The Expressive Power of a Class of Normalizing Flow Models

Zhifeng Kong
z4kong@eng.ucsd.edu
University of California San Diego

Kamalika Chaudhuri
kamalika@cs.ucsd.edu
University of California San Diego

Abstract

Normalizing flows have received a great deal of recent attention as they allow flexible generative modeling as well as easy likelihood computation. While a wide variety of flow models have been proposed, there is little formal understanding of the representation power of these models. In this work, we study some basic normalizing flows and rigorously establish bounds on their expressive power. Our results indicate that while these flows are highly expressive in one dimension, in higher dimensions their representation power may be limited, especially when the flows have moderate depth.

1 Introduction

Normalizing flows are a class of deep generative models that aspire to learn an invertible transformation to convert a pre-specified distribution, such as a Gaussian, to the distribution of the input data. These models offer flexible generative modeling – as the invertible transformation can be implemented by deep neural networks – and easy likelihood computation in equation (3) that follows from the invertibility of the transformation [Rezende and Mohamed, 2015].

Due to these advantages and their empirical success, a number of flow models have been proposed [Dinh et al., 2014, Germain et al., 2015, Uria et al., 2016, Kingma et al., 2016, Tomczak and Welling, 2016, Dinh et al., 2016, Papamakarios et al., 2017, Huang et al., 2018, Berg et al., 2018, Grathwohl et al., 2018, Behrmann et al., 2018, Jaini et al., 2019, Ho et al., 2019]. However, the expressive power offered by different kinds of flow models – what kind of distributions they can map between, and with what complexity – remains not well-understood, which makes it challenging to select the right flow model for specific tasks. Obviously, due to their invertible nature, a normalizing flow can only transform a distribution to one with a homeomorphic support [Armstrong, 2013]. However, even within such distributions, it remains unclear whether a simple distribution supported on $\mathbb{R}^d$ could be transformed or approximated via a normalizing flow from a Gaussian.

In this work, we carry out a rigorous analysis of the expressive power of planar flows, Sylvester flows, and Householder flows – the most basic classes of normalizing flows. The main challenge in analyzing the expressive power of any flow model class is *invertibility*. There is a body of prior work that analyzes the universal approximation properties of standard neural networks; however, analyzing the approximation properties of *invertible* mappings between distributions is a completely different problem. Just because a function class $\mathcal{F}$ is a universal approximator does not mean that the set of all its invertible functions can transform between arbitrary distributions; dually, even if functions in $\mathcal{F}$ have limited expressivity, it is possible that its invertible subset is an universal approximator in transforming between distributions [Villani, 2008]. Additionally, universal approximation properties are often proved by construction via non-invertible functions [Lu et al., 2017, Lin and Jegelka, 2018] and hence these constructions cannot to be used to establish properties of the corresponding flows.

This work gets around this challenge by studying properties of input-output distribution pairs directly, instead of considering the transformation class itself. In particular, we consider both a local and global analysis of properties of planar flows, their higher dimensional generalization – Sylvester flows, and Householder flows. First, we analyze the local topology – namely, the directional derivatives of the induced density. Second, we seek to bound the global total variation distance between the input and output distributions that can be achieved by each planar flow or Householder flow under certain conditions.

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

Using these two kinds of analysis, we make three main contributions in this paper.

First, we show that in one dimension, even planar flows are highly expressive. In particular, they can transform a source distribution supported on $\mathbb{R}$ to an arbitrarily-accurate approximation of any target distribution supported on a finite union of intervals. The conclusion holds even if we restrict to planar flows with ReLU non-linearity and Gaussian source distributions. This indicates that planar flows in one dimension are universal approximators.

We next turn our attention to general d-dimensional spaces, and we look at what kinds of distributions may be expressed by a Sylvester flow model acting on a Gaussian, mixtures of Gaussian (MoG) distributions, or product (Prod) distributions. We show that when the non-linearity is a ReLU function, Sylvester flows of any depth cannot in general exactly transform between certain standard classes of distributions. In particular, ReLU Sylvester flows cannot exactly transform any mixture of k Gaussian distributions or product distributions into another one – no matter what the depth is – except under very special circumstances.

Finally, we consider the approximation capability of normalizing flow models in d-dimensional space. Here, we focus on local planar flows with a class of local non-linearities – including common non-linearities such as tanh, arctan and sigmoid – and Householder flows. We show that in these cases, provided certain conditions hold, transforming a source distribution into a target may require flows of inordinately large depth. In particular, if the target distribution $p(z)$ is constant in a ball centered at the origin and proportional to $\exp(-\|x\|_2^{1/\tau})$ outside the ball, then p may require local planar flows with depth $\Omega\left(d^{1/\tau-1}\right)$ to transform from an arbitrary source distribution (that is not too close). A similar conclusion holds for Householder flows when the target distribution is close to the standard Gaussian distribution. These results indicate that when local planar flows with certain non-linearities and Householder flows have moderate depth, they may have poor approximation power.

1.1 Related Work

There is a body of work on analyzing the approximation properties of neural networks [Cybenko, 1989, Hornik et al., 1989, Hornik, 1991, Montufar et al., 2014, Telgarsky, 2015, Lu et al., 2017, Hanin, 2017, Raghu et al., 2017]. Most of these results apply to feed-forward neural networks including non-invertible functions. Therefore, their universal approximation properties do not directly translate to normalizing flows.

The work most related to ours shows that a residual network (ResNet) in which each block is a single-neuron hidden layer with ReLU activation is a universal approximator in the space of Lebesgue integrable functions from $\mathbb{R}^d$ to $\mathbb{R}^d$ [Lin and Jegelka, 2018]. This is related to us because the set of all such ResNets with T invertible blocks is exactly T-layer ReLU planar flows. However, their construction that establishes this property is based on non-invertible mappings, consequently, their universal approximation result does not extend to planar flows.

