Journal of Machine Learning Research 24 (2023) 1-68 Submitted 4/22; Revised 8/23; Published 8/23

Universal Approximation Property of Invertible Neural Networks

Isao Ishikawa[†] ISHIKAWA.ISAO.ZX@EHIME-U.AC.JP
Center for Data Science
Ehime University
Ehime, JAPAN
Center for Advanced Intelligence Project
RIKEN
Tokyo, JAPAN

Takeshi Teshima[†*] TAKESHI.TESHIMA@A.RIKEN.JP
The University of Tokyo
Tokyo, JAPAN
Center for Advanced Intelligence Project
RIKEN
Tokyo, JAPAN

Koichi Tojo KOICHI.TOJO@RIKEN.JP
Center for Advanced Intelligence Project
RIKEN
Tokyo, JAPAN

Kenta Oono KENTA.OONO@A.RIKEN.JP
Center for Advanced Intelligence Project
RIKEN
Tokyo, JAPAN

Masahiro Ikeda MASAHIRO.IKEDA@RIKEN.JP
Center for Advanced Intelligence Project
RIKEN
Tokyo, JAPAN

Masashi Sugiyama SUGI@K.U-TOKYO.AC.JP
Center for Advanced Intelligence Project
RIKEN
Tokyo, JAPAN
The University of Tokyo
Tokyo, JAPAN

Editor: Joan Bruna

Abstract

Invertible neural networks (INNs) are neural network architectures with invertibility by design. Thanks to their invertibility and the tractability of their Jacobians, INNs have

∗. This work was done when the author was with The University of Tokyo and RIKEN.
†. These authors contributed equally to this work.

various machine learning applications such as probabilistic modeling, generative modeling, and representation learning. However, their attractive properties often come at the cost of restricting the layer design, which poses a question on their representation power: can we use these models to approximate sufficiently diverse functions? To answer this question, we have developed a general theoretical framework to investigate the representation power of INNs, building on a structure theorem of differential geometry. The framework simplifies the approximation problem of diffeomorphisms, which enables us to show the universal approximation properties of INNs. We apply the framework to two representative classes of INNs, namely Coupling-Flow-based INNs (CF-INNs) and Neural Ordinary Differential Equations (NODEs), and elucidate their high representation power despite the restrictions on their architectures.

Keywords: invertible neural network, normalizing flow, universal approximation property, coupling flow, neural ordinary differential equation

1. Introduction

Invertible neural networks (INNs) are neural network architectures with invertibility by design. They are often endowed with tractable algorithms to compute the inverse map and the Jacobian determinant, such as their explicit formulas. These characteristics of INNs have enabled a series of new techniques in various machine learning tasks, for example, generative modeling (Dinh et al., 2017; Kingma and Dhariwal, 2018; Oord et al., 2018; Jacobsen et al., 2018; Behrmann et al., 2019; Kim et al., 2019; Zhou et al., 2019), probabilistic inference (Bauer and Mnih, 2019; Ward et al., 2019; Louizos and Welling, 2017), solving inverse problems (Ardizzone et al., 2019), feature extraction and manipulation (Kingma and Dhariwal, 2018; Nalisnick et al., 2019; Izmailov et al., 2020; Teshima et al., 2020b), quantum field theory (Albergo et al., 2019), modeling non-linear dynamics (Bevanda et al., 2022a,b), and 3D point cloud generation (Yang et al., 2019; Kim et al., 2020; Kimura et al., 2021).

INNs have been realized by the careful design of the special invertible layers called the *flow layers*. Examples of flow layer designs include *coupling flows* (CFs; Papamakarios et al., 2021; Kobyzev et al., 2021) and *neural ordinary differential equations* (NODEs; Chen et al., 2018). CFs employ a highly restricted network architecture in which only some of the input variables undergo some transformations, and the rest of the input variables become the output as-is without being transformed (Section 2.1.1). Also, NODEs offer flow layers by indirectly modeling an invertible function by transforming an input vector through an ordinary differential equation (ODE). To construct more flexible INNs, multiple such flow layers are composed as well as invertible affine transformation layers. Moreover, a variety of CF layer designs have been proposed to construct CF-INNs with high representation power, such as the affine coupling flow (Dinh et al., 2015, 2017; Kingma and Dhariwal, 2018; Papamakarios et al., 2017; Kingma et al., 2016), the neural autoregressive flow (Huang et al., 2018; Cao et al., 2019; Ho et al., 2019), and the polynomial flow (Jaini et al., 2019), each demonstrating enhanced empirical performance.

However, despite the diversity of flow-layer designs (Papamakarios et al., 2021; Kobyzev et al., 2021), and their popularity in practice, the theoretical understanding of the representation power of INNs has been limited. Indeed, the most basic property as a function approximator, namely the *universal approximation property* (or *universality* for short) (Cybenko, 1989; Hornik et al., 1989; Funahashi, 1989), has not been elucidated until recently

(Teshima et al., 2020a,c). The universality can be crucial when INNs are used to learn an invertible transformation such as feature extraction (Nalisnick et al., 2019) or independent component analysis (Teshima et al., 2020b) because, informally speaking, lack of universality implies that there exists an invertible transformation, even among well-behaved ones, that the INN can never approximate. The lack of universality could hinder the model's ability to generalize well across various tasks, potentially making it less reliable for function approximation in certain scenarios.

In this work, we show the high representation power of some representative architectures of CF-based INNs and NODE-based INNs by showing their universal approximation properties for a fairly large class of *diffeomorphisms*, namely smooth invertible maps with smooth inverse. The present article is an extended version of Teshima et al. (2020a) and Teshima et al. (2020c), but with substantial extensions. First, we extend the theoretical framework of Teshima et al. (2020a) by taking into account the approximation of the *derivatives* in addition to the function values. Investigating the representation power to approximate the derivatives can be important in providing machine learning methods with theoretical guarantees. For example, in Teshima et al. (2020b, Appendix C.7.), the Sobolev norm has been used to characterize the approximation error of an invertible model.

By such an extension, we also strengthen the theoretical guarantees for the distributional approximation using INNs. Whereas the preliminary version of the framework in Teshima et al. (2020a) could only guarantee the approximation capability in terms of the weak convergence topology, the present framework can elucidate the universality in terms of the *total variation* distance of distributions. Approximation in the total variation distance is a stronger notion that can be useful in providing machine learning algorithms with theoretical guarantees. See Remark 52 in Section 5.

The difficulty in proving the universality of INNs comes from two complications. (i) Only function composition can be leveraged to make accurate approximators (for example, a linear combination of sub-networks is not allowed, as opposed to standard fully-connected neural networks). (ii) INNs have architecture-specific inflexibility: CF layers have restricted function forms, and NODE layers can only model functions that can be realized by differential equations. We overcome these complications by problem reduction: we decompose a general diffeomorphism into much simpler ones by using a structural theorem of differential geometry that untangles the structure of a certain diffeomorphism group. By showing that CF layers and NODE layers can approximate the simple components of the target diffeomorphism, we prove the universality results.

We first provide a general theorem that shows the equivalence of the universality for certain diffeomorphism classes, which can be used to reduce the approximation of a general diffeomorphism to that of a much simpler one. Then, by leveraging this problem reduction, we show that certain examples of the CF layer designs and the NODEs result in universal approximators for a general class of diffeomorphisms.

Our contributions. Our contributions are summarized as follows.

1. We present a theorem to show the equivalence of universal approximation properties for certain classes of functions (Theorem 24 and Theorem 25). The result enables the reduction of the task of proving the universality for general diffeomorphisms to that for much simpler coordinate-wise ones. It generalizes and unifies the equivalence theorems

previously shown by Teshima et al. (2020a) and Teshima et al. (2020c), while removing some restrictions that these two previous studies imposed.

2. We relate functional universality (that is, universality for approximating functions) to distributional universality (that is, universality for approximating distributions by pushforward). We introduce a new type of functional approximation property, namely *Sobolev universality*, which is a stronger notion of what has been previously considered by Teshima et al. (2020a) and Teshima et al. (2020c). Then, we show Sobolev universality implies the distributional universality in terms of the *weak* topology (Corollary 39) and the topology induced by the *total variation* norm (Corollary 41) under appropriate assumptions.

3. We show that the INNs based on certain CF architectures have the Sobolev universality, implying they may be more suitable choices for obtaining theoretical guarantees in the machine learning tasks that require the approximation of derivatives.

Notation We list the mathematical notations we use in this paper in the notation tables in Appendix. We also summarize several mathematical notions and their properties in Appendix A.

2. Preliminaries and Related Work

In this section, we describe the models analyzed in this study, the notion of universality, and related work.

2.1 Invertible Layers

We introduce several invertible layers we consider in this paper, which constitute invertible neural networks.

2.1.1 Coupling-flow Based Invertible Neural Networks (CF-INNs)

We fix $d \in \mathbb{N}$ and assume $d \geq 2$. For a vector $\boldsymbol{x} \in \mathbb{R}^d$ and $k \in [d-1]$, we define $\boldsymbol{x}_{\leq k}$ as the vector $(x_1, \ldots, x_k)^\top \in \mathbb{R}^k$ and $\boldsymbol{x}_{>k}$ the vector $(x_{k+1}, \ldots, x_d)^\top \in \mathbb{R}^{d-k}$.

Definition 1 (Coupling flows) *We define a coupling flow (CF) (Papamakarios et al., 2021) $h_{k,\tau,\theta}$ by $h_{k,\tau,\theta}(\boldsymbol{x}_{\leq k}, \boldsymbol{x}_{>k}) = (\boldsymbol{x}_{\leq k}, \tau(\boldsymbol{x}_{>k}, \theta(\boldsymbol{x}_{\leq k})))$, where $k \in [d-1]$, $\theta \colon \mathbb{R}^k \to \mathbb{R}^l$ and $\tau : \mathbb{R}^{d-k} \times \mathbb{R}^l \to \mathbb{R}^{d-k}$ are maps, and $\tau(\cdot, \theta(\boldsymbol{y}))$ is an invertible map for any $\boldsymbol{y} \in \mathbb{R}^k$.*

One of the most standard types of CFs is *affine coupling flows* (Dinh et al., 2017; Kingma and Dhariwal, 2018; Kingma et al., 2016; Papamakarios et al., 2017).

Definition 2 (Affine coupling flows) *We define an affine coupling (ACF) flow by the map $\Psi_{k,s,t}$ from $\mathbb{R}^d$ to $\mathbb{R}^d$ such that*

$$\Psi_{k,s,t}(\boldsymbol{x}_{\leq k}, \boldsymbol{x}_{>k}) = (\boldsymbol{x}_{\leq k}, \boldsymbol{x}_{>k} \odot \exp(s(\boldsymbol{x}_{\leq k})) + t(\boldsymbol{x}_{\leq k})),$$

where $k \in [d-1]$, $\odot$ is the Hadamard product, exp *is applied in an element-wise manner, and $s, t : \mathbb{R}^k \to \mathbb{R}^{d-k}$ are maps.*

The maps s and t are typically parametrized by neural networks.

Definition 3 (Single-coordinate affine coupling flows) *Let $\mathcal{H}$ be a set of functions from $\mathbb{R}^{d-1}$ to $\mathbb{R}$. We define the set of* $\mathcal{H}$-single-coordinate affine coupling flows *as a subclass of ACFs by $\mathcal{H}$-ACF $:= \{\Psi_{d-1,s,t} : s, t \in \mathcal{H}\}$.*

$\mathcal{H}$-ACF is the least expressive flow design appearing in this paper. However, we show in Section 4.1 that it can form a CF-INN with universality. Later, we require various regularity conditions on $\mathcal{H}$ depending on the type of universality we want to show.

2.1.2 Neural ordinary differential equations (NODEs)

Here, we define the family of NODEs considered in the present paper. NODE is based on the following fact that any *autonomous* ODE (that is, an ODE is defined by a time-invariant vector field) with a Lipschitz continuous vector field has a solution and that the solution is unique:

Fact 4 (Existence and uniqueness of a global solution to an ODE) *Let $f \in \mathrm{Lip}$. Then, a solution $z\colon \mathbb{R} \to \mathbb{R}^d$ to the following ODE exists, and it is unique:*

$$z(0) = \boldsymbol{x}, \quad \dot{z}(t) = f(z(t)), \quad t \in \mathbb{R}, \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^d$, and $\dot{z}$ denotes the derivative of z (see Derrick and Janos (1976) for example).

In view of Fact 4, we use the following notation.

Definition 5 (Autonomous-ODE flow endpoints) *For $f \in \mathrm{Lip}$, $\boldsymbol{x} \in \mathbb{R}^d$, and $t \in \mathbb{R}$, we define*

$$\mathrm{IVP}[f](\boldsymbol{x}, t) := z(t),$$

where $z : \mathbb{R} \to \mathbb{R}^d$ is the unique solution to Equation (1). Then, for $\mathcal{F} \subset \mathrm{Lip}$, we define

$$\Psi(\mathcal{F}) := \{\mathrm{IVP}[f](\cdot, 1) \mid f \in \mathcal{F}\}.$$

See, for example, (Li et al., 2022). Note that the elements of $\Psi(\mathcal{F})$ are invertible.

2.2 Invertible Neural Networks (INNs)

We consider the INN architectures constructed by composing flow layers, defined as follows.

Definition 6 (INNs) *Let $\mathcal{G}$ be a set consisting of bijective maps on $\mathbb{R}^d$. We define the set of INNs based on $\mathcal{G}$ as*

$$\mathrm{INN}_{\mathcal{G}} := \{W_1 \circ g_1 \circ \cdots \circ W_n \circ g_n : \ n \in \mathbb{N}, g_i \in \mathcal{G}, W_i \in \mathrm{Aff}\}.$$

Remark 7 *Previous studies such as Kingma and Dhariwal (2018) used* GL *(see Table 3 for its definition) in place of* Aff *in the definition of* $\mathrm{INN}_{\mathcal{G}}$*. This difference is not a problem in most cases. For example, if there exist finite elements of $\mathcal{G}$ such that their composition equals the map $x \mapsto x + b$ for an arbitrary vector $b \in \mathbb{R}^d$, then, replacing* Aff *with* GL *does not change the function set* $\mathrm{INN}_{\mathcal{G}}$*. In fact, when $\mathcal{G}$ contains $\mathcal{H}$-ACF with minimal requirements on $\mathcal{H}$, we can further reduce the set of linear transformations for INNs from* Aff *to the symmetric group $\mathfrak{S}_d$, that is, the permutations of variables. See Appendix E.1 for details.*

2.3 Universal Approximation Properties

Here, we clarify the notions of universality in this paper. The definitions use general topological terms, generalizing the L^p-universality and sup-universality in Teshima et al. (2020a,c).

2.3.1 Functional universality

We define the notion of universality for sets of functions, which is a key notion in this paper. Roughly speaking, a model class is universal for a set of target functions if one can always find a model in the proximity of any target function. The notion of proximity is stated in general terms of topology.

Definition 8 (General functional universality) *Let U be a subset of $\mathbb{R}^m$ and let $\mathcal{F}_0$ be an $\mathbb{R}^n$-valued function space on U with some topology and let $\mathcal{F} \subset \mathcal{F}_0$ be a subset. Let $\mathcal{M}$ be a model, which is a set of measurable maps from $\mathbb{R}^m$ to $\mathbb{R}^n$. We say that $\mathcal{M}$ is an $\mathcal{F}_0$-universal approximator for $\mathcal{F}$ (or has an $\mathcal{F}_0$-universal approximation property for $\mathcal{F}$), if $\{g|_U : g \in \mathcal{M}\}$ is a subset of $\mathcal{F}_0$ and its closure contains $\mathcal{F}$.*

It is well-known that 2-layer neural networks with suitable activation functions are universal, namely, they can approximate any continuous functions on any compact set in $\mathbb{R}^d$ (see, for example, Cybenko 1989). In the manner of Definition 8, we can translate this fact into the $C^0(\mathbb{R}^d)$-universal approximation property of 2-layer neural networks for $C^0(\mathbb{R}^d)$, where we equip $C^0(\mathbb{R}^d)$ with the topology with seminorms composed of the sup norms on compact sets.

As an example of $\mathcal{F}_0$, we typically use the $\mathbb{R}^n$-valued *local Sobolev space* $W^{r,p}_{\mathrm{loc}}(U, \mathbb{R}^n)$, which is roughly speaking the space of r-times (weakly-) differentiable measurable functions f such that for any compact set $K \subset U$, $\|f\|_{K,r,p} < \infty$, where

$$\|f\|_{K,r,p} := \begin{cases} \displaystyle\sum_{|\alpha| \leq r} \left(\int_K \|\partial^\alpha f(x)\|^p dx \right)^{1/p} & \text{if } p < \infty, \\ \displaystyle\sum_{|\alpha| \leq r} \mathrm{ess.sup}_{x \in K} \|\partial^\alpha f(x)\| & \text{if } p = \infty. \end{cases}$$

Formally, we define the local Sobolev space as follows.

Definition 9 (McDuff and Salamon, 2004, Appendix B) *Let U be a subset of $\mathbb{R}^m$, r a non-negative integer, and $p \in [1, \infty]$. We define the local Sobolev space $W^{r,p}_{\mathrm{loc}}(U, \mathbb{R}^n)$ by*

$$W^{r,p}_{\mathrm{loc}}(U, \mathbb{R}^n) := \varprojlim_V W^{r,p}(V, \mathbb{R}^n),$$

where the right hand side is explicitly defined as the following set:

$$\left\{ (f_V)_V \in \prod_{V \subset U:\ \text{bounded open} \atop \overline{V} \subset U} W^{r,p}(V, \mathbb{R}^n) : f_{V_1}|_{V_2} = f_{V_2} \text{ if } V_2 \subset V_1 \right\}.$$

Here, $W^{r,p}(V,\mathbb{R}^n)$ is the $\mathbb{R}^n$-valued Sobolev space on V. The local Sobolev space is equipped with the relative topology of the product of the Sobolev spaces. We denote $W^{0,p}_{\mathrm{loc}}(U,\mathbb{R}^n)$ by $L^p_{\mathrm{loc}}(U,\mathbb{R}^n)$.

Remark 10 *The approximation in terms of the topology of $W^{r,p}_{\mathrm{loc}}(U,\mathbb{R}^n)$ is equivalent to the usual notion of approximation in terms of the norms. Namely, a function f on U is in the closure of a model $\mathcal{M}$ in the sense of the topology of the local Sobolev space $W^{r,p}_{\mathrm{loc}}(U,\mathbb{R}^n)$ if and only if for any compact set K and $\varepsilon > 0$, there exists $g \in \mathcal{M}$ such that $\|f - g|_U\|_{K,r,p} < \varepsilon$.*

The ordinary function classes can be considered to be subsets of $W^{r,p}_{\mathrm{loc}}(U,\mathbb{R}^n)$ as in the following proposition.

Proposition 11 *Let $U \subset \mathbb{R}^m$ be an open subset, $p \in [1,\infty]$, and $f : U \to \mathbb{R}^n$ be a measurable mapping.*

1. *If $r = 0$ and f is L^∞, then $(f|_V)_V \in W^{0,p}_{\mathrm{loc}}(U,\mathbb{R}^n)$.*
2. *If $r \geq 1$ and f is locally $C^{r-1,1}$ (see Table 3 for the definition), then $(f|_V)_V \in W^{r,p}_{\mathrm{loc}}(U,\mathbb{R}^n)$.*

Proof The first statement is easily shown by noting that L^∞ is locally L^p. The second statement follows from Remark 2.12 of Ern and Guermond (2021) and induction on r. ■

In words, according to Proposition 11, we can embed a set of suitable functions on U into $W^{r,p}_{loc}(U,\mathbb{R}^n)$ via the correspondence $f \mapsto (f|_V)_V$. Therefore, usual models, for example, Multilayer perceptron (MLP) with rectifier linear unit (ReLU) activation functions, are contained in $W^{1,p}_{\mathrm{loc}}$ as they are usually locally Lipschitz (note that locally $C^{0,1}$ means locally Lipschitz).

We call $W^{r,p}_{\mathrm{loc}}(U,\mathbb{R}^n)$-universality the *Sobolev* universality and introduce a special notion for simplicity:

Definition 12 ($W^{r,p}$-universality and L^p-universality) *Notations are as in Definition 8. Let r be a non-negative integer and let $p \in [1,\infty]$. We say a model $\mathcal{M}$ is a $W^{r,p}$-universal approximator for $\mathcal{F}$ (or has a $W^{r,p}$-universal approximation property for $\mathcal{F}$) if the model $\mathcal{M}$ is a $W^{r,p}_{\mathrm{loc}}(U,\mathbb{R}^n)$-universal approximator for $\mathcal{F}$. In the case of $r = 0$, we use L^p- instead of $W^{0,p}$-, for example, we say an L^p-universal approximator instead of a $W^{0,p}$-universal approximator.*

We note that the $W^{r,p}$-universal approximation property implies the $W^{r',p'}$-universal approximation property if $r \geq r'$ and $p \geq p'$.

Remark 13 *If $\mathcal{F}^0$ in Definition 8 is the space of locally bounded measurable maps with seminorms of* sup *(not ess.sup) norms on compact sets, a model with $\mathcal{F}^0$-universal approximation property is called a* sup*-universal approximator. The notion of* sup*-universality was introduced in Teshima et al. (2020a) and Teshima et al. (2020c) and is a slightly different concept from L^∞-universality. We mainly deal with L^∞-universality in this paper.*

2.3.2 Distributional universality

We define the notion of distributional universality. Distributional universality has been used as a notion of theoretical guarantees in the literature on normalizing flows, that is, probability distribution models constructed using INNs (Kobyzev et al., 2021). We here provide a generalized version of the classical distributional universality as follows:

Definition 14 (General distributional universality) *Let $\mathcal{M}$ be a model which is a set of measurable maps from $\mathbb{R}^m$ to $\mathbb{R}^n$. Let $\mathcal{P}_0$ be a set of probability measures on $\mathbb{R}^n$ with some topology. Let $\mathcal{Q} \subset \mathcal{P}_0$ be a subset. Fix probability measure μ_0 on $\mathbb{R}^m$. We say that a model $\mathcal{M}$ is a* $(\mathcal{P}_0, \mu_0)$-distributional universal approximator *for $\mathcal{Q}$ (or* has the $(\mathcal{P}_0, \mu_0)$-distributional universal approximation property *for $\mathcal{Q}$) if $\{g_*\mu_0 : g \in \mathcal{M}\} \subset \mathcal{P}_0$ and the closure of the set $\{g_*\mu_0 : g \in \mathcal{M}\}$ in $\mathcal{P}_0$ contains $\mathcal{Q}$. Here, $g_*\mu_0$ denotes the pushforward of μ_0 by g.*

Remark 15 *When $\mathcal{P}_0 = \mathcal{Q} = \mathcal{P}^{\mathrm{w}}$ (see Table 3 for the definition of $\mathcal{P}^{\mathrm{w}}$), $(\mathcal{P}_0, \mu_0)$-distributional universality for $\mathcal{Q}$ is equivalent to the sequential convergence, that is, the existence of a sequence $\{g_i\}_{i=1}^{\infty} \subset \mathcal{M}$ for each $\nu \in \mathcal{P}$ such that $g_{i*}\mu_0$ converges to ν in distribution as $i \to \infty$.*

Remark 16 *The distributional universality described in Definition 14 is a generalized notion considered in existing work. For example, the distributional universality in Jaini et al. (2019) is rephrased as a $(\mathcal{P}^{\mathrm{w}}, \nu)$-distributional universal approximation property for $\mathcal{P}_{\mathrm{ab}}$ for any $\nu \in \mathcal{P}_{\mathrm{ab}}$ in our terminology. Teshima et al. (2020a) extended the definition by Jaini et al. (2019). Their distributional universality is a $(\mathcal{P}^{\mathrm{w}}, \nu)$-distributional universal approximation property for $\mathcal{P}$ for any $\nu \in \mathcal{P}_{\mathrm{ab}}$. It is worth noting that these two concepts of distributional universal approximation are equivalent. This is essential because absolutely continuous probability measures are dense in the set of all the probability measures. We prove this fact as Lemma C.1 in Appendix C.1.*

The different notions of universality are interrelated. Most importantly, the L^p-universality for a certain function class implies the distributional universality (see Proposition 38). Moreover, if a model $\mathcal{M}$ is a sup-universal approximator for $\mathcal{F}$, it is also an L^p-universal approximator for $\mathcal{F}$ for any $p \in [1, \infty]$.

2.4 Related Work

Several studies showed the functional or distributional universality of INNs other than CF-INNs and NODEs. They are not competitive with but complementary to ours as their problem settings are different from ours in target models and evaluation norms. Gopal (2021) proposed a type of INNs named Exact-Lipschitz Flows (ELF) and proved their functional universality (more specifically, sup-universality in our terminology). Kong and Chaudhuri (2021) showed the universality of residual flows in terms of the maximum mean discrepancy (MMD). They quantitatively evaluated the number of layers needed to approximate a target function with prescribed precision.

Another line of work is to study the expressive power of specific forms of CF-INNs and NODEs. Huang et al. (2021) introduced Convex Potential Flows, which is a parameterization of invertible models inspired by the optimal transport theory. They proved its distributional

universality. Ruiz-Balet and Zuazua (2021) analyzed a NODE coming from the following form:

$$\dot{x}(t) = W(t)\sigma(A(t)x(t) + b(t)),$$

where A, W, and b are time-dependent matrices and a vector. They showed that, despite the restricted form, the flow generated by the ODE above has the L^2-universal approximation property. It is an interesting research direction to develop a general theory to broaden the applicability of our results to models like theirs.

Since the publication of our previous work (Teshima et al., 2020a,c), several researchers have studied the universality of INNs based on our theory. Puthawala et al. (2022) showed that injective flows between $\mathbb{R}^n$ and $\mathbb{R}^m$ ($n \leq m$) universally approximate measures supported on the images of *extendable embeddings*, which is a composition of a full-rank linear transformation followed by a diffeomorphism, in terms of the Wasserstein distance. Their results were built on our previous result of the sup-universality of neural autoregressive flows. Abe et al. (2021) proposed a novel network architecture called *Abelian group networks* that employs INNs as building blocks. They proved that Abelian group networks have a functional universal approximation property for Abelian Lie group operations on a Euclidean space. They essentially used the universality of INNs in the proof of the theorem. Also, concurrently with the present work, Lyu et al. (2022) showed the universality of CF-INNs in the C^k-norm, that is, a notion of universality taking into account the approximation of derivatives. Their result on the *C^k-universality*, namely Theorem 3.5 in Lyu et al. (2022), can be reproduced as a special case in our Theorem 24 by selecting $p = \infty$ and $\mathcal{G}$ to be a set of diffeomorphisms. While their proof has the advantage of being more concise thanks to focusing on this special case, they require the models to be smooth everywhere. On the other hand, our result can accommodate those flow layers which are not smooth everywhere, for example, CF layers with ReLU activation functions that are prevalent in applications. On a more technical side, our result provides a finer understanding of the diffeomorphism group $\mathrm{Diff}_{\mathrm{c}}^r$, which allows us to provide a theoretical guarantee of NODE-based INNs. More concretely, their proof directly uses the fact that the elements of $\mathrm{Diff}_{\mathrm{c}}^r$ can be decomposed into near-Id diffeomorphisms, while our Theorem 24 indicates that $\mathrm{Diff}_{\mathrm{c}}^r$ can be decomposed into the elements of Ξ^r, which can be further decomposed into near-Id diffeomorphisms.

As for theoretical limitations of INNs, Okuno and Imaizumi (2021) showed the lower bound (in a minimax sense) of estimation risks in non-parametric regression problems for estimating invertible functions on a plane. Although they constructed an estimator that achieved the lower bound, it is not known whether INNs of any kind can achieve this optimality.

3. General Framework

In this section, we present the main results (Theorems 24 and 25) of this paper on the universality of INNs. The main theorem breaks down the functional universality for a general class of diffeomorphisms into that for a much simpler class of diffeomorphisms. We also explain the implication of the main theorem to the distributional universality. The results in this section are derived and stated in a general setup so that it is not limited to the representation power analyses of specific INN architectures.

3.1 Equivalence of Universal Approximation Properties

Our first main theorem allows us to lift a universality result for a restricted set of diffeomorphisms to the universality for a fairly general class of diffeomorphisms by showing a certain equivalence of universalities. Thanks to this problem reduction, we can essentially circumvent the major complication in proving the universality of CF-INNs, namely that only function composition can be leveraged to make complex approximators (for example, a linear combination is not allowed).

We define the following classes of invertible functions: C^r-diffeomorphisms $\mathcal{D}^r$, flow endpoints Ξ^r, triangular transformations $\mathcal{T}^\infty$, and single-coordinate transformations $\mathcal{S}^r_{\mathrm{c}}$. Our main theorem later reveals an equivalence of $W^{r,p}$-universality for these classes.

First, we define the set of C^r-diffeomorphisms.

Definition 17 (C^r-diffeomorphisms: $\mathcal{D}^r$) *Let $0 \leq r \leq \infty$. For each open subset $U \subset \mathbb{R}^d$, we define $\mathcal{D}^r_U$ to be the set of maps from U to $\mathbb{R}^d$ which are C^r-diffeomorphisms from U to their images. We denote $\mathcal{D}^r := \sqcup_U \mathcal{D}^r_U$ (the formal disjoint union of the sets), where $U \subset \mathbb{R}^d$ runs over the set of all open subsets which are C^r-diffeomorphic to $\mathbb{R}^d$. Let $s \leq r$. We say that a model $\mathcal{M}$ is a $W^{s,p}$-universal approximator for $\mathcal{D}^r$ if $\mathcal{M}$ is a $W^{s,p}$-universal approximator for $\mathcal{D}^r_U$ for any open subset $U \subset \mathbb{R}^d$ that is C^r-diffeomorphic to $\mathbb{R}^d$.*

We require the domain U to be C^r-diffeomorphic to $\mathbb{R}^d$ for technical reasons. However, this constraint would not be too strong: the entire $\mathbb{R}^d$, any open convex set, and, more generally, any star-shaped open set, all satisfy this condition. In addition, it is known that if $d \geq 5$, any connected and simply connected open subset in $\mathbb{R}^d$ is always C^∞-diffeomorphic to $\mathbb{R}^d$.

Remark 18 *Although we required the domain U to be C^r-diffeomorphic to $\mathbb{R}^d$ in Definition 17, we may replace "C^r-diffeomorphic" with "C^∞-diffeomorphic" if $r > 0$. In fact, in the case of $r > 0$, it is known that an open subset of $\mathbb{R}^d$ is C^r-diffeomorphic to $\mathbb{R}^d$ if and only if C^∞-diffeomorphic (Hirsch, 1976, p.50, Theorem 2.7). On the other hand, in the case of $r = 0$, there exists an open subset that is homeomorphic to $\mathbb{R}^d$ but not diffeomorphic to $\mathbb{R}^d$.*

Before going to the second class, we define the set of *compactly-supported* diffeomorphisms on $\mathbb{R}^d$ as its container.

