Moving forward with the help of backward error analysis: improving symplectic and other numerical methods

Farkas Miklós Seminar on Applied Analysis, BME

Donát M. Takács

Budapest University of Technology and Economics,
Faculty of Mechanical Engineering, Department of Energy Engineering

and

HUN-REN Centre for Energy Research, Fusion Plasma Physics Department

`takacs@energia.bme.hu`

2025. 11. 13.

Overview

1. An introduction to backward error analysis (BEA) for ODEs
2. BEA-based compensation I. – Modifying system parameters
3. BEA-based compensation II. – Choosing a better coordinate system

Backward error analysis for ODEs

Notation, terminology

Autonomous, d-dimensional ODE (system), initial value problem:

$$\dot{\mathbf{y}}(t) = \mathbf{f}\big(\mathbf{y}(t)\big), \qquad \mathbf{y}(0) = \mathbf{y_0}, \tag{1}$$

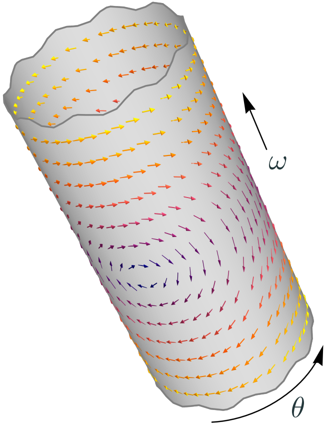
where $\mathbf{y} : \mathbb{R} \to \mathbb{R}^d$ is the (unknown) solution, $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^d$ is the (sufficiently smooth) generating vector field (the "system" being simulated), and $\mathbf{y_0} \in \mathbb{R}^d$ is the initial condition.

Explicit, one-step numerical *method* on a uniform grid with time step Δt:

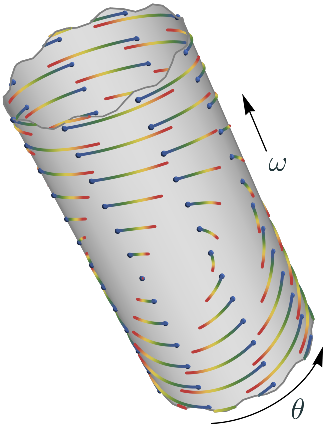
$$\mathbf{y}^{j+1} = \Phi_{\Delta t}(\mathbf{y}^j), \quad j = 0, 1, 2, \ldots \tag{2}$$

where $t^j := j\Delta t$, with $j = 0, 1, 2, \ldots$ as the index of the time step, $\mathbf{y}^j$ is the approximate solution at t^j. Designed to solve (1) approximately.

Flow of the vector field (integral curves at all points in state space, parametrised by t – a one-parameter map): $\varphi_t : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$



Vector field $\mathbf{f}$

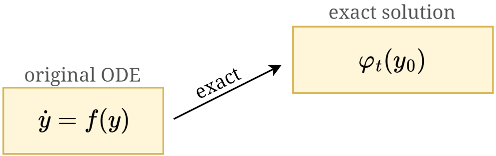


Flow φ_t

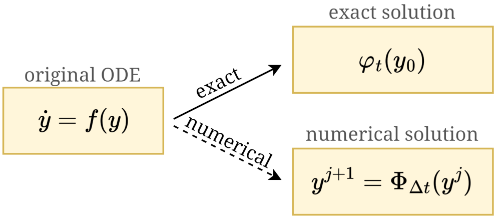
original ODE

$$\boxed{\dot{y} = f(y)}$$

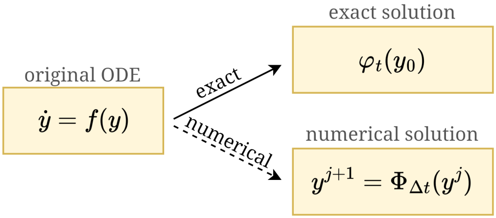
- The flow corresponds to the exact solution of the original problem



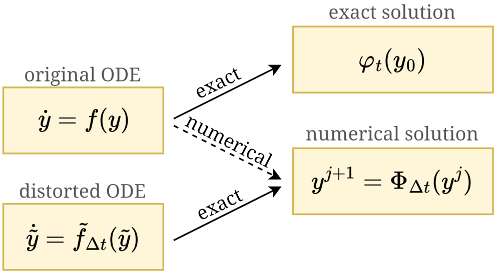
- The flow corresponds to the exact solution of the original problem
- The numerical solution is only approximate



- The flow corresponds to the exact solution of the original problem
- The numerical solution is only approximate
- *What is the numerical solution an exact solution of?*

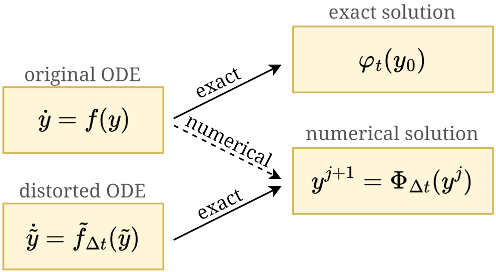


- The flow corresponds to the exact solution of the original problem
- The numerical solution is only approximate
- *What is the numerical solution an exact solution of?*
- BEA: constructing the *distorted* ODE



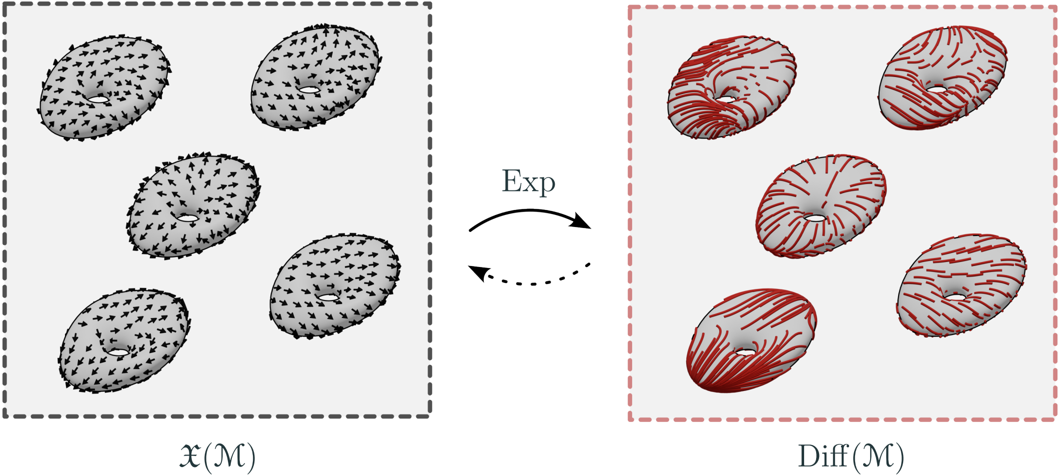
Backward error analysis for ODEs

- Backward error analysis: treating the approximate solution of $\mathbf{f}$ by $\Phi_{\Delta t}$ as an *exact solution* of a certain system nearby to the original one, i.e. $\tilde{\mathbf{y}}\left(t^j\right) = \mathbf{y}^j$.
- This other system is described by its so-called *modified equations* or *distorted equations*.
- Corresponding distorted vector field (DVF): $\tilde{\mathbf{f}}$
- $\tilde{\mathbf{f}}$ is an underlying, continuous-time counterpart of the discrete-time numerical method



