Application of Uniformly Scaling Flows

Faried Abu Zaid [□△1] Daniel Neider [□△23] Emmanuel Müller [△23]
Mustafa Yalciner [□23] Tim Katzke [△23] Giorgia Pitteri [△4]

[1] appliedAI Institute for Europe [2] TU Dortmund [3] Center for Trustworthy Data Science and Security [4] appliedAI Initiative for Applied Artificial Intelligence

appliedAI institute for europe TransferLab technische universität dortmund

Uniformly Scaling Flows

tl;dr: Uniformly scaling flows with p-radial monotonic base distributions "linearize" typicality.

A **probabilistic flow model** is a diffeomorphism $f : \mathbb{R}^d \to \mathbb{R}^d$ that transforms a (usually simple) base distribution B into a (usually more complex) data distribution D, i.e. $D = f(B)$ and

$$p_D(f) = p_B(f^{-1}(x)) \left|\det \frac{\partial f^{-1}}{\partial x}\right|.$$

We say that a flow f is **uniformly scaling**, if it has constant Jacobian determinant, i.e. there exists some $c \in \mathbb{R}_{>0}$ such that $\left|\det \frac{\partial f}{\partial x}\right| = c$ for all $x \in \mathbb{R}^d$.
Uniformly scaling flows are long known. In fact, the arguably first normalizing flow architecture NICE is uniformly scaling [1]. However, they have some intriguing properties that haven't been leveraged much to the best of our knowledge.

p-Radial Base Distributions

We call an absolutely continuous distribution B **p-radial**, if there is a function $g : \mathbb{R}_+ \to \mathbb{R}_+$ such that $p_B(x) = g(|x|_p)$. If g is additionally strictly monotonically decreasing, then we say that B is p-radial monotonic.

Radial distributions can be defined by starting from the distribution of the p-norm. Let ρ be a probability density on $\mathbb{R}_+$. We call $R_{\rho,p,d}$ the p-radial distribution over $\mathbb{R}^d$ with p-norm distribution ρ, which is given by the pdf $p_{R_{\rho,p,d}}(x) = \rho(|x|_p) \left|\frac{\partial V_p^d(r)}{\partial r}(|x|_p)\right|^{-1}$, where $V_p^d(r)$ = volume of d-dim. L^p ball of radius r

An **upper density level set** (UDL) w.r.t. B is a set that can be written as $\{x \mid p_B(x) > t\}$ for some $t \in \mathbb{R}$. We denote UDL of probability q by $\mathrm{UDL}_B(q)$.

The following observation is crucial for our applications. If B is p-radial monotonic, then by choosing $r(q) = \mathrm{quantile}_{|B|_p}(q)$ we obtain that $\mathrm{UDL}_B(q) = \mathbb{B}_p^d(r(q))$, i.e. upper density level sets are L^p-balls.

Linearizing Typicality

If f is a uniformly scaling flow on $\mathbb{R}^d$, then f preserves density level sets. Hence, if B is a p-radial monotonic distribution over $\mathbb{R}^d$, then there is a function $r : [0, 1) \to \mathbb{R}_+$ such that

$$\mathrm{UDL}_{f(B)}(q) = f\left(\mathbb{B}_p^d(r(q))\right).$$

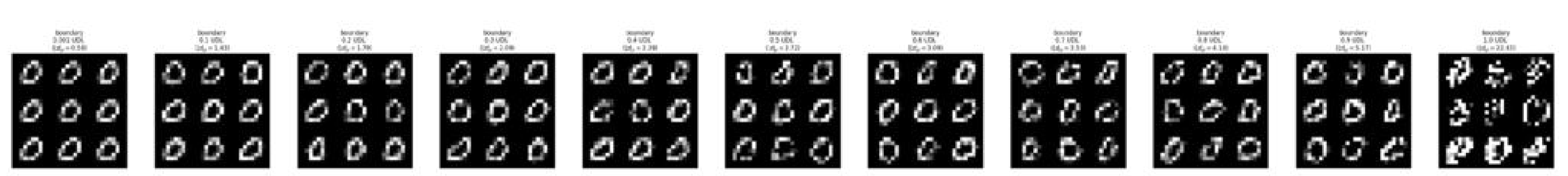


Figure 1. Samples from the boundary of UDLs with a given probability.

Neuro-Symbolic Verification□

tl;dr: Verification on density level sets via SMT and abstract interpretation through u.s. flows.

Formal verification has emerged as a promising method to ensure the safety and reliability of neural networks [2]. Currently, the two major approaches are satisfiability modulo theory (SMT) and abstract interpretation (A.I.). Naively verifying a property on the entire input space implies that the safety of the neural network is checked even for inputs that do not occur in the real-world and have no meaning at all, often resulting in spurious errors.

