SA-NODEs and Universal Approximation of Dynamical Systems

Lorenzo Liverani
Joint work with Z. Li, K. Liu, E. Zuazua

Friedrich-Alexander-Universität
DYNAMICS, CONTROL,
MACHINE LEARNING
AND NUMERICS

August 19, 2024

Data-Driven Modeling

Data:

- Time-dependent
- Scalar
- Vector
- Discrete
- Continuous
- ...

$\longrightarrow$

Model: A mathematical object fitting the data

- Least squares
- Manifold learning
- ODE system
- PDE system
- ...

Our Setting

Data:

$\mathcal{D} = \{z_k(t)\}_{k=1}^{N} \subset \mathbb{R}^d$, for $t \in [0, T]$

In practice:

$\mathcal{D} = \{z_k(t_l)\}_{k,l} \subset \mathbb{R}^d$, for $k = 1, \ldots, N$, $l = 1, \ldots, M$

Model:

Our Setting

Data:

$\mathcal{D} = \{z_k(t)\}_{k=1}^{N} \subset \mathbb{R}^d$, for $t \in [0, T]$

In practice:

$\mathcal{D} = \{z_k(t_l)\}_{k,l} \subset \mathbb{R}^d$, for $k = 1, \ldots, N$, $l = 1, \ldots, M$

Model: System of ODEs

$$\begin{cases} \dot{x} = \sum_{i=1}^{r} W_i \sigma(A^i x + A^{i,0} - W_0), \\ x(0) = x_0. \end{cases}$$

Our Setting

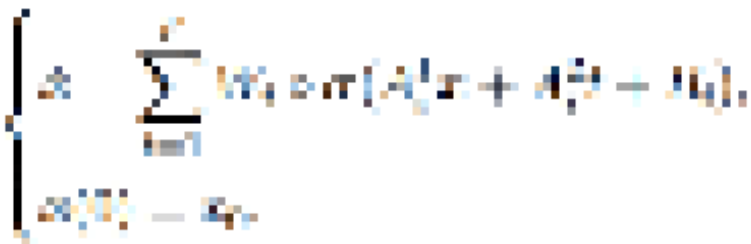
Data:

$\mathcal{D} = \{z_k(t)\}_{k=1}^{N} \subset \mathbb{R}^d$, for $t \in [0, T]$

In practice:

$\mathcal{D} = \{z_k(t_l)\}_{k,l} \subset \mathbb{R}^d$, for $k = 1, \ldots, N$, $l = 1, \ldots, M$

Model: System of ODEs



Goal: Trajectory tracking

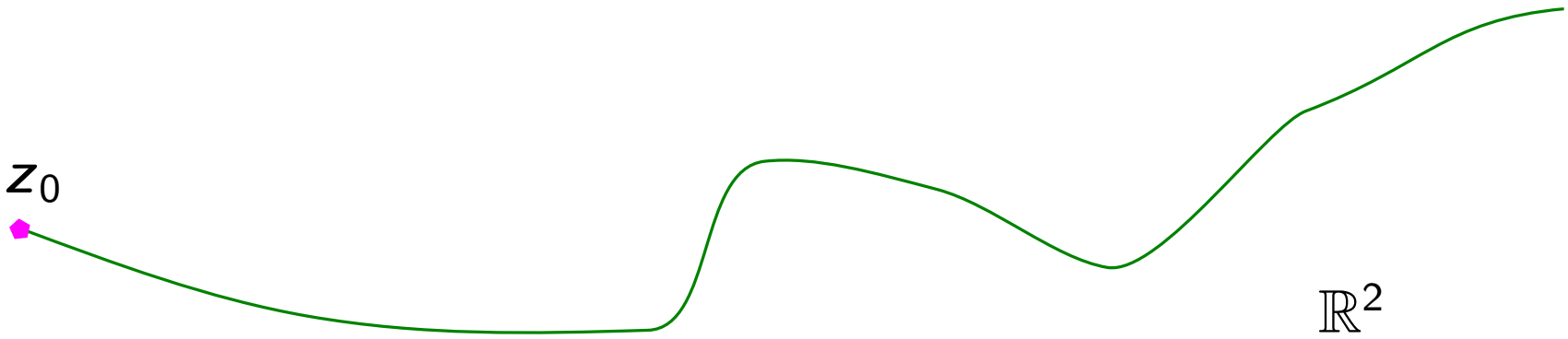


Figure: We start with a curve (trajectory) in $\mathbb{R}^2$

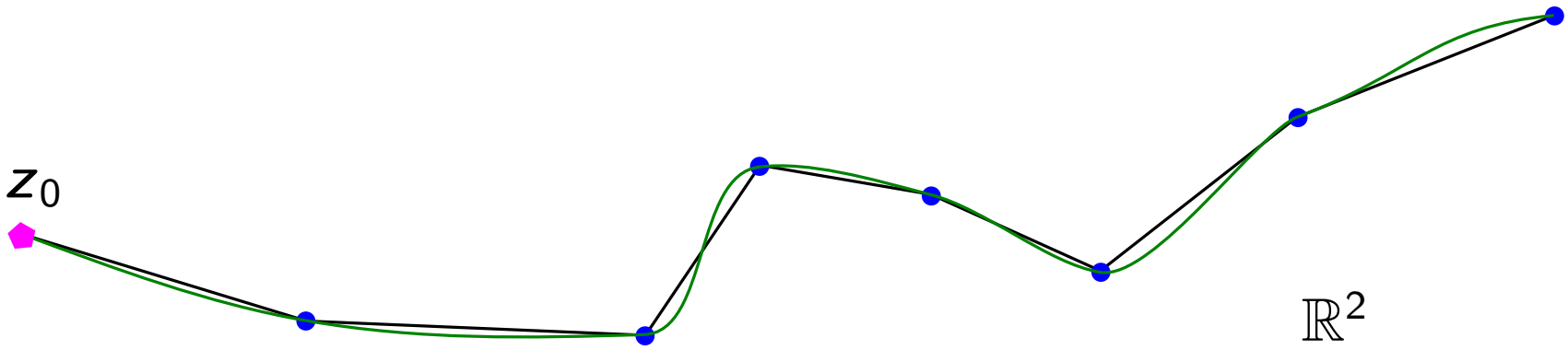


Figure: In reality we only have some points $\boldsymbol{z}(t_l)$

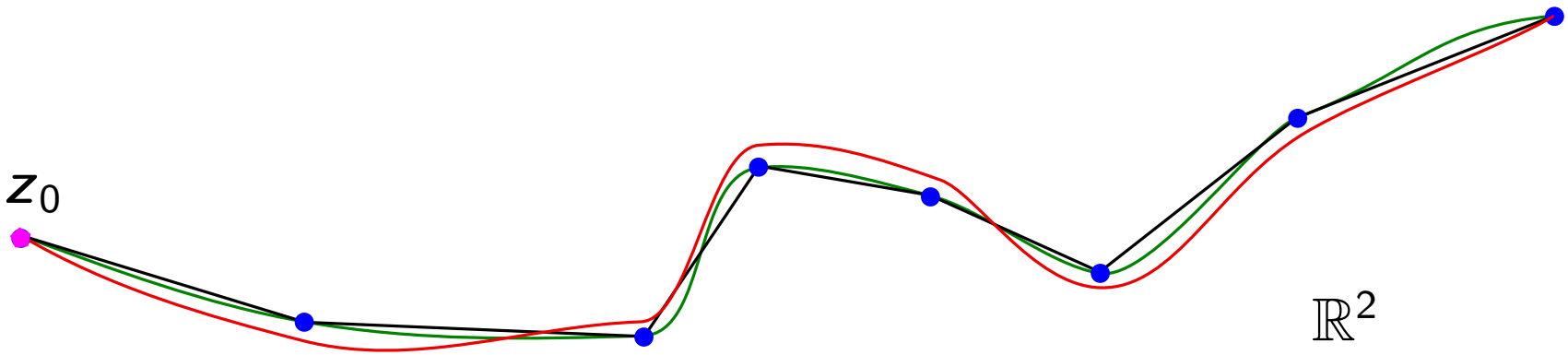


Figure: We seek a dynamical system such that the solution starting at z_0 is close to the measured trajectory

