A. Mathematical Proof

A.1. Proof of Thm. 5

Lemma 3. *Consider $\dot{x} = f(x)$ and $\dot{y} = f(y)$, where f is L-Lipschitz continuous. Then*

$$\|x(t) - y(t)\| \leq \exp(Lt)\|x(0) - y(0)\|. \tag{10}$$

Proof. Since

$$x(t) = x(0) + \int_0^t f(x(\tau))d\tau$$
$$y(t) = y(0) + \int_0^t f(y(\tau))d\tau,$$

triangular inequality and Lipschitz continuity give

$$\begin{aligned}&\|x(t) - y(t)\| \\ &\leq \|x(0) - y(0)\| + \int_0^t \|f(x(\tau)) - f(y(\tau))\|d\tau \\ &\leq \|x(0) - y(0)\| + \int_0^t L\|x(\tau) - y(\tau)\|d\tau.\end{aligned}$$

Gronwall inequality thus gives Eq. (10). □

Lemma 4. *Consider $\dot{x} = f(x)$, with $x(0) = x_0$ and L-Lipschitz continuous f. Then*

$$\|x(h) - x_0\| \leq \frac{\exp(Lh) - 1}{L}\|f(x_0)\|. \tag{11}$$

(Note $\frac{\exp(Lh)-1}{L} = \mathcal{O}(h)$.)

Proof. Note

$$\begin{aligned}x(h) &= x_0 + \int_0^h f(x(\tau))d\tau \\ &= x_0 + \int_0^h f(x(\tau)) - f(x_0) + f(x_0)d\tau.\end{aligned}$$

Let $D(t) := x(t) - x_0$. Then triangular inequality and Lipschitz continuity of f give

$$D(h) \leq \int_0^h LD(\tau) + \|f(x_0)\|d\tau$$

Gronwall lemma thus yields

$$D(h) \leq \exp(Lh)D(0) + \frac{\exp(Lh) - 1}{L}\|f(x_0)\|.$$

Since $D(0) = 0$, Eq. (11) is proved. □

Proof of Thm. 5. Let $E_n = \|x(nh) - x_n\|$ denote the prediction accuracy, and $\phi^h_{x_0}$ be the h-time flow map of the latent dynamics, i.e., $\phi^h_{x_0} := x(h)$ where $x(\cdot)$ satisfies $\dot{x} = f(x)$ subject to $x(0) = x_0$. Then

$$x((n+1)h) - x_{n+1} = x((n+1)h) - \phi^h_{x_n} + \phi^h_{x_n} - x_{n+1},$$

and therefore

$$E_{n+1} \leq \|x((n+1)h) - \phi^h_{x_n}\| + \|\phi^h_{x_n} - x_{n+1}\|.$$

The first term is exactly $\|\phi^h_{x(nh)} - \phi^h_{x_n}\|$, and by Lemma. 3, it is bounded by

$$\|\phi^h_{x(nh)} - \phi^h_{x_n}\| \leq \exp(Lh)\|x(nh) - x_n\| = \exp(Lh)E_n.$$

For the second term, Taylor expansion gives

$$\phi^h_{x_n} = x_n + hf(x_n) + h^2/2f'(\phi^\xi_{x_n})f(\phi^\xi_{x_n})$$

for some $\xi \in [0, h]$, and therefore

$$\begin{aligned}\|\phi^h_{x_n} - x_{n+1}\| &= \|h(f(x_n) - \tilde{f}(x_n)) + h^2/2f'(\phi^\xi_{x_n})f(\phi^\xi_{x_n})\| \\ &\leq h\delta + h^2/2\|f'(\phi^\xi_{x_n})\|\|f(\phi^\xi_{x_n})\|.\end{aligned}$$

Note $\|f'\| \leq L$ as f is $\mathcal{C}^1$ and L-Lipschitz. For the $f(\phi^\xi_{x_n})$ factor, note Lemma. 4 gives

$$\|\phi^\xi_{x_n} - x_n\| \leq \frac{\exp(L\xi) - 1}{L}\|f(x_n)\|,$$

and therefore

$$\begin{aligned}\|f(\phi^\xi_{x_n})\| &= \|f(x_n) + f(\phi^\xi_{x_n}) - f(x_n)\| \\ &\leq \|f(x_n)\| + \|f(\phi^\xi_{x_n}) - f(x_n)\| \\ &\leq \|f(x_n)\| + L\|\phi^\xi_{x_n} - x_n\| \\ &\leq \exp(L\xi)\|f(x_n)\|\end{aligned}$$

Since $0 \leq \xi \leq h$, $\exp(L\xi)$ is bounded. Moreover, $f(x_n)$ is bounded because f is Lipschitz and therefore continuous and x_n is assumed to be bounded. Therefore, there exists constant C such that

$$\|f'(\phi^\xi_{x_n})\|\|f(\phi^\xi_{x_n})\| \leq C$$

Summarizing both terms, we have

$$E_{n+1} \leq E_n \exp(Lh) + h\delta + Ch^2/2.$$

Mathematical induction thus gives

$$\begin{aligned}E_N &\leq E_0\exp(Lh)^N \\ &\quad + \left(\exp(Lh)^{N-1} + \exp(Lh)^{N-2} + \cdots + 1\right)(h\delta + Ch^2/2) \\ &= E_0\exp(LT) + \frac{\exp(LT) - 1}{\exp(Lh) - 1}(h\delta + Ch^2/2) \\ &\leq E_0\exp(LT) + \frac{\exp(LT) - 1}{Lh}(h\delta + Ch^2/2) \\ &= \frac{\exp(LT) - 1}{L}(\delta + Ch/2).\end{aligned}$$

□

A.2. Proof of Thm. 3

Definition 5 (Diophantine condition)**.** *A frequency vector $\boldsymbol{\omega} = \{\omega_1, \omega_2, \ldots, \omega_d\}$ satisfies Diophantine condition if and only there exists positive constants γ, ν such that $\boldsymbol{\omega}$ satisfies (γ, ν)-Diophantine condition.*

Definition 6 ((γ, ν)-Diophantine set)**.** *For a set $\Omega \subseteq \mathbb{R}^d$, the corresponding (γ, ν)-Diophantine set is defined as*

$$\Omega^*(\gamma, \nu) \overset{def}{=} \{\boldsymbol{\omega} \in \Omega : \boldsymbol{\omega} \text{ satisfy } (\gamma, \nu)\text{-Diophantine condition}\}.$$