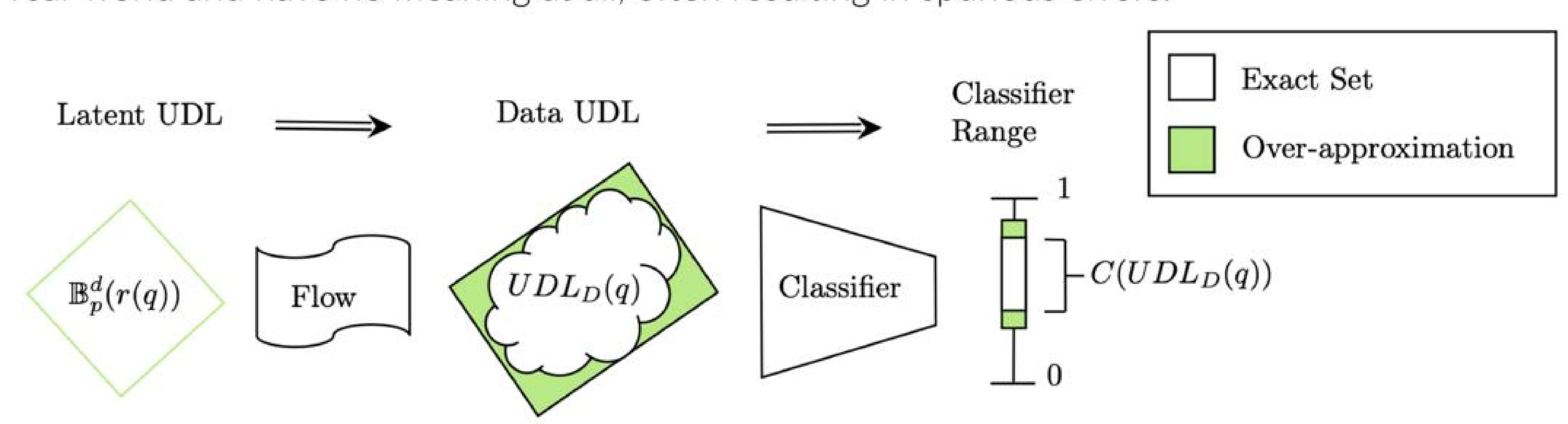


Figure 2. Verifying a classier via abstract interpretation on an UDL of a uniformly scaling flow model.

VeriFlow

Let f be a network that is purely build from the layer types (masked) additive coupling, additive autoregression, masked additive convolution, and LU layers. If the first three layer types only use piece-wise affine conditioning networks, then f is a uniformly scaling piece-wise affine flow. In particular, any density p_D defined by f has the following properties:

1. If $\log p_B(\cdot)$ is piece-wise affine, then $\log p_D$ is piece-wise affine.
2. For any p-radial monotonic base distribution B there is a function $r : [0, 1) \to \mathbb{R}_+$ such that $\mathrm{UDL}_{f(B)}(q) = f(\mathbb{B}_p^d(r(q)))$.
3. Computing log-densities has the same computational complexity as sampling.

Figure 3. Instances of low classification confidence. Found without (left) and with (right) leveraging a flow model.

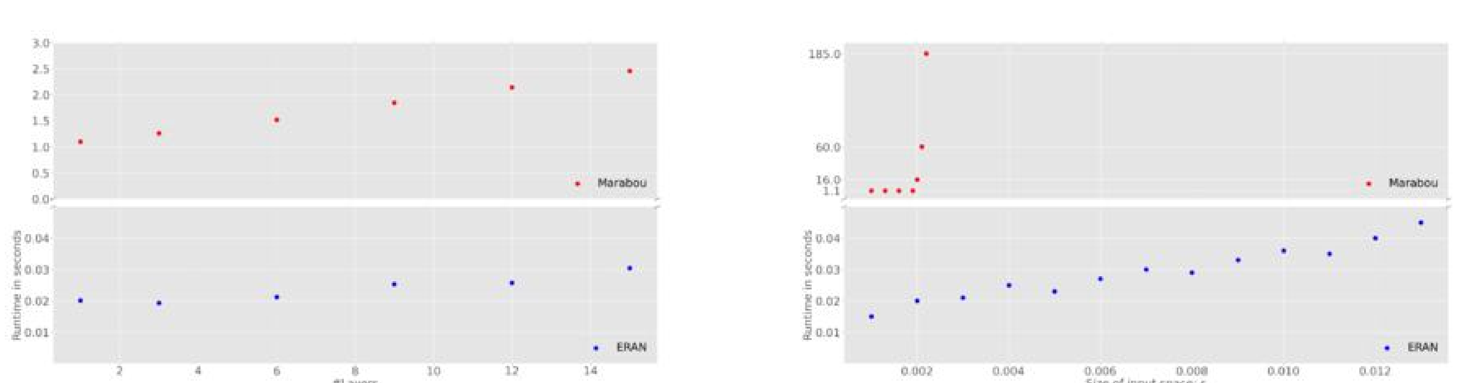


Figure 4. Verification runtime for different network depth and search space size with Marabou (SMT) and Eran (A.I.).