There has also been some recent related work on the expressive power of generative networks. In particular, it was proved by construction that when the output dimension is equal to the input dimension, deep neural networks can approximately transform Gaussians to uniform distributions and vice versa [Bailey and Telgarsky, 2018]. However, their constructions are again based on non-invertible functions, and hence their results do not extend to normalizing flows.

Finally, there is also a body of empirical work on different kinds of normalizing flows; a more detailed discussion of these works is presented in Section 6.

2 Preliminaries

2.1 Definitions and Notation

Suppose d is the data dimension. Let $z \in \mathbb{R}^d$ be a random variable with density $q_z : \mathbb{R}^d \to \{0\} \cup \mathbb{R}^+$. Then, an invertible function $f : \mathbb{R}^d \to \mathbb{R}^d$ is called a *normalizing flow* if f is differentiable almost everywhere (*a.e.*) and the determinant of the Jacobian matrix of f does not equal to zero:

$$\det J_f(z) \neq 0 \ (a.e.)$$

where $J_f(z)_{ij} = \frac{\partial f_i}{\partial z_j}$, $\forall i, j \in \{1, \cdots, d\}$. If we apply a flow f over z, we obtain a new random variable $y = f(z)$, whose density q_y can be written through the change-of-variable formula:

$$q_y(y) = \frac{q_z(z)}{|\det J_f(z)|} \tag{1}$$

or

$$\log q_y(y) = \log q_z(z) - \log |\det J_f(z)| \tag{2}$$

For conciseness, we write $q_y = f \# q_z$ in such context. In particular, if the flow f is composed of T *simple* flows $f_t, t = 1 \cdots, T$:

$$f = f_T \circ f_{T-1} \circ \cdots \circ f_1$$

then according to the chain rule of the Jacobian matrix, we have

$$\log q_y(y) = \log q_z(z) - \sum_{t=1}^{T} \log |\det J_{f_t}(z_{t-1})| \tag{3}$$

where $z_0 = z,\ z_t = f_t(z_{t-1}),\ t = 1, \cdots, T$.

Two simple flows are defined below [Rezende and Mohamed, 2015]:

Planar Flows. Given the scaling vector $u \in \mathbb{R}^d$, tangent vector $w \in \mathbb{R}^d$, shift $b \in \mathbb{R}$, and non-linearity $h : \mathbb{R} \to \mathbb{R}$, a planar flow f_{pf} on $\mathbb{R}^d$ is defined by

$$f_{\text{pf}}(z) = z + uh(w^\top z + b) \tag{4}$$

Radial Flows. Given the smoothing factor $a \in \mathbb{R}^+$, scaling factor $b \in \mathbb{R}$, and center $z_0 \in \mathbb{R}^d$, a radial flow f_{rf} on $\mathbb{R}^d$ is defined by

$$f_{\text{rf}}(z) = z + \frac{b}{a + \|z - z_0\|_2}(z - z_0) \tag{5}$$

A geometric intuition between planar and radial flows is shown in Section A.1. Planar flows can be generalized to a higher dimension below [Berg et al., 2018]:

Sylvester Flows. Given the flow dimension $m < d$, scaling matrix $A \in \mathbb{R}^{d\times m}$, tangent matrix $B \in \mathbb{R}^{d\times m}$, shift vector $b \in \mathbb{R}^d$, and non-linearity $h : \mathbb{R} \to \mathbb{R}$, a Sylvester flow f_{syl} on $\mathbb{R}^d$ is defined by

$$f_{\text{syl}}(z) = z + Ah(B^\top z + b) \tag{6}$$

where h maps coordinate-wise.

In addition, Householder matrices can also be used to construct flows [Tomczak and Welling, 2016]:

Householder Flows. Given a unit reflection vector $v \in \mathbb{R}^d$, a Householder flow f_{hh} on $\mathbb{R}^d$ is defined by

$$f_{\text{hh}}(z) = z - 2vv^\top z \tag{7}$$

For conciseness, we denote these flows by *base* flows.

2.2 Problem Statement

In this paper, we study the expressivity of base flows in Section 2.1: given an input distribution q, we hope to understand when a flow f composed of a finite number of base flows can transform q into any target distribution p or its approximation on $\mathbb{R}^d$. Formally, suppose f is composed of T base flows in the same class. We propose to answer the following two questions:

Q1 (Exact transformation): Under what conditions is it possible to *exactly* transform q into p with a finite number of base flows? That is, $f\#q = p,\ (a.e.)$.

Q2 (Approximation): Since sometimes it may not be possible to exactly transform q into p, when is it possible to *approximate* p in total variation distance (which is equal to half of the ℓ_1 distance)? How many layers of base flows do we need? That is, given $\epsilon > 0$, is there a bound for T such that

$$\|f\#q - p\|_1 \le \epsilon$$

2.3 Additional Definitions and Notations

The determinant of the Jacobian matrix of a planar flow f_{pf}, a Sylvester flow f_{syl}, and a Householder flow f_{hh} can be easily calculated by

$$\begin{aligned} \det J_{f_{\text{pf}}}(z) &= 1 + u^\top w h'(w^\top z + b) \\ \det J_{f_{\text{syl}}}(z) &= \det(I_m + diag(h'(B^\top z + b))B^\top A) \\ \det J_{f_{\text{hh}}}(z) &= -1 \end{aligned} \tag{8}$$

In this paper, we consider three types of non-linearities h: $\text{ReLU}(x) = \max(x, 0)$, general differentiable functions, and local non-linearities (see Section 5 for detail) including $\tanh(x)$, $\arctan(x)$ and $\text{sigmoid}(x) = 1/(1 + \exp(-x))$. Specifically, let $h = \text{ReLU}$ and $1\{\cdot\}$ be the indicator function, then $\det J_{f_{\text{pf}}}$ is equal to

$$\det J_{f_{\text{pf}}}(z) = 1 + u^\top w \cdot 1\{w^\top z + b \ge 0\} \tag{9}$$

A ReLU planar/Sylvester flow is invertible under certain bounds on its parameters as ReLU is Lipschitz.

