The Learned Hamiltonian Is a Perturbation: A KAM-Theoretic Account of Long-Horizon Error in Hamiltonian Neural Networks

Illia

Erlangen, Germany | *draft manuscript*

June 15, 2026

Abstract

A Hamiltonian neural network (HNN) is trained to match a vector field, yet it is used to integrate trajectories over long horizons, where it is observed to be far more stable than a black-box integrator. The standard explanation—"HNNs conserve energy"—is incomplete: energy conservation neither bounds the trajectory error nor explains its structure, and the network does not in fact conserve the true energy. We start from a sharper observation: the learned flow is the *exact* Hamiltonian flow of a learned Hamiltonian $H_\theta = H + \mathcal{R}$. Long-horizon fidelity is therefore not a question of vector-field accuracy but of how the phase portrait of H_θ relates to that of H—precisely the perturbation theory of Hamiltonian systems, i.e. Kolmogorov–Arnold–Moser (KAM) and Nekhoroshev theory. This reframing makes falsifiable, quantitative predictions that an energy-conservation story cannot: (i) rollout error is not uniform but *localizes at resonances*, where the network manufactures islands and chaos absent from an integrable target; (ii) the width of these spurious islands obeys a $\sqrt{\varepsilon}$ law in the training residual; (iii) the HNN advantage is *confinement*, diagnosed by a bounded-versus-secular error-growth law rather than instantaneous accuracy; and (iv) KAM's non-degeneracy hypothesis predicts that HNNs are *more* fragile on "simple" isochronous systems such as the harmonic oscillator. We confirm (i)–(iii) on an integrable twist system whose true dynamics contains *no* chaos, so that every island the network produces is unambiguously its own error. We close by noting that the L^2 vector-field loss is blind to the analytic-norm and resonant structure that KAM says controls long-horizon fidelity, and propose two KAM-informed training objectives.

1 Introduction

Structure-preserving neural networks—Hamiltonian neural networks (HNNs) [12], symplectic networks [13], and their relatives—are now a standard tool for learning dynamics. Their empirical selling point is *long-horizon stability*: where a generic regression model fed to an integrator drifts off the energy surface within a few periods, an HNN produces bounded, recurrent rollouts over thousands of periods. The usual rationale is that the network "conserves energy."

This rationale is incomplete in three ways. First, energy conservation is a single scalar constraint; it does not bound the *trajectory* error, and a model can conserve energy while being qualitatively wrong about the orbit. Second, it says nothing about *where* in phase space the model fails, yet HNN error is strongly non-uniform (Fig. 2a). Third, it is not even true as stated: an HNN does not conserve the true energy H; it conserves a *different* function H_θ that it learned.

We take that third point as the starting observation. Because the learned vector field is the symplectic gradient of a scalar H_θ, the learned dynamics is the *exact* Hamiltonian flow of H_θ, and $H_\theta = H + \mathcal{R}$ is a perturbation of the true Hamiltonian. The long-horizon question—does the learned orbit stay near the true orbit, stay bounded, drift?—is then a question about a near-integrable Hamiltonian system, and the mature theory that answers such questions is KAM and Nekhoroshev theory [1, 2, 3, 5, 6].

Read through this lens, long-horizon stability is not vector-field accuracy but *preservation of the topology of the flow*, and the theory predicts a specific, structured *failure* mode rather than uniform error. Our contributions are:

- We frame an HNN as a learned near-integrable Hamiltonian system, with the training residual playing the role of the perturbation, and identify long-horizon stability with KAM torus persistence rather than with Grönwall-type vector-field bounds (§2–3).
- From this we derive four predictions—resonance localization of error, a $\sqrt{\varepsilon}$ island-width law, confinement diagnosed by error-growth law, and degeneracy-induced fragility—and confirm the first three on an integrable twist system with no intrinsic chaos (§4).

- We observe that the standard L^2 vector-field loss is in the wrong norm for long-horizon fidelity and propose analytic-norm and action-variance objectives motivated directly by KAM (§3, §3.5).
- We make explicit the connection to backward error analysis of symplectic integrators: an HNN is a *learned shadow Hamiltonian*, and the Nekhoroshev apparatus developed there transfers (§6).

