

---

# Stabilized Neural Differential Equations for Learning Constrained Dynamics

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

Many successful methods to learn dynamical systems from data have recently been introduced. However, assuring that the inferred dynamics preserve known constraints, such as conservation laws or restrictions on the allowed system states, remains challenging. We propose *stabilized neural differential equations* (SNDEs), a method to enforce arbitrary manifold constraints for neural differential equations. Our approach is based on a stabilization term that, when added to the original dynamics, renders the constraint manifold provably asymptotically stable. Due to its simplicity, our method is compatible with all common neural ordinary differential equation (NODE) models and broadly applicable. In extensive empirical evaluations, we demonstrate that SNDEs outperform existing methods while extending the scope of which types of constraints can be incorporated into NODE training.

## 1 Introduction

Advances in machine learning have recently spurred hopes of displacing or at least enhancing the process of scientific discovery by inferring natural laws directly from observational data. In particular, there has been a surge of interest in data-driven methods for learning dynamical laws in the form of differential equations directly from data [1, 2, 3, 4, 5, 6]. Assuming there is a ground truth system with dynamics governed by an ordinary differential equation

$$\frac{du(t)}{dt} = f(u(t), t) \quad (\text{with initial condition } u(0) = u_0) \quad (1)$$

with  $u(t) \in \mathbb{R}^n$  and  $f : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n$ , the question is whether we can learn  $f$  from (potentially noisy and irregularly sampled) observations  $(t_i, u(t_i))_{i=1}^N$ .

Neural ordinary differential equations (NODEs) provide a prominent and successful method for this task, which leverages machine learning by directly parameterizing the vector field  $f$  of the ODE as a neural network [1] (see also Kidger [7] for an overview). A related approach is called universal differential equations (UDEs) [2] and combines mechanistic or process-based model components with universal function approximators, typically also neural networks. In this paper, we will refer to these methods collectively as *neural differential equations* (NDEs), meaning any ordinary differential equation model in explicit form, where the right-hand side is either partially or entirely parameterized by a neural network. Due to the use of flexible neural networks, NDEs have certain universal approximation properties [8, 9], which are often interpreted as “in principle an NDE can learn any vector field  $f$ ” [10]. While this can be a desirable property in terms of applicability, in typical settings one often has prior knowledge about the dynamical system that should be incorporated.

Like in other areas of machine learning – particularly deep learning – inductive biases can substantially aid generalization, learning speed and stability, as well as successful training in the low data regime. Learning dynamics from data is no exception [11]. In scientific applications, physical priors are often

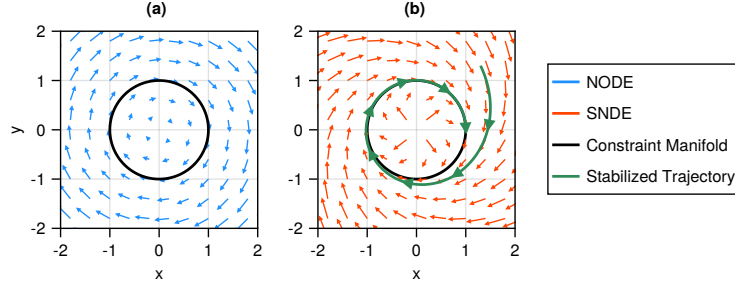


Figure 1: Sketch of the basic idea of stabilized neural differential equations, showing a simple example of a constraint manifold  $\mathcal{M}$  (black circle), an unstabilized vector field ((a), blue arrows) and the corresponding stabilized vector field ((b), red arrows). The stabilization pushes any trajectory starting away from (but near) the manifold ((b), green line) to converge to it at a rate  $\gamma$  (see Section 3).

not only a natural source for inductive biases, but can even impose hard constraints on the allowed dynamics. For instance, when observing mechanical systems, a popular approach is to directly parameterize either the Lagrangian or Hamiltonian via a neural network [12, 13, 14]. Constraints such as energy conservation can then be “baked into the model”, in the sense that the parameterization of the vector field is designed to only represent functions that satisfy the constraints. Finzi et al. [15] build upon these works and demonstrate how to impose explicit constraints in second-order ODEs.

In this work, we propose a stabilization technique to enforce arbitrary, even time-dependent manifold constraints for any class of NDEs, not limited to second-order systems and not requiring observations in particular (canonical) coordinates. It is compatible with all common explicit differential equation solvers as well as adjoint sensitivity methods. All code is publicly available at [anonymized].

## 2 Background and Related Work

A first order neural differential equation is typically given as

$$\frac{du(t)}{dt} = f_{\theta}(u, t) \quad u(0) = u_0, \quad (2)$$

where  $u : \mathbb{R} \rightarrow \mathbb{R}^n$  and the vector field  $f_{\theta} : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n$  is at least partially parametrized by a neural network with parameters  $\theta \in \mathbb{R}^d$ . We focus our attention on ground truth dynamics  $f$  in Equation (1) that are continuous in  $t$  and Lipschitz continuous in  $u$  such that the existence of a unique (local) solution to the initial value problem is guaranteed by the Picard-Lindelöf theorem. As universal function approximators [16, 17], neural networks  $f_{\theta}$  can in principle approximate any such  $f$  to arbitrary precision, i.e., the problem of learning  $f$  is realizable.

In practice, the parameters  $\theta$  are optimized via stochastic gradient descent by integrating a trajectory  $\hat{u}(t) = \text{ODESolve}(u_0, f_{\theta}, t)$  and taking gradients of the loss  $\mathcal{L}(u, \hat{u})$ . Computing these gradients with respect to  $\theta$  can be achieved using adjoint sensitivity analysis (*optimize-then-discretize*) or automatic differentiation of the solver operations (*discretize-then-optimize*) [1, 7]. While these have different (dis)advantages [18, 7], we use the adjoint sensitivity method for all experiments due to reportedly improved stability [1]. We also use the standard squared loss  $\mathcal{L}(u, \hat{u}) = \|u - \hat{u}\|_2^2$ .

Our stabilization approach is also related to the practice of index reduction in the context of differential algebraic equations (DAEs). We refer the interested reader to Appendix A for a brief overview of these connections.

**Related work.** We focus on NODEs as they can handle arbitrary non-linear vector fields  $f$  and outperform traditional ODE parameter estimation techniques. In particular, they do not require a pre-specified parameterization of  $f$  in terms of a small set of semantically meaningful parameters. The original NODE model [1] has quickly been extended to augmented variants that work for second order systems [10, 19], irregularly-sampled observations [20], Bayesian NODEs [21], partial differential equations [22] and more—see, e.g., [7] for an overview. All such variants can in principle be stabilized via our approach. We focus primarily on the standard NODE approach to demonstrate the impact

of stabilization rather than comparing different NODE methods. In our empirical evaluation, we therefore use vanilla NODE [1] or second-order NODE (SONODE) [19] implementations.

Originally, NODEs were introduced as the infinite depth limit of residual neural networks (typically for classification), where there is no single true underlying dynamic law, but “some” vector field  $f$  is learned that allows subsequent (linear) separation of the inputs into different classes. A number of techniques have been introduced to restrict the number of function evaluations needed during training to improve efficiency, which typically also results in relatively simple learned dynamics [23, 24, 25, 26, 27]. These are orthogonal to our method and are mentioned here only for completeness, as they could also be viewed as “regularized” or “stabilized” NODEs from a different perspective.

A large body of related work has focused on Hamiltonian or Lagrangian dynamics with conserved, time-independent first integrals, such as energy conservation, as constraints on the dynamics. Greydanus et al. [12] assume second-order Hamiltonian dynamics where  $u(t) = (q(t), p(t))$  consists of canonical coordinates and proposed to directly parameterize and learn the Hamiltonian  $\mathcal{H}$  (instead of  $f$ ) from which the (autonomous) vector field can then be derived via  $f(q, p) = (\frac{\partial \mathcal{H}}{\partial p}, -\frac{\partial \mathcal{H}}{\partial q})$ . This approach has been extended to also work on certain more general coordinates (e.g., angles) or when only velocities are observed instead of momenta [28], to be agnostic to the coordinate system altogether by modeling the underlying coordinate-free symplectic two-form directly [29], studied extensively with respect to the importance of symplectic integrators [30], and adapted specifically to robotic systems measured in terms of their SE(3) pose and generalized velocity [31]. Recently, Gruver et al. [32] have shown that what makes Hamiltonian neural nets work in practice is not so much the built in energy conservation or symplectic structure, but rather the fact that they inherently assume that the system is governed by a single second-order DE. Chen et al. [33] provides a recent overview of learning Hamiltonian dynamics using neural architectures. A related line of work instead assumes second-order Lagrangian dynamics and parameterizes the inertia matrix and divergence of the potential [13] or any generic Lagrangian function [14], which again uniquely determine the dynamics  $f$  via the Euler-Lagrange equations.

