Relaxing Bijectivitiy Constraints with Continuously Indexed Normalising Flows

ICML 2020

Rob Cornish, Anthony Caterini, George Deligiannidis, Arnaud Doucet

University of Oxford

July 12-18, 2020

Motivation

The following densities were learned using a Gaussian prior with a 10-layer Residual Flow [Chen et al., 2019] (.5M parameters) trained to convergence.

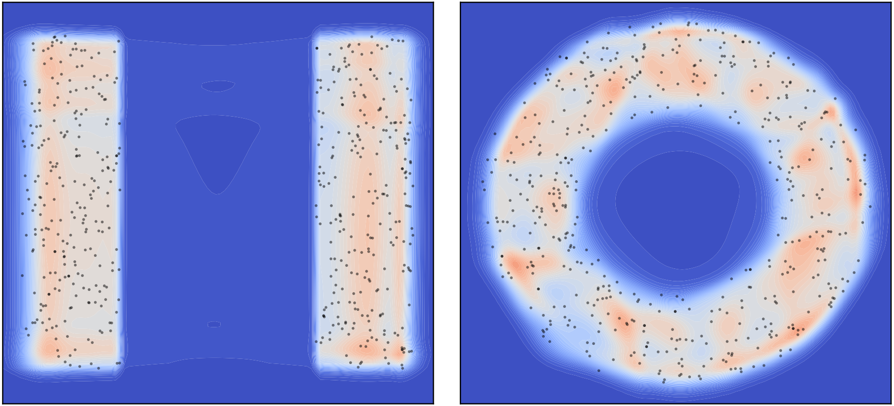


Figure 1: Darker regions indicate lower density. Data shown in black.

Why Does This Occur?

Normalising Flows (NFs) define the following process:

$$Z \sim P_Z, \qquad X := f(Z),$$

where f is a diffeomorphism.

Why Does This Occur?

Normalising Flows (NFs) define the following process:

$$Z \sim P_Z, \qquad X := f(Z),$$

where f is a diffeomorphism.

Hence the support of X will share the same topological properties as the support of Z, i.e.

- Number of connected components
- Number of "holes"
- How they are "knotted"
- etc.

Problem

This suggests a problem when the support of the prior P_Z is simple (e.g. a Gaussian): we usually can't then reproduce the target exactly.

Problem

This suggests a problem when the support of the prior P_Z is simple (e.g. a Gaussian): we usually can't then reproduce the target exactly.

Moreover, to approximate the target closely, our flow must approach non-invertibility.

Our Proposal: Continuously Indexed Flows

Continuously indexed flows (CIFs) instead use the process

$$Z \sim P_Z, \qquad U \mid Z \sim P_{U|Z}(\cdot \mid Z), \qquad X := F(Z; U),$$

where U is a continuous index variable, and each $F(\cdot; u)$ is a normalising flow.

Our Proposal: Continuously Indexed Flows

Continuously indexed flows (CIFs) instead use the process

$$Z \sim P_Z, \qquad U \mid Z \sim P_{U|Z}(\cdot \mid Z), \qquad X := F(Z; U),$$

where U is a continuous index variable, and each $F(\cdot; u)$ is a normalising flow.

Any existing normalising flow can be used to construct F.

Our Proposal: Continuously Indexed Flows

Continuously indexed flows (CIFs) instead use the process

$$Z \sim P_Z, \qquad U \mid Z \sim P_{U|Z}(\cdot \mid Z), \qquad X := F(Z; U),$$

where U is a continuous index variable, and each $F(\cdot; u)$ is a normalising flow.

Any existing normalising flow can be used to construct F.

A continuous index means the density of X is no longer tractable, but can be trained via a natural ELBO objective instead.

Benefits

Intuitively, CIFs can "clean up" mass that would otherwise be misplaced by a single bijection.

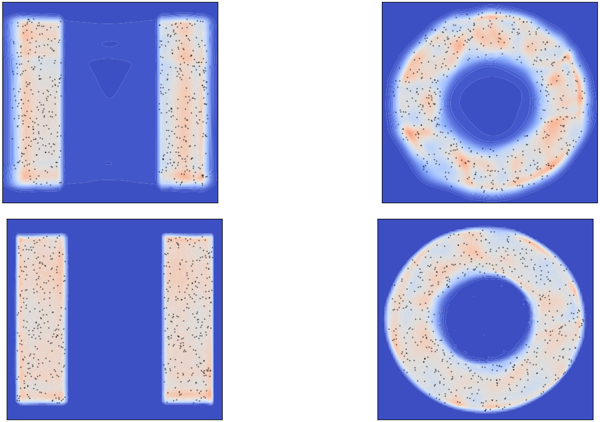


Figure 2: 10-layer Residual Flow (top) and Continuously-Indexed Residual Flow (bottom). Both use .5M parameters.

