Para-CFlows: C^k-universal Diffeomorphism Approximators as Superior Neural Surrogates

Junlong Lyu*
Huawei Noah's Ark Lab
Hong Kong SAR, China
lyujunlong@huawei.com

Zhitang Chen*
Huawei Noah's Ark Lab
Hong Kong SAR, China
chenzhitang2@huawei.com

Wenjing Cun
Huawei Noah's Ark Lab
China
cunwenjing@huawei.com

Shengyu Zhu
Huawei Noah's Ark Lab
China
zhushengyu@huawei.com

Yanhui Geng
Huawei Noah's Ark Lab
Hong Kong SAR, China
geng.yanhui@huawei.com

Zhijie Xu
Huawei
China
xuzhijie@huawei.com

Yongwei Chen
Huawei
China
chenyongwei@huawei.com

Abstract

Invertible neural networks based on Coupling Flows (CFlows) have various applications such as image synthesis and data compression. The approximation universality for CFlows is of paramount importance to ensure the model expressiveness. In this paper, we prove that CFlows can approximate any diffeomorphism in C^k-norm if its layers can approximate certain single-coordinate transforms. Specifically, we derive that a composition of affine coupling layers and invertible linear transforms achieves this universality. Furthermore, in parametric cases where the diffeomorphism depends on some extra parameters, we prove the corresponding approximation theorems for parametric coupling flows named Para-CFlows. In practice, we apply Para-CFlows as a neural surrogate model in contextual Bayesian optimization tasks, to demonstrate its superiority over other neural surrogate models in terms of optimization performance and gradient approximations. Code will be avaliable at https://gitee.com/mindspore/models/tree/master/research/bo/paracflow.

1 Introduction

Invertible neural networks (INNs) such as coupling flows are firstly introduced as a class of generative models with a tractable likelihood [14, 28, 44], and have shown their usefulness and powerfulness in various machine learning tasks such as inverse problems [2], probabilistic inference [32] and feature extraction [25] in recent years. With plenty of successful applications of INNs, one would wonder if such a type of models have the universal expressiveness. As most generative models mainly concern about the transform between distributions, existing works such as [22, 26] focused on the expressiveness from the distribution perspective. However, the expressiveness from the distribution perspective does not result in the expressiveness from the mapping perspective, as there are a large (or even infinite) number of functions mapping the given source μ to the given target ν . In many

*Equal contribution.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

applications, knowing the distributional universality is not yet enough. One may be interested in knowing if the optimal transport [45], which finds emerging applications in many fields, e.g., machine learning [35], wireless communication [33] and economics [18], can be approximated by invertible neural networks. Therefore, beyond the distributional universality, it is also important to investigate the universality from the mapping perspective. As INNs are always invertible, it is natural to consider their approximation ability to diffeomorphisms. Diffeomorphism plays an important role in mathematics, physics and engineering domains with applications in fluid dynamics [16], wave propagation [20], robot controls [37] etc. A natural question comes to the surface: can all diffeomorphisms be approximated by INNs? Besides, whether the INNs can approximate the derivatives in the meanwhile is also interesting and important, e.g., it provides theoretical guarantee for black box optimization tasks based on surrogate models when gradients are utilized.

More importantly, in many real-world problems, e.g., 3D Euclidean groups, ODE systems and invertible PDE systems with time t as its parameter, diffeomorphisms are usually described as a parametric type: a parametric diffeomorphism is a function $f(\boldsymbol{y}, \boldsymbol{x}) : \mathbb{R}^{m+d} \to \mathbb{R}^d$ such that for any fixed parameter $\boldsymbol{y}_0 \in \mathbb{R}^m$, $f(\boldsymbol{y}_0, \cdot)$ is a diffeomorphism between $\mathbb{R}^d$ and $\mathbb{R}^d$. It is important to know if they can also be approximated by INNs with additional inputs of parameters. Existing works including [42] have been proposed to investigate the universality of INNs over diffeomorphisms, however it is not able to handle C^k-diffeomorphisms when $k = d + 1$ with d as the dimension of the space. In this paper, we address this limitation and more importantly, inspired by the contextualization way proposed in [3], we generalize this structural theorem to the parametric diffeomorphism situation, by concatenating parameters with the variable to be transformed, forming a new subgroup over a higher dimensional space, and provide a complete proof. This structural theorem shows that any compact-supported diffeomorphisms can be decomposed into finitely many compositions of single-coordinate transforms, which we can easily prove to be C^k-approximated by, e.g., affine coupling flows together with dimension augmentation ([15], [46]).

Lastly, we suggest using `Para-CFlow`, an affine coupling flow model with dimension augmentation and contextualization, as a superior neural surrogate for black box optimization. The superiority lies in: 1) it uses INN architecture to preserve full rank of the action; 2) it has C^k-universality when the target system satisfies certain conditions. Specifically, 2) facilitates much more efficient surrogate and gradient-descent based black box optimization compared to other neural surrogates.

Overall, our contribution is three-fold: 1) We improve the results in [42] to higher-order derivatives with a simpler proof, and give an upper bound on number of layers needed for approximating a certain diffeomorphism; 2) We generalize coupling flows to parametric coupling flows and prove their C^k-universal approximation to parametric diffeomorphisms using a novel proving technique; 3) We propose a practical neural network structure of the parametric affine coupling flows and verify the advantage of using this for contextual Bayesian Optimization (BO) tasks.

2 Preliminary

Here we introduce some prior knowledge on (parametric) diffeomorphism, INNs and universality, as well as existing works on universality of INNs. As our theoretical results are motivated by and improve upon [42], some of our notations and preliminaries are adopted from them. In what follows, we always assume $k, m, n, d \in \mathbb{N}^+$, where k represents the derivative order, and m, n, d represent dimensionalities for some Euclidean spaces. All vectors are supposed to be row vectors.

2.1 C^k-diffeomorphism groups on $\mathbb{R}^d$

C^k-diffeomorphisms (group). Consider an invertible map f from $\mathbb{R}^d$ to $\mathbb{R}^d$. f is said to be a C^k-diffeomorphism, if f has up to k-th continuous derivatives, and $\det Df(\boldsymbol{x}) \neq 0$ for all $\boldsymbol{x}$. One can easily verify that, given $f, g : C^k$ diffeomorphisms from $\mathbb{R}^d$ to $\mathbb{R}^d$, $f \circ g$ and f^{-1} is still a C^k-diffeomorphism from $\mathbb{R}^d$ to $\mathbb{R}^d$. If we denote $\text{Diff}^k(\mathbb{R}^d) \triangleq \{f : f \text{ is a } C^k \text{ diffeomorphism over } \mathbb{R}^d\}$, we see that $\text{Diff}^k(\mathbb{R}^d)$ has a natural group structure with composition as its group operator.

Compactly supported C^k-diffeomorphisms. In real applications, finite data cannot cover the whole space $\mathbb{R}^d$. Besides, existing approximation theories only guarantee the capability over some bounded compact set K. Therefore we consider compactly supported functions or diffeomorphisms. A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be compactly supported, if there exists a compact set K such

that $f(\boldsymbol{x}) = 0$, for all $\boldsymbol{x} \notin K$. Similarly, a diffeomorphism f is said to be compactly supported, if there exists a compact set K such that $f(\boldsymbol{x}) = \boldsymbol{x}$, for all $\boldsymbol{x} \notin K$, resulting in that all the non-diagonal components of Df are compactly supported. We define $\text{Diff}_c^k(\mathbb{R}^d) \triangleq \{f \in \text{Diff}^k(\mathbb{R}^d) : f \text{ compactly supported}\}$. Obviously $\text{Diff}_c^k(\mathbb{R}^d)$ is a subgroup of $\text{Diff}^k(\mathbb{R}^d)$.

C^k-parametric diffeomorphisms. A parametric diffeomorphism is a family of diffeomorphisms with some parameter α: $\{f_\alpha\}_{\alpha \in A}$, where for any given α, f_α is a $\mathbb{R}^d \to \mathbb{R}^d$ diffeomorphism. Usually the parameter is described by some vector $\boldsymbol{y} \in \mathbb{R}^m$, and $f_{\boldsymbol{y}}$ varies continuously or smoothly w.r.t. $\boldsymbol{y}$. We denote $f(\boldsymbol{y}, \boldsymbol{x}) = f_{\boldsymbol{y}}(\boldsymbol{x})$. We can embed it into a higher dimensional diffeomorphism $F(\boldsymbol{y}, \boldsymbol{x}) = \big(\boldsymbol{y}, f(\boldsymbol{y}, \boldsymbol{x})\big)$. One can verify directly that $F \in \text{Diff}^k(\mathbb{R}^{m+d})$ given that (1) $f(\boldsymbol{y}_0, \boldsymbol{x}) \in \text{Diff}^k(\mathbb{R}^d)$ for any fixed $\boldsymbol{y}_0$; (2) f is k-th differentiable w.r.t. $(\boldsymbol{y}, \boldsymbol{x})$. Obviously $\text{Diff}_c^{k,m,d} \triangleq \{F \in \text{Diff}_c^k(\mathbb{R}^{m+d}) : F(\boldsymbol{y}, \boldsymbol{x}) = (\boldsymbol{y}, f(\boldsymbol{y}, \boldsymbol{x})) \text{ with } \boldsymbol{y} \in \mathbb{R}^m, \boldsymbol{x} \in \mathbb{R}^d\}$, whose elements keep the first m coordinates and change the last d ones, is a subgroup of $\text{Diff}_c^k(\mathbb{R}^{m+d})$.

2.2 INNs based on parametric coupling flows

Here we introduce the classical INNs and investigate the space generated by them.

Invertible linear transforms. First, let us define the invertible linear transforms (ILT):

$$\text{ILT}_d \triangleq \{\mathcal{L} : \mathcal{L}\boldsymbol{x}^T = A\boldsymbol{x}^T + \boldsymbol{b}^T, A \in \text{GL}_d(\mathbb{R}), \boldsymbol{b} \in \mathbb{R}^d\}$$

and the parametric case where $\boldsymbol{y}$ are parameters, here $\text{GL}_d(\mathbb{R})$ includes all $d \times d$ regular matrices.

$$\text{ILT}_{m,d} \triangleq \{\mathcal{L} : \mathcal{L}\begin{pmatrix}\boldsymbol{y}^T \\ \boldsymbol{x}^T\end{pmatrix} = \begin{pmatrix} I_m & 0 \\ B & A\end{pmatrix}\begin{pmatrix}\boldsymbol{y}^T \\ \boldsymbol{x}^T\end{pmatrix} + \begin{pmatrix}\boldsymbol{0} \\ \boldsymbol{b}^T\end{pmatrix}, A \in \text{GL}_d(\mathbb{R}), \boldsymbol{b} \in \mathbb{R}^d\, B \in \mathbb{R}^{d\times m}\}.$$

Coupling flows. We now define invertible coupling flows [34], a specific type of nonlinear transforms:

$$\Phi_{d,i,\phi} : \mathbb{R}^d \longrightarrow \mathbb{R}^d, \quad (\boldsymbol{x}_{\leq i}, \boldsymbol{x}_{>i}) \longmapsto (\boldsymbol{x}_{\leq i}, \phi(\boldsymbol{x}_{\leq i}, \boldsymbol{x}_{>i}))\,,$$

where $\phi(\boldsymbol{x}_{\leq i}, \cdot) : \mathbb{R}^{d-i} \to \mathbb{R}^{d-i}$ is a diffeomorphism for each fixed $\boldsymbol{x}_{\leq i}$. Specifically, when $\phi(\boldsymbol{x}_{\leq i}, \boldsymbol{x}_{>i}) = \boldsymbol{x}_{>i} \odot \exp(\sigma(\boldsymbol{x}_{\leq i})) + t(\boldsymbol{x}_{\leq i})$, it is the so-called affine coupling flow [13] and we denote $\Phi_{d,i,\sigma,t} = \Phi_{d,i,\phi}$ for such ϕ. σ, t are some functions with $d - i$ output units, typically modeled with deep neural networks. $\odot$ represents the point-wise product. $\boldsymbol{x}_{\leq i} = (x_1, \cdots, x_i)$, $\boldsymbol{x}_{>i} = (x_{i+1}, \cdots, x_d)$ for $\boldsymbol{x} = (x_1, \cdots, x_d)$. Similarly, for parametric cases, we have:

$$\Phi_{d,i,m,\phi} : \mathbb{R}^{m+d} \longrightarrow \mathbb{R}^{m+d}, \quad (\boldsymbol{y}, \boldsymbol{x}_{\leq i}, \boldsymbol{x}_{>i}) \longmapsto (\boldsymbol{y}, \boldsymbol{x}_{\leq i}, \phi(\boldsymbol{y}, \boldsymbol{x}_{\leq i}, \boldsymbol{x}_{>i}))\,, \tag{1}$$

and specifically, $\phi(\boldsymbol{y}, \boldsymbol{x}_{\leq i}, \boldsymbol{x}_{>i}) = \boldsymbol{x}_{>i} \odot \exp(\sigma(\boldsymbol{y}, \boldsymbol{x}_{\leq i})) + t(\boldsymbol{y}, \boldsymbol{x}_{\leq i})$ for affine-type coupling flows and we denote $\Phi_{d,i,m,\sigma,t} = \Phi_{d,i,m,\phi}$ for such ϕ.