Instead of adapting the neural net architecture to satisfy certain properties by design, Lim and Kasim [34] take a different approach in which they still learn  $f$  directly and manually craft different types of regularization terms added to the loss that aim at enforcing different constraints or conservation laws. While similar to our approach in that no special architecture is required for different constraints, the key difference is that their approach requires crafting specific loss terms for different types of dynamics. Moreover, tuning the regularization parameter can be rather difficult. Lou et al. [35] develop “Manifold ODE”, a method that directly adjusts the forward mode integration and backward mode adjoint gradient computation to ensure that the trajectory is confined to a given manifold. Their method not only requires an entirely new training procedure, but also relies on an explicit chart representation of the manifold, which can be cumbersome to define in practice.

The work most closely related to ours is by Finzi et al. [15]. The present work differs in a number of ways with the key advances of our approach being that SNDEs (a) are applicable to any type of ODE, allowing us to go beyond second-order systems with primarily Hamiltonian or Lagrangian type constraints, (b) are compatible with hybrid models, i.e., the UDE approach where part of the dynamics is assumed to be known and only the remaining unknown part is learned while still constraining the overall dynamics, and (c) can incorporate any type of manifold constraints. Regarding (c), we can for instance also enforce time-dependent first integrals, which do not correspond to constants of motion or conserved quantities arising directly from symmetries in the Lagrangian.

In this work, we vastly broaden the scope of learning constrained dynamics by demonstrating the effectiveness of our approach on both first- and second-order systems including chaotic and non-chaotic as well as autonomous and non-autonomous examples. We cover constraints arising from holonomic restrictions on system states, conservation laws, and constraints imposed by controls.

### 3 Stabilized Neural Differential Equations

**General approach.** Given  $m < n$  explicit constraints, we require that solution trajectories of the NDE in Equation (2) are confined to an  $(n - m)$ -dimensional submanifold of  $\mathbb{R}^n$  defined by

$$\mathcal{M} = \{u \in \mathbb{R}^n; g(u) = 0\}, \quad (3)$$

where  $g: \mathbb{R}^n \rightarrow \mathbb{R}^m$  is a smooth function with  $0 \in \mathbb{R}^m$  being a regular value of  $g$ .<sup>1</sup> In other words, we have an NDE on a manifold

$$\dot{u} = f_\theta(u, t) \quad \text{with} \quad g(u) = 0. \quad (4)$$

Any non-autonomous system can equivalently be represented by an autonomous system by adding time as an additional coordinate with constant derivative 1 and initial condition  $t_0 = 0$ . Without loss of generality, from now on, we will consider only autonomous systems for ease of notation. We highlight again that our method applies equally to non-autonomous systems.

While there are methods that aim at constraining neural network outputs to lie on a pre-specified manifold, the added difficulty in our setting is that we learn the vector field  $f$ , but constrain the solution trajectory  $u$  that solves a given initial value problem for the ODE defined by  $f$ . Inspired by Chin [36], we propose the following stabilization of the vector field Equation (4)

$$\dot{u} = f_\theta(u) - \gamma F(u)g(u), \quad [\text{general SNDE}] \quad (5)$$

where  $\gamma \geq 0$  is a scalar parameter of our method and  $F: \mathbb{R}^n \rightarrow \mathbb{R}^{n \times m}$  is a so-called *stabilization matrix*. We call Equation (5) a *stabilized neural differential equation* (SNDE) and say that it is stabilized with respect to the invariant manifold  $\mathcal{M}$ . We illustrate the main idea in Figure 1. While even small deviations from the solenoidal (divergence free) vector field can lead to (potentially accumulating) constraint violations (left), our stabilization adjusts the vector field near the invariant manifold to render it asymptotically stable while leaving the vector field on  $\mathcal{M}$  unaffected (right).

**Theoretical guarantees.** First, note that ultimately we still want  $f_\theta$  to approximate the assumed ground truth dynamics  $f$ . However, Equation (5) explicitly modifies the right hand side of the NDE. The following theorem provides necessary and sufficient conditions under which  $f_\theta$  can still learn the correct dynamics when using a different right hand side.

**Theorem 1** (adapted from Chin [36]). *Consider an NDE*

$$\dot{u} = f_\theta(u) \quad (6)$$

*on an invariant manifold  $\mathcal{M} = \{u \in \mathbb{R}^n; g(u) = 0\}$ . A vector field  $\dot{u} = h_\theta(u)$  admits all solutions of Equation (6) on  $\mathcal{M}$  if and only if  $h_\theta|_{\mathcal{M}} = f_\theta|_{\mathcal{M}}$ .*

Since  $g(u) = 0$  on  $\mathcal{M}$ , the second term on the right-hand side of Equation (5) vanishes on  $\mathcal{M}$ . Therefore the SNDE Equation (5) admits all solutions of the constrained NDE Equation (4). Next, we will show that under mild conditions, the additional stabilization term in Equation (5) “nudges” the solution trajectory to lie on the constraint manifold in the sense that  $\mathcal{M}$  is asymptotically stable.

**Theorem 2** (adapted from Chin [36]). *Suppose the stabilization matrix  $F(u)$  is chosen such that the matrix  $G(u)F(u)$ , where  $G(u) = g_u$  is the Jacobian of  $g$  at  $u$ , is symmetric positive definite with the smallest eigenvalue  $\lambda(u)$  satisfying  $\lambda(u) > \lambda_0 > 0$  for all  $u$ . Assume further that there is a positive number  $\gamma_0$  such that*

$$\|G(u)f_\theta(u)\| \leq \gamma_0 \|g(u)\| \quad (7)$$

*for all  $u$  near  $\mathcal{M}$ . Then the invariant manifold  $\mathcal{M}$  is asymptotically stable in the SNDE Equation (5) if  $\gamma \geq \gamma_0/\lambda_0$ .*

*Proof.* Consider the Lyapunov function  $V(u) = \frac{1}{2}g^T(u)g(u)$ . Then (omitting arguments)

$$\frac{d}{dt}V(t) = \frac{1}{2} \frac{d}{dt} \|g(u(t))\|^2 = g^T \frac{dg}{dt} = g^T \frac{dg}{du} \dot{u} = g^T G(f_\theta - \gamma Fg), \quad (8)$$

where we substitute Equation (5) for  $\dot{u}$ . With Equation (7), we have  $g^T G f_\theta \leq \gamma_0 g^T g$  and since the eigenvalues of  $GF$  are assumed to be at least  $\lambda_0 > 0$  we have  $g^T GFg \geq \lambda_0 g^T g$ . Hence

$$\frac{d}{dt}V \leq (\gamma_0 - \gamma\lambda_0) \|g\|^2, \quad (9)$$

so the manifold  $\mathcal{M}$  is asymptotically stable whenever  $\gamma_0 - \gamma\lambda_0 \leq 0$ .<sup>2</sup>  $\square$

<sup>1</sup>The preimage theorem ensures that  $\mathcal{M}$  is indeed an  $n - m$ -dimensional submanifold of  $\mathcal{M}$ .

<sup>2</sup>We note that there is a minor error in the proof of Theorem 2 in Chin [36], which we corrected here.

156 When  $f_\theta(u)$  and  $g(u)$  are given,  $\mathcal{M}$  is asymptotically stable in the SNDE Equation (5) as long as

$$\gamma \geq \frac{\|G(u)f_\theta(u)\|}{\lambda_0\|g(u)\|}. \quad (10)$$

157 To summarize, the general form of the SNDE Equation (5) has the following important properties.

- 158 1. The SNDE admits all solutions of the constrained NDE Equation (4) on  $\mathcal{M}$ .
- 159 2.  $\mathcal{M}$  is asymptotically stable in the SNDE for sufficiently large values of  $\gamma$ .

160 The stabilization hyperparameter  $\gamma$ , with units of inverse time, determines the rate of relaxation to  
 161 the invariant manifold. In the limit  $\gamma \rightarrow \infty$ , the SNDE Equation (5) is equivalent to a Hessenberg  
 162 index-2 DAE, see Appendix A for more details.

163 **Practical implementation.** This leaves us with finding a concrete instantiation of the stabilization  
 164 matrix  $F(u)$  that should (a) satisfy that  $F(u)G(u)$  is symmetric positive definite with the small-  
 165 est eigenvalue bounded away from zero near  $\mathcal{M}$ , (b) efficiently computable, (c) compatible with  
 166 gradient-based optimization of  $\theta$  as part of an NDE. In our experiments, we use the Moore-Penrose  
 167 pseudoinverse of the Jacobian of  $g$  at  $u$  as the stabilization matrix

$$F(u) = G^+(u) = G(u)^T (G(u)G(u)^T)^{-1} \in \mathbb{R}^{n \times m}. \quad (11)$$

168 Let us analyze the properties of this choice. Regarding the requirements (b) and (c), the pseudoinverse  
 169 can be computed efficiently via a singular value decomposition with highly optimized implementations  
 170 in all common numerical linear algebra libraries (including deep learning frameworks) and does not  
 171 interfere with gradient-based optimization. In particular, the computational cost for the pseudoinverse  
 172 is  $\mathcal{O}(m^2n)$ , i.e., it scales well with the problem size. The quadratic scaling in the number of constraints  
 173 is often tolerable in practice, where the constraint manifold is typically low-dimensional. Moreover,  
 174 the Jacobian  $G(u)$  of  $g$  can be obtained via auto-differentiation in the respective frameworks.