We make a few additional definitions here. $\mathcal{N}$ denotes a Gaussian distribution on $\mathbb{R}^d$:

$$\mathcal{N}(x; \mu, \Sigma) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)}{(2\pi)^{d/2}\sqrt{\det \Sigma}}$$

The set **supp** p denotes the support of distribution p:

$$\textbf{supp}\ p = \{x \in \mathbb{R}^d : p(x) > 0\}$$

For vectors $w_i \in \mathbb{R}^d, 1 \le i \le k$, the **span** of them denotes the subspace spanned by $\{w_i\}_{i=1}^k$:

$$\textbf{span}\{w_1, \cdots, w_k\} = \left\{ \sum_{i=1}^k \alpha_i w_i :\ \alpha_i \in \mathbb{R}, 1 \le i \le k \right\}$$

The **span** of a set of matrices is defined as the span of the union of their column vectors. For any differentiable function $g : \mathbb{R}^d \to \mathbb{R}$ and direction $\delta \in \mathbb{R}^d \setminus \{0\}$, its corresponding directional derivative is defined by

$$\lim_{\alpha \to 0} \frac{g(x + \alpha\delta) - g(x)}{\alpha} = \nabla_x g(x)^\top \delta$$

2.4 Challenges

The main challenge in analyzing whether a class of flows can universally approximate any target distribution when applied to a fixed source is *invertibility*. To understand this, suppose $\mathcal{F}, \mathcal{C}$ are function classes and $\mathcal{I}$ is the set of all invertible functions.

Even if $\mathcal{F}$ can approximate any function in $\mathcal{C}$, it might not hold that the invertible functions in $\mathcal{F}$ can approximate any invertible function in $\mathcal{C}$. This is because the set of invertible functions $\mathcal{I}$ might have no interior in

$\mathcal{C}$: for any invertible function, it is possible to modify it slightly to make it non-invertible – and hence the approximation to an invertible function $c \in \mathcal{C}$ may be a non-invertible function $f \in \mathcal{F}$ (see **Lemma** 4, [Mulansky and Neamtu, 1998]). For instance, it was shown that a certain ResNet ($\mathcal{F}$) is a universal approximator in $\mathcal{C} = \ell_1(\mathbb{R}^d)$ [Lin and Jegelka, 2018], and its invertible function subset ($\mathcal{F} \cap \mathcal{I}$) is exactly the set of transformations composed of finitely many ReLU planar flows. However, since the universal approximation property was proved by construction using the non-invertible trapezoid functions, this result does not translate to ReLU planar flows.

Dually, if $\mathcal{F}$ has limited expressivity, it might still happen that functions in $\mathcal{F} \cap \mathcal{I}$ can approximate or even express transformations between arbitrary pairs of distributions. This is because a small subset of functions $\mathcal{T}$ (for instance, increasing triangular maps [Villani, 2008]) is enough to transform between distributions. Therefore, if $\mathcal{F} \cap \mathcal{I}$ is dense in $\mathcal{T}$, then it is expressive. It is however challenging to find all such dense sets $\mathcal{T}$.

3 The $d = 1$ case

In this section, we discuss the universal approximation properties of Sylvester flows when the data dimension $d = 1$. In this case, a Sylvester flow is identical to a planar flow. However, the one-dimensional case is not trivial and requires delicate design. For both general and ReLU non-linearity cases, we demonstrate they are able to achieve universal approximation.

3.1 General Smooth Non-linearity

Suppose the flow f is a single planar flow with an arbitrary smooth non-linearity h. It is straightforward to show by construction that if $\textbf{supp}\ p = \textbf{supp}\ q = \mathbb{R}$, then there exists a planar flow that exactly transforms q into p. (See **Lemma** A.1). Using these exact transformations, we can approximate any density supported on a finite union of intervals when the input distribution is supported on $\mathbb{R}$ (e.g. a Gaussian).

Theorem 3.1 (Universal Approximation)**.** *Let p, q be densities on $\mathbb{R}$ such that p is supported on a finite union of intervals and $\textbf{supp}\ q = \mathbb{R}$. Then, for any $\epsilon > 0$, there exists a planar flow f_{pf} such that $\|f_{\text{pf}} \# q - p\|_1 \leq \epsilon$.*

Since in **Theorem** 3.1, the support of p might not be $\mathbb{R}$, we are unable to achieve exact transformation between p and q. However, approximation is possible in that we can transform q into $\tilde{p}$, a distribution supported on $\mathbb{R}$ but approximates p in ℓ_1 norm. To achieve this, we construct such $\tilde{p}$ that satisfying $\tilde{p} \approx p$ on $\textbf{supp}\ p$ and $\tilde{p} \approx 0$ on $\overline{\textbf{supp}\ p}$. An example is shown in Figure 1, where $p(x) = \frac{3}{4}\min((|x|-1)^2, (|x|-3)^2)$ for $1 \leq |x| \leq 3$ and $p(x) = 0$ elsewhere.

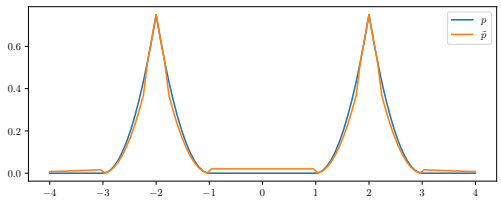

Figure 1: Target distribution p and its approximation $\tilde{p}$ with $\textbf{supp}\ \tilde{p} = \mathbb{R}$.

3.2 ReLU Non-linearity

Since the ReLU activation has been proven to be expressive and is popular in recent neural network models [He et al., 2016, Lin and Jegelka, 2018], we provide a universal approximation result for planar flows with ReLU non-linearity.