Definition 7 (Diophantine set)**.** *For a set $\Omega \subseteq \mathbb{R}^d$, the corresponding Diophantine set is defined as*

$$\Omega^* \overset{def}{=} \bigcup_{\gamma>0, \nu>0} \Omega^*(\gamma, \nu).$$

Theorem 6. *For any bounded domain $\Omega \subseteq \mathbb{R}^d$, there exists $C > 0$, such that the Lebesgue measure of the complementary of (γ, ν)-Diophantine set with $\nu \geq d$ is bounded from above,*

$$\lambda(\Omega \backslash \Omega^*(\gamma, \nu)) \leq C \cdot \gamma.$$

Proof. See for instance (Hairer et al., 2006). □

Theorem 7. *For any bounded domain $\Omega \subseteq \mathbb{R}^d$, Diophantine frequencies exist almost everywhere.*

Proof. Since $\lambda(\Omega \backslash \Omega^*) \leq \lambda(\Omega \backslash \Omega^*(\gamma, \nu)), \forall \gamma > 0, \nu > 0$, Thm. 6 gives, $\forall \gamma > 0$,

$$\lambda(\Omega \backslash \Omega^*) \leq C \cdot \gamma,$$

meaning that Diophantine frequencies exist *almost everywhere* in Ω. □

Remark 7. *Even Diophantine frequencies exist almost everywhere in bounded domain $\Omega \subseteq \mathbb{R}^d$, $\Omega \backslash \Omega^*$ is still an open and dense set in $\mathbb{R}^d$ (see for instance Hairer et al., 2006).*

Lemma 5 (Cauchy's inequality)**.** *Suppose that f is a holomorphic function on a closed ball $\overline{\mathcal{B}_r(\theta^*)} \subset \mathbb{C}$ with $r > 0$. If $|f(\theta)| \leq M$ for all θ on the boundary of $\mathcal{B}_r(\theta^*)$, then for all $n \geq 0$,*

$$\left| f^{(n)}(\theta^*) \right| \leq \frac{n! M}{r^n}.$$

Proof. See for instance (Stein & Shakarchi, 2010). □

Definition 8 (average over angles)**.** *Assume $F(\boldsymbol{\theta})$ is periodic in each argument, i.e., $F : \mathbb{T}^d \to \mathbb{R}$, then the (angle) average of F is defined as*

$$\overline{F} = \frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} F(\boldsymbol{\theta}) \, d\boldsymbol{\theta}. \tag{12}$$

Definition 9 (complex extension of $\mathbb{T}^d$)**.** *The complex extension of $\mathbb{T}^d$ of width ρ is defined as*

$$\mathcal{B}_\rho(\mathbb{T}^d) = \left\{ \boldsymbol{\theta} \in \mathbb{T}^d + i\mathbb{R}^d ; \|Im\theta\| < \rho \right\}. \tag{13}$$

Definition 10. *For an analytic function $\boldsymbol{f}(\cdot) = [f_1(\cdot), \ldots, f_d(\cdot)] \in \mathbb{C}^d$, we define the following norm*

$$\|\boldsymbol{f}\|_{\infty, S} := \sum_{i=1}^{d} \sup_{x \in S} |f_i(x)|. \tag{14}$$

Lemma 6. *Suppose $\boldsymbol{\omega} \in \mathbb{R}^d$ satisfies the (γ, ν)-Diophantine condition and $G(\boldsymbol{\theta}) \in \mathbb{R}$ is a bounded and analytic function on $\mathcal{B}_\rho(\mathbb{T}^d)$. Then, with $\overline{G}$ being the average of $G(\boldsymbol{\theta})$, the PDE*

$$DF(\boldsymbol{\theta}) \cdot \boldsymbol{\omega} + G(\boldsymbol{\theta}) = \overline{G} \tag{15}$$

has a unique real analytic solution $F(\cdot)$ with $\overline{F} = 0$. Moreover, for every positive $\delta < \min(\rho, 1)$, F is bounded on $\mathcal{B}_{\rho-\delta}(\mathbb{T}^d)$ by

$$\begin{cases} \|F\|_{\infty, \mathcal{B}_{\rho-\delta}(\mathbb{T}^d)} \leq \kappa_0 \delta^{-\alpha+1} \|G\|_{\infty, \mathcal{B}_\rho(\mathbb{T}^d)}, \\ \|\partial_{\boldsymbol{\theta}} F\|_{\infty, \mathcal{B}_{\rho-\delta}(\mathbb{T}^d)} \leq \kappa_1 \delta^{-\alpha} \|G\|_{\infty, \mathcal{B}_\rho(\mathbb{T}^d)}, \end{cases}$$

with $\alpha = \nu + d + 1$ and $\kappa_0 = \nu^{-1} 8^d 2^\nu \nu!$, $\kappa_1 = \nu^{-1} 8^d 2^{\nu+1} (\nu+1)!$.

Proof. See for instance (Hairer et al., 2006). □

Lemma 7. *Consider a nearly integrable system with generating function $S(\boldsymbol{I}_0, \boldsymbol{\varphi}_1) = S_0(\boldsymbol{I}_0) + \varepsilon S_1(\boldsymbol{I}_0, \boldsymbol{\varphi}_1)$. Suppose S_0 and S_1 are analytic and bounded in a complex neighborhood of $\mathcal{D}_1 \subseteq \mathbb{R}^d$ and $\mathcal{D} = \mathcal{D}_1 \times \mathbb{T}^d$ respectively. Then, there exists a real analytic canonical transformation $(\boldsymbol{J}, \boldsymbol{\theta}) \leftrightarrow (\boldsymbol{I}, \boldsymbol{\varphi})$ generated by $\mathcal{T}(\boldsymbol{J}, \boldsymbol{\varphi}) = \boldsymbol{J} \cdot \boldsymbol{\varphi} + \varepsilon \mathcal{T}_1(\boldsymbol{J}, \boldsymbol{\varphi})$, such that the generating function in $\boldsymbol{J}, \boldsymbol{\theta}$ variables takes the form of*

$$\tilde{S}(\boldsymbol{J}_0, \boldsymbol{\theta}_1) = \tilde{S}_0(\boldsymbol{J}_0) + \varepsilon^2 \tilde{R}_2(\boldsymbol{J}_0, \boldsymbol{\theta}_1, \varepsilon). \tag{16}$$