175 Regarding requirement (a), the pseudoinverse  $G(u)^+$  is an orthogonal projection onto the tangent  
 176 space  $T_u\mathcal{M}$  of the manifold at  $u$ . Hence, locally in a neighborhood of  $u \in \mathcal{M}$ , we consider the  
 177 stabilization matrix as a projection back onto the invariant manifold  $\mathcal{M}$ , see Figure 1. In particular,  
 178  $G(u)$  has full rank for  $u \in \mathcal{M}$  and  $G^+G = G^T(GG^T)^{-1}G$  is symmetric and positive definite near  
 179  $\mathcal{M}$ . From here on, we thus consider the following specific form for the SNDE Equation (5),

$$\dot{u} = f_\theta(u) - \gamma G^+(u)g(u). \quad [\text{practical SNDE}] \quad (12)$$

180 The only parameter of SNDE is  $\gamma$ , which intuitively determines the “strength of nudging the trajectory  
 181 back towards  $\mathcal{M}$ ”. Here,  $\gamma$  is neither a Lagrangian parameter (corresponding to a constraint on  $\theta$ ), nor  
 182 a regularization parameter (to overcome an ill-posedness by regularization). Therefore, there is no  
 183 “correct” value for  $\gamma$ . In particular, Theorem 1 holds for all  $\gamma$ , which implies that we can also include  
 184 or remove the stabilization term at any time during training. For example, it may be beneficial to start  
 185 training without stabilization until  $f_\theta$  is close to  $f$  and then switch on stabilization. Theorem 2 only  
 186 requires  $\gamma$  to be “sufficiently large”.

## 187 4 Results

188 We now demonstrate the effectiveness of SNDEs on examples that cover autonomous first- and  
 189 second-order systems with a conserved first integral of motion or holonomic constraints, a non-  
 190 autonomous first-order system with a conserved quantity, a non-autonomous controlled first-order  
 191 system with a time-dependent constraint stemming from the control, and a chaotic second-order  
 192 system with a conservation law. As a metric for predicted state  $\hat{u}(t)$  versus ground truth  $u(t)$ , we use  
 193 the relative error  $\|u(t) - \hat{u}(t)\|_2 / \|u(t)\|_2$ , and analogous relative errors for the constraints  $g(u)$ .

194 We further demonstrate empirically that SNDEs are insensitive to the specific choice of  $\gamma$  over a large  
 195 range (beyond a minimum value) and provide more intuition about choosing  $\gamma$  and the computational  
 196 implications in Appendix B. Hence, SNDEs are easy to use in practice across a wide variety of settings  
 197 with minimal to no tuning. Appendix C provides runtime comparisons between SNDEs and vanilla  
 198 NODEs showing that SNDEs only incur moderate overhead in terms of computational requirements  
 199 compared to vanilla NODEs. Finally, Appendix D provides results on additional experiments due to  
 200 space constraints.

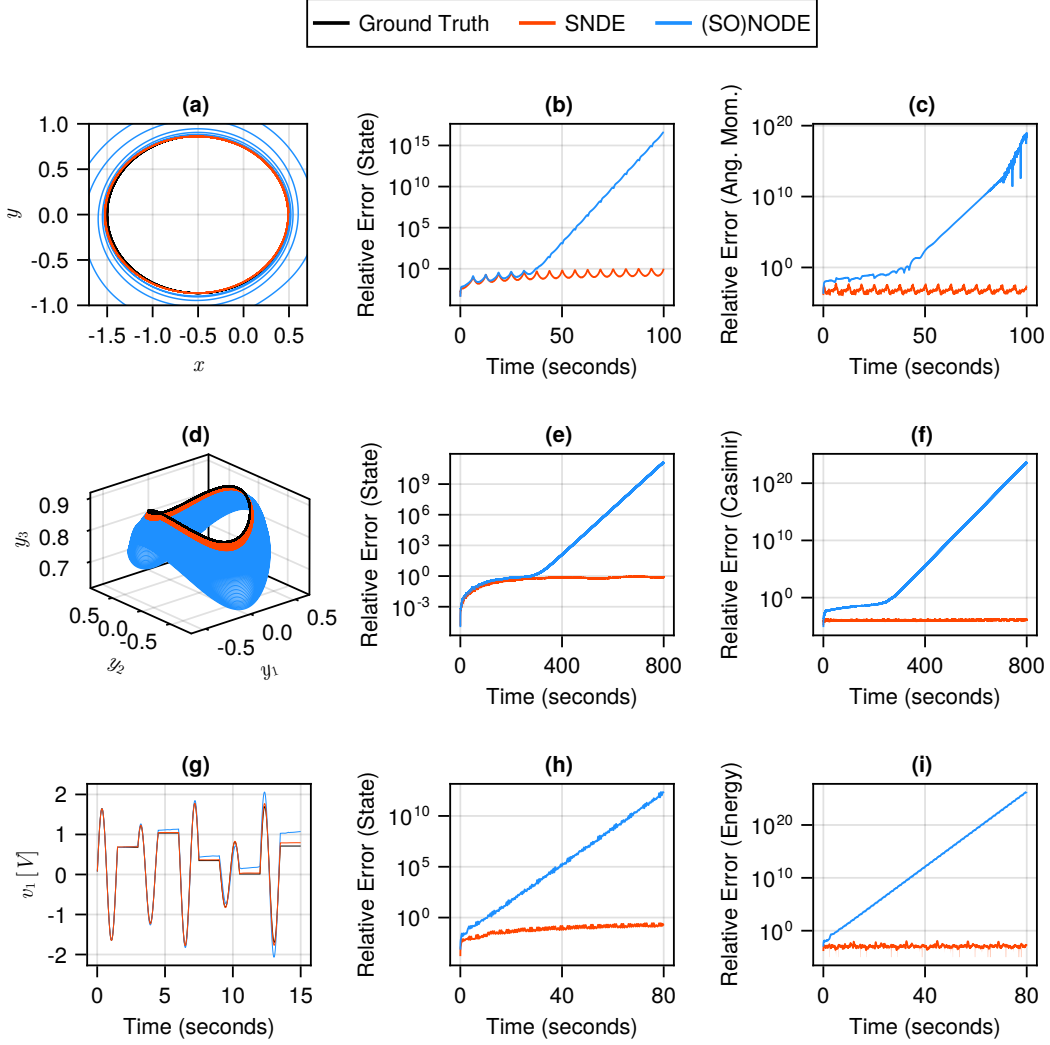


Figure 2: **Top row:** Results for the two-body problem (a-c). **Middle row:** Results for the rigid body rotation (d-f). **Bottom row:** Results for the DC-to-DC converter (g-i), where we show the voltage  $v_1$  across the first capacitor during a single test trajectory in (g), highlighting that vanilla NODE (blue) accumulates errors with each application of the switch. In all settings, the vanilla NODE (blue) quickly drifts from the manifold and subsequently diverges exponentially in relative error, while the SNDE (red) is confined to the manifold with accurate predictions over a long horizon.

#### 201 4.1 Two-Body Problem [second-order, autonomous, non-chaotic, conservation law]

202 The motion of two (pointlike) bodies attracting each other with a force inversely proportional to their  
 203 square distance (e.g., gravitational interaction of unit mass objects in the non-relativistic limit) can be  
 204 written in Cartesian coordinates as

$$\ddot{x} = -\frac{x}{(x^2 + y^2)^{3/2}}, \quad \ddot{y} = -\frac{y}{(x^2 + y^2)^{3/2}}, \quad (13)$$

205 where one body is fixed at the origin and  $x, y$  are the coordinates of the other body in the plane of its  
 206 orbit [37]. We stabilize the dynamics with respect to the conserved angular momentum  $L$ , yielding

$$\mathcal{M} = \{(x, y) \in \mathbb{R}^2; x\dot{y} + y\dot{x} - L_0 = 0\}. \quad (14)$$

207 Again, we train on 40 trajectories with initial conditions  $(x, y, \dot{x}, \dot{y}) = (1 - e, 0, 0, \sqrt{1 - e/(1 + e)})$ , where  
 208 the eccentricity  $e$  is sampled uniformly via  $e \sim U(0.5, 0.7)$ . Each trajectory consists of a single  
 209 period of the orbit sampled with a timestep of  $\Delta t = 0.1$ .