Definition 19 (Compactly supported diffeomorphism: $\mathrm{Diff}^r_{\mathrm{c}}$) *We say a diffeomorphism f on $\mathbb{R}^d$ is* compactly supported *if there exists a compact subset $K \subset \mathbb{R}^d$ such that for any $x \notin K$, $f(x) = x$. We use $\mathrm{Diff}^r_{\mathrm{c}}$ to denote the set of all compactly supported C^r-diffeomorphisms ($1 \leq r \leq \infty$) from $\mathbb{R}^d$ to $\mathbb{R}^d$. We regard $\mathrm{Diff}^r_{\mathrm{c}}$ as a group whose group operation is function composition. For $f \in \mathrm{Diff}^r_{\mathrm{c}}$, we define $\mathrm{supp} f \subset \mathbb{R}^d$ by the closure of the set $\{x \in \mathbb{R}^d : f(x) \neq x\}$, which is compact by definition.*

Our second class is a subset Ξ^r of $\mathrm{Diff}^r_{\mathrm{c}}$ consisting of *flow endpoints*.

Definition 20 (Flow endpoints: Ξ^r) *Let $1 \leq r \leq \infty$. Let $\Xi^r \subset \mathrm{Diff}^r_{\mathrm{c}}$ be the set of diffeomorphisms g of the form $g(\boldsymbol{x}) = \Phi(\boldsymbol{x}, 1)$ for some map $\Phi : \mathbb{R}^d \times U \to \mathbb{R}^d$ such that*

- *$U \subset \mathbb{R}$ is an open interval containing $[0, 1]$,*
- *$\Phi(\boldsymbol{x}, 0) = \boldsymbol{x}$,*

- $\Phi(\cdot, t) \in \mathrm{Diff}_{\mathrm{c}}^r$ *for any* $t \in U$,
- $\Phi(\boldsymbol{x}, s+t) = \Phi(\Phi(\boldsymbol{x}, s), t)$ *for any* $s, t \in U$ *with* $s+t \in U$,
- Φ *is* C^r *on* $\mathbb{R}^d \times U$, *and*
- *there exists a compact subset* $K_\Phi \subset \mathbb{R}^d$ *such that* $\cup_{t\in U}\mathrm{supp}\Phi(\cdot, t) \subset K_\Phi$.

Remark 21 *Definition 20 is the same as Definition 7 of Teshima et al. (2020c). A similar definition of flow endpoints can be found in Definition 9 of Teshima et al. (2020a). The difference between Definition 20 and the one of Teshima et al. (2020a) mainly lies in the last two conditions. Technically, these two conditions are used in Theorem 44 for showing that the partial derivative of* Φ *in* t *at* $t = 0$ *is Lipschitz continuous. We can prove the universality of CF-INNs without these two conditions, as done in Teshima et al. (2020a).*

Finally, we define two subclasses of $\mathcal{D}_{\mathbb{R}^d}^r$ as follows:

Definition 22 (Triangular transformations: $\mathcal{T}^\infty$) *We define* $\mathcal{T}^\infty$ *as the set of all* increasing triangular C^∞*-maps from* $\mathbb{R}^d$ *to* $\mathbb{R}^d$. *Here, we say a map* $\tau = (\tau_1, \ldots, \tau_d) : \mathbb{R}^d \to \mathbb{R}^d$ *is increasing triangular if each* $\tau_k(\boldsymbol{x})$ *depends only on* $\boldsymbol{x}_{\leq k}$ *and is strictly increasing with respect to* x_k.

Definition 23 (Single-coordinate transformations: $\mathcal{S}_{\mathrm{c}}^r$) *We define* $\mathcal{S}_{\mathrm{c}}^r$ *as the set of all compactly-supported* C^r*-diffeomorphisms* τ *satisfying* $\tau(\boldsymbol{x}) = (x_1, \ldots, x_{d-1}, \tau_d(\boldsymbol{x}))$, *that is, those which alter only the last coordinate.*

Note that for any $r \geq 1$, we have

$$\begin{array}{ccccc} \mathcal{D}_{\mathbb{R}^d}^0 & \supset & \mathrm{Diff}_{\mathrm{c}}^0 & & \\ \cup & & \cup & & \\ \mathcal{D}_{\mathbb{R}^d}^r & \supset & \mathrm{Diff}_{\mathrm{c}}^r & \supset & \Xi^r \\ \cup & & \cup & & \\ \mathcal{T}^\infty & \supset & \mathcal{S}_{\mathrm{c}}^\infty & & \end{array}$$

Remark that τ_d for $\tau \in \mathcal{S}_{\mathrm{c}}^r$ $(r \geq 0)$ is strictly increasing with respect to x_d since the C^r-diffeomorphism τ is compactly supported. Among the above classes of invertible functions, $\mathcal{D}^r$ is our main approximation target, and it is a fairly large class. The class $\mathcal{T}^\infty$ relates to the distributional universality as we will see in Proposition 38. The class $\mathcal{S}_{\mathrm{c}}^\infty$ is a much simpler class of diffeomorphisms that we use as a stepladder for showing the universality for $\mathcal{D}^r$.

Now we are ready to state the first main theorem. It reveals an equivalence among the universalities for $\mathcal{D}^r$, Ξ^∞, $\mathcal{T}^\infty$, and $\mathcal{S}_{\mathrm{c}}^\infty$, under mild regularity conditions. We can use the theorem to lift up the universality for $\mathcal{S}_{\mathrm{c}}^\infty$ to that for $\mathcal{D}^r$.

Theorem 24 (Equivalence for Sobolev universality) *Let* $p \in [1, \infty]$ *and let* $r \geq 0$ *be a nonnegative integer. Let* $\mathcal{G}$ *be a set of invertible functions from* $\mathbb{R}^d$ *to* $\mathbb{R}^d$. *Suppose one of the following two conditions hold:*

(A) When $p < \infty$, all elements of $\mathcal{G}$ are C^r and piecewise C^{r+1}-diffeomorphisms if $r \geq 1$ or piecewise C^1-diffeomorphisms if $r = 0$.

(B) When $p = \infty$, the following two conditions hold: (i) all elements of $\mathcal{G}$ are locally $C^{r-1,1}$ if $r \geq 1$ or locally L^∞ if $r = 0$ and (ii) their inverse image of a nullset is again a nullset.

Then, the following statements are equivalent:

1) $\mathrm{INN}_{\mathcal{G}}$ *is a $W^{r,p}$-universal approximator for $\mathcal{D}^{\max(r,1)}$,*

2) $\mathrm{INN}_{\mathcal{G}}$ *is a $W^{r,p}$-universal approximator for Ξ^∞,*

3) $\mathrm{INN}_{\mathcal{G}}$ *is a $W^{r,p}$-universal approximator for $\mathcal{T}^\infty$, and*

4) $\mathrm{INN}_{\mathcal{G}}$ *is a $W^{r,p}$-universal approximator for $\mathcal{S}_{\mathrm{c}}^\infty$.*

As for the sup-universality (Remark 13), we have a similar result:

Theorem 25 *Suppose the assumptions of Theorem 24 hold. In addition, suppose that $r = 0$ and that $\mathcal{G}$ consists of locally bounded measurable mappings. Then, the equivalence in Theorem 24 is valid if we replace "$W^{r,p}$-" with "*sup-*".*

For the definitions of the piecewise C^r-diffeomorphisms, locally $C^{r-1,1}$, and locally L^∞, see Appendix A. The regularity conditions in (A) and (B) assure that the functional composition within $\mathcal{G}$ is compatible with approximations (see Appendix B for details). These conditions are usually satisfied. The key step of the proof of this theorem is a decomposition of f into flow endpoints, which is realized by relying on a structure theorem of $\mathrm{Diff}_{\mathrm{c}}^\infty$ (Fact 30 in Appendix D) attributed to Herman (1973), Thurston (1974), Epstein (1970), and Mather (1974, 1975).

If we impose some restriction on the dimension d, we have stronger results:

Theorem 26 (Higher dimensional case) *The notation is as in Theorem 24.*

1. *Under (A), if $r = 0$ and $d \geq 2$, the statements 1)–4) are equivalent to the statement:*
   - $\mathrm{INN}_{\mathcal{G}}$ *is an L^p-universal approximator for $C^0(U, \mathbb{R}^d)$ for any open subset of $U \subset \mathbb{R}^d$ (U is not necessarily homeomorphic to $\mathbb{R}^d$).*
2. *Under (B), if $r = 0$ and $d \geq 7$, the statements 1)–4) are equivalent to the statement:*
   - $\mathrm{INN}_{\mathcal{G}}$ *is an L^∞-universal approximator for $\mathcal{D}^0$.*

In parallel to Theorem 25, we have a similar result for the sup-universality.

Theorem 27 *Suppose the assumptions of Theorem 26 hold. In addition, suppose that $r = 0$ and that $\mathcal{G}$ consists of locally bounded measurable mappings. Then, the equivalence in the statement 2 in Theorem 26 is valid if we replace "$W^{r,p}$-" with "*sup-*".*

We provide the proof of Theorem 26 and Theorem 27 in Appendix D.4.

Theorem 25 and Theorem 27 strengthen Theorem 1 in Teshima et al. (2020a) which provides the equivalence of the universality between $\mathcal{S}_{\mathrm{c}}^\infty$ and $\mathcal{D}^2$ instead of $\mathcal{D}^1$ (or $\mathcal{D}^0$ if $d \geq 7$).

3.2 Proof of Theorem 24 and Theorem 25

Theorem 24 and Theorem 25 are the consequence of Lemma 28, Lemma 29, Lemma 31, Lemma 33, and Lemma 36 we below show. More precisely, the proof is carried out by decomposing an approximation in $\mathcal{D}_U^\infty$ of $f : U \to \mathbb{R}^d$ into simpler functions step by step:

$$
\begin{array}{lll}
 & f & \\
 & \wr\!\!\!\sim \text{Lemma 28 (§3.2.1)} & \\
 & \tilde{f} \in \mathcal{D}_U^\infty & \\
 & \| \text{Lemma 29; the equality holds on } K \text{(§3.2.2)} & \\
W & \circ \quad h & (W \in \mathrm{Aff}, \quad h \in \mathrm{Diff}_{\mathrm{c}}^\infty) \\
 & \| \text{Lemma 31 (§3.2.3)} & \\
g_1 & \circ \cdots \circ \quad g_\ell & (g_1, \dots g_\ell \in \Xi^\infty) \\
 & \| \text{Lemma 33 (§3.2.4)} & \\
 & h_1 \quad \circ \cdots \circ \quad h_m & (h_1, \dots, h_m \text{: near-Id's}) \\
 & \| \text{Lemma 36 (§3.2.5)} & \\
 & \tau_1 \circ \sigma_1 \quad \circ \cdots \circ \quad \tau_n \circ \sigma_n & (\tau_1, \dots, \tau_n \in \mathfrak{S}_d, \quad \sigma_1, \dots, \sigma_n \in \mathcal{S}_{\mathrm{c}}^\infty)
\end{array}
$$

To have the diagram above work, we need one more proposition to ensure that the approximation of composite functions is the composition of the approximated functions (Proposition 37). See Section 3.2.7 for the complete proof. The proofs of some of the auxiliary lemmas are in Appendix D and Proposition 37 in Appendix B.

3.2.1 Approximation of $\mathcal{D}^{\max(r,1)}$ by $\mathcal{D}^\infty$

This reduction is a direct consequence of the following lemma:

Lemma 28 *For any $r \geq 1$, any $1 \leq p \leq \infty$, and any open subset $U \subset \mathbb{R}^d$, $\mathcal{D}_U^\infty$ is a $W^{r,p}$-universal approximator for $\mathcal{D}_U^r$.*

Proof Since the universal approximation property only considers approximation on compact sets by definition, $W^{r,\infty}$-universal approximation implies $W^{r,p}$-universal approximation for any $p \geq 1$. Therefore, it follows Hirsch (1976, Theorem 2.7, p.50). ■

3.2.2 From $\mathcal{D}^\infty$ to $\mathrm{Diff}_{\mathrm{c}}^\infty$

Let $f \in \mathcal{D}^\infty$. If we fix a compact set K of the domain U of f, we can find a compactly supported diffeomorphism identical with f on K: there exists $h \in \mathrm{Diff}_{\mathrm{c}}^\infty$ and $W \in \mathrm{Aff}$ such that

$$f|_K = W \circ h|_K. \tag{2}$$

It follows from the following lemma in the case of $r = \infty$:

Lemma 29 *Assume $r \geq 2$. Let $U \subset \mathbb{R}^d$ be an open set C^r-diffeomorphic to $\mathbb{R}^d$, $K \subset U$ a compact set, and $f \in \mathcal{D}^r_U$. Then, there exist $h \in \mathrm{Diff}^r_{\mathrm{c}}$ and an affine transform $W \in \mathrm{Aff}$ such that $W \circ h|_K = f|_K$.*

Proof See Appendix D.1. ■

3.2.3 From $\mathrm{Diff}^\infty_{\mathrm{c}}$ to Ξ^∞

The set $\mathrm{Diff}^r_{\mathrm{c}}$ constitutes a group whose group operation is the function composition. Moreover, $\mathrm{Diff}^r_{\mathrm{c}}$ is a topological group with respect to the *Whitney topology* (Haller, 1995, Proposition 1.7.(9)). Then there is a crucial structure theorem of $\mathrm{Diff}^r_{\mathrm{c}}$ attributed to Herman, Thurston (Thurston, 1974), Epstein (Epstein, 1970), and Mather (Mather, 1974, 1975):

Fact 30 *Assume $1 \leq r \leq \infty$ and $r \neq d+1$. Then, the group $\mathrm{Diff}^r_{\mathrm{c}}$ is simple, that is, any normal subgroup $H \subset \mathrm{Diff}^r_{\mathrm{c}}$ is either $\{\mathrm{Id}\}$ or $\mathrm{Diff}^r_{\mathrm{c}}$.*

The assertion is proven in Mather (1975) for the connected component containing Id, instead of the entire set of compactly-supported C^r-diffeomorphisms when the domain space is a general manifold instead of $\mathbb{R}^d$. In the special case of $\mathbb{R}^d$, the connected component containing Id is known to be $\mathrm{Diff}^r_{\mathrm{c}}$ itself (Haller, 1995, Example 1.15), hence Fact 30 follows. For details, see (Haller, 1995, Corollary 3.5 and Example 1.15). Also, Banyaga (1997) is an introductory monograph that explains the simplicity of $\mathrm{Diff}^\infty_{\mathrm{c}}$.

We use Fact 30 to prove that a compactly supported diffeomorphism can be represented as a composition of flow endpoints in $\mathrm{Diff}^r_{\mathrm{c}}$. Thanks to Lemma 28, we only consider $\mathrm{Diff}^\infty_{\mathrm{c}}$, and thus we do not need to consider the condition $r \neq d+1$ in Fact 30.

Lemma 31 *If $s \neq d+1$, the set of compactly supported diffeomorphisms $\mathrm{Diff}^s_{\mathrm{c}}$ coincides with the set of finite compositions of the elements of Ξ^r. More specifically, we have*

$$\mathrm{Diff}^r_{\mathrm{c}} = \{g_1 \circ \cdots \circ g_n : n \geq 1, g_1, \ldots, g_n \in \Xi^r\}.$$

Proof See Appendix D.2. ■

3.2.4 From Ξ^∞ to near-Id's

First, we provide the definition of near-Id's.

Definition 32 (near-Id elements) *Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be a differentiable map. We say f is* near-Id *if, for any $x \in \mathbb{R}^d$, the Jacobian Df of f at x satisfies*

$$\|Df(x) - I\|_{\mathrm{op}} < 1,$$

where I is the unit matrix.

Then, the decomposition from flow endpoints to near-Id's follows from the following lemma:

Lemma 33 *Let $r \geq 1$. For any $f \in \Xi^r$, there exist finite elements $g_1, \ldots, g_k \in \mathrm{Diff}^r_{\mathrm{c}}$ such that $f = g_k \circ \cdots \circ g_1$ and g_i is C^r-near-Id for any $i \in [k]$.*

Proof Let Φ be a flow associated with f. Since $\Phi(\cdot, 0)$ is the identity function and Φ is continuous on $\mathbb{R}^d \times U$, we can take a sufficiently large n such that $\tilde{h} := \Phi(\cdot, 1/n)$ is near-Id. By the additive property of Φ, we have

$$f = \underbrace{\tilde{h} \circ \cdots \circ \tilde{h}}_{n \text{ times}},$$

which completes the proof. ■

3.2.5 From Near-Id's to $\mathcal{S}_{\mathrm{c}}^{\infty}$

First, we introduce two elementary lemmas.

Lemma 34 *Let $A = (a_{i,j})_{i,j=1,\dots,d}$ be a matrix. If $\|A - I_d\|_{\mathrm{op}} < 1$, then for $k = 1, \dots, d$, the k-th trailing principal submatrix $A_k := (a_{i+k-1,j+k-1})_{i,j=1,\dots,d-(k-1)}$ of A is invertible. Here I_d is a unit matrix of degree d.*

Proof Let $v \in \mathbb{R}^{d-k+1}$ with $\|v\| = 1$, and put $w := (0, \dots, 0, v) \in \mathbb{R}^d$. Then we have $1 > \|(A - I_d)w\|^2 \geq \|(A_k - I_k)v\|^2$. Thus $\|A_k - I_k\| < 1$. Since $\sum_{r=0}^{\infty}(I_k - A_k)^r$ absolutely converges, and it is identical to the inverse of A_k, we have that A_k is invertible. ■

For $a \in \mathbb{N}$, we denote the set of a-by-a real-valued matrices by $M(a, \mathbb{R})$.

Lemma 35 *Let $1 \leq r \leq \infty$ and $f \colon \mathbb{R}^d \to \mathbb{R}^d$ a compactly supported C^r-diffeomorphism. We write $f = (f_1, \dots, f_d)$ with $f_i \colon \mathbb{R}^d \to \mathbb{R}$. For $k \in [d]$, let $\Delta_k^f(\boldsymbol{x}) \in M(d - (k-1), \mathbb{R})$ be the k-th trailing principal submatrix of the Jacobian matrix of f, whose (i, j) component is given by $\left(\frac{\partial f_{i+k-1}}{\partial x_{j+k-1}}(\boldsymbol{x})\right)$ $(i, j = 1, \cdots, d - (k-1))$. We assume*

$$\det \Delta_k^f(x) \neq 0 \text{ for any } k \in [d] \text{ and } x \in \mathbb{R}^d.$$

Then there exist compactly supported C^r-diffeomorphisms $F_1, \dots, F_d : \mathbb{R}^d \to \mathbb{R}^d$ in the forms of

$$F_i(\boldsymbol{x}) := (x_1, \dots, x_{i-1}, h_i(\boldsymbol{x}), x_{i+1}, \dots, x_d)$$

for some $h_i \colon \mathbb{R}^d \to \mathbb{R}$ such that the identity holds:

$$f = F_1 \circ \cdots \circ F_d.$$

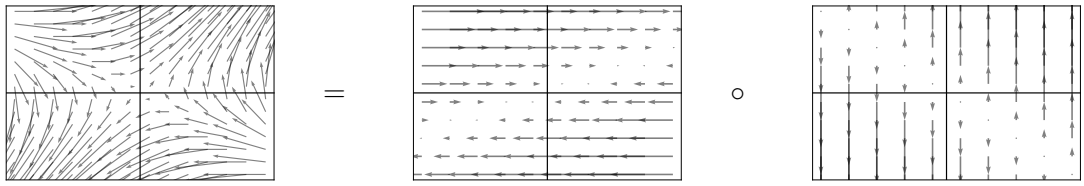
Proof See Appendix D.3 (see also Figure 1). ■

Now, we apply the Lemma 35 together with Lemma 34 to decompose near-Id elements into $\mathcal{S}_{\mathrm{c}}^r$ and permutations as follows:

Lemma 36 *Let $1 \leq r \leq \infty$. Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be a compactly supported C^r-near-Id map. Then there exist $\tau_1, \dots, \tau_n \in \mathcal{S}_{\mathrm{c}}^r$, and permutations of variables $\sigma_1, \dots, \sigma_n \in \mathfrak{S}_d$, such that*

$$f = \tau_1 \circ \sigma_1 \circ \cdots \circ \tau_n \circ \sigma_n.$$

Proof Combining Lemma 34, and Lemma 35 below, we have the assertion. ■



$$f = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = f_2 \circ f_1 \qquad f_2\begin{pmatrix} x_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} ax_1 + b(\frac{y_2 - cx_1}{d}) \\ y_2 \end{pmatrix} \qquad f_1\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ cx_1 + dx_2 \end{pmatrix}$$

Figure 1: An illustrative example for Lemma 35. This example shows the decomposition of a near-Id transformation that is linear on the unit square (denoted by f) into coordinate-wise ones (f_1 and f_2). The arrows indicate the transportation of the positions. The figure is taken from Teshima et al. (2020a, Figure 1) with the authors' permission.

3.2.6 Compatibility of compositions and approximations

Now, we provide a general result of the compatibility of composition and approximation. It enables the component-wise approximation, that is, approximating a composition of some transformations by approximating each constituent and composing them. The justification of this procedure is not trivial and requires a fine mathematical argument.

Proposition 37 *Let $r \geq 1$ and $p \in [1, \infty]$. Let $\mathcal{G}$ be the set of $\mathbb{R}^d$-valued mappings and $\mathcal{G}_0$ be a subset of $\mathcal{G}$. Assume either of the following conditions:*

1. *$1 \leq p \leq \infty$, $\mathcal{G}$ is composed of C^r and piecewise C^{r+1} diffeomorphisms on $\mathbb{R}^d$, and $\mathcal{G}_0$ is the subset composed of linearly increasing mappings.*
2. *$p = \infty$, $\mathcal{G}$ is composed of locally $C^{r-1,1}$-mappings whose inverse image of nullsets are again nullsets, and $\mathcal{G}_0$ is C^r-mappings.*

Then, for any $k \geq 1$, the map

$$\mathcal{G}^k \longrightarrow \mathcal{G}; (f_1, \ldots, f_k) \mapsto f_1 \circ \cdots \circ f_k \tag{3}$$

is continuous at any point of $\mathcal{G}_0^k$ with respect to the relative topology of $W^{r,p}_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^d)^k$. If $\mathcal{G} \subset B_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^d)$ and the subset $\mathcal{G}_0 \subset \mathcal{G}$ is composed of continuous mapping, we have the same continuity result of the composition with respect to the topology of $B_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^d)^k$.

Proof See Appendix B. ■

3.2.7 Overall Proof

Proof [Proof of Theorem 24 and 25] First, we prove the equivalence of statements 1) and 2). Let $f \in \mathcal{D}_U^{\max\{1,r\}}$ and fix compact subset $K \subset U_f$. By Lemma 28, we may assume $f \in \mathcal{D}_U^{\infty}$. Thus, in light of Lemma 29 and Lemma 31, there exist $W \in \mathrm{Aff}$ and $g_1, \ldots, g_m \in \Xi^{\infty}$ such

that $f(x) = W \circ g_1 \circ \cdots \circ g_m(x)$ for all $x \in K$. Since W and g_i's satisfy the condition to apply Proposition 37, and are linearly increasing (see Remark A.6), we obtain the equivalence of statements 1) and 2).

Next, we prove the equivalence of statements 1), 3), and 4). Since we have $\mathcal{S}_{\mathrm{c}}^{\infty} \subset \mathcal{T}^{\infty} \subset \mathcal{D}_{\mathbb{R}^d}^{\max\{1,r\}}$, it is sufficient to prove that the universal approximation property for $\mathcal{S}_{\mathrm{c}}^{\infty}$ implies that for $\mathcal{D}_{U}^{\max\{1,r\}}$ for any open subset $U \subset \mathbb{R}^d$ which is C^{∞}-diffeomorphic to $\mathbb{R}^d$. The strategy is similar to the flow endpoint case in the previous paragraph. By Lemma 28, we may assume $f \in \mathcal{D}_{U}^{\infty}$. Using Lemma 36 and Lemma 31, for any $f \in \mathcal{D}_{U}^{\infty}$ and a compact subset $K \subset U_f$, there exist $W_1, \ldots, W_k \in \mathrm{Aff}$ and $\tau_1, \ldots, \tau_k \in \mathcal{S}_{\mathrm{c}}^{\infty}$ such that $f(x) = W_1 \circ \tau_1 \circ \cdots \circ W_k \circ \tau_k(x)$ for all $x \in K$. Again, we use Proposition 37 to prove the claim.

■

3.3 Implications of the Main Theorem for Distributional Universality

Next, we give two consequences of Theorem 24 (namely, Corollary 39 and Corollary 41). We first note the relationship between functional universality (Definition 8) and distributional universality (Definition 14).

Proposition 38 *Let $p \in [1, \infty]$. An L^p-universal approximator for $\mathcal{T}^{\infty}$ is a $(\mathcal{P}^{\mathrm{w}}, \nu)$-distributional universal approximator for $\mathcal{P}$ for any $\nu \in \mathcal{P}_{\mathrm{ab}}$*

The proof is based on the existence of a triangular map connecting two absolutely continuous distributions (Bogachev et al., 2005). See Appendix C.1 for details. Note that the previous studies (Jaini et al., 2019; Huang et al., 2018) have discussed the distributional universality of some flow architectures essentially via showing the sup-universality for $\mathcal{T}^{\infty}$. Proposition 38 clarifies that the weaker notion of L^p-universality is sufficient for the distributional universality since sup-universality implies L^p-universality.

Proposition 38 can be combined with both cases of (A) and (B) in Theorem 24, namely, we have the following corollary:

Corollary 39 (Sobolev universality implies weak topology universality)
Notations and assumptions are as in Theorem 24. Then, if $\mathrm{INN}_{\mathcal{G}}$ is a $W^{r,p}$-universal approximator for $\mathcal{S}_{\mathrm{c}}^{\infty}$, then it is a $(\mathcal{P}^{\mathrm{w}}, \nu)$-distributional universal approximator for $\mathcal{P}$ for any $\nu \in \mathcal{P}_{\mathrm{ab}}$.

If the model can also universally approximate the derivatives, then it is guaranteed to have a stronger distributional universality in terms of the total variation distance, as we see in the following proposition:

Proposition 40 *Let $r \geq 1$. Let $\mathcal{F}_0 := W_{\mathrm{loc}}^{0,\infty}(\mathbb{R}^d, \mathbb{R}^d) \cap W_{\mathrm{loc}}^{1,1}(\mathbb{R}^d, \mathbb{R}^d)$, where we define the topology $\mathcal{F}_0$ to be the weakest topology such that the inclusion maps $\imath_0 : \mathcal{F}_0 \hookrightarrow W_{\mathrm{loc}}^{0,\infty}(\mathbb{R}^d, \mathbb{R}^d)$ and $\imath_1 : \mathcal{F}_0 \hookrightarrow W_{\mathrm{loc}}^{1,1}(\mathbb{R}^d, \mathbb{R}^d)$ are both continuous. Suppose any element in model $\mathcal{M}$ is locally $C^{0,1}$ and a piecewise C^1-diffeomorphism. If $\mathcal{M}$ is an $\mathcal{F}_0$-universal approximator for $\mathcal{T}^{\infty}$, then $\mathcal{M}$ is a $(\mathcal{P}^{\mathrm{TV}}, \nu)$-distributional universal approximator for $\mathcal{P}_{\mathrm{ab}}$ for any $\nu \in \mathcal{P}_{\mathrm{ab}}$.*

Since $W^{1,\infty}_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^d)$ is continuously included in the space $\mathcal{F}_0$ defined in Proposition 40, we immediately have

Corollary 41 (Sobolev universality implies total variation universality) *Notation is the same as Theorem 24. Assume that any element of $\mathcal{G}$ is locally $C^{0,1}$ and a piecewise C^1-diffeomorphism. Then, if $\mathrm{INN}_{\mathcal{G}}$ is a $W^{1,\infty}$-universal approximator for $\mathcal{S}^{\infty}_{\mathrm{c}}$, then so is a $(\mathcal{P}^{\mathrm{TV}}, \nu)$-distributional universal approximator for $\mathcal{P}_{\mathrm{ab}}$ for any $\nu \in \mathcal{P}_{\mathrm{ab}}$.*

We defer their proofs to Appendix C.2.

4. Application of the General Framework

In this section, we show several crucial results for the universalities of INNs with certain flow layers. The results are proved by using the general framework developed in Section 3.

4.1 Affine Coupling Flows (ACFs)

Here, we reveal the L^p-universality of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$. This result reframes that of Teshima et al. (2020a), which answered a previously unsolved question for the distributional universality of ACF-based invertible neural networks. In this subsection, we always assume $d \geq 2$ since CF-INNs are only defined for $d \geq 2$ (see Section 2.1.1).

Theorem 42 (L^p-universality of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$) *Let $p \in [1, \infty)$. Assume that $\mathcal{H}$ is an L^∞-universal approximator for $C^0(\mathbb{R}^{d-1})$ and that it consists of piecewise C^1-functions. Then, $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$ is an L^p-universal approximator for $C^0(U, \mathbb{R}^d)$ for any open subset $U \subset \mathbb{R}^d$.*

We remark that the universality is still valid if we restrict the affine layers of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$ to elements in $\mathfrak{S}_d$, i.e., the permutations of variables. For the definition of piecewise C^1-functions, see Appendix A. We provide the proof of Theorem 42 by combining Theorem 24 with a slightly general result, which is an L^p-universal approximation property of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$ for $\mathcal{S}^0_{\mathrm{c}}$, in Appendix E.2. Examples of $\mathcal{H}$ satisfying the condition of Theorem 42 include MLP models with ReLU activation (LeCun et al., 2015) and a linear-in-parameter model with smooth universal kernels (Micchelli et al., 2006).

By combining Theorem 24, Theorem 42, and Proposition 38, we can affirmatively answer a previously unsolved problem (Papamakarios et al., 2021, p.13), the distributional universality of CF-INN based on ACFs, and we can confirm the theoretical plausibility of using them for normalizing flows.

Theorem 43 (Distributional universality of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$) *Under the conditions of Theorem 42, $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$ is a $(\mathcal{P}^{\mathrm{w}}, \nu)$-distributional universal approximator for $\mathcal{P}$ for any $\nu \in \mathcal{P}_{\mathrm{ab}}$.*

4.2 Neural Ordinary Differential Equations (NODEs)

The following shows that the INNs based on NODEs can approximate diffeomorphisms with respect to the $W^{r,\infty}$-norm. We denote by $\mathrm{Lip} \cap C^r$ the space of Lipschitz and C^r maps from $\mathbb{R}^d$ to $\mathbb{R}^d$ and we equip it with the relative topology of $W^{r,\infty}_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^d)$.

