Variational Principle and Variational Integrators for Neural Symplectic Forms

Yuhan Chen [*1] **Baige Xu** [*2] **Takashi Matsubara** [*3] **Takaharu Yaguchi** [*2]

Abstract

In this study, we investigate the variational principle for neural symplectic forms, thereby designing the variational integrators for this model. In recent years, neural networks models for physical phenomena have been attracting much attention. In particular, the neural symplectic form is a method that can model general Hamiltonian systems, which are not necessary in the canonical form. In this paper, we make the following two contributions regarding this model. Firstly, we show that this model is derived from a variational principle and hence admits the Noether theorem. Secondly, when the trained models are used for simulations, they must be discretized using numerical integrators; however, unless carefully designed, numerical integrators destroy physical laws. We propose variational integrators for the neural symplectic forms, which are numerical integrators that preserve the laws of physics.

1. Introduction

In recent years, research on deep learning techniques to formulate models from observed data of physical phenomena have been attracting much attention. The techniques include the Lagrangian neural network (Cranmer et al., 2020) and the Hamiltonian neural network (Greydanus et al., 2019). In particular, following the Hamilton mechanics, the Hamiltonian neural network(HNN) uses the following Hamilton equation for modeling the dynamics:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} O & I \\ -I & O \end{pmatrix} \nabla H, \tag{1}$$

where q is the generalized coordinates and p is the generalized momenta. The Hamiltonian H is the total energy of the system and this function is learned from the data in HNN. By employing the equation of this form as the model, HNN ensures the energy conservation law, thereby improving the long-term prediction. Nonetheless, the above form, which is called the canonical form of the Hamilton equation, is only satisfied in the Darboux coordinates, that is, the generalized coordinate and the generalized momenta. This means that the training data for HNN is usually unavailable since the analytic form of the generalized momenta is in general unknown without the detailed knowledge on the target system. To overcome this limitation, the neural symplectic form (NSF) was proposed (Chen et al., 2021). NSF uses the coordinate-free expression of the Hamilton equation and can be trained by using the observation data in general coordinate systems.

We make two contributions to this model. Firstly, we show that this model is derived from the variational principle. In NSF, the 1-form that derives the symplectic 2-form is modeled by neural networks. It was explained in Chen et al. (2021) that in this way, models other than Hamiltonian systems would not be explored, and hence this model is trained efficiently. However, what is more important is that this model always yields Hamiltonian systems, which implies that a variational principle exists in this model. The variational principle is a fundamental principle of analytical mechanics. In fact, most of the important physical laws are derived from this principle. For example, the Noether theorem, which states that if the system has a symmetry there exists a conservation law corresponding to this symmetry, follows from the variational principle. We show that a variational principle for NSF certainly exists, and this model admits the Noether theorem.

Meanwhile, a primal application of the deep physical models is physical simulations, and to perform the simulations, these models must be discretized. However, the application of typical numerical integrators destroys the Hamilton structure of the models along with the physical properties such as the energy conservation law (Hairer et al.,

[*]Equal contribution [1]Graduate School of System Informatics, Kobe University, 1-1 Rokkodai-cho, Nada-ku, Kobe, 657-8501, Japan [2]Graduate School of Science, Kobe University, 1-1 Rokkodai-cho, Nada-ku, Kobe, 657-8501, Japan [3]Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama-cho, Toyonaka-shi, Osaka, 560-8531 Japan. Correspondence to: Yuhan Chen <193x226x@stu.kobe-u.ac.jp>, Baige Xu <baigexu@stu.kobe-u.ac.jp>, Takashi Matsubara <matsubara@sys.es.osaka-u.ac.jp>, Takaharu Yaguchi <yaguchi@pearl.kobe-u.ac.jp>.

Workshop on New Frontiers in Learning, Control, and Dynamical Systems at the International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA, 2023.

2013). The discrete-time models have been considered, as such models do not require further discretization for simulations (see, e.g., Matsubara et al. (2020); Saemundsson et al. (2020); Desai & Roberts (2020); Xiong et al. (2020); Jin et al. (2020); Santos et al. (2022); Offen & Ober-Blöbaum (2022).) On the other hand, discrete-time models have a disadvantage of being unable to be used in the simulations with time steps other than that used in the training process. As a result, studies on the methods to discretize continuous-time deep physical models for simulation are important. In addition, such methods are also useful for designing discrete-time models.

The numerical integrators that preserve the mechanical structure are called the structure-preserving integrators or the geometric integrators. Among them, symplectic integrators are the most important. They are designed so that the symplectic form is preserved, thereby preserving the physical properties. However, most symplectic integrators are designed for the Hamilton equation of the canonical form, and they may not be available for NSF because the Hamilton equations of the neural symplectic forms are not canonical. Another way to design a symplectic integrator is the variational integrator for Lagrangian mechanics (Marsden & West, 2001). By discretizing the variational principle, the variational integrator discretizes the Euler–Lagrange equation while preserving various conservation laws. In this paper, we show that this integrator is available for NSF.

