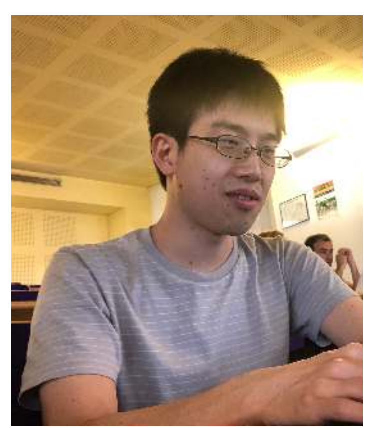
On the Universality of Invertible Neural Networks

Takeshi Teshima[1 2], Isao Ishikawa[3 2]

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






Joint work with Koichi Tojo, Kenta Oono, Masahiro Ikeda, and Masashi Sugiyama.

Today's talk structure

Part 1

Introduction.

Overview of what we did and why it's important.

Part 2

Details of the theory.

Theoretical preliminaries and proof machinery.

Self-introduction

Takeshi Teshima (https://takeshi-teshima.github.io)

Ph.D. candidate @ UTokyo
(advisor: Prof. Sugiyama)

GRADUATE SCHOOL OF FRONTIER SCIENCES
THE UNIVERSITY OF TOKYO

Supported by:
RIKEN JRA Program
and Masason Foundation.

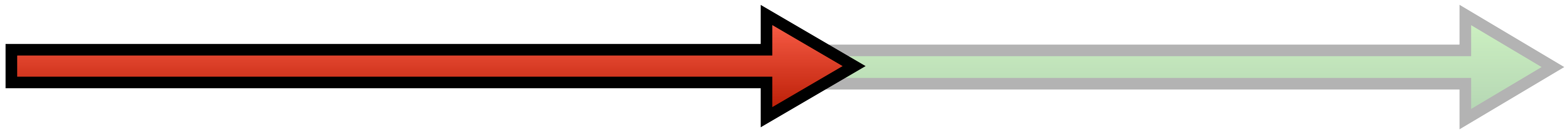
Programs for Junior Scientists RIKEN

Recent Research Interests:

General methodology of machine learning.
In particular: "Causality for machine learning"

- Causal mechanism transfer ← I used INNs in this work
- Causal data augmentation

Today's talk structure



Part 1

Introduction.

Overview of what we did and why it's important.

Part 2

Details of the theory.

Theoretical preliminaries and proof machinery.

Invertible Neural Networks (INNs)

Goal

Understand theoretical props of **invertible neural networks (INNs)**.

Invertible Neural Networks (INNs) generated by $\mathscr{G}$

Compositions of **flow maps/layers** $\mathscr{G}$ and **affine transforms** Aff.

$$f = g_1 \circ W_1 \circ \cdots \circ g_k \circ W_k \quad (g_i \in \mathscr{G}, W_i \in \mathrm{Aff})$$

$\mathscr{G}$ is parametrized ("trainable") but **designed to be invertible**.

($\mathscr{G}$ is often rather simple → Composed to model complex f)

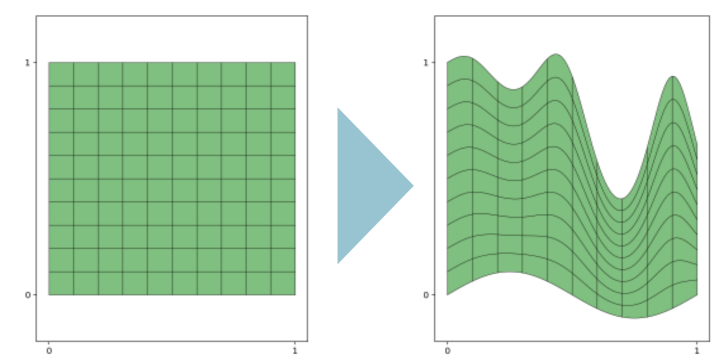
Example (Designs of flow layers $\mathscr{G}$)

- Coupling-based flow layers [DKB14, PNRML19, KPB19]
- Neural ordinary differential equations [CRBD18]

Example1: Coupling Flows

Coupling flows (CFs) [DKB14, PNRML19, KPB19]

$$\begin{bmatrix} x_{\le k} \\ x_{>k} \end{bmatrix} \mapsto \begin{bmatrix} x_{\le k} \\ \mathcal{T}(x_{>k},\ \theta(x_{\le k})) \end{bmatrix}$$



Idea: Keep some dimensions unchanged. (Strong constraint!)

CF-INN = Coupling-flow based INN.

Affine-coupling flows (ACFs) [DKB14,DSB17,KD18]

One of the simplest CFs using **coordinate-wise affine transformation**:

$$\Psi_{s,t,k} : \begin{bmatrix} x_{\le k} \\ x_{>k} \end{bmatrix} \mapsto \begin{bmatrix} x_{\le k} \\ x_{>k} \odot \exp(s(x_{\le k})) + t(x_{\le k}) \end{bmatrix}$$

Example 2: Neural Ordinary Differential Equations

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]

$\mathbf{x}$ — Solve for $\mathbf{z}(t)$:

$$\begin{cases} \mathbf{z}(0) = \mathbf{x}, \\ \dot{\mathbf{z}}(t) = f(\mathbf{z}(t)) \ (t \in \mathbb{R}) \,. \end{cases}$$

→ $\mathbf{z}(1)$

NODE layers [CRBD18]

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

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

Applications of INNs

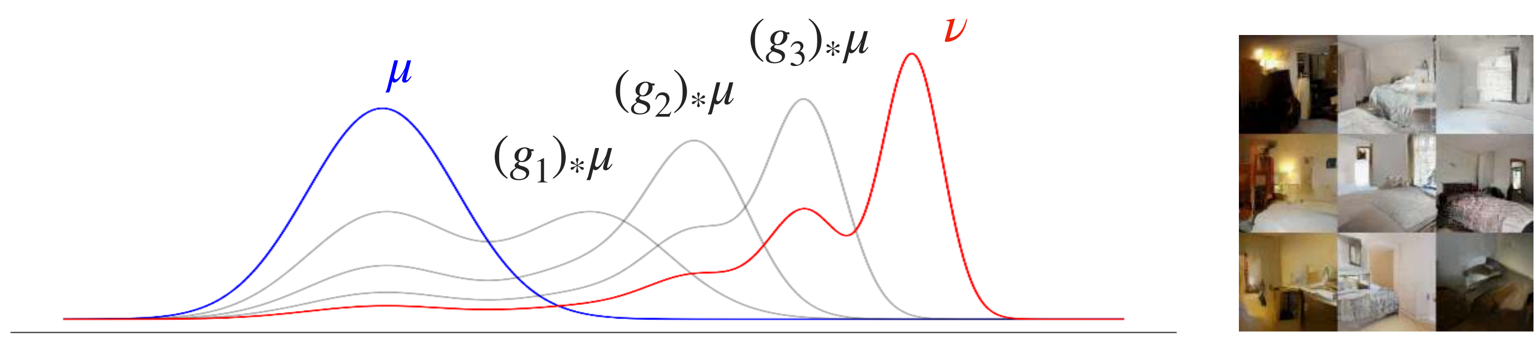
Useful properties of INNs (for nicely designed $\mathscr{G}$)

✓ **Explicit and efficient invertibility**.

✓ **Tractability** of Jacobian determinant (for nicely designed $\mathscr{G}$).

