Bergische Universität Wuppertal

Fakultät für Mathematik und Naturwissenschaften

Institute of Mathematical Modelling, Analysis and Computational Mathematics (IMACM)

Preprint BUW-IMACM 26/01

Luca Di Persio, Matthias Ehrhardt and Youness Outaleb

Stochastic Port-Hamiltonian Neural Networks: Universal Approximation with Passivity Guarantees

March 5, 2026

http://www.imacm.uni-wuppertal.de

Stochastic Port-Hamiltonian Neural Networks: Universal Approximation with Passivity Guarantees

Luca Di Persio [a], Matthias Ehrhardt [b], Youness Outaleb [c]

[a] *Department of Computer Science - College of Mathematics*
University of Verona, Strada le Grazie 15 - 37134 Verona, Italy

[b] *University of Wuppertal, Chair of Applied and Computational Mathematics,*
Gaußstrasse 20, 42119 Wuppertal, Germany

[c] *Doctoral School in Mathematics, University of Trento*
Via Sommarive, 14 - 38123 Povo, Italy

Abstract

Stochastic port-Hamiltonian systems represent open dynamical systems with dissipation, inputs, and stochastic forcing in an energy based form. We introduce stochastic port-Hamiltonian neural networks, SPH-NNs, which parameterize the Hamiltonian with a feedforward network and enforce skew symmetry of the interconnection matrix and positive semidefiniteness of the dissipation matrix. For Itô dynamics we establish a weak passivity inequality in expectation under an explicit generator condition, stated for a stopped process on a compact set. We also prove a universal approximation result showing that, on any compact set and finite horizon, SPH-NNs approximate the coefficients of a target stochastic port-Hamiltonian system with C^2 accuracy of the Hamiltonian and yield coupled solutions that remain close in mean square up to the exit time. Experiments on noisy mass spring, Duffing, and Van der Pol oscillators show improved long horizon rollouts and reduced energy error relative to a multilayer perceptron baseline.

Key words: (Discrete) stochastic port-Hamiltonian system; port-Hamiltonian Neural Networks; Passivity; Interconnection.

1 Introduction

Recently, there has been growing interest in using Hamiltonian-based frameworks to model complex physical systems, especially in the field of machine learning for dynamical systems. These approaches are appealing because they integrate structural and physical constraints, such as energy conservation and passivity, into data-driven models. However, despite their universal approximation capabilities, traditional *neural networks* (NNs) cannot generally enforce fundamental physical laws. Consequently, they may produce predictions that violate energy conservation or stability constraints, resulting in unphysical behaviors such as uncontrolled energy growth or dissipation in conservative systems.

Email addresses: luca.dipersio@univr.it (Luca Di Persio), ehrhardt@uni-wuppertal.de (Matthias Ehrhardt), youness.outaleb@unitn.it (Youness Outaleb).

1.1 Learning Physics Under Uncertainty

There is a particular need for data-efficient and physically reliable learning algorithms in domains where safety, interpretability, and extrapolation are essential. Examples range from robotics and autonomous systems, where inaccurate energy accounting can lead to catastrophic controller failures, to molecular simulation, where violations of stochastic energetics distort statistical observables.

In these applications, two goals emerge: learned models should respect intrinsic geometric or energetic properties (e.g., symplecticity, skew-symmetry, positive definiteness, and passivity) that are known a priori. However, environmental disturbances, parameter variability, and measurement noise must also be taken into account. Deterministic models alone are insufficient when uncertainty alters qualitative behavior.

Port-Hamiltonian theory provides a unifying language that fulfills both criteria. It builds upon classical Hamil-

tonian mechanics by explicitly distinguishing between storage (energy), dissipation (entropy production), and interaction (ports) via Dirac structures. When Brownian noise is present, *stochastic port-Hamiltonian systems* (SPHS) are obtained, in which noise enters through additional stochastic ports. However, only a few machine learning papers have addressed the full stochastic, energy-aware setting to date.

1.2 Structure-Preserving Machine Learning

Many techniques have been proposed to embed physical structure into learning algorithms. One such technique is *physics-informed neural networks* (PINNs), which constrain the solutions of partial differential equations (PDEs) by adding the residuals of the governing equations to the loss function [15,10]. However, vanilla PINNs are not well-suited for long-time integration of Hamiltonian flows because they do not guarantee symplecticity. *Hamiltonian neural networks* (HNNs) address this issue for conservative, deterministic systems by learning a scalar Hamiltonian H_θ, whose gradients reproduce Hamilton's equations [8].

Port-Hamiltonian neural networks (PHNNs) are an extension of HNNs that allow for non-conservative interconnections and damping matrices. The early work of van der Schaft and Jeltsema [17] inspired data-driven identification methods. Regarding model reduction, Breiten and Schulze proposed a linear-quadratic Gaussian (LQG) balanced truncation method for linear port-Hamiltonian descriptor systems that preserves the port-Hamiltonian structure [2]. While not a learning algorithm, this method demonstrates how to combine stochastic optimal control ideas with structure preservation, offering a complementary approach to noise-aware modeling. However, these contributions typically address deterministic settings and often either overlook noise or add it after training. For an overview of recent machine learning techniques for deterministic port-Hamiltonian systems (PHSs), see Cherifi [3].

1.3 Related Work on Stochastic Dynamics

Over the last five years, the intersection of *stochastic differential equations* (SDEs) and deep learning has given rise to a vibrant research field. This field is driven by the need to model noisy physical systems, financial time series, and irregular biomedical records. Below, we review the relevant literature on structure preservation, emphasizing how each method addresses – or fails to address – energy consistency, passivity, and uncertainty.

In their work, Kidger and Lyons [11] cast the drift and diffusion of an Itô SDE as neural networks trained by stochastic adjoint methods, forming the Neural SDE framework. Subsequent refinements have improved training stability by expanding state spaces [7] and by using robust solvers such as Milstein nets [18]. However, these black-box approaches, while expressive, disregard the geometric constraints of Hamiltonian flows.

Inspired by deterministic HNNs, Kong *et al.* [13] introduced stochastic HNNs (SHNNs), which add isotropic Langevin noise to the canonical equations. Di Persio *et al.* [6] developed the Port-Hamiltonian Neural Network (PHNN) framework and extended this concept by providing a systematic way to model both energy dissipation and external interactions (ports). This overcomes the limitations of gradient-based damping and enables the simulation of interconnected stochastic systems. These Hamiltonian-inspired architectures are complemented by *stochastic PINNs* (sPINNs) [14], which extend the PINN residual loss to stochastic dynamics. However, residual-based training alone does not guarantee structure preservation (e.g., passivity or energy consistency) without additional constraints.

Score-based generative models offer an alternative approach. By training a time-reversed SDE to match the data distribution, one obtains a controllable forward SDE. Zhang *et al.* [16] embedded port-Hamiltonian priors into the score network via energy regularizers. Their sampler enforces passivity only in distribution. In contrast, our parameterization (Section 3) enforces the structural constraints $J = -J^\top$ and $R \succeq 0$ exactly, and admits localized weak-passivity bounds under an explicit generator inequality (Corollary 5).

Current stochastic-dynamic learning methods either ignore the energy structure, are restricted to conservative noise models, or treat passivity as an afterthought. The proposed SPH-NN, which integrates port-Hamiltonian geometry with flexible neural parameterization, improves long-horizon rollouts and energy-related metrics on the benchmarks considered here relative to a standard MLP baseline.

1.4 Present Work

This paper introduces the Stochastic Port-Hamiltonian Neural Network (SPH-NN), which closes the existing gap. Starting from the axiomatic definition of SPHS [4,5], we parameterize the Hamiltonian H_θ with a feedforward network and construct the drift in port-Hamiltonian form. In the deterministic case this yields the standard energy balance

$$\dot{H}_\theta = u^\top y_\theta - \nabla H_\theta^\top R_\theta \nabla H_\theta.$$

In the Itô stochastic case, the expected energy involves an additional Itô correction $\frac{1}{2}\operatorname{Tr}(\sigma\sigma^\top \nabla^2 H_\theta)$, and weak passivity follows under an explicit generator inequality (Corollary 5).

The dissipative co-matrix $R(x)$ and the interconnection matrix $J(x)$ are assigned to parameterizations that are

tractable and maintain positive semi-definiteness and skew-symmetry, respectively. The stochastic signature $\sigma(x)$ can be prescribed or learned, enabling fidelity to noisy data.

The result is a learned SDE that retains the port-Hamiltonian structural constraints exactly and admits weak-passivity bounds in expectation (Corollary 5), together with universal approximation guarantees (Theorem 4). Figures 1 and 2 illustrate the benefits of conservative and damped systems affected by nontrivial noise.

1.5 Contributions

This paper makes three contributions. First, we present the first end-to-end differentiable architecture that incorporates the complete stochastic port-Hamiltonian formalism formulated in the Itô sense within a neural network. Our construction guarantees skew-symmetry and positive semi-definiteness by design, and we study weak passivity in the sense of Definition 1 (see Corollary 7, Proposition 8, and Section 4).

Second, we prove a universal approximation theorem on compact sets: for any compact $\mathcal{K}$ and finite horizon T, the SPH-NN parameterization approximates the SPHS coefficients (including C^2-accuracy of H on $\mathcal{K}$) and the corresponding coupled solutions are close up to the exit time from $\mathcal{K}$ (Theorem 4). We furthermore derive weak-passivity bounds in expectation via an explicit generator inequality: a localized bound on $\mathcal{K}$ for the stopped process, including an additive term that captures the Itô correction and approximation error (Corollary 5), a strict localized version without the additive term under $r \leq -\delta$ on $\mathcal{K}$ (Corollary 7), and a non-localized weak-passivity result under the additional global condition $\hat{r} \leq 0$ on $\mathbb{R}^n$ plus integrability assumptions (Proposition 8). To the best of our knowledge, this is the first rigorous approximation result coupling stochastic dynamics with a port-Hamiltonian structure.

Third, using canonical benchmarks, we demonstrate that SPH-NNs achieve lower long-term energy drift and more accurate rollouts than *multilayer perceptrons* (MLPs). In this context, a rollout starts from an initial condition and simulates the learned dynamics forward over many time steps to produce a predicted trajectory, and the *rollout error* quantifies the deviation of this trajectory from the true one over the time horizon.

1.6 Organization of the Paper

Section 2 reviews the port-Hamiltonian and stochastic foundations. Section 3 introduces the SPH-NN architecture and training procedure. Section 4 presents numerical experiments and ablation studies. Section 6 discusses limitations and perspectives, and Section 7 concludes.

2 Theoretical Background

This section introduces the foundational concepts that underpin our proposed approach. First, we review port-Hamiltonian systems.

2.1 Port-Hamiltonian Systems

Port-Hamiltonian systems (PHSs) are a powerful, structured framework for modeling physical systems involving energy storage, dissipation, and environmental interaction. PHSs are particularly well-suited for multi-physics applications where different physical domains are interconnected via energy exchanges.

A standard PHS is described by the following state-space representation:

$$\dot{x} = \big[J(x) - R(x)\big]\nabla H(x) + G(x)u. \tag{1}$$

Here, $x \in \mathbb{R}^n$ denotes the state vector, $H(x)\colon \mathbb{R}^n \to \mathbb{R}$ is the Hamiltonian function representing the total stored energy in the system, $J(x) = -J(x)^\top$ is a skew-symmetric matrix encoding the structure of the internal energy-conserving interconnections, and $R(x)$ is a positive semidefinite matrix modeling dissipative effects. The influence of external inputs $u \in \mathbb{R}^m$ is captured by the control input matrix $G(x)$.

A key property of PHSs is *passivity*, which ensures that the system cannot generate energy internally. Instead, the system can only store, dissipate, or exchange energy with its environment. Passivity plays a central role in analyzing system stability and designing robust, physically consistent controllers. Port-Hamiltonian models naturally extend to networked and distributed systems, making them valuable tools for theoretical analysis and practical engineering applications.

2.2 Stochastic Extension

Uncertainties, measurement noise, and unmodeled external disturbances can significantly impact the dynamics of many real-world systems. To account for these effects, the port-Hamiltonian framework can be extended to include stochastic perturbations, resulting in a *stochastic port-Hamiltonian system* (SPHS) [4].