The main contributions of this paper include:

- Finding a variational principle for NSF, which implies that the Noether theorem holds for this model.
- Developing variational integrators for NSF, which are numerical integrators that admits the energy conservation law and the discrete Noether theorem.

2. Related work

The related work includes the following studies: as the discrete-time models mentioned above, there have been many studies of discrete-time models, such as Matsubara et al. (2020); Saemundsson et al. (2020); Desai & Roberts (2020); Xiong et al. (2020); Jin et al. (2020); Santos et al. (2022); Offen & Ober-Blöbaum (2022). In addition, other methods such as (Rath et al., 2021) using Gaussian process regression. However, to the best of the authors' knowledge, there has been no research on the variational principle or variational integrators for neural network models that represent the Hamilton equation.

3. Variational Integrator

The variational integrator was proposed by West–Marsden as a discretization method for the Euler–Lagrange equation (Marsden & West, 2001). In the Lagrangian mechanics, the Euler–Lagrange equation, which is the equation of motion, is derived from the variational principle (Abraham & Marsden, 2008). In addition, most properties, such as the energy conservation law and the Noether theorem, of this equation are also derived from this principle (Marsden & Ratiu, 2013). The variational integrator uses the discrete variational principle to derive numerical integrators, thereby preserving the physical properties.

More precisely, given a Lagrangian $L(q, \dot{q})$, the Euler–Lagrange equation is derived by computing the variation of the action integral $\int L(q, \dot{q})\mathrm{d}t$. Instead of computing the variation of this integral, to derive the variational integrator, the variation of the action sum $\sum_n L_\mathrm{d}(q^{(n)}, q^{(n+1)})\Delta t$ is typically computed. Δt is the time step size and $q^{(n)}$ is an approximation of $q(n\Delta t)$. $L_\mathrm{d}(q^{(n)}, q^{(n+1)})$ is any discretization of the Lagrangian $L(q, \dot{q})$. This discrete variational calculus gives a discrete approximation of the equations of motion, which is used as a numerical integrator. It is known that numerical integrators by this process admit the energy conservation law and the discrete Noether theorem. See Marsden & West (2001) for details.

4. Neural Symplectic Forms

Sympletic 2-Form on manifolds In this section, we briefly explain NSF. In terms of geometry, the Hamilton equation is defined as a flow on a symplectic manifold, which is a pair of a manifold and a symplectic 2-form. Because a flow is defined in a coordinate-free form, the Hamilton equation can be defined in a coordinate-free manner as well (Abraham & Marsden, 2008; Marsden & Ratiu, 2013). NSF uses this coordinate-free form of the Hamilton equation.

Suppose that the phase space is $\mathcal{M} = \mathbb{R}^{2N}$. The differential 0-form on $\mathbb{R}^{2N}$ denotes the function mapping from $\mathbb{R}^{2N}$ to $\mathbb{R}$. The differential 1-form on $\mathbb{R}^{2N}$ is a linear function defined at each point $u \in \mathbb{R}^{2N}$, and it maps a vector $v \in \mathbb{R}^{2N}$ to a real number. In general, because any linear function from $v \in \mathbb{R}^{2N}$ to $\mathbb{R}$ can be expressed as an inner product with a vector, any differential 1-form can be expressed as a vector field depending on u. For differential forms, a differential operation called the exterior derivative d is defined. The exterior derivative maps a differential k-form to a differential $(k+1)$-form. This operator has a remarkable property: $\mathrm{d}\mathrm{d} = 0$. A differential 2-form ω is a skew-symmetric bilinear function $\omega : \mathbb{R}^{2N} \times \mathbb{R}^{2N} \to \mathbb{R}$ that is defined at each $u \in \mathbb{R}^{2N}$. Using an appropriate matrix, any skew-symmetric bilinear function can be expressed as $\omega_u(v_1, v_2) = {v_1}^T W_u v_2,\ u \in \mathbb{R}^{2N},\ (v_1, v_2) \in \mathbb{R}^{2N} \times \mathbb{R}^{2N}$. where the subscript u denotes that ω or the matrix W depends on u. Some definitions related to differential 2-form are as follows:

Definition 4.1. A differential form ω is closed if $\mathrm{d}\omega$ =0.

Definition 4.2. A differential 2-form ω is non-degenerate if the skew matrix associated with ω is non-degenerate.

Definition 4.3. A symplectic 2-form is a closed and non-degenerate differential 2-form.