with $\varepsilon^2 \tilde{R}_2$ being a higher-order perturbation to a new integrable system $\tilde{S}_0$. Moreover, this result is constructive: the transformation $\boldsymbol{J}, \boldsymbol{\theta} \leftrightarrow \boldsymbol{I}, \boldsymbol{\varphi}$ is given by $\mathcal{T}$ through

$$\begin{cases} \boldsymbol{I}_i = \partial_2 \mathcal{T}(\boldsymbol{J}_i, \boldsymbol{\varphi}_i), \\ \boldsymbol{\theta}_i = \partial_1 \mathcal{T}(\boldsymbol{J}_i, \boldsymbol{\varphi}_i). \end{cases} \forall i = 0, 1 \tag{17}$$

and $\mathcal{T}_1$ is the solution to the PDE

$$\partial_2 \mathcal{T}_1(\boldsymbol{J}, \boldsymbol{\varphi}) \cdot \boldsymbol{\omega}(\boldsymbol{J}) + G_1(\boldsymbol{J}, \boldsymbol{\varphi}) = \overline{G_1}(\boldsymbol{J}), \tag{18}$$

where $\boldsymbol{\omega}(\cdot) = \nabla S_0(\cdot)$, $G_1(\boldsymbol{J}, \boldsymbol{\varphi}) = S_1(\boldsymbol{J}, \boldsymbol{\varphi} + h \nabla S_0(\boldsymbol{J}))$, and $\overline{G_1}(\boldsymbol{J})$ is its angle average.

Proof. The generating function S in $\boldsymbol{J},\boldsymbol{\theta}$ variables ($\tilde{S}$) can be converted in the following using Eq. (17),

$$\begin{aligned}
&\tilde{S}\left(\boldsymbol{J}_0,\boldsymbol{\theta}_1\right)=S\left(\boldsymbol{I}_0,\boldsymbol{\varphi}_1\right)\\
=&S\left(\boldsymbol{J}_0+\varepsilon\partial_2\mathcal{T}_1\left(\boldsymbol{J}_0,\boldsymbol{\varphi}_0\right),\boldsymbol{\varphi}_1\right)\\
=&S_0\left(\boldsymbol{J}_0+\varepsilon\partial_2\mathcal{T}_1\left(\boldsymbol{J}_0,\boldsymbol{\varphi}_0\right)\right)\\
&+\varepsilon S_1\left(\boldsymbol{J}_0+\varepsilon\partial_2\mathcal{T}_1\left(\boldsymbol{J}_0,\boldsymbol{\varphi}_0\right),\boldsymbol{\varphi}_0+h\partial_1 S(\boldsymbol{I}_0,\boldsymbol{\varphi}_1)\right)\\
&+\mathcal{O}\left(\varepsilon^2\right)\\
=&S_0\left(\boldsymbol{J}_0\right)+\varepsilon\partial_2\mathcal{T}_1\left(\boldsymbol{J}_0,\boldsymbol{\varphi}_0\right)\cdot\nabla S_0\left(\boldsymbol{J}_0\right)\\
&+\varepsilon S_1\left(\boldsymbol{J}_0,\boldsymbol{\varphi}_0+h\nabla S(\boldsymbol{J}_0)\right)+\mathcal{O}\left(\varepsilon^2\right)\\
=&S_0\left(\boldsymbol{J}_0\right)+\varepsilon\big\{\underline{\partial_2\mathcal{T}_1\left(\boldsymbol{J}_0,\boldsymbol{\varphi}_0\right)\cdot\nabla S_0\left(\boldsymbol{J}_0\right)}\\
&\underline{+S_1\left(\boldsymbol{J}_0,\boldsymbol{\varphi}_0+h\nabla S(\boldsymbol{J}_0)\right)}\big\}+\mathcal{O}\left(\varepsilon^2\right)
\end{aligned}\tag{19}$$

with h the constant in Eq. (4). Collecting all $\mathcal{O}(\varepsilon^2)$ terms, denoting them by a remainder term $\tilde{R}_2$, and converting all angles in $\tilde{R}_2$ to $\boldsymbol{\theta}_1$, we have

$$\begin{aligned}
&\tilde{S}\left(\boldsymbol{J}_0,\boldsymbol{\theta}_1\right)\\
=&S_0\left(\boldsymbol{J}_0\right)+\varepsilon\big\{\underline{\partial_2\mathcal{T}_1\left(\boldsymbol{J}_0,\boldsymbol{\varphi}_0\right)\cdot\nabla S_0\left(\boldsymbol{J}_0\right)}\\
&\underline{+S_1\left(\boldsymbol{J}_0,\boldsymbol{\varphi}_0+h\nabla S_0(\boldsymbol{J}_0)\right)}\big\}+\varepsilon^2\tilde{R}\left(\boldsymbol{J}_0,\boldsymbol{\theta}_1,\varepsilon\right).
\end{aligned}\tag{20}$$

As long as the terms underlined in Eq. (20) add up to a function of $\boldsymbol{J}_0$ only, $\tilde{S}\left(\boldsymbol{J}_0,\boldsymbol{\theta}_1\right)$ won't have angle dependence till the $\mathcal{O}\left(\varepsilon^2\right)$ term. This leads to a solvability requirement. More precisely, let $G_1(\boldsymbol{J}_0,\boldsymbol{\varphi}_0)=S_1(\boldsymbol{J}_0,\boldsymbol{\varphi}_0+h\nabla S_0(\boldsymbol{J}_0))$ and $\overline{G_1}(\boldsymbol{J}_0)$ be its angle average. Then the PDE

$$\partial_2\mathcal{T}_1\left(\boldsymbol{J}_0,\boldsymbol{\varphi}_0\right)\cdot\nabla S_0\left(\boldsymbol{J}_0\right)+G_1\left(\boldsymbol{J}_0,\boldsymbol{\varphi}_0\right)=\overline{G_1}(\boldsymbol{J}_0)\tag{21}$$

has a solution $\mathcal{T}_1$, and it makes the underlined terms $\overline{G_1}(\boldsymbol{J}_0)$. Therefore, $\mathcal{T}_1$ and hence the generating function $\mathcal{T}$ can be solved for from Eq. (21). The generating function $\tilde{S}$ in $\boldsymbol{J},\boldsymbol{\theta}$ variables takes the form

$$\tilde{S}\left(\boldsymbol{J}_0,\boldsymbol{\theta}_1\right)=\tilde{S}_0\left(\boldsymbol{J}_0\right)+\varepsilon^2\tilde{R}_2\left(\boldsymbol{J}_0,\boldsymbol{\theta}_1,\varepsilon\right),$$

with $\tilde{S}_0(\boldsymbol{J}_0)=S_0(\boldsymbol{J}_0)+\varepsilon\overline{G}(\boldsymbol{J}_0)$. □

