \exists g \in G,\ s.t.\, \Lambda(g, x) = y.$$

Example 1.1.9. *The action of a Lie group* G *on itself by left multiplication is transitive.*

1.1.5 Homogeneous spaces

Given a Lie group G and a subgroup H we can define an equivalence relation on G:

$$g \sim \tilde{g} \Leftrightarrow \exists\, \tilde{h} \in \mathrm{H}\,\text{s.t.}\tilde{g} = g\tilde{h}.$$

The equivalence classes, $[g] = g \cdot \mathrm{H}$, are called left-cosets

$$[g] = \{gh \,|\, h \in \mathrm{H}\}.$$

One can prove that if H is a closed subgroup then the quotient G/H (i.e. G/ $\sim$) is a manifold called homogeneous space.

Recall that for G/H to be a group H needs to be a normal subgroup i.e. $g\mathrm{H}g^{-1} = H$ for all $g \in \mathrm{G}$.

In a homogeneous space the action $\Lambda : \mathrm{G} \times \mathrm{G/H} \to \mathrm{G/H}$, $\Lambda(g, [\tilde{g}]) = [g\tilde{g}]$ is transitive. In fact for any $[g_1]$ and $[g_2]$ in G/H it exists $g \in \mathrm{G}$ such that $g = g_2 g_1^{-1}$ and $\Lambda(g, [g_1]) = [g_2]$.

Example 1.1.10. Relevant homogeneous spaces. *Prove that the sphere is an homogeneous space* $S^2 = \mathrm{SO}(3)/\mathrm{SO}(2)$.

More in general $\mathrm{SO}(n)/\mathrm{SO}(p)$ *for* $p < n$ *is another interesting homogeneous space called Stiefel manifold and can be identified with the set of all* $n \times p$ *matrices with* p *orthonormal columns.*

Analogously $\mathrm{O}(n)/(\mathrm{O}(p)\times\mathrm{O}(n-p))$ *is the homogeneous space also known as Grassmann manifold.*

Definition 1.1.11. *Given* $x \in \mathcal{M}$ *and* Λ *a Lie group action on* $\mathcal{M}$ *the isotropy subgroup of* $x \in \mathcal{M}$ *is*

$$\mathrm{G}_x = \{g \in G \,|\, \Lambda(g, x) = x\}.$$

Recall that if $\mathrm{H} \subset \mathrm{G}$ is a subgroup of G (a Lie group) and H is topologically closed then H is a Lie subgroup see chapter II in [Ol].

This implies that G_x is a Lie subgroup.

Theorem 1.1.12. *A Lie group* G *acts globally and transitively on* $\mathcal{M}$ *if and only if* $\mathcal{M} \simeq \mathrm{G/H}$ *is isomorphic to the homogeneous space obtained as* G/H *with* $H = G_x$ *the isotropy subgroup of any chosen* $x \in \mathcal{M}$.

So any transitive Lie group action corresponds to a homogeneous space and viceversa.

1.1.6 Lie algebra of a Lie group

A Lie algebra $\mathfrak{g}$ ia a vector space with a bracket operation:

- $[\cdot, \cdot] : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ is a bilinear map
- is skew-symmetric: $[u, v] = -[v, u]$, $\forall\, u, v \in \mathfrak{g}$
- satisfies the Jacobi identity: $[u, [v, w]] + [w, [u, v]] + [v, [w, u]] = 0\ \forall u, v, w\ \in \mathfrak{g}$.

The Lie algebra of a given Lie group G is the tangent space at the identity element e, $\mathfrak{g} := \mathrm{T}_e G$.

This means that $\mathfrak{g}$ is a linear vector space, for matrix Lie groups the Lie algebra is typically a linear vector subspace of $M^{n\times n}$.

One can verify that

- the Lie algebra of $(\mathbf{R}, +)$ is $\mathbf{R}$
- the Lie algebra of $\mathrm{GL}(n)$ is $\mathfrak{gl}(n) = M^{n\times n}$,
- the Lie algebra of $\mathrm{SL}(n)$ is $\mathfrak{sl}(n) = \{A \in M^{n\times n} | \mathrm{trace}(A) = 0\}$,
- the Lie algebra of $\mathrm{SO}(n)$ is $\mathfrak{so}(n) = \{A \in M^{n\times n} | A^T + A = O\}$,
- the Lie algebra of $\mathrm{Sp}(2r)$ is $\mathfrak{sp}(2r) = \{A \in M^{2r\times 2r} | AJ + JA^T = O\}$,

  Bevis. Consider $A(t) \in \mathrm{Sp}(2r)$ $A(0) = I$, we get $V = \frac{d}{dt}A(t)\big|_{t=0} \in \mathfrak{sp}(2r)$. We differentiate $A(t)JA(t)^T = J$ and obtain

$$\dot{A}JA^T + AJ\dot{A}^T = O.$$

  Setting $t = 0$ we obtain

$$VJ + JV^T = O.$$

□

We now consider an alternative definition of the Lie algebra of a Lie group, we will see how this is equivalent to the definition given above, and will naturally obtain the definition of the bracket operation on $\mathfrak{g}$.

Consider the left multiplication of the Lie group $L_g : \mathrm{G} \to \mathrm{G}$, $L_g(\tilde{g}) = g\tilde{g}$, consider also the derivative mapping or differential of L_g, $d\,L_g : T_xG \to T_{gx}G$.

Definition 1.1.13. *A vector field on G,* $\mathbf{v}$*, is left invariant if*

$$d\,L_g(\mathbf{v}) = \mathbf{v}.$$

Since for a left invariant vector field $\mathbf{v}(e) = A$ implies $d\,L_g(\mathbf{v}(e)) = \mathbf{v}(g)$, once we know how the vector field looks like at the identity, e, we know how it looks like everywhere. For this reason we can identify the tangent space at the identity of the Lie group G, i.e. the Lie algebra $\mathfrak{g}$, with the set of left invariant vector fields. An analogous definition and identification can be given for right invariant vector fields.

The Lie algebra can be also used to describe the tangent space to G at any point. Here is the case of the orthogonal group.

Example 1.1.14. *Consider* $\gamma(t) \in \mathrm{O}(n)$. $\gamma(t)^T\gamma(t) = I$ *assume* $\gamma(0) = Q$ *and* $\dot{\gamma}(0) = W$. *By differentiating with respect to* t *and setting* $t = 0$ *we obtain* $W^TQ + Q^TW = O$. *Set* $A := Q^TW$ *and substitute* $W = QA$ *in the previous equation, obtaining* $A^T + A = O$. *So we obtain a characterization of the tangent space of* $\mathrm{O}(n)$ *at* Q *by means of* $\mathfrak{so}(n)$*:*

$$T_Q\,\mathrm{O}(n) = \{W = QA \,|\, A \in \mathfrak{so}(n)\}.$$

Analogous results can be obtained for other matrix Lie groups.

Recall form chapter I in [Ol]: given a mapping $F : \mathcal{M} \to \mathcal{N}$ the derivative mapping $d\,F$ does not map vector fields to vector fields unless F is one-to-one. Assume F is one-to-one and $\mathbf{v}$ a vector field on $\mathcal{M}$ and $d\,F(\mathbf{v})$ a vector field on $\mathcal{N}$, one can prove that

$$F(\exp(t\mathbf{v})x) = \exp(td\,F(\mathbf{v}))F(x). \tag{1.2}$$

If $\mathbf{w}$ is also a vector field on $\mathcal{M}$ than one can also prove that the Lie bracket of vector fields is invariant under dF, i.e.

$$dF([\mathbf{v},\mathbf{w}]) = [dF(\mathbf{v}), dF(\mathbf{w})].$$

If $\mathcal{M} = \mathrm{G}$ (a Lie group) the right multiplication $R_g : \mathrm{G} \to \mathrm{G}$ is one-to-one. Assume $\mathbf{v}$ and $\mathbf{w}$ are right invariant vector fields: $dR_g(\mathbf{v}) = \mathbf{v}$ and $dR_g(\mathbf{w}) = \mathbf{w}$, then

$$dR_g([\mathbf{v},\mathbf{w}]) = [dR_g(\mathbf{v}), dR_g(\mathbf{w})] = [\mathbf{v},\mathbf{w}],$$

so that also the Lie bracket $[\mathbf{v},\mathbf{w}]$ of the two right invariant vector fields is right invariant. The Lie algebra is closed under the Lie bracket of vector fields. So the Lie bracket of vector fields is the bracket operation of the Lie algebra $\mathfrak{g}$ of a Lie group G. Recall that the Lie bracket of vector fields is bilinear skew-symmetric and satisfies the Jacobi identity, see [Ol] chapter I.

Example 1.1.15. *Consider $\mathfrak{gl}(n)$ and the bracket of vector fields by a similar argument to that used in example 1.1.14 we obtain that a left invariant vector field $\mathbf{v}_A$ on $\mathrm{GL}(n)$ has coordinates given by the matrix product XA with $A \in \mathfrak{gl}(n)$ and $X \in \mathrm{GL}(n)$. Consider two left invariant vector fields $\mathbf{v}_A$ and $\mathbf{v}_B$ and write them as derivation operators:*

$$\mathbf{v}_A(X) = \sum_{i,j,k=1}^{n} x_{i,k}a_{k,j}\frac{\partial}{\partial x_{i,j}}, \quad \mathbf{v}_B(X) = \sum_{i,j,k=1}^{n} x_{i,k}b_{k,j}\frac{\partial}{\partial x_{i,j}},$$

here we denote with $a_{i,j} = (A)_{i,j}$ the (i,j)-element of the matrix A. Computing the Lie bracket of the two vector fields we obtain

$$\begin{aligned}[\mathbf{v}_A,\mathbf{v}_B](X) &= \sum_{i,j,k=1}^{n} x_{i,k}a_{k,j}\sum_{s=1}^{n} b_{j,s}\frac{\partial}{\partial x_{i,s}} - \sum_{l,r,s=1}^{n} x_{l,r}b_{r,s}\sum_{j=1}^{n} a_{s,j}\frac{\partial}{\partial x_{l,j}} \\ &= \sum_{i,k,s=1}^{n} x_{i,k}(AB-BA)_{k,s}\frac{\partial}{\partial x_{i,s}}.\end{aligned}$$

The Lie bracket of the two vector fields is a vector field with coordinates the matrix commutator of A and B: $[A,B] = AB - BA$. The situation is analogous for the case of right invariant vector fields a part from a change of sign.

1.1.7 The exponential map

Consider $\mathbf{v}$ a right invariant vector field on G and the right multiplication of the Lie group R_g. Using (1.2) we obtain

$$R_g(\exp(t\,\mathbf{v})e) = \exp(t\,dR_g(\mathbf{v}))g$$

and further using the right invariance of $\mathbf{v}$ on the right hand side we get

$$(\exp(t\,\mathbf{v})e)g = \exp(t\,\mathbf{v})g,$$

so that the left multiplication of g by the flow through e of $\mathbf{v}$ is equal to the flow through g of $\mathbf{v}$. We therefore can identify the flow of the right invariant vector field to be the corresponding one parameter subgroup[2] of G:

$$\exp(t\,\mathbf{v}) := \exp(t\,\mathbf{v})e.$$

[2]One parameter subgroup: a subgroup depending on one parameter, in this case t.

We can also define the exponential map $\exp : \mathfrak{g} \to \mathrm{G}$ as

$$\mathbf{v} \in \mathfrak{g} \mapsto \exp(t\mathbf{v})e|_{t=1} \in \mathrm{G}.$$

Example 1.1.16. *The flow of a right invariant vector field*

$$\mathbf{v}_A = \sum_{i,j,k} a_{i,j} x_{k,j} \frac{\partial}{\partial x_{i,j}},$$

is $\gamma(t)$ such that $\dot{\gamma} = \mathbf{v}_A(\gamma(t))$, in coordinates

$$\dot{\gamma}_{i,j} = \sum_{i,j=1}^{n} \left(\sum_{k=1}^{n} a_{i,k} \gamma_{k,j}(t) \right), \quad \text{i.e.} \quad \dot{\gamma} = A\gamma,$$

and $\gamma(t) = \exp(tA)\gamma(0)$. For the left invariant vector fields the flow is instead of the type $\eta(t) = \eta(0)\exp(tA)$.

Exercise 1.1.17. *Show that $\exp(O) = e$, where O is the zero element in $\mathfrak{g}$ and e is the identity element of* G. *Show that the derivative mapping of* exp *at O is the identity mapping in $\mathfrak{g}$.*

The results of the previous exercise guarantee that exp is a local diffeomorphism from a neighborhood of $O \in \mathfrak{g}$ to a neighborhood of $e \in \mathrm{G}$. This follows from the inverse function theorem. (See also [HLW] chapter IV.6 on this topic.)

The exponential mapping can be used to put local coordinates on the Lie group by means of the Lie algebra.

Theorem 1.1.18. *Let* G *be a connected Lie group with Lie algebra $\mathfrak{g}$. Every group element can be written as a product of exponentials:*

$$g = \exp(V_1)\exp(V_2)\cdots\exp(V_k),$$

for $V_1, \ldots, V_k \in \mathfrak{g}$.

1.1.8 Some properties of the exponential in matrix Lie groups

We want to consider

$$\frac{d}{dt}\exp(\sigma(t))e\Big|_{t=0},$$

where $\sigma(t)$ is a curve in $\mathfrak{gl}(n)$. We proceed giving two Lemmas which are used for this aim.

Lemma 1.1.19. Variation of constants formula.
The solution of the differential equation

$$\dot{u} = Au + w, \quad u(0) = u_0,$$

where A is a $m \times m$ constant matrix and $u_0, w \in \mathbf{R}^m$ are fixed, is

$$u(t) = e^{tA}u_0 + \int_0^t e^{(t-x)A} w dx.$$

Bevis. To find the solution of the considered differential equation we use the integrating factor e^{-xA}, we obtain

$$e^{-xA}\dot{u}(x) - Ae^{-xA}u(x) = e^{-xA}w.$$

We now integrate between 0 and t and obtain

$$e^{-tA}u(t) - u(0) = \int_0^t e^{-xA}wdx,$$

and multiplying on both sides with e^{tA} we obtain the result. □

Corollary 1.1.20. *If $w \in \mathbf{R}^m$ and A is a $m \times m$ matrix we have that*

$$\int_0^t e^{(t-x)A}wdx = \left.\frac{e^{tz}-1}{z}\right|_{z=A} w.$$

Bevis. We expand the integral at the left hand side of the equality by using the Taylor series of the exponential mapping and we obtain

$$\int_0^t e^{(t-x)A}wdx = \sum_{i=0}^{\infty}\int_0^t \frac{(t-x)^i}{i!}A^i wdx,$$

and since

$$\int_0^t \frac{(t-x)^i}{i!}wdx = \frac{t^{i+1}}{(i+1)!}w,$$

we obtain

$$\int_0^t e^{(t-x)A}wdx = \sum_{i=0}^{\infty} A^i\frac{t^{i+1}}{(i+1)!}w = \sum_{k=1}^{\infty} A^{k-1}\frac{t^k}{k!}w.$$

Now one can verify that

$$\frac{e^{tz}-1}{z} = \sum_{k=1}^{\infty} z^{k-1}\frac{t^k}{k!},$$

(use the expansion for e^{tz}), which implies that

$$\int_0^t e^{(t-x)A}wdx = \left.\frac{e^{tz}-1}{z}\right|_{z=A} w.$$

□

This Lemma is used in the proof of the next Lemma.

Lemma 1.1.21. *Assume $\sigma(t)$ is a $n \times n$ matrix for each t then we have that*

$$\left(\frac{d}{dt}e^{\sigma(t)}\right)e^{-\sigma(t)} = \left.\frac{e^z-1}{z}\right|_{z=\mathrm{ad}_\sigma}(\dot{\sigma}), \tag{1.3}$$

where for two $n \times n$ matrices B and C we have $\mathrm{ad}_B(C) = [B,C] = BC - CB$, where $[\cdot,\cdot]$ is the matrix commutator.