Anomaly Detection△

tl;dr: Uniformly scaling flows are also "better" DeepSVDDs. Ongoing research.

DeepSVDD is a popular anomaly detection method, where we learn to map the data into a hyper-sphere in a latent space using an arbitrary neural network [3]. Uniformly scaling flows turn out to be quite directly related to this class of models. Also, they bear certain advantages over the use of arbitrary networks.

Comparing DeepSVDD and Uniformly Scaling Flows

The objective of a DeepSVDD, $\min_\theta \mathbb{E}_X \left[|g_\theta(x) - c|^2\right] + \lambda|\theta|^2$, has pathological optimal solutions (e.g. for $c = 0$, simply $\theta = 0$). As ad-hoc counter measure, one removes all bias terms and sets $c \neq 0$ in practice.

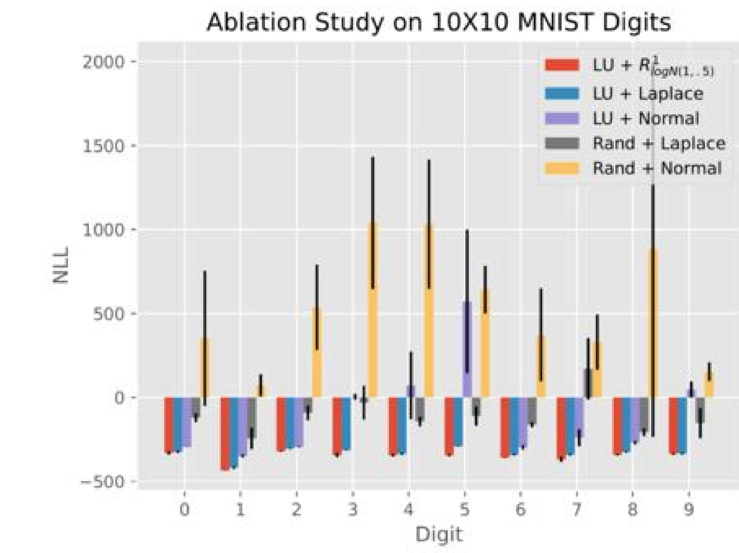
Uniformly scaling flows can be seen as DeepSVDD variant with a more principled optimization objective. As one can compute for $B = \mathcal{N}\left(c, \frac{1}{2}I\right)$ the objective of the flow with a Bayesian prior on the parameters,

$$\min_\theta \mathbb{E}_X\left[-\log p_D(x \mid \theta)p_{\text{prior}}(\theta)\right] = \min_\theta \mathbb{E}_X\left[|f_\theta^{-1}(x) - c|^2\right] + \psi_{\det}(\theta) + \psi_{prior}(\theta)$$
$$= \min_\theta \mathrm{KL}\left(f_\theta^{-1}(X) \mid \mathcal{N}\left(c, \frac{1}{2}I\right)\right) + \psi_{\text{prior}}(\theta)$$

is a variant of the DeepSVDD objective with a different "regularization" term. However, this variation ensures that a loss of 0 can only be achieved if $f^{-1}(X) = B$.

However, exploding determinants is a related pathology of flow models that can still occur. We propose to assume a symmetrized log-normal prior on the diagonal entries of LU-layers in our architecture, which we show to induce a log-normal prior on the determinant of the flow:

$$p\left(\left|\det \frac{\partial LUx + b}{\partial x}\right|\right) = p\left(e^{\sum_{i=1}^d \ln|u_{ii}|}\right)$$



Benchmark		
Dataset	Base Distribution	NLL
MNIST	Normal	-1679.132
	Laplace	-1879.869
	$\mathbf{R}_{\log N(1,.5),1,784}$	**-2013.079**
Circles	Normal	-0.910
	Laplace	**-1.113**
	$R_{\log N(1,.5),1,2}$	-0.604
Moons	Normal	-1.899
	Laplace	**-2.279**
	$R_{\log N(1,.5),1,2}$	-1.084
Blobs	**Normal**	**2.287**
	Laplace	2.500
	$R_{\log N(1,.5),1,2}$	2.596

References

[1] Dinh, L., Krueger, D., and Bengio, Y. NICE: Non-linear Independent Components Estimation. In Bengio, Y. and LeCun, Y., (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, .

[2] Xie, X., Kersting, K., and Neider, D. Neuro-Symbolic Verification of Deep Neural Networks. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence* International Joint Conferences on Artificial Intelligence.

[3] Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. Deep One-Class Classification. In *Proceedings of the 35th International Conference on Machine Learning* PMLR pp. 4393–4402.