Research Article

Herbert Egger*, Oliver Habrich and Vsevolod Shashkov

On the Energy Stable Approximation of Hamiltonian and Gradient Systems

<https://doi.org/10.1515/cmam-2020-0025>

Received February 29, 2020; revised August 27, 2020; accepted November 16, 2020

Abstract: A general framework for the numerical approximation of evolution problems is presented that allows to preserve an underlying dissipative Hamiltonian or gradient structure exactly. The approach relies on rewriting the evolution problem in a particular form that complies with the underlying geometric structure. The Galerkin approximation of a corresponding variational formulation in space then automatically preserves this structure which allows to deduce important properties for appropriate discretization schemes including projection based model order reduction. We further show that the underlying structure is preserved also under time discretization by a Petrov–Galerkin approach. The presented framework is rather general and allows the numerical approximation of a wide range of applications, including nonlinear partial differential equations and port-Hamiltonian systems. Some examples will be discussed for illustration of our theoretical results, and connections to other discretization approaches will be highlighted.

Keywords: Hamiltonian Systems, Gradient Systems, Nonlinear Partial Differential Equations, Entropy Methods, Galerkin Approximation

MSC 2010: 37K05, 37L65, 47J35, 65J08

1 Introduction

The modeling of dynamical systems often leads to problems with a generalized Hamiltonian or gradient structure and various examples have been studied intensively in the literature. In this paper, we consider abstract evolution problems of the form

$$\mathcal{C}(u)\partial_t u = -\mathcal{H}'(u) + f(u), \quad (1.1)$$

which include Hamiltonian and gradient systems as special cases. Here $\mathcal{H}: \mathbb{V} \rightarrow \mathbb{R}$ represents a continuously differentiable energy or storage functional with derivative denoted by $\mathcal{H}': \mathbb{V} \rightarrow \mathbb{V}'$ and $\mathbb{V}'$. The functions $f: \mathbb{V} \rightarrow \mathbb{V}'$ and $\mathcal{C}: \mathbb{V} \rightarrow \mathcal{L}(\mathbb{V}, \mathbb{V}')$ are assumed to be sufficiently smooth, defined on some Banach space $\mathbb{V}$ with dual $\mathbb{V}'$, and $\mathcal{L}(\mathbb{V}, \mathbb{V}')$ denoting the space of linear bounded operators mapping from $\mathbb{V}$ to $\mathbb{V}'$. We further assume that $\mathcal{C}(u): \mathbb{V} \rightarrow \mathbb{V}'$ is positive semi-definite for any $u \in \mathbb{V}$. For every point in time, equation (1.1) is thus understood in the sense of linear functionals in $\mathbb{V}'$. As we will see below, the particular form (1.1), with the derivative of the Hamiltonian appearing explicitly in the equation, is canonical for the approximation by Galerkin methods. One might also replace $\mathcal{C}(u)\partial_t u$ and $f(u)$ by more general expressions $\mathcal{C}(u, \partial_t u)$ and $f(u, \partial_t u)$.

Equation (1.1) is rather general and may represent finite- or infinite-dimensional systems, including ordinary and partial differential equations and differential-algebraic problems. Questions of existence and uniqueness of solutions can hardly be addressed on this general level. However, properties of sufficiently reg-

*Corresponding author: Herbert Egger, Department of Mathematics, TU Darmstadt, Darmstadt, Germany, e-mail: egger@mathematik.tu-darmstadt.de

Oliver Habrich, Vsevolod Shashkov, Department of Mathematics, TU Darmstadt, Darmstadt, Germany, e-mail: habrich@mathematik.tu-darmstadt.de, shashkov@mathematik.tu-darmstadt.de

ular solutions can be derived, if they exist. Any classical solution $u \in C^1(0, T; \mathbb{V})$ of system (1.1), for instance, can be shown to satisfy

$$\frac{d}{dt} \mathcal{H}(u(t)) = \langle \mathcal{H}'(u(t)), \partial_t u(t) \rangle = \langle f(u(t)), \partial_t u(t) \rangle - \langle \mathcal{C}(u(t)) \partial_t u(t), \partial_t u(t) \rangle \quad (1.2)$$

for all $0 \leq t \leq T$; here $\langle \cdot, \cdot \rangle$ denotes the duality product on $\mathbb{V}' \times \mathbb{V}$. The physical meaning of this *energy-dissipation identity* is that the total energy $\mathcal{H}(u)$ of the system changes in time only due to the work done by external forces and internal dissipation. Through integration in time, one can obtain a corresponding identity in integral form

$$\mathcal{H}(u(t)) = \mathcal{H}(u(s)) + \int_s^t \langle f(u(r)), \partial_t u(r) \rangle - \langle \mathcal{C}(u(r)) \partial_t u(r), \partial_t u(r) \rangle dr, \quad (1.3)$$

which holds for all $0 \leq s \leq t \leq T$ and which remains valid for less regular solutions, e.g. $u \in W^{1,p}(0, T; \mathbb{V})$; in this case, equation (1.2) is valid only for a.e. $0 \leq t \leq T$. Let us note that, in the derivation of (1.2) and (1.3), only the variational form

$$\langle \mathcal{C}(u(t)) \partial_t u(t), v \rangle = -\langle \mathcal{H}'(u(t)), v \rangle + \langle f(u(t)), v \rangle \quad (1.4)$$

of the problem with $v = \partial_t u(t)$ was used and that the validity of (1.4) for all $v \in \mathbb{V}$ yields an equivalent variational or weak form of the system (1.1) under consideration.

We briefly mention two important particular cases of problem (1.1) that will be covered automatically by our considerations. If $\mathcal{C}(u) = -\mathcal{C}(u)^*$ is skew-adjoint and $f(u) \equiv 0$, then equation (1.1) models a Hamiltonian system and the energy $\mathcal{H}(u)$ is preserved for all time. On the other hand, if $\mathcal{C}(u) = \mathcal{C}(u)^*$ is positive definite and $f(u) \equiv 0$, then the problem describes a gradient system and energy decays until a steady state is reached.

Energy identities (1.2) and (1.3) play a fundamental role in the analysis of evolution problem (1.1), and they encode important physical principles. Therefore, much research has been devoted to the construction and analysis of numerical methods that still satisfy similar identities even after discretization. In [19], the *discrete derivative methods*, also often called *discrete gradient methods*, were introduced, which, applied to (1.1), lead to time-stepping schemes of the form

$$\bar{\mathcal{C}}(u^n, u^{n-1}) \frac{u^n - u^{n-1}}{\tau} = -\bar{\mathcal{H}}'(u^n, u^{n-1}) + \bar{f}(u^n, u^{n-1}), \quad n \geq 1.$$

Particular approximations $\bar{\mathcal{C}}(u^n, u^{n-1})$, $\bar{\mathcal{H}}'(u^n, u^{n-1})$, $\bar{f}(u^n, u^{n-1})$ result in the *discrete derivative* and the *average vector field* methods studied in [19, 26]; second order convergence in time was established for these methods. The *average vector field collocation* methods, which were proposed in [12, 20] for the numerical integration of Hamiltonian systems, yield corresponding methods of higher order. Their application to port-Hamiltonian systems was investigated in [10], and the generalization to gradient and Hamiltonian systems on Riemannian manifolds was studied in [8, 9, 21]. The general ideas underlying the discrete gradient approach can also be utilized for the space discretization of nonlinear evolution problems; we refer to the monograph [17] for details. Many further results concerning the structure-preserving discretization of Hamiltonian systems can be found in [23, 25] as well as the references given there.