Theorem 44 (Universality of NODEs) *Let $r \geq 0$. Assume $\mathcal{H} \subset \mathrm{Lip} \cap C^r$ is a $W^{r,\infty}$-universal approximator for $\mathrm{Lip} \cap C^r$. Then, $\mathrm{INN}_{\Psi(\mathcal{H})}$ is a $W^{r,\infty}$-universal approximator for $\mathcal{D}^{\max(r,1)}$.*

Theorem 44 is shown by applying Theorem 24 in combination with Lemma 29 (Appendix D.1) to approximate the elements of Ξ^∞ by NODEs. A proof is in Appendix F. We remark that the universality in this theorem still holds if we restrict the affine layers of $\mathrm{INN}_{\Psi(\mathcal{H})}$ to identity except the last one, which is denoted by W_1 in Definition 6 (see Proposition F.2. Examples of $\mathcal{H}$ include the MLP with finite weights and Lipschitz-continuous activation functions such as ReLU activation (LeCun et al., 2015; Chen et al., 2018), as well as the *Lipschitz Networks* (Anil et al., 2019, Theorem 3).

4.3 Sum-of-Squares Polynomial Flows (SoS Flows)

The sum-of-squares polynomial flow (SoS flow) (Jaini et al., 2019) is an important example of the flow layer for INNs (see also Section E.4). Here, we consider a special class of SoS flow layers $\mathcal{H}$-SoS where only the last dimension is converted (for the general description of SoS flow layers, see Section E.4).

Definition 45 *Let $\mathcal{H}$ be a set of measurable functions on $\mathbb{R}^{d-1}$. For $c \in \mathbb{R}$ and $h_1, \dots, h_k \in \mathcal{H}$, let*

$$g(\boldsymbol{x}; c, h_1, \dots, h_k) := c + \int_0^{x_d} \sum_{l=0}^{k} h_l(\boldsymbol{x}_{\leq d-1}) u^l du.$$

Then, we define $\mathcal{H}$-SoS to be the set of all maps of the form $\boldsymbol{x} \mapsto (\boldsymbol{x}_{\leq d-1}, g(\boldsymbol{x}; c, h_1, \dots, h_k))$ where $k \geq 1$, $c \in \mathbb{R}$, and $h_1, \dots, h_k \in \mathcal{H}$.

Although the universality for SoS based INN was proved in Jaini et al. (2019), we prove a much stronger universality for the architecture (Proposition E.10):

Theorem 46 *Let $r \geq 0$ and let $\mathcal{H}$ be a set of measurable functions on $\mathbb{R}^{d-1}$. Assume that all elements of $\mathcal{H}$ are locally $C^{r-1,1}$ if $r \geq 1$ or locally L^∞ if $r = 0$ and that $\mathcal{H}$ is a $W^{r,\infty}$-universal approximator for the set of $(d-1)$-variable polynomials. Then, $\mathrm{INN}_{\mathcal{H}\text{-SoS}}$ is a $W^{r,\infty}$-universal approximator for $\mathcal{D}^{\max(r,1)}$.*

This theorem immediately follows from Proposition E.10 and Theorem 24. As a direct corollary of Theorem 46, Corollary 39, and Proposition 40, we have the following.

Corollary 47 *Let us use the same notation as in Theorem 46. Then, $\mathrm{INN}_{\mathcal{H}\text{-SoS}}$ is a $(\mathcal{P}^{\mathrm{w}}, \nu)$-distributional universal approximator for $\mathcal{P}$ for any $\nu \in \mathcal{P}_{\mathrm{ab}}$. Moreover, if $r \geq 1$, $\mathrm{INN}_{\mathcal{H}\text{-SoS}}$ is a $(\mathcal{P}^{\mathrm{TV}}, \nu)$-distributional universal approximator for $\mathcal{P}_{\mathrm{ab}}$ for any $\nu \in \mathcal{P}_{\mathrm{ab}}$.*

4.4 Other Examples of Flow Layers

Theorem 42 can be interpreted as providing a convenient criterion to check the universality of a CF-INN: if the flow architecture $\mathcal{G}$ contains ACFs (or even just $\mathcal{H}$-ACF with sufficiently expressive $\mathcal{H}$) as special cases, then $\mathrm{INN}_{\mathcal{G}}$ is an L^p-universal approximator for $C^0(U, \mathbb{R}^d)$ for any open subset $U \subset \mathbb{R}^d$. Such examples of $\mathcal{G}$ include the *nonlinear squared flow* (Ziegler

and Rush, 2019), *Flow++* (Ho et al., 2019), and the *neural autoregressive flow* (Huang et al., 2018).

The result may not immediately apply to the typical *Glow* (Kingma and Dhariwal, 2018) architecture for image data that uses the 1x1 invertible convolution layers and convolutional neural networks for the coupling layers. However, the Glow architecture for non-image data (Ardizzone et al., 2019; Teshima et al., 2020b) can also be interpreted as $\mathrm{INN}_{\mathcal{G}}$ with ACF layers, and hence it is an L^p-universal approximator for $C^0(U, \mathbb{R}^d)$ for any open subset $U \subset \mathbb{R}^d$.

5. Integral Probability Metrics

Proposition 40 implies the universality of INNs with respect to the total variation (TV) topology. Here, we consider how the theoretical guarantees in the TV topology can be transported to other notions of closeness, namely those of integral probability metrics (IPMs).

We say a measurable set $A \subset \mathbb{R}^n$ is a *continuity set* of a measure μ if the boundary ∂A of A is a null set, that is, $\mu(\partial A) = 0$. We say a measurable set $A \subset \mathbb{R}^n$ is a *non-null set* of a measure μ if $\mu(A) \neq 0$. For any measurable subset $K \subset \mathbb{R}^n$ and any probability measure η on $\mathbb{R}^n$, let us define the truncated measure $\eta|_K := \eta(\cdot \cap K)/\eta(K)$ if $\eta(K) > 0$ and $\eta|_K := \mathbf{0}$ if $\eta(K) = 0$, where $\mathbf{0}$ is a constant zero measure. To state the results, we define the following notion of universality.

Definition 48 (Compact distributional universality) *Let $\mathcal{M}$ be a model which is a set of measurable maps from $\mathbb{R}^m$ to $\mathbb{R}^n$. Let $\mathcal{P}_0$ be a set of probability measures on $\mathbb{R}^n$ with some topology. Let $\mathcal{Q}$ be a subset of $\mathcal{P}_0$. Fix a probability measure μ_0 on $\mathbb{R}^m$. We say that a model $\mathcal{M}$ is a* $(\mathcal{P}_0, \mu_0)$-compact-distributional universal approximator *for $\mathcal{Q}$ (or* has the $(\mathcal{P}_0, \mu_0)$-compact-distributional universal approximation property *for $\mathcal{Q}$) if for any $\nu \in \mathcal{Q}$ and any non-null compact continuity set $K \subset \mathbb{R}^n$ of ν, $\{(g_*\mu_0)|_K : g \in \mathcal{M}\} \setminus \{\mathbf{0}\}$ is a subset of $\mathcal{P}_0$ and if its closure (in $\mathcal{P}_0$) contains $\nu|_K$.*

Note that if ν is compactly supported and K is such that $\operatorname{supp} \nu \subset K^\circ$, where K° denotes the interior of K, then K is a continuity set of ν. Also, in this case, $\nu|_K = \nu$. Therefore, practically, given a compact distributional universality of a model $\mathcal{M}$ and a compactly supported approximation target $\nu \in \mathcal{Q}$, one can regard it as an approximation guarantee for ν by taking a sufficiently large K so that it covers any practically relevant range of values as well as $\operatorname{supp} \nu$.

Remark 49 *Let $\mathcal{P}_0$ be a set of probability measures on $\mathbb{R}^n$ with some topology. For $\mu \in \mathcal{P}_0$, a compact continuity set K of μ, and a neighborhood V of $\mu|_K$ with $\mu|_K \neq \mathbf{0}$, we define*

$$W_\mu(K, V) := \{\nu \in \mathcal{P}_0 : \nu|_K \in V\}.$$

We define a new topology of $\mathcal{P}_0$ via the neighborhoods of μ's by those generated by $W_\mu(K, V)$'s. We denote by $\mathcal{P}_0^\tau$ the set $\mathcal{P}_0$ equipped with the topology above. By definition, the truncation $\cdot|_K : \mathcal{P}_0^\tau \to \mathcal{P}_0 \cup \{\mathbf{0}\}$ for any compact continuity set of μ is continuous at any μ satisfying $\mu|_K \neq \mathbf{0}$, where the topology of $\mathcal{P}_0 \cup \{\mathbf{0}\}$ is the direct sum topology. Conversely, $\mathcal{P}_0^\tau$ is characterized as the set $\mathcal{P}_0$ equipped with the weakest topology such that the above truncations

are continuous. If we impose that the topology of $\mathcal{P}_0$ is stronger than $\mathcal{P}_0^\tau$, namely the truncation $\cdot|_K$ is continuous at μ for any continuity set K of μ with respect to the topology of $\mathcal{P}_0$. Under this assumption, the compact distributional universality in Definition 48 is rephrased as the $(\mathcal{P}_0^\tau, \mu_0)$-distributional universality for $\mathcal{Q}$. Moreover, we may immediately prove that $(\mathcal{P}_0, \mu_0)$-distributional universality implies compact distributional universality. In the case of $\mathcal{P}_0 = \mathcal{P}^{\mathrm{w}}$, thanks to the portmanteau lemma, we may prove that the topology of $\mathcal{P}_0$ is stronger than $\mathcal{P}_0^\tau$, namely the truncation $\cdot|_K$ is continuous at μ for any continuity set K of μ.

IPMs are defined as follows.

Definition 50 (Integral probability metric; Müller 1997) *Let $\mathcal{X}$ be a measurable space, μ and ν be probability measures on $\mathcal{X}$, and $\mathcal{F}$ be $\mathbb{R}$-valued bounded measurable functions on $\mathcal{X}$. Then, the* integral probability metric *(IPM) based on $\mathcal{F}$ is defined as*

$$\mathrm{IPM}_{\mathcal{F}}(\mu, \nu) := \sup_{f \in \mathcal{F}} \left| \int_{\mathcal{X}} f d\mu - \int_{\mathcal{X}} f d\nu \right|$$

For a comprehensive review on IPMs, see, for example, Sriperumbudur et al. (2009).

By selecting appropriate $\mathcal{F}$, various distance measures in probability theory and statistics can be obtained as special cases of the IPM. In the following, assume that $\mathcal{X}$ is equipped with a distance metric ρ and that the σ-algebra is the Borel σ-algebra induced by the metric topology of ρ. Let $\|f\|_{\mathrm{Lip}} := \sup_{x,y \in \mathcal{X}, x \neq y} \frac{|f(x)-f(y)|}{\rho(x,y)}$ and $\|f\|_{\mathrm{BL}} := \|f\|_{\sup} + \|f\|_{\mathrm{Lip}}$. Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) induced by a positive semidefinite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and let $\|\cdot\|_{\mathcal{H}}$ be its RKHS norm.

Definition 51 (Sriperumbudur et al. 2009) *We define the following metrics.*

- Dudley metric*: $\mathcal{F}_{\mathrm{Dud}} = \{f : \|f\|_{\mathrm{BL}} \leq 1\}$ yields the Dudley metric* $\mathrm{IPM}_{\mathcal{F}_{\mathrm{Dud}}}(\mu, \nu)$.
- Wasserstein distance*: if $\mathcal{X}$ is separable, then $\mathcal{F}_{W_1} = \{f : \|f\|_{\mathrm{Lip}} \leq 1\}$ yields the* 1*-Wasserstein distance* $\mathrm{IPM}_{\mathcal{F}_{W_1}}(\mu, \nu)$ *for $\mu, \nu \in \mathcal{P}_{W_1} = \{\nu' : \int \rho(x, y) d\nu'(x) < \infty, \forall y \in \mathcal{X}\}$.*
- Total variation distance*: $\mathcal{F}_{\mathrm{TV}} = \{f : \|f\|_{\sup} \leq 1\}$ yields the total variation distance* $\mathrm{IPM}_{\mathcal{F}_{\mathrm{TV}}}(\mu, \nu)$.
- Maximum mean discrepancy *(MMD): selecting $\mathcal{F}_{\mathrm{MMD}} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$ yields the MMD* $\mathrm{IPM}_{\mathcal{F}_{\mathrm{MMD}}}(\mu, \nu)$.

We use $\mathcal{P}^{\mathrm{Dud}}$, $\mathcal{P}^{W_1}$, and $\mathcal{P}^{\mathrm{MMD}}$, to denote $\mathcal{P}$ equipped with the induced topology of $\mathrm{IPM}_{\mathcal{F}_{\mathrm{Dud}}}(\cdot, \cdot)$, $\mathrm{IPM}_{\mathcal{F}_{W_1}}(\cdot, \cdot)$, *and* $\mathrm{IPM}_{\mathcal{F}_{\mathrm{MMD}}}(\cdot, \cdot)$, *respectively.*

Note that, if $(\mathcal{X}, \rho)$ is separable (such as $\mathcal{X} = \mathbb{R}^d$), then the convergence in the Dudley metric is equivalent to the convergence in the weak topology (Dudley, 2002, Theorem 11.3.3.).

Remark 52 *If we interpret $\mathcal{F}$ in Definition 51 as a family of* statistics*, that is, functions that take random variables as the arguments, we can interpret an approximation guarantee in terms of an IPM as an approximation guarantee for the expectation of the statistics computed*

from these distributions. More concretely, once we obtain an approximation guarantee such as $\mathrm{IPM}_{\mathcal{F}}(\mu, \nu) < \varepsilon$ *where* ν *is an approximation target,* μ *is a model, and* $\varepsilon > 0$*, then we can deduce that* $|\mathbb{E}_{X\sim\mu}[f(X)] - \mathbb{E}_{Y\sim\nu}[f(Y)]| < \varepsilon$*, where* $\mathbb{E}$ *denotes the expectation, holds uniformly over the class of statistics* $f \in \mathcal{F}$*. If, moreover, we have a theoretical guarantee that* $|\int f d\mu - \sum_{i=1}^{N} f(X_i)| < \varepsilon'$ *for* $\{X_i\}_{i=1}^{N} \overset{i.i.d.}{\sim} \mu$*, where i.i.d. stands for* independently and identically distributed*, with high probability for some* $f \in \mathcal{F}$*, then we can combine these inequalities to provide an upper bound on* $|\sum_{i=1}^{N} f(X_i) - \mathbb{E}_{Y\sim\nu}[f(Y)]|$*, that is, the error of* Monte Carlo *approximation based on the samples generated by the model* μ *that approximated the target distribution* ν*.*

Depending on the IPM, we have different families of statistics, $\mathcal{F}$*, over which we can obtain such theoretical guarantees. In the case of the Dudley metric corresponding to the weak convergence topology, we can obtain such an approximation guarantee over the class of (uniformly) bounded and Lipschitz-continuous (and hence measurable) functions* f *with a uniformly bounded Lipschitz constant. In the case of the total variation, the guarantee is stronger, and we can obtain the guarantee over the class of (uniformly) bounded measurable functions* f*.*

We have the following elementary relations that can be easily shown from the definitions.

Proposition 53 *We have the following inequalities:*

$$\mathrm{IPM}_{\mathcal{F}_{\mathrm{Dud}}}(\mu, \nu) \leq \mathrm{IPM}_{\mathcal{F}_{\mathrm{TV}}}(\mu, \nu),$$
$$\mathrm{IPM}_{\mathcal{F}_{\mathrm{MMD}}}(\mu, \nu) \leq \left(\sup_{x\in\mathcal{X}} k(x, x)\right)^{\frac{1}{2}} \mathrm{IPM}_{\mathcal{F}_{\mathrm{TV}}}(\mu, \nu).$$

Proof The first inequality follows from $\mathcal{F}_{\mathrm{Dud}} \subset \mathcal{F}_{\mathrm{TV}}$, which holds by definition. The second inequality follows from the Cauchy-Schwarz inequality:

$$\|f\|_{\sup} = \sup_{x\in\mathcal{X}} |f(x)| = \sup_{x\in\mathcal{X}} |\langle f, k(x, \cdot)\rangle_{\mathcal{H}}| \leq \|f\|_{\mathcal{H}} \left(\sup_{x\in\mathcal{X}} k(x, x)\right)^{\frac{1}{2}},$$

where $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ denotes the inner product of $\mathcal{H}$. ■

We also have the following relation between the total variation distance and the 1-Wasserstein distance for $\mathcal{X} = \mathbb{R}^d$.

Lemma 54 *Let* $\mu, \nu \in \mathcal{P}$*, and let* K *be a compact non-null set of* ν*. If* $\mathrm{IPM}_{\mathcal{F}_{\mathrm{TV}}}(\mu, \nu) < \nu(K)$*, then*

$$\mathrm{IPM}_{\mathcal{F}_{W_1}}(\mu|_K, \nu|_K) \leq \frac{4 \cdot \mathrm{diam}(K)}{\nu(K)} \cdot \frac{\mathrm{IPM}_{\mathcal{F}_{\mathrm{TV}}}(\mu, \nu)}{\nu(K) - \mathrm{IPM}_{\mathcal{F}_{\mathrm{TV}}}(\mu, \nu)}, \tag{4}$$

where $\mathrm{diam}(K)$ *denotes the diameter of* K*.*

We defer the proof of Lemma 54 to the bottom part of this subsection, and we first display the following proposition to collect Corollary 53 and Lemma 54.

Proposition 55 *Let $\mathcal{Q} \subset \mathcal{P}$ and $\mu \in \mathcal{P}$. Assume that $\mathcal{M}$ is a $(\mathcal{P}^{\mathrm{TV}}, \mu)$-distributional universal approximator for $\mathcal{Q}$. Then, we have the following.*

(a) $\mathcal{M}$ is a $(\mathcal{P}^{\mathrm{Dud}}, \mu)$-distributional universal approximator for $\mathcal{Q}$,

(b) If $\sup_{x \in \mathbb{R}^d} k(x,x) < \infty$, then $\mathcal{M}$ is a $(\mathcal{P}^{\mathrm{MMD}}, \mu)$-distributional universal approximator for $\mathcal{Q}$,

(c) $\mathcal{M}$ is a $(\mathcal{P}^{W_1}, \mu)$-compact-distributional universal approximator for $\mathcal{Q}$.

The condition part of Proposition 55 is covered by the conclusion part of Theorem C.4, where $\mathcal{Q}$ and μ are arbitrary $\mathcal{Q} \subset \mathcal{P}_{\mathrm{ab}}$ and $\mu \in \mathcal{P}_{\mathrm{ab}}$. Therefore, we can immediately obtain the theoretical guarantee of distribution approximation using INNs with respect to these IPMs given a Sobolev universality of $\mathcal{M}$.

Proof [Proof of Proposition 55] The first two immediately follow from Corollary 53. The final assertion follows from Lemma 54. To show the final assertion, one needs to show that, for any $\nu \in \mathcal{Q}$, any non-null compact continuity set $K \subset \mathbb{R}^d$ of ν, and any $\varepsilon > 0$, there exists $g \in \mathcal{M}$ such that $\mathrm{IPM}_{\mathcal{F}_{W_1}}((g_*\mu)|_K, \nu|_K)$. By the assumption that $\mathcal{M}$ is a $(\mathcal{P}^{\mathrm{TV}}, \mu)$-distributional universal approximator for $\mathcal{Q}$, there exists $g \in \mathcal{M}$ such that both $\mathrm{IPM}_{\mathcal{F}_{\mathrm{TV}}}(g_*\mu, \nu) < \nu(K)$ and the right-hand side of Equation (4) in Lemma 54 is smaller than ε, so that $\mathrm{IPM}_{\mathcal{F}_{W_1}}((g_*\mu)|_K, \nu|_K) < \varepsilon$. ■

To prove Lemma 54, we use the following well-known inequality between the Wasserstein distance and the total variation distance.

Fact 56 (Villani, 2009, Theorem 6.15) *Let $(\mathcal{X}, \rho)$ be a separable complete metric space that is bounded with diameter R, and μ and ν be probability measures on $\mathcal{X}$. Then, we have* $\mathrm{IPM}_{\mathcal{F}_{W_1}}(\mu, \nu) \leq R \cdot \mathrm{IPM}_{\mathcal{F}_{\mathrm{TV}}}(\mu, \nu)$.

Lemma 54 is an immediate corollary of this fact. Note that

$$\mathrm{IPM}_{\mathcal{F}_{\mathrm{TV}}}(\mu, \nu) = 2 \sup_A |\mu(A) - \nu(A)|$$

holds, where $\sup_A$ denotes the supremum over all measurable subsets of the underlying space.

Proof [Proof of Lemma 54] Since $(K, \|\cdot\|)$ is a separable complete metric space, we have, by applying Fact 56 with $\mu|_K$ and $\nu|_K$,

$$\begin{aligned}
\mathrm{IPM}_{\mathcal{F}_{W_1}}(\mu|_K, \nu|_K) &= \sup_{f \in \mathcal{F}_{W_1}} \left| \int_{\mathbb{R}^d} f d(\mu|_K) - \int_{\mathbb{R}^d} f d(\nu|_K) \right| \\
&= \sup_{f \in \mathcal{F}_{W_1}|_K} \left| \int_K f d(\mu|_K) - \int_K f d(\nu|_K) \right| \\
&\leq \mathrm{diam}(K) \cdot 2 \cdot \sup_{A'} |(\mu|_K)(A') - (\nu|_K)(A')| =: (\mathrm{RHS}),
\end{aligned}$$

where $\sup_{A'}$ denotes the supremum over all measurable subsets of K, and $\mathcal{F}_{W_1}|_K := \{f|_K : f \in \mathcal{F}_{W_1}\}$. Now, since we have $\nu(K) - \mu(K) \leq |\mu(K) - \nu(K)| \leq \mathrm{IPM}_{\mathcal{F}_{\mathrm{TV}}}(\mu, \nu)$, we obtain

$\mu(K) \geq \nu(K) - \mathrm{IPM}_{\mathcal{F}_{\mathrm{TV}}}(\mu, \nu) > 0$. Thus, $\mu|_K(\cdot) = \mu(\cdot \cap K)/\mu(K)$, and hence the right-hand side (RHS) is further bounded as

$$\begin{aligned}(\mathrm{RHS}) &= 2 \cdot \mathrm{diam}(K) \sup_A |\mu(A \cap K)/\mu(K) - \nu(A \cap K)/\nu(K)| \\ &\leq 2 \cdot \mathrm{diam}(K) \sup_A |\mu(A)/\mu(K) - \nu(A)/\nu(K)|,\end{aligned}$$

where $\sup_A$ denotes the supremum over all measurable subsets of $\mathbb{R}^d$, and the inequality holds since $\sup_A$ runs through all the measurable subsets of the form $A \cap K$ as well. Now,

$$\begin{aligned}\left|\frac{\mu(A)}{\mu(K)} - \frac{\nu(A)}{\nu(K)}\right| &\leq \left|\frac{\mu(A)}{\mu(K)} - \frac{\nu(A)}{\mu(K)}\right| + \left|\frac{\nu(A)}{\mu(K)} - \frac{\nu(A)}{\nu(K)}\right| \\ &= \frac{|\mu(A) - \nu(A)|}{\mu(K)} + |\nu(K) - \mu(K)| \frac{\nu(A)}{\mu(K)\nu(K)} \\ &\leq \frac{\nu(K) + \nu(A)}{\mu(K)\nu(K)} \mathrm{IPM}_{\mathcal{F}_{\mathrm{TV}}}(\mu, \nu).\end{aligned}$$

Therefore, we have

$$\mathrm{IPM}_{\mathcal{F}_{W_1}}(g_*\mu|_K, \nu) \leq \frac{4 \cdot \mathrm{diam}(K)}{\nu(K)} \frac{\mathrm{IPM}_{\mathcal{F}_{\mathrm{TV}}}(g_*\mu, \nu)}{\nu(K) - \mathrm{IPM}_{\mathcal{F}_{\mathrm{TV}}}(g_*\mu, \nu)},$$

where we used $\mu(K) \geq \nu(K) - \mathrm{IPM}_{\mathcal{F}_{\mathrm{TV}}}(\mu, \nu) > 0$ and $\nu(K) + \nu(A) \leq 2$. ■

6. Conclusion

In this paper, we provided a general framework to analyze the theoretical representation power of a family of invertible function models. The key idea is to simplify the problem of approximating a general C^r-diffeomorphism by decomposing it into a finite set of simpler invertible maps by using the structure theorem of the diffeomorphism group.

The general framework was applied to two representative architectures of INNs: the CF-INNs and the NODEs, and we showed the high representation power of these architectures contrary to their apparent limitations on expressiveness.

For future work, it is important to quantitatively evaluate how many flow layers are required to approximate a given target map to assess the efficiency of the approximation. It includes exploring efficient approximation of well-behaved target functions (for example, the subset of $\mathcal{D}^1$ consisting of bi-Lipschitz diffeomorphisms). Also, comparing the approximation efficiency of different flow layer designs is an important issue. We expect that answering these questions provides principled design choices of invertible models tailored for a given task.

Acknowledgments

We would like to thank the action editor and two anonymous referees for their constructive comments that led to an improved version of the paper. We would also like to thank Prof. Taiji

Suzuki for his valuable comments and fruitful discussions on distributional universality. TT was supported by RIKEN Junior Research Associate Program and Masason Foundation. II and MI were supported by CREST: JPMJCR1913. II was supported by ACTX: JPMJAX2004. MS was supported by KAKENHI 20H04206.

The following is supplementary material for "Universal approximation property of invertible neural networks." We provide proof for the statements in the paper. Table 1 is the list of abbreviations we use in the paper. Tables 2 and 3 summarize the symbols we employed in the paper.

Abbreviation	Meaning
INN	Invertible neural network
CF-INN	Invertible neural network based on coupling flow
IAF	Inverse autoregressive flow
DSF	Deep sigmoidal flow
SoS	Sum-of-squares polynomial flow
MLP	Multi-layer perceptron
NODE	Neural ordinary differential equation

Table 1: Abbreviations in the paper

Notation	Meaning
$\mathbb{R}$	Set of all real numbers
$\mathbb{N}$	Set of all positive integers
$[n]$	Set $\{1, 2, \dots, n\}$
$\|\cdot\|$	Euclidean norm
$\|\cdot\|_{\mathrm{op}}$	Operator norm
$\|\cdot\|_{K,0,p}$	L^p-norm ($p \in [1, \infty)$) on a subset $K \subset \mathbb{R}^d$
$\mathbf{1}_A$	Indicator (characteristic) function of A
Id	Identity map
supp	Support of a map or measure
$Df(x)$	Jacobian matrix of f at x

Table 2: Notation table (part 1 of 2)

Appendix A. Locally bounded maps and piecewise diffeomorphisms

In this section, we provide the notions of locally-ness and piecewise-ness. These notions are used to state the regularity conditions on the invertible layers $\mathcal{G}$ in Theorem 24 and to prove the results in Section B.

A.1 Definition of locally-ness

Here, we provide the definition of "locally" for functions.

Definition A.1 (locally bounded maps) *Let* $\mathbf{P}$ *be a property of functions such as boundedness. Let* f *be a map from* $\mathbb{R}^m$ *to* $\mathbb{R}^n$*. We say* f *is* locally $\mathbf{P}$ *if for each point* $\boldsymbol{x} \in \mathbb{R}^m$*, there exists an open neighborhood* U *of* $\boldsymbol{x}$ *such that* f *has property* $\mathbf{P}$ *on* U*.*

Notation	Meaning
CF, $h_{k,\tau,\theta}$	Coupling flow
ACF, $\Psi_{k,s,t}$	Affine coupling flow
$\mathcal{H}$	Generic notation for a set of functions from $\mathbb{R}^{d-1}$ to $\mathbb{R}$
$\mathcal{H}$-ACF, $\Psi_{d-1,s,t}$	$\mathcal{H}$-single-coordinate affine coupling flows ($s,t \in \mathcal{H}$)
IVP$[f](\boldsymbol{x},t)$	The (unique) solution to an initial value problem evaluated at t
$\Psi(\mathcal{F})$	Set of NODEs obtained from the Lipschitz continuous vector fields $\mathcal{F}$
$\mathcal{G}$	Generic notation for a set of invertible functions
$\mathrm{INN}_{\mathcal{G}}$	Set of all invertible neural networks based on $\mathcal{G}$
$d \in \mathbb{N}$	Dimensionality of the input/output Euclidean space
$\ell \in \{0\} \cup \mathbb{N}$	Differentiability of the model
$\mathcal{D}^r$	Set of all C^r-diffeomorphisms with C^r-diffeomorphic domains
$\mathrm{Diff}^r_{\mathrm{c}}$ $(1 \leq r \leq \infty)$	Group of compactly-supported C^r-diffeomorphisms (on $\mathbb{R}^d$)
Ξ^r	Set of all flow endpoints in $\mathrm{Diff}^r_{\mathrm{c}}$
$\mathcal{T}^\infty$	Set of all C^∞-increasing triangular mappings
$\mathcal{S}^r_{\mathrm{c}}$	Set of all C^r-single-coordinate transformations
$\mathfrak{S}_d$	Set of all permutations of variables of $\mathbb{R}^d$
GL	Set of all regular real matrices of size d
Aff	Set of all affine transformations, that is, $\{\boldsymbol{x} \mapsto A\boldsymbol{x}+b : A \in \mathrm{GL}, b \in \mathbb{R}^d\}$
C^r	r-times continuously differentiable
$C^{r,\alpha}$	C^r and any k-th derivative with $\lvert k\rvert = r$ is α-Hölder continuous
$C^r(\mathbb{R}^m)$	Set of all C^r functions on $\mathbb{R}^m$ equipped with local Sobolev topology
$C^\infty_c(\mathbb{R}^d)$	Set of all compactly-supported C^∞ functions on $\mathbb{R}^d$
$B_{\mathrm{loc}}(\mathbb{R}^d,\mathbb{R}^m)$	Set of all locally bounded measurable maps from $\mathbb{R}^d$ to $\mathbb{R}^m$
$C^r(U,\mathbb{R}^n)$	Set of all $\mathbb{R}^n$-valued C^r maps on U
$W^{r,p}_{\mathrm{loc}}(U,\mathbb{R}^n)$	$\mathbb{R}^n$-valued local Sobolev space on U
$L^p_{\mathrm{loc}}(U,\mathbb{R}^n)$	$\mathbb{R}^n$-valued local Lebesgue space on U (equal to $W^{0,p}_{\mathrm{loc}}(U,\mathbb{R}^n)$)
Lip	Set of all Lipschitz continuous maps from $\mathbb{R}^d$ to $\mathbb{R}^d$
$\mathrm{Lip} \cap C^r$	Set of all Lipschitz and C^r maps from $\mathbb{R}^d$ to $\mathbb{R}^d$ with $W^{r,\infty}_{\mathrm{loc}}$-topology
$\mathcal{P}$	Set of all probability measures on $\mathbb{R}^d$
$\mathcal{P}_{\mathrm{ab}}$	Set of all absolutely continuous probability measures on $\mathbb{R}^d$
$\mathcal{P}^{\mathrm{w}}$	$\mathcal{P}$ equipped with the weak convergence topology
$\mathcal{P}^{\mathrm{TV}}$	$\mathcal{P}$ equipped with the total variation topology

Table 3: Notation table (part 2 of 2)

The boundedness is a typical example of **P**. We easily see that a continuous function is locally bounded.