2 The learned flow is Hamiltonian for a perturbed Hamiltonian

Let $z = (q, p) \in \mathbb{R}^{2d}$ with the canonical symplectic matrix $\mathbb{J} = \left(\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}\right)$. The true system has Hamiltonian H and flow generated by the vector field $X_H(z) = \mathbb{J}\nabla H(z)$. An HNN is a scalar network $H_\theta : \mathbb{R}^{2d} \to \mathbb{R}$ whose induced vector field is

$$X_{H_\theta}(z) \;=\; \mathbb{J}\nabla H_\theta(z), \tag{1}$$

trained so that $X_{H_\theta} \approx X_H$ on a sampling measure μ.

Observation 1. *The learned rollout is the exact Hamiltonian flow of H_θ. Hence it conserves H_θ exactly and preserves the symplectic form $\omega = \sum_i dq_i \wedge dp_i$ exactly, for any integration error-free realization. Writing $\mathcal{R} := H_\theta - H$, the learned system is the perturbed Hamiltonian system*

$$H_\theta \;=\; H + \mathcal{R}, \qquad X_{H_\theta} = X_H + X_{\mathcal{R}}. \tag{2}$$

This is elementary, but it relocates the problem. The training objective is, up to weighting,

$$\mathcal{L}(\theta) = \left\|X_{H_\theta} - X_H\right\|^2_{L^2(\mu)} = \left\|X_{\mathcal{R}}\right\|^2_{L^2(\mu)}, \tag{3}$$

i.e. an L^2 bound on the *perturbation's vector field*. Define the residual scale

$$\varepsilon \;:=\; \left\|X_{\mathcal{R}}\right\|_{\mathrm{rms}} / \left\|X_H\right\|_{\mathrm{rms}}. \tag{4}$$

The HNN is thus a *learned near-integrable system* with perturbation parameter ε, and the long-horizon question is the structural stability of the invariant foliation of H under the perturbation $\mathcal{R}$. This is exactly the domain of KAM and Nekhoroshev theory, and the remainder of the paper draws out the consequences.

3 Consequences of KAM/Nekhoroshev for HNNs

3.1 Persistence: stability is topology-preservation, not vector-field accuracy

For H integrable with action–angle variables (I, φ) and frequencies $\omega(I) = \partial H_0/\partial I$, the KAM theorem states that if H satisfies a non-degeneracy condition (Kolmogorov: $\det(\partial\omega/\partial I) \neq 0$) and ε is sufficiently small, then every torus with *Diophantine* frequency,

$$|\langle k, \omega\rangle| \;\geq\; \frac{\kappa}{|k|^\tau} \quad \forall\, k \in \mathbb{Z}^d \setminus \{0\}, \tag{5}$$

persists as an invariant torus of the perturbed system H_θ, merely deformed by $O(\varepsilon)$; the persisting tori fill phase space up to a set of measure $O(\sqrt{\varepsilon})$ [1, 2, 3, 6].

The consequence for HNNs is the central conceptual point of this paper. On the positive-measure majority of initial conditions lying on Diophantine tori, the learned trajectory remains on an invariant torus of the *learned* flow *for all time*: it is bounded, recurrent, and non-drifting. The guarantee is *topological*—the orbit stays on a torus—not *metric*—the vector field is accurate. Contrast the naive error bound from Grönwall's inequality,

$$\|\delta z(t)\| \;\leq\; \frac{\varepsilon_{\mathrm{abs}}}{L}\left(e^{Lt} - 1\right), \tag{6}$$

with L a Lipschitz constant: this is exponential in t and useless beyond $O(1/L)$ horizons. KAM replaces (6) on the protected set by an $O(\varepsilon)$-*forever* bound. This is the precise sense in which the symplectic inductive bias buys long-horizon stability: not by making ε smaller, but by changing the *character* of error propagation from secular to bounded.