In a recent paper [13], we studied the systematic approximation of dissipative nonlinear dynamical systems of another canonical form, namely

$$\mathcal{Q}(u) \partial_t u = \mathcal{A}(u),$$

by means of Galerkin approximation in space and discontinuous Galerkin discretization in time. In this system, the operator $\mathcal{Q}(u)$ is related to the derivative of the energy functional and $\mathcal{A}(u)$ encodes the dissipation terms, which essentially drive the evolution. In the current manuscript, we follow a similar route but address problems of a different form, namely (1.1), in which dissipative terms are encoded in the operator $\mathcal{C}(u)$ in front of the time derivative, while the derivative $\mathcal{H}'(u)$ of the energy functional appears as a driving force on the right-hand side.

As a first step of our analysis, we will show that the particular structure of the problem is preserved exactly by Galerkin projection of the corresponding variational formulation (1.4) in space. Such discretization strategies are frequently used for the semi-discretization of evolutionary partial differential equations, see e.g.

[16, 30], and also for model order reduction of dynamical systems [4, 7]. Let us mention [1, 5, 6, 11, 28] as particular examples addressing the structure-preserving model reduction of dissipative Hamiltonian systems. Note that the structure preservation under Galerkin projection in space does in general not hold for other, seemingly equivalent formulations of the evolution problem, like

$$\partial_t u = \mathcal{B}(u) \nabla_u \mathcal{H}(u) + g(u), \quad (1.5)$$

which are often considered in the literature; see [10, 19, 20, 26] for instance. In a finite-dimensional setting, the gradient $\nabla_u \mathcal{H}(u)$ may be identified with the derivative $\mathcal{H}'(u)$. If $\mathcal{B}(u)$ is invertible, we may transform this equation into

$$-\mathcal{B}(u)^{-1} \partial_t u = -\nabla_u \mathcal{H}(u) - \mathcal{B}(u)^{-1} g(u), \quad (1.6)$$

which, apart from the identification of $\nabla_u \mathcal{H}(u)$ with $\mathcal{H}'(u)$ is of the particular form (1.1) with $\mathcal{C}(u) = -\mathcal{B}(u)^{-1}$ and $f(u) = -\mathcal{B}(u)^{-1} g(u)$. The equivalence of (1.5) and (1.6), however, is usually lost after approximation. To see this, assume $g(u) = 0$ for simplicity, and consider the Galerkin approximation in space realized by some projection $\mathcal{P}$. Then projection of (1.5) leads to $\mathcal{P}\mathcal{B}(u)^{-1} \partial_t u = \mathcal{P}\nabla_u \mathcal{H}(u)$, while that of (1.6) yields $\mathcal{P}\partial_t u = \mathcal{P}\mathcal{B}(u) \nabla_u \mathcal{H}(u)$. Equivalence of the two formulations would require that $\mathcal{B}(u)$ commutes with $\mathcal{P}$, which is usually not the case. As we will see below, the particular form (1.1) of the problem allows for a natural approximation by Galerkin projection and therefore seems *canonical* for a structure-preserving discretization.

As a second step of our analysis, we will show that the energy identities in the integral form (1.3) remain valid after time-discretization of (1.1) or its variational form (1.4) by certain Petrov–Galerkin methods. This is again strongly linked to the particular structure of the evolution problem under investigation, which should therefore be taken into account already in the modeling stage. We will demonstrate by examples that this can be achieved without difficulties for typical problems of rather different origin. We will also discuss in detail the connection of our approach to the *discrete derivative* and the *average vector field collocation* methods mentioned earlier, which can in fact be viewed as special instances or inexact realizations of the methods proposed in this paper. Our considerations may therefore provide further insight also into the analysis and generalization of these methods and the construction of new schemes.

The remainder of the manuscript is organized as follows. In Section 2, we discuss the space discretization of problem (1.1) by Galerkin approximation of the corresponding variational form (1.4). In addition, we discuss inexact variants of the resulting methods, which may be more convenient for a practical realization. In Section 3, we then study the time discretization by a Petrov–Galerkin approach, and we highlight the connection to other methods that have been discussed in the literature. Again, some level of inexactness is allowed in the realization which may facilitate the actual numerical implementation. The wide applicability of our methods will be demonstrated in Section 4, where we discuss some typical test problems in finite and infinite dimensions. Corresponding numerical results are presented in Section 5 for illustration of our theoretical results.

2 Space Discretization

Let $\mathbb{V}_h \subset \mathbb{V}$ be a closed subspace of the Hilbert space $\mathbb{V}$. We consider the Galerkin approximation of the variational principle (1.4) in $\mathbb{V}_h$ given by

$$\langle \mathcal{C}(u_h(t)) \partial_t u_h(t), v_h \rangle = -\langle \mathcal{H}'(u_h(t)), v_h \rangle + \langle f(u_h(t)), v_h \rangle. \quad (2.1)$$

This equation is assumed to hold for all $v_h \in \mathbb{V}_h$ and for all $0 \leq t \leq T$. As a direct consequence of the particular problem structure, we obtain the following structure result.

Theorem 1. *Let $u_h \in C^1([0, T]; \mathbb{V}_h)$ satisfy (2.1) for all $0 \leq t \leq T$ and $v_h \in \mathbb{V}_h$. Then the discrete energy dissipation relation*

$$\frac{d}{dt} \mathcal{H}(u_h(t)) = \langle f(u_h(t)), \partial_t u_h(t) \rangle - \langle \mathcal{C}(u_h(t)) \partial_t u_h(t), \partial_t u_h(t) \rangle \quad (2.2)$$

holds for all $0 \leq t \leq T$, and as a direct consequence, one also has the integral identity

$$\mathcal{H}(u_h(t)) = \mathcal{H}(u_h(s)) + \int_s^t \langle f(u_h(r)), \partial_t u_h(r) \rangle - \langle \mathcal{C}(u_h(r)) \partial_t u_h(r), \partial_t u_h(r) \rangle dr$$

for all $0 \leq s \leq t \leq T$. The second identity remains valid for non-smooth solutions, e.g., if $u_h \in H^1(0, T; \mathbb{V}_h)$, while the first identity holds only for a.e. $0 \leq t \leq T$ in that case.

Proof. The first identity follows by formal differentiation of $\mathcal{H}(u(t))$ with respect to time and use of identity (2.1) with $v_h = \partial_t u_h(t)$; this is possible since $\partial_t u_h(t) \in \mathbb{V}_h$ is an admissible test function. The second identity then simply follows by integration of the first relation with respect to time. $\square$

After semi-discretization in space, equation (2.1) amounts to a finite-dimensional system of ordinary or differential-algebraic equations. Sufficient conditions on $\mathcal{C}(u)$ and $\mathcal{H}(u)$ guaranteeing the existence and uniqueness of solutions for the corresponding initial value problems could be stated, but for sake of generality, we do not address the well-posedness of the discretized problem but rather derive properties of semi-discrete solutions, assuming that they exist.

Remark 1. Let us emphasize that the structure preservation under Galerkin projection is a direct consequence of the particular *canonical* form of the evolution problem under consideration. This would not be the case if the discretization was based on a different formulation; compare with (1.5) and the corresponding remarks in the introduction.

Remark 2. A similar result can be obtained if certain approximations for the individual terms in the discrete variational problem (2.1) are used. As an example, we may consider an inexact Galerkin approximation of the form

$$\langle \mathcal{C}_h(u_h(t)) \partial_t u_h(t), v_h \rangle = -\langle \mathcal{H}'_h(u_h(t)), v_h \rangle + \langle f_h(u_h(t)), v_h \rangle.$$

Energy identities similar to (2.2) or (1.3) then hold with $\mathcal{H}(u)$, $\mathcal{C}(u)$ and $f(u)$ replaced appropriately by $\mathcal{H}_h(u_h)$, $\mathcal{C}_h(u_h)$ and $f_h(u_h)$. One could also utilize a discretization dependent duality product $\langle \cdot, \cdot \rangle_h$ on $\mathbb{V}'_h \times \mathbb{V}_h$ and consider non-conforming Galerkin approximations with spaces $\mathbb{V}_h \not\subset \mathbb{V}$; we refer to [17] for further considerations in this direction and to Sections 4 and 5 for particular examples.