A.2 Definition and properties of piecewise C^r-mappings

In this section, we define the notion of piecewise properties of functions, for example, piecewise C^r-functions. Examples of piecewise C^r-diffeomorphisms appearing in this paper include the $\mathcal{H}$-ACF with $\mathcal{H}$ being MLPs with ReLU activation. We first introduce the notion of *piecewise properties*.

Definition A.2 *Let* **P** *be a property of functions such as continuous, C^r, $C^{r,\alpha}$, and Lipschitz. Let $f : \mathbb{R}^m \to \mathbb{R}^n$ be a map. We say f is a* piecewise **P**-map *if there exists a mutually disjoint family of (at most countable) open subsets $\{V_i\}_{i\in I}$ such that*

- $\mathrm{vol}(\mathbb{R}^m \setminus U_f) = 0$,
- *for any $i \in I$, there exists an open subset W_i containing the closure $\overline{V_i}$ of V_i, and a map $\tilde{f}_i : W_i \to \mathbb{R}^n$ with the property* **P** *such that $\tilde{f}_i|_{V_i} = f|_{V_i}$, and*
- *for any compact subset K, $\#\{i \in I : V_i \cap K \neq \emptyset\} < \infty$.*

where $\#(\cdot)$ denotes the cardinality of a set, and we define

$$U_f := \bigsqcup_{i\in I} V_i.$$

Although there exist several definitions of piecewise functions, we introduce a generalized definition for our purpose. We remark that we here do not assume that piecewise C^r-maps are continuous everywhere, and thus they might have discontinuous points. We also remark that piecewise continuous mappings are essentially locally bounded in the sense that for any compact subset $K \subset \mathbb{R}^d$, $\mathrm{ess.sup}_K \|f\| = \|f\|_{K\cap U_f,0,\infty} < \infty$.

We define the notion of piecewise C^r-diffeomorphisms as follows.

Definition A.3 (Piecewise C^r-diffeomorphisms) *Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be a piecewise C^r-map. We say f is a* piecewise C^r-diffeomorphism *if we can choose $\{V_i\}_{i\in I}$ and $\{\tilde{f}_i : W_i \to \mathbb{R}^d\}_{i\in I}$ in Definition A.2 so that they additionally satisfy the following conditions:*

1. *the image of a nullset (that is, a Lebesgue-measurable subset of $\mathbb{R}^d$ whose measure is 0) via f is also a nullset,*
2. *$f|_{U_f}$ is injective,*
3. *for $i \in I$, $\tilde{f}_i$ is a C^r-diffeomorphism from W_i onto $\tilde{f}_i(W_i)$,*
4. *$\mathrm{vol}\left(\mathbb{R}^d \setminus f(U_f)\right) = 0$, and*
5. *for any compact subset K, $\#\{i \in I : f(V_i) \cap K \neq \emptyset\} < \infty$.*

We summarize the basic properties of piecewise C^r-diffeomorphisms in the proposition below. Note that for a piecewise C^r-diffeomorphism f, Df is defined almost everywhere since its value is determined on U_f (hence so is its determinant $|Df|$).

Proposition A.4 (Basic Properties of Piecewise C^r-diffeomorphisms) *Let $r \geq 1$ be a positive integer. Let f be a piecewise C^r-diffeomorphism. Then, we have the following:*

1. *There exists a piecewise C^r-diffeomorphism $f^\dagger$ such that $f(f^\dagger(x)) = x$ for $x \in U_{f^\dagger}$ and $f^\dagger(f(y)) = y$ for $y \in U_f$.*
2. *For any $h \in L^1$, we have $\int h(x)dx = \int h(f(x))|Df(x)|dx$.*
3. *For any compact subset K, $f^{-1}(K) \cap U_f$ is a bounded subset.*
4. *For any nullset F, then $f^{-1}(F)$ is also a nullset.*
5. *For any measurable set E and any compact set K, $f^{-1}(E \cap K)$ has a finite volume.*
6. *For any piecewise C^r-map (resp. piecewise Lipschitz map, piecewise C^r-diffeomorphism) g , the composition $g \circ f$ is also a piecewise C^r-map (resp. piecewise Lipschitz map, piecewise C^r-diffeomorphism).*

Proof Let $\{V_i\}_{i\in I}$ and $\{\tilde{f}_i : W_i \to \mathbb{R}^d\}_{i\in I}$ be as in Definition A.3.

Proof of 1 : First we note that since $f|_{V_i}$ is a restriction of the diffeomorphism $\tilde{f}_i$, $f(V_i)$ is an open set and $f|_{V_i}^{-1}$ is a well-defined C^r-function on $f(V_i)$. We also note that since $f|_{U_f}$ is injective, we have $f(U_f) = \bigsqcup_{i\in I} f(V_i)$. Fix $a \in \mathbb{R}^d$. We define $f^\dagger(x) = a$ for $x \in \mathbb{R}^d \setminus f(U_f)$ and define $f^\dagger(x) := f|_{V_i}^{-1}(x)$ for $x \in f(V_i)$. Then, $f^\dagger$ is a piecewise C^r-mapping with respect to the family of pairwise disjoint open subsets $\{f(V_i)\}_{i\in I}$, and satisfies the conditions for a piecewise C^r-diffeomorphism.

Proof of 2 : It follows by the following computation:

$$\begin{aligned}\int h(x)dx &= \int_{f(U_f)} h(x)dx \\ &= \sum_{i\in I} \int_{f(V_i)} h(x)dx \\ &= \sum_{i\in I} \int_{V_i} h(f(x))|Df(x)|dx = \int h(f(x))|Df(x)|dx.\end{aligned}$$

Proof of 3 It suffices to show that $f^{-1}(K) \cap U_f$ is covered by finitely many compact subsets. We remark that only finitely many V_i's intersect with $f^{-1}(K)$. If not, infinitely many $f(V_i)$'s intersect with $f(f^{-1}(K)) = K$, which contradicts the definition of piecewise C^r-diffeomorphisms. Let $I_0 \subset I$ be a finite subset composed of $i \in I$ such that V_i intersects with $f^{-1}(K)$. For $i \in I_0$, we define a compact subset $F_i := \tilde{f}_i^{-1}(\tilde{f}_i(\overline{V_i}) \cap K)$. Then we see that $f^{-1}(K) \cap U_f$ is contained in $\cup_{i\in I_0} F_i$.

Proof of 4 : It suffices to show that for any compact subset K, the volume of $f^{-1}(F) \cap K$ is zero. By applying 2 to the case $h = \mathbf{1}_F$, we see that

$$\int_{f^{-1}(F)} |Df(x)|dx = 0.$$

For $n > 0$, let $E_n := f^{-1}(F) \cap K \cap \{x \in \mathbb{R}^d : |Df(x)| \geq 1/n\}$. Then we have

$$\frac{\mathrm{vol}(E_n)}{n} \leq \int_{E_n} |Df(x)|dx \leq \int_{f^{-1}(F)} |Df(x)|dx = 0,$$

thus $\mathrm{vol}(K \cap f^{-1}(F)) = \lim_{n\to\infty} \mathrm{vol}(E_n) = 0$

Proof of 5 : By applying 2 to the case $h = \mathbf{1}_{E\cap K}$, we see that

$$\int_{f^{-1}(E\cap K)} |Df(x)|dx = \mathrm{vol}(E \cap K).$$

Let F be a closure of $f^{-1}(K) \cap U_f$. By 3, F is a compact subset. Let $I_0 := \{i \in I : F \cap V_i \neq \emptyset\}$ be a finite subset. Then we have

$$\begin{aligned} C &:= \inf_{f^{-1}(K)\cap U_f} |Df| \\ &\geq \inf_{i\in I_0} \inf_{F\cap \overline{V_i}} |D\tilde{f}_i| > 0. \end{aligned}$$

Thus,

$$\int_{f^{-1}(E\cap K)\cap U_f} |Df(x)|dx \geq C\mathrm{vol}(f^{-1}(E \cap K)),$$

where the last equality follows from $\mathrm{vol}(f^{-1}(E\cap K)\backslash U_f) = 0$. Thus we have $\mathrm{vol}(f^{-1}(E\cap K)) < \infty$

Proof of 6 : We first assume that g is a piecewise C^r-mapping and prove that $g \circ f$ is a piecewise C^r-mapping. We denote by $\{V_i\}_{i\in I}$, $\{V'_j\}_{j\in J}$ the disjoint open-set families associated with f and g, respectively. Let $V_{ij} := f^{-1}(f(V_i)\cap V'_j)\cap U_f$. We prove $\{V_{ij}\}_{(i,j)\in I\times J}$ is the open-set family associated with $g \circ f$ (that is, $\{V_{ij}\}$ satisfies the conditions of Definition A.2). Let $U_{g\circ f} := \cup_{i,j} V_{ij} = f^{-1}(U_g \cap f(U_f)) \cap U_f$. Then, we have

$$\mathbb{R}^d \setminus U_{g\circ f} = f^{-1}((\mathbb{R}^d \setminus U_g) \cup (\mathbb{R}^d \setminus f(U_f))) \cup (\mathbb{R}^d \setminus U_f).$$

Since $\mathrm{vol}(\mathbb{R}^d \setminus U_g) = 0$ and $\mathrm{vol}(\mathbb{R}^d \setminus f(U_f)) = 0$, we have

$$\mathrm{vol}(f^{-1}((\mathbb{R}^d \setminus U_g) \cup (\mathbb{R}^d \setminus f(U_f)))) = 0$$

by 4 of Proposition A.4. In addition, since $\mathrm{vol}(\mathbb{R}^d \setminus U_f) = 0$, we have $\mathrm{vol}(\mathbb{R}^d \setminus U_{g\circ f}) = 0$. That is, the first condition is satisfied. For the second condition, we denote by $\tilde{f}_i$ (resp. $\tilde{g}_j$) the extension of $f|_{V_i}$ (resp. $g|_{V'_j}$). Then, $\tilde{g}_j \circ \tilde{f}_i$ is an extension of $g \circ f|_{V_{ij}}$ on each V_{ij}. Finally, to prove the third condition, we take an arbitrary compact subset K and prove that $\#\{(i, j) \in I \times J : K \cap V_{ij} \neq \emptyset\} < \infty$. Indeed, since f is a piecewise C^r-diffeomorphism, $f(U_f \cap K)$ is a bounded subset by 3 of Proposition A.4. Hence, $M := \overline{f(U_f \cap K)}$ is compact. Since f is a piecewise C^r-diffeomorphism, we have

$$\#\{i \in I \mid M \cap f(V_i) \neq \emptyset\} < \infty.$$

Similarly, since g is a piecewise C^r-mapping, we have

$$\#\{j \in J \mid M \cap V'_j \neq \emptyset\} < \infty.$$

Therefore, the number of pairs (i, j) satisfying $M \cap f(V_i) \cap V'_j \neq \emptyset$ is also finite. Note that $U_f \cap K \cap V_i \cap f^{-1}(V_j) = K \cap V_{ij}$. Therefore, by applying the inverse of f (see 1 of

Proposition A.4), we obtain $\#\{(i,j) \mid K \cap V_{ij} \neq \emptyset\} < \infty$. It means the third condition is satisfied. Combining the above discussions so far, we conclude that $g \circ f$ is a piecewise C^r-mapping. In the case where g is a piecewise Lipschitz, the proof is the same as above.

Next, we prove that $f \circ g$ is a piecewise C^r-diffeomorphism when g is a piecewise C^r-diffeomorphism. We check the conditions in Definition A.3. The first, second, and third conditions follow by definition. For the third condition, since

$$\mathbb{R}^d \setminus (g \circ f(U_{g \circ f})) = \big(\mathbb{R}^d \setminus g(U_g)\big) \cup \big(\mathbb{R}^d \setminus g\big(f(U_f)\big) \subset \mathbb{R}^d \setminus g(f(U_f) \cap U_g),$$

it suffices to show that the volume of $\mathbb{R}^d \setminus g(f(U_f) \cap U_g)$ is zero. In fact, by the injectivity of g on U_g, we have

$$g(f(U_f) \cap U_g) = g(U_g) \setminus g(U_g \setminus f(U_f)).$$

Thus, we have

$$\mathbb{R}^d \setminus g(f(U_f) \cap U_g) = (\mathbb{R}^d \setminus g(U_g)) \cup g(U_g \setminus f(U_f)).$$

By definition of C^r-diffeomorphism, we conclude $\mathbb{R}^d \setminus g(f(U_f) \cap U_g)$ is a null set. For the fourth condition, let K be a compact subset. Let K be a compact set. Suppose $(i,j) \in I \times J$ satisfies $K \cap (g \circ f)(V_{ij}) \neq \emptyset$. Since $f(V_{ij}) \subset V_j'$, we have

$$K \cap g(V_j') \neq \emptyset. \tag{5}$$

Since g is a piecewise C^r-diffeomorphism, there exist finitely many j's satisfying (5). On the other hand, by applying the inverse of g, we have $g^{-1}(K) \cap U_g \cap f(V_{ij}) \neq \emptyset$, which implies

$$\overline{g^{-1}(K) \cap U_g} \cap f(V_i) \neq \emptyset. \tag{6}$$

Note that $\overline{g^{-1}(K) \cap U_g}$ is compact. Therefore, using the fact that f is a piecewise C^r-diffeomorphism, we see that there exist finitely many $i \in I$ satisfying (6). Therefore, we have $\#\{(i,j) \in I \times J \mid K \cap (g \circ f)(V_{ij}) \neq \emptyset\} < \infty$.

■

For a measurable mapping $f : \mathbb{R}^m \to \mathbb{R}^n$ and $R > 0$, we define a measurable set

$$\mathcal{L}(R; f) := \{x \in \mathbb{R}^m : \|f(x) - f(y)\| > R\|x - y\| \text{ for some } y \in U_f\}.$$

Then, we have the following proposition:

Proposition A.5 *Let $f : \mathbb{R}^m \to \mathbb{R}^n$ be a piecewise Lipschitz function. Assume f is linearly increasing, namely, there exists $a, b > 0$ such that $\|f(x)\| < a\|x\| + b$ for any $x \in \mathbb{R}^m$. Then for any compact subset $K \subset \mathbb{R}^m$, $\mathrm{vol}(\mathcal{L}(R; f) \cap K) \to 0$ as $R \to \infty$.*

Proof Let $\{V_i\}_{i \in I}$ be the disjoint family of open sets associated with f satisfying the properties of Definition A.2. Let B be an m-dimensional open ball of radius r containing K. Fix an arbitrary $\varepsilon > 0$. Let $C := \sup_{x \in \overline{B}} \|f(x)\|$. Because the linearly increasing condition of f implies its locally boundedness, we have $C < \infty$. For $\delta > 0$, we define

$$W_\delta := \{x \in \overline{B} : \mathrm{dist}\,(x, \partial U_f \cup \partial B)) < \delta\},$$

where $\mathrm{dist}(x, S) := \inf_{y \in S}\{\|x - y\|\}$. By the continuity of the Lebesgue measure, we have $\lim_{\delta \to 0} \mathrm{vol}(W_\delta) = 0$. Therefore, we can choose $\delta > 0$ so that $\mathrm{vol}(W_\delta) < \varepsilon$ holds.

We claim that

$$L := \sup_{(x,y)\in K\times(\mathbb{R}^m\setminus B)} \frac{\|f(x)-f(y)\|}{\|x-y\|}$$

is finite. In fact, let $r' := \inf_{(x,y)\in K\times(\mathbb{R}^m\setminus B)} \|x-y\|$. Then for $x \in K$ and $y \notin B$, we have

$$\begin{aligned}
\frac{\|f(x)-f(y)\|}{\|x-y\|} &\leq \frac{\|f(x)\|+\|f(y)\|}{\|x-y\|} \\
&\leq \frac{a\|x\|+a\|y\|+2b}{\|x-y\|} \\
&\leq \frac{a\|x\|+a(\|x-y\|+\|x\|)+2b}{\|x-y\|} \\
&\leq a+\frac{2a\|x\|+2b}{\|x-y\|} \\
&< a+\frac{2ar+2b}{r'}.
\end{aligned}$$

Thus, L is finite.

Due to the piecewise Lipschitz-ness of f, $\overline{B}$ intersects with finitely many V_i's. It implies that $f|_{B\setminus W_{\delta/2}}$ is a Lipschitz function. Put $L_\delta > 0$ as the Lipschitz constant of $f|_{B\setminus W_{\delta/2}}$.

For any $R > \max(L, L_\delta, 4C/\delta)$, we claim that $\mathcal{L}(R;f)\cap K$ is contained in W_δ. To prove it, we show that $x \notin \mathcal{L}(R;f)$ when $x \in K \setminus W_\delta$. Take arbitrary $y \in \mathbb{R}^m$. (Case 1) When $y \notin B$, since $x \in K$, we have $\frac{\|f(x)-f(y)\|}{\|x-y\|} \leq L$ by the definition of L. (Case 2) When $y \in B \setminus W_{\delta/2}$, since $x \in K \setminus W_\delta \subset B \setminus W_{\delta/2}$, we have $\frac{\|f(x)-f(y)\|}{\|x-y\|} \leq L_\delta$ by the definition of L_δ. (Case 3) When $y \in B \cap W_{\delta/2}$, we have $\|x-y\| \geq \frac{\delta}{2}$ because $x \notin W_\delta$. Thus,

$$\frac{\|f(x)-f(y)\|}{\|x-y\|} \leq \frac{\|f(x)\|+\|f(y)\|}{\delta/2} \leq \frac{C+C}{\delta/2} \leq \frac{4C}{\delta}.$$

Combining these three cases, we conclude that $x \notin \mathcal{L}(R;f)$. Thus we have $\mathrm{vol}(\mathcal{L}(R;f)\cap K) < \varepsilon$, namely, we conclude $\mathrm{vol}(\mathcal{L}(R;f)\cap K) \to 0$ as $R \to \infty$. ■

Remark A.6 *The linearly increasing condition is important to prove our main theorem. Our approximation targets are compactly supported diffeomorphisms, affine transformations, and the discontinuous ACFs appeared in Section E.2.1, all of which satisfy the linearly increasing condition.*

Appendix B. Compatibility of approximation and composition

In this section, we prove Proposition 37. It enables the component-wise approximation, that is, approximating a composition of some transformations by approximating each constituent and composing them. The justification of this procedure is not trivial and requires a fine mathematical argument. The results here build on the terminologies and the propositions for piecewise C^1-diffeomorphisms presented in Section A.

Lemma B.1 *Let $p = [1, \infty)$. Let $m \geq 1$ and let $\mathcal{F}$ be the set of $\mathbb{R}^m$-valued piecewise Lipschitz mappings. Let $\mathcal{G}$ be the set of piecewise C^1-diffeomorphisms on $\mathbb{R}^d$. Let $\mathcal{F}_0 \subset \mathcal{F}$ and $\mathcal{G}_0 \subset \mathcal{G}$ be the subsets composed of linearly increasing mappings. Here, a function f on $\mathbb{R}^d$ is linearly increasing if there exists $a, b > 0$ such that $\|f(x)\| < a\|x\| + b$ for all $x \in \mathbb{R}^d$. Then, the map*

$$\mathcal{C} : \mathcal{F} \times \mathcal{G}^k \longrightarrow \mathcal{F}; (h, f_1, \ldots, f_k) \mapsto h \circ f_1 \circ \cdots \circ f_k \tag{7}$$

is continuous at any point of $\mathcal{F}_0 \times \mathcal{G}_0^k$ with respect to the relative topology of $W^{0,p}_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^m) \times W^{0,p}_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^d)^k$.

Proof Since $\mathcal{C}(\mathcal{F}_0 \times \mathcal{G}_0) \subset \mathcal{F}_0$ (see the statement 6 of Proposition A.4), the lemma follows from the case $k = 1$ via the mathematical induction. Thus, we only treat the case $k = 1$. Let $(F_2, G_2) \in \mathcal{F}_0 \times \mathcal{G}_0$. Then, it suffices to show that for any $\varepsilon > 0$ and compact set $K \subset \mathbb{R}^d$, there exist $\delta > 0$ and compact set $K_0 \subset \mathbb{R}^d$ such that for any $(F_1, G_1) \in \mathcal{F} \times \mathcal{G}$ satisfying $\|G_2 - G_1\|_{0,p,K_0}, \|F_2 - F_1\|_{0,p,K_0} < \delta$, we have

$$\|F_2 \circ G_2 - F_1 \circ G_1\|_{0,p,K} < \varepsilon.$$

Fix arbitrary $\varepsilon > 0$ and compact set $K \subset \mathbb{R}^d$. Put $K' := \overline{G_2(K \cap U_{G_2})}$. Then, since $G_2(K \cap U_{G_2})$ is bounded (see the remark under Definition A.2), K' is compact. We claim that there exists $R > 0$ such that

$$\mathrm{vol}(G_2^{-1}\left(\mathcal{L}(R; F_2) \cap K'\right))^{1/p} < \frac{\varepsilon}{3\underset{K'}{\mathrm{ess.sup}}\|F_2\|},$$

which can be confirmed as follows. Take an increasing sequence $R_n > 0$ ($n \geq 1$) satisfying $\lim_{n\to\infty} R_n = \infty$. Let $B_n := \mathcal{L}(R_n; F_2) \cap K'$ and $A_n := G_2^{-1}(B_n)$. Then, from Proposition A.5, we have $\mathrm{vol}(B_n) \to 0$, which implies $\mathrm{vol}(\bigcap_{n=1}^{\infty} B_n) = 0$. By Proposition A.4 (4), we have $\mathrm{vol}(\bigcap_{n=1}^{\infty} A_n) = \mathrm{vol}(G_2^{-1}(\bigcap_{n=1}^{\infty} B_n)) = 0$. By Proposition A.4 (5), we have $\mathrm{vol}(A_1) = \mathrm{vol}(G_2^{-1}(B_1)) < \infty$. Recall that if a decreasing sequence $\{S_n\}_{n=1}^{\infty}$ of measurable sets satisfies $\mathrm{vol}(S_1) < \infty$ and $\mathrm{vol}(\bigcap_{n=1}^{\infty} S_n) = 0$, then $\lim_{n\to\infty} \mathrm{vol}(S_n) = 0$. Therefore, we obtain $\lim_{n\to\infty} \mathrm{vol}(A_n) = 0$, and we have the assertion of the claim.

Take $G_1 \in \mathcal{G}$ such that

$$\|G_2 - G_1\|_{0,p,K_0} < \frac{\varepsilon}{3R}.$$

Put $S := G_2^{-1}\left(\mathcal{L}(R; F_2) \cap K'\right)$, and define a compact subset $K'' := \overline{(G_1^{\dagger})^{-1}(K) \cap U_{G_1^{\dagger}}}$. Here, the compactness of K'' follows from Proposition A.4 (3). Next, we take $F_1 \in \mathcal{F}$ such that

$$\|F_2 - F_1\|_{p,K''} < \frac{\varepsilon}{3 \underset{(G_1^{\dagger})^{-1}(K)}{\mathrm{ess.sup}} |\det(DG_1^{\dagger})|}$$

where $G_1^{\dagger}$ is a piecewise C^1-diffeomorphism defined by Proposition A.4 (1). Therefore, if we take

$$\delta := \min\left(\frac{\varepsilon}{3\underset{K'}{\mathrm{ess.sup}}\|F_2\|}, \frac{\varepsilon}{3R}\right)$$

and $K_0 := K \cup K''$, then we have

$$\begin{aligned}
&\|F_2 \circ G_2 - F_1 \circ G_1\|_{0,p,K} \\
&\leq \|F_2 \circ G_2 - F_2 \circ G_1\|_{0,p,K_0} + \|F_2 \circ G_1 - F_1 \circ G_1\|_{0,p,K_0} \\
&\leq \|(F_2 \circ G_2 - F_2 \circ G_1)\mathbf{1}_S\|_{0,p,K} + \left\|(F_2 \circ G_2 - F_2 \circ G_1)\mathbf{1}_{K\setminus S}\right\|_{0,p,K} \\
&\quad + \operatorname*{ess.sup}_{(G_1^\dagger)^{-1}(K)} |\det(DG_1^\dagger)\, \|F_2 - F_1\|_{0,p,K} \\
&< \varepsilon.
\end{aligned}$$

■

Lemma B.2 *Let $m \geq 1$ and let $\mathcal{F} := W^{0,\infty}(\mathbb{R}^d, \mathbb{R}^m)$. Let $\mathcal{G}$ be a subset $W^{0,\infty}(\mathbb{R}^d, \mathbb{R}^d)$ whose inverse images of any null sets are again null sets. Let $\mathcal{F}_0 \subset \mathcal{F}$ and $\mathcal{G}_0 \subset \mathcal{G}$ be the subsets composed of continuous mappings. Then, the map*

$$\mathcal{C} : \mathcal{F} \times \mathcal{G}^k \longrightarrow \mathcal{F}; (h, f_1, \ldots, f_k) \mapsto h \circ f_1 \circ \cdots \circ f_k \tag{8}$$

is continuous at any point of $\mathcal{F}_0 \times \mathcal{G}_0^k$ with respect to the relative topology of $W^{0,\infty}_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^m) \times W^{0,\infty}_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^d)^k$.

Proof Since $\mathcal{C}(\mathcal{F}_0 \times \mathcal{G}_0) \subset \mathcal{F}_0$ (see the statement 6 of Proposition A.4), the proposition follows from the case $k = 1$ via the mathematical induction. Thus, we only treat the case $k = 1$. Let $(F_2, G_2) \in \mathcal{F}_0 \times \mathcal{G}_0$. Then, it suffices to show that for any $\varepsilon > 0$ and compact set $K \subset \mathbb{R}^d$, there exist $\delta > 0$ and compact set $K_0 \subset \mathbb{R}^d$ such that for any $(F_1, G_1) \in \mathcal{F} \times \mathcal{G}$ satisfying $\|G_2 - G_1\|_{0,\infty,K_0}, \|F_2 - F_1\|_{0,\infty,K_0} < \delta$, we have

$$\|F_2 \circ G_2 - F_1 \circ G_1\|_{0,\infty,K} < \varepsilon.$$

Take any positive number $\epsilon > 0$ and compact set $K \subset \mathbb{R}^d$. Put $r := \max_K |G_2|$ (note that G_2 is continuous) and $K' := \{x \in \mathbb{R}^d : |x| \leq r + 1\}$. Let $F_1 \in \mathcal{F}$ satisfying

$$\mathrm{vol}\{x \in K' : |F_2(x) - F_1(x)| > \epsilon/2\} = 0.$$

Since any continuous map is uniformly continuous on a compact set, we can take a positive number $\delta > 0$ such that for any $x, y \in K'$ with $|x - y| < \delta$,

$$|F_2(x) - F_2(y)| < \frac{\varepsilon}{2}.$$

From the assumption, we can take $G_1 \in \mathcal{G}$ satisfying

$$\mathrm{vol}\{x \in K : |G_2(x) - G_1(x)| > \min\{1, \delta\}\} = 0.$$

Since

$$|F_2 \circ G_2(x) - F_1 \circ G_1(x)| \leq |F_2(G_2(x)) - F_2(G_1(x))| + |F_2(G_1(x)) - F_1(G_1(x))|,$$

we see that the set of $x \in K$ such that $\varepsilon < |F_2 \circ G_2(x) - F_1 \circ G_1(x)|$ is a null set. Thus, we have

$$\|F_2 \circ G_2 - F_1 \circ G_1\|_{0,\infty,K} < \varepsilon.$$

■

Let $B_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^m)$ be the linear space composed of locally bounded measurable maps from $\mathbb{R}^d$ to $\mathbb{R}^m$. We equip B_{loc} with the topology generated by the seminorms $\{\|\cdot\|_{\sup,K}\}_K$, where K runs on the set of compact subsets of $\mathbb{R}^d$, and define for any $h \in B_{\mathrm{loc}}$,

$$\|h\|_{\sup,K} := \sup_{x \in K} \|h(x)\|.$$

Then, we provide a similar result for the *sup*-norm case as follows:

Lemma B.3 *Let $m \geq 1$ and let $\mathcal{F} := B_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^m)$ and $\mathcal{G}$ be a subset $B_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^d)$. Let $\mathcal{F}_0 \subset \mathcal{F}$ and $\mathcal{G}_0 \subset \mathcal{G}$ be the subsets composed of continuous mappings. Then, the map*

$$\mathcal{C} : \mathcal{F} \times \mathcal{G}^k \longrightarrow \mathcal{F}; (h, f_1, \ldots, f_k) \mapsto h \circ f_1 \circ \cdots \circ f_k \tag{9}$$

is continuous at any point of $\mathcal{F}_0 \times \mathcal{G}_0^k$ with respect to the relative topology of $W^{0,\infty}_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^m) \times W^{0,\infty}_{\mathrm{loc}}(\mathbb{R}^d, \mathbb{R}^d)^k$.

Proof We may assume $k = 1$ and let $(F_2, G_2) \in \mathcal{F}_0 \times \mathcal{G}_0$ as in the proof of Lemma B.2. Take any positive number $\epsilon > 0$ and compact set $K \subset \mathbb{R}^d$. Put $r := \max_{k \in K} |G_2(k)|$ and $K' := \{x \in \mathbb{R}^d : |x| \leq r + 1\}$. Let $F_1 \in \mathcal{F}$ satisfying

$$\sup_{x \in K'} |F_2(x) - F_1(x)| \leq \frac{\epsilon}{2}.$$

Since any continuous map is uniformly continuous on a compact set, we can take a positive number $\delta > 0$ such that for any $x, y \in K'$ with $|x - y| < \delta$,

$$|F_2(x) - F_2(y)| < \frac{\varepsilon}{2}.$$

Let $G_1 \in \mathcal{G}$ satisfying

$$\sup_{x \in K} |G_2(x) - G_1(x)| \leq \min\{1, \delta\}.$$

Then, it is clear that $G_2(K) \subset K'$ by the definition of K'. Moreover, we have $G_1(K) \subset K'$. In fact, we have

$$|G_1(k)| \leq \sup_{x \in K} |G_2(x) - G_1(x)| + |G_2(k)| \leq 1 + r \quad (k \in K).$$

Then for any $x \in K$, we have

$$\begin{aligned} |F_2 \circ G_2(x) - F_1 \circ G_1(x)| &\leq |F_2(G_2(x)) - F_2(G_1(x))| + |F_2(G_1(x)) - F_1(G_1(x))| \\ &< \epsilon. \end{aligned}$$

■

Now, we provide the proof of Proposition 37:
Proof [Proof of Proposition 37] The Leibniz rule and the chain rule hold for weak derivatives under the present condition (see McDuff and Salamon 2004, Exercise B.1.2 and Ziemer 1989, Theorem 2.1.11). Thus, it follows from Lemmas B.1 and B.2. The last statement follows from Lemma B.3 in the same way. ■

Appendix C. Proof of Distributional Universalities

C.1 Proof of Proposition 38: From L^p-universality to distributional universality

Here, we prove Proposition C.2, which corresponds to Proposition 38 in the main text. We first include proof that any probability measure on $\mathbb{R}^m$ is arbitrarily approximated by an absolutely continuous probability measure in the weak convergence topology.