3.2 Resonances and the $\sqrt{\varepsilon}$ law

KAM protects Diophantine tori; it does not protect resonant ones. Where $\langle k, \omega(I)\rangle = 0$ for some low-order k, the small divisors in (5) fail, the torus is destroyed (Poincaré–Birkhoff), and is replaced by a chain of islands enclosing a thin stochastic layer. For an isolated $m:n$ resonance the dynamics reduces, to leading order, to a pendulum whose separatrix half-width in action scales as the square root of the resonant Fourier coefficient $\mathcal{R}_{mn}$ of the perturbation:

$$w \;\sim\; \sqrt{|\mathcal{R}_{mn}|} \;\sim\; \sqrt{\varepsilon}. \tag{7}$$

For an HNN this predicts: (i) rollout error concentrates near low-order resonances of the true system; (ii) if the target is integrable, the network *manufactures* resonant islands and chaos that do not exist in the true dynamics; (iii) their width scales as $\sqrt{\varepsilon}$ in the training residual; and (iv) *which* resonance is excited is selected by the Fourier content of $\mathcal{R}$, i.e. by what the network got wrong, not merely by how much. We confirm (i)–(iv) in §4.

3.3 Nekhoroshev: exponentially slow drift of the true invariants

Where invariant tori do not separate the energy shell (e.g. $d \geq 3$, where Arnold diffusion permits slow transport along resonance webs) or off the protected set, confinement is not absolute but the Nekhoroshev theorem still bounds drift: under a steepness condition on H,

$$\|I(t) - I(0)\| \;\leq\; C\,\varepsilon^{\,b} \qquad \text{for} \qquad |t| \;\leq\; T_0\,\exp\!\big(c\,\varepsilon^{-a}\big), \tag{8}$$

for positive constants a, b, c [5, 8]. The implication for HNNs is that the true energy and actions along the learned flow are *not* conserved—only H_θ is—but they remain within $O(\varepsilon^b)$ for times *exponentially long* in $1/\varepsilon$. The empirically observed bounded, non-secular energy behaviour of HNNs is thus the expected Nekhoroshev signature, and the correct contrast with a black-box net is in the *growth law* (bounded versus secular), not in the instantaneous error. We confirm this in §4 (Fig. 3).

3.4 Degeneracy is a liability

KAM requires non-degeneracy. Kolmogorov's condition $\det(\partial\omega/\partial I) \neq 0$ (or, more sharply, Rüssmann transversality [7]) demands that frequencies vary with the action—a *twist*. Degenerate systems violate this. The isotropic harmonic oscillator $H_0 = \frac{1}{2}|p|^2 + \frac{1}{2}|q|^2$ is the extreme case: it is isochronous (all frequencies equal and amplitude-independent) and superintegrable, so *every* torus is resonant (1:1) and the foliation by periodic orbits is non-generic. An arbitrarily small generic perturbation then has no protective KAM cloud to fall back on and reorganizes the global phase portrait at first order.

This is made precise by Eliasson, Fayad and Krikorian [4]: a Diophantine invariant torus is *non-isolated*—accumulated by a positive-measure set of KAM tori, hence stable—if and only if its Birkhoff normal form is non-degenerate (a Rüssmann-type transversality condition). A degenerate normal form yields no surrounding cloud of invariant tori, and the torus loses its protection.

Prediction 1. *For fixed training residual ε, an HNN is* more *fragile on degenerate (isochronous / superintegrable) systems than on non-degenerate twist systems. In particular, the linear harmonic oscillator—usually treated as the easy case—is the adversarial case for HNN long-horizon structural stability. Long-horizon HNN reliability should therefore not be benchmarked solely on harmonic oscillators, and an explicit twist (anharmonicity) is protective.*

We flag Prediction 1 as a theoretical consequence and a design caution; the twist systems we use below are precisely the non-degenerate regime where KAM applies, and they exhibit the protected behaviour the theory predicts.

3.5 The objective ignores what KAM says matters

KAM smallness is measured in an *analytic* (or Gevrey) norm that weights Fourier modes by $e^{|k|\sigma}$, and torus destruction is driven by *specific* resonant Fourier coefficients $\mathcal{R}_{mn}$ (7). The L^2 vector-field loss (3) is blind to both: it can be small while $\mathcal{R}$ carries high-frequency content that destroys tori, and it weights every resonance equally regardless of order. This suggests two KAM-informed objectives, which we propose and leave to future empirical work:

1. **Analytic-norm regularizer.** Penalize high-frequency content of the residual (or of the learned field), e.g. $\sum_k e^{|k|\sigma}|\hat{\mathcal{R}}_k|^2$ estimated on a Fourier feature basis, pushing $\mathcal{R}$ toward the smoothness regime KAM tolerates rather than minimizing pointwise error alone.
2. **Action-variance penalty.** Along short rollouts, penalize $\mathrm{Var}_t(\hat{I})$ for surrogate actions $\hat{I}$ (e.g. single-degree-of-freedom energies of a normal form). This directly penalizes destruction of the invariants that label the tori—the actual long-horizon failure mode—rather than instantaneous field error.