3 Time Discretization

We now turn to the time discretization. Since the variational form (1.4) of the problem is inherited by Galerkin approximation in space, see the previous section, the following arguments also cover problems that have already been semi-discretized in space.

Let $T_N = \{0 = t^0 < \dots < t^N = T\}$ be a partition of the interval $[0, T]$; set $\tau^n = t^n - t^{n-1}$, and denote by $P_k(T_N; \mathbb{V})$, $k \geq 0$, the space of piecewise polynomial functions over the partition T_N with values in $\mathbb{V}$. For the time discretization of (1.4), we consider the following Petrov–Galerkin approach: find

$$u_N \in P_{k+1}(T_N; \mathbb{V}) \cap H^1(0, T; \mathbb{V})$$

satisfying

$$\int_{t^{n-1}}^{t^n} \langle \mathcal{C}(u_N(t)) \partial_t u_N(t), \tilde{v}_N(t) \rangle dt = \int_{t^{n-1}}^{t^n} \langle f(u_N(t)), \tilde{v}_N(t) \rangle - \langle \mathcal{H}'(u_N(t)), \tilde{v}_N(t) \rangle dt \quad (3.1)$$

for all $\tilde{v}_N \in P_k(T_N; \mathbb{V})$ and all time slabs $1 \leq n \leq N$. We note that the ansatz functions u_N are piecewise polynomials of degree $k+1$ and globally continuous in time, while the test functions $\tilde{v}_N$ are piecewise polynomials of degree k and may be discontinuous at time points t_n , $n = 1, \dots, N-1$; therefore, (3.1) is a Petrov–Galerkin approximation [2, 30]. Solvability of system (3.1) can again be certainly obtained under appropriate structural assumptions on the operators $\mathcal{C}(u)$ and the functional $\mathcal{H}(u)$, which however depend on the particular

examples to be studied. We again stay on the general level and do not consider issues of well-posedness, but rather prove properties of discrete solutions.

Remark 3. It suffices to consider scalar-valued test functions $\tilde{v}_N \in P_k(T_h; \mathbb{R})$ in the discrete variational problem (3.1), in which case one obtains the equivalent formulation

$$\int_{t^{n-1}}^{t^n} f(u_N(t)) \tilde{v}_N(t) - \mathcal{C}(u_N(t)) \partial_t u_N(t) \tilde{v}_N(t) dt = \int_{t^{n-1}}^{t^n} \mathcal{H}'(u_N(t)) \tilde{v}_N(t) dt, \quad (3.2)$$

which now has to be understood as an equation in $\mathbb{V}'$; compare with system (1.1) and its equivalent variational formulation (1.4) on the continuous level. We thus obtain a time discretization of (1.1) as well. This observation will be useful for our discussions later on.

With arguments similar to the previous sections, we now obtain the following result.

Theorem 2. Let $u_N \in P_{k+1}(T_N; \mathbb{V}) \cap H^1(0, T; \mathbb{V})$ solve (3.1) for all $\tilde{v}_N \in P_k(T_N; \mathbb{V})$ and all $1 \leq n \leq N$ (or equivalently (3.2) for all $\tilde{v}_N \in P_k(T_N; \mathbb{R})$). Then

$$\mathcal{H}(u_N(t^n)) = \mathcal{H}(u_N(t^m)) + \int_{t^m}^{t^n} \langle f(u_N(t)), \partial_t u_N(t) \rangle - \langle \mathcal{C}(u_N(t)) \partial_t u_N(t), \partial_t u_N(t) \rangle dt$$

for all time instances $0 \leq t_m \leq t_n \leq T$ of the respective time grid T_N .

Proof. From the fundamental theorem of calculus, we can deduce

$$\mathcal{H}(u_N(t^n)) = \mathcal{H}(u_N(t^{n-1})) + \int_{t^{n-1}}^{t^n} \langle \mathcal{H}'(u_N(t)), \partial_t u_N(t) \rangle dt = (i).$$

By choice of the trial and test spaces, the time derivative $\tilde{v}_N = \partial_t u_N$ is an admissible test function for problem (3.1), and we can replace the right-hand side (i) by

$$(i) = \int_{t^{n-1}}^{t^n} \langle f(u_N(t)), \partial_t u_N(t) \rangle - \langle \mathcal{C}(u_N(t)) \partial_t u_N(t), \partial_t u_N(t) \rangle dt.$$

This proves the result for $m = n - 1$; the general case then follows by induction over n . $\square$

Remark 4. Similar to the corresponding result for the semi-discretization in space, the proof and validity of the discrete energy identity above strongly relies on the particular structure of problem (1.1) and its variational formulation (1.4). Further, note that the energy identity here only holds at specific points in time and, in general, we do not have a pointwise energy identity like (1.2) or (2.2) for the time discrete approximation.

Let us next comment on the connection to other approximation schemes that have been proposed for the time discretization of Hamiltonian and gradient systems.

Remark 5. Consider the case $k = 0$ in the Petrov–Galerkin method (3.1). Then u_N is piecewise linear in time, and $\tilde{v}_N$ is piecewise constant. Using the abbreviation $u^n = u_N(t^n)$, scheme (3.1) can be written equivalently as

$$\bar{\mathcal{C}}(u^n, u^{n-1}) \frac{u^n - u^{n-1}}{\tau^n} = -\bar{\mathcal{H}}'(u^n, u^{n-1}) + \bar{f}(u^n, u^{n-1}),$$

with averages

$$\bar{\mathcal{C}}(u^n, u^{n-1}) = \frac{1}{\tau^n} \int_{t^{n-1}}^{t^n} \mathcal{C}(u^n(s)) ds, \quad \bar{\mathcal{H}}'(u^n, u^{n-1}) = \frac{1}{\tau^n} \int_{t^{n-1}}^{t^n} \mathcal{H}'(u^n(s)) ds, \quad \bar{f}(u^n, u^{n-1}) = \frac{1}{\tau^n} \int_{t^{n-1}}^{t^n} f(u^n(s)) ds;$$

here we used $u^n(s) = u^{n-1} + \frac{t-t^{n-1}}{\tau^n}(u^n - u^{n-1})$ for abbreviation. For $k = 0$, method (3.1) thus coincides with a discrete gradient method as outlined in the introduction; see [10, 19, 21, 26] for further details.

