KAM Theory Meets Statistical Learning Theory: Hamiltonian Neural Networks with Non-zero Training Loss

Yuhan Chen[1*], **Takashi Matsubara**[2*], **Takaharu Yaguchi**[1*],

[1]Kobe University, 1–1 Rokkodai, Nada, Kobe, Hyogo, Japan 657–8501
[2]Osaka University, 1–3 Machikaneyama, Toyonaka, Osaka, Japan 560–8531
193x226x@stu.kobe-u.ac.jp, matsubara@sys.es.osaka-u.ac.jp, yaguchi@pearl.kobe-u.ac.jp

Abstract

Many physical phenomena are described by Hamiltonian mechanics using an energy function (the Hamiltonian). Recently, the Hamiltonian neural network, which approximates the Hamiltonian as a neural network, and its extensions have attracted much attention. This is a very powerful method, but its use in theoretical studies remains limited. In this study, by combining the statistical learning theory and Kolmogorov–Arnold–Moser (KAM) theory, we provide a theoretical analysis of the behavior of Hamiltonian neural networks when the learning error is not completely zero. A Hamiltonian neural network with non-zero errors can be considered as a perturbation from the true dynamics, and the perturbation theory of the Hamilton equation is widely known as the KAM theory. To apply the KAM theory, we provide a generalization error bound for Hamiltonian neural networks by deriving an estimate of the covering number of the gradient of the multi-layer perceptron, which is the key ingredient of the model. This error bound gives a sup-norm bound on the Hamiltonian that is required in the application of the KAM theory.

Introduction

Many physical phenomena are described by energy-based theories, such as Hamiltonian mechanics (e.g., Furihata and Matsuo (2010)). The governing equation of Hamiltonian mechanics is

$$\frac{\mathrm{d}u}{\mathrm{d}t} = S\frac{\partial H}{\partial u}, \tag{1}$$

where $u : t \in \mathbb{R} \mapsto u(t) \in \mathbb{R}^N$, $H : u \in \mathbb{R}^N \mapsto H(u) \in \mathbb{R}$, and S is a skew-symmetric matrix. H represents the energy function of the system. In recent years, there has been a lot of research on predicting the corresponding physical phenomena by learning the energy function H in such equations with a neural network H_{NN} (e.g., Chen et al. (2020); Cranmer et al. (2020); Greydanus, Dzamba, and Yosinski (2019); Matsubara, Ishikawa, and Yaguchi (2020); Zhong, Dey, and Chakraborty (2020)); however, to the best of our knowledge, theoretical analysis of such models has not been performed sufficiently, except for SympNet (Jin et al. 2020b) for the Hamilton equation, where the universal approximation theorems for discrete-time neural network models are provided.

In this paper, we focus on theories of the properties of the most fundamental model, comprising Hamiltonian neural networks (HNNs) (Greydanus, Dzamba, and Yosinski 2019)

$$\frac{\mathrm{d}u}{\mathrm{d}t} = S\frac{\partial H_{\mathrm{NN}}}{\partial u} \tag{2}$$

and their extensions in practical situations, where the learning error is not completely zero. In this case, the trained model can be regarded as a perturbed Hamiltonian system due to the modeling error of the energy function. In addition, S is a general skew-symmetric matrix and hence (2) can model Hamiltonian partial differential equations (Matsubara, Ishikawa, and Yaguchi 2020).

In mathematical physics, perturbed Hamiltonian systems are well studied. For example, because whether the solar system will continue to exist in the future is of fundamental interest in astronomy, the stability of the solar system under perturbations is a very important issue that has been studied for a long time (e.g., Féjoz (2013); Laskar (1996)). The Kolmogorov–Arnold–Moser (KAM) theory gives an answer to questions of this type; essentially, periodic motions of such systems are stable under small perturbations. The stability of periodic motions is of particular importance in science. In addition to the stability of celestial systems, the recursive nature of physical phenomena is also of interest in physics. For example, in the famous numerical experiments using the Korteweg–De Vries (KdV) equation by Zabusky and Kruskal (1965), it was confirmed that the waveform given as the initial condition initially collapsed due to severe oscillations, and then returned to its original shape after a long time. Whether such phenomena can be reproduced by deep learning models is an important problem that greatly affects the usefulness of deep physical models.

In this paper, we give an answer to this question by combining the KAM theory and statistical learning theory. Because trained models are also perturbed Hamiltonian systems, it is expected that the periodic behaviors of the systems are preserved even if the loss function does not vanish completely provided it is sufficiently small. However, this expectation cannot be proved in a straightforward way, because the application of the KAM theory requires that *the energy function be close enough to that of the true dynamics in the whole phase space.* Noticing that the error of the

*These authors contributed equally.

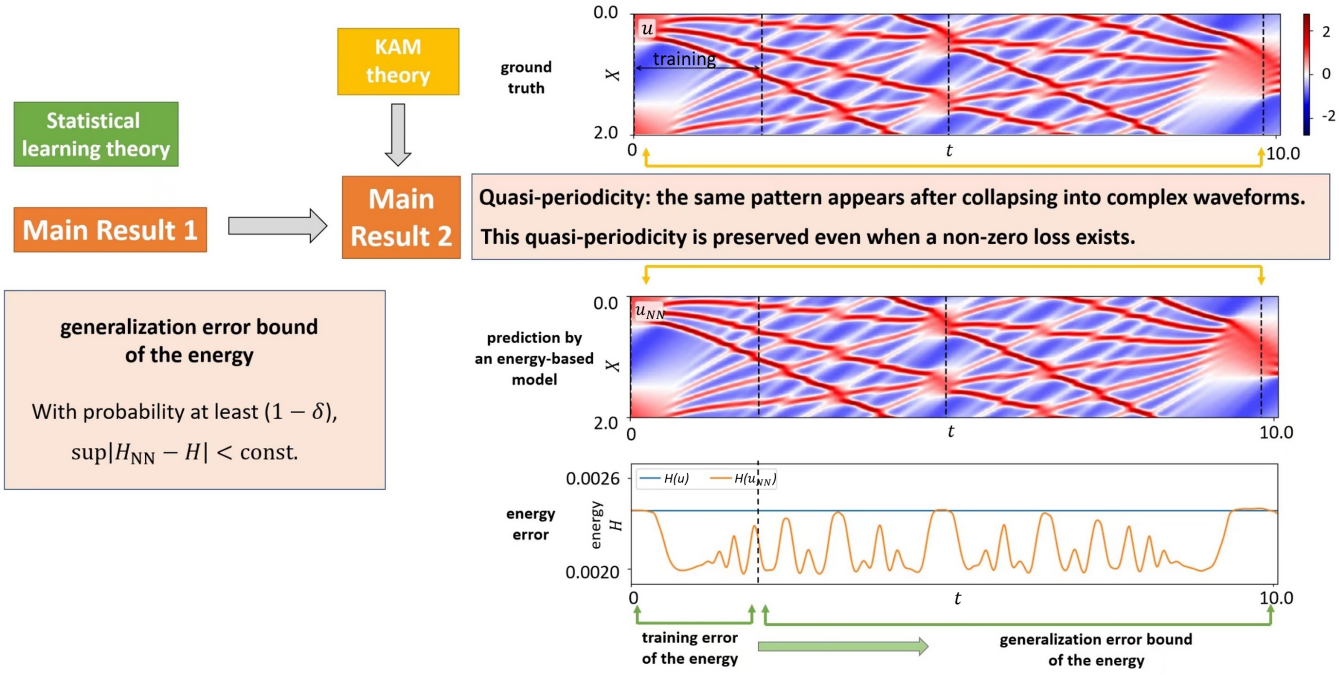


Figure 1: Outline of our main theorem. The first main result is the generalization error of the energy function, which is proved by the statistical learning theory. By combining the first result and the KAM theorem, we prove that the quasi-periodic behaviors of target systems are preserved.

energy function in the whole space is essentially the generalization error, we overcome this difficulty by combining the KAM theory with statistical learning theory. This illustrates that combinations of statistical learning theory and dynamical systems theory can lead to powerful results. Indeed, combinations of this kind may be applicable to other stability results in dynamical system contexts, for example, the stability of solitary wave solutions.