Usages of INNs

- Approximate distributions (normalizing flows).

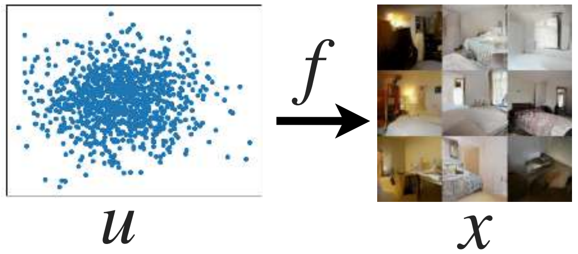


[KD18]

- Approximate invertible maps (feature extraction & manipulation).

[DSB17]

Application 1: Distribution Modeling



[KD18]

Normalizing Flows

Express x as a transformation f of a real vector u sampled from p_u:

$$x = f(u) \text{ where } u \sim p_u$$

Examples

- Generative modeling [DSB17,KD18,OLB+18,KLSKY19,ZMWN19]
- Probabilistic inference [BM19,WSB19,LW17,AKRK19]
- Semi-supervised learning [IKFW20]

Training by Maximum Likelihood (Invertibility+Tractable Jacobian!)

By change of variables formula:

↓ easily invertible

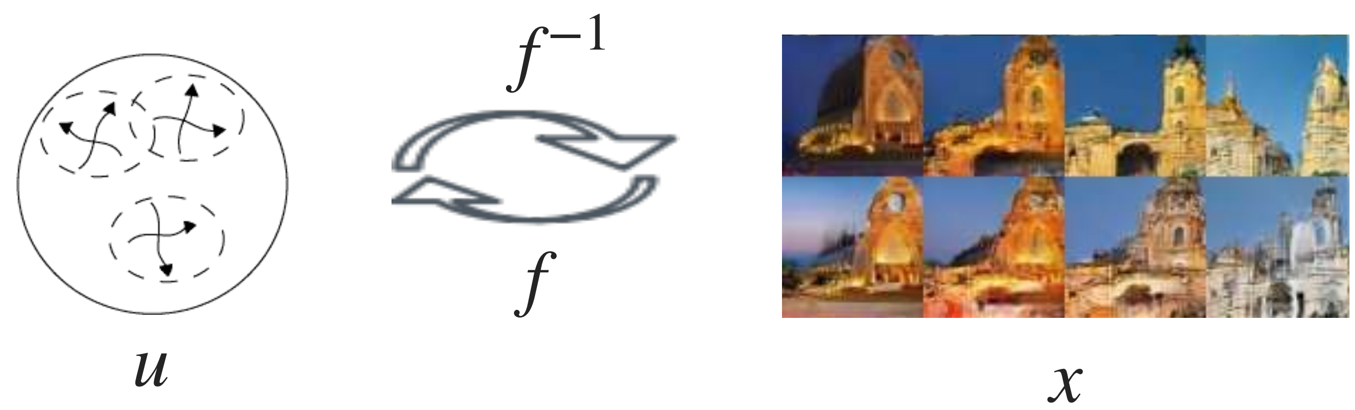
$$\log p_x(x) = \log p_u(f^{-1}(x)) + \log \left| \det J_{f^{-1}}(x) \right|$$

($J_{f^{-1}}$: Jacobian of f^{-1})

↑ known ↑ tractable

Application 2: Function Modeling

Feature Extraction & Manipulation



[DSB17]

1. Extract latent representation u from x by f.
2. Modify u in the latent space (e.g., interpolation).
3. Map back to the data space by f^{-1}.

Examples

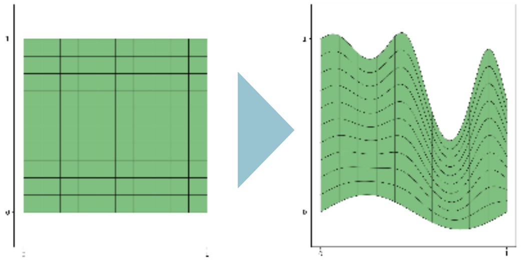
- Generative modeling [DSB17,KD18,OLB+18,KLSKY19,ZMWN19]
- Semi-supervised learning [IKFW20]
- Transfer learning [TSS20]

How flexible are INNs?

INN f is used for **distribution modeling** (application 1) and **invertible function modeling** (application 2).

BUT...

$\mathcal{G}$ relies on special designs to maintain good properties.
(e.g., CF layers keep some dimensions unchanged)



Complications

- The layers have clever specific designs (e.g., ACFs).
- Function composition is the only way to build complex models. (Operations such as addition or multiplications are not allowed.)

Research question

Can these INNs have sufficient representation power?

(Restricted function form → restricted representation power?)

This talk is based on the following papers

Paper 1: Coupling-based invertible neural networks are universal diffeomorphism approximators (NeurIPS 2020) [TIT+20] NEURAL INFORMATION PROCESSING SYSTEMS **Oral paper!**

- Proposed **a general theoretical framework** to analyze the representation power (universalities) of invertible models.
- Analyzed **CF-INNs** (**ACFs** and more advanced ones).

Paper 2: Universal Approximation Property of Neural Ordinary Differential Equations (NeurIPS 2020 DiffGeo4DL Workshop) [TTI+20] NEURAL INFORMATION PROCESSING SYSTEMS

- Analyzed **NODEs**, building on the general framework.
- (with minor modification to the general framework)

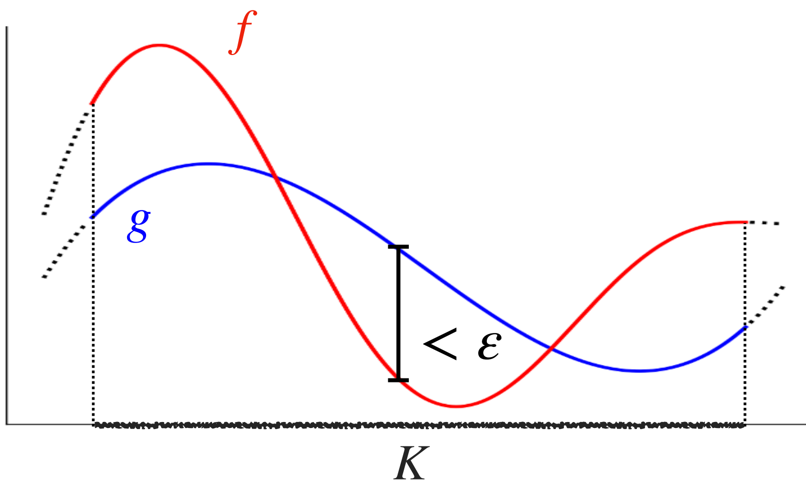
What is "representation power"?

Here,

"Representation power" = Universal approximation property.