Remark 6. If $f(u) = 0$ and $\mathcal{C}(u)$ is anti-symmetric, then exact energy conservation is implied by Theorem 2. In this case, the proof essentially relies on the exact realization of the basic identity

$$\mathcal{H}(u_N(t^n)) - \mathcal{H}(u_N(t^{n-1})) = \int_{t^{n-1}}^{t^n} \langle \mathcal{H}'(u_N(t)), \partial_t u_N(t) \rangle dt$$

and the anti-symmetry of the operator $\mathcal{C}(u)$, which allows for some degree of inexactness in the realization. Using a quadrature rule (γ_j, ω_j) , $0 \leq j \leq J$, to approximate the integral on the left-hand side of (3.2), for instance, leads to

$$\int_{t^{n-1}}^{t^n} \mathcal{C}(u_N(t)) \partial_t u_N(t) \tilde{v}_N(t) dt \approx \tau_n \sum_{j=0}^J \mathcal{C}(u_N(t_j^n)) \partial_t u_N(t_j^n) \tilde{v}_N(t_j^n) \omega_j,$$

with intermediate time points $t_j^n = t^{n-1} + \gamma_j(t^n - t^{n-1})$. For convenience, we here employed formulation (3.2) with scalar-valued test functions $\tilde{v}_N \in P_k(T_N; \mathbb{R})$. A similar approximation can be used for the term $\int_{t^{n-1}}^{t^n} f(u_N(t)) \tilde{v}_N(t) dt$ if $f(u) \neq 0$. Note, however, that an exact integration of the nonlinear term involving $\mathcal{H}'(u_N)$ is always required to guarantee exact energy conservation. When choosing $\tilde{v}_N(t)$ as Lagrange polynomials for the quadrature points t_j^n , $j = 0, \dots, J = k$, this immediately leads to an *average vector field collocation method* discussed in [10, 20, 21]. If $\mathcal{C}(u) = \mathcal{C}$ is independent of u and t , and $f(u) = f(t)$ is a polynomial of degree $\leq k + 1$ in time, then exact integration of the left-hand side of (3.2) can be achieved by the Gauss quadrature rule. In that case, the Petrov–Galerkin approximations and the average vector field collocation methods coincide. The Petrov–Galerkin scheme of order k is actually closely related to the Lobatto-IIIa method with $s = k + 1$ stages, which can be obtained by appropriate inexact numerical integration of all terms in (3.1). The proposed Petrov–Galerkin approach can thus be used to derive well-known time-discretization methods in a systematic manner, and it provides a framework to generalize these methods to a wider class of problems.

4 Examples

We now illustrate the viability of the proposed discretization strategies in space and time by discussing their application for some typical problems we have in mind.

4.1 Magneto-Quasistatics

As a first test problem, we consider the equations of magneto-quasistatics, which arise in the eddy current approximation of Maxwell's equations at low frequency [3]. The magnetic flux density $\mathbf{b} = \text{curl } \mathbf{a}$ is then usually represented by a magnetic vector potential $\mathbf{a}$, which is assumed to satisfy

$$\sigma \partial_t \mathbf{a} + \text{curl}(\nu(\text{curl } \mathbf{a}) \text{curl } \mathbf{a}) = \mathbf{j} \quad \text{in } \Omega.$$

Here σ denotes the electric conductivity, ν is the incremental magnetic reluctivity tensor, and $\mathbf{j}$ a given source current density, and $\Omega \subset \mathbb{R}^3$ some bounded Lipschitz domain.

For ease of presentation, we further assume that σ is bounded and uniformly positive. Then $\mathbf{a}$ is determined uniquely without *gauging*, and $\mathbf{e} = -\partial_t \mathbf{a}$ amounts to the electric field. Moreover, we consider homogeneous boundary conditions $\mathbf{a} \times \mathbf{n} = 0$ on $\partial\Omega$, where $\mathbf{n}$ is the outward pointing unit normal vector on $\partial\Omega$. The variational formulation for the above problem then reads

$$\int_{\Omega} \sigma \partial_t \mathbf{a}(t) \cdot \mathbf{w} dx + \int_{\Omega} (\nu(\text{curl } \mathbf{a}(t)) \cdot \text{curl } \mathbf{a}(t)) \cdot \text{curl } \mathbf{w} dx = \int_{\Omega} \mathbf{j}(t) \cdot \mathbf{w} dx,$$

which is supposed to hold for all t of relevance and for all functions $\mathbf{w} \in H_0(\text{curl}; \Omega)$ having a square integrable weak curl and satisfying zero boundary conditions $\mathbf{w} \times \mathbf{n} = 0$ on $\partial\Omega$; see [3, 27] for details. We assume that there exists a magnetic energy density $\eta: \mathbb{R}^3 \rightarrow \mathbb{R}$ such that $\nabla_{\mathbf{b}} \eta(\mathbf{b}) = \nu(\mathbf{b}) \cdot \mathbf{b}$; in other words,

$$\eta(\mathbf{b}) = \int_0^{\mathbf{b}} (\nu(\mathbf{b}) \cdot \mathbf{b}) \cdot d\mathbf{b};$$

for $\nu(\mathbf{b}) = \nu_0 \mathcal{I}$, with scalar constant ν_0 and $\mathcal{I}$ denoting the identity matrix, we would obtain the usual term $\eta(\mathbf{b}) = \frac{\nu_0}{2} |\mathbf{b}|^2$. The expression $\mathcal{H}(\mathbf{a}) = \int_{\Omega} \eta(\text{curl } \mathbf{a}) \, dx$ now defines the magnetic energy of the system. By differentiation in time, we obtain

$$\frac{d}{dt} \mathcal{H}(\mathbf{a}(t)) = \int_{\Omega} (\nu(\text{curl } \mathbf{a}(t)) \cdot \text{curl } \mathbf{a}(t)) \cdot \text{curl } \partial_t \mathbf{a}(t) \, dx = \int_{\Omega} \mathbf{j}(t) \cdot \partial_t \mathbf{a}(t) \, dx - \int_{\Omega} \sigma \partial_t \mathbf{a}(t) \cdot \partial_t \mathbf{a}(t) \, dx,$$

which expresses the fact that the magnetic energy of the system is only altered by the power supplied through excitation currents and the dissipation caused by eddy currents.

By setting $\mathbb{V} = H_0(\text{curl}; \Omega)$ and $u = \mathbf{a}$, the magneto-quasistatic problem can be seen to fit into the abstract framework of the previous sections, with $\mathcal{H}(u) = \int_{\Omega} \eta(\mathbf{a}) \, dx$, derivative $\mathcal{H}'(u) = \text{curl}(\nu(\text{curl } \mathbf{a}) \cdot \text{curl } \mathbf{a})$, which has to be understood in a distributional sense, generalized time derivative $\mathcal{C}(u) \partial_t u = \sigma \partial_t \mathbf{a}$ and forcing term $f(u) = \mathbf{j}$.

We then consider the Galerkin approximation in space using an appropriate finite element space $\mathbb{V}_h \subset H_0(\text{curl}; \Omega)$; see [27] for details. Following our previous considerations, this leads to a semi-discretization which automatically inherits the energy identity derived above. After choice of a basis for the space $\mathbb{V}_h$, the semi-discrete problem can be cast into a system of ordinary differential equations

$$M_{\sigma} \partial_t a(t) + K_{\nu}(a(t)) a(t) = j(t),$$

and the time discretization of this system by a Petrov–Galerkin approximation, as proposed in Section 3, leads to a fully discrete scheme which satisfies a corresponding energy identity in integral form (1.3) at discrete points t^n of the time mesh. We thus obtain a discretization scheme which automatically inherits the physical principle of energy conservation.

Remark 7. Before closing this section, let us briefly comment on some obvious generalizations. Without any complications, one can consider other types of boundary conditions and a field dependent conductivity $\sigma = \sigma(\partial_t \mathbf{a})$. If σ vanishes identically on a subdomain $\Omega_{nc} \subset \Omega$, then one has to restrict $\mathbf{a}$ in Ω_{nc} by some gauging conditions [3, 27]; these can be treated, e.g., as additional constraints with arguments similar to those in Section 4.2 below.