Importantly, for the neural network models to be close to the true dynamics, we need a universal approximation theorem and also a generalization error bound. In this paper, we also provide such results for HNNs.

Regarding the generalization error bound, because the derivative of a multi-layer perceptron is used in HNNs, a bound for the derivative is required. To this end, we estimated the covering number of the derivative of multi-layer perceptrons. An L^∞ bound on the error in the Hamiltonian is also provided, which is required for application of the KAM theory.

In addition, we show a universal approximation theorem for a model with the coordinate transformation

$$\frac{\mathrm{d}x}{\mathrm{d}t} = (\frac{\partial u}{\partial x})^{-1} S(\frac{\partial u}{\partial x})^{-\top} \frac{\partial H}{\partial x}. \tag{3}$$

This model is indispensable in practice; to apply HNNs, the data are given in the canonical coordinate system because the equation of motion is in the form of the Hamilton equation (1) only when the state variables are represented by the canonical coordinate system. However, this coordinate system depends on an unknown Lagrangian and hence the energy function. Hence, the coordinate system must be also learned from data by using, for example, neural networks. In addition, this model can be extended to represent other energy-based physical models beyond the Hamilton equation (Matsubara, Ishikawa, and Yaguchi 2020).

The main contributions of this paper are as follows.

1. **Combination of the KAM theory and statistical learning theory for HNNs with non-zero training loss to prove the existence of quasi-periodic behaviors (see Figure 1).**
2. **Derivation of a generalization error bound for HNNs.**
3. **Development of a universal approximation theorem for HNNs and other energy-based physical models with coordinate transformations.**

Related Work

Many studies of neural network models for phenomena that can be modeled by energy-based equation (1) have been put forward. Among them, the most basic studies are neural ordinary differential equations (Chen et al. 2018) and HNNs (Greydanus, Dzamba, and Yosinski 2019). In particular, extensions of HNNs have been intensively developed.

Describing them all is beyond the scope of this paper, but some examples are given here. In Toth et al. (2019) HNNs were extended to latent variable models. Other studies, such as DiPietro, Xiong, and Zhu (2020); Xiong et al. (2021); Zhong, Dey, and Chakraborty (2020), focused on the symplectic structure of the Hamilton equation. For Noether's

theorem, which is a fundamental theorem in classical mechanics, several studies (Bharadwaj, Li, and Demanet 2020; Bondesan and Lamacraft 2019; Finzi et al. 2020) developed methods related to symmetry and conservation laws. In addition, a discrete-time model that preserves the energy behaviors was constructed in Matsubara, Ishikawa, and Yaguchi (2020). In Galioto and Gorodetsky (2020), HNNs were combined with a Bayesian approach.

Methods applied to the framework of classical mechanics other than Hamiltonian mechanics include those in Cranmer et al. (2020); Desai and Roberts (2020); Sæmundsson et al. (2019), which are methods for Lagrangian formalism. In Jin, Lin, and Li (2020), reinforcement learning was applied to the variational principle. A simplified model formed by introducing constraints was proposed in Finzi, Wang, and Wilson (2020). In Jin et al. (2020a), HNNs were extended to the Poisson system, which is a wider class of mechanical equations. There are also a number of proposals that integrate them with more advanced deep learning techniques, namely, graph networks (Sanchez-Gonzalez et al. 2019), recurrent neural networks (Chen et al. 2020), and normalizing flows (Li et al. 2020). As an application-oriented approach, Feng et al. (2020) designed a microscopic model for structural analysis.

However, as far as the authors know, there is no theoretical research other than the universal approximation theorems for Hamiltonian mechanics in SympNet (Jin et al. 2020b), in which a certain kind of neural network is shown to have universal approximation properties for symplectic maps. The difference between their results and ours is that (1) we analyze the behaviors of a HNN with non-zero training loss by a combination of the KAM theory and statistical learning theory, (2) we provide a generalization error bound for NHHs, and (3) the universal approximation theorems in Jin et al. (2020b) are for discrete-time models, while ours are for continuous-time models.

Meanwhile, as an existing energy-based model, Hopfield neural network is known. Both Hopfield neural network and Hamiltonian network are derived from energy-based theories, and their dynamics are described by (1). Hamiltonian neural network is associated with a skew-symmetric matrix S and is a model of an energy-preserving, continuous-time, and deterministic physics phenomenon. Its output is the time-series of the state. Hopfield network is associated with a negative definite matrix S and exhibits a dynamics which is often energy-dissipating, discrete-time, and stochastic. It is a machine learning tool rather than a physical model, and its equilibrium point is treated as its output. Because their outputs are different, their theoretical properties should be discussed separately.

Brief Introduction to Hamiltonian Systems and the KAM Theory

We briefly introduce some properties of Hamiltonian systems.

Theorem 1 (Darboux)**.** *By an appropriate coordinate transformation, the matrix S is transformed into the normal form*

$$\begin{pmatrix} O & I \\ -I & O \end{pmatrix}.$$

Definition 1. *The function $\omega : (v, w) \in \mathbb{R}^N \times \mathbb{R}^N \mapsto \omega(v, w) \in \mathbb{R}$*

$$\omega(v, w) = v^\top S^{-1} w$$

is called the symplectic form. Using the symplectic form associates a vector field X_F with each function $F : \mathbb{R}^N \to \mathbb{R}$ by requiring

$$\omega(X_F, w) = \frac{\partial F}{\partial u} \cdot w \qquad \text{for all } w.$$

For two functions F, G, the following operation is called the Poisson bracket:

$$\{F, G\} = \omega(X_F, X_G). \tag{4}$$

Definition 2. *A Hamiltonian system for which the state variable is $N = 2M$ dimensional is integrable in the sense of Liouville if this Hamiltonian system has the first integrals (i.e., conserved quantities) $F_1, F_2, \ldots, F_M$ with $\nabla F_1(u), \nabla F_2(u), \ldots, \nabla F_M(u)$ independent at each u and for all i, j:*

$$\{F_i, F_j\} = 0.$$

For integrable systems, Theorem 2 is known.

Theorem 2 (Liouville–Arnold)**.** *Suppose that for an integrable Hamiltonian system, constants $c_1, \ldots, c_M$ exist such that $K = \cap_{i=1}^M F_i^{-1}(c_i)$ is connected and compact. Then, there exists a neighborhood $\mathcal{N}$ comprising K, $\mathcal{U} \subset \mathbb{R}^n$ and a coordinate transform*

$$\phi : (\theta, J) \in \mathbb{T}^n \times \mathcal{U} \to \phi(\theta, J) \in \mathcal{N} \tag{5}$$

such that the transformed system is the Hamilton equation of which Hamiltonian $H \circ \phi$ depends only on J.