A typical SPHS is modeled by introducing a stochastic term driven by a Brownian motion W_t into the deterministic formulation (1):

$$\begin{aligned} dX_t = \big[J(X_t) - R(X_t)\big]\nabla H(X_t)\,dt \\ + g(X_t)u_t\,dt + \sigma(X_t)\,dW_t, \end{aligned} \tag{2}$$

where $X_t \in \mathbb{R}^n$ denotes the system state at time t, $J(X_t)$ is a skew-symmetric interconnection matrix,

$R(X_t)$ is a positive semidefinite dissipation matrix, $H(X_t)$ is the Hamiltonian representing stored energy, $y(X_t) = g(X_t)^\top \nabla H(X_t)$ is the output, and $\sigma(X_t)$ is the diffusion matrix that characterizes how noise enters the system.

This stochastic formulation preserves the key structural components of PHSs, namely the separation of conservative, dissipative, and external effects, while accommodating random fluctuations. This extension is essential for accurately modeling physical systems subject to uncertainty, and ensures that important properties, such as passivity and energy consistency, are retained in expectation or in a probabilistic sense.

2.3 Energy Principles and Passivity

The principle of energy conservation and balance is a central concept in PHS theory. The Hamiltonian function H represents the system's total stored energy, while the dissipation matrix $R(x)$ models energy loss due to resistive effects. In deterministic PHSs, the system's structure ensures that it cannot generate energy internally, thereby satisfying the principle of *passivity*.

In the Itô interpretation, applying Itô's formula to $H(X_t)$ produces an additional second-order term $\frac{1}{2}\operatorname{Tr}(\sigma\sigma^\top \nabla^2 H)$ in the energy balance, so noise may inject or remove energy in expectation even when $R \equiv 0$. Consequently, weak passivity is not automatic from $J = -J^\top$ and $R \succeq 0$ alone, and it requires an explicit generator inequality that balances dissipation and the Itô correction (see Corollary 5 and Proposition 8).

Several formal notions of stochastic passivity have been proposed in the literature, see for example, [4,5]. For the sake of completeness, we recall the standard definitions of *strong* and *weak* passivity.

Definition 1 (Strong and Weak Passivity) *Let $H\colon \mathcal{X} \to \mathbb{R}$ denote the Hamiltonian of a SPHS. The system interacts with its environment through port variables (u, y), where the input is $u\colon [0,\infty) \to \mathbb{R}^m$ and the output is $y\colon [0,\infty) \to \mathbb{R}^m$. Then:*

(1) The system is strongly passive [4] if, for all sample paths and all $t \geq 0$,

$$H(X_t) \leq H(X_0) + \int_0^t u(s)^\top y(s)\, ds. \tag{3}$$

(2) The system is weakly passive [4] if the inequality (3) holds in expectation:

$$\mathbb{E}\big[H(X_t)\big] \leq \mathbb{E}\big[H(X_0)\big] + \mathbb{E}\left[\int_0^t u(s)^\top y(s)\, ds\right]. \tag{4}$$

These definitions ensure that, on average, the energy within the system does not exceed the initial energy plus the cumulative power supplied through the input ports. In a stochastic setting, these passivity constraints are crucial for ensuring stability and limiting the system's response under random perturbations.

2.4 Learning a Hamiltonian from Data

A key concept in structure-preserving machine learning is learning the underlying Hamiltonian function directly from the data. This captures the physics that govern the system. In the deterministic setting, the dynamics of many physical systems can be described by *Hamilton's equations*:

$$\dot{\mathbf{q}} = \frac{\partial \mathcal{H}_\theta}{\partial \mathbf{p}}, \qquad \dot{\mathbf{p}} = -\frac{\partial \mathcal{H}_\theta}{\partial \mathbf{q}}, \tag{5}$$

where $\mathcal{H}_\theta$ is a neural network parameterization of the Hamiltonian function, and $(\mathbf{q}, \mathbf{p})$ denote the generalized coordinates and momenta.

Given observed trajectory data $\{(\mathbf{q}, \mathbf{p}, \dot{\mathbf{q}}, \dot{\mathbf{p}})\}$, one can train the neural network $\mathcal{H}_\theta$ by minimizing the discrepancy between the predicted and true time derivatives using the following loss function [8]:

$$L_{\text{HNN}} = \left\| \dot{\mathbf{q}} - \frac{\partial \mathcal{H}_\theta}{\partial \mathbf{p}} \right\|^2 + \left\| \dot{\mathbf{p}} + \frac{\partial \mathcal{H}_\theta}{\partial \mathbf{q}} \right\|^2. \tag{6}$$

This objective function (6) motivates the learned model to satisfy Hamilton's equations. It embeds the system's symplectic structure, which improves long-term predictive stability and energy conservation compared to standard regression techniques. This methodology forms the basis of the *Hamiltonian Neural Network* (HNN) framework, introduced in [8].

To extend this idea to stochastic systems, we consider the drift component of a SPHS (2), given by

$$\big[J(X_t) - R(X_t)\big] \nabla H_\theta(X_t), \tag{7}$$

while a diffusion term $\sigma(X_t)\, dW_t$ accounts for stochastic perturbations (noise), and is usually modeled as Brownian motion W_t. Given sampled trajectories of the system state X_t, we define a training objective that captures the drift structure:

$$\arg\min_\theta \left\| \dot{X}_t - \big[J(X_t) - R(X_t)\big] \nabla H_\theta(X_t) \right\|^2. \tag{8}$$

In the stochastic setting, X_t is an Itô process that does not admit a classical derivative $\frac{dX_t}{dt}$. We can interpret $\dot{X}$ in two different ways:

(1) Drift-based interpretation: $\dot{X}$ could be read as the empirical velocity computed from finite-difference estimates of trajectory increments,

$$\dot{X}_t \approx \frac{x_{t+\Delta t} - x_t}{\Delta t}, \tag{9}$$

where x_t are data points. This approach is used in [8] when analytic derivatives are unavailable. Instead, the authors used finite-difference approximations of $\dot{\mathbf{q}}, \dot{\mathbf{p}}$, where q is the position, p is the momentum and $X_t = (\mathrm{q}, \mathrm{p})$.

(2) Conditional expectation interpretation: we match the expected increment per unit time,

$$\dot{X}_t \approx \mathbb{E}\left[\frac{X_{t+\Delta} - X_t}{\Delta}\middle| X_t\right]. \tag{10}$$

This approach is used in [1] where SDEs are learned by regressing Kramers-Moyal conditional moments.

The formulation (8) is the cornerstone of our stochastic port-Hamiltonian neural network (SPH-NN) framework. It allows for the learning of energy-consistent dynamics under uncertainty by expressing the discrepancy in the drift term as an explicit loss function.

2.5 The SDE Stability Lemma

Several of our theoretical results depend on the continuity of SDE solutions with respect to small perturbations in the drift and diffusion terms. In stochastic calculus, it is well-known that the strong solution of an SDE depends continuously on its coefficients (and initial data) under standard conditions (Lipschitz continuity). If two SDEs have drift and diffusion coefficients that are close to each other (e.g., in sup norm) and share the same Brownian motion and initial state, then their solution processes will also be close to each other.

Lemma 2 *Let $\mathcal{K} \subset \mathbb{R}^n$ be compact. Let $J\colon \mathcal{K} \to \mathbb{R}^{n\times n}$ and $\sigma\colon \mathcal{K} \to \mathbb{R}^{n\times d}$ be continuous and assume $J(x) = -J(x)^\top$ for all $x \in \mathcal{K}$. Let $H \in C^2(\mathbb{R}^n)$. Assume there exist $r \in \mathbb{N}$ and a continuous map $D\colon \mathcal{K} \to \mathbb{R}^{r\times n}$ and define $R(x) = D(x)^\top D(x)$ for $x \in \mathcal{K}$. Assume the activation G is l-finite in the sense of [9] for some $l \in \mathbb{N}$ with $l > 2$. Throughout, $\|\cdot\|$ denotes the Euclidean norm for vectors and the induced operator norm for matrices, and $\|\cdot\|_F$ denotes the Frobenius norm.*

Then, for any $\varepsilon > 0$ there exist feedforward neural networks with activation G

$$\hat{A}\colon \mathbb{R}^n \to \mathbb{R}^{n\times n}, \quad \hat{D}\colon \mathbb{R}^n \to \mathbb{R}^{r\times n},$$
$$\hat{\sigma}\colon \mathbb{R}^n \to \mathbb{R}^{n\times d}, \quad \hat{H}\colon \mathbb{R}^n \to \mathbb{R},$$

such that, defining $\hat{J}(x) = \hat{A}(x) - \hat{A}(x)^\top$ and $\hat{R}(x) = \hat{D}(x)^\top \hat{D}(x)$, we have:

(1) $\hat{J}(x) = -\hat{J}(x)^\top$ and $\hat{R}(x) \succeq 0$ for all $x \in \mathbb{R}^n$.
(2)

$$\sup_{x\in\mathcal{K}} \Big(\|\hat{J}(x) - J(x)\| + \|\hat{R}(x) - R(x)\| + \|\hat{\sigma}(x) - \sigma(x)\| \Big) + \|\hat{H} - H\|_{C^2(\mathcal{K})} \le \varepsilon,$$

where

$$\|\hat{H} - H\|_{C^2(\mathcal{K})} = \sup_{x\in\mathcal{K}} |\hat{H}(x) - H(x)| + \sup_{x\in\mathcal{K}} \|\nabla\hat{H}(x) - \nabla H(x)\| + \sup_{x\in\mathcal{K}} \|\nabla^2\hat{H}(x) - \nabla^2 H(x)\|.$$

PROOF. Since $\mathcal{K}$ is compact it is closed, and each scalar entry of J, D, σ is bounded on $\mathcal{K}$. By the Tietze extension theorem applied to each scalar entry (componentwise), extend J, D, σ from $\mathcal{K}$ to continuous maps on $\mathbb{R}^n$ (still denoted J, D, σ). Use the norm bounds

$$\|M\| \le \|M\|_F \le \sqrt{pq}\max_{i,j}|M_{ij}| \quad \text{for } M \in \mathbb{R}^{p\times q}.$$

Let $M_D = \sup_{x\in\mathcal{K}} \|D(x)\| < \infty$ and set

$$\eta_J = \frac{\varepsilon}{4},\ \eta_\sigma = \frac{\varepsilon}{4},\ \eta_H = \frac{\varepsilon}{4},\ \eta_D = \min\Big\{1, \frac{\varepsilon}{4(2M_D+1)}\Big\}.$$

Choose

$$\alpha_A = \frac{\eta_J}{2n},\ \alpha_D = \frac{\eta_D}{\sqrt{rn}},\ \alpha_\sigma = \frac{\eta_\sigma}{\sqrt{nd}},\ \alpha_H = \frac{\eta_H}{1+\sqrt{n}+n}.$$

By [9, Corollary 3.4] with $m = 0$, choose networks $\hat{A}$, $\hat{D}$, $\hat{\sigma}$ such that on $\mathcal{K}$,

$$\max_{i,j}\Big|\hat{A}_{ij}(x) - \frac{1}{2}J_{ij}(x)\Big| \le \alpha_A,$$
$$\max_{i,j}\big|\hat{D}_{ij}(x) - D_{ij}(x)\big| \le \alpha_D,$$
$$\max_{i,j}\big|\hat{\sigma}_{ij}(x) - \sigma_{ij}(x)\big| \le \alpha_\sigma.$$

Then for $x \in \mathcal{K}$, we have

$$\Big\|\hat{A}(x) - \frac{1}{2}J(x)\Big\| \le \Big\|\hat{A}(x) - \frac{1}{2}J(x)\Big\|_F \le n\alpha_A = \frac{\eta_J}{2},$$

$$\|\hat{D}(x) - D(x)\| \le \|\hat{D}(x) - D(x)\|_F \le \sqrt{rn}\alpha_D = \eta_D,$$
$$\|\hat{\sigma}(x) - \sigma(x)\| \le \|\hat{\sigma}(x) - \sigma(x)\|_F \le \sqrt{nd}\alpha_\sigma = \eta_\sigma.$$

By [9, Corollary 3.4] with $m = 2$, choose $\hat{H}$ such that for all multi-indices β with $|\beta| \le 2$,

$$\sup_{x\in\mathcal{K}}|\partial^\beta\hat{H}(x) - \partial^\beta H(x)| \le \alpha_H.$$

Hence

$$\sup_{x\in\mathcal{K}}|\hat{H}(x)-H(x)|\le\alpha_H,$$
$$\sup_{x\in\mathcal{K}}\|\nabla\hat{H}(x)-\nabla H(x)\|\le\sqrt{n}\alpha_H,$$
$$\sup_{x\in\mathcal{K}}\|\nabla^2\hat{H}(x)-\nabla^2 H(x)\|\le n\alpha_H,$$

so $\|\hat{H}-H\|_{C^2(\mathcal{K})}\le(1+\sqrt{n}+n)\alpha_H=\eta_H$.

