The Expressive Power of a Class of Normalizing Flow Models

Medha Agarwal and Garrett Mulcahy

October 25, 2022

Contents

Estimating KR maps using input convex neural networks

Setup

- The goal is to describe a function class $\mathcal{F}$ such that for each $T \in \mathcal{F}$, T^j is a function of $x_{1:j}$ and is a monotonically increasing function of x_j.
- Let each $T \in \mathcal{F}$ be indexed by its model parameters θ s.t. $\forall\, \theta \in \Theta$, $T(\theta) \in \mathcal{F}$.
- Then the optimization objective as a function of model parameters is

$$KL(\theta) = -\frac{1}{n}\sum_{j=1}^{n}\left(\log q \circ T(\theta)(\mathbf{x}_j) + \sum_{i=1}^{d}\log(\nabla T(\theta))_i(\mathbf{x}_j)\right). \tag{1}$$

- Now, since $T(\theta)$ is a triangular map, it can be expressed as

$$S(\theta)(\mathbf{x}) = \left(T^1(x_1), T^2(x_{1:2}), \ldots, T^j(x_{1:j}), \ldots, T^d(\mathbf{x})\right)^T.$$

Input Convex Neural Network

- Input convex neural networks (ICNN) proposed by Amos et al. (2017) are scalar-valued neural networks $f(x, y; \theta)$ with inputs x and y, and are defined by the model parameters θ.
- $f(x, y; \theta)$ is a convex function of y.
- Now each component T^j can be modelled as the partial derivative (in jth input) of a partial ICNN which takes as input $x_{1:j}$ and is convex in the input x_j.
- That is, if $f^j(x_{1:j-1}, x_j; \theta_j)$ is the PICNN convex in x_j, then

$$T^j(x_{1:j}) := x_j + \frac{\partial f^j(x_j, x_{1:j-1}; \theta^j)}{\partial x_j}.$$

- This implies that $\theta = (\theta_1, \ldots, \theta_d)$ fully parameterizes the KR map, though some constraints need to be imposed on Θ that we will discuss in next slide.

Input Convex Neural Network

- PICNN $f(x, y; \theta)$ is described by two set of hidden layers $\{u_i\}_{i=1}^k$ and $\{z_i\}_{i=1}^k$ that correspond to the x-path and y-path respectively.
- The architecture of a K-layer PICNN is given by the following following recurring hidden units

$$u_{i+1} = \tilde{g}_i(\tilde{W}_i u_i + \tilde{b}_i)$$

$$z_{i+1} = g_i \left(W_i^{(z)} \left(z_i \circ [W_i^{(zu)} u_i + b_i^{(zu)}] \right) + W_i^{(y)} \left(y \circ [W_i^{(yu)} u_i + b_i^{(yu)}] \right) + W_i^{(u)} u_i + b_i \right),$$

where $\tilde{g}_i$ and g_i are non-linear activation functions. The final scalar-valued output of the PICNN is $f(x, y; \theta) = z_K$.

Training

Proposition 1, Amos et al. (2017)

The function f is convex in y provided that all $\{W_i^{(z)}\}_{i=1}^{k-1}$ are non-negative, and all functions $\tilde{g}_i$ are convex and non-decreasing.

- The proof follows the simple idea that the convex combinations and compositions of convex functions is also convex.
- To ensure that f^j is a convex function of x_j, we need to constrain all its entries of $W^{(z)}$ to be non-negative.
- The learning process optimizes the parameters θ such that $KL(\theta)$ from (1) is minimized.
- A regularization term is added to ensure that the $W^{(z)}$ weight matrix entries for all layers of all d PICNN is positive giving

$$-\frac{1}{n}\sum_{i=1}^{n}\left(\log g \circ T(\theta)(\mathbf{x}_i) + \sum_{j=1}^{d}\log(\nabla T(\theta))_j(\mathbf{x}_i)\right) + \lambda\sum_{j=1}^{d}\sum_{k=1}^{K}\| \max(-(W_j)_k^{(z)}, 0)\|_F^2 , \tag{2}$$

where λ is the regularization tuning parameter.

Motivation for studying expressive power

Motivation

Universality

- Universal approximation results are often asymptotic in nature and prove to be of little help in practice.
- Careful analysis of expressive power is required as a function of problem dimension and desired accuracy
  [see Lu et al. (2017), Lin and Jegelka (2018) for neural networks].

Invertibility

- Analyzing expressive power of the subset of invertible functions in $\mathcal{F}$ is a different problem than analyzing $\mathcal{F}$. Let $\mathcal{C} := \{f \in \mathcal{F} : f^{-1}\text{exits}\}$.
  - If $\mathcal{F}$ is a universal approximator $\not\Longrightarrow$ $\mathcal{C}$ can transform between any two distributions.
  - If $\mathcal{F}$ has limited expressivity $\not\Longrightarrow$ $\mathcal{C}$ cannot transform between any two distributions.