Bevis. Consider $B(s,t) = \left(\frac{d}{dt}e^{s\sigma(t)}\right)e^{-s\sigma(t)}$. By differentiating with respect to s we obtain

$$\begin{aligned}\frac{\partial}{\partial s}B(s,t) &= \left(\frac{d}{dt}\left(\sigma(t)e^{s\sigma(t)}\right)\right)e^{-s\sigma(t)} - \left(\frac{d}{dt}e^{s\sigma(t)}\right)e^{-s\sigma(t)}\sigma(t)\\ &= \dot{\sigma}(t)e^{s\sigma(t)}e^{-s\sigma(t)} + \sigma(t)\left(\frac{d}{dt}e^{s\sigma(t)}\right)e^{-s\sigma(t)} - B(s,t)\sigma(t)\\ &= \dot{\sigma}(t) + [\sigma(t), B(s,t)].\end{aligned}$$

This means that

$$\frac{\partial}{\partial s}B(s,t) = \mathrm{ad}_{\sigma}(B) + \dot{\sigma},$$

and we have $B(0,t) = O$. Note that ad_σ is a linear operator acting on $n \times n$ matrices, and can be represented as a $n^2 \times n^2$ matrix. Then taking $A = \mathrm{ad}_{\sigma(t)}$, in Lemma 1.1.19 and Corollary 1.1.20 we have

$$B(s,t) = \left.\frac{e^{sz}-1}{z}\right|_{z=\mathrm{ad}_{\sigma(t)}}(\dot{\sigma}(t)).$$

□

From Lemma 1.1.21 we have that

$$\frac{d}{dt}e^{\sigma(t)} = \left.\frac{e^z-1}{z}\right|_{z=\mathrm{ad}_\sigma}(\dot{\sigma})\cdot e^{\sigma(t)},$$

and for ease of notation we define

$$\begin{aligned}\mathrm{dexp}_{\sigma(t)}(\dot{\sigma}(t)) : &= \left.\frac{e^z-1}{z}\right|_{z=\mathrm{ad}_\sigma}(\dot{\sigma})\\ &= \sum_{k=1}^{\infty}\frac{1}{k!}\mathrm{ad}_\sigma^{k-1}(\dot{\sigma}(t))\\ &= \dot{\sigma}(t) + \frac{1}{2}[\sigma(t),\dot{\sigma}(t)] + \frac{1}{3!}[\sigma(t),[\sigma(t),\dot{\sigma}(t)]] + \dots\end{aligned}$$

1.2 Integration methods on manifolds

1.2.1 Introduction and motivation

We are interested in deriving *intrinsic* numerical integration methods for the problem

$$\dot{y} = F(y) \tag{1.4}$$

$$y(t_0) = y_0, \tag{1.5}$$

with $y_0 \in \mathcal{M}$, $\mathcal{M}$ a smooth manifold, and F a vector field on $\mathcal{M}$, i.e. $F(y(t)) \in T_{y(t)}\mathcal{M}$ for all t. Using a classical Runge-Kutta or multi-step method to approximate this problem does not automatically produce approximations on $\mathcal{M}$. Our aim is to design numerical methods which by construction produce approximations on the manifold. We call them intrinsic because we use only operations which make sense in the manifold setting.

As an example consider the following differential equation on the orthogonal group

$$\dot{Y} = A(Y)\cdot Y, \quad Y(0) = Y_0, \tag{1.6}$$

where Y and $A(Y)$ are $n \times n$ matrices, $A(Y)$ is skew-symmetric for all Y and Y_0 is an orthogonal matrix. The solution of (1.6) is an orthogonal matrix in fact if we take the derivative w.r.t. time of $Y(t)^T Y(t)$ we obtain

$$\begin{aligned} \frac{d}{dt} Y(t)^T Y(t) &= \dot{Y}^T Y + Y^T \dot{Y} = Y^T A(Y)^T Y + Y^T A(Y) Y \\ &= -Y^T A(Y) Y + Y^T A(Y) Y = 0, \end{aligned}$$

which means that $Y(t)^T Y(t)$ is constant and therefore

$$Y(t)^T Y(t) = Y_0^T Y_0 = I, \quad \forall t,$$

i.e. $Y(t)$ is an orthogonal matrix for all t. The format (1.6) is a consequence of the characterization of the tangent space discussed in example 1.1.14 and is valid in general for $A(Y)$ belonging to the Lie algebra $\mathfrak{g}$ of a Lie group G and $Y_0 \in \mathrm{G}$.

This type of differential equations arise in rigid body dynamics, structural mechanics, and many other fields of science and engineering. Often it is important to compute numerical approximations of the solution which belong to the Lie group.

Example 1.2.1. The rigid body equations *Euler's theorem states that the general displacement of a rigid body (or coordinate frame) with one point fixed is a rotation about some axis. This leads to Euler's equations for the free rigid body,*

$$\dot{\boldsymbol{\pi}} = \mathrm{skew}(T^{-1}\boldsymbol{\pi})\,\boldsymbol{\pi}, \tag{1.7}$$

where

$$\mathrm{skew}(\boldsymbol{v}) = \begin{pmatrix} 0 & v_3 & -v_2 \\ -v_3 & 0 & v_1 \\ v_2 & -v_1 & 0 \end{pmatrix},$$

and

$$T = \begin{pmatrix} I_1 & 0 & 0 \\ 0 & I_2 & 0 \\ 0 & 0 & I_3 \end{pmatrix}.$$

Here π is the angular momentum, I_1, I_2, I_3 are the principal moments of inertia of the body. According to Euler's theorem there exists a 3×3 rotation matrix $Q(t)$ such that $\pi = Q\pi_0$. Such a matrix must then satisfy the following differential equation

$$\dot{Q} = \mathrm{skew}(T^{-1}\boldsymbol{\pi})\,Q. \tag{1.8}$$

Correct and efficient numerical integration of equations (1.7) and (1.8) is of interest for applications in molecular dynamics and celestial mechanics, for example.

Suppose we apply the Forward Euler's method to (1.6), i.e.

$$Y_{n+1} = Y_n + hA(Y_n)Y_n, \;\; n = 0, 1, 2 \ldots$$

and assume $Y_0^T Y_0 = I$. We want to know if this implies that also $Y_1^T Y_1 = I$. By direct calculation we obtain

$$\begin{aligned} Y_1^T Y_1 &= (Y_0^T + hY_0^T A(Y_0)^T)(Y_0 + hA(Y_0)Y_0) \\ &= I + hY_0^T A(Y_0)Y_0 + hY_0^T A(Y_0)^T Y_0 + h^2 Y_0^T A(Y_0)^T A(Y_0) Y_0 \\ &= I - h^2 Y_0^T A(Y_0)^2 Y_0, \end{aligned}$$

in general we can not expect the term $Y_0^T A(Y_0)^2 Y_0$ to vanish and therefore we can not expect Y_1 to be orthogonal.

We want to find strategies alternative to the forward Euler's method which maintain the orthogonality under time discretization. Given a $n \times n$ matrix B with constant entries we know that

$$\exp(tB) = e^{tB} = \sum_{k=0}^{\infty} \frac{t^k}{k!} B^k$$

is the solution of the matricial differential equation

$$\dot{Y} = BY, \quad Y(0) = I.$$

Observe that if B is a skew-symmetric matrix the above equation is a special case of equation (1.6) and $\exp(tB)$ is an orthogonal matrix. We consider the following alternative to the forward Euler method for the numerical integration of equation (1.6),

$$Y_{n+1} = \exp(hA(Y_n))Y_n. \tag{1.9}$$

Provided $Y_0^T Y_0 = I$ and exploiting the orthogonality of $\exp(hA(Y_0))$ we can verify that Y_1 is an orthogonal matrix, in fact

$$Y_1^T Y_1 = Y_0^T \exp(hA(Y_0))^T \exp(hA(Y_0)) Y_0 = I,$$

in other words the proposed method produces numerical approximations to the solution of (1.6) which belong to the set of orthogonal matrices.

The method (1.9) is known as Lie-Euler method and has order 1. In fact if we consider the Taylor expansion of $Y(h)$ and $Y_1 = Y_1(h)$ around zero, we can easily verify that they coincide up to second order in h. We have

$$Y(h) = Y_0 + hA(Y_0)Y_0 + \frac{h^2}{2!}\ddot{Y}(0) + \dots$$

and

$$Y_1(h) = Y_1(0) + h\dot{Y}_1(0) + \frac{h^2}{2!}\ddot{Y}_1(0) + \dots.$$

Now $Y_1(0) = Y_0$ and since

$$\frac{d}{dh} Y_1(h)\Big|_{h=0} = \frac{d}{dh} \exp(hA(Y_0))Y_0\Big|_{h=0} = A(Y_0) \exp(hA(Y_0))Y_0|_{h=0} = A(Y_0)Y_0,$$

we easily see that the two Taylor expansions coincide up to terms of order at least 2 in h.

In the next sections we will generalize this method following two different strategies which lead to two different classes of methods.

1.2.2 Methods based on frame vector fields

Definition 1.2.2. *A set of vector fields $\{\mathcal{E}_1, \dots \mathcal{E}_d\}$ on the manifold $\mathcal{M}$ of dimension $m \leq d$ is a set of frame vector fields if*

$$T_x\mathcal{M} = \text{span}\{\mathcal{E}_1|_x, \dots \mathcal{E}_d|_x\}, \quad \forall\, x \in \mathcal{M}.$$

Given any vector field F on $\mathcal{M}$ we have

$$F(y) = \sum_{i=1}^{d} f_i(y)\mathcal{E}_i(y).$$

Definition 1.2.3. *We denote with F_p the vector field*

$$F_p(x) = \sum_{i=1}^{d} f_i(p)\mathcal{E}_i(x)$$

we say that F_p is the vector field F frosen at the point p.

Given at $\mathcal{M}$ is a manifold with a set of frame vector fields we can define intrinsic Runge-Kutta like methods as follows:

Commutator-free method

for $r = 1 : s$ do

$Y_r = \exp(\sum_{k=1}^{s} \alpha_{rJ}^k F_k) \cdots \exp(\sum_{k=1}^{s} \alpha_{r1}^k F_k)(p)$

$F_r = hF_{Y_r} = h\sum_{i=1}^{d} f_i(Y_r)\mathcal{E}_i$

end

$y_1 = \exp(\sum_{k=1}^{s} \beta_J^k F_k) \cdots \exp(\sum_{k=1}^{s} \beta_1^k F_k)p$

Here n counts the number of time steps and h is the step-size of integration. The integrator has s stages and parameters α_{rJ}^k, β_J^k. Each new stage value is obtained as a composition of J exponentials of linear combinations of vector fields frozen at the previously computed stage values.

In the following tableaus we report the coefficients of a method of order 3 and a method of order 4. The method of order 3 requires the computation of one exponential of each internal stage value and the composition of two exponentials for updating the solution. In the order 4 method the first three stage values require one exponential each, while the fourth stage and the solution update require two exponentials.

$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{3} & \frac{1}{3} & & \\ \frac{2}{3} & 0 & \frac{2}{3} & \\ \hline & \frac{1}{3} & 0 & 0 \\ & -\frac{1}{12} & 0 & \frac{3}{4} \end{array} \qquad \begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & \frac{1}{2} & 0 & 0 & \\ & -\frac{1}{2} & 0 & 1 & \\ \hline & \frac{1}{4} & \frac{1}{6} & \frac{1}{6} & -\frac{1}{12} \\ & -\frac{1}{12} & \frac{1}{6} & \frac{1}{6} & \frac{1}{4} \end{array}$$

Example 1.2.4. *Let $\mathcal{M}$ be a manifold acted upon transitively by a Lie group G. Denote with $\Lambda : \mathrm{G} \times \mathcal{M} \to \mathcal{M}$ the Lie group action. Suppose $E_1, \ldots, E_d$ a basis of the Lie algebra then $F_{E_1}, \ldots F_{E_d}$ obtained by*

$$F_{E_i}(x) = \frac{d}{dt}\Lambda(\exp(tE_i), x)\Big|_{t=0},$$

are a set of frame vector fields.

In particular for matrix Lie groups consider the vector field $A(Y)Y$ of equation (1.6). Here $A(y) \in \mathfrak{g}$ and $A(Y) = \sum_{i=1}^{d} a_i(y)E_i$ with $E_1, \ldots, E_d$ a basis of the Lie algebra[3]. The vector field frosen at a point $P \in \mathrm{G}$ is simply $A(P)Y$.

[3]Say for $\mathfrak{so}(n)$ a basis is given by the matrices of rank 2 of the type $\mathbf{e}_i\mathbf{e}_j^T - \mathbf{e}_j\mathbf{e}_i^T$ with $\mathbf{e}_i, \mathbf{e}_j \in \mathbf{R}^n$ canonical vectors, and $i = 1, \ldots, n$, $j \leq i - 1$.

Note that at each stage in the commutator-free methods we allow for the composition of at most J exponentials. As, in general, the computation of matrix exponentials is a computationally demanding task, it is of interest to find methods in this class which require a minimal number of exponentials, see [CMO] for details.

Given the above format for the methods the challenge is to find parameters

$$\alpha^k_{J,i}, \ldots, \alpha^k_{1,i}, \quad i,k = 1, \ldots, s, \qquad \beta^k_J, \ldots, \beta^k_1, \quad k = 1, \ldots, s$$

such that the formulae above produce a method of a desired order. The order theory for these methods can be developed as usual by requiring that

$$\frac{d^r}{dh^r} Y_1 \Big|_{h=0} = \frac{d^r}{dh^r} Y(h) \Big|_{h=0}, \quad r = 1, \ldots, p.$$

The number of order conditions for order p is higher than for classical Runge-Kutta methods. A complete treatment of this subject can be found in [O].

If we assume $J = i - 1$ and allow for just one of the α values to be different from zero in each exponential, we obtain the Crouch and Grossman methods, [CG].

Crouch and Grossman

for $i = 1 : s$ do

$Y_i = \exp(a_{i,i-1} F_{i-1}) \cdots \exp(a_{i,1} F_1) Y_n$

$F_r = h F_{Y_r} = h \sum_{i=1}^d f_i(Y_r) \mathcal{E}_i$

end

$Y_{n+1} = \exp(b_s F_s) \cdots \exp(b_1 F_1) Y_n.$

Crouch and Grossman pioneered the field of integration methods on manifolds (and Lie Group methods) in the nineties. These methods are defined by a tableau similar to the classical Runge-Kutta Butcher tableau, but also in this case the methods require extra order conditions compared to classical Runge-Kutta methods. An example of a method of order 3 is the following:

$$\begin{array}{c|ccc} 0 & & & \\ -\frac{1}{24} & -\frac{1}{24} & & \\ \frac{17}{24} & \frac{161}{24} & -6 & \\ \hline & 1 & -\frac{2}{3} & \frac{2}{3} \end{array}.$$

In the case of matrix Lie groups, equation (1.6), using the results of example 1.2.4 and applying this method we obtain:

$$\begin{aligned} Y_1 &= Y_n \\ Y_2 &= \exp(-h/24\, A(Y_1)) Y_n \\ Y_3 &= \exp(-6\, hA(Y_2)) \exp(161/24\, hA(Y_1)) Y_n \\ \\ Y_{n+1} &= \exp(2/3\, hA(Y_3)) \exp(-2/3\, hA(Y_2)) \exp(hA(Y_1)) Y_n, \end{aligned}$$

this method requires 6 exponentials. If we use the commutator-free method of the same order we obtain:

$$\begin{aligned} Y_1 : &= Y_n \\ Y_2 &= \exp(h/3\, A(Y_1)) Y_1 \\ Y_3 &= \exp(2/3 hA(Y_2)) Y_1 \\ \\ Y_{n+1} &= \exp(-1/2 hA(Y_1) + 3/4 hA(Y_3)) Y_2, \end{aligned}$$

requiring three exponentials.

1.2.3 RK-MK methods

Assume the manifold $\mathcal{M}$ is acted upon transitively by a Lie group G. We use the action and perform the following change of variables for (1.4):

$$y(t) = \Lambda(\exp(\sigma(t)), y_0), \quad \sigma(t) \in \mathfrak{g},$$

valid in a neighborhood of $y_0 \in \mathcal{M}$. The idea is to derive an equation for σ in the Lie algebra and then to integrate this equation with a classical Runge-Kutta method and obtain an approximation $\tilde{\sigma} \approx \sigma(h)$ which still belongs to the same Lie algebra and which, via exponentiation, and the Lie group action, generates

$$y(h) \approx y_1 = \Lambda(\exp(\tilde{\sigma}), y_0).$$

Consider the mapping $\Lambda_{y_0} : \mathrm{G} \to \mathcal{M}$, defined by $\Lambda_{y_0}(g) = \Lambda(g, y_0)$. By differentiation we obtain

$$F(y(t)) = \frac{d}{dt}y(t) = \frac{d}{dt}\Lambda(\exp(\sigma(t)), y_0) = \frac{d}{dt}\Lambda_{y_0}(\exp(\sigma(t)))$$

and further, from the definition of differential (derivative mapping), [Ol] chapter I, we obtain

$$F(y(t)) = d\Lambda_{y_0}(\frac{d}{dt}\exp(\sigma(t))).$$

By assuming now without substantial loss of generality that we work with matrix Lie groups, from Lemma 1.9 we obtain

$$F(y(t)) = d\Lambda_{y_0}(d\mathrm{R}_{\exp(\sigma)} d\exp_\sigma(\dot{\sigma})) = (d\Lambda_{y_0} \circ d\mathrm{R}_{\exp(\sigma)})(d\exp_\sigma(\dot{\sigma})), \tag{1.10}$$

where R_g is the right multiplication by g in the Lie group.