Definition (informal) [C89,HSW89]

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

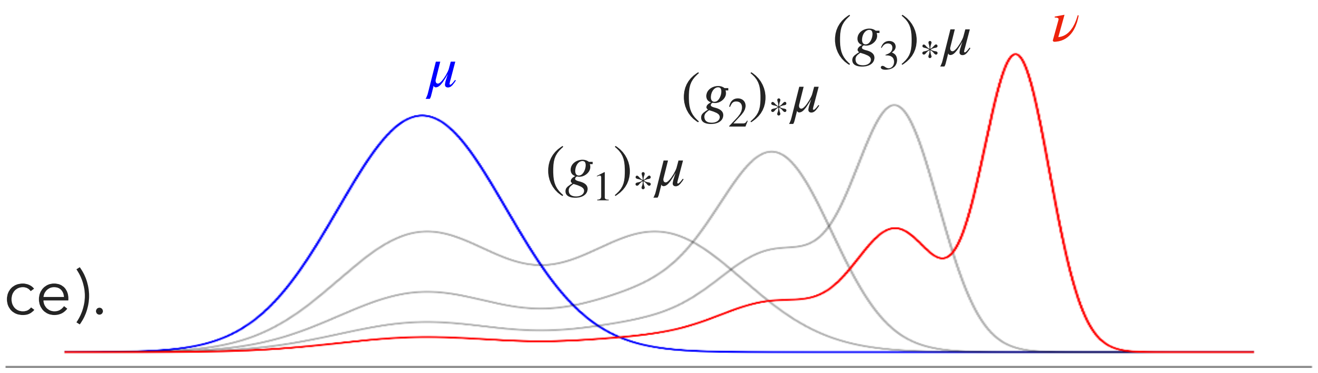


Definition (informal)

A model is a **distributional universal approximator** if it can transform one distribution arbitrarily close to any distribution.

$$(g_n)_*\mu \xrightarrow[n\to\infty]{} \nu$$

(**weak** convergence).



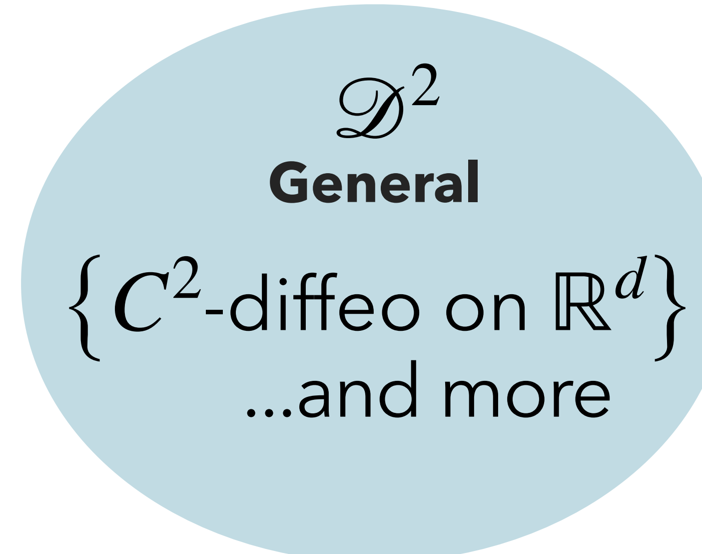
Target Class and General Framework

Definition (Approximation target $\mathscr{D}^2$)

Fairly **large set** of smooth invertible maps.

$$\mathscr{D}^2 := \{C^2\text{-diffeo of the form } f : U_f \to f(U_f)\}$$

$(U_f \subset \mathbb{R}^d$: open C^2-diffeo to $\mathbb{R}^d)$

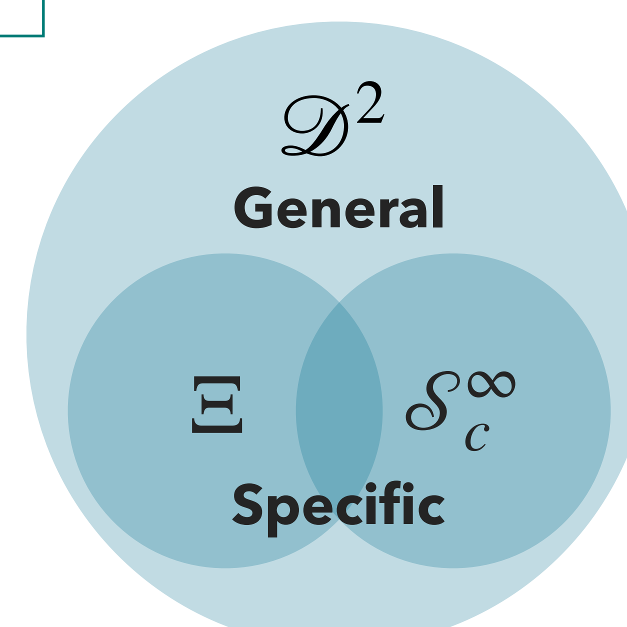


Paper 1 Result (Theoretical Framework) (under mild regularity conditions)

[L^p-/sup-univ. for $\mathscr{D}^2$] $\Leftrightarrow$ [L^p-/sup-univ. for Ξ] $\Leftrightarrow$ [L^p-/sup-univ. for $\mathscr{S}_c^\infty$] $\Rightarrow$ [Distributional-univ.]

Ξ : **"flow endpoints"**

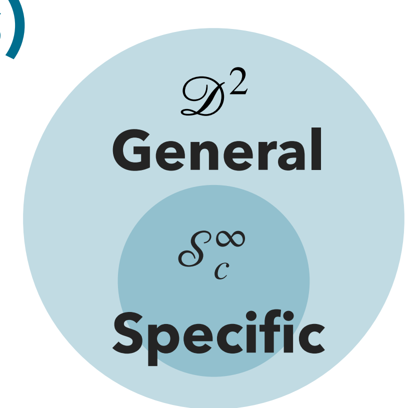
Application of a structure theorem in differential geometry



Preview of Results

L^p-/sup-univ. for $\mathscr{D}^2$ ⇔ L^p-/sup-univ. for Ξ ⇔ L^p-/sup-univ. for $\mathscr{S}_c^\infty$ ⇒ Distributional-univ.

Paper 1 Result (Examples of Universal Coupling Flows)



- **Sum-of-squares polynomial flow** (SoS-flow) [JSY19]
- **Deep sigmoidal flow** (DSF; aka. NAF) [HKLC18]

yield sup-**univ. INNs for $\mathscr{S}_c^\infty$ (and hence for $\mathscr{D}^2$, and also Dist-univ.)**.

(stronger than in [JSY19, HKLC18]).

Paper 1 Result (Affine Coupling Flows yield universal INNs)

Affine Coupling Flows yield L^p-univ. INNs for $\mathscr{S}_c^\infty$
(and hence for $\mathscr{D}^2$, and also Dist-univ.).

Preview of Results

L^p-/$\sup$-univ. for $\mathscr{D}^2$	$\Leftrightarrow$	L^p-/$\sup$-univ. for Ξ	$\Leftrightarrow$	L^p-/$\sup$-univ. for $\mathcal{S}_c^\infty$	$\Rightarrow$	Distributional-univ.

Paper 2 Result (Analysis of NODEs)

NODEs yield $\sup$-univ. INNs for Ξ
(and hence $\sup$-univ. for $\mathscr{D}^2$. Also Dist-univ.).

Overview and Recap

What did we do?

Theoretically investigated:
Are our INNs expressive enough?

INNs = Invertible neural networks

Why important?

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

What is the result?

"Coupling-based INNs (CF-INNs)" and "NODE-based INNs (NODE-INNs)" are "universal function approximators" despite their special architectures.

Message

CF-INNs and NODE-INNs can be relied on in modeling invertible functions and probability distributions.

Today's talk structure

Part 1

Introduction.

Overview of what we did and why it's important.

Part 2

Details of the theory.

Theoretical preliminaries and proof machinery.