Correspondence between the Lie algebra and the Lie group

Lie algebra of vector fields with Lie derivative $\overset{?}{\Leftrightarrow}$ Lie group of smooth maps?



The distorted vector field

The distorted vector field (DVF) $\tilde{\mathbf{f}}$:

- ...Does it exist?

The distorted vector field

The distorted vector field (DVF) $\tilde{\mathbf{f}}$:

- ...Does it exist?
- usually, it does [1], but...

The distorted vector field

The distorted vector field (DVF) $\tilde{\mathbf{f}}$:

- ...Does it exist?
- usually, it does [1], but...
- it is not autonomous: $\tilde{\mathbf{f}}(\mathbf{y}, t)$

The distorted vector field

The distorted vector field (DVF) $\tilde{\mathbf{f}}$:

- ...Does it exist?
- usually, it does [1], but...
- it is not autonomous: $\tilde{\mathbf{f}}(\mathbf{y}, t)$
- or, it can be approximated via an asymptotic series of an autonomous $\tilde{\mathbf{f}}(\mathbf{y})$ as:

$$\tilde{\mathbf{f}}(\tilde{\mathbf{y}}) = \mathbf{f}(\tilde{\mathbf{y}}) + \Delta t\, \mathbf{f}_1(\tilde{\mathbf{y}}) + \Delta t^2\, \mathbf{f}_2(\tilde{\mathbf{y}}) + \Delta t^3\, \mathbf{f}_3(\tilde{\mathbf{y}}) + \cdots \tag{3}$$

The distorted vector field

The distorted vector field (DVF) $\tilde{\mathbf{f}}$:

- ...Does it exist?
- usually, it does [1], but...
- it is not autonomous: $\tilde{\mathbf{f}}(\mathbf{y}, t)$
- or, it can be approximated via an asymptotic series of an autonomous $\tilde{\mathbf{f}}(\mathbf{y})$ as:

$$\tilde{\mathbf{f}}(\tilde{\mathbf{y}}) = \mathbf{f}(\tilde{\mathbf{y}}) + \Delta t\, \mathbf{f}_1(\tilde{\mathbf{y}}) + \Delta t^2\, \mathbf{f}_2(\tilde{\mathbf{y}}) + \Delta t^3\, \mathbf{f}_3(\tilde{\mathbf{y}}) + \cdots \tag{3}$$

- Truncated version is “good enough” for “long enough”

Nice properties of the DVF

$$\tilde{\mathbf{f}}(\tilde{\mathbf{y}}) = \mathbf{f}(\tilde{\mathbf{y}}) + \Delta t\, \mathbf{f}_1(\tilde{\mathbf{y}}) + \Delta t^2\, \mathbf{f}_2(\tilde{\mathbf{y}}) + \Delta t^3\, \mathbf{f}_3(\tilde{\mathbf{y}}) + \cdots \tag{3}$$

Nice properties of the DVF include:

- If $\Phi_{\Delta t}$ has order p, then $\mathbf{f}_j(\mathbf{y}) \equiv \mathbf{0}$ for $j = 1, \ldots, p-1$
- If $\Phi_{\Delta t}$ is a symmetric method, then $\mathbf{f}_j(\mathbf{y}) \equiv \mathbf{0}$ for all odd j
- Transfer of structure-preserving properties:
  - If $\Phi_{\Delta t}$ exactly conserves a first integral $I(\mathbf{y})$, then the distorted equation also has $I(\mathbf{y})$ as a first integral
  - If $\Phi_{\Delta t}$ is symplectic when applied to a Hamiltonian system of the form $\mathbf{f} = \mathcal{J}_\mathrm{c}\, \mathbf{D}H(\mathbf{y})$, then the distorted equation is also Hamiltonian
    $\rightarrow$ with a *distorted Hamiltonian*: $\tilde{\mathbf{f}} = \mathcal{J}_\mathrm{c}\, \mathbf{D}\tilde{H}(\mathbf{y})$
    Central result for symplectic methods: [2]
  - etc.

Fruitful for the analysis of structure-preserving numerical methods

Example

Example: explicit Euler (EE) method, system: $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}(t))$

Numerical method:

$$\mathbf{y}^{j+1} = \mathbf{y}^{j} + \Delta t \, \mathbf{f}\big(\mathbf{y}^{j}\big) =: \Phi_{\Delta t}\big(\mathbf{y}^{j}\big) \tag{4}$$

Looking for the DVF in the form (ansatz):

$$\tilde{\mathbf{f}}(\tilde{\mathbf{y}}) = \mathbf{f}(\tilde{\mathbf{y}}) + \Delta t \, \mathbf{f}_1(\tilde{\mathbf{y}}) + \Delta t^2 \, \mathbf{f}_2(\tilde{\mathbf{y}}) + \cdots \tag{3}$$

Taylor expansion of the distorted solution $\tilde{\mathbf{y}}$ around t:

$$\tilde{\mathbf{y}}(t + \Delta t) = \tilde{\mathbf{y}}(t) + \Delta t \frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t}(t) + \frac{\Delta t^2}{2!} \frac{\mathrm{d}^2\mathbf{y}}{\mathrm{d}t^2}(t) + \cdots \tag{5}$$

Expressing this via $\tilde{\mathbf{f}}(\tilde{\mathbf{y}})$ (chain rule):

$$\tilde{\mathbf{y}}(t + \Delta t) = \tilde{\mathbf{y}}(t) + \Delta t \, \tilde{\mathbf{f}}\big(\tilde{\mathbf{y}}(t)\big) + \frac{\Delta t^2}{2!} \Big(\mathbf{D}\tilde{\mathbf{f}}\tilde{\mathbf{f}}\Big)\big(\tilde{\mathbf{y}}(t)\big) + \cdots \tag{6}$$