Using the same set of frame vector fields defined in example 1.2.4 in the previous section, $F_{E_i}(x) = d\Lambda_x(E_i)$, and the linearity of the map $d\Lambda_x$

$$F(x) = \sum_{i=1}^{d} f_i(x)F_{E_i}(x) = d\Lambda_x(\sum_{i=1}^{d} f_i(x)E_i),$$

here we can define $f(x) := \sum_{i=1}^{d} f_i(x)E_i \in \mathfrak{g}$. Setting $x = \Lambda_{y_0}(\exp(\sigma(t)))$ we obtain the following expression for the left hand side of (1.10)

$$F(y(t)) = F(\Lambda_{y_0}(\exp(\sigma(t)))) = d\Lambda_{\Lambda_{y_0}(\exp(\sigma))}\left(f(\Lambda_{y_0}(\exp(\sigma)))\right.\!.$$

Now we have that

$$d\Lambda_{y_0} \circ d\mathrm{R}_{\exp(\sigma)} = d(\Lambda_{y_0} \circ \mathrm{R}_{\exp(\sigma)}) = d\Lambda_{\Lambda_{y_0}(\exp(\sigma))},$$

see [Ol] chapter I. By substituting at the right hand side of (1.10) we obtain

$$d\Lambda_{\Lambda_{y_0}(\exp(\sigma))}(\sum_{i=1}^{d} f_i(y(t))E_i) = d\Lambda_{\Lambda_{y_0}(\exp(\sigma))}(d\exp_\sigma(\dot{\sigma})),$$

which is fulfilled if

$$d\exp_\sigma(\dot{\sigma}) = f(\Lambda_{y_0}(\exp(\sigma)),$$

and which in turn gives the differential equation for σ in the Lie algebra $\mathfrak{g}$:

$$\dot{\sigma} = d\exp_\sigma^{-1}(f(\Lambda_{y_0}(\exp(\sigma))). \tag{1.11}$$

Here dexp_σ is an invertible map provided $\|\sigma\| < \pi$, see [HLW] (Lemma 4.2 chap. III, 4.1) for details, and the inverse is given by

$$\mathrm{dexp}^{-1}_{\sigma(t)}(u) := \left.\frac{z}{e^z-1}\right|_{z=\mathrm{ad}_\sigma}(u) = \sum_{k=0}^{\infty}\frac{B_k}{k!}\mathrm{ad}^k_\sigma(u)\,. \tag{1.12}$$

Here the coefficients B_k are called Bernoulli numbers, the first four of them are

$$B_0 = 1,\ \ B_1 = -\frac{1}{2},\ \ B_2 = \frac{1}{6},\ \ B_3 = 0.$$

We solve numerically with a Runge-Kutta method equation (1.11) in place of applying the same method directly to equation (1.4). The approximation $\tilde{\sigma} \approx \sigma(h)$ is then used to construct $y_1 \approx y(h)$ via exponentiation and the Lie group action, i.e.

$$y_1 = \Lambda(\exp(\tilde{\sigma}), y_0).$$

This procedure gives the Runge-Kutta Munthe-Kaas (RKMK) methods, and has been originally presented in [MK], in an equivalent but different way.

Assuming $\sigma(h) = \tilde{\sigma} + \mathcal{O}(h^{p+1})$, (which means that the Runge-Kutta method considered is of order p) the local error $\|y(h) - y_1\| = \|\Lambda(e^\sigma, y_0) - \Lambda(e^{\tilde{\sigma}}, y_0)\|$ is also $\mathcal{O}(h^{p+1})$. This can be shown using the Taylor expansion of the exponential and of the action. In other words if we apply a RK method of order p to (1.11) we obtain an approximation of the same order for (1.4).

In the practical implementation of this strategy, as $\sigma(h) = \mathcal{O}(h)$, (remember that $\sigma(0) = 0$) the infinite series defining $\mathrm{dexp}^{-1}_\sigma$, (1.12), can be truncated to the right order of consistency in h, including just a small number of terms. We here report the algorithm for the RK-MK methods applied to (1.4).

RKMK

for $i = 1 : s$ do

$\quad \sigma_i = h\sum_{j=1}^{i-1} a_{i,j}\hat{\mathrm{dexp}}^{-1}_{\sigma_j}(f(\Lambda_{Y_n}(\exp(\sigma_j))))$

end

$\tilde{\sigma} = h\sum_{i=1}^{s} b_i\hat{\mathrm{dexp}}^{-1}_{\sigma_i}(f(\Lambda_{Y_n}(\exp(\sigma_i))))$

$Y_1 = \Lambda_{Y_n}(\exp(\tilde{\sigma}))$,

where we have denoted with $\hat{\mathrm{dexp}}^{-1}_\sigma$ the truncation (to the correct consistency order) of the expansion (1.12). Here $a_{i,j}$ and b_i $i = 1, \ldots, s$ $j = 1, \ldots, s$ are the parameters of a classical Runge-Kutta method.

Alternatively the algorithm can be written in the following format.

RKMK

for $i = 1 : s$ do

$\quad \sigma_i = h\sum_{j=1}^{i-1} a_{i,j}\tilde{K}_j$

$\quad K_i = f(\Lambda_{Y_n}(\exp(\sigma_i)))$

$\quad \tilde{K}_i = \hat{\mathrm{dexp}}^{-1}_{\sigma_i}(K_i)$

end

$\tilde{\sigma} = h\sum_{i=1}^{s} b_i\tilde{K}_i$

$Y_1 = \Lambda_{Y_n}(\exp(\tilde{\sigma}))$.

We stress once more that for the RKMK methods the Runge-Kutta parameters coincide with the parameters of the classical Runge-Kutta methods and no extra order conditions are produced in this case. We can for example consider the Butcher tableau for the Heuns method of order 3, see [HNW] p. 135. This method applied to the ordinary differential equation $\dot{y} = g(y)$, $y(0) = y_0$ takes the following form:

$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{3} & \frac{1}{3} & & \\ \frac{2}{3} & & \frac{2}{3} & \\ \hline & \frac{1}{4} & 0 & \frac{3}{4} \end{array}, \qquad \begin{aligned} Y_1 &= y_n \\ Y_2 &= y_n + \tfrac{h}{3} g(Y_1) \\ Y_3 &= y_n + \tfrac{2}{3} h g(Y_2) \\ y_{n+1} &= y_n + \tfrac{h}{4}(g(Y_1) + 3g(Y_3)). \end{aligned}$$

We use now this method on a matrix Lie group equation in the RK-MK fashion. The Lie group action is the action of G on itself by matrix-matrix multiplication, and the exponential is the matrix exponential. We obtain

$$\begin{aligned} \sigma_1 &= 0 \\ K_1 &= A(Y_n) \\ \tilde{K}_1 &= \mathrm{dexp}_O^{-1}(K_1) = K_1 \\ \sigma_2 &= \tfrac{h}{3} K_1 \\ K_2 &= A(\exp(\sigma_2) Y_n) \\ \tilde{K}_2 &= K_2 - \tfrac{1}{2}[\sigma_2, K_2] + \tfrac{1}{12}[\sigma_2, [\sigma_2, K_2]] \\ \sigma_3 &= \tfrac{2}{3} h \tilde{K}_2 \\ K_3 &= A(\exp(\sigma_3) Y_n) \\ \tilde{K}_3 &= K_3 - \tfrac{1}{2}[\sigma_3, K_3] + \tfrac{1}{12}[\sigma_3, [\sigma_3, K_3]] \\ \tilde{\sigma} &= \tfrac{h}{4} \tilde{K}_1 + \tfrac{3}{4} h \tilde{K}_3 \\ Y_{n+1} &= \exp(\tilde{\sigma}) Y_n. \end{aligned}$$

The described method requires the computation of 4 matrix commutators and 3 matrix exponentials per time step. It is possible to reduce the number of commutators to 1 for methods of order 3 by using techniques described in [MKO].

1.2.4 Magnus methods for linear systems of ODEs

A complete treatment of Lie group methods includes the methods based on Magnus expansion. We refer to [HLW] p.121-123, for this topic.

1.2.5 Further implementation details for the implementation of Lie group methods

We have seen that he Commutator-free and Crouch and Grossmann methods and RK-MK methods, applied to differential equations on manifolds acted upon by Lie groups, produce numerical approximations which belong to the manifold.

We leave as an exercise to the reader to show with similar arguments that the methods based on Magnus series are also producing approximations of linear differential equations on the Lie group G which belong to G.

For techniques for the approximation of the matrix exponential in a Lie algebraic setting see [CI1], [CI2], [IZ].

In Lie group methods we exploit the crucial property that the exponential mapping is a local diffeomorphism form a neighborhood of $O \in \mathfrak{g}$ to put local coordinates on the Lie group G. This can be done also in other ways. In the case of quadratic Lie groups, i.e.

$$G := \{y \in GL(n) \,|\, yPy^T = P\}, \quad \mathfrak{g} := \{x \in \mathfrak{gl}(n) \,|\, xP + Px^T = O\},$$

where P is a fixed $n \times n$ invertible matrix (like for example $\mathrm{SO}(n)$: $P = I$, $\mathrm{SP}(2r)$: $P = J$) it is possible to use alternatively the Cayley transformation as a coordinate map:

$$\mathrm{cay}(A) = (I - \frac{1}{2}A)^{-1}(I + \frac{1}{2}A),$$

see for example [LS] and [DLP]. The mapping $\mathrm{dexp}|_A^{-1}$ is replaced by

$$\mathrm{dcay}|_A^{-1} : \mathfrak{g} \to \mathfrak{g}, \quad \mathrm{dcay}|_A^{-1}(B) = (I - \frac{1}{2}A)B(I + \frac{1}{2}A).$$

Alternative coordinate mappings valid for any Lie group are the so called second kind coordinates (skc) [V], and can be obtained by taking a basis $E_1, \ldots, E_d$ of $\mathfrak{g}$ as a starting point and considering the following composition of exponentials:

$$\mathrm{skc}(A) = \exp(a_1 E_1) \ldots \exp(a_d E_d), \quad A = \sum_{i=1}^{d} a_i E_i.$$

For further details on integration methods using these coordinates see [OM].

1.2.6 Applications of Lie group methods

Isospectral flows

Isospectral flows arise in a wide range of applications for example in certain control problems and in medical imaging. See p. 107 in [HLW].

1.2.7 Stiefel manifolds

Differential equations on Stiefel manifolds arise in many applications as for example statistical signal processing and neural networks, multivariate data analysis, and mechanical systems.

Definition 1.2.5. *The Stiefel manifold is the set*

$$\mathrm{St}(n,p) := \{X \in M^{n\times p} \quad \text{s.t.} \quad X^T X = I_{p\times p} \in M^{p\times p}\}$$

where $M^{n\times p}$ is the linear vector space of $n \times p$ matrices and we assume $p \leq n$, and $I_{p\times p}$ denotes the $p \times p$ identity matrix.

One can show that $\mathrm{St}(n,p)$ is a differentiable manifold.

Note that with $n = p$, $\mathrm{St}(n,n) = SO(n)$, while, for $p = 1$, $\mathrm{St}(n,1)$ is the unit sphere in R^n. We want to show that any differential equation on the $\mathrm{St}(n,p)$ can be written in the form

$$\dot{Y} = A(Y) \cdot Y, \quad Y(0) = Y_0 \in \mathrm{St}(n,p), \tag{1.13}$$

with $A(Y) \in \mathfrak{so}(n)$. To prove this fact we start by characterizing the tangent space of $\mathrm{St}(n,p)$ by means of $\mathfrak{so}(n)$.

Consider the curve $X(t) \in \mathrm{St}(n,p)$, by definition we have that for each t, $X(t)^T X(t) = I_{p\times p}$.

Proposition 1.2.6. *For any smooth curve $X(t) \in \mathrm{St}(n,p)$ one can find a smooth curve $\mathbf{X}(t) = (X(t), X(t)^\perp) \in SO(n)$ such that $X(t) = \mathbf{X} \cdot I_{n,p}$, where*

$$I_{n,p} := \begin{bmatrix} I_{p\times p} \\ O_{(n-p)\times p} \end{bmatrix},$$

and $O_{(n-p)\times p}$ is the $(n-p) \times p$ zero matrix.

Proof omitted

Consider $V \in \mathrm{T}_P\mathrm{St}(n,p)$ with $P \in \mathrm{St}(n,p)$, assume $X(t) \in \mathrm{St}(n,p)$ is a smooth curve such that $X(0) = P$, and $\dot{X}(0) = V$ then, from Proposition 1.2.6, we have

$$V = \frac{d}{dt}(X(t), X(t)^\perp)I_{n,p}\Big|_{t=0} = (\dot{X}(t), \dot{X}(t)^\perp)\Big|_{t=0} I_{n,p} = C(P)I_{n,p},$$

with $C(P) \in T_{(P,P^\perp)}SO(n)$, (assuming $\dot{X}^\perp(0) = P^\perp$). Now Proposition **??** guarantees the existence of an $A(P) \in \mathfrak{so}(n)$ such that $C(P) = A(P) \cdot (P, P^\perp)$, which implies that

$$V = A(P) \cdot (P, P^\perp)I_{n,p} = A(P) \cdot P.$$

A direct consequence of this characterization of $\mathrm{T}_P\mathrm{St}(n,p)$ by means of $\mathfrak{so}(n)$ is that any differential equation on $\mathrm{St}(n,p)$ can be written in the form (1.13).

We can now apply Lie group integration methods to (1.13). The strategy is as usual to assume $Y(t) = \exp(\sigma(t))Y_0$ with $\sigma(t) \in \mathfrak{so}(n)$, and applying a classical Runge-Kutta method to the equation for σ, which also in this case is (1.11). Finally the RKMK methods applied to (1.13) assume exactly the same format as given in section 1.2.3, with the difference that the stage values now are $Y_i \in \mathrm{St}(n,p)$ while, as before, the σ_i belong to the Lie algebra.

The reader can show that the straightforward use of the commutator-free methods to (1.13) produces numerical approximations of the solution which belong to the Stiefel manifold.

Example 1.2.7. *Equation (1.7) is a differential equation on the sphere of radius* $\sqrt{\pi_0^T\pi_0}$ *in* $\mathbf{R}^3$. *The use of Lie group methods on this problem guarantees that the numerical solution* $\pi(t)$ *has constant Euclidean norm, i.e.* $\pi(t)^T\pi(t) = \pi_0^T\pi_0$, *for all* t.

Bibliografi

[CI1] E. Celledoni and A. Iserles, *Approximating the exponential from a Lie algebra to a Lie group.* Math. Comp. 69 (2000), no. 232, 1457–1480.

[CI2] E. Celledoni and A. Iserles, *Methods for the approximation of the matrix exponential in a Lie-algebraic setting.* IMA J. Numer. Anal. 21 (2001), no. 2, 463–488.

[CMO] E. Celledoni, A. Marthinsen and B. Owren, *Commutator-free Lie group methods*, FCGS, *19* (2003), 341–352.

[CG] P.E. Crouch, and R. Grossman, *Numerical integration of ordinary differential equations on manifolds*, J. Nonlinear Sci. **3** (1993), 1–33.

[DLP] F. Diele and L. Lopez and R. Peluso *The Cayley Transform in the Numerical Solution of Unitary Differential Systems*, Journal of Appl. Num. Math., (1998) **8**: 317–334.

[HNW] E. Hairer, S.P. Nørsett and G. Wanner *Solving Ordinary Differential Equations I*, Springer series in Computational Mathematics, Springer, (2000), second edition.

[HLW] E. Hairer, C. Lubich and G. Wanner *Geometric Numerical Integration*, Springer series in Computational Mathematics, Springer, (2002), first edition.

[IZ] A. Iserles and A. Zanna *Efficient computation of the matrix exponential by Generalized Polar Decompositions.* SIAM J. Num. Anal., vol. 42, nr. 5, pp. 2218–2256, (2005).

[LS] D. Lewis and J.C. Simo *Conserving algorithms for the dynamics of Hamiltonian systems on Lie groups.* J. Nonlinear Sci. **4** (1994), 253–299.

[MK] H. Munthe-Kaas, *High order Runge-Kutta methods on manifolds*, Appl. Num. Math., **29** (1999), 115–127.

[MKO] H. Munthe-Kaas and B. Owren, *Computations in a free Lie algebra* R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci. **357** (1999), 957–981.

[Ol] P. Olver, *Equivalence Invariants and symmetries* Cambridge University Press, Cambridge, (1995).

[O] B. Owren, *Order conditions for commutator-fee Lie group methods* J. of Phys. A: Math. Gen., **39** (2006), 5585–5599.