210 The top row of Figure 2 shows that SNDE achieves stable long-term prediction over multiple orbits,  
 211 whereas unstabilized NODEs exponentially diverge from the correct orbit.

## 212 4.2 Motion of a Rigid Body [first-order, autonomous, non-chaotic, holonomic constraint]

213 The angular momentum vector  $y = (y_1, y_2, y_3)^T$  of a rigid body with arbitrary shape and mass  
 214 distribution satisfies Euler’s equations of motion

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{pmatrix} = \begin{pmatrix} 0 & -y_3 & y_2 \\ y_3 & 0 & -y_1 \\ -y_2 & y_1 & 0 \end{pmatrix} \begin{pmatrix} y_1/I_1 \\ y_2/I_2 \\ y_3/I_3 \end{pmatrix}, \quad (15)$$

215 where the coordinate axes are the principal axes of the body,  $I_1, I_2, I_3$  are the principal moments of  
 216 inertia, and the origin of the coordinate system is fixed at the body’s centre of mass [37]. The motion  
 217 of  $y$  conserves the Casimir function  $C(y) = \frac{1}{2}(y_1^2 + y_2^2 + y_3^2)$ , which is equivalent to conservation  
 218 of angular momentum in the orthogonal body frame and constitutes a holonomic constraint on the  
 219 allowed states of the system. We therefore have the manifold

$$\mathcal{M} = \{(y_1, y_2, y_3) \in \mathbb{R}^3; y_1^2 + y_2^2 + y_3^2 - C_0 = 0\}. \quad (16)$$

220 We train on 40 trajectories with initial conditions  $(y_1, y_2, y_3) = (\cos(\phi), 0, \sin(\phi))$ , where  $\phi$  is drawn  
 221 from a uniform distribution  $\phi \sim U(0.5, 1.5)$ . Each trajectory consists of a 15 seconds sample with a  
 222 timestep of  $\Delta t = 0.1$  seconds.

223 Again, the middle row in Figure 2 demonstrates that unlike vanilla NODE, SNDE manages to stabilize  
 224 the predicted dynamics over a long time horizon in this first-order system.

## 225 4.3 DC-to-DC Converter [first-order, non-autonomous, non-chaotic, conservation law]

226 We now consider an idealized DC-to-DC converter [38, 39] illustrated in  
 227 Figure 3 with dynamics

$$C_1 \dot{v}_1 = (1 - u)i_3, \quad C_2 \dot{v}_2 = ui_3, \quad L_3 \dot{i}_3 = -(1 - u)v_1 - uv_2, \quad (17)$$

228 where  $v_1, v_2$  are state voltages across capacitors  $C_1, C_2$ , respectively,  $i_3$  is  
 229 the state current across an inductor  $L_3$ , and  $u \in \{0, 1\}$  is a control input (a  
 230 switch) that can be used to transfer energy between the two capacitors via  
 231 the inductor. The total energy in the circuit,  $E = \frac{1}{2}(C_1 v_1^2 + C_2 v_2^2 + L_3 i_3^2)$ ,  
 232 is conserved, yielding the manifold

$$\mathcal{M} = \{(v_1, v_2, i_3) \in \mathbb{R}^3; C_1 v_1^2 + C_2 v_2^2 + L_3 i_3^2 - E_0 = 0\}. \quad (18)$$

233 We train on 40 trajectories integrated over 10 seconds with a timestep of  $\Delta t = 0.1$  seconds, where  
 234  $C_1 = 0.1, C_2 = 0.2, L_3 = 0.5$ , and a switching period of 3 seconds, i.e., the switch is toggled every  
 235 1.5 seconds. The initial conditions for  $(v_1, v_2, i_3)$  are drawn from a uniform distribution  $U(0, 1)$ .

236 The bottom row of Figure 2 shows the voltage across  $C_1$  over multiple switching events (g), with  
 237 NODE (blue) accumulating errors every time the switch is applied, whereas SNDE remains accurate.  
 238 Panels (h,i) show the familiar exponentially accumulating errors for vanilla NODE versus constant  
 239 relative errors for SNDE.

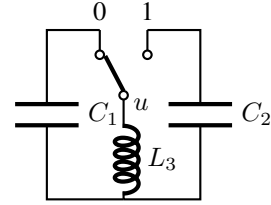


Figure 3: Idealized schematic of a DC-to-DC converter.

## 240 4.4 Controlled Robot Arm [first-order, non-autonomous, non-chaotic, time-dependent control]

241 Next, we apply SNDEs to solve a data-driven inverse kinematics problem [40], that is, learning the  
 242 dynamics of a robot arm that satisfy a prescribed path  $p(t)$ . We consider an articulated robot arm  
 243 consisting of three connected segments of fixed length 1 illustrated in Figure 4(a). Assuming one  
 244 end of the first segment is fixed at the origin and the robot arm is restricted to move in a plane, the  
 245 endpoint  $e(\theta)$  of the last segment is given by

$$e(\theta) = \begin{pmatrix} \cos(\theta_1) + \cos(\theta_2) + \cos(\theta_3) \\ \sin(\theta_1) + \sin(\theta_2) + \sin(\theta_3) \end{pmatrix}, \quad (19)$$

246 where  $\theta_j$  is the angle of the  $j$ -th segment with respect to the horizontal and  $\theta = (\theta_1, \theta_2, \theta_3)$ . The  
 247 problem consists of finding the motion of the three segments  $\theta(t)$  such that the endpoint  $e(\theta)$  follows

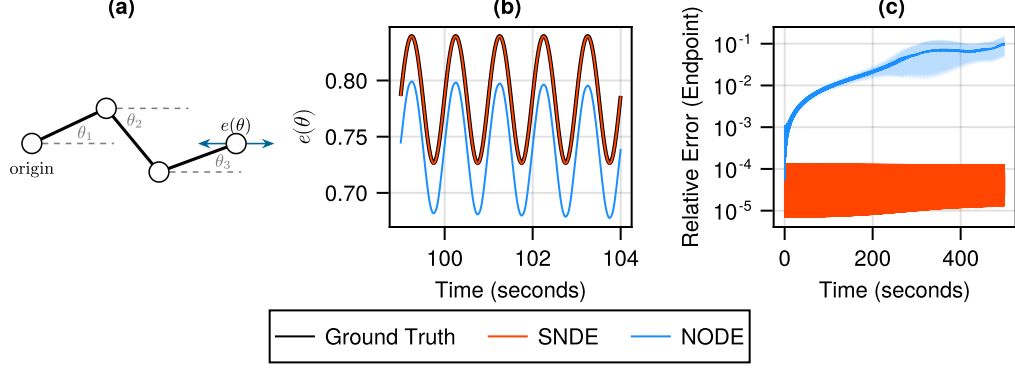


Figure 4: Controlled robot arm. **(a)** Schematic of the robot arm. **(b)** Snapshot of a single test trajectory. After 100 seconds the NODE (blue) has drifted significantly from the prescribed control while the SNDE (red) accurately captures the ground truth dynamics (black). **(c)** Relative error in the endpoint  $e(\theta)$  averaged over 100 test trajectories. NODE (blue) accumulates errors and leaves the prescribed path, while SNDE (red) remains accurate. Shadings in (c) are 95% confidence intervals.

a prescribed path  $p(t)$  in the plane, i.e.,  $e(\theta) = p(t)$ . Minimizing  $\|\dot{\theta}(t)\|$ , it can be shown [41] that the optimal path satisfies

$$\dot{\theta} = e'(\theta)^T (e'(\theta)e'(\theta)^T)^{-1} \dot{p}(t), \quad (20)$$

where  $e'$  is the Jacobian of  $e$ . These will be our ground truth equations of motion.

We stabilize the SNDE with respect to the (time-dependent) manifold

$$\mathcal{M} = \{(\theta, t) \in \mathbb{S} \times \mathbb{R}; e(\theta) - p(t) = 0\}. \quad (21)$$

In particular, we prescribe the path

$$p(t) = e_0 - \begin{pmatrix} \sin(2\pi t)/2\pi \\ 0 \end{pmatrix}, \quad (22)$$

where  $e_0$  is the initial position of the endpoint, such that  $e(\theta)$  traces a line back and forth on the  $x$ -axis. We train on 40 trajectories, integrated over 5 seconds with timestep  $\Delta t = 0.1$  and initial conditions  $(\theta_1, \theta_2, \theta_3) = (\theta_0, -\theta_0, \theta_0)$ , where  $\theta_0$  is drawn from a uniform distribution  $\theta_0 \sim U(\pi/4, \pi/8)$ . Additionally we provide the network with  $\dot{p}$ , the time derivative of the prescribed control.