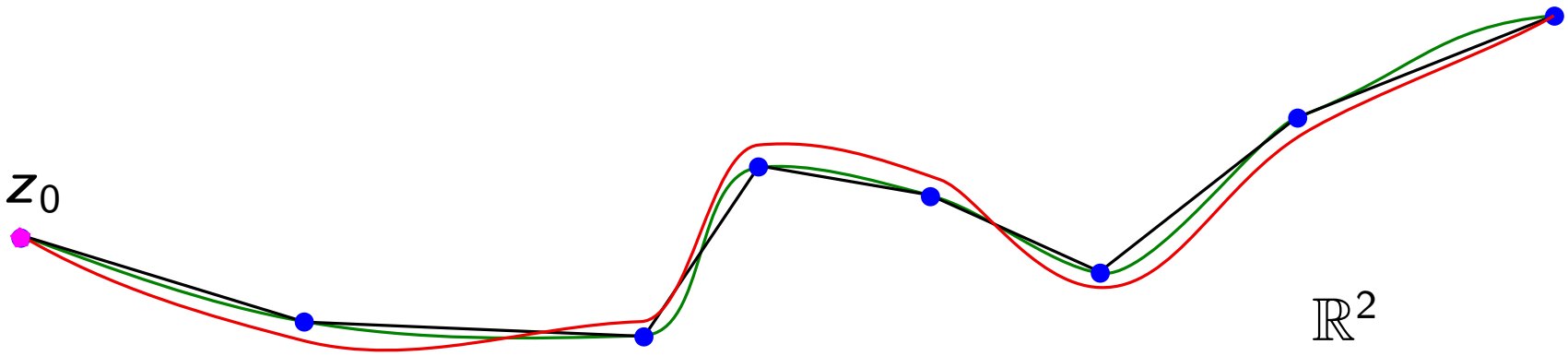
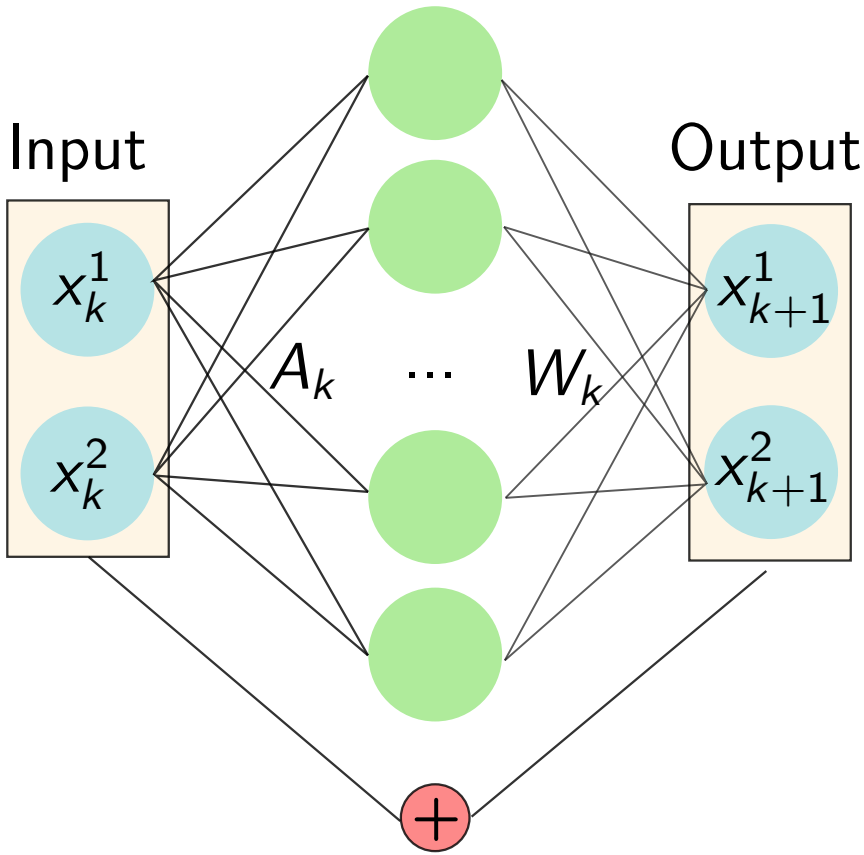


Figure: We seek a dynamical system such that the solution starting at z_0 is close to the wanted trajectory

- We are interested in the case data comes from a dynamical system

$$\begin{cases} \dot{z}(t) = f(z(t), t) \\ z(0) = z_0 \end{cases}$$

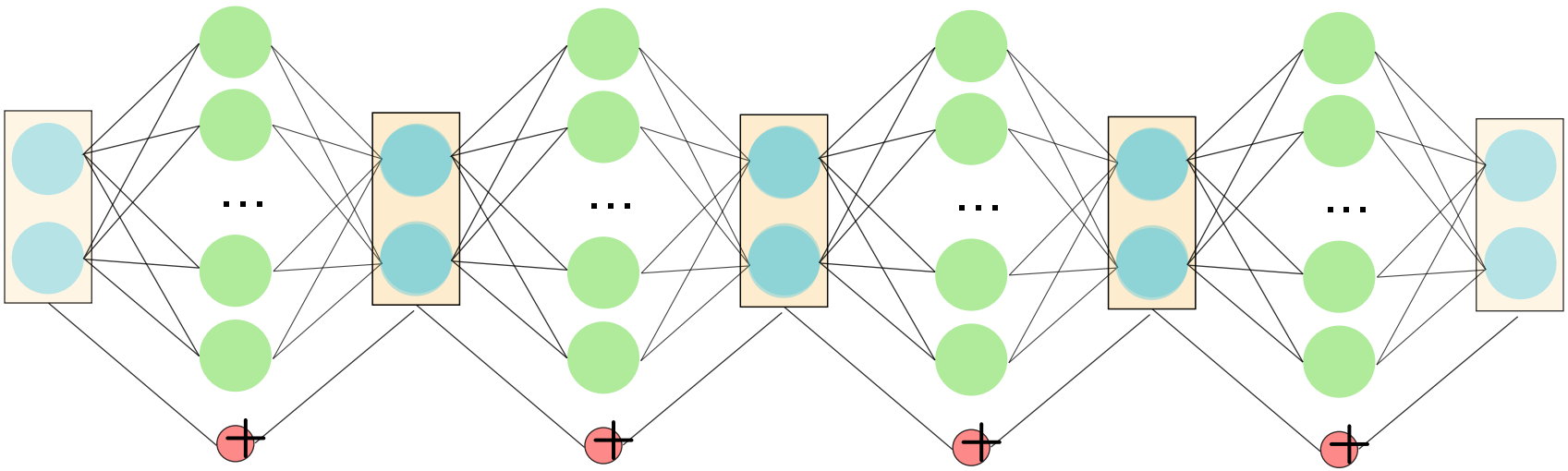
Neural ODEs



ResNets block: $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + W_k \boldsymbol{\sigma}(A_k \boldsymbol{x}_k + B_k)$

$\boldsymbol{\sigma}$ is an activation function (e.g. ReLU)

Neural ODEs



$$\begin{cases} \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + W_k \boldsymbol{\sigma}(A_k \boldsymbol{x}_k + B_k) \\ \boldsymbol{x}_0 = \boldsymbol{x}_0 \end{cases}$$

$$\begin{cases} \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + W_k \boldsymbol{\sigma}(A_k \boldsymbol{x}_k + B_k) \\ \boldsymbol{x}_0 = \boldsymbol{x}_0 \end{cases}$$

$$\downarrow$$

$$\begin{cases} \dot{\boldsymbol{x}}(t) = W(t) \boldsymbol{\sigma}(A(t) \boldsymbol{x}(t) + B(t)) \\ \boldsymbol{x}(0) = \boldsymbol{x}_0 \end{cases}$$

We use the following, equivalent notation to highlight the number P of neurons

$$\begin{cases} \dot{\boldsymbol{x}} = \sum_{i=1}^{P} W_i(t) \circ \boldsymbol{\sigma}(A_i(t) \boldsymbol{x} + B_i(t)) \\ \boldsymbol{x}(0) = x_0 \end{cases}$$

Here $\circ$ is the Hadamard product $(a, b) \circ (c, d) = (ac, bd)$

NODE Results

NODE were first introduced by Chen et al. (2018)

- Controllability: Find $W_i(t)$, $A_i(t)$ and $B_i(t)$ to drive an initial input to a final output (Geshkovski, Ruiz-Balet, Zuazua, Cheng et al., ...)
- Approximation: Relation between the parameters with the approximation properties of NODEs (Alvarez-Lopez, ...)
- Long-time behavior (Geshkovski, Zuazua, ...)
- Formal limit of ResNets (Massaroli, Sander et al., ...)

A time issue

$$\begin{cases} \dot{\boldsymbol{x}} = \sum_{i=1}^{P} W_i(t) \circ \boldsymbol{\sigma}(A_i(t)\boldsymbol{x} + B_i(t)) \\ \boldsymbol{x}(0) = x_0 \end{cases}$$

Training a NODE entails finding functions $A_i(t)$, $W_i(t)$ and $B_i(t)$ which depend on time $\rightsquigarrow$ In practice we find the values of these functions on a set of time steps

- The number of parameters scales as the number of time steps $\Rightarrow$ High complexity
- Impossible to make predictions

Our Setting

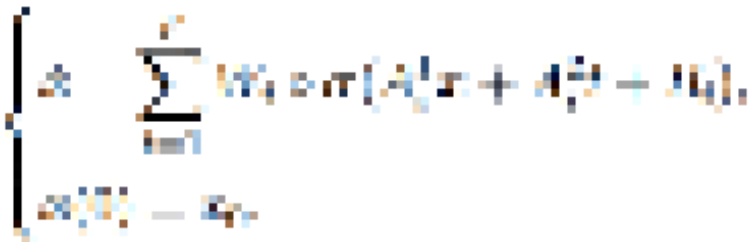
Data:

$\mathcal{D} = \{z_k(t)\}_{k=1}^{N} \subset \mathbb{R}^d$, for $t \in [0, T]$

In practice:

$\mathcal{D} = \{z_k(t_l)\}_{k,l} \subset \mathbb{R}^d$, for $k = 1, \ldots, N$, $l = 1, \ldots, M$

Model: System of ODEs



Goal: Trajectory tracking

Our Setting

Data:

$\mathcal{D} = \{z_k(t)\}_{k=1}^N \subset \mathbb{R}^d$,
for $t \in [0, T]$

In practice:
$\mathcal{D} = \{z_k(t_l)\}_{k,l} \subset \mathbb{R}^d$,
for $k = 1, \dots, N$,
$l = 1, \dots, M$

Model: System of ODEs

$$\begin{cases} \dot{x} = \sum_{i=1}^{P} W_i \circ \sigma(A_i^1 x + A_i^2 t + B_i) \\ x(0) = z_0 \end{cases}$$