4.2 Cahn–Hilliard Equation

A simple mathematical model for the phase separation in binary fluids is given by the Cahn–Hilliard equation; see e.g. [15]. After elimination of the chemical potential, the simplest form of the problem can be phrased as

$$\partial_t u = \Delta(-\gamma \Delta u + \psi'(u)) \quad \text{in } \Omega, \quad t > 0.$$

Here u represents the difference of the phase fractions of the two components, ψ is a double well potential with two minima, and $\gamma > 0$ is a positive constant. When complementing the system with homogeneous Neumann boundary conditions

$$0 = \partial_n u = \partial_n \Delta u \quad \text{on } \partial\Omega, \quad t > 0,$$

which implies that $\partial_n \psi'(u) = \psi''(u) \partial_n u = 0$, one can see that $\frac{d}{dt} \int_{\Omega} u(t) \, dx = \int_{\Omega} \partial_t u(t) \, dx = 0$, i.e., the total mass in the system does not change over time. For ease of notation, we will assume that $\int_{\Omega} u(t) \, dx = 0$ in the sequel. We further denote by $(-\Delta)^{-1}$ the solution operator for the Neumann problem

$$-\Delta w = f \quad \text{in } \Omega, \quad \partial_n w = 0 \quad \text{on } \partial\Omega.$$

For any sufficiently regular f with zero average, this equation has a unique weak solution $w \in H^1(\Omega)$ with $\int_{\Omega} w \, dx = 0$. By formally applying this operator to the Cahn–Hilliard equation stated above, we obtain the equivalent system

$$\begin{aligned} (-\Delta)^{-1} \partial_t u &= \gamma \Delta u - \psi'(u) && \text{in } \Omega, \\ 0 &= \partial_n u && \text{on } \partial\Omega, \end{aligned}$$

which will be the basis for further considerations. The weak form of this problem reads

$$\langle (-\Delta)^{-1} \partial_t u(t), v \rangle = -\langle \gamma \nabla u(t), \nabla v \rangle - \langle \psi'(u(t)), v \rangle,$$

which is assumed to hold for any sufficiently regular test function v and for any time t under consideration. Here $\langle a, b \rangle = \int_{\Omega} ab \, dx$ amounts to the L^2 -scalar product on Ω .

We next recall that, besides the conservation of mass, the solutions of the Cahn–Hilliard problem also are governed by decay of the free energy $\mathcal{H}(u) = \int_{\Omega} \frac{\gamma}{2} |\nabla u|^2 + \psi(u) \, dx$. By formal differentiation in time, we obtain

$$\frac{d}{dt} \mathcal{H}(u(t)) = \langle \gamma \nabla u(t), \nabla \partial_t u(t) \rangle + \langle \psi'(u(t)), \partial_t u(t) \rangle = -\langle (-\Delta)^{-1} \partial_t u(t), \partial_t u(t) \rangle.$$

In the second step, we used the weak form of the problem with test function $v = \partial_t u(t)$. Due to the Poincaré inequality, the operator $(-\Delta)^{-1}$ is positive definite on the space of functions with vanishing average, and hence the free energy $\mathcal{H}(u(t))$ is monotonically decreasing until the system reaches a steady state. The Cahn–Hilliard problem in its second equivalent form thus has exactly the canonical structure (1.1), with $\mathcal{C}(u) = (-\Delta)^{-1}$ independent of u , $f(u) \equiv 0$, and $\mathbb{V} = \{v \in H^1(\Omega) : \int_{\Omega} v \, dx = 0\}$. Discretizations which inherit the underlying energy-dissipation structure can thus be formally obtained with the arguments outlined in Sections 2 and 3.

Let us note that a standard Galerkin approximation of the weak form of the Cahn–Hilliard system leads to an approximation u_h with values in $\mathbb{V}_h \subset \mathbb{V}$ characterized by a discrete variational principle of the form

$$\langle (-\Delta)^{-1} \partial_t u_h(t), v_h \rangle = -\langle \gamma \nabla u_h(t), \nabla v_h \rangle - \langle \psi'(u_h(t)), v_h \rangle.$$

While theoretically sound, such a method is not very useful in practice since the application of the inverse Laplacian $(-\Delta)^{-1}$ can in general not be realized exactly. To overcome this problem, one may utilize a discrete approximation $(-\Delta_h)^{-1}$, which is defined via the solution $w_h = (-\Delta_h)^{-1} f \in \mathbb{V}_h$ of the discretized problem

$$\langle \nabla w_h, \nabla \eta_h \rangle = \langle f, \eta_h \rangle \quad \text{for all } \eta_h \in \mathbb{V}_h.$$

Again, a zero average condition for the right-hand side f and the solution w_h allow to guarantee existence of a unique solution. The approximation u_h for the Cahn–Hilliard system is then characterized by the discrete variational principle

$$\langle (-\Delta_h)^{-1} \partial_t u_h(t), v_h \rangle = -\langle \gamma \nabla u_h(t), \nabla v_h \rangle - \langle \psi'(u_h(t)), v_h \rangle,$$

which is again required to hold for all $v_h \in \mathbb{V}_h$ and all relevant t . With similar reasoning as on the continuous level, one can verify the energy identity

$$\frac{d}{dt} \mathcal{H}(u_h(t)) = -\langle (-\Delta_h)^{-1} \partial_t u_h(t), \partial_t u_h(t) \rangle,$$

which shows that the energy of semi-discrete solutions again decreases until a steady state is reached. This approach corresponds to an inexact Galerkin approximation in space, as discussed in Remark 2. Further approximations, e.g., numerical quadrature for the nonlinear terms, may be used to facilitate the numerical implementation; see Section 5.

For the subsequent time discretization, we can employ a Petrov–Galerkin approximation as outlined in Section 3. Since the dissipation term here is linear and the force term vanishes, this coincides with a particular average vector field collocation method based on Gauss quadrature; see [21], and refer to Remark 6 for discussion in this direction.

4.3 Constrained Hamiltonian Systems

As another relevant class of applications, we consider finite-dimensional Hamiltonian systems with holonomic constraints given by

$$\partial_t q = H_p(p, q), \quad (4.1)$$

$$\partial_t p = -H_q(p, q) + f(p, q) + g_q(p, q)^\top \lambda, \quad (4.2)$$

$$0 = g(q). \quad (4.3)$$

Such systems arise, for instance, in the modeling of multibody dynamics: see, e.g., [14, Chapter 1], [25, Chapter 7] or [29, Chapter 13]. Here p and q are, respectively, the vectors of generalized coordinates and momenta, $H(p, q)$ is the Hamiltonian, i.e., the energy or storage functional, $f(p, q)$ denotes external forces, λ is the vector of Lagrange multipliers for the constraints, and $g_q(p, q)^\top \lambda$ are the corresponding constraint forces. We assume in the following that p, q, λ are real-valued vectors and use subscripts to denote partial derivatives with respect to these variables. The symbol $\langle \cdot, \cdot \rangle$ then stands for the Euclidean scalar product on $\mathbb{R}^n$.

By differentiating (4.3) with respect to time, we obtain the linearized constraint

$$0 = g_q(q) \partial_t q, \quad (4.4)$$

which is equivalent to the nonlinear constraint (4.3) up to a constant factor that can be fixed by an appropriate initial condition. We further introduce the principal function of the Lagrange multiplier $\Lambda(t) = \int_0^t \lambda(s) ds$. The variational form of system (4.1), (4.2) complemented by the linearized constraint (4.4) can then be written as

$$\langle \partial_t q(t), v \rangle = \langle H_p(p(t), q(t)), v \rangle, \quad (4.5)$$

$$\langle \partial_t p(t), w \rangle = -\langle H_q(p(t), q(t)), w \rangle + \langle f(p(t), q(t)), w \rangle + \langle g_q(q(t))^\top \partial_t \Lambda(t), w \rangle, \quad (4.6)$$

$$0 = \langle g_q(q(t)) \partial_t q(t), \eta \rangle. \quad (4.7)$$

These identities are assumed to hold for all test vectors v, w, η of appropriate size, and any time $t > 0$. For every sufficiently smooth solution (p, q, Λ) of (4.5)–(4.7), we then get

$$\begin{aligned} \frac{d}{dt} H(p(t), q(t)) &= \langle H_p(p(t), q(t)), \partial_t p(t) \rangle + \langle H_q(p(t), q(t)), \partial_t q(t) \rangle \\ &= \langle \partial_t q(t), \partial_t p(t) \rangle - \langle \partial_t p(t), \partial_t q(t) \rangle + \langle f(p(t), q(t)), \partial_t q(t) \rangle + \langle g_q(q(t))^\top \partial_t \Lambda(t), \partial_t q(t) \rangle \\ &= \langle f(p(t), q(t)), \partial_t q(t) \rangle. \end{aligned}$$

In the last step, we used that $\langle g_q(q)^\top \partial_t \Lambda, \partial_t q \rangle = \langle g_q(q) \partial_t q, \partial_t \Lambda \rangle = 0$, which follows from testing equation (4.7) with $\eta = \partial_t \Lambda$. This identity describes that the total energy of the system can only change due to the work done by the external forces.

