Coupling-based Invertible Neural Networks Are Universal Diffeomorphism Approximators

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

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

NEURAL INFORMATION PROCESSING SYSTEMS
arXiv:2006.11469

High-level summary

What did we do? Theoretically investigated: "Are popular INNs expressive enough?"

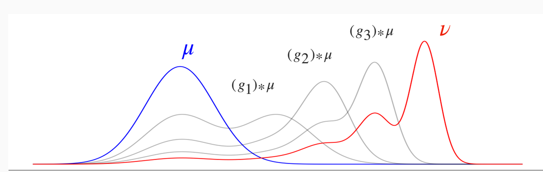
What are INNs? INNs (= Invertible neural networks) are special neural networks with invertibility by design. Enables various applications.

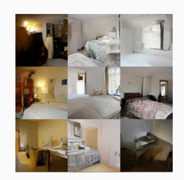
Why important? Models without a representation power guarantee are hard to rely on.

What is the result? "Coupling-based INNs (CF-INNs)" are "universal function approximators" despite their restricted architectures.

Usages of INNs

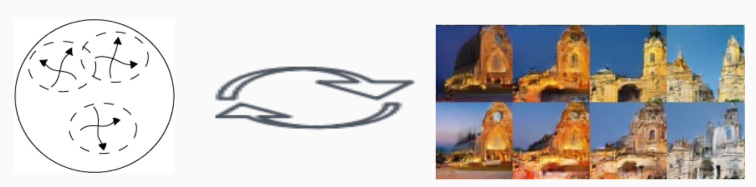
- Modelling distributions (a.k.a. normalizing flows)





[KD18]

- Modelling invertible maps (feature extraction & manipulation).



[DSB17]

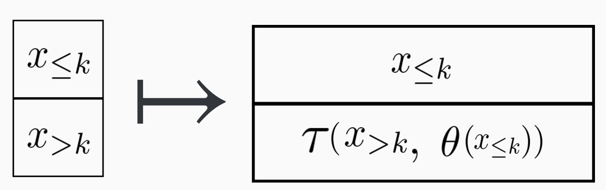
Message

"CF-INNs" can model various invertible functions and probability distributions. They can be relied on in many applications.

Preliminary: CF-INN

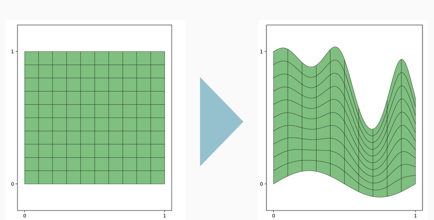
Coupling flows

Idea: Keep some dimensions unchanged.



Why CF-INN (=Coupling-flow based INN)?

- Popular in application (see next section)
- Lack of theory about its representative power



Definition (Invertible neural networks; Sec. 2.1 Definition 1)

$\mathcal{G}$: a set of **bijections** on $\mathbb{R}^d$.

The set of **invertible neural networks based on** $\mathcal{G}$ is

$$\mathrm{INN}_{\mathcal{G}} := \{W_1 \circ g_1 \circ \ldots \circ W_r \circ g_r \mid r > 0,\ g_i \in \mathcal{G},\ W_i \in \mathrm{Aff}\}.$$

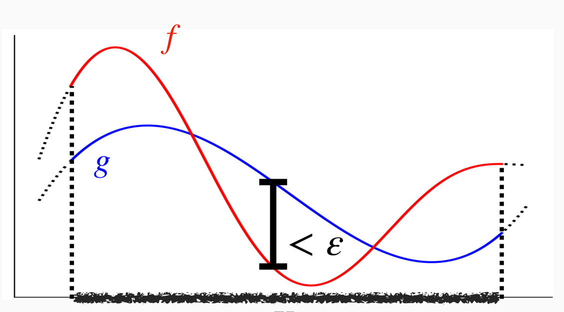
Research question

Restricted function form → restricted representation power?

Preliminary: Universality

Definition (informal, Sec. 2.2.)

sup- (L^p-) universal approximator: the model can approximate any target function w.r.t. sup- (L^p-) norm on a compact set.