Goal: Trajectory tracking

Semi-autonomous NODEs

$$\dot{x} = \sum_{i=1}^{P} W_i \circ \sigma(A_i^1 x + A_i^2 t + B_i), \quad x(t) \in \mathbb{R}^d, t \in [0, T]$$

- σ is the activation function (e.g. ReLU)
- $W_i \in \mathbb{R}^d$, $A_i^1 \in \mathbb{R}^{d \times d}$, $A_i^2 \in \mathbb{R}^d$ $B_i \in \mathbb{R}^d$
- Parameters independent of time $\rightsquigarrow$ Total: $Pd(d+3)$

Motivation

Universal Approximation property (UAP) of ReLU:

Theorem (Pinkus, 1999)

Fix a compact set $X \subseteq \mathbb{R}^{d+1}$. Let σ be a non-polynomial continuous function. For any function $g \in \mathcal{C}(X; \mathbb{R}^d)$ and $\varepsilon > 0$, $\exists P$ and parameters $(W_i, A_i, B_i) \in \mathbb{R}^d \times \mathbb{R}^{(d+1)\times d} \times \mathbb{R}^d$, for $i = 1, \dots, P$, such that, calling

$$f_\Theta(x) = \sum_{i=1}^{P} W_i \circ \boldsymbol{\sigma}(A_i x + B_i), \quad \forall x \in X,$$

it holds

$$\|g - f_\Theta\|_{\mathbb{L}^\infty(X; \mathbb{R}^d)} \leq \varepsilon.$$

Motivation

Using the UAP for $f(x, t)$:

$$f(x, t) \sim \sum_{i=1}^{P} W_i \circ \sigma(A_i^1 x + A_i^2 t + B_i)$$

$$\downarrow$$

$$\begin{cases} \dot{z} = f(z, t) \\ z(0) = z_0 \end{cases} \sim \begin{cases} \dot{x} = \sum_{i=1}^{P} W_i \circ \sigma(A_i^1 x + A_i^2 t + B_i) \\ x(0) = z_0 \end{cases}$$

Universal Approximation Property

Theorem (Li, Liu, L., Zuazua)

Let f be uniformly Lipschitz in z with respect to t. For any compact set $K \subseteq \mathbb{R}^d$ and any $\varepsilon > 0$, there exists a constant $P_{\varepsilon,T,K,f}$ such that for any $P \geq P_{\varepsilon,T,K,f}$, there exist parameters $(W_i, A_i^1, A_i^2, B_i) \in \mathbb{R}^d \times \mathbb{R}^{d\times d} \times \mathbb{R}^d \times \mathbb{R}^d$, for $i = 1, \dots, P$, such that

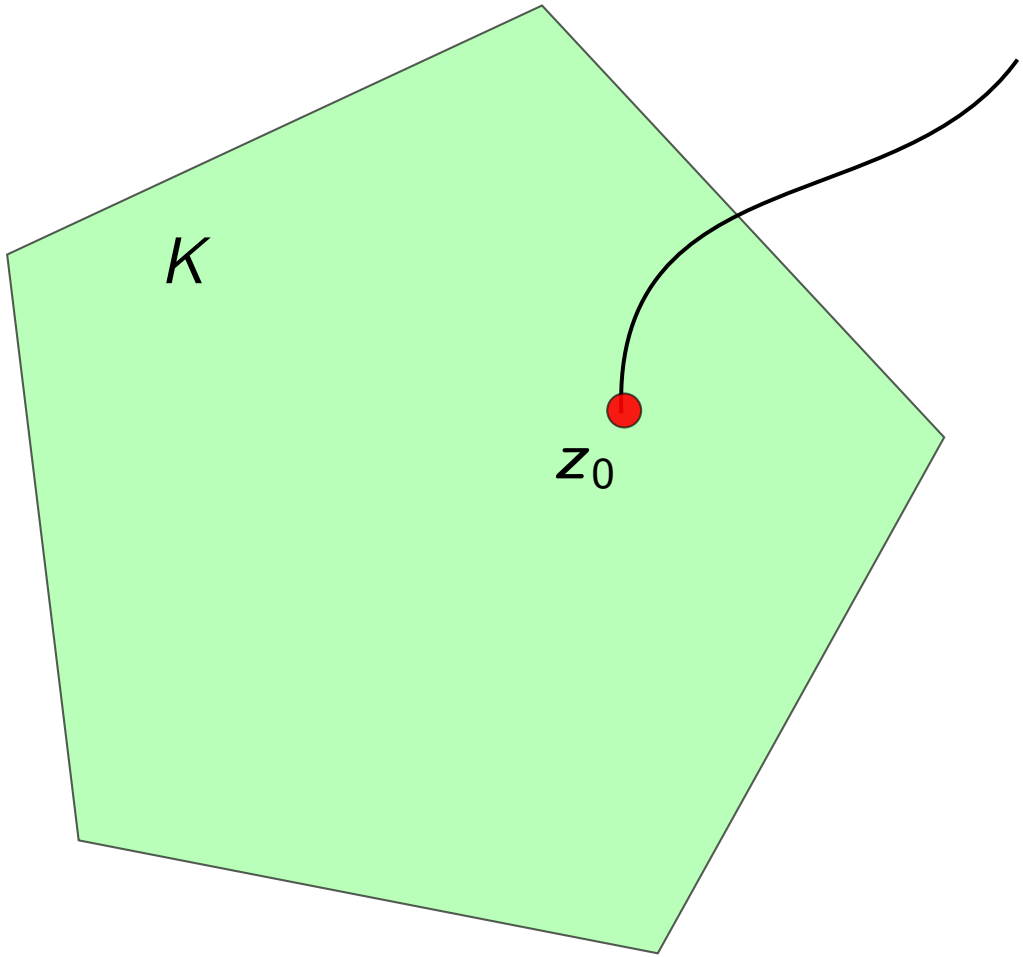
$$\|z_{z_0}(\cdot) - x_{z_0}(\cdot)\|_{\mathbb{L}^\infty([0,T];\mathbb{R}^d)} \leq \varepsilon, \quad \forall z_0 \in K$$

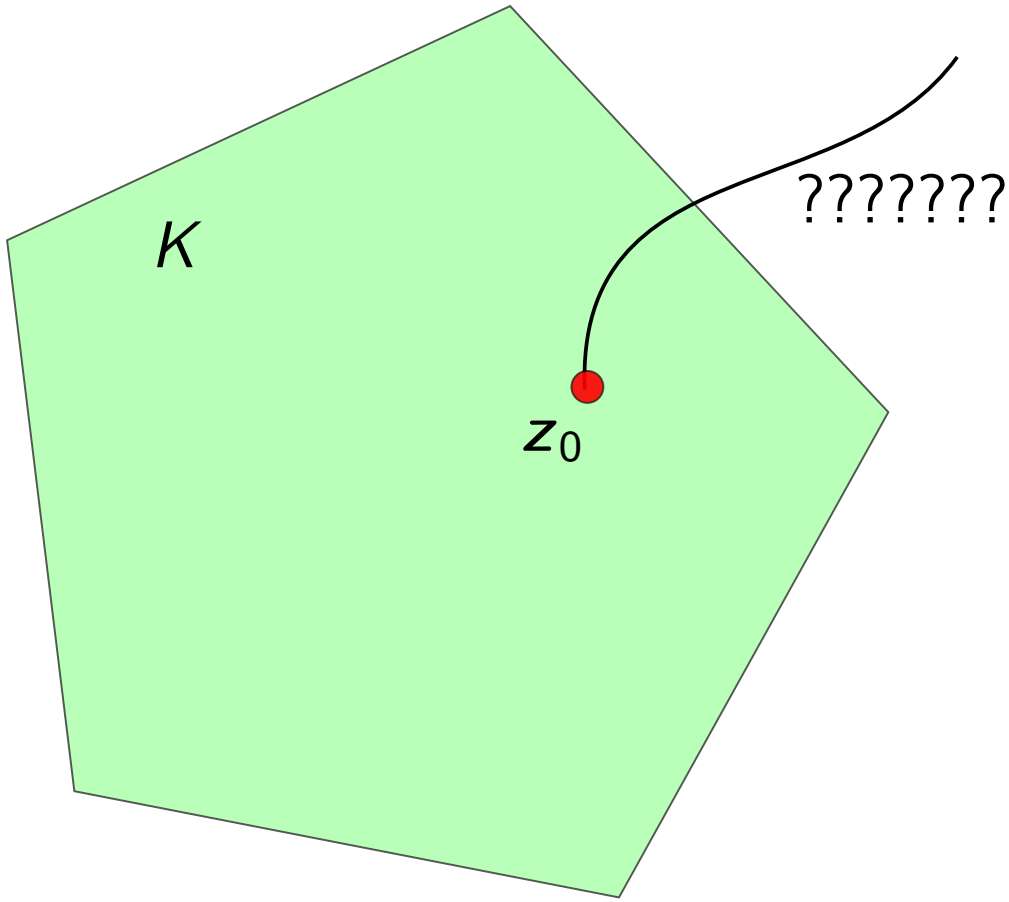
$\rightsquigarrow$ We approximate the **global** flow of the ODE for initial data starting in K