Example (cont.)

$$\tilde{\mathbf{y}}(t+\Delta t) = \tilde{\mathbf{y}}(t) + \Delta t\,\tilde{\mathbf{f}}\big(\tilde{\mathbf{y}}(t)\big) + \frac{\Delta t^2}{2!}\Big(\mathbf{D}\tilde{\mathbf{f}}\tilde{\mathbf{f}}\Big)\big(\tilde{\mathbf{y}}(t)\big) + \cdots \tag{6}$$

Substituting the ansatz:

$$\tilde{\mathbf{y}}(t+\Delta t) = \tilde{\mathbf{y}}(t) + \Delta t\,\mathbf{f}\big(\tilde{\mathbf{y}}(t)\big) + \Delta t^2\left(\mathbf{f}_1 + \frac{1}{2}\mathbf{D}\mathbf{f}\mathbf{f}\right)\big(\tilde{\mathbf{y}}(t)\big) + \cdots \tag{7}$$

Set $t = t^j$, apply the defining condition $\tilde{\mathbf{y}}\big(t^j\big) = \mathbf{y}^j$, $\forall j$,

$$\mathbf{y}^{j+1} = \mathbf{y}^j + \Delta t\,\mathbf{f}\big(\mathbf{y}^j\big) + \Delta t^2\left(\mathbf{f}_1 + \frac{1}{2}\mathbf{D}\mathbf{f}\mathbf{f}\right)\big(\mathbf{y}^j\big) + \cdots \tag{8}$$

and compare with the formula for the EE method:

$$\mathbf{y}^{j+1} = \mathbf{y}^j + \Delta t\,\mathbf{f}\big(\mathbf{y}^j\big) \tag{4}$$

$\Rightarrow \mathbf{f}_1 = -\frac{1}{2}\mathbf{D}\mathbf{f}\mathbf{f}$, $\mathbf{f}_2 = \ldots$, etc.

Example (cont.)

DVF of the EE method in general (consistent, first-order):

$$\tilde{\mathbf{f}} = \mathbf{f} - \frac{\Delta t}{2}\mathbf{D}\mathbf{f}\mathbf{f} + \frac{\Delta t^2}{12}\mathbf{D}^2(\mathbf{f},\mathbf{f}) + \frac{\Delta t^2}{3}\mathbf{D}\mathbf{f}\mathbf{D}\mathbf{f}\mathbf{f} + \cdots \tag{9}$$

Mass-spring system (harmonic oscillator):

$$m\ddot{x}(t) + kx(t) = 0 \quad \Leftrightarrow \quad \underbrace{\begin{pmatrix}\dot{x}\\ \dot{v}\end{pmatrix}}_{\dot{\mathbf{y}}} = \underbrace{\begin{pmatrix}0 & 1\\ -\omega^2 & 0\end{pmatrix}}_{\mathbf{A}}\underbrace{\begin{pmatrix}x\\ v\end{pmatrix}}_{\mathbf{y}}, \quad \text{with } \omega := \sqrt{\frac{k}{m}} \tag{10}$$

DVF of this system simulated via the EE method:

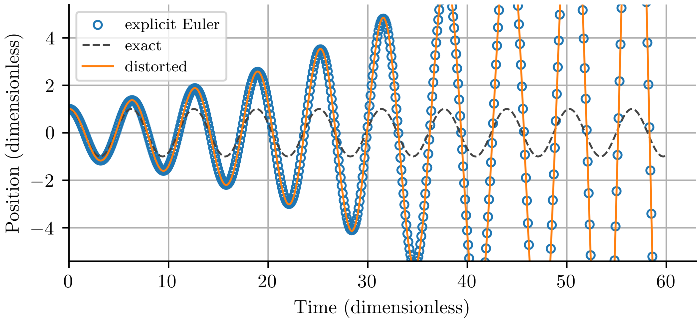
$$\dot{\tilde{x}} = \tilde{v} + \frac{\Delta t}{2}\omega^2\tilde{x} - \frac{\Delta t^2}{3}\omega^2\tilde{v} + \mathcal{O}\big(\Delta t^3\big), \tag{11}$$

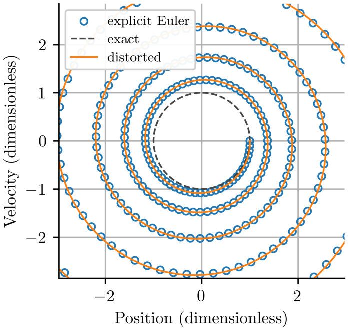
$$\dot{\tilde{v}} = -\omega^2\tilde{x} + \frac{\Delta t}{2}\omega^2\tilde{v} + \frac{\Delta t^2}{3}\omega^4\tilde{x} + \mathcal{O}\big(\Delta t^3\big). \tag{12}$$

Example (cont.)

Solution of DVF truncated up to $\mathcal{O}\left(\Delta t^2\right)$, position:

$$\tilde{x}(t) = C_1\, \mathrm{e}^{\frac{\Delta t}{2}\omega^2 t} \cos\left[\left(\omega - \frac{\Delta t^2}{3}\omega^3\right)t\right] + C_2\, \mathrm{e}^{\frac{\Delta t}{2}\omega^2 t} \sin\left[\left(\omega - \frac{\Delta t^2}{3}\omega^3\right)t\right], \qquad (13)$$





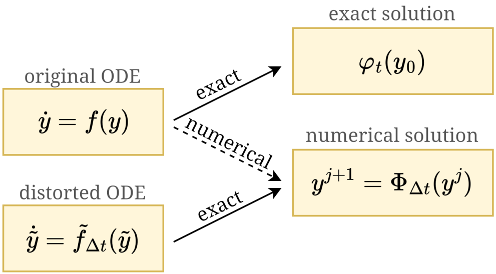
→ notice: numerical anti-dissipation, shift in frequency (dispersion error)

BEA in practice: how to do it?