4 Experimental setup

Testbed. We deliberately choose an *integrable* system with genuine twist, two uncoupled hardening oscillators,

$$H(x, y, p_x, p_y) = \tfrac{1}{2}(p_x^2 + p_y^2) + \tfrac{1}{2}(x^2 + y^2) + \tfrac{\beta}{4}(x^4 + y^4), \qquad \beta = 1. \tag{9}$$

It has two independent integrals $h_x = \frac{1}{2}p_x^2 + \frac{1}{2}x^2 + \frac{\beta}{4}x^4$ and h_y (with $h_x + h_y = H$); its frequencies harden with amplitude, so it is Kolmogorov non-degenerate; and it contains *no chaos anywhere*. This is the key design choice: any island or chaotic layer the network produces is unambiguously an artifact of its own error, not a feature of the target. We work on the energy surface $E = 0.5$.

Models. The HNN is a scalar MLP $[4, 128, 128, 1]$ with tanh activations; its field is $\mathbb{J}\nabla H_\theta$ via automatic differentiation. A plain-MLP baseline $[4, 128, 128, 4]$ regresses the field directly. Both are trained with Adam on 4000 phase-space samples of the true field with light additive noise. The fully trained HNN reaches residual $\varepsilon \approx 0.004$ (relative RMS 0.8%); for visualization of the resonance structure we also use a deliberately under-trained network with $\varepsilon \approx 0.07$.

Integration and diagnostics. True and controlled sections use a symplectic leapfrog integrator (the relevant Hamiltonians are separable); the learned flow uses RK4 with a closed-form input-gradient (gradient check error 10^{-10}). To probe scaling laws at exactly known perturbation strength we use a controlled family $H_\varepsilon = H + \varepsilon\, x^2y^2$, whose coupling drives the 1:1 resonance and remains separable. This is legitimate because the relevant scaling laws are properties of the *perturbed Hamiltonian* and do not depend on whether the perturbation is produced by hand or by a network; the trained network produces a perturbation of the same type (Fig. 1, right).

5 Results

The network manufactures resonant chaos (Fig. 1). The true section is a clean family of nested invariant curves with no chaotic region. The under-trained HNN reproduces the surviving (deformed) tori but develops a prominent resonant island chain surrounded by a thin stochastic layer—structure that is absent from the integrable target. The surviving curves are the protected Diophantine tori; the broken chain is a low-order resonance. This is exactly the persistence/destruction dichotomy of §3.1–3.2, here produced entirely by the network's residual.

Error localizes at resonances (Fig. 2a). Measuring the drift of the true action h_x along the flow as a function of the initial torus shows the error is *not* uniform: it is essentially zero on Diophantine tori and rises sharply at the 1:1 resonance, with the twin libration peaks (and the node at the elliptic fixed point) characteristic of an isolated resonance. The contrast is two to three orders of magnitude. The L^2 loss, being uniform over phase space, cannot predict *where* the model fails; KAM can.

The $\sqrt{\varepsilon}$ law (Fig. 2b). Across the controlled family, the island half-width follows a power law in the perturbation size with exponent 0.59, consistent with the KAM/pendulum prediction of 1/2 (7) given finite-time integration, an FWHM (rather than separatrix) width estimate, and a coordinate (rather than action) measure. The under-trained network of Fig. 1 ($\varepsilon \approx 0.07$) produces islands of comparable width (≈ 0.17), matching the law's ≈ 0.19. A quantitative KAM prediction thus holds for a learned model.

Bounded versus secular error growth (Fig. 3). Evaluating the true energy along each learned trajectory, the HNN remains within $O(\varepsilon)$ of the initial energy in a bounded oscillation, while the plain MLP—of comparable instantaneous accuracy—drifts far off the energy surface. The decisive difference is qualitative: bounded versus secular growth, exactly the distinction (6) versus (8) predicts. Instantaneous field accuracy does not separate the two models; the error-growth law does.