After rearrangement of terms, one can see that system (4.1), (4.2) with linearized constraint (4.4) is again of the abstract form $\mathcal{C}(u) \partial_t u = -\mathcal{H}'(u) + f(u)$ with state vector $u = (p, q, \Lambda)$, energy functional $\mathcal{H}(u) = H(p, q)$ and operators

$$\mathcal{C}(u) = \begin{pmatrix} 0 & -I & 0 \\ I & 0 & -g_q(q)^\top \\ 0 & g_q(q) & 0 \end{pmatrix} \quad \text{and} \quad f(u) = \begin{pmatrix} 0 \\ f(p, q) \\ 0 \end{pmatrix}.$$

Hence all results about the approximation by Galerkin methods in space and the Petrov–Galerkin method in time obtained in Sections 2 and 3 apply immediately. In particular, the Galerkin projection of (4.5)–(4.7) into a conforming subspace $\mathbb{V}_h \subset \mathbb{V}$ automatically inherits the energy identity stated above. Our results in particular also cover model order reduction approaches based on Galerkin projection [7]. Furthermore, the energy identity in its integral form also remains valid after time-discretization if the appropriate Petrov–Galerkin approximation is used; see Theorem 2 and Remark 3.

Remark 8. In the Petrov–Galerkin scheme (3.1), we may use test functions of the form $\bar{v}(t) = 1|_{[t^{n-1}, t^n]} \bar{v}$ with $\bar{v} \in \mathbb{V}$ arbitrary. For the system under investigation, this yields

$$0 = \int_{t^{n-1}}^{t^n} g_q(q(t)) \partial_t q(t) dt = \int_{t^{n-1}}^{t^n} \frac{d}{dt} g(q(t)) dt = g(q(t^n)) - g(q(t^{n-1})).$$

The original constraint (4.3) thus remains valid for all time points t^n of the time grid T_N if it was valid at initial time t^0 ; we only used that the piecewise constant functions in time are elements of the test space $\bar{V}_N$ in (3.1). Hence, in contrast to other time discretization schemes, the proposed Petrov–Galerkin approximation of (4.1), (4.2) and (4.4) does not suffer from the well-known *drift-off phenomenon*; see, e.g., [14, Chapter 5] or [29, Chapter 14] for details. Let us note that some artificial drift-off may be observed in practice due to inexact integration of (3.1), which can be removed by standard stabilization or projection steps; see [14, Chapter 5] and [29, Chapter 14]. An alternative approach to avoid drift-off in index-2 formulations of multibody dynamics is the GGL approach [18], also see [29, Chapter 14], which however requires to extend the system by additional unknowns. Another approach by “minimal extension” is discussed in [24, Section 6.4]. We are not aware, however, of a provably energy conserving time discretization strategy for these approaches.

5 Numerical Tests

For illustration of our theoretical results and in order to demonstrate the practicability of the proposed methods, we now briefly discuss their application and numerical realization for the test problems mentioned in the previous section.

5.1 Magneto-Quasistatics

We consider a geometry $\Omega = \Omega_{xy} \times \mathbb{R}$ with infinite extent in the z -direction and assume that $\mathbf{j} = (0, 0, j_z)$ and that j_z and σ are independent of z . Then solutions of the eddy current problem can be found in the form $\mathbf{a} = (0, 0, a_z)$ with a_z independent of z , and the equations of magneto-quasistatics reduce to

$$\sigma \partial_t a_z = \operatorname{div}(\nu(\nabla a_z) \nabla a_z) - j_z \quad \text{in } \Omega_{xy}.$$

We assume that $\Omega_{xy} \subset \mathbb{R}^2$ is bounded and complement the system by homogeneous boundary conditions; see Figure 1 for an illustration of the geometric arrangement.

As magnetic reluctivity function for our tests, we take $\nu(\mathbf{b}) = \nu_0(1 - \frac{\alpha}{\beta + |\mathbf{b}|^2})$ with constants $0 < \alpha < \beta$ leading to the energy density $u(\mathbf{b}) = \frac{\nu_0}{2}(|\mathbf{b}|^2 - \alpha \log(\beta + |\mathbf{b}|^2))$. The excitation currents are defined as

$$j_z(x, y, t) = I(\chi_{\Omega_+} - \chi_{\Omega_-}) \sin(\omega t)$$

with I, ω prescribed, and $\Omega_{\pm}$ denoting the support of positive and negative part of the impressed current; see Figure 1. In our simulations, we use the parameters $\alpha = 0.9, \beta = 1, \sigma = 1, \omega = 2\pi$ and $I = 1$. For discretization of the problem, we chose a Galerkin approximation in space with piecewise linear finite elements and a Petrov–Galerkin approximation of orders $k = 1, 2$ in time with constant time step size $\tau^n = \tau > 0$.

In the middle of Figure 1, we show the potential a_z and the magnetic field vectors $\mathbf{b}$ at time $t = \frac{1}{2}$. In the right plot, we display the evolution of the total energy $H^n = \mathcal{H}(a_h(t^n))$ as well as the time accumulated supplied and dissipated energy of the system. Note that according to the discrete energy balance

$$H^n = H^0 + H_{\text{supp}}^n - H_{\text{loss}}^n,$$

with symbols $H_{\text{supp}}^n = \tau \sum_{k=1}^n P_{\text{supp}}^k$ and $H_{\text{loss}}^n = \tau \sum_{k=1}^n P_{\text{loss}}^k$ denoting the total accumulated supply and loss energy, respectively, where

$$P_{\text{supp}}^k = \frac{1}{\tau} \int_{t^{k-1}}^{t^k} \int_{\Omega_{xy}} j_z(t) \partial_t a_h(t) dx dt \quad \text{and} \quad P_{\text{loss}}^k = \frac{1}{\tau} \int_{t^{k-1}}^{t^k} \int_{\Omega_{xy}} \sigma |\partial_t a_h(t)|^2 dx dt$$

Figure 1: (a) Geometric setup of the model problem. (b) Potential a_z and magnetic flux $\mathbf{b}$ (arrows) for $t = 1/2$. (c) Energy $H(t)$ (red, dashed) at time t and total supply $H_{\text{supply}}(t)$ (black, dotted) as well as total loss $H_{\text{loss}}(t)$ (blue, solid) up to time t .

