Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Seminar for Applied Mathematics SAM

Minimal l2 Norm Discrete Multiplier Method

E. Schulz and A. Wan

Research Report No. 2022-37
August 2022

Seminar für Angewandte Mathematik
Eidgenössische Technische Hochschule
CH-8092 Zürich
Switzerland

MINIMAL ℓ^2 NORM DISCRETE MULTIPLIER METHOD

ERICK SCHULZ[1] AND ANDY T. S. WAN[2]

Abstract. We introduce an extension to the Discrete Multiplier Method (DMM) [1], called Minimal ℓ_2 Norm Discrete Multiplier Method (MN-DMM), where conservative finite difference schemes for dynamical systems with multiple conserved quantities are constructed procedurally, instead of analytically as in the original DMM. For large dynamical systems with multiple conserved quantities, MN-DMM alleviates difficulties that can arise with the original DMM at constructing conservative schemes which satisfies the discrete multiplier conditions. In particular, MN-DMM utilizes the right Moore-Penrose pseudoinverse of the discrete multiplier matrix to solve an underdetermined least-square problem associated with the discrete multiplier conditions. We prove consistency and conservative properties of the MN-DMM schemes. We also introduce two variants – Mixed MN-DMM and MN-DMM using Singular Value Decomposition – and discuss their usage in practice. Moreover, numerical examples on various problems arising from the mathematical sciences are shown to demonstrate the wide applicability of MN-DMM and its relative ease of implementation compared to the original DMM.

1. Introduction

In recent decades, numerical methods which preserve intrinsic geometric structures of dynamical systems have gained considerable interest. Geometric numerical integrators are numerical methods which preserve underlying geometric features of solutions between successive time steps. An extensive summary of relevant literature is presented in [2]. In addition to striving for the traditional goals of high order accuracy, stability and ease of implementation, geometric numerical integrators seek to respect inherent geometric structures of dynamical systems to provide more accurate and stable solutions over *long-term* integration. Examples of geometric numerical integrators include symplectic integrators which preserve the symplectic two-forms associated with Hamiltonian flows [2], variational integrators which mimic the action principles of Lagrangian systems at the discrete level [3] and Lie group integrators which compose discrete Lie group actions to approximate underlying continuous Lie group flows [4].

Another important class of geometric numerical integrators is conservative integrators, which preserve conserved quantities or invariants associated with the underlying dynamics. In general, similar to symplectic methods [5], conservative numerical methods can have favorable long-term stability properties [6] over traditional numerical methods. Typically, such quantities include energy and momentum, but nontrivial time-dependent conserved quantities can also exist in dissipative

[1] Seminar in Applied Mathematics, ETH Zürich, Switzerland (erick.schulz@sam.math.ethz.ch).

[2] Department of Mathematics & Statistics, University of Northern British Columbia, Canada (andy.wan@unbc.ca).

systems [1]. Unfortunately, traditional numerical methods do not in general preserve all forms of conserved quantities. For instance, the barrier theorem by [7] states that no consistent Runge-Kutta method can preserve all polynomial invariants. Thus, non-traditional numerical methods are needed to preserve general forms of conserved quantities. Within the literature of geometric numerical integration, there are a few general classes of conservative integrators, such as projection methods [2], discrete gradient methods [8] and more recently Discrete Multiplier Methods (DMM) [1], which we briefly review next.

With projection methods, it is customary to first employ a traditional explicit integrator to advance one step in time, then to subsequently project the resulting numerical approximation onto the level set of the conserved quantities by solving a constrained optimization problem [2]. As discussed in [6], while projection methods are general conservative integrators, the projection step can become problematic if the level set of the conserved quantities contain connected components which are nearby each other. Indeed, if the time step size is not sufficiently small to account for small distances between neighboring connected components, the projection step may bring the numerical solution to the wrong connected component, leading to incorrect long-term trajectories.

The discrete gradient method exploits the fact that dynamical systems with conserved quantities can be expressed in a skew-symmetric gradient form [8]. This can then be used to derive conservative schemes using discrete gradient approximations. While the discrete gradient method is best suited for dynamical systems which naturally comes in such a skew-symmetric gradient representation, such as Hamiltonian systems, transforming a general dynamical system and utilizing the resulting skew-symmetric gradient form is not always straightforward in practice. Specifically, one drawback of the discrete gradient method is that the rank of the skew-symmetric tensor increases with the number of conserved quantities, making its applicability impractical for large dynamical systems with multiple invariants.

The Discrete Multiplier Method (DMM) was introduced in [1] as a new class of general conservative integrators that can preserve multiple conserved quantities of arbitrary forms up to machine precision. The main idea behind DMM is to discretize the so-called conservation law multiplier associated with the conserved quantities in such a way that discrete chain rules and other compatibility conditions are satisfied. In contrast to the discrete gradient method, DMM can work directly with the desired dynamical system, without having to reformulate the differential equations. Moreover, DMM requires only working with the so-called discrete multiplier matrix, whose number of rows increases with the number of conserved quantities while retaining a *constant tensor rank of two.* Such conservative integrators have recently been applied to a wide range of problems from the mathematical sciences, including many-body systems [9], vortex-blob models [10], and piecewise smooth systems [11]. In addition, for some applications such as Hamiltonian Monte Carlo [12], the gradient-free nature of DMM is advantageous over other conservative methods which require computation of the gradients of the conserved quantities.

Despite the wide applicability of DMM, there remains practical challenges when applying DMM on large dynamical systems with multiple conserved quantities. Specifically for each dynamical system, DMM proceeds in two main stages: First, derive an analytic conservative scheme using DMM; Second, solve the associated implicit conservative scheme. In this work, we extend the DMM framework by implicitly defining conservative schemes via a Moore-Penrose pseudoinverse of the associated discrete multiplier matrix. In doing so, DMM conservative schemes are constructed *procedurally* and solved *simultaneously*, without the need to first derive analytic conservative schemes. This extension of DMM widens its applicability to more complex dynamical systems and semi-discretizations of partial differential equations with conserved quantities.

This paper is organized as follows. In Section 2, we give a brief overview of the background material of DMM and introduce the relevant notations used throughout the paper. We then introduce the *Minimal ℓ_2 Norm Discrete Multiplier Method* (MN-DMM) in Section 3. Consistency and conservative properties of the implicitly defined schemes are established. In Section 4, we discuss practical issues that can arise in solving the MN-DMM schemes using the *Direct MN-DMM algorithm* via fixed point iterations. We prove convergence under appropriate conditions. Furthermore, we introduce in Section 4 two variants of the Direct MN-DMM algorithm, called *Mixed MN-DMM algorithm* and *Mixed MN-DMM algorithm using Singular Value Decomposition.* These two variants alleviate potential drawbacks with the Direct MN-DMM algorithm. In Section 5, numerical comparisons between the various MN-DMM approaches and traditional methods are presented for five examples chosen from a wide range of applications in the mathematical sciences. These includes Lotka–Volterra systems, the planar restricted three-body problem, the Lorenz system, the spherical point vortex problem and the evolution of geodesic curves in Schwarzschild geometry.

2. Background material

We retain most of the notations of the Discrete Multiplier Method from [1]. For details on the theoretical developments of DMM, see [1] and [6]. In this section, we summarize the content of these articles by stating some basic definitions along with a few necessary results.

2.1. Notation.

Throughout this paper, the integers $m, n, p, r \in \mathbb{N}$ are strictly positive. We denote open subsets of $\mathbb{R}^n$ by U, $U^{(1)}$, $U^{(2)}$, etc. We write U^r for the Cartesian product of r copies of U.

By $f \in C^p(U \to \mathbb{R}^m)$, we mean that the function f from U to $\mathbb{R}^m$ is at least p times continuously differentiable. We use a **bold** font to distinguish vector quantities from scalars. The Jacobian matrix of a differentiable vector-valued function $\boldsymbol{f}$ is denoted by $\partial_{\boldsymbol{x}}\boldsymbol{f} := \left[\partial f_i / \partial x_j\right]$.

Let $I \subset \mathbb{R}$ be an open time interval. We adopt Newton's notation $\dot{\boldsymbol{x}}$ for the time derivative of a curve $\boldsymbol{x} \in C^1(I \to U)$. If $\boldsymbol{x} \in C^p(I \to U)$, then $\boldsymbol{x}^{(p)}$ stands for its p-th time derivative. We use $D_t\boldsymbol{\psi}$ to distinguish the total time derivative of a vector-valued function $\boldsymbol{\psi} \in C^1(I \times U \to \mathbb{R}^m)$ from its partial time derivative $\partial_t\boldsymbol{\psi}$.

The vector space of $m \times n$ real matrices is written as $M_{m\times n}(\mathbb{R})$. It is equipped with the operator norm, which we denote by $\|\cdot\|_{m\times n}$. A superscript '$\top$' indicates the transpose of a matrix quantity, e.g. $\Lambda^\top$. We indicate the dependence of strictly positive constants in parentheses, e.g. $C(\Lambda)$. These constants are generic and should generally not be considered equal between different results.

2.2. Review on conservation law multipliers.

Next, we briefly review the theory of conservation law multipliers for first-order quasi-linear systems of ordinary differential equations—recall all quasi-linear systems can be made first-order by adding more variables. More precisely, for $p = 1, 2, \ldots$ and a source function $\boldsymbol{f} \in C^{p-1}(I \times U \to \mathbb{R}^n)$ with Lipschitz continuity in U, consider the continuous dynamical system $\boldsymbol{F} : I \times U \times U^{(1)} \to \mathbb{R}^n$ given by the initial value problem

$$\begin{aligned} \boldsymbol{F}(t, \boldsymbol{x}(t), \dot{\boldsymbol{x}}(t)) := \dot{\boldsymbol{x}}(t) - \boldsymbol{f}(t, \boldsymbol{x}(t)) &= \boldsymbol{0}, \\ \boldsymbol{x}(t^0) &= \boldsymbol{x}^0. \end{aligned} \tag{2.1}$$

It is a classical result that there exists a unique solution $\boldsymbol{x}(t) = (x_1(t), ..., x_n(t))$ of class C^p in a neighborhood of any initial condition $(t^0, \boldsymbol{x}^0) \in I \times U$. For simplicity, we will always assume from now on that I is a maximal interval of existence.

A function $\boldsymbol{\psi} \in C^1(I \times U \to \mathbb{R}^m)$ is called a conserved quantity of $\boldsymbol{F}$ if

$$D_t \boldsymbol{\psi}(t, \boldsymbol{x}(t)) = \boldsymbol{0} \tag{2.2}$$

for all $\boldsymbol{x} \in C^p(I \to U)$ such that $\boldsymbol{F}(t, \boldsymbol{x}(t), \dot{\boldsymbol{x}}(t)) = \boldsymbol{0}$. In other words, a conserved quantity remains constant along *solutions* of (2.1). In principle, a conserved quantity can depend on higher-order time derivatives of $\boldsymbol{x}$ also, but these can always be reformulated as $\boldsymbol{\psi}\,(t, \boldsymbol{x})$ by substituting the relation $\dot{\boldsymbol{x}} = \boldsymbol{f}(t, \boldsymbol{x}(t))$ and its differential consequences, as shown in [1, Sec. 3.1]. Thus, without loss of generality, we can focus on conserved quantities of such form.

We say that a matrix-valued function $\Lambda \in C(I \times U \times U^{(1)} \to M_{m \times n}(\mathbb{R}))$ is a conservation law multiplier of $\boldsymbol{F}$ if there exists $\boldsymbol{\psi} \in C^1(I \times U \to \mathbb{R}^m)$ satisfying

$$\Lambda(t, \boldsymbol{x}(t), \dot{\boldsymbol{x}}(t)) \boldsymbol{F}(t, \boldsymbol{x}(t), \dot{\boldsymbol{x}}(t)) = D_t \boldsymbol{\psi}(t, \boldsymbol{x}(t)) \tag{2.3}$$

for all $\boldsymbol{x} \in C^1(I \to U)$. We insist that (2.3) must hold for all *arbitrary* differentiable functions—not only for solutions of $\boldsymbol{F}$ as previously required for (2.2).

In general, there can be many different conservation law multipliers satisfying (2.3) for the same $\boldsymbol{\psi}$. However, the following theorem guarantees that there is a one-to-one correspondence between conservation law multipliers of the form $\Lambda(t, \boldsymbol{x})$ and zero-order conserved quantities of $\boldsymbol{F}$, cf. [1, Thm. 4].

Theorem 2.1. *Let $\boldsymbol{\psi} \in C^1(I \times U \to \mathbb{R}^m)$. There exists a unique conservation law multiplier $\Lambda \in C(I \times U \to M_{m \times n}(\mathbb{R}))$ of $\boldsymbol{F}$ associated with the function $\boldsymbol{\psi}$ if and only if $\boldsymbol{\psi}$ is a conserved quantity of $\boldsymbol{F}$. If so, the correspondence identities*

$$\Lambda(t, \boldsymbol{x}) = \partial_{\boldsymbol{x}} \boldsymbol{\psi}(t, \boldsymbol{x}), \tag{2.4a}$$

$$\Lambda(t, \boldsymbol{x}) \boldsymbol{f}(t, \boldsymbol{x}) = -\partial_t \boldsymbol{\psi}(t, \boldsymbol{x}), \tag{2.4b}$$

are satisfied for any arbitrary function $\boldsymbol{x} \in C^1(I \to U)$.

We will commonly refer to (2.4a) and (2.4b) as *multiplier conditions*. Importantly, (2.4a) explicitly characterizes the conservation law multiplier.

For the purpose of deriving conservative schemes, there is some freedom in choosing the dimension of $\boldsymbol{\psi}$. For a given dynamical system, the components of the vector-valued function $\boldsymbol{\psi}$ consist of known conserved quantities of interest. How many are to be preserved using DMM is up to the one's discretion. In practice, there are typically much fewer conserved quantities than the dimension of $\boldsymbol{F}$. We thus take for granted the following assumption.