Remark 8. *Note that boundedness of $\tilde{R}_2$ requires some extra condition on ∇S_0 (being Diophantine at some point; see Lemma. 9 for details). With bounded $\tilde{R}_2$, the generating function $S\left(\cdot,\cdot\right)$ in $\boldsymbol{I},\boldsymbol{\varphi}$ variables is near integrable of order $\mathcal{O}\left(\varepsilon\right)$ while $\tilde{S}\left(\cdot,\cdot\right)$ in $\boldsymbol{J},\boldsymbol{\theta}$ variables is near integrable of order $\mathcal{O}\left(\varepsilon^2\right)$. Therefore, under the transformation $\mathcal{T}$, we get a 'better' set of variables $\boldsymbol{J},\boldsymbol{\theta}$ instead of $\boldsymbol{I},\boldsymbol{\varphi}$, as the $\boldsymbol{J},\boldsymbol{\theta}$ dynamics is closer to being integrable, hence the dynamics of $\boldsymbol{J},\boldsymbol{\theta}$ can be estimated for longer time.*

Remark 9. *As angles satisfy periodic boundary conditions, $\mathcal{T}_1$ and S_1 can be expanded in Fourier series*

$$\begin{cases}\mathcal{T}_1\left(\boldsymbol{J},\boldsymbol{\theta}\right)=\sum\limits_{\boldsymbol{k}\in\mathbb{Z}^d}t_{\boldsymbol{k}}\left(\boldsymbol{J}\right)\cdot e^{i(\boldsymbol{k}\cdot\boldsymbol{\theta})},\\ S_1\left(\boldsymbol{J},\boldsymbol{\theta}\right)=\sum\limits_{\boldsymbol{k}\in\mathbb{Z}^d}s_{\boldsymbol{k}}\left(\boldsymbol{J}\right)\cdot e^{i(\boldsymbol{k}\cdot\boldsymbol{\theta})}.\end{cases}\tag{22}$$

Plugging Eq. (22) into Eq. (18), we have

$$t_{\boldsymbol{k}}\cdot\left(\boldsymbol{k}\cdot\boldsymbol{\omega}\left(\boldsymbol{J}\right)\right)+s_{\boldsymbol{k}}=0.$$

Noting that if $\boldsymbol{\omega}$ doesn't satisfy Diophantine condition, $\boldsymbol{k}\cdot\boldsymbol{\omega}$ can be small and may even vanish for some $\boldsymbol{k}\in\mathbb{Z}^d$, meaning that under some circumstances, the transformation constructed by $\mathcal{T}\left(\boldsymbol{J},\boldsymbol{\varphi}\right)=\boldsymbol{J}\cdot\boldsymbol{\varphi}+\varepsilon\mathcal{T}_1\left(\boldsymbol{J},\boldsymbol{\varphi}\right)$ is no longer of near identity ($Id+\mathcal{O}(\varepsilon)$) as $\mathcal{T}_1$ is not of order $\mathcal{O}\left(1\right)$ any more.

Lemma 8. *Consider a nearly integrable system with generating function $S\left(\boldsymbol{I}_0,\boldsymbol{\varphi}_1\right)=S_0\left(\boldsymbol{I}_0\right)+\varepsilon S_1\left(\boldsymbol{I}_0,\boldsymbol{\varphi}_1\right)$. Suppose S_0 and S_1 are analytic and bounded in a complex neighborhood of $\mathcal{D}_1\subseteq\mathbb{R}^d$ and $\mathcal{D}=\mathcal{D}_1\times\mathbb{T}^d$ respectively. Then, there exists a real analytic canonical transformation $(\boldsymbol{J},\boldsymbol{\theta})\leftrightarrow(\boldsymbol{I},\boldsymbol{\varphi})$ generated by $\mathcal{T}\left(\boldsymbol{J},\boldsymbol{\varphi}\right)=\boldsymbol{J}\cdot\boldsymbol{\varphi}+\sum_{k=1}^{N-1}\varepsilon^k\cdot\mathcal{T}_k\left(\boldsymbol{J},\boldsymbol{\varphi}\right)$, such that the dynamics produced by the original generating function S rewritten in $\boldsymbol{J},\boldsymbol{\theta}$ variables corresponds to a transformed generating function $\tilde{S}$ given by*

$$\tilde{S}\left(\boldsymbol{J}_0,\boldsymbol{\theta}_1\right)=\tilde{S}_0\left(\boldsymbol{J}_0\right)+\varepsilon^N\tilde{R}_N\left(\boldsymbol{J}_0,\boldsymbol{\theta}_1,\varepsilon\right),\tag{23}$$

where $\varepsilon^N\tilde{R}_N$ is a high-order perturbation to a new integrable system $\tilde{S}_0(\boldsymbol{J}_0)$. Here, $\boldsymbol{J},\boldsymbol{\theta}\leftrightarrow\boldsymbol{I},\boldsymbol{\varphi}$ is defined by $\mathcal{T}$ through

$$\begin{cases}\boldsymbol{I}_i=\partial_2\mathcal{T}\left(\boldsymbol{J}_i,\boldsymbol{\varphi}_i\right),\\ \boldsymbol{\theta}_i=\partial_1\mathcal{T}\left(\boldsymbol{J}_i,\boldsymbol{\varphi}_i\right),\end{cases}\quad\forall i=0,1.\tag{24}$$

