Universal Approximation Property of Neural Ordinary Differential Equations

Takeshi Teshima[1 2], Koichi Tojo[2], Masahiro Ikeda[2],
Isao Ishikawa[3 2], Kenta Oono[1],

[1]The University of Tokyo, Japan
[2]RIKEN, Japan
[3]Ehime University, Japan

NEURAL INFORMATION PROCESSING SYSTEMS

High-level summary

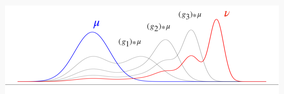
Research question "How expressive are NODEs?"
NODE = Neural Ordinary Differential Equations [CRBD18]

Why important?
- Strong (sup-norm) guarantee for a large class of invertible maps.
- cf. Previous result: Universality for $C^0(\mathbb{R}^n, \mathbb{R}^m)$ w.r.t. L^p-norm. [LLS20]

What is the result? Universality of NODE + Affine transform for a large class of diffeomorphisms w.r.t. sup-norm.

Usages of invertible neural networks

- Modelling distributions (a.k.a. normalizing flows)



[KD18]

- Modelling invertible maps (feature extraction & manipulation).

[DSB17]

Message

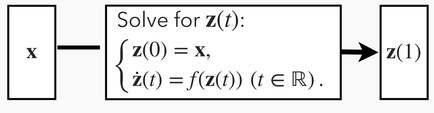
NODE-based invertible neural networks have high representation power for approximating diffeomorphisms. They can be relied on in modeling invertible maps and distributions.

Preliminary: NODEs

NODE layer

$$\mathrm{Lip}(\mathbb{R}^d) := \{f\colon \mathbb{R}^d \to \mathbb{R}^d \mid f \text{ is Lipschitz}\}$$

For each $f \in \mathrm{Lip}(\mathbb{R}^d)$, we define an invertible map $\mathbf{x} \mapsto \mathbf{z}(1)$ via an initial value problem: [DJ76]



Definition (NODE layers)

Then, for $\mathscr{H} \subset \mathrm{Lip}(\mathbb{R}^d)$, consider the set of NODEs:

$$\mathrm{NODEs}(\mathscr{H}) := \{\mathbf{x} \mapsto \mathbf{z}(1) \mid f \in \mathscr{H}\}$$

Definition (Invertible neural networks based on $\mathscr{H}$-NODE)

$$\mathrm{INN}_{\mathscr{H}\text{-NODE}} := \{W \circ \psi_k \circ \cdots \circ \psi_1 \mid \psi_1, \ldots, \psi_k \in \mathrm{NODEs}(\mathscr{H}), W \in \mathrm{Aff}, k \in \mathbb{N}\}$$

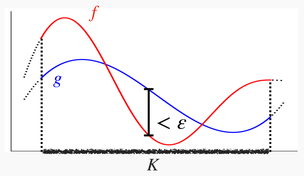
Research question

Restricted functions → restricted representation power?

Preliminary: Universality

Definition (informal; Sec. 2.2.) [C89]

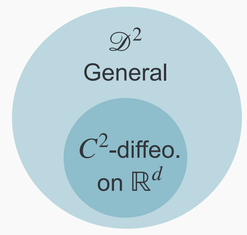
sup-**universal approximator**: the model can approximate any target function w.r.t. sup-norm on a compact set.



Approximation target

Definition (approximation target $\mathscr{D}^2$; Sec. 3.1.)

$\mathscr{D}^2 := \{f : U_f \to f(U_f) \mid f : C^2\text{-diffeo}^1,\ U_f \subset \mathbb{R}^d \text{: open} \cong \mathbb{R}^d\}$
(C^2-diffeo.)



$\mathscr{D}^2$ is **fairly large**

Result: Universality of NODEs

Theorem (Sec. 3.2, Theorems 2, 3)

Assume $\mathscr{H}$ is a sup-universal approximator for $\mathrm{Lip}(\mathbb{R}^d)$.

$\mathrm{INN}_{\mathscr{H}\text{- ACF}}$ is a sup-universal approximator for $\mathscr{D}^2$.

Examples for $\mathscr{H}$: multi-layer perceptron [LBH15], Lipschitz Networks [ALG19].

Implication

High representation power of NODEs.

Notation, terminology, and abbreviations

1. Diffeomorphism (diffeo) = smooth function with a smooth inverse.
2. Compactly supported (invertible map) = identity map outside some compact set.
3. $C^0(\mathbb{R}^n, \mathbb{R}^m)$ = Continuous maps from $\mathbb{R}^n$ to $\mathbb{R}^m$.