The variables J and θ are called action-angle variables. Theorems 1 and 2 roughly mean that integrable Hamiltonian systems can be written in the following form:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} \theta \\ J \end{pmatrix} = \begin{pmatrix} O & I \\ -I & O \end{pmatrix}\begin{pmatrix} \frac{\partial \tilde{H}}{\partial \theta} \\ \frac{\partial \tilde{H}}{\partial J} \end{pmatrix}.$$

Further, because $\tilde{H} = H \circ \phi$ depends on I only, it holds that

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} \theta \\ J \end{pmatrix} = \begin{pmatrix} O & I \\ -I & O \end{pmatrix}\begin{pmatrix} 0 \\ \frac{\partial \tilde{H}}{\partial J} \end{pmatrix} = \begin{pmatrix} \frac{\partial \tilde{H}}{\partial J} \\ 0 \end{pmatrix}.$$

This shows that J is constant, and hence θ moves on the torus at a constant velocity. Because the velocities are typically not co-related to each other, the dynamics are "quasi-periodic." See, for example, Scott Dumas (2014) for more details.

As seen above, integrable Hamiltonian systems are quasi-periodic. Note that general Hamiltonian systems are not necessarily quasi-periodic and neither are HNNs. However, for the HNNs that are trained to model integrable systems, the quasi-periodic behaviors are preferably maintained. When

the modeling error is sufficiently small, this is considered as a perturbation problem. The perturbation theory of Hamiltonian systems has been investigated from various perspectives. For example, perturbed integrable Hamiltonian systems are in general no longer integrable; hence, approximation of integrable Hamiltonian systems by integrable neural network models appears to be difficult. Fortunately, however, the KAM theory shows that even though the perturbed system is not integrable, it maintains the quasi-periodic behaviors described above under certain conditions.

The KAM theorem has many variants under various conditions. The following variant (Scott Dumas 2014) is typical:

Theorem 3 (KAM Theorem)**.** *Let θ and J be the action-angle variables for a C^∞ integrable Hamiltonian $H_0 : \mathbb{R}^{2M} \to \mathbb{R}$ with $M \geq 2$. If H_0 is non-degenerate, that is,*

$$\det \frac{\partial^2 H_0}{\partial J^2} \neq 0, \tag{6}$$

for the perturbed system $H(\theta, J) = H_0(J) + \varepsilon F(\theta, J, \varepsilon)$ by $F \in C^\infty$, there exists ε_0 such that if $\varepsilon F < \varepsilon_0$, there exists a set of M-dimensional tori that are invariant under the perturbed flow. For each invariant torus, the flow of the perturbed system H is quasi-periodic. In addition, the set of invariant tori is large in the sense that its measure becomes full as $\varepsilon \to 0$.

Remark 1. *The last sentence – the set of invariant tori is large in the sense that its measure becomes full as $\varepsilon \to 0$ – corresponds to the non-existence of so-called resonance. If the perturbation added to the system is in resonance with the original system, the perturbation may grow rapidly and the behavior of the system may change significantly. This statement assures that for small perturbations, such a situation almost never occurs.*

Remark 2. *It may be difficult to check whether the target system is integrable by using given data. One possibility is application of the Koopman operator, which makes it possible to find the conserved quantities that the given data may admit. If a sufficient number of conserved quantities exist, it is highly likely that the target system is integrable.*

Main Results

Universal Approximation Properties of HNNs

For HNNs to be close to the true dynamics, a universal approximation theorem and a generalization error analysis are needed. First, we show universal approximation theorems.

We first define some notation to describe the theorem. $C^m(X)$ with the topology of the Sobolev space $W^{p,m}(X)$ is denoted by $S_p^m(X)$, where $W^{p,m}(X)$ is a space of functions that admit weak derivatives up to the mth order that bounds L^p-norms. Hence, $S_p^m(X)$ is the space of functions in $W^{p,m}(X)$ with (usual) derivatives; for details on the Sobolev theory, see Adams and Fournier (2003). L^p-norms of functions are denoted by $\|\cdot\|_{L^p}$, and those of vectors by $\|\cdot\|_p$.

Universal approximation theorem for HNNs The following theorem shows the universal approximation property of general energy-based physical models, which include HNNs (Matsubara, Ishikawa, and Yaguchi 2020).

Theorem 4. *Let $H : \mathbb{R}^N \to \mathbb{R}$ be an energy function with the equation*

$$\frac{\mathrm{d}u}{\mathrm{d}t} = G\frac{\partial H}{\partial u},$$

where $u : t \in \mathbb{R} \mapsto u(t) \in \mathbb{R}^N$ and G is a non-degenerate $N \times N$ matrix. Suppose that the state space K of this system is compact and the right-hand side $G\partial H/\partial u$ is Lipschitz continuous. If the activation function $\sigma \neq 0$ belongs to $S_2^1(\mathbb{R})$, then for any $\varepsilon > 0$ there exists a neural network H_{NN} for which

$$\left\| G\frac{\partial H}{\partial u} - G\frac{\partial H_{\mathrm{NN}}}{\partial u} \right\|_2 < \varepsilon$$

holds. In addition, if the energy function is C^∞, the function can be approximated by a C^∞ neural network provided that the activation function is sufficiently smooth.

To prove this theorem, we use the following theorem and the lemma, both of which were shown in Hornik, Stinchcombe, and White (1990).

Theorem 5 (Hornik et al., 1990)**.** *Let $\Sigma(\sigma)$ be the space of the neural networks with the activation function σ. If the activation function $\sigma \neq 0$ belongs to $S_p^m(\mathbb{R}, \lambda)$ for an integer $m \geq 0$, then $\Sigma(\sigma)$ is m-uniformly dense in $C^\infty(K)$, where K is any compact subset of $\mathbb{R}^N$.*

Lemma 1 (Hornik et al., 1990)**.** *Under the same assumption, $\Sigma(\sigma)$ is also dense in $S_p^m(\mathbb{R}, \lambda)$.*

From these it follows that if the activation function σ of the hidden layer is in $S_p^m(\mathbb{R}, \lambda)$ and does not vanish everywhere, then for any sufficiently smooth function, there exists a neural network that approximates the function and its derivatives up to the order m arbitrarily well on compact sets. This theorem has also been extended to the functions of multiple outputs; see Hornik, Stinchcombe, and White (1990).

Proof of Theorem 4. Because the target equation is determined only by the gradient of H, any function obtained by shifting H by a constant gives the same equation. Hence, we choose and fix an energy function H that yields the target equation. Because $G\partial H/\partial u$ is Lipschitz continuous and hence continuous on the phase space K, this function is bounded and square-integrable. Thus, $G\partial H/\partial u \in S_2^0(K)$, which means H is in $S_2^1(K)$. Therefore, from Lemma 1 and the assumption that the activation function is in $S_2^1(\mathbb{R})$, for each ε, there exists a neural network that approximates H in $S_2^1(K)$:

$$\|H - H_{\mathrm{NN}}\|_2^2 + \left\| \frac{\partial H}{\partial u} - \frac{\partial H_{\mathrm{NN}}}{\partial u} \right\|_2^2 < \frac{\varepsilon^2}{\|G\|_2^2},$$

which gives

$$\left\| G\frac{\partial H}{\partial u} - G\frac{\partial H_{\mathrm{NN}}}{\partial u} \right\|_2^2 \leq \|G\|_2^2 \left\| \frac{\partial H}{\partial u} - \frac{\partial H_{\mathrm{NN}}}{\partial u} \right\|_2^2 < \varepsilon^2.$$

□

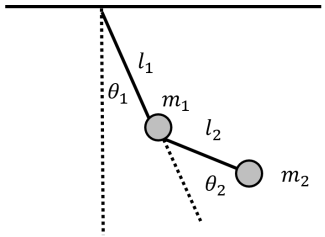
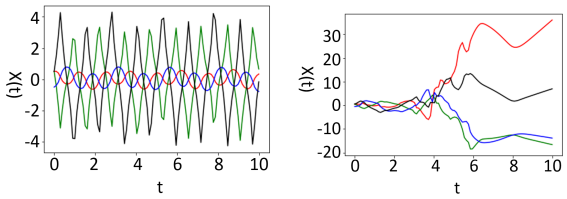
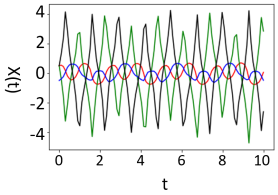


Figure 2: Double pendulum used as the target in the experiment for the illustration of the model with the coordinate transformations.