Proof. Apply $\mathcal{T}(\boldsymbol{J},\boldsymbol{\varphi})=\boldsymbol{J}\cdot\boldsymbol{\varphi}+\varepsilon\mathcal{T}_1+\varepsilon\mathcal{T}_2+\ldots+\varepsilon^{N-1}\mathcal{T}_{N-1}$ to $\tilde{S}(\boldsymbol{J}_0,\boldsymbol{\theta}_1)=S(\boldsymbol{I}_0,\boldsymbol{\varphi}_1)$ like in the proof of Lemma. 7, Taylor expand, and put all $\mathcal{O}(\varepsilon^N)$ terms into $\tilde{R}_N$. Then we have

$$\begin{aligned}
&\tilde{S}(\boldsymbol{J}_0,\boldsymbol{\theta}_1)=\ S_0(\boldsymbol{J}_0)\\
&+\varepsilon\left(\partial_2\mathcal{T}_1(\boldsymbol{J}_0,\boldsymbol{\varphi}_0)\cdot\boldsymbol{\omega}(\boldsymbol{J}_0)+G_1(\boldsymbol{J}_0,\boldsymbol{\varphi}_0)\right)\\
&+\varepsilon^2\left(\partial_2\mathcal{T}_2(\boldsymbol{J}_0,\boldsymbol{\varphi}_0)\cdot\boldsymbol{\omega}(\boldsymbol{J}_0)+G_2(\boldsymbol{J}_0,\boldsymbol{\varphi}_0)\right)\\
&+\ldots\\
&+\varepsilon^N\tilde{R}_N\left(\boldsymbol{J}_0,\boldsymbol{\theta}_1,\varepsilon\right),
\end{aligned}\tag{25}$$

for some functions $G_1,\cdots,G_{N-1}$ periodic in the angles, and $\boldsymbol{\omega}(\cdot):=\nabla S(\cdot)$. Similar to G_1 in the proof of Lemma. 7, $G_2,G_3,\cdots$ can be explicitly computed from Taylor expansions, but our proof does not require their specific expressions. Lemma. 7 solved for $\mathcal{T}_1$ (periodic in angles) by making the underlined expression independent of the angles.

Repeating a similar procedure at different orders of ε, $\mathcal{T}_i$ (periodic) can be obtained for $i = 1, 2, \cdots$. Specifically, $\mathcal{T}_i$ satisfies the PDE

$$\partial_2 \mathcal{T}_i(\boldsymbol{J}_0, \boldsymbol{\varphi}_0) \cdot \boldsymbol{\omega}(\boldsymbol{J}_0) + G_i(\boldsymbol{J}_0, \boldsymbol{\varphi}_0) = \overline{G_i}(\boldsymbol{J}_0) \tag{26}$$

(The existence of the solution to this is proved in Lemma. 6). In general,

$$\begin{aligned}\tilde{S}\left(\boldsymbol{J}_0, \boldsymbol{\theta}_1\right) &= S_0\left(\boldsymbol{J}_0\right) + \sum_{k=1}^{N-1} \varepsilon^k \overline{G_k}\left(\boldsymbol{J}_0\right) + \varepsilon^N \tilde{R}_N\left(\boldsymbol{J}_0, \boldsymbol{\theta}_1\right) \\ &= \tilde{S}_0\left(\boldsymbol{J}_0\right) + \varepsilon^N \tilde{R}_N\left(\boldsymbol{J}_0, \boldsymbol{\theta}_1\right),\end{aligned}$$

with $\tilde{S}_0\left(\boldsymbol{J}_0\right) = S_0\left(\boldsymbol{J}_0\right) + \sum_{k=1}^{N-1} \varepsilon^k \overline{G_k}\left(\boldsymbol{J}_0\right)$. □

Note that $\tilde{R}_N$ is not necessarily uniformly bounded in the whole data domain in Lemmas. 7 and 8, but in most cases, the uniform boundedness can be established (Lemma. 9) and under that circumstance, we will be able to quantitatively estimate the $\boldsymbol{J}, \boldsymbol{\theta}$ dynamics (Lemma. 10).

Lemma 9. *Consider a nearly integrable system with generating function $S\left(\boldsymbol{I}_0, \boldsymbol{\varphi}_1\right) = S_0\left(\boldsymbol{I}_0\right) + \varepsilon S_1\left(\boldsymbol{I}_0, \boldsymbol{\varphi}_1\right)$. Suppose S_0 and S_1 are analytic and bounded in a complex neighborhood of $\mathcal{D}_1$ and $\mathcal{D} = \mathcal{D}_1 \times \mathbb{T}^d \subseteq \mathbb{R}^d \times \mathbb{T}^d$ respectively. There exists a real analytic symplectic change of coordinates of order $\mathcal{O}(\varepsilon)$: $(\boldsymbol{I}, \boldsymbol{\varphi}) \leftrightarrow (\boldsymbol{J}, \boldsymbol{\theta})$ and under this transformation, the generating function in $\boldsymbol{J}, \boldsymbol{\theta}$ takes the form*

$$\tilde{S}\left(\boldsymbol{J}_0, \boldsymbol{\theta}_1\right) = \tilde{S}_0\left(\boldsymbol{J}_0\right) + \varepsilon^N \tilde{R}_N\left(\boldsymbol{J}_0, \boldsymbol{\theta}_1, \varepsilon\right).$$

Suppose that $\omega(\boldsymbol{J}^*)$ satisfies the (γ, ν)-Diophantine condition for some $\boldsymbol{J}^* \in \mathcal{D}_1$. Then, for any fixed $N \geq 2$, there exist positive constants $\varepsilon_0, c, C, \rho$ such that if $\varepsilon \leq \varepsilon_0$, then

$$\left\|\tilde{R}_N(\cdot, \cdot)\right\|_{\infty, \overline{\mathcal{B}_{2\delta}(\boldsymbol{J}^*)} \times \overline{\mathcal{B}_\rho(\mathbb{T}^d)}} \leq C$$

with $\delta = c\left(N^2 \left|\log \varepsilon\right|\right)^{-\nu-1}$.

Proof. Applying the canonical transformation $\mathcal{T}$ constructed in Lemma. 8, $\exists \rho', C' > 0$ such that

$$\left\|\tilde{R}_N\left(\boldsymbol{J}^*, \cdot\right)\right\|_{\infty, \mathcal{B}_{\rho'}(\mathbb{T}^d)} \leq C'(N, d, \gamma, \nu) \tag{27}$$