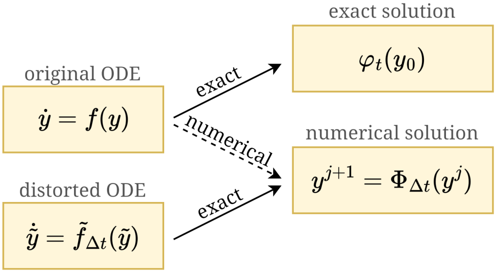
- Some more examples of this approach based on Taylor series: [3]
- Equivalent, but different approaches also exist [4, 5, 6] $\rightarrow$ sometimes better suited for proofs
- Usually, the calculation quickly becomes tedious
- Quite algorithmic: original computer algebra implementation [3], extended in [7]

BEA-based compensation I. – Modifying system parameters

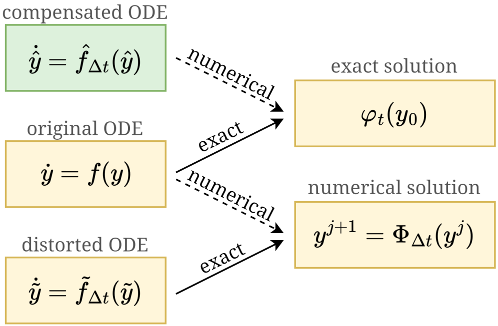
- BEA: tool for the analysis of structure-preserving numerical methods



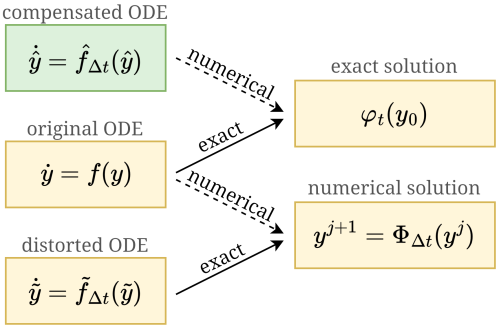
- BEA: tool for the analysis of structure-preserving numerical methods
- Can it also be applied constructively?



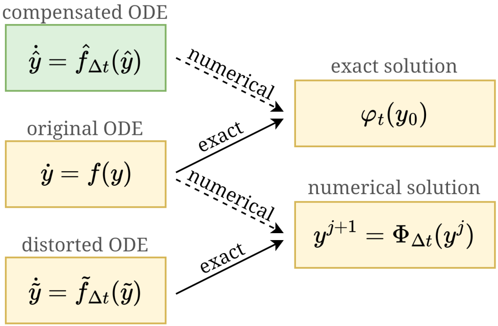
- BEA: tool for the analysis of structure-preserving numerical methods
- Can it also be applied constructively?
- *Compensating the system being simulated*



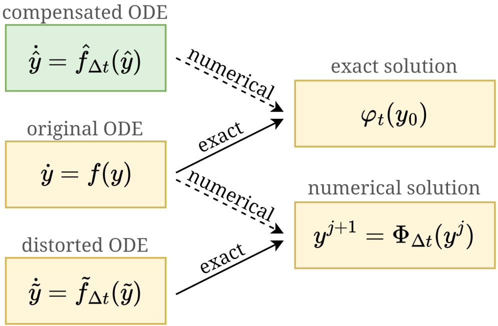
- BEA: tool for the analysis of structure-preserving numerical methods
- Can it also be applied constructively?
- *Compensating the system being simulated*
- Compensation purely through system parameters?



- BEA: tool for the analysis of structure-preserving numerical methods
- Can it also be applied constructively?
- *Compensating the system being simulated*
- Compensation purely through system parameters?
- Potential advantages:
    - Rectification of existing methods
    - No need to modify existing software implementations



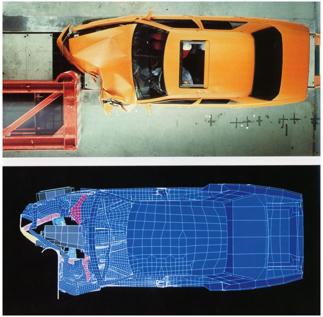
- BEA: tool for the analysis of structure-preserving numerical methods
- Can it also be applied constructively?
- *Compensating the system being simulated*
- Compensation purely through system parameters?
- Potential advantages:
    - Rectification of existing methods
    - No need to modify existing software implementations
- Existing similar approaches [8, 9, 10] modify the vector field directly, not the parameters



Analysis and improvement of the Newmark method

- Demonstration: compensating the Newmark method [7]
- Newmark method: widely used time integration method for dynamics, available in commercial engineering software (ANSYS, Abaqus)
- Time-dependent finite element method (FEM) simulations for dynamics (*direct time integration*)
  - Wave propagation, crash simulations...
- Generally, numerical dissipation is present (absent only in special cases) – can be a (dis)advantage
- BEA-based compensation: opportunity to achieve better results with existing software

→ More accurate / less computationally intensive, more reliable numerical simulations



Analysis and improvement of the Newmark method

ODE: second-order linear system, corresponding to the semi-discrete equation of motion from the FEM model, with time-dependent external excitation

$$\mathbf{M}\ddot{\mathbf{q}}(t) + \mathbf{C}\dot{\mathbf{q}}(t) + \mathbf{K}\mathbf{q}(t) = \mathbf{F}(t), \quad \mathbf{q}(0) = \mathbf{q}_0, \, \dot{\mathbf{q}}(0) = \mathbf{v}_0. \tag{14}$$

To solve this numerically, the Newmark method is applied:

$$\mathbf{M}\mathbf{a}^{j+1} + \mathbf{C}\mathbf{v}^{j+1} + \mathbf{K}\mathbf{q}^{j+1} = \mathbf{F}^{j+1}, \tag{15}$$

$$\mathbf{q}^{j+1} = \mathbf{q}^{j} + \Delta t \mathbf{v}^{j} + \frac{\Delta t^2}{2}\left[(1-2\beta)\,\mathbf{a}^{j} + 2\beta\mathbf{a}^{j+1}\right], \tag{16}$$

$$\mathbf{v}^{j+1} = \mathbf{v}^{j} + \Delta t\left[(1-\gamma)\,\mathbf{a}^{j} + \gamma\mathbf{a}^{j+1}\right], \tag{17}$$

where γ, β are parameters of the Newmark method. $\gamma = 1/2 \to$ symmetric, second-order

Analysis and improvement of the Newmark method

Rewriting (14) to an autonomous, first-order form is needed for BEA:

$$\underbrace{\begin{pmatrix} \dot{\tau} \\ \dot{\mathbf{q}} \\ \dot{\mathbf{v}} \end{pmatrix}}_{\dot{\mathbf{y}}} = \underbrace{\begin{pmatrix} 1 \\ \mathbf{v} \\ -\mathbf{M}^{-1}\left(\mathbf{C}\mathbf{v} + \mathbf{K}\mathbf{q} - \mathbf{F}(\tau)\right) \end{pmatrix}}_{\mathbf{f}(\mathbf{y})}, \tag{18}$$

$\Rightarrow$ the corresponding DVF $\tilde{\mathbf{f}}$ has been derived (margin on this slide too narrow)

$\Rightarrow$ remarkably, the resulting DVF can be re-written in a second-order form with *distorted matrices*

$$\mathbf{M}\ddot{\mathbf{q}}(t) + \tilde{\mathbf{C}}\dot{\mathbf{q}}(t) + \tilde{\mathbf{K}}\mathbf{q}(t) = \tilde{\mathbf{F}}(t), \tag{19}$$

$\Rightarrow$ this means that the Newmark method simulates a very similar system with different parameters: opportunity for compensation!