Using the above, the Hamilton equation can be expressed as

$$\frac{\mathrm{d}u}{\mathrm{d}t} = X_H,\ \omega(X_H, v) = \mathrm{d}H(v) \text{ for all } v \in \mathbb{R}^{2N}. \qquad (2)$$

By replacing the 2-form ω with the skew-symmetric matrix, this equation becomes

$$\frac{\mathrm{d}u}{\mathrm{d}t} = X_H,\ {X_H}^T W v = v \cdot \nabla H \Leftrightarrow \frac{\mathrm{d}u}{\mathrm{d}t} = W_u^{-T} \nabla H.$$

Neural Symplectic Form The equation (2) holds regardless of the coordinate system. Therefore, by using (2) as a model, as long as the given data is described by the Hamilton equation, it is possible to learn both the symplectic 2-form and the Hamiltonian that define the Hamilton equation, no matter what coordinate system the data is given in.

To use (2) as a model, the symplectic 2-form or the matrix W must be learned from data. Although a differential 2-form corresponds to a skew-symmetric matrix, not all skew-symmetric matrices define a symplectic 2-form; symplectic 2-forms must be closed; in other words, it must be in $\mathrm{Ker}\ \mathrm{d}$. The 2-form modeled by NSF is guaranteed to be closed in the following way. Actually, according to the de Rham theorem, when the phase space is $\mathbb{R}^{2N}$, it holds that $\mathrm{Im}\ \mathrm{d} = \mathrm{Ker}\ \mathrm{d}$. The difference between these two spaces $\mathrm{Im}\ \mathrm{d}/\mathrm{Ker}\ \mathrm{d}$ is called the cohomology space. The de Rham theorem states that the cohomology space is isomorphic to the homology space, which is roughly a space of spatial holes. Because $\mathbb{R}^{2N}$ contains no holes, the homology space must vanish, and hence $\mathrm{Im}\ \mathrm{d} = \mathrm{Ker}\ \mathrm{d}$ holds. Even when the phase space has a hole, in many cases the space can be embedded in a large Euclid space without holes, and the model would work on such a space.

For this reason, instead of learning it directly, the symplectic 2-form is obtained by learning the differential 1-form θ and computing its exterior derivative $\hat{\omega} = d\theta$ by automatic differentiation. By doing so, the property $\mathrm{dd} = 0$ of the exterior derivative guarantees that the learned 2-form is closed. The following is the model of NSF:

$$\hat{\omega} = \mathrm{d}\theta_{\mathrm{NN}}, \quad \frac{\mathrm{d}u}{\mathrm{d}t} = \hat{X}_{H_{\mathrm{NN}}},$$
$$\hat{\omega}(\hat{X}_{H_{\mathrm{NN}}}, v) = \mathrm{d}H_{\mathrm{NN}}(v) \text{ for all } v \in \mathbb{R}^{2N}. \qquad (3)$$

As mentioned above, the 1-form θ_{NN} and the 2-form $\hat{\omega}$ can be represented as a vector Y_{NN} and a matrix $\hat{W}_u$. Thus the model expressed in terms of vectors and a matrix, without using the differential forms is given as

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \hat{W}_u^{-T} \nabla H_{\mathrm{NN}}(u),$$
$$(\hat{W}_u)_{i,j} = \frac{\partial (Y_{\mathrm{NN}})_i}{\partial u_j} - \frac{\partial (Y_{\mathrm{NN}})_j}{\partial u_i}. \qquad (4)$$

5. Variational principle for neural symplectic forms

In NSF, skew-symmetric matrices that do not correspond to symplectic forms are not explored. Hence, in Chen et al. (2021) it is explained that this model can be trained much more efficiently than learning general skew-symmetric matrices; however, more important fact is that the model (3) is always Hamiltonian and hence there should be a variational principle that derives this model.

As is well-known the Hamilton equation (1) in the canonical form is derived from the variational principle using the following action integral (Abraham & Marsden, 2008):

$$\int_0^T (p \cdot \dot{q} - H(q,p))\,\mathrm{d}t \qquad T \in \mathbb{R}. \qquad (5)$$

We investigate a similar principle for the NSF. First, we focus on the first term $p\mathrm{d}q$ of the action integral (5). Taking the exterior derivative of it gives the symplectic form: $\mathrm{d}(p\mathrm{d}q) = -\mathrm{d}q \wedge \mathrm{d}p$, which is similar to $-\theta_{\mathrm{NN}}$ in NSF. Therefore, instead of (5), we should consider the following integral:

$$\int_0^T (-\theta_{\mathrm{NN}}(\dot{u}) - H_{\mathrm{NN}})\,\mathrm{d}t, \qquad (6)$$

In fact, the variational principle associated with this integral derives NSF.