Figure 4 shows that the unconstrained NODE drifts substantially from the prescribed path, while the SNDE implements the control to a high degree of accuracy and without drift.

#### 4.5 Double Pendulum [second-order, autonomous, chaotic, conservation law]

Finally, we apply stabilization to the chaotic dynamics of the frictionless double pendulum system. The total energy  $E$  of the system is conserved [42], yielding the manifold,

$$\mathcal{M} = \{(\theta_1, \theta_2, \omega_1, \omega_2) \in \mathbb{S}^2 \times \mathbb{R}^2; E(\theta_1, \theta_2, \omega_1, \omega_2) - E_0 = 0\}, \quad (23)$$

where  $\theta_i$  is the angle of the  $i$ -th arm with the vertical and  $\omega_i = \dot{\theta}_i$ . We refer the reader to, for example, Arnold [42] (or the excellent wikipedia entry) for the lengthy equations of motion and expression for the total energy. For simplicity we take  $m_1 = m_2 = 1$  kg,  $l_1 = l_2 = 1$  m, and  $g = 9.81$  ms<sup>-2</sup>. We train on 40 trajectories, each consisting of 10 seconds equally sampled with  $\Delta t = 0.05$ , and with initial conditions  $(\theta_1, \theta_2, \omega_1, \omega_2) = (\phi, \phi, 0, 0)$ , where  $\phi$  is drawn randomly from a uniform distribution  $\phi \sim U(\pi/4, 3\pi/4)$ . We emphasize that this is a highly limited amount of data when it comes to describing the chaotic motion of the double pendulum system, intended to highlight the effect of stabilization in the low-data regime.

Figure 5(a,b) shows that while initially SNDE only marginally outperforms vanilla NODE in terms of the relative error of the state, the longer term relative error in energy is substantially larger for NODE than for SNDE. A certain relative error in state is in fact unavoidable for chaotic systems.



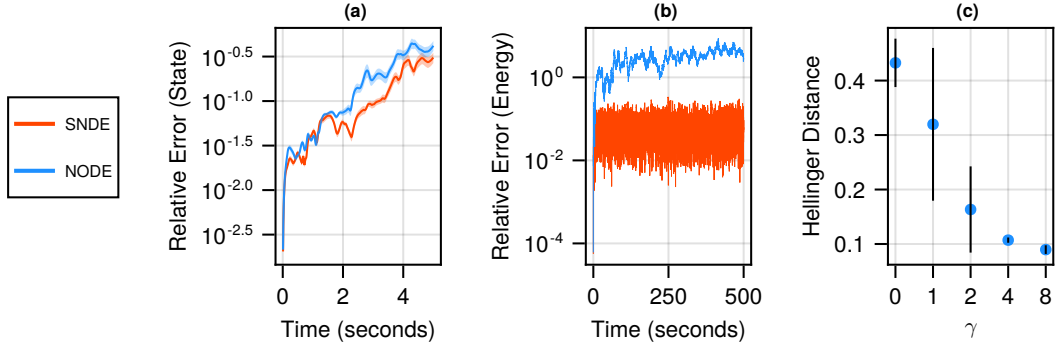


Figure 5: Results for the double pendulum. (a) Relative error in the state over 300 short test trials, shown with 95% confidence intervals (shaded). Compared to the SNDE, the NODE diverges rapidly as it begins to accumulate errors in the energy. (b) Relative error in the energy averaged over 5 long test trials. (c) Comparison of the double pendulum’s invariant measure estimated by the NODE/SNDE versus ground truth, with 95% confidence intervals.

In addition to predicting individual trajectories of the double pendulum, we also consider an additional important task: learning the invariant measure of this chaotic system. This can be motivated in rough analogy to climate predictions, where one also focuses on long-term prediction of the invariant measure of the system, as opposed to predicting individual trajectories in the sense of weather forecasting, which must break down after a short time due to the highly chaotic underlying dynamics. With an analogy to hybrid models of the Earth’s climate, we choose a slightly different training strategy than before, namely a hybrid setup in line with the UDE approach mentioned above, in which the dynamics of the first arm  $\theta_1$  are known, while the dynamics of the second arm  $\theta_2$  are inferred from data. We train on a *single trajectory* of duration 60 seconds with  $\Delta t = 0.05$ . For each model, we then integrate ten trajectories of duration one hour – far longer than the observed data – each with initial conditions drawn from the same invariant set. An invariant measure is estimated from each long trajectory (see Appendix E) and compared with the ground truth via the Hellinger distance.

Figure 5(c) shows that as we set  $\gamma$  to non-zero values, the accuracy in learning the double pendulum’s invariant measure increases dramatically due to stabilization, demonstrating that the ‘climate’ of this system is captured much more accurately by SNDE than by NODE.

## 5 Conclusion

We have introduced stabilized neural differential equations (SNDEs), a method for learning ordinary differential equation systems from observational data, subject to arbitrary explicit constraints such as those imposed by physical conservation laws. Our approach is based on a stabilization term that can be computed efficiently for arbitrary constraint functions and provably renders the invariant manifold asymptotically stable while allowing for all trajectories of the ground truth dynamics. A key benefits of our stabilization are its simplicity and generality, which make it compatible with all common NODE architectures and training methods without requiring any changes to the architecture. Crucially, SNDEs vastly broaden the scope of which constrained dynamics can be learned. We demonstrate their consistent efficacy in a range of settings including first- and second-order, autonomous and non-autonomous (controlled) systems, with constraints stemming from holonomic constraints, conserved first integrals of motion, as well as time-dependent restrictions on the system state. SNDEs are robust with respect to the only tuneable parameter and only incur moderate computational overhead compared to vanilla NODEs.

The current key limitations and simultaneously interesting directions for future work include generalizations to partial differential equations, allowing observations and constraints to be provided in different coordinates, and scaling the method to high-dimensional settings such as learning dynamics from pixel observations, for example in fluid dynamics or climate modelling. Finally, we emphasize that high-dimensional, non-linear dynamics may not be identifiable from just a small number of solution trajectories. Hence, care must be taken when using learned dynamics in high-stakes scenarios (e.g., human robot interactions), especially when going beyond the training distribution.

## References

- [1] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. 2018. doi: 10.48550/ARXIV.1806.07366. URL <https://arxiv.org/abs/1806.07366>. 1, 2, 3
- [2] Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. Universal differential equations for scientific machine learning. 2020. doi: 10.48550/ARXIV.2001.04385. URL <https://arxiv.org/abs/2001.04385>. 1
- [3] Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016. doi: 10.1073/pnas.1517384113. URL <https://www.pnas.org/doi/abs/10.1073/pnas.1517384113>. 1
- [4] Miles Cranmer. Interpretable machine learning for science with pysr and symbolicregression.jl, 2023. 1
- [5] Hananeh Aliee, Fabian J Theis, and Niki Kilbertus. Beyond predictions in neural odes: Identification and interventions. *arXiv preprint arXiv:2106.12430*, 2021. 1
- [6] Sören Becker, Michal Klein, Alexander Neitz, Giambattista Parascandolo, and Niki Kilbertus. Predicting ordinary differential equations with transformers. In *International Conference on Machine Learning (ICML)*, 2023. 1
- [7] Patrick Kidger. On neural differential equations. 2022. doi: 10.48550/ARXIV.2202.02435. URL <https://arxiv.org/abs/2202.02435>. 1, 2
- [8] Takeshi Teshima, Koichi Tojo, Masahiro Ikeda, Isao Ishikawa, and Kenta Oono. Universal approximation property of neural ordinary differential equations. *arXiv preprint arXiv:2012.02414*, 2020. 1
- [9] Han Zhang, Xi Gao, Jacob Unterman, and Tom Arodz. Approximation capabilities of neural odes and invertible residual networks. In *International Conference on Machine Learning*, pages 11086–11095. PMLR, 2020. 1
- [10] Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented neural odes. *Advances in neural information processing systems*, 32, 2019. 1, 2
- [11] Hananeh Aliee, Till Richter, Mikhail Solonin, Ignacio Ibarra, Fabian Theis, and Niki Kilbertus. Sparsity in continuous-depth neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1
- [12] Sam Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks, 2019. 2, 3
- [13] Michael Lutter, Christian Ritter, and Jan Peters. Deep lagrangian networks: Using physics as model prior for deep learning, 2019. 2, 3
- [14] Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks, 2020. 2, 3
- [15] Marc Finzi, Ke Alexander Wang, and Andrew Gordon Wilson. Simplifying hamiltonian and lagrangian neural networks via explicit constraints, 2020. 2, 3
- [16] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989. 2
- [17] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989. 2
- [18] Yingbo Ma, Vaibhav Dixit, Michael J Innes, Xingjian Guo, and Chris Rackauckas. A comparison of automatic differentiation and continuous sensitivity analysis for derivatives of differential equation solutions. In *2021 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–9. IEEE, 2021. 2