Lemma C.1 *Let $\mu \in \mathcal{P}$ be an arbitrary probability measure. Then there exists a sequence $\{\mu_n\}_{n=1}^{\infty} \subset \mathcal{P}_{\mathrm{ab}}$ of absolutely continuous probability measures such that μ_n weakly converges to μ.*

Proof Let ϕ be a compactly-supported positive bounded C^∞ function such that $\int_{\mathbb{R}^m} \phi(x)dx = 1$ and $\mathrm{supp}(\phi) \subset \{x \in \mathbb{R}^m : \|x\| \leq B\}$ where $B > 0$. For $t > 0$, put $\phi_t(x) := t^{-m}\phi(x/t)$. We define

$$w_t(x) = \int_{\mathbb{R}^m} \phi_t(x - y)d\mu(y).$$

We prove that the absolutely continuous measure $w_t dx$ weakly converges to μ as $t \to 0$. In fact, given an L-Lipschitz continuous function f such that, we have

$$\begin{aligned}
\left| \int_{\mathbb{R}^m} f w_t dx - \int f d\mu \right| &= \left| \int \int_{\mathbb{R}^m} \left(f(y + tx) - f(y)\right) \phi(x) dx d\mu(y) \right| \\
&\leq \int \int_{\mathbb{R}^m} |f(y + tx) - f(y)| \phi(x) dx d\mu(y) \\
&\leq \int \int_{\mathbb{R}^m} Lt\|x\| \phi(x) dx d\mu(y) \\
&\leq LBt.
\end{aligned}$$

Therefore, as $t \to 0$, we have

$$\int_{\mathbb{R}^m} f w_t dx \to \int f d\mu,$$

therefore, $\left\{ w_{\frac{1}{n}} dx \right\}_n$ weakly converges to μ. ■

First, note that the larger p, the stronger the notion of L^p-universality: if a model $\mathcal{M}$ is an L^p-universal approximator for $\mathcal{F}$, it is also an L^q-universal approximator for $\mathcal{F}$ for all $1 \leq q \leq p$. In particular, we use this fact with $q = 1$ in the following proof.

Proposition C.2 (Proposition 38 in the main text) *Let $p \in [1, \infty)$. Suppose $\mathcal{M}$ is an L^p-universal approximator for $\mathcal{T}^\infty$. Then $\mathcal{M}$ is a $(\mathcal{P}^{\mathrm{w}}, \mu)$-distributional universal approximator for $\mathcal{P}$ for any $\mu \in \mathcal{P}_{\mathrm{ab}}$.*

Proof By Lemma C.1, it suffices to prove that $\mathcal{M}$ is a $(\mathcal{P}^{\mathrm{w}}, \mu)$-distributional universal approximator for $\mathcal{P}_{\mathrm{ab}}$ for any $\mu \in \mathcal{P}_{\mathrm{ab}}$. We denote by BL_1 the set of bounded Lipschitz functions $f\colon \mathbb{R}^d \to \mathbb{R}$ satisfying $\|f\|_{\sup,\mathbb{R}^d} + L_f \leq 1$, where L_f denotes the Lipschitz constant of f. Let $\mu, \nu \in \mathcal{P}_{\mathrm{ab}}$ be absolutely continuous probability measures, and take any $\varepsilon > 0$. By Theorem 11.3.3 in Dudley (2002), it suffices to show that there exists $g \in \mathcal{M}$ such that

$$\beta(g_*\mu, \nu) := \sup_{f \in \mathrm{BL}_1} \left| \int_{\mathbb{R}^d} f \, dg_*\mu - f \, d\nu \right| < \varepsilon.$$

Let $p, q \in L^1(\mathbb{R}^d)$ be the density functions of μ and ν respectively. Let $\phi \in L^1(\mathbb{R}^d)$ be a positive C^∞-function such that $\int_{\mathbb{R}^d} \phi(x)dx = 1$ (for example, the density function of the standard Gaussian distribution), and for $t > 0$, put $\phi_t(x) := t^{-d}\phi(x/t)$. We define $\mu_t := \phi_t * pdx$ and $\nu_t := \phi_t * qdx$. Since both $\|\phi_t * p - p\|_{1,\mathbb{R}^d}$ and $\|\phi_t * q - q\|_{1,\mathbb{R}^d}$ converge to 0 as $t \to 0$, there exists $t_0 > 0$ such that for any continuous mapping $G : \mathbb{R}^d \to \mathbb{R}^d$,

$$\left| \int_{\mathbb{R}^d} f \, dG_*\mu_{t_0} - f \, dG_*\mu \right| < \frac{\|f\|_{\mathbb{R}^d,0,\infty}\,\varepsilon}{5}, \quad \left| \int_{\mathbb{R}^d} f \, d\nu_{t_0} - f \, d\nu \right| < \frac{\|f\|_{\mathbb{R}^d,0,\infty}\,\varepsilon}{5}.$$

By using Lemma C.3 below, there exists $T \in \mathcal{T}^\infty$ such that $T_*\mu_{t_0} = \nu_{t_0}$. Let $K \subset \mathbb{R}^d$ be a compact subset such that

$$1 - \mu_{t_0}(K) < \frac{\varepsilon}{5}.$$

By the assumption, there exists $g \in \mathcal{M}$ such that

$$\int_K |T(x) - g(x)|dx < \frac{\varepsilon}{5 \sup_{x \in K} |\phi_{t_0} * p(x)|}.$$

Thus for any $f \in \mathrm{BL}_1$, we have

$$\begin{aligned}
&\left| \int_{\mathbb{R}^d} f \, dg_*\mu - f \, d\nu \right| \\
&\leq \left| \int_{\mathbb{R}^d} f \, dg_*\mu_{t_0} - f \, dg_*\mu \right| + \left| \int_{\mathbb{R}^d} f \, d\nu_{t_0} - f \, d\nu \right| \\
&\quad + \left| \int_{\mathbb{R}^d \setminus K} f \circ T \, d\mu_{t_0} \right| + \left| \int_{\mathbb{R}^d \setminus K} f \circ g \, d\mu_{t_0} \right| + \int_K |f(T(x)) - f(g(x))| \, d\mu_{t_0}(x) \\
&< \frac{\|f\|_{\mathbb{R}^d,0,\infty}\,\varepsilon}{5} + \frac{\|f\|_{\mathbb{R}^d,0,\infty}\,\varepsilon}{5} + \frac{\|f\|_{\mathbb{R}^d,0,\infty}\,\varepsilon}{5} + \frac{\|f\|_{\mathbb{R}^d,0,\infty}\,\varepsilon}{5} + \frac{L_f\varepsilon}{5} \\
&\leq \varepsilon,
\end{aligned}$$

where L_f is the Lipschitz constant of f. Here we used $\|f\|_{\mathbb{R}^d,0,\infty} + L_f \leq 1$. Therefore, we have $\beta(g_*\mu, \nu) < \varepsilon$. ■

The following lemma is essentially due to (Hyvärinen and Pajunen, 1999).

Lemma C.3 *Let μ be a probability measure on $\mathbb{R}^d$ with a C^∞ density function p. Let $U := \{x \in \mathbb{R}^d : p(x) > 0\}$. Then there exists a diffeomorphism $T : U \to (0,1)^d$ such that its Jacobian is an upper triangular matrix with positive diagonals, and $T_*\mu = \mathrm{U}(0,1)^d$. Here, $\mathrm{U}(0,1)^d$ is the uniform distribution on $[0,1]^d$.*

Proof Let $q_i(x_1, \ldots, x_i) := \int_{\mathbb{R}^{d-i}} p(x_1, \ldots, x_{i+1}, \ldots, x_d)\, dx_{i+1} \ldots dx_d$. Then we define $T : U \to (0,1)^d$ by

$$T(x_1, \ldots, x_d) := \left(\int_{-\infty}^{x_i} \frac{q_i(x_1, \ldots, x_{i-1}, y)}{q_{i-1}(x_1, \ldots, x_{i-1})} dy \right)_i .$$

Then we see that T is a diffeomorphism, and its Jacobian is upper triangular with positive diagonal elements. Moreover, by direct computation, we have $T_* d\mu = U(0,1)$. ■

C.2 Proof of Proposition 40: From Sobolev Universality to Distributional Universality in the Total Variation Metric

In this section, we prove Proposition 40. Recall the definition of the total variation distance:

$$\|\nu - \mu\|_{\mathrm{TV}} := \sup_A |\nu(A) - \mu(A)|,$$

where the supremum is taken over all measurable sets of the underlying space.

Here, we restate the proposition.

Theorem C.4 (Proposition 40 in the main text) *Let $r \geq 1$. Let*

$$\mathcal{F}_0 := W^{0,\infty}_{\mathrm{loc}}(U, \mathbb{R}^d) \cap W^{1,1}_{\mathrm{loc}}(U, \mathbb{R}^d).$$

We define the topology of $\mathcal{F}_0$ as the weakest topology such that the inclusion maps $\imath_0 : \mathcal{F}_0 \hookrightarrow W^{0,\infty}_{\mathrm{loc}}(U, \mathbb{R}^d)$ and $\imath_1 : \mathcal{F}_0 \hookrightarrow W^{1,1}_{\mathrm{loc}}(U, \mathbb{R}^d)$ are both continuous. Suppose any element in the model $\mathcal{M}$ is locally $C^{0,1}$ and a piecewise C^1-diffeomorphism. If $\mathcal{M}$ is an $\mathcal{F}_0$-universal approximator for $\mathcal{T}^\infty$, then $\mathcal{M}$ is a $(\mathcal{P}^{\mathrm{TV}}, \mu)$-distributional universal approximator for $\mathcal{P}_{\mathrm{ab}}$ for any $\mu \in \mathcal{P}_{\mathrm{ab}}$.

Proof Let μ, $\nu \in \mathcal{P}_{\mathrm{ab}}$. Take any $\varepsilon > 0$. It is enough to show that there exists $f \in \mathcal{M}$ such that

$$2\|\nu - f_*\mu\|_{\mathrm{TV}} < \epsilon,$$

where $\|\cdot\|_{\mathrm{TV}}$ is the total variation norm. By Lemmas C.3 and C.5, we can assume that there exist a positive smooth function w satisfying $d\mu(x) = w(x)dx$ and $g \in \mathcal{T}^\infty$ such that $\nu = g_*\mu$ and $g(\mathbb{R}^d) = \mathbb{R}^d$. We fix a large compact set $K' \subset \mathbb{R}^d$ such that

$$\int_{\mathbb{R}^d \setminus K'} dg_*\mu < \frac{\varepsilon}{4}.$$

We fix an "inverse" $f^\dagger$ of the piecewise C^1-diffeomorphism f as in 1 in Proposition A.4. We may assume $f^\dagger(K') \subset f^{-1}(K')$ if we take a suitable $f^\dagger$. Note that $f^{-1}(K') \setminus f^\dagger(K')$ is a nullset. Then, we can write $d(f_*\mu)(x) = w(f^\dagger(x))J_{f^\dagger}(x)dx$ and $d(g_*\mu)(x) = w(g^{-1}(x))J_{g^{-1}}(x)dx$.

By Lemma C.6 below, there exists a compact subset $K \subset \mathbb{R}^d$ such that $f^{-1}(K') \subset K$ for any $f \in \mathcal{M}$ satisfying $\|f - g\|_{K,0,\infty} < \varepsilon$.

Since g is a diffeomorphism, there exists $M_0 > 0$ such that $|J_g(g^{-1}(k'))|^{-1} < M_0$ for any $k' \in K'$. Moreover, since the function $J_g(g^{-1}(\cdot))$ is Lipschitz on $g(K) \cup K'$, we can take $M_1 > 0$ satisfying $|J_g(g^{-1}(x)) - J_g(g^{-1}(y))| < M_1|x - y|$ for any $x, y \in g(K) \cup K'$. Since the function w is Lipschitz on $g^{-1}(K') \cup K$, we can take $L_0 > 0$ satisfying $|w(x) - w(y)| < L_0|x - y|$ for any $x, y \in g^{-1}(K') \cup K$. Since g^{-1} is Lipschitz on $g(K) \cup K'$, we can take $L_1 > 0$ satisfying $|g^{-1}(x) - g^{-1}(y)| < L_1|x - y|$ for any $x, y \in g(K) \cup K'$.

From the assumption, we can take $f \in \mathcal{M}$ satisfying

$$\|f - g\|_{K,0,\infty} < \frac{\varepsilon}{16M_0L_0 \max\{M_1, L_1\} \max\{\mathrm{vol}(K'), \mathrm{vol}(K), 1\}},$$
$$\|f - g\|_{K,1,1} < \frac{\varepsilon}{16M_0 \max_{x\in K} |w(x)|}.$$

Then, since the total variation distance of probability measures is given by half the L^1-norm of the Radon-Nikodym derivative, we have

$$\begin{aligned}
&2\|g_*\mu - f_*\mu\|_{TV} \\
&\leq \int_{K'} |w(f^\dagger(x))J_{f^\dagger}(x) - w(g^{-1}(x))J_{g^{-1}}(x)|dx + \int_{\mathbb{R}^d\setminus K'} df_*\mu + \int_{\mathbb{R}^d\setminus K'} dg_*\mu \\
&\leq 2\int_{K'} |w(f^\dagger(x))J_{f^\dagger}(x) - w(g^{-1}(x))J_{g^{-1}}(x)|dx + 2\int_{\mathbb{R}^d\setminus K'} dg_*\mu \\
&\leq 2\int_{K'} |w(f^\dagger(x)) - w(g^{-1}(x))||J_{g^{-1}}(x)|dx + 2\int_{K'} |J_{f^\dagger}(x) - J_{g^{-1}}(x)||w(f^\dagger(x))|dx + \frac{\varepsilon}{2}.
\end{aligned}$$

As for the second equality, we use

$$\begin{aligned}
\int_{\mathbb{R}^d\setminus K'} df_*\mu &= 1 - \int_{K'} df_*\mu \\
&\leq \int_{K'} |w(f^\dagger(x))J_{f^\dagger}(x) - w(g^{-1}(x))J_{g^{-1}}(x)|dx + 1 - \int_{K'} dg_*\mu \\
&= \int_{K'} |w(f^\dagger(x))J_{f^\dagger}(x) - w(g^{-1}(x))J_{g^{-1}}(x)|dx + \int_{\mathbb{R}^d\setminus K'} dg_*\mu
\end{aligned}$$

The first term is estimated as follows:

$$\begin{aligned}
&\int_{K'} |w(f^\dagger(x)) - w(g^{-1}(x))||J_{g^{-1}}(x)|dx \\
&\leq L_0 M_0 \int_{K'} |f^\dagger(x) - g^{-1}(x)|dx \\
&= L_0 M_0 \int_{K'} |g^{-1}(g \circ f^\dagger(x)) - g^{-1}(f \circ f^\dagger(x))|dx \\
&\leq L_0 M_0 L_1 \int_{K'} |g(f^\dagger(x)) - f(f^\dagger(x))|dx \\
&\leq L_0 M_0 L_1 \operatorname{vol}(K') \sup_{k' \in f^{-1}(K')} |g(k') - f(k')| \\
&\leq L_0 M_0 L_1 \operatorname{vol}(K') \sup_{k \in K} |g(k) - f(k)| \\
&< \frac{\varepsilon}{8}.
\end{aligned}$$

Here, we used the fact $f^\dagger(K') \subset K$ in the second-to-last inequality and the bound for $\|f - g\|_{K,0,\infty}$ in the last inequality.

Similarly, the second term is bounded as follows:

$$\begin{aligned}
&\int_{K'} |J_{f^\dagger}(x) - J_{g^{-1}}(x)||w(f^\dagger(x))|dx \\
&= \int_{f^\dagger(K')} |J_f(x)^{-1} - J_g(g^{-1} \circ f(x))^{-1}||w(x)|J_f(x)dx \\
&\leq \int_{f^\dagger(K')} |1 - J_f(x)J_g(g^{-1} \circ f(x))^{-1}||w(x)|dx \\
&= \int_{f^\dagger(K')} |J_g(g^{-1} \circ f(x))^{-1}||J_g(g^{-1} \circ f(x)) - J_f(x)||w(x)|dx \\
&\leq M_0 \max_{x \in K} |w(x)| \int_{f^\dagger(K')} |J_g(g^{-1} \circ f(x)) - J_f(x)|dx \\
&= M_0 \max_{x \in K} |w(x)| \left[\int_{f^\dagger(K')} |J_g(g^{-1} \circ f(x)) - J_g(g^{-1} \circ g(x))| + |J_g(x) - J_f(x)|dx \right] \\
&\leq M_0 \max_{x \in K} |w(x)| \left[M_1 \int_{f^\dagger(K')} |f(x) - g(x)|dx + \int_{f^\dagger(K')} |J_g(x) - J_f(x)|dx \right] \\
&\leq M_0 \max_{x \in K} |w(x)| \left[M_1 \int_{x \in K} |f(x) - g(x)|dx + \int_{x \in K} |J_g(x) - J_f(x)|dx \right] \\
&< \frac{\varepsilon}{16} + \frac{\varepsilon}{16} = \frac{\varepsilon}{8}.
\end{aligned}$$

Again, we used $f^\dagger(K) \subset K$ in the second-to-last inequality. In the last inequality, we used the bound for $\|f - g\|_{K,0,\infty}$ for the first term and the bound for $\|f - g\|_{K,1,1}$ for the second term, respectively. ■

Lemma C.5 *Let μ be an absolutely continuous probability measure on $\mathbb{R}^d$. For any $\varepsilon > 0$, there exists an absolutely continuous probability measure ν such that $d\nu(x) = w(x)dx$ for some $w \in C^\infty(\mathbb{R}^d)$ with $w > 0$ and $\|\mu - \nu\|_{\mathrm{TV}} < \varepsilon$.*

Proof Let $p \in L^1(\mathbb{R}^d)$ be the density function of μ. Let $\phi \in L^1(\mathbb{R}^d)$ be a positive C^∞ function satisfying $\int_{\mathbb{R}^d} \phi(x)dx = 1$. For $t > 0$, put $\phi_t(x) := t^{-d}\phi(x/t)$. Then we have

$$2\|\mu - \nu\|_{TV} = \|p - \phi_t * p\|_{L^1(\mathbb{R}^d)} \to 0 \quad (t \to +0).$$

■

Lemma C.6 *Let the model $\mathcal{M}$ be as in Theorem C.4 and let g be a homeomorphism from $\mathbb{R}^d$ to $\mathbb{R}^d$. Let $K' \subset \mathbb{R}^d$ be a compact set and $\varepsilon > 0$. Then, there exists a compact subset $K \subset \mathbb{R}^d$ such that $f^{-1}(K') \subset K$ for any $f \in \mathcal{M}$ satisfying $\|f - g\|_{K,0,\infty} < \varepsilon$.*

Proof We may assume $K' = \overline{B(0, L)}$ for sufficiently large $L > 0$ such that $L \geq \varepsilon$. Since g is a homeomorphism, there exists sufficiently large $R > 0$ such that $\overline{B(0, R)} \supset g^{-1}(B(0, L + 2\varepsilon))$, that is, $g(\overline{B(0, R)}) \supset B(0, L + 2\varepsilon)(\supset K')$. We denote $K := \overline{B(0, R)}$. Suppose $f \in \mathcal{M}$ satisfies $\|f - g\|_{K,0,\infty} < \varepsilon$. Then, we have $f(\partial K) \cap K' = \emptyset$ for any f. Thus, we see that $K' \subset f(B(0, R)) \cup (\mathbb{R}^d \setminus f(K))$. Since K' is connected, we see that either $K' \subset f(K)$ or $K' \subset \mathbb{R}^d \setminus f(K)$. Suppose $K' \subset \mathbb{R}^d \setminus f(K)$. On the other hand, since $0 \in K' \subset g(K)$, there exists $x \in K$ such that $g(x) = 0$. Since $f(K) \cap K' = \emptyset$, we have

$$L < |f(x) - 0| = |f(x) - g(x)| < \varepsilon,$$

which is a contradiction. Therefore, we conclude $K' \subset f(K)$. Since f is a diffeomorphism, we have $f^{-1}(K') \subset K$. ■

Appendix D. Proofs for the equivalence of the universality

D.1 Proof of Lemma 29

Proof [Proof of Lemma 29] We denote the injections of U and $f(U)$ into $\mathbb{R}^d$ by $\iota_1 \colon U \hookrightarrow \mathbb{R}^d$ and $\iota_2 \colon f(U) \hookrightarrow \mathbb{R}^d$, respectively. Since U is C^r-diffeomorphic to $\mathbb{R}^d$ and f is C^r-diffeomorphic, $f(U)$ is also C^r-diffeomorphic to $\mathbb{R}^d$. Thus, $f(U)$ is C^∞-diffeomorphic to $\mathbb{R}^d$ by Hirsch (1976, p.50, Theorem 2.7). By applying Corollary D.2 below to $\iota_1 \circ f^{-1}|_{f(U)} \colon f(U) \to \mathbb{R}^d$ and the injection ι_2, we can obtain C^r-diffeomorphisms $F_1 \colon f(U) \to \mathbb{R}^d$ and $F_2 \colon f(U) \to \mathbb{R}^d$ such that $F_1|_{f(K)} = f^{-1}|_{f(K)}$ and $F_2|_{f(K)} = \mathrm{Id}_{f(K)}$, where $\mathrm{Id}_{f(K)}$ denotes the identity map on $f(K)$. Let $F := F_2 \circ F_1^{-1} \colon \mathbb{R}^d \to \mathbb{R}^d$. By definition, we have $F|_K = f|_K$.

Take a sufficiently large open ball B centered at 0 such that $K \subset \frac{1}{2}B$. Let $W \in \mathrm{Aff}$ such that $W^{-1}(x) = DF(0)^{-1}(x - F(0))$. Then by Lemma D.1 below, we conclude that there exists a compactly supported diffeomorphism $h \colon \mathbb{R}^d \to \mathbb{R}^d$ such that $W \circ h|_K = F|_K = f|_K$. ■

Here, we remark that Lemma D.1 below is a modified version of Lemma D.1 in Bernard et al. (2018), with a correction to make it explicit that the extended diffeomorphism is compactly supported. Their Lemma D.1 does not explicitly state that it is compactly supported, but by Theorem 1.4 in Section 8 of Hirsch (1976), it can be shown that the diffeomorphism is compactly supported. We provide the proof as follows:

Lemma D.1 *Let $r \geq 2$ be an integer, R a positive scalar, and $B_R \subset \mathbb{R}^d$ an open ball of radius R with origin 0, and let $f : B_R \to f(B_R) \subset \mathbb{R}^d$ be a C^r-diffeomorphism onto its image such that $f(0) = 0$ and $Df(0) = I$. Let $\varepsilon \in (0, R/2)$. Then there exists $h \in \mathrm{Diff}^r_{\mathrm{c}}$ such that $f(x) = h(x)$ for any $x \in B_{R-\varepsilon}$.*

Proof Put $\delta := \varepsilon/(2R - \varepsilon)$, and define $I_\delta := (-\delta, 1 + \delta)$. We define $F : B_{R-\frac{\varepsilon}{2}} \times I_\delta \to \mathbb{R}^d$ by

$$F(x,t) := \begin{cases} \frac{f(tx)}{t} & \text{if } t \neq 0, \\ x & \text{if } t = 0. \end{cases}$$

Here F is C^r, C^1 with respect to x, t, respectively. Let

$$U := \left\{ (F(x,t), t) : (x,t) \in B_{R-\frac{\varepsilon}{2}} \times I_\delta \right\} \subset \mathbb{R}^d \times \mathbb{R}$$

and let $F^\dagger : U \to B_{R-\frac{\varepsilon}{2}}$ such that $F(F^\dagger(x,t),t) = x$ for any $(x,t) \in U$. Here, $F^\dagger$ is the first component of the inverse of the map $(x,t) \mapsto (F(x,t),t)$ from $B_{R-\frac{\varepsilon}{2}} \times I_\delta$ onto U. We note that U is a bounded open subset in $\mathbb{R}^d \times \mathbb{R}$. Fix a compactly supported C^∞-function ϕ on $\mathbb{R}^d \times I_\delta$ such that for $(x,t) \in F\big(\overline{B_{R-\varepsilon}} \times [0,1]\big) \times [0,1]$, $\phi(x,t) = 1$, and for $(x,t) \notin U$, $\phi(x,t) = 0$. Then we define $H : \mathbb{R}^d \times I_\delta \to \mathbb{R}^d$ by

$$H(x,t) := \begin{cases} \phi(x,t) \frac{\partial F}{\partial t}(F^\dagger(x,t),t) & (x,t) \in U, \\ 0 & \text{otherwise.} \end{cases}$$

Since $F^\dagger$ is C^1 and for fixed $t \in I_\delta$, $\frac{\partial F}{\partial t}(\cdot, t)$ is C^r, there exists $L > 0$ such that for any $t \in I_\delta$, $\|H(x,t) - H(y,t)\| < L\|x - y\|$ with $x, y \in \mathbb{R}^d$. Thus the differential equation

$$\frac{dz}{dt} = H(z,t), \quad z(0) = x$$

has a unique solution $\phi_x(t)$. Then $h(x) := \phi_x(1)$ is the desired extension. ■

As a corollary, we can prove a C^r-version of Theorem 3.3 in Bernard et al. (2015):

Corollary D.2 *Let $r \geq 2$ be a positive integer and $f \in \mathcal{D}^r_U$. Assume U is C^∞-diffeomorphic to $\mathbb{R}^d$. Then, for any compact $K \subset U$, there exists a C^r-diffeomorphism F from U to $\mathbb{R}^d$ with $F(U) = \mathbb{R}^d$ such that*

$$F|_K = f|_K.$$

Proof Fix a C^r-diffeomorphism $g : U \to \mathbb{R}^d$. Let $\varepsilon > 0$ and take a sufficiently large R such that $g^{-1}(B_{R-\varepsilon})$ contains K, where B_R is the open ball of radius R with origin 0. By using Lemma D.1, there exists $h \in \mathrm{Diff}^r_{\mathrm{c}}$ and $W \in \mathrm{Aff}$ such that $h(x) = W \circ f \circ g^{-1}(x)$ for all $x \in B_{R-\varepsilon}$. As h is surjective mapping, $F := W^{-1} \circ h \circ g$ is the desired C^r-diffeomorphism from U onto $\mathbb{R}^d$. ■

D.2 Proof of Lemma 31

Proof [Proof of Lemma 31] Put $H^r := \{g_1 \circ \cdots \circ g_n : n \geq 1, g_1, \ldots, g_n \in \Xi^r\}$. First, we prove that H^r forms a subgroup of $\mathrm{Diff}^r_{\mathrm{c}}$. By definition, for any $g, h \in H^r$, it holds that $g \circ h \in H^r$. Also, H^r is closed under inversion; to see this, it suffices to show that Ξ^r is closed under inversion. Let $g = \Phi(\cdot, 1) \in \Xi^r$. Consider the map $\phi : \mathbb{R}^d \times U \to \mathbb{R}^d$ defined by $\phi(x, t) := \Phi(\cdot, t)^{-1}(x)$. It is easy to confirm that ϕ satisfies the conditions of Definition 20, hence $g^{-1} = \phi(\cdot, 1)$ is an element of Ξ^r. Note that ϕ is confirmed to be C^r on $\mathbb{R}^d \times U$ by applying the inverse function theorem (for example, Lang, 1985, Theorem 1 of Chapter I, Section 5) to $(t, \boldsymbol{x}) \mapsto (t, \Phi(\boldsymbol{x}, t))$.

Next, we prove that H^r is normal. To show that the subgroup generated by Ξ^r is normal, it suffices to show that Ξ^r is closed under conjugation. Take any $g \in \Xi^r$ and $h \in \mathrm{Diff}^r_{\mathrm{c}}$, and let Φ be a flow associated with g. Then, the function $\Phi' : \mathbb{R}^d \times U \to \mathbb{R}^d$ defined by $\Phi'(\cdot, s) := h^{-1} \circ \Phi(\cdot, s) \circ h$ is a flow associated with $h^{-1} \circ g \circ h$ satisfying the conditions in Definition 20, which implies $h^{-1} \circ g \circ h \in \Xi^r$, that is, Ξ^r is closed under conjugation.

Next, we prove that H^r is non-trivial by constructing an element of Ξ^r that is not the identity element. First, consider the case $d = 1$. Let $\tilde{v} : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a non-constant C^∞-function such that $\mathrm{supp}\, \tilde{v} \subset [0, 1]$ and $\tilde{v}^{(k)}(0) = 0$ for any $k \in \mathbb{N}$. Then define $v : \mathbb{R} \to \mathbb{R}$ by

$$v(x) = \begin{cases} \tilde{v}(|x|)\frac{x}{|x|} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0, \end{cases}$$

which is a C^∞-function on $\mathbb{R}$ with a compact support. Since v is Lipschitz continuous and C^∞, there exists $\mathrm{IVP}[v]$ that is a C^∞-function over $\mathbb{R} \times \mathbb{R}$; see Fact 4 and Hartman (2002, Chapter V, Corollary 4.1). Let $K_v \subset \mathbb{R}$ be a compact subset that contains $\mathrm{supp}\, v$. Then, by considering the ordinary differential equation by which $\mathrm{IVP}[v]$ is defined, we see that $\bigcup_{t \in \mathbb{R}} \mathrm{supp}\, \mathrm{IVP}[v](\cdot, t) \subset K_v$ and also that $\mathrm{IVP}[v](x, 0) = x$. We also have $\mathrm{IVP}[v](x, s + t) = \mathrm{IVP}[v](\mathrm{IVP}[v](x, s), t)$ for any $s, t \in \mathbb{R}$. In particular, we have $\mathrm{IVP}[v](\cdot, s)^{-1} = \mathrm{IVP}[v](\cdot, -s)$ for any $s \in \mathbb{R}$. Therefore, we have $\mathrm{IVP}[v](\cdot, 1) \in \Xi^r$. Since $v \not\equiv 0$, $\mathrm{IVP}[v](\cdot, 1)$ is not an identity map and thus Ξ^r is not trivial. Next, we consider the case $d \geq 2$. Take a C^∞-function $\phi \colon \mathbb{R} \to \mathbb{R}$ with $\mathrm{supp}\, \phi = [1, 2]$ and a nonzero skew-symmetric matrix A (that is, $A^\top = -A$) of size d, and let $X(x) := \phi(\|x\|)A$. We define a C^∞-map $\Phi \colon \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ by

$$\Phi(x, t) := \exp(tX(x))x.$$

Since $\exp(tX(x))$ is an orthogonal matrix for any $t \in \mathbb{R}$ and $x \in \mathbb{R}^d$, Φ is a C^∞-flow on $\mathbb{R}^d$. Now, it is enough to show that there exists a compact set $K_\Phi \subset \mathbb{R}^d$ satisfying $\cup_{t \in \mathbb{R}} \mathrm{supp}\, \Phi(\cdot, t) \subset K_\Phi$. Let $K_\Phi := \{x \in \mathbb{R}^d \mid \|x\| \leq 2\}$. Then the inclusion $\mathrm{supp}\, \Phi(\cdot, t) \subset K_\Phi$ holds for any $t \in \mathbb{R}$ since $X(x) = 0$ for $x \in \mathbb{R}^d \setminus K_\Phi$. ■

D.3 Proof of Lemma 35

Proof [Proof of Lemma 35] The proof is based on induction. Suppose that f is in the form of

$$f(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x}), x_{m+1}, \ldots, x_d).$$

By means of induction with respect to m, we prove that there exist compactly supported C^r-diffeomorphisms $F_1, \ldots, F_m : \mathbb{R}^d \to \mathbb{R}^d$ in the forms of $F_i(\boldsymbol{x}) := (x_1, \ldots, x_{i-1}, h_i(\boldsymbol{x}), x_{i+1}, \ldots, x_d)$ for some $h_i : \mathbb{R}^d \to \mathbb{R}$ such that $f = F_1 \circ \cdots \circ F_m$.

