The Expressive Power of a Class of Normalizing Flow Models

Zhifeng Kong, Kamalika Chaudhuri

UC San Diego
Overview

Background

Problem Statement

Challenges

Results

One-dimensional Universal Approximation
High-dimensional Exact Transformation
Approximation Capacity for Large $d$

Conclusions
Background: Normalizing Flows

Figure: A normalizing flow that transforms $p_0(z_0) \rightarrow p_K(z_K)$. 

- $p_K = f \# p_0$, where $f = f_K \circ \cdots \circ f_1$.  
- Each $f_i$ is simple, invertible, and parameterized.  
- Solve MLE ⇒ a generative model with computable likelihood.

---

1 Figure by Lilian Weng, https://lilianweng.github.io/lil-log/assets/images/normalizing-flow.png

Zhifeng Kong, Kamalika Chaudhuri  
The Expressive Power of a Class of Normalizing Flow Models
Basic Flows

- **Planar flows** ([Rezende and Mohamed, 2015]):
  \[ f_{\text{pf}}(z) = z + uh(w^\top z + b); \quad u, w \in \mathbb{R}^d, b \in \mathbb{R} \]

- **Radial flows** ([Rezende and Mohamed, 2015]):
  \[ f_{\text{rf}}(z) = z + \frac{b}{a + \|z - z_0\|} (z - z_0); \quad z_0 \in \mathbb{R}^d, a, b \in \mathbb{R} \]

**Figure:** Planar flows versus radial flows.
Basic Flows

- **Sylvester flows (Berg et al., 2018):**

  \[ f_{\text{syl}}(z) = z + Ah(B^\top z + b); \quad A, B \in \mathbb{R}^{d \times m}, b \in \mathbb{R}^m \]

  Sylvester flows are matrix-form generalization of planar flows. We say \( f_{\text{syl}} \) has \( m \) neurons.

- **Householder flows (Tomczak and Welling, 2016):**

  \[ f_{\text{hh}}(z) = z - 2vv^\top z; \quad v \in \mathbb{R}^d, \|v\| = 1 \]
Problem Statement

- Setting: \( f \) is composed of \( T \) basic normalizing flows.

- \( q \) – source(input) distribution, \( p \) – target distribution.

- \( \mathcal{Q}_1 \) – Exact transformation: when does it satisfy

\[
p = f \# q \ (a.e.)
\]

- \( \mathcal{Q}_2 \) – Approximation: given \( \epsilon > 0 \), is there a bound on \( T \) s.t.

\[
\| f \# q - p \|_1 \leq \epsilon
\]
Challenges – Invertibility

A normalizing flow is an *invertible* function!

Suppose $\mathcal{F}$ is a function class and $\mathcal{I} = \{\text{all invertible functions}\}$.

- $\mathcal{F}$ is a universal approximator $\nRightarrow \mathcal{F} \cap \mathcal{I}$ can transform between arbitrary distributions.
  E.g. piecewise constant functions

- $\mathcal{F}$ has limited expressivity $\nRightarrow \mathcal{F} \cap \mathcal{I}$ is not a universal approximator in transforming distributions.
  E.g. triangular transformations (Villani, 2008)

Our technique: directly look at input-output distribution pairs.
Results for Universal Approximation ($d = 1$)

If the non-linearity $h = \text{ReLU}$:

**Theorem**

If $\text{supp } p$ is a finite union of intervals, then $\forall \epsilon > 0$, $\exists$ a flow $f$ composed of finitely many ReLU planar flows and a Gaussian distribution $q_N$ such that $\| f # q_N - p \|_1 \leq \epsilon$.

Figure: Approximation with 50 (left) and 300 (right) planar flows.
Results for Exact Transformation \((d > 1)\)

If the non-linearity \(h\) is a smooth function:

**Theorem**

\[
\text{Let } f \text{ be composed of Sylvester flows and } p = f \# q. \text{ Let } L(z) = \log p(f(z)) - \log q(z). \text{ Then, } \dim \{\nabla_z L\} \leq \text{Num(neurons of } f)\]

Figure: \(q, p = f_{\text{syl}} \# q, \text{ and } L\).
Results for Exact Transformation ($d > 1$)

**Corollary ($\mathcal{N} \leftrightarrow \mathcal{N}$)**

Let $p \sim \mathcal{N}(0, \Sigma_p)$, $q \sim \mathcal{N}(0, \Sigma_q)$ and $f$ is composed of Sylvester flows. If $\text{Num(neurons of } f) < \frac{1}{2} \text{rank}(\Sigma_q^{-1} - \Sigma_p^{-1})$ then $p \neq f \# q$. 
Results for Exact Transformation \((d > 1)\)

If the non-linearity \(h = \text{ReLU}\):

**Theorem**

\(\text{Let } f \text{ be composed of finitely many ReLU Sylvester flows and } p = f \# q, \text{ then } J_f(z)\top \nabla_z \log p(f(z)) = \nabla_z \log q(z) \text{ a.e.}\)

\[\begin{align*}
\text{Figure: } q, p = f_{\text{syl}} \# q, \text{ and mapped peaks.}
\end{align*}\]
Corollary (MoG $\leftrightarrow$ MoG, Prod $\leftrightarrow$ Prod)

Suppose $p, q$ are either a pair of

1. mixture of Gaussian distributions:
   \[ p(z) = \sum_{i=1}^{r_p} w_p^i N(z; \mu_p^i, \Sigma_p), \quad q(z) = \sum_{j=1}^{r_q} w_q^j N(z; \mu_q^j, \Sigma_q), \]
   or

2. product distributions:
   \[ p(z) \propto \prod_{i=1}^{d} g(z_i)^{r_p}; \quad q(z) \propto \prod_{i=1}^{d} g(z_i)^{r_q}. \]

Then, generally there does not exist flow $f$ composed of finitely many ReLU Sylvester flows such that $p = f \# q$. 
Results for Approximation Capacity (Large $d$)

**Definition (minimum depth)**

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 $\epsilon$ is

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

**Assumption**

$$\|p - q\|_1 = \Theta(1)$$
Results for Approximation Capacity (Large $d$)

**Definition (local planar flow)**

$f_{pf}(z) = z + uh(w^\top z + b)$ is local if $\|u\|, \|w\| \leq 1$, and $\|h\|_\infty, |h'(x)(1 + |x|)|$ are bounded. (e.g. $h = \arctan, \text{sigmoid}, \tanh$, etc.)

**Theorem (planar flow $\ell_1$-approximation lower bound)**

Let $\mathcal{F}_{lpf}$ be the set of local planar flows. For any $\tau > 0$, there exists a distribution $p$ on $\mathbb{R}^d$ and $\epsilon = \Theta(1)$ such that

$$T_\epsilon(p, q, \mathcal{F}_{lpf}) = \tilde{\Omega}(d^\tau)$$
Results for Approximation Capacity (Large $d$)

Theorem (Householder flow $\ell_1$-approximation lower bound)

Let $\mathcal{F}_{hh}$ be the set of Householder flows. For any $\tau > 0$, there exists a distribution $p$ on $\mathbb{R}^d$ and $\epsilon = \Theta(1)$ such that

$$T_{\epsilon}(p, q, \mathcal{F}_{hh}) = \Omega(d^{\tau})$$
Conclusions

Takeaways:

▶ On one dimension, planar flows are universal approximators.

▶ On higher dimensions, both exact transformation and approximation for basic flows may be hard.

Open problems:

▶ What distributions are these basic flows good at transforming between?

▶ What is the expressive power of deep Sylvester flows with other non-linearities?
References I


THANK YOU