Problem Setup

Problem Setup

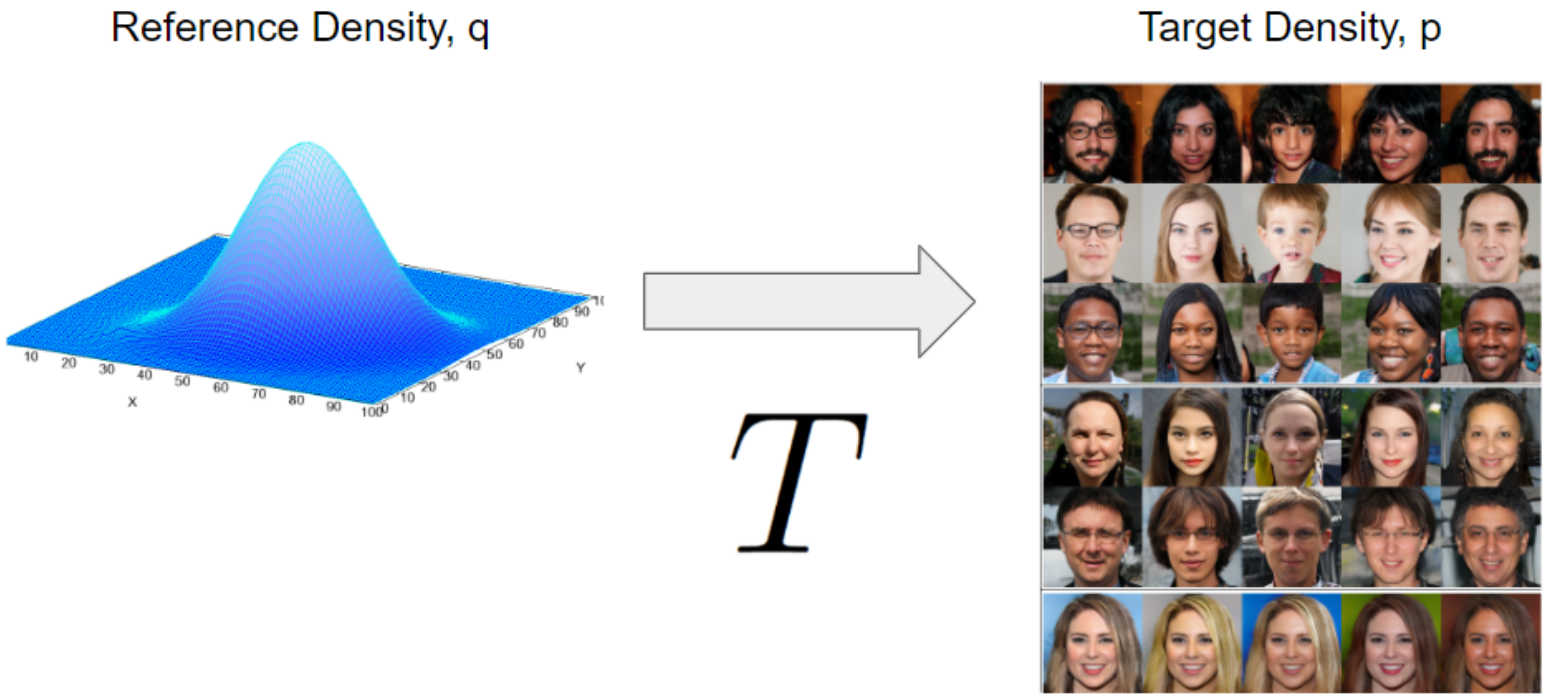


Figure: Normalizing Flow Schematic Diagram

Notations

- Let μ, ν be probability measures on $\mathbb{R}^d$ with Lebesgue densities p and q, respectively.
- We call μ and p the **target measure/density**, ν and q that **reference measure/density**
- Unlike previous notation, Kong and Chaudhuri (2020) construct flow that transports q to p, we seek f such that i.e. $f_{\#}q = p$.
- Using change of variable formula

$$p(x) = \frac{q(f^{-1}(x))}{\left|\det J_f(f^{-1}(x))\right|} \Leftrightarrow q(x) = p(f(x))\left|\det J_f(x)\right|$$

Sylvester Flows

Given a positive integer $m < d$, $A \in \mathbb{R}^{d \times m}$, $B \in \mathbb{R}^{d \times m}$, $b \in \mathbb{R}^m$, $h : \mathbb{R} \to \mathbb{R}$, define **Sylvester flow** $T_{syl} : \mathbb{R}^d \to \mathbb{R}^d$ by

$$T_{syl}(z) = z + Ah(B^T z + b).$$

We compute

$$J_{T_{syl}}(z) = \mathrm{Id}_d + A\mathrm{diag}(h'(B^T z + b))B^T,$$

By Sylvester's determinant identity (Kobyzev et al. (2020)), we get that

$$\begin{aligned} \det J_{T_{syl}}(z) &= \det(\mathrm{Id}_d + A\mathrm{diag}(h'(B^T z + b))B^T) \\ &= \det(\mathrm{Id}_m + \mathrm{diag}(h'(B^T z + b))B^T A). \end{aligned}$$

Proof of Universal Approximation Results, $d = 1$

General Universal Approximation

Theorem 3.1, Kong and Chaudhuri (2020)

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

Steps:

(1) Approx p with $\tilde{p}$ in L^1 such that **supp** $\tilde{p} = \mathbb{R}$.

(2) We ensure that $\tilde{p} \approx p$ on **supp** p and $\tilde{p} \approx 0$ on $\overline{\textbf{supp}}\, p$.