In the case of $m = 1$, the above is clear. Assume that the statement is true in the case of any $k < m$. Define

$$\begin{aligned} F(x_1, \ldots, x_d) &:= (x_1, \ldots, x_{m-1}, f_m(\boldsymbol{x}), x_{m+1}, \ldots, x_d), \\ \tilde{f} &:= f \circ F^{-1}. \end{aligned}$$

Note that F is a compactly supported C^r-diffeomorphism from $\mathbb{R}^d$ to $\mathbb{R}^d$. In fact, compactly supportedness and surjectivity of F comes from the compactly supportedness of f. Moreover, since we have $\det DF_x = \frac{\partial f_m}{\partial x_m}(x) \neq 0$ for any $x \in \mathbb{R}^d$ by the assumption on f, F is injective and is a C^r-diffeomorphism from $\mathbb{R}^d$ to $\mathbb{R}^d$ by inverse function theorem. Therefore, $\tilde{f}$ is also a C^r-diffeomorphism from $\mathbb{R}^d$ to $\mathbb{R}^d$. We show that $\tilde{f}$ is of the form $\tilde{f}(\boldsymbol{x}) = (g_1(\boldsymbol{x}), \cdots, g_{m-1}(\boldsymbol{x}), x_m, \cdots, x_d)$ for some C^r-functions $g_i \colon \mathbb{R}^d \to \mathbb{R}$ $(i = 1, \cdots, m-1)$ satisfying $\det \Delta_k^{\tilde{f}}(x) \neq 0$ for any $x \in \mathbb{R}^d$ and $k \in [d]$. From Lemma D.3, there exist $g_i, h \in C^r(\mathbb{R}^d)$ $(i = 1, \cdots, m)$ such that

$$\begin{aligned} f^{-1}(\boldsymbol{x}) &= (g_1(\boldsymbol{x}), \cdots, g_m(\boldsymbol{x}), x_{m+1}, \cdots, x_d) \\ F^{-1}(\boldsymbol{x}) &= (x_1, \cdots, x_{m-1}, h(\boldsymbol{x}), x_{m+1}, \cdots, x_d). \end{aligned}$$

Then we have

$$\begin{aligned} \tilde{f}^{-1}(\boldsymbol{x}) = F \circ f^{-1}(\boldsymbol{x}) &= (g_1(\boldsymbol{x}), \cdots, g_{m-1}(\boldsymbol{x}), f_m(f^{-1}(\boldsymbol{x})), x_{m+1}, \cdots, x_d) \\ &= (g_1(\boldsymbol{x}), \cdots, g_{m-1}(\boldsymbol{x}), x_m, \cdots, x_d). \end{aligned}$$

Therefore, from Lemma D.3, $\tilde{f}$ is of the following form

$$\tilde{f}(x) = f \circ F^{-1}(x) = (f_1 \circ F^{-1}(x), \cdots, f_{m-1} \circ F^{-1}(x), x_m, \cdots, x_d).$$

Moreover, by the form of F^{-1} and f, we have $D\tilde{f}(x) = Df(F^{-1}(x)) \circ DF^{-1}(x)$ and

$$Df = \begin{pmatrix} A & \\ & I \end{pmatrix}, \quad D(F^{-1}) = \begin{pmatrix} I_{m-1} & & \\ \frac{\partial h}{\partial x_1} & \cdots & \frac{\partial h}{\partial x_d} \\ & & I_{d-m} \end{pmatrix}$$

for some $A \in M(m, \mathbb{R})$ with all the trailing principal minors nonzero. Therefore, we obtain $\det \Delta_k^f(x) \neq 0$ for any $x \in \mathbb{R}^d$ and $k \in [d]$. Here, by the assumption of the induction, there exist compactly supported C^r-diffeomorphisms $F_i \colon \mathbb{R}^d \to \mathbb{R}^d$ and $h_i \in C^r(\mathbb{R}^d)$ $(i = 1, \cdots, m-1)$ such that

$$\tilde{f} = F_1 \circ \cdots \circ F_{m-1}, \ F_i(\boldsymbol{x}) = (x_1, \cdots x_{i-1}, h_i(x), x_{i+1}, \cdots, x_d).$$

Thus $f = \tilde{f} \circ F$ has the desired form. ■

Lemma D.3 *Let $1 \leq r \leq \infty$ and $f\colon \mathbb{R}^d \to \mathbb{R}^d$ C^r-diffeomorphism of the form*

$$f(\boldsymbol{x}) := (f_1(\boldsymbol{x}), \cdots, f_m(\boldsymbol{x}), x_{m+1}, \cdots, x_d),$$

where $f_i\colon \mathbb{R}^d \to \mathbb{R}$ belongs to $C^r(\mathbb{R}^d)$ $(i = 1, \cdots, m)$. Then the inverse map f^{-1} becomes of the form

$$f^{-1}(\boldsymbol{x}) = (g_1(\boldsymbol{x}), \cdots, g_m(\boldsymbol{x}), x_{m+1}, \cdots x_d),$$

where $g_i : \mathbb{R}^d \to \mathbb{R}$ belongs to $C^r(\mathbb{R}^d)$ for $i = 1, \cdots, m$.

Proof We write $f^{-1}(\boldsymbol{x}) = (h_1(\boldsymbol{x}), \cdots, h_d(\boldsymbol{x}))$, where $h_i \in C^r(\mathbb{R}^d)$ $(i = 1, \cdots, d)$. Then by the definition of the inverse map, the identity

$$(x_1, \cdots, x_d) = f \circ f^{-1}(\boldsymbol{x}) = (f_1(h_1(\boldsymbol{x})), \cdots, f_m(h_m(\boldsymbol{x})), h_{m+1}(\boldsymbol{x}), \cdots, h_d(\boldsymbol{x}))$$

holds for any $\boldsymbol{x} \in \mathbb{R}^d$, which implies that we obtain $h_i(x) = x_i$ $(i = m+1, \cdots, d)$. This completes the proof of the lemma. ■

D.4 Proof of Theorem 26

Here, we provide the proof of Theorem 26.

D.4.1 Proof of the statement 1

Here, we prove a key ingredient (Corollary D.5) for the generalized equivalence of the statement 1 of Theorem 26. The statement is the direct consequence of Corollary D.5. In this section, we always assume $p \in [1, \infty)$. For any finite subset $S \subset \mathbb{R}^d$, we denote by $\mathrm{Map}(S, \mathbb{R}^d)$ the set of maps from S to $\mathbb{R}^d$ and equip it with the supremum topology. Then, for any finite subset $S \subset \mathbb{R}^d$, a set of bijections $\mathcal{M}$, and a subset $\mathcal{F} \subset \mathrm{Map}(S, \mathbb{R}^d)$, $\mathcal{M}$ is an L^∞-universal approximator for $\mathcal{F}$ if $\mathcal{M}$ is a $\mathrm{Map}(S, \mathbb{R}^d)$-universal approximator for $\mathcal{F}$.

First, we prove the following lemma, which is essentially proved by Li et al. (2022).

Lemma D.4 *Let $\mathcal{M}$ be a set of bijections from $\mathbb{R}^d$ to $\mathbb{R}^d$. We assume that $\mathcal{M}$ satisfies the following three conditions:*

(1) all functions of $\mathcal{M}$ are locally Lipschitz.

(2) for any finite subset $S \subset \mathbb{R}^d$, $\mathcal{M}$ is an L^∞-universal approximator for the set of all injections from S to $\mathbb{R}^d$.

(3) $\mathcal{M}$ is an L^p-universal approximator for the subset

$$\left\{ f : [0,1]^d \to \mathbb{R}^d : f(x_1, \ldots, x_d) = (f_i(x_i))_{i=1}^d \text{ and } f_i \text{ is nondecreasing} \right\}.$$

Then, $\mathcal{M} \circ \mathcal{M} := \{g \circ f : g, f \in \mathcal{M}\}$ is an L^∞-universal approximator for $C^0([0,1]^d, \mathbb{R}^d)$, where $C^0(U, V)$ is the set of continuous maps from U to V.

Proof Let $\varepsilon > 0$ be a positive number. Let $f \in C^0([0,1]^d, \mathbb{R}^d)$, m be a positive integer, and $K \subset [0,1]^d$. For any $\alpha \in \mathbb{Z}^d_{\geq 0}$ with $|\alpha| \geq m$, let

$$\Delta_\alpha := \prod_{i=1}^{d} \left[\frac{\alpha_i - 1}{m}, \frac{\alpha_i}{m} \right) \subset \mathbb{R}^d,$$
$$p_\alpha := \left(\frac{\alpha_1 - 1}{m}, \ldots, \frac{\alpha_m - 1}{m} \right).$$

Put $y_\alpha := f(p_\alpha)$. We define

$$H_m(x_1, \ldots, x_m) := \left(\sum_{k=0}^{m} \frac{k}{m} \mathbf{1}_{[k/m, k+1/m)}(x_i) \right)_{i=1}^{m}.$$

By (2), there exists $\psi_m \in \mathcal{M}$ such that

$$\|\psi_m(p_\alpha) - y_\alpha\| < 1/m$$

for any α with $|\alpha| \leq m$. Since f is continuous, we see that

$$\sup_{|\alpha| \leq m} \sup_{x \in \Delta_\alpha} \|\psi_m(p_\alpha) - f(x)\| < \varepsilon/2$$

if we take m sufficiently large. Let L_m be the Lipschitz constant for $\psi_m|_K$. by (3), there exists $g_m \in \mathcal{M}$ such that

$$\|g_m - H_m\|_{K,0,p} < \frac{\varepsilon}{2L_m}.$$

Therefore, we have

$$\begin{aligned} \|\psi_m \circ g_m - f\|_{K,0,p} &\leq \|\psi_m \circ g_m - \psi_m \circ H_m\|_{K,0,p} + \|\psi_m \circ H_m - f\|_{K,0,p} \\ &\leq L_m \|g_m - H_m\|_{K,0,p} + \sup_{|\alpha| \leq m} \sup_{x \in \Delta_\alpha} \|\psi_m(p_\alpha) - f(x)\| \\ &< \varepsilon. \end{aligned}$$

■

Then, we have the following corollary:

Corollary D.5 *Assume $d \geq 2$. Let $U \subset \mathbb{R}^d$ be an open subset. Then, $\mathcal{D}^\infty_{\mathbb{R}^d}$ is an L^p-universal approximator for $C^0(U, \mathbb{R}^d)$ for any open subset $U \subset \mathbb{R}^d$.*

Proof it suffices to show that for any $f \in C^0(U, \mathbb{R}^d)$, $\varepsilon > 0$, and compact subset $K \subset U$, there exists $g \in \mathcal{D}^\infty_{\mathbb{R}^d}$ such that

$$\|g - f\|_{K,0,\infty} < \varepsilon.$$

We may assume $U = \mathbb{R}^d$ and $K = [0,1]^d$ by extending f as a continuous function on $\mathbb{R}^d$. Then, we can easily see that $\mathcal{D}^\infty_{\mathbb{R}^d}$ satisfies the three conditions in Lemma D.4. The assumption $d \geq 2$ is used here for the second condition of Lemma D.4. Thus, the assertion

follows from Lemma D.4. ■

Proof [Proof of the statement 1 of Theorem 26] It is a direct consequence of Corollary D.5. ■

Remark D.6 *Brenier and Gangbo (2003) proves a similar result on the approximation of L^∞-function on bounded domain with a diffeomorphism in L^p-space.*

D.4.2 Proof of the statement 2

Here, we prove the statement 2 of Theorem 26. The key ingredients are the annulus theorem by Kirby (1969) and the approximation result by Connell (1963). We describe the precise statement of the Annulus theorem:

Fact 7 (Annulus Theorem) *Let $f, g : S^{d-1} \to \mathbb{R}^d$ be locally flat embeddings with $f(S^{d-1})$ inside the bounded component of $\mathbb{R}^d \setminus g(S^{d-1})$. Then, the closed region that is bounded by $f(S^{d-1})$ and $g(S^{d-1})$ is homeomorphic to $S^{d-1} \times [0, 1]$.*

Here, a map $h : S^{d-1} \to \mathbb{R}^d$ is locally flat if for any $x \in h(S^{d-1})$, there exists a open set $U \subset \mathbb{R}^d$ with isomorphism $\iota : U \cong \mathbb{R}^d$ such that $x \in U$ and $\iota(U \cap h(S^{d-1})) = \mathbb{R}^{d-1} \times \{0\}$. This fact proved by Radó (1925) for $d = 2$, Moise (1952) for $d = 3$, Quinn (1982) for $d = 4$, and Kirby (1969) for $d \geq 5$.

By means of Fact 7, we obtain the following lemma:

Lemma D.8 *Let $U \subset \mathbb{R}^d$ be an open set homeomorphic to $\mathbb{R}^d$, $K \subset U$ a compact set, and $f \in \mathcal{D}^0_U$. Then, there exist $h \in \mathrm{Diff}^0_{\mathrm{c}}$ and an affine transform $W \in \mathrm{Aff}$ such that $W \circ h|_K = f|_K$.*

Proof It suffices to prove that there exists $h \in \mathrm{Diff}^0_{\mathrm{c}}$ such that $f|_K = h|_K$. We may assume that f is orientation preserving by multiplying suitable , $U = \mathbb{R}^d$ and $K = D_r$ $(r > 0)$, where $D_r := \{x \in \mathbb{R}^d : |x| \leq r\}$ be the closed ball centered at 0 of radius r. Fix sufficiently large $R > 0$ such that $f(D_r) \subset D_R$. Then, by Fact 7, there exists homomorphism $\Phi := (\phi_1, \phi_2) : D_R \setminus f(B_r) \cong \partial D_R \times [r, R]$, where B_r is the open ball centered at 0 of radius r and ϕ_2 satisfies that for any $x \in D_R \setminus f(B_r)$, $\phi_2(x) = r$ (resp. R) if and only if $x \in f(\partial D_r)$ (resp. ∂D_R). Define

$$F : \partial D_R \times [r, R] \to \mathbb{R}^d; (x, t) \to \Phi^{-1}\left(\phi_1\big(f(rR^{-1}x)\big), t\right).$$

Since f is orientation preserving and F provides isotopy between $f(rR^{-1}\cdot) : \partial D_R \to f(\partial D_r)$ and $F(\cdot, R) : \partial D_R \to \partial D_R$, $F(\cdot, R)$ is also orientation preserving. It is known that the orientation preserving map on the sphere is isotopic to the identity (see, for example, Livingston 2021, Theorem 1). Let $\tilde{F} : \partial D_R \times [0, 1] \to \partial D_R$ be the isotopy between $F(\cdot, R)$ and the identity on ∂D_R. Then, the desired compactly support homeomorphism $h \in \mathrm{Diff}^0_{\mathrm{c}}$ is defined as follows:

$$h(x) := \begin{cases} f(x) & \text{if } x \in D_r, \\ F(Rx/|x|, |x|) & \text{if } x \in D_R \setminus D_r, \\ R^{-1}|x|\tilde{F}(Rx/|x|, |x| - R) & \text{if } x \in D_{R+1} \setminus D_R, \\ x & \text{if } x \notin D_{R+1}. \end{cases}$$

■

Then, we have the following result:

Lemma D.9 *Assume $d \geq 7$. For any open subset $U \subset \mathbb{R}^d$ homeomorphic to $\mathbb{R}^d$, $\mathcal{D}^\infty_{\mathbb{R}^d}$ is a L^∞-universal approximator for $\mathcal{D}^0_U$.*

Proof By Lemma D.8, it suffices to show that $\mathcal{D}^\infty_{\mathbb{R}^d}$ is a L^∞-universal approximator for $\mathrm{Diff}^0_{\mathrm{c}}$. It follows from Connell (1963, Theorem 4) as a compactly supported homeomorphism is stable (we say a homeomorphism is stable if the homeomorphism is a finite composition of homeomorphisms each of which is the identity on a nonempty open subset). ■

Proof [Proof of the statement 2] It is a direct consequence of Lemma D.9. ■

Appendix E. Universality of coupling-flow based INNs

In this section, we give the proof for the universal approximation properties of certain CF-INNs.

E.1 Using permutation matrices instead of Aff in the definition of $\mathrm{INN}_{\mathcal{G}}$

In terms of representation power, there is no essential difference if we substitute the general linear group in Definition 6 with the permutation group. It comes from the fact that one can express the elementary operation matrices using affine coupling flows and permutations. More formally, we have the following proposition.

Proposition E.1 *Assume that $\mathcal{H}$ includes all the functions $\mathbb{R}^{d-1} \to \mathbb{R}$ of the following forms: $x \mapsto -x \cdot e_i$, $x \mapsto x \cdot e_i$, and $x \mapsto b$ (constant map), where $b \in \mathbb{R}^{d-1}$ and $i = 1, \cdots, d-1$. Then, we have*

$$\mathrm{INN}_{\mathcal{H}\text{-ACF}} = \{W_1 \circ g_1 \circ \cdots \circ W_n \circ g_n \ : \ g_i \in \mathcal{H}\text{-ACF}, W_i \in \mathfrak{S}_d\}, \tag{10}$$

where $\mathfrak{S}_d$ is the permutation group of degree d.

Proof Since the multiplication of any permutation matrix is an affine transformation, the right-hand side of (10) is included in the left-hand side.

We prove the converse inclusion. Since any translation operator (that is, the addition of a constant vector) can be easily represented by the elements of $\mathcal{H}$-ACF and permutations, it is enough to show that any element of $\mathrm{GL}(d, \mathbb{R})$ can be realized by a finite composition of elements of $\mathcal{H}$-ACF and $\mathfrak{S}_d$. To show that, it is sufficient to consider only the elementary matrices. Row switching comes from $\mathfrak{S}_d$. Moreover, element-wise sign flipping can be described by a composition of finite elements of $\mathcal{H}$-ACF. To see this, first, observe that

$$\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

holds. Here, the linear transforms

$$\begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

are realized by the $\mathcal{H}$-ACF layers

$$(x, y) \mapsto (x, y - x), \quad (x, y) \mapsto (x, y + x),$$

respectively. Now, any lower triangular matrix with positive diagonals can be described by a composition of finite elements of $\mathcal{H}$-ACF. Therefore, any diagonal matrix whose components are ± 1 can be described by a composition of elements in $\mathcal{H}$-ACF and $\mathfrak{S}_d$. Therefore, any affine transform is an element of the right-hand side of (10). ■

This result implies that employing Aff in Definition 6 instead of the permutation matrices is not an essential requirement for the universal approximation properties to hold. For this reason, we believe that the empirically reported difference in the performances of Glow (Kingma and Dhariwal, 2018) and RealNVP (Dinh et al., 2017) is mainly in the efficiency of approximation rather than the capability of approximation.

E.2 Affine coupling flows (ACFs)

In this section, we provide the proof details of Theorem 42 in the main text. We note that we are assuming $d \geq 2$.

E.2.1 Proof of Theorem 42: L^p-universality of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$

In this subsection, we prove the following lemma to construct an approximator for an arbitrary element of $\mathcal{S}_{\mathrm{c}}^0$ (hence for $\mathcal{S}_{\mathrm{c}}^\infty$) within $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$. It is based on Lemma E.3 proved in Section E.2.2, which corresponds to a special case. Figure 2 illustrates the proof technique for Lemma E.2.

Lemma E.2 (L^p-universality of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$ for compactly supported $\mathcal{S}_{\mathrm{c}}^\infty$) *Let $p \in [1, \infty)$. Assume $\mathcal{H}$ is an L^∞-universal approximator for $C_c^\infty(\mathbb{R}^{d-1})$ and that it consists of piecewise C^1-functions. Let $f \in \mathcal{S}_{\mathrm{c}}^0$, $\varepsilon > 0$, and $K \subset \mathbb{R}^d$ be a compact subset. Then, there exists $g \in \mathrm{INN}_{\mathcal{H}\text{-ACF}}$ such that $\|f - g\|_{K,0,p} < \varepsilon$.*

Proof Since we can take $a > 0$, $b \in \mathbb{R}$ satisfying $aK + b \subset [0, 1]^d$, it is enough to prove the assertion for the case $K = [0, 1]^d$.

Next, we show that we can assume that for any $(\boldsymbol{x}, y) \in \mathbb{R}^d$, $u(\boldsymbol{x}, 0) = 0$ and $u(\boldsymbol{x}, 1) = 1$ for any $\boldsymbol{x} \in \mathbb{R}^{d-1}$. Since $u(\boldsymbol{x}, \cdot)$ is a homeomorphism, we have $u(\boldsymbol{x}, 0) \neq u(\boldsymbol{x}, 1)$ for any $x \in \mathbb{R}$. By the continuity of f, either of $u(\boldsymbol{x}, 0) > u(\boldsymbol{x}, 1)$ for all $\boldsymbol{x} \in [0, 1]^{d-1}$ or $u(\boldsymbol{x}, 0) < u(\boldsymbol{x}, 1)$ for all $x \in [0, 1]^{d-1}$ holds. Without loss of generality, we assume the latter case holds (if the former one holds, we just switch $u(\boldsymbol{x}, 0)$ and $u(\boldsymbol{x}, 1)$). We define $s(\boldsymbol{x}) = -\log(u(\boldsymbol{x}, 1) - u(\boldsymbol{x}, 0))$ and $t(\boldsymbol{x}) = -u(\boldsymbol{x}, 0)(u(\boldsymbol{x}, 1) - u(\boldsymbol{x}, 0))^{-1}$. By direct computation, we have

$$\Psi_{d-1,s,t} \circ f(\boldsymbol{x}, y) = \left(\boldsymbol{x}, \frac{u(\boldsymbol{x}, y) - u(\boldsymbol{x}, 0)}{u(\boldsymbol{x}, 1) - u(\boldsymbol{x}, 0)}\right) =: (\boldsymbol{x}, u_0(\boldsymbol{x}, y)).$$

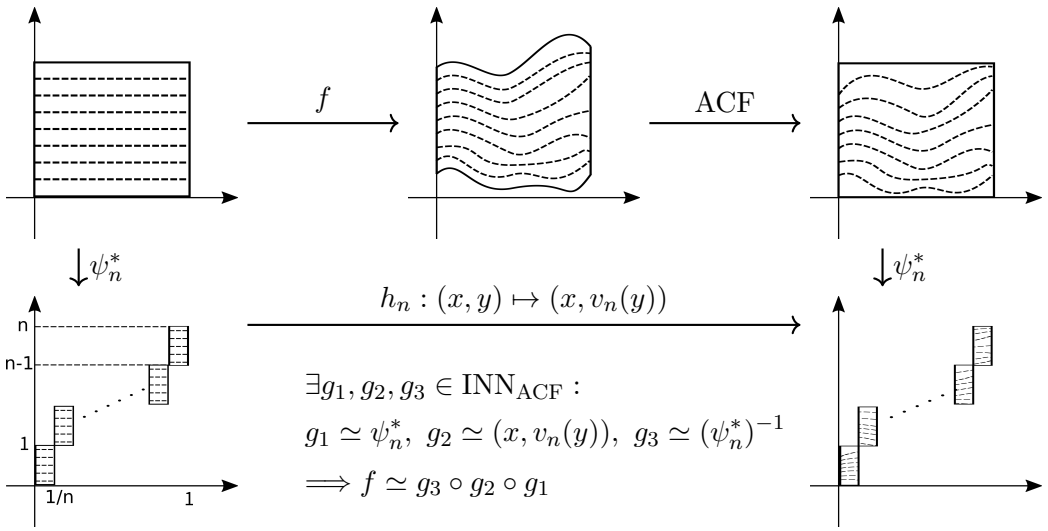


Figure 2: Illustration of the proof technique for Lemma E.2. The symbol $\simeq$ indicates approximation to arbitrary precision. The figure is taken from Teshima et al. (2020a, Figure 2) with the authors' permission.

In particular, $\Psi_{s,t} \circ f(\boldsymbol{x}, 0) = (\boldsymbol{x}, 0)$ and $\Psi_{s,t} \circ s(\boldsymbol{x}, 1) = (\boldsymbol{x}, 1)$ hold. , and the map $y \mapsto u_0(\boldsymbol{x}, y)$ is a diffeomorphism for each $\boldsymbol{x}$. Thus if we prove the existence of an approximator for $\Psi_{s,t} \circ f$, by Proposition B.1, we can arbitrarily approximate f itself.

For $\underline{k} := (k_1, \ldots, k_{d-1}) \in \mathbb{Z}^{d-1}$ and $n \in \mathbb{N}$, we define $(\underline{k})_n := \sum_{i=1}^{d} k_i n^{i-1} \in \{0, \ldots, n^d - 1\}$, that is, $\underline{k}$ is the n-adic expansion of $(\underline{k})_n$. For any $n \in \mathbb{N}$, define the following discontinuous ACF: $\psi_n \colon [0,1]^d \to [0,1]^{d-1} \times [0, n^d]$ by

$$\psi_n(\boldsymbol{x}, y) := \left(\boldsymbol{x}, y + \sum_{k_1, \cdots, k_{d-1} = 0}^{n-1} (\underline{k})_n 1_{\Delta^n_{\underline{k}+1}}(\boldsymbol{x}) \right),$$

where $\underline{k} := (k_1, \ldots, k_d)$ and $\underline{k} + 1 := (k_1 + 1, \ldots, k_d + 1)$. We take an increasing function $v_n \colon \mathbb{R} \to \mathbb{R}$ that is smooth outside finite points such that

$$v_n(z) := \begin{cases} u\left(\frac{k_1}{n}, \cdots, \frac{k_{d-1}}{n}, z - (\underline{k})_n\right) + (\underline{k})_n & \text{if } z \in [(\underline{k})_n, (\underline{k})_n + 1) \\ z & \text{if } z \notin [0, n^d). \end{cases}$$

We consider maps h_n on $[0,1]^{d-1} \times [0, n^d]$ and $f_n : [0,1]^d \to [0,1]^d$ defined by

$$h_n(\boldsymbol{x}, z) := (\boldsymbol{x}, v_n(z)),$$
$$f_n := \psi_n^{-1} \circ h_n \circ \psi_n.$$

Then we have the following claim.
Claim. For all $k_1, \cdots, k_{d-1} = 0, \cdots, n-1$, we have

$$f_n(\boldsymbol{x}, y) = \left(\boldsymbol{x}, u\left(\frac{k_1}{n}, \ldots, \frac{k_{d-1}}{n}, y\right)\right)$$

on $\prod_{i=1}^{d-1}[\frac{k_i}{n}, \frac{k_i+1}{n}) \times [0,1)$.
In fact, we have

$$\begin{aligned} f_n(\boldsymbol{x}, y) &= \psi_n^{-1} \circ h_n \circ \psi_n(\boldsymbol{x}, y) \\ &= \psi_n^{-1} \circ h_n(\boldsymbol{x}, y + (\underline{k})_n) \\ &= \psi_n^{-1}(\boldsymbol{x}, v_n(y + (\underline{k})_n)) \\ &= \psi_n^{-1}\left(\boldsymbol{x}, u\left(\frac{k_1}{n}, \ldots, \frac{k_{d-1}}{n}, y\right) + (\underline{k})_n\right) \\ &= \left(\boldsymbol{x}, u\left(\frac{k_1}{n}, \ldots, \frac{k_{d-1}}{n}, y\right)\right). \end{aligned}$$

Therefore, the claim above has been proved. Hence we see that $\|f - f_n\|_{K,0,\infty} \to 0$ as $n \to \infty$. By Lemma E.3 below and the universal approximation property of $\mathcal{H}$, for any compact subset K and $\varepsilon > 0$, there exist $g_1, g_2, g_3 \in \mathrm{INN}_{\mathcal{H}\text{-ACF}}$ such that $\left\|g_1 - \psi_n^{-1}\right\|_{K,0,p} < \varepsilon$, $\|g_2 - h_n\|_{K,0,p} < \varepsilon$, and $\|g_3 - \psi_n\|_{K,0,p} < \varepsilon$. Thus by Proposition B.1, for any compact K and $\varepsilon > 0$, there exists $g \in \mathrm{INN}_{\mathcal{H}\text{-ACF}}$ such that $\|g - f\|_{K,0,p} < \varepsilon$. ■

Proof [Proof of Theorem 42] The assertion follows from Lemma E.2 and Theorem 26. ■

E.2.2 Special case: Approximation of coordinate-wise independent transformation

In this section, we show the lemma claiming that special cases of single-coordinate transformations, namely coordinate-wise independent transformations, can be approximated by the elements of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$ given sufficient representational power of $\mathcal{H}$.

Lemma E.3 *Let $p \in [1, \infty)$. Assume $\mathcal{H}$ is an L^∞-universal approximator for $C_c^\infty(\mathbb{R}^{d-1})$ and that it consists of piecewise C^1-functions. Let $u : \mathbb{R} \to \mathbb{R}$ be a continuous and increasing function. Let $f : \mathbb{R}^d \to \mathbb{R}^d; (\boldsymbol{x}, y) \mapsto (\boldsymbol{x}, u(y))$ where $\boldsymbol{x} \in \mathbb{R}^{d-1}$ and $y \in \mathbb{R}$. For any compact subset $K \subset \mathbb{R}^d$ and $\varepsilon > 0$, there exists $g \in \mathrm{INN}_{\mathcal{H}\text{-ACF}}$ such that $\|f - g\|_{K,0,p} < \varepsilon$.*

Proof We may assume without loss of generality, in light of Lemma E.5, that u is a C^∞-diffeomorphism on $\mathbb{R}$ and that the inequality $u'(y) > 0$ holds for any $y \in \mathbb{R}$. Furthermore, we may assume that u is compactly supported, that is, $u(y) = y$ outside a compact subset of $\mathbb{R}$, without loss of generality because we can take a compactly supported diffeomorphism $\tilde{u}$ and $a, b \in \mathbb{R}$ $(a \neq 0)$ such that $a\tilde{u} + b = u$ on any compact set containing K by Lemma 29, and the scaling a and the offset b can be realized by the elements of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$.