(a) Ground truth (b) Naive HNN model



(c) HNN model with a coordinate transformation

Figure 3: Examples of the orbits predicted by a HNN and the model with coordinate transformations. Each component of $x(t) = [q_1(t), v_1(t), q_2(t), v_2(t)]$ is represented as red (q_1), green (v_1), blue (q_2), and black (v_2).

HNNs with a coordinate transformation The practical use of HNNs is hampered by the fact that the state variables must be represented by a specific coordinate, such as the generalized momentum; however, the derivation of the generalized momentum requires the energy function, which is unknown. For example, the double pendulum in Figure 2 exhibits the dynamics shown in Figure 3. These are predicted by the models that are trained from the data of the state variables and their derivatives, not those of the generalized momenta. HNNs failed to solve such problems because the data were not given in the canonical coordinate system. Based on this, we here consider a model with a coordinate transformation, such as the transformations that appear, for example, in Rana et al. (2020); Jin et al. (2020a).

Suppose that, although the given data point $x(t)$ is not represented by the canonical coordinate system, the data point $x(t)$ can be transformed into the canonical coordinate system by an unknown transformation $u(t) = u_{\mathrm{NN}}[x(t)]$. By substituting $u = u_{\mathrm{NN}}(x)$ into the model equation (2), we obtain

$$\frac{\mathrm{d}x}{\mathrm{d}t} = (\frac{\partial u_{\mathrm{NN}}}{\partial x})^{-1} S (\frac{\partial u_{\mathrm{NN}}}{\partial x})^{-\top} \frac{\partial H_{\mathrm{NN}}}{\partial x}. \tag{7}$$

We show that model (7) admits the same energetic property as the original equation and also the universal approximation property.

Theorem 6. *The model* (7) *admits the energy conservation law in the sense that* $\mathrm{d}H_{\mathrm{NN}}/\mathrm{d}t = 0$.

Proof. By substituting the equation, we obtain

$$\begin{aligned}\frac{\mathrm{d}H_{\mathrm{NN}}}{\mathrm{d}t} &= \frac{\partial H_{\mathrm{NN}}}{\partial x}^{\top} \frac{\mathrm{d}x}{\mathrm{d}t} \\ &= \frac{\partial H_{\mathrm{NN}}}{\partial x}^{\top} \frac{\partial u_{\mathrm{NN}}}{\partial x}^{-1} S \frac{\partial u_{\mathrm{NN}}}{\partial x}^{-\top} \frac{\partial H_{\mathrm{NN}}}{\partial x} = 0\end{aligned}$$

because S is skew-symmetric and hence for any vector v, $v^\top S v = 0$. □

Theorem 7. *Let* $H : \mathbb{R}^N \to \mathbb{R}$ *be an energy function for the equation*

$$\frac{\mathrm{d}x}{\mathrm{d}t} = (\frac{\partial u}{\partial x})^{-1} S (\frac{\partial u}{\partial x})^{-\top} \frac{\partial H}{\partial x},$$

where $x : t \in \mathbb{R} \mapsto x(t) \in \mathbb{R}^N$, $u : x \in \mathbb{R}^N \mapsto u(x) \in \mathbb{R}^N$, *and* S *is an* $N \times N$ *matrix. Suppose that the phase space* K *of this system is compact and the right-hand side* $\partial H/\partial u$ *is Lipschitz continuous. Suppose also that* u *is a* C^1*-diffeomorphism. If the functions* $\sigma \neq 0$ *and* $\rho \neq 0$ *belong to* $S_2^1(\mathbb{R})$, *then for any* $\varepsilon > 0$, *there exist neural networks* H_{NN} *with the activation functions* σ *and* u_{NN} *with* ρ *for which*

$$\left\| (\frac{\partial u}{\partial x})^{-1} S (\frac{\partial u}{\partial x})^{-\top} \frac{\partial H}{\partial x} - (\frac{\partial u_{\mathrm{NN}}}{\partial x})^{-1} S (\frac{\partial u_{\mathrm{NN}}}{\partial x})^{-\top} \frac{\partial H_{\mathrm{NN}}}{\partial x} \right\|_2 < \varepsilon$$

holds.

Proof. We need to prove the approximation property for $(\partial u/\partial x)^{-1}$. From the assumption that $\rho \neq 0$ is in $S_2^1(\mathbb{R})$, there exists a function u_{NN} that approximates $\partial u/\partial x$. Because the determinant function of matrices is continuous, it is deduced that $\det \partial u_{\mathrm{NN}}/\partial x \neq 0$ and hence $(\partial u_{\mathrm{NN}}/\partial x)^{-1}$ exists. Because the matrix inverse is also continuous, $(\partial u_{\mathrm{NN}}/\partial x)^{-1}$ is also approximated by u_{NN}. □

Generalization Error Analysis of HNNs

Next, we derive a generalization error bound for the standard HNN (2) by employing a technique from statistical learning theory. More precisely, we adjust the technique so that an estimate on the energy gradient can be obtained.

Remark 3. *Although the standard HNN without the coordinate transformations is considered below, the results can be extended to the general energy-based model with the coordinate transformations if the matrix* $(\partial u/\partial x)^{-1}$ *is bounded.*

In statistical learning theory, generalization error bounds are typically obtained by using the Rademacher complexities. See, for example, Bousquet, Boucheron, and Lugosi (2004); Giné and Nickl (2016); Shalev-Shwartz and Ben-David (2014); Steinwart and Christmann (2008) for details.

Definition 3. *For a set $V \subset \mathbb{R}^n$,*

$$\mathcal{R}_n(V) := \frac{1}{n}\mathbb{E}_{\sigma\sim\{-1,1\}^n} \sup_{v\in V} \sum_{i=1}^{n} \sigma_i v_i$$

is called the Rademacher complexity of V.

Lemma 2. *Let X and Y be arbitrary spaces, $\mathcal{F} \subset \{f : X \to Y\}$ be a hypotheses class, and $L : Y \times Y \to [0, c]$ be a loss function. For a given data set $(x_i, y_i) \in X \times Y$ $(i = 1, \dots, n)$, let $\mathcal{G}$ be defined by $\{(x_i, y_i) \in X \times Y \mapsto L[y_i, h(x_i)] \in \mathbb{R} \mid h \in \mathcal{F}, i = 1, \dots, n\}$. Then, for any $\delta > 0$ and any probability measure P, we obtain, with a probability of at least $(1 - \delta)$ with respect to the repeated sampling of P^n-distributed training data, the following:*

$$E[L(Y, h(X))] \leq$$
$$\frac{1}{n}\sum_{i=1}^{n} L(y_i, h(x_i)) + 2\mathcal{R}_n(\mathcal{G}) + 3c\sqrt{\frac{2\ln\frac{4}{\delta}}{n}}$$

for all $h \in \mathcal{F}$.

The Rademacher complexity is known to be bounded by using the covering number.