Single-coordinate affine coupling flows (SACFs) denotes the flows with only the last coordinate changed. Let $\mathcal{H}_d$ be a set of functions from $\mathbb{R}^d$ to $\mathbb{R}$. We define

$$\mathcal{H}\text{-SACF}_d \triangleq \{\Phi_{d,d-1,\sigma,t} : \sigma, t \in \mathcal{H}_{d-1}\}, \text{and } \mathcal{H}\text{-SACF}_{m,d} \triangleq \{\Phi_{d,d-1,m,\sigma,t} : \sigma, t \in \mathcal{H}_{m+d-1}\}.$$

Note that any multi-coordinates affine coupling flows can be represented by finite composition of SACFs and invertible linear transforms, it suffices to just consider the universality of SACFs.

Invertible neural networks. Now let us combine linear invertible transform layers and some coupling flow layers to construct our INNs. Let $\mathcal{G}$ be a set consisting of invertible coupling flows. We define the set of INNs based on $\mathcal{G}$ and the parametric case:

$$\mathcal{G}\text{-INN}_d \triangleq \{g_s \circ W_s \circ \cdots \circ g_1 \circ W_1 : s \in \mathbb{N}, g_i \in \mathcal{G}, W_i \in \text{ILT}_d\},$$
$$\mathcal{G}\text{-INN}_{m,d} \triangleq \{g_s \circ W_s \circ \cdots \circ g_1 \circ W_1 : s \in \mathbb{N}, g_i \in \mathcal{G}, W_i \in \text{ILT}_{m,d}\}.$$

When $\mathcal{G}$ contains $\mathcal{H}$-SACF$_d$ (or $\mathcal{H}$-SACF$_{m,d}$ in parametric cases), it is equivalent to replacing ILT$_d$ (or ILT$_{m,d}$) by the symmetric group S_d containing all the permutation over d coordinates operating on $\boldsymbol{x}$. This type of networks are well known as Real-NVPs [13]. One can also define other types of coupling flows. Nevertheless, the theoretical guarantee of any coupling flow can be verified by simply checking their universality to single-coordinate transforms, as our main result stated in Thm. 3.5.

2.3 Different types of universality and their relations

Here we define the functional universality. See Appx. A for distributional universality and more.

Definition 2.1. *($L^p/L^\infty/C^k$-universality [9]). Let $\mathcal{M}$ be a set of measurable mappings from $\mathbb{R}^n$ to $\mathbb{R}^d$. Let $p \in [1, \infty)$, $k \in \mathbb{N}^+$ and let $\mathcal{F}$ be a set of measurable mappings $f : U_f \to \mathbb{R}^d$, where U_f is a measurable subset of $\mathbb{R}^n$ which may depend on f. We say that $\mathcal{M}$ has L^p (or L^∞, C^k)-universality for $\mathcal{F}$, if for any $f \in \mathcal{F}$ any $\epsilon > 0$, and any compact subset $K \subseteq U_f$, there exists a $g \in \mathcal{M}$ such that, $\|f - g\|_{L^p(K)} < \epsilon$ (or $\|f - g\|_{L^\infty(K)}, \|f - g\|_{C^k(K)} < \epsilon$, definition of these norms in Appx. A).*

If $\mathcal{M}$ has the C^k-universality, $\mathcal{M}$ has the L^∞-universality because $C^k(K)$ is dense in $L^\infty(K)$ for any compact K. Also, if $\mathcal{M}$ has the L^∞-universality, $\mathcal{M}$ has the L^p-universality for any $1 \leq p < \infty$. If an L^p-universality is satisfied for some $1 < p < \infty$, then for any $1 \leq q < p$, L^q-universality is ensured. Finally, L^1-universality implies distributional universality but the reverse is not true.

2.4 Related works

Coupling (or triangular) flows. Authors of [24, 6] proved that the distributional universality of a flow family h can be deduced, if h is dense in the set of all monotone functions by pointwise convergence topology. In particular, [22] proved the distributional universality for neural autoregressive flows, and [26] proved that for sum-of-square flows. [42] generalized it to L^p universality with $1 \leq p < \infty$ for affine coupling flows, and L^∞ universality for autoregressive flows and sum-of-square flows.

Non-coupling flows. The expressiveness of general non-coupling flows is not well studied, as most existing works restrict the form of nonlinearity to certain types, e.g., planar and radial flows [36], Sylvester flows [43], in order to easily compute the determinant of the Jacobian matirx and the inverse maps. [29] gave a first study over some specific distributions, while the universal distributional results still remains unknown. Another type, iResNet [5], was proposed based on residual network (ResNet) [21] to improve nonlinearity as well as the log determinant computation efficiency. [46] proved that iResNet, capped by linear layers or with extra dimensions, has C^0-universality.

Continuous time flows. The ODE-based method is also a major class of flow models as introduced in [10, 19, 11, 38]. [11] gave counterexamples for the C^0-universality of neural ODEs, however its distributional universality is not yet addressed. An "augmented" neural ODEs was proposed by [15] and then analyzed by [46]. They showed that embedding the original d dimensional neural ODE into a $d + 1$ dimensional space can bypass the original counterexamples, achieving a C^0-universality.

3 Main results

In this section, we present our main results, following the architecture similar to [42]. Section 3.1 provides a proof for non-parametric cases, showing that: 1) the C^k-universality over the $\text{Diff}_c^k(\mathbb{R}^d)$ can be achieved when the C^k-universality over a simple subgroup of $\text{Diff}_c^k(\mathbb{R}^d)$ is ensured; 2) affine coupling layers with one zero-padding has C^k-universality over such a simple subgroup. In addition, we estimate the number of layers needed for approximate certain diffeomorphism in Thm 3.3. In Section 3.2, we generalize these results to parametric cases with a novel proof technique.

3.1 Non-parametric cases

Here we outline the main steps of our proof, and the complete proof is available in Appx. B.

Recall the space $\text{Diff}_c^k(\mathbb{R}^d)$: All the $C^k, \mathbb{R}^d \to \mathbb{R}^d$ diffeomorphisms which are compactly supported. There are two advantages of choosing $\text{Diff}_c^k(\mathbb{R}^d)$ instead of $\text{Diff}^k(\mathbb{R}^d)$: firstly the compactly supported property greatly simplifies the structure of this group; secondly for any diffeomorphism $F \in \text{Diff}^k(\mathbb{R}^d)$ and any compact set K, there exists a compactly supported diffeomorphism $f \in \text{Diff}_c^k(\mathbb{R}^d)$ such that $F|_K = f|_K$, i.e., $F(\boldsymbol{x}) = f(\boldsymbol{x}),$ for all $\boldsymbol{x} \in K$ (See Lemma 6 in [42]).

Directly proving the universality to a general diffeomorphism is difficult. It is beneficial to decompose the original hard problem into a series of simpler ones as follows.

Definition 3.1. *A diffeomorphism $\tau \in \text{Diff}_c^k(\mathbb{R}^d)$ is in single-coordinate transforms $S_c^{k,d}$, if $\tau(\boldsymbol{x}) = (x_1, x_2, \cdots, x_{i-1}, \tau_i(\boldsymbol{x}), x_{i+1}, \cdots, x_d)$, i.e., only one coordinate is altered, and $\tau_i(\boldsymbol{x}) = \tau_i(x_1, \cdots, x_d)$ is a function from $\mathbb{R}^d$ to $\mathbb{R}$ which is monotonic for x_i .*

The theorem as follows shows that, all diffeomorphisms in $\text{Diff}_c^k(\mathbb{R}^d)$ can be constructed from $S_c^{k,d}$.

Theorem 3.2. *For any subgroup $H \subseteq \text{Diff}_c^k(\mathbb{R}^d)$ s.t. $S_c^{k,d} \subseteq H$, $H = \text{Diff}_c^k(\mathbb{R}^d)$. That is, for any $f \in \text{Diff}_c^k(\mathbb{R}^d)$, there exists $\tau_1, \tau_2, \cdots, \tau_s \in S_c^{k,d}, s \in \mathbb{N}$ s.t. $f = \tau_s \circ \tau_{s-1} \circ \cdots \circ \tau_1$.*

Proof. The following two steps are sketched: 1) If $H \subseteq \text{Diff}_c^k(\mathbb{R}^d)$ and $S_c^{k,d} \subseteq H$, H contains all near-identity diffeomorphisms (Def. B.3). The detail is stated in Cor. B.6. 2) If $H \subseteq \text{Diff}_c^k(\mathbb{R}^d)$ contains all near-identity diffeomorphisms, $H = \text{Diff}_c^k(\mathbb{R}^d)$. The detail is stated in Lem. B.4. □

We further investigate the number of single-coordinate transforms needed to represent an arbitrary diffeomorphism in Thm. 3.3 as follows. The definition of $\text{dist}_{\text{Diff}_c^1(\mathbb{R}^d)}$ and the proof is provided in Appx. B.2

Theorem 3.3. *Given $f \in \text{Diff}_c^k(\mathbb{R}^d)$, we denote $\text{dist}_{\text{Diff}_c^1(\mathbb{R}^d)}(f, I) = \ell$ for some $\ell > 0$, the minimal length of paths between f and the identity map I lying in $\text{Diff}_c^1(\mathbb{R}^d)$ with the C^1-norm induced metric. Then f can be decomposed into $d(d-1)\ell$ single-coordinate transforms.*

Thm. 3.2 decomposes the original general complex diffeomorphism into finitely-many simple diffeomorphisms. It is natural to ask: if we can approximate all the $S_c^{k,d}$ well, can we also approximate $\text{Diff}_c^k(\mathbb{R}^d)$ well? Thm. 3.4 gives a positive answer to the question.

Theorem 3.4. *(Approximation for composition.) Suppose G is a group of diffeomorphisms over $\mathbb{R}^d$. Given a set of diffeomorphisms F over $\mathbb{R}^d$, denote $\mathcal{F}$ as the semigroup generated by F. If G has C^k-universality for F, G has C^k-universality for $\mathcal{F}$.*

Proof. We only need to prove the following statement: If G has C^k-universality for $f_1, f_2 \in \mathcal{F}$, then G has C^k-universality for $f_1 \circ f_2 \in \mathcal{F}$. By the property of semigroups, we know that G has C^k-universality for the whole $\mathcal{F}$. Details can be found in Appx. B.3. □

Combining Thm. 3.2 and 3.4, noticing that $S_c^{k,d}$ is inverse-invariant (for any $\tau \in S_c^{k,d}, \tau^{-1} \in S_c^{k,d}$), we can immediately obtain following corollary.

Corollary 3.5. *Suppose G is a group of diffeomorphisms over $\mathbb{R}^d$, if G has C^k-universality for $S_c^{k,d}$, then G has C^k-universality for $\text{Diff}_c^k(\mathbb{R}^d)$.*

As stated in Cor. 3.5, any INNs that approximate $S_c^{k,d}$ will eventually approximate $\text{Diff}_c^k(\mathbb{R}^d)$ in C^k-norm, which is an extension from the L^p and L^∞ universality stated in [42]. Furthermore, we generalize this result to parametric cases in the next section.

Affine-Coupling Flows. Now we only need to prove the universality of $\mathcal{G}$-INN$_d$ over $S_c^{k,d}$, when $\mathcal{G}$ contains $\mathcal{H}$-SACF$_d$. However, even for such a simple class of diffeomorphisms, $\mathcal{G}$-INN$_d$ is not able to approximate it very well: existing works [42] shows that $\mathcal{G}$-INN$_d$ has L^p universality over $S_c^{k,d}$. Whereas they gave a negative conjecture for L^∞ universality, let alone the C^k-universality.

However, despite the lack of easily computed log-determinant for density evaluation, dimension augmentation has been proven to reach better approximation ability for Neural ODEs [15], [46]. Similarly, when the single-transform is embedded into a higher dimension Euclidean space by padding zeros, the approximation ability becomes much stronger, achieving C^k-universality, far beyond L^p universality. We first define the canonical immersion ι_d, and submersion π_d:

$$\iota_d \colon (x_1, \cdots, x_d) \mapsto (x_1 \cdots, x_d, 0), \quad \pi_d \colon (x_1, \cdots, x_d, x_{d+1}) \mapsto (x_1 \cdots, x_d).$$

Thm. 3.6 says by raising the dimension by 1, we can achieve C^k -universality. Proof is in Appx. B.4.