Assumption 2.2. We suppose that $m < n$ and assume that $\Lambda(t, \boldsymbol{x})$ has full row rank within $I \times U$.

Remark 2.3. Notice that $\Lambda(t, \boldsymbol{x})$ having full row rank in Assumption 2.2 is equivalent to the conserved quantities being linearly independent on $I \times U$.

2.3. **Review of DMM.** The idea behind DMM is to provide a discrete framework which preserves the structure of the continuous theory of conservation law multipliers from Section 2.2. Specifically, it establishes discrete analogues of the multiplier conditions (2.4a) and (2.4b).

Let $t^0 < t^1 < ... < t^k < ...$ be a sequence in I having a largest time step of size $\tau = \sup_k (t^{k+1} - t^k) < \infty$. Let W be a finite dimensional normed vector space. A r-step function $\boldsymbol{g}^\tau \in C^{p+q}(I \times U^{r+1} \to W)$ is said to be consistent of order q to a function $\boldsymbol{g} \in C^{p+q}(I \times U \times U^{(1)} \times ... \times U^{(q)} \to W)$ if for any $\boldsymbol{x} \in C^{p+q}(I \to U)$, there exists a constant $C(\boldsymbol{g}, \boldsymbol{x}) > 0$ independent of τ such that

$$\left\| \boldsymbol{g}(t^k, \boldsymbol{x}(t^k), ..., \boldsymbol{x}^{(p)}(t^k)) - \boldsymbol{g}^\tau(t^k, \boldsymbol{x}(t^{k+1}), ..., \boldsymbol{x}(t^{k-r+1})) \right\|_W \leq C(\boldsymbol{g}, \boldsymbol{x})\, \tau^q. \tag{2.5}$$

If so, we simply write $\boldsymbol{g}^\tau = \boldsymbol{g} + \mathcal{O}(\tau^q)$. This definition is general enough to provide a notion of consistency for both vector-valued and matrix-valued quantities. In the following sections, we will encounter $W = \mathbb{R}^m$ and $M_{m \times n}(\mathbb{R})$.

Denote the approximation at time t_k of the exact solution $\boldsymbol{x}(t^k)$ by $\boldsymbol{x}^k$. Let $\boldsymbol{F}^\tau$ be a consistent r-step function to $\boldsymbol{F}$ and suppose that $\boldsymbol{\psi}^\tau$ is a consistent $(r-1)$-step function to $\boldsymbol{\psi}$. We say that the r-step method $\boldsymbol{F}^\tau$ is conservative in $\boldsymbol{\psi}^\tau$ if

$$\boldsymbol{\psi}^\tau(t^k, \boldsymbol{x}^k, ..., \boldsymbol{x}^{k-r+1}) = \boldsymbol{\psi}^\tau(t^{k+1}, \boldsymbol{x}^{k+1}, ..., \boldsymbol{x}^{k-r+2}) \tag{2.6}$$

whenever $\boldsymbol{x}^{k+1}$ satisfies $\boldsymbol{F}^\tau\left(t^k, \boldsymbol{x}^{k+1}, ..., \boldsymbol{x}^{k-r+1}\right) = \boldsymbol{0}$.

When $D_t^\tau \boldsymbol{\psi}$ is an r-step function consistent to $D_t \boldsymbol{\psi}$, we say that it is constant compatible with $\boldsymbol{\psi}^\tau$ if $D_t^\tau \boldsymbol{\psi}\left(t^k, ..., \boldsymbol{x}^{k+1}, ..., \boldsymbol{x}^{k-r+1}\right) = \boldsymbol{0}$ implies that (2.6) holds.

Assumption 2.4. Henceforth, we will always suppose that $\boldsymbol{f}^\tau$, $D_t^\tau \boldsymbol{x}$, $D_t^\tau \boldsymbol{\psi}$, $\partial_t^\tau \boldsymbol{\psi}$ and Λ^τ are r-step functions consistent of order q respectively to $\boldsymbol{f}$, $\dot{\boldsymbol{x}}$, $D_t \boldsymbol{\psi}$, $\partial_t \boldsymbol{\psi}$ and Λ, where Λ is a conservation law multiplier of $\boldsymbol{F}$ associated with the conserved quantity $\boldsymbol{\psi}$. We assume that $D_t^\tau \boldsymbol{\psi}$ is constant compatible with a discrete $(r-1)$-step function $\boldsymbol{\psi}^\tau$.

The following theorem is the heart of DMM, cf. [1, Thm. 4.5].

Theorem 2.5. *Let $\boldsymbol{f}^\tau_{\text{DMM}}$ be a r-step function consistent of order q to $\boldsymbol{f}$. Under assumptions 2.2 and 2.4, if the discrete compatibility conditions*

$$\Lambda^\tau D_t^\tau \boldsymbol{x} = D_t^\tau \boldsymbol{\psi} - \partial_t^\tau \boldsymbol{\psi}, \tag{2.7a}$$

$$\Lambda^\tau \boldsymbol{f}^\tau_{\text{DMM}} = -\partial_t^\tau \boldsymbol{\psi}, \tag{2.7b}$$

hold for all $(t^k, \boldsymbol{x}^{k+1}, ..., \boldsymbol{x}^{k-r+1}) \in I \times U^{r+1}$ satisfying

$$\boldsymbol{F}^\tau_{\text{DMM}}(t^k, \boldsymbol{x}^{k+1}, ..., \boldsymbol{x}^{k-r+1}) = \boldsymbol{0},$$

where

$$\boldsymbol{F}^\tau_{\text{DMM}} := D_t^\tau \boldsymbol{x} - \boldsymbol{f}^\tau_{\text{DMM}},$$

then the r-step method defined by (2.5) *is conservative in $\boldsymbol{\psi}^\tau$. Moreover, it is consistent of at least order q to $\boldsymbol{F}$, and for any sufficiently differentiable arbitrary function $\boldsymbol{x}$, the discrete quantities satisfy*

$$\Lambda^\tau D_t^\tau \boldsymbol{x} - D_t^\tau \boldsymbol{\psi} - \partial_t^\tau \boldsymbol{\psi} = \mathcal{O}(\tau^q),$$
$$\Lambda^\tau \boldsymbol{f}^\tau_{\text{DMM}} + \partial_t^\tau \boldsymbol{\psi} = \mathcal{O}(\tau^q).$$

Remark 2.6. The discrete compatibility condition (2.7a) corresponds *implicitly* to (2.4a) by the chain rule:

$$\Lambda(t, \boldsymbol{x})\dot{\boldsymbol{x}} = (\partial_{\boldsymbol{x}} \boldsymbol{\psi}) D_t \boldsymbol{x} = (\partial_{\boldsymbol{x}} \boldsymbol{\psi}) D_t \boldsymbol{x} + \partial_t \boldsymbol{\psi} - \partial_t \boldsymbol{\psi} = D_t \boldsymbol{\psi} - \partial_t \boldsymbol{\psi}.$$

3. Minimal ℓ_2 Norm DMM

So far, the construction of conservative schemes using DMM reduces to satisfying the discrete multiplier conditions (2.7a) and (2.7b). While (2.7a) can be resolved through the use of discrete chain rules as described in [1, Thm. 22], resolving (2.7b) relies on the local solvability of Λ^τ. As discussed in [1, Thm. 20], (2.7b) can be satisfied by locally inverting an $m \times m$ submatrix $\tilde{\Lambda}^\tau$ of Λ^τ for a given dynamical system with m conserved quantities. However, this traditional approach of DMM has two main drawbacks.

(1) First, analytical matrix inversion of a submatrix of Λ^τ becomes difficult, if not impractical, as m increases. As it will be highlighted in examples later in Section 5, even a small number of conserved quantities can pose a significant challenge to construct conservative schemes using the traditional DMM approach.

(2) Second, due to the local nature of the rank of the submatrix $\tilde{\Lambda}^\tau$, its invertibility may vary depending on the phase space region where the conservative scheme is to be evaluated, making the traditional DMM approach cumbersome to implement for dynamical systems with complex phase spaces.

Indeed, there are alternate techniques, such as the *method of undetermined coefficients* used in [1] and [9], that could alleviate some of these difficulties, but it still relies on the need to construct analytic conservative scheme, which can be difficult to apply for large dynamical systems with multiple conserved quantities.

In this section, we tackle the problem of solving (2.7b) systematically for large dynamical systems with multiple conserved quantities, without the need to construct analytic conservative schemes. The proposed new approach paves the way for the procedural construction of globally defined conservative schemes using DMM. Thus, this leads to a promising starting point for conservative discretizations of large dynamical systems which can only be evaluated procedurally and also in semi-discretization of partial differential equations.

3.1. Minimal ℓ^2 Norm Discrete Multiplier Method.

We define the *Minimal ℓ^2 Norm Discrete Multiplier Method*, or *Minimal Norm DMM* (MN-DMM), as the conservative scheme

$$\boldsymbol{f}^\tau_{\text{MN}} := \boldsymbol{f}^\tau - (\Lambda^\tau)^+(\Lambda^\tau \boldsymbol{f}^\tau + \partial_t^\tau \boldsymbol{\psi}), \tag{3.1}$$

where $\boldsymbol{f}^\tau$ is *any* consistent scheme to $\boldsymbol{f}$ and $(\Lambda^\tau)^+ = (\Lambda^\tau)^\top(\Lambda^\tau(\Lambda^\tau)^\top)^{-1}$ is the unique *right* Moore-Penrose pseudoinverse of Λ^τ. By construction, it can be readily check that $\boldsymbol{f}^\tau_{\text{MN}}$ satisfies (2.7b). We will discuss the theoretical analysis of the implicit scheme $\boldsymbol{f}^\tau_{\text{MN}}$ shortly in Section 3.2, where we will show that it is indeed conservative and well-defined for sufficiently small τ. Different algorithmic choices for the practical evaluation of the second term on the right-hand side of (3.1) will be discussed in Section 4.

Let us motivate the expression of (3.1) in two ways and the reasons for its name[1]:

(I) First, one can view (3.1) as "projecting" an r-step scheme $\boldsymbol{f}^\tau$ consistent to $\boldsymbol{f}$ onto a scheme satisfying (2.7b), hence resulting in a conservative scheme implicitly. To better see this, suppose that the vector of conserved quantities

[1] For a general introduction to both *under*determined and *over*determined ℓ^2 minimization problems, see the first chapters of the monograph [13, Chap. 1 & 2], where orthogonal projections, normal equations and the Moore-Penrose inverse are studied in detail.

$\boldsymbol{\psi}$ is independent of time explicitly, i.e. $\partial_t^\tau \boldsymbol{\psi} = \mathbf{0}$. Then, satisfying condition (2.7b) is equivalent to asking for the numerical scheme $\boldsymbol{f}_{\text{MN}}^\tau$ to be in $\ker(\Lambda^\tau) = \text{Ran}((\Lambda^\tau)^\top)^\perp$. In other words, we seek to find a numerical scheme that is orthogonal to the row space of the discrete multiplier matrix Λ^τ. Since the projection operator onto the row space of Λ^τ can be expressed as [13, Eq. 1.2.29]

$$P_{\text{Ran}((\Lambda^\tau)^\top)} = (\Lambda^\tau)^+\Lambda^\tau,$$

the projection operator onto its orthogonal complement is then given by

$$P_{\ker(\Lambda^\tau)} = P_{\text{Ran}((\Lambda^\tau)^\top))^\perp} = I_{n\times n} - P_{(\Lambda^\tau)^\top} = I_{n\times n} - (\Lambda^\tau)^+\Lambda^\tau,$$

where $I_{n\times n}$ denotes the $n \times n$ identity matrix. So in the case of time-independent conserved quantities, the MN-DMM scheme is equivalent to $\boldsymbol{f}_{\text{MN}}^\tau = P_{\ker(\Lambda^\tau)}\boldsymbol{f}^\tau$, which automatically satisfies the discrete multiplier condition (2.7b). Indeed, this follows by definition, since

$$\Lambda^\tau P_{\ker(\Lambda^\tau)} = \Lambda^\tau(I_{n\times n} - (\Lambda^\tau)^+\Lambda^\tau) = \Lambda^\tau - \Lambda^\tau = 0_{m\times n}.$$

(II) Alternatively, we can also take the point of view that any scheme satisfying (2.7b) solves an undetermined linear system. Specifically, the general solution to (2.7b) is given by $\boldsymbol{f}_0^\tau + \boldsymbol{f}_P^\tau$, where $\boldsymbol{f}_0^\tau \in \ker(\Lambda^\tau)$ and $\boldsymbol{f}_P^\tau$ is any particular solution of (2.7b). Note that by direct substitution, one particular choice is provided by

$$\boldsymbol{f}_P^\tau := (\Lambda^\tau)^+ \left(-\partial_t^\tau \boldsymbol{\psi}\right).$$

Since $\ker(\Lambda^\tau) = \text{Ran}(P_{\ker(\Lambda^\tau)})$, then for any consistent $\boldsymbol{f}^\tau$, $\boldsymbol{f}_0^\tau = P_{\ker(\Lambda^\tau)}\boldsymbol{f}^\tau$ and we arrive at the MN-DMM scheme:

$$\boldsymbol{f}_0^\tau + \boldsymbol{f}_P^\tau = \boldsymbol{f}^\tau - (\Lambda^\tau)^+(\Lambda^\tau\boldsymbol{f}^\tau + \partial_t^\tau\boldsymbol{\psi}) = \boldsymbol{f}_{\text{MN}}^\tau.$$