Definition 4. *Let V and V' be subsets of $\mathbb{R}^n$. V is r-covered by V' with respect to the metric function defined by the norm $\|\cdot\|$ if for all $v \in V$, there exists a $v' \in V'$ such that $\|v - v'\| < r$. The covering number $N(r, V, \|\cdot\|)$ of V is the minimum number of elements of a set that r covers V. $N(r, V, \|\cdot\|)$ is also denoted by $N(r, V)$ if the metric is clear from the context.*

Lemma 3. *If $\sqrt{\log N(c2^{-k}, V)} \leq \alpha + k\beta$ for some α and β, then $\mathcal{R}_n(V) \leq 6c(\alpha + 2\beta)/n$.*

Thus, if the covering number is estimated for a HNN, the bound on the generalization error is obtained. To this end, we suppose that the model is trained by minimizing the p-norm of the error in the right-hand side of the model. More precisely, for the hypothesis $h : u_j \mapsto S\frac{\partial H_{\mathrm{NN}}(u_j)}{\partial u}$, we consider the loss function

$$L[\nabla H(u_j), h(u_j)] = \left\| \frac{\partial H(u_j)}{\partial u} - \frac{\partial H_{\mathrm{NN}}(u_j)}{\partial u} \right\|_p^p, \quad (8)$$

where u_i are training data. We denote the Lipschitz constant of the loss function associated with the above by ρ_p. Of course, $p = 2$ is typically used; however, we show below that $p > 2M$ is useful to obtain an L^∞ bound on the Hamiltonian.

Remark 4. *The training can be performed also by using the symplectic gradient:*

$$\left\| S\frac{\partial H(u_j)}{\partial u} - S\frac{\partial H_{\mathrm{NN}}(u_j)}{\partial u} \right\|_p^p. \quad (9)$$

In that case, the results will be slightly modified using the norm of S; however, we omit this for simplicity.

A bound of the covering number is derived as follows.

Theorem 8. *Suppose that the hypotheses class $\mathcal{F}$ consists of multi-layer perceptrons f_{NN} that have ρ_{σ_j}-Lispchitz activation functions $\sigma_j (j = 1, \dots, n_l)$, for which the derivatives are ρ'_j-Lipschitz continuous and bounded by $\sup|\sigma'_j| < c_{\sigma_j}$. Suppose also that the matrices $A_j^\top (j = 1, \dots, n_l + 1)$ in the linear layers in the perceptrons have the bounded norm $|A_j^\top| < c_{A_j}$:*

$$\mathcal{F} = \{f_{\mathrm{NN}}(u) \mid$$
$$A_{n_l+1}\sigma_{n_l}(A_{n_l}\sigma_{n_{l-1}}[\cdots\sigma_1(A_1u + b_1)\cdots] + b_{n_l}) + b_{n_l+1}\},$$

where b_j's are vectors. Let $\mathcal{G}$ be defined by $\{L[\nabla H(u_i), h(u_i)] \mid h \in \mathcal{F}\}$ with the ρ_p-Lipschitz continuous loss function L. In addition, suppose that the phase space is compact so that the data $u_i (i = 1, \dots, n)$ are in a bounded set with the bound $\|u_i\| < c_u$. Then, the covering number of $\mathcal{G}$ is estimated by

$$N(\varepsilon, \mathcal{G}) \leq$$
$$\Big(\frac{2\rho_p c_u c_{A_{n_{l+1}}} \rho'_{\sigma_{n_l}} (\prod_{j=1}^{n_l-1} \rho_{\sigma_j})(\prod_{j=1}^{n_l-1} c_{\sigma_j})(\prod_{j=1}^{n_l} c_{A_j})^2}{\varepsilon} + 1\Big)^n.$$

To prove this theorem, we use the following lemmas, which are typically used to estimate the covering numbers (Shalev-Shwartz and Ben-David 2014).

Lemma 4. *Let B be a unit ball in $\mathbb{R}^n$. Then, $N(\varepsilon, B, \|\cdot\|_2) \leq \left(\frac{2}{\varepsilon} + 1\right)^n$.*

Lemma 5. *Suppose that functions $\phi_i : \mathbb{R} \to \mathbb{R}$, $i = 1, 2, \dots, n$ are ρ-Lipschitz continuous. Then, for $V \subset V^n$, $N(\varepsilon, \vec{\phi} \circ V) \leq N(\frac{\varepsilon}{\rho}, V)$, where for $v \in \mathbb{R}^n$, $\vec{\phi}(v) := [\phi_1(v_1), \dots, \phi_n(v_n)]$, $\vec{\phi} \circ V := \{\vec{\phi}(v) \mid v \in V\}$.*

Proof of Theorem 8. To simplify the discussion, we will estimate the covering number of the following perceptron

$$f_{\mathrm{NN}}(u) = A_3\sigma_2[A_2\sigma_1(A_1u + b_1) + b_2] + b_3.$$

Because the proof for general cases is exactly the same, we need to estimate the covering number of the gradient of f_{NN}, which is written as

$$\nabla f_{\mathrm{NN}}(u) = A_1^\top (D\sigma_1) A_2^\top (D\sigma_2) A_3^\top,$$

where $D\sigma_2$ and $D\sigma_1$ are Jacobian matrices. These Jacobian matrices are evaluated at $u = A_2\sigma_1(A_1u + b_1) + b_2$ and $A_1u + b_1$, respectively. We first estimate the covering number associated with $D\sigma_2$. $D\sigma_2$ has the same architecture as a multi-layer perceptron, except that the last activation function is replaced by the differential σ'_2 of σ_2:

$$D\sigma_2 = \sigma'_2[A_2\sigma_1(A_1u + b_1) + b_2].$$

Assuming that the input data are in the ball B_{c_u} with radius c_u, we obtain

$$N(\varepsilon, B_{c_u}) \leq \left(\frac{2c_u}{\varepsilon} + 1\right)^n.$$

Then, because the norms of A_1, A_2 are bounded by c_{A_1}, c_{A_2}, matrix multiplications by these matrices are c_{A_1}- and c_{A_2}-Lipschitz continuous, respectively. In addition, σ_1 is ρ_1-Lipschitz and σ_2' is ρ_2'-Lipschitz continuous. Therefore, the covering number associated with $D\sigma_2$ is estimated by

$$N(\varepsilon, D\sigma_2) \leq \left(\frac{2\rho_1\rho_2' c_{A_1} c_{A_2} c_u}{\varepsilon} + 1\right)^n.$$

Finally, because σ_1' is assumed to be bounded by c_{σ_1}, the norms of the matrices other than $D\sigma_2$ in ∇f_{NN} are bounded as follows: $\|A_1^\top\| < c_{A_1}$, $\|A_2^\top\| < c_{A_2}$, $\|A_3^\top\| < c_{A_3}$, $\|D\sigma_1\| < c_{\sigma_1}$. Because the loss function is assumed to be ρ_p-Lipschitz continuous, we obtain the estimate

$$N(\varepsilon, \mathcal{G}) \leq \left(\frac{2\rho_p\rho_1\rho_2' c_{\sigma_1} (c_{A_1} c_{A_2})^2 c_{A_3} c_u}{\varepsilon} + 1\right)^n.$$

□

L^∞ Estimate on the Error in the Hamiltonian

The generalization error analysis in Theorem 8 shows that, at a certain probability, the expectation of the loss function can be bounded. If this bound certainty holds and if the training is performed by minimizing the p-norm with $p > 2M$, we can derive an L^∞ estimate on the Hamiltonian for the standard HNN (2) by applying the Poincaré inequality and the Sobolev inequality under Assumption 1.

Assumption 1 There exists a density f_P for measure P with $\inf f_P > 0$.