Distorted matrices

$$\tilde{\mathbf{C}} = \mathbf{C} + \mathbf{W}(\Delta t, \gamma)\left(\mathbf{M}^{-1}\mathbf{K} - \mathbf{M}^{-1}\mathbf{C}\mathbf{M}^{-1}\mathbf{C}\right) + \Delta t^2\left(\eta - \tfrac{1}{12}\right)\mathbf{K}\mathbf{M}^{-1}\mathbf{C}, \qquad (20)$$

$$\tilde{\mathbf{K}} = \mathbf{K} - \mathbf{W}(\Delta t, \gamma)\mathbf{M}^{-1}\mathbf{C}\mathbf{M}^{-1}\mathbf{K} + \Delta t^2\left(\eta - \tfrac{1}{12}\right)\mathbf{K}\mathbf{M}^{-1}\mathbf{K}, \qquad (21)$$

$$\begin{aligned}\tilde{\mathbf{F}}(t) = \mathbf{F}(t) - \mathbf{W}(\Delta t, \gamma)\left(\mathbf{M}^{-1}\mathbf{C}\mathbf{M}^{-1}\mathbf{F}(t) - \mathbf{M}^{-1}\mathbf{F}'(t)\right) + \\ + \Delta t^2\left(\eta - \tfrac{1}{12}\right)\left(\mathbf{K}\mathbf{M}^{-1}\mathbf{F}(t) - \mathbf{F}''(t)\right),\end{aligned} \qquad (22)$$

up to $\mathcal{O}\left(\Delta t^2\right)$, with

$$\eta = \tfrac{1}{2}\gamma - \beta - \tfrac{1}{12}, \quad \mathbf{W}(\Delta t, \gamma) = \Delta t\left(\gamma - \tfrac{1}{2}\right)\mathbf{M} - \Delta t^2\left(\left(\gamma - \tfrac{1}{2}\right)^2 + \tfrac{1}{12}\right)\mathbf{C}, \qquad (23)$$

and distorted initial conditions

$$\mathbf{q}(0) = \mathbf{q}_0, \qquad (24)$$

$$\begin{aligned}\dot{\mathbf{q}}(0) = \mathbf{v}_0 + \Delta t^2\eta\left(-\mathbf{M}^{-1}\mathbf{C}\mathbf{M}^{-1}\mathbf{K}\mathbf{q}_0 + \left(\mathbf{M}^{-1}\mathbf{K} - \mathbf{M}^{-1}\mathbf{C}\mathbf{M}^{-1}\mathbf{C}\right)\mathbf{v}_0 + \right. \\ \left. + \mathbf{M}^{-1}\mathbf{C}\mathbf{M}^{-1}\mathbf{F}(0) - \mathbf{M}^{-1}\mathbf{F}'(0)\right).\end{aligned} \qquad (25)$$

Compensated system:

$$\mathbf{M}\ddot{\mathbf{q}}(t) + \widehat{\mathbf{C}}\dot{\mathbf{q}}(t) + \widehat{\mathbf{K}}\mathbf{q}(t) = \widehat{\mathbf{F}}(t), \qquad \mathbf{q}(0) = \mathbf{q}_0, \quad \dot{\mathbf{q}}(0) = \mathbf{v}_0, \tag{26}$$

Two compensations introduced:

- Eliminating numerical damping (not shown here)
- **Achieving fourth-order accuracy**
  - for $\gamma = 1/2$, only the $\mathcal{O}\left(\Delta t^2\right)$ terms need to be eliminated

Fourth-order compensation of the Newmark method

Result of derivation for the fourth-order compensation:

$$\widehat{\mathbf{C}} = \mathbf{C} + \frac{1}{12}\Delta t^2 \left(\mathbf{C}\mathbf{M}^{-1}\mathbf{K} + \mathbf{K}\mathbf{M}^{-1}\mathbf{C} - \mathbf{C}\mathbf{M}^{-1}\mathbf{C}\mathbf{M}^{-1}\mathbf{C}\right), \tag{27}$$

$$\widehat{\mathbf{K}} = \mathbf{K} + \frac{1}{12}\Delta t^2 \left(\mathbf{K}\mathbf{M}^{-1}\mathbf{K} - \mathbf{C}\mathbf{M}^{-1}\mathbf{C}\mathbf{M}^{-1}\mathbf{K}\right), \tag{28}$$

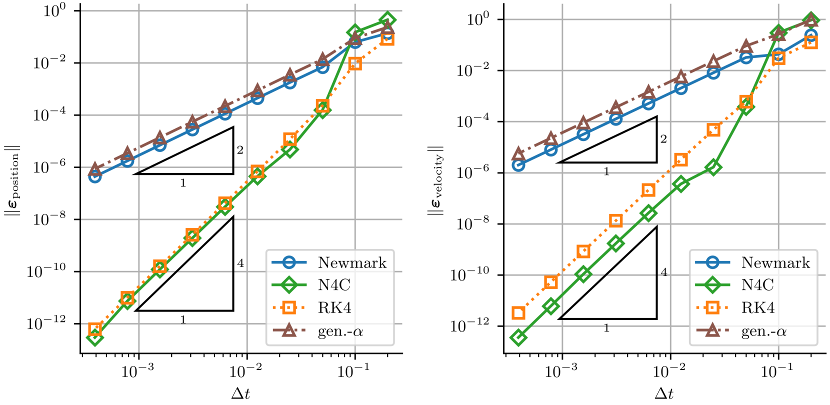
$$\widehat{\mathbf{F}}(t) = \mathbf{F}(t) + \frac{1}{12}\Delta t^2 \left(\mathbf{C}\mathbf{M}^{-1}\left(\mathbf{C}\mathbf{M}^{-1}\mathbf{F}(t) - \mathbf{F}'(t)\right) - \mathbf{K}\mathbf{M}^{-1}\mathbf{F}(t) + \mathbf{F}''(t)\right), \tag{29}$$

and $\gamma = 1/2$, $\beta = 1/6$ is required.