Theorem 5.1. *The model* (3) *is derived by requiring the derivative of the action integral* (6) *to vanish with* $u(0)$ *and* $u(T)$ *fixed.*

Proof. Roughly, ignoring the higher-order terms, for any Δu, we get

$$\int_0^T (-\theta_{\mathrm{NN}}(\dot{u}) - \theta_{\mathrm{NN}}(\dot{u} + \Delta\dot{u})$$
$$-H_{\mathrm{NN}}(u + \Delta u))\,\mathrm{d}t - \int_0^T (-\theta_{\mathrm{NN}}(\dot{u}) - H_{\mathrm{NN}})\,\mathrm{d}t$$
$$= \int_0^T \left[\left(\dot{u}^T (\frac{\partial \theta_{\mathrm{NN}}}{\partial u}^T - \frac{\partial \theta_{\mathrm{NN}}}{\partial u}) - \frac{\partial H_{\mathrm{NN}}}{\partial u} \right) \Delta u \right] \mathrm{d}t.$$

For the right-hand side to vanish for any Δu, $\dot{u}^T (\frac{\partial \theta_{\mathrm{NN}}}{\partial u}^T - \frac{\partial \theta_{\mathrm{NN}}}{\partial u}) - \frac{\partial H_{\mathrm{NN}}}{\partial u}$ must vanish, which results in (4). See Appendix A for details. □

This theorem means that the laws of physics, particularly, the Noether theorem holds for NSF. Therefore, if the action integral (3) admits a symmetry, NSF has a corresponding conservation law.

6. Variational integrators for neural symplectic forms

Note that all quantities needed to define the action integral (6) are available in the NSF model (3); hence the action integral (6) can be used to derive variational integrators.

Although any discretization can be employed, we discretize the above action integral (6), for example, as

$$\sum_{n=0}^{N}\Bigg(-\theta_{\mathrm{NN}}(\frac{u^{(n+1)}-u^{(n)}}{\Delta t};\frac{u^{(n+1)}+u^{(n)}}{2}) -H_{\mathrm{NN}}(\frac{u^{(n+1)}+u^{(n)}}{2})\Bigg)\Delta t, \tag{7}$$

where Δt is the time step size and $u^{(n)}$ is an approximation of $u(n\Delta t)$. $\theta_{\mathrm{NN}}(v_1;v_2)$ denotes the linear function $v_1 \in \mathbb{R}^{2N} \mapsto \theta_{\mathrm{NN}}(v_1;v_2) \in \mathbb{R}$ at each point $v_2 \in \mathbb{R}^{2N}$.

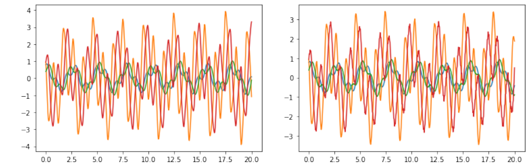
Taking the discrete variation of the action sum (7) with respect to the infinitesimal perturbations $\Delta u^{(n)}, \Delta u^{(n+1)}$, and considering any variation to be zero, we require the following equality:

$$\begin{aligned}
&\frac{1}{\Delta t}\Bigg(D_1\theta_{\mathrm{NN}}(\frac{u^{(n+1)}-u^{(n)}}{\Delta t};\frac{u^{(n+1)}+u^{(n)}}{2}) \\
&\quad -D_1\theta_{\mathrm{NN}}(\frac{u^{(n)}-u^{(n-1)}}{\Delta t};\frac{u^{(n)}+u^{(n-1)}}{2})\Bigg) \\
&+\frac{1}{2}\Bigg(-D_2\theta_{\mathrm{NN}}(\frac{u^{(n+1)}-u^{(n)}}{\Delta t};\frac{u^{(n+1)}+u^{(n)}}{2}) \\
&\quad - DH_{\mathrm{NN}}(\frac{u^{(n+1)}+u^{(n)}}{2}) \\
&\quad -D_2\theta_{\mathrm{NN}}(\frac{u^{(n)}-u^{(n-1)}}{\Delta t};\frac{u^{(n)}+u^{(n-1)}}{2}) \\
&\quad -DH_{\mathrm{NN}}(\frac{u^{(n)}+u^{(n-1)}}{2})\Bigg) = 0,
\end{aligned}$$

where DH_{NN} is the derivative of the neural network H_{NN}, and $D_1\theta_{\mathrm{NN}}$ and $D_2\theta_{\mathrm{NN}}$ are derivatives with respect to the first and second variables of the neural network θ_{NN}, respectively. See Appendix B for the detailed derivation of the above numerical scheme.

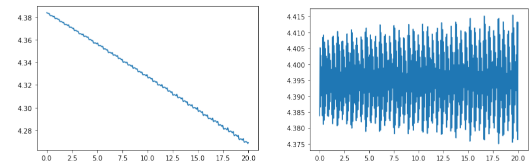
7. Numerical Experiments

We tested the proposed method in the following numerical experiment. Firstly, we trained the NSF on a Hamiltonian system, a double pendulum (see Appendix C for details), using the code and data published in Chen et al. (2021)[1]. The Hamiltonian H_{NN} and the 1-form θ_{NN} were modeled by using a neural network with two hidden layers of 200 units and the tanh activation function. We trained the model using the Adam optimizer with a learning rate of 10^{-3} for 2000 iterations. All computations are performed by using NVIDIA A100.