Going Deeper

What happens when we model a complicated target using a normalising flow?

Theorem: If the prior Z has non-homeomorphic support to a target $X_\star$, then a sequence of flows $f_n(Z) \to X_\star$ in distribution only if

$$\max\left\{\operatorname{Lip} f_n, \operatorname{Lip} f_n^{-1}\right\} \to \infty.$$

Going Deeper

What happens when we model a complicated target using a normalising flow?

Theorem: If the prior Z has non-homeomorphic support to a target $X_\star$, then a sequence of flows $f_n(Z) \to X_\star$ in distribution only if

$$\max\left\{\operatorname{Lip} f_n, \operatorname{Lip} f_n^{-1}\right\} \to \infty.$$

Implications for Residual Flows

For residual flows [Chen et al., 2019],

$$\max\left\{\operatorname{Lip} f_n, \operatorname{Lip} f_n^{-1}\right\} \leq \max\left\{1+\kappa, (1-\kappa)^{-1}\right\}^L < \infty,$$

where $\kappa \in (0, 1)$ is fixed and L is the number of layers.

Hence the previous theorem guarantees we cannot have $f_n(Z) \to X_\star$ in distribution regardless of training time, neural network size, etc.

Implications for Other Flows

For most other flows, $\max\left\{\operatorname{Lip} f_n, \operatorname{Lip} f_n^{-1}\right\}$ is unconstrained [Behrmann et al., 2020].

However, we can still only have $f_n(Z) = X_\star$ exactly if the supports of Z and $X_\star$ are homeomorphic.

It seems reasonable to hope for better performance if we can generalise our model class so that $f_n(Z) = X_\star$ is at least possible.

Continuously Indexed Flows

Recap: Continuously-indexed flows (CIFs) use the process

$$Z \sim P_Z, \qquad U \mid Z \sim P_{U|Z}(\cdot \mid Z), \qquad X := F(Z; U),$$

where U is a continuous index variable, and each $F(\cdot; u)$ is a normalising flow.

Continuously Indexed Flows

Recap: Continuously-indexed flows (CIFs) use the process

$$Z \sim P_Z, \qquad U \mid Z \sim P_{U|Z}(\cdot \mid Z), \qquad X := F(Z; U),$$

where U is a continuous index variable, and each $F(\cdot; u)$ is a normalising flow.

This is compatible with all existing normalising flows: take

$$F(z; u) = f\left(e^{-s(u)} \odot z - t(u)\right).$$

where f is a standard flow.

Multi-layer CIFs

An L-layer CIF is obtained by

$$\begin{aligned} & Z_0 \sim P_{Z_0}, \\ U_1 \sim P_{U_1|Z_0}(\cdot|Z_0), \qquad & Z_1 = F_1(Z_0; U_1), \\ & \cdots \\ U_L \sim P_{U_L|Z_{L-1}}(\cdot|Z_{L-1}), \qquad & X = F_L(Z_{L-1}; U_L). \end{aligned}$$

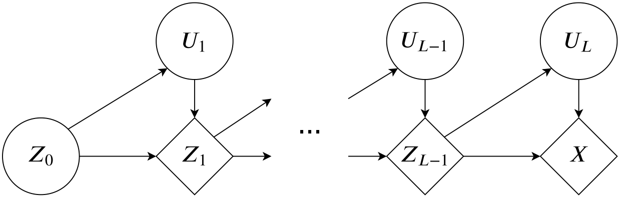


Figure 3: Graphical multi-layer CIF generative model.

Training and inference

The marginal p_X is intractable, but the joint $p_{X,U_{1:L}}$ has a closed-form.

Training and inference

The marginal p_X is intractable, but the joint $p_{X,U_{1:L}}$ has a closed-form.

Given an inference model $q_{U_{1:L}|X}$, we can use the ELBO for training:

$$\mathcal{L}(x) := \mathbb{E}_{u_{1:L} \sim q_{U_{1:L}|X}(\cdot|x)} \left[\log \frac{p_{X,U_{1:L}}(x, u_{1:L})}{q_{U_{1:L}|X}(u_{1:L}|x)} \right] \leq \log p_X(x).$$

Training and inference

The marginal p_X is intractable, but the joint $p_{X,U_{1:L}}$ has a closed-form.

Given an inference model $q_{U_{1:L}|X}$, we can use the ELBO for training:

$$\mathcal{L}(x) := \mathbb{E}_{u_{1:L} \sim q_{U_{1:L}|X}(\cdot|x)} \left[\log \frac{p_{X,U_{1:L}}(x, u_{1:L})}{q_{U_{1:L}|X}(u_{1:L}|x)} \right] \leq \log p_X(x).$$

At test time, we can estimate $\log p_X(x)$ to arbitrary precision using an m-sample IWAE estimate with $m \gg 1$.