Once calculated, these can be used through the entire computation.

Fourth-order compensation of the Newmark method

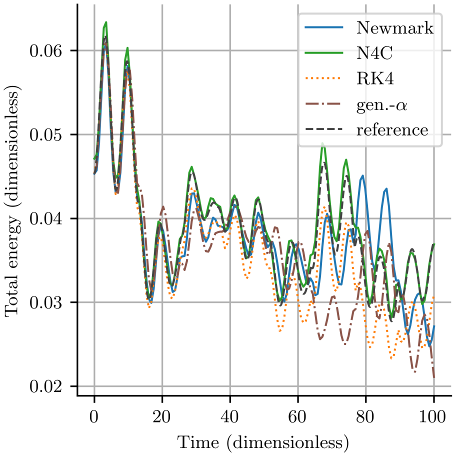
Convergence with Δt is indeed fourth-order:



Derivatives of $\mathbf{F}(t)$ can also be calculated/estimated numerically with an at least second-order accurate formula.

Fourth-order compensation of the Newmark method

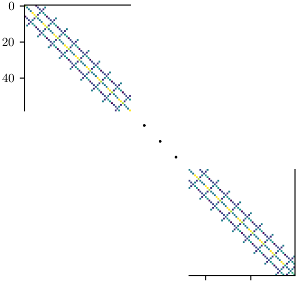
Non-continuous excitation (square wave), total energy of the system:



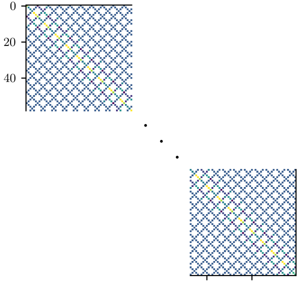
Structure of the compensated matrices

$\mathbf{M}$ does not change in the compensated system, but $\mathbf{K} \to \widehat{\mathbf{K}}$ does – how does this change the sparsity structure of the FEM matrices?

$\mathbf{K}$

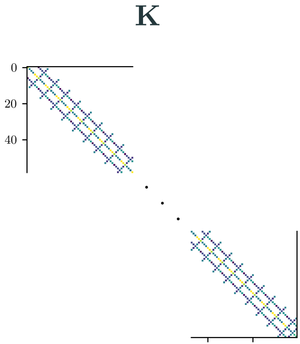


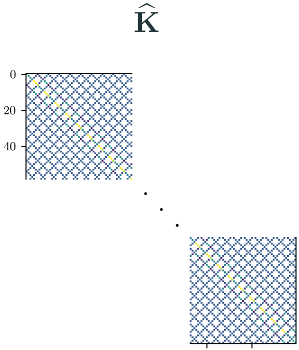
$\widehat{\mathbf{K}}$



Structure of the compensated matrices

$\mathbf{M}$ does not change in the compensated system, but $\mathbf{K} \to \widehat{\mathbf{K}}$ does – how does this change the sparsity structure of the FEM matrices?

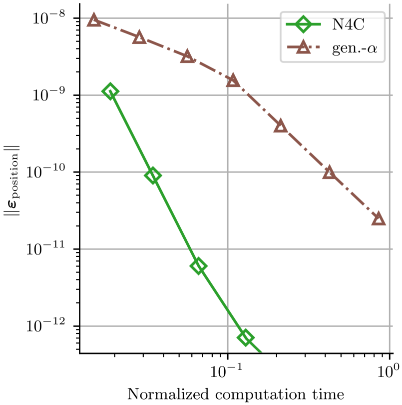


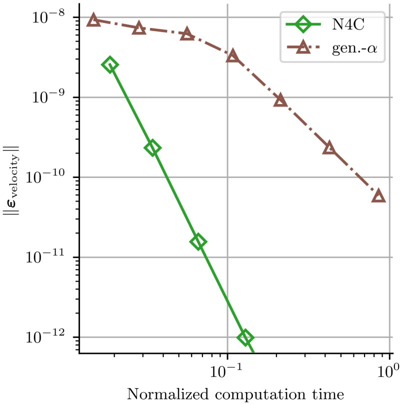


The derived DVF is underdetermined up to a matrix multiplier of $\mathbf{M}$
$\to$ opportunity for a better structure

Structure of the compensated matrices

Computation time vs. accuracy (~ 750 DoF)





Open questions, future work

- How does the computational time / accuracy tradeoff scale with the number of DoFs? (ongoing work with D. Borza)
- How do the eigenfrequencies change from $\mathbf{K}$ to $\widehat{\mathbf{K}}$?
- Can this approach be extended for more complex (nonlinear) excitation?
- Can a more favourable rewriting into second-order form be achieved with respect to the sparsity structure?
- Extension of this approach to extensions of the Newmark method: HHT-α, generalised-α

More details on this topic: [7] D. M. Takács and T. Fülöp. **"Improving the accuracy of the Newmark method through backward error analysis".** In: *Computational Mechanics* 75.5 (2025), pp. 1585–1606.

BEA-based compensation II. – Choosing a better coordinate system

Accuracy of a numerical method and the applied coordinate system

- We have seen so far that, through BEA, the DVF corresponding to a numerical method as applied to a system can be constructed
- For structure-preserving methods: the DVF is associated with a system of the same structure (detailed above)
- Specifically, for symplectic numerical methods: the DVF is Hamiltonian; there is a distorted Hamiltonian $\tilde{H}$ associated
- But: is the DVF invariant to a coordinate transformation of the original system?
  - Is the distorted Hamiltonian?
- Can this be exploited to achieve better accuracy?