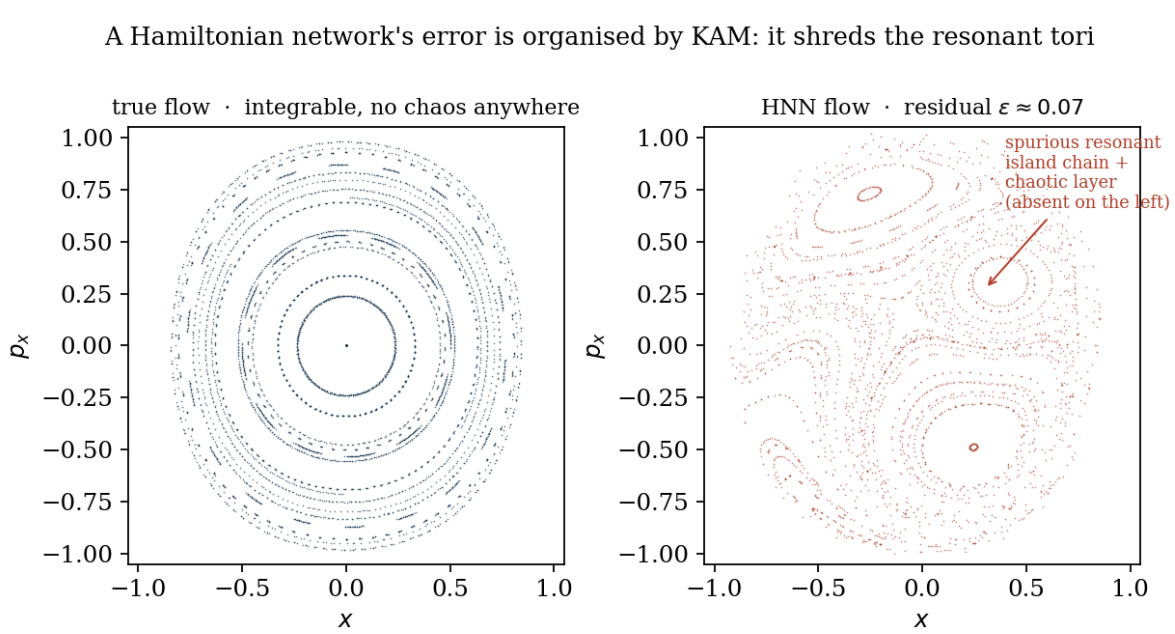


Figure 1: The network manufactures resonant chaos. Poincaré section ($y = 0$, $p_y > 0$) at $E = 0.5$. *Left:* the true integrable flow—clean nested tori, no chaos anywhere. *Right:* the flow of an HNN with residual $\varepsilon \approx 0.07$—the deformed Diophantine tori survive, but a resonant island chain and a thin chaotic layer have appeared that exist nowhere in the true system. The break is the KAM dichotomy realized by learning error; which resonance breaks is selected by the Fourier content of the residual.