Moreover, such a particular choice for $\boldsymbol{f}_P^\tau$ has the *minimal ℓ^2 norm* in the sense that for any consistent scheme $\boldsymbol{f}^\tau$, the MN-DMM scheme $\boldsymbol{f}_{\text{MN}}^\tau$ is the closest scheme to $\boldsymbol{f}^\tau$ in the ℓ^2 norm satisfying (2.7b):

$$\boldsymbol{f}_{\text{MN}}^\tau = \underset{\Lambda^\tau\tilde{\boldsymbol{f}}^\tau = -\partial_t^\tau\boldsymbol{\psi}}{\text{argmin}} \left\| \boldsymbol{f}^\tau - \tilde{\boldsymbol{f}}^\tau \right\|_2. \tag{3.2}$$

Indeed, this follows from the observation that for any $\tilde{\boldsymbol{f}}^\tau$ satisfying (2.7b),

$$\begin{aligned}\|\boldsymbol{f}^\tau - \tilde{\boldsymbol{f}}^\tau\|_2^2 &= \|(\boldsymbol{f}^\tau - \boldsymbol{f}_{\text{MN}}^\tau) + (\boldsymbol{f}_{\text{MN}}^\tau - \tilde{\boldsymbol{f}}^\tau)\|_2^2 \\ &= \|\boldsymbol{f}^\tau - \boldsymbol{f}_{\text{MN}}^\tau\|_2^2 + \|\boldsymbol{f}_{\text{MN}}^\tau - \tilde{\boldsymbol{f}}^\tau\|_2^2 \geq \|\boldsymbol{f}^\tau - \boldsymbol{f}_{\text{MN}}^\tau\|_2^2.\end{aligned}$$

Orthogonality in the second equality follows from $(\boldsymbol{f}_{\text{MN}}^\tau - \tilde{\boldsymbol{f}}^\tau) \in \ker(\Lambda^\tau)$ by (2.7b). Moreover, $(\boldsymbol{f}^\tau - \boldsymbol{f}_{\text{MN}}^\tau) \perp \ker(\Lambda^\tau)$, since for any $\boldsymbol{f}_0^\tau \in \ker(\Lambda^\tau)$,

$$\begin{aligned}(\boldsymbol{f}^\tau - \boldsymbol{f}_{\text{MN}}^\tau) \cdot \boldsymbol{f}_0^\tau &= (\Lambda^\tau)^+(\Lambda^\tau\boldsymbol{f}^\tau + \partial_t^\tau\boldsymbol{\psi}) \cdot \boldsymbol{f}_0^\tau \\ &= (\Lambda^\tau(\Lambda^\tau)^\top)^{-1}(\Lambda^\tau\boldsymbol{f}^\tau + \partial_t^\tau\boldsymbol{\psi}) \cdot (\Lambda^\tau\boldsymbol{f}_0^\tau) = \mathbf{0}.\end{aligned}$$

3.2. **Theory.**

Before we prove that (3.1) defines a consistent conservative scheme, we will need show that the pseudoinverses $(\Lambda^\tau)^+$ is well-defined and are uniformly bounded for small enough τ.

Recall the hypotheses of assumptions 2.2 and 2.4 introduced in Section 2.3:

- $\Lambda(t, \boldsymbol{x})$ has full row rank in $I \times U$,
- $\Lambda^\tau = \Lambda + \mathcal{O}(\tau^q)$.

These hypotheses imply that for any $\boldsymbol{x} \in C^{p+q}(I \to \mathbb{R}^n)$, the rows of Λ^τ are linearly independent for τ small enough.

Lemma 3.1. *Under assumptions 2.2 and 2.4, for any $\boldsymbol{x} \in C^{p+q}(I \to \mathbb{R}^n)$, there exists $\tau_0 > 0$ such that Λ^τ has for full row rank whenever $\tau < \tau_0$.*

Proof. We argue by contradiction. Suppose that for some $\boldsymbol{x} \in C^{p+q}\,(I \to U)$, there exists a sequence of *unit* vectors $(\boldsymbol{u}_k)_{k\in\mathbb{N}} \in \mathbb{R}^n$ and a sequence of parameters $(\tau_k)_{k\in\mathbb{N}} \in \mathbb{R}$ with $\tau_k \longrightarrow 0$ as $k \to \infty$, such that $(\Lambda^{\tau_k})^\top \boldsymbol{u}_k = \boldsymbol{0}$ for all $k \in \mathbb{N}$. Then, since the unit sphere $S^{n-1} = \{\boldsymbol{x} \in \mathbb{R}^n | \|\boldsymbol{x}\| = 1\}$ is compact in $\mathbb{R}^m$, it follows from the hypothesis that Λ has full row rank and the extreme value theorem that a positive lower bound $\alpha := \min_{\|\boldsymbol{v}\|=1} \|\Lambda^\top \boldsymbol{v}\|$ is achieved. Since Λ^τ is consistent with Λ, we therefore obtain the contradiction that

$$0 < \alpha \le \|\Lambda^\top \boldsymbol{u}_k\| \le \|(\Lambda^\top - (\Lambda^{\tau_k})^\top)\boldsymbol{u}_k\| \le \|\Lambda - \Lambda^{\tau_k}\|_{m\times n} \longrightarrow 0, \quad \text{as } k \to \infty.$$

□

As a consequence, the Moore-Penrose right inverse of Λ^τ is well-defined and given by [13, Eq. 1.2.27]

$$(\Lambda^\tau)^+ = (\Lambda^\tau)^\top (\Lambda^\tau (\Lambda^\tau)^\top)^{-1}, \qquad \tau < \tau_0. \tag{3.3}$$

Moreover, we see that $\boldsymbol{f}^\tau_{\text{MN}}$ is a well-defined r-step function in the sense of Section 2.3. Indeed, it is clear from $(\Lambda^\tau(\Lambda^\tau)^\top)^{-1} = (\det(\Lambda^\tau(\Lambda^\tau)^\top))^{-1}\mathrm{adj}(\Lambda^\tau(\Lambda^\tau)^\top)$ that the matrix $(\Lambda^\tau)^+$ is of class C^l whenever $\Lambda^\tau \in C^l$, $l \in \mathbb{N}$.

Thanks to the next lemma, the expression (3.3) is also useful in proving that for any $\boldsymbol{x} \in C^{p+q}\,(I \to U)$, the parametrized family $\{(\Lambda^\tau)^+\}_\tau$ is eventually uniformly bounded in τ as $\tau \to 0$.

Lemma 3.2. *Under assumptions 2.2 and 2.4, for any $\boldsymbol{x} \in C^{p+q}(I \to U)$, there exists a parameter $\tau^0 > 0$ and a constant $C(\Lambda, \boldsymbol{x}) > 0$ independent of τ such that whenever $0 < \tau < \tau^0$, the inverse $(\Lambda^\tau(\Lambda^\tau)^\top)^{-1}$ exists and satisfies*

$$\|(\Lambda^\tau(\Lambda^\tau)^\top)^{-1}\|_{m\times m} \le C(\Lambda, \boldsymbol{x}). \tag{3.4}$$

Proof. We see from combining the estimates

$$\begin{aligned}\|\Lambda^\tau(\Lambda^\tau)^\top - \Lambda\Lambda^\top\|_{m\times m} &\le \|\Lambda^\tau(\Lambda^\tau)^\top - \Lambda^\tau\Lambda^\top\|_{m\times m} + \|\Lambda^\tau\Lambda^\top - \Lambda\Lambda^\top\|_{m\times m} \\ &\le (\,\|\Lambda^\tau\|_{m\times n} + \|\Lambda^\top\|_{n\times m})\,\|\Lambda^\tau - \Lambda\|_{m\times n}\end{aligned}$$

and

$$\|\Lambda^\tau\|_{m\times n} \le \|\Lambda^\tau - \Lambda\|_{m\times n} + \|\Lambda\|_{m\times n}, \tag{3.5}$$

that $\Lambda^\tau(\Lambda^\tau)^\top = \Lambda\Lambda^\top + \mathcal{O}(\tau^q)$. Therefore, the desired conclusion follows from a well-known result concerning perturbation of regular matrices [14, Thm. 1.5], which in the current setting states that if there exists $\tau_0 > 0$ such that

$$\|\Lambda^\tau(\Lambda^\tau)^\top - \Lambda\Lambda^\top\|_{m\times m} < \|(\Lambda\Lambda^\top)^{-1}\|^{-1}_{m\times m},$$

then the matrices $\Lambda^\tau\,(\Lambda^\tau)^\top$ are invertible on the interval $(0, \tau_0)$ and bounded by a constant independent of τ. □

Corollary 3.3. *Under assumptions 2.2 and 2.4, for any $\boldsymbol{x} \in C^{p+q}(I \to U)$, there exists a parameter $\tau^0 > 0$ and a constant $C(\Lambda, \boldsymbol{x}) > 0$ independent of τ such that*

$$\|(\Lambda^\tau)^+\|_{n\times m} \le C(\Lambda, \boldsymbol{x}), \qquad \tau < \tau_0.$$

Proof. Based on (3.3), we find that for any $\boldsymbol{x} \in C^{p+q}(I \to U)$, we have

$$\|(\Lambda^\tau)^+\|_{n\times m} = \|(\Lambda^\tau)^\top(\Lambda^\tau(\Lambda^\tau)^\top)^{-1}\|_{n\times m} \leq \|(\Lambda^\tau)^\top\|_{m\times m}\|(\Lambda^\tau(\Lambda^\tau)^\top)^{-1}\|_{m\times m}$$

for τ small enough. In particular, τ_0 can be chosen as in the proof of Lemma 3.2. □

The next theorem shows that upon satisfying the discrete multiplier condition (2.7a), which as previously mentioned can be resolved by discrete chain rules [1], the MN-DMM indeed leads to a conservative scheme.

Theorem 3.4. *Under assumptions 2.2 and 2.4, suppose that the discrete quantities of Section 2.3 satisfy the compatibility condition* (2.7a) *for all* $(t^k, \boldsymbol{x}^{k+1}, ..., \boldsymbol{x}^{k-r+1}) \in I \times U^{r+1}$ *such that*

$$\boldsymbol{F}^\tau_{\text{MN}}(t^k, \boldsymbol{x}^{k+1}, ..., \boldsymbol{x}^{k-r+1}) = \boldsymbol{0}, \tag{3.6}$$

where

$$\boldsymbol{F}^\tau_{\text{MN}} := D^\tau_t \boldsymbol{x} - \boldsymbol{f}^\tau_{\text{MN}}.$$

Then, the r*-step method defined by* (3.6) *is conservative in* $\boldsymbol{\psi}^\tau$*. Moreover, it is consistent of at least order* q *to the function* $\boldsymbol{F}$ *defined in* (2.1)*, and for any* $\boldsymbol{x} \in C^{p+q}(I \to \mathbb{R}^n)$ *the discrete quantities satisfy*

$$\Lambda^\tau D^\tau_t \boldsymbol{x} - D^\tau_t \boldsymbol{\psi} - \partial^\tau_t \boldsymbol{\psi} = \mathcal{O}(\tau^q), \tag{3.7a}$$

$$\Lambda^\tau \boldsymbol{f}^\tau_{\text{MN}} + \partial^\tau_t \boldsymbol{\psi} = \mathcal{O}(\tau^q). \tag{3.7b}$$

Proof. Our goal is to resort to Theorem 2.5. Two ingredients are required.

First, we need to confirm that the discrete function $\boldsymbol{f}^\tau_{\text{MN}}$ verifies the second discrete multiplier condition (2.7b). This holds by construction. Since by definition $\Lambda^\tau(\Lambda^\tau)^+ = I_{m\times m}$, multiplying both sides of (3.1) by Λ^τ immediately yields

$$\Lambda^\tau \boldsymbol{f}^\tau_{\text{MN}} = \Lambda^\tau \boldsymbol{f}^\tau - \Lambda^\tau(\Lambda^\tau)^+(\Lambda^\tau \boldsymbol{f}^\tau + \partial^\tau_t \boldsymbol{\psi}) = -\partial^\tau_t \boldsymbol{\psi}.$$

Second, we need to show that $\boldsymbol{f}^\tau_{\text{MN}} = \boldsymbol{f} + \mathcal{O}(\tau^q)$. Since the triangle inequality yields

$$\|\boldsymbol{f} - \boldsymbol{f}^\tau_{MN}\| \leq \|\boldsymbol{f} - \boldsymbol{f}^\tau\| + \|(\Lambda^\tau)^+\|_{n\times m}\|\Lambda^\tau \boldsymbol{f}^\tau + \partial^\tau_t \psi\|, \tag{3.8}$$

it follows from Corollary 3.3 that we only need to verify that $\Lambda^\tau \boldsymbol{f}^\tau + \partial^\tau_t \psi = \mathcal{O}(\tau^q)$.