Hamiltonian system (autonomous): described by the *Hamiltonian* $H(\mathbf{q}, \mathbf{p})$, $H : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

Equations of motion:

$$\underbrace{\begin{pmatrix} \dot{\mathbf{q}} \\ \dot{\mathbf{p}} \end{pmatrix}}_{\dot{\mathbf{y}}} = \underbrace{\begin{pmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{0} \end{pmatrix}}_{\mathcal{J}_{\mathrm{c}}} \underbrace{\begin{pmatrix} H_q(\mathbf{q}, \mathbf{p}) \\ H_p(\mathbf{q}, \mathbf{p}) \end{pmatrix}}_{\mathbf{D}H(\mathbf{y})}, \tag{30}$$

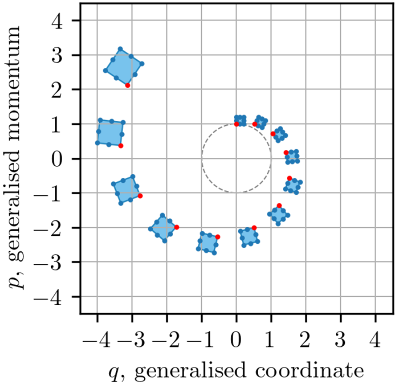
Symplectic property:

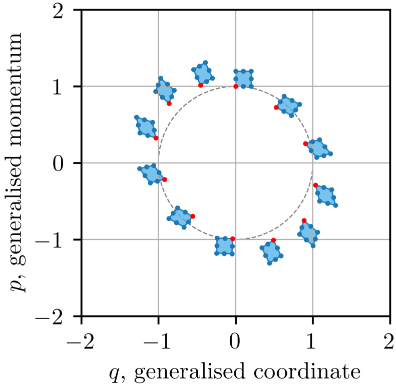
$$\left(\frac{\partial \varphi_t}{\partial \mathbf{y}}\right)^{\mathsf{T}} \mathcal{J}_{\mathrm{c}}^{-1} \left(\frac{\partial \varphi_t}{\partial \mathbf{y}}\right) = \mathcal{J}_{\mathrm{c}}^{-1} \tag{31}$$

For a numerical method:

$$\left(\frac{\partial \Phi_{\Delta t}}{\partial \mathbf{y}}\right)^{\mathsf{T}} \mathcal{J}_{\mathrm{c}}^{-1} \left(\frac{\partial \Phi_{\Delta t}}{\partial \mathbf{y}}\right) = \mathcal{J}_{\mathrm{c}}^{-1} \tag{32}$$

Explicit Euler vs. Symplectic Euler method, harmonic oscillator:





To canonical transformations

$$\begin{pmatrix}\mathbf{q}\\ \mathbf{p}\end{pmatrix} \mapsto \begin{pmatrix}\bar{\mathbf{q}}\\ \bar{\mathbf{p}}\end{pmatrix} = \begin{pmatrix}\mathbf{Q}(\mathbf{q},\mathbf{p})\\ \mathbf{P}(\mathbf{q},\mathbf{p})\end{pmatrix}, \tag{33}$$

the Hamiltonian is invariant, i.e.

$$\bar{H}\big(\bar{\mathbf{q}},\bar{\mathbf{p}}\big) := H\big(\mathbf{Q}^{-1}(\bar{\mathbf{q}},\bar{\mathbf{p}}),\mathbf{P}^{-1}(\bar{\mathbf{q}},\bar{\mathbf{p}})\big); \qquad \bar{H} = H \tag{34}$$

From here on, we restrict ourselves to the subset of canonical transformations induced by a coordinate transformation $\bar{\mathbf{q}} = \mathbf{Q}(\mathbf{q})$:

$$\bar{q}^\alpha = Q^\alpha(\mathbf{q}); \quad \bar{p}_\alpha = p_i \left(\frac{\partial \left(Q^{-1}\right)^i}{\partial \bar{q}^\alpha}\right) \circ \mathbf{Q}(\mathbf{q}) \tag{35}$$

Does this change of coordinates change the numerical results in a meaningful way?

Distorted Hamiltonian of a symplectic method

Symplectic Euler (SE) method (first-order):

$$\begin{aligned} \mathbf{q}^{j+1} &= \mathbf{q}^{j} + \Delta t H_p\left(\mathbf{q}^{j+1}, \mathbf{p}^{j}\right), \\ \mathbf{p}^{j+1} &= \mathbf{p}^{j} - \Delta t H_q\left(\mathbf{q}^{j+1}, \mathbf{p}^{j}\right). \end{aligned} \tag{36}$$

Distorted Hamiltonian from BEA:

$$\tilde{H} = H - \frac{\Delta t}{2} H_p H_q + \mathcal{O}\left(\Delta t^2\right), \tag{37}$$

Similarly, in the transformed coordinate system:

$$\tilde{\bar{H}} = \bar{H} - \frac{\Delta t}{2} \bar{H}_{\bar{p}} \bar{H}_{\bar{q}} + \mathcal{O}\left(\Delta t^2\right). \tag{38}$$

$\rightarrow$ are these actually the same, i.e., is $\tilde{H}$ invariant to coordinate transformations?

Distorted Hamiltonian of a symplectic method

$$
\begin{tikzcd}
H \arrow[r, dashed, "\text{discretisation}"] \arrow[d, shift right, "\mathbf{Q}"'] & \tilde{H} \arrow[d, dash, dashed, "?"] \\
\bar{H} \arrow[r, dashed, "\text{discretisation}"] \arrow[u, shift right, "\mathbf{Q^{-1}}"'] & \tilde{\bar{H}}
\end{tikzcd}
$$

Distorted Hamiltonian of a symplectic method

$$\tilde{\bar{H}} = \bar{H} - \frac{\Delta t}{2} \bar{H}_{\bar{p}} \bar{H}_{\bar{q}} + \mathcal{O}\left(\Delta t^2\right). \tag{38}$$

Let us consider the second term in (38), $H_p H_q$ (first elementary Hamiltonian):

$$\bar{H}_{\bar{p}} \bar{H}_{\bar{q}} \equiv \frac{\partial \bar{H}}{\partial \bar{p}_\alpha} \frac{\partial \bar{H}}{\partial \bar{q}^\alpha} = \ldots = \underbrace{\frac{\partial H}{\partial p_i} \frac{\partial H}{\partial p_k} p_l \frac{\partial \left(Q^{-1}\right)^l}{\partial \bar{q}^\beta} \frac{\partial^2 Q^\beta}{\partial q^k \, \partial q^i}}_{=: \Xi_{H_p H_q, \mathbf{Q}}} + \frac{\partial H}{\partial p_i} \frac{\partial H}{\partial q^i}. \tag{39}$$

$\rightarrow$ necessary condition for the invariance of the distorted Hamiltonian:

$$\Xi_{H_p H_q, \mathbf{Q}} = 0 \tag{40}$$

- trivially fulfilled for affine coordinate transformations, can also be fulfilled nontrivially
- non-invariant in general
- can this be exploited to eliminate the first-order term in $\tilde{H}$ to raise accuracy?