(a) The Heun method (b) Variational integrator

Figure 1. Simulated orbits. Each component of $u(t) = (\theta_1(t), \phi_1(t), \theta_2(t), \phi_2(t))$ is represented: blue (θ_1), orange (ϕ_1), green (θ_2), and red (ϕ_2).The horizontal axis represents the time and the vertical axis represents the values of the variables.



(a) The Heun method (b) Variational integrator

Figure 2. The evolution of the energy predicted by the proposed and the Heun method. The horizontal axis represents time and the vertical axis represents energy.

Next, the model was discretized using the proposed method. The proposed method is symmetric and hence has a second-order accuracy (Hairer et al., 2013). Therefore, we compared it with the Heun method, which also has a second-order accuracy. The time step Δt was set to 0.04. The calculated trajectories and the energies learned by the neural network are shown in Figures 1 and 2, respectively. The trajectories are not significantly different; in fact, NSF preserves the laws of physics through the variational principle, and hence even the Heun method yields good results. However, the energy graphs show that the energy by the Heun method gradually increases, while that by the proposed method oscillates, neither diverging nor decaying. This confirms that the proposed variational integrator does not destroy the energy conservation law.

8. Concluding Remarks

In recent years, the neural network modeling of physical phenomena has been widely studied. Among them, NSF

[1] https://github.com/YuhanChen0805/neural_symplectic_form (MIT License)

is a practical method in that the model can be trained from data given in arbitrary coordinate systems. In this study, we have investigated the variational principle that derives NSF. This shows that NSF admits the physical properties, particularly the Noether theorem. In addition, taking the variation of the discretized action integral expressed by the neural networks, we have proposed variational integrators for NSF. The numerical experiment has confirmed the conservation of energy. Meanwhile, only the simplest action sum is used in this study. Future work includes the development of more accurate numerical integrators by discretizing the action integral in a more sophisticated way.

Acknowledgements

This work is supported by JST CREST Grant Number JPMJCR1914, JST PRESTO Grant Number JPMJPR21C7 and JSPS KAKENHI Grant Number 20K11693 and JSPS 23KJ1555.

References

Abraham, R. and Marsden, J. E. *Foundations of Mechanics*. American Mathematical Soc., 2008.

Chen, Y., Matsubara, T., and Yaguchi, T. Neural symplectic form: Learning hamiltonian equations on general coordinate systems. In *Advances in Neural Information Processing Systems*, 2021.

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

Desai, S. and Roberts, S. VIGN: variational integrator graph networks. *CoRR*, abs/2004.13688, 2020.

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

Hairer, E., Lubich, C., and Wanner, G. *Geometric numerical integration: Structure-preserving algorithms for ordinary differential equations*. Springer Series in Computational Mathematics. Springer, Berlin, Germany, 2002 edition, March 2013.

Jin, P., Zhang, Z., Zhu, A., Tang, Y., and Karniadakis, G. E. Sympnets: Intrinsic structure-preserving symplectic networks for identifying hamiltonian systems. *Neural networks : the official journal of the International Neural Network Society*, 132:166–179, 2020.

Marsden, J. E. and Ratiu, T. S. *Introduction to Mechanics and Symmetry: A Basic Exposition of Classical Mechanical Systems*. Springer Science & Business Media, March 2013.

Marsden, J. E. and West, M. Discrete mechanics and variational integrators. *Acta Numerica*, 10:357–514, 2001.

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

Offen, C. and Ober-Blöbaum, S. Symplectic integration of learned Hamiltonian systems. *Chaos*, 32(1):013122, January 2022. doi: 10.1063/5.0065913.

Rath, K., Albert, C. G., Bischl, B., and von Toussaint, U. Symplectic Gaussian process regression of maps in Hamiltonian systems. *Chaos*, 31(5):053121, May 2021. doi: 10.1063/5.0048129.

Saemundsson, S., Terenin, A., Hofmann, K., and Deisenroth, M. Variational integrator networks for physically structured embeddings. In *International Conference on Artificial Intelligence and Statistics*, pp. 3078–3087. PMLR, 2020.

Santos, S., Ekal, M., and Ventura, R. Symplectic Momentum Neural Networks – Using Discrete Variational Mechanics as a prior in Deep Learning. *arXiv e-prints*, art. arXiv:2201.08281, January 2022.