Theorem 3.6. *For any compact set $K \subseteq \mathbb{R}^d$, and any $\tau \in S_c^{k,d}$, $\tau : (x_1, x_2, \cdots, x_{d-1}, x_d) \longrightarrow (x_1, x_2, \cdots, x_{d-1}, \tau_d(\boldsymbol{x}))$, there exists $\tilde{\tau}$ in $\mathcal{G}$-INN$_{d+1}$ with $\mathcal{G} \supseteq \mathcal{H}$-SACF$_{d+1}$, such that $\|\tau - \pi_d \circ \tilde{\tau} \circ \iota_d\|_{C^k(K)} < \epsilon$, given that $\mathcal{H}_d$ has C^k-universality over functions $K \to \mathbb{R}$.*

3.2 Parametric cases

Following Section 3.1, we generalize the results to parametric cases. We use a novel proof technique, by concatenating parameters with the primary input to be transformed, forming a new subgroup over a higher dimensional space, and exploiting the structure of this subgroup.

We denote $S_c^{k,m,d} \triangleq S_c^{k,m+d} \cap \text{Diff}_c^{k,m,d}$, which is the set of transforms acting on a single coordinate of $\boldsymbol{x}$. Recall the definition of compactly supported parametric diffeomorphisms group in Section 2.1, where $\boldsymbol{y}$ acts as parameters. One might raise a question: why not directly use the result in non-parametric case, as all the elements in $\text{Diff}_c^{k,m,d}$ are also in $\text{Diff}_c^{k}(\mathbb{R}^{m+d})$? Note that the decomposition theorem only ensures that $S_c^{k,m+d}$ can generate $\text{Diff}_c^{k}(\mathbb{R}^{m+d})$ and then $\text{Diff}_c^{k,m,d}$, it does not guarantee that $S_c^{k,m,d}$ can generate $\text{Diff}_c^{k,m,d}$. Without such a guarantee, we have to use operators that may alter coordinates of both $\boldsymbol{y}$ and $\boldsymbol{x}$, and the coordinates of $\boldsymbol{y}$ in the output layer should be exactly the same as the coordinates of $\boldsymbol{y}$ in the input layer, which restricts the flexibility and increases the approximation difficulty. Thus we derive new dedicated theorems for parametric cases.

Theorem 3.7. *For any subgroup $H \subseteq \text{Diff}_c^{k,m,d}$ s.t. $S_c^{k,m,d} \subseteq H$, $H = \text{Diff}_c^{k,m,d}$. That is, for any $f \in \text{Diff}_c^{k,m,d}$, there exists $\tau_1, \tau_2, \cdots, \tau_s \in S_c^{k,m,d}$, $s \in \mathbb{N}$ s.t. $f = \tau_s \circ \tau_{s-1} \circ \cdots \circ \tau_1$.*

Proof. The following two steps are sketched: 1) If $H \subseteq \text{Diff}_c^{k,m,d}$ and $S_c^{k,m,d} \subseteq H$, H contains all near-identity diffeomorphisms (Def. B.3). The detail is stated in Cor. B.6. 2) If $H \subseteq \text{Diff}_c^{k,m,d}$ contains all near-identity diffeomorphisms, $H = \text{Diff}_c^{k,m,d}$ The detail is stated in Lem. B.10. □

Given the above theorem, we easily obtain the C^k-universality for $\text{Diff}_c^{k,m,d}$ following Section 3.1.

Theorem 3.8. *Suppose G is a group of diffeomorphisms over $\mathbb{R}^{m+d}$. If G has C^k-universality for $S_c^{k,m,d}$, then G has C^k-universality for $\text{Diff}_c^{k,m,d}$.*

A proof of $\mathcal{G}$-INN$_{m,d}$ universality over $S_c^{k,m,d}$, when $\mathcal{G}$ contains $\mathcal{H}$-SACF$_{m,d}$, follows Thm. 3.6 in Section 3.1, and thus the detail of the proof is skipped here.

4 `Para-CFlows` as superior neural surrogates for black box optimization

In real applications such as control and optimization of telecommunication networks, power grid systems, and electronic systems, we aim to optimize a complicated system under various contexts. Specifically, we are interested in an unknown function $b(\boldsymbol{a}, \boldsymbol{c}) : \mathcal{A} \times \mathcal{C} \to \mathbb{R}$, where $\mathcal{A} \subseteq \mathbb{R}^{d_a}, \mathcal{C} \subseteq \mathbb{R}^{d_c}$ are connected and positive measurable. $\boldsymbol{a}$ is a set of actions one can tune and $\boldsymbol{c}$ is a context vector representing exogenous stimuli. In most cases, we are interested in optimizing the function b, i.e. finding $\boldsymbol{a}^* = \arg\max_{\boldsymbol{a}} b(\boldsymbol{c}, \boldsymbol{a})$ given a context $\boldsymbol{c}$. Our task is to quickly find $\boldsymbol{a}^*$ at each $\boldsymbol{c}$ using as few queries as possible. In this paper, we advocate to use `Para-CFlows`, affine coupling flow models with dimension augmentation and parameterization (context information inputted as parameters), to model the relationship between b and $\boldsymbol{a}, \boldsymbol{c}$. Specifically, `Para-CFlows` are constructed as follows:
Step 1: We first map $\boldsymbol{a}$ into a higher dimensional space $\mathbb{R}^d$ $(d > d_a)$: $\phi_0 : \boldsymbol{a} \mapsto (\boldsymbol{a}, \boldsymbol{a}\boldsymbol{W}) \triangleq \tilde{\boldsymbol{a}}^{(0)}$, where $\boldsymbol{W} \in \mathbb{R}^{d_a \times (d-d_a)}$. By construction, $\text{rank}(\phi_0) = d_a$.
Step 2: We further compose ϕ_0 with a series of invertible affine coupling layers $\phi_i, 1 \leq i \leq N$ operating in $\mathbb{R}^d$ (Eq. (1)). For any $\boldsymbol{x} \in \mathbb{R}^d$, let $d' = \max\{d_a, [\frac{d}{2}]\}$, we define

$$\phi_{\boldsymbol{c},i}(\boldsymbol{x}) = (\boldsymbol{x}_{\leq d'}, \boldsymbol{x}_{>d'} \odot \exp(\sigma_i(\boldsymbol{c}, \boldsymbol{x}_{\leq d'})) + t_i(\boldsymbol{c}, \boldsymbol{x}_{\leq d'}))\,,$$

σ_i, t_i are functions with $d - d'$ outputs, implemented by any nonlinear functions like DNNs.
Step 3: We write $\tilde{\boldsymbol{a}}^{(i)} = \phi_i P_i(\tilde{\boldsymbol{a}}^{(i-1)}), 1 \leq i \leq N$, where P_i are any fixed random permutation, and

$$\tilde{\boldsymbol{a}}^{(N)} = (\phi_{\boldsymbol{c},N} P_N) \circ \cdots \circ (\phi_{\boldsymbol{c},1} P_1) \circ \phi_0(\boldsymbol{a}) \triangleq \phi(\boldsymbol{c}, \boldsymbol{a}), \tag{2}$$

which is followed by a simple function $\hat{b} : \mathbb{R}^d \to \mathbb{R}$ represented by a simple neural network that shall be trained. Then the whole model is trained by minimizing the Mean Square Error (MSE). $\min_{\sigma_i, t_i, \hat{b}} \frac{1}{M} \sum_{1 \leq j \leq M} \left(b_j - \hat{b}(\phi(\boldsymbol{c}_j, \boldsymbol{a}_j))\right)^2$ with $b_j = b(\boldsymbol{c}_j, \boldsymbol{a}_j)$ are trainning samples.

Although it is a straight-forward extension of coupling flows and a similar one was found in [3], our key contribution mainly lie in the theoretical part: for the first time, we prove its C^k-universality to

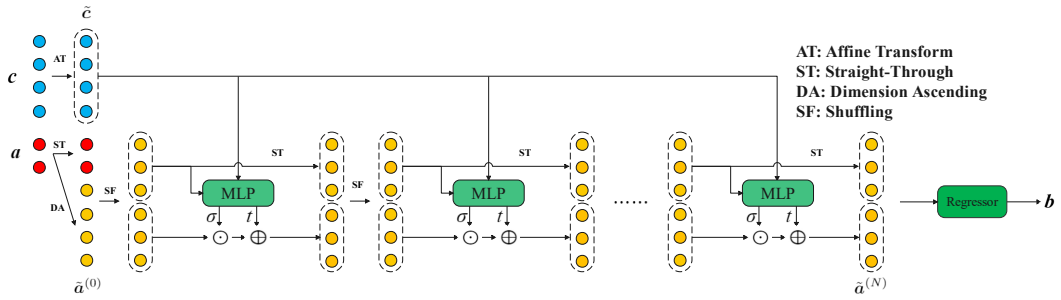


Figure 1: Network structure of `Para-CFlow`

parametric diffeomorphisms and with this property, we advocate to apply it to address an vital action sensitivity issue in black box optimization other than computer vision tasks where flow models are most widely applied.

We now analyze the superiority of `Para-CFlows` as the surrogate for black box optimization.

Property A: `Para-CFlows` **preserves full information of actions.** Existing methods [30, 12, 1, 17, 23] for BO do not handle well the weak action sensitivity issue of many real system optimization tasks [40], where the actions/configuration parameters' effect is overwhelmed by the highly volatile system state and thus the surrogate model $\hat{f}(\boldsymbol{c}, \boldsymbol{a})$ can easily degenerate to $\hat{f}(\boldsymbol{c})$ due to a small sample size, especially when $d_{\boldsymbol{c}} \gg d_{\boldsymbol{a}}$. On the contrary, `Para-CFlow`, by construction, preserves all the information (the rank) of $\boldsymbol{a}$ up to the last feature extraction layer before the final regression layer.

Property B: `Para-CFlows` **has C^k universality to all rank-preserving systems.** Based on the universality of parametric coupling flows in Section 3, we further investigate what kind of objective function b can be approximated by `Para-CFlows`. We give the necessary condition to b as follows.

Assumption 4.1. *b is an observation of the inner system state $\Phi : b(\boldsymbol{c}, \boldsymbol{a}) = l(\Phi(\boldsymbol{c}, \boldsymbol{a}))$ and the system is `rank-preserving`, i.e. the inner state $\Phi(\boldsymbol{c}, \boldsymbol{a}) : \mathcal{C} \times \mathcal{A} \to \mathbb{R}^{d_\phi}$ is a differentiable function and preserves full information of actions, i.e., rank $\left(\frac{\partial \Phi}{\partial \boldsymbol{a}}\right) = d_a$.*

Under Assumption 4.1, we can get the following result, which ensures the expressiveness of our architecture,

Theorem 4.2. *For any system $\Phi : \mathcal{C} \times \mathcal{A} \to \mathbb{R}^{d_\phi}$ satisfying Asmp. 4.1, if $d \geq 2d_\phi$ and Φ have up to k-th order derivatives, then for any $\epsilon > 0$, we can always find a ϕ in the form* (2)*, such that $\|\Phi(\boldsymbol{c}, \boldsymbol{a}) - \pi\phi(\boldsymbol{c}, \boldsymbol{a})\|_{C^k(\mathcal{C}\times\mathcal{A})} \leq \epsilon$, given that $\sigma_i, t_i, \hat{b}$ are C^k-universal function approximator. Here π is the canonical submersion: $\mathbb{R}^d \to \mathbb{R}^{d_\phi}$, $(x_1, x_2, \cdots, x_{d_\phi}, \cdots, x_d) \mapsto (x_1, x_2, \cdots, x_{d_\phi})$.*

Readers can check Appx. B.6 for the proof. Thm 4.2 shows that the inner system Φ can be uniformly approximated by the feature extraction layer ϕ of `Para-CFlow`. The observation function $l(\cdot)$ can then be approximated by $\hat{b}$ as a result of the universality of Multi-Layer Perceptons (MLPs).

The C^k-approximability property of `Para-CFlow` are of vital importance for black box optimization. Let us just focus on the exploitation phase, i.e., $\boldsymbol{a}^* = \arg\max \hat{b}(\phi(\boldsymbol{c}, \boldsymbol{a}))$. When the action space is continuous and high dimensional (though in this paper, we focus on the context dominating tasks where $d_{\boldsymbol{c}} \gg d_{\boldsymbol{a}}$), one is not able to find the optima easily by brute-force search. Many existing works utilize meta-heuristic such as Evolutionary Algorithm (EA) of which the performance heavily depends on the population size and the number of iterations, and thus it is computation demanding and time consuming. Instead, one can apply Gradient Descent (GD) to the surrogate if it is differentiable, i.e. $\boldsymbol{a}^{t+1} \leftarrow \boldsymbol{a}^t + \eta \nabla_{\boldsymbol{a}} \hat{b}(\phi(\boldsymbol{c}, \boldsymbol{a}))|_{\boldsymbol{a}=\boldsymbol{a}^t}$. However, for the GD method to work, it requires that the surrogate approximates not only the function value but also its first-order derivative. However function value approximation does not necessarily guarantee derivative approximation. A simple illustrative example is $f_n(x) = g(x) + \sin(nx)/n$. One can easily observe that $f_n(x) \to g(x)$, but $f_n'(x) = g'(x) + \cos(nx)$ will never converge to $g'(x)$ when $n \to \infty$. With this argument, we advocate `Para-CFlow` as an adequate choice for surrogate modeling for continuous action space BO. We also find it suitable to approximate Q-function in Reinforcement Learning (RL), especially

for DPG [39], DDPG[31] and references therein where one need to estimate the gradient of the Q-function w.r.t. the action to train the actor network. Extension to DRL will be left for future work.