Suppose the one-dimensional ReLU flow has the form $f(z) = f_{\text{pf}}(z) = z + uh(wz + b)$, where $h = \text{ReLU}$. Since ReLU is linear on both $\mathbb{R}^-$ and $\mathbb{R}^+$, we assign $u = \pm 1$ for concreteness. In addition, to ensure the transformation is strictly increasing, we require $uw > -1$. Different from the general non-linearity case, the determinant of $\det J_f$ in (9) indicates that a ReLU planar flow keeps a halfspace of $\mathbb{R}$ and applies linear scaling transformation to the other halfspace.

Given that the input distribution q is Gaussian, we prove it is possible to approximate any density supported on a finite union of intervals in ℓ_1 norm using a finite number of ReLU planar flows.

Theorem 3.2 (Universal Approximation)**.** *Let p be a density on $\mathbb{R}$ supported on a finite union of intervals. Then, for any $\epsilon > 0$, there exists a flow f composed of finitely many ReLU planar flows and a Gaussian distribution $q_{\mathcal{N}}$ such that $\|f \# q_{\mathcal{N}} - p\|_1 \leq \epsilon$.*

There are two steps in the proof. First, we show that Gaussian distributions can be exactly transformed to tail-consistent piecewise Gaussian distributions (see **Definition** A.3, **Definition** A.4 for formal definitions and **Lemma** A.3). An example of a tail-consistent piecewise Gaussian distribution of three pieces is shown in Figure 2: the distribution is composed of three Gaussian pieces in full lines of three colors, where the dashed lines are corresponding prolongations. Then, the area below yellow lines (—/- -) is equal to the area below the blue dashed line (- -), and the area below the green full line (—) is equal to the area below the yellow dashed line (- -).

In the second step, we show that tail-consistent piece-

wise distributions can approximate any piecewise constant distribution supported on a finite union of compact intervals (see **Lemma** A.4). Notice that piecewise constant functions supported on a finite union of compact intervals can approximate any Lebesgue-integrable function [Lin and Jegelka, 2018], so do densities supported on a finite union of intervals. Therefore, the universal approximation property of ReLU planar flows (**Theorem** 3.2) is obtained.

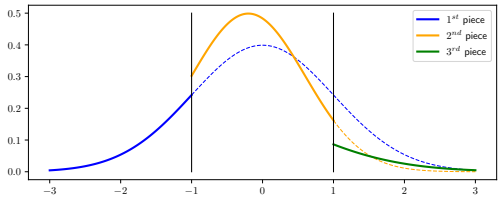

Figure 2: A tail-consistent piecewise Gaussian distribution in $\mathcal{PW}(3, \mathcal{G})$.

In Figure 3, two examples are presented on approximating the same target distribution p with different number of ReLU planar flows. As illustrated, the approximation almost reaches perfection when we choose a larger number of ReLU planar flows.

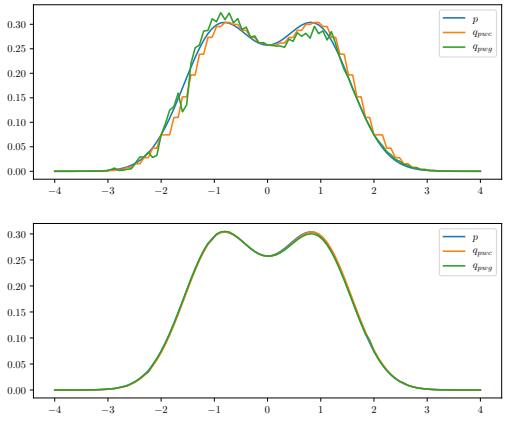

Figure 3: Target distribution p, its piecewise constant distribution approximation q_{pwc} of 50 (top)/300 (bottom) pieces, and its tail-consistent piecewise Gaussian distribution approximation q_{pwg} generated by 50 (top)/300 (bottom) ReLU planar flows over a Gaussian.

Remark. *Since we can transform the standard Gaussian distribution $\mathcal{N}(0, 1)$ to any other Gaussian distribution using a scaling function, which can be achieved by two ReLU planar flows and a shift, we can further assign the input distribution $q_{\mathcal{N}}$ in* ***Theorem*** *3.2 to be the standard Gaussian distribution.*

4 Exact Transformation for $d > 1$

In this section, we consider the exact transformation question when the data dimension $d > 1$. We study two cases where the flow is composed of a finite number of Sylvester flows with (i) ReLU non-linearity and (ii) general non-linearity. We specifically show how the topology matching conditions yield negative results to the exact transformation question (that is, to show there does not exist such flow that can transform between certain distributions).

Our results are based on the following key observation for a flow $f : \mathbb{R}^d \to \mathbb{R}^d$. For almost every $z \in \mathbb{R}^d$ there exists a subspace $\mathcal{V}(z) \subset \mathbb{R}^d$ such that for any $v \in \mathcal{V}$ and small $\alpha > 0$, $\det J_f(z) = \det J_f(z + \alpha v)$. We call $\mathcal{V}$ the complementary subspace of f at z. This observation can be used to determine what class of distributions flows can transform between. By letting $\alpha \to 0$, we can focus on properties of small neighbourhoods around z, which we call *topology matching*.

4.1 ReLU Non-linearity

We begin with constructing a topology matching condition for ReLU Sylvester flows: $f(z) = f_{\text{syl}}(z) = Z + A\ \text{ReLU}(B^\top z + b)$. (8) shows that for a single ReLU Sylvester flow, if $B^\top z + b \neq 0$, then $\det J_f(z') = \det J_f(z)$ when z' is close to z. This statement can be further generalized: if f is a flow composed of a finite number of ReLU Sylvester flows, for almost every $z \in \mathbb{R}^d$, the determinant of the Jacobian of f is a constant near z. Based on this observation, we conclude that the complementary subspace $\mathcal{V}(z) = \mathbb{R}^d$, *a.e.* (see **Lemma** A.5). Using this property, we construct the topology matching condition in the following theorem.