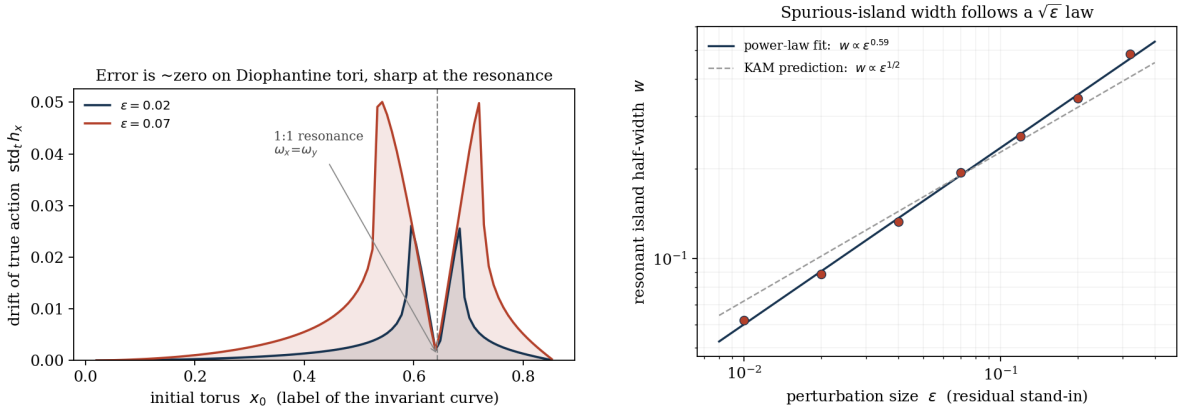
6 Related work

HNNs [12], symplectic networks [13], and physics-informed networks [14] establish *that* a structural prior improves long-horizon behaviour. Our contribution is a perturbation-theoretic account of *why and when*, together with predictions of *how* such models fail.

The closest prior art is the backward-error analysis of symplectic integrators. A symplectic one-step method conserves a *modified* (shadow) Hamiltonian $\tilde{H} = H + O(\Delta t^p)$ exactly, and its long-time energy behaviour is governed by Nekhoroshev-type estimates on $\tilde{H}$ [9, 8]. An HNN is the *learned* analogue of this construction: H_θ plays the role of $\tilde{H}$, with the discretization error $O(\Delta t^p)$ replaced by the learning residual $\mathcal{R}$. We import this lineage and add (i) the resonance-localization and $\sqrt{\varepsilon}$ predictions specialized to learned residuals (whose spectral content is data- and architecture-dependent rather than fixed by a method), (ii) the degeneracy/non-degeneracy design implication tied to the stability characterization of [4], and (iii) the analytic-norm and action-variance objectives. To our knowledge these transfers, and their empirical confirmation for learned Hamiltonians, are new. The dynamical-systems ingredients are classical: KAM [1, 2, 3, 6], the non-degeneracy theory of [7, 4], Nekhoroshev stability [5], resonance overlap [10], and the Hénon–Heiles model of induced chaos [11].

7 Limitations

The hypotheses of KAM—analyticity or high smoothness, ε below an explicit threshold, and non-degeneracy—are *not* guaranteed for a given trained network: tanh/ReLU models are only finitely smooth, the residual need not be below KAM thresholds, and real targets may be far from integrable. We therefore claim KAM/Nekhoroshev as the governing *asymptotic regime* that predicts the correct *qualitative laws*—resonance localization, $\sqrt{\varepsilon}$ island width, and bounded-versus-secular growth, all confirmed empirically—not that any particular network provably satisfies KAM's premises. In $d \geq 3$ degrees of freedom, invariant tori no longer separate the energy shell and Arnold diffusion permits slow transport along resonance webs; HNN confinement then weakens to the Nekhoroshev-exponential timescale of (8) rather than holding forever, and our two-degree-of-freedom experiments do not probe this. Estimating the residual's analytic norm or its resonant Fourier coefficients in high dimension is nontrivial, and the proposed objectives await empirical validation. Finally, our measured scaling exponent (0.59) deviates from 1/2; we attribute this to finite integration time and to measuring an FWHM coordinate width rather than a separatrix action width, and do not claim exactly 1/2.



(a) Error localizes at the resonance.

(b) Island width follows $\sqrt{\varepsilon}$.

Figure 2: Quantitative KAM signatures. (a) Drift of the true action $\mathrm{std}_t\, h_x$ along the learned/perturbed flow versus the initial torus x_0, for the controlled family. The error is $\sim 10^{-4}$ on Diophantine tori and spikes by 2–3 orders of magnitude at the 1:1 resonance, with the characteristic twin libration peaks straddling the elliptic centre; the $\varepsilon = 0.07$ peaks are visibly wider than $\varepsilon = 0.02$. (b) Resonant-island half-width versus perturbation size: a power law with exponent 0.59, consistent with the KAM/pendulum prediction $w \propto \sqrt{\varepsilon}$ up to finite-time and coordinate corrections (dashed: slope 1/2).

8 Conclusion

An HNN does not learn the true Hamiltonian; it learns a nearby one, and its long-horizon behaviour is consequently a perturbation-theory question. KAM and Nekhoroshev theory then explain the empirically observed stability—as preservation of the topology of the flow and as confinement of the true invariants—and, crucially, *predict structured failure*: error that localizes at resonances, spurious islands whose width scales as $\sqrt{\varepsilon}$, and heightened fragility on degenerate systems. None of these follow from an "energy conservation" account, and we confirm the first three on a system engineered to have no chaos of its own. The same lens yields concrete guidance—analytic-norm and action-variance objectives, and caution against degenerate benchmarks—and places HNNs within the mature backward-error-analysis theory of symplectic integrators as *learned shadow Hamiltonians*.