With the aforementioned properties, our proposed `Para-CFlow` model can serve as a superior neural surrogates and in the setting of BO, we can easily adopt existing SOTA methods of Bayesian deep learning and deep ensembles to facilitate the uncertainty modeling with `Para-CFlow`.

5 Experiments

In this section, we conduct experiments to verify the expressiveness of our proposed `Para-CFlow` and demonstrate its advantage as neural surrogate models in application to black-box optimizations.

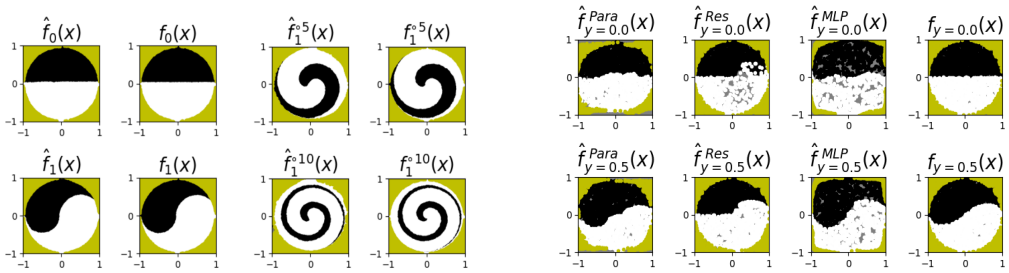
5.1 Learning diffeomorphism "Taiji"

We demonstrate the expressiveness of our model to learn a parametric diffeomorphism with a synthetic task called "Taiji". We define our target parametric diffeomorphism:

$$f_y : (\rho\cos\theta, \rho\sin\theta) \mapsto (\rho\cos\tilde{\theta}, \rho\sin\tilde{\theta}), \text{where } \tilde{\theta} = \theta + y\arccos(\min(\rho, 1)). \tag{3}$$

On the unit circle $\rho = 1$, it is not differentiable, thus we expect the trained model to have irregular derivatives over the unit circle, while maintain accurate derivatives elsewhere.

First, we design a task with sufficient amount of data to verify the C^1 approximation ability of `Para-CFlows`. We generate samples $\{(\boldsymbol{x}_i, y_i)\}_{1\le i\le 30000}$ with $\boldsymbol{x}_i \sim \mathcal{U}[-1,1]^2$ and $y_i \sim \mathcal{U}[0,1]$ independently. Under the polar coordinate representation, i.e., $\boldsymbol{x}_i = (\rho_i\cos\theta_i, \rho_i\sin\theta_i)$, we calculate the target $f_{y_i}(\boldsymbol{x}_i) = (\rho_i\cos\tilde{\theta}_i, \rho_i\sin\tilde{\theta}_i)$ according to Eq. (3). We use a 6-layers affine coupling flows, each layer composed with a random permute. The coupling functions σ, t are implemented by 1-hidden-layer Multi-Layer Perceptron (`MLP`) with hidden-unit number comparable to input dimension. We pad zeros to the input $\boldsymbol{x}$, and use the first two output dimensions to compute the loss.

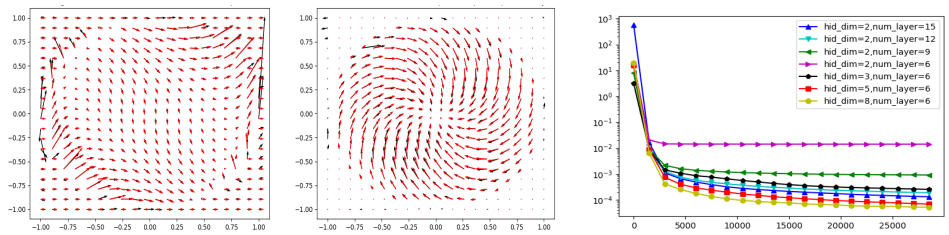


(a) Trained $\hat{f}_y(\boldsymbol{x})$ v.s. Target $f_y(\boldsymbol{x})$ (b)

Figure 2: 2(a): learning "Taiji" diffeomorphism tasks in Section 5.1. Black region, $x_2 > 0 \cap \rho \le 1$; white region, $x_2 \le 0 \cap \rho \le 1$; yellow region, $\rho > 1$. $\rho = \sqrt{x_1^2 + x_2^2}$ and $\boldsymbol{x} = (x_1, x_2)$. $\hat{f}_1^{\circ N}$ denotes the N-fold composition $\hat{f}_1 \circ \cdots \circ \hat{f}_1$. 2(b): Comparing `Para-CFlow` with `Resnet` and `MLP` with small sample size and parameter $y^{(i)} = y, 1 \le i \le 100$. Grey color area has no point resides.

On the left panel of Fig. 2(a), we observe good consistency between the trained model $\hat{f}_y(\boldsymbol{x})$ and $f_y(\boldsymbol{x})$, when $y \in [0,1]$. Note that we only use samples with $y_i \in [0,1]$ for training, and thus the model generalizability to $y > 1$ is not guaranteed. However, as the true diffeomorphism satisfies $f_{y_1} \circ f_{y_2} = f_{y_1+y_2}$, we iterate the trained model recurrently and expect the final output to achieve the effect for $y > 1$. The results on the right panel of Fig. 2(a) justify that the cumulative error from composition does not grow fast, and our model can approximate complex diffeomorphisms with sufficient layers. In Fig. 3(a) and Fig. 3(b) we plot the derivatives of `Para-CFlow`, compared to the true one. The non-differentiability of the groundtruth on the unit circle results in irregular behaviors of trained model near the unit circle; elsewhere it shows good consistency.

To further verify our theorem stated in Section 3, specifically the effect of padding zero(s), we test the same model under different hidden dimensions d_{hid} and number of stacking layers in Fig. 3(c).



(a) $\frac{\partial \hat{f}^{(1)}(y,\boldsymbol{x})}{\partial \boldsymbol{x}}$ v.s. $\frac{\partial f^{(1)}(y,\boldsymbol{x})}{\partial \boldsymbol{x}}$ (b) $\frac{\partial \hat{f}(y,\boldsymbol{x})}{\partial y}$ v.s. $\frac{\partial f(y,\boldsymbol{x})}{\partial y}$ (c)

Figure 3: 3(a) and 3(b): Derivatives of the trained model $\hat{f}_y(\boldsymbol{x}) = (\hat{f}^{(1)}(y,\boldsymbol{x}), \hat{f}^{(2)}(y,\boldsymbol{x}))$ with black arrows and the groundtruth $f_y(\boldsymbol{x}) = (f^{(1)}(y,\boldsymbol{x}), f^{(2)}(y,\boldsymbol{x}))$ with red arrows, all at $y = 1$. 3(c): MSE under different hidden dimensions and number of layers.

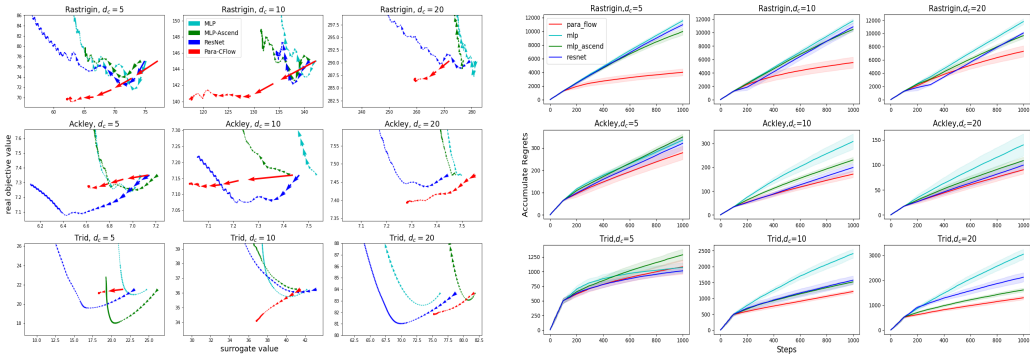
Here hidden dimension equals to the input dimension (2 in this experiment) plus the number of padding zero(s). As we expected, according to our theorems, with padding zero(s) `Para-CFlow` enjoys universality. We also investigate the expressiveness when hidden dimensions = 2. When we increase the number of stacking layers, MSE also decrease, which justifies the L^p universality result in [42]. However, increasing the hidden dimension is much more efficient in terms of improving expressiveness than increasing the number of stacking layers. Such results support our theory of zero-padding. More results of different architecture settings are available in Appx. D.1.

Next we design a more challenging task to compare `Para-CFlow` against a simple `MLP` with a hidden-layer-setting as $(128, 64, 32)$, and Residual Nets (`Resnet`) [21], that concatenates $\boldsymbol{y}$ and $\boldsymbol{x}$ as the network's input and then appends $\boldsymbol{x}$ to the output of each layer forward for prediction. Unlike the previous experiment where a scalar y is used, we generate 3000 samples of a 100-dimensional parameter vector $\boldsymbol{y}_i = (y_i^{(1)}, \cdots, y_i^{(100)})$ where $y_i^{(j)} \sim \mathcal{N}(y_i, 0.16)$ and $y_i \sim \mathcal{U}[0,1]$. Then $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is fed to each model to learn the target $f_{\text{avg}(\boldsymbol{y}_i)}(\boldsymbol{x}_i)$. Fig. 2(b) shows that `Para-CFlow` learn a much more continuous mapping than `MLP` and `Resnet`. While `Resnet` learns relatively better the outer part than `MLP`, similar to `MLP` it suffers from bad mapping in the inner part, i.e. the grey area inside the unit circle is much larger than `Para-Flow`, which represents bad continuity of the learnt mapping.

5.2 Application to contextual BO

In this section,we ensemble `Para-CFlows` as a novel surrogate model for BO with three well-known benchmark functions: `Rastrigin`, `Ackley` and `Trid` [41]. Note that the original benchmark tasks are **contextless**, here we simply set the first d_c dimensions as the context vector sampled from a multivariate uniform distribution and leave the last dimension as the action to optimize. To construct context-dominating tasks we set $d_c \gg 1$. In all benchmarks above, the contexts $\boldsymbol{c}$ are uniformly sampled at random from $[-3, 3]^{d_c}$ sequentially while the action a is optimized by gradient descent in $[-3, 3]$ over surrogate models. The used ensemble-like surrogate models for contextual BO are listed as follows: (1) `Para-CFlow`: Ensemble of `Para-CFlows` for prediction with uncertainty; (2) `MLP`: Ensemble of `MLPs` for uncertainty modelling by concatenating context $\boldsymbol{c}$ and a to obtain input $(\boldsymbol{c}, a) \in \mathbb{R}^{d_c+1}$; (3) `MLP-Ascend`: Similar to `MLP`, the difference is that it firstly raises both $\boldsymbol{c}$ and a to the same hidden dimension $\max(5, d_c)$ and then concatenates them as inputs to a `MLP`; (4) `Resnet`: Ensemble of multiple `Resnet` that concatenates $\boldsymbol{c}$ and a as the network's input and then appends a to the output of each layer forward for prediction. More experimental settings are available in Appx. D.