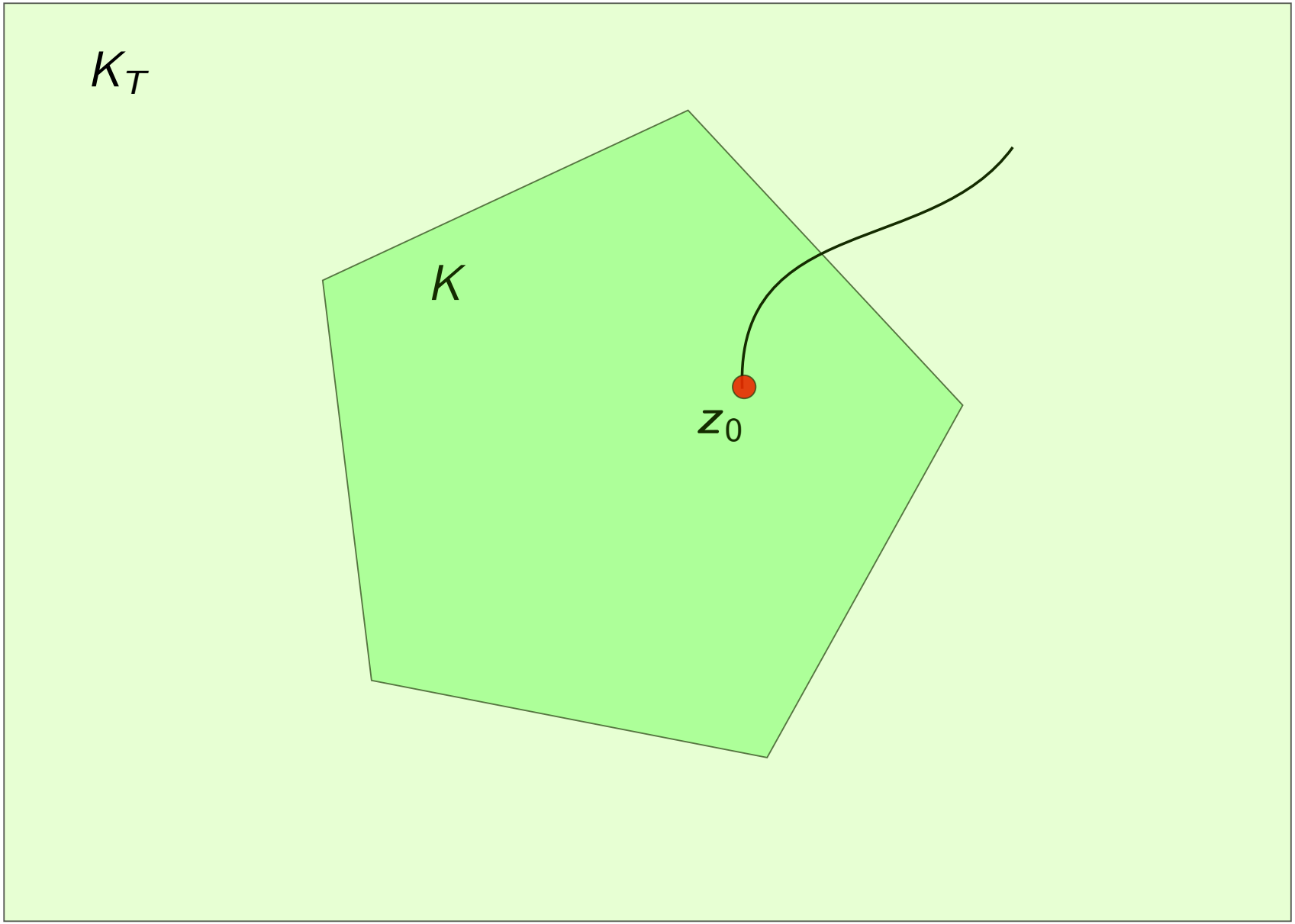
Proof

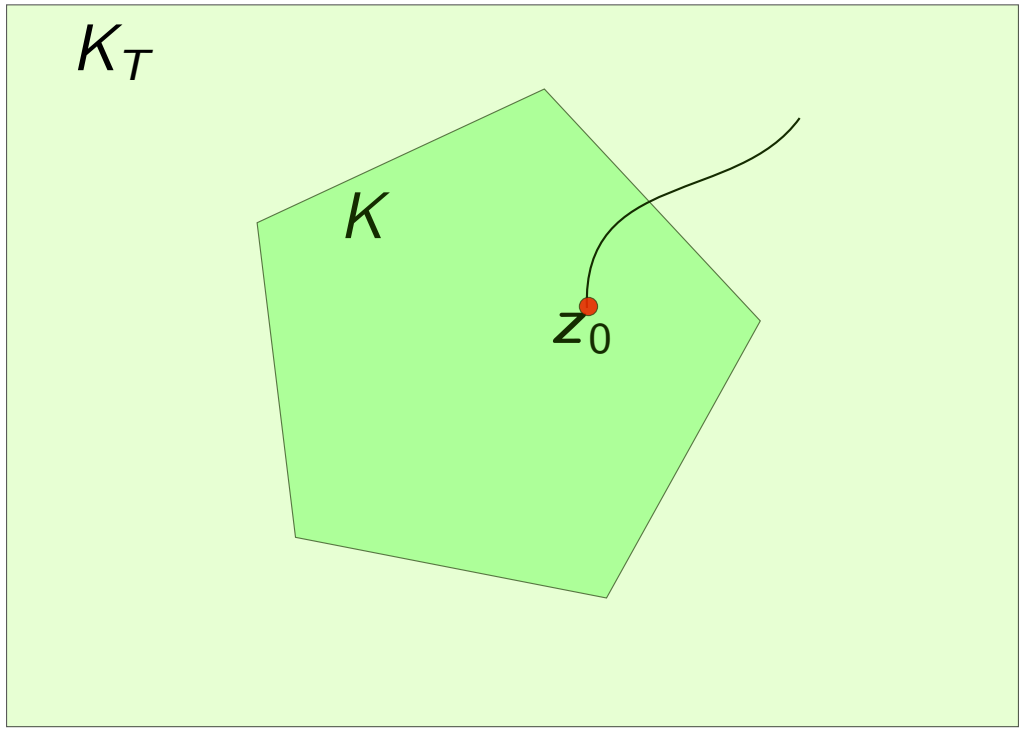
Idea: Apply UAP to f and use Gronwall inequality

Problem: For UAP we need to fix a compact set


K
z_0


K
???????
z_0


K_T
K
z_0



$$K_T := \left\{ \|x\| \leq \sup_{z \in K} \left(\|z\| + T + \int_0^T \|f(0,s)\| ds \right) \exp(LT) \right\}$$

where L is the Lipschitz constant of f

$\rightsquigarrow$ Every solution of $\dot{x} = f_1(x,t)$, for $t \in [0,T]$ and $x_0 \in K$ and $\|f_1 - f\|_\infty \leq 1$ stays inside K_T

Proof

Apply UAP in K_T to f, obtaining

$$f_\Theta(x,t) = \sum_{i=1}^{P} W_i \circ \sigma(A_i^1 x + A_i^2 t + B_i)$$

$\rightsquigarrow$ By Lipschitz continuity + UAP:

$$\begin{aligned}
&\|\boldsymbol{z}_{z_0}(t) - \boldsymbol{x}_{z_0}(t)\| \\
&= \left\| z_0 + \int_0^t f(\boldsymbol{z}_{z_0}(s), s) ds - z_0 - \int_0^t f_\Theta(\boldsymbol{x}_{z_0}(s), s) ds \right\| \\
&\leq \int_0^t \|f(\boldsymbol{z}_{z_0}(s), s) - f(\boldsymbol{x}_{z_0}(s), s) + f(\boldsymbol{x}_{z_0}(s), s) - f_\Theta(\boldsymbol{x}_{z_0}(s), s)\| \, ds \\
&\leq L \int_0^t \|\boldsymbol{z}_{z_0}(s) - \boldsymbol{x}_{z_0}(s)\| ds + \varepsilon t
\end{aligned}$$

for any $t \leq T$

Proof

Apply UAP in K_T to f, obtaining

$$f_\Theta(x,t) = \sum_{i=1}^{P} W_i \circ \sigma(A_i^1 x + A_i^2 t + B_i)$$

$\rightsquigarrow$ By Lipschitz continuity + UAP:

$$\begin{aligned}
&\|z_{z_0}(t) - x_{z_0}(t)\| \\
&= \left\| z_0 + \int_0^t f(z_{z_0}(s), s) ds - z_0 - \int_0^t f_\Theta(x_{z_0}(s), s) ds \right\| \\
&\leq \int_0^t \| f(z_{z_0}(s), s) - f(x_{z_0}(s), s) + f(x_{z_0}(s), s) - f_\Theta(x_{z_0}(s), s) \| \, ds \\
&\leq L \int_0^t \| z_{z_0}(s) - x_{z_0}(s) \| ds + \varepsilon t
\end{aligned}$$

for any $t \leq T$

Proof

Apply UAP in K_T to f, obtaining

$$f_\Theta(x,t) = \sum_{i=1}^{P} W_i \circ \sigma(A_i^1 x + A_i^2 t + B_i)$$

$\rightsquigarrow$ By Lipschitz continuity + UAP:

$$\begin{aligned}
&\|\boldsymbol{z}_{z_0}(t) - \boldsymbol{x}_{z_0}(t)\| \\
&= \left\| z_0 + \int_0^t f(\boldsymbol{z}_{z_0}(s), s) ds - z_0 - \int_0^t f_\Theta(\boldsymbol{x}_{z_0}(s), s) ds \right\| \\
&\leq \int_0^t \|f(\boldsymbol{z}_{z_0}(s), s) - f(\boldsymbol{x}_{z_0}(s), s) + f(\boldsymbol{x}_{z_0}(s), s) - f_\Theta(\boldsymbol{x}_{z_0}(s), s)\| \, ds \\
&\leq L \int_0^t \|\boldsymbol{z}_{z_0}(s) - \boldsymbol{x}_{z_0}(s)\| ds + \varepsilon t
\end{aligned}$$

for any $t \leq T$

Proof

We have obtained

$$\|z_{z_0}(t) - x_{z_0}(t)\| \leq L \int_0^t \|z_{z_0}(s) - x_{z_0}(s)\| ds + \varepsilon t$$