References

[1] A. N. Kolmogorov. On the conservation of conditionally periodic motions under small perturbations of the Hamiltonian. *Dokl. Akad. Nauk SSSR*, 98:527–530, 1954.

[2] V. I. Arnold. Proof of a theorem of A. N. Kolmogorov on the preservation of conditionally periodic motions. *Russian Math. Surveys*, 18(5):9–36, 1963.

[3] J. Moser. On invariant curves of area-preserving mappings of an annulus. *Nachr. Akad. Wiss. Göttingen Math.-Phys. Kl. II*, pages 1–20, 1962.

[4] L. H. Eliasson, B. Fayad, and R. Krikorian. Around the stability of KAM tori. *Duke Math. J.*, 164(9):1733–1775, 2015.

[5] N. N. Nekhoroshev. An exponential estimate of the time of stability of nearly integrable Hamiltonian systems. *Russian Math. Surveys*, 32(6):1–65, 1977.

[6] J. Pöschel. A lecture on the classical KAM theorem. *Proc. Sympos. Pure Math.*, 69:707–732, 2001.

[7] H. Rüssmann. Invariant tori in non-degenerate nearly integrable Hamiltonian systems. *Regul. Chaotic Dyn.*, 6(2):119–204, 2001.

[8] G. Benettin and A. Giorgilli. On the Hamiltonian interpolation of near-to-the-identity symplectic mappings with application to symplectic integration algorithms. *J. Stat. Phys.*, 74:1117–1143, 1994.

[9] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration*, 2nd ed. Springer, 2006.

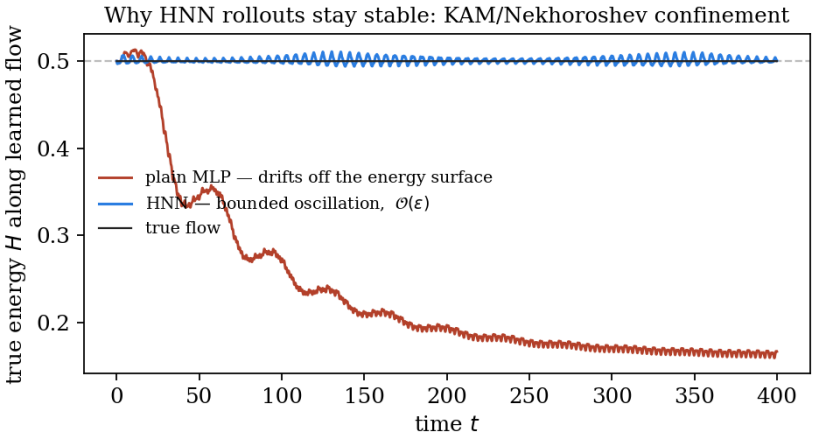


Figure 3: Confinement, not energy conservation *per se*. True energy H evaluated along each learned flow. The HNN (blue) stays within $O(\varepsilon)$ of E_0 in a bounded oscillation with no secular trend; the plain MLP (red), trained to comparable field MSE, drifts $\sim$68% off the energy surface. The HNN advantage is the *growth law* (bounded versus secular), the empirical signature of KAM/Nekhoroshev confinement.

[10] B. V. Chirikov. A universal instability of many-dimensional oscillator systems. *Phys. Rep.*, 52(5):263–379, 1979.

[11] M. Hénon and C. Heiles. The applicability of the third integral of motion: some numerical experiments. *Astron. J.*, 69:73–79, 1964.

[12] S. Greydanus, M. Dzamba, and J. Yosinski. Hamiltonian neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[13] P. Jin, Z. Zhang, A. Zhu, Y. Tang, and G. E. Karniadakis. SympNets: intrinsic structure-preserving symplectic networks for identifying Hamiltonian systems. *Neural Networks*, 132:166–179, 2020.

[14] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks. *J. Comput. Phys.*, 378:686–707, 2019.