Remark 5. *The condition $p > 2M$ is not required in practice because of the well-known equivalence of the norms in finite dimensional spaces; for example, if the standard 2-norm is small enough, then the p-norm is also small. However, when the dimension $2M$ is large, the 2-norm needs to be very small to bound the p-norm because the constant in the inequality used to bound the p-norm depends on the dimension. Therefore, it is preferable to minimize the p-norm in such cases.*

Theorem 9 (Poincaré inequality)**.** *Suppose that $1 \leq p \leq \infty$ and $\Omega \subset \mathbb{R}^{2M}$ is bounded. Then there exists a constant c_{p} such that, for any $H \in S_p^1(\Omega)$,*

$$\int_\Omega |H(u) - \bar{H}|^p \mathrm{d}u \leq c_{\mathrm{p}} \left\| \frac{\partial H}{\partial u} \right\|_p^p, \ \bar{H} = \frac{1}{\int_\Omega \mathrm{d}u} \int_\Omega H(u)\mathrm{d}u.$$

The constant c_{p} is called the Poincaré constant.

Theorem 10 (Sobolev inequalities)**.** *There exist constants c_1, c_2 such that, if $lp > 2M$,*

$$\|e\|_{L^\infty}(\mathbb{R}^{2M}) \leq c\|e\|_{W^{p,l}}(\mathbb{R}^{2M}),$$
$$\|e\|_{L^\infty}(\mathbb{T}^{2M}) \leq c\|e\|_{W^{p,l}}(\mathbb{T}^{2M}).$$

By using these inequalities along with the invariance of the Hamilton equation under the constant shift of the energy function, we obtain an error bound on the Hamiltonian.

Lemma 6. *Among the energy functions that yield the target Hamilton equation, we choose the one for which*

$$\int H(u)\mathrm{d}u = \int H_{\mathrm{NN}}(u)\mathrm{d}u \tag{10}$$

holds, so that the error function has zero mean: $e(u) := H(u) - H_{\mathrm{NN}}(u)$, $\bar{e}(u) := 0$. Then,

$$\int_\Omega |e(u)|^p \mathrm{d}u \leq c_{\mathrm{p}} \left\| \frac{\partial e}{\partial u} \right\|_{L^p}^p.$$

From the above estimate, we obtain

$$\int \left\| \frac{\partial H(u)}{\partial u} - \frac{\partial H_{\mathrm{NN}}(u)}{\partial u} \right\|_p^p \mathrm{d}P$$
$$\leq \frac{1}{n} \sum_{i=1}^n L[Y_i, h(X_i)] + 2\mathcal{R}_n(\mathcal{G}) + 3c\sqrt{\frac{2\ln\frac{4}{\delta}}{n}}.$$

By using the density f_P for the measure P, we obtain

$$\inf f_P \int \left\| \frac{\partial H(u)}{\partial u} - \frac{\partial H_{\mathrm{NN}}(u)}{\partial u} \right\|_p^p \mathrm{d}u$$
$$\leq \int \left\| \frac{\partial H(u)}{\partial u} - \frac{\partial H_{\mathrm{NN}}(u)}{\partial u} \right\|_p^p \mathrm{d}P,$$

which gives us

$$\int \left\| \frac{\partial H(u)}{\partial u} - \frac{\partial H_{\mathrm{NN}}(u)}{\partial u} \right\|_p^p \mathrm{d}u$$
$$\leq \frac{1}{\inf f_P} \int \left\| \frac{\partial H(u)}{\partial u} - \frac{\partial H_{\mathrm{NN}}(u)}{\partial u} \right\|_p^p \mathrm{d}P$$
$$\leq \frac{1}{\inf f_P} \left(\frac{1}{n} \sum_{i=1}^n L[Y_i, h(X_i)] + 2\mathcal{R}_n(\mathcal{G}) + 3c\sqrt{\frac{2\ln\frac{4}{\delta}}{n}} \right).$$

We note that the left-hand side is the Sobolev norm of the error in $W^{p,l}$; then, under the assumption that $p > 2M$, we can use the Sobolev inequality to obtain

$$(\sup_u \|H(u) - H_{\mathrm{NN}}(u)\|)^p \leq c^p \left\| \frac{\partial H(u)}{\partial u} - \frac{\partial H_{\mathrm{NN}}(u)}{\partial u} \right\|_p^p$$
$$\leq \frac{c^p}{\inf f_P} \left(\frac{1}{n} \sum_{i=1}^n L[Y_i, h(X_i)] + 2\mathcal{R}_n(\mathcal{G}) + 3c\sqrt{\frac{2\ln\frac{4}{\delta}}{n}} \right),$$

which ensure that H and H_{NN} are close in terms of the function values.

KAM Theory for HNNs

The universal approximation property shown in the previous sections guarantees that the value of MSE can be made arbitrarily small by training; however, in actual training, a finite error remains. In this section, as an application of the generalization bound, we apply the KAM theory to theoretically investigate the trained standard HNN model (2) in such cases by assuming that the target system is integrable.

We make a few assumptions that are needed for the application of the KAM theory.

Assumption 2 The dimension of the phase space is assumed to be $2M$ with $M \geq 2$.

Assumption 3 The target system is an integrable Hamiltonian system with the conserved quantities $F_1, \ldots, F_M$. The

series $c_1, \ldots, c_M$ exists such that $K = \cap_{i=1}^{M} F_i^{-1}(c_i)$ is connected and compact.

Under the above assumptions, from the Liouville–Arnold theorem there exist a neighborhood $\mathcal{N}$ of $K, \mathcal{U} \subset \mathbb{R}^n$ and a coordinate transform

$$\phi : (\theta, J) \in \mathbb{T}^n \times \mathcal{U} \to \phi(\theta, J) \in \mathcal{N}, \tag{11}$$

such that the transformed system is the Hamilton equation. Following the usual setting of the KAM theorem, we consider the target system and the Hamiltonian equation in the transformed coordinate $\mathbb{T}^n \times \mathcal{U}$.

Assumption 4 The Hamiltonian $H : \mathbb{T}^n \times \mathcal{U} \to \mathbb{R}$ of the target system is C^∞ and non-degenerate. The activation functions of the HNNs used are in C^∞.

Assumption 5 From the generalization error analysis in the previous section, we have essentially shown that if $p > 2M$, with at least probability $1 - \delta$, it holds that

$$\sup |H(u) - H_{\mathrm{NN}}(u)| < c_1 L_{\mathrm{train}} + c_2 R_n + c_3 \sqrt{\frac{\ln \frac{1}{\delta}}{n}}$$

with constants c_1, c_2, and c_3, where R_n is a bound on the Rademacher complexity. We assume that the training was performed with $p > 2M$ and the above statement certainly holds.

Using these assumptions, we obtain Theorem 10.

Theorem 11. *Let the threshold of the KAM theorem be ε_0 and δ be*

$$\delta = \exp\left(-n\left(\frac{\varepsilon_0 - c_1 L_{\mathrm{train}} - c_2 R_n}{c_3}\right)^2\right).$$

Under the above assumptions, with a probability of at least $(1 - \delta)$, a set of invariant tori exists for the trained model H_{NN}.

Proof. It is confirmed by a straightforward calculation that if δ is given as described above, it holds that $\sup |H(u) - H_{\mathrm{NN}}(u)| < \varepsilon_0$, and hence the assumption of the KAM theorem is satisfied. □

Remark 6. *As mentioned in Remark 1, the KAM theorem also shows that the invariant tori become larger when the perturbation becomes smaller. Hence, if the generalization error is small, the size of the tori is expected to be large.*

Note that general Hamiltonian systems, and hence general HNNs, are not quasi-periodic. Therefore, a model that approximates a quasi-periodic Hamilton equation may be (in some sense) *approximately* quasi-periodic, but it is not necessarily *strictly* quasi-periodic. This theorem states that *the trained model can be strictly quasi-periodic even if the training loss does not completely vanish.*

Numerical Example: Learning the Zabusky and Kruskal Experiment As a numerical experiment, we trained a HNN[1] so that the dynamics of the KdV equation is learned by using the data from the experiment by Zabusky and Kruskal (1965), in which a nontrivial recurrence of initial states is reported.