Consider the estimate

$$\begin{aligned}\|\Lambda^\tau \boldsymbol{f}^\tau + \partial^\tau_t \boldsymbol{\psi}\| &\leq \|\Lambda^\tau \boldsymbol{f}^\tau - \Lambda^\tau \boldsymbol{f}\| + \|\Lambda^\tau \boldsymbol{f} - \Lambda \boldsymbol{f}\| + \|\Lambda \boldsymbol{f} + \partial^\tau_t \boldsymbol{\psi}\| \\ &\leq \|\Lambda^\tau\|_{m\times n}\|\boldsymbol{f}^\tau - \boldsymbol{f}\| + \|\Lambda^\tau - \Lambda\|_{m\times n}\|\boldsymbol{f}\| + \|\Lambda \boldsymbol{f} + \partial^\tau_t \boldsymbol{\psi}\|.\end{aligned} \tag{3.9}$$

The key observation is that since Λ is a conservation law multiplier of $\boldsymbol{F}$ associated to $\boldsymbol{\psi}$ by hypothesis, it satisfies the correspondence identity (2.4b), i.e. $\Lambda \boldsymbol{f} = -\partial_t \boldsymbol{\psi}$. Introducing $\partial_t \boldsymbol{\psi}$ in the last term of (3.9) yields

$$\|\Lambda \boldsymbol{f} + \partial^\tau_t \boldsymbol{\psi}\| = \|\partial_t \boldsymbol{\psi} - \partial^\tau_t \boldsymbol{\psi}\|. \tag{3.10}$$

Upon inserting (3.10) in (3.9), then (3.9) in (3.8), the proof follows by consistency of $\boldsymbol{f}^\tau$, Λ^τ and $\partial^\tau_t \boldsymbol{\psi}$ to $\boldsymbol{f}$, Λ and $\partial_t \boldsymbol{\psi}$, respectively. □

4. Practical Implementations

As the MN-DMM scheme (3.6) is implicitly defined, we turn to an iterative fixed point algorithm in order to converge to the desired conservative scheme and simultaneously solve the associated nonlinear equations. Following [1], we will focus on one-step conservative methods constructed by using divided differences for

$$\boldsymbol{\psi}^\tau(t^k, \boldsymbol{x}^k) := \boldsymbol{\psi}(t^k, \boldsymbol{x}^k), \tag{4.1a}$$

$$D_t^\tau \boldsymbol{x}(t^k, \boldsymbol{x}^{k+1}, \boldsymbol{x}^k) := \frac{\boldsymbol{x}^{k+1} - \boldsymbol{x}^k}{t^{k+1} - t^k}, \tag{4.1b}$$

$$D_t^\tau \boldsymbol{\psi}(t^k, \boldsymbol{x}^{k+1}, \boldsymbol{x}^k) := \frac{\boldsymbol{\psi}(t^k, \boldsymbol{x}^{k+1}) - \boldsymbol{\psi}(t^k, \boldsymbol{x}^k)}{t^{k+1} - t^k}, \tag{4.1c}$$

which are used throughout in the numerical results presented in Section 5.

It is clear that these discrete quantities are consistent single-step functions of at least first-order to their continuous counterpart. Conveniently, constant compatibility of $D_t^\tau \boldsymbol{\psi}$ with $\boldsymbol{\psi}^\tau$ is immediate. We refer to [1] for the derivation of a single-step function Λ^τ using discrete chain rules that satisfy condition (2.7a).

Some higher-order multi-step DMM schemes were constructed in [6]. Also note that first-order symmetric schemes can turn out to be high-order as well [2, Chapter II.3, Theorem 3.2], which was studied in the conservative DMM schemes for many-body problems [9], vortex blob methods [10], and Hamiltonian Monte Carlo methods [12].

4.1. **Analytic expressions of MN-DMM for small number of conserved quantities.**

For a small number of conserved quantities, it is in fact analytically tractable to write out the expressions of MN-DMM given by (3.1). For ease of future reference, we write out the explicit MN-DMM schemes for preserving one and two conserved quantities, i.e. $m = 1$ and 2. Specifically, the case $m = 1$ leads to a simple way to enable conservation for an arbitrary consistent scheme and in a gradient-free manner. For instance, this could be highly relevant to physical systems where energy conservation is important, such as for Hamiltonian systems.

4.1.1. *Analytic expression for m=1.* For a single scalar conserved quantity, the discrete multiplier matrix $\Lambda^\tau \in M_{1\times n}(\mathbb{R})$ is the row vector

$$\Lambda^\tau(t^k, \boldsymbol{x}^{k+1}, \boldsymbol{x}^k) := \frac{\Delta\psi}{\Delta\boldsymbol{x}}^\top(t^k, \boldsymbol{x}^{k+1}, \boldsymbol{x}^k),$$

where $\dfrac{\Delta\psi}{\Delta\boldsymbol{x}}$ denotes the column vector of partial divided differences of ψ with respect to $\boldsymbol{x}$ for a specific permutation of S_{n+1} satisfying the discrete chain rule (2.7a)[2]. Since $\Lambda^\tau(\Lambda^\tau)^\top = \left\|\dfrac{\Delta\psi}{\Delta\boldsymbol{x}}\right\|_2^2$ is a scalar quantity in this case, we see that the MN-DMM scheme of (3.1) for $m = 1$ is given by

$$\boldsymbol{f}^\tau_{\text{MN}} := \boldsymbol{f}^\tau - \frac{1}{\left\|\frac{\Delta\psi}{\Delta\boldsymbol{x}}\right\|_2^2}\left(\frac{\Delta\psi}{\Delta\boldsymbol{x}}^\top \boldsymbol{f}^\tau + \partial_t^\tau\psi\right)\frac{\Delta\psi}{\Delta\boldsymbol{x}}, \tag{4.2}$$

[2] Details on divided difference calculus and explicit formulas for $\frac{\Delta\psi}{\Delta\boldsymbol{x}}$ are in Appendix B of [1].

where we have suppressed the arguments $(t^k, \boldsymbol{x}^{k+1}, \boldsymbol{x}^k)$ for clarity. As seen in Section 3.1, for time-independent ψ, (4.2) can be viewed as subtracting off the projection of the scheme $\boldsymbol{f}^\tau$ onto the orthogonal complement of the discrete multiplier's kernel. Since the kernel is in this case the multi-dimensional plane perpendicular to the vector $\frac{\Delta\psi}{\Delta\boldsymbol{x}}$, its orthogonal complement is simply the span of the latter, and the resulting scheme reads

$$\boldsymbol{f}^\tau_{\text{MN}} := \boldsymbol{f}^\tau - \alpha\frac{\Delta\psi}{\Delta\boldsymbol{x}}, \qquad \text{where } \alpha := \frac{1}{\left\|\frac{\Delta\psi}{\Delta\boldsymbol{x}}\right\|_2^2}\frac{\Delta\psi}{\Delta\boldsymbol{x}}^\top\boldsymbol{f}^\tau. \tag{4.3}$$

In other words, $\alpha\frac{\Delta\psi}{\Delta\boldsymbol{x}}$ is the scalar projection of $\boldsymbol{f}^\tau$ onto $\frac{\Delta\psi}{\Delta\boldsymbol{x}}$ and $\boldsymbol{f}^\tau_{\text{MN}}$ is the vector projection of $\boldsymbol{f}^\tau$ onto the kernel of $\Lambda^\tau = \frac{\Delta\psi}{\Delta\boldsymbol{x}}^\top$, as discussed in Section 3.1. The expression in (4.3) conveys how MN-DMM schemes arise as ℓ^2-projections. It also demonstrates the ease with which a consistent scheme can be amended to a consistent conservative one.

4.1.2. *Analytic expression for m=2.* For two conserved quantities, the discrete multiplier matrix $\Lambda^\tau \in M_{2\times n}(\mathbb{R})$ is given by

$$\Lambda^\tau(t^k, \boldsymbol{x}^{k+1}, \boldsymbol{x}^k) := \begin{pmatrix} \dfrac{\Delta\psi_1}{\Delta\boldsymbol{x}}^\top(t^k, \boldsymbol{x}^{k+1}, \boldsymbol{x}^k) \\ \dfrac{\Delta\psi_2}{\Delta\boldsymbol{x}}^\top(t^k, \boldsymbol{x}^{k+1}, \boldsymbol{x}^k) \end{pmatrix},$$

where $\dfrac{\Delta\psi_i}{\Delta\boldsymbol{x}}$ again denotes the column vector of partial divided differences of ψ_i with respect to $\boldsymbol{x}$, similar to the $m = 1$ case. With $\Lambda^\tau(\Lambda^\tau)^\top$ now being a 2×2 matrix, the MN-DMM scheme of (3.1) for $m = 2$ takes the explicit form

$$\boldsymbol{f}^\tau_{\text{MN}} := \boldsymbol{f}^\tau - \frac{\begin{pmatrix}\frac{\Delta\psi_1}{\Delta\boldsymbol{x}} & \frac{\Delta\psi_2}{\Delta\boldsymbol{x}}\end{pmatrix}}{\det(\Lambda^\tau(\Lambda^\tau)^\top)}\begin{pmatrix}\left\|\frac{\Delta\psi_2}{\Delta\boldsymbol{x}}\right\|_2^2 & -\frac{\Delta\psi_1}{\Delta\boldsymbol{x}}^\top\frac{\Delta\psi_2}{\Delta\boldsymbol{x}} \\ -\frac{\Delta\psi_2}{\Delta\boldsymbol{x}}^\top\frac{\Delta\psi_1}{\Delta\boldsymbol{x}} & \left\|\frac{\Delta\psi_1}{\Delta\boldsymbol{x}}\right\|_2^2\end{pmatrix}\begin{pmatrix}\frac{\Delta\psi_1}{\Delta\boldsymbol{x}}^\top\boldsymbol{f}^\tau + \partial_t^\tau\psi_1 \\ \frac{\Delta\psi_2}{\Delta\boldsymbol{x}}^\top\boldsymbol{f}^\tau + \partial_t^\tau\psi_2\end{pmatrix}, \tag{4.4}$$

where $\det(\Lambda^\tau(\Lambda^\tau)^\top) = \left\|\frac{\Delta\psi_1}{\Delta\boldsymbol{x}}\right\|_2^2\left\|\frac{\Delta\psi_2}{\Delta\boldsymbol{x}}\right\|_2^2 - \left(\frac{\Delta\psi_2}{\Delta\boldsymbol{x}}^\top\frac{\Delta\psi_1}{\Delta\boldsymbol{x}}\right)^2$.

In principle, MN-DMM schemes for other small m values can also be written out analytically. However, for practical implementations involving $m > 2$, we instead refer to Section 4.2 to 4.4 for more general algorithms that implicitly construct conservative schemes without resorting to analytic computations.

4.2. **Direct MN-DMM Algorithm.** For an arbitrary number m of conserved quantities, we now present a fixed-point iteration algorithm associated with the scheme (3.6) introduced in Theorem 3.4, where consistency and conservative properties were shown. Before we compare different ways of computing the pseudoinverse expression involved in (3.6), let us describe how to solve the implicit scheme

$$\mathbf{0} = \boldsymbol{F}^\tau_{MN}(t^k, \boldsymbol{x}^{k+1}, \boldsymbol{x}^k) = D_t^\tau\boldsymbol{x} - \boldsymbol{f}^\tau_{MN}(t^k, \boldsymbol{x}^{k+1}, \boldsymbol{x}^k).$$

More explicitly, this is equivalent to the equations

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + (t^{k+1} - t^k)\boldsymbol{f}^\tau_{MN}(t^k, \boldsymbol{x}^{k+1}, \boldsymbol{x}^k). \tag{4.5}$$

We will employed a fixed-point iteration to solve for $\boldsymbol{x}^{k+1}$ in (4.5), which can also be viewed as a predictor-corrector method. For brevity and clarity, we shall consider an uniform time step[3] $\tau = t^{k+1} - t^k$ for all k, and denote the unknown vector as $\boldsymbol{x} := \boldsymbol{x}^{k+1}$ and the fixed vector as $\boldsymbol{y} := \boldsymbol{x}^k$. To bootstrap the fixed point iteration, we first compute an initial guess $\boldsymbol{x}^{(0)} = \phi(t^k, \boldsymbol{y})$ using any sufficiently accurate explicit time-stepping scheme $\phi : I \times U \to U$ depending on $\boldsymbol{f}$, t^k and $\boldsymbol{y}$, such as explicit Runge-Kutta schemes. From this initial guess or predictor, subsequent iterates $\boldsymbol{x}^{(i)}$ are then improved or corrected using the implicit MN-DMM scheme by iterating the fixed point iteration of (4.5) given by

$$\boldsymbol{x}^{(i)} := \boldsymbol{y} + \tau \boldsymbol{f}^{\tau}_{\scriptscriptstyle MN}(t^k, \boldsymbol{x}^{(i-1)}, \boldsymbol{y})$$

until a desired tolerance δ is reached. More explicitly, short-handing the notations

$$\begin{aligned} A(\boldsymbol{x}) &:= \Lambda^{\tau}(t^k, \boldsymbol{x}, \boldsymbol{y}), && \text{(Discrete multiplier matrix)} \\ \boldsymbol{s}(\boldsymbol{x}) &:= \boldsymbol{f}^{\tau}(t^k, \boldsymbol{x}, \boldsymbol{y}), && \text{(Discrete source term)} \\ \boldsymbol{r}(\boldsymbol{x}) &:= A(\boldsymbol{x})\boldsymbol{s}(\boldsymbol{x}) + \partial_t^{\tau} \boldsymbol{\psi}(t^k, \boldsymbol{x}, \boldsymbol{y}), && \text{(Residual of (2.7b))} \end{aligned}$$

and using the absolute error of the conserved quantities as the tolerance criteria, we arrive at the *Direct MN-DMM Algorithm*, or *MN-DMM Algorithm*:

Algorithm 1 Direct MN-DMM

```
$\boldsymbol{x}^{(0)} \leftarrow \phi(t^k, \boldsymbol{y})$
**repeat** $i = 1, 2, \dots$
    $\boldsymbol{x}^{(i)} \leftarrow \boldsymbol{y} + \tau \left( \boldsymbol{s}(\boldsymbol{x}^{(i-1)}) - A^{+}(\boldsymbol{x}^{(i-1)})\, \boldsymbol{r}(\boldsymbol{x}^{(i-1)}) \right)$
**until** $\left| \boldsymbol{\psi}(\boldsymbol{x}^{(i)}) - \boldsymbol{\psi}(\boldsymbol{x}^{0}) \right| < \delta$
**return** $\boldsymbol{x}^{(i)}$
```

A Banach fixed point argument shows that Algorithm 1 converges.

Theorem 4.1. *If for sufficiently small* τ*, the collection of functions* $\{\boldsymbol{s},\ \boldsymbol{A}^{+},\ \boldsymbol{r}\}_{\tau}$ *are locally Lipschitz continuous with Lipschitz constants independent of* τ*, then under the hypotheses of assumptions 2.2 and 2.4, there exists* $\tau_* > 0$ *such that Algorithm 1 converges whenever* $\tau < \tau_*$.