(3) Find f_{pf} such that $(f_{pf})_{\#}q = \tilde{p}$. This gives

$$\begin{aligned}\|(f_{pf})_{\#}q - p\|_1 &\leq \|(f_{pf})_{\#}q - \tilde{p}\|_1 + \|\tilde{p} - p\|_1 \\ &= \|\tilde{p} - p\|_1\end{aligned}$$

Approx Step

Our goal is the following construction

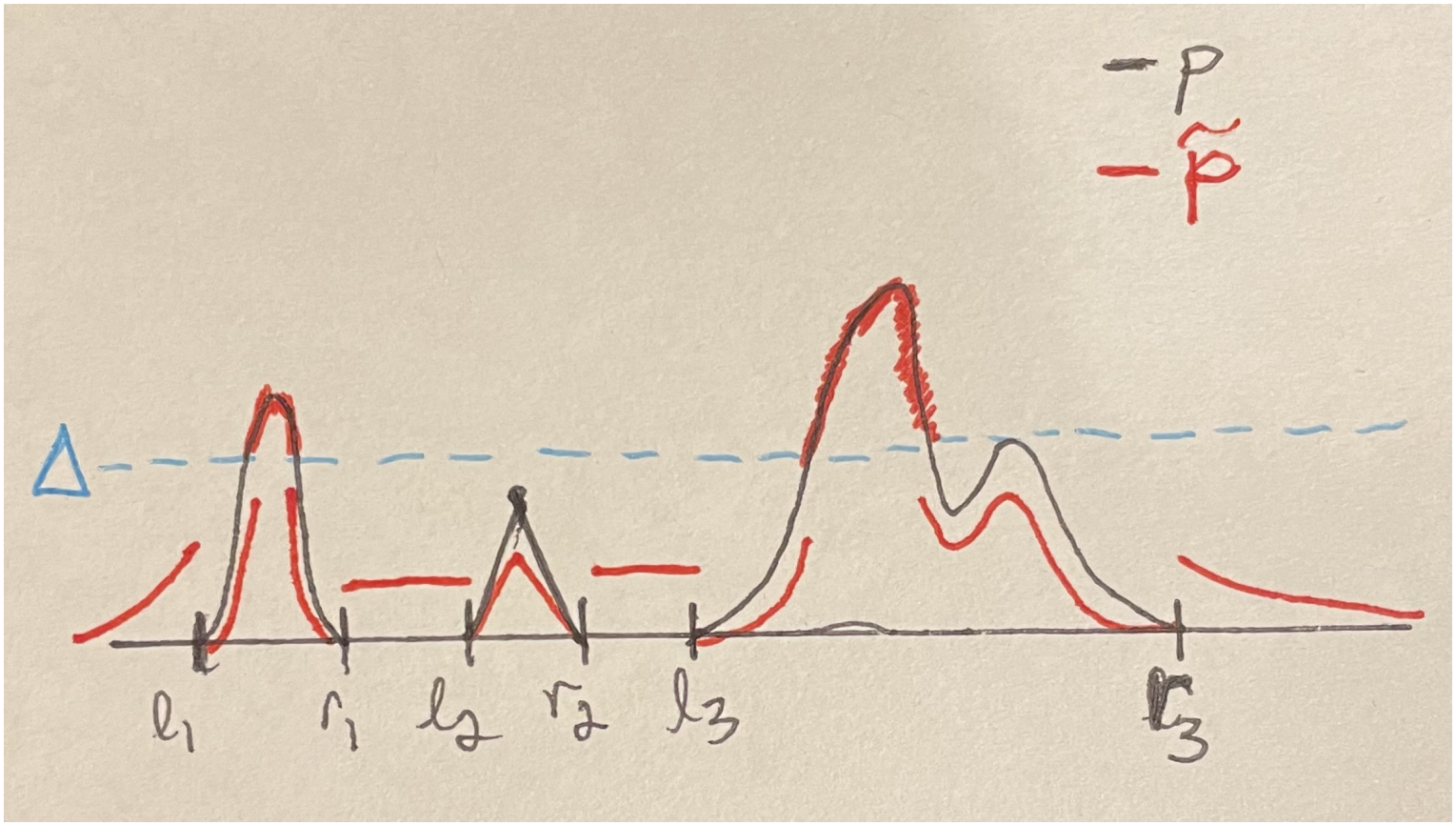


Figure: Approx of p with $\tilde{p}$ such that $\mathbf{supp}\ \tilde{p} = \mathbb{R}$

Approx Step

- WLOG insist **supp** $p = \bigcup_{j=1}^{n}(l_i, r_i)$ with $-\infty < l_i < r_i < \infty$ for all i
- Set cutoff height as

$$\Delta = \frac{2}{\sum_{j=1}^{n} r_i - l_i}$$

- Let $m(\Delta)$ be the measure of the set of points x such that $p(x) \geq \Delta$.
- Then observe

$$\begin{aligned} 1 = \int p(x)dx &= \int_{m(\Delta)} p(x)dx + \int_{m(\Delta)^C} p(x)dx \\ &\geq \int_{m(\Delta)} p(x)dx \\ &\geq \Delta m(\Delta) \\ \implies m(\Delta) &\leq \frac{1}{\Delta} \end{aligned}$$

- Define

$$\gamma = \int_{\{0<p(x)<\Delta\}} p(x)dz \leq 1$$

Approx Step

Now we will construct a $\tilde{p}$ that is supported on $\mathbb{R}$ but approximates p up to ϵ error. Let $\epsilon > 0$, define $\tilde{p}$ as

- If $p(x) \geq \Delta$, set $\tilde{p}(x) = p(x)$
- If $0 < p(x) < \Delta$, set $\tilde{p}(x) = \left(1 - \frac{\epsilon}{2}\right) p(x)$
- If $x \in [r_i, l_{i+1}]$ then $\tilde{p}(x) = \frac{\epsilon\gamma}{2n(l_{i+1} - r_i)}$
- If $x \leq l_1$ or $x \geq r_n$, let $\tilde{p}$ be the tail of a Gaussian that integrates to $\frac{\epsilon\gamma}{4n}$

Is $\tilde{p}$ a distribution?