Next, define $\hat{J}=\hat{A}-\hat{A}^\top$ and $\hat{R}=\hat{D}^\top\hat{D}$. Then, $\hat{J}$ is skew-symmetric and $\hat{R}\succeq 0$ on $\mathbb{R}^n$. For $x\in\mathcal{K}$, since $J(x)=-J(x)^\top$,

$$\begin{aligned}\|\hat{J}(x)-J(x)\|&=\Big\|\Big(\hat{A}(x)-\frac{1}{2}J(x)\Big)-\Big(\hat{A}(x)-\frac{1}{2}J(x)\Big)^\top\Big\|\\&\le 2\Big\|\hat{A}(x)-\frac{1}{2}J(x)\Big\|\le\eta_J.\end{aligned}$$

Also, for $x\in\mathcal{K}$,

$$\begin{aligned}\|\hat{R}(x)-R(x)\|&=\|\hat{D}(x)^\top\hat{D}(x)-D(x)^\top D(x)\|\\&\le(\|\hat{D}(x)\|+\|D(x)\|)\,\|\hat{D}(x)-D(x)\|.\end{aligned}$$

Since $\|\hat{D}(x)\|\le\|D(x)\|+\eta_D\le M_D+1$, we get

$$\|\hat{R}(x)-R(x)\|\le(2M_D+1)\eta_D\le\frac{\varepsilon}{4}.$$

Therefore,

$$\begin{aligned}&\sup_{x\in\mathcal{K}}\Big(\|\hat{J}(x)-J(x)\|+\|\hat{R}(x)-R(x)\|\\&\qquad+\|\hat{\sigma}(x)-\sigma(x)\|\Big)+\|\hat{H}-H\|_{C^2(\mathcal{K})}\\&\le\eta_J+\frac{\varepsilon}{4}+\eta_\sigma+\eta_H=\varepsilon.\end{aligned}$$

Proposition 3 *Fix $T>0$ and a compact set $\mathcal{K}\subset\mathbb{R}^n$. Let $b,\hat{b}\colon[0,T]\times\mathbb{R}^n\to\mathbb{R}^n$ be measurable and continuous in x, and let $\sigma,\hat{\sigma}\colon\mathbb{R}^n\to\mathbb{R}^{n\times d}$ be continuous. Let X, $\widehat{X}$ be strong solutions on $[0,T]$ of*

$$dX_t=b(t,X_t)\,dt+\sigma(X_t)\,dW_t,\quad X_0=x_0\in\mathcal{K},$$

$$d\widehat{X}_t=\hat{b}(t,\widehat{X}_t)\,dt+\hat{\sigma}(\widehat{X}_t)\,dW_t,\quad \widehat{X}_0=x_0,$$

driven by the same Brownian motion W. Assume:

(1) There exists $L<\infty$ such that for all $x,y\in\mathcal{K}$ and all $t\in[0,T]$,

$$|b(t,x)-b(t,y)|+\|\sigma(x)-\sigma(y)\|_F\le L|x-y|.$$

(2)

$$\alpha=\sup_{t\in[0,T]}\sup_{x\in\mathcal{K}}|b(t,x)-\hat{b}(t,x)|<\infty,$$
$$\beta=\sup_{x\in\mathcal{K}}\|\sigma(x)-\hat{\sigma}(x)\|_F<\infty.$$

Define the joint exit time

$$\tau_{\mathcal{K}}=\inf\{t\ge 0: X_t\notin\mathcal{K}\ \text{or}\ \widehat{X}_t\notin\mathcal{K}\}.$$

With $\Delta_t=X_{t\wedge\tau_{\mathcal{K}}}-\widehat{X}_{t\wedge\tau_{\mathcal{K}}}$ and $F(t)=\mathbb{E}\big[\sup_{0\le r\le t}|\Delta_r|^2\big]$, for all $t\in[0,T]$,

$$F(t)\le(4T+16)L^2\int_0^t F(s)\,ds+(4T^2+16T)(\alpha+\beta)^2,$$

hence

$$F(T)\le(4T^2+16T)\,\exp\big((4T+16)L^2T\big)(\alpha+\beta)^2.$$

In particular, for every $\varepsilon>0$,

$$\begin{aligned}&\mathbb{P}\Big(\sup_{0\le t\le T\wedge\tau_{\mathcal{K}}}|X_t-\widehat{X}_t|>\varepsilon\Big)\\&\le\frac{(4T^2+16T)\,\exp\big((4T+16)L^2T\big)}{\varepsilon^2}\times(\alpha+\beta)^2.\end{aligned}$$

PROOF. Since $X,\widehat{X}$ are continuous adapted processes and $\mathcal{K}\times\mathcal{K}$ is closed, $\tau_{\mathcal{K}}$ is a stopping time and $X_{t\wedge\tau_{\mathcal{K}}},\widehat{X}_{t\wedge\tau_{\mathcal{K}}}\in\mathcal{K}$ for all $t\in[0,T]$. In particular, the stopped integrands below are predictable and bounded, hence square-integrable, so the Itô integrals are well-defined square-integrable martingales.

For $t\in[0,T]$,

$$\begin{aligned}\Delta_t&=\int_0^{t\wedge\tau_{\mathcal{K}}}\big(b(s,X_s)-\hat{b}(s,\widehat{X}_s)\big)\,ds\\&\quad+\int_0^{t\wedge\tau_{\mathcal{K}}}\big(\sigma(X_s)-\hat{\sigma}(\widehat{X}_s)\big)\,dW_s.\end{aligned}$$

Set

$$A_t=\int_0^{t\wedge\tau_{\mathcal{K}}}\big(b(s,X_s)-\hat{b}(s,\widehat{X}_s)\big)\,ds,$$
$$M_t=\int_0^{t\wedge\tau_{\mathcal{K}}}\big(\sigma(X_s)-\hat{\sigma}(\widehat{X}_s)\big)\,dW_s,$$

so that $\Delta_t=A_t+M_t$ and

$$\sup_{0\le r\le t}|\Delta_r|^2\le 2\sup_{0\le r\le t}|A_r|^2+2\sup_{0\le r\le t}|M_r|^2.$$

For $s < \tau_{\mathcal{K}}$ we have $X_s, \widehat{X}_s \in \mathcal{K}$, so

$$\begin{aligned}|b(s,X_s) - \hat{b}(s,\widehat{X}_s)| &\le |b(s,X_s) - b(s,\widehat{X}_s)| \\ &\quad + |b(s,\widehat{X}_s) - \hat{b}(s,\widehat{X}_s)| \\ &\le L|\Delta_s| + \alpha,\end{aligned}$$

and

$$\begin{aligned}\|\sigma(X_s) - \hat{\sigma}(\widehat{X}_s)\|_F &\le \|\sigma(X_s) - \sigma(\widehat{X}_s)\|_F \\ &\quad + \|\sigma(\widehat{X}_s) - \hat{\sigma}(\widehat{X}_s)\|_F \\ &\le L|\Delta_s| + \beta.\end{aligned}$$

For the drift term, for every $r \in [0,t]$,

$$\begin{aligned}|A_r| &\le \int_0^{r\wedge\tau_{\mathcal{K}}} \big|b(s,X_s) - \hat{b}(s,\widehat{X}_s)\big|\, ds \\ &\le \int_0^{t\wedge\tau_{\mathcal{K}}} \big(L|\Delta_s| + \alpha\big)\, ds,\end{aligned}$$

so by Cauchy–Schwarz,

$$\begin{aligned}\sup_{0\le r\le t}|A_r|^2 &\le \Big(\int_0^{t\wedge\tau_{\mathcal{K}}} \big(L|\Delta_s| + \alpha\big)\, ds\Big)^2 \\ &\le t\int_0^{t\wedge\tau_{\mathcal{K}}} \big(L|\Delta_s| + \alpha\big)^2\, ds.\end{aligned}$$

Taking expectations and using $(a+b)^2 \le 2a^2 + 2b^2$ and $|\Delta_s|^2 \le \sup_{0\le u\le s}|\Delta_u|^2$ gives

$$\begin{aligned}\mathbb{E}\Big[\sup_{0\le r\le t}|A_r|^2\Big] &\le t\int_0^t \mathbb{E}\big[(L|\Delta_s| + \alpha)^2\big]\, ds \\ &\le t\int_0^t \big(2L^2F(s) + 2\alpha^2\big)\, ds \\ &\le 2TL^2\int_0^t F(s)\, ds + 2T^2\alpha^2.\end{aligned}$$

For the martingale term, Doob's L^2 maximal inequality and Itô isometry yield

$$\begin{aligned}\mathbb{E}\Big[\sup_{0\le r\le t}|M_r|^2\Big] &\le 4\,\mathbb{E}|M_t|^2 \\ &= 4\,\mathbb{E}\int_0^{t\wedge\tau_{\mathcal{K}}} \|\sigma(X_s) - \hat{\sigma}(\widehat{X}_s)\|_F^2\, ds \\ &\le 4\int_0^t \mathbb{E}\big[(L|\Delta_s| + \beta)^2\big]\, ds \\ &\le 4\int_0^t \big(2L^2F(s) + 2\beta^2\big)\, ds \\ &\le 8L^2\int_0^t F(s)\, ds + 8T\beta^2.\end{aligned}$$

Combining the two estimates, we get

$$\begin{aligned}F(t) &= \mathbb{E}\Big[\sup_{0\le r\le t}|\Delta_r|^2\Big] \\ &\le 2\mathbb{E}\Big[\sup_{0\le r\le t}|A_r|^2\Big] + 2\mathbb{E}\Big[\sup_{0\le r\le t}|M_r|^2\Big] \\ &\le (4T+16)L^2\int_0^t F(s)\, ds + 4T^2\alpha^2 + 16T\beta^2 \\ &\le (4T+16)L^2\int_0^t F(s)\, ds + (4T^2+16T)(\alpha+\beta)^2.\end{aligned}$$

Gronwall's inequality yields the bound on $F(T)$. Finally, Markov's inequality gives

$$\begin{aligned}&\mathbb{P}\Big(\sup_{0\le t\le T\wedge\tau_{\mathcal{K}}}|X_t - \widehat{X}_t| > \varepsilon\Big) \\ &\qquad = \mathbb{P}\Big(\sup_{0\le r\le T}|\Delta_r|^2 > \varepsilon^2\Big) \le \frac{F(T)}{\varepsilon^2},\end{aligned}$$

which yields the stated probability estimate.

3 Stochastic Port-Hamiltonian Neural Networks

We now formalize our proposed model: a class of neural networks–based stochastic port-Hamiltonian systems (SPHS), referred to as a stochastic port-Hamiltonian neural networks (SPH-NNs). The key idea is to use a neural network to parameterize the system's Hamiltonian. This allows us to learn energy-consistent dynamics directly from data while enforcing key port-Hamiltonian structural constraints. Weak-passivity bounds in expectation follow under explicit generator inequalities (Corollary 5).

Given a time series $\{X_t\}_{t=0}^T$, sampled at intervals with a step size Δ, we fit the drift

$$\dot{X}_t \approx \big[J(X_t) - R(X_t)\big]\,\nabla H_\theta(X_t) \tag{11}$$

by minimizing the squared error between the model prediction and a data-driven target for $\dot{X}$. We consider two targets:

(1) **Increment-Based (IB):**

$$\widehat{\dot{X}}_t^{\,\mathrm{IB}} = \frac{X_{t+\Delta} - X_t}{\Delta}, \tag{12}$$

(2) **Conditional Expectation (CE):**

$$\widehat{\dot{X}}^{\,\mathrm{CE}}(x) \approx \mathbb{E}\left[\frac{X_{t+\Delta} - X_t}{\Delta}\,\middle|\, X_t = x\right], \tag{13}$$

estimated by local regression of increments against state.