By the Gronwall inequality

$$\|z_{z_0} - x_{z_0}\|_{\mathbb{L}^\infty([0,T];\mathbb{R}^d)} \leq \varepsilon T e^{LT}$$

Approximation Rate

- Quantitative version of the previous result, with respect to the number p of neurons

Theorem (Li, Liu, L., Zuazua)

Let $\boldsymbol{f} \in \mathcal{H}^k_{\text{loc}}(\mathbb{R}^d \times [0,T]; \mathbb{R}^d)$, for $k > (d+1)/2 + 2$. Fix any compact set $K \subseteq \mathbb{R}^d$. Then, for any $P \in \mathbb{N}_+$, there exist parameters $(W_i, A_i^1, A_i^2, B_i) \in \mathbb{R}^d \times \mathbb{R}^{d\times d} \times \mathbb{R}^d \times \mathbb{R}^d$, for $i = 1, \dots, P$, such that

$$\sup_{t\in[0,T]} \int_K \|z_{z_0}(t) - \boldsymbol{x}_{z_0}(t)\|^2 dz_0 \leq \frac{C_{T,K,f}}{P},$$

where $C_{T,K,f}$ is a constant independent of P

The Barron Space

Fix a compact set $X \subseteq \mathbb{R}^n$. The Barron space is

$$\mathcal{S}_B(X) := \Big\{ f \in \mathcal{C}(X;\mathbb{R}) \,\Big|\, \exists \mu \in \mathcal{P}(\mathbb{R}^{n+2}) \text{ s.t. } f(x) = \int_{\mathbb{R}^{n+2}} w\sigma(\langle a, x\rangle + b) d\mu(w,a,b),\ \forall x \in X \Big\}$$

The Barron norm:

$$\|f\|_{\mathcal{S}_B(X)} = \inf_{\mu \text{ satisfies above}} \int_{\mathbb{R}^{n+2}} |w|(\|a\|_{\ell^1} + |b|) d\mu(w,a,b)$$

The definition extends immediately to $f \in \mathcal{C}(X, \mathbb{R}^d)$

A quantitative lemma

Lemma (E, Ma, Wu (2022))

Let $f \in \mathcal{S}_B^d(X)$. For any $P \geq 1$, there exists $(W_i, A_i, B_i) \in \mathbb{R}^d \times \mathbb{R}^{d\times n} \times \mathbb{R}^d$, for $i = 1, \ldots, P$, such that

$$\left\| f(\cdot) - \sum_{i=1}^{P} W_i \circ \boldsymbol{\sigma}(A_i \cdot + B_i) \right\|_{\mathbb{L}^2(X;\mathbb{R}^d)}^2 \leq \frac{3\|f\|_{\mathcal{S}_B^d(X)}^2}{P}.$$

Moreover,

$$\left\| \sum_{i=1}^{P} |W_i| \circ (\|A_i\|_{\ell^1} + |B_i|) \right\| \leq 2\|f\|_{\mathcal{S}_B^d(X)}.$$

Proof

Step 1. f belongs to the Barron space. This follows since $f \in \mathcal{H}^k_{\text{loc}}(\mathbb{R}^d \times [0,T]; \mathbb{R}^d)$, for $k > (d+1)/2 + 2$

Step 2. Thanks to the Lemma, we find parameter $\Theta = (W_i, A_i^1, A_i^2, B_i)_{i=1}^P$ such that

$$\|f(\cdot,\cdot) - f_\Theta(\cdot,\cdot)\|^2_{\mathbb{L}^2(X;\mathbb{R}^d)} \leq \frac{3\|f\|^2_{\mathcal{S}^d_B(X)}}{P}$$

$$\left\| \sum_{i=1}^P |W_i| \circ (\|A_i^1\|_{\ell^1} + |A_i^2| + |B_i|) \right\| \leq 2\|f\|_{\mathcal{S}^d_B(X)}$$

Accordingly, it is not difficult to show

$$\|f_\Theta(x,t) - f_\Theta(y,t)\| \leq 2\|f\|_{\mathcal{S}^d_B}\|x-y\| \quad \forall (x,y) \in \mathbb{R}^d, t \in [0,T]$$

Proof

Step 3. We have

$$
\begin{aligned}
&\|\boldsymbol{z}_{z_0}(t) - \boldsymbol{x}_{z_0}(t)\|^2 \\
&= \left\| \int_0^t f(\boldsymbol{z}_{z_0}(s), s) - f_\Theta(\boldsymbol{z}_{z_0}(s), s) + f_\Theta(\boldsymbol{z}_{z_0}(s), s) - f_\Theta(\boldsymbol{x}_{z_0}(s), s) ds \right\|^2 \\
&\leq 2t \int_0^t \|f(\boldsymbol{z}_{z_0}(s), s) - f_\Theta(\boldsymbol{z}_{z_0}(s), s)\|^2 ds \\
&\qquad + 2t \int_0^t \|f_\Theta(\boldsymbol{z}_{z_0}(s), s) - f_\Theta(\boldsymbol{x}_{z_0}(s), s)\|^2 ds
\end{aligned}
$$

Proof

Step 4. Recall:

$$\|z_{z_0}(t) - x_{z_0}(t)\|^2 \leq 2t \int_0^t \|f(z_{z_0}(s), s) - f_\Theta(z_{z_0}(s), s)\|^2 ds + 2t \int_0^t \|f_\Theta(z_{z_0}(s), s) - f_\Theta(x_{z_0}(s), s)\|^2 ds$$

We integrate in space:

$$\int_K \|z_{z_0}(t) - x_{z_0}(t)\|^2 dz_0 \leq 2t \int_0^t \int_K \|f(z_{z_0}(s), s) - f_\Theta(z_{z_0}(s), s)\|^2 dz_0 ds + 2t \int_0^t \int_K \|f_\Theta(z_{z_0}(s), s) - f_\Theta(x_{z_0}(s), s)\|^2 dz_0 ds$$

Proof

Step 4. Recall:

$$\|z_{z_0}(t) - x_{z_0}(t)\|^2 \leq 2t \int_0^t \|f(z_{z_0}(s), s) - f_\Theta(z_{z_0}(s), s)\|^2 ds + 2t \int_0^t \|f_\Theta(z_{z_0}(s), s) - f_\Theta(x_{z_0}(s), s)\|^2 ds$$

We integrate in space:

$$\int_K \|z_{z_0}(t) - x_{z_0}(t)\|^2 dz_0 \leq 2t \int_0^t \int_K \|f(z_{z_0}(s), s) - f_\Theta(z_{z_0}(s), s)\|^2 dz_0 ds + 2t \int_0^t \int_K \|f_\Theta(z_{z_0}(s), s) - f_\Theta(x_{z_0}(s), s)\|^2 dz_0 ds$$

Proof

Step 5. Call

$$Q(t) = \int_K \|z_{z_0}(t) - x_{z_0}(t)\|^2 dz_0$$

Since

$$\|f_\Theta(x,t) - f_\Theta(y,t)\| \leq 2\|f\|_{\mathcal{S}_B^d}\|x-y\|, \quad \forall (x,y) \in \mathbb{R}^d, t \in [0,T]$$

We have

$$2t \int_0^t \int_K \|f_\Theta(z_{z_0}(s),s) - f_\Theta(x_{z_0}(s),s)\|^2 dz_0 ds$$
$$\leq 8t\|f\|^2_{\mathcal{S}_B^d(X)} \int_0^t Q(s) ds$$

Proof

Step 6. Let $\phi_s \colon \mathbb{R}^d \to \mathbb{R}^d$, $z_0 \mapsto z_{z_0}(s)$

$$\int_K \|f(z_{z_0}(s), s) - f_\Theta(z_{z_0}(s), s)\|^2 dz_0$$

$$= \int_K \|f(\phi_s(z_0), s) - f_\Theta(\phi_s(z_0), s)\|^2 dz_0$$

$$\text{"} = \text{"} \int_{\phi_s(K)} \|f(x, s) - f_\Theta(x, s)\|^2 \frac{1}{|\det(\nabla \phi_s(z_0))|} dx$$

$$\leq e^{sLd} \int_{\phi_s(K)} \|f(x, s) - f_\Theta(x, s)\|^2 dx$$

Now we can use

$$\|f(\cdot, \cdot) - f_\Theta(\cdot, \cdot)\|^2_{\mathbb{L}^2(X; \mathbb{R}^d)} \leq \frac{3\|f\|^2_{\mathcal{S}^d_B(X)}}{P}$$

Proof

Step 6. (Continued) We finally find

$$2t \int_0^t \int_K \|f(z_{z_0}(s), s) - f_\Theta(z_{z_0}(s), s)\|^2 dz_0 ds \leq \frac{6t\|f\|^2_{\mathcal{S}^d_B(X)}}{P} e^{tLd}$$

Conclusion.