Second-order accuracy with the symplectic Euler method

Condition for second-order accuracy of SE in the transformed coordinate system:

$$\bar{H}_{\bar{p}}\bar{H}_{\bar{q}} \equiv 0 \quad \Leftrightarrow \quad \Xi_{H_p H_q,\mathbf{Q}} = -\frac{\partial H}{\partial p_i}\frac{\partial H}{\partial q^i} \tag{41}$$

- trivially fulfilled in fully cyclic (action-angle) coordinates.
  - achieving this is usually not possible
- are there any other, non-trivial $\mathbf{Q}$ coordinate transformations that achieve this?
  - not guaranteed, but sometimes possible

Demonstration: harmonic oscillator, symplectic Euler method

Harmonic oscillator:

$$H(q,p) = \frac{1}{2}p^2 + \frac{1}{2}q^2, \tag{42}$$

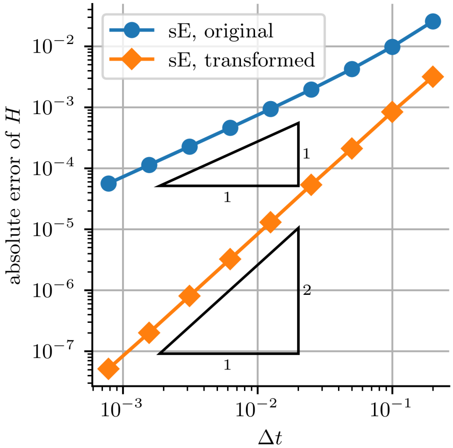
Condition (41) for second-order SE in this case, after calculations:

$$\left(1 - q^2\right)\left(\frac{\partial Q}{\partial q}\right)^{-1}\frac{\partial^2 Q}{\partial q^2} + q = 0. \tag{43}$$

Appropriate coordinate transformation fulfilling this condition:

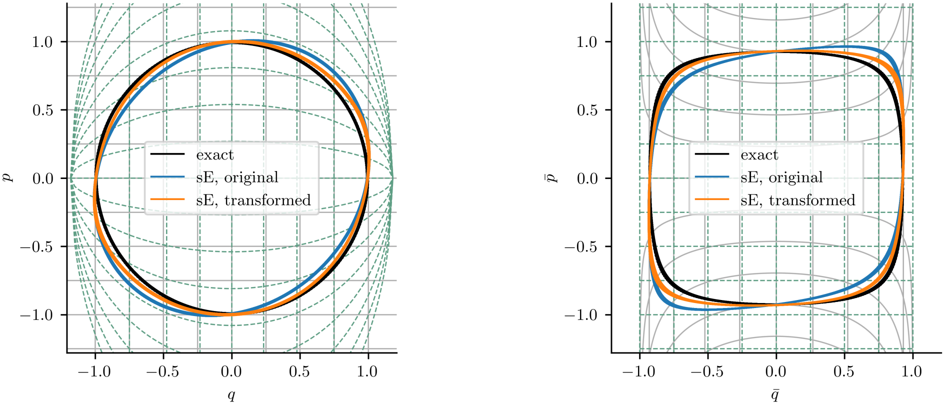
$$Q(q) = \frac{2}{\pi}\left(\hat{q}\sqrt{1-\hat{q}^2} + \arcsin(\hat{q})\right), \qquad \text{where } \hat{q} = \frac{q}{1+2\Delta t^2}, \tag{44}$$

Demonstration



Demonstration

Trajectory of the numerical results in the two coordinate systems:



Open questions, future work, other results

- When does a better coordinate system exist in general?
- Extension of this approach beyond the SE method? (Størmer–Verlet might be a good candidate)
- Extension to the entire class of canonical transformations?
- Additional result not explored here: preservation of first integrals in the SE method

More details on this topic: [11] D. M. Takács and T. Fülöp. **"On the coordinate system-dependence of the accuracy of symplectic methods".** In: *Journal of Numerical Analysis and Approximation Theory* (2025). In press.

[1] P. C. Moan. ***On rigorous modified equations for discretizations of ODEs***. Tech. rep. Technical Report 2005-3, Geometric Integration Preprint Server, 2005.

[2] G. Benettin and A. Giorgilli. **"On the Hamiltonian interpolation of near-to-the identity symplectic mappings with application to symplectic integration algorithms".** In: *Journal of Statistical Physics* 74 (1994), pp. 1117–1143.

[3] E. Hairer and C. Lubich. **"Asymptotic Expansions and Backward Analysis for Numerical Integrators".** In: *Dynamics of Algorithms*. Springer New York, 2000, pp. 91–106.

[4] S. Reich. **"Backward error analysis for numerical integrators".** In: *SIAM Journal on Numerical Analysis* 36.5 (1999), pp. 1549–1570.

[5] O. Gonzalez, D. J. Higham, and A. M. Stuart. **"Qualitative properties of modified equations".** In: *IMA Journal of Numerical Analysis* 19.2 (1999), pp. 169–190.

[6] P. C. Moan. **"On modified equations for discretizations of ODEs".** In: *Journal of Physics A: Mathematical and General* 39.19 (2006), p. 5545.

[7] D. M. Takács and T. Fülöp. **"Improving the accuracy of the Newmark method through backward error analysis".** In: *Computational Mechanics* 75.5 (2025), pp. 1585–1606.

[8] B. A. Shadwick, J. C. Bowman, and P. J. Morrison. **"Exactly conservative integrators".** In: *SIAM Journal on Applied Mathematics* 59.3 (1998), pp. 1112–1133.

[9] X. Shang and H. C. Öttinger. **"Structure-preserving integrators for dissipative systems based on reversible–irreversible splitting".** In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 476.2234 (Feb. 2020), p. 20190446.

[10] P. Chartier, E. Hairer, and G. Vilmart. **"Numerical integrators based on modified differential equations".** In: *Mathematics of computation* 76.260 (2007), pp. 1941–1953.

[11] D. M. Takács and T. Fülöp. **"On the coordinate system-dependence of the accuracy of symplectic methods".** In: *Journal of Numerical Analysis and Approximation Theory* (2025). In press.

Thank you for your kind attention!

Non-invariance of the distorted Hamiltonian