Self-introduction

Isao Ishikawa

Assistant professor
@Center for Data Science, Ehime University

CDSE
Center for Data Science, Ehime University

EHIME UNIVERSITY

Supported by CREST JPMJCR1913

Recent Research Interests:

Mathematical analysis of theoretical backgrounds of machine learning and data analysis

- Analysis of representation power of neural networks
- Data analysis via Koopman operator

Contents of Part 2

1. Idea of proof
2. Notion of universalities
3. Machinery for proof
   i) Compatibility of approximation and composition
   ii) Structure theorem of diffeomorphism group
4. Proof outline of universality of NODE
5. Proof of results in paper 1

Idea of Proof

Difficulty

- We cannot use **techniques of functional analysis**!
  - INNs and $\mathscr{D}^2$ are **not** linear spaces

    Recall : $\mathscr{D}^2 := \left\{ C^2\text{-diffeo of the form } f : U_f \to f(U_f) \right\}$

    ($U_f \subset \mathbb{R}^d$: open C^2-diffeo to $\mathbb{R}^d$)
  - Existing methods do not work….(e.g. Hahn-Banach theorem, Fourier transform, Stone-Weirestrass theorem, e.t.c)

Idea

- Utilize a concrete structure of the **diffeomorphism group** !

L^p-Universal approximators

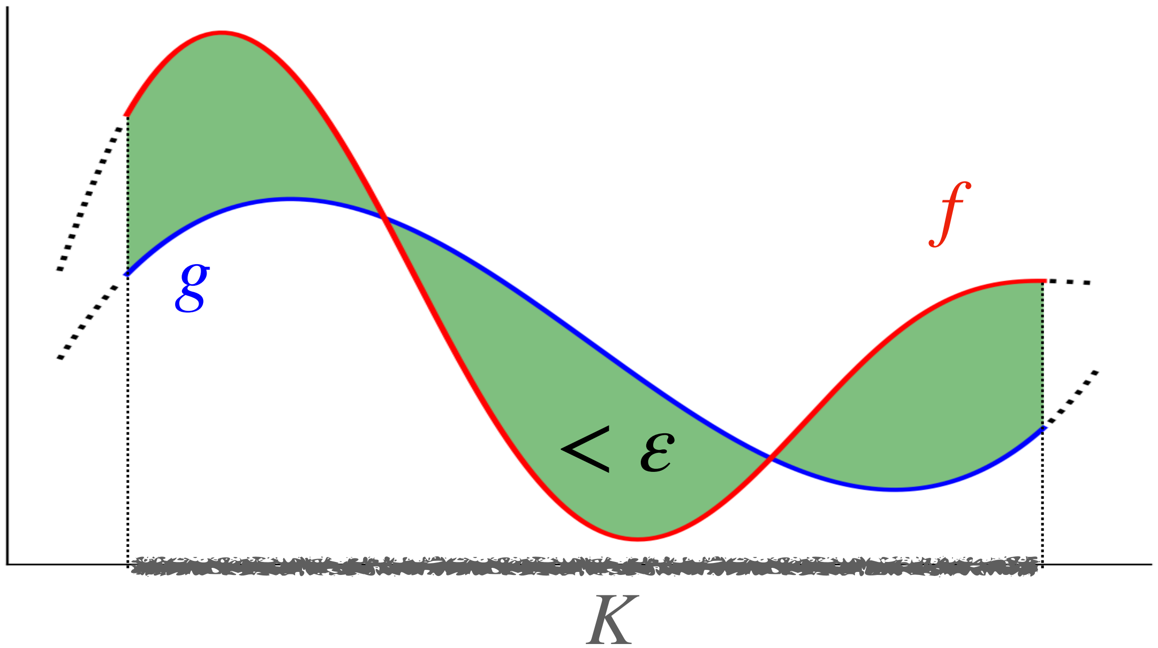
$\mathcal{M}$: model, set of measurable bijection from $\mathbb{R}^d$ to $\mathbb{R}^d$ (e.g. INNs)

$\mathcal{F}$: target functions $f: U_f \to f(U_f)$ (e.g. $\mathcal{D}^2$)

$\mathcal{M}$ is an **L^p-universal approximator** for $\mathcal{F}$ **if**

$\forall f \in \mathcal{F}, \forall \varepsilon > 0, \forall K \subset U_f$: compact , $\exists g \in \mathcal{M}$

$$\int_K |f(x) - g(x)|^p dx < \varepsilon$$



sup-Universal approximators

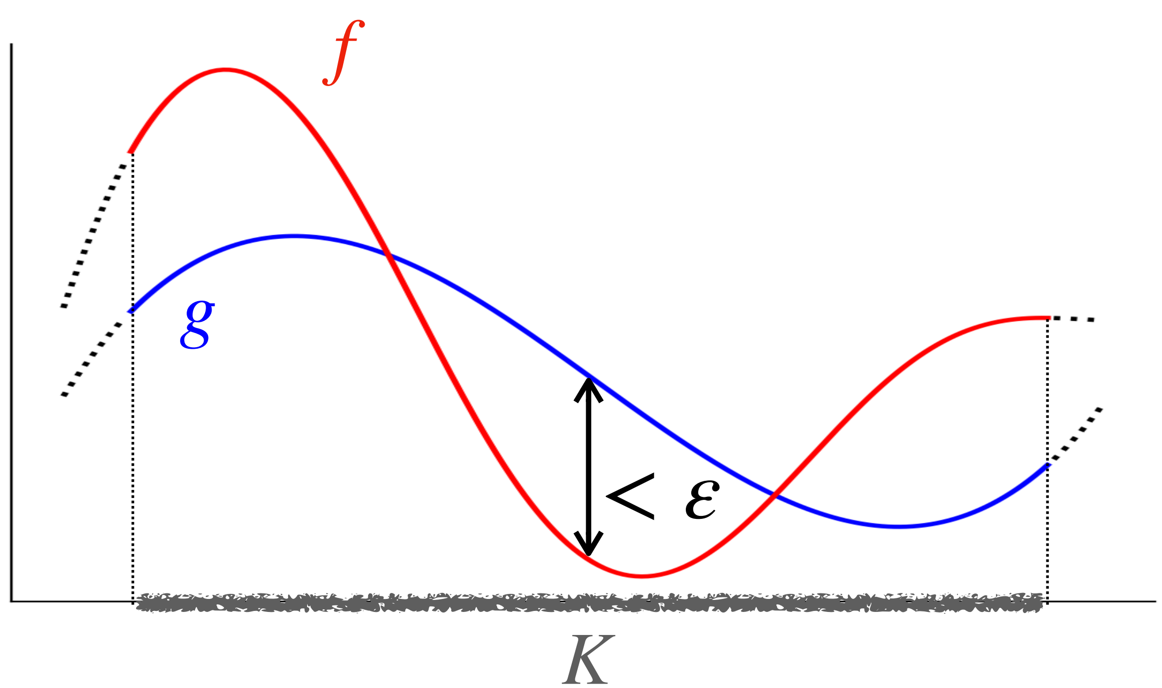
$\mathscr{M}$: model, set of measurable bijection from $\mathbb{R}^d$ to $\mathbb{R}^d$ (e.g. INNs)

$\mathscr{F}$: target functions $f: U_f \to f(U_f)$ (e.g. $\mathscr{D}^2$)