$$Q(t) \leq 8t\|f\|^2_{\mathcal{S}^d_B(X)} \int_0^t Q(s)ds + \frac{6t\|f\|^2_{\mathcal{S}^d_B(X)}}{P} e^{tLd}$$

and by Gronwall we conclude

Transport Equations

By the method of characteristics, SA-NODEs can be applied to approximate in a data-driven manner transport equations

$$\begin{cases} \partial_t \rho(x,t) + \mathrm{div}_x(\rho(x,t) f(x,t)) = 0 \\ \rho(\cdot,0) = \rho_0 \in \mathcal{M}(\mathbb{R}^d) \end{cases}$$

with the **neural transport equation**

$$\begin{cases} \partial_t \rho(x,t) + \mathrm{div}_x \left(\rho(x,t) \sum_{i=1}^{P} W_i \circ \boldsymbol{\sigma}(A_i^1 \boldsymbol{x} + A_i^2 t + B_i) \right) = 0 \\ \rho(\cdot,0) = \rho_0 \in \mathcal{M}(\mathbb{R}^d) \end{cases}$$

Transport Equations

Theorem (Li, Liu, L., Zuazua)

Let ρ_0 be a probability measure supported in a compact set K such that $\rho_0 \in \mathbb{L}^2(K)$. Then, for any $P \in \mathbb{N}_+$, there exist parameters $\Theta = \{(W_i, A_i^1, A_i^2, B_i)\}_{i=1}^P$ such that

$$\sup_{t\in[0,T]} \mathbb{W}_1(\rho(\cdot,t), \rho_\Theta(\cdot,t)) \leq \frac{C_{T,f,\rho_0}}{\sqrt{P}},$$

where C_{T,f,ρ_0} is a constant independent of P, $\mathbb{W}_1(\cdot,\cdot)$ is the Wasserstein-1 distance, and $\rho(\cdot,t)$ (resp. $\rho_\Theta(\cdot,t)$) is the solution of the transport equation (resp. the Neural transport equation) at the time $t \in [0,T]$.

Proof

By our assumptions

$$\rho(\cdot, t) = \phi_t \# \rho_0, \quad \rho_\Theta(\cdot, t) = \phi_{\Theta,t} \# \rho_0, \quad \forall t \in [0, T]$$

where $\#$ is the pushforward. Besides

$$\begin{aligned}
\mathbb{W}_1(\rho(\cdot, t), \rho_\Theta(\cdot, t)) &= \sup_{\mathrm{Lip}(g) \leq 1} \int_{\mathbb{R}^d} g(x) d\left(\rho(x, t) - \rho_\Theta(x, t)\right) \\
&= \sup_{\mathrm{Lip}(g) \leq 1} \int_K g(\phi_t(z)) - g(\phi_{\Theta,t}(z)) d\rho_0(z) \\
&\leq \int_K \|\phi_t(z) - \phi_{\Theta,t}(z)\| d\rho_0(z)
\end{aligned}$$

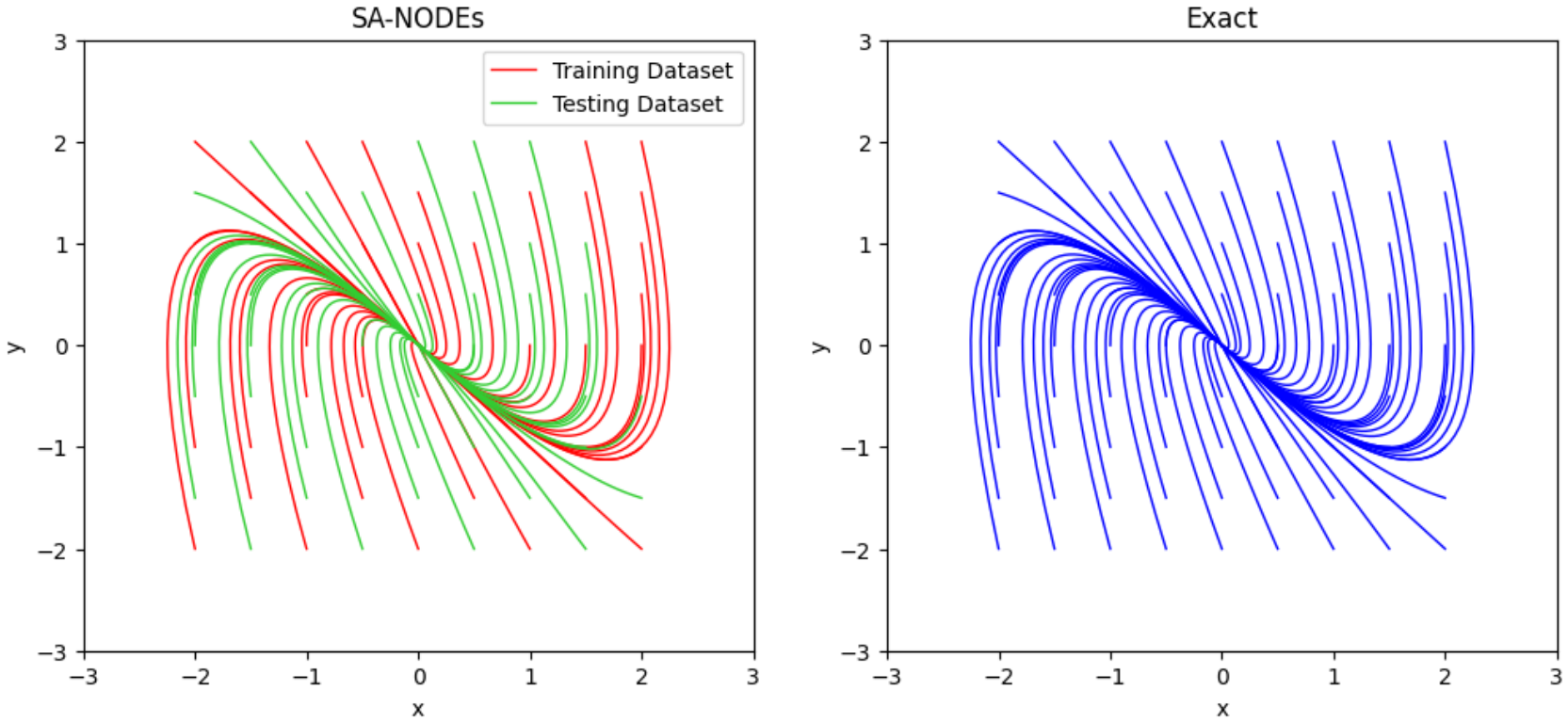
Proof

Finally, we have

$$\int_K \|\phi_t(z) - \phi_{\Theta,t}(z)\| d\rho_0(z)$$
$$\leq \|\rho_0\|^{1/2}_{\mathbb{L}^2(K)} \left(\int_K \|\phi_t(z) - \phi_{\Theta,t}(z)\|^2 dz \right)^{1/2}$$

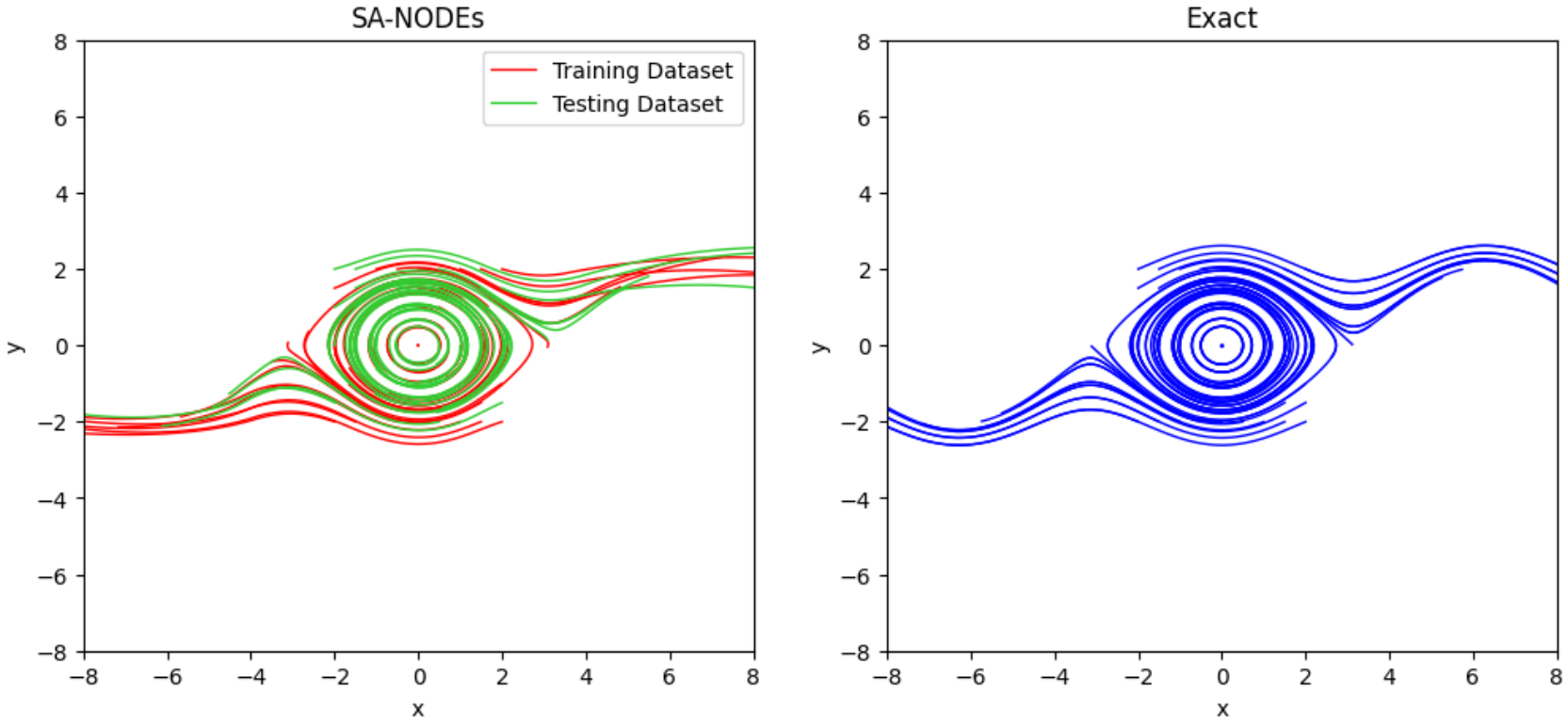
and we conclude applying the previous theorem