Firstly, we investigate the action sensitivity of each model trained with 100 random samples. The dimensionality of context is set to 5, 10, 20, which is significantly greater than that of action. The action sensitivity is calculated as the ratio of Permutation Feature Importance (`PFI`) [8] of the action w.r.t. that of the context, i.e. $\texttt{PFI}(a)/\max_i(\texttt{PFI}(c_i))$ and reported in Tab.1. We observe that `Para-CFlow` and `Resnet` have much higher action sensitivities thanks to the model architecture, while `MLP` and `MLP-Ascend` are dominated by contexts as expected. We further investigate whether different surrogates and their gradients manage to guide the real objective function optimization. Denote each trajectory as $\{(b^t, \hat{b}^t, a^t)_{t=1}^T | \mathbf{c}\}$, where $(a^t)_{t=1}^T$ is a T-step gradient trajectory over the



(a) (b)

Figure 4: 4(a): `quiver plot` of $(\hat{b}^t)_{t=1}^T$ (x-axis) v.s. $(b^t)_{t=1}^T$ (y-axis) along each averaged gradient trajectory over 10 random initialization $\times$ 10 random contexts. 4(b):The mean and standard deviation of the cumulative regret under 10 independent trials with the context dimension as 5, 10 and 20.

surrogate $\hat{b}(\boldsymbol{c}, a)$ and $b(\boldsymbol{c}, a)$ is the corresponding real objective value . We visualize the `quiver plot` of $(\hat{b}^t)_{t=1}^T$ (x-axis) v.s. $(b^t)_{t=1}^T$ (y-axis) along each gradient trajectory for all compared surrogates in Fig.4(a). We can easily observe that `Para-CFlow` outperforms other surrogate models as the red arrows move towards the bottom left corner which exactly means the gradient trajectory on `Para-CFlow` indeed minimizes the real objective function, and thus the result supports our main results of C^k approximability, whereas the compared trajectories move towards the top left corner after a few steps probably because those surrogates suffer from more severe local oscillation w.r.t. a, comparing to `Para-CFlow`, resulting in unexpected gradient explosion. We then compare different surrogate models' performances on BO. The mean and standard deviation of the cumulative regrets for different surrogate models are reported in Fig. 4(b) where we observe improvement using `Para-CFlow` as the neural surrogates, especially for the `Rastrigin` benchmark.

Table 1: Comparison of action sensitivity over 10 independent trails

	Rastrigin			**Ackley**			**Trid**		
	$d_c = 5$	$d_c = 10$	$d_c = 20$	$d_c = 5$	$d_c = 10$	$d_c = 20$	$d_c = 5$	$d_c = 10$	$d_c = 20$
Para-CFlow	**2.62±0.71**	**4.30±1.16**	**7.14±1.49**	**3.62±1.21**	**4.47±0.95**	**4.37±1.64**	**1.61±0.44**	1.10±0.51	1.34±0.68
MLP	0.89±0.16	0.70±0.14	0.71±0.13	0.91±0.17	0.79±0.15	0.51±0.16	0.71±0.20	0.64±0.19	0.47±0.20
MLP-Ascend	0.90±0.24	0.80±0.21	0.15±0.16	0.95±0.20	0.88±0.10	0.58±0.14	0.81±0.29	0.64±0.19	0.19±0.15
Resnet	1.65±0.22	1.60±0.22	2.32±0.34	2.03±0.38	1.88±0.25	3.44±1.02	1.29±0.32	**1.77±0.39**	**1.74±0.56**

6 Conclusion

In this paper, we firstly prove the equivalence between the C^k-universality over compactly-supported diffeomorphisms and that over single-coordinate transforms, resulting in the C^k universal approximation ability of affine coupling flows. Furthermore, we generalize the main theorems to parametric cases and propose a practical model called `Para-CFlows` which could serve as a good neural surrogate for gradient-based search and optimization. With good capabilities both in universal approximation and robust sensing of critical features in parametric diffeomorphisms, we empirically exhibit the advantages of `Para-CFlow` using various benchmarks. However, due to the dimension augmentation (for expressiveness purpose), `Para-CFlow` is not able to benefit from fast log-determinant estimation if it is applied to generative model tasks which will be studied in our future work.

References

[1] Robin Allesiardo, Raphaël Féraud, and Djallel Bouneffouf. A neural networks committee for the contextual bandit problem. In *ICONIP*, 2014.

[2] Lynton Ardizzone, Jakob Kruse, Sebastian J. Wirkert, Daniel Rahner, Eric Pellegrini, Ralf S. Klessen, Lena Maier-Hein, Carsten Rother, and U. Köthe. Analyzing inverse problems with invertible neural networks. *ArXiv*, abs/1808.04730, 2019.

[3] Lynton Ardizzone, Carsten Lüth, Jakob Kruse, Carsten Rother, and U. Köthe. Guided image generation with conditional invertible neural networks. *ArXiv*, abs/1907.02392, 2019.

[4] A. Banyaga. *The Structure of Classical Diffeomorphism Groups*. Springer US, 1997.

[5] Jens Behrmann, David Kristjanson Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. In *ICML*, 2019.

[6] Vladimir Bogachev, Alexander V. Kolesnikov, and Kirill Medvedev. Triangular transformations of measures. *Sbornik Mathematics*, 196:309–335, 2004.

[7] W. M. Boothby. An introduction to differentiable manifolds and riemannian geometry. 1975.

[8] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2004.

[9] Haim Brezis. Functional analysis, sobolev spaces and partial differential equations. 2010.

[10] Changyou Chen, Chunyuan Li, Liquan Chen, Wenlin Wang, Yunchen Pu, and Lawrence Carin. Continuous-time flows for efficient inference and density estimation. In *ICML*, 2018.

[11] Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David Kristjanson Duvenaud. Neural ordinary differential equations. In *NeurIPS*, 2018.

[12] Youngseog Chung, Ian Char, Willie Neiswanger, Kirthevasan Kandasamy, Andrew, Oakleigh Nelson, Mark D. Boyer, Egemen Kolemen, and Jeff G. Schneider. Offline contextual bayesian optimization for nuclear fusion. *ArXiv*, abs/2001.01793, 2020.

[13] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. abs/1605.08803, 2016.

[14] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *ArXiv*, abs/1605.08803, 2017.

[15] Emilien Dupont, A. Doucet, and Yee Whye Teh. Augmented neural odes. In *NeurIPS*, 2019.

[16] David G. Ebin and Jerrold E. Marsden. Groups of diffeomorphisms and the motion of an incompressible fluid. *Annals of Mathematics*, 92:102, 1970.

[17] Raphaël Féraud, Robin Allesiardo, Tanguy Urvoy, and Fabrice Clérot. Random forest for the contextual bandit problem. In *AISTATS*, 2016.

[18] Alfred Galichon. Optimal transport methods in economics. 2016.

[19] Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Kristjanson Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *ArXiv*, abs/1810.01367, 2019.

[20] David F. Griffiths, Andrew M. Stuart, and H. C. Yee. Numerical wave propagation in an advection equation with a nonlinear source term. *SIAM Journal on Numerical Analysis*, 29:1244–1260, 1992.

[21] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[22] Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron C. Courville. Neural autoregressive flows. *ArXiv*, abs/1804.00779, 2018.

[23] Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *LION*, 2011.

[24] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks : the official journal of the International Neural Network Society*, 12 3:429–439, 1999.

[25] Pavel Izmailov, P. Kirichenko, Marc Finzi, and Andrew Gordon Wilson. Semi-supervised learning with normalizing flows. *ArXiv*, abs/1912.13025, 2020.

[26] Priyank Jaini, Kira A. Selby, and Yaoliang Yu. Sum-of-squares polynomial flow. *ArXiv*, abs/1905.02325, 2019.

[27] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[28] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, 2018.

[29] Zhifeng Kong and Kamalika Chaudhuri. The expressive power of a class of normalizing flow models. *ArXiv*, abs/2006.00392, 2020.

[30] Andreas Krause and Cheng Soon Ong. Contextual gaussian process bandit optimization. In *NIPS*, 2011.

[31] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Manfred Otto Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *CoRR*, abs/1509.02971, 2016.

[32] Christos Louizos and Max Welling. Multiplicative normalizing flows for variational bayesian neural networks. In *ICML*, 2017.

[33] Mohammad Mozaffari, Walid Saad, Mehdi Bennis, and Mérouane Debbah. Wireless communication using unmanned aerial vehicles (uavs): Optimal transport theory for hover time optimization. *IEEE Transactions on Wireless Communications*, 16:8052–8066, 2017.

[34] George Papamakarios, Eric T. Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.*, 22:57:1–57:64, 2021.

[35] Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. 2019.

[36] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *ICML*, 2015.

[37] Elon D. Rimon and Daniel E. Koditschek. The construction of analytic diffeomorphisms for exact robot navigation on star worlds. *Proceedings, 1989 International Conference on Robotics and Automation*, pages 21–26 vol.1, 1989.

[38] Hadi Salman, Payman Yadollahpour, Tom Fletcher, and Nematollah Batmanghelich. Deep diffeomorphic normalizing flows. *ArXiv*, abs/1810.03256, 2018.

[39] David Silver, Guy Lever, Nicolas Manfred Otto Heess, Thomas Degris, Daan Wierstra, and Martin A. Riedmiller. Deterministic policy gradient algorithms. In *ICML*, 2014.

[40] Miguel Suau, Alexandros Agapitos, David Lynch, Derek Farrell, Mingqi Zhou, and Aleksandar Milenovic. Offline contextual bandits for wireless network optimization. *ArXiv*, abs/2111.08587, 2021.

[41] Sonja Surjanovic and Derek Bingham. Optimization test functions and datasets. http://www.sfu.ca/~ssurjano/optimization.html, 2017.

[42] Takeshi Teshima, Isao Ishikawa, Koichi Tojo, Kenta Oono, Masahiro Ikeda, and Masashi Sugiyama. Coupling-based invertible neural networks are universal diffeomorphism approximators. *ArXiv*, abs/2006.11469, 2020.

[43] Rianne van den Berg, Leonard Hasenclever, Jakub M. Tomczak, and Max Welling. Sylvester normalizing flows for variational inference. *ArXiv*, abs/1803.05649, 2018.

[44] Aäron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis C. Cobo, Florian Stimberg, Norman Casagrande, Dominik Grewe, Seb Noury, Sander Dieleman, Erich Elsen, Nal Kalchbrenner, Heiga Zen, Alex Graves, Helen King, Tom Walters, Dan Belov, and Demis Hassabis. Parallel wavenet: Fast high-fidelity speech synthesis. In *ICML*, 2018.

[45] Cédric Villani. Optimal transport: Old and new. 2008.

[46] Han Zhang, Xi Gao, Jacob Unterman, and Tom Arodz. Approximation capabilities of neural ordinary differential equations. *ArXiv*, abs/1907.12998, 2019.

A Different functional norms and corresponding function spaces

Here we introduce some preliminary definitions on function norms and functions spaces involved in this paper, and the definition for distributional universality.

For a measurable mapping $f : \mathbb{R}^n \to \mathbb{R}^d$ and a subset $K \subseteq \mathbb{R}^n$, we define:

$$\|f\|_{L^p(K)} \triangleq \left(\int_K \|f(\boldsymbol{x})\|^p d\boldsymbol{x}\right)^{1/p}, 1 \leq p < \infty,$$
$$\|f\|_{L^\infty(K)} \triangleq \lim_{p\to\infty} \|f\|_{L^p(K)} \overset{f \text{ cont.}}{=\!=\!=} \sup_{\boldsymbol{x}\in K} \|f(\boldsymbol{x})\|,$$

where $\|\cdot\|$ can be any norm on $\mathbb{R}^d$, as norms on the finite-dimensional vector space are all equivalent. For simplicity, we choose the maximum norm on $\mathbb{R}^d$, i.e., $\|(x_1, x_2, \cdots, x_d)\| = \max_{1\leq i\leq d} |x_i|$.

We can also consider the norm containing derivative information when f is k-th differentiable:

$$\|f\|_{C^k(K)} \triangleq \sum_{|\boldsymbol{\alpha}|\leq k} \|D^{\boldsymbol{\alpha}} f\|_{L^\infty(K)},$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_d) \in \mathbb{N}^d$, $|\boldsymbol{\alpha}| = \sum_{i=1}^d \alpha_i$ and $D^{\boldsymbol{\alpha}} f = \frac{\partial^{|\boldsymbol{\alpha}|} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$.

With these norms, we can define corresponding function spaces:

$$L^p(K) \triangleq \{\text{Domain}(f) = K : \|f\|_{L^p(K)} < \infty\}$$
$$L^\infty(K) \triangleq \{\text{Domain}(f) = K : \|f\|_{L^\infty(K)} < \infty\}$$
$$C^k(K) \triangleq \{\text{Domain}(f) = K : \|f\|_{C^k(K)} < \infty\}$$

when K is compact, $C^k(K) \subseteq L^\infty(K) \subseteq L^p(K)$, and an important fact is that $C^k(K)$ is dense in $L^p(K)$ under L^p norm. Thus the universality over $C^k(K)$ in C^k-norm is stronger than the universality over $L^\infty(K)$ in L^∞ norm, further stronger than the universality over $L^p(K)$ in L^p norm.