$\mathscr{M}$ is an **sup-universal approximator** for $\mathscr{F}$ **if**

$\forall f \in \mathscr{F}$, $\forall \varepsilon > 0$, $\forall K \subset U_f$: compact , $\exists g \in \mathscr{M}$

$$\sup_{x \in K} |f(x) - g(x)| < \varepsilon$$



Relation of universalities

Proposition

A model $\mathscr{M}$ is a sup-universal approximator for a target $\mathscr{F}$

$$\Downarrow$$

A model $\mathscr{M}$ is an L^p-universal approximator a target $\mathscr{F}$

Compatibility of approximations and compositions

- Is a composition of approximations an approximation of the composition ?
- We may reduce the problem to approximations of small constituents

Proposition

$\mathscr{M}$: a set of piecewise C^1-diffeomorphisms

$F_1, \ldots, F_r$: **linearly increasing** piecewise C^1-diffeomorphims

Assume $\exists H_i \in \mathscr{M}$ such that

$$H_i \approx F_i \text{ (L^p-approximation on any compact sets)}$$

Then, for compact set $K \subset \mathbb{R}^d$, there exist $G_1, \ldots, G_r \in \mathscr{M}$ such that

$$G_r \circ \cdots \circ G_1 \approx F_r \circ \cdots \circ F_1 \text{ (L^p-approximation on K)}$$

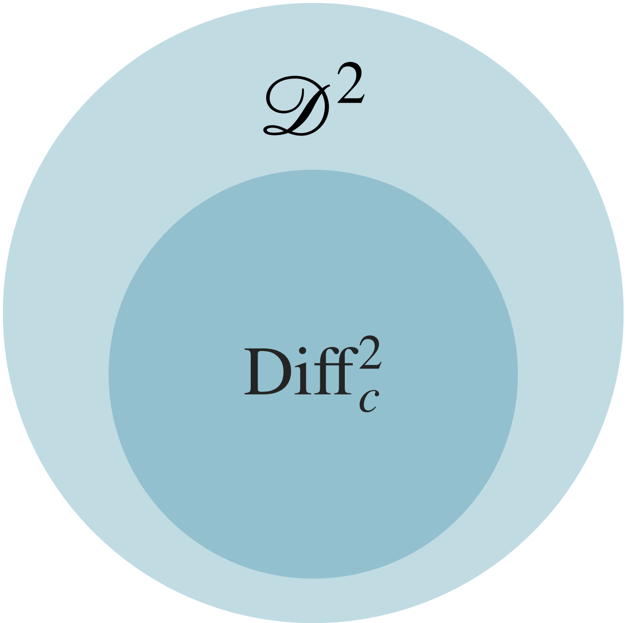
Remark

If $\mathscr{M}$ is composed of **locally bounded** maps and F_i's are **continuous**, we have a similar proposition for $\sup$-universal approximators.

A structure theorem of diffeomorphism groups

Definition (compactly supported diffeomorphisms)

Diff_c^2: the set of C^2-diffeomorphisms $f : \mathbb{R}^d \to \mathbb{R}^d$ such that $f(x) = x$ outside a compact subset ($U_f = \mathbb{R}^d$).



Theorem (Herman, Thurston, Epstein, and Mather)

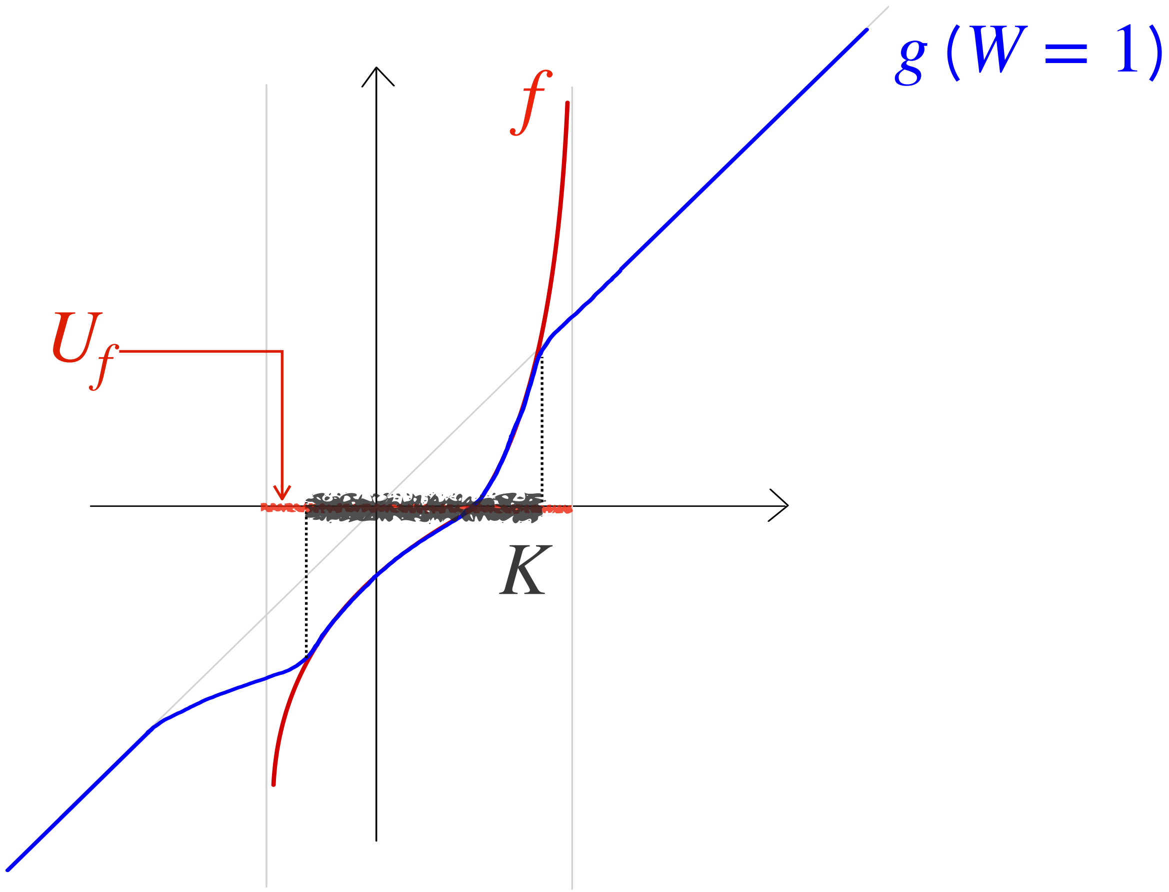
Diff_c^2 is a **simple group** (does not have any proper normal subgroup except $\{\mathrm{Id}\}$)

Extension of $\mathscr{D}^2$

Proposition

For $f \in \mathscr{D}^2$ ($f : U_f \to \mathbb{R}^d$) and compact subset $K \subset U_f$, there exist an affine transform $W \in \mathrm{Aff}$ and $g \in \mathrm{Diff}^2_c$ such that

$$f|_K = W \circ g|_K.$$



Flow endpoints

Definition (flow endpoints Ξ)

w.r.t. Whitney topology.

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

$\Xi := \{\text{flow endpoints}\}$

$\forall s,t,\ s+t \in [0,1],\ \phi(s+t) = \phi(s) \circ \phi(t)$

Proposition

The set of finite compositions of flow endpoints (the group generated by Ξ) is a **nontrivial normal subgroup** of Diff^2_c.