The corresponding SPH-NN objectives are

$$\mathcal{L}_{\mathrm{IB}}(\theta) = \frac{1}{T}\sum_{t=0}^{T-1}\Big\|\widehat{\dot{X}}_t^{\mathrm{IB}} - \big[J(X_t) - R(X_t)\big]\nabla H_\theta(X_t)\Big\|^2, \tag{14}$$

$$\mathcal{L}_{\mathrm{CE}}(\theta) = \frac{1}{T}\sum_{t=0}^{T-1}\Big\|\widehat{\dot{X}}^{\mathrm{CE}}(X_t) - \big[J(X_t) - R(X_t)\big]\nabla H_\theta(X_t)\Big\|^2. \tag{15}$$

We also parameterize the diffusion $B_\theta(x)$ and train it using the negative log-likelihood under an Euler-Maruyama discretization. With an Euler-Maruyama discretization, we obtain the discrete-time update

$$X_{k+1} \approx X_k + f(X_k)\,\Delta t + B(X_k)\sqrt{\Delta t}\,\xi_k, \quad \xi_k \sim \mathcal{N}(0, I),$$

such that

$$\Delta X_k := X_{k+1} - X_k \sim \mathcal{N}\big(f(X_k)\Delta t, B(X_k)B(X_k)^\top \Delta t\big),$$

with mean $\mu_\theta(X_k) = f(X_k)\Delta t$ and covariance $\Sigma_\theta(X_k) = B_\theta(X_k)B_\theta(X_k)^\top \Delta t$.

Given data (X_k, X_{k+1}), the network is trained by *negative log-likelihood* (NLL):

$$\mathcal{L}_{\mathrm{NLL}}(\theta) = -\sum_k \log \mathcal{N}\big(\Delta X_k \mid \mu_\theta(X_k), \Sigma_\theta(X_k)\big). \tag{16}$$

The NLL objective (16) is used to fit the full conditional law of the increments ΔX_k by learning the mean $\mu_\theta(x) = f_\theta(x)\Delta t$ and the covariance $\Sigma_\theta(x) = B_\theta(x)B_\theta(x)^\top \Delta t$ that best explain the data. Since the Gaussian model arises from the Euler-Maruyama scheme, it should be interpreted as an approximation of the continuous-time transition density whose fidelity improves as $\Delta t \to 0$ (under standard regularity assumptions).

The SPH-NN architecture remains the same in all experiments, except for the varying training objective. Table 1 summarizes the three resulting variants and their associated losses.

For all variants, the gradient of the loss function ∇H_θ, or the derivative of the loss function with respect to the parameters, is computed by automatic differentiation. The parameters are then optimized with Adam. The system state, $x = [q, p] \in \mathbb{R}^n$, serves as the input to the network. Here, q and p represent the generalized positions and momenta, respectively. The network architecture consists of two hidden layers, each with 64 neurons and Tanh activation functions. These are followed by a linear output layer that produces a scalar-valued Hamiltonian. The network is trained using the Adam optimizer [12], with gradients computed via automatic differentiation. In the benchmarks below, the structure-preserving formulation improves long-horizon behavior relative to an unconstrained MLP baseline.

Variant	Objective
SPH-NN-IB	Increment-based loss $\mathcal{L}_{\mathrm{IB}}$
SPH-NN-CE	Conditional-expectation loss $\mathcal{L}_{\mathrm{CE}}$
SPH-NN-NLL	Euler-Maruyama neg. log-likelihood $\mathcal{L}_{\mathrm{NLL}}$

Table 1
SPH-NN variants trained with three objectives.

In the stochastic setting with diffusion, individual trajectories neither conserve energy nor satisfy strict passivity, because random perturbations can inject or remove energy. The evolution of energy is determined by the balance between dissipation and noise-induced effects. The SPH-NN satisfies key properties similar to those of port-Hamiltonian systems. When dissipation is absent ($R(x) = 0$), the Itô correction can inject energy in expectation. In Section 4 we report the empirical mean energy over the rollout horizon $[0, T]$, where it remains bounded.

4 Experimental Results

We evaluate the proposed structure-preserving learning approach using three stochastic port-Hamiltonian benchmarks: a linear mass-spring oscillator, the Duffing oscillator, and a stochastic Van der Pol (VDP) oscillator. In Table 2, we compare a multilayer perceptron baseline (Baseline) against the three SPH-NN variants: SPH-NN-IB/CE/NLL. Unless otherwise stated, we fix the analytical interconnection matrix J (and g, when present). For mass-spring and Duffing we learn H_θ and compare it against the analytic Hamiltonian. For Van der Pol, which is not canonically Hamiltonian in 2D and for which the choice of storage is non-unique, we fix $H(X_t) = \frac{1}{2}X_t^\top I X_t$ (cf. [4]) and learn only the nonconservative term, reporting energy relative to this storage. When learning dissipation, we parameterize $R_\theta = D_\theta^\top D_\theta$. Trajectories are generated by simulating the ground-truth SDEs using the Euler-Maruyama method.

4.1 Mass-spring oscillator

Consider the mass–spring system

$$m\ddot{x} = -kx + F,$$

Table 2
One-step drift error (True-MSE) and mean absolute rollout errors on the deterministic mean dynamics over $t \in [0, 20]$ from $(q_0, p_0) = (1, 0)$. The energy error uses the ground-truth Hamiltonian H.

System	Model	True-MSE	Mean $\lvert\Delta q\rvert$	Mean $\lvert\Delta p\rvert$	Mean $\lvert\Delta H\rvert$
Mass-spring	Baseline	3.63×10^{-3}	0.259	0.269	0.557
Mass-spring	SPH-NN-IB	2.68×10^{-3}	0.065	0.049	0.065
Mass-spring	SPH-NN-CE	1.14×10^{-3}	0.059	0.057	0.011
Mass-spring	SPH-NN-NLL	2.65×10^{-3}	0.065	0.052	0.055
Duffing	Baseline	5.67×10^{-3}	0.331	0.393	0.083
Duffing	SPH-NN-IB	2.60×10^{-3}	0.018	0.017	0.005
Duffing	SPH-NN-CE	3.23×10^{-3}	0.030	0.022	0.011
Duffing	SPH-NN-NLL	2.76×10^{-3}	0.020	0.039	0.007
Van der Pol	Baseline	2.53×10^{-3}	0.247	0.330	0.349
Van der Pol	SPH-NN-IB	2.87×10^{-5}	0.037	0.051	0.053
Van der Pol	SPH-NN-CE	2.20×10^{-4}	0.011	0.017	0.018
Van der Pol	SPH-NN-NLL	1.50×10^{-4}	0.015	0.020	0.023

where x is the position of the system, m its mass, F the applied force and k the stiffness of the spring. Defining $p = m\dot{x}$ as the momentum and $q = x$. $X = (p, q)$ defines a PHS with respect to the energy

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

of the form

$$\begin{cases} \dot{X} &= J\partial_x H(X) + gF, \\ y &= g^\top \partial_x H(X), \end{cases}$$

with

$$J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad g = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \partial_x H(X) = \begin{pmatrix} kq \\ \frac{p}{m} \end{pmatrix}.$$

We can generalize the system to a SPHS in Itô form

$$\begin{pmatrix} dq_t \\ dp_t \end{pmatrix} = \begin{pmatrix} \frac{p_t}{m} - \frac{kq_t}{2m} \\ -kq_t - \frac{k}{2m}p_t + F \end{pmatrix} dt + \begin{pmatrix} \frac{p_t}{m} \\ -kq_t \end{pmatrix} dW_t,$$

where W_t is a standard Brownian motion, and q_t and p_t denote respectively the position and the momentum. In what follows we consider $F = 0$ which yields an undamped mass-spring oscillator.

Figure 1 shows that the baseline drift accumulates phase error and yields a large energy mismatch on long rollouts, while all SPH variants remain close to the invariant orbit.

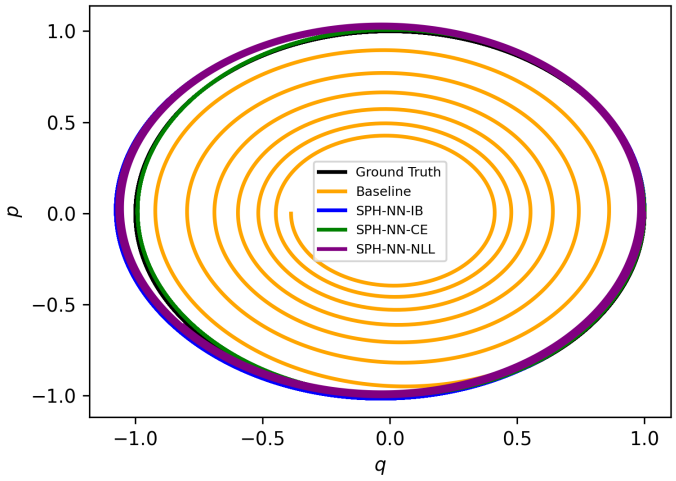

Fig. 1. Phase space - Mass spring oscillator.

4.2 Duffing oscillator

Consider the undamped Duffing system

$$\ddot{q} = q - q^3 + F,$$

where q is the position and F the applied force. Defining $p = \dot{q}$ as the momentum and $X = (q, p)$, X defines a PHS with respect to the energy

$$H(q, p) = \frac{1}{2}p^2 - \frac{1}{2}q^2 + \frac{1}{4}q^4,$$

of the form

$$\begin{cases} \dot{X} &= J\partial_x H(X) + gF, \\ y &= g^\top \partial_x H(X) \end{cases}$$

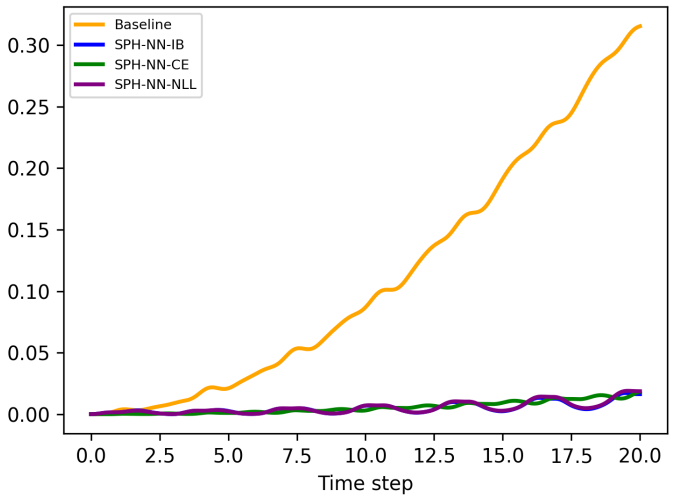

Fig. 2. Mean squared error of the rollout - Mass-spring oscillator: the baseline error grows quickly, while the SPH methods keep the error low and stable.

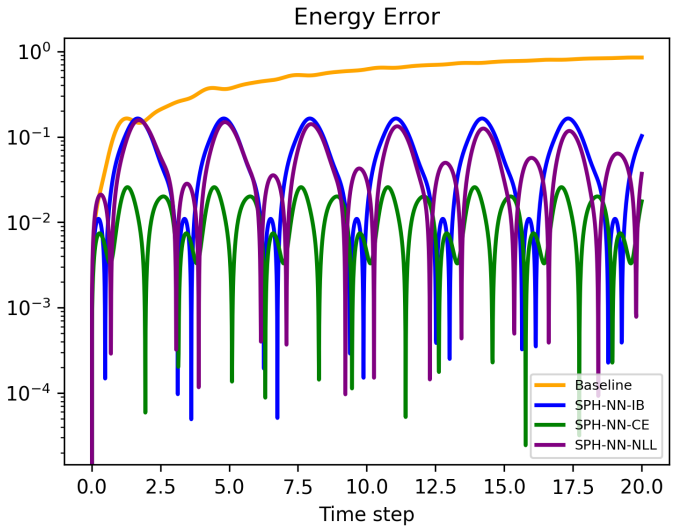

Fig. 3. Energy error - Mass-spring oscillator: SPH-NN models significantly reduce energy error compared to the baseline, with the CE loss achieving the most accurate and stable energy behavior over time.