Definition A.1. *(Distributional universality). Let $\mathcal{M}$ be a set a set of measurable mappings from $\mathbb{R}^n$ to $\mathbb{R}^d$. We say that $\mathcal{M}$ is a distributional universal approximator if for any absolutely continuous probability measure μ over $\mathbb{R}^n$ w.r.t. Lebesgue measure, and any probability measure ν over $\mathbb{R}^d$, there exists a sequence $\{g_i\}_{i=1}^\infty \subseteq \mathcal{M}$ such that $(g_i)_*\mu$ converges to ν in distribution as $i \to \infty$, where $(g_i)_*\mu(A) \triangleq \mu\left(g_i^{-1}(A)\right)$ for any measurable set A.*

B Proofs

B.1 Proofs for Theorem 3.2

Definition B.1. *Isotopies. An isotopy between two diffeomorphisms $\phi_0, \phi_1 \in$ Diff$_c^k\left(\mathbb{R}^d\right)$ is a C^k-map $H : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ such that the mapping $h_t : \mathbb{R}^d \to \mathbb{R}^d$ defined by $h_t(x) = H(t,x)$ for all $t \in [0,1]$ satisfies $h_0 = \phi_0, h_1 = \phi_1$ and $h_t \in$ Diff$_c^k\left(\mathbb{R}^d\right)$ for all $t \in [0,1]$. It turns out that $t \to h_t$ is a continuous path in the group Diff$_c^k\left(\mathbb{R}^d\right)$ joining ϕ_0 to ϕ_1.*

Proposition B.2. *(Proposition 1.2.1 in [4]) The group Diff$_c^k\left(\mathbb{R}^d\right)$ is connected. Moreover, the group Iso$^k\left(\mathbb{R}^d\right)$ of diffeomorphisms with compact supports which are isotopic to the identity map I through isotopies coincide with Diff$_c^k\left(\mathbb{R}^d\right)$. Here the identity map I means $I(x) = x$ for all $x \in \mathbb{R}^d$.*

Definition B.3. *(δ, k)-near-identity C^k-diffeomorphisms. Let $B_{\delta,k}$ be the C^k-norm ball with radius δ and centered at identity map $I(\boldsymbol{x}) = \boldsymbol{x}$, that is to say, $B_{\delta,k} = \{f \in$ Diff$_c^k\left(\mathbb{R}^d\right) : \sup_{|\alpha|\leq k} \|D^{|\alpha|}(f - I)\|_{L^\infty} < \delta\}$. A diffeomorphism $\phi \in$ Diff$_c^k\left(\mathbb{R}^d\right)$ is said to be (δ, k)-near-identity, if $\phi \in B_{\delta,k}$.*

Lemma B.4. *(Lemma 2.1.8 in [4]) For any diffeomorphism $f \in$ Diff$_c^k\left(\mathbb{R}^d\right)$ and any $\delta > 0$, there exists a finite sequence of (δ, k)-near-identity diffeomorphisms $g_1, \cdots, g_s$ such that $f = g_s \circ g_{s-1} \circ \cdots \circ g_1$.*

Proof. Note that there exists an isotopy h_t from I to f such that $h_0 = I$ and $h_1 = f$. We rewrite $f = h_1 = \left(h_1 \circ h_{(s-1)/s}^{-1}\right) \circ \left(h_{(s-1)/s} \circ h_{(s-2)/s}^{-1}\right) \circ \cdots \circ \left(h_{1/s} \circ h_0^{-1}\right)$ and let $g_i = h_{i/s} \circ h_{(i-1)/s}^{-1}$, we can see that $f = g_s \circ g_{s-1} \circ \cdots \circ g_1$. Take s large enough, we can make $h_{i/s}$ and $h_{(i-1)/s}$ close enough such that $h_{i/s} \circ h_{(i-1)/s}^{-1}$ is (δ, k)-near-identity. □

Theorem B.5. *There exists a $\delta_0 > 0$, such that for any $\delta < \delta_0$ and any $f \in \mathit{Diff}_c^1\left(\mathbb{R}^d\right)$ that is $(\delta, 1)$-near-identity, f can be written as $g \circ h$ with $h\left(\boldsymbol{x}, y\right) = \left(\boldsymbol{x}, \tilde{h}\left(\boldsymbol{x}, y\right)\right)$ and $g\left(\boldsymbol{x}, y\right) = \left(\tilde{g}\left(\boldsymbol{x}, y\right), y\right)$ for $\boldsymbol{x} \in \mathbb{R}^{d-1}, y \in \mathbb{R}$. If $f \in \mathit{Diff}_c^k\left(\mathbb{R}^d\right)$, so are g and h. Further more, g satisfies $\left(\tilde{\delta}, 1\right)$-near-identity for $\tilde{\delta} = \frac{\delta}{1-\delta} > 0$.*

Proof. Let $\pi_i : \mathbb{R}^d \to \mathbb{R}$ denote the projection onto the ith coordinate. Suppose $f : \mathbb{R}^d \to \mathbb{R}^d$ is compactly supported and sufficiently C^k-close to the identity. Then for any point $(\boldsymbol{x}, y) = (x_1, \cdots, x_{d-1}, y)$, the map $f_x : \mathbb{R} \to \mathbb{R}$ given by $f_{\boldsymbol{x}}\left(y\right) = \pi_n f\left(\boldsymbol{x}, y\right)$ is a diffeomorphism: surjectivity follows from the fact that f has compact support, which means $\lim_{y \to \pm\infty} f_{\boldsymbol{x}}\left(y\right) = \pm\infty$ and by the continuity of $f_{\boldsymbol{x}}$; injectivity follows from the fact that if $f_{\boldsymbol{x}}\left(y_1\right) = f_{\boldsymbol{x}}\left(y_2\right)$ for some $y_1 \neq y_2$, then $f_{\boldsymbol{x}}$ must has zero derivatives at some point $y \in (y_1, y_2)$, but the derivative of $f_{\boldsymbol{x}}$ respect to y is near 1, a contradiction.

Now given f, define h and $g : \mathbb{R}^{d-1} \times \mathbb{R} \to \mathbb{R}^{d-1} \times \mathbb{R}$ by

$$\begin{aligned} &h\left(\boldsymbol{x}, y\right) = \left(\boldsymbol{x}, f_{\boldsymbol{x}}\left(y\right)\right), \text{ and} \\ &g\left(\boldsymbol{x}, y\right) = \left(g_1\left(\boldsymbol{x}, y\right), g_2\left(\boldsymbol{x}, y\right), \cdots, g_{d-1}\left(\boldsymbol{x}, y\right), y\right), \end{aligned}$$

where $g_i\left(\boldsymbol{x}, y\right) = \pi_i\left(f\left(\boldsymbol{x}, f_{\boldsymbol{x}}^{-1}\left(y\right)\right)\right) \in \mathbb{R}$. Obviously $g, h \in \mathrm{Diff}_c^k\left(\mathbb{R}^d\right)$ given $f \in \mathrm{Diff}_c^k\left(\mathbb{R}^d\right)$ and $f = g \circ h$. Also we observe that, f is $(\delta, 1)$-near-identity, thus

$$\sup_{\boldsymbol{x}, y} |\frac{\partial}{\partial y} f_{\boldsymbol{x}}\left(y\right) - 1| < \delta, \ \sup_{\boldsymbol{x}, y} |\frac{\partial}{\partial x_i} f_{\boldsymbol{x}}\left(y\right)| < \delta,$$

$$\sup_{\boldsymbol{x}, y} |\frac{\partial}{\partial y} f_{\boldsymbol{x}}^{-1}\left(y\right)| = \sup_{\boldsymbol{x}, y} |\frac{\partial}{\partial y} f_{\boldsymbol{x}}\left(y\right)|^{-1} < \frac{1}{1-\delta}.$$

Then we have

$$\begin{aligned} 0 =& \frac{d}{dx_i} y = \frac{d}{dx_i}\left(f_{\boldsymbol{x}}^{-1}\left(f_{\boldsymbol{x}}\left(y\right)\right)\right) \\ =& (\frac{\partial}{\partial x_i} f_{\boldsymbol{x}}^{-1})\left(f_{\boldsymbol{x}}\left(y\right)\right) + (\frac{\partial}{\partial y} f_{\boldsymbol{x}}^{-1})\left(f_{\boldsymbol{x}}\left(y\right)\right) \cdot \left(\frac{\partial}{\partial x_i} f_{\boldsymbol{x}}\left(y\right)\right), \end{aligned}$$

and further

$$\begin{aligned} |\frac{\partial}{\partial x_i} f_{\boldsymbol{x}}^{-1}\left(y\right)| =& |\left(\frac{\partial}{\partial y} f_{\boldsymbol{x}}^{-1}\right)\left(y\right) \cdot \left(\left(\frac{\partial}{\partial x_i} f_{\boldsymbol{x}}\right)\left(f_{\boldsymbol{x}}^{-1}\left(y\right)\right)\right)| \\ \leq& \sup |\frac{\partial}{\partial y} f_{\boldsymbol{x}}^{-1}| \cdot \sup |\frac{\partial}{\partial x_i} f_{\boldsymbol{x}}| < \frac{\delta}{1-\delta}. \end{aligned}$$

If we denote $f_j\left(\boldsymbol{x}, y\right) = \pi_j f\left(\boldsymbol{x}, y\right)$, we get

$$\begin{aligned} &|\frac{d}{dx_i} g_j\left(\boldsymbol{x}, y\right) - \delta_{i,j}| \\ =& |\tfrac{\partial}{\partial x_i} f_j\left(\boldsymbol{x}, f_{\boldsymbol{x}}^{-1}\left(y\right)\right) - \delta_{i,j} + \tfrac{\partial}{\partial y}\left(f_j\left(\boldsymbol{x}, f_{\boldsymbol{x}}^{-1}\left(y\right)\right)\right) \cdot \tfrac{\partial}{\partial x_i} f_{\boldsymbol{x}}^{-1}\left(y\right)| \\ <& \delta + \delta \cdot \frac{\delta}{1-\delta} = \frac{\delta}{1-\delta}, \end{aligned}$$

which proved that g is $\left(\frac{\delta}{1-\delta}, 1\right)$-near-identity. Here $\delta_{i,j}$ are the kronecker symbols and we notice $|\frac{\partial}{\partial x_i} f_j - \delta_{i,j}| < \delta$. □

Corollary B.6. *There is a $0 < \delta < \frac{1}{d-1}$ such that for any f that is $(\delta, 1)$-near-identity, f can be written as $f_1 \circ f_2 \circ \cdots f_n$, $f_i \in S_c^{1,d}$ with $f_i(\boldsymbol{x}) = \left(x_1, x_2, \cdots, x_{i-1}, \tilde{f}_i(\boldsymbol{x}), x_{i+1}, \cdots, x_d\right)$ for $\boldsymbol{x} = (x_1, \cdots, x_d) \in \mathbb{R}^d$. If f is in $\mathit{Diff}_c^k(\mathbb{R}^d)$ for some $k >= 1$, then so are f_i.*

Proof. By taking δ small enough, we can make $\tilde{\delta} = \delta/(1-\delta)$ small enough, thus the g in Thm. B.5 can be further decomposed. A simple observation shows that if f already preserves some coordinates, then so does g. If we set $\delta_i = \frac{\delta_{i-1}}{1-\delta_{i-1}}$ and $\delta_1 = \delta < \frac{1}{n-1}$, by noticing $\frac{1}{\delta_{n-1}} = \frac{1}{\delta_1} - n + 2 > 1$, we can make δ_{n-1} small enough, thus the decomposition can always be continued until all the coordinates have been decomposed. □

Proof of Thm. 3.2. It can be immediately proved by Lem. B.4 and Cor. B.6 . □

B.2 Definition of $\mathrm{dist}_{\mathrm{Diff}_c^1(\mathbb{R}^d)}$ and proof of Theorem 3.3

Definition B.7. *For any f and $g \in \mathit{Diff}_c^1(\mathbb{R}^d)$, let $\pi(f, g, \epsilon)$ denotes the set of partitions between f and g with maximum jump less than ϵ,*

$$\pi(f, g, \epsilon) := \{(f_1, \cdots, f_n) : f_1 = f, f_n = g, \|f_i \circ f_{i+1}^{-1} - I\|_{C^k} < \epsilon, 1 \le i < n, n \in \mathbb{N}^+\}$$

then we define

$$\mathit{dist}_{\mathit{Diff}_c^1(\mathbb{R}^d)}(f, g) := \limsup_{\epsilon \to 0} \min_{(f_1, \cdots, f_n) \in \pi(f, g, \epsilon)} \sum_{i=1}^{n-1} \|f_i \circ f_{i+1}^{-1} - I\|_{C^k}.$$

Proof of Thm. 3.3 . By Lem. B.4, f can be decomposed into s_1 many $(\frac{1}{d-1}, 1)$-near-identity diffeomorphisms with $s_1 \approx (d-1)\ell$; by Cor. B.6, each $(\frac{1}{d-1}, 1)$-near-identity diffeomorphism can be decomposed into at most d single-coordinate transforms, thus $s \le s_1 \cdot d \approx d(d-1)\ell$. The evaluation for ℓ is difficult, but a lower bound is straightfoward: $\ell \ge \|f - I\|_{C^1}$. When f is not far from I, the lower bound is adequate. More details refer to [4]. □

B.3 Proof for Theorem 3.4

Proof. Take any positive number $1 > \tilde{\epsilon} > 0$ and compact set $K \in \mathbb{R}^d$. Put $r \triangleq \max_{\boldsymbol{x} \in K} \|f_1(\boldsymbol{x})\|$ and $K' \triangleq \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| \le r + 1\}$. Let $g_2 \in G$ satisfying

$$\|f_2 - g_2\|_{C^k(K')} < \tilde{\epsilon}.$$

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

$$\sup_{|\alpha| \le k} \|D^\alpha f_2(\boldsymbol{x}) - D^\alpha f_2(\boldsymbol{y})\| < \tilde{\epsilon}.$$

From the assumption, we can take $g_1 \in G$ satisfying

$$\|f_1 - g_1\|_{C^k(K)} < \min\{1, \delta\}.$$

Then it is clear that $f_1(K) \subseteq K'$ by the definition of K'. Moreover, we have $g_1(K) \subseteq K'$. In fact, we have

$$\|g_1(\boldsymbol{y})\| \le \sup_{\boldsymbol{x} \in K} \|f_1(\boldsymbol{x}) - g_1(\boldsymbol{x})\| + |f_1(\boldsymbol{y})| \le 1 + r$$

for any $\boldsymbol{y} \in K'$.