$$\begin{aligned}
\|\tilde{p}\|_1 &= \int_{\{p(x) \geq \Delta\}} p(x) + \int_{\{0 < p(x) < \Delta\}} \left(1 - \frac{\epsilon}{2}\right) p(x) + \int_{\{p(x)=0\}} \tilde{p}(x) \\
&= \left(1 - \frac{\epsilon}{2}\right) + \frac{\epsilon}{2} \int_{\{p(x) \geq \Delta\}} p(x) + \sum_{i=1}^{n-1} \frac{\epsilon\gamma}{2n(l_{i+1} - r_i)} (l_{i+1} - r_i) + 2\left(\frac{\epsilon\gamma}{4n}\right) \\
&= \left(1 - \frac{\epsilon}{2}\right) + \frac{\epsilon}{2}(1 - \gamma) + \frac{\epsilon\gamma(n-1)}{2n} + \frac{\epsilon\gamma}{2n} \\
&= 1.
\end{aligned}$$

Approx Step

Does $\tilde{p}$ approx p in L^1?

$$
\begin{aligned}
\|p - \tilde{p}\|_1 &= \int_{\{p(x) \geq \Delta\}} |p - \tilde{p}| + \int_{\{0 < p(x) < \Delta\}} |p - \tilde{p}| + \int_{\{p(x) = 0\}} |p - \tilde{p}| \\
&= \int_{\{0 < p(x) < \Delta\}} |p - \tilde{p}| + \int_{\{p(x) = 0\}} \tilde{p} \\
&= \frac{\epsilon}{2} \int_{0 < \gamma < \Delta} p + \sum_{j=1}^{n-1} \frac{\epsilon\gamma}{2n(l_{i+1} - r_i)} (l_{i+1} - r_i) + 2\left(\frac{\epsilon\gamma}{4n}\right) \\
&= \frac{\gamma\epsilon}{2} + \frac{\epsilon\gamma(n-1)}{2n} + \frac{\epsilon\gamma}{2n} \\
&= \gamma\epsilon \\
&\leq \epsilon.
\end{aligned}
$$

Exact Planar Flow

- We have $\tilde{p}$, q with $\textbf{supp}\ \tilde{p} = \textbf{supp}\ q = \mathbb{R}$
- Let $\Phi_{\tilde{p}}, \Phi_q$ denote distribution functions of $\tilde{p}$, q (resp)
- Set $f = \Phi_{\tilde{p}}^{-1} \circ \Phi_q$, which is continuous on $\mathbb{R}$
- Via argument from Week 2, f transports q to p
- Let $h(z) = f(z) - z$, then

$$f_{pf}(z) = z + h(1 \cdot z + 0)$$

Proof of Theorem 3.1

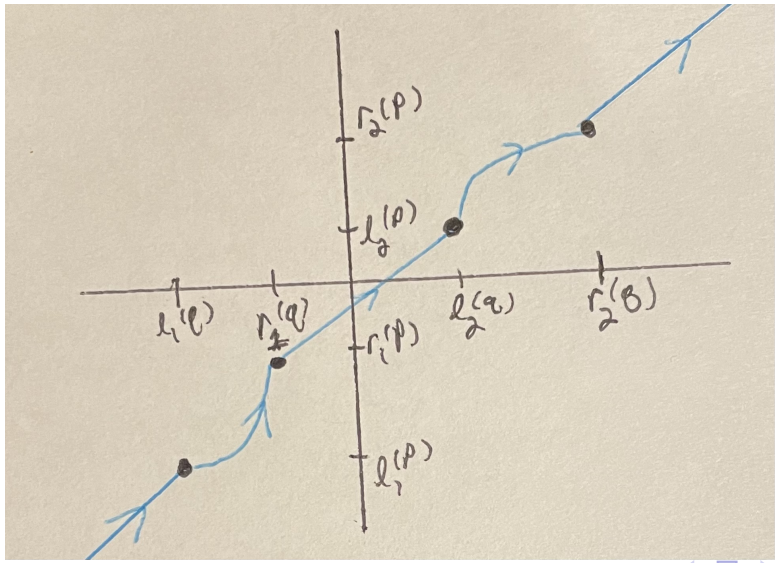
Lemma A.1, Possible Transformations (single flow)

If p and q are densities on R supported on n non-intersecting intervals:

$$\textbf{supp } p = \bigcup_{i=1}^{n} (l_i^{(p)}, r_i^{(p)}) \quad \text{and} \quad \textbf{supp } q = \bigcup_{i=1}^{n} (l_i^{(q)}, r_i^{(q)})$$

and if $\Phi_q(r_i^{(q)}) = \Phi_p(r_i^{(p)})$ for all $i = 1, \ldots, n$, then there exists a planar flow f such that $f_{\#}q = p$, a.e.

Picture proof!



ReLU Universal Approximation

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 $\|f_{\#}q_{\mathcal{N}} - p\|_1 \leq \epsilon$.

Some definitions

Some definitions first:

Definition A.2, Piecewise Distributions in $\mathcal{C}$

Let $\mathcal{C}_0$ be the set of distributions with continuous densities. Suppose $\mathcal{C} \subset \mathcal{C}_0$, then we define $\mathcal{PW}(n, \mathcal{C})$ to be the set of all distributions p on $\mathbb{R}$ satisfying: there exists real numbers $-\infty = t_0 < t_1 < \ldots, < t_{n-1} < t_n = \infty$ such that for any $i = 1, \ldots, n$, on the ith interval $\{t_{i-1}, t_i\}$, p is equal to some distribution $p_i \in \mathcal{C}$. For conciseness, we say p is described by $\{p_{i+1}, t_i\}_{i=0}^{n-1}$. We define $\mathcal{PW}(n) = \mathcal{PW}(n, \mathcal{C}_0)$. If $n' > n$, then $\mathcal{PW}(n) \subset \mathcal{PW}(n')$.