τ	Lobatto ($s = 2$)	PG ($k = 1$)	Lobatto ($s = 3$)	PG ($k = 2$)
0.1	$1.1000 \cdot 10^{-3}$	$8.7137 \cdot 10^{-14}$	$9.2094 \cdot 10^{-6}$	$1.7073 \cdot 10^{-13}$
0.05	$2.7303 \cdot 10^{-4}$	$5.9014 \cdot 10^{-13}$	$1.6664 \cdot 10^{-6}$	$1.4951 \cdot 10^{-12}$
0.01	$1.0740 \cdot 10^{-5}$	$1.4769 \cdot 10^{-12}$	$2.0862 \cdot 10^{-9}$	$2.2324 \cdot 10^{-12}$

Table 1: Energy error $\max_n |H^n - H_{\text{supply}}^n + H_{\text{loss}}^n|$ obtained for the Lobatto-IIIA method with $s = 2$ and $s = 3$ stages and the corresponding Petrov–Galerkin methods of order $k = 1$ and $k = 2$.

are the average supply and loss power at time step k . For a better evaluation, we list in Table 1 the maximal discrepancy in the above energy identity obtained with the proposed Petrov–Galerkin time discretization and the Crank–Nicolson scheme.

As predicted, the Petrov–Galerkin method preserves the energy identity on the discrete level up to round-off errors which may accumulate with the number of time steps. In contrast to that, the Crank–Nicolson scheme is not able to exactly reproduce the underlying energy-dissipation structure on the discrete level.

5.2 Cahn–Hilliard Equation

As a second test problem, we consider the Cahn–Hilliard equation with a logarithmic potential

$$\psi(x) = \frac{\theta}{2} \left[(1+x) \ln \left(\frac{1+x}{2} \right) + (1-x) \ln \left(\frac{1-x}{2} \right) \right] + \frac{\theta_c}{2} (1-x^2)$$

and Neumann boundary conditions; see Section 4.2. For our computations, we use $\Omega = (0, 1)^2$ as a computational domain and choose $\gamma = 0.002$, $\theta = 0.3$, $\theta_c = 0.8$. For discretization, we use the proposed Galerkin Petrov–Galerkin approximation with piecewise linear finite elements in space and piecewise linear approximation in time. The fully discrete solution is characterized by

$$\left\langle (-\Delta_h)^{-1} \frac{u_h^n - u_h^{n-1}}{\tau^n}, v_h \right\rangle_h = - \left\langle \gamma \nabla \frac{u_h^n + u_h^{n-1}}{2}, \nabla v_h \right\rangle - \langle \psi'(u_h^n, u_h^{n-1}), v_h \rangle_h.$$

Here we used

$$u_h^n = u_{h,N}(t^n) \quad \text{and} \quad \bar{\psi}'(u_h^n, u_h^{n-1}) = \frac{1}{\tau} \int_{t^{n-1}}^{t^n} \psi'(u_h(t^{n-1} + s\tau)) ds$$

for abbreviation. Note that this average vector field approximation automatically results from the Petrov–Galerkin time integration of lowest order. In the terms characterized by $\langle \cdot, \cdot \rangle_h$, we use numerical integration by the vertex rule; see Remark 2. The inverse of the discrete Laplacian is defined by $w_h = (-\Delta_h)^{-1}f$ with $\langle \nabla w_h, \nabla \eta_h \rangle = \langle f, \eta_h \rangle_h$ for all η_h and zero average conditions. The assembly of the inverse Laplacian is circumvented by introducing a discrete potential μ_h in the numerical realization; cf. Section 4.2.

Figure 2: (a)–(c) Snapshot of the phase fraction u at time $t = 0, 0.3, 2$. (d) Evolution of the free energy (blue, solid), dissipated energy (black, dotted), total energy (red, dashed).

For our computations, we utilize a uniform triangulation with mesh size $h = \frac{1}{64}$ and a time step $\tau = 4 \cdot 10^{-4}$. The nonlinear system in every time step is solved by a simplified Newton iteration with tolerance 10^{-12} . The nodal values of the initial datum $u_h(0)$ are chosen randomly around 0 with standard deviation of 10^{-3} .

In Figure 2, we display some snapshots of the numerical solution, and we depict the evolution of the discrete free energy, which is defined by

$$H_h^n = \int_{\Omega} \frac{\gamma}{2} |\nabla u_h(t^n)|^2 + I_h \psi(u_h(t^n)) \, dx.$$

Here $I_h: H^1(\Omega) \rightarrow V_h$ denotes the piecewise linear interpolation for which we observe that $\int_{\Omega} I_h u \, dx = \langle u, 1 \rangle_h$ exactly amounts to numerical integration by the vertex rule. As a consequence, the discrete energy identities derived in Sections 2 and 3 carry over almost verbatim, which yields to the conclusion that the decay of the discrete free energy should be compensated exactly by the total accumulated dissipation. In our computational tests, this can be observed up to machine accuracy.

5.3 Constrained Hamiltonian Dynamics

As a simple example for a constrained dynamical system, we consider a pendulum with a mass $m = 1$ at position (x, y) swinging at a fixed distance ℓ around the origin. The equations of motion can then be written as

$$\dot{x} = v, \quad \dot{v} = -2x\lambda, \quad \dot{y} = w, \quad \dot{w} = -g - 2y\lambda, \quad 0 = x^2 + y^2 - \ell^2,$$

which comprises a differential-algebraic system of index 3; see [22, 24] for details. This example fits into the class of problems considered in Section 4.3 with $q = (x, y)$, $p = (v, w)$, constraint $g(q) = x^2 + y^2 - \ell^2$, and energy $H(x, y, v, w) = \frac{1}{2}(v^2 + w^2) + gy$, consisting of a kinetic and a potential contribution. Following the considerations presented in Section 4.3, we obtain the canonical form of the system, which here has the

form

$$\begin{pmatrix} 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & -2x \\ 0 & 1 & 0 & 0 & -2y \\ 0 & 0 & 2x & 2y & 0 \end{pmatrix} \begin{pmatrix} \dot{v} \\ \dot{w} \\ \dot{x} \\ \dot{y} \\ \dot{\Lambda} \end{pmatrix} = - \begin{pmatrix} v \\ w \\ 0 \\ g \\ 0 \end{pmatrix}.$$

In the derivation, the algebraic constraint was differentiated once, and the Lagrange multiplier was integrated. The resulting system turns out to be an equivalent index-1 reformulation of the original problem. In compact notation, the above system can be written as $\mathcal{C}(u)\partial_t u = -\mathcal{H}'(u)$ with anti-symmetric matrix $\mathcal{C}(u)$ and energy derivative vector $\mathcal{H}'(u)$. The above index-1 formulation therefore has the canonical form (1.1) with force vector $f(u) \equiv 0$, and we obtain exact conservation of the energy for all time.

For the time-discretization, we use a Petrov–Galerkin method as outlined in Section 4.3. Since $\mathcal{C}(u)$ and $\mathcal{H}'(u)$ both depend affine linearly on u , all time integrals can be evaluated exactly. The approximation of lowest order with piecewise linear functions in time and constant time step τ then leads to a recursion of the form

$$\bar{\mathcal{C}}(u^n, u^{n+1}) \frac{u^{n+1} - u^n}{\tau} = -\bar{\mathcal{H}}'(u^n, u^{n+1}), \quad n \geq 0, \quad (5.1)$$

with $\bar{\mathcal{C}}(u^n, u^{n+1}) = \frac{1}{2}(\mathcal{C}(u^n) + \mathcal{C}(u^{n+1}))$ and $\bar{\mathcal{H}}'(u^n, u^{n+1}) = \frac{1}{2}(\mathcal{H}'(u^n) + \mathcal{H}'(u^{n+1}))$. As before, $u^n = u(t^n)$ denote the values of the discrete solution at the time points t^n .

For our numerical tests, we chose the initial conditions as $x_0 = -1$, $y_0 = 1$, $v_0 = 0$, $w_0 = 0$ and $\Lambda_0 = 0$. We further set $g = 1$ and choose a time step size of $\tau = 0.2$. The nonlinear systems at every time step are solved with the Newton iteration. Three to four iterations were usually required to achieve the prescribed tolerance of 10^{-12} .