Corollary

For $g \in \mathrm{Diff}^2_c$, there exist **finite** flow endpoints $g_1, \ldots, g_m \in \Xi$ such that

$$g = g_1 \circ \cdots \circ g_m.$$

In particular, L^p-/sup-univ. for $\mathscr{D}^2$ $\iff$ L^p-/sup-univ. for Ξ

Paper 2: Universality of NODE

$f \in \mathscr{D}^2$: target, $K \subset U_f$: compact

$$f|_K$$

$\|$ « Extend $f|_K$

$\exists W \circ h$ (Aff & compactly supported C^2-diffeomorphism)

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

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

$\wr\wr$

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

Paper 2 Result (Analysis of NODEs)

NODEs yield $\sup$-univ. INNs for Ξ

(and hence $\sup$-univ. for $\mathscr{D}^2$. Also Dist-univ.).

Proof outline of result in Paper 1

L^p-/sup-univ. for $\mathscr{D}^2$ $\iff$ L^p-/sup-univ. for $\mathcal{S}_c^\infty$

$\mathcal{S}_c^\infty := \{\tau : \text{compactly supported } \tau(\mathbf{x}, y) = (\mathbf{x}, u(\mathbf{x}, y))\} \subset \mathrm{Diff}_c^2$

$u : \mathbb{R}^{d-1} \to \mathbb{R}, \quad (\mathbf{x}, y) \in \mathbb{R}^{d-1} \times \mathbb{R}$

$f|_K$ — $f \in \mathscr{D}^2$: target, $K \subset U_f$: compact

|| « Extend $f|_K$

$\exists W \circ h$ (Aff & compactly supported C^2-diffeomorphism)

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

$\exists h_1 \circ h_2 \circ \cdots$ (**flow endpoints** Ξ)

||

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

||

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

Decompose $f|_K$ into simpler mappings

Nearly Id

Definition (nearly-Id elements)

$g \in \mathrm{Diff}_c^2$: **nearly-Id element** **if** $\|dg(x) - I\| < 1$ for $x \in \mathbb{R}^d$

Proposition

For a flow endpoint $g \in \mathrm{Diff}_c^2$, there exist nearly-Id elements $g_1, \ldots, g_m \in \mathrm{Diff}_c^2$ such that

$$g = g_1 \circ \cdots \circ g_m.$$

$\because$ $g = \phi(1)$ ($\phi : [0,1] \to \mathrm{Diff}_c^2$: "additive" and continuous)

Then, $g = \phi(1/m)^m$ and $\phi(1/m) \to \mathrm{Id}$ as $m \to \infty$

Thus, we define $g_1 = g_2 = \ldots = g_m = \phi(1/m)$ for sufficiently large m ∎

Decomposition of nearly Id's

Proposition

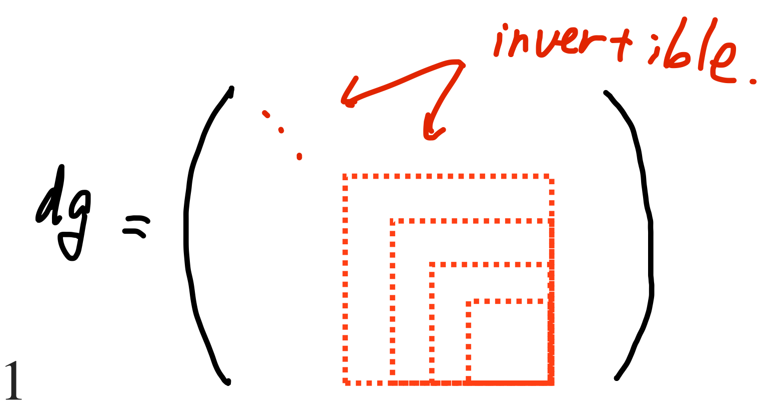
For a nearly-Id element $g \in \mathrm{Diff}_c^2$, there exist $\tau_1, \ldots, \tau_d \in \mathcal{S}_c^2$ and $\sigma_1, \ldots, \sigma_d \in \mathfrak{S}_d$ such that

$$g = \sigma_1 \circ \tau_1 \circ \cdots \circ \sigma_m \circ \tau_m.$$

Lemma for this proposition

For $g = (g_i)_{i=1}^d \in \mathrm{Diff}_c^2$, if for any $k = 1,\ldots,d$, the submatrix of its jacobian

$$\left(\frac{\partial g_{i+k-1}}{\partial x_{j+k-1}}(x) \right)_{i,j=1,\ldots,d-k-1}$$



is invertible for all x, then there exit $\tau_1, \ldots, \tau_d \in \mathcal{S}_c^2$ and $\sigma_1, \ldots, \sigma_d \in \mathfrak{S}_d$ such that

$$g = \sigma_1 \circ \tau_1 \circ \cdots \circ \sigma_m \circ \tau_m.$$

Proof outline of result in Paper 1

L^p-/sup-univ. for $\mathscr{D}^2$ $\iff$ L^p-/sup-univ. for $\mathscr{S}_c^\infty$

$\mathscr{S}_c^\infty := \{\tau : \text{compactly supported } \tau(\mathbf{x}, y) = (\mathbf{x}, u(\mathbf{x}, y))\} \subset \mathrm{Diff}_c^2$

$u : \mathbb{R}^{d-1} \to \mathbb{R}, \quad (\mathbf{x}, y) \in \mathbb{R}^{d-1} \times \mathbb{R}$

$f|_K$ — $f \in \mathscr{D}^2$: target, $K \subset U_f$: compact

|| « Extend $f|_K$

$\exists W \circ h$ (Aff & compactly supported C^2-diffeomorphism)

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

$\exists h_1 \circ h_2 \circ \cdots$ (**flow endpoints** Ξ)

||

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

||

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

Decompose $f|_K$ into simpler mappings

How the result can be used

You show

sup-univ. for $\mathscr{S}_c^\infty$

$\Downarrow$

L^p-univ. for $\mathscr{S}_c^\infty$

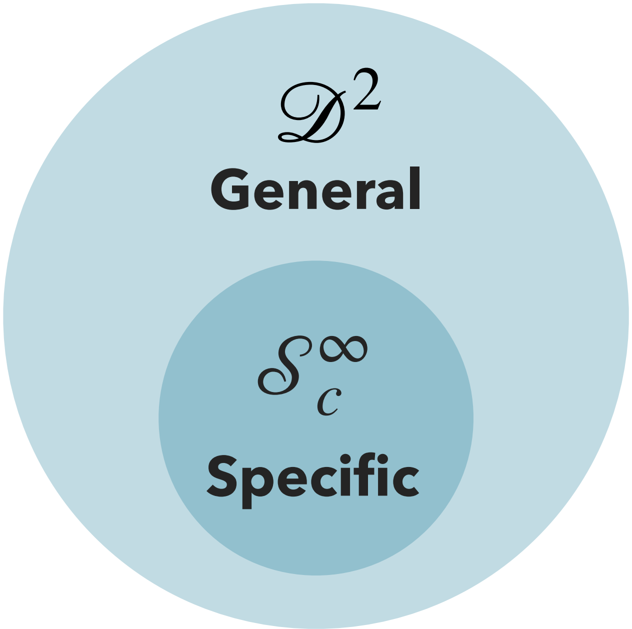
$\Longrightarrow$

You get

sup-univ. for $\mathscr{D}^2$

$\Downarrow$

L^p-univ. for $\mathscr{D}^2$



Upgrade Existing Guarantees

Regrading guarantees for existing INN architectures:

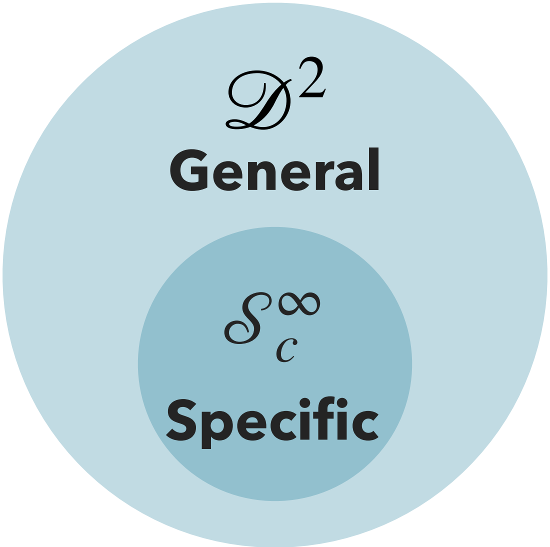
- **Sum-of-squares polynomial flow** (SoS-flow)
- **Deep sigmoidal flow** (DSF; aka. NAF)

Previously known/claimed [JSY19, HKLC18]:

$\sup$-universality for $\mathscr{S}_c^\infty$

$\Downarrow$

$\sup$-universality for $\mathscr{D}^2$



Distributional universal approximators

Definition (distributional universal approximator)

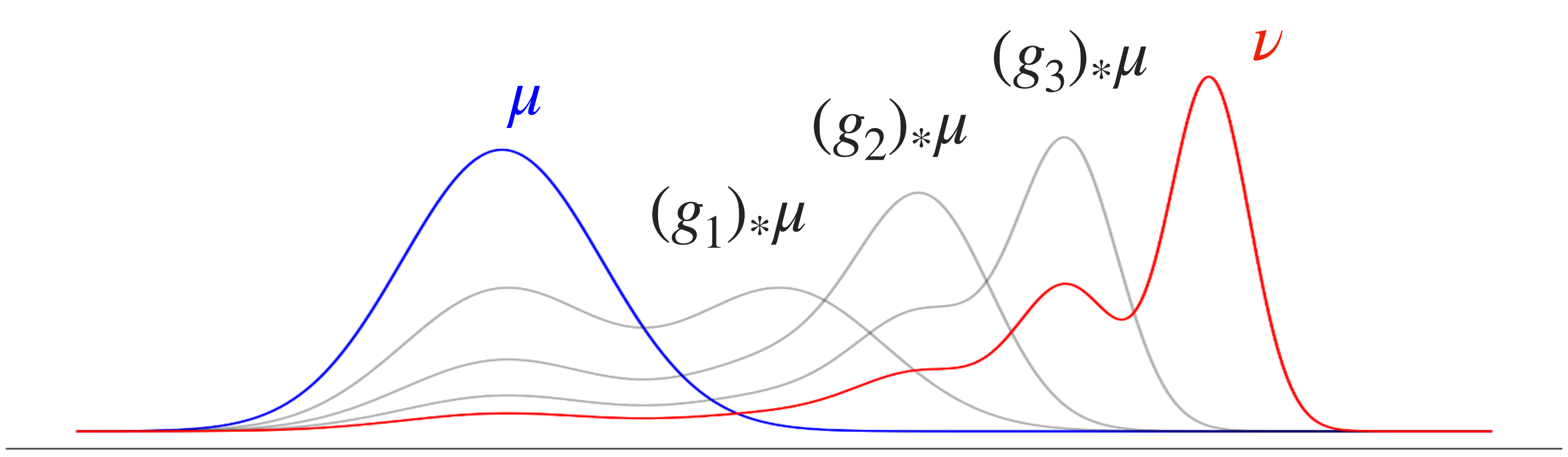
$\mathscr{M}$: model, set of measurable bijection from $\mathbb{R}^d$ to $\mathbb{R}^d$ (e.g. INNs)

$\mathscr{P}$: absolutely continuous probability measures

$\mathscr{M}$ is a **distributional universal approximator if**

$\forall \mu, \nu \in \mathscr{P}, \ \exists \{g_n\}_{n=1}^{\infty} \subset \mathscr{M}$

$$(g_n)_* \mu \underset{n \to \infty}{\longrightarrow} \nu \quad \text{(\textbf{weak} convergence).}$$



Relation of universalities

Proposition

A model $\mathscr{M}$ is a L^p-universal approximator for a target $\mathscr{D}^2$

$$\Downarrow$$

A model $\mathscr{M}$ is a distributional universal approximator

In summary, we obtain

[L^p-/sup-univ. for $\mathscr{D}^2$] $\Leftrightarrow$ [L^p-/sup-univ. for Ξ] $\Leftrightarrow$ [L^p-/sup-univ. for $\mathscr{S}_{\mathrm{c}}^{\infty}$] $\Rightarrow$ [Distributional-univ.]

Universality of CF-INNs

$\mathscr{H}$: functions on $\mathbb{R}^{d-1}$ (e,g, MLPs)

$\mathrm{INN}_{\mathscr{H}\text{-ACF}}$ is an INN with the flow layers composed of

$$\Psi_{d-1,s,t}(\mathbf{x}, y) := \left(\mathbf{x}, e^{s(\mathbf{x})}y + t(\mathbf{x})\right)$$

$$(\mathbf{x}, y) \in \mathbb{R}^{d-1} \times \mathbb{R},\ s, t \in \mathscr{H}$$