- [19] Alexander Norcliffe, Cristian Bodnar, Ben Day, Nikola Simidjievski, and Pietro Liò. On second order behaviour in augmented neural odes. *Advances in Neural Information Processing Systems*, 33:5911–5921, 2020. 2, 3
- [20] RT Chen, D Duvenaud, and Y Rubanova. Latent odes for irregularly-sampled time series. *Advances in Neural Information Processing Systems*, 32, 2019. 2
- [21] Raj Dandekar, Karen Chung, Vaibhav Dixit, Mohamed Tarek, Aslan Garcia-Valadez, Krishna Vishal Vemula, and Chris Rackauckas. Bayesian neural ordinary differential equations. *arXiv preprint arXiv:2012.07244*, 2020. 2
- [22] Maximilian Gelbrecht, Niklas Boers, and Jürgen Kurths. Neural partial differential equations for chaotic systems. *New Journal of Physics*, 23(4):043005, 2021. 2
- [23] Chris Finlay, Jörn-Henrik Jacobsen, Levon Nurbekyan, and Adam Oberman. How to train your neural ode: the world of jacobian and kinetic regularization. In *International conference on machine learning*, pages 3154–3164. PMLR, 2020. 3
- [24] Arnab Ghosh, Harkirat Behl, Emilien Dupont, Philip Torr, and Vinay Namboodiri. Steer: Simple temporal regularization for neural ode. *Advances in Neural Information Processing Systems*, 33:14831–14843, 2020. 3
- [25] Jacob Kelly, Jesse Bettencourt, Matthew J Johnson, and David K Duvenaud. Learning differential equations that are easy to solve. *Advances in Neural Information Processing Systems*, 33: 4370–4380, 2020. 3
- [26] Avik Pal, Yingbo Ma, Viral Shah, and Christopher V Rackauckas. Opening the blackbox: Accelerating neural differential equations by regularizing internal solver heuristics. In *International Conference on Machine Learning*, pages 8325–8335. PMLR, 2021. 3
- [27] Patrick Kidger, Ricky TQ Chen, and Terry J Lyons. "hey, that's not an ode": Faster ode adjoints via seminorms. In *ICML*, pages 5443–5452, 2021. 3
- [28] Yaofeng Desmond Zhong, Biswadip Dey, and Amit Chakraborty. Symplectic ode-net: Learning hamiltonian dynamics with control. *arXiv preprint arXiv:1909.12077*, 2019. 3
- [29] Yuhan Chen, Takashi Matsubara, and Takaharu Yaguchi. Neural symplectic form: learning hamiltonian equations on general coordinate systems. *Advances in Neural Information Processing Systems*, 34:16659–16670, 2021. 3
- [30] Aiqing Zhu, Pengzhan Jin, and Yifa Tang. Deep hamiltonian networks based on symplectic integrators. *arXiv preprint arXiv:2004.13830*, 2020. 3
- [31] Thai Duong and Nikolay Atanasov. Hamiltonian-based neural ode networks on the se (3) manifold for dynamics learning and control. *arXiv preprint arXiv:2106.12782*, 2021. 3
- [32] Nate Gruver, Marc Finzi, Samuel Stanton, and Andrew Gordon Wilson. Deconstructing the inductive biases of hamiltonian neural networks. *arXiv preprint arXiv:2202.04836*, 2022. 3
- [33] Zhijie Chen, Mingquan Feng, Junchi Yan, and Hongyuan Zha. Learning neural hamiltonian dynamics: A methodological overview. *arXiv preprint arXiv:2203.00128*, 2022. 3
- [34] Yi Heng Lim and Muhammad Firmansyah Kasim. Unifying physical systems' inductive biases in neural ode using dynamics constraints. *arXiv preprint arXiv:2208.02632*, 2022. 3
- [35] Aaron Lou, Derek Lim, Isay Katsman, Leo Huang, Qingxuan Jiang, Ser Nam Lim, and Christopher M De Sa. Neural manifold ordinary differential equations. *Advances in Neural Information Processing Systems*, 33:17548–17558, 2020. 3
- [36] Hong Sheng Chin. *Stabilization methods for simulations of constrained multibody dynamics*. PhD thesis, University of British Columbia, 1995. URL <https://open.library.ubc.ca/collections/ubctheses/831/items/1.0080031>. 4, 14

- [37] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration*, volume 31 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 2006. ISBN 3-540-30663-3; 978-3-540-30663-4. Structure-preserving algorithms for ordinary differential equations. 6, 7
- [38] N.E. Leonard and P.S. Krishnaprasad. Control of switched electrical networks using averaging on lie groups. In *Proceedings of 1994 33rd IEEE Conference on Decision and Control*, volume 2, pages 1919–1924 vol.2, 1994. doi: 10.1109/CDC.1994.411098. 7
- [39] Simone Fiori. Manifold calculus in system theory and control—fundamentals and first-order systems. *Symmetry*, 13(11), 2021. ISSN 2073-8994. doi: 10.3390/sym13112092. URL <https://www.mdpi.com/2073-8994/13/11/2092>. 7
- [40] Suhan Park, Mathew Schwartz, and Jaeheung Park. Node ik: Solving inverse kinematics with neural ordinary differential equations for path planning, 2022. 7
- [41] Ernst Hairer. Solving differential equations on manifolds. 2011. URL <https://www.unige.ch/~hairer/poly-sde-mani.pdf>. 8
- [42] Vladimir Igorevich Arnold. *Mathematical methods of classical mechanics*, volume 60. Springer Science & Business Media, 2013. 8
- [43] C. W. Gear. Differential-algebraic equation index transformations. *SIAM Journal on Scientific and Statistical Computing*, 9(1):39–47, 1988. doi: 10.1137/0909004. URL <https://doi.org/10.1137/0909004>. 14
- [44] K. E. Brenan, S. L. Campbell, and L. R. Petzold. *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*. Society for Industrial and Applied Mathematics, 1995. doi: 10.1137/1.9781611971224. URL <https://epubs.siam.org/doi/abs/10.1137/1.9781611971224>. 14
- [45] Uri M. Ascher and Linda R. Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. Society for Industrial and Applied Mathematics, USA, 1st edition, 1998. ISBN 0898714125. 14
- [46] Linda Petzold. Differential/algebraic equations are not ode’s. *SIAM Journal on Scientific and Statistical Computing*, 3(3):367–384, 1982. doi: 10.1137/0903023. URL <https://doi.org/10.1137/0903023>. 14
- [47] J. Baumgarte. Stabilization of constraints and integrals of motion in dynamical systems. *Computer Methods in Applied Mechanics and Engineering*, 1(1):1–16, 1972. ISSN 0045-7825. doi: [https://doi.org/10.1016/0045-7825\(72\)90018-7](https://doi.org/10.1016/0045-7825(72)90018-7). URL <https://www.sciencedirect.com/science/article/pii/0045782572900187>. 14
- [48] Suyong Kim, Weiqi Ji, Sili Deng, Yingbo Ma, and Christopher Rackauckas. Stiff neural ordinary differential equations. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(9), 09 2021. ISSN 1054-1500. doi: 10.1063/5.0060697. URL <https://doi.org/10.1063/5.0060697.093122>. 14
- [49] Ludwig Arnold, Christopher KRT Jones, Konstantin Mischaikow, Geneviève Raugel, and Ludwig Arnold. *Random dynamical systems*. Springer, 1995. 16, 17
- [50] Mickaël D Chekroun, Eric Simonnet, and Michael Ghil. Stochastic climate dynamics: Random attractors and time-dependent invariant measures. *Physica D: Nonlinear Phenomena*, 240(21): 1685–1700, 2011. 16, 17
- [51] David Diego, Kristian Agasøster Haaga, and Bjarte Hannisdal. Transfer entropy computation using the perron-frobenius operator. *Phys. Rev. E*, 99:042212, Apr 2019. doi: 10.1103/PhysRevE.99.042212. URL <https://link.aps.org/doi/10.1103/PhysRevE.99.042212>. 17
- [52] Kristian Agasøster Haaga, George Datseris, Inga Kottlarz, Alistair White, HeineRugland, and Steven G. Johnson. Juliadynamics/complexitymeasures.jl: v2.7.2, April 2023. URL <https://doi.org/10.5281/zenodo.7862020>. 17