[OM] B. Owren and A. Marthinsen, *Integration methods based on canonical coordinates of the second kind.* Numer. Math. (2001) **87**: 763–790.

[V] V.S. Varadarajan, An introduction to harmonic analysis on semisimple Lie groups. Cambridge Studies in Advanced Mathematics, 16. Cambridge University Press, Cambridge, 1989.

Kapittel 2

IMEX methods and exponential integrators

2.1 Integration methods and exponential integrators for PDEs

Consider a semi-linear PDE, our main example in this section is the viscous Burgers equation

$$u_t + uu_x = \nu u_{xx}, \qquad \begin{array}{rcccl} u(x,0) & = & u_0(x), & t \geq 0 & x \in [0,1], \\ u(0,t) & = & u(1,t) & = & 0. \end{array} \tag{2.1}$$

Another example is the KdV equation

$$u_t + uu_x = \nu u_{xxx}, \qquad \begin{array}{rcccl} u(x,0) & = & u_0(x), & t \geq 0 & x \in [0,2\pi], \\ u(0,t) & = & u(2\pi,t). & & \end{array} \tag{2.2}$$

Assume we discretize these equations with finite differences, finite elements or spectral elements in space and we generate a system of ODEs with n components of the type

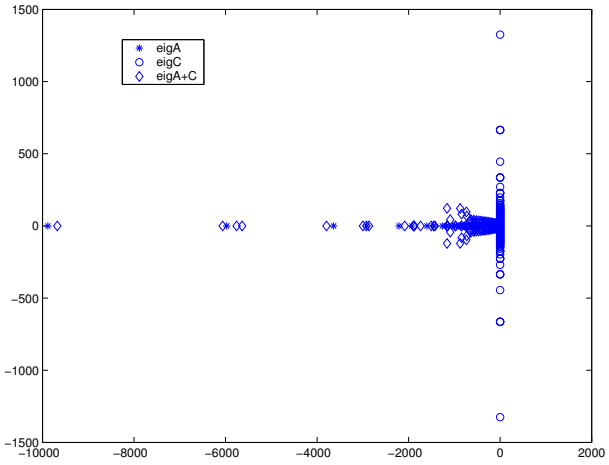
$$\dot{y} = N(y) + A \cdot y, \tag{2.3}$$

where $N(y)$ is a nonlinear term in the right hand side and $A \cdot y$ is a linear autonomous term. If, for example, we discretize using central finite differences in space, considering $n+2$ equispaced nodes in the space domain $[0,1]$ (including the boundary), we have

$$A = \frac{\nu}{\Delta x^2} \begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & \ddots & O & \\ & \ddots & \ddots & \ddots & \\ & O & \ddots & \ddots & 1 \\ & & & 1 & -2 \end{pmatrix},$$

and the coefficient $\frac{\nu}{\Delta x^2} = \nu(n+1)^2$, dictates the stiffness of the linear part of the problem. In particular A has real and negative eigenvalues $-4\frac{\nu}{\Delta x^2}\sin(\frac{k\pi}{2(n+1)})^2$ some are near the origin and some have big absolute value. The stiffness of the linear part of the problem grows when $\Delta x = \frac{1}{n+1}$ goes to zero.

The finite differences discretization of a convection term of the type $uu_x = \frac{1}{2}\frac{d}{dx}(u^2)$, would give a semidiscretized convection term in (2.3) of the type $D \cdot \mathrm{diag}(y) \cdot y = C(y) \cdot y$,



Figur 2.1: Discretization of the convection C and diffusion A operators.Discretization with a spectral method. Polynomial of degree 64, Gauss Lobatto Legendre nodes, $\nu = 1$ diffusion dominated problem.

where

$$D = \frac{1}{\Delta x}\begin{pmatrix} 1 & & & \\ -1 & 1 & & O \\ & \ddots & \ddots & \\ & O & \ddots & \ddots \\ & & & -1 \quad 1 \end{pmatrix}, \quad \text{or} \quad D = \frac{1}{2\Delta x}\begin{pmatrix} 0 & 1 & & \\ -1 & 0 & \ddots & O \\ & \ddots & \ddots & \ddots \\ & O & \ddots & \ddots \quad 1 \\ & & & -1 \quad 0 \end{pmatrix}.$$

Here we have used upwinding in the first case and central differences in the second case. The first operator has eigenvalues all equal to $\frac{a}{\Delta x}$, the second operator is skew-symmetric (as the continuous convection operator) and its eigenvalues are pure imaginary and appear in complex conjugate couples on the imaginary axis.

An appropriate modification of the above discretized linear convection operator leads to the semidiscretization of the nonlinear convection uu_x in (2.1) giving

$$N(y) = C(y) \cdot y, \quad C(y) = \text{diag}(y)C,$$

where $\text{diag}(y)$ denotes the diagonal matrix with the entries of y on the diagonal. In figure 2.1 we report the spectrum of A, C, (linear case), and $A + C$ for a discretization in space of (2.1) obtained using spectral methods, with $\nu = 1$.

The spectrum of A is on the real axis and the spectrum of C is on the imaginary axis. The minimal rectangle including all the eigenvalues of A and C contains also the spectrum of $A + C$. We shall consider numerical methods for the integration of these problems using separate strategies for the convection and the diffusion terms in the equation.

Let us consider the implicit Euler method for the integration of (2.3) in time. We have

$$y_{n+1} = y_n + hAy_n + hN(y_{n+1}) \Rightarrow (I - hA)y_{n+1} - hN(y_{n+1}) = y_n,$$

where I is the $n \times n$ identity matrix. The obtained expression is implicit in y_{n+1}, this requires the use of a Newton iteration in the implementation.

An alternative to this method is the following strategy

$$y_{n+1} = y_n + hAy_{n+1} + hN(y_n) \Rightarrow y_{n+1} = (I - hA)^{-1}(y_n + hN(y_n)). \tag{2.4}$$

The implementation of this method requires the solution of a linear system per time step. The computational burden per time step is much lower now than for the previous method.

Exercise 2.1.1. *Prove that the method (2.4) has local order* 2.

The method (2.4) belongs to the class of implicit-explicit methods (IMEX).

2.1.1 Higher order IMEX methods

We will now outline a systematic procedure to generate high order IMEX methods. Assume

$$y(t) = W(t)z(t), \tag{2.5}$$

where $W(t) \in M^{n\times n}$ is satisfying the differential equation

$$\dot{W} = AW, \;\; W(0) = I,$$

and $z(t) \in \mathbf{R}^n$ with $z(0) = y_0$. By differentiation with respect to t we obtain

$$\dot{W}z + W\dot{z} = \dot{y} \Rightarrow AWz + W\dot{z} = AWz + N(Wz),$$

which leads to the system of equations

$$\begin{array}{rclcrcl} \dot{W} & = & AW, & & W(0) & = & I, \\ \dot{z} & = & W^{-1}N(Wz), & & z(0) & = & y_0. \end{array} \tag{2.6}$$

The system (2.6) is naturally partitioned into a diffusion part (the first matricial equation) and a convection part. We can now apply a partitioned method using an implicit method for the linear diffusion part and an explicit method for the nonlinear convection part.

If we for example apply an implicit Euler method for the diffusion and an explicit Euler method for the convection we obtain

$$W_1 = (I - hA)^{-1}, \quad z_1 = z_0 + hW_0^{-1}N(y_0),$$

which by taking $y_1 = W_1 z_1$ gives the method (2.4). Note that if $W_1 = W(h) + \mathcal{O}(h^{p+1})$ and $z_1 = z(h) + \mathcal{O}(h^{p+1})$ one has that

$$y_1 = W_1 z_1 = W(h)z(h) + \mathcal{O}(h^{p+1}) = y(h) + \mathcal{O}(h^{p+1}).$$

This proves that the simple IMEX method (2.4) is a first order method. Therefore a partitioned method of order p for (2.6) generates a method $y_n = W_n z_n \approx y(t_n)$ which is also of order p for (2.3). Note that in this case not all the extra order conditions for Partitioned Runge-Kutta methods must be satisfied. In the system (2.6) the equation for W can be solved independently from the equation for z, but vice versa the numerical integration of z requires the use of the approximation of $W(c_i h)$ to a sufficiently high order of accuracy.

We now construct a method of order 2. We can consider the partitioned method of order 2 defined by the following Butcher tableaus

$$\begin{array}{c|cc} 0 & 0 & \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array} \qquad \begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} \end{array}.$$

Applying this method on the partitioned system and then defining $y_1 = W_1 z_1$ one obtains the second order method

$$\begin{aligned} Y_{1/2} &= (I - \tfrac{h}{2}A)^{-1}(y_0 + \tfrac{h}{2}N(y_0) \\ y_1 &= (I + \tfrac{h}{2}A)(I - h/2A)^{-1}y_0 + h(I + \tfrac{h}{2}A)N(Y_{1/2}). \end{aligned}$$

The following IMEX method of order 2 has been proposed by Ascher and collaborators [ARS]:

$$\begin{array}{c|cc} 0 & & \\ \gamma & \gamma & \\ 1 & \delta & 1-\delta \\ \hline & 1 & 1-\gamma \quad \gamma \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & & \\ \gamma & 0 & \gamma & \\ 1 & 0 & 1-\gamma & \gamma \\ \hline & 0 & 1-\gamma & \gamma \end{array},$$

with $\gamma = \frac{2-\sqrt{2}}{2}$ and $\delta = -\frac{2\sqrt{2}}{3}$. The implicit method to the right is stiffly accurate and L-stable. An IMEX method of order 3 can be derived by starting from the partitioned method

$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{2} & \frac{1}{2} & & \\ 1 & -1 & 2 & \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{2}{3} \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & & \\ \frac{1}{2} & -\frac{\beta}{2} & \frac{1+\beta}{2} & \\ 1 & \frac{3+5\beta}{2} & -(1+3\beta) & \frac{1+\beta}{2} \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{2}{3} \end{array},$$

with $\beta = \frac{\sqrt{3}}{3}$ the method has order 3 and the implicit method on the right is A-stable, this method has been proposed by Griepentrog in 1978.

Another example of a method of order three is given by the following two methods:

$$\begin{array}{c|ccc} 0 & & & \\ \gamma & \gamma & & \\ 1-\gamma & \gamma-1 & 2(1-\gamma) & \\ \hline & 0 & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & & \\ \gamma & 0 & \gamma & \\ 1-\gamma & 0 & 1-2\gamma & \gamma \\ \hline & 0 & \frac{1}{2} & \frac{1}{2} \end{array},$$

with $\gamma = \frac{3+\sqrt{3}}{6}$, [ARS].

Exercise 2.1.2. *Consider the following IMEX method based on the combination of the implicit trapezoidal rule and an Adams Bashford method of order* 2.

$$y_{n+1} = (I - \frac{h}{2}A)^{-1}(I + \frac{h}{2}A)y_n + \frac{h}{2}(I - \frac{h}{2}A)^{-1}(3N(y_n) - N(y_{n-1})).$$

Prove that the method has order 2.

More information on IMEX methods based on multi-step formulae can be found in [ARW].

2.1.2 Exponential integrators

Consider the partitioned system (2.6) equivalent to (2.3), we want now to use the exact integration of the matricial equation in the system as a part of the methods. Since this equation is linear with constant coefficients we can take for example

$$W_1 = W(h) = \exp(hA),$$

which combined with an explicit Euler method for the second part of the system gives

$$z_1 = z_0 + hW(0)^{-1}N(y_0) = y_0 + hN(y_0), \tag{2.7}$$

and finally

$$y_1 = W_1 z_1 = \exp(hA)(y_0 + hN(y_0)).$$

This method has order 1 and is known as Lawson-Euler method. Following a strategy similar to the one considered in the previous section, one can construct higher order methods. These methods are called integrating factor methods. These methods and the IMEX methods fit in the more general setting of exponential integrators whose format is reported here below.

Exponential integrators

for $i = 1 : s$ do

$Y_i = \tilde{\varphi}_i(hA)y_n + h\sum_{j=1}^{i} a_{i,j}(hA)N(Y_j)$

end

$y_{n+1} = \tilde{\varphi}_{s+1}(hA)y_n + h\sum_{i=1}^{s} b_i(hA)N(Y_i)$.

In the method $a_{i,j}(z)$, $b_i(z)$, $\tilde{\varphi}_i(z)$ are analytic functions of z, with $z = hA$. We require that $a_{i,j}(0) = a_{i,j}$ $b_i(0) = b_i$ and $\varphi_i(0) = 1$, where $a_{i,j}$ and b_i are the coefficients of a classical Runge-Kutta method. This guarantees that when $A = O$, i.e. $z = 0$, the resulting method is a classical Runge-Kutta method applied to the nonlinear part of (2.3). We can represent each exponential integrator by using a Butcher-like tableau as follows

$$\begin{array}{c|ccc|c} c_1 & a_{1,1}(z) & \dots & a_{1,s}(z) & \tilde{\varphi}_1(z) \\ \vdots & \vdots & \dots & \vdots & \vdots \\ c_s & a_{s,1}(z) & \dots & a_{s,s}(z) & \tilde{\varphi}_s(z) \\ \hline & b_1(z) & \dots & b_s(z) & \tilde{\varphi}_{s+1}(z) \end{array}.$$

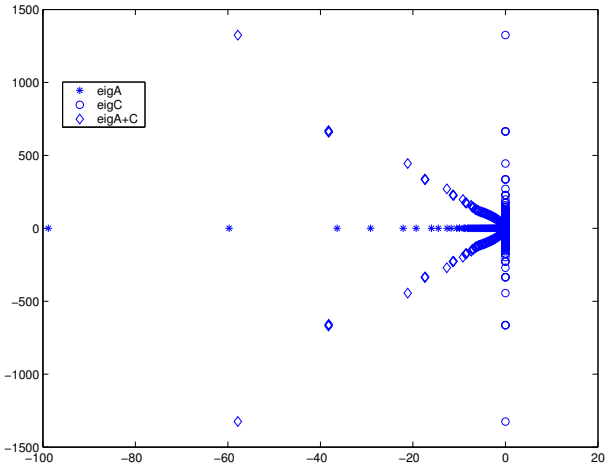
The order conditions for exponential integrators can be derived directly from the general format for the methods, for more details on this subject see [BOS] and [HO], where the convergence of the methods for semi-linear parabolic partial differential equations is also studied.

We give here to examples of exponential integrators of order 4. The first example is due to Cox and Matthews, they called their class of exponential integrators exponential time differencing methods (ETD) [CMt]:

$$\begin{array}{c|cccc|c} 0 & & & & & 1 \\ \frac{1}{2} & \frac{1}{2}\varphi_1(\frac{z}{2}) & & & & e^{z/2} \\ \frac{1}{2} & 0 & \frac{1}{2}\varphi_1(\frac{z}{2}) & & & e^{z/2} \\ 1 & \frac{1}{2}\varphi_1(\frac{z}{2})(e^{z/2}-1) & 0 & \frac{1}{2}\varphi_1(\frac{z}{2}) & & e^{z} \\ \hline & \varphi_1 - 3\varphi_2 + 4\varphi_3 & 2\varphi_2 - 4\varphi_3 & 2\varphi_2 - 4\varphi_3 & -\varphi_2 + 4\varphi_3 & e^{z} \end{array}.$$

where

$$\varphi_l(z) = \frac{1}{(l-1)!}\int_0^1 e^{(1-\theta)z}\theta^{l-1}\,d\theta, \quad \varphi_{l+1}(z) = \frac{\varphi_l(z) - \frac{1}{l!}}{z},\ \varphi_l(0) = \frac{1}{l!},$$



Figur 2.2: Discretization of the convection C and diffusion A operators.Discretization with a spectral method. Polynomial of degree 64, Gauss Lobatto Legendre nodes, $\nu = 0.01$ convection dominated problem.

$\varphi_0(z) = e^z$. The second example is a method derived by Lawson

$$\begin{array}{c|cccc|c} 0 & & & & & 1 \\ \frac{1}{2} & \frac{1}{2}e^{(\frac{z}{2})} & & & & e^{z/2} \\ \frac{1}{2} & 0 & \frac{1}{2} & & & e^{z/2} \\ 1 & 0 & 0 & e^{(\frac{z}{2})} & & e^z \\ \hline & \frac{1}{6}e^z & \frac{1}{3}e^{\frac{z}{2}} & \frac{1}{3}e^{\frac{z}{2}} & \frac{1}{6} & e^z \end{array}.$$

For an overview and some history as well as more details on the relationships of exponential integrators and Lie group methods see [MW]. For a Matlab package implementing exponential integrators see [BSW], for results on Krylov subspace methods for the computation of the matrix exponential see for example [HL].