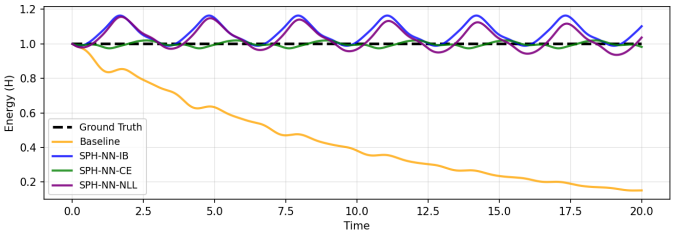

Fig. 4. Energy evolution - Mass-spring oscillator: the baseline loses energy steadily, while SPH models keep energy close to the true constant value.

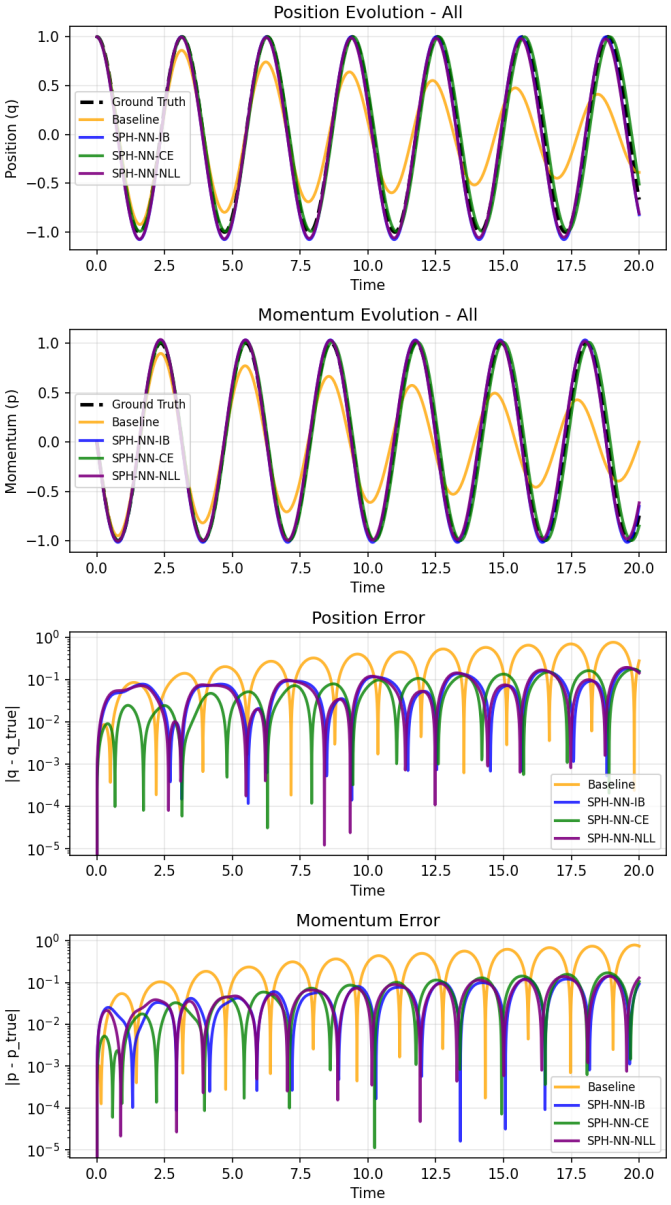

Fig. 5. Position and momentum time evolution - Mass-spring oscillator: the baseline drifts and loses amplitude, while SPH methods track the true oscillation and keep errors much smaller.

with

$$J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad g = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \partial_x H(X) = \begin{pmatrix} -q + q^3 \\ p \end{pmatrix}.$$

We can generalize the system to a SPHS in Itô form

$$\begin{pmatrix} dq_t \\ dp_t \end{pmatrix} = \begin{pmatrix} p_t \\ q_t - q_t^3 + F \end{pmatrix} dt + \sigma \, dW_t,$$

where W_t is a standard two dimensional Brownian motion.

Figure 6 shows that SPH variants preserve the correct phase-space geometry and reduce long-horizon error growth. Table 2 indicates that SPH-NN-IB is the most accurate in this benchmark, reducing mean rollout errors by more than an order of magnitude relative to the baseline and reducing mean energy error from 0.083 to 0.005.

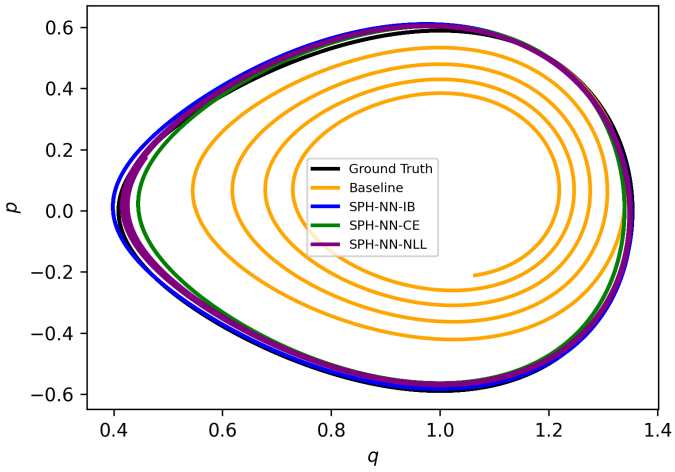

Fig. 6. Phase space - Duffing oscillator: the baseline spirals inward, while SPH methods stay close to the true closed trajectory.

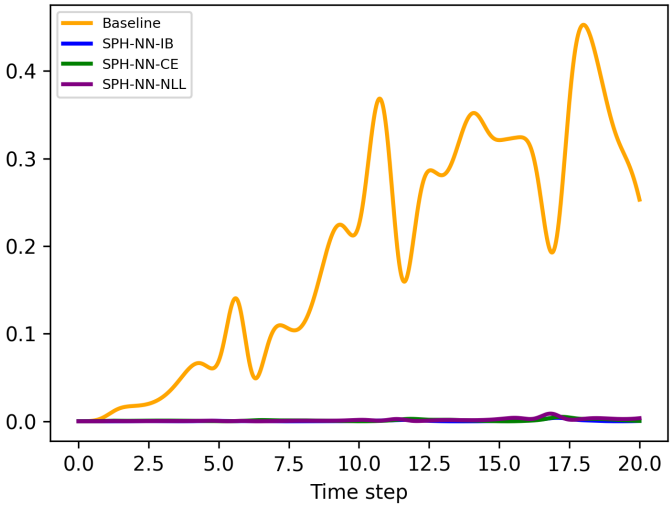

Fig. 7. Mean squared error of the rollout - Duffing oscillator.

4.3 Van der Pol oscillator

In its standard 2D form, the Van der Pol oscillator is not conservative (hence not canonical Hamiltonian), but it can be written in a (stochastic) port-Hamiltonian form by choosing an energy-like storage function H and encoding the nonconservative dynamics in the dissipation

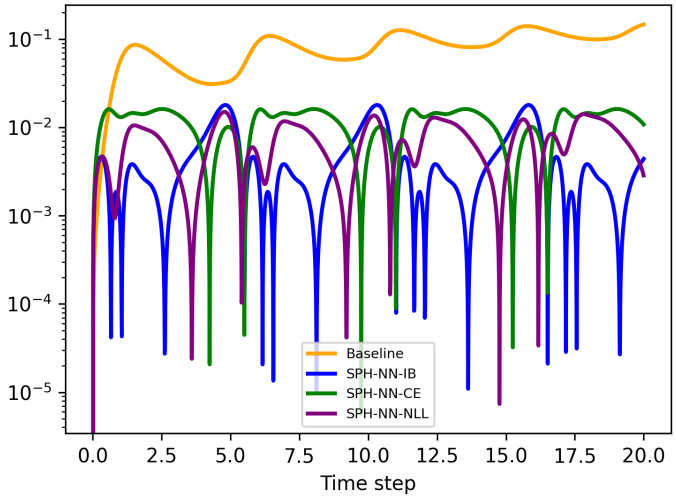

Fig. 8. Energy error - Duffing oscillator.

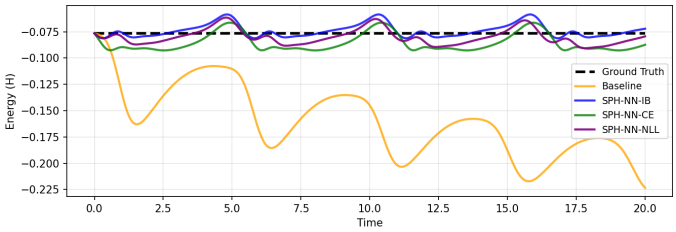

Fig. 9. Energy evolution - Duffing oscillator.

term R. Consider a stochastic Van der Pol oscillator [4]

$$\begin{cases} dx_1(t) &= x_2(t)\,dt \\ dx_2(t) &= \Big(\mu\big(1-x_1(t)^2\big)x_2(t) - x_1(t) \\ &\qquad +\tfrac{1}{2}\xi(x_2(t))\xi'(x_2(t))\Big)\,dt + \xi(x_2(t))\,dW_t. \end{cases}$$

Here $x(t) = (x_1(t), x_2(t))^\top \in \mathbb{R}^2$ is the state, $t \geq 0$ is time, $\mu \in \mathbb{R}$ is the van der Pol parameter, W_t is a one-dimensional Brownian motion, $\xi\colon \mathbb{R} \to \mathbb{R}$ is the diffusion coefficient (noise amplitude) acting on the second component.

Since the Van der Pol oscillator is neither conservative nor canonically Hamiltonian in two dimensions, the storage function used in the port-Hamiltonian representation is not uniquely defined. Following [4], we fix:

$$H(X_t) = \frac{1}{2} X_t^\top I X_t,$$

where X_t is the state variable and I is the 2×2 identity matrix.

Figure 11 shows that the SPH variant learns the limit cycle and reduces long-horizon mismatch, whereas the baseline has phase and amplitude errors.

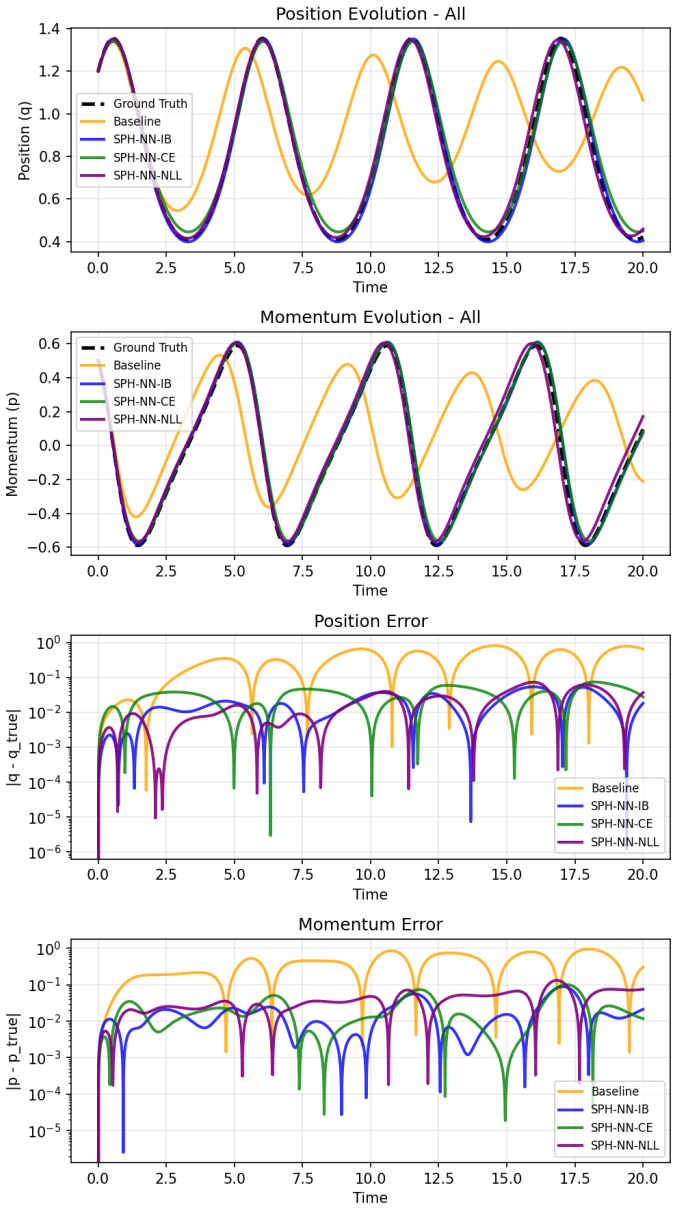

Fig. 10. Position and momentum evolution - Duffing oscillator.