Proof. Denote $\boldsymbol{G}^{\tau}(\boldsymbol{z}) := \boldsymbol{y} + \tau\, \boldsymbol{F}^{\tau}(\boldsymbol{z})$ and $\boldsymbol{F}^{\tau}(\boldsymbol{z}) := \boldsymbol{s}(\boldsymbol{z}) - A^{+}(\boldsymbol{z})\, \boldsymbol{r}(\boldsymbol{z})$. Then the above algorithm is equivalent to the fixed point iteration

$$\boldsymbol{x}^{(i+1)} = \boldsymbol{G}^{\tau}(\boldsymbol{x}^{(i)}), \qquad \boldsymbol{x}^{(0)} := \phi(t^k, \boldsymbol{y}).$$

By continuity, it follows by Lemma 3.2 and consistency of $\boldsymbol{s}$ and $\boldsymbol{r}$ to their continuous counterpart that there exists $\tau_0 > 0$ and an open ball $\mathcal{B} \subset \mathbb{R}^n$ of radius $\epsilon > 0$ centered at $\boldsymbol{y}$ over which the restrictions of the discrete functions are Lipschitz continuous and $M := \sup_{\tau < \tau_0} \sup_{\boldsymbol{z} \in \mathcal{B}} \|\boldsymbol{F}^{\tau}(\boldsymbol{z})\| < \infty$. In particular, $\boldsymbol{G}^{\tau}(\mathcal{B}) \subset \mathcal{B}$ for $\tau < \min\{\tau_0, \epsilon/M\}$. In fact, the Lipschitz continuity hypothesis guarantees that

$$\|\boldsymbol{G}^{\tau}(\boldsymbol{z}_1) - \boldsymbol{G}^{\tau}(\boldsymbol{z}_2)\| = \tau \|\boldsymbol{F}^{\tau}(\boldsymbol{z}_1) - \boldsymbol{F}^{\tau}(\boldsymbol{z}_2)\| \leq \tau\, L \|\boldsymbol{z}_1 - \boldsymbol{z}_1\| \tag{4.6}$$

where $L > 0$ is the Lipschitz constant of $\boldsymbol{F}^{\tau}$ over $\mathcal{B}$. We conclude that $\boldsymbol{G}^{\tau} : \mathcal{B} \to \mathcal{B}$ is a contraction for $\tau < \tau_* := \min\{\tau_0, \epsilon/M, 1/L\}$, and thus Algorithm 1 converges by the Banach fixed point theorem. □

[3] Similar results can be derived with variable time steps by replacing τ with $\tau_k := t^{k+1} - t^k$ and ensuring $\tau := \sup_k (t^{k+1} - t^k)$ satisfies the contraction criteria in the fixed point iteration.

The main drawback of the Direct MN-DMM Algorithm is the need to compute analytically an inverse matrix within the pseudoinverse of $A(\boldsymbol{x})$. This can be alleviated by introducing auxiliary variables, as we discuss next.

4.3. **Mixed MN-DMM.** In order to solve (4.5) without having to invert $\Lambda^\tau \left(\Lambda^\tau\right)^\top$ explicitly for the computation of $\boldsymbol{f}^\tau_{MN}$, one option is to consider the mixed formulation of 4.5,

$$D_t^\tau \boldsymbol{x} + {\Lambda^\tau}^\top \boldsymbol{g} = \boldsymbol{f}^\tau, \tag{4.7a}$$

$$\Lambda^\tau {\Lambda^\tau}^\top \boldsymbol{g} = \Lambda^\tau \boldsymbol{f}^\tau + \partial_t^\tau \boldsymbol{\psi}, \tag{4.7b}$$

where the matrix inversion is replaced with solving the linear system (4.7b). Denoting $B(\boldsymbol{x}) := A(\boldsymbol{x})A(\boldsymbol{x})^\top$, equations (4.7a) and (4.7b) are equivalent to

$$\boldsymbol{x} = \boldsymbol{y} + \tau(\boldsymbol{s}(\boldsymbol{x}) - A(\boldsymbol{x})^\top \boldsymbol{g}), \tag{4.8a}$$

$$B(\boldsymbol{x})\boldsymbol{g} = \boldsymbol{r}(\boldsymbol{x}). \tag{4.8b}$$

To solve (4.8a) and (4.8b), we again propose a fixed point iteration type algorithm, which we referred to as the *Mixed MN-DMM Algorithm*:

Algorithm 2 Mixed MN-DMM

```
1: $\boldsymbol{x}^{(0)} \leftarrow \phi(t^k, \boldsymbol{y})$
2: **repeat** $i = 1, 2, \ldots$
3:     $\boldsymbol{g} \leftarrow \text{Solve}\left(B(\boldsymbol{x}^{(i-1)})\boldsymbol{g} = \boldsymbol{r}(\boldsymbol{x}^{(i-1)})\right)$
4:     $\boldsymbol{x}^{(i)} \leftarrow \boldsymbol{y} + \tau(\boldsymbol{s}(\boldsymbol{x}^{(i-1)}) - A(\boldsymbol{x}^{(i-1)})^\top \boldsymbol{g})$
5: **until** $\left|\boldsymbol{\psi}(\boldsymbol{x}^{(i)}) - \boldsymbol{\psi}(\boldsymbol{x}^{0})\right| < \delta$
6: **return** $\boldsymbol{x}^{(i)}$
```

Notice that for any accurate enough initial guess $\boldsymbol{x}^{(0)}$, standard arguments for perturbation of matrices that we have previously used in Lemma 3.2 guarantees that B will be invertible for sufficiently small τ. In other words, Algorithm 2 is iteratively solving normal equations of the second kind

$$B(\boldsymbol{x})\boldsymbol{g} = \boldsymbol{r}, \qquad\qquad A^\top \boldsymbol{g} = \boldsymbol{f}^\tau - \boldsymbol{f}^\tau_{\text{MN}},$$

associated with the *under*determined minimization problem (3.2), see for example [13, Eq. 1.1.20]. Moreover, in line 4 of Algorithm 2, we have the freedom to choose any state of the art linear solver for this type of equation. However, it is well-known that forming $B(\boldsymbol{x}) = A(\boldsymbol{x})A(\boldsymbol{x})^\top$ explicitly may lead to loss of accuracy and large condition numbers. Taking this possibility into account, we propose next using matrix decomposition techniques that are better suited to tackle such instances.

4.4. **Mixed MN-DMM using Singular Value Decomposition.** As discussed, the matrix B can be ill-conditioned in practice, and we will see this in some numerical examples of Section 5. Appealing to the Singular Value Decomposition (SVD) [15],

$$A = U\Sigma V^\top$$

can alleviated this issue[4], though at additional costs of computing such decomposition. Recall here that $U \in M_{m\times m}(\mathbb{R})$ and $V \in M_{n\times n}(\mathbb{R})$ are orthogonal

[4] Indeed, QR decomposition is another possibility as well.

matrices and the non-zero block of $\Sigma = \begin{pmatrix}\Sigma_m & 0_{m\times(n-m)}\end{pmatrix}$ is the diagonal matrix $\Sigma_m := \text{diag}(\sigma_1, ..., \sigma_m)$, where $\sigma_1 \geq ... \geq \sigma_m$ are the real eigenvalues of B. Thus, the multiplication of the right Moore-Penrose inverse on $\boldsymbol{f}^{\tau}_{MN}$ can be computed using

$$A^+ = V\Sigma^+U^\top, \qquad \Sigma^+ := \begin{pmatrix}\Sigma_m^{-1}\\ 0_{(n-m)\times m}\end{pmatrix},$$

which can be done in a sequential manner involving only matrix–vector products. We refer this approach as the *Mixed MN-DMM Algorithm using SVD*:

Algorithm 3 Mixed MN-DMM using SVD

```
$\boldsymbol{x}^{(0)} \leftarrow \phi(t^k, \boldsymbol{y})$
**repeat** $i = 1, 2, \ldots$
    $[U, \Sigma, V] \leftarrow \text{SVD}(A(\boldsymbol{x}^{(i-1)}))$
    $\boldsymbol{a} \leftarrow U^\top \boldsymbol{r}(\boldsymbol{x}^{(i-1)})$
    $\boldsymbol{b} \leftarrow \Sigma^+ \boldsymbol{a}$
    $\boldsymbol{x}^{(i)} \leftarrow \boldsymbol{y} + \tau\left(\boldsymbol{s}(\boldsymbol{x}^{(i-1)}) - V\boldsymbol{b}\right)$
**until** $\left|\boldsymbol{\psi}(\boldsymbol{x}^{(i)}) - \boldsymbol{\psi}(\boldsymbol{x}^{0})\right| < \delta$
**return** $\boldsymbol{x}^{(i)}$
```

The main advantage of this approach is that the product $B(\boldsymbol{x}) = A(\boldsymbol{x})A(\boldsymbol{x})^\top$ does not need to be assembled at each iteration, thus potentially improving the accuracy of the solution for poorly conditioned problems. However, the main drawback is that computing the SVD decomposition of $A(\boldsymbol{x}) \in M_{m\times n}(\mathbb{R})$ at each iteration requires additional costs. Nevertheless, Algorithm 3 can yield more accurate numerical solutions when A is poorly conditioned, which opens the possibility to future improvements along this direction.

5. Numerical results

With the theoretical results now established and practical implementation discussed, we now present several numerical examples to illustrate the MN-DMM approach and its two variants. The examples were chosen from a wide variety of physical problems, such as biological systems, chaotic systems, classical mechanics, fluid dynamics and geodesic flows. Moreover, they are roughly ordered at increasing difficulty in deriving analytic conservative schemes using the original DMM approach. In contrast, the MN-DMM approach only requires knowledge of the divided difference expressions within the discrete multiplier matrix to construct the conservative schemes, which can readily be systematized using modern computer algebra packages.

For the following examples, we have chosen to compare the MN-DMM method with two traditional methods, namely the standard 4th-order Runge-Kutta method and the 2nd-order symplectic Implicit Midpoint method. While these choices do not form an exhaustive comparison, they do highlight the large difference at preserving multiple conserved quantities across a wide variety of examples.

For the implicit schemes, such as Implicit Midpoint method and MN-DMM schemes, we have used the improved Euler's method to obtain an initial guess for the fixed point iteration employed to solve the nonlinear or implicitly defined equations. For the choice of $\boldsymbol{f}^{\tau}$ for the MN-DMM in these tests, the improved Euler method was also chosen. For the sake of reproducibility, we have listed within each example their relevant problem parameters, time step size τ, final time T, error tolerance for conserved quantities δ, error tolerance for residual of the nonlinear equations ϵ, and maximum number of fixed point iterations used per time step K. In all subsequent tables, we compare their maximum error in conserved quantities, as well as the mean fixed point iterations (FPIs) used for the implicit methods. Moreover, we compare the largest condition number of $\kappa(B)$ or $\kappa(A)$ encountered during simulation, for the Mixed MN-DMM and Mixed MN-DMM using SVD respectively. Note that for systems with a single conserved quantity, $\kappa(B) = 1 = \kappa(A)$, since $\Lambda^\tau(\Lambda^\tau)^\top$ is a scalar quantity as discussed in Section 4.1.1.

5.1. **Lotka-Volterra systems.** As a first simple example, we illustrate MN-DMM for the two and three species Lotka-Volterra system, with one and two conserved quantities respectively. We first recall their definitions and conserved quantities.

In [1, Example 5.2.1], analytic DMM schemes were derived for the two-species Lotka-Volterra system given by

$$\boldsymbol{F}(\boldsymbol{x}, \dot{\boldsymbol{x}}) := \begin{pmatrix} \dot{x} - x(a - by) \\ \dot{y} - y(dx - c) \end{pmatrix}, \tag{5.1}$$

for positive constants a, b, c, d. It is well-known that this system has a conserved quantity of the form

$$\psi(\boldsymbol{x}) := a \log y - by + c \log x - dx. \tag{5.2}$$

Using $\tau = 0.1, T = 10000, \delta = 1 \times 10^{-15}, \epsilon = 1 \times 10^{-15}, K = 20$ and initial conditions $\boldsymbol{x}^0 = (0.3, 0.7)^\top$ with $(a, b, c, d) = (1, 2, 3, 4)$, we obtain the results showed in Table 1, which confirms the machine precision accuracy of the MN-DMM at preserving the conserved quantity ψ of the two species Lotka-Volterra system. We also note that both the Implicit Midpoint method and MN-DMM methods utilized a similar number of fixed point iterations, with about 11 $\sim$ 12 mean FPIs.

Numerical Method	$\|\|\psi - \psi^0\|\|_\infty$	Mean FPIs	$\|\|\kappa(\cdot)\|\|_\infty$
RK4	1.279×10^{-1}	–	–
Implicit Midpoint	1.825×10^{-1}	12.069	–
MN-DMM	3.553×10^{-15}	11.649	–
Mixed MN-DMM	4.441×10^{-15}	11.678	1.000
Mixed MN-DMM (SVD)	3.553×10^{-15}	11.666	1.000

Table 1. Two-species Lotka-Volterra system with $\psi(x, y) = x - \log x + y - 2 \log y$.

Moreover, Figure 1 shows the trajectories of the Implicit Midpoint method and RK4 method drifting away from the level set of ψ. In contrast, the MN-DMM results show machine precision accuracy at remaining on the level set of ψ.