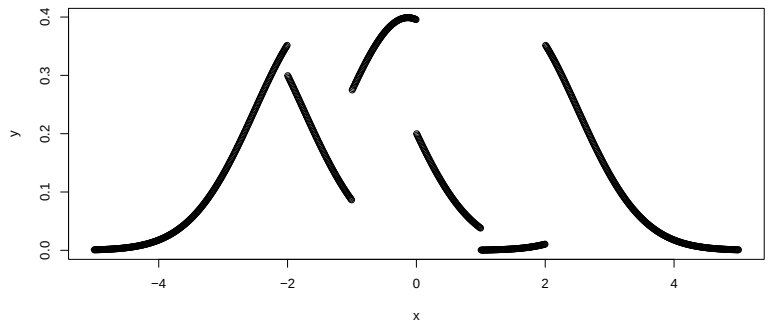


Figure: PW(6,$\mathcal{C}$) where $\mathcal{C}$ is Gaussian probability distributions

Some definitions

Definition A.3, Piecewise Gaussian Distributions

Let $\mathcal{G}$ be the set of Gaussian distributions $\{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma > 0\}$. We define the set of piecewise Gaussian distributions to be $\mathcal{PW}(n, \mathcal{G})$.

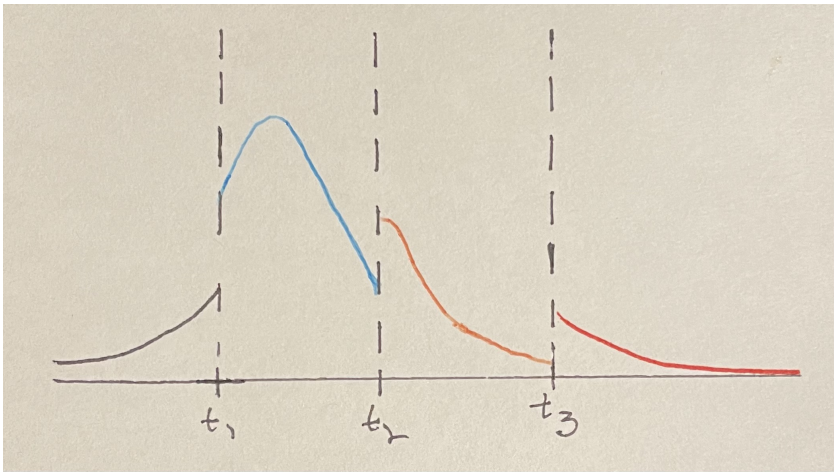


Figure: PW(4,$\mathcal{G}$) where $\mathcal{G}$ is Gaussian probability distributions

Some definitions

Definition A.4, Tail-consistency

Suppose $p \in \mathcal{PW}(n)$ is described by $\{p_i, t_i\}_{i=0}^{n-1}$. We say p is tail-consistent w.r.t. t_k if

$$\sum_{i=1}^{k} \int_{t_{i-1}}^{t_i} p_i(z)dz + \int_{t_k} p_{k+1}(z)dz = 1\,.$$

If p is tail-consistent w.r.t. t_k for any $k = 1, \ldots, n-1$, we say p is tail-consistent.