The KdV equation is derived from the energy function

$$H(u) = \int \left[\frac{1}{6}\alpha u^3 - \frac{1}{2}\beta\left(\frac{\partial u}{\partial x}\right)^2\right] \mathrm{d}x.$$

In fact, under the periodic boundary condition, the variational derivative is

$$\frac{\delta H}{\delta u} = \int \left[\frac{1}{2}\alpha u^2 + \beta \frac{\partial^2 u}{\partial x^2}\right] \mathrm{d}x,$$

and the KdV equation is defined as a Hamiltonian equation:

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\frac{\delta H}{\delta u} = \alpha u \frac{\partial u}{\partial x} + \beta \frac{\partial^3 u}{\partial x^3}.$$

For spatial discretization, we used the forward and backward difference operators,

$$D_{\mathrm{f}} := \frac{1}{\Delta x}\begin{pmatrix} -1 & 1 & \cdots & 0 & 0 \\ 0 & -1 & \cdots & 0 & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & \cdots & -1 & 1 \\ 1 & 0 & \cdots & 0 & -1 \end{pmatrix} \text{ and}$$

$$D_{\mathrm{b}} := \frac{1}{\Delta x}\begin{pmatrix} 1 & 0 & \cdots & 0 & -1 \\ -1 & 1 & \cdots & 0 & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & -1 & 1 \end{pmatrix},$$

respectively. The central difference operator D is their mean, specifically $D = \frac{1}{2}(D_{\mathrm{f}} + D_{\mathrm{b}})$ and that for the second derivative is $D_2 = D_{\mathrm{f}} D_{\mathrm{b}} = D_{\mathrm{b}} D_{\mathrm{f}}$. Using these difference operators, the equation is semi-discretized as

$$H(u) = \sum_x \left[\frac{1}{6}\alpha u^3 - \frac{1}{2}\beta \frac{(D_{\mathrm{f}} u)^2 + (D_{\mathrm{b}} u)^2}{2}\right] \Delta x,$$

$$\frac{\mathrm{d}u}{\mathrm{d}t} = D\frac{\partial H}{\partial u} = D\left(\frac{1}{2}\alpha u^2 + \beta D_2 u\right).$$

Following Zabusky and Kruskal (1965), we set the parameters to $\alpha = -1.0$ and $\beta = -0.022^2$, set the width of phase space to 2.0, and used the initial condition $u(0, x)$ to $u(0, x) = \cos(x\pi)$. We discretized the system with the spatial and temporal mesh sizes of $\Delta x = 0.1$ and $\Delta t = 0.01$. We obtained an orbit for 200 time steps from the initial condition using the fifth-order Dormand–Prince method with the absolute and relative tolerances of 10^{-10} and 10^{-8}.

We performed the experiments on an NVIDIA TITAN V with double precision. We employed a three-layered convolutional neural network with kernel sizes of 3, 1, and 1. The number of hidden channels was 200, the number of output channels was 1, the activation function was the $\tanh$ function, and each weight parameter was initialized as a random orthogonal matrix. We summed up the output in the spatial direction and obtained the global energy. We used the whole orbit at every iteration, and minimized the mean squared error of the time derivative as the loss function using the Adam

[1] We use the HNN code for the KdV equation provided by https://github.com/tksmatsubara/discrete-autograd (MIT License).

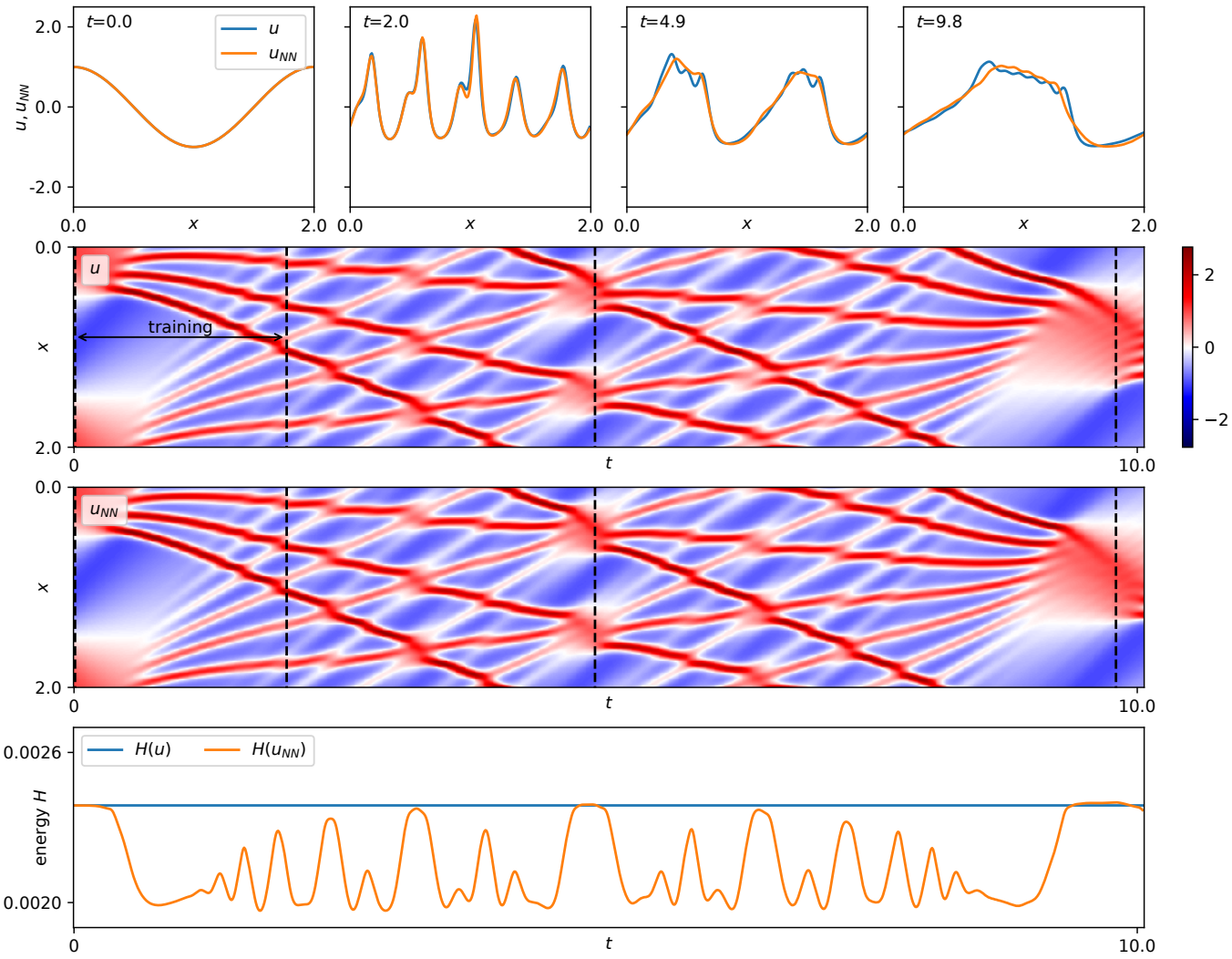


Figure 4: Results of training in the Zabusky and Kruskal experiment (Zabusky and Kruskal 1965). (top panels) The predicted states at t =0.0, 2.0, 4.8, and 9.8. (second panel) The true dynamics u. (third panel) The dynamics u_{NN} modeled by a neural network. (bottom panel) The energy function H given the true dynamics u and modeled dynamics u_{NN}.

optimizer with a learning rate of 10^{-3} for 10,000 iterations; the error reached a maximum of 1.37×10^{-3}. Given the true dynamics u, the absolute error between the energy function H and the neural network H_{NN} was 1.31×10^{-4} on average and 2.51×10^{-4} at most.