Inference model

To obtain an efficient inference model $q_{U_{1:L}|X}$, we exploit the conditional independence structure of $p_{U_{1:L}|X}$ from the forward model:

$$\begin{aligned} & Z_L = X, \\ U_L \sim q_{U_L|Z_L}(\cdot|Z_L), \qquad & Z_{L-1} = F_L^{-1}(Z_L; U_L), \\ & \cdots \\ U_1 \sim q_{U_1|Z_1}(\cdot|Z_1), \qquad & Z_0 = F_1^{-1}(Z_1; U_1). \end{aligned}$$

In other words

$$q_{U_{1:L}|X}(U_{1:L}|X) := \prod_{\ell=1}^{L} q_{U_\ell|Z_\ell}(U_\ell|Z_\ell).$$

Inference model

To obtain an efficient inference model $q_{U_{1:L}|X}$, we exploit the conditional independence structure of $p_{U_{1:L}|X}$ from the forward model:

$$\begin{aligned} & Z_L = X, \\ U_L \sim q_{U_L|Z_L}(\cdot|Z_L), \qquad & Z_{L-1} = F_L^{-1}(Z_L; U_L), \\ & \cdots \\ U_1 \sim q_{U_1|Z_1}(\cdot|Z_1), \qquad & Z_0 = F_1^{-1}(Z_1; U_1). \end{aligned}$$

In other words

$$q_{U_{1:L}|X}(U_{1:L}|X) := \prod_{\ell=1}^{L} q_{U_\ell|Z_\ell}(U_\ell|Z_\ell).$$

This naturally shares weights between the forward and inverse models, since the same F_ℓ are used in both cases.

Intuition

Intuitively, the additional flexibility afforded by $P_{U|Z}$ allows a CIF to "clean up" mass that would be misplaced by a single bijection

Intuition

Intuitively, the additional flexibility afforded by $P_{U|Z}$ allows a CIF to "clean up" mass that would be misplaced by a single bijection

Proposition: Under mild conditions on the target and F, there exists $P_{U|Z}$ such that the model X has the same support as the target.

Intuition

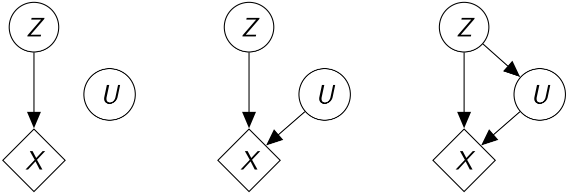
Intuitively, the additional flexibility afforded by $P_{U|Z}$ allows a CIF to "clean up" mass that would be misplaced by a single bijection

Proposition: Under mild conditions on the target and F, there exists $P_{U|Z}$ such that the model X has the same support as the target.

Proposition: If $F(z; \cdot)$ is surjective for each z, there exists $P_{U|Z}$ such that X matches the target distribution exactly.

Comparison with related models

CIFs may be understood as a hybrid between standard normalising flow and VAE density models:



NF VAE CIF

In all cases $X = F(Z; U)$ for some family of bijections F

Experimental Results

Table 1: Test set bits per dimension. Lower is better.

	MNIST	CIFAR-10
ResFlow (small)	1.074	3.474
ResFlow (big)	1.018	3.422
CIF-ResFlow	**0.922**	**3.334**

Note that these ResFlows were smaller than those used by Chen et al. [2019].

Experimental Results

Table 1: Test set bits per dimension. Lower is better.

	MNIST	CIFAR-10
ResFlow (small)	1.074	3.474
ResFlow (big)	1.018	3.422
CIF-ResFlow	**0.922**	**3.334**

Note that these ResFlows were smaller than those used by Chen et al. [2019].

We obtained similar improvements on several other problems and flow models

Experimental Results

Table 1: Test set bits per dimension. Lower is better.

	MNIST	CIFAR-10
ResFlow (small)	1.074	3.474
ResFlow (big)	1.018	3.422
CIF-ResFlow	**0.922**	**3.334**

Note that these ResFlows were smaller than those used by Chen et al. [2019].

We obtained similar improvements on several other problems and flow models

Thank you!



Figure 4: Joint work with Anthony Caterini, George Deligiannidis, and Arnaud Doucet

References

Rob Cornish, Anthony L Caterini, George Deligiannidis, and Arnaud Doucet. Relaxing bijectivity constraints with continuously-indexed normalising flows. In *International Conference on Machine Learning*, 2020.

Tian Qi Chen, Jens Behrmann, David K Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. In *Advances in Neural Information Processing Systems*, pages 9913–9923, 2019.

Jens Behrmann, Paul Vicol, Kuan-Chieh Wang, Roger B. Grosse, and Jörn-Henrik Jacobsen. On the invertibility of invertible neural networks, 2020.