2.1.3 Methods for convection dominated problems

When the viscosity parameter in (2.1) is small, the convection term prevails compared to the diffusion term. In figure 2.1.3 we can see how the eigenvalues of the operators A, C (linear convection) and $A + C$ are placed in the complex plane for a convection dominated problem, see figure 2.1 for comparison. In this case it is useful to consider methods which allow for the accurate integration of the convection part rather than the diffusion. Assume the nonlinear term can be written in the form

$$N(y) = C(y) \cdot y,$$

and

$$y(t) = W(t)z(t), \tag{2.8}$$

where $W(t) \in M^{n\times n}$ is satisfying the differential equation

$$\dot{W} = C(y)W, \;\; W(0) = I,$$

and $z(t) \in \mathbf{R}^n$ with $z(0) = y_0$. By differentiation with respect to t we obtain

$$\dot{W}z + W\dot{z} = \dot{y} \Rightarrow C(Wz)Wz + W\dot{z} = AWz + C(Wz)Wz,$$

which leads to the system of equations

$$\begin{array}{rclcrcl} \dot{W} & = & C(Wz)W, & & W(0) & = & I, \\ \dot{z} & = & W^{-1}AWz, & & z(0) & = & y_0. \end{array} \tag{2.9}$$

Using a Lie-Euler method for the convection and an implicit Euler for the diffusion we obtain the following first order method

$$y_{n+1} = \exp(hC(y_n))y_n + hAy_{n+1}. \tag{2.10}$$

To generalize this procedure to higher order one can consider a combination of a Lie group or Commutator-Free method for the convection and an implicit method for the diffusion. We can for example obtain a second order method using the partitioned Runge-Kutta method of order 2 already considered in subsection 2.1.1, i.e.

$$\begin{array}{c|cc} 0 & 0 & \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array} \qquad \begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} \end{array}.$$

We obtain

$$\begin{array}{rcl} y_0 & = & W_0 z_0 \\ W_{\frac{1}{2}} & = & \exp(\frac{h}{2}C(y_0)) \\ Z_{\frac{1}{2}} & = & y_0 + \frac{h}{2}W_{\frac{1}{2}}^{-1}AW_{\frac{1}{2}}Z_{\frac{1}{2}} \\ W_1 & = & \exp(hC(W_{\frac{1}{2}}Z_{\frac{1}{2}})) \\ z_1 & = & y_0 + \frac{h}{2}W_{\frac{1}{2}}^{-1}AW_{\frac{1}{2}}Z_{\frac{1}{2}}, \end{array}$$

which, setting $Y_{\frac{1}{2}} = W_{\frac{1}{2}}Z_{\frac{1}{2}}$ and $y_1 = W_1 z_1$, becomes

$$\begin{array}{rclcrcl} W_{\frac{1}{2}} & = & \exp(\frac{h}{2}C(y_0)) & & Y_{\frac{1}{2}} & = & W_{\frac{1}{2}}y_0 + \frac{h}{2}AY_{\frac{1}{2}} \\ W_1 & = & \exp(\frac{h}{2}C(Y_{\frac{1}{2}})) & & y_1 & = & W_1 y_0 + hW_1 W_{\frac{1}{2}}^{-1}AY_{\frac{1}{2}}. \end{array}$$

We will discuss how to achieve an accurate approximation of terms of the type $W_i\, y_0$ at the end of this subsection.

Operator integrating factor splitting method

Another way of achieving higher order is to linearize the problem using extrapolation an then use an implicit method, say a BDF formula, for the diffusion and an explicit Runge-Kutta method of high order for the convection in (2.9), see [MPR] for details. This approach is known as the Operator Integrating Factor (OIF) Splitting method. We here outline this procedure for order 2.

Assume we are given the values of the numerical solution $y(0) = y_0$ and $y_1 = y(h) + \mathcal{O}(h^3)$. Consider the polynomial of degree 1

$$p_1(t) = -\frac{t - t_1}{h}y_0 + \frac{t - t_0}{h}y_1,$$

with $t_0 = 0$, $t_1 = h$, $p_1(t_0) = y_0$ and $p_1(t_1) = y_1$. We consider the linearized system of equations

$$\begin{array}{rclcrcl} \dot{\tilde{W}} & = & C(p_1(t))\tilde{W}, & & \tilde{W}(0) & = & I, \\ \dot{\tilde{z}} & = & \tilde{W}^{-1}A\tilde{W}\tilde{z}, & & \tilde{z}(0) & = & y_0. \end{array} \tag{2.11}$$

Note that the semi-discretized convection operator $C(y)$ is in some cases a linear function of y, (see for instance the case of finite differences discretizations considered at the beginning of this section).

Using the Taylor expansion of W solution of (2.9) and $\tilde{W}$ (2.11), and assuming for simplicity the linearity of $C(y)$ as function of y, we have

$$W(h)-\tilde{W}(h) = (I-I)+h(C(y_0)-C(p_1(0)))+\frac{h^2}{2!}(C(\dot{y}(0))+C(y_0)^2-C(\dot{p}_1(0))-C(y_0)^2)+\mathcal{O}(h^3),$$

which gives

$$W(h) - \tilde{W}(h) = \frac{h^2}{2!}(C(\dot{y}(0) - \frac{y_1 - y_0}{h})) + \mathcal{O}(h^3).$$

By expanding $y(h)$ in a Taylor series around zero we obtain

$$\dot{y}(0) = \frac{y(h) - y_0}{h} + \mathcal{O}(h^2) = \frac{y(h) - y_1 + y_1 - y_0}{h} + \mathcal{O}(h^2) = \frac{y_1 - y_0}{h} + \mathcal{O}(h^2),$$

and as a consequence $W(h) - \tilde{W}(h) = \mathcal{O}(h^3)$ and similarly $W(2h) - \tilde{W}(2h) = \mathcal{O}(h^3)$, $z(h) - \tilde{z}(h) = \mathcal{O}(h^3)$ and $z(2h) - \tilde{z}(2h) = \mathcal{O}(h^3)$. We conclude that the solution of (2.11) is an approximation of order 3 in h for the solution of (2.9) on the interval $[0, 2h]$. If we apply a numerical integration method of order 2 to (2.11) we obtain an approximation of (2.9) of the same order of accuracy.

In the system (2.11) the equation for $\tilde{W}$ can be solved independently form the equation for $\tilde{z}$ and we assume to compute $\tilde{W}(c_i h)$ for opportune c_i very accurately. We consider a BDF method of order 2 for the integration of the equation for z in (2.11), we obtain

$$\frac{3}{2}\tilde{z}_2 = 2\tilde{z}_1 - \frac{1}{2}\tilde{z}_0 + h\tilde{W}(2h)^{-1}A\tilde{W}(2h)\tilde{z}_2,$$

and by setting $\tilde{y}_2 = \tilde{W}(2h)\tilde{z}_2$ we get

$$\frac{3}{2}\tilde{y}_2 = 2\tilde{W}(h)y_1 - \frac{1}{2}\tilde{W}(2h)y_0 + hA\tilde{y}_2.$$

In the BDF2-OIF method above we need to compute $\tilde{W}(h)y_1$ and $\tilde{W}(2h)y_0$. Each of these problems corresponds to the solution of a differential equation of the type

$$\dot{v} = C(p_1(t))v, \quad v(0) = b, \tag{2.12}$$

where $b = y_0$ and $t \in [0, 2h]$ or $b = y_1$ and $t \in [0, h]$. Equation (2.12) is a semi-discretized pure convection problem arising from a PDE of the type

$$u_t + v(x,t)u_x = 0, \quad x \in [0,1], \tag{2.13}$$

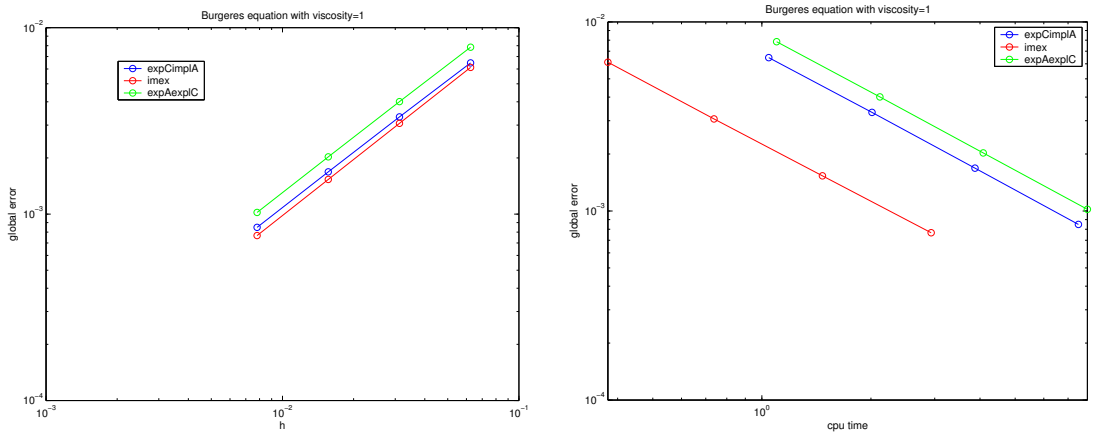
with $u(x,0) = \beta(x)$ and $\beta(x_i) = b_i$ and $v(x_i,t) = p_{1,i}(t)$, where x_i are the nodes on the grid of the chosen discretization and $p_{1,i}$ is the i-th component of the polynomial p_1.

Such problems can be solved computing characteristics. In fact we have that the solution of (2.13) is

$$u(x,t) = u(X(t),0) = \beta(X(t)),$$

where $X(t)$ is the solution of the differential equation

$$\begin{array}{lcl} \dot{X} & = & -v(X(t),t) \\ X(0) & = & x, \end{array} \tag{2.14}$$



Figur 2.3: Discretization of the Burgers equation with a spectral method. Polynomial of degree 64, Gauss Lobatto Legendre nodes. Methods: imex (red), expAexplC (green) and expCimplA (blue). $\nu = 1$. Global error y-axis versus step size x-axis (left). Global error y-axis versus CPU time x-axis (right). The imex is the best in this case.

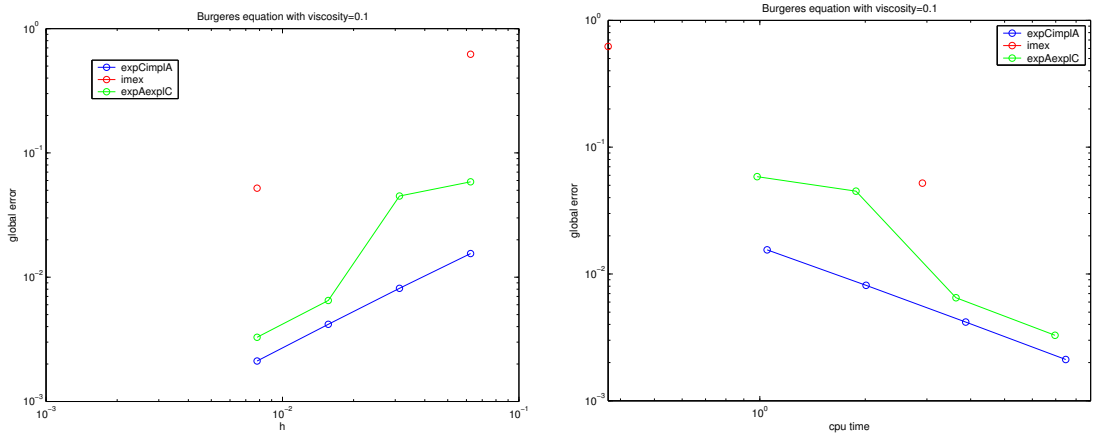
for any $x \in [0, 1]$, (this can be shown by differentiating with respect to t the solution $u(X(t), 0)$). Therefore the accurate numerical approximation of (2.12) can be done by the direct numerical integration of the system with a Runge-Kutta method of high order, (Runge-Kutta 4 is a good choice), or by the accurate numerical approximation of the characteristic trajectories satisfying (2.14). The second choice gives rise to a semi-Lagrangian method. For related topics see also [Ce].

2.1.4 Numerical comparison of the various methods of order 1

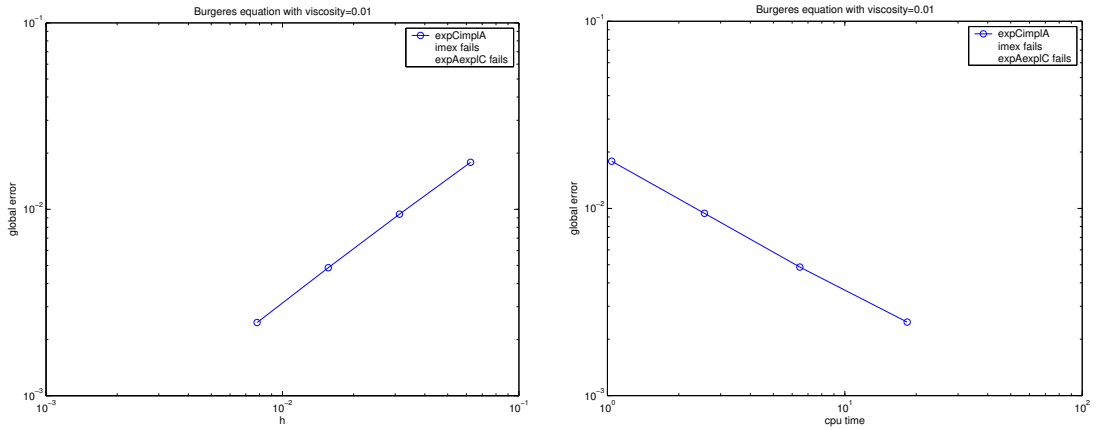
We report here the comparison of the different described approaches for the numerical integration of the Burgers equation (2.1) with different values of ν, ($\nu = 1$, diffusion dominated, $\nu = 0.1$ and $\nu = 0.01$, convection dominated). We consider spectral methods for the discretization in space with a polynomial of degree 64 based on Gauss Lobatto Legendre points.

In the first three experiments we compare the IMEX method (2.4), imex in the plots, with the exponential integrators (2.7) and (2.10), expAexplC and expCimplA respectively in the plots.

We present plots of the global error versus the time step, for different step sizes, and global error versus the CPU time for the same experiments.



Figur 2.4: Discretization of the Burgers equation with a spectral method. Polynomial of degree 64, Gauss Lobatto Legendre nodes. Methods: imex (red), expAexplC (green) and expCimplA (blue). $\nu = 0.1$. Global error y-axis versus step size x-axis (left). Global error y-axis versus CPU time x-axis (right). The expCiplA is the best in this case.



Figur 2.5: Discretization of the Burgers equation with a spectral method. Polynomial of degree 64, Gauss Lobatto Legendre nodes. Methods: expCimplA (blue) the other methods fail on this experiment. $\nu = 0.01$. Global error y-axis versus step size x-axis (left). Global error y-axis versus CPU time x-axis (right). The expCimplA is the best in this case.

Bibliografi

[ARS] U. M. Ascher, S. J. Ruuth and R. Spiteri *Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations.* Appl. Numer. Math., **25** (1997), 151–167.

[ARW] U. M. Ascher, S. J. Ruuth and B. T. R. Wetton, *Implicit-explicit methods for time-dependent partial differential equations*, SIAM J. Num. Anal., **32** (1995), 797–823.

[BOS] H. Berland, B. Owren and B. Skaflestad *B-series and order conditions for exponential integrators*, SIAM J. Numer. Anal. **43** (2005), 1715–1727.

[BSW] H. Berland, B. Skaflestad, W. Wright *EXPINT – A MATLAB package for exponential integrators*, Preprint in numerics, 4/2005, Department of Mathematical Sciences, NTNU (to appear in ACM/TOMS).

[CI1] E. Celledoni and A. Iserles, *Approximating the exponential from a Lie algebra to a Lie group.* Math. Comp. 69 (2000), no. 232, 1457–1480.

[CI2] E. Celledoni and A. Iserles, *Methods for the approximation of the matrix exponential in a Lie-algebraic setting.* IMA J. Numer. Anal. 21 (2001), no. 2, 463–488.

[CMO] E. Celledoni, A. Marthinsen and B. Owren, *Commutator-free Lie group methods*, FCGS, *19* (2003), 341–352.

[Ce] E. Celledoni, *Eulerian and semi-Lagrangian commutator-free exponential integrators*, to appear in CRM Proceedings vol. 39, *19* (2004).

[1] E. Celledoni and B. K. Kometa, *Semi-Lagrangian exponential integrators for convection dominated problems*, Journal of Scientific Computing, *41*, issue 1, p.139.

[CK] E. Celledoni and B. K. Kometa, *Order theory for semi-Lagrangian exponential integrators* , NTNU report nr 4, 2009.