One of the simplest CF-INN

Theorem

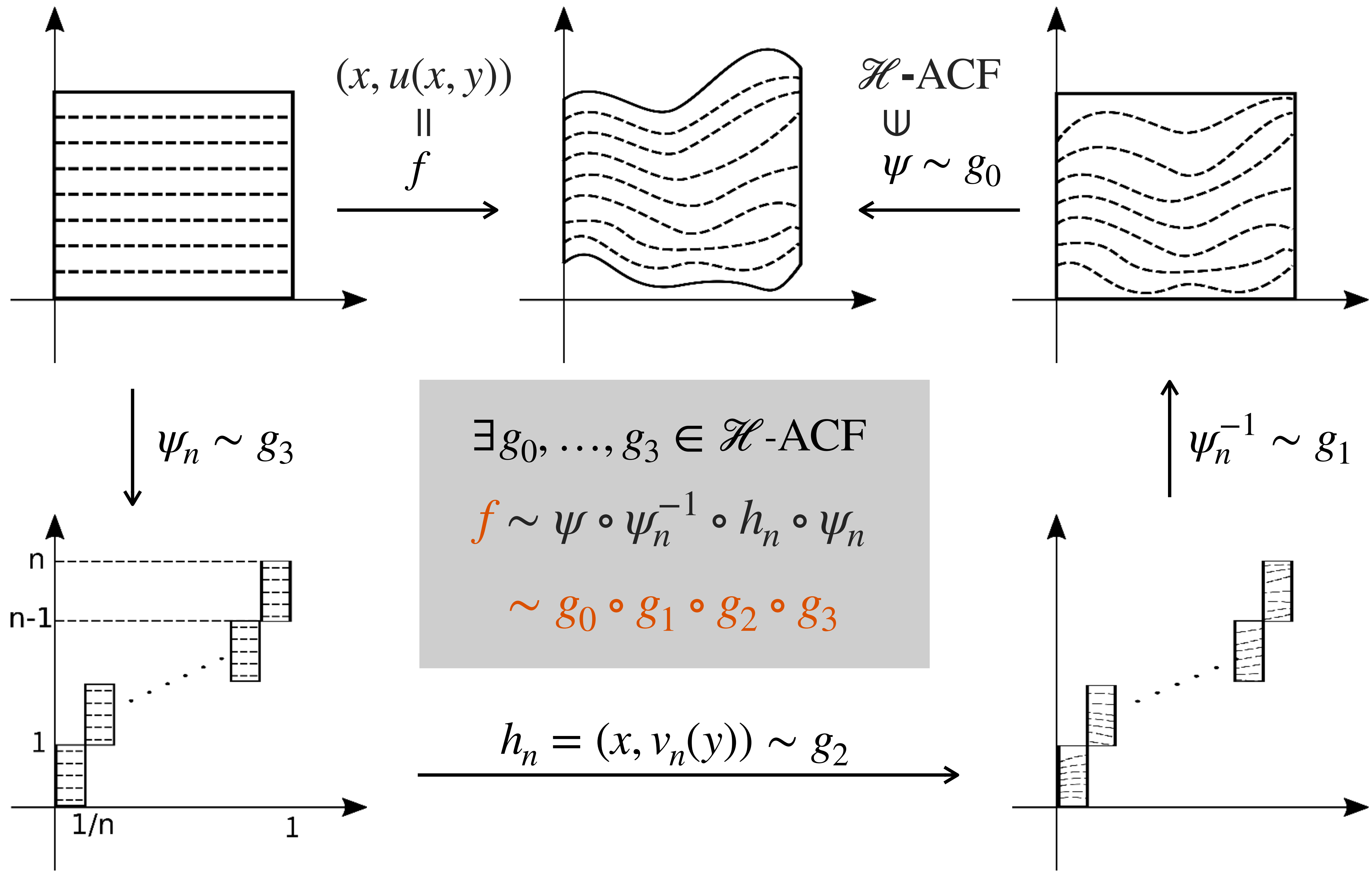
Assume $\mathscr{H}$ arbitrarily approximates any element in $C_c^\infty(\mathbb{R}^{d-1})$, and is composed of piecewise C^1-functions (e.g. MLPs with ReLU activation, RKHS with Gaussian kernel, e.t.c).

Then, $\mathrm{INN}_{\mathscr{H}\text{-ACF}}$ is an L^p-universal approximator for $\mathscr{S}_c^\infty$

Universality of CF-INNs

- We may assume $K = [0,1]^2$

For detail, look at our paper !



$$\psi_n := \Psi_{d-1,1,t_n} \qquad t_n := \sum_{k=0}^{n-1} k\mathbf{1}_{[k/n,(k+1)/n)} \qquad v_n(y) = \begin{cases} u(k/n, y) + k & y \in [k, k+1), \\ y & \text{otherwise}. \end{cases}$$

Universality of CF-INNs

L^p-/sup-univ. for $\mathscr{D}^2$ ⇔ L^p-/sup-univ. for Ξ ⇔ L^p-/sup-univ. for $\mathscr{S}_c^\infty$ ⇒ Distributional-univ.

Paper 1 Result (Affine Coupling Flows yield universal INNs)

Affine Coupling Flows yield L^p-univ. INNs for $\mathscr{S}_c^\infty$ (and hence for $\mathscr{D}^2$, and also Dist-univ.).

Remark

The representation power of invertible neural networks based on affine coupling flow is empirically known, and they were **conjectured** distributional universal approximator. We **affirmatively** answer this question.

Conclusion & Future Work

Conclusion

- Proposed a general theoretical framework to analyze the representation power (universalities) of invertible models.
- Guarantee the representation power of CF-INNs as an L^p -universal approximator.
- Guarantee the representation power of NODE-INNs as a $\sup$ -universal approximator.

Future work

- Quantitative analysis: Estimate the number of layers required for the approximation given the smoothness of the target.

Our papers are available at

[1] https://papers.nips.cc/paper/2020/hash/2290a7385ed77cc5592dc2153229f082-Abstract.html

[2] http://arxiv.org/abs/2012.02414

Message

CF-INNs and NODE-INNs can be relied on in modeling invertible functions and probability distributions.

Appendix

References

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

[HSW89] Hornik, K., Stinchcombe, M., & White, H. (1989).
Multilayer feedforward networks are universal approximators.
Neural Networks, 2(5), 359–366.

[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.

[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.

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

[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].

[KPB19] Kobyzev, I., Prince, S., & Brubaker, M. A. (2019).
Normalizing flows: An introduction and review of current methods.
ArXiv:1908.09257 [Cs, Stat].

References

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

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

[AKRK19] Ardizzone, L., Kruse, J., Rother, C., & Köthe, U. (2019).
Analyzing inverse problems with invertible neural networks.
7th International Conference on Learning Representations.

[BM19] Bauer, M., & Mnih, A. (2019).
Resampled priors for variational autoencoders.
In Proceedings of machine learning research, 89, 66–75.

[LW17] Louizos, C., & Welling, M. (2017).
Multiplicative normalizing flows for variational Bayesian neural networks.
In Proceedings of the 34th International Conference on Machine Learning,70, 2218–2227.

[NMT+19] Nalisnick, E. T., Matsukawa, A., Teh, Y. W., Görür, D., & Lakshminarayanan, B. (2019).
Hybrid models with deep and invertible features.
In Proceedings of the 36th International Conference on Machine Learning, 97, 4723–4732.

[IKFW20] Izmailov, P., Kirichenko, P., Finzi, M., & Wilson, A. G. (2020).
Semi-supervised learning with normalizing flows.
Proceedings of the 37th International Conference on Machine Learning.

References

[OLB+18] Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., Driessche, G., Lockhart, E., Cobo, L., Stimberg, F., Casagrande, N., Grewe, D., Noury, S., Dieleman, S., Elsen, E., Kalchbrenner, N., Zen, H., Graves, A., King, H., … Hassabis, D. (2018). Parallel WaveNet: Fast high-fidelity speech synthesis. Proceedings of the 35th International Conference on Machine Learning, 80, 3918–3926.

[TSS20] Teshima, T., Sato, I., & Sugiyama, M. (2020). Few-shot domain adaptation by causal mechanism transfer. Proceedings of the 37th International Conference on Machine Learning.

[KLSKY19] Kim, S., Lee, S.-G., Song, J., Kim, J., & Yoon, S. (2019). FloWaveNet: A generative flow for raw audio. In Proceedings of the 36th International Conference on Machine Learning, 97, 3370–3378.

[ZMWN19] Zhou, C., Ma, X., Wang, D., & Neubig, G. (2019). Density matching for bilingual word embedding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 1588–1598.

[WSB19] Ward, P. N., Smofsky, A., & Bose, A. J. (2019). Improving exploration in soft-actor-critic with normalizing flows policies. ArXiv:1906.02771 [Cs, Stat].

References

[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.

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

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

[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.

[PPMF20] Pumarola, A., Popov, S., Moreno-Noguer, F., & Ferrari, V. (2020).
C-Flow: Conditional Generative Flow Models for Images and 3D Point Clouds.
2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7946–7955.