Extending to three-species, the Lotka-Volterra system takes the general form

$$\boldsymbol{F}(\boldsymbol{x}, \dot{\boldsymbol{x}}) := \begin{pmatrix} \dot{x} - x(a_{11}(x - \xi_1) + a_{12}(y - \xi_2) + a_{13}(z - \xi_3)) \\ \dot{y} - x(a_{21}(x - \xi_1) + a_{22}(y - \xi_2) + a_{23}(z - \xi_3)) \\ \dot{z} - x(a_{31}(x - \xi_1) + a_{32}(y - \xi_2) + a_{33}(z - \xi_3)) \end{pmatrix}, \tag{5.3}$$

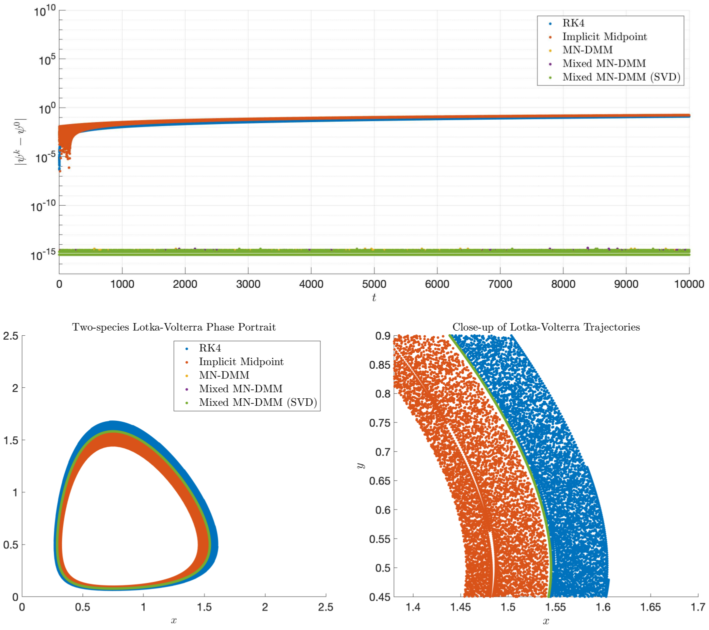


FIGURE 1. Comparison of error in $\psi(\boldsymbol{x})$ and trajectories for the two-species Lotka-Volterra problem.

where $A = [a_{ij}]$ is a real-valued interaction matrix and $\boldsymbol{\xi} = (\xi_1, \xi_2, \xi_3)^\top$ is a fixed point of the system. [16] showed that there are two conserved quantities

$$\boldsymbol{\psi}(\boldsymbol{x}) := \begin{pmatrix} d_1(x - \xi_1 \log x) + d_2(y - \xi_2 \log y) + d_3(z - \xi_3 \log z) \\ x^{\eta_1} y^{\eta_2} z^{\eta_3} \end{pmatrix}, \tag{5.4}$$

if the diagonal matrix $D := \mathrm{diag}(d_1, d_2, d_3)$ and vector $\boldsymbol{\eta} := (\eta_1, \eta_2, \eta_3)^\top$ satisfies

$$DA + A^\top D = 0, \qquad \boldsymbol{\eta}^\top A = \mathbf{0}. \tag{5.5a}$$

In [1, Example 5.2.2], analytic DMM schemes were derived for a special three-species system with a specific $A, D, \boldsymbol{\xi}, \boldsymbol{\eta}$. Here we compare results using MN-DMM for the following example satisfying (5.5a),

$$A = \begin{pmatrix} 0 & 3 & -2 \\ -3 & 0 & 1 \\ 2 & -1 & 0 \end{pmatrix}, \ \boldsymbol{\xi} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \ D = \mathrm{diag}(1,1,1), \ \boldsymbol{\eta} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

Using $\tau = 0.05, T = 30000, \delta = 1 \times 10^{-15}, \epsilon = 1 \times 10^{-15}, K = 20$ and initial conditions $\boldsymbol{x}^0 = (0.2, 0.5, 0.3)^\top$, we obtain the result listed in Table 2.

Similar to the two-species case, Table 2 shows machine precision accuracy at preserving the two conserved quantities $\boldsymbol{\psi}(\boldsymbol{x})$ for the MN-DMM results. While MN-DMM did required a mean FPI of ~ 12 over the Implicit Midpoint method 's mean FPI of ~ 9, the MN-DMM results are the only methods not exhibiting large deviation of the level sets of $\boldsymbol{\psi}(\boldsymbol{x})$ as shown in Figure 2, in contrast to the Implicit Midpoint and RK4 method. Moreover, as this example involves more than

Numerical Method	$\lVert\psi_1-\psi_1^0\rVert_\infty$	$\lVert\psi_2-\psi_2^0\rVert_\infty$	Mean FPIs	$\lVert\kappa(\cdot)\rVert_\infty$
RK4	3.893×10^{-2}	1.478×10^{-4}	–	–
Implicit Midpoint	3.701×10^{0}	1.350×10^{-3}	8.957	–
MN-DMM	3.553×10^{-15}	1.003×10^{-15}	12.205	–
Mixed MN-DMM	3.553×10^{-15}	1.003×10^{-15}	12.249	2.243×10^{6}
Mixed MN-DMM (SVD)	2.665×10^{-15}	1.003×10^{-15}	12.216	1.309×10^{3}

TABLE 2. Three-species Lotka-Volterra system with
$$\boldsymbol{\psi}(\boldsymbol{x}) = \begin{pmatrix} x-\log x+y-2\log y+z-3\log z \\ xy^2z^3 \end{pmatrix}.$$

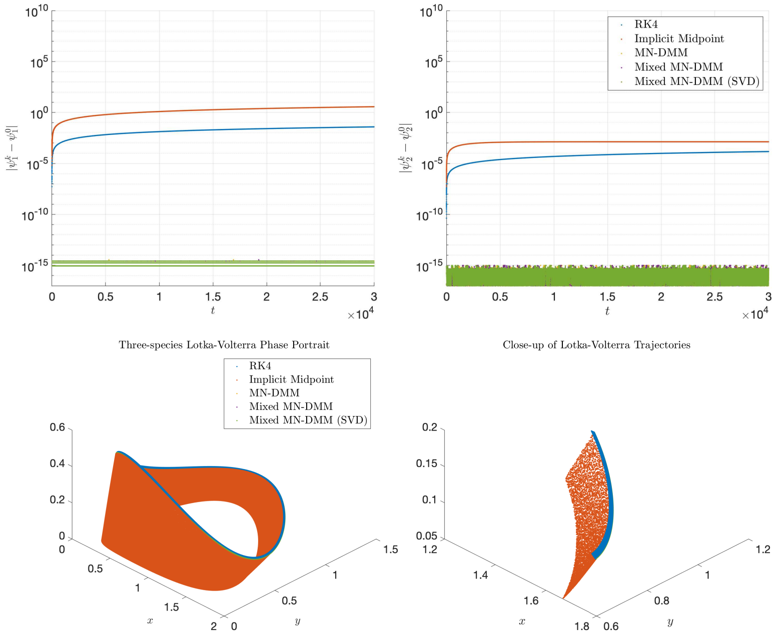


FIGURE 2. Comparison of error in $\boldsymbol{\psi}(\boldsymbol{x})$ and trajectories for the three-species Lotka-Volterra problem.

one conserved quantity, Table 2 now shows a smaller condition number for the associated linear system when SVD is used in the Mixed MN-DMM approach.

5.2. **Planar restricted three-body problem.**

In [1, Example 5.3], the planar restricted 3-body problem involving the Arenstorf orbit parameters was considered and an analytic DMM scheme was derived, albeit with much effort using divided difference calculus. Here we consider the same example but with much less effort to derive the conservative scheme using the MN-DMM approach.

For completeness, we first briefly recall the planar restricted three-body problem, which describes the gravitational motion of three bodies in a plane with a negligible mass, such as the Earth–Moon–Satellite system. The equations of motions are

$$\boldsymbol{F}(\boldsymbol{x},\dot{\boldsymbol{x}}) := \begin{pmatrix} \dot{x}_1 - y_1 \\ \dot{x}_2 - y_2 \\ \dot{y}_1 - \left(x_1 + 2y_2 - \dfrac{\alpha(x_1-\beta)}{((x_1-\beta)^2+x_2^2)^{\frac{3}{2}}} - \dfrac{\beta(x_1+\alpha)}{((x_1+\alpha)^2+x_2^2)^{\frac{3}{2}}} \right) \\ \dot{y}_2 - \left(x_2 - 2y_1 - \dfrac{\alpha x_2}{((x_1-\beta)^2+x_2^2)^{\frac{3}{2}}} - \dfrac{\beta x_2}{((x_1+\alpha)^2+x_2^2)^{\frac{3}{2}}} \right) \end{pmatrix}, \tag{5.6}$$

where $\boldsymbol{x} = (x_1, x_2, y_1, y_2)$ are the relative positions and momenta of the satellite to the center of mass between the Earth and Moon, with α, β being relative masses of the two bodies satisfying $\alpha + \beta = 1$. It is well-known that (5.6) has a conserved quantity called the Jacobi integral J given by,

$$J(\boldsymbol{x}) = \frac{x_1^2 + x_2^2 - y_1^2 - y_2^2}{2} + \frac{\alpha}{((x_1-\beta)^2+x_2^2)^{\frac{1}{2}}} + \frac{\beta}{((x_1+\alpha)^2+x_2^2)^{\frac{1}{2}}}.$$

We consider the Arenstorf orbit period $P = 17.0652165601579625588917206249$ and parameter $\alpha = 0.012277471$ were used with initial conditions,

$$\boldsymbol{x}^0 = (0.994, 0, 0, -2.00158510637908252240537862224)^\top.$$

Using the solver parameters $T = P \times 1.015, \tau = T \times 10^{-6}, \delta = 1 \times 10^{-15}, \epsilon = 1 \times 10^{-15}, K = 20$, we obtained the error in the Jacobi integral in Table 3.

Numerical Method	$\|J-J^0\|_\infty$	Mean FPIs	$\|\kappa(\cdot)\|_\infty$
RK4	5.793×10^{-8}	–	–
Implicit Midpoint	1.921×10^{2}	2.468	–
MN-DMM	6.639×10^{-14}	17.310	–
Mixed MN-DMM	6.639×10^{-14}	17.310	1.000
Mixed MN-DMM (SVD)	6.639×10^{-14}	17.310	1.000

Table 3. Planar restricted three-body problem with conserved quantity $J(\boldsymbol{x})$

As Figure 3 illustrates, all methods were able to reproduce the Arenstorf orbit qualitatively over one period P. However, shortly after one period, the Implicit Midpoint method results in a nonphysical trajectory, with several orders of magnitude jump in the error of the Jacobi integral due to the nonconvergence of its fixed point iterations. While the results from the MN-DMM approach do not show an exact periodic orbit, their trajectories beyond one period are close to that of the RK4 method, which is expected due to its higher order accuracy than the presented MN-DMM methods.

5.3. **Lorenz system.** In [1], analytic DMM scheme was also derived for time-dependent conserved quantities for dissipative systems, such as the damped harmonic oscillator. As another interesting example with time-dependent conserved quantities, we consider the Lorenz system for $\boldsymbol{x} = (x, y, z)$,

$$\boldsymbol{F}(\boldsymbol{x},\dot{\boldsymbol{x}}) := \begin{pmatrix} \dot{x} - \sigma(y-x) \\ \dot{y} - x(\rho - z) - y \\ \dot{z} - xy - \beta z \end{pmatrix}$$

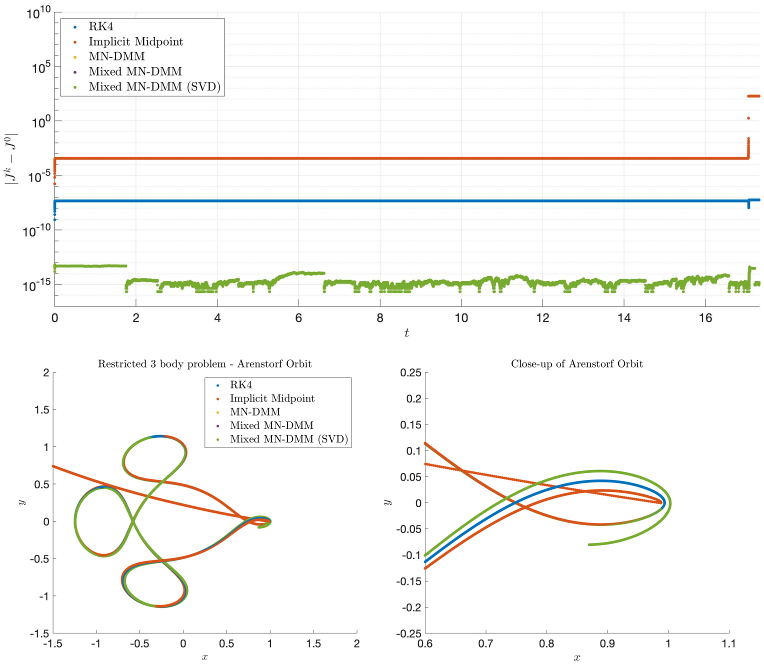


FIGURE 3. Comparison of error in Jacobi integral and trajectories for the Arenstorf orbit.

which has six conserved quantities over different sets of positive parameters σ, ρ, β in nonchaotic regime [17, 18]. Specifically, for the parameters $\sigma = 1/3, \rho = 400$ and $\beta = 0$, [18] showed that there exists a conserved quantity of the form,

$$\psi(t, \boldsymbol{x}) = \left(x^4 - \frac{4}{3}x^2 z - \frac{4}{9}y^2 - \frac{8}{9}xy + \frac{1600}{3}x^2 \right) e^{4t/3}.$$

Using $\tau = 0.001, T = 5, \delta = 1 \times 10^{-15}, \epsilon = 1 \times 10^{-15}, K = 20$ and initial conditions $\boldsymbol{x}^0 = (0.1, 0, 0)^\top$, we obtain the error in $\psi(t, \boldsymbol{x})$.

Numerical Method	$\left\|\psi-\psi^0\right\|_\infty$	Mean FPIs	$\left\|\kappa(\cdot)\right\|_\infty$
RK4	2.916×10^{-3}	–	–
Implicit Midpoint	7.971×10^{1}	18.601	–
MN-DMM	4.425×10^{-8}	19.990	–
Mixed MN-DMM	4.425×10^{-8}	19.990	1.000
Mixed MN-DMM (SVD)	4.425×10^{-8}	19.990	1.000

TABLE 4. Lorenz system with time-dependent conserved quantity $\psi(t, \boldsymbol{x}) = \left(x^4 - \frac{4}{3}x^2 z - \frac{4}{9}y^2 - \frac{8}{9}xy + \frac{1600}{3}x^2\right) e^{4t/3}$.