[CMt] S.M. Cox and P.C. Matthews *Exponential time differencing for stiff systems*, J. of Comp. Phys. , **176** (2002), 430–455.

[CG] P.E. Crouch, and R. Grossman, *Numerical integration of ordinary differential equations on manifolds*, J. Nonlinear Sci. **3** (1993), 1–33.

[HNW] E. Hairer, S.P. Nørsett and G. Wanner *Solving Ordinary Differential Equations I*, Springer series in Computational Mathematics, Springer, (2000), second edition.

[HW] E. Hairer and G. Wanner *Solving Ordinary Differential Equations II*, Springer series in Computational Mathematics, Springer, (1996), second edition.

[HLW] E. Hairer, C. Lubich and G. Wanner *Geometric Numerical Integration*, Springer series in Computational Mathematics, Springer, (2002), first edition.

[HL] M. Hochbruck and C. Lubich *On Krylov subspace approximations of the matrix exponential operator*, SIAM J. Num. Anal., **34** (1997), 1911–1925.

[HLS] M. Hochbruck, C. Lubich and H. Selhofer *Exponential integrators for large systems of differential equations*, SIAM J. Sci. Comput., **19** (1998), 1552–1574.

[HO] M. Hochbruck and A. Ostermann *Explicit exponential Runge-Kutta methods for semilinear parabolic problems* , SIAM J. Num. Anal., **43** (2005), 1069–1090.

[INMZ] A. Iserles, H. Munthe-Kaas, S.P. Nørsett, and A. Zanna *Lie-group methods* , Acta Numerica, vol.9 215–365, (2000). Cambridge Univ. Press, Cambridge, 2000.

[IZ] A. Iserles and A. Zanna *Efficient computation of the matrix exponential by Generalized Polar Decompositions.* SIAM J. Num. Anal., vol. 42, nr. 5, pp. 2218–2256, (2005).

[KT] A. K. Kassam and L. N. Trefethen, *Fourth-order time stepping for stiff PDEs*, SIAM J. Sci. Comput. **26**, (2005) 1214-1233.

[MPR] Y. Maday, A. T. Patera and E. M. Rønquist, *An operator integration factor splitting method for time dependent problems: Application to incompressible fluid flows*, J. of Sci. Comp., **5** (1990), 263–292.

[MW] B.V. Minchev and W. M. Wright, *A review of exponential integrators for first order semi-linear problems*, Preprint in numerics 2/2005, Department of Mathematical Sciences, NTNU.

[MK] H. Munthe-Kaas, *High order Runge-Kutta methods on manifolds*, Appl. Num. Math., **29** (1999), 115–127.

[MKO] H. Munthe-Kaas and B. Owren, *Computations in a free Lie algebra* R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci. **357** (1999), 957–981.

[Ol] P. Olver, *Equivalence Invariants and symmetries* Cambridge University Press, Cambridge, (1995).

[O] B. Owren, *Order conditions for commutator-fee Lie group methods* J. of Phys. A: Math. Gen., **39** (2006), 5585–5599

[Pi] O. Pirroneau, *On the transport-diffusion algorithm and its applications to the Navier-Stokes equations* , Numer. Math. 38, 309-332 (1982).

[XK] D. Xiu and G.E. Karniadakis, *A semi-Lagrangian high-order method for Navier-Stokes equations*, J. of Comput. Phys. 172, 658-684 (2001).

Kapittel 3

Energy preserving methods and multi-symplectic methods

3.1 Introduction

In chapter VI of the Geometric Numerical Integration book, [7], you find the treatment of the preservation of invariants in the numerical integration of ODEs. An invariant (first integral) of the flow of a differential equation

$$\dot{y} = f(y), \quad y(0) = y_0 \in \mathbf{R}^n,$$

is a function $I : \mathbf{R}^n \to \mathbf{R}$ which is constant along solutions of the differential equations i.e.

$$\frac{d\, I(y(t))}{dt} = \nabla\, I(y(t))^T f(y(t)) = 0, \quad \forall t > 0.$$

We distinguish between linear, quadratic and polynomial invariants of degree higher than 2. Linear invariants are of the type

$$I(y) = d^T y, \quad d \in \mathbf{R}^n.$$

Quadratic invariants are of the type

$$Q(y) = y^T C y, \quad C = C^T.$$

An example of a polynomial invariant of degree higher than 2 is the determinant of a matrix valued solution of a matricial ODE of the type

$$\dot{y} = A(y)y, \quad y(0) = I, \quad \text{trace}(A) = 0, \quad \det(y) = 1,$$

and A $n \times n$, $n \geq 3$.

It is possible to show that all Runge-Kutta methods preserve all linear invariants, and that a special class of Runge-Kutta methods preserve any quadratic invariant, see [7].

But no Runge-Kutta method can preserve all cubic/polynomial invariants.

Example 3.1.1. *We consider the following simple ODE*

$$\begin{aligned} \dot{x} &= x \\ \dot{y} &= y \\ \dot{z} &= -2z \end{aligned}$$

with initial value $x(0) = x_0$, $y(0) = y_0$, $z(0) = z_0$. *We observe that*

$$x(t)y(t)z(t) = e^t x_0 e^t y_0 e^{-2t} z_0 = x_0 y_0 z_0$$

is a cubic invariant of the flow. The equation can be written in the form

$$\dot{\mathbf{x}} = A\,\mathbf{x}, \quad \mathbf{x}(0) = \mathbf{x}_0,$$

where A *is the diagonal matrix with diagonal entries* $1, 1, -2$. *Applying a Runge-Kutta method to this problem amounts to compute*

$$\mathbf{x}_1 = R(hA)\mathbf{x}_0,$$

where $R(z)$ *is the stability function of the Runge-Kutta method,* $R(0) = 1$ *and* $R'(0) = 1$, *and* $R(z)$ *is a rational function. In particular multiplying together the components of* $\mathbf{x}_1$ *we obtain*

$$x_1 y_1 z_1 = R(h)^2 R(-2h) x_0 y_0 z_0.$$

The invariant is preserved by the Runge-Kutta method if and only if

$$R(h)^2 R(-2h) = 1, \quad \forall h.$$

Now we observe that this can happen only if $R(z) = e^z$, *in fact: assume*

$$\psi(z) := \log(R(z)) = \sum_{k=0}^{\infty} \psi_k\, z^k,$$

so $R(z) = e^{\psi(z)}$. *By the properties of logarithms we have*

$$\log(R(z)^2 R(-2z)) = 2\log(R(z)) + \log(R(-2z)) = 2\psi(z) + \psi(-2z) = 0.$$

So expanding the series and collecting terms we get

$$2\psi(z) + \psi(-2z) = \sum_{k=0}^{\infty} \psi_k\, z^k (2 - 2^k) = 0,$$

since $(2 - 2^k) \neq 0$ *for* $k \geq 2$, ψ_k *must be zero for* $k \geq 2$. *Also since* $R(0) = 1 = e^{\psi(0)}$ $\psi_0 = 0$ *and*

$$\psi(z) = \psi_1\, z,$$

finally $R'(0) = \psi_1 e^{\psi_1 z}\big|_{z=0} = 1$ *gives* $\psi_1 = 1$ *so*

$$R(z) = e^z.$$

3.2 Discrete Gradients

We consider a general ODE problem with an invariant I. It is always possible to write the ODE in the following form

$$\dot{y} = f(y) = S(y)\nabla I(y), \quad y(0) = y_0, \tag{3.1}$$

where $S(y)$ is a skew-symmetric matrix. A choice for $S(y)$, under the assumption that $\nabla I(y)$ does not vanish, is

$$S(y) = \frac{1}{\|\nabla I\|_2}(f(y)\nabla I(y)^T - \nabla I(y) f(y)^T),$$

[9], see the same reference for a discussion on the boundedness of $S(y)$ defined above around non degenerate fixed points of $\nabla I(y)$, (i.e. equilibria of $f(y)$)????.

For problems reformulated in the form (3.1) it is is always possible to build an energy-preserving method by approximating ∇I by a so called discrete gradient.

Definition 3.2.1. *(Gonzalez 96, [6]).*
If I is differentiable then $\bar{\nabla} I : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}^n$ is a discrete gradient if I is continuous and

$$\begin{cases} \bar{\nabla} I(u,v)^T (v-u) &= I(v) - I(u) \\ \bar{\nabla} I(u,u) &= \nabla I(u) \end{cases}$$

Example 3.2.2. *Examples of discrete gradients*

1.
$$\bar{\nabla} I(u,v) = \int_0^1 \nabla I\left((1-\xi)u + \xi v\right) d\xi$$
is the average vector field discrete gradient;

2.
$$\bar{\nabla} I(u,v) = \nabla I(\frac{u+v}{2}) + \frac{I(v) - I(u) - \nabla I(\frac{u+v}{2}^T (v-u)}{\|v-u\|^2}(v-u),$$
this discrete gradient is due to Gonzalez.

Theorem 3.2.3. *A numerical integrator having the format*

$$y_{n+1} = y_n + h\bar{S}(y_n, y_{n+1})\bar{\nabla} I(y_n, y_{n+1}),$$

where $\bar{S}(y_n, y_{n+1}) \approx S(y)$ is a skew-symmetric matrix and $\bar{\nabla} I(y_n, y_{n+1}) \approx \nabla I$ in a neighborhood of y_n, is an energy-preserving integrator.

Proof Using the definition of discrete gradient and the skew-symmetry of $\bar{S}(y_n, y_{n+1})$ we get

$$I(y_{n+1}) - I(y_n) = \bar{\nabla} I(y_n, y_{n+1})^T (y_{n+1} - y_n) = h\bar{\nabla} I(y_n, y_{n+1})^T \bar{S}(y_n, y_{n+1})\bar{\nabla} I(y_n, y_{n+1}) = 0.$$

Example 3.2.4. *The method*

$$y_{n+1} = y_n + hS(\frac{y_n + y_{n+1}}{2}) \int_0^1 \nabla I\left((1-\xi)y_n + \xi y_{n+1}\right) d\xi,$$

fits the framework of the previous theorem and is an energy preserving method. By using Taylor expansion it is possible to show that this method has order 2.

3.3 Preservation of the energy of ODEs in canonical Hamiltonian form

Consider the ODE $\dot{y} = f(y)$, $y(0) = y_0$, we call the integration method

$$\frac{y_{n+1} - y_n}{\Delta t} = \int_0^1 f((1-\xi)y_n + \xi y_{n+1})\, d\xi, \tag{3.2}$$

"Average Vector Field method", [13].
Consider canonical Hamiltonian systems in the form

$$\dot{y} = f(y) = J^{-1} \nabla H(y), \quad y(0) = y_0. \tag{3.3}$$

Theorem 3.3.1. *Let y_n be the solution of the average vector field (AVF) method (3.17) applied to equation (3.3). Then the energy H is preserved exactly :*

$$H(y_{n+1}) = H(y_n).$$

Proof H is preserved since

$$\dot{H} = (\nabla H)^T J^{-1} \nabla H = 0. \tag{3.4}$$

It is possible to prove that this method has order 2 by usual Taylor expansion.

Corollary 3.3.2. *Given $b_1, \dots, b_s$ and $c_1, \dots, c_s$ defining a quadrature formula of polynomial order $m-1$. Consider an ODE of the type (3.3) where H is a polynomial of degree m. Then the s-stages Runge-Kutta method*

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i \, f(y_n + (y_{n+1} - y_n)c_i)$$

preserves H and has order $\min(m, 2)$.

Proof For polynomial Hamiltonians of degree m, $f = J^{-1}\nabla H$ has polynomial components of degree at most $m-1$ and

$$h \sum_{i=1}^{s} b_i \, f(y_n + (y_{n+1} - y_n)c_i) = \int_0^1 f((1-\xi)y_n + \xi y_{n+1}) \, d\xi.$$

So the given quadrature formula coincides with the exact integral, and the given method coincides with the AVF method.

3.4 Hamiltonian PDEs and preservation of energy

This part is taken from [5]. We consider evolutionary PDEs with independent variables $(x,t) \in \mathbf{R}^d \times \mathbf{R}$, functions u belonging to a Banach space $\mathcal{B}$ with values[1] $u(x,t) \in \mathbf{R}^m$, and PDEs of the form

$$\dot{u} = \mathcal{D} \frac{\delta \mathcal{H}}{\delta u}, \tag{3.5}$$

where $\mathcal{D}$ is a constant linear differential operator, the dot denotes $\frac{\partial}{\partial t}$, and

$$\mathcal{H}[u] = \int_\Omega H(x; u^{(n)}) \, dx \tag{3.6}$$

where Ω is a subset of $\mathbf{R}^d \times \mathbf{R}$, and $dx = dx_1 dx_2 \dots dx_d$. $\frac{\delta \mathcal{H}}{\delta u}$ is the variational derivative of $\mathcal{H}$ in the sense that

$$\frac{d}{d\epsilon} \mathcal{H}[u + \epsilon v]\big|_{\epsilon=0} = \int_\Omega \frac{\delta \mathcal{H}}{\delta u} v \, dx, \tag{3.7}$$

for all $u, v \in \mathcal{B}$ (cf. [12]). For example, if $d = m = 1$,

$$\mathcal{H}[u] = \int_\Omega H(x; u, u_x, u_{xx}, \dots) \, dx, \tag{3.8}$$

then

$$\frac{\delta \mathcal{H}}{\delta u} = \frac{\partial H}{\partial u} - \partial_x \left(\frac{\partial H}{\partial u_x} \right) + \partial_x^2 \left(\frac{\partial H}{\partial u_{xx}} \right) - \cdots, \tag{3.9}$$

when the boundary terms are zero.

[1] Although it is generally real-valued, the function u may also be complex-valued, for example, the nonlinear Schrödinger equation.

Similarly, for general d and m, we obtain

$$\frac{\delta \mathcal{H}}{\delta u_l} = \frac{\partial H}{\partial u_l} - \sum_{k=1}^{d} \frac{\partial}{\partial x_k}\left(\frac{\partial H}{\partial u_{l,k}}\right) + \ldots, \quad l = 1, \ldots, m. \tag{3.10}$$

We consider Hamiltonian systems of the form (3.5), where $\mathcal{D}$ is a constant skew symmetric operator (cf. [12]) and $\mathcal{H}$ the energy (Hamiltonian). In this case, we prefer to designate the differential operator in (3.5) with $\mathcal{S}$ instead of $\mathcal{D}$. The PDE preserves the energy because $\mathcal{S}$ is skew-adjoint with respect to the L_2 inner product, i.e.

$$\int_\Omega u\mathcal{S}u\, dx = 0, \quad \forall u \in \mathcal{B}. \tag{3.11}$$

The system (3.5) has $\mathcal{I} : \mathcal{B} \to \mathbf{R}$ as an integral if $\dot{\mathcal{I}} = \int_\Omega \frac{\delta \mathcal{I}}{\delta u}\mathcal{S}\frac{\delta \mathcal{H}}{\delta u}\, dx = 0$.
Integrals $\mathcal{C}$ with $\mathcal{D}\frac{\delta \mathcal{C}}{\delta u} = 0$ are called Casimirs.

Besides PDEs of type (3.5) where $\mathcal{D}$ is skew-adjoint, we also consider PDEs of type (3.5) where $\mathcal{D}$ is a constant negative (semi)definite operator with respect to the L_2 inner product, i.e.

$$\int_\Omega u\mathcal{D}u\, dx \leq 0, \quad \forall u \in \mathcal{B}. \tag{3.12}$$

In this case, we prefer to designate the differential operator $\mathcal{D}$ with $\mathcal{N}$ and the function $\mathcal{H}$ is a Lyapunov function, since then the system (3.5), i.e.

$$\dot{u} = \mathcal{N}\frac{\delta \mathcal{H}}{\delta u}, \tag{3.13}$$

has $\mathcal{H}$ as a Lyapunov function, i.e. $\dot{\mathcal{H}} = \int_\Omega \frac{\delta \mathcal{H}}{\delta u}\mathcal{N}\frac{\delta \mathcal{H}}{\delta u}\, dx \leq 0$. We will refer to systems (3.5) with a skew-adjoint $\mathcal{S}$ and an energy $\mathcal{H}$ as conservative and to systems (3.5) with a negative (semi)definite operator $\mathcal{N}$ and a Lyapunov function $\mathcal{H}$ as dissipative.

Conservative PDEs (3.5) can be semi-discretised in "skew-gradient" form

$$\dot{u} = \overline{\mathcal{S}}\nabla\overline{\mathcal{H}}(u), \qquad \overline{\mathcal{S}}^T = -\overline{\mathcal{S}}, \tag{3.14}$$

when $\mathcal{D} = \mathcal{S}$ is skew-adjoint. $u \in \mathbf{R}^k$, and here, and in the following, we will always denote the discretisations with bars. $\overline{\mathcal{H}}$ is chosen in such a way that $\overline{\mathcal{H}}\Delta x$ is an approximation to $\mathcal{H}$.