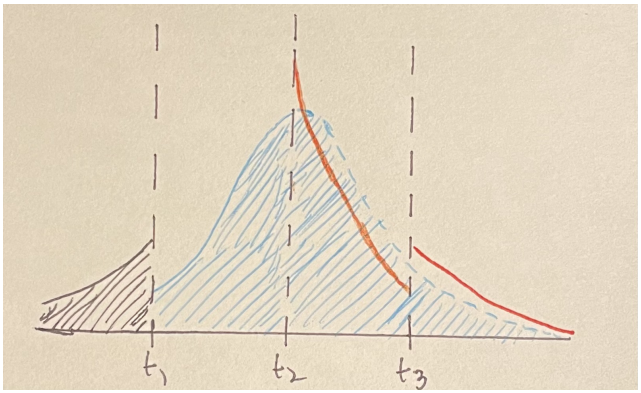


Figure: Tail-consistent PWG w.r.t. t_2

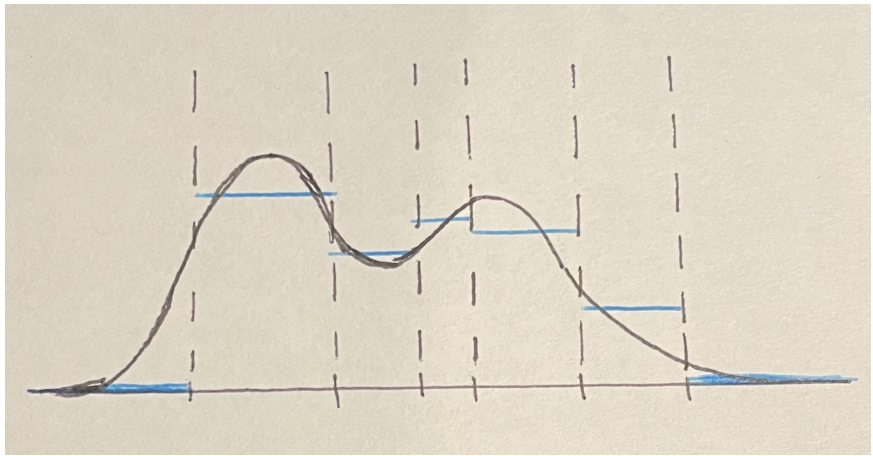
Sketch

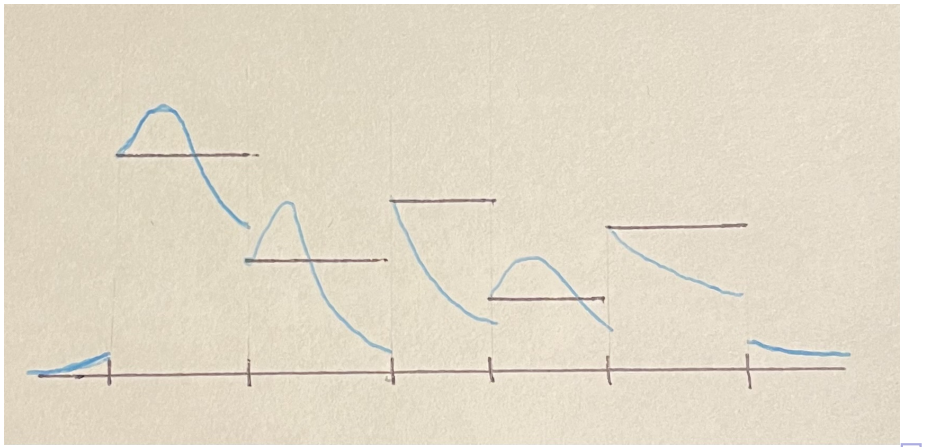
We want f composed of finitely many planar flows such that $\|(f_{\#})q_{\mathcal{N}} - p\|_1 \leq \epsilon$

(1) There exists q_{pwc} such that $\|p - q_{pwc}\|_1 \leq \epsilon/2$ (Lin and Jegelka, 2018)

(2) There exists tail-consistent $q_{pwg} \in \mathcal{PW}(n, \mathcal{G})$ such that $\|q_{pwg} - q_{pwc}\|_1 \leq \epsilon/2$

(3) There exists a flow f as previously described such that $q_{pwg} = (f_{\#})q_{\mathcal{N}}$





Technical Lemma (Inductive Step)

Lemma A.2, Possible Transformations (single flow)

Let $p, q \in \mathcal{PW}(n, \mathcal{G})$ where p is given by $\{p_{i+1}, t_i\}_{i=0}^{n-1}$, q is given by $\{q_{i+1}, t_i\}_{i=0}^{n-1}$ and $p_i = q_i$ for $i < n$. Then there exists a ReLU planar flow f such that $f_{\#}q = p$.

Idea: Select parameters so that flow is constant on $(-\infty, t_{n-1}]$ and shifts/scales appropriately on $[t_n, \infty)$

We have $p(y) = q(y)$ for $y < t_{n-1}$. Now, on $[t_{n-1}, \infty)$ assume $q \sim \mathcal{N}(\mu_n, \sigma_n^2)$ and $p \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$. Let f be a ReLU planar flow

$$f(z) = z + uh(wz + b)$$

with $u = \text{sgn}(\hat{\sigma} - \sigma_n)$, $w = \left|1 - \frac{\hat{\sigma}}{\sigma_n}\right|$, $b = -wt_{n-1}$. Observe

$$f(z) = z + \left(\frac{\hat{\sigma}}{\sigma_n} - 1\right) h(z - t_{n-1}) = \begin{cases} z & \text{if } z \leq t_{n-1} \\ \frac{\hat{\sigma}}{\sigma_n} z + \left(\frac{\hat{\sigma}}{\sigma_n} - 1\right) t_{n-1} & \text{if } z > t_{n-1} \end{cases}$$

Technical Lemma

We compute

$$f^{-1}(y) = \begin{cases} \frac{\sigma_n}{\hat{\sigma}} y - \left(1 - \frac{\sigma_n}{\hat{\sigma}}\right) t_{n-1} & \text{if } y > t_{n-1} \end{cases}$$

which gives

$$\det J_{f^{-1}}(y) = \begin{cases} \frac{\sigma_n}{\hat{\sigma}} & \text{if } y > t_{n-1} \end{cases}$$

For $y > t_{n-1}$

$$\begin{aligned} (f_{\#} q) &= q(f^{-1}(y)) \left| \det J_{f^{-1}}(y) \right| \\ &= \frac{\sigma_n}{\hat{\sigma}} \mathcal{N}\left(\frac{\sigma_n}{\hat{\sigma}} y - \left(1 - \frac{\sigma_n}{\hat{\sigma}}\right) t_{n-1}; \mu_n, \sigma_n^2 \right) \\ &= \mathcal{N}(y; \tilde{\mu}, \hat{\sigma}^2) \end{aligned}$$

We get on $[t_{n+1}, \infty)$ that $f_{\#} q \sim \mathcal{N}(\tilde{\mu}, \hat{\sigma}^2)$ for some $\tilde{\mu}$. But since

$$\int_{t_{n-1}}^{\infty} \mathcal{N}(\tilde{\mu}, \hat{\sigma}^2) = \int_{t_{n-1}}^{\infty} \mathcal{N}(\hat{\mu}, \hat{\sigma}^2),$$

so $\tilde{\mu} = \hat{\mu}$ as desired.

Transport $\mathcal{N}$ to $\mathcal{PW}(n, \mathcal{G})$

Lemma A.3, Possible Transformations (flows)

Let $p \in \mathcal{PW}(n, \mathcal{G})$, if p is tail-consistent, then there exists $n-1$ ReLU planar flows $\{f_t\}_{t=1}^{n-1}$ and a Gaussian distribution $q_{\mathcal{N}}$ such that $(f_{n-1} \circ \cdots \circ f_1)_{\#} q_{\mathcal{N}} = p$.