Theorem 4.1 (Topology Matching for ReLU Sylvester flows)**.** *Suppose distribution q is defined on $\mathbb{R}^d$, and flow f is composed of finitely many ReLU Sylvester flows on $\mathbb{R}^d$. Let $p = f\#q$. Then, there exists a zero-measure closed set $\Omega \subset \mathbb{R}^d$ such that $\forall z \in \mathbb{R}^d \setminus \Omega$, we have*

$$J_f(z)^\top \nabla_z \log p(f(z)) = \nabla_z \log q(z)$$

Intuitively, the local directional derivatives of the logarithm of the density are preserved. As a special case, if z satisfies $\nabla_z q(z) = 0$ (which means that z is a local minima, local maxima, or saddle point of q), then $p(f(z))$ must also have zero gradient at z. For instance, suppose p is the standard Gaussian distribution on $\mathbb{R}^2$ and q is a mixture of two Gaussian distributions on $\mathbb{R}^2$ with two peaks. Since only at the origin does p have zero gradient, we conclude there does not exist a planar flow that transforms q to p. Additional examples are illustrated in Figure 6 in the Appendix.

The proof of **Theorem** 4.1 follows from (2), the Taylor expansion of f, and the observation that $\mathcal{V}(z) = \mathbb{R}^d$ *a.e.*. Notably, the conclusion holds for any number of ReLU Sylvester flows. Using this condition, we show in the following corollaries that it is unlikely for finitely many ReLU Sylvester flows to transform between mixture of Gaussian (MoG) or product (Prod) distributions unless special conditions are satisfied.

Corollary 4.1.1 (MoG↛MoG)**.** *(See formal version in **Corollary** A.5.1) Suppose p, q are mixture of Gaussian distributions on $\mathbb{R}^d$ in the following form:*

$$p(z) = \sum_{i=1}^{r_p} w_p^i \mathcal{N}(z; \mu_p^i, \Sigma_p),\ \ q(z) = \sum_{j=1}^{r_q} w_q^j \mathcal{N}(z; \mu_q^j, \Sigma_q)$$

Then, there generally does not exist flow f composed of finitely many ReLU Sylvester flows such that $p = f\#q$.

Corollary 4.1.2 (Prod↛Prod)**.** *(See formal version in **Corollary** A.5.2) Suppose p and q are product distributions in the following form:*

$$p(z) \propto \prod_{i=1}^{d} g(z_i)^{r_p};\ \ q(z) \propto \prod_{i=1}^{d} g(z_i)^{r_q}$$

where $r_p, r_q > 0, r_p \neq r_q$, and g is a smooth function. Then, there generally does not exist flow f composed of finitely many ReLU Sylvester flows such that $p = f\#q$.

Given our negative results, the reader might wonder what distributions can be transformed by ReLU Sylvester flows. We show that certain linear transformations can be exactly expressed (see **Theorem** A.6, **Corollary** A.6.1 and **Corollary** A.6.2).

4.2 General Smooth Non-linearity

In this section, we construct a topology matching condition for Sylvester flows with general non-linearities. Suppose f is a Sylvester flow $f(z) = z + Ah(B^\top z + b)$ with flow dimension m, where h is an arbitrary smooth function. Analogous to **Theorem** 4.1, there exists a $d - m$ dimensional complementary subspace of f at every point $z \in \mathbb{R}^d$: $\mathcal{V}(z) = \mathbf{span}\{B\}^\perp$. Using this property, we are able to establish the topology matching condition for a single Sylvester flow (see **Lemma** A.7). Then, we generalize this result to n layers of Sylvester flows in the following theorem.

Theorem 4.2 (Topology Matching for Sylvester flows)**.** *Suppose distribution q is defined on $\mathbb{R}^d$, and n Sylvester flows $\{f_i\}_{i=1}^n$ on $\mathbb{R}^d$ have flow dimensions $\{m_i\}_{i=1}^n$, tangent matrices $\{B_i\}_{i=1}^n$, and smooth non-linearities. Let $f = f_n \circ \cdots \circ f_1$ and $p = f\#q$. Then $\forall z \in \mathbb{R}^d$, we have*

$$\nabla_z \log p(f(z)) - \nabla_z \log q(z) \in \boldsymbol{span}\{B_1, B_2, \cdots, B_n\}$$

When the sum of flow dimensions of $\{f_i\}_{i=1}^n$ is strictly less than the data dimension d, $\mathbf{span}\{B_1, B_2, \cdots, B_n\}$ is a strict subspace of $\mathbb{R}^d$. Under this situation, we show in the following corollary that transformation between Gaussian distributions might be impossible with a bounded number of Sylvester flows.

Corollary 4.2.1 ($\mathcal{N} \nrightarrow \mathcal{N}$)**.** *(See formal version in **Corollaries** A.7.1 and A.7.2) Let $p \sim \mathcal{N}(0, \Sigma_p), q \sim \mathcal{N}(0, \Sigma_q)$ be two Gaussian distributions on $\mathbb{R}^d$, and $\Sigma_q^{-1} - \Sigma_p^{-1}$ has high rank. Then, with a limited number of planar or Sylvester flows that have smooth non-linearities, it is impossible to transform q to p.*

Additional experiments are demonstrated in Figure 7 in the Appendix. We also construct a topology matching condition for radial flows in **Theorem** A.8, and compare that result with **Theorem** 4.2.

5 Approximation Capacity for Large d

In this section, we provide a partially negative answer to the universal approximation question for certain normalizing flows by showing that approximations in these cases may require very deep flows. In particular, we study local planar flows and Householder flows with specific target distributions.

Given an input distribution q and a target distribution p on $\mathbb{R}^d$, our goal is to lower bound the depth T of a normalizing flow that can transform q to an approximation of p. This is formally defined below.