Approximate S with respect to angle variables using Fourier series $\hat{S}_m$ till term $m \propto |\log \varepsilon|$ such that the error is of order $\mathcal{O}(\varepsilon^N)$ in a complex neighborhood of the torus $\left\{\boldsymbol{J} = \boldsymbol{J}^*, \boldsymbol{\varphi} \in \mathbb{T}^d\right\}$. Since $|\boldsymbol{k} \cdot \boldsymbol{\omega}\left(\boldsymbol{J}^*\right)| \geq \gamma \|\boldsymbol{k}\|_1^{-\nu}$, $\forall \boldsymbol{k} \in \mathbb{Z}^d$, then $\exists$ sufficiently small $c > 0$ such that

$$\left|\boldsymbol{k} \cdot \boldsymbol{\omega}\left(\boldsymbol{J}\right)\right| \geq \frac{1}{2}\gamma \|\boldsymbol{k}\|_1^{-\nu}, \ \|\boldsymbol{k}\|_1 \leq Nm \tag{28}$$

for all $\boldsymbol{J} \in \mathcal{B}_{2\delta}(\boldsymbol{J}^*)$ with $\delta = c(N^2 |\log \varepsilon|)^{-\nu-1}$. As the Fourier coefficients of $\hat{S}_m$ vanishes for $\|\boldsymbol{k}\|_1 > Nm$, thus according to condition Eq. (28) and combining $S = \hat{S}_m + \mathcal{O}(\varepsilon^N)$, $\exists\, \rho'' > 0$ and $C'' > 0$, such that

$$\left\|\tilde{R}_N\left(\boldsymbol{J}, \cdot\right)\right\|_{\infty, \mathcal{B}_{\rho''}(\mathbb{T}^d)} \leq C''(N, d, \gamma, \nu) \tag{29}$$

for all $\|\boldsymbol{J} - \boldsymbol{J}^*\| \leq 2\delta$.

In general, $\exists C$ and ε independent ρ such that

$$\left\|\tilde{R}_N(\cdot, \cdot)\right\|_{\infty, \overline{\mathcal{B}_{2\delta}(\boldsymbol{J}^*)} \times \overline{\mathcal{B}_\rho(\mathbb{T}^d)}} \leq C(N, d, \gamma, \nu).$$

(for the specific forms of C', C'', which are lengthy but obtainable using tools of Fourier series and Cauchy's inequality, see for instance (Hairer et al., 2006)). □

Lemma 10. *Consider a nearly integrable system with generating function $S\left(\boldsymbol{I}_0, \boldsymbol{\varphi}_1\right) = S_0\left(\boldsymbol{I}_0\right) + \varepsilon S_1\left(\boldsymbol{I}_0, \boldsymbol{\varphi}_1\right)$. Suppose S_0 and S_1 are analytic and bounded in a complex neighborhood of $\mathcal{D}_1$ and $\mathcal{D} = \mathcal{D}_1 \times \mathbb{T}^d \subseteq \mathbb{R}^d \times \mathbb{T}^d$ respectively. Then there exists a real analytic near identity symplectic change of coordinates $(\boldsymbol{I}, \boldsymbol{\varphi}) \mapsto (\boldsymbol{J}, \boldsymbol{\theta})$ of order $\mathcal{O}(\varepsilon)$ and under this transformation, the generating function $\tilde{S}$ in $\boldsymbol{J}, \boldsymbol{\theta}$ variables takes the form*

$$\tilde{S}\left(\boldsymbol{J}, \boldsymbol{\theta}\right) = \tilde{S}_0\left(\boldsymbol{J}\right) + \varepsilon^N \tilde{R}_N\left(\boldsymbol{J}, \boldsymbol{\theta}, \varepsilon\right).$$

where $\tilde{S}_0$ only depends on actions. Suppose that $\boldsymbol{\omega}\left(\boldsymbol{J}^*\right)$ satisfies the (γ, ν)-Diophantine condition for some $\boldsymbol{J}^* \in \mathcal{D}_1$. Then, for any fixed $N \geq 2$, $\exists$ positive constants $\varepsilon_0, c, C, \rho$ such that if $\varepsilon \leq \varepsilon_0$, the dynamics of $\boldsymbol{J}, \boldsymbol{\theta}$ (generated by $\tilde{S}$) with $\|\boldsymbol{J}_0 - \boldsymbol{J}^*\|_2 \leq c\left|\log \varepsilon\right|^{-\nu-1}$ satisfies

$$\begin{cases} \|\boldsymbol{J}_n - \boldsymbol{J}_0\|_2 \leq Cnh\varepsilon^N, \\ \|\boldsymbol{\theta}_n - \tilde{\boldsymbol{\omega}}\left(\boldsymbol{J}_0\right) nh - \boldsymbol{\theta}_0\|_2 \\ \qquad \leq C\left(n^2h^2 + nh\left|\log \varepsilon\right|^{\nu+1}\right)\varepsilon^N. \end{cases} \tag{30}$$

Here $\boldsymbol{\omega}(\cdot) = \nabla S_0(\cdot)$ and $\tilde{\boldsymbol{\omega}}(\cdot) = \nabla \tilde{S}_0(\cdot)$.

Proof. According to Lemma. 9, $\exists c > 0, \rho > 0, C' > 0$ such that for $\delta = c\left(N^2 \left|\log \varepsilon\right|\right)^{-\nu-1}$, $\boldsymbol{J} \in \overline{\mathcal{B}_\delta\left(\boldsymbol{J}^*\right)}$ and $\boldsymbol{\theta} \in \mathcal{B}_\rho\left(\mathbb{T}^d\right)$, $\left|\tilde{R}_N\left(\boldsymbol{J}, \boldsymbol{\theta}\right)\right| \leq C'$. As $\forall \boldsymbol{J} \in \overline{\mathcal{B}_\delta\left(\boldsymbol{J}^*\right)}$, $\overline{\mathcal{B}_\delta\left(\boldsymbol{J}\right)} \subset \overline{\mathcal{B}_{2\delta}\left(\boldsymbol{J}^*\right)}$, $\left|\tilde{R}_N\left(\tilde{\boldsymbol{J}}, \boldsymbol{\theta}\right)\right| \leq C'$ for all $\tilde{\boldsymbol{J}} \in \overline{\mathcal{B}_\delta\left(\boldsymbol{J}\right)}$ and $\boldsymbol{\theta} \in \mathcal{B}_\rho\left(\mathbb{T}^d\right)$. Using Cauchy's inequality (Lemma. 5), we have

$$\left\|\partial_2 \tilde{R}_N\right\|_{\infty, \overline{\mathcal{B}_\delta(\boldsymbol{J}^*)} \times \overline{\mathcal{B}_\rho(\mathbb{T}^d)}} \leq C' \tag{31}$$

and

$$\left\|\partial_1 \tilde{R}_N\right\|_{\infty, \overline{\mathcal{B}_\delta(\boldsymbol{J}^*)} \times \overline{\mathcal{B}_\rho(\mathbb{T}^d)}} \leq \frac{C'}{\delta}. \tag{32}$$

