HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference

