The Expressive Power of a Class of Normalizing Flow Models

Zhifeng Kong, Kamalika Chaudhuri

UC San Diego

Overview

Background: Normalizing Flows

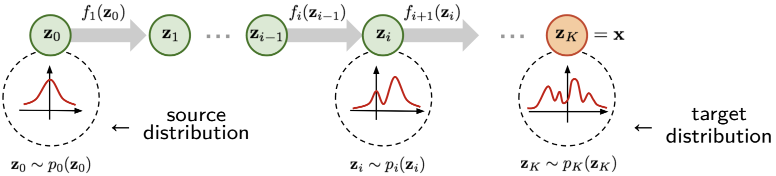

Figure: A normalizing flow that transforms $p_0(z_0) \to p_K(z_K)$. [1]

- $p_K = f\#p_0$, 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

Basic Flows

- Planar flows (Rezende and Mohamed, 2015):

$$f_{\text{pf}}(z) = z + uh(w^\top z + b);\ u, w \in \mathbb{R}^d, b \in \mathbb{R}$$

- Radial flows (Rezende and Mohamed, 2015):

$$f_{\text{rf}}(z) = z + \frac{b}{a + \|z - z_0\|}(z - z_0);\ z_0 \in \mathbb{R}^d, a, b \in \mathbb{R}$$

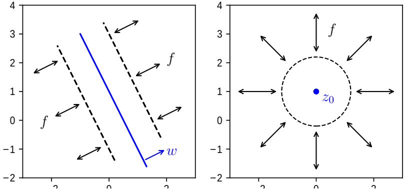

Figure: Planar flows versus radial flows.

Basic Flows

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

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

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

- Householder flows (Tomczak and Welling, 2016):

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

Problem Statement

- Setting: f is composed of T basic normalizing flows.
- q – source(input) distribution, p – target distribution.
- Q_1–*Exact transformation*: when does it satisfy

$$p = f\#q \ (a.e.)$$

- Q_2–*Approximation*: given > 0, is there a bound on T s.t.

$$f\#q - p \ _1 \leq$$

Challenges – *Invertibility*

A normalizing flow is an *invertible* function!

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

- F is a universal approximator ; $F \cap I$ can transform between arbitrary distributions.
  E.g. piecewise constant functions
- F has limited expressivity ; $F \cap I$ 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 = ReLU:

Theorem

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

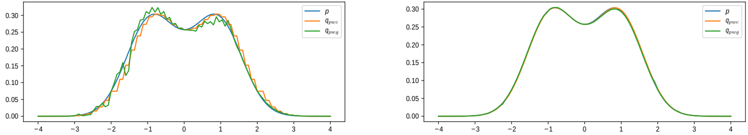

Figure: Approximation with 50 (left) and 300 (right) planar flows.

Results for Exact Transformation ($d > 1$)

If the non-linearity h is a smooth function:

Theorem

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

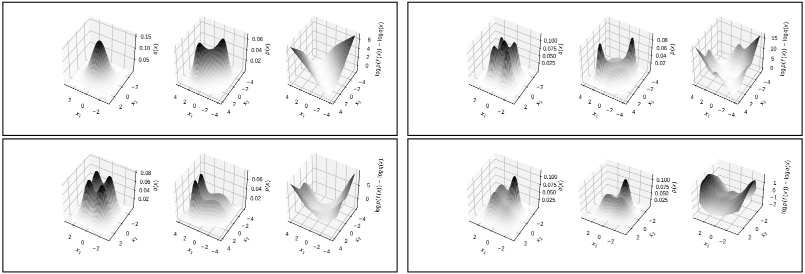

Figure: q, $p = f_{\text{syl}}\#q$, and L.

Results for Exact Transformation ($d > 1$)

Corollary (N 9 N)

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

Results for Exact Transformation ($d > 1$)

If the non-linearity h = ReLU:

Theorem

Let f be composed of finitely many ReLU Sylvester flows and $p = f\#q$, then $J_f(z) \quad _z \log p(f(z)) = \quad _z \log q(z)$ a.e.

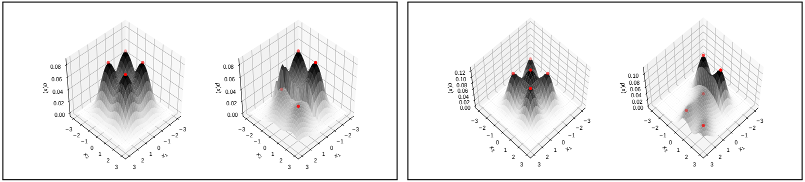

Figure: q, $p = f_{\text{syl}}\#q$, and mapped peaks.

Results for Exact Transformation ($d > 1$)

Corollary (MoG 9 MoG, Prod 9 Prod)

Suppose p, q are either a pair of

1. *mixture of Gaussian distributions:*
   $p(z) = \sum_{i=1}^{r_p} w_p^i N(z; \mu_p^i, \Sigma_p),\ q(z) = \sum_{j=1}^{r_q} w_q^j N(z; \mu_q^j, \Sigma_q)$,
   or
2. *product distributions:*
   $p(z) \quad \prod_{i=1}^{d} g(z_i)^{r_p};\ q(z) \quad \prod_{i=1}^{d} g(z_i)^{r_q}$.

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 R^d, $\quad > 0$, and F be a set of normalizing flows. Then, the minimum number of flows in F required to transform q to an approximation of p to within $\quad$ is

$$T\ (p, q, F) = \inf\{n : \quad \{f_i\}_{i=1}^{n} \quad F \text{ such that } (f_1 \circ \cdots \circ f_n)\#q - p \ _1 \leq \quad \}$$

Assumption

$$p - q \ _1 = \Theta(1)$$

Results for Approximation Capacity (Large d)

Definition (local planar flow)

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

Theorem (planar flow $_1$-approximation lower bound)

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

$$T\;(p, q, F_{lpf}) = \tilde{\Omega}\,(d^{\tau}\,)$$

Results for Approximation Capacity (Large d)

Theorem (Householder flow $_1$-approximation lower bound)

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

$$T\ (p, q, F_{hh}) = \Omega\,(d^{\tau}\,)$$

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

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

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

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

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

THANK YOU