Then for any $\boldsymbol{x} \in K$, we have

$$\begin{aligned} &\|f_2 \circ f_1 - g_2 \circ g_1\| \\ \le &\|f_2 \circ f_1 - f_2 \circ g_1\| + \|f_2 \circ g_1 - g_2 \circ g_1\| < 2\tilde{\epsilon}. \end{aligned}$$

Now let us consider the cumulative error for derivatives. We have

$$\begin{aligned}
&\|D\left(f_2 \circ f_1\right)-D\left(g_2 \circ g_1\right)\| \\
\leq & \|D\left(f_2 \circ f_1\right)-D\left(f_2 \circ g_1\right)\|+\|D\left(f_2 \circ g_1\right)-D\left(g_2 \circ g_1\right)\| \\
= & \|\left(D f_2\right) \circ f_1 \cdot D f_1-\left(D f_2\right) \circ g_1 \cdot D g_1\|+\|\left(D f_2\right) \circ g_1 \cdot D g_1-\left(D g_2\right) \circ g_1 \cdot D g_1\| \\
\leq & \|\left(D f_2\right) \circ f_1 \cdot D\left(f_1-g_1\right)\|+\|\left(D\left(f_2-g_2\right)\right) \circ g_1 \cdot D g_1\|+\|\left(\left(D f_2\right) \circ f_1-\left(D f_2\right) \circ g_1\right) \cdot D g_1\| \\
< & \left(\|D f_2\|+2\|D g_1\|\right) \tilde{\epsilon} \\
< & C\left(f_1, f_2\right) \tilde{\epsilon}
\end{aligned}$$

by noticing that

$$\|D g_1(\boldsymbol{x})\| \leq\|D f_1(\boldsymbol{x})\|+\tilde{\epsilon} \leq\|D f_1(\boldsymbol{x})\|+1 .$$

Higher order derivatives can be estimated following the same procedure with more complex computations and reusing of triangular inequality. We can finally arrive at

$$\|f_2 \circ f_1(\boldsymbol{x})-g_2 \circ g_1(\boldsymbol{x})\|_{C^k(K)}<\tilde{C}\left(f_1, f_2\right) \tilde{\epsilon}$$

with $\tilde{C}\left(f_1, f_2\right)$ only depends on f_1 and f_2 and their derivatives, doesn't depend on $\tilde{\epsilon}$ because f_1, f_2 are compactly supported, which means they have finite high order derivatives over $\mathbb{R}^d$.

Thus we take $\tilde{\epsilon}=\frac{\epsilon}{\tilde{C}\left(f_1, f_2\right)}$, then $\|f_2 \circ f_1(\boldsymbol{x})-g_2 \circ g_1(\boldsymbol{x})\|_{C^k(K)} \leq \epsilon$, and then finished our proof. □

B.4 Proof for Theorem 3.6

Proof. Note that

$$\iota_d \circ \tau \circ \pi_d\left(x_1, x_2, \cdots, x_d, 0\right)=\left(x_1, x_2, \cdots, \tau_d(\boldsymbol{x}), 0\right) .$$

This can be decomposed into three small steps:

$$\left(x_1, x_2, \cdots, x_d, 0\right) \stackrel{\phi_1}{\rightarrow}\left(x_1, x_2, \cdots, x_d, \tau_d(\boldsymbol{x})\right) \stackrel{\phi_2}{\rightarrow} \quad\left(x_1, x_2, \cdots, \tau_d(\boldsymbol{x}), x_d\right) \stackrel{\phi_3}{\rightarrow}\left(x_1, x_2, \cdots, \tau_d(\boldsymbol{x}), 0\right) .$$

Next let us approximate ϕ_1, ϕ_2, ϕ_3 using the elements in $\mathcal{G}$-INN$_{d+1}$. By definition, ϕ_1 can be written as $\Phi_{d+1, d, \sigma, t}$ with σ to be any function, $t(\boldsymbol{x})=\tau_d(\boldsymbol{x})$. By assumption, $\mathcal{H}$ has C^k-universality for t, thus we know $\mathcal{G}$-INN$_{d+1}$ has universality for ϕ_1. ϕ_2 is just a permutation which is already in our layers. ϕ_3 can be written as $\Phi_{d+1, d, \sigma, t}$ with $\sigma=0$, $t(\boldsymbol{x})=-\tau_d^{-1}(\boldsymbol{x})$. Here $\tau_d^{-1}(\boldsymbol{x})$ is the inverse of $\tau_d(\boldsymbol{x})$ w.r.t. x_d because $\tau_d(\boldsymbol{x})$ is a monotonic function w.r.t. x_d. Thus, we claim that $\mathcal{G}$-INN$_{d+1}$ has universality for ϕ_3. By Thm. 3.4, we know that $\iota_d \circ \tau \circ \pi_d$ can be arbitrarily approximated by $\mathcal{G}$-INN$_{d+1}$.

Thus, for any $\epsilon>0$, there exists a $\tilde{\tau} \in \mathcal{G}$-INN$_{d+1}$ such that

$$\|\iota_d \circ \tau \circ \pi_d-\tilde{\tau}\|_{C^k(K \times \mathbb{R})}<\epsilon,$$

and furthermore,

$$\begin{aligned}
&\|\tau-\pi_d \circ \tilde{\tau} \circ \iota_d\|_{C^k(K)} \\
= & \|\left(\pi_d \circ \iota_d\right) \circ \tau \circ\left(\pi_d \circ \iota_d\right)-\pi_d \circ \tilde{\tau} \circ \iota_d\|_{C^k(K)} \\
\leq & \|\pi_d\| \cdot\|\iota_d \circ \tau \circ \pi_d-\tilde{\tau}\|_{C^k(K \times \mathbb{R})} \cdot\|\iota_d\|<\epsilon .
\end{aligned}$$

□

B.5 Proofs for Theorem 3.7

Definition B.8. *Isotopies along Diff$_c^{k,m,d}$. An isotopy between two diffeomorphisms $\phi_0, \phi_1 \in$ Diff$_c^{k,m,d}$ is a C^k-map $H:[0,1] \times \mathbb{R}^{m+d} \rightarrow \mathbb{R}^{m+d}$ such that the mapping $h_t: \mathbb{R}^d \rightarrow \mathbb{R}^d$ defined by $h_t(x)=H(t, x)$ for all $t \in[0,1]$, $h_0=\phi_0, h_1=\phi_1$ and $h_t \in$ Diff$_c^{k,m,d}$ for all $t \in[0,1]$. It turns out that $t \rightarrow h_t$ is a continuous path in the group Diff$_c^{k,m,d}$ joining ϕ_0 to ϕ_1.*

Theorem B.9. *The group Diff$_c^{k,m,d}$ is connected. Moreover, the group of diffeomorphisms with compact supports which are isotopic to the identity map I through compactly supported isotopies coincide with Diff$_c^{k,m,d}$. Here the identity map I means $I(x)=x$ for all $x \in \mathbb{R}^{m+d}$.*

Proof. For arbitrary $f \in \text{Diff}_c^{k,m,d}$, we have a compactly supported C^k-isotopy

$$H : \mathbb{R}^{m+d} \times [\epsilon, 1] \to \mathbb{R}^{m+d},$$

given by the proportionately contraction

$$H\left(\left(\boldsymbol{y}, \boldsymbol{x}\right), t\right) \triangleq tf\left(\boldsymbol{y}/t, \boldsymbol{x}/t\right), \boldsymbol{y} \in \mathbb{R}^m, \boldsymbol{x} \in \mathbb{R}^d.$$

It is obvious that $H\left(\left(\boldsymbol{y}, \boldsymbol{x}\right), t\right)$ always lie in $\text{Diff}_c^{k,m,d}$, the first m-coordinate corresponding to y are fixed. By choosing ϵ small enough, we can achieve $h_\epsilon\left(\boldsymbol{y}, \boldsymbol{x}\right) = H\left(\left(\boldsymbol{y}, \boldsymbol{x}\right), \epsilon\right) \in \left(\text{Diff}_c^{0,m,d}\right)_0$, the C^0-connected neighborhood of identity I.

Thus, there is a prolongation of $H : \mathbb{R}^{m+d} \times [0, 1] \to \mathbb{R}^{m+d}$ that is a compactly supported C^0-isotopy from I to f with $H_t = I,$ for all $t \in [0, \epsilon/2]$. Since compactly supported C^k-isotopy space with first m-coordinate fixed is dense in compactly supported C^0-isotopy space with first m-coordinate fixed, and since H is already C^k on $\mathbb{R}^{m+d} \times ([0, \epsilon/2] \cup [\epsilon, 1])$, we can always find a compactly supported C^k-isotopy $\tilde{H}$ from I to f, and thus f is isotopic to the identity I in $\text{Diff}_c^{k,m,d}$. □

Follow the same proof in Lem. B.4, we can get the following lemma.

Theorem B.10. *For any diffeomorphism $f \in \text{Diff}_c^{k,m,d}$ and any $\delta > 0$ of the identity, there exists a finite sequence of (δ, k)-near-identity diffeomorphisms $g_1, \cdots, g_s \in \text{Diff}_c^{k,m,d}$ such that $f = g_s \circ g_{s-1} \circ \cdots \circ g_1$.*

We remark that Thm. B.5 and Cor. B.6 need no modification and can be directly applied to here.

Proof. of Thm. 3.7.

It is immediately proved by Lem. B.10 and Cor. B.6 by replac int d in Cor. B.6 with $m + d$. □

B.6 Proofs for Theorem 4.2

Proof. By the famous Constant Rank Theorem [7], there exists a parametric ($\boldsymbol{c}$ plays the role of parameter) diffeomorphism $f_{\boldsymbol{c}}(\cdot) : \mathbb{R}^{d_\phi} \to \mathbb{R}^{d_\phi}$ such that

$$f_{\boldsymbol{c}}(\boldsymbol{a}, \mathbf{0}_{1\times(d_\phi - d_a)}) = \Phi(\boldsymbol{c}, \boldsymbol{a}), \text{ for all } (\boldsymbol{c}, \boldsymbol{a}) \in \mathcal{C} \times \mathcal{A}.$$

Then we can approximate $f_{\boldsymbol{c}}$ by Thm. 3.6, 3.8, given that $d' = \max\{d_a, [\frac{d}{2}]\} \geq d_\phi$. □

C Learning invertible map under dimension augmentation

In this section, we analysis how to make the affine coupling flow with dimension-augmentation invertible. For simplicity we only consider the non-parametric case. The input is $\boldsymbol{x} \in \mathbb{R}^d$, and after dimension-augmentation $\boldsymbol{x} \to (\boldsymbol{x}, \mathbf{0})$ and some invertible map layers f, output is $F(\boldsymbol{x}) = f(\boldsymbol{x}, \mathbf{0}) \in \mathbb{R}^{d+r}$, where r is the augmented dimension. Note that and F is not a surjective map to $\mathbb{R}^{d+r}$, and $\text{Range}(F) \in \mathbb{R}^{d+r}$ is a manifold of dimension $\mathbb{R}^d$. Given any $\boldsymbol{y} \in \mathbb{R}^{d+r}$, it's not always true that we can find $\boldsymbol{x} \in \mathbb{R}^d$ such that $F(\boldsymbol{x}) = \boldsymbol{y}$ unless $\boldsymbol{y} \in \text{Range}(F)$.