Xiong, S., Tong, Y., He, X., Yang, S., Yang, C., and Zhu, B. Nonseparable Symplectic Neural Networks. *arXiv e-prints*, art. arXiv:2010.12636, October 2020.

A. The variational principle for the neural symplectic forms

We consider the variation of the action integral (6) of NSF with infinitesimal perturbations Δu with respect to u with the both ends fixed $\Delta u(0) = \Delta u(T) = 0$:

$$\int_0^T \left(-\theta_{\mathrm{NN}}(u+\Delta u)\cdot(\dot{u}+\Delta\dot{u}) - H_{\mathrm{NN}}(u+\Delta u)\right)\mathrm{d}t - \int_0^T \left(-\theta_{\mathrm{NN}}(u)\cdot(\dot{u}) - H_{\mathrm{NN}}(u)\right)\mathrm{d}t.$$

Applying the Taylor expansion to the above equation, we get

$$\int_0^T \left(-\theta_{\mathrm{NN}}(u)\cdot\Delta\dot{u} - \frac{\partial\theta_{\mathrm{NN}}}{\partial u}\Delta u\cdot\dot{u} - \frac{\partial H_{\mathrm{NN}}}{\partial u}\Delta u\right)\mathrm{d}t,$$

where higher-order terms of Δu are omitted. Next, by the integration by parts we obtain

$$\int_0^T \left(\frac{\mathrm{d}}{\mathrm{d}t}\theta_{\mathrm{NN}}(u)\cdot\Delta u - \dot{u}^T\frac{\partial\theta_{\mathrm{NN}}}{\partial u}\Delta u - \frac{\partial H_{\mathrm{NN}}}{\partial u}\Delta u\right)\mathrm{d}t,$$

where we used $\Delta u(0) = \Delta u(T) = 0$. Then, from the chain rule, we have

$$\int_0^T \left[\left(\dot{u}^T\left(\frac{\partial\theta_{\mathrm{NN}}}{\partial u}^T - \frac{\partial\theta_{\mathrm{NN}}}{\partial u}\right) - \frac{\partial H_{\mathrm{NN}}}{\partial u}\right)\Delta u\right]\mathrm{d}t.$$

For this variation to be zero for any Δu, the following equation must be satisfied

$$\dot{u}^T\left(\frac{\partial\theta_{\mathrm{NN}}}{\partial u}^T - \frac{\partial\theta_{\mathrm{NN}}}{\partial u}\right) - \frac{\partial H_{\mathrm{NN}}}{\partial u} = 0$$
$$\iff \quad \left(\frac{\partial\theta_{\mathrm{NN}}}{\partial u}^T - \frac{\partial\theta_{\mathrm{NN}}}{\partial u}\right)^T\dot{u} = \frac{\partial H_{\mathrm{NN}}}{\partial u}.$$

It follows that

$$\dot{u} = \tilde{W}_u^{-\top}\nabla H_{\mathrm{NN}}(u), \quad (\tilde{W}_u)_{i,j} = \frac{\partial(\theta_{\mathrm{NN}})_i}{\partial u_j} - \frac{\partial(\theta_{\mathrm{NN}})_j}{\partial u_i},$$

which is same as the equation of NSF (4).

B. Derivation of the variational integrator for the neural symplectic forms

We compute the variation of discretized action integral (7) with respect to the infinitesimal perturbations $\Delta u^{(n)}$ of $u^{(n)}$ under the assumption that $\Delta u^{(0)} = \Delta u^{(N)} = 0$:

$$\sum_{n=0}^{N}\Bigg(-\theta_{\mathrm{NN}}\Big(\frac{u^{(n+1)}+\Delta u^{(n+1)}-u^{(n)}-\Delta u^{(n)}}{\Delta t};\frac{u^{(n+1)}+\Delta u^{(n+1)}+u^{(n)}+\Delta u^{(n)}}{2}\Big) - H_{\mathrm{NN}}\Big(\frac{u^{(n+1)}+\Delta u^{(n+1)}+u^{(n)}+\Delta u^{(n)}}{2}\Big)\Bigg)\Delta t$$
$$-\sum_{n=0}^{N}\Bigg(-\theta_{\mathrm{NN}}\Big(\frac{u^{(n+1)}-u^{(n)}}{\Delta t};\frac{u^{(n+1)}+u^{(n)}}{2}\Big) - H_{\mathrm{NN}}\Big(\frac{u^{(n+1)}+u^{(n)}}{2}\Big)\Bigg)\Delta t.$$