Definition 5.1. *Let p, q be two distributions on $\mathbb{R}^d$, $\epsilon > 0$, and $\mathcal{F}$ be a set of normalizing flows. Then, the minimum number of flows in $\mathcal{F}$ required to transform q to an approximation of p to within ϵ is*

$$T_\epsilon(p, q, \mathcal{F}) = \inf\{n : \ \exists\{f_i\}_{i=1}^n \in \mathcal{F} \text{ such that } \|(f_1 \circ \cdots \circ f_n)\#q - p\|_1 \leq \epsilon\}$$

To achieve this goal, we look at the maximum ℓ_1 norm distance reduction of a normalizing flow f towards p:

$$\mathcal{L}(p, f) = \sup_{q' \text{ is a density on } \mathbb{R}^d} \|p - q'\|_1 - \|p - f\#q'\|_1$$

We first show a surprisingly concise upper bound $\hat{\mathcal{L}}$ of $\mathcal{L}$. This bound is used in proving **Theorem** 5.2 and **Theorem** 5.3 in this section.

Lemma 5.1. *$\mathcal{L}(p, f) \leq \hat{\mathcal{L}}(p, f)$, where*

$$\hat{\mathcal{L}}(p, f) = \int_{\mathbb{R}^d} |\det J_f(z)| p(f(z)) - p(z)| \, dz$$

Then, we naturally obtain a lower bound of T:

$$T_\epsilon(p, q, \mathcal{F}) \geq \frac{\|p - q\|_1 - \epsilon}{\sup_{f \in \mathcal{F}} \mathcal{L}(p, f)} \geq \frac{\|p - q\|_1 - \epsilon}{\sup_{f \in \mathcal{F}} \hat{\mathcal{L}}(p, f)}$$

Next, we make the following assumption on q:

Assumption 1. $\|p-q\|_1 = \Theta(1)$.

This assumption holds when the input distribution q is a random initialization (that is, q is chosen arbitrarily without any prior knowledge on p). Then, under **Assumption** 1, there exists $\epsilon > 0$ (e.g. $\epsilon = \frac{1}{2}\|p-q\|_1$) such that

$$T_\epsilon(p, q, \mathcal{F}) = \Omega\left(\frac{1}{\sup_{f\in\mathcal{F}} \hat{\mathcal{L}}(p, f)}\right)$$

In the rest of this section, we use this lower bound on T to construct results for local planar flows and Householder flows with specific target distributions.

5.1 Local Planar Flows

In this section, we look at a specific group of planar flows, which we call the *local* planar flows. A c_h-local planar flow is defined below.

Definition 5.2. *A non-linearity h is called c_h-local if there is a constant $c_h \in \mathbb{R}$ satisfying for any $x \in \mathbb{R}$, (i) $|h(x)| \le c_h$, and (ii) $|h'(x)| \le c_h/(1+|x|)$. A planar flow $f(z) = z + uh(w^\top z + b)$ is called c_h-local if h is c_h-local, $\|u\|_2 \le 1$, and $\|w\|_2 \le 1$.*

Many popular non-linearities are c_h-local, such as tanh ($c_h = 2$), sigmoid ($c_h = 1$), and arctan ($c_h = \pi/2$).

Geometrically, a local planar flow applies non-linear scaling on the region near the $d-1$ dimensional subspace $\{z : w^\top z + b = 0\}$ in $\mathbb{R}^d$, while having little effect on regions far away from the subspace (almost a constant shift). This observation leads to the intuition that one layer of local planar flow can only affect a small volume of the whole space, so a large number of layers is needed to approximate the target distribution if **supp** p is a large region. In the following theorem, we show for certain p, T goes up polynomially in the data dimension d with adjustable degrees.

Theorem 5.2 (ℓ_1 norm approximation lower bound for local planar flows)**.** *Let p be a distribution on $\mathbb{R}^d$ ($d > 2$) such that for $\tau \in (0, 1)$:*

- *$p = \mathcal{O}(p_1)$, where density p_1 satisfies*

$$p_1(z) \propto \exp(-\|z\|_2^\tau)$$

- *$\|\nabla p\|_2 = \mathcal{O}(\|\nabla p_2\|_2)$, where density p_2 satisfies*

$$p_2(z) \propto \begin{cases} \exp(-d) & \|z\|_2 \le d^{\frac{1}{\tau}} \\ \exp(-\|z\|_2^\tau) & \|z\|_2 > d^{\frac{1}{\tau}} \end{cases}$$

*Suppose $\mathcal{F}$ is the set of all c_h-local planar flows. Then, under **Assumption** 1, there exists $\epsilon = \Theta(1)$ such that*

$$T_\epsilon(p, q, \mathcal{F}) = \Omega\left(\min\left((\log d)^{-\frac{1}{\tau}} d^{\left(\frac{1}{\tau}-\frac{1}{2}\right)}, d^{\left(\frac{1}{\tau}-1\right)}\right)\right)$$

This indicates that if the target distribution p has specifically bounded values and gradients, a large number of local planar flows is needed to approximate p starting with a distribution q that obeys **Assumption** 1. The number T is polynomial in d with adjustable degrees, so it can be incredibly large as d gets large.

A concrete example that satisfies the condition in **Theorem** 5.2 is when $p(z)$ is equal to the p_2 in the statement. This satisfies the first condition because $\exp(-d) \le \exp(-\|z\|_2^\tau)$ in the ball centered at the origin with radius $d^{1/\tau}$, and the integration of p_1 in this ball is $o(1)$ (see proof of **Lemma** A.9). Then, taking for instance $\tau = 0.2$, the lower bound on T becomes $\Omega(d^4)$, which is incredibly large in practical scenarios.

To prove **Theorem** 5.2, we first show that $\hat{\mathcal{L}}(p, f)$ is upper bounded by an integration of two terms. We then present **Lemma** A.9 and **Lemma** A.10 to bound these two terms separately.

5.2 Householder Flows

In this section, we look at Householder flows. Since a Householder matrix does not change the ℓ_2 norm of any vector, it is possible to upper bound $\mathcal{L}$ when the target distribution p is almost symmetric, according to **Lemma** 5.1. If p is a standard Gaussian distribution, we have $\mathcal{L} = 0$, indicating that Householder flows cannot transform any different distribution to a standard Gaussian distribution. In the following theorem, we provide a concise bound on T when p is very close to the standard Gaussian distribution, where there is only a small perturbation on its covariance matrix.