Numerical Results



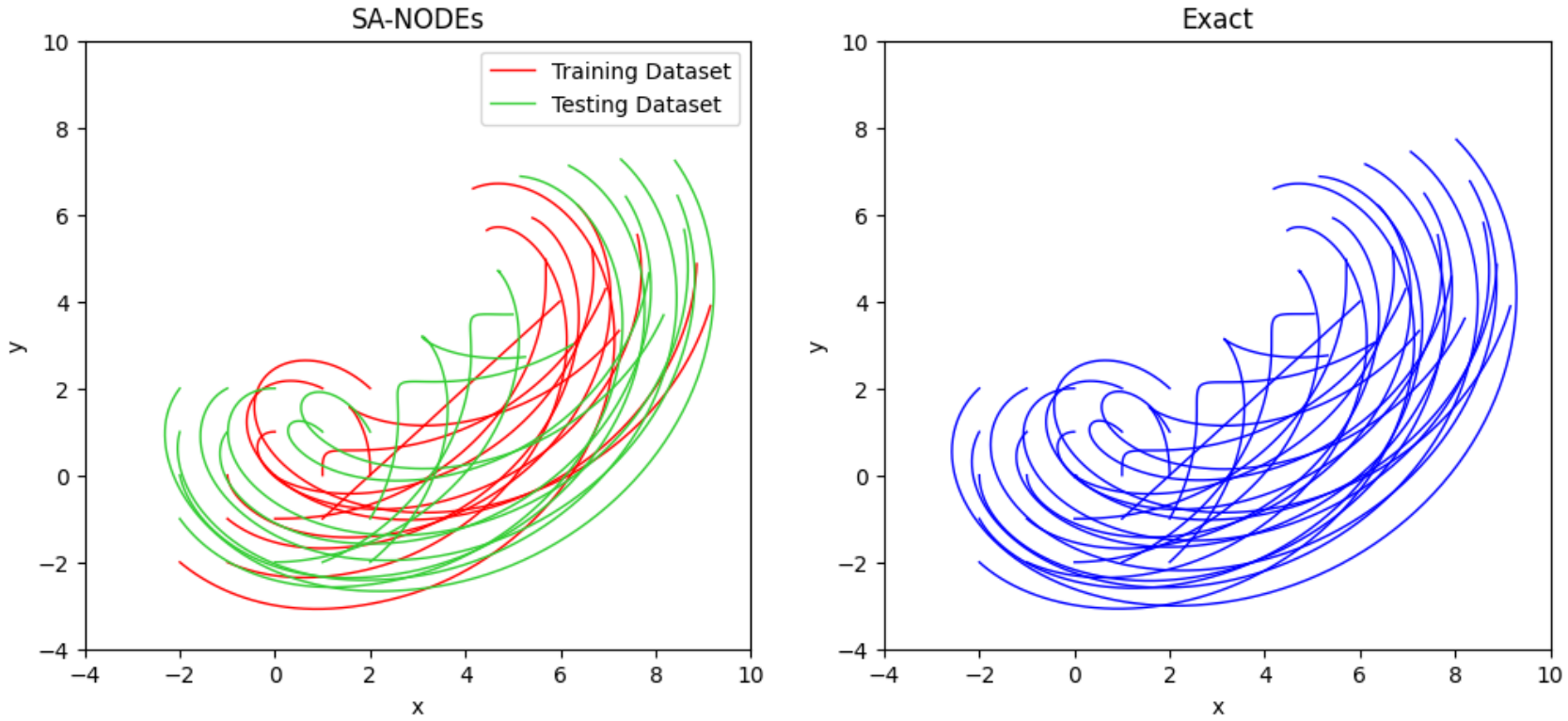
$$\begin{cases} \dot{z}_1 = z_2 \\ \dot{z}_2 = -2z_1 - 3z_2 \end{cases}$$

Numerical Results



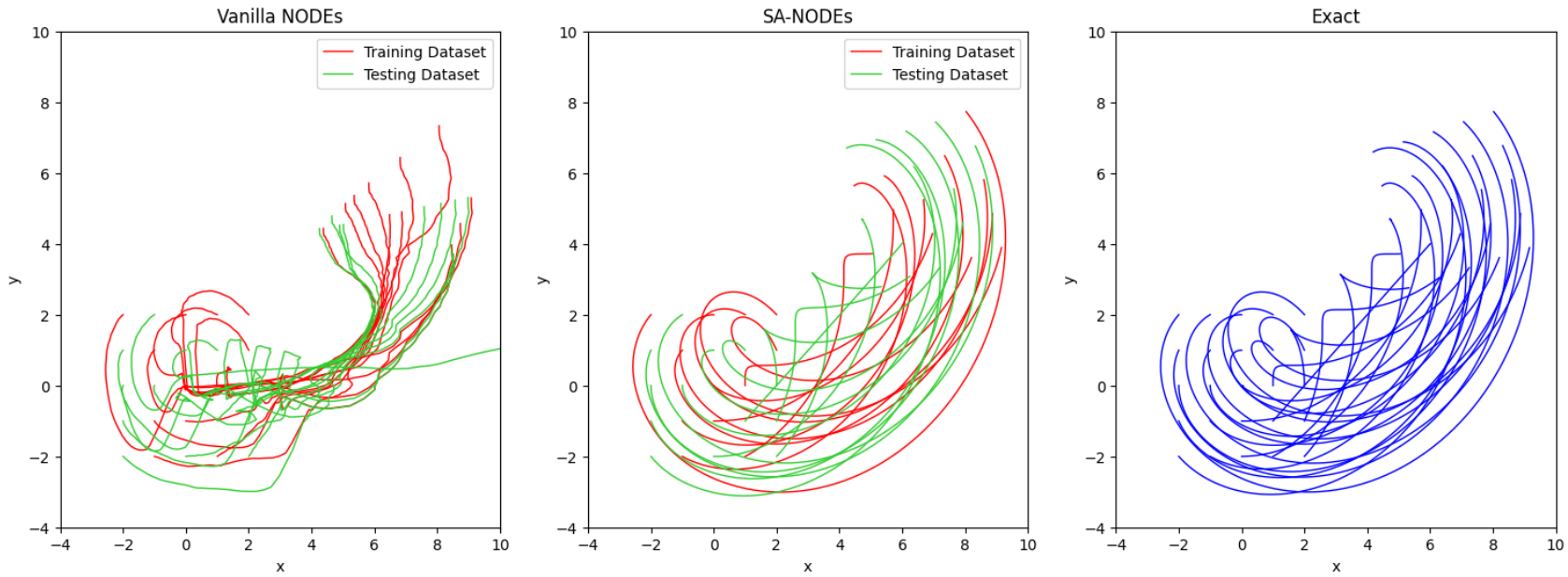
$$\begin{cases} \dot{z}_1 = z_2 \\ \dot{z}_2 = -\sin(z_1) \end{cases}$$

Numerical Results



$$\begin{cases} \dot{z}_1 = t - z_2 \\ \dot{z}_2 = z_1 - t \end{cases}$$

Comparison with vanilla NODEs



Comparison with vanilla NODEs

P	Neural ODEs	Autonomous Case	Non-Autonomous Case	DoF
100	Vanilla NODEs	2.60e-01	3.66e+00	50000
	SA-NODEs	4.65e-02	7.78e-02	1200
500	Vanilla NODEs	9.21e-02	2.54e+00	250000
	SA-NODEs	2.16e-02	7.35e-02	6000
1000	Vanilla NODEs	1.38e-01	2.37e+00	500000
	SA-NODEs	1.58e-02	6.73e-02	12000

Comparison of errors and degrees of freedom (DoF) between vanilla NODEs and SA-NODEs. We tested on the systems

$$\begin{cases} \dot{z}_1 = z_2 \\ \dot{z}_2 = -2z_1 - 3z_2 \end{cases} \quad \text{and} \quad \begin{cases} \dot{z}_1 = t - z_2 \\ \dot{z}_2 = z_1 - t \end{cases}$$

$\rightsquigarrow$ We obtain better accuracy with less parameters

Numerical Results: Transport Equations

Case study: Doswell frontogenesis

Wikipedia: Frontogenesis is a meteorological process of tightening of horizontal temperature gradients to produce fronts. In the end, two types of fronts form: cold fronts and warm fronts

$$\partial_t \rho(x, y, t) + \operatorname{div}\left(\rho(x, y, t)\left(-y g(r), x g(r)\right)\right) = 0$$

where $(x, y, t) \in \mathbb{R}^2 \times [0, T]$ and $g(r) = \frac{1}{r}\, \overline{v}\, \operatorname{sech}^2 r \tanh r$