Applying the Taylor expansion to the 1-form θ_{NN} and the Hamiltonian H_{NN}, we get

$$\sum_{n=0}^{N}\Bigg(-D_1\theta_{\mathrm{NN}}\Big(\frac{u^{(n+1)}-u^{(n)}}{\Delta t};\frac{u^{(n+1)}+u^{(n)}}{2}\Big)\frac{\Delta u^{(n+1)}-\Delta u^{(n)}}{\Delta t} - D_2\theta_{\mathrm{NN}}\Big(\frac{u^{(n+1)}-u^{(n)}}{\Delta t};\frac{u^{(n+1)}+u^{(n)}}{2}\Big)\frac{\Delta u^{(n+1)}+\Delta u^{(n)}}{2} - DH_{\mathrm{NN}}\Big(\frac{u^{(n+1)}+u^{(n)}}{2}\Big)\frac{\Delta u^{(n+1)}+\Delta u^{(n)}}{2}\Bigg)\Delta t,$$

where DH_{NN} is the derivative of H_{NN}; $D_1\theta_{\mathrm{NN}}$ and $D_2\theta_{\mathrm{NN}}$ are derivatives with respect to the first and second variables of the θ_{NN}, respectively. The higher-order terms of $\Delta u^{(n)}$'s are omitted. Since θ_{NN} is a 1-form, θ_{NN} defines a linear map that maps a vector v to a real number $\theta_{\mathrm{NN}}(v;u)$ for each u. Because this is a linear map, there exists a vector $\vec{\theta}_{\mathrm{NN}}(u)$ such that $\theta_{\mathrm{NN}}(v;u) = \vec{\theta}_{\mathrm{NN}}(u)\cdot v$. Note that although $\vec{\theta}$ is a linear map with respect to the vector v, $\vec{\theta}$ can be nonlinearly dependent on u. By using this expression, $D_1\theta_{\mathrm{NN}}$ and $D_2\theta_{\mathrm{NN}}$ are given as

$$D_1\theta_{\mathrm{NN}}(v;u) = \frac{\partial}{\partial v}\vec{\theta}_{\mathrm{NN}}(u) = \vec{\theta}_{\mathrm{NN}}(u)$$
$$D_2\theta_{\mathrm{NN}}(v;u) = \frac{\partial}{\partial u}\vec{\theta}_{\mathrm{NN}}(u)\cdot v = J^T v,$$

where J is the Jacobian matrix: $J = \frac{\partial\vec{\theta}_{\mathrm{NN}}}{\partial u}$.

Rearranging the above equation, we get

$$\sum_{n=0}^{N}\Bigg[\Bigg(\frac{1}{\Delta t}D_1\theta_{\rm NN}(\frac{u^{(n+1)}-u^{(n)}}{\Delta t};\frac{u^{(n+1)}+u^{(n)}}{2}) -\frac{1}{2}D_2\theta_{\rm NN}(\frac{u^{(n+1)}-u^{(n)}}{\Delta t};\frac{u^{(n+1)}+u^{(n)}}{2}) -\frac{1}{2}DH_{\rm NN}(\frac{u^{(n+1)}+u^{(n)}}{2})\Bigg)\Delta u^{(n)}\Bigg]\Delta t$$
$$+\sum_{n=1}^{N+1}\Bigg[\Bigg(-\frac{1}{\Delta t}D_1\theta_{\rm NN}(\frac{u^{(n)}-u^{(n-1)}}{\Delta t};\frac{u^{(n)}+u^{(n-1)}}{2}) -\frac{1}{2}D_2\theta_{\rm NN}(\frac{u^{(n)}-u^{(n-1)}}{\Delta t};\frac{u^{(n)}+u^{(n-1)}}{2}) -\frac{1}{2}DH_{\rm NN}(\frac{u^{(n)}+u^{(n-1)}}{2})\Bigg)\Delta u^{(n)}\Bigg]\Delta t.$$

By using the assumption that the both ends are fixed $\Delta u^{(0)}=\Delta u^{(N)}=0$, the second term can be rewritten as:

$$\sum_{n=1}^{N+1}\Bigg[\Bigg(-\frac{1}{\Delta t}D_1\theta_{\rm NN}(\frac{u^{(n)}-u^{(n-1)}}{\Delta t};\frac{u^{(n)}+u^{(n-1)}}{2}) -\frac{1}{2}D_2\theta_{\rm NN}(\frac{u^{(n)}-u^{(n-1)}}{\Delta t};\frac{u^{(n)}+u^{(n-1)}}{2}) -\frac{1}{2}DH_{\rm NN}(\frac{u^{(n)}+u^{(n-1)}}{2})\Bigg)\Delta u^{(n)}\Bigg]\Delta t$$
$$=\sum_{n=0}^{N}\Bigg[\Bigg(-\frac{1}{\Delta t}D_1\theta_{\rm NN}(\frac{u^{(n)}-u^{(n-1)}}{\Delta t};\frac{u^{(n)}+u^{(n-1)}}{2}) -\frac{1}{2}D_2\theta_{\rm NN}(\frac{u^{(n)}-u^{(n-1)}}{\Delta t};\frac{u^{(n)}+u^{(n-1)}}{2}) -\frac{1}{2}DH_{\rm NN}(\frac{u^{(n)}+u^{(n-1)}}{2})\Bigg)\Delta u^{(n)}\Bigg]\Delta t.$$