Table 4 indicates that machine precision accuracy for the time-dependent conserved quantity was not obtained using the MN-DMM approach. This is due to the stiffness of the problem as indicated by the high average number of FPIs for both

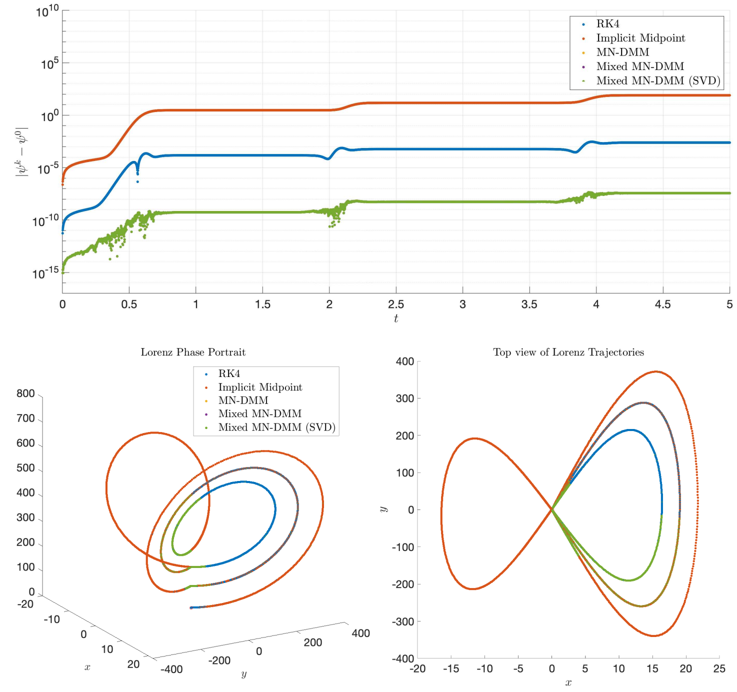


FIGURE 4. Comparison of error in $\psi(t, \boldsymbol{x})$ and trajectories for the Lorenz system.

the Implicit Midpoint method and the MN-DMM approach. Moreover, it can be observed in Figure 4 that fast transient dynamics occurs when the solution loops back toward the origin on the xy–plane, corresponding to the three apparent "jumps" in the error of ψ. Nevertheless, the Implicit Midpoint method has the largest error of $\sim 10^1$ in the conserved quantity, leading to an incorrect transient part of its trajectory located in the $x < 0$ region, as depicted in Figure 4. In contrast, the RK4 method and the MN-DMM approach have respective errors of $\sim 10^{-3}$ and $\sim 10^{-8}$ in the time-dependent conserved quantity ψ, with their trajectories remaining in the $x > 0$ region.

5.4. ***N*-point vortex problem on the unit sphere.** In [9, Example 4.5], analytic DMM scheme was derived for the classical N-point vortex problem on the unit sphere, which is an idealized model of approximating the solution to the incompressible Euler's equation on the unit sphere, given by

$$\boldsymbol{F}(\boldsymbol{x}, \dot{\boldsymbol{x}}) := \dot{\boldsymbol{x}}_i - \frac{1}{4\pi} \sum_{j=1, j\neq i}^{N} \Gamma_j \frac{\boldsymbol{x}_j \times \boldsymbol{x}_i}{1 - \boldsymbol{x}_i \cdot \boldsymbol{x}_j} = \boldsymbol{0}, \tag{5.7}$$

where $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ with $\boldsymbol{x}_i \in \mathbb{S}^2$ being the position of the i-th point vortex on the unit sphere and Γ_i being the vortex strength of the i-th vortex. The point vortex equations on the unit sphere (5.7) possess four conserved quantities, given by the momentum vector $\boldsymbol{P} \in \mathbb{R}^3$ and the Hamiltonian H, which are

$$\boldsymbol{P}(\boldsymbol{x}) := \sum_{i=1}^{N} \Gamma_i \boldsymbol{x}_i, \qquad H(\boldsymbol{x}) := -\frac{1}{4\pi} \sum_{1 \le i < j \le N} \Gamma_i \Gamma_j \log(1 - \boldsymbol{x}_i \cdot \boldsymbol{x}_j). \tag{5.8}$$

An analytic DMM scheme was derived in [9] with significant computation effort to verify the discrete multiplier conditions, in contrast to the MN-DMM approach. Using $N = 100$ randomly generated vortices and the solver parameters $\tau = 0.1, T = 200, \delta = 1 \times 10^{-15}, \epsilon = 1 \times 10^{-15}$ and $K = 20$, we obtain the error in four conserved quantities given in Table 5.

Numerical Method	$\lVert \boldsymbol{P} - \boldsymbol{P}^0 \rVert_\infty$	$\lVert H - H^0 \rVert_\infty$	Mean FPIs	$\lVert \kappa(\cdot) \rVert_\infty$
RK4	3.022×10^{-16}	1.360×10^{-6}	–	–
Implicit Midpoint	3.193×10^{-16}	1.240×10^{-7}	20.000	–
MN-DMM	2.705×10^{-16}	1.025×10^{-15}	4.670	–
Mixed MN-DMM	3.243×10^{-16}	1.022×10^{-15}	4.652	11.58
Mixed MN-DMM (SVD)	3.243×10^{-16}	1.022×10^{-15}	4.652	3.403

Table 5. Point vortices on the unit sphere with conserved quantities $\boldsymbol{P}(\boldsymbol{x})$ and $H(\boldsymbol{x})$.

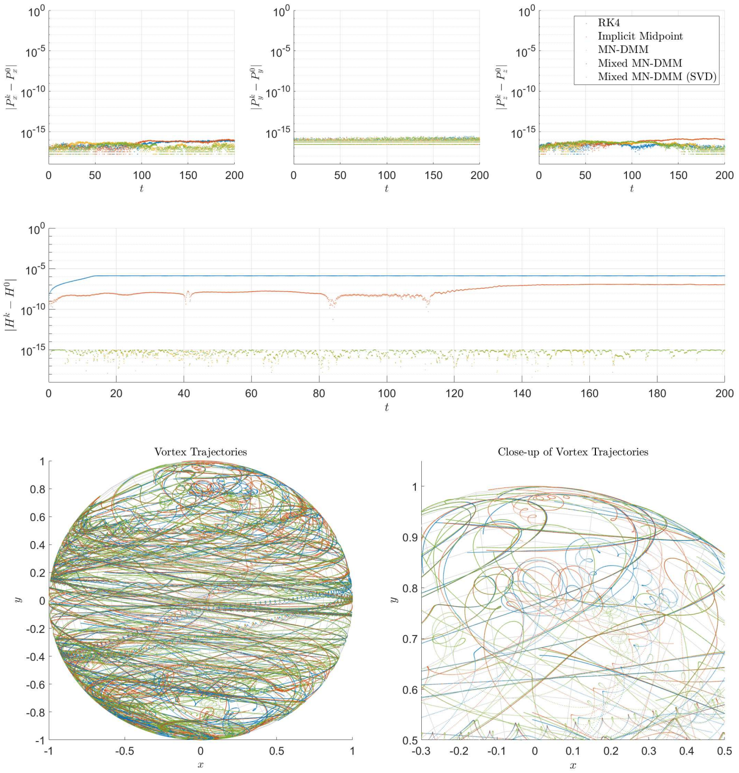


Figure 5. Comparison of error in conserved quantities and trajectories for the point vortex problem.

Table 5 indicates that this problem is relatively well-conditioned, with the MN-DMM approach converging faster than the Implicit Midpoint method. Moreover, as Figure 5 illustrates, all methods can preserves the momentum vector $\boldsymbol{P}$ up to machine precision. This is expected since the conserved quantities $\boldsymbol{P}$ are linear invariants, see [2]. On the other hand, only the MN-DMM approach is able to preserve the Hamiltonian H. While both the RK4 and Implicit Midpoint methods have error in Hamiltonian of $10^{-7} \sim 10^{-6}$, the observed trajectories are in stark contrast to the MN-DMM ones on a relatively short integration time of $T = 200$. This is consistent with the observations made in [9, Example 4.5] using the analytic DMM scheme for this problem. Thus, for larger number of vortices and longer term integration, large deviation in trajectories are likely to occur when the error in Hamiltonian is not close to machine precision.

5.5. **Geodesic curve on Schwarzschild Geometry.** For the final example, we apply the MN-DMM approach to solve for geodesic curves on an n-dimensional pseudo-Riemannian manifold. Specifically, we study geodesics for the Schwarzschild metric via the evolution of test particles in a spherically symmetric gravitational field. We refer to [19–21] for details on the following system. Recall that geodesic curves locally satisfy the first order system of ordinary differential equations

$$\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y},\dot{\boldsymbol{x}},\dot{\boldsymbol{y}}) := \begin{pmatrix} \left[\dot{x}^l - y^l\right]_{1\le l\le n} \\ \left[\dot{y}^l + \sum\limits_{j,k=1}^{n} \Gamma^l_{j,k}(\boldsymbol{x}) y^j y^k\right]_{1\le l\le n} \end{pmatrix} = \boldsymbol{0}, \tag{5.9}$$

where $\Gamma^i_{j,k}$ are Christoffel symbols of the second kind, cf. [19, Chap.3] or [21, Chap. 3]. A well-known conserved quantity is the speed [19, Chap. 5.4] given by

$$S(\boldsymbol{x},\boldsymbol{y}) = \sum_{i,j=1}^{n} g_{ij}(\boldsymbol{x}) y^i y^j \tag{5.10}$$

where $g_{ij}(\boldsymbol{x})$ denotes the Riemannian metric tensor [19, Sec. 3.8 and Appendix B]. As a concrete example, we consider the Schwarzschild metric, which is a radially symmetric solution to Einstein's equation in vacuum. In Schwarzschild coordinates $\boldsymbol{x} = (t, r, \theta, \phi)$ and $\boldsymbol{y} = (t', r', \theta', \phi')$, it is represented by the diagonal matrix

$$g(\boldsymbol{x}) = \operatorname{diag}\left(1 - \frac{r_s}{r}, -\left(1 - \frac{r_s}{r}\right)^{-1}, -r^2, -r^2 \sin^2\theta\right), \tag{5.11}$$

where $r_s = \frac{2GM}{c^2}$ is the Schwarzschild radius. In this setting, there are five conserved quantities. Indeed, using the spherical symmetries of this metric, it can be shown that the energy E and angular momentum $\boldsymbol{L}$ are conserved:

$$\begin{aligned} E(\boldsymbol{x},\boldsymbol{y}) &= \left(1 - \frac{r_s}{r}\right) t', \\ \boldsymbol{L}(\boldsymbol{x},\boldsymbol{y}) &= \begin{pmatrix} r^2 \sin(\theta)\phi' \\ r^2(\cos(\phi)\theta' - \cos(\theta)\sin(\phi)\phi') \\ r^2(\sin(\phi)\theta' - \cos(\theta)\cos(\phi)\phi') \end{pmatrix}. \end{aligned}$$

Moreover, the expression in (5.10) reduces to

$$S(\boldsymbol{x},\boldsymbol{y}) = \left(1 - \frac{r_s}{r}\right) t'^2 - \left(1 - \frac{r_s}{r}\right)^{-1} r'^2 - r^2\theta'^2 - r^2 \sin^2\theta\phi'^2.$$

Due to the complexity of these expressions, significant computation effort would be required to derive an analytic DMM scheme for (5.9) to preserve these five conserved quantities. In contrast, the MN-DMM approach requires relatively minimal effort to implement. We compare their numerical results using the solver parameters $\tau = 1/3, T = 200, \delta = 1 \times 10^{-15}, \epsilon = 1 \times 10^{-15}, K = 20$. We have set G, M, c to unity for simplicity, and used the initial conditions

$$\boldsymbol{x}^0 = \left(0, 37.338379348829989, \pi/2, 3.006861595479139\right)^\top,$$
$$\boldsymbol{y}^0 = \left(1, -0.990937492340824, 0, 0.003597472991852\right)^\top.$$

As Table 6 shows, the MN-DMM schemes are the only ones able to preserve all five conserved quantities up to machine precision. In contrast, the RK4 method was unstable at $\tau = 1/3$ and the Implicit Midpoint method had errors in the conserved quantities between $10^{-4} \sim 10^{-3}$. Due to the intricate short time dynamics of passing near the Schwarzschild radius, both the Implicit Midpoint method and MN-DMM required a similar number of fixed point iterations of $18 \sim 19$, with the maximum of 20. Also, the condition number for the Mixed MN-DMM using SVD approach is nearly seven orders of magnitude smaller than the Mixed MN-DMM.

Numerical Method	$\|S-S^0\|_\infty$	$\|E-E^0\|_\infty$	$\|\boldsymbol{L}-\boldsymbol{L}^0\|_\infty$	Mean FPIs	$\|\kappa(\cdot)\|_\infty$
RK4	NaN	NaN	NaN	–	–
Implicit Midpoint	2.590×10^{-4}	3.624×10^{-4}	4.590×10^{-3}	18.863	–
MN-DMM	7.896×10^{-15}	1.221×10^{-15}	1.579×10^{-14}	19.142	–
Mixed MN-DMM	4.816×10^{-15}	9.992×10^{-16}	8.464×10^{-15}	19.273	1.023×10^{13}
Mixed MN-DMM (SVD)	9.867×10^{-15}	1.332×10^{-15}	1.921×10^{-14}	19.347	5.062×10^{5}

Table 6. Geodesic curves on Schwarzschild Geometry with conserved quantities $S, E, \boldsymbol{L}$ ($\tau = 1/3$)

From Figure 6, we see that the Implicit Midpoint method and the different MN-DMM methods show out-going trajectories even at a relatively large time step of $\tau = 1/3$. The unstable RK4 results indicate that preserving the conserved quantities near the Schwarzschild radius is critical at predicting the correct long-term trajectories. Moreover, Figure 6 illustrates the geodesic curves of the Implicit Midpoint method predicts entirely wrong long term trajectory, while RK4 predicts a nonphysical outcome of ending inside the black hole.