Definition (informal, Sec 2.2) [HKLC18,JSY19]

A model is a **distributional universal approximator** if it can transform one distribution arbitrarily close to any distribution w.r.t weak convergence.

Overview of main results

To answer the research question, we show two types of representation power: (1) sup/Lp-universality for $\mathcal{D}^2$. (2) Distributional universality.

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

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

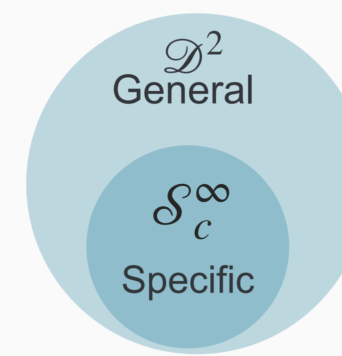
$\mathcal{D}^2$ is **fairly large**: $\{C^2\text{-diffeo. on } \mathbb{R}^d\} \subset \mathcal{D}^2$

Result 1: General approximation theory for $\mathcal{D}^2$

Theorem (Sec. 3.1. Theorem 1)

sup- (L^p-) universal approximator **for** $\mathcal{S}_c^\infty$

$\Longrightarrow$ sup- (L^p-) universal approximator **for** $\mathcal{D}^2$.



Definition (special subset of $\mathcal{D}^2$: single-coordinate transformations)

$\mathcal{S}_c^\infty$: **compactly supported**[2] C^∞-diffeo on $\mathbb{R}^d$ of the form:

$$\tau(\mathbf{x}, y) = (\mathbf{x},\ u(\mathbf{x}, y)) \qquad (\mathbf{x}, y) \in \mathbb{R}^{d-1} \times \mathbb{R}$$

Application

- Usage: (1) Show universality for $\mathcal{S}_c^\infty$ using its rich properties. (2) Upgrade to $\mathcal{D}^2$.
- Example: sum-of-squares polynomial flow (SoS[JSY]) and deep sigmoidal flow (DSF [HKLC18]) yield sup-universal CF-INNs.

Result 2: Result for affine coupling flows (ACFs)

Definition ($\mathcal{H}$**-single-coordinate** ACFs; Sec. 2.1) [DSB17,KD18]

$\mathcal{H}$: a set of real-valued functions on $\mathbb{R}^{d-1}$, e.g., MLP[4], RKHS[5]

$$\mathcal{H}\text{-ACF} := \left\{\Psi_{s,t}(\mathbf{x}, y) =^2 \left(\mathbf{x}, e^{s(\mathbf{x})}y + t(\mathbf{x})\right) \;\middle|\; s, t \in \mathcal{H}\right\}.$$

Theorem (Sec. 3.2, Theorems 2, 3)

Assume $\mathcal{H}$ is a sup-universal approximator for $C_c^\infty(\mathbb{R}^{d-1})$.[3]

$\mathrm{INN}_{\mathcal{H}\text{-ACF}}$ is

(1) an L^p-universal approximator for $\mathcal{D}^2$.

(2) a distributional universal approximator

Implication

- Useful criterion: "if my CF architecture contains ACFs (as special cases), then they are also (L^p-/dist.) universal."
- Affirmative answer to an unsolved conjecture.[PNRML19]

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. $C_c^\infty(\mathbb{R}^{d-1})$ = smooth functions on $\mathbb{R}^{d-1}$ with a compact support.
4. MLP = Multi-Layer Perceptron
5. RKHS = Reproducing Kernel Hilbert Space

Detail 1: Proof outline of Result 1

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

$f|_K$

$=$

$\exists W \circ h$ (Aff & Diff_c^2)

$\|$ « **structure theorem of diffeomorphism group**

$\exists h_1 \circ h_2 \circ \cdots$ (**flow endpoints**) ← Detail 2

$=$

$\exists g_1 \circ g_2 \circ \cdots$ (nearly Ids)

$=$

$\tau_1 \circ \sigma_1 \circ \cdots$ (permutations & $\mathcal{S}_c^\infty$)

Detail 2: "Structure theorem"

Our analysis is via Diff_c^2, which is a broad class but can be conveniently represented by elements of well-understood behavior (flow endpoints).