- 449 [53] Harald Cramér. *Mathematical methods of statistics*, volume 26. Princeton university press,  
450 1999. 17
- 451 [54] James H Verner. Numerically optimal runge–kutta pairs with interpolants. *Numerical Algo-*  
452 *rithms*, 53(2-3):383–396, 2010. doi: 10.1007/s11075-009-9290-3. URL [https://doi.org/](https://doi.org/10.1007/s11075-009-9290-3)  
453 [10.1007/s11075-009-9290-3](https://doi.org/10.1007/s11075-009-9290-3). 17
- 454 [55] Christopher Rackauckas and Qing Nie. DifferentialEquations.jl—a performant and feature-rich  
455 ecosystem for solving differential equations in Julia. *Journal of Open Research Software*, 5(1),  
456 2017. 17
- 457 [56] Michael Innes, Elliot Saba, Keno Fischer, Dhairya Gandhi, Marco Concetto Rudilosso,  
458 Neethu Mariya Joy, Tejan Karmali, Avik Pal, and Viral Shah. Fashionable modelling with flux.  
459 *CoRR*, abs/1811.01457, 2018. URL <https://arxiv.org/abs/1811.01457>. 17
- 460 [57] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 17
- 461 [58] Ch Tsitouras. Runge–kutta pairs of order 5 (4) satisfying only the first column simplifying  
462 assumption. *Computers & Mathematics with Applications*, 62(2):770–775, 2011. 17

## A Differential Algebraic Equations

A differential algebraic equation (DAE) in its most general, implicit form is

$$F(t, x, \dot{x}) = 0, \quad (24)$$

where  $x \in \mathbb{R}^n$  and  $F: \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ . When  $\partial F / \partial \dot{x}$  is nonsingular, Equation (24) is an implicit ODE and by the implicit function theorem may be written as an explicit ODE in the form  $\dot{x} = f(x, t)$  [43]. In the more interesting case of singular  $\partial F / \partial \dot{x}$ , an important special case of Equation (24) is given by semi-explicit DAEs in Hessenberg form, for example,

$$\dot{y} = f(t, y, z) \quad (25a)$$

$$0 = g(t, y, z), \quad (25b)$$

where  $x = (y, z)$ . We call  $y$  the *differential variables*, since their derivatives appear in the equations, and  $z$  the *algebraic variables*, since their derivatives do not. The semi-explicit form of Equation (25) highlights the connection between certain classes of DAEs and ODEs subject to constraints.

It is generally possible to differentiate the constraints Equation (25b) a number of times and substitute the result into Equation (25a) to obtain a mathematically equivalent ODE. The number of differentiations required to do so is the *differential index* of the DAE, and corresponds loosely to the “distance” of the DAE from an equivalent ODE. The differential index – and related measures of index not directly based on differentiation – is used extensively to classify DAEs, especially in the context of numerical methods for their solution [44]. Each differentiation of the constraints reduces the index of the system by one (ODEs have index 0).

Of particular interest in the context of this paper are constrained ODEs of the form

$$\dot{u} = f(t, u) \quad (26a)$$

$$0 = g(t, u), \quad (26b)$$

where  $u \in \mathbb{R}^n$ ,  $f: \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ , and  $g: \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^m$ . Equation (26) can be written as a semi-explicit Hessenberg index-2 DAE, with  $u$  as the differential variables and  $m$  Lagrange multipliers as the algebraic variables, for example,

$$\dot{u} = f(t, u) - D(u)\lambda \quad (27a)$$

$$0 = g(t, u), \quad (27b)$$

where  $\lambda \in \mathbb{R}^m$  and  $D(u)$  is any bounded matrix function such that  $GD$ , where  $G = g_u$  is the Jacobian of  $g$ , is boundedly invertible for all  $t$  [45]. We can therefore approach the task of solving the constrained ODE Equation (26) from the perspective of solving the Hessenberg index-2 DAE Equation (27).

DAEs are not ODEs, however, and a number of additional complications are encountered when we seek numerical solutions [46]. Generally, the higher the index, the harder it is to solve a given DAE. For this reason, it is common to first perform an index reduction (i.e. differentiate the constraints) before applying numerical methods. However, the numerical solution of the resulting index-reduced system may exhibit *drift off* from the invariant manifold defined by the original constraints. For this reason, Baumgarte [47] proposed a stabilization procedure for index-reduced DAEs that renders the invariant manifold asymptotically stable. Baumgarte’s stabilization is, in turn, a special case of the stabilization procedure later proposed by Chin [36] and adapted by us in this paper. We emphasize, however, that our stabilization procedure addresses a different application and problem than these related methods; while Baumgarte and Chin sought to stabilize drift off from the invariant manifold due to discretization error in an index-reduced DAE, we seek to constrain some learned dynamics imperfectly approximated by a neural network.

Finally, one may ask why a neural network could not be incorporated directly into Equation (27) and the resulting index-2 DAE solved directly. While possible in principle, DAEs require implicit numerical methods, with the result that the computational cost of computing gradients of solutions – whether via automatic differentiation or adjoint sensitivity analysis – scales with the cube of the system size [48], rather than the efficient linear scaling when computing gradients of explicit solvers.

Table 1: Training time of (SO)NODEs vs SNDEs. All experiments are trained for 1,000 epochs on an Intel(R) Xeon(R) CPU E5-2667 v3 @ 3.20GHz. Statistics are calculated over 5 random seeds.

Model	$\gamma$	Training Time (seconds)					
		Two-Body Problem		Rigid Body		DC-to-DC Converter	
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
NODE	-	10,580	271	9,730	71	14,000	316
SNDE	0.1	12,060	206	12,000	283	18,060	524
	1	12,180	194	12,500	126	18,020	549
	2	12,160	102	12,340	301	18,280	240
	4	12,980	147	14,160	280	18,020	354
	8	14,000	268	15,320	376	18,200	482
	16	14,260	162	15,680	519	-	-
	32	15,300	167	17,140	680	-	-

## 504 B The Choice of $\gamma$ and Runtime Implications

505 We assess the computational cost of SNDEs compared to vanilla (SO)NODEs. SNDEs require the  
 506 computation of (and backpropagation through) the pseudoinverse of the Jacobian of the constraint  
 507 function  $g$ . Additionally, as  $\gamma$  is increased, SNDEs may require more solver steps at a given error  
 508 tolerance, with the SNDE eventually becoming stiff for sufficiently large  $\gamma$ . Naively, one may thus  
 509 expect a noticeable increase in runtime. However, as described in Section 2, the computational cost of  
 510 training NDEs also depends on the “complexity” of the learned dynamics, which in turn determines  
 511 how many function evaluations are required by the solver. This leads to nontrivial interactions  
 512 between the added computation of enforcing constraints and the thereby potentially regularized  
 513 “simpler” dynamics, which may require fewer function evaluations by the solver.

514 In Table 1, we report comparisons of training times between NODEs and SNDEs for different values  
 515 of  $\gamma$  for three settings. SNDEs take roughly 1.2 to 1.8 times longer to train, with smaller values of  $\gamma$   
 516 incurring less overhead. Overall, this is a manageable increase for most relevant scenarios.

517 To complement these results, Figure 6 shows that the relative error remains almost unchanged for  
 518 a large range of  $\gamma$  values, that is, beyond a certain minimum value SNDEs are not sensitive to the  
 519 specific choice of  $\gamma$ . Even a value of  $\gamma = 1$  works well in the settings we have considered, indicating  
 520 that we can typically get away with runtime increases of a factor of 1.2. However, larger values of  $\gamma$   
 521 only lead to slightly increased training times, while potentially enforcing the constraints to a higher  
 522 degree of accuracy.

523 Finally, we also show inference times in Table 2. Here, the trend reverses and larger values of  $\gamma$  lead  
 524 to lower inference times. This is because the solver requires fewer steps (has higher acceptance rates  
 525 of proposed step sizes) for stronger stabilization. Hence, while predictive performance is largely  
 526 unaffected, one can use the specific choice of  $\gamma$  as a tuning knob that trades off training time versus  
 527 inference time.

## 528 C Runtime Evaluation

529 After the main paper deadline, we found it more natural to merge appendices Appendix B and  
 530 Appendix C into one. This section is now empty, since runtime evaluations and comparisons with  
 531 vanilla (SO)NODEs are already discussed in Appendix B.