Lemma 3.4.1. *Let*

$$\mathcal{H}[u] = \int_\Omega H(x; u^{(n)})dx, \tag{3.15}$$

and let $\overline{\mathcal{H}}\Delta x$ be any consistent (finite difference) approximation to $\mathcal{H}$ (where $\Delta x := \Delta x_1 \Delta x_2 \ldots \Delta x_d$). Then the discrete analogue of the variational derivative $\dfrac{\delta \mathcal{H}}{\delta u}$ is given by $\nabla\overline{\mathcal{H}}$.

The proof is given in the appendix.

It is worth noting that the above lemma also applies directly when the approximation to $\mathcal{H}$ is obtained by a spectral discretization, since such an approximation can be viewed as a finite difference approximation where the finite difference stencil has the same number of entries as the number of grid points on which it is defined.

The operator ∇ is the standard gradient, which replaces the variational derivative because we are now working in a finite (although large) number of dimensions (cf. e.g. (3.10)).

When dealing with (semi-)discrete systems we use the notation $u_{j,n}$ where the index j corresponds to increments in space and n to increments in time. That is, the point $u_{j,n}$ is the discrete equivalent of $u(a + j\Delta x, t_0 + n\Delta t)$ where $x \in [a, b]$ and where t_0 is the initial time. In most of the equations we present, one of the indices is held constant, in which case, for simplicity, we drop it from the notation. For example, we use u_j to refer to the values of u at different points in space and at a fixed time level.

Theorem 3.4.2. *Let $\overline{\mathcal{S}}$ (resp. $\overline{\mathcal{N}}$) be any consistent constant skew (resp. negative-definite) matrix approximation to $\mathcal{S}$ (resp. $\mathcal{N}$). Let $\overline{\mathcal{H}}\Delta x$ be any consistent (finite difference) approximation to $\mathcal{H}$. Finally, let*

$$f(u) := \overline{\mathcal{S}}\nabla\overline{\mathcal{H}}(u) \qquad (\textit{resp. } f(u) := \overline{\mathcal{N}}\nabla\overline{\mathcal{H}}(u)), \tag{3.16}$$

and let u_n be the solution of the average vector field (AVF) method

$$\frac{u_{n+1} - u_n}{\Delta t} = \int_0^1 f((1-\xi)u_n + \xi u_{n+1})\, d\xi, \tag{3.17}$$

applied to equation (3.16). Then the semidiscrete energy $\overline{\mathcal{H}}$ is preserved exactly (resp. dissipated monotonically):

$$\overline{\mathcal{H}}(u_{n+1}) = \overline{\mathcal{H}}(u_n) \qquad (\textit{resp } \overline{\mathcal{H}}(u_{n+1}) \leq \overline{\mathcal{H}}(u_n)).$$

$\overline{\mathcal{H}}$ is preserved since

$$\dot{\overline{\mathcal{H}}} = \left(\nabla\overline{\mathcal{H}}\right)^T \overline{\mathcal{S}}\nabla\overline{\mathcal{H}} = 0. \tag{3.18}$$

Discretisations of this type can be given for pseudospectral, finite-element, Galerkin and finite-difference methods (cf. [10, 11]); for simplicity's sake, we will concentrate on finite-difference methods, though we include one example of a pseudospectral method for good measure.

The AVF method was recently [13] shown to preserve the energy $\overline{\mathcal{H}}$ exactly for any vector field f of the form $f(u) = \overline{\mathcal{S}}\nabla\overline{\mathcal{H}}(u)$, where $\overline{\mathcal{H}}$ is an arbitrary function, and $\overline{\mathcal{S}}$ is any **constant** skew matrix [2]. The AVF method is related to discrete gradient methods (cf. [9]).

If $\mathcal{D}$ is a constant negative-definite operator, then the dissipative PDE (3.5) can be discretized in the form

$$\dot{u} = \overline{\mathcal{N}}\nabla\overline{\mathcal{H}}(u), \tag{3.19}$$

where $\overline{\mathcal{N}}$ is a negative (semi)definite matrix and $\overline{\mathcal{H}}$ is a discretisation as above.

That is, $\overline{\mathcal{H}}$ is a Lyapunov-function for the semi-discretized system, since

$$\dot{\overline{\mathcal{H}}} = \left(\nabla\overline{\mathcal{H}}\right)^T \overline{\mathcal{N}}\nabla\overline{\mathcal{H}} \leq 0. \tag{3.20}$$

The AVF method (3.17) again preserves this structure, i.e. we have

$$\overline{\mathcal{H}}(u_{n+1}) \leq \overline{\mathcal{H}}(u_n), \tag{3.21}$$

and $\overline{\mathcal{H}}$ is a Lyapunov function for the discrete system. Taking the scalar product of (3.17) with $\int_0^1 \nabla\overline{\mathcal{H}}((1-\xi)u_n + \xi u_{n+1})\, d\xi$ on both sides of the equation yields

$$\frac{1}{\Delta t}\int_0^1 (u_{n+1} - u_n) \cdot \nabla\overline{\mathcal{H}}((1-\xi)u_n + \xi u_{n+1})\, d\xi \leq 0, \tag{3.22}$$

i.e.

$$\frac{1}{\Delta t}\int_0^1 \frac{d}{d\xi}\overline{\mathcal{H}}((1-\xi)u_n + \xi u_{n+1})\, d\xi \leq 0, \tag{3.23}$$

and therefore

$$\frac{1}{\Delta t}(\overline{\mathcal{H}}(u_{n+1}) - \overline{\mathcal{H}}(u_n)) \leq 0. \tag{3.24}$$

Our purpose is to show that the procedure described above, namely

[2] The relationship of (3.17) to Runge-Kutta methods was explored in [4].

1. Discretize the energy functional $\mathcal{H}$ using any (consistent) approximation $\overline{\mathcal{H}}\Delta x$
2. Discretize $\mathcal{D}$ by a constant skew-symmetric (resp. negative (semi)definite) matrix
3. Apply the AVF method

can be generally applied and leads, in a systematic way, to energy-preserving methods for conservative PDEs and energy-dissipating methods for dissipative PDEs. We shall demonstrate the procedure by going through several well-known nonlinear and linear PDEs step by step. In particular we give examples of how to discretise nonlinear conservative PDEs (in subsection 2.1), linear conservative PDEs (in subsection 2.2), nonlinear dissipative PDEs (in subsection 3.1), and linear dissipative PDEs (in subsection 3.2).

3.5 Conservative PDEs

Example 3.5.1. *Sine-Gordon equation:*
Continuous:

$$\frac{\partial^2\varphi}{\partial t^2} = \frac{\partial^2\varphi}{\partial x^2} - \alpha\sin\varphi. \tag{3.25}$$

The Sine-Gordon equation is of type (3.5) with

$$\mathcal{H} = \int \left[\frac{1}{2}\pi^2 + \frac{1}{2}\left(\frac{\partial\varphi}{\partial x}\right)^2 + \alpha\left(1-\cos\varphi\right)\right] dx, \tag{3.26}$$

where $u := \begin{pmatrix} \varphi \\ \pi \end{pmatrix}$ and

$$\mathcal{S} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{3.27}$$

(Note that it follows that $\pi = \frac{\partial\varphi}{\partial t}$.)
Boundary conditions: periodic, $u(-20,t) = u(20,t)$.
Semi-discrete: finite differences[3]

$$\overline{\mathcal{H}}_{fd} = \sum_j \left[\frac{1}{2}\pi_j^2 + \frac{1}{2(\Delta x)^2}(\varphi_{j+1}-\varphi_j)^2 + \alpha\left(1-\cos\varphi_j\right)\right]. \tag{3.28}$$

$$\overline{\mathcal{S}} = \begin{pmatrix} 0 & \mathrm{id} \\ -\mathrm{id} & 0 \end{pmatrix}. \tag{3.29}$$

The resulting system of ordinary differential equations is

$$\begin{bmatrix} \dot{\boldsymbol{\varphi}} \\ \dot{\boldsymbol{\pi}} \end{bmatrix} = \overline{\mathcal{S}}\nabla\overline{\mathcal{H}}_{fd} = \begin{bmatrix} \boldsymbol{\pi} \\ \frac{1}{\Delta x^2}L\boldsymbol{\varphi} - \alpha\sin\boldsymbol{\varphi} \end{bmatrix}, \tag{3.30}$$

where L is the circulant matrix

$$L = \begin{bmatrix} -2 & 1 & & 1 \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ 1 & & 1 & -2 \end{bmatrix}.$$

[3]Summations of the form $\sum_j$ mean $\sum_{j=0}^{N-1}$ unless stated otherwise.

We have used the bold variables $\boldsymbol{\varphi}$ and $\boldsymbol{\pi}$ for the finite dimensional vectors $[\varphi_1, \varphi_2, \ldots, \varphi_N]^\top$, *et cetera*, which replace the functions π and φ in the (semi-) discrete case. Where necessary, we will write $\boldsymbol{\varphi}_n$, *et cetera* to denote the vector $\boldsymbol{\varphi}$ at time $t_0 + n\Delta t$.

The integral in the AVF method can be calculated exactly to give[4]

$$\frac{1}{\Delta t}\begin{bmatrix} \boldsymbol{\varphi}_{n+1} - \boldsymbol{\varphi}_n \\ \boldsymbol{\pi}_{n+1} - \boldsymbol{\pi}_n \end{bmatrix} = \begin{bmatrix} (\boldsymbol{\pi}_{n+1} + \boldsymbol{\pi}_n)/2 \\ L(\boldsymbol{\varphi}_{n+1} + \boldsymbol{\varphi}_n)/2 - \alpha(\cos\boldsymbol{\varphi}_{n+1} - \cos\boldsymbol{\varphi}_n)/(\boldsymbol{\varphi}_{n+1} - \boldsymbol{\varphi}_n) \end{bmatrix}. \tag{3.31}$$

Semi-discrete: spectral discretization

Instead of using finite differences for the discretization of the spatial derivative in (3.26), one may use a spectral discretization. This can be thought of as replacing φ with its Fourier series, truncated after N terms, where N is the number of spatial intervals, and differentiating the Fourier series. This can be calculated, using the discrete Fourier transform[5] (DFT), as $\mathcal{F}_N^{-1} d_N \mathcal{F}_N \boldsymbol{\varphi}$ where $\mathcal{F}_N$ is the matrix of DFT coefficients with entries given by $[\mathcal{F}_N]_{n,k} = \omega_N^{nk}$, $\omega_N = e^\theta$ and $\theta = i2\pi/l$ where $l = b - a$ is the extent of the spatial domain; that is $l/N = \Delta x$. Additionally, $[\mathcal{F}_N^{-1}]_{n,k} = \omega_N^{-nk}$ and d_N is a diagonal matrix whose (non-zero) entries are the wave-numbers $\theta_k = i2\pi k/l$, $k = 1, \ldots, N$, i.e., $[d_N]_{k,k} = \theta_k$. (For more details on properties of the DFT and its application to spectral methods see [2] and [15].)

$$\overline{\mathcal{H}}_{sp} = \sum_j \left[\frac{1}{2}\pi_j^2 + \frac{1}{2}\left[\mathcal{F}_N^{-1} d_N \mathcal{F}_N \boldsymbol{\varphi}\right]_j^2 + \alpha(1 - \cos\varphi_j) \right], \tag{3.32}$$

$$\overline{\mathcal{S}} = \begin{pmatrix} 0 & \mathrm{id} \\ -\mathrm{id} & 0 \end{pmatrix}. \tag{3.33}$$

The resulting system of ODEs is then given by

$$\begin{bmatrix} \dot{\boldsymbol{\varphi}} \\ \dot{\boldsymbol{\pi}} \end{bmatrix} = \overline{\mathcal{S}} \nabla \overline{\mathcal{H}}_{sp} = \begin{bmatrix} \boldsymbol{\pi} \\ -(\mathcal{F}_N^{-1} D_N \mathcal{F}_N)^\top (\mathcal{F}_N^{-1} d_N \mathcal{F}_N \boldsymbol{\varphi}) - \alpha \sin\boldsymbol{\varphi} \end{bmatrix}, \tag{3.34}$$

where $[D_N]_{n,k} = \theta_k$. Again, the integral in the AVF method can be calculated exactly to give

$$\frac{\boldsymbol{\varphi}_{n+1} - \boldsymbol{\varphi}_n}{\Delta t} = (\boldsymbol{\pi}_{n+1} + \boldsymbol{\pi}_n)/2, \tag{3.35}$$

$$\frac{\boldsymbol{\pi}_{n+1} - \boldsymbol{\pi}_n}{\Delta t} = -(\mathcal{F}_N^{-1} D_N \mathcal{F}_N)^\top (\mathcal{F}_N^{-1} d_N \mathcal{F}_N)(\boldsymbol{\varphi}_{n+1} + \boldsymbol{\varphi}_n)/2 - \alpha(\cos\boldsymbol{\varphi}_{n+1} - \cos\boldsymbol{\varphi}_n)/(\boldsymbol{\varphi}_{n+1} - \boldsymbol{\varphi}_n). \tag{3.36}$$

Initial conditions and numerical data for both discretizations:

Spatial domain, number N of spatial intervals, and time-step size Δt used were [6]

$$x \in [-20, 20], \qquad N = 200, \qquad \Delta t = 0.01, \qquad \text{parameter: } \alpha = 1.$$

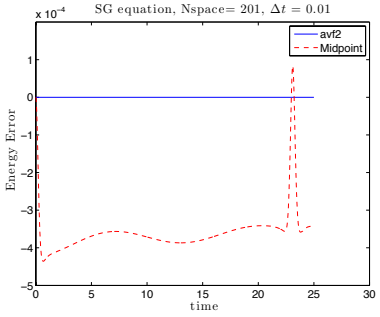
Initial conditions:

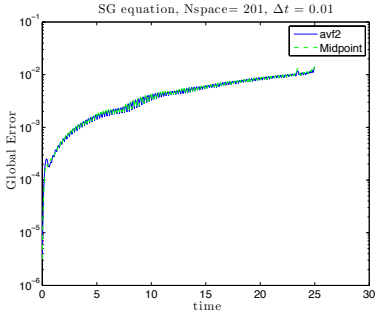
$$\left.\begin{aligned} \varphi(x,0) &= 0, \\ \pi(x,0) &= \frac{8}{\cosh(2x)}. \end{aligned}\right\} \quad \text{Right-moving kink and left-moving anti-kink solution.} \tag{3.37}$$

[4] For numerical computations, care must be taken to avoid problems when the difference $\boldsymbol{\varphi}_{n+1} - \boldsymbol{\varphi}_n$ in the denominator of (3.31) becomes small. We used the sum-to-product identity $\cos a - \cos b = -2\sin((a+b)/2)\sin((a-b)/2)$ to give a more numerically amenable expression.

[5] In practice, one uses the fast Fourier transform algorithm to calculate the DFTs in $\mathcal{O}(N \log N)$ operations.

[6] Here and below, if $x \in [a, b]$, then $\Delta x = \frac{b-a}{N}$, and $x_j = a + j\Delta x$, $j = 0, 1, \ldots, N$.





Figur 3.1: Sine-Gordon equation with finite differences semi-discretization: Energy error (left) and global error (right) vs time, for AVF and implicit midpoint integrators.

Numerical comparisons of the AVF method with the well known (symplectic) implicit midpoint integrator[7] are given in figure 3.1 for the finite differences discretization.

Example 3.5.2. *Korteweg-de Vries equation:*

Continuous:

$$\frac{\partial u}{\partial t} = -6u\frac{\partial u}{\partial x} - \frac{\partial^3 u}{\partial x^3}, \tag{3.38}$$

$$\mathcal{H} = \int \left[\frac{1}{2}(u_x)^2 - u^3\right] dx, \tag{3.39}$$

$$\mathcal{S} = \frac{\partial}{\partial x}. \tag{3.40}$$

Boundary conditions: periodic, $u(-20,t) = u(20,t)$.

Semi-discrete:

$$\overline{\mathcal{H}} = \sum_j \left[\frac{1}{2(\Delta x)^2}(u_{j+1} - u_j)^2 - u_j^3\right], \tag{3.41}$$

$$\overline{\mathcal{S}} = \frac{1}{2\Delta x}\begin{bmatrix} 0 & -1 & & & 1 \\ 1 & 0 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 0 & -1 \\ -1 & & & 1 & 0 \end{bmatrix}. \tag{3.42}$$

Initial conditions and numerical data:

$$x \in [-20, 20], \qquad N = 400, \qquad \Delta t = 0.001.$$

3.6 Dissipative PDEs

Example 3.6.1. *Heat equation:*

[7] Recall that the implicit midpoint integrator is given by $\frac{u_{n+1}-u_n}{\Delta t} = f\left(\frac{u_n+u_{n+1}}{2}\right)$.