Definition (compactly supported C^2-diffeomorphisms)

Diff_c^2: compactly-supported C^2-diffeomorphism on $\mathbb{R}^d$

Definition (Flow endpoints)

$h \in \mathrm{Diff}_c^2$ is a **flow endpoint if** there exists a **continuous** and **"additive"** map $\phi : [0,1] \to \mathrm{Diff}_c^2$ such that $\phi(0) = \mathrm{Id}$ and $\phi(1) = h$.

Theorem (Herman, Thurston, Epstein, and Mather)

Diff_c^2 is a **simple group** (= its normal subgroup is either $\{\mathrm{Id}\}$ or itself).

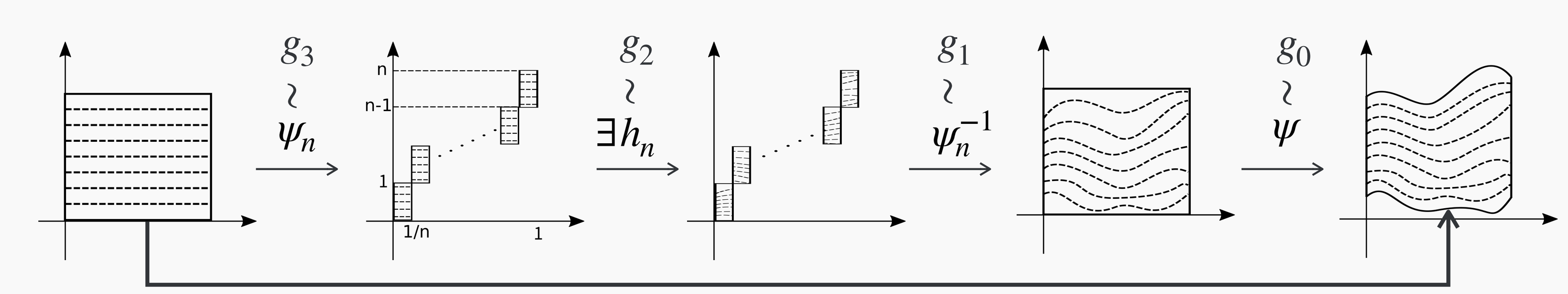
Proposition $S := \{h_1 \circ \cdots \circ h_m \mid h_1, \ldots, h_m : \text{flow endpoints}\}$.

S is a **nontrivial normal subgroup** of Diff_c^2.

Conclusion $S = \mathrm{Diff}_c^2$.

Detail 3: Proof outline of Result 2 ($d = 2$ case)

$\exists g_0, \ldots, g_3 \in \mathcal{H}\text{-ACF}$ such that $f \sim \psi \circ \psi_n^{-1} \circ h_n \circ \psi_n \sim g_0 \circ g_1 \circ g_2 \circ g_3$



$f \in \mathcal{S}_c^\infty$ $\quad$ $(\, t_n := \sum_{k=0}^{n-1} k\mathbf{1}_{[k/n,(k+1)/n)},\ \psi_n := \Psi_{1,t_n}\,)$

References

[DKB14] Dinh, L., Krueger, D., & Bengio, Y. (2014). "NICE: Non-linear independent components estimation."ArXiv:1410.8516 [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)

[HKLC18] Huang, C.-W., Krueger, D., Lacoste, A., & Courville, A. (2018). "Neural autoregressive flows." Proceedings of the 35th International Conference on Machine Learning, 80, 2078–2087.

[JSY19] Jaini, P., Selby, K. A., & Yu, Y. (2019). "Sum-of-squares polynomial flow." Proceedings of the 36th International Conference on Machine Learning, 97, 3009–3018.

[PNRML19] Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., & Lakshminarayanan, B. (2019)."Normalizing flows for probabilistic modeling and inference." ArXiv:1912.02762 [Cs, Stat].