To further study the differences between these methods at predicting the correct geodesic curves, we decrease their time step size and compare their long term trajectories and error in conserved quantities.

In Figure 7 and Figure 8 with $\tau = 1/3 \times 2^{-3}$, we see that both the Implicit Midpoint method and RK4 method does not preserve conserved quantities up to machine precision, with RK4 still predicting nonphysical results. At $\tau = 1/3 \times 2^{-5}$, both the Implicit Midpoint method and RK4 method now predict outgoing trajectories, albeit incorrect long term trajectories. Finally at $\tau = 1/3 \times 2^{-7}$, the Implicit Midpoint method still predicts incorrect long term trajectory. Meanwhile, the RK4 method is now able to mimic machine precision accuracy for the conserved quantities due to its higher order accuracy and much small τ. Thus, the long term trajectories of the RK4 method now agrees with the MN-DMM ones obtained using much larger τ. This final example highlights that conservative integration techniques, such as the MN-DMM, can be useful in intricate short-term dynamics, where machine precision level accuracy in conserved quantities can lead to more accurate long term predictions.

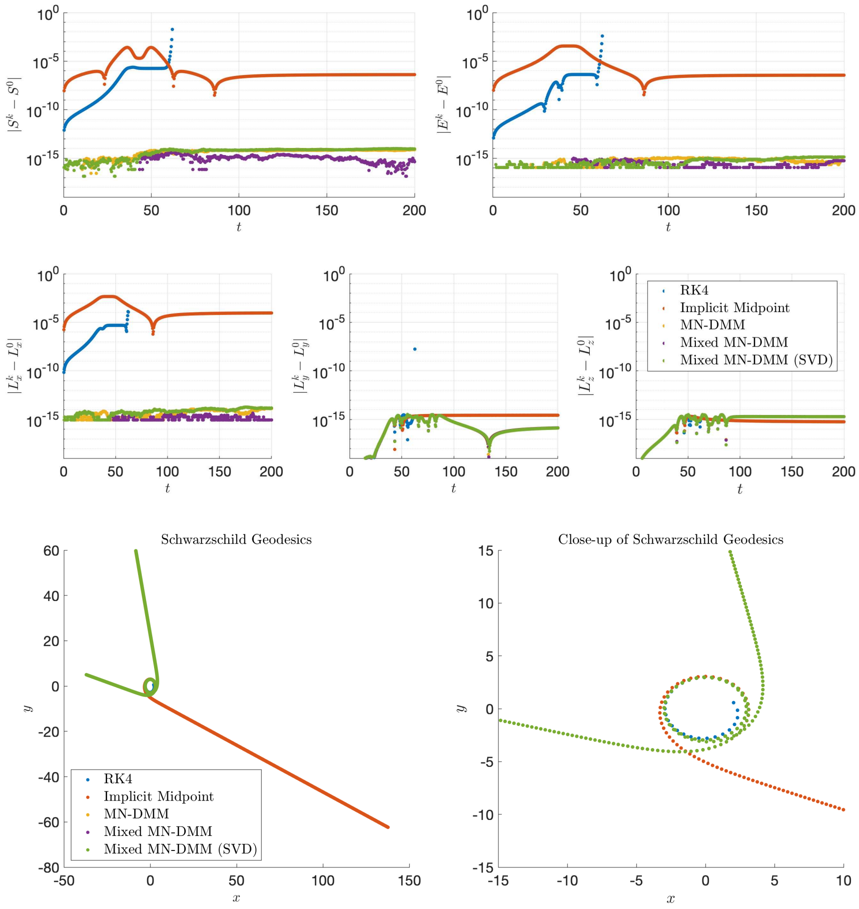


FIGURE 6. Comparison of error in conserved quantities and geodesics using $\tau = 1/3$.

6. Acknowledgement

The authors thank Jean-Christophe Nave and Gantumur Tsogtgerel for their valuable discussions in the initial stage of this work at McGill University. Andy T.S. Wan acknowledges funding support from the Discovery Grant program of the Natural Sciences and Engineering Research Council of Canada, (RGPIN-2019-07286) and Early Career Researcher Supplement (DGECR-2019-00467).

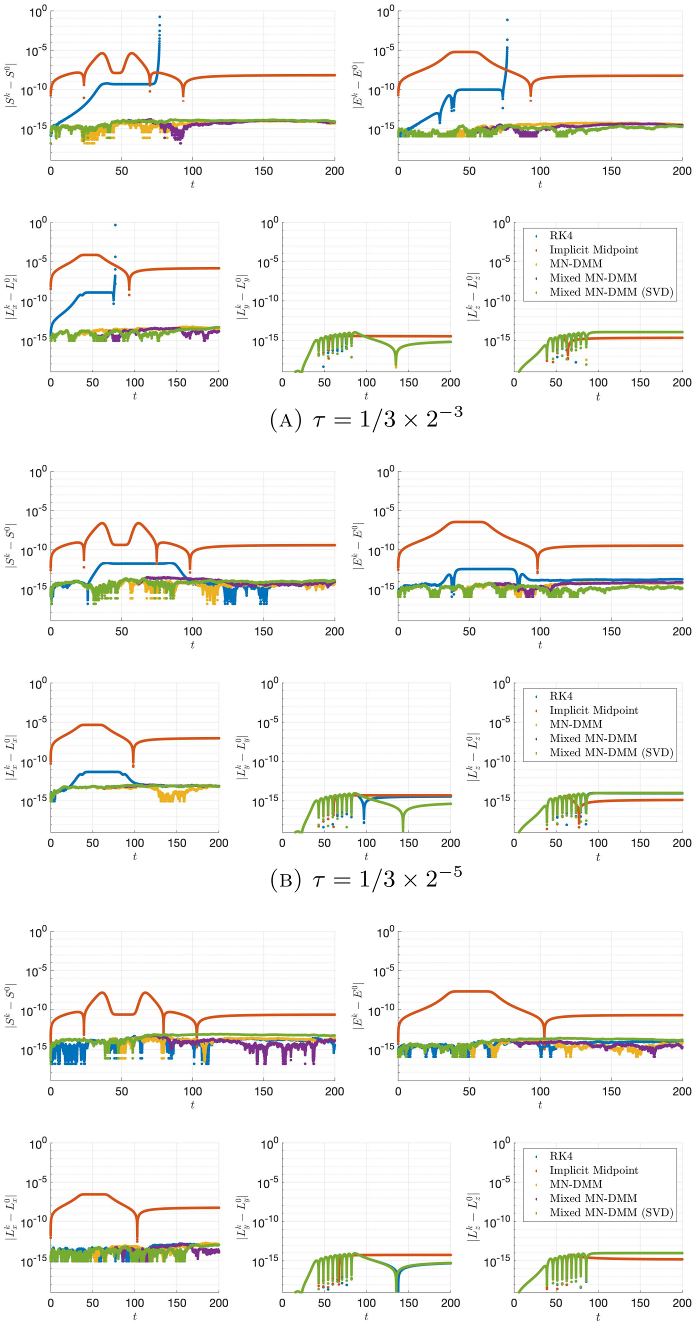
References

[1] Andy T. S. Wan, Alexander Bihlo, and Jean-Christophe Nave. Conservative methods for dynamical systems. *SIAM J. Numer. Anal.*, 55(5):2255–2285, 2017.

[2] E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*. Springer, Berlin, 2006.

[3] J. E. Marsden and M. West. Discrete Mechanics and Variational Integrators. *Acta Numerica*, 10:357–514, 2001.

[4] Arieh Iserles, Hans Z. Munthe-Kaas, Syvert P. Nørsett, and Antonella Zanna. Lie-group methods. *Acta Numerica*, 9:215–365, 2000.

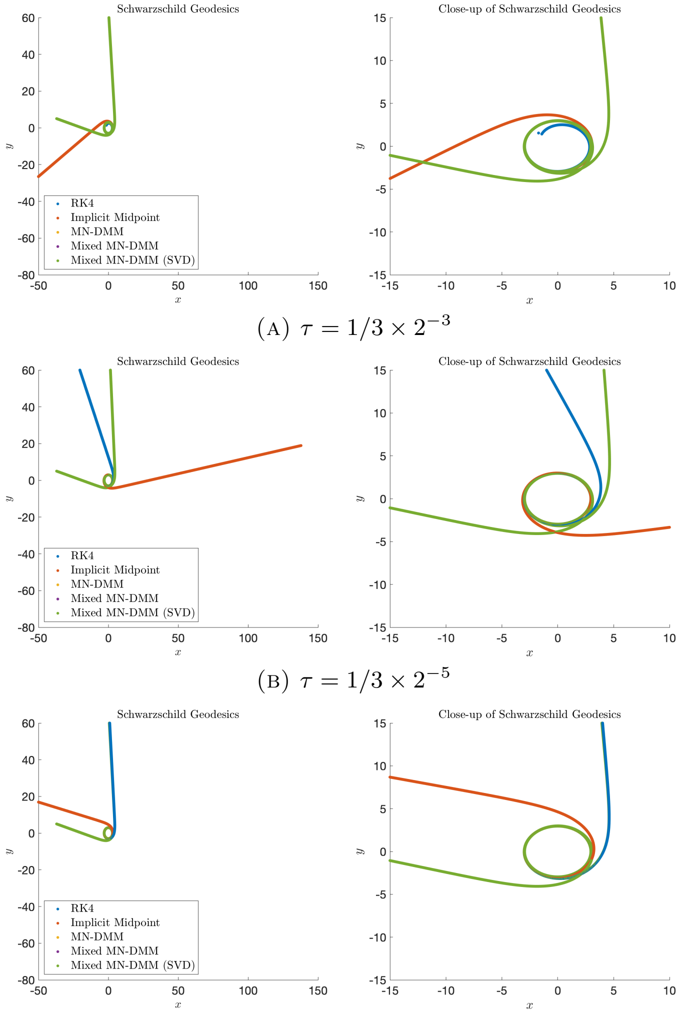


(A) $\tau = 1/3 \times 2^{-3}$

(B) $\tau = 1/3 \times 2^{-5}$

(C) $\tau = 1/3 \times 2^{-7}$

FIGURE 7. Comparison of error in conserved quantities for various τ.



(A) $\tau = 1/3 \times 2^{-3}$

(B) $\tau = 1/3 \times 2^{-5}$

(C) $\tau = 1/3 \times 2^{-7}$

FIGURE 8. Comparison of geodesics curves for various τ.

[5] M.P. Calvo and E. Hairer. Accurate long-term integration of dynamical systems. *Appl. Numer. Math.*, 18:95–105, 1995.

[6] Andy T. S. Wan and Jean-Christophe Nave. On the arbitrarily long-term stability of conservative methods. *SIAM J. Numer. Anal.*, 56(5):2751–2775, 2018.

[7] Feng Kang and Shang Zai-jiu. Volume-preserving algorithms for source-free dynamical systems. *Numerische Mathematik*, 71(4):451–463, 1995.

[8] G. R. W. Quispel R. I. McLachlan and N. Robidoux. Geometric integration using discrete gradients. *Phil. Trans. R. Soc. Lond.*, 357:1021–1045, 1999.

[9] Andy T. S. Wan, Alexander Bihlo, and Jean-Christophe Nave. Conservative integrators for many–body problems. *Journal of Computational Physics*, 466:111417, 2022.

[10] Cem Gormezano, Jean-Christophe Nave, and Andy T. S. Wan. Conservative Integrators for Vortex Blob Methods on the Plane. *Journal of Computational Physics*, 469:111357, 2022.

[11] Anil Hirani, Andy T. S. Wan, and Nikolas Wojtalewicz. Conservtive Integrators for Piecewise Smooth Dynamics with Transversal Dynamics. arxiv:2106.07484, 2021.

[12] Geoffrey McGregor and Andy T. S. Wan. Conservtive Hamiltonian Monte Carlo. arxiv:2206.06901, 2022.

[13] Åke Björck. *Numerical methods for least squares problems*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996.

[14] Alfio Quarteroni, Riccardo Sacco, and Fausto Saleri. *Numerical Mathematics*, volume 37 of *Texts in Applied Mathematics*. Springer-Verlag, Berlin, second edition, 2007.

[15] Lloyd N. Trefethen and David Bau. *Numerical Linear Algebra*. SIAM, 1997.

[16] R. Schimming. Conservation laws for Lotka–Volterra models. *Math. Methods Appl. Sci.*, 26(17):1517–1528, 2003.

[17] Mark J. Ablowitz and Harvey Segur. *Solitons and the inverse scattering transform*. SIAM: Studies in applied and numerical mathematics, 1981.

[18] M. Kus. Integrals of motion for the Lorenz system. *J. Phys. A: Math. Gen.*, 16(18):L689–L691, 1983.

[19] Sean M. Carroll. *Spacetime and Geometry: An Introduction to General Relativity*. Cambridge University Press, 2019.

[20] Leonor Godinho and José Natário. *An introduction to Riemannian geometry*. Universitext. Springer, Cham, 2014. With applications to mechanics and relativity.

[21] Barrett O'Neill. *Semi-Riemannian geometry with applications to relativity*. Academic press, 1983.

7. Appendix

[1]Seminar in Applied Mathematics, Swiss Federal Institute of Technology Zurich
Current address: Rämistrasse 101, CH-8092 Zurich, Switzerland
Email address: `erick.schulz@sam.math.ethz.ch`

[2]Department of Mathematics and Statistics, University of Northern British Columbia
Current address: 3333 University Way, Prince George, BC V2N 4Z9, Canada
Email address: `andy.wan@unbc.ca`