Theorem 5.3 (ℓ_1 norm approximation lower bound for Householder flows)**.** *Let p be a Gaussian distribution $\mathcal{N}(0, I+S)$ on $\mathbb{R}^d$ ($d > 2$), where $|S_{ij}| \le d^{-(2+\kappa)}$ for some $\kappa > 0$ and any $1 \le i, j \le d$. Suppose $\mathcal{F}$ is the set of all Householder flows. Then, under **Assumption** 1, there exists $\epsilon = \Theta(1)$ such that*

$$T_\epsilon(p, q, \mathcal{F}) = \Omega\left(d^\kappa\right)$$

This indicates that we need a large number of Householder flows to approximate a distribution close to the standard Gaussian distribution, starting with a distribution q that obeys **Assumption** 1. The number T is also polynomial in the data dimension d with adjustable degrees, so it could be large as well. The bound is computed from $\hat{\mathcal{L}}$, where $|\det J_f(z)| = 1$ for a Householder flow f.

6 Additional Related Work

6.1 Normalizing Flows

It was shown that transforming a simple distribution to a complicated one by composing many simple transformations can be used to solve density estimation problems [Tabak et al., 2010, Tabak and Turner, 2013]. These transformations are called *normalizing flows*. Two basic normalizing flows (planar and radial flows) were introduced [Rezende and Mohamed, 2015]. Due to their empirical success, there has been a growing body of work on other kinds of normalizing flows. Two categories of normalizing flows have been developed.

Triangular flows. It was proven that increasing triangular functions can transform between arbitrary distributions [Villani, 2008]. Therefore, triangular flows composed of fixed classes of increasing triangular functions are expected to enjoy good expressive power. In addition, the determinant of the Jacobian matrix of an increasing triangular function is easy to compute. These two benefits have led to the development of a large family of triangular flows [Dinh et al., 2014, Germain et al., 2015, Uria et al., 2016, Kingma et al., 2016, Dinh et al., 2016, Papamakarios et al., 2017, Huang et al., 2018, Jaini et al., 2019]. Among these flows, IAF [Kingma et al., 2016], NAF [Huang et al., 2018] and SOS flows [Jaini et al., 2019] were shown to have the universal approximation property.

Non-triangular flows. It is possible to calculate the determinant of the Jacobian matrix and the inverse of a well designed non-triangular function. Several flows parameterized by matrices were inspired by results from linear algebra and thus enjoy this property [Tomczak and Welling, 2016, Hasenclever et al., 2017, Ho et al., 2019, Berg et al., 2018], where the last one is a matrix-form generalization of the planar flow. Moreover, a recent non-triangular flow, the iResNet [Behrmann et al., 2018], in the form of residual networks (ResNet) [He et al., 2016], was designed with an efficient log-det approximator. It was further improved in residual flows with an unbiased approximator [Chen et al., 2019]. However, the expressivity of these flows still remain unknown, even though the iResNet is expressed by powerful neural networks.

6.2 Continuous Time Flows

It is possible, from the infinitesimal point of view, to generalize the discrete update of finite flows to continuous update of infinite flows. Infinite flows are described by a differential equation instead of a sequence of transformations in the finite flow context [Chen et al., 2017, Grathwohl et al., 2018, Chen et al., 2018, Salman et al., 2018, Zhang et al., 2018]. The neural ODEs [Chen et al., 2018] is one significant work in this class, but its expressivity still lacks understanding. A counter-example was provided on the expressivity of the neural ODEs [Dupont et al., 2019]. However, this does not rigorously imply that neural ODEs are not universal approximators because (i) the failure in exact transformation does not imply the impossibility in approximation, and (ii) universal transformation does not necessarily need universal function representation.

To tackle the problem of such counter-example, additional p dimensions were introduced to "augment" the neural ODEs [Dupont et al., 2019]. By solving a $d+p$ dimensional augmented ODE and extracting the first d dimensions, the expressivity of the neural ODEs is enhanced. It was further shown that the augmented neural ODEs is a universal approximator in the continuous function space when $p = 1$ [Zhang et al., 2019]. Nevertheless, in the context of normalizing flows, every transformation has to be invertible, so the change of dimension strategy, as well as its universal approximation property, does not apply to normalizing flows.

7 Conclusions

Normalizing flows are a class of deep generative models that offer flexible generative modeling as well as easy likelihood computation. While there has been a great deal of prior empirical work on different normalizing flow models, not much is (formally) known about their expressive power; we provide one of the first systematic studies on non-triangular flows. Our results demonstrate that one needs to be careful while designing normalizing flow models as well as their non-linearities in high dimensional space. In particular, we show that Sylvester flows, a universal approximator in one dimension, are unable to exactly transform between two (even simple) distributions unless rigorous conditions are satisfied. Additionally, a prohibitively large number of layers of planar or Householder flows are required to reduce the ℓ_1 distance between input and output distributions under certain conditions.

There are a large number of open problems. Some unresolved problems towards expressivity of simple flows include (i) are certain combinations of tangent matrices or non-linearities useful, (ii) can normalizing flows composed of finitely many ($\geq d$) Sylvester flows with arbitrary non-linearities (or other simple flows) transform between any pair of input-output distributions in high dimensional space, (iii) are such normalizing flows universal approximators in converting distributions, and (iv) what class of distributions are easy or hard for normalizing flows composed of Sylvester flows or other simple flows to transform between. A final open problem is to look at other, more general classes of flows, and provide upper and lower bounds on their expressive power under different non-linearities.

Acknowledgements

We thank NSF under IIS 1617157 for research support.

References

[Armstrong, 2013] Armstrong, M. A. (2013). *Basic topology*. Springer Science & Business Media.

[Bailey and Telgarsky, 2018] Bailey, B. and Telgarsky, M. J. (2018). Size-noise tradeoffs in generative networks. In *Advances in Neural Information Processing Systems*, pages 6489–6499.