Proceed by induction on n.
Case $n = 1$: p is a Gaussian, so pick $q_{\mathcal{N}} = p$.
$n \implies (n+1)$ For $p = \{p_{i+1}, t_i\}_{i=0}^{n} \in \mathcal{PW}(n+1, \mathcal{G})$ set $p' = \{p_{i+1}, t_i\}_{i=0}^{n-1} \in \mathcal{PW}(n, \mathcal{G})$ by tail-consistency. IH gives $f_1, \ldots, f_{n-1}$ and $q_{\mathcal{N}}$ such that

$$(f_{n-1} \circ \cdots \circ f_1)_{\#} q_{\mathcal{N}} = p'$$

Notice that p' and p only differ in $[t_n, \infty)$. Using Lemma A.2 gives f_n such that $(f_n)_{\#} p' = p$. Then $\{f_t\}_{t=1}^{n}$ is as desired.

Approx by $\mathcal{PW}(n, \mathcal{G})$

Lemma A.4

Given any piecewise constant distribution q_{pwc} supported on a finite union of compact intervals, for all $\epsilon > 0$ there exists a tail-consistent piecewise Gaussian distribution q_{pwg} such that $\|q_{pwc} - q_{pwg}\|_1 \leq \epsilon$.

Picture construction of piecewise Gaussian approximation

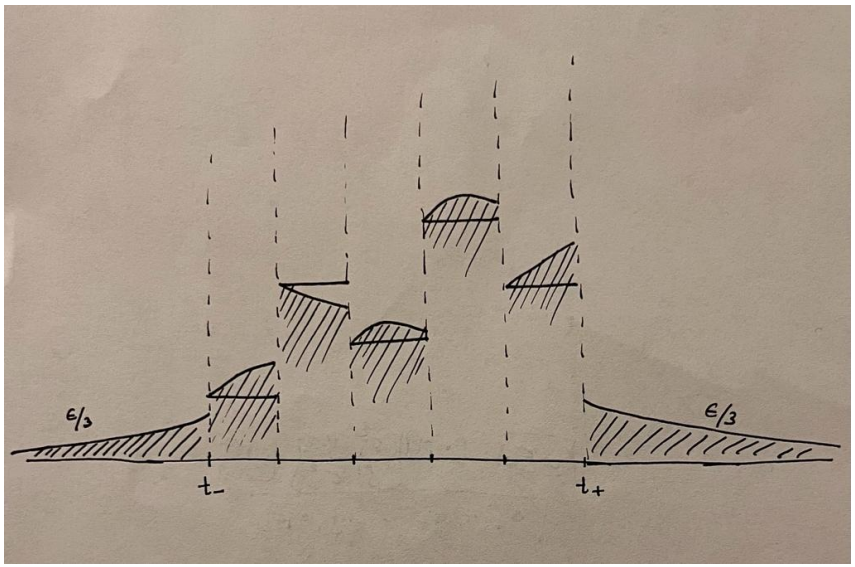


Figure: Piecewise Gaussian approximation of piecewise uniform

Issues with Lemma A.4

- Proof is incorrect as the piecewise-Gaussian approximation of the piece-wise constant density is not "tail consistent".
- Tail consistency condition: consider the ith interval $[t, t + \delta_i]$, the Gaussian $\mathcal{N}(\mu, \sigma^2)$ in this interval should satisfy

$$\int_t^\infty \mathcal{N}(x; \mu, \sigma^2)dx = \left(1 - \frac{2\epsilon}{3}\right) \int_t^\infty q_{pwc}(x)dx$$

- Tail consistency not satisfied at the base case. If $t = t_-$ and $\delta_i = t_+ - t_-$, then

$$\int_{t_-}^\infty \mathcal{N}(x; \mu, \sigma^2)dx = \left(1 - \frac{2\epsilon}{3}\right)$$

- Remedy not immediate
- Lemma A.3 requires tail consistency to construct f such that $f_\# q_\mathcal{N} = p_{pwg}$.

Universal Approximation Results, $d > 1$

Structure of Arguments

Kong and Chaudhuri (2020) prove several non-approximation results using the following structure

(1) Establish a "topology matching condition"

(2) Exploit structure of flow to demonstrate some degenerate Jacobian structure

(3) Use (1) and (2) to show that if $p = f_{\#}q$ then p must be similar to q

Topology Matching Condition

Planar Flow Topology Matching

Suppose distribution q is defined on $\mathbb{R}^d$ and a Sylvester flow on $\mathbb{R}^d$ has tangent matrix B and smooth non-linearity. Let $p = f_{\#}q$. Then for all $z \in \mathbb{R}^d$ we have

$$\nabla_z \log p(f(z)) - \nabla_z \log q(z) \in \text{span}(B).$$

We have $f(z) = z + Ah(B^T z + b)$. Let $\alpha \in \mathbb{R}$ and $w \in \text{span}\{B\}^{\perp}$, then

$$\begin{aligned} f(z + \alpha w) &= z + \alpha w + h(B^T z + \alpha B^T w + b) \\ &= z + \alpha w + h(B^T z + b) \\ &= f(z) + \alpha w. \end{aligned}$$

This gives

$$\det J_f(z) = \det J_f(z + \alpha w)$$

Topology Matching Condition

By change of variable formula for pushfowards,

$$\log p(f(z)) = \log q(z) - \log \det J_f(z)$$
$$\log p(f(z + \alpha w)) = \log(q(z + \alpha w)) - \log \det J_f(z + \alpha w).$$

Subtracting and dividing by α gives

$$\frac{\log p(f(z) + \alpha w) - \log p(f(z))}{\alpha} = \frac{\log q(z + \alpha w) - \log q(z)}{\alpha}$$