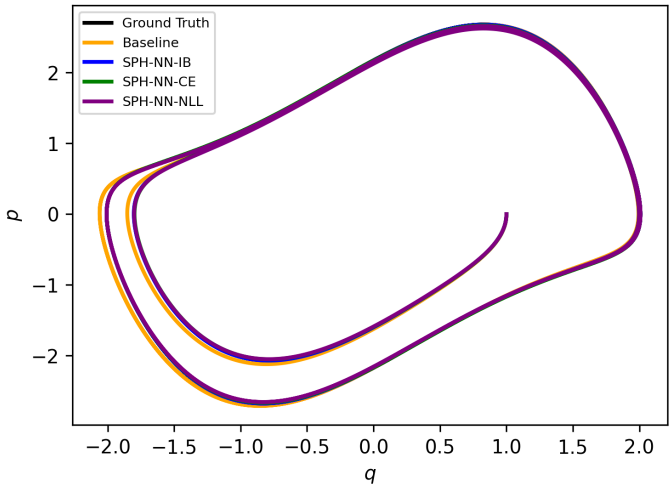

Fig. 11. Phase space - Van der Pol.

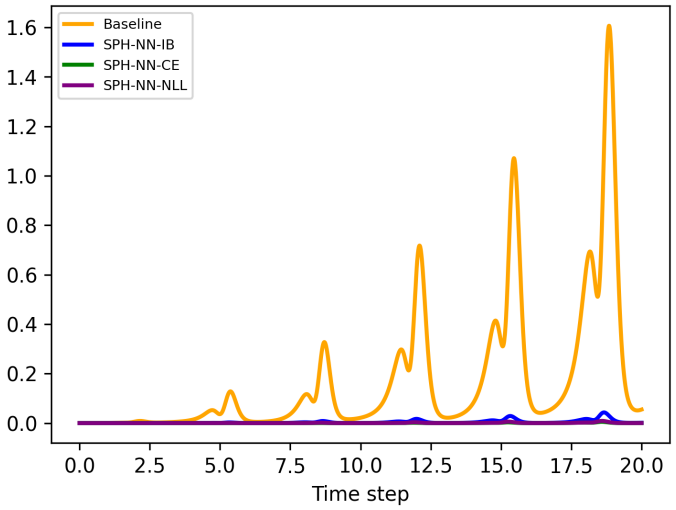

Fig. 12. Mean squared error of the rollout - Van der Pol: the baseline error explodes with large spikes, while SPH methods keep the error low and stable across the rollout.

5 Universal Approximation Theorem of SPHS

In this section, we prove an approximation theorem on compact domains and finite horizons. More precisely, for any compact set $\mathcal{K}$ and $T > 0$, the SPH-NN parameterization can approximate the coefficients of a given SPHS uniformly on $\mathcal{K}$ (including C^2-accuracy of the Hamiltonian). Moreover, under a natural coupling driven by the same Brownian motion, the corresponding solution processes remain close up to the joint exit time from $\mathcal{K}$ (Theorem 4). We then derive weak-passivity bounds in expectation via generator inequalities (Corollary 5).

Theorem 4 (UAT for SPHS) *Fix $T > 0$. Let $\mathcal{K} \subset \mathbb{R}^n$ and $U \subset \mathbb{R}^m$ be compact, and let $u \in C([0,T], U)$. Let $g\colon \mathbb{R}^n \to \mathbb{R}^{n\times m}$ be continuous, locally Lipschitz and assumed known. Let $J, R\colon \mathbb{R}^n \to \mathbb{R}^{n\times n}$ and $\sigma\colon \mathbb{R}^n \to \mathbb{R}^{n\times d}$ be continuous, locally Lipschitz, and let $H \in C^2(\mathbb{R}^n)$.*

Assume that for all $x \in \mathcal{K}$,

$$J(x) = -J(x)^\top,\ R(x) = R(x)^\top \succeq 0,$$

and assume there exist $r \in \mathbb{N}$ and a continuous map $D\colon \mathcal{K} \to \mathbb{R}^{r\times n}$ such that $R(x) = D(x)^\top D(x)$ for all $x \in \mathcal{K}$. Assume the activation G is l-finite in the sense of [9] for some $l \in \mathbb{N}$ with $l > 2$.

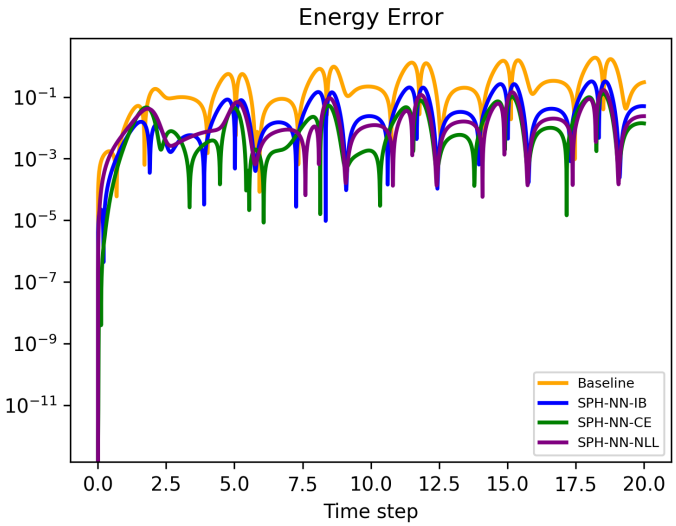

Fig. 13. Energy error - Van der Pol oscillator.

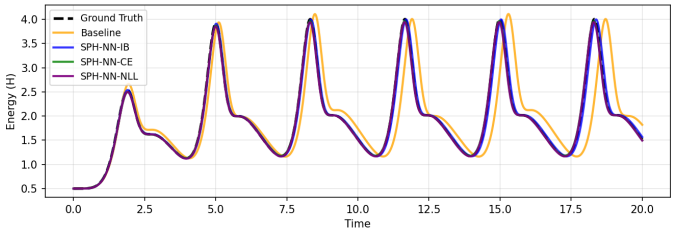

Fig. 14. Energy evolution - Van der Pol oscillator: SPH methods match the true energy oscillations closely, while the baseline slowly drifts and becomes misaligned.

Let X be a strong solution on $[0,T]$ of

$$dX_t = \Big(\big[J(X_t) - R(X_t)\big]\nabla H(X_t) + g(X_t)u_t\Big)\,dt + \sigma(X_t)\,dW_t, \qquad X_0 = x_0 \in \mathcal{K}.$$

Then, for any $\varepsilon > 0$ and $\delta > 0$ there exist feedforward neural networks

$$\hat{A}\colon \mathbb{R}^n \to \mathbb{R}^{n\times n}, \quad \hat{D}\colon \mathbb{R}^n \to \mathbb{R}^{r\times n},$$
$$\hat{H}\colon \mathbb{R}^n \to \mathbb{R}, \quad \hat{\sigma}\colon \mathbb{R}^n \to \mathbb{R}^{n\times d},$$

and defining

$$\mathcal{N}_\theta^J(x) = \hat{A}(x) - \hat{A}(x)^\top, \quad \mathcal{N}_\theta^R(x) = \hat{D}(x)^\top \hat{D}(x),$$
$$\mathcal{N}_\theta^H(x) = \hat{H}(x), \quad \mathcal{N}_\theta^\sigma(x) = \hat{\sigma}(x).$$

The following hold:

(1) $\mathcal{N}_\theta^J(x) = -(\mathcal{N}_\theta^J(x))^\top$ and $\mathcal{N}_\theta^R(x) \succeq 0$ for all $x \in \mathbb{R}^n$, and

$$\sup_{x\in\mathcal{K}} \Big(\|\mathcal{N}_\theta^J(x) - J(x)\| + \|\mathcal{N}_\theta^R(x) - R(x)\| + \|\mathcal{N}_\theta^\sigma(x) - \sigma(x)\|\Big) + \|\mathcal{N}_\theta^H - H\|_{C^2(\mathcal{K})} \le \varepsilon.$$

(2) Let $\widehat{X}$ be a strong solution on $[0,T]$ of

$$d\widehat{X}_t = \big((\mathcal{N}_\theta^J(\widehat{X}_t) - \mathcal{N}_\theta^R(\widehat{X}_t))\nabla\mathcal{N}_\theta^H(\widehat{X}_t) + g(\widehat{X}_t)u_t\big)\,dt + \mathcal{N}_\theta^\sigma(\widehat{X}_t)\,dW_t, \quad \widehat{X}_0 = x_0,$$

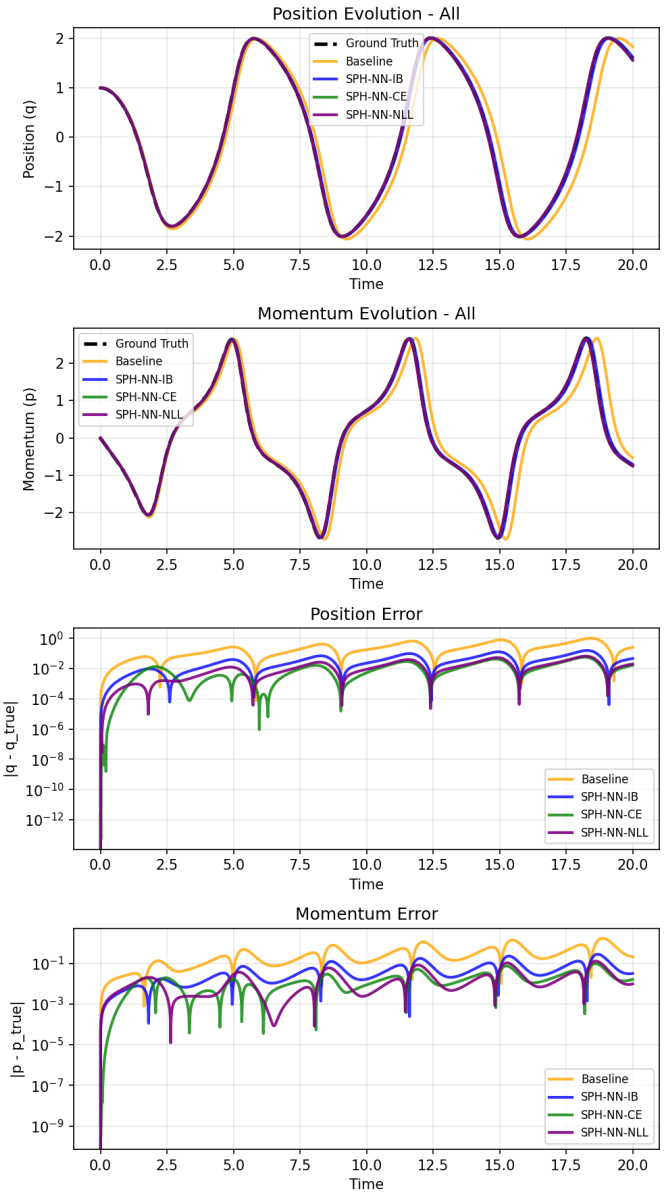

Fig. 15. Position and momentum evolution - Van der Pol oscillator.

driven by the same W, and define

$$\tau_{\mathcal{K}} = \inf\{t \geq 0 \colon X_t \notin \mathcal{K} \text{ or } \widehat{X}_t \notin \mathcal{K}\}.$$

Then we have

$$\mathbb{P}\Big(\sup_{0\leq t\leq T\wedge\tau_{\mathcal{K}}} |X_t - \widehat{X}_t| \leq \varepsilon\Big) \geq 1-\delta.$$

PROOF. Choose $\eta > 0$ such that $\eta \leq \varepsilon$ and $\eta \leq 1$ (specified below). Apply Lemma 2 on $\mathcal{K}$ with tolerance η to (J, D, σ, H) to obtain networks $\hat{A}$, $\hat{D}$, $\hat{H}$, $\hat{\sigma}$ such that item (1) holds with η in place of ε, hence with ε since $\eta \leq \varepsilon$.