Figur 3.2: Korteweg-de Vries equation: Energy error (left) and global error (right) vs time, for AVF and implicit midpoint integrators.

Continuous:
The heat equation

$$\frac{\partial u}{\partial t} = u_{xx}, \tag{3.43}$$

is a dissipative PDE and can be written in the form (3.5), i.e.

$$\frac{\partial u}{\partial t} = \mathcal{N}_1 \frac{\delta \mathcal{H}_1}{\delta u}, \qquad \frac{\partial u}{\partial t} = \mathcal{N}_2 \frac{\delta \mathcal{H}_2}{\delta u}, \tag{3.44}$$

with the Lyapunov functions $\mathcal{H}_1(u) = \int_0^1 \frac{1}{2} u_x^2 \, dx$ and $\mathcal{H}_2(u) = \int_0^1 \frac{1}{2} u^2 \, dx$ and the operators $\mathcal{N}_1 = -1$ and $\mathcal{N}_2 = \partial_x^2$, respectively.
Boundary conditions: $u(0,t) = u(1,t) = 0$.
Semi-discrete:

$$\overline{\mathcal{H}}_1 = \frac{1}{2(\Delta x)^2} \left[u_1^2 + \sum_{j=2}^{N-1} (u_j - u_{j-1})^2 + u_{N-1}^2 \right] \tag{3.45}$$

and

$$\overline{\mathcal{H}}_2 = \sum_{j=1}^{N-1} \frac{1}{2} (u_j)^2, \tag{3.46}$$

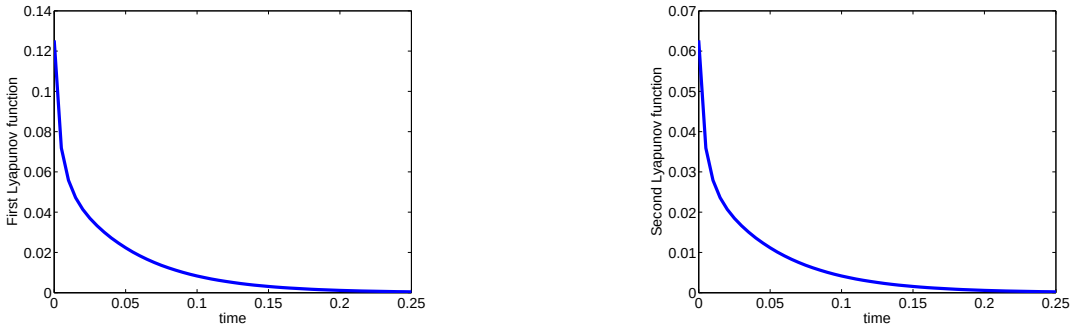
as well as

$$\overline{\mathcal{N}}_2 = \frac{1}{(\Delta x)^2} \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix} \tag{3.47}$$

and the obvious discretisation of $\mathcal{N}_1$. With these choices, both discretisations yield identical semi-discrete equations of motion and therefore $\overline{\mathcal{H}}_1$ and $\overline{\mathcal{H}}_2$ are simultaneously Lyapunov functions of the semi-discrete system and therefore, the AVF integrator preserves both Lyapunov functions.
Initial conditions and numerical data:

$$x \in [0,1], \qquad N = 50, \qquad \Delta t = 0.0025. \tag{3.48}$$

Initial condition: $u(x,0) = x(1-x)$.
This system is numerically illustrated in Figure 3.3, where the monotonic decrease of the Lyapunov functions for the heat equation in (3.45) and (3.46) is shown.



Figur 3.3: Heat equation: plots of Lyapunov functions $\overline{\mathcal{H}}_1\Delta x$ (left) and $\overline{\mathcal{H}}_2\Delta x$ (right) vs time, AVF integrator.

3.7 Multi-symplectic partial differential equations

The introduction on multi-symplectic PDEs is mainly based on [14] and the part on multi-symplectic methods is taken from [3].

3.7.1 Conservation laws for partial differential equations

Given a PDE with solution $u = u(x,t)$, $x \in \Omega \subset \mathbf{R}$, $t \in [0,T]$, belonging to suitable function-space X, we say that we have a conservation law for the PDE if there exist two functions $E : \mathbf{R} \to \mathbf{R}$ and $F : \mathbf{R} \to \mathbf{R}$ such that

$$\frac{\partial E(u(x,t))}{\partial t} + \frac{\partial F(u(x,t))}{\partial x} = 0. \tag{3.49}$$

The function E is called local density and F is called local flux of the conserved quantity. Integrating the conservation law on the space domain we obtain

$$\frac{\partial}{\partial t}\int_\Omega E(u(x,t))\,dx + F(u(x,t))|_\Gamma = 0,$$

where Γ is the boundary of Ω.

Example 3.7.1. *Consider the inviscid Burgers equation*

$$u_t + \frac{\partial}{\partial x}(\frac{1}{2}u^2) = 0, \quad x \in [0,1], \quad t \geq 0,$$

with homogeneous Dirichlet boundary conditions. By taking $E(u(x,t)) = u(x,t)$ *and* $F(u(x,t)) = \frac{1}{2}u^2(x,t)$ *we can write the equation in the form (3.49). The conserved quantity is*

$$\frac{\partial}{\partial t}\int_0^1 u(x,t)\,dx = -\frac{1}{2}u^2\Big|_0^1 = 0.$$

3.8 Multi-symplectic PDEs

These are PDEs which can be written in the form

$$Kz_t + Lz_x = \nabla S(z),$$

where $z = z(x,t)$ $x \in \Omega \subset \mathbf{R}$, $t \in [0,T]$, and K and L are $d \times d$ skew-symmetric matrices and are constant, $S : \mathbf{R}^d \to \mathbf{R}$.

These PDEs admit an energy and a momentum conservation law.

- **Energy**: take $E(z) = S(z) + \frac{1}{2} z_x^T L z$ (energy density) and $F(z) = -\frac{1}{2} z_t^T L z$ (energy flux) then (3.49) is satisfied. One can easily verify this fact by direct calculation:

$$\frac{\partial E(u(x,t))}{\partial t} + \frac{\partial F(u(x,t))}{\partial x} =$$

$$\nabla S(z)^T z_t + \frac{1}{2} z_{xt}^T L z + \frac{1}{2} z_x^T L z_t - \frac{1}{2} z_{tx}^T L z - \frac{1}{2} z_t^T L z_x =$$

  and using the multisymplectic PDE we get

$$= z_t^T K^T z_t + z_x^T L^T z_t + z_x^T L z_t = 0.$$

- **Momentum**: take $I(z) = -\frac{1}{2} z_x^T K z$ (momentum density) and $G(z) = S(z) + \frac{1}{2} z_t^T K z$ (momentum flux) then (3.49) is satisfied. The proof is similar to the previous one.

Example 3.8.1. *Consider the scalar nonlinear wave equation:*

$$u_{tt} = u_{xx} - V'(u),$$

with $V : \mathbf{R} \to \mathbf{R}$, $x \in \Omega \subset \mathbf{R}$, $t \in [0,T]$, with periodic boundary conditions. A possible choice for $V(u)$ is $V(u) = \frac{1}{4} u^4$, and $V'(u) = u^3$.

A Hamiltonian formulation of this equation can be obtained by considering the Hamiltonian function

$$H[u,v] = \int_\Omega \left(\frac{1}{2} v^2 + \frac{1}{2} u_x^2 + V(u) \right) dx,$$

computing the variational derivative one obtains

$$\begin{aligned} u_t &= \frac{\delta H}{\delta v}, \\ v_t &= -\frac{\delta H}{\delta u}, \end{aligned}$$

where $\frac{\delta H}{\delta v} = v$ and $\frac{\delta H}{\delta v} = u_{xx} - V'(u)$, i.e.

$$\begin{aligned} u_t &= v, \\ v_t &= u_{xx} - V'(u). \end{aligned}$$

For a multi-symplectic formulation consider $v := u_t$ and $w := u_x$ and $z = [u, v, w]^T$, while $S(z) = \frac{1}{2}(v^2 - w^2) + V(u)$ is just a reformulation of the integrand of $H[u,v]$ in the new dependent variables of the problem. The wave equation can then be put in multi-symplectic form by taking

$$K = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad L = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}.$$

Example 3.8.2. *The KdV equation*

$$u_t = -6uu_x - u_{xxx},$$

with periodic boundary conditions. Consider the new unknowns $u = \phi_x$, $v = u_x$, $w = \frac{1}{2}\phi_t + v_x + 3u^2$ *and* $z = [\phi, u, v, w]^T$. *The multi-symplectic formulation is achieved by taking* $S(z) = \frac{1}{2}v^2 - uw + u^3$ *and the two* 4×4 *matrices:*

$$K = \begin{bmatrix} 0 & 1/2 & 0 & 0 \\ -1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad L = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix}.$$

There is a third conservation law for multi-symplectic PDEs this is the multi-symplectic conservation law. Consider the variational equation

$$K\psi_t + L\psi_x = \nabla^2 S(z)\psi, \quad \psi = \psi(x,t),$$

which is obtained taking the first variation of the multi-symplectic PDE.

Consider also the following two skew-inner-products:

$$\omega(U,V) := U^T K^T V, \quad \gamma(U,V) := U^T L^T V.$$

For any $U = U(x,t)$ and $V = V(x,t)$ solutions of the variational equation we have the following conservation law (multi-symplectic conservation law) :

$$\frac{\partial \omega(U,V)}{\partial t} + \frac{\partial \gamma(U,V)}{\partial x} = 0,$$

this is readily seen by taking the derivatives of $\omega(U,V)$ and $\gamma(U,V)$ and using the variational equation:

$$\frac{\partial (U^T K^T V)}{\partial t} + \frac{\partial (U^T L^T V)}{\partial x} = (U_t^T K^T + U_x^T L^T)V + U^T(K^T V_t + L^T V_x) = U^T \nabla^2 S(z) V - U^T \nabla^2 S(z) V.$$

3.9 Multi-symplectic methods

Assume $z_i^j \approx z(x_i, t_j)$ where x_i $i = 1, \ldots, N$ and t_j, $j = 0, 1, \ldots$ are the nodes of a discretization grid in space-time. Assume also $\partial_t^{i,j} \approx \partial_t$ and $\partial_x^{i,j} \approx \partial_x$ are suitable finite difference approximations of the derivatives in time and space. We define the following discretization method based on these difference operators:

$$K\partial_t^{i,j} z_i^j + L\partial_x^{i,j} z_i^j = \left(\nabla_z S(z_i^j)\right)_{i,j}.$$

By using the same discretization operators we obtain the discrete variational equation

$$K\partial_t^{i,j}\psi_i^j + L\partial_x^{i,j}\psi_i^j = \left(\nabla_z^2 S(z_i^j)\right)\psi_i^j, \quad \psi_i^j \approx \psi(x_i, t_j),$$

here $\nabla_z^2 S(z_i^j)$ is the Hessian of $S(z_i^j)$ with respect to the variables z_i^j. Assume U_i^j and V_i^j are any two solutions of the discrete variational equation then we say that the method is multi-symplectic if the following discrete multi-symplectic conservation law is fulfilled

$$\partial_t^{i,j}\omega_i^j + \partial_x^{i,j}\gamma_i^j = 0, \quad \omega_i^j := {U_i^j}^T K^T V_i^j, \quad \gamma_i^j := {U_i^j}^T L^T V_i^j.$$

As an example we consider the Preismann box scheme (or concatenated midpoint rule).

The method is derived as follows. One starts by performing a semi-discretization of the multi-symplectic PDE in time using the midpoint rule:

$$\frac{d}{dx}Lz^{j+\frac{1}{2}} = \nabla_z S(z^{j+\frac{1}{2}}) - K\frac{z^{j+1}-z^j}{\Delta t},$$

the analogous semi-discrete conservation law is

$$\frac{d}{dx}\gamma(U^{j+\frac{1}{2}}, U^{j+\frac{1}{2}}) + \frac{\omega(U^{j+1}, V^{j+1}) - \omega(U^j, V^j)}{\Delta t} = 0.$$

By using the midpoint rule in the x-direction we now obtain

$$K\frac{z_{i+\frac{1}{2}}^{j+1} - z_{i+\frac{1}{2}}^{j}}{\Delta t} + L\frac{z_{i+1}^{j+\frac{1}{2}} - z_i^{j+\frac{1}{2}}}{\Delta x} = \nabla_z S\big(z_{i+\frac{1}{2}}^{j+\frac{1}{2}}\big),$$

where

$$z_{i+\frac{1}{2}}^{j} = \frac{1}{2}\big(z_i^j + z_{i+1}^j\big), \quad z_i^{j+\frac{1}{2}} = \frac{1}{2}\big(z_i^j + z_i^{j+1}\big),$$

and

$$z_{i+\frac{1}{2}}^{j+\frac{1}{2}} = \frac{1}{4}\big(z_i^j + z_{i+1}^j + z_i^{j+1} + z_{i+1}^{j+1}\big).$$

One can prove, [3] that the solutions of the corresponding discrete variational equation fulfill the following discrete multi-symplectic conservation law

$$\frac{\omega_{i+\frac{1}{2}}^{j+1} - \omega_{i+\frac{1}{2}}^{j}}{\Delta t} + \frac{\gamma_{i+1}^{j+\frac{1}{2}} - \gamma_i^{j+12}}{\Delta x} = 0, \quad \omega_i^j = {U_i^j}^T K^T V_i^j, \quad \gamma_i^j := {U_i^j}^T L^T V_i^j,$$

which is obtained discretizing in space with the midpoint rule the semi-discrete multi-symplectic conservation law.

Multi-symplectic schemes have been shown to be superior to other schemes when considering their linear stability properties (Von Neuman stability) [1]. For more information on multi-symplectic integration see also [8].

Bibliografi

[1] U. M. Ascher, R.I. McLachlan, *On symplectic and multisymplectic schemes for the KdV equation.* J. Sci. Comput. 25 (2005), no. 1-2, 83–104.

[2] W.L. Briggs and V.E. Henson, *The DFT: An Owner's Manual for the Discrete Fourier Transform*, SIAM, 1995.

[3] T. J. Bridges and S. Reich, *Multi-symplectic integrators: numerical schemes for Hamiltonian PDEs that conserve symplecticity*, Phys. Lett. A 284 (2001), no. 4-5, 184–193.

[4] E. Celledoni, R.I. McLachlan, D.I. McLaren, B. Owren, G.R.W. Quispel, and W.M. Wright, *Energy-preserving Runge-Kutta methods*, M2AN, vol 43 (4), 645-649.

[5] E. Celledoni, F. Grimm, R.I. McLachlan, D.I. McLaren, D.R.J. O'Neale, B. Owren, G.R.W. Quispel, *Preserving energy resp. dissipation in numerical PDEs, using the "average vector field"method.* NTNU Reports nr. 7/2009. Submitted.

[6] O. Gonzalez, *Time integration and discrete Hamiltonian systems*, J. Nonlinear Science, vol 6:pp 449-467, 1996.

[7] E. Hairer, C. Lubich and G. Wanner, *Geometric Numerical Integration*, Springer, II edition.

[8] B. Leimkuhler and S. Reich, *Simulating Hamiltonian dynamics.* Cambridge Monographs on Applied and Computational Mathematics, 14. Cambridge University Press.

[9] R.I. McLachlan, G.R.W. Quispel, and N. Robidoux. *Geometric integration using discrete gradients*, Phil. Trans. Roy. Soc. A, 357 (1999), pp. 1021–1045.

[10] R.I. McLachlan and N. Robidoux. *Antisymmetry, pseudospectral methods, weighted residual discretizations, and energy conserving partial differential equations*, Preprint, 2000.

[11] R.I. McLachlan and N. Robidoux. *Antisymmetry, pseudospectral methods, and conservative PDEs*, in International Conference on Differential Equations, Vol. 1, 2 (Berlin, 1999), World Sci. Publ., River Edge, NJ, 2000, pp. 994–999. .

[12] P. Olver, *Applications of Lie Groups to Differential Equations, Second Edition*, Springer-Verlag, New York, 1993.

[13] G.R.W. Quispel and D.I. McLaren. *A new class of energy-preserving numerical integration methods*, J. Phys. A: Math. Theor., 41 (2008), 045206 (7pp).

[14] B.N. Ryland, *Multisymplectic integration* PhD thesis, Massey University, Palmerston North, New Zealand, available online http://muir.massey.ac.nz/handle/10179/809.

[15] L.N. Trefethen, *Spectral Methods in MATLAB*, SIAM, 2000.