Letting $\alpha \to 0$ gives

$$(\nabla_z \log p(f(z))) \cdot w = (\nabla_z \log q(z)) \cdot w$$
$$(\nabla_z \log p(f(z)) - \nabla_z \log q(z)) \cdot w = 0.$$

As this holds for all $w \in \text{span}(B)^\perp$, we have that

$$\nabla_z \log p(f(z)) - \nabla_z \log q(z) \in \left(\text{span}(B)^\perp\right)^\perp = \text{span}(B).$$

Planar Flow cannot map Gaussian to Gaussian

Planar Flow cannot map $\mathcal{N} \to \mathcal{N}$

Let $p \sim \mathcal{N}(0, \Sigma_p)$, $q \sim \mathcal{N}(0, \Sigma_q)$ be two Gaussian distributions on $\mathbb{R}^d$. If there exists a planar flow f on $\mathbb{R}^d$ with smooth non-linearity such that $p = f_{\#}q$, then $\text{rank}(\Sigma_q - \Sigma_p) \leq 1$.

Planar flow has form $f(z) = z + uh(w^T z + b)$, so by topology matching condition

$$\nabla_z \log p(f(z)) - \nabla_z \log q(z) \in \text{span}\{w\}$$

Since $p(z) \propto \exp\left(-\frac{1}{2} z^T \Sigma_p^{-1} z\right)$, using change of variables formula

$$\log p(z) = -\frac{1}{2} z^T \Sigma_p^{-1} z - C$$

and thus

$$\nabla_z \log p(f(z)) = -\Sigma_p^{-1} f(z)$$
$$\nabla_z \log q(z) = -\Sigma_q^{-1} z.$$

For all $w^\perp \in \text{span}\{w\}^\perp$ we have

$$f(z)^T \Sigma_p^{-1} w^\perp = z^T \Sigma_q^{-1} w^\perp$$

Planar Flow cannot map Gaussian to Gaussian

Since $f(z) - z = h(w^T z + b)u$ we get

$$z^T(\Sigma_q^{-1} - \Sigma_p^{-1})w^\perp = h(w^T z + b)u^T \Sigma_p^{-1} w^\perp$$

Setting $z = 0$ gives

$$h(b)u^T \Sigma_p^{-1} w^\perp = 0.$$

Setting $z = w^\perp$ gives

$$\begin{aligned}(w^\perp)^T(\Sigma_q^{-1} - \Sigma_p^{-1})w^\perp &= h(w^T w^\perp + b)u^T \Sigma_p^{-1} w^\perp \\ &= h(b)u^T \Sigma_p^{-1} w^\perp \\ &= 0.\end{aligned}$$

Thus for all $w \in \text{span}\{w\}^\perp$

$$(w^\perp)^T(\Sigma_q^{-1} - \Sigma_p^{-1})w^\perp = 0.$$

Planar Flow cannot map Gaussian to Gaussian

It remains to consider $w^T(\Sigma_q^{-1} - \Sigma_p^{-1})w$

- If $w^T(\Sigma_q^{-1} - \Sigma_p^{-1})w = 0$ then $\Sigma_q^{-1} = \Sigma_p^{-1}$
- If $w^T(\Sigma_q^{-1} - \Sigma_p^{-1})w < 0$, repeat next analysis with $\Sigma_p^{-1} - \Sigma_q^{-1}$
- If $w^T(\Sigma_q^{-1} - \Sigma_p^{-1})w > 0$, then $\Sigma_q^{-1} - \Sigma_p^{-1}$ is PSD and we can write

$$\Sigma_q^{-1} - \Sigma_p^{-1} = Q^T \Lambda Q$$

For all $w^\perp \in \text{span}\{w\}^\perp$

$$\Lambda^{1/2} Q w^\perp = 0,$$

so rank $\Lambda^{1/2} = 1$ and thus rank$(\Sigma_q^{-1} - \Sigma_p^{-1}) = 1$.

Other negative results

Planar ReLU

Let p and q be mixture of Gaussians on $\mathbb{R}^d$. In general, it is impossible to find f composed of finitely many ReLU Slyvester flows such that $p = f_{\#}q$.

Radial Flows

Let $p \sim \mathcal{N}(0, \Sigma_p)$, $q \sim \mathcal{N}(0, \Sigma_q)$. If there is a radial flow on $\mathbb{R}^d$ such that $p = f_{\#}q$ then $\Sigma_q = \Sigma_p$.

Positive Results with Linear Maps

Linear Transformations

For any $A \in \mathbb{R}^{d \times d}$, the linear transformation $g(z) = Az$ can be generated by $(4d - 4)$ ReLU planar flows and d Householder flows.

Proof sketch

- Write $A = LUP$
- P can be written as product of d Householder matrices
- L and U can each be written as product of $(d - 1)$ matrices of form $I + uw^T$
- A matrix of form $I + uw^T$ can be learned with two ReLU planar flows

$$f_1(z) = z + h(w^T z)u$$
$$f_2(z) = z - h(-w^T z)u,$$

so $f_2 \circ f_1(z) = z + uw^T z = (I + uw^T)z$

References

Amos, B., Xu, L., and Kolter, J. Z. (2017). Input convex neural networks. In *International Conference on Machine Learning*, pages 146–155. PMLR.

Kobyzev, I., Prince, S. J., and Brubaker, M. A. (2020). Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979.

Kong, Z. and Chaudhuri, K. (2020). The expressive power of a class of normalizing flow models. *International Conference on Artificial Intelligence and Statistics*, 108:3599–3609.

Lin, H. and Jegelka, S. (2018). Resnet with one-neuron hidden layers is a universal approximator. *Advances in neural information processing systems*, 31.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. (2017). The expressive power of neural networks: A view from the width. *Advances in neural information processing systems*, 30.