Define

$$b(t,x) = (J(x) - R(x))\nabla H(x) + g(x)u_t,$$
$$\hat{b}(t,x) = (\mathcal{N}_\theta^J(x) - \mathcal{N}_\theta^R(x))\nabla\mathcal{N}_\theta^H(x) + g(x)u_t.$$

(Here $\|\cdot\|$ denotes the operator norm and $\|\cdot\|_F$ denotes the Frobenius norm.)

Since J, R, g, σ are locally Lipschitz and ∇H is locally Lipschitz, their restrictions to the compact set $\mathcal{K}$ are Lipschitz and bounded. Since u is bounded on $[0,T]$, there exists $L < \infty$ such that for all $x, y \in \mathcal{K}$ and $t \in [0,T]$,

$$|b(t,x) - b(t,y)| + \|\sigma(x) - \sigma(y)\|_F \leq L|x-y|.$$

Set

$$\alpha = \sup_{t\in[0,T]}\sup_{x\in\mathcal{K}} |b(t,x) - \hat{b}(t,x)|,\ \beta = \sup_{x\in\mathcal{K}} \|\sigma(x) - \mathcal{N}_\theta^\sigma(x)\|_F.$$

Let

$$M_{JR} = \sup_{x\in\mathcal{K}} \|J(x) - R(x)\|,\ M_1 = \sup_{x\in\mathcal{K}} \|\nabla H(x)\|.$$

For $x \in \mathcal{K}$, we obtain

$$\begin{aligned}|b(t,x) - \hat{b}(t,x)| &\leq \|J(x) - R(x)\|\|\nabla H(x) - \nabla\mathcal{N}_\theta^H(x)\| \\ &\quad + \big(\|\mathcal{N}_\theta^J(x) - J(x)\| + \|\mathcal{N}_\theta^R(x) - R(x)\|\big) \\ &\quad \times \|\nabla\mathcal{N}_\theta^H(x)\|.\end{aligned}$$

By Lemma 2 (with tolerance η), on $\mathcal{K}$ we have

$$\|\nabla H(x) - \nabla\mathcal{N}_\theta^H(x)\| \leq \eta,$$
$$\|\mathcal{N}_\theta^J(x) - J(x)\| + \|\mathcal{N}_\theta^R(x) - R(x)\| \leq \eta.$$

Also, using Lemma 2 and $\eta \leq 1$ gives $\|\nabla\mathcal{N}_\theta^H(x)\| \leq \|\nabla H(x)\| + \eta \leq M_1 + 1$, and hence

$$\alpha \leq (M_{JR} + M_1 + 1)\eta.$$

Moreover, Lemma 2 gives $\|\sigma(x) - \mathcal{N}_\theta^\sigma(x)\| \leq \eta$ on $\mathcal{K}$, and $\|A\|_F \leq \sqrt{d}\|A\|$ for $A \in \mathbb{R}^{n\times d}$, so $\beta \leq \sqrt{d}\eta$. Let

$$C_1 = M_{JR} + M_1 + 1 + \sqrt{d},$$

so $\alpha + \beta \leq C_1\eta$.

We apply Proposition 3 to get

$$\begin{aligned}&\mathbb{P}\Big(\sup_{0\leq t\leq T\wedge\tau_{\mathcal{K}}} |X_t - \widehat{X}_t| > \varepsilon\Big) \\ &\leq \frac{(4T^2 + 16T)\exp((4T+16)L^2T)}{\varepsilon^2} \times (C_1\eta)^2.\end{aligned}$$

Choose $\eta > 0$ such that $\eta \leq \varepsilon$, $\eta \leq 1$, and

$$(4T^2 + 16T)\exp((4T+16)L^2T)(C_1\eta)^2 \leq \delta\varepsilon^2.$$

Then the probability above is at most δ, which gives item (2).

Corollary 5 (Localized weak passivity) *Assume the hypotheses of Theorem 4. Define*

$$y(x) = g(x)^\top\nabla H(x)$$

and assume there exists $c_0 \geq 0$ such that for all $x \in \mathcal{K}$,

$$r(x) = \frac{1}{2}\operatorname{Tr}(\sigma(x)\sigma(x)^\top\nabla^2 H(x)) - \nabla H(x)^\top R(x)\nabla H(x) \leq c_0.$$

Fix $\varepsilon > 0$ and choose the networks from Theorem 4 with a coefficient tolerance $\eta > 0$ small enough. Next, define

$$\hat{y}(x) = g(x)^\top\nabla\mathcal{N}_\theta^H(x),$$

$$\begin{aligned}\hat{r}(x) = \frac{1}{2}\operatorname{Tr}\Big(\mathcal{N}_\theta^\sigma(x)(\mathcal{N}_\theta^\sigma(x))^\top\nabla^2\mathcal{N}_\theta^H(x)\Big) \\ - \nabla\mathcal{N}_\theta^H(x)^\top\mathcal{N}_\theta^R(x) \ \times \nabla\mathcal{N}_\theta^H(x).\end{aligned}$$

Then:

(1) $\hat{r}(x) \leq c_0 + \varepsilon$ for all $x \in \mathcal{K}$.

(2) Let $\widehat{X}$ be the learned solution from Theorem 4 and let

$$\tau_{\mathcal{K}}^{\widehat{X}} = \inf\{t \geq 0 \colon \widehat{X}_t \notin \mathcal{K}\}.$$

Then for all $t \in [0,T]$,

$$\begin{aligned}\mathbb{E}\big[\mathcal{N}_\theta^H(\widehat{X}_{t\wedge\tau_{\mathcal{K}}^{\widehat{X}}})\big] &\leq \mathcal{N}_\theta^H(x_0) + \mathbb{E}\int_0^{t\wedge\tau_{\mathcal{K}}^{\widehat{X}}} u_s^\top\hat{y}(\widehat{X}_s)\,ds \\ &\quad + (c_0 + \varepsilon)t.\end{aligned}$$

PROOF. Let η be the coefficient tolerance on $\mathcal{K}$ from Theorem 4, so on $\mathcal{K}$,

$$\|\mathcal{N}_\theta^\sigma - \sigma\| \le \eta, \quad \|\mathcal{N}_\theta^R - R\| \le \eta,$$
$$\|\nabla\mathcal{N}_\theta^H - \nabla H\| \le \eta, \quad \|\nabla^2\mathcal{N}_\theta^H - \nabla^2 H\| \le \eta.$$

Define the finite constants

$$M_\sigma = \sup_{x\in\mathcal{K}} \|\sigma(x)\|, \quad M_R = \sup_{x\in\mathcal{K}} \|R(x)\|,$$
$$M_1 = \sup_{x\in\mathcal{K}} \|\nabla H(x)\|, \quad M_2 = \sup_{x\in\mathcal{K}} \|\nabla^2 H(x)\|.$$

Then on $\mathcal{K}$,

$$\|\mathcal{N}_\theta^\sigma\| \le M_\sigma + \eta, \quad \|\mathcal{N}_\theta^R\| \le M_R + \eta,$$
$$\|\nabla\mathcal{N}_\theta^H\| \le M_1 + \eta, \quad \|\nabla^2\mathcal{N}_\theta^H\| \le M_2 + \eta.$$

Moreover,

$$\begin{aligned}\|\mathcal{N}_\theta^\sigma(\mathcal{N}_\theta^\sigma)^\top - \sigma\sigma^\top\| &\le (\|\mathcal{N}_\theta^\sigma\| + \|\sigma\|)\|\mathcal{N}_\theta^\sigma - \sigma\| \\ &\le (2M_\sigma + \eta)\eta \ \text{ on } \mathcal{K}.\end{aligned}$$

Using $|\operatorname{Tr}(AB)| \le n\|A\|\|B\|$, for $x \in \mathcal{K}$,

$$\begin{aligned}|\hat{r}(x) - r(x)| \le{}& \frac{1}{2}\Big|\operatorname{Tr}\big((\mathcal{N}_\theta^\sigma(\mathcal{N}_\theta^\sigma)^\top - \sigma\sigma^\top)\nabla^2\mathcal{N}_\theta^H\big)\Big| \\ &+ \frac{1}{2}\Big|\operatorname{Tr}\big(\sigma\sigma^\top(\nabla^2\mathcal{N}_\theta^H - \nabla^2 H)\big)\Big| \\ &+ \Big|\nabla\mathcal{N}_\theta^{H\top}\mathcal{N}_\theta^R\nabla\mathcal{N}_\theta^H - \nabla H^\top R\nabla H\Big|.\end{aligned}$$

Also, on $\mathcal{K}$,

$$\begin{aligned}&\Big|\nabla\mathcal{N}_\theta^{H\top}\mathcal{N}_\theta^R\nabla\mathcal{N}_\theta^H - \nabla H^\top R\nabla H\Big| \\ &\le \eta\|\nabla\mathcal{N}_\theta^H\|^2 + \|R\|\|\nabla\mathcal{N}_\theta^H - \nabla H\|\|\nabla\mathcal{N}_\theta^H + \nabla H\|.\end{aligned}$$

Combining these bounds and using $\|\nabla\mathcal{N}_\theta^H\| \le M_1 + \eta$ and $\|\nabla^2\mathcal{N}_\theta^H\| \le M_2 + \eta$ yields

$$\sup_{x\in\mathcal{K}} |\hat{r}(x) - r(x)| \le C_2\eta,$$

for some finite constant C_2 depending only on n, M_σ, M_R, M_1, M_2. Choose η small enough so that $C_2\eta \le \varepsilon$. Since $r(x) \le c_0$ on $\mathcal{K}$, we get $\hat{r}(x) \le c_0 + \varepsilon$ on $\mathcal{K}$, proving item (1).

Let $\hat{b}(t,x) = (\mathcal{N}_\theta^J(x) - \mathcal{N}_\theta^R(x))\nabla\mathcal{N}_\theta^H(x) + g(x)u_t$. For $x \in \mathcal{K}$, the generator applied to $\mathcal{N}_\theta^H$ is given by

$$\begin{aligned}(\hat{\mathcal{L}}_t[\mathcal{N}_\theta^H])(x) &= \nabla\mathcal{N}_\theta^H(x)^\top\hat{b}(t,x) \\ &\quad + \frac{1}{2}\operatorname{Tr}\big(\mathcal{N}_\theta^\sigma(x)(\mathcal{N}_\theta^\sigma(x))^\top\nabla^2\mathcal{N}_\theta^H(x)\big) \\ &= \nabla\mathcal{N}_\theta^H(x)^\top(\mathcal{N}_\theta^J(x) - \mathcal{N}_\theta^R(x))\nabla\mathcal{N}_\theta^H(x) \\ &\quad + u_t^\top\hat{y}(x) + \frac{1}{2}\operatorname{Tr}\big(\mathcal{N}_\theta^\sigma(x)(\mathcal{N}_\theta^\sigma(x))^\top\nabla^2\mathcal{N}_\theta^H(x)\big).\end{aligned}$$

Since $\mathcal{N}_\theta^J(x)$ is skew-symmetric, we have

$$\nabla\mathcal{N}_\theta^H(x)^\top\mathcal{N}_\theta^J(x)\nabla\mathcal{N}_\theta^H(x) = 0,$$

hence

$$(\hat{\mathcal{L}}_t[\mathcal{N}_\theta^H])(x) = u_t^\top\hat{y}(x) + \hat{r}(x) \le u_t^\top\hat{y}(x) + c_0 + \varepsilon \quad \text{for } x \in \mathcal{K}.$$

Apply Itô's formula to $\mathcal{N}_\theta^H(\widehat{X}_{t\wedge\tau_{\mathcal{K}}^{\widehat{X}}})$. Since the process is stopped on $\mathcal{K}$ and the coefficients are continuous, the stochastic integral has zero expectation. Taking expectations gives

$$\mathbb{E}[\mathcal{N}_\theta^H(\widehat{X}_{t\wedge\tau_{\mathcal{K}}^{\widehat{X}}})] = \mathcal{N}_\theta^H(x_0) + \mathbb{E}\int_0^{t\wedge\tau_{\mathcal{K}}^{\widehat{X}}} (\hat{\mathcal{L}}_s[\mathcal{N}_\theta^H])(\widehat{X}_s)\,ds.$$

For $s < \tau_{\mathcal{K}}^{\widehat{X}}$, $\widehat{X}_s \in \mathcal{K}$, so $(\hat{\mathcal{L}}_s[\mathcal{N}_\theta^H])(\widehat{X}_s) \le u_s^\top\hat{y}(\widehat{X}_s) + c_0 + \varepsilon$. Since $t \wedge \tau_{\mathcal{K}}^{\widehat{X}} \le t$, we obtain item (2).