We thus obtain

$$\sum_{n=0}^{N}\Bigg\{\Bigg[\frac{1}{\Delta t}\Bigg(D_1\theta_{\rm NN}(\frac{u^{(n+1)}-u^{(n)}}{\Delta t};\frac{u^{(n+1)}+u^{(n)}}{2}) -D_1\theta_{\rm NN}(\frac{u^{(n)}-u^{(n-1)}}{\Delta t};\frac{u^{(n)}+u^{(n-1)}}{2})\Bigg) +\frac{1}{2}\Bigg(-D_2\theta_{\rm NN}(\frac{u^{(n+1)}-u^{(n)}}{\Delta t};\frac{u^{(n+1)}+u^{(n)}}{2}) -DH_{\rm NN}(\frac{u^{(n+1)}+u^{(n)}}{2}) -D_2\theta_{\rm NN}(\frac{u^{(n)}-u^{(n-1)}}{\Delta t};\frac{u^{(n)}+u^{(n-1)}}{2}) -DH_{\rm NN}(\frac{u^{(n)}+u^{(n-1)}}{2})\Bigg)\Bigg]\Delta u^{(n)}\Bigg\}\Delta t.$$

By requiring the variation with respect to any $\Delta u^{(n)}$ to be zero, we get

$$\frac{1}{\Delta t}\Bigg(D_1\theta_{\rm NN}(\frac{u^{(n+1)}-u^{(n)}}{\Delta t};\frac{u^{(n+1)}+u^{(n)}}{2}) -D_1\theta_{\rm NN}(\frac{u^{(n)}-u^{(n-1)}}{\Delta t};\frac{u^{(n)}+u^{(n-1)}}{2})\Bigg) +\frac{1}{2}\Bigg(-D_2\theta_{\rm NN}(\frac{u^{(n+1)}-u^{(n)}}{\Delta t};\frac{u^{(n+1)}+u^{(n)}}{2}) -DH_{\rm NN}(\frac{u^{(n+1)}+u^{(n)}}{2}) -D_2\theta_{\rm NN}(\frac{u^{(n)}-u^{(n-1)}}{\Delta t};\frac{u^{(n)}+u^{(n-1)}}{2}) -DH_{\rm NN}(\frac{u^{(n)}+u^{(n-1)}}{2})\Bigg)=0,$$

which is the proposed variational integrator.

C. The equation of motion of the double pendulum and the necessity of the neural symplectic forms

In this section, we show the equation of the double pendulum used in the experiment. We also explain why it is difficult to learn this equation with a Hamiltonian neural network and why NSF is necessary.

The equation of the double pendulum used in the experiment is as follows

$$\frac{\mathrm{d}\theta_1}{\mathrm{d}t}=\phi_1,\quad \frac{\mathrm{d}\theta_2}{\mathrm{d}t}=\phi_2,$$
$$\frac{\mathrm{d}\phi_1}{\mathrm{d}t}=\Big(g(\sin\theta_2\sin(\theta_1-\theta_2)-\frac{m_1+m_2}{m_2}\sin(\theta_1)) -(l_1\theta_1^2\cos(\theta_1-\theta_2)+l_2\theta_2^2)\sin(\theta_1-\theta_2)\Big) /l_1(\frac{m_1+m_2}{m_2}-\cos^2(\theta_1-\theta_2)),$$
$$\frac{\mathrm{d}\phi_2}{\mathrm{d}t}=\Big(\frac{g(m_1+m_2)}{m_2}(\sin\theta_1\cos(\theta_1-\theta_2)-\sin(\theta_2)) -(\frac{l_1(m_1+m_2)}{m_2}\theta_1^2+l_2\theta_2^2\cos(\theta_1-\theta_2))\sin(\theta_1-\theta_2)\Big) /l_2(\frac{m_1+m_2}{m_2}-\cos^2(\theta_1-\theta_2)).$$

This equation can be rewritten into the standard Hamiltonian equation by using generalized momentum; however, the analytic expression of the generalized momentum is given by the following very complex expression

$$p_1=(m_1+m_2)l_1^2\phi_1+m_2l_1l_2\phi_2\cos(\theta_1-\theta_2),$$
$$p_2=m_2l_2^2\phi_2+m_2l_1l_2\phi_1\cos(\theta_1-\theta_2).$$

Without detailed prior knowledge, such an expression cannot be known in advance; however, the Hamiltonian neural network assumes the canonical form of the Hamilton equation and hence requires the generalized momenta as data. Hence, the Hamiltonian neural network cannot be applied to learning the dynamics of which the generalized momenta are unknown from real data.