[Behrmann et al., 2018] Behrmann, J., Duvenaud, D., and Jacobsen, J.-H. (2018). Invertible residual networks. *arXiv preprint arXiv:1811.00995*.

[Berg et al., 2018] Berg, R. v. d., Hasenclever, L., Tomczak, J. M., and Welling, M. (2018). Sylvester normalizing flows for variational inference. *arXiv preprint arXiv:1803.05649*.

[Chen et al., 2017] Chen, C., Li, C., Chen, L., Wang, W., Pu, Y., and Carin, L. (2017). Continuous-time flows for efficient inference and density estimation. *arXiv preprint arXiv:1709.01179*.

[Chen et al., 2019] Chen, R. T., Behrmann, J., Duvenaud, D., and Jacobsen, J.-H. (2019). Residual flows for invertible generative modeling. *arXiv preprint arXiv:1906.02735*.

[Chen et al., 2018] Chen, T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. (2018). Neural ordinary differential equations. In *Advances in neural information processing systems*, pages 6571–6583.

[Cybenko, 1989] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.

[Dinh et al., 2014] Dinh, L., Krueger, D., and Bengio, Y. (2014). Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*.

[Dinh et al., 2016] Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2016). Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*.

[Dupont et al., 2019] Dupont, E., Doucet, A., and Teh, Y. W. (2019). Augmented neural odes. *arXiv preprint arXiv:1904.01681*.

[Germain et al., 2015] Germain, M., Gregor, K., Murray, I., and Larochelle, H. (2015). Made: Masked autoencoder for distribution estimation. In *International Conference on Machine Learning*, pages 881–889.

[Grathwohl et al., 2018] Grathwohl, W., Chen, R. T., Betterncourt, J., Sutskever, I., and Duvenaud, D. (2018). Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*.

[Hanin, 2017] Hanin, B. (2017). Universal function approximation by deep neural nets with bounded width and relu activations. *arXiv preprint arXiv:1708.02691*.

[Hasenclever et al., 2017] Hasenclever, L., M. Tomczak, J., van den Berg, R., and Welling, M. (2017). Variational inference with orthogonal normalizing flows. In *Workshop on Bayesian Deep Learning (NIPS 2017)*, Long Beach, CA, USA.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[Ho et al., 2019] Ho, J., Chen, X., Srinivas, A., Duan, Y., and Abbeel, P. (2019). Flow++: Improving flow-based generative models with variational dequantization and architecture design. *arXiv preprint arXiv:1902.00275*.

[Hornik, 1991] Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257.

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

[Huang et al., 2018] Huang, C.-W., Krueger, D., Lacoste, A., and Courville, A. (2018). Neural autoregressive flows. *arXiv preprint arXiv:1804.00779*.

[Jaini et al., 2019] Jaini, P., Selby, K. A., and Yu, Y. (2019). Sum-of-squares polynomial flow. *arXiv preprint arXiv:1905.02325*.

[Kingma et al., 2016] Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, pages 4743–4751.

[Lin and Jegelka, 2018] Lin, H. and Jegelka, S. (2018). Resnet with one-neuron hidden layers is a universal approximator. In *Advances in Neural Information Processing Systems*, pages 6169–6178.

[Lu et al., 2017] Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. (2017). The expressive power of neural

networks: A view from the width. In *Advances in neural information processing systems*, pages 6231–6239.

[Lütkepohl, 1996] Lütkepohl, H. (1996). *Handbook of matrices*, volume 1. Wiley Chichester.

[Montufar et al., 2014] Montufar, G. F., Pascanu, R., Cho, K., and Bengio, Y. (2014). On the number of linear regions of deep neural networks. In *Advances in neural information processing systems*, pages 2924–2932.

[Mulansky and Neamtu, 1998] Mulansky, B. and Neamtu, M. (1998). Interpolation and approximation from convex sets. *Journal of approximation theory*, 92(1):82–100.

[Muleshkov and Nguyen, 2017] Muleshkov, A. and Nguyen, T. (2017). Easy proof of the jacobian for the n-dimensional polar coordinates.

[Neuman, 2013] Neuman, E. (2013). Inequalities and bounds for the incomplete gamma function. *Results in Mathematics*, 63(3-4):1209–1214.

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

[Raghu et al., 2017] Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., and Dickstein, J. S. (2017). On the expressive power of deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2847–2854. JMLR. org.

[Rezende and Mohamed, 2015] Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538.

[Salman et al., 2018] Salman, H., Yadollahpour, P., Fletcher, T., and Batmanghelich, K. (2018). Deep diffeomorphic normalizing flows. *arXiv preprint arXiv:1810.03256*.

[Sherman and Morrison, 1950] Sherman, J. and Morrison, W. J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127.

[Tabak and Turner, 2013] Tabak, E. G. and Turner, C. V. (2013). A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164.

[Tabak et al., 2010] Tabak, E. G., Vanden-Eijnden, E., et al. (2010). Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233.

[Telgarsky, 2015] Telgarsky, M. (2015). Representation benefits of deep feedforward networks. *arXiv preprint arXiv:1509.08101*.

[Tomczak and Welling, 2016] Tomczak, J. M. and Welling, M. (2016). Improving variational auto-encoders using householder flow. *arXiv preprint arXiv:1611.09630*.

[Uria et al., 2016] Uria, B., Côté, M.-A., Gregor, K., Murray, I., and Larochelle, H. (2016). Neural autoregressive distribution estimation. *The Journal of Machine Learning Research*, 17(1):7184–7220.

[Villani, 2008] Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

[Zhang et al., 2019] Zhang, H., Gao, X., Unterman, J., and Arodz, T. (2019). Approximation capabilities of neural ordinary differential equations. *arXiv preprint arXiv:1907.12998*.

[Zhang et al., 2018] Zhang, L., Wang, L., et al. (2018). Monge-ampere flow for generative modeling. *arXiv preprint arXiv:1809.10188*.