## 532 D Additional Experiments

### 533 D.1 Stable Time

534 The unstabilized, vanilla NODEs of Figure 2 are characterized by an initial drift from the invariant  
 535 manifold that gives way to a subsequent rapid divergence. In practice, however, certain test trials  
 536 may remain stable for significantly longer than others. In this section, we characterize the *stable time*

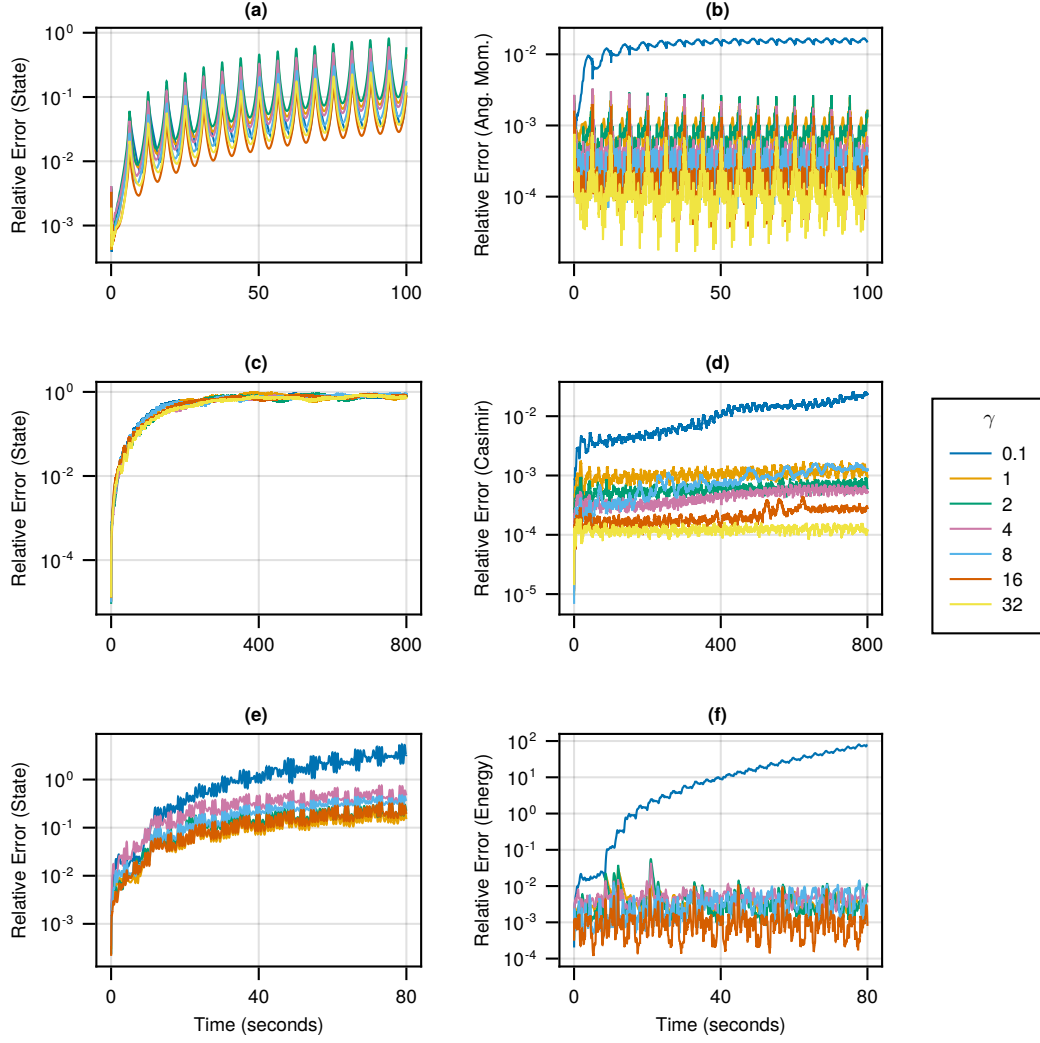


Figure 6: Effect of  $\gamma$  on relative errors. **Top row:** Two-body problem (a-b). **Middle row:** Rigid body (c-d). **Bottom row:** DC-to-DC converter (e-f). As in Figure 2, all relative errors are averaged over 100 test trials. Beyond a certain value, SNDEs are not highly sensitive to the choice of  $\gamma$ , although larger values may enforce the constraints more accurately.

537  $T_{\text{stab}}$  of an individual test trial as the time elapsed until the relative error  $E(t)$  in the predicted system  
538 state  $\hat{u}(t)$  exceeds a given threshold value  $E_{\text{stab}}$ , i.e.

$$T_{\text{stab}} = \max \{t \mid E(t) < E_{\text{stab}}\}. \quad (28)$$

539 Taking  $E_{\text{stab}} = 10^3$ , Table 3 shows  $T_{\text{stab}}$  for the the two-body problem, rigid body, and DC-to-DC  
540 converter experiments. Across several thousand test trials in this paper, we did not observe a single  
541 SNDE model diverge.

## 542 E Invariant Measure

543 Given that the double pendulum is a chaotic system, predictions of individual trajectories will break  
544 down after short times. We therefore also quantify the performance of NODEs and SNDEs in terms  
545 of their ability to capture the double pendulum's invariant measure. We refer to Arnold et al. [49] and  
546 Chekroun et al. [50] for detailed definitions of invariant measures; in short, a measure  $\mu$  is said to be  
547 invariant under some flow  $\Phi$  if  $\mu(\Phi^{-1}(t)(\mathcal{A})) = \mu(\mathcal{A})$  for all measurable sets  $\mathcal{A}$ . Invariant measures  
548 are commonly used to characterise the long-term dynamical characteristics of chaotic dynamical



Table 2: Inference time and (adaptive) solver statistics of (SO)NODEs vs SNDEs for the two-body problem experiment. Inference time statistics are calculated using the same 100 test initial conditions as in Figure 2, integrated for 20 seconds (short enough so that the NODE solution does not diverge). Solver step statistics are reported for a single test trial, intended to illustrate the observed trends in inference time. SNDEs are cheaper at inference time due to significantly fewer rejected solver steps.

Model		Inference Time (seconds)		Solver Steps		
Type	$\gamma$	Median	Mean	Accepted	Rejected	RHS Evaluations
NODE	-	2.44	$2.51 \pm 0.04$	2,379	3,343	34,335
SNDE	0.1	2.14	$2.17 \pm 0.03$	2,040	2,874	29,487
	1	2.19	$2.19 \pm 0.03$	2,054	2,704	28,551
	2	2.22	$2.27 \pm 0.04$	2,061	2,541	27,615
	4	2.07	$2.15 \pm 0.05$	2,180	2,382	27,375
	8	2.05	$2.07 \pm 0.04$	2,355	1,877	25,395
	16	1.98	$2.03 \pm 0.05$	2,677	1,437	24,687
	32	1.96	$1.99 \pm 0.04$	3,219	1,029	25,491

Table 3: Stable time of NODEs for the same 100 test trials as shown in Figure 2. SNDE models (not shown) did not diverge during any trial.

Experiment	Trial Length (seconds)	NODE Stable Time (seconds)				
		Min.	Max.	Median	Mean	Std. Dev.
Two-Body Problem	200.0	52.0	182.8	121.7	117.3	30.3
Rigid Body	1600.0	341.7	1600.0	1600.0	1349.9	468.4
DC-to-DC Converter	160.0	19.3	160.0	113.9	110.49	45.6

systems (see, for example, Arnold et al. [49], Chekroun et al. [50] for a discussion of the invariant measure of the paradigmatic Lorenz-63 system). Since the double pendulum is an ergodic system, averages over long times approximate ensemble averages. We can therefore obtain a sample of the invariant measure numerically by integrating the system for a very long time. Concretely, we estimate the invariant measure from a single long trajectory using an algorithm due to Diego et al. [51], implemented in ComplexityMeasures.jl [52], based on a numerical estimate of the transfer operator. We then use the Hellinger distance [53] to compare the resulting probability distribution with the ground truth value for the double pendulum.

## F Architecture and Training

Training trajectories are generated using the 9(8) explicit Runge-Kutta algorithm due to Verner [54], implemented in DifferentialEquations.jl [55] as Vern9, with absolute and relative tolerances of  $10^{-24}$ . Each trajectory is split into non-overlapping chunks of 3 timesteps each, with all chunks then randomized and split into training and validation sets in the ratio 75:25.

Networks are implemented in Flux.jl [56] and consist of fully-connected dense layers with ReLU activation functions. All experiments are trained for 1,000 epochs using the AdamW optimizer [57] with weight decay of  $10^{-6}$  and an exponentially decaying learning rate schedule. During training, trajectories are integrated using the 5(4) explicit Runge-Kutta algorithm due to Tsitouras [58], implemented in DifferentialEquations.jl [55] as Tsit5, with absolute and relative tolerances of  $10^{-6}$ .

The stabilization hyperparameter  $\gamma$  as well as network sizes and learning rates are optimized for each experiment and are summarized in Table 4.

In Figure 2 and Figure 4, average relative errors are calculated over 100 test trials with initial conditions drawn from the same distribution as the training trajectories. In Figure 5, average relative errors are calculated over 300 test trials.

Table 4: Additional hyperparameters.

	Experiment				
	Two-Body Problem	Rigid Body	DC-to-DC Converter	Robot Arm	Double Pendulum
$\gamma$	8	32	8	16	16
Hidden Layers	2	2	2	2	2
Hidden Width	128	64	64	128	128
Max LR	$10^{-3}$	$10^{-4}$	$5 \times 10^{-3}$	$10^{-3}$	$10^{-2}$
Min LR	$10^{-5}$	$10^{-5}$	$10^{-5}$	$10^{-5}$	$10^{-4}$