Figure 3: (a) Snapshot of the position (x, y) at time $t = 0, 0.2, \dots, 3$. (b) Evolution of kinetic (blue, dashed), potential (red, dotted) and total energy (black, solid) of the system. (c) Evolution of the position x (blue, dashed), y (red, dotted) and length (black, solid) of the pendulum.

In Figure 3, we display some snapshots of the numerical solution and the evolution of the energy and the position as well as length over time. As predicted by our theoretical results, scheme (5.1) is exactly energy preserving and, despite the fact that the constraint was differentiated, does not suffer from the *drift-off phenomenon*; see Remark 8. In our computations, the length of the pendulum as well as the total energy of the system are conserved up to machine precision even over hundreds of periods.

6 Discussion

We proposed an abstract framework for the numerical approximation of evolution problems with an underlying Hamiltonian or gradient structure, and we showed that this underlying geometric structure can be preserved exactly under discretization with Galerkin methods in space and Petrov–Galerkin approximation in time if the evolution problem was written in a *canonical form* already on the continuous level. We also illustrated that some inexactness in the numerical realization of the Galerkin approximations is possible which may facilitate the numerical realization. The discrete derivative and average vector field collocation methods could be interpreted as such inexact realizations of the Petrov–Galerkin time-discretization. The viability of the proposed approach and the practicability of the resulting discretization schemes was demonstrated for some typical test problems. We note that, in case of complex non-quadratic Hamiltonians, the efficient numerical handling of the term $\langle \mathcal{H}'(u), v \rangle$ arising in (1.4) and the corresponding discrete equations is somewhat subtle and may deserve further considerations.

Funding: The work of the author was supported by the “Excellence Initiative” of the German Federal and State Governments via the Center of Computational Engineering GSC 233 at TU Darmstadt and by the German Research Foundation (DFG) via grants TRR 146 project C3, TRR 154 project C04, and Eg-331/2-1 within SPP 2256.

References

- [1] B. M. Afkham and J. S. Hesthaven, Structure-preserving model-reduction of dissipative Hamiltonian systems, *J. Sci. Comput.* **81** (2019), no. 1, 3–21.
- [2] G. Akrivis, C. Makridakis and R. H. Nochetto, Galerkin and Runge–Kutta methods: Unified formulation, a posteriori error estimates and nodal superconvergence, *Numer. Math.* **118** (2011), no. 3, 429–456.
- [3] A. Alonso Rodríguez and A. Valli, *Eddy Current Approximation of Maxwell Equations. Theory, Algorithms and Applications*, Model. Simul. Appl. 4, Springer, Milan, 2010.
- [4] A. C. Antoulas, *Approximation of Large-Scale Dynamical Systems*, Adv. Des. Control 6, Society for Industrial and Applied Mathematics, Philadelphia, 2005.
- [5] C. Beattie and S. Gugercin, Interpolatory projection methods for structure-preserving model reduction, *Systems Control Lett.* **58** (2009), no. 3, 225–232.
- [6] C. Beattie and S. Gugercin, Structure-preserving model reduction for nonlinear port-Hamiltonian systems, in: *50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)*, IEEE Press, Piscataway (2011), 6564–6569.
- [7] P. Benner, V. Mehrmann and D. C. Sorensen, *Dimension Reduction of Large-Scale Systems*, Lect. Notes Comput. Sci. Eng. 45, Springer, Berlin, 2005.
- [8] E. Celledoni, S. L. Eidnes, B. Owren and T. R. Ringholm, Dissipative numerical schemes on Riemannian manifolds with applications to gradient flows, *SIAM J. Sci. Comput.* **40** (2018), no. 6, A3789–A3806.
- [9] E. Celledoni, S. L. Eidnes, B. Owren and T. R. Ringholm, Energy-preserving methods on Riemannian manifolds, *Math. Comp.* **89** (2020), no. 322, 699–716.
- [10] E. Celledoni and E. H. Hoiseth, Energy-preserving and passivity-consistent numerical discretization of port-Hamiltonian systems, preprint (2017), <https://arxiv.org/abs/1706.08621>.
- [11] S. Chaturantabut, C. Beattie and S. Gugercin, Structure-preserving model reduction for nonlinear port-Hamiltonian systems, *SIAM J. Sci. Comput.* **38** (2016), no. 5, B837–B865.
- [12] D. Cohen and E. Hairer, Linear energy-preserving integrators for Poisson systems, *BIT* **51** (2011), no. 1, 91–101.

- [13] H. Egger, Structure preserving approximation of dissipative evolution problems, *Numer. Math.* **143** (2019), no. 1, 85–106.
- [14] E. Eich-Soellner and C. Führer, *Numerical Methods in Multibody Dynamics*, Eur. Consort. Math. Ind., B. G. Teubner, Stuttgart, 1998.
- [15] C. M. Elliott, The Cahn–Hilliard model for the kinetics of phase separation, in: *Mathematical Models for Phase Change Problems* (Óbidos 1988), Internat. Ser. Numer. Math. 88, Birkhäuser, Basel (1989), 35–73.
- [16] A. Ern and J.-L. Guermond, *Theory and Practice of Finite Elements*, Appl. Math. Sci. 159, Springer, New York, 2004.
- [17] D. Furihata and T. Matsuo, *Discrete Variational Derivative Method*, Chapman & Hall/CRC Numer. Anal. Sci. Comput., CRC Press, Boca Raton, 2011.
- [18] C. W. Gear, B. Leimkuhler and G. K. Gupta, Automatic integration of Euler–Lagrange equations with constraints, *J. Comput. Appl. Math.* **12–13** (1985), 77–90.
- [19] O. Gonzalez, Time integration and discrete Hamiltonian systems, *J. Nonlinear Sci.* **6** (1996), no. 5, 449–467.
- [20] E. Hairer, Energy-preserving variant of collocation methods, *JNAIAM. J. Numer. Anal. Ind. Appl. Math.* **5** (2010), no. 1–2, 73–84.
- [21] E. Hairer and C. Lubich, Energy-diminishing integration of gradient systems, *IMA J. Numer. Anal.* **34** (2014), no. 2, 452–461.
- [22] E. Hairer, C. Lubich and M. Roche, *The Numerical Solution of Differential-Algebraic Systems by Runge–Kutta methods*, Lecture Notes in Math. 1409, Springer, Berlin, 1989.
- [23] E. Hairer, C. Lubich and G. Wanner, *Geometric Numerical Integration*, Springer Ser. Comput. Math. 31, Springer, Heidelberg, 2010.
- [24] P. Kunkel and V. Mehrmann, *Differential-Algebraic Equations. Analysis and Numerical Solution*, EMS Textbk. Math., European Mathematical Society, Zürich, 2006.
- [25] B. Leimkuhler and S. Reich, *Simulating Hamiltonian Dynamics*, Cambridge Monogr. Appl. Comput. Math. 14, Cambridge University, Cambridge, 2004.
- [26] R. I. McLachlan, G. R. W. Quispel and N. Robidoux, Geometric integration using discrete gradients, *R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci.* **357** (1999), no. 1754, 1021–1045.
- [27] P. Monk, *Finite Element Methods for Maxwell’s Equations*, Numer. Math. Sci. Comput., Oxford University, New York, 2003.
- [28] R. V. Polyuga, Discussion on: “Passivity and structure preserving order reduction of linear port-Hamiltonian systems using Krylov subspaces” [mr2731511], *Eur. J. Control* **16** (2010), no. 4, 407–409.
- [29] K. Strehmel and R. Weiner, *Numerik gewöhnlicher Differentialgleichungen*, Teubner Math. Textb., B. G. Teubner, Stuttgart, 1995.
- [30] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, 2nd ed., Springer Ser. Comput. Math. 25, Springer, Berlin, 2006.