Using the true model and the trained neural network, we also obtained orbits for 1100 time steps from the same initial condition, as shown in the second and third panels of Fig. 4, respectively. In the top panels, blue and orange lines denote the true state u and the state predicted by the trained neural network u_{NN} at t =0.0, 2.0, 4.8, and 9.8. The bottom panel shows the energy function H given the predicted states u and u_{NN}. Due to the non-zero training error, more waves incur a larger error. Nonetheless, at around $t = 9.8$, the true model and learned neural network reproduce $\sin$ waves, which are given as the initial condition, and the energy error is restored to zero; they exhibit quasi-periodic behaviors.

Concluding Remarks

We analyzed the behavior of HNNs with non-zero learning errors by combining the KAM theory and statistical machine learning. We investigated the approximation properties of deep energy-based models, including HNNs. More precisely, we proved the persistence of the quasi-periodic behaviors of integrable Hamiltonian systems with a high probability even when the loss function is not perfectly zero. Further, we provided a generalization error bound and universal approximation theorems for HNNs to ensure that the loss function can be sufficiently small for application of the KAM theorem. Meanwhile, in the recent research on this type of model, numerically integrated gradients are often used for training. Similar results should be obtained for such cases; however, rigorous discussion is needed.

Acknowledgments

Funding in direct support of this work: JST CREST Grant Number JPMJCR1914, JST PRESTO Grant Number JPMJPR21C7 and JSPS KAKENHI Grant Number 20K11693.

References

Adams, R. A.; and Fournier, J. J. F. 2003. *Sobolev Spaces*. Elsevier.

Bharadwaj, P.; Li, M.; and Demanet, L. 2020. SymAE: An autoencoder with embedded physical symmetries for passive time-lapse monitoring. In *SEG Technical Program Expanded Abstracts 2020*. Society of Exploration Geophysicists.

Bondesan, R.; and Lamacraft, A. 2019. Learning Symmetries of Classical Integrable Systems. In *ICML 2019 Workshop on Theoretical Physics for Deep Learning*.

Bousquet, O.; Boucheron, S.; and Lugosi, G. 2004. Introduction to Statistical Learning Theory. In Bousquet, O.; von Luxburg, U.; and Rätsch, G., eds., *Advanced Lectures on Machine Learning*, 169–207. Berlin, Heidelberg: Springer Berlin Heidelberg.

Chen, T. Q.; Rubanova, Y.; Bettencourt, J.; Duvenaud, D.; Chen, R. T. Q.; Rubanova, Y.; Bettencourt, J.; and Duvenaud, D. 2018. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chen, Z.; Zhang, J.; Arjovsky, M.; and Bottou, L. 2020. Symplectic Recurrent Neural Networks. In *International Conference on Learning Representations (ICLR)*.

Cranmer, M.; Greydanus, S.; Hoyer, S.; Battaglia, P.; Spergel, D.; and Ho, S. 2020. Lagrangian Neural Networks. *ICLR 2020 Deep Differential Equations Workshop*.

Desai, S.; and Roberts, S. 2020. VIGN: Variational Integrator Graph Networks. *arXiv:2004.13688*.

DiPietro, D. M.; Xiong, S.; and Zhu, B. 2020. Sparse Symplectically Integrated Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Féjoz, J. 2013. On "Arnold's theorem" on the stability of the solar system. *Discrete & Continuous Dynamical Systems*, 33(8): 3555–3565.

Feng, Y.; Wang, H.; Yang, H.; and Wang, F. 2020. Time-Continuous Energy-Conservation Neural Network for Structural Dynamics Analysis. *arXiv:2012.14334*.

Finzi, M.; Stanton, S.; Izmailov, P.; and Wilson, A. G. 2020. Generalizing Convolutional Neural Networks for Equivariance to Lie Groups on Arbitrary Continuous Data. In *International Conference on Machine Learning (ICML)*, 3165–3176.

Finzi, M.; Wang, K. A.; and Wilson, A. G. 2020. Simplifying Hamiltonian and Lagrangian Neural Networks via Explicit Constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Furihata, D.; and Matsuo, T. 2010. *Discrete Variational Derivative Method: A Structure-Preserving Numerical Method for Partial Differential Equations*. Chapman and Hall/CRC.

Galioto, N.; and Gorodetsky, A. A. 2020. Bayesian Identification of Hamiltonian Dynamics from Symplectic Data. In *IEEE Conference on Decision and Control (CDC)*, 1190–1195.

Giné, E.; and Nickl, R. 2016. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press.

Greydanus, S.; Dzamba, M.; and Yosinski, J. 2019. Hamiltonian Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Jin, P.; Zhang, Z.; Kevrekidis, I. G.; and Karniadakis, G. E. 2020a. Learning Poisson systems and trajectories of autonomous systems via Poisson neural networks. *arXiv:2012.03133*.

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

Jin, Z.; Lin, J. Y.-Y.; and Li, S.-F. 2020. Learning Principle of Least Action with Reinforcement Learning. *arXiv:2011.11891*.

Laskar, J. 1996. Large scale chaos and marginal stability in the solar system. volume 64, 115–162. Chaos in gravitational N-body systems (La Plata, 1995).

Li, S.-H.; Dong, C.-X.; Zhang, L.; and Wang, L. 2020. Neural Canonical Transformation with Symplectic Flows. *Physical Review X*, 10(2): 021020.

Matsubara, T.; Ishikawa, A.; and Yaguchi, T. 2020. Deep Energy-Based Modeling of Discrete-Time Physics. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Rana, M. A.; Li, A.; Fox, D.; Boots, B.; Ramos, F.; and Ratliff, N. 2020. Euclideanizing Flows: Diffeomorphic Reduction for Learning Stable Dynamical Systems. In *Conference on Learning for Dynamics and Control (L4DC)*, volume 120.

Sæmundsson, S.; Hofmann, K.; Terenin, A.; and Deisenroth, M. P. 2019. Variational integrator networks for physically meaningful embeddings. In *Artificial Intelligence and Statistics (AISTATS)*, volume 108, 3078–3087.

Sanchez-Gonzalez, A.; Bapst, V.; Cranmer, K.; and Battaglia, P. 2019. Hamiltonian Graph Networks with ODE Integrators. *arXiv:1909.12790*.

Scott Dumas, H. 2014. *KAM Story, The: A Friendly Introduction To The Content, History, And Significance Of Classical Kolmogorov-Arnold-Moser Theory*. World Scientific Publishing Company.

Shalev-Shwartz, S.; and Ben-David, S. 2014. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

Steinwart, I.; and Christmann, A. 2008. *Support Vector Machines*. Springer Science & Business Media.

Toth, P.; Rezende, D. J.; Jaegle, A.; Racanière, S.; Botev, A.; and Higgins, I. 2019. Hamiltonian Generative Networks. In *International Conference on Learning Representations (ICLR)*.

Xiong, S.; Tong, Y.; He, X.; Yang, C.; Yang, S.; and Zhu, B. 2021. Nonseparable Symplectic Neural Networks. In *International Conference on Learning Representations (ICLR)*.

Zabusky, N. J.; and Kruskal, M. D. 1965. Interaction of "Solitons" in a Collisionless Plasma and the Recurrence of Initial States. *Phys. Rev. Lett.*, 15(6): 240–243.

Zhong, Y. D.; Dey, B.; and Chakraborty, A. 2020. Symplectic ODE-Net: Learning Hamiltonian Dynamics with Control. In *ICLR*.