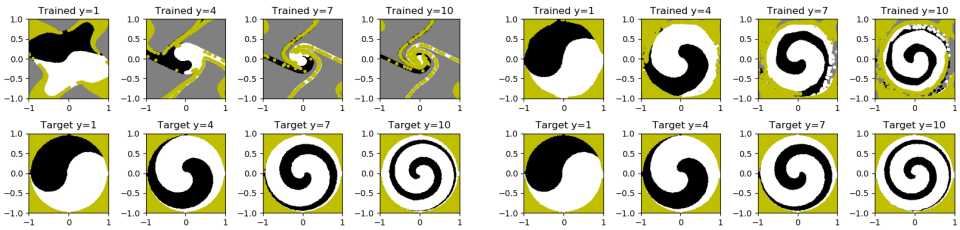
To handle this problem, we need to make sure that $\text{Range}(F)$ is tractable for easy sampling. The easiest way is to make $\text{Range}(F) \approx \mathbb{R}^d \times \{0\}^r$. Note that in this way, for any $\boldsymbol{y} \in \mathbb{R}^d$, we can always find $\boldsymbol{x} \in \mathbb{R}^d$ such that $f(\boldsymbol{x}, \mathbf{0}) = f^{-1} = (\boldsymbol{y}, \mathbf{0})$, and note that f is an invertible map, which means $(\boldsymbol{x}, \mathbf{0}) = f^{-1}(\boldsymbol{y}, \mathbf{0})$. Such a f, by Thm. 3.6, Cor. 3.5 and Thm. 3.3, can be approximated arbitrary well in C^k norm by a fixed-layer-number affine-coupling flow $\hat{f}$.

To summarize, if we want to learn an invertible map from $\boldsymbol{x} \in \mathbb{R}^d$ to $\boldsymbol{y} \in \mathbb{R}^d$, we first augment $\boldsymbol{x}$ and $\boldsymbol{y}$ to $(\boldsymbol{x}, 0) \in \mathbb{R}^{d+r}$ and $(\boldsymbol{y}, 0) \in \mathbb{R}^{d+r}$; then apply affine-coupling flow to learn the map $\hat{f}$ from $(\boldsymbol{x}, 0)$ to $(\boldsymbol{y}, 0)$. After that, we apply $\hat{f}^{-1}$ to $(\boldsymbol{y}, 0)$, and get the initial-domain samples.

D Experiments

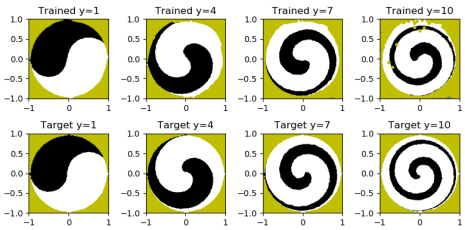
D.1 Taiji

In this section we show the learning results of `Para-CFlows` on the Taiji tasks. The experimental setting is stated in Section 5.1. We compare the effects of hidden dimensions d_{hid} and number of layers N_{layer}. As we see, when $d_{hid} = 2, N_{layer} = 6$, the learned result is far from correct; when $d_{hid} = 3, N_{layer} = 6$ or $d_{hid} = 2, N_{layer} = 9$, the results are imperfect; Other situations are almost the same. We can also refer to Fig. 3(c) for the MSE under different setting, where we can see when $d_{hid} = 5, N_{layer} = 6$, it can do better than $d_{hid} = 5, N_{layer} = 15$. Thus we believe that raising to higher dimensions can do better than stacking more layers under similar extra parameter size.

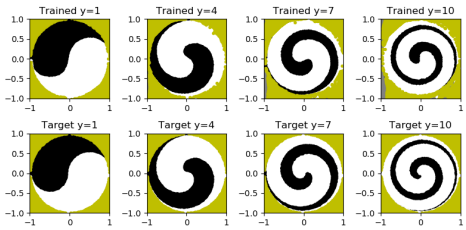


(a) $d_{hid} = 2, N_{layer} = 6$

(b) $d_{hid} = 3, N_{layer} = 6$



(c) $d_{hid} = 5, N_{layer} = 6$



(d) $d_{hid} = 8, N_{layer} = 6$

Figure 5: Different training results under different hidden dimensions

D.2 BO test

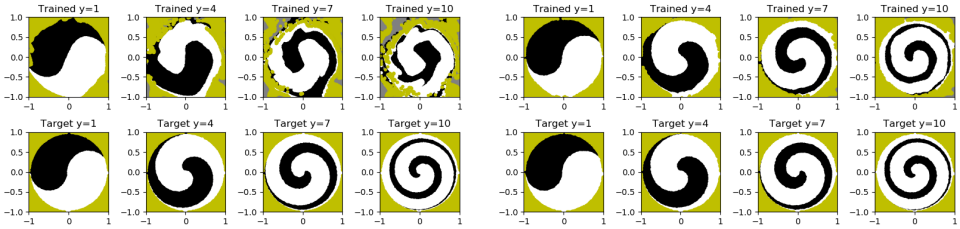
D.2.1 Experimental settings

In our experiments, we use 3 well-known benchmark functions: `Ackley`, `Trid` and `Rastrigin` [41]. It is noted that the original benchmark tasks are **contextless**. Here, we simply set the first d_c dimensions as the context vector and leave the last one dimension as the action to be optimized. For our constructed optimization problems, `Ackley` has the action dimension that is coupled with all context dimensions, `Trid`'s action is coupled with just one context dimension, and all the dimensions of `Rastrigin` are independent with each other. To construct context-dominating tasks where the reward depends much more on the context than the action, we set $d_c \gg 1$.

In the training of all the surrogate models, we consistently set the number of batch size to 64, number of the epochs to 200, and the learning rate to 0.01. For cases when the dimensionality of context is 5, 10 and 20, we implement each of the aforementioned surrogate models in Sec. 5.2 with similar sizes (number of trainable parameters) and the corresponding specifications are posted as in Tab.2, Tab. 3, and Tab. 4, respectively.

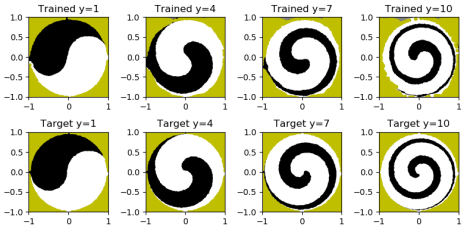
D.2.2 KT

To further verify the sensitiveness, we train all models with uniformly randomly sampled data (both context and action) of various sizes and then test the trained models on 10,000 testing contexts, each with the action space sweeped. We calculate the Kendall's Tau (KT) score [27] between $f_{\boldsymbol{c}}(a)$ and



(a) $d_{hid} = 2, N_{layer} = 9$

(b) $d_{hid} = 3, N_{layer} = 12$



(c) $d_{hid} = 5, N_{layer} = 15$

Figure 6: Different training results under different number of stacking layers

Table 2: When dimensionality of context is 5: #hidden-layers and #hidden-nodes represent the number of hidden layers and nodes for state network in `Para_CFlow` and base network in other models, respectively. #flows is the number of modules in a flow-based model. #param is the total number of trainable parameters in a neural surrogate model.

Method	#hidden-layers	#hidden-nodes	#flows	#parameters
Para-CFlow	1	64	3	1428
MLP	2	32	0	1313
MLP-Ascend	2	32	0	1481
Resnet	2	32	0	1346

Table 3: When dimensionality of context is 10: The meaning of the header is the same as in Tab. 2.

Method	#hidden-layers	#hidden-nodes	#flows	#parameters
Para-CFlow	0	128	3	5337
MLP	2	64	0	4993
MLP-Ascend	2	64	0	5669
Resnet	2	64	0	5058

Table 4: When dimensionality of context is 20: The meaning of the header is the same as in Tab. 2.

Method	#hidden-layers	#hidden-nodes	#flows	#parameters
Para-CFlow	1	64	3	12723
MLP	3	64	0	9793
MLP-Ascend	3	64	0	11469
Resnet	3	64	0	9922

$\hat{f}_c(a)$, which measures the action order consistency between the groundtruth and the prediction. As shown in Fig. 7, `Para-CFlow` exhibits higher KT scores than the other models under higher dimensional context, demonstrating that it indeed preserves critical information of the action.

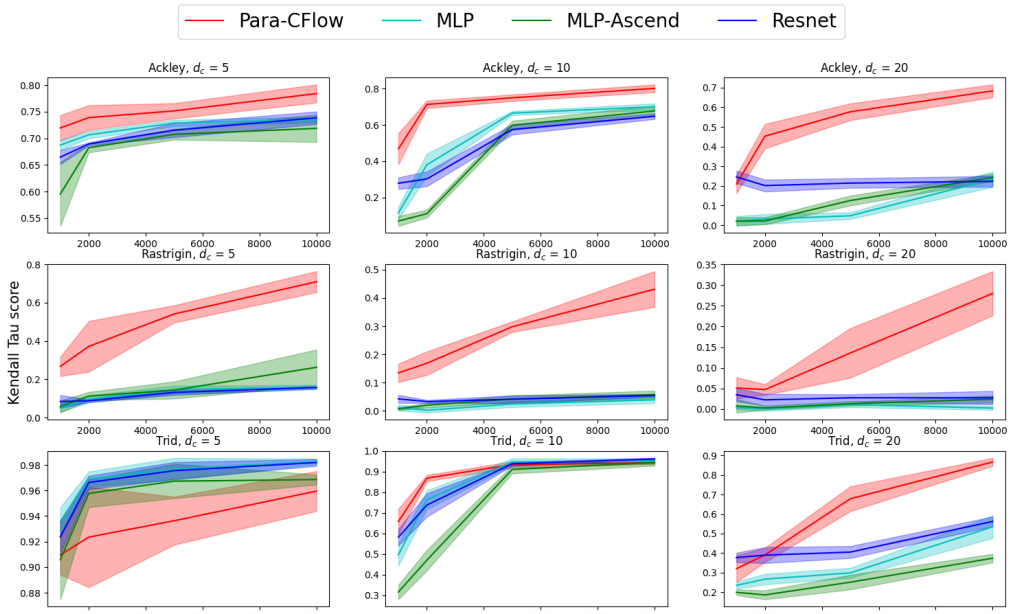


Figure 7: KT scores calculated from 5 independent trials with the context dimension as 5, 10 and 20.

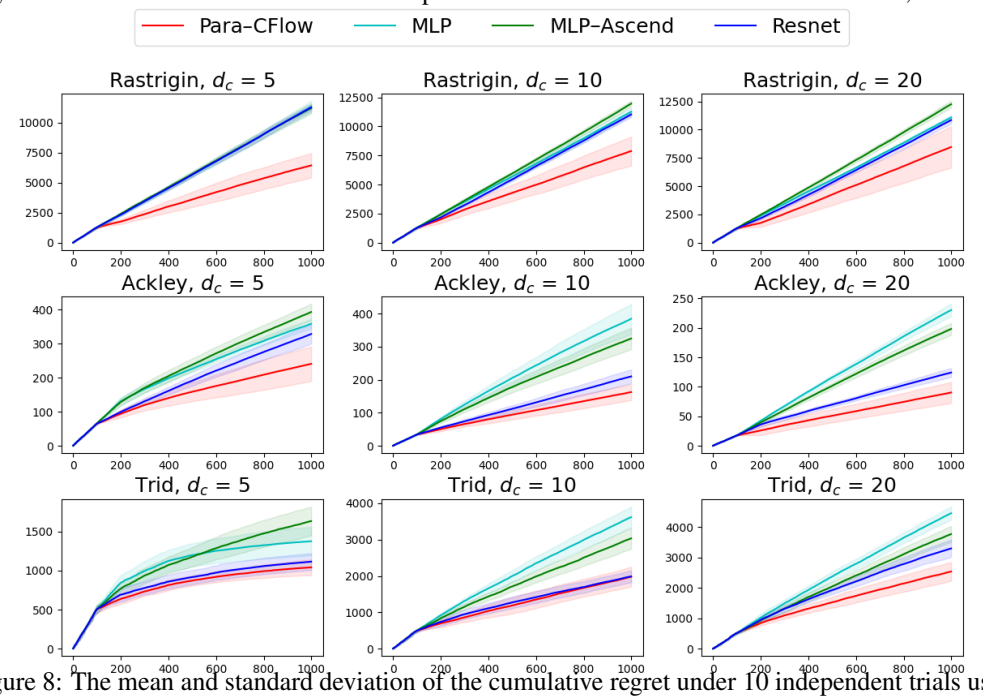


Figure 8: The mean and standard deviation of the cumulative regret under 10 independent trials using *discrete search* with different neural surrogate models for context of dimensionality 5, 10, and 20 on `Rastrigin`, `Ackley` and `Trid`, respectively.

D.2.3 Discrete search

The cumulative regrets of BO using grid search (That is, we traversal all a in $\texttt{linespace}(-3, 3, 100)$ in each surrogate-prediction state) with different neural surrogate models for context of dimensionality 5, 10 and 20 are shown in Fig. 8. Comparing this result with 4(b), we can see that `Para-CFlow` has lower cumulative regrets when using Gradient Decent than grid search, but the results for other three models do not support this. Besides, even in grid search, `Para-CFlow` still does the best among different models, which means it preserves better action sensitivities.