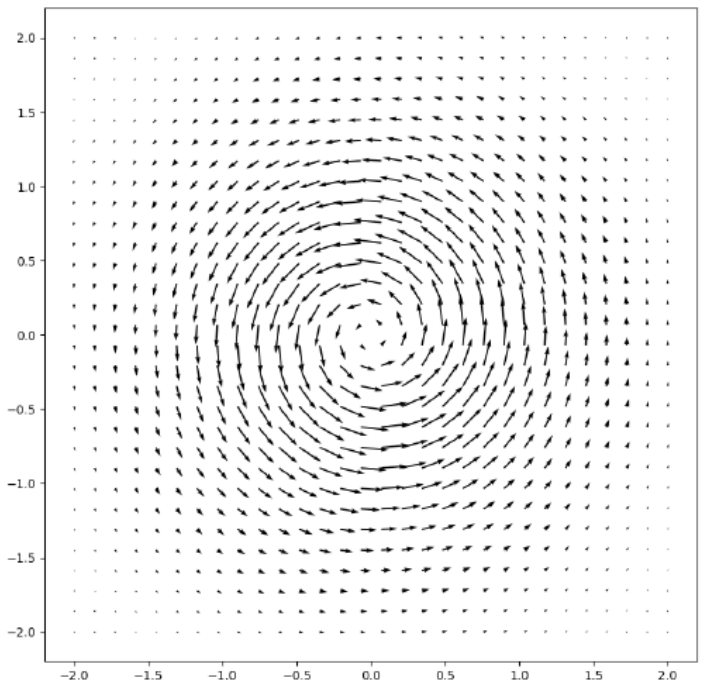
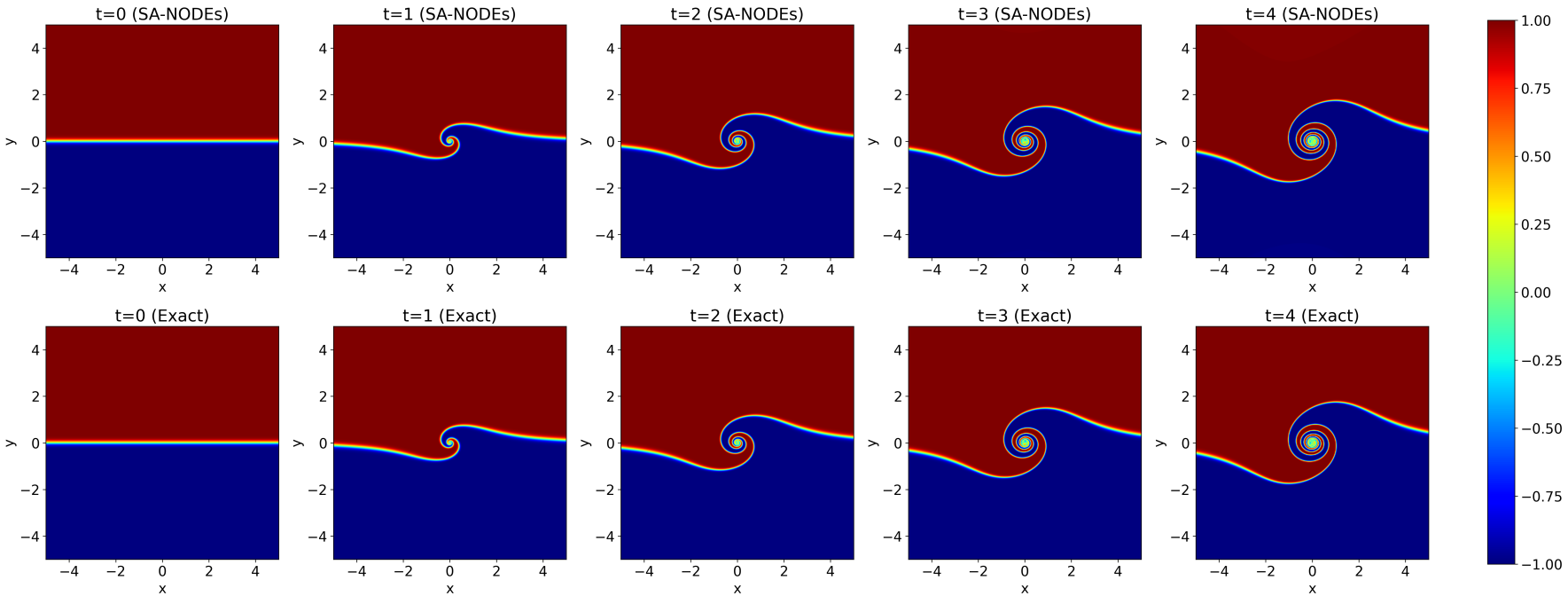


FIGURE 1. The Doswell vector field (11).

Figure: Credits: Manea, Zuazua (2023)

SA-NODEs approximation



Thank you for your attention!

Training SA-NODEs: Theory

Data: $\mathcal{D} = \{z_i(t)\}_{i=1}^N \subset \mathbb{R}^d$, for $t \in [0, T]$

We want to solve the optimal control problem

$$\inf_{\Theta} \int_0^T \int_K \|z_{z_0}(t) - x_{z_0}(t)\|^2 dz_0 dt$$

$$\text{where } \begin{cases} \dot{x}_{z_0} = f_\Theta(x_{z_0}, t), \; x_{z_0}(0) = z_0 \in K \\ \left\| \sum_{i=1}^P |W_i| \circ \|A_i^1\|_{\ell^2} \right\| \leq 2\|f\|_{\mathcal{S}^d_{\mathrm{B}(X)}} \end{cases}$$

We relax the problem and instead minimize

$$L(\Theta) = \int_0^T \int_K \|z_{z_0}(t) - x_{z_0}(t)\|^2 dz_0 dt + \lambda \left\| \sum_{i=1}^P |W_i| \circ \|A_i^1\|_{\ell^2} \right\|$$

Training SA-NODEs: Theory

Theorem

For any $(\Theta, x, t) \in \mathbb{R}^{Pd(d+3)} \times \mathbb{R}^d \times [0, T]$, *let* $g(\Theta) = \left\| \sum_{i=1}^{P} |W_i| \circ \|A_i^1\|_{\ell^2} \right\|$. *It holds that*

$$\nabla L(\Theta) = \int_0^T \int_K \frac{\partial f_\Theta}{\partial \Theta}(\Theta, \boldsymbol{x}_{z_0}(t), t)^\top \boldsymbol{a}_{z_0}(t) dz_0 dt + \lambda \nabla g(\Theta)$$

where $\boldsymbol{a}_{z_0}$ *satisfies the adjoint equation*

$$\begin{cases} -\dot{\boldsymbol{a}}_{z_0}(t) = \frac{\partial f_\Theta}{\partial x}(\Theta, \boldsymbol{x}_{z_0}(t), t)^\top \boldsymbol{a}_{z_0}(t) + 2(\boldsymbol{x}_{z_0}(t) - \boldsymbol{z}_{z_0}(t)) \\ \boldsymbol{a}_{z_0}(T) = 0 \end{cases}$$

for $t \in [0, T]$, $z_0 \in K$.

Training SA-NODEs: Practice

Data: $\mathcal{D} = \{z_k(t_l)\}_{k,l} \subset \mathbb{R}^d$, for $k = 1, \ldots, N$, $l = 1, \ldots, M$

Loss Function:

$$\hat{L}(\Theta) = \frac{1}{NM} \sum_{k=1}^{N} \sum_{l=1}^{M} (z_k(t_l) - x_k(t_l, \Theta))^2 + \lambda \left\| \sum_{i=1}^{P} |W_i| \circ \|A_i^1\|_{\ell^2} \right\|$$

$\rightsquigarrow$ Stochastic gradient descent

Training SA-NODEs: Transport

Transport equation

$$\begin{cases} \partial_t \rho(x,t) + \mathrm{div}_x(\rho(x,t) f(x,t)) = 0 \\ \rho(\cdot,0) = \rho_0 \in \mathcal{M}(\mathbb{R}^d) \end{cases}$$

$$\downarrow$$

Characteristic system

$$\begin{cases} \dfrac{dx}{dt} = f(x,t) \\ \dfrac{d\rho}{dt} = -\mathrm{div}_x(f(x,t))\rho \end{cases}$$

Training SA-NODEs: Transport

Transport equation

$$\begin{cases} \partial_t \rho(x,t) + \mathrm{div}_x(\rho(x,t) f(x,t)) = 0 \\ \rho(\cdot,0) = \rho_0 \in \mathcal{M}(\mathbb{R}^d) \end{cases}$$

$$\downarrow$$

Approximated characteristic system

$$\begin{cases} \dfrac{d\boldsymbol{x}}{dt} = \displaystyle\sum_{i=1}^{P} W_i \circ \boldsymbol{\sigma}(A_i^1 \boldsymbol{x} + A_i^2 t + B_i) \\ \dfrac{d\rho}{dt} = -\mathrm{div}_x \left(\displaystyle\sum_{i=1}^{P} W_i \circ \boldsymbol{\sigma}(A_i^1 \boldsymbol{x} + A_i^2 t + B_i) \right) \rho \end{cases}$$

Training SA-NODEs: Transport

Transport equation

$$\begin{cases} \partial_t \rho(x,t) + \mathrm{div}_x(\rho(x,t) f(x,t)) = 0 \\ \rho(\cdot,0) = \rho_0 \in \mathcal{M}(\mathbb{R}^d) \end{cases}$$

$$\downarrow$$

Approximated characteristic system

$$\begin{cases} \dfrac{d\boldsymbol{x}}{dt} = \displaystyle\sum_{i=1}^{P} W_i \circ \boldsymbol{\sigma}(A_i^1 \boldsymbol{x} + A_i^2 t + B_i) \\ \dfrac{d\rho}{dt} = -\mathrm{div}_x \left(\displaystyle\sum_{i=1}^{P} W_i \circ \boldsymbol{\sigma}(A_i^1 \boldsymbol{x} + A_i^2 t + B_i) \right) \rho \end{cases}$$

We need to learn only the first equation!