Remark 6 *Corollary 5 is a localized statement: it applies to the stopped process $\widehat{X}_{t\wedge\tau_{\mathcal{K}}^{\widehat{X}}}$ and includes the additive term $(c_0 + \varepsilon)t$, which accumulates linearly in time. It provides an a priori energy/passivity control on a prescribed finite horizon and within the region $\mathcal{K}$ where the model is required to operate.*

Corollary 7 *Assume the hypotheses of Theorem 4 and use the notation of Corollary 5. Assume that there exists $\delta > 0$ such that for all $x \in \mathcal{K}$,*

$$r(x) = \frac{1}{2}\operatorname{Tr}(\sigma\sigma^\top\nabla^2 H)(x) - \nabla H(x)^\top R(x)\nabla H(x) \le -\delta.$$

Choose the networks from Theorem 4 with tolerance $\eta > 0$ small enough so that $\sup_{x\in\mathcal{K}} |\hat{r}(x) - r(x)| \le \frac{\delta}{2}$ (for example $C_2\eta \le \frac{\delta}{2}$ with C_2 as in the proof of Corollary 5). Then $\hat{r}(x) \le 0$ on $\mathcal{K}$ and, for all $t \in [0,T]$,

$$\mathbb{E}[\mathcal{N}_\theta^H(\widehat{X}_{t\wedge\tau_{\mathcal{K}}^{\widehat{X}}})] \le \mathcal{N}_\theta^H(x_0) + \mathbb{E}\int_0^{t\wedge\tau_{\mathcal{K}}^{\widehat{X}}} u_s^\top\hat{y}(\widehat{X}_s)\,ds.$$

In particular, the learned SPHS is weakly passive on $\mathcal{K}$ (up to the exit time) in the sense of Definition 1.

PROOF. By Theorem 4 and the estimate in the proof of Corollary 5, choose $\eta > 0$ such that $\sup_{x\in\mathcal{K}} |\hat{r}(x) - r(x)| \le \frac{\delta}{2}$. Since $r(x) \le -\delta$ on $\mathcal{K}$, we obtain $\hat{r}(x) \le -\frac{\delta}{2} \le 0$ for all $x \in \mathcal{K}$. Apply Itô's formula to $\mathcal{N}_\theta^H(\widehat{X}_{t\wedge\tau_{\mathcal{K}}^{\widehat{X}}})$. As in Corollary 5, the local martingale term has zero expectation, hence

$$\mathbb{E}\big[\mathcal{N}_\theta^H(\widehat{X}_{t\wedge\tau_{\mathcal{K}}^{\widehat{X}}})\big] = \mathcal{N}_\theta^H(x_0) + \mathbb{E}\int_0^{t\wedge\tau_{\mathcal{K}}^{\widehat{X}}} u_s^\top \hat{y}(\widehat{X}_s)\,ds + \mathbb{E}\int_0^{t\wedge\tau_{\mathcal{K}}^{\widehat{X}}} \hat{r}(\widehat{X}_s)\,ds.$$

For $s \le \tau_{\mathcal{K}}^{\widehat{X}}$ we have $\widehat{X}_s \in \mathcal{K}$, hence $\hat{r}(\widehat{X}_s) \le 0$ and therefore the last term is ≤ 0. Dropping it proves the claim.

Proposition 8 *Assume the hypotheses of Theorem 4 and let $\widehat{X}$ denote the corresponding learned SPHS solution on $[0,T]$ with deterministic initial condition x_0. Let*

$$\begin{cases} \hat{y}(x) & := g(x)^\top \nabla\mathcal{N}_\theta^H(x), \\ \hat{r}(x) & := \frac{1}{2}\operatorname{Tr}\Big(\mathcal{N}_\theta^\sigma(x)\big(\mathcal{N}_\theta^\sigma(x)\big)^\top \nabla^2\mathcal{N}_\theta^H(x)\Big) \\ & \quad -\nabla\mathcal{N}_\theta^H(x)^\top \mathcal{N}_\theta^R(x)\nabla\mathcal{N}_\theta^H(x) \end{cases}$$

Assume that $\mathcal{N}_\theta^H \in C^2(\mathbb{R}^n)$ and that the following two conditions hold:

(i) $\hat{r}(x) \le 0$ for all $x \in \mathbb{R}^n$.
(ii) For every $t \in [0,T]$,

$$\begin{cases} \mathbb{E}\int_0^t \Big\|\big(\nabla\mathcal{N}_\theta^H(\widehat{X}_s)\big)^\top \mathcal{N}_\theta^\sigma(\widehat{X}_s)\Big\|^2 ds < \infty, \\ \mathbb{E}\int_0^t \big|u_s^\top \hat{y}(\widehat{X}_s)\big|\,ds < \infty, \end{cases}$$

and $\mathbb{E}\big[|\mathcal{N}_\theta^H(\widehat{X}_t)|\big] < \infty$.

Then, for all $t \in [0,T]$,

$$\mathbb{E}\big[\mathcal{N}_\theta^H(\widehat{X}_t)\big] \le \mathcal{N}_\theta^H(x_0) + \mathbb{E}\int_0^t u_s^\top \hat{y}(\widehat{X}_s)\,ds.$$

In particular, the learned SPHS is weakly passive without localization (on $[0,T]$) in the sense of Definition 1 (with storage $\mathcal{N}_\theta^H$ and output $\hat{y}$). If the assumptions hold for every finite horizon $T > 0$, then the same inequality holds for all $t \ge 0$.

PROOF. By Itô's formula and the SPHS structure,

$$\mathcal{N}_\theta^H(\widehat{X}_t) = \mathcal{N}_\theta^H(x_0) + \int_0^t u_s^\top \hat{y}(\widehat{X}_s)\,ds + \int_0^t \hat{r}(\widehat{X}_s)\,ds + M_t,$$

where

$$M_t = \int_0^t \big(\nabla\mathcal{N}_\theta^H(\widehat{X}_s)\big)^\top \mathcal{N}_\theta^\sigma(\widehat{X}_s)\,dW_s.$$

Assumption (ii) implies that M_t is a square-integrable martingale, hence $\mathbb{E}[M_t] = 0$ and $\mathbb{E}[|M_t|] \le (\mathbb{E}[M_t^2])^{\frac{1}{2}} < \infty$. Moreover, by (ii), $\mathbb{E}\int_0^t |u_s^\top \hat{y}(\widehat{X}_s)|\,ds < \infty$ and $\mathbb{E}[|\mathcal{N}_\theta^H(\widehat{X}_t)|] < \infty$, so

$$\mathbb{E}\Big[\Big|\int_0^t \hat{r}(\widehat{X}_s)\,ds\Big|\Big] \le \mathbb{E}[|\mathcal{N}_\theta^H(\widehat{X}_t)|] + |\mathcal{N}_\theta^H(x_0)| + \mathbb{E}\int_0^t |u_s^\top \hat{y}(\widehat{X}_s)|\,ds + \mathbb{E}[|M_t|] < \infty.$$

Taking expectations in the Itô identity yields

$$\mathbb{E}[\mathcal{N}_\theta^H(\widehat{X}_t)] = \mathcal{N}_\theta^H(x_0) + \mathbb{E}\int_0^t u_s^\top \hat{y}(\widehat{X}_s)\,ds + \mathbb{E}\int_0^t \hat{r}(\widehat{X}_s)\,ds.$$

Using (i) we have $\hat{r}(\widehat{X}_s) \le 0$ a.s. for all s, hence $\int_0^t \hat{r}(\widehat{X}_s)\,ds \le 0$ a.s. and therefore $\mathbb{E}\int_0^t \hat{r}(\widehat{X}_s)\,ds \le 0$. Discarding this non-positive term yields the claim.

6 Discussion

Numerical evidence indicates that incorporating the complete port-Hamiltonian structure into the learning architecture provides clear practical benefits. Over extended periods, energy remains within one order of magnitude of the actual value, phase-space trajectories remain on the correct invariant manifold, and the model is less susceptible to noise than an unconstrained multilayer perceptron baseline.

However, several challenges remain. Training costs increase disproportionately with state dimension because each iteration requires Hessian-vector products. Scalable parameterizations or sparsity priors are necessary for high-dimensional systems. Additionally, the present work assumes full-state observations. Extending the formulation to partial and irregular data will likely necessitate a latent SDE approach. Finally, although the model is less fragile than unconstrained baselines, careful tuning of learning rates and loss weights is still necessary. This suggests that automatic hyperparameter scheduling is a promising area for future research.

7 Conclusion

This paper introduces a unified framework for stochastic port-Hamiltonian neural networks, combining data-

driven learning with a structure-preserving parameterization of stochastic port-Hamiltonian systems. The Hamiltonian, interconnection, dissipation, and diffusion components can, in principle, be learned from data. In our experiments, we fix the analytical interconnection J (and g, if applicable) and focus on learning H_θ, though the theory covers the general case in which all coefficients are learned. A universal approximation theorem demonstrates the expressive power of the architecture. Experiments on noisy mass-spring, Duffing, and Van der Pol oscillators demonstrate an order-of-magnitude improvement in long-term accuracy compared to a baseline multilayer perceptron.

These results suggest that combining geometric structure with neural function approximation is a viable approach for creating reliable, data-driven models of complex physical systems subject to uncertainty. These results also motivate future work on scalable variants, partial-observation settings, and integration with feedback control.

Acknowledgements

M. Ehrhardt acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 531152215 – CRC 1701.

References

[1] L. Boninsegna, F. Nüske, and C. Clementi. Sparse learning of stochastic dynamical equations. *The Journal of Chemical Physics*, 148(24), 2018.

[2] T. Breiten and P. Schulze. Structure-preserving linear quadratic Gaussian balanced truncation for port-Hamiltonian descriptor systems. *Linear Algebra and its Applications*, 704:146–191, 2025.

[3] K. Cherifi. An overview on recent machine learning techniques for Port Hamiltonian systems. *Physica D: Nonlinear Phenomena*, 411:132620, 2020.

[4] F. Cordoni, L. Di Persio, and R. Muradore. Stochastic port-Hamiltonian systems. *Journal of Nonlinear Science*, 32(6):91, 2022.

[5] F. Cordoni, L. Di Persio, and R. Muradore. Weak energy shaping for stochastic controlled port-Hamiltonian systems. *SIAM Journal on Control and Optimization*, 61(5):2902–2926, 2023.

[6] L. Di Persio, M. Ehrhardt, Y. Outaleb, and S. Rizzotto. Port-Hamiltonian neural networks: From theory to simulation of interconnected stochastic systems. *arXiv preprint arXiv:2509.06674*, September 2025.

[7] E. Dupont, A. Doucet, and Y. W. Teh. Augmented neural ODEs. *Advances in Neural Information Processing Systems*, 32, 2019.

[8] S. Greydanus, M. Dzamba, and J. Yosinski. Hamiltonian neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[9] K. Hornik, M. Stinchcombe, and H. White. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural networks*, 3(5):551–560, 1990.

[10] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

[11] P. Kidger, J. Morrill, J. Foster, and T. Lyons. Neural controlled differential equations for irregular time series. *Advances in Neural Information Processing Systems*, 33:6696–6707, 2020.

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412(6), 2014.

[13] L. Kong, J. Sun, and C. Zhang. SDE-Net: Equipping deep neural networks with uncertainty estimates. *arXiv preprint arXiv:2008.10546*, 2020.

[14] J. O'Leary, J. A. Paulson, and A. Mesbah. Stochastic physics-informed neural ordinary differential equations. *Journal of Computational Physics*, 468:111466, 2022.

[15] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

[16] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[17] A. Van Der Schaft and D. Jeltsema. Port-Hamiltonian systems theory: An introductory overview. *Foundations and Trends® in Systems and Control*, 1(2-3):173–378, 2014.

[18] X. Zhang, W. Wei, Z. Zhang, L. Zhang, and W. Li. Milstein-driven neural stochastic differential equation model with uncertainty estimates. *Pattern Recognition Letters*, 174:71–77, 2023.