Detail 1: Proof outline

$f \in \mathscr{D}^2$: target, $K \subset U_f$: compact ⇝ **Decompose** $f|_K$ into **simpler** maps

$$f|_K$$
$$= \exists W \circ h \ (\mathrm{Aff} \ \& \ \mathrm{Diff}_c^2)$$
$$\| \ll \textbf{structure theorem of diffeomorphism group}$$
$$\mathrm{IVP}[\exists f_1] \circ \mathrm{IVP}[f_2] \circ \cdots \circ \mathrm{IVP}[f_i] \circ \cdots \ (\textbf{flow endpoints})$$
$$\wr\!\wr \ \ll \mathscr{H} \text{ is sup-universal } (f_i \approx g_i \Rightarrow \mathrm{IVP}[f_i] \approx \mathrm{IVP}[g_i])$$
$$\mathrm{IVP}[\exists g_i] \ (g_i \in \mathscr{H})$$

Detail 2

Detail 2: "Structure theorem"

Our analysis is via Diff_c^r, which is a broad class but can be conveniently represented by elements of well-understood behavior (flow endpoints).

Definition (compactly supported C^r-diffeomorphisms) $1 \le r \le \infty$

Diff_c^r: compactly-supported C^r-diffeomorphism on $\mathbb{R}^d$

Definition (Flow endpoints S^r)

$h \in \mathrm{Diff}_c^r$ **is a flow endpoint if** $\exists \Phi : \mathbb{R}^d \times U \to \mathbb{R}^d$ **such that** $h = \Phi(\cdot, 1)$ and

(1) $\Phi(\cdot, 0) = \mathrm{Id}$ **(2)** $U \subset \mathbb{R}$ is open and $[0,1] \subset U$ **(3)** $\Phi(\,\cdot\,, t) \in \mathrm{Diff}_c^r$ $(\forall t \in U)$

(4) $\Phi(x, s+t) = \Phi(\Phi(x, s), t)$ **(5)** $\Phi \in C^r(\mathbb{R}^d \times U)$ **(6)** $\exists K_\Phi \subset \mathbb{R}^d$ s.t. $\bigcup_{t \in U} \mathrm{supp}\Phi(\,\cdot\,, t) \subset K_\Phi$.

Theorem (Herman, Thurston, Epstein, and Mather) [TIT+20] $r \neq d+1$

Diff_c^r is a **simple group** (= its normal subgroup is either $\{\mathrm{Id}\}$ or itself).

Proposition $H^r := \{h_1 \circ \cdots \circ h_m \mid h_1, \ldots, h_m \in S^r\}$.

H^r is a **nontrivial normal subgroup** of Diff_c^r.

Conclusion $H^r = \mathrm{Diff}_c^r$.

References

[TIT+20] Teshima, T., Ishikawa. I., Tojo. K., Ohno. K., Ikeda. M., Sugiyama. M, (2020). "Coupling-based invertible neural networks are universal diffeomorphism Approximators." ArXiv:2006.11469 [Cs.LG].

[KD18] Kingma, D. P., & Dhariwal, P. (2018). "Glow: Generative flow with invertible 1x1 convolutions." In Advances in Neural Information Processing Systems 31, 10215–10224.

[DSB17] Dinh, L., Sohl-Dickstein, J., & Bengio, S. (2017). "Density estimation using Real NVP." Fifth International Conference on Learning Representations (ICLR)

[LLS20] Q. Li, T. Lin, and Z. Shen. (2020). Deep learning via dynamical systems: an approximation perspective. arXiv:1912.10382 [cs, math, stat].

[LBH15] Y. LeCun, Y. Bengio, and G. Hinton. (2015). "Deep learning." Nature, 521(7553), 436–444.

[ALG19] C. Anil, J. Lucas, and R. Grosse. (2019). "Sorting out Lipschitz function approximation." Proceedings of the 36th International Conference on Machine Learning, PMLR 97, 291–301.

[DJ76] W. Derrick and L. Janos. (1976). "A global existence and uniqueness theorem for ordinary differential equations." Canadian Mathematical Bulletin, 19(1), 105–107.

[CRBD18] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. (2018). "Neural ordinary differential equations." Advances in Neural Information Processing Systems 31, 6571–6583.

[C89] Cybenko, G. (1989). "Approximation by superpositions of a sigmoidal function." Mathematics of Control, Signals, and Systems, 2, 303–314.