Plug Eq. (31) in the dynamics of $\boldsymbol{J}, \boldsymbol{\theta}$

$$\begin{cases} \boldsymbol{J}_i = \boldsymbol{J}_{i+1} + h\partial_2 \tilde{S}\left(\boldsymbol{J}_i, \boldsymbol{\theta}_{i+1}\right), \\ \boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + h\partial_1 \tilde{S}\left(\boldsymbol{J}_i, \boldsymbol{\theta}_{i+1}\right), \end{cases} \tag{33}$$

we have

$$\begin{aligned} &\|\boldsymbol{J}_{i+1} - \boldsymbol{J}_i\|_2 \leq C' h\varepsilon^N, \, \forall i \in \mathbb{N} \\ \Longrightarrow \; &\|\boldsymbol{J}_n - \boldsymbol{J}_0\|_2 \leq C' nh\varepsilon^N. \end{aligned}$$

for the $\boldsymbol{J}$ sequence. On the other hand, for $\boldsymbol{\theta}$ sequence, plug Eq. (32) in Eq. (33), we have

$$\|\boldsymbol{\theta}_{i+1} - (\boldsymbol{\theta}_i + h\tilde{\omega}\left(\boldsymbol{J}_i\right))\|_2 \leq \frac{C'}{\delta} h\varepsilon^N. \tag{34}$$

Since $\tilde{\omega}$ is analytic on a bounded domain, $\tilde{\omega}$ is Lipschitz. Thus, changing $\boldsymbol{J}_i$ in Eq. (34) to $\boldsymbol{J}_0$, $\exists C''$ such that

$$\|\boldsymbol{\theta}_{i+1} - (\boldsymbol{\theta}_i + h\tilde{\omega}\left(\boldsymbol{J}_0\right))\|_2 \leq C'' nh^2\varepsilon^N + \frac{C'}{\delta} h\varepsilon^N.$$

Therefore, letting $C = \max(C', C'')$, we have

$$\begin{aligned} &\|\boldsymbol{\theta}_n - (\boldsymbol{\theta}_0 + nh\tilde{\omega}\left(\boldsymbol{J}_0\right))\|_2 \\ \leq &\sum_{i=0}^{n-1} \|\boldsymbol{\theta}_{i+1} - (\boldsymbol{\theta}_i + h\tilde{\omega}\left(\boldsymbol{J}_0\right))\|_2 \\ \leq &Cnh\left(nh + \frac{1}{\delta}\right)\varepsilon^N \\ \leq &Cnh\left(nh + |\log\varepsilon|^{\nu+1}\right)\varepsilon^N, \end{aligned}$$

and Eq. (30) is proved. □

Proof of Thm. 3. Since it is assumed that analytic S_h and S_h^θ satisfy

$$\sum_{i=1,2} \left\|\partial_i S_h^\theta\left(\cdot,\cdot\right) - \partial_i S_h\left(\cdot,\cdot\right)\right\|_\infty \leq C_1\varepsilon$$

on a bounded domain $\mathcal{D}$, S_h^θ is an $\mathcal{O}(\varepsilon)$ perturbation of S_h (note they can also be different by an $\mathcal{O}(1)$ constant, but adding a constant to a generating function does not change its induced dynamics, and we thus assume without loss of generality that there is no such constant difference). Therefore, as S_h is integrable, S_h^θ can be written as $S_h^\theta(\boldsymbol{I}_0, \boldsymbol{\varphi}_1) = S_h(\boldsymbol{I}_0) + \varepsilon S_1(\boldsymbol{I}_0, \boldsymbol{\varphi}_1)$ for some function S_1 modeling the (normalized) perturbation, and is thus nearly integrable. The latent dynamics, i.e., the exact solution of the integrable $S_h(\boldsymbol{I}_0)$ with initial condition $\boldsymbol{I}_0, \boldsymbol{\varphi}_0$ is

$$\begin{cases} \boldsymbol{I}\left(t\right) = \boldsymbol{I}_0, \\ \boldsymbol{\varphi}\left(t\right) = \left(\boldsymbol{\varphi}_0 + \boldsymbol{\omega}\left(\boldsymbol{I}_0\right)t\right) \mod 2\pi. \end{cases}$$

Applying Lemma. 10 with $N \geq 3$ (so that the $(nh)^2\varepsilon^N$ term is of order $\mathcal{O}(\varepsilon)$ when $nh\varepsilon = \mathcal{O}(1)$) , there exists a near identity canonical transformation $(\boldsymbol{I}, \boldsymbol{\varphi}) \mapsto (\boldsymbol{J}, \boldsymbol{\theta})$ of order ε such that the solution of S_h^θ in $\boldsymbol{J}, \boldsymbol{\theta}$ variable satisfies

$$\begin{cases} \|\boldsymbol{J}_n - \boldsymbol{J}_0\|_2 \leq C'\varepsilon, \\ \|\boldsymbol{\theta}_n - (\boldsymbol{\theta}_0 + \tilde{\boldsymbol{\omega}}\left(\boldsymbol{J}_0\right)nh)\|_2 \leq C'\varepsilon nh, \end{cases} \tag{35}$$

for some constant C' (ε independent) $\forall n \leq h^{-1}\varepsilon^{-1}$ (so that $nh\varepsilon$ is of order $\mathcal{O}(1)$) with $\tilde{\boldsymbol{\omega}}\left(\cdot\right)$ defined in Lemma. 10. Note that the canonical transformation holds for all $(\boldsymbol{I}_i, \boldsymbol{\varphi}_i) \leftrightarrow (\boldsymbol{J}_i, \boldsymbol{\theta}_i)$, we have $\|\boldsymbol{I}_i - \boldsymbol{J}_i\|_2 \leq k\varepsilon$, $\|\boldsymbol{\varphi}_i - \boldsymbol{\theta}_i\|_2 \leq k\varepsilon$, $\forall i \in \mathbb{N}$ for some constant $k > 0$ and $\|\tilde{\boldsymbol{\omega}}\left(\boldsymbol{J}_0\right) - \boldsymbol{\omega}\left(\boldsymbol{I}_0\right)\|_2 \leq k'\varepsilon$ for some positive constants k'. Applying triangular inequality, $\exists\, C > 0$, such that

$$\begin{cases} \|\boldsymbol{I}_n - \boldsymbol{I}_0\|_2 \leq C\varepsilon, \\ \|\boldsymbol{\varphi}_n - (\boldsymbol{\varphi}_0 + \boldsymbol{\omega}\left(\boldsymbol{I}_0\right)nh)\|_2 \leq C\varepsilon nh, \end{cases}$$

for $n \leq h^{-1}\varepsilon^{-1}$. □

Remark 10. *$h^{-1}\varepsilon^{-1}$ is actually a conservative bound for n and one can extend the bound to be $h^{-1}\varepsilon^{(1-N)/2}$, $\forall N \geq 3$. Since N can be arbitrary, even if ε cannot be made infinitesimal, as long as it is below a threshold, the time of validity of the error bound in Thm. 3 can be extended to arbitrarily long.*

B. Experimental Details

Like most neural-network based algorithms for learning dynamics, the full potential of GFNN is achieved in the data rich regime. When preparing training data, we not only prepared in an unbiased way, but also emphasized on fair comparisons so that each of the existing methods is given the same or more training data.