Fix $\delta \in (0,1)$. We define the following functions:

$$\begin{aligned}
\psi_0(\boldsymbol{x}, y) &:= (\boldsymbol{x}_{\leq d-2}, u'(y)x_{d-1}, y) \\
&= (\boldsymbol{x}_{\leq d-2}, \exp(\log u'(y))x_{d-1}, y), \\
\psi_1(\boldsymbol{x}, y) &:= \left(\boldsymbol{x}_{\leq d-2}, x_{d-1} + \delta^{-1}(u(y) - y), y\right), \\
\psi_2(\boldsymbol{x}, y) &:= (\boldsymbol{x}_{\leq d-2}, x_{d-1}, y + \delta x_{d-1}), \\
\psi_3(\boldsymbol{x}, y) &:= \left(\boldsymbol{x}_{\leq d-2}, x_{d-1} - \delta^{-1}(y - u^{-1}(y)), y\right),
\end{aligned}$$

where we denote $\boldsymbol{x} = (x_1, \ldots, x_{d-1}) \in \mathbb{R}^{d-1}$. First, we show that $\|f - \psi_3 \circ \psi_2 \circ \psi_1 \circ \psi_0\|_{K,0,\infty} \to 0$ as $\delta \to 0$. By direct computation, we have

$$\begin{aligned}
\psi_3 \circ \psi_2 \circ \psi_1(\boldsymbol{x}, y) &= \psi_3 \circ \psi_2(\boldsymbol{x}_{\leq d-2}, x_{d-1} + \delta^{-1}(u(y) - y), y) \\
&= \psi_3(\boldsymbol{x}_{\leq d-2}, x_{d-1} + \delta^{-1}(u(y) - y), y + \delta(x_{d-1} + \delta^{-1}(u(y) - y))) \\
&= \psi_3(\boldsymbol{x}_{\leq d-2}, x_{d-1} + \delta^{-1}(u(y) - y), \delta x_{d-1} + u(y)) \\
&= (\boldsymbol{x}_{\leq d-2}, x_{d-1} - \delta^{-1}(\delta x_{d-1} + u(y) - u^{-1}(\delta x_{d-1} + u(y))), \delta x_{d-1} + u(y)) \\
&= (\boldsymbol{x}_{\leq d-2}, \delta^{-1}u^{-1}(\delta x_{d-2} + u(y)) - \delta^{-1}y, u(y) + \delta x_{d-1}),
\end{aligned}$$

where $\boldsymbol{x} = (x_1, \ldots, x_{d-1}) \in \mathbb{R}^{d-1}$. Since $u \in C^\infty([-r, r])$ where $r = \max_{(\boldsymbol{x},y)\in K} |y|$, by applying Taylor's theorem, there exists a function $R(\boldsymbol{x}, y; \delta)$ and $C = C([-r, r], u) > 0$ such that

$$u^{-1}(u(y) + \delta x) = y + u'(y)^{-1}\delta x + R(\boldsymbol{x}, y; \delta)(\delta x)^2 \quad \text{and} \quad \sup_{\delta\in(0,1)} |R(\boldsymbol{x}, y; \delta)| \leq C$$

for all $(\boldsymbol{x}, y) \in K$. Therefore, we have

$$\psi_3 \circ \psi_2 \circ \psi_1 \circ \psi_0(\boldsymbol{x}, y) = (\boldsymbol{x}, u(y)) + \delta(R(\boldsymbol{x}, u'(y)x_{d-1}; \delta)\boldsymbol{x}_{\leq d-1}, u'(y)x_{d-1}).$$

For any compact subset K, the last term uniformly converges to 0 as $\delta \to 0$ on K.

Assume δ is taken to be small enough. Now, we approximate $\psi_3 \circ \cdots \circ \psi_0$ by the elements of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$. Since u is a compactly-supported C^∞-diffeomorphism on $\mathbb{R}$, the functions $(\boldsymbol{x}_{\leq d-2}, y) \mapsto \log u'(y)$, $(\boldsymbol{x}_{\leq d-2}, y) \mapsto u(y) - y$, and $(\boldsymbol{x}_{\leq d-2}, y) \mapsto y - u^{-1}(y)$, each appearing in ψ_0, ψ_1, ψ_3, respectively, belong to $C_c^\infty(\mathbb{R}^{d-1})$. On the other hand, ψ_2 can be realized by $\mathrm{GL} \subset \mathrm{Aff}$. Therefore, combining the above with the fact that $\mathcal{H}$ is a L^∞-universal approximator for $C_c^\infty(\mathbb{R}^{d-1})$, we have that for any compact subset $K' \subset \mathbb{R}^d$ and any $\varepsilon > 0$, there exist $\phi_0, \ldots, \phi_3 \in \mathrm{INN}_{\mathcal{H}\text{-ACF}}$ such that $\|\psi_i - \phi_i\|_{K',0,\infty} < \varepsilon$. In particular, we can find $\phi_0, \ldots, \phi_3 \in \mathrm{INN}_{\mathcal{H}\text{-ACF}}$ such that $\|\psi_i - \phi_i\|_{K',0,p} < \varepsilon$.

Now, recall that $\mathcal{H}$ consists of piecewise C^1-functions as well as ψ_i $(i = 0, \ldots, 3)$. Moreover, ψ_0, ψ_1, ψ_3 are compactly supported while $\psi_2 \in \mathrm{GL}$, hence they are Lipschitz continuous outside a bounded open subset. Therefore, by Proposition B.1, we have the assertion of the lemma.

■

The following Lemma E.5 is used above when reducing the approximation problem from $\mathcal{S}_{\mathrm{c}}^2$ to $\mathcal{S}_{\mathrm{c}}^\infty$.

Definition E.4 *We say that a map $f : \mathbb{R}^d \to \mathbb{R}$ is* last-increasing *(resp.* last-non-decreasing*) if, for any $(a_1, \dots, a_{d-1}) \in \mathbb{R}^{d-1}$, the function $f(a_1, \dots, a_{d-1}, x)$ is strictly increasing (resp. non-decreasing) with respect to x.*

Lemma E.5 *Let $r \geq 0$ be an integer and let $p \in [1, \infty]$. Let $\tau \colon \mathbb{R}^d \to \mathbb{R}$ be a last-non-decreasing measurable function. We assume that τ is locally $C^{r-1,1}$-function if $r \geq 1$ or locally L^∞ if $r = 0$. Then for any compact subset $K \subset \mathbb{R}^d$ and any $\varepsilon > 0$, there exists a last-increasing C^∞-function $\tilde{\tau} \colon \mathbb{R}^d \to \mathbb{R}$ satisfying*

$$\|\tau - \tilde{\tau}\|_{K,r,p} < \varepsilon.$$

Proof Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a compactly supported non-negative C^∞-function with $\int |\phi(x)| dx = 1$ such that for any $(a_1, \dots, a_{d-1}) \in \mathbb{R}^{d-1}$, the function $\phi(a_1, \dots, a_{d-1}, x)$ of x is even and decreasing on $\{x > 0 : \phi(a_1, \dots, a_{d-1}, x) > 0\}$. For $t > 0$, we define $\phi_t(x) := t^{-d}\phi(x/t)$. Then we see that $\tau_t := \phi_t * \tau$ is a C^∞-function. We take any $\boldsymbol{a} \in \mathbb{R}^{d-1}$. We verify that $\tau_t(\boldsymbol{a}, x_d)$ is strictly increasing with respect to x_d. Take any $x_d, x_d' \in \mathbb{R}$ satisfying $x_d > x_d'$. Since τ is strictly increasing, we have

$$\tau_t(\boldsymbol{a}, x_d) - \tau_t(\boldsymbol{a}, x_d') = \int_{\mathbb{R}^d} \phi_t(x)(\tau((\boldsymbol{a}, x_d) - x) - \tau((\boldsymbol{a}, x_d') - x))dx > 0.$$

Thus for any $(a_1, \dots, a_{d-1}) \in \mathbb{R}^{d-1}$, the C^∞-function $\tau_t(a_1, \dots, a_{d-1}, x)$ is strictly increasing with respect to x.

Assume $p < \infty$. Take any compact subset $K \subset \mathbb{R}^d$. We show $\|\tau_t - \tau\|_{K,r,p} \to 0$ as $t \to 0$. We prove τ_t converges τ as $t \to 0$. Take $R > 0$ satisfying $K \subset B(R) := \{x \in \mathbb{R}^d : |x| \leq R\}$. We assume $0 < t < 1$. Then we have $\phi_t * \tau = \phi_t * (\mathbf{1}_{B(R+1)}\tau)$. Since we have $\mathbf{1}_{B(R+1)}\tau \in L^p(\mathbb{R}^d)$, we obtain

$$\begin{aligned}\|\phi_t * \tau - \tau\|_{K,r,p} &= \sum_{|\alpha| \leq r} \|\phi_t * (\mathbf{1}_{B(R+1)}\partial_\alpha \tau) - \mathbf{1}_{B(R+1)}\partial_\alpha \tau\|_{K,0,p} \\ &= \sum_{|\alpha| \leq r} \|\phi_t * (\mathbf{1}_{B(R+1)}\partial_\alpha \tau) - \mathbf{1}_{B(R+1)}\partial_\alpha \tau\|_{\mathbb{R}^d,0,p} \to 0 \quad (t \to 0).\end{aligned}$$

Here, we use a property of the mollifier ϕ_t (see Theorem 8.14 in Folland 1999 for example).

In the case of $p = \infty$, by direct computation, we have

$$|\tau_t - \tau|_{K,r,\infty} \leq C \sum_{|\alpha| \leq r} \sup_{(x,y) \in \mathrm{supp}(\phi) \times K} |\partial_\alpha \tau(y - tx) - \partial_\alpha \tau(y)| \to 0 \quad (t \to 0).$$

Here $C := \sup_{x \in \mathbb{R}^d} |\phi(x)|$. Thus in both cases above, By taking sufficiently small t, we obtain the desired C^∞-function $\tilde{\tau} = \tau_t$. ■

E.3 Neural autoregressive flows (NAFs)

In this section, we prove that *neural autoregressive flows* (Huang et al., 2018) yield sup-universal approximators for $\mathcal{S}_{\mathrm{c}}^1$ (hence for $\mathcal{S}_{\mathrm{c}}^\infty$). The proof is not merely an application of

a known result in Huang et al. (2018), but it requires additional non-trivial consideration to enable the adoption of Lemma 3 in Huang et al. (2018) as it is applicable only for those smooth mappings that match certain boundary conditions.

Definition E.6 *A* deep sigmoidal flow *(DSF; a special case of neural autoregressive flows) is a flow layer $g = (g_1, \dots, g_d) \colon \mathbb{R}^d \to \mathbb{R}^d$ of the following form (Huang et al., 2018, Equation (8)):*

$$g_k(\boldsymbol{x}) := \sigma^{-1}\left(\sum_{j=1}^{n} w_{k,j}(\boldsymbol{x}_{\leq k-1}) \cdot \sigma\left(\frac{x_k - b_{k,j}(\boldsymbol{x}_{\leq k-1})}{\tau_j(\boldsymbol{x}_{\leq k-1})}\right)\right),$$

where σ is the sigmoid function, $n \in \mathbb{N}$, $w_j, b_j, \tau_j \colon \mathbb{R}^{k-1} \to \mathbb{R}$ $(j \in [n])$ are neural networks such that $b_j(\cdot) \in (r_0, r_1)$, $\tau_j(\cdot) \in (0, r_2)$, $w_j(\cdot) > 0$, and $\sum_{j=1}^{n} w_j(\cdot) = 1$ $(r_0, r_1 \in \mathbb{R}$, $r_2 > 0)$. We define DSF *to be the set of all possible DSFs.*

Proposition E.7 (Universality of INNs based on DSF) *The elements of* DSF *are locally bounded, and* $\mathrm{INN}_{\mathrm{DSF}}$ *is a* sup*-universal approximator for $\mathcal{S}^1_{\mathrm{c}}$.*

Proof The elements of DSF are continuous, hence locally bounded. Let $s = (s_1, \cdots, s_d) \in \mathcal{S}^1_{\mathrm{c}}$. Take any compact set $K \subset \mathbb{R}^d$ and $\epsilon > 0$. Since K is compact, there exist $r_0, r_1 \in \mathbb{R}$ such that $K \subset [r_0, r_1]^d$. Put $r'_0 = r_0 - 1$, $r'_1 = r_1 + 1$. We take a C^1-function $b \colon (r'_0, r'_1) \to \mathbb{R}$ satisfying

1. $b|_{[r_0, r_1]} = 0$,
2. $b|_{(r'_0, r_0)}$ and $b|_{(r_1, r'_1)}$ are strictly increasing,
3. $\lim_{x \to r'_0 + 0} b(x) = -\infty$ and $\lim_{x \to r'_1 - 0} b(x) = \infty$,
4. $\lim_{x \to r'_0 + 0} \frac{d(\sigma \circ b)}{dx}(x)$ and $\lim_{x \to r'_1 - 0} \frac{d(\sigma \circ b)}{dx}(x)$ exist in $\mathbb{R}$,

where σ is the sigmoid function. For each $k \in [d]$, we define a C^1-map $\tilde{s}_k \colon [r'_0, r'_1]^{k-1} \times (r'_0, r'_1) \times [r'_0, r'_1]^{d-k} \to \mathbb{R}$, which is strictly increasing with respect to x_k, by

$$\tilde{s}_k(x) := s_k(x) + b(x_k) \quad (x = (x_1, \cdots, x_d)).$$

Moreover, we define a map $S \colon [r'_0, r'_1]^d \to [0, 1]^d$ by

$$\begin{aligned} S_k|_{[r'_0, r'_1]^{k-1} \times (r'_0, r'_1) \times [r'_0, r'_1]^{d-k}} &= \sigma \circ \tilde{s}_k, \\ S_k(x_1, \cdots, x_{k-1}, r'_0, x_{k+1}, \cdots, x_d) &= 0, \\ S_k(x_1, \cdots, x_{k-1}, r'_1, x_{k+1}, \cdots, x_d) &= 1, \end{aligned}$$

where we write $S = (S_1, \cdots, S_d)$. Then, by Lemma E.8, S satisfies the assumptions of Lemma 3 in Huang et al. (2018). Since $S([r_0, r_1]^d) \subset (0, 1)^d$ is compact, there exists a positive number $\delta > 0$ such that

$$S([r_0, r_1]^d) + B(\delta) := \{S(x) + v \ : \ x \in [r_0, r_1]^d, v \in B(\delta)\} \subset [\delta, 1 - \delta]^d,$$

where $B(\delta) := \{x \in \mathbb{R}^d : |x| \leq \delta\}$. Let $L > 0$ be a Lipschitz constant of $\sigma^{-1}\colon (0,1)^d \to \mathbb{R}^d$ on $[\delta, 1-\delta]^d$. By Lemma 3 in Huang et al. (2018), there exists $g \in \mathrm{INN}_{\mathrm{DSF}}$ such that

$$\|S - \sigma \circ g\|_{[r_0', r_1']^d, 0, \infty} < \min\left\{\delta, \frac{\epsilon}{L}\right\}.$$

As a result, $\sigma \circ g([r_0, r_1]^d) \subset S([r_0, r_1]^d) + B(\delta) \subset [\delta, 1-\delta]^d$. Then we obtain

$$\begin{aligned}\|s - g\|_{K,0,\infty} \leq \|s - g\|_{[r_0,r_1]^d,0,\infty} &= \|\sigma^{-1} \circ \sigma \circ s - \sigma^{-1} \circ \sigma \circ g\|_{[r_0,r_1]^d,0,\infty} \\ &\leq L\|S - \sigma \circ g\|_{[r_0,r_1]^d,0,\infty} \\ &< \epsilon.\end{aligned}$$

■

Lemma E.8 *We denote by $\mathcal{T}^1$ the set of all C^1-increasing triangular mappings from $\mathbb{R}^d$ to $\mathbb{R}^d$. For $s = (s_1, \cdots, s_d) \in \mathcal{T}^1$, we define a map $S\colon [r_0', r_1']^d \to [0,1]^d$ as in the proof of Proposition E.7. Then S is a C^1-map.*

Proof It is enough to show that $S_d\colon [r_0', r_1']^d \to [0,1]$ is a C^1-function. We prove that for any $i \in [d]$, the i-th partial derivative of S_d exists and that it is continuous on $[r_0', r_1']^d$. First, for $i \in [d-1]$, we consider the i-th partial derivative.
Claim 1.

$$\frac{\partial S_d}{\partial x_i}(x) = \begin{cases} \frac{d\sigma}{dx}(s_i(x) + b(x_d))\frac{\partial s_d}{\partial x_i}(x) & (x \in [r_0', r_1']^{d-1} \times (r_0', r_1')) \\ 0 & (x_d = r_0', r_1') \end{cases}$$

In fact, for $x \in [r_0', r_1']^{d-1} \times (r_0', r_1')$, we have

$$\frac{\partial S_d}{\partial x_i}(x) = \frac{\partial(\sigma \circ \tilde{s_d})}{\partial x_i}(x) = \frac{d\sigma}{dx}(s_d(x) + b(x_d))\left(\frac{\partial s_d}{\partial x_i}(x) + 0\right).$$

For $x = (x_{\leq d-1}, r_0')$, we have

$$\begin{aligned}\frac{\partial S_d}{\partial x_i}(x) &= \lim_{h \to 0} \frac{S_d(x_{\leq i-1}, x_i + h, x_{i+1}, \cdots, x_{d-1}, r_0') - S_d(x_{\leq d-1}, r_0')}{h} \\ &= \lim_{h \to 0} \frac{0 - 0}{h} = 0\end{aligned}$$

Here, note that by the definition of S_d, the notation $S_d(x_{\leq i-1}, x_i + h, x_{i+1}, \cdots, x_{d-1}, r_0')$ makes sense even if $x_i = r_0'$ or $x_i = r_1'$. We can verify the case $x = (x_{\leq d-1}, r_1')$ similarly.

Next, we show that $\frac{\partial S_d}{\partial x_i}$ is continuous. We take any $x_{\leq d-1} \in [r_0', r_1']^{d-1}$. Since we have $\lim_{x \to r_0'} b(x) = -\infty$, $\lim_{x \to r_1'} b(x)$, $\lim_{x \to \pm\infty} \frac{d\sigma}{dx}(x) = 0$, and $|\frac{\partial s_d}{\partial x_I}(x)| < \infty$ $(x \in [r_0', r_1']^d)$, we obtain

$$\begin{aligned}\lim_{x \to (x_{d-1}, r_0')} \frac{d\sigma}{dx}(s_i(x) + b(x_d))\frac{\partial s_d}{\partial x_i}(x) = 0, \\ \lim_{x \to (x_{d-1}, r_1')} \frac{d\sigma}{dx}(s_i(x) + b(x_d))\frac{\partial s_d}{\partial x_i}(x) = 0.\end{aligned}$$

Therefore, the partial derivative $\frac{\partial S_d}{\partial x_i}(x)$ is continuous on $[r'_0, r'_1]^d$ for $i \in [d-1]$.

Next, we consider the d-th derivative of S_d.

Claim 2.

$$\frac{\partial S_d}{\partial x_d}(x) = \begin{cases} \frac{d\sigma}{dx}(s_d(x) + b(x_d)) \left(\frac{\partial s_d}{\partial x_d}(x) + \frac{db}{dx}(x_d) \right) & (x \in [r'_0, r'_1]^{d-1} \times (r'_0, r'_1)) \\ e^{s_d(x_{\leq d-1}, r'_0)} \lim_{x \to r'_0 + 0} \frac{d(\sigma \circ b)}{dx}(x) & (x_d = r'_0) \\ e^{-s_d(x_{\leq d-1}, r'_1)} \lim_{x \to r'_1 - 0} \frac{d(\sigma \circ b)}{dx}(x) & (x_d = r'_1) \end{cases}$$

We verify Claim 2. Since it is clear for the case $x \in [r'_0, r'_1]^{d-1} \times (r'_0, r'_1)$ by the definition of S_k, we consider the case $x_d = r'_0, r'_1$.

Subclaim. For $x'_{\leq d-1} \in [r'_0, r'_1]^{d-1}$,

$$\lim_{x \to (x'_{\leq d-1}, r'_0)} \frac{\sigma(s_d(x) + b(x_d))}{\sigma(b(x_d))} = e^{s_d(x'_{\leq d-1}, r'_0)}$$

$$\lim_{x \to (x'_{\leq d-1}, r'_1)} \frac{\sigma(s_d(x) + b(x_d)) - 1}{\sigma(b(x_d)) - 1} = e^{-s_d(x'_{\leq d-1}, r'_1)}$$

We verify this subclaim. From $\lim_{x \to r'_0} b(x) = -\infty$, we have

$$\begin{aligned} \frac{\sigma(s_d(x) + b(x_d))}{\sigma(b(x_d))} &= \frac{1 + e^{-b(x_d)}}{1 + e^{-s_d(x) - b(x_d)}} = \frac{e^{b(x_d)} + 1}{e^{b(x_d)} + e^{-s_d(x)}} \\ &\to \frac{1}{e^{-s_d(x'_{\leq d-1}, r'_0)}} = e^{s_d(x'_{\leq d-1}, r'_0)} \quad (x \to (x'_{\leq d-1}, r'_0)) \end{aligned}$$

Similarly, from $\lim_{x \to r'_1} b(x) = \infty$, we have

$$\begin{aligned} \frac{\sigma(s_d(x) + b(x_d)) - 1}{\sigma(b(x_d)) - 1} &= e^{-s_d(x)} \frac{1 + e^{-b(x_d)}}{1 + e^{-s_d(x) - b(x_d)}} \\ &\to e^{-s_d(x_{\leq d-1}, r'_1)} \quad (x \to (x'_{\leq d-1}, r'_1)). \end{aligned}$$

Therefore, our subclaim has been proved. By using L'Hôpital's rule, we have

$$\lim_{h \to +0} \frac{\sigma(b(r'_0 + h))}{h} = \lim_{x \to r'_0} \frac{d(\sigma \circ b)}{dx}(x), \quad \lim_{x \to r'_1} \frac{\sigma(b(r'_1 + h)) - 1}{h} = \lim_{x \to r'_1} \frac{d(\sigma \circ b)}{dx}(x).$$

Then, from Subclaim, we obtain

$$\begin{aligned} \frac{\partial S_d}{\partial x_d}(x_{\leq d-1}, r'_0) &= \lim_{h \to +0} \frac{\sigma(s_d(x_{\leq d-1}, r'_0 + h) + b(r'_0 + h)) - 0}{h} \\ &= \lim_{h \to +0} \frac{\sigma(s_d(x_{\leq d-1}, r'_0 + h) + b(r'_0 + h))}{\sigma(b(r_0 + h))} \cdot \frac{\sigma(b(r'_0 + h))}{h} \\ &= e^{s_d(x_{\leq d-1}, r'_0)} \lim_{x \to r'_0 + 0} \frac{d(\sigma \circ b)}{dx}(x), \\ \frac{\partial S_d}{\partial x_d}(x_{\leq d-1}, r'_1) &= \lim_{h \to -0} \frac{\sigma(s_d(x_{\leq d-1}, r'_1 + h) + b(r'_1 + h)) - 1}{h} \\ &= \lim_{h \to -0} \frac{\sigma(s_d(x_{\leq d-1}, r'_1 + h) + b(r'_1 + h)) - 1}{\sigma(b(r'_1 + h)) - 1} \cdot \frac{\sigma(b(r'_1 + h)) - 1}{h} \\ &= e^{s_d(x_{\leq d-1}, r'_1)} \lim_{x \to r'_1} \frac{d(\sigma \circ b)}{dx}(x). \end{aligned}$$

Therefore, Claim 2 was proved.

Finally, we verify $\frac{\partial S_d}{\partial x_d}(x)$ is continuous on $[r'_0, r'_1]^d$. Fix $x'_{\leq d-1} \in [r'_0, r'_1]^{d-1}$. Since we have $\lim_{x \to (x'_{\leq d-1}, r'_0)} \frac{d\sigma}{dx}(\sigma_d(x) + b(x_d)) \frac{\partial s_d}{\partial x_d}(x) = 0$, from Claim 2, it is enough to show the following:

Claim 3.

$$\lim_{x \to (x'_{\leq d-1}, r'_0)} \frac{d\sigma}{dx}(s_d(x) + b(x_d)) \frac{db}{dx}(x_d) = e^{s_d(x_{\leq d-1}, r'_0)} \lim_{x \to r'_0 + 0} \frac{d(\sigma \circ b)}{dx}(x),$$
$$\lim_{x \to (x'_{\leq d-1}, r'_1)} \frac{d\sigma}{dx}(s_d(x) + b(x_d)) \frac{db}{dx}(x_d) = e^{-s_d(x_{\leq d-1}, r'_1)} \lim_{x \to r'_1 - 0} \frac{d(\sigma \circ b)}{dx}(x).$$

We verify Claim 3. We have

$$\begin{aligned} \frac{d\sigma}{dx}(s_d(x) + b(x_d)) \frac{db}{dx}(x_d) &= \frac{\frac{d\sigma}{dx}(s_d(x) + b(x_d))}{\frac{d\sigma}{dx}(b(x_d))} \frac{d\sigma}{dx}(b(x_d)) \frac{db}{dx}(x_d) \\ &= \frac{\frac{d\sigma}{dx}(s_d(x) + b(x_d))}{\frac{d\sigma}{dx}(b(x_d))} \frac{d(\sigma \circ b)}{dx}(x_d). \end{aligned}$$

Since we have $\frac{d\sigma}{dx}(x) = \sigma(x)(1 - \sigma(x))$, from Subclaim above, Claim 3 follows from

$$\begin{aligned} \frac{\frac{d\sigma}{dx}(s_d(x) + b(x_d))}{\frac{d\sigma}{dx}(b(x_d))} &= \frac{\sigma(s_d(x) + b(x_d))}{\sigma(b(x_d))} \cdot \frac{1 - \sigma(s_d(x) + b(x_d))}{1 - \sigma(b(x_d))} \\ &\to \begin{cases} e^{s_d(x'_{\leq d-1}, r'_0)} & (x \to (x'_{\leq d-1}, r'_0)) \\ e^{-s_d(x'_{\leq d-1}, r'_1)} & (x \to (x'_{\leq d-1}, r'_1)) \end{cases}. \end{aligned}$$

Therefore, we proved the continuity of $\frac{\partial S_d}{\partial x_d}(x)$. ■

E.4 Sum-of-squares polynomial flows (SoS flows)

In this section, we prove that *sum-of-squares polynomial flows* (Jaini et al., 2019) yield CF-INNs with the sup-universal approximation property for $\mathcal{S}_c^1$ (hence for $\mathcal{S}_c^\infty$). Even though Jaini et al. (2019) claimed the distributional universality of the SoS flows by providing a proof sketch based on the univariate Stone-Weierstrass approximation theorem, we regard the sketch to be invalid or at least incomplete as it does not discuss the smoothness of the coefficients, in other words, whether the polynomial coefficients can be realized by continuous functions. Here, we provide complete proof that takes an alternative route to prove the sup-universality of the SoS flows via the multivariate Stone-Weierstrass approximation theorem.

A *sum-of-squares polynomial flow* (SoS flow) is a flow layer $g = (g_1, \ldots, g_d) \colon \mathbb{R}^d \to \mathbb{R}^d$ of the following form(Jaini et al., 2019, Equation (9)):

$$g_k(\boldsymbol{x}) := \mathfrak{B}_{2r+1}(x_k; C_k(\boldsymbol{x}_{\leq k-1})),$$
$$\mathfrak{B}_{2r+1}(z; (c, \boldsymbol{a})) := c + \int_0^z \sum_{b=1}^{B} \left(\sum_{l=0}^{r} a_{l,b} u^l \right)^2 du,$$

where $r \in \mathbb{N} \cup \{0\}$, $B \in \mathbb{N}$, $c \in \mathbb{R}$, $\boldsymbol{a} \in \mathbb{R}^{B(r+1)}$, and $C_k \colon \mathbb{R}^{k-1} \to \mathbb{R}^{B(r+1)+1}$ is a certain map, for example, a neural network.

Here, we consider a small class of SoS flows as follows:

Definition E.9 *Let $\mathcal{H}$ be a function on $\mathbb{R}^{d-1}$. For $c \in \mathbb{R}$ and $h_1, \ldots, h_r \in \mathcal{H}$, Let*

$$\tilde{\mathfrak{B}}(\boldsymbol{x}; c, h_1, \ldots, h_r) := c + \int_0^{x_d} \left(\sum_{l=0}^{r} h_l(\boldsymbol{x}_{\leq d-1}) u^l \right)^2 du.$$

Then, we define the set $\mathcal{H}$-SoS as a subset consisting of $\tilde{\mathfrak{B}}(\cdot; h_1, \ldots, h_r)$ where $r \geq 1$ and h_i's are elements of $\mathcal{H}$.

Then, we have the following proposition:

Proposition E.10 *Let $r \geq 0$. Let $\mathcal{H} \subset C^r(\mathbb{R}^{d-1})$ and assume that $\mathcal{H}$ is a $W^{r,\infty}$-universal approximator for the set of $(d-1)$-variable polynomials. Then, $\mathrm{INN}_{\mathcal{H}\text{-SoS}}$ is a $W^{r,\infty}$-universal approximator for $\mathcal{S}_c^{r+1}$.*

Proof We only illustrate the proof in the cases of $r = 0$ and $r = 1$. The general cases follow from a similar argument with the Leibniz rule and chain rule.