More specifically, for each experiment, the training set contains a number of sequences starting with different initial conditions. When training for predicting Hamiltonian dynamics (continuous), each sequence in the training set is stroboscopically sampled from simulated ground truth, which is obtained using high-order numerical integrator with sufficiently small timestep $\tau \ll h$. For each experiment, the data set with sequences of length 2 will be denoted as $\mathcal{D}_2$, and the data set with sequences of length 5 will be denoted as $\mathcal{D}_5$. VFNN, HNN, SRNN (seq_len=2), and GFNN are trained with the same data set $\mathcal{D}_2$, while SRNN (seq_len=5) is trained with $\mathcal{D}_5$. Note the number of flow maps (ϕ) needed for each sequence in $\mathcal{D}_5$ is 4, while the number of maps for each sequence of $\mathcal{D}_2$ is 1. Therefore, for fairness, the number of sequences in $\mathcal{D}_2$, $n_{train}(\mathcal{D}_2)$, is set to be four times $n_{train}(\mathcal{D}_5)$ in most examples (exceptions will be explained).

All experiments are performed with PyTorch (CUDA) on a machine with GeForce RTX 3070 graphic card, AMD

Ryzen 7 3700X 8-Core Processor, 16 GB memory and the Linux distribution of openSUSE Leap 15.2.

Codes are provided.

B.1. 2-Body Problem

The step size of each data sequence in $\mathcal{D}_2$ and $\mathcal{D}_5$ is $h = 0.1$. The ground truth trajectory is simulated using a $4th$ order symplectic integrator with step size 10^{-3}. The initial condition of each data sequence is uniformly drawn from the orbits with semi-major axis $a \in (0.8, 1.2)$, eccentricity $e \in (0, 0.05)$. In terms of the number of samples, $n_{train}(\mathcal{D}_2) = 100,000$, $n_{train}(\mathcal{D}_5) = 100,000$. Note SRNN (seq_len=5) is provided more training data $n_{train}(\mathcal{D}_5)$ than described above, which would be $25,000$ instead, because less training data didn't provide good performance. The time derivative data of the vector field based methods (VFNN, HNN) are generated using (1st-order) finite difference.

S_h^θ is represented using multilayer perceptron (MLP), with 5 layers and 200, 100, 50, 20 neurons in hidden layers. The Adam optimizer is utilized with batch size 200. The model is trained for more than 20 epochs with initial learning rate 0.01. HNN, SRNN, SympNets are trained by their provided codes. HNN, SRNN are trained under default training setups and SympNets is trained using LA-SympNets with 30 layers and 10 sublayers (deeper than their default setups for improved performance).

B.2. Hénon-Heiles System

The step size of each data sequence in $\mathcal{D}_2$ and $\mathcal{D}_5$ is $h = 0.5$. The ground truth trajectory is simulated using a $4th$ order symplectic integrator with step size 10^{-3}. The initial condition of each data sequence is drawn from a centered Gaussian perturbation of states along one orbit randomly with variance 0.01^2. In terms of the number of samples, $n_{train}(\mathcal{D}_2) = 100,000$, $n_{train}(\mathcal{D}_5) = 25,000$. The data sets for the regular motion experiment and for the chaotic dynamics experiment are generated separably around a trajectory with energy level $\frac{1}{12}$ and $\frac{1}{6}$ respectively. The time derivative data of the vector field based methods (VFNN, HNN) are generated using (1st-order) finite difference.

The MLP that represents S_h^θ has 5 layers and 200, 100, 50, 20 neurons in hidden layers. The Adam optimizer is utilized with batch size 200. The model is trained for more than 20 epochs with initial learning rate 0.01. HNN, SRNN are trained by their provided codes under default training setups.

B.3. PCR3BP

The step size of each data sequence in $\mathcal{D}_2$ and $\mathcal{D}_5$ is $h = 0.1$. The ground truth trajectory is simulated using RK4 with step size 10^{-3}. The initial condition of each data sequence is drawn from a centered Gaussian perturbation of states along one orbit randomly with variance 0.05^2. In terms of the number of samples, $n_{train}(\mathcal{D}_2) = 100,000$, $n_{train}(\mathcal{D}_5) = 25,000$. The time derivative data of the vector field based methods (VFNN, HNN) are generated using (1st-order) finite difference.

The MLP that represents S_h^θ has 5 layers and 200, 100, 50, 20 neurons in hidden layers. The Adam optimizer is utilized with batch size 200. The model is trained for more than 20 epochs with initial learning rate 0.01. HNN, SRNN are trained by their provided codes under default training setups.

B.4. Standard Map

The step size of each data sequence in $\mathcal{D}_2$ and $\mathcal{D}_5$ is $h = 1$. The ground truth map is directly evolved from the discrete-time evolution map Eq. (9). The initial condition of each data sequence is drawn from a Gaussian perturbation of states along one orbitrandomly with variance 0.5^2. In terms of the number of samples in training / testing data, $n_{train}(\mathcal{D}_2) = 1,000,000$, $n_{train}(\mathcal{D}_5) = 250,000$. The data sets of the regular motion experiment and for the chaotic dynamics experiment are generated separably with $K = 0.6$ and $K = 1.2$ and correspondingly different initial conditions respectively. The time derivative data of the vector field based methods (VFNN, HNN) are generated using (1st-order) finite difference (with $\Delta t = 1$).

The MLP that represents S_h^θ has 5 layers and 500, 500, 200, 20 neurons in hidden layers. The Adam optimizer is utilized with batch size 1000. The model is trained for more than 20 epochs with initial learning rate 0.001. HNN, SRNN are trained by their provided codes under default training setups.