The L^∞-universality follows from the Stone-Weierstrass approximation theorem as in the below. Let $s = (s_1, \ldots, s_d) \in \mathcal{S}_c^1$, a compact subset $K \subset \mathbb{R}^d$, and $\epsilon > 0$ be given. Then, there exists $R > 0$ such that $K \subset [-R, R]^d$. Since $s_d(\boldsymbol{x})$ is strictly increasing with respect to x_d and s is C^1, we have $\eta(\boldsymbol{x}) := \frac{\partial s_d}{\partial x_d}(\boldsymbol{x}) > 0$ and η is continuous. Therefore, we can apply the Stone-Weierstrass approximation theorem (Folland, 1999, Corollary 4.50) to $\sqrt{\eta(\boldsymbol{x})}$: for any $\delta > 0$, there exists a polynomial $\pi(x_1, \ldots, x_d)$ such that $\left\| \sqrt{\eta} - \pi \right\|_{[-R,R]^d, 0, \infty} < \delta$. Then, by rearranging the terms, there exist $r \in \mathbb{N}$ and polynomials $\xi_l(x_1, \ldots, x_{d-1})$ such that $\pi(x_1, \ldots, x_d) = \sum_{l=0}^{r} \xi_l(x_1, \ldots, x_{d-1}) x_d^l$. Now, define

$$\begin{aligned}
\tilde{g}_d(\boldsymbol{x}) &:= s_d(\boldsymbol{x}_{\leq d-1}, 0) + \int_0^{x_d} (\pi(\boldsymbol{x}_{\leq d-1}, u))^2 du \\
&= s_d(\boldsymbol{x}_{\leq d-1}, 0) + \int_0^{x_d} \left(\sum_{l=0}^{r} \xi_l(x_1, \ldots, x_{d-1}) u^l \right)^2 du
\end{aligned}$$

and $\tilde{g}(\boldsymbol{x}) := (x_1, \ldots, x_{d-1}, \tilde{g}_d(\boldsymbol{x}))$. Then,

$$\begin{aligned}
\|s - \tilde{g}\|_{K,0,\infty} &= \sup_{\boldsymbol{x} \in K} |s_d(\boldsymbol{x}) - \tilde{g}_d(\boldsymbol{x})| \\
&= \sup_{\boldsymbol{x} \in K} \left| s_d(\boldsymbol{x}_{\leq d-1}, 0) + \int_0^{x_d} \eta(\boldsymbol{x}_{\leq d-1}, u) du - \tilde{g}_d(\boldsymbol{x}) \right| \\
&= \sup_{\boldsymbol{x} \in K} \left| \int_0^{x_d} (\sqrt{\eta(\boldsymbol{x}_{\leq d-1}, u)}^2 - \pi(\boldsymbol{x}_{\leq d-1}, u)^2) du \right| \\
&\leq R \cdot \sup_{\boldsymbol{x} \in [-R,R]^d} \left| \sqrt{\eta(\boldsymbol{x})}^2 - \pi(\boldsymbol{x})^2 \right| \\
&= R \cdot \sup_{\boldsymbol{x} \in [-R,R]^d} |\sqrt{\eta(\boldsymbol{x})} + \pi(\boldsymbol{x})| \cdot |\sqrt{\eta(\boldsymbol{x})} - \pi(\boldsymbol{x})| \\
&\leq R \left(\sup_{\boldsymbol{x} \in [-R,R]^d} 2\sqrt{\eta(\boldsymbol{x})} + \delta \right) \delta,
\end{aligned}$$

where we used

$$\begin{aligned}\sup_{\boldsymbol{x}\in[-R,R]^d} |\sqrt{\eta(\boldsymbol{x})}+\pi(\boldsymbol{x})| &\le \sup_{\boldsymbol{x}\in[-R,R]^d} |2\sqrt{\eta(\boldsymbol{x})}| + |\sqrt{\eta(\boldsymbol{x})}-\pi(\boldsymbol{x})| \\ &\le \sup_{\boldsymbol{x}\in[-R,R]^d} 2\sqrt{\eta(\boldsymbol{x})}+\delta.\end{aligned}$$

It is straightforward to show that there exists $g \in \mathrm{SoS}$ such that $\|\tilde{g}-g\|_{K,0,\infty} < \frac{\epsilon}{2}$ by approximating each of $s_d(\boldsymbol{x}_{\le d-1})$ and ξ_l on K using neural networks. Finally, take δ to be small enough so that $\|s-\tilde{g}\|_{K,0,\infty} < \frac{\epsilon}{2}$ holds.

Next, we consider the $W^{1,\infty}$-universality. We use the same notations as above. We note that since $s \in \mathcal{S}_c^2$, we have $\eta \in C^1$, and η is positive and continuous. This enables us to apply the Stone-Weierstrass approximation theorem (Peet, 2009, Theorem 5) to $\sqrt{\eta(\boldsymbol{x})}$: for any $\delta > 0$, there exists a polynomial $\pi(x_1,\dots,x_d)$ such that $\|\sqrt{\eta}-\pi\|_{[-R,R]^d,1,\infty} < \delta$. We define $\tilde{g}_d$ and $\tilde{g}$ as above. Then we have

$$\begin{aligned}\|s-\tilde{g}\|_{K,1,\infty} &= \left\| \int_0^{x_d} (\sqrt{\eta(\boldsymbol{x}_{\le d-1},u)}^2 - \pi(\boldsymbol{x}_{\le d-1},u)^2) du \right\|_{K,1,\infty} \\ &\le \sup_{\boldsymbol{x}\in K} \left| \int_0^{x_d} (\sqrt{\eta(\boldsymbol{x}_{\le d-1},u)}^2 - \pi(\boldsymbol{x}_{\le d-1},u)^2) du \right| \\ &\quad + \sup_{\boldsymbol{x}\in K} \sum_{i=1}^{d-1} \left| \partial_{x_i} \int_0^{x_d} (\sqrt{\eta(\boldsymbol{x}_{\le d-1},u)}^2 - \pi(\boldsymbol{x}_{\le d-1},u)^2) du \right| \\ &\quad + \sup_{\boldsymbol{x}\in K} \left| \partial_{x_d} \int_0^{x_d} (\sqrt{\eta(\boldsymbol{x}_{\le d-1},u)}^2 - \pi(\boldsymbol{x}_{\le d-1},u)^2) du \right| \\ &=: I + II + III.\end{aligned}$$

In a similar manner as above, we have $I \le R\left(\sup_{\boldsymbol{x}\in[-R,R]^d} 2\sqrt{\eta(\boldsymbol{x})}+\delta\right)\delta$. We note that since $\eta \in C^1$ and η is positive and continuous, we have $\|\sqrt{\eta}\|_{[-R,R]^d,1,\infty} < \infty$. A direct computation gives

$$\begin{aligned}II &= 2\sup_{\boldsymbol{x}\in K} \sum_{i=1}^{d-1} \left| \int_0^{x_d} \left\{ \sqrt{\eta(\boldsymbol{x}_{\le d-1},u)} \partial_{x_i} \sqrt{\eta(\boldsymbol{x}_{\le d-1},u)} - \pi(\boldsymbol{x}_{\le d-1},u) \partial_{x_i} \pi(\boldsymbol{x}_{\le d-1},u) \right\} du \right| \\ &\le 2\sup_{\boldsymbol{x}\in K} \sum_{i=1}^{d-1} \left| \int_0^{x_d} \left\{ \sqrt{\eta(\boldsymbol{x}_{\le d-1},u)} - \pi(\boldsymbol{x}_{\le d-1},u) \right\} \partial_{x_i} \sqrt{\eta(\boldsymbol{x}_{\le d-1},u)} du \right| \\ &\quad + 2\sup_{\boldsymbol{x}\in K} \sum_{i=1}^{d-1} \left| \int_0^{x_d} \pi(\boldsymbol{x}_{\le d-1},u) \partial_{x_i} \left\{ \sqrt{\eta(\boldsymbol{x}_{\le d-1},u)} - \pi(\boldsymbol{x}_{\le d-1},u) \right\} du \right| \\ &\le 2(d-1)R(2\|\sqrt{\eta}\|_{[-R,R]^d,1,\infty}+\delta)\|\sqrt{\eta}-\pi\|_{[-R,R]^d,1,\infty} \\ &\le 2(d-1)R(2\|\sqrt{\eta}\|_{[-R,R]^d,1,\infty}+\delta)\delta.\end{aligned}$$

A simple computation gives

$$III = \sup_{\boldsymbol{x}\in K} \left|\sqrt{\eta(\boldsymbol{x})}+\pi(\boldsymbol{x})\right| \left|\sqrt{\eta(\boldsymbol{x})}-\pi(\boldsymbol{x})\right| \le \left(\sup_{\boldsymbol{x}\in[-R,R]^d} 2\sqrt{\eta(\boldsymbol{x})}+\delta \right)\delta.$$

In a similar manner as above, we can see that there exists $g \in \mathrm{SoS}$ such that $\|\tilde{g} - g\|_{K,1,\infty} < \frac{\epsilon}{2}$. Finally, taking δ to be small enough so that $\|s - \tilde{g}\|_{K,1,\infty} < \frac{\epsilon}{2}$ holds, the assertion is proved. ■

Appendix F. Universality of NODE-based INNs

Here, we provide a proof of Theorem 44:

Proof [Proof of Theorem 44] By Theorem 24, we only consider an approximation of the elements of Ξ^∞. Let $g \in \Xi^\infty$. Then, by Definition 20, there exists $f \in \mathrm{Lip} \cap \infty$ such that

$$f(\cdot) := \left.\frac{\partial \Phi(\cdot, t)}{\partial t}\right|_{t=0}.$$

for some flow Φ Therefore, g is arbitrarily approximated by an element of $\mathrm{INN}_{\Psi(\mathcal{H})}$ by Lemma F.1. ■

The following lemma, used in the above proof, allows us to approximate an autonomous ODE flow endpoint by approximating the differential equation. See Definition 5 for the definition of $\Psi(\cdot)$.

Lemma F.1 (Approximation of Autonomous-ODE flow endpoints) *Let $r \geq 0$. Assume $\mathcal{H} \subset \mathrm{Lip} \cap C^r$ is a $W^{r,\infty}$-universal approximator for $\mathrm{Lip} \cap C^r$. Then, $\Psi(\mathcal{H})$ is a $W^{r,\infty}$-universal approximator for $\Psi(\mathrm{Lip} \cap C^r)$.*

Proof We first treat the case of $r > 0$. By combining the fact that the map

$$(\boldsymbol{x}, f) \mapsto \mathrm{IVP}[f](\boldsymbol{x}, 1)$$

is C^r map (Theorem B.3 (ii) in Duistermaat and Kolk 2000) with the Berge maximum theorem (Aliprantis and Border, 2006), we see that for any compact set $K \subset \mathbb{R}^d$ and $F \in \mathrm{Lip} \cap C^r$ we see that the map

$$f \mapsto \|\mathrm{IVP}[f](\cdot, 1) - \mathrm{IVP}[F](\cdot, 1)\|_{K,r,\infty} = \sum_{|\alpha| \leq r} \sup_{\boldsymbol{x} \in K} \|\mathrm{IVP}[f](\boldsymbol{x}, 1) - \mathrm{IVP}[F](\boldsymbol{x}, 1)\|$$

is continuous. Therefore, the $W^{r,\infty}$-universality of $\Psi(\mathcal{H})$ for $\Psi(\mathrm{Lip} \cap C^r)$ follows from that of $\mathcal{H}$ for $\mathrm{Lip} \cap C^r$.

We next treat the case of $r = 0$. Let $\phi \in \Psi(\mathrm{Lip})$. Then, by definition, there exists $F \in \mathrm{Lip}$ such that $\phi = \mathrm{IVP}[F](\cdot, 1)$. Let L_F denote the Lipschitz constant of F. In the following, we approximate $\mathrm{IVP}[F](\cdot, 1)$ by approximating F using an element of $\mathcal{H}$.

Let $\varepsilon > 0$, and let $K \subset \mathbb{R}^d$ be a compact subset of $\mathbb{R}^d$. We show that there exists $f \in \mathcal{H}$ such that $\|\mathrm{IVP}[F](\cdot, 1) - \mathrm{IVP}[f](\cdot, 1)\|_{K,0,\infty} < \varepsilon$. Note that $\mathrm{IVP}[f](\cdot, \cdot)$ is well-defined because $\mathcal{H} \subset \mathrm{Lip}$. Define

$$K' := \left\{ \boldsymbol{x} \in \mathbb{R}^d \;\middle|\; \inf_{\boldsymbol{y} \in \mathrm{IVP}[F](K,[0,1])} \|\boldsymbol{x} - \boldsymbol{y}\| \leq 2e^{L_F} \right\}.$$

Then, K' is compact. This follows from the compactness of $\mathrm{IVP}[F](K,[0,1])$: (i) K' is bounded since $\mathrm{IVP}[F](K,[0,1])$ is bounded, and (ii) it is closed since the function $\boldsymbol{x} \mapsto \min_{\boldsymbol{y}\in\mathrm{IVP}[F](K,[0,1])} \|\boldsymbol{x}-\boldsymbol{y}\|$ is continuous and hence K' is the inverse image of a closed interval $[0, 2e^{L_F}]$ by a continuous map.

Since $\mathcal{H}$ is assumed to be an L^∞-universal approximator for Lip, for any $\delta > 0$, we can take $f \in \mathcal{H}$ such that $\|f - F\|_{K',0,\infty} < \delta$. Let δ be such that $0 < \delta < \min\{\varepsilon/(2e^{L_F}), 1\}$, and take such an f.

Fix $\boldsymbol{x}_0 \in K$ and define $\Delta_{\boldsymbol{x}_0}(t) := \|\mathrm{IVP}[F](\boldsymbol{x}_0, t) - \mathrm{IVP}[f](\boldsymbol{x}_0, t)\|$. Let $B := \delta e^{L_F}$ and we show that

$$\Delta_{\boldsymbol{x}_0}(t) < 2B$$

holds for all $t \in [0,1]$. We prove this by contradiction. Suppose that there exists t' for which the inequality does not hold. Then, the set $\mathcal{T} := \{t \in [0,1] | \Delta_{\boldsymbol{x}_0}(t) \geq 2B\}$ is not empty and thus $\tau := \inf \mathcal{T} \in [0,1]$. For this τ, we show both $\Delta_{\boldsymbol{x}_0}(\tau) \leq B$ and $\Delta_{\boldsymbol{x}_0}(\tau) \geq 2B$. First, we have

$$\begin{aligned}
\Delta_{\boldsymbol{x}_0}(\tau) &= \|\mathrm{IVP}[F](\boldsymbol{x}_0, \tau) - \mathrm{IVP}[f](\boldsymbol{x}_0, \tau)\| \\
&= \left\| \boldsymbol{x}_0 + \int_0^\tau F(\mathrm{IVP}[F](\boldsymbol{x}_0, t))dt - \boldsymbol{x}_0 - \int_0^\tau f(\mathrm{IVP}[f](\boldsymbol{x}_0, t))dt \right\| \\
&\leq \left\| \int_0^\tau (F(\mathrm{IVP}[F](\boldsymbol{x}_0, t)) - F(\mathrm{IVP}[f](\boldsymbol{x}_0, t)))dt \right\| \\
&\quad + \left\| \int_0^\tau (F(\mathrm{IVP}[f](\boldsymbol{x}_0, t)) - f(\mathrm{IVP}[f](\boldsymbol{x}_0, t)))dt \right\|.
\end{aligned}$$

The last term can be bounded as

$$\left\| \int_0^\tau (F(\mathrm{IVP}[f](\boldsymbol{x}_0, t)) - f(\mathrm{IVP}[f](\boldsymbol{x}_0, t)))dt \right\| \leq \int_0^\tau \delta dt$$

because of the following argument. If $\tau = 0$, then both sides are equal to zero, hence it holds with equality. If $\tau > 0$, then for any $t < \tau$, we have $\mathrm{IVP}[f](\boldsymbol{x}_0, t) \in K'$ because $t < \tau$ implies $\Delta_{\boldsymbol{x}_0}(t) \leq 2B$. In this case, $\|F - f\|_{K',0,\infty} < \delta$ implies the inequality. Therefore, we have

$$\Delta_{\boldsymbol{x}_0}(\tau) \leq L_F \int_0^\tau \Delta_{\boldsymbol{x}_0}(t)dt + \int_0^\tau \delta dt.$$

Now, by applying Grönwall's inequality (Gronwall, 1919), we obtain

$$\Delta_{\boldsymbol{x}_0}(\tau) \leq \delta\tau e^{L_F\tau} \leq B.$$

On the other hand, by the definition of $\mathcal{T}$ and the continuity of $\Delta_{\boldsymbol{x}_0}(\cdot)$, we have $\Delta_{\boldsymbol{x}_0}(\tau) \geq 2B$. These two inequalities contradict.

Therefore, $\|\mathrm{IVP}[F](\cdot, 1) - \mathrm{IVP}[f](\cdot, 1)\|_{K,0,\infty} = \sup_{\boldsymbol{x}_0 \in K} \Delta_{\boldsymbol{x}_0}(1) \leq 2B = 2\delta e^{L_F}$ holds. Since $\delta < \varepsilon/(2e^{L_F})$, the right-hand side is smaller than ε. ■

When we construct a NODE to approximate a diffeomorphism, we may insert any invertible affine map between flow layers by definition (see Definition 6). However, we actually need an affine layer only in the last layer to obtain a universality of NODE, namely we have the following proposition:

Proposition F.2 *The notation is as in Theorem 44. Then, the subset*

$$\{W \circ g_1 \circ \cdots \circ g_k : k \geq 0, W \in \mathrm{Aff}, g_1, \ldots, g_k \in \Psi(\mathcal{H})\}$$

of $\mathrm{INN}_{\Psi(\mathcal{H})}$ *has a* $W^{r,\infty}$*-universal approximation property for* $\mathcal{D}^{\max\{r,1\}}$*, where* $\mathcal{H}$ *is a subset of* $\mathrm{Lip} \cap C^r$ *as in Theorem 44.*

Proof Let $F \in \mathcal{D}^{\max\{r,1\}}$. Take any compact set $K \subset U$ and $\varepsilon > 0$. First, thanks to Lemma 28 and 29, there exists a $G \in \mathrm{Diff}_{\mathrm{c}}^{\infty}$ and an affine transform $W \in \mathrm{Aff}$ such that

$$W \circ G|_K = F|_K.$$

Then, we use Lemma 31 to show that there exists a finite set of flow endpoints (Definition 20) $g_1, \ldots, g_k \in \Xi^{\infty}$ such that

$$G = g_k \circ \cdots \circ g_1.$$

We now construct $f_j \in \mathrm{Lip}$ such that $g_j = \mathrm{IVP}[f_j](\cdot, 1)$. By Definition 20, for each g_j $(1 \leq j \leq k)$, there exists an associated flow Φ_j. Now, define

$$f_j(\cdot) := \left.\frac{\partial \Phi_j(\cdot, t)}{\partial t}\right|_{t=0}.$$

Then, $f_j \in \mathrm{Lip}$ because it is a compactly-supported C^{∞}-map: it is compactly supported since there exists a compact subset $K_j \subset \mathbb{R}^d$ containing the support of $\Phi(\cdot, t)$ for all t, and hence $\Phi(\cdot, t) - \Phi(\cdot, 0)$ is zero in the complement of K_j.

Now, $\Phi_j(\boldsymbol{x}, t) = \mathrm{IVP}[f_j](\boldsymbol{x}, t)$ since, by additivity of the flows,

$$\begin{aligned}\frac{\partial \Phi_j}{\partial t}(\boldsymbol{x}, t) &= \lim_{s \to 0} \frac{\Phi_j(\boldsymbol{x}, t+s) - \Phi_j(\boldsymbol{x}, t)}{s} = \lim_{s \to 0} \frac{\Phi_j(\Phi_j(\boldsymbol{x}, t), s) - \Phi_j(\Phi_j(\boldsymbol{x}, t), 0)}{s} \\ &= \left.\frac{\partial \Phi_j(\Phi_j(\boldsymbol{x}, t), s)}{\partial s}\right|_{s=0} = f_j(\Phi_j(\boldsymbol{x}, t)),\end{aligned}$$

and hence it is a solution to the initial value problem that is unique. As a result, we have $g_j = \Phi_j(\cdot, 1) = \mathrm{IVP}[f_j](\cdot, 1)$.

By combining Lemma B.1 and Lemma F.1, there exist $\phi_1, \ldots, \phi_k \in \Psi(\mathcal{H})$ such that

$$\|g_k \circ \cdots \circ g_1 - \phi_k \circ \cdots \circ \phi_1\|_{K,r,\infty} < \frac{\varepsilon}{\|W\|_{\mathrm{op}}},$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm. Therefore, we have that $W \circ \phi_k \circ \cdots \circ \phi_1 \in \mathrm{INN}_{\Psi(\mathcal{H})}$ satisfies

$$\begin{aligned}\|F - W \circ \phi_k \circ \cdots \circ \phi_1\|_{K,r,\infty} &= \|W \circ G - W \circ \phi_k \circ \cdots \circ \phi_1\|_{K,r,\infty} \\ &\leq \|W\|_{\mathrm{op}} \|g_k \circ \cdots \circ g_1 - \phi_k \circ \cdots \circ \phi_1\|_{K,r,\infty} \\ &< \varepsilon\end{aligned}$$

■

References

K. Abe, T. Maehara, and I. Sato. Abelian neural networks. *arXiv:2102.12232*, 2021.

M. S. Albergo, G. Kanwar, and P. E. Shanahan. Flow-based generative models for Markov chain Monte Carlo in lattice field theory. *Physical Review D: Particles and Fields*, 100(3): 034515, 2019.

C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 3rd edition, 2006.

C. Anil, J. Lucas, and R. Grosse. Sorting out Lipschitz function approximation. In *Proceedings of the 36th International Conference on Machine Learning*, pages 291–301, 2019.

L. Ardizzone, J. Kruse, C. Rother, and U. Köthe. Analyzing inverse problems with invertible neural networks. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

A. Banyaga. *The Structure of Classical Diffeomorphism Groups*. Springer US, Boston, MA, 1997.

M. Bauer and A. Mnih. Resampled priors for variational autoencoders. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 66–75, 2019.

J. Behrmann, W. Grathwohl, R. T. Q. Chen, D. Duvenaud, and J.-H. Jacobsen. Invertible residual networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 573–582, 2019.

P. Bernard, L. Praly, and V. Andrieu. On diffeomorphism extension. Research Report, MINES ParisTech, 2015.

P. Bernard, V. Andrieu, and L. Praly. Expressing an observer in preferred coordinates by transforming an injective immersion into a surjective diffeomorphism. *SIAM Journal on Control and Optimization*, 56(3):2327–2352, 2018.

P. Bevanda, M. Beier, S. Kerz, A. Lederer, S. Sosnowski, and S. Hirche. Diffeomorphically learning stable Koopman operators. *in IEEE Control Systems Letters*, 6:3427–3432, 2022a.

P. Bevanda, J. Kirmayr, S. Sosnowski, and S. Hirche. Learning the Koopman eigendecomposition: A diffeomorphic approach. *American Control Conference*, pages 2736–2741, 2022b. doi: 10.23919/ACC53348.2022.9867829.

V. I. Bogachev, A. V. Kolesnikov, and K. V. Medvedev. Triangular transformations of measures. *Sbornik: Mathematics*, 196(3):309–335, 2005.

Y. Brenier and W. Gangbo. l^papproximation of maps by diffeomorphisms. *Calculus of Variations and Partial Differential Equations*, 16(2):147–164, Feb 2003. ISSN 1432-0835. doi: 10.1007/s005260100144. URL `https://doi.org/10.1007/s005260100144`.

N. D. Cao, W. Aziz, and I. Titov. Block neural autoregressive flow. In *Proceedings of The 35th Conference on Uncertainty in Artificial Intelligence*, 2019.

R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems 31*, pages 6571–6583, 2018.

E. H. Connell. Approximating stable homeomorphisms by piecewise linear ones. *The Annals of Mathematics*, 78(2):326, 1963.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2:303–314, 1989.

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

L. Dinh, D. Krueger, and Y. Bengio. NICE: Non-linear independent components estimation. In *Workshop Track Proceedings of the 3rd International Conference on Learning Representations*, 2015. arXiv:1410.8516.

L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real NVP. In *Conference Track Proceedings of the 5th International Conference on Learning Representations*, 2017.

R. M. Dudley. *Real Analysis and Probability*. Cambridge studies in advanced mathematics. Cambridge University Press, Cambridge, 2nd edition, 2002.

J. J. Duistermaat and J. A. C. Kolk. *Lie Groups*. Universitext. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000.

D. B. A. Epstein. The simplicity of certain groups of homeomorphisms. *Compositio Mathematica*, 22(2):165–173, 1970.

A. Ern and J.-L. Guermond. *Finite Elements I: Approximation and Interpolation*. Number 72 in Texts in Applied Mathematics. Springer International Publishing, Cham, 2021.

G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Number 125 in Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts Book. Wiley, New York, 2nd edition, 1999.

K.-I. Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2(3):183–192, 1989.

A. Gopal. ELF: Exact-Lipschitz based universal density approximator flow. *arXiv:2112.06997*, 2021.

T. H. Gronwall. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, 20(4):292–296, 1919.

S. Haller. *Groups of Diffeomorphisms*. Magister, University of Vienna, Vienna, Austria, 1995.

P. Hartman. *Ordinary Differential Equations*, volume 38 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, 2002.

M. R. Herman. Sur le groupe des difféomorphismes du tore. *Annales de l'Institut Fourier*, 23 (2):75–86, 1973.

M. W. Hirsch. *Differential Topology*, volume 33 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1st edition, 1976.

J. Ho, X. Chen, A. Srinivas, Y. Duan, and P. Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2722–2730, 2019.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

C.-W. Huang, D. Krueger, A. Lacoste, and A. Courville. Neural autoregressive flows. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2078–2087, 2018.

C.-W. Huang, R. T. Q. Chen, C. Tsirigotis, and A. Courville. Convex potential flows: Universal probability distributions with optimal transport and convex optimization. In *International Conference on Learning Representations*, 2021.

A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

P. Izmailov, P. Kirichenko, M. Finzi, and A. G. Wilson. Semi-supervised learning with normalizing flows. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

J.-H. Jacobsen, A. W. M. Smeulders, and E. Oyallon. I-RevNet: Deep invertible networks. In *6th International Conference on Learning Representations, Conference Track Proceedings*, 2018.

P. Jaini, K. A. Selby, and Y. Yu. Sum-of-squares polynomial flow. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3009–3018, 2019.

H. Kim, H. Lee, W. H. Kang, J. Y. Lee, and N. S. Kim. SoftFlow: Probabilistic framework for normalizing flow on manifolds. In *Advances in Neural Information Processing Systems 33*, pages 16388–16397, 2020.

S. Kim, S.-G. Lee, J. Song, J. Kim, and S. Yoon. FloWaveNet: A generative flow for raw audio. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3370–3378, 2019.

T. Kimura, T. Matsubara, and K. Uehara. ChartPointFlow for topology-aware 3D point cloud generation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1396–1404, 2021.

D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems 31*, pages 10215–10224, 2018.

D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems 29*, pages 4743–4751, 2016.

R. C. Kirby. Stable homeomorphisms and the annulus conjecture. *Annals of Mathematics*, 89(3):575–582, 1969. ISSN 0003486X. URL `http://www.jstor.org/stable/1970652`.

I. Kobyzev, S. J. Prince, and M. A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43 (11):3964–3979, 2021. doi: 10.1109/TPAMI.2020.2992934.

Z. Kong and K. Chaudhuri. Universal approximation of residual flows in maximum mean discrepancy. *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021.

S. Lang. *Differential Manifolds*. Springer-Verlag, New York, NY, USA, 1985.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Q. Li, T. Lin, and Z. Shen. Deep learning via dynamical systems: An approximation perspective. *Journal of the European Mathematical Society*, 2022. doi: 10.4171/JEMS/1221.

C. Livingston. Connected sums of codimension two locally flat submanifolds, 2021.

C. Louizos and M. Welling. Multiplicative normalizing flows for variational Bayesian neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2218–2227, 2017.

J. Lyu, Z. Chen, C. Feng, W. Cun, S. Zhu, Y. Geng, Z. Xu, and Y. Chen. Universality of parametric coupling flows over parametric diffeomorphisms. *arXiv:2202.02906*, 2022.

J. N. Mather. Commutators of diffeomorphisms. *Commentarii mathematici Helvetici*, 49(1): 512–528, 1974.

J. N. Mather. Commutators of diffeomorphisms: II. *Commentarii Mathematici Helvetici*, 50 (1):33–40, 1975.

D. McDuff and D. Salamon. *J-holomorphic Curves and Symplectic Topology*. Number 52 in Colloquium Publications. American Mathematical Society, Providence, RI, USA, 2004.

C. A. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, 2006.

E. E. Moise. Affine structures in 3-manifolds: V. the triangulation theorem and hauptvermutung. *Annals of Mathematics*, 56(1):96–114, 1952. ISSN 0003486X. URL `http://www.jstor.org/stable/1969769`.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

E. T. Nalisnick, A. Matsukawa, Y. W. Teh, D. Görür, and B. Lakshminarayanan. Hybrid models with deep and invertible features. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4723–4732, 2019.

A. Okuno and M. Imaizumi. Minimax analysis for inverse risk in nonparametric planer invertible regression. *arXiv:2112.00213*, 2021.

A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis. Parallel WaveNet: Fast high-fidelity speech synthesis. In *Proceedings of the 35th international conference on machine learning*, pages 3918–3926, 2018.

G. Papamakarios, T. Pavlakou, and I. Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems 30*, pages 2338–2347, 2017.

G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57), 2021. URL `http://jmlr.org/papers/v22/19-1028.html`.

M. Peet. Exponentially stable nonlinear systems have polynomial Lyapunov functions on bounded regions. *IEEE Transactions on Automatic Control*, 54(5):979–987, 2009.

M. A. Puthawala, M. Lassas, I. Dokmanić, and M. V. de Hoop. Universal joint approximation of manifolds and densities by simple injective flows. *OpenReview ICLR 2022 Submission*, 2022.

F. Quinn. Ends of maps. III. Dimensions 4 and 5. *Journal of Differential Geometry*, 17(3): 503 – 521, 1982. doi: 10.4310/jdg/1214437139. URL `https://doi.org/10.4310/jdg/1214437139`.

T. Radó. Über den begriff der riemannschen fläche. *Acta Szeged*, 2:101–121, 1925.

D. Ruiz-Balet and E. Zuazua. Neural ODE control for classification, approximation and transport. *arXiv:2104.05278*, 2021.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet. On integral probability metrics, ϕ-divergences and binary classification. *arXiv:0901.2698*, 2009.

T. Teshima, I. Ishikawa, K. Tojo, K. Oono, M. Ikeda, and M. Sugiyama. Coupling-based invertible neural networks are universal diffeomorphism approximators. In *Advances in Neural Information Processing Systems 33*, pages 3362–3373, 2020a.

T. Teshima, I. Sato, and M. Sugiyama. Few-shot domain adaptation by causal mechanism transfer. In *Proceedings of the 37th International Conference on Machine Learning*, 2020b.

T. Teshima, K. Tojo, M. Ikeda, I. Ishikawa, and K. Oono. Universal approximation property of neural ordinary differential equations. *ICML Workshop, Differential Geometry meets Deep Learning*, 2020c. arXiv: 2012.02414.

W. Thurston. Foliations and groups of diffeomorphisms. *Bulletin of the American Mathematical Society*, 80(2):304–307, 1974.

C. Villani. *Optimal Transport: Old and New*. Number 338 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009.

P. N. Ward, A. Smofsky, and A. J. Bose. Improving exploration in soft-actor-critic with normalizing flows policies. *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2019. arXiv:1906.02771.

G. Yang, X. Huang, Z. Hao, M.-Y. Liu, S. Belongie, and B. Hariharan. Pointflow: 3D point cloud generation with continuous normalizing flows. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4540–4549, 2019.

C. Zhou, X. Ma, D. Wang, and G. Neubig. Density matching for bilingual word embedding. In *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)*, pages 1588–1598, 2019.

Z. Ziegler and A. Rush. Latent normalizing flows for discrete sequences. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7673–7682, 2019.

W. P. Ziemer. *Weakly Differentiable Functions: Sobolev Spaces and Functions of Bounded Variation*. Number 120 in Graduate Texts in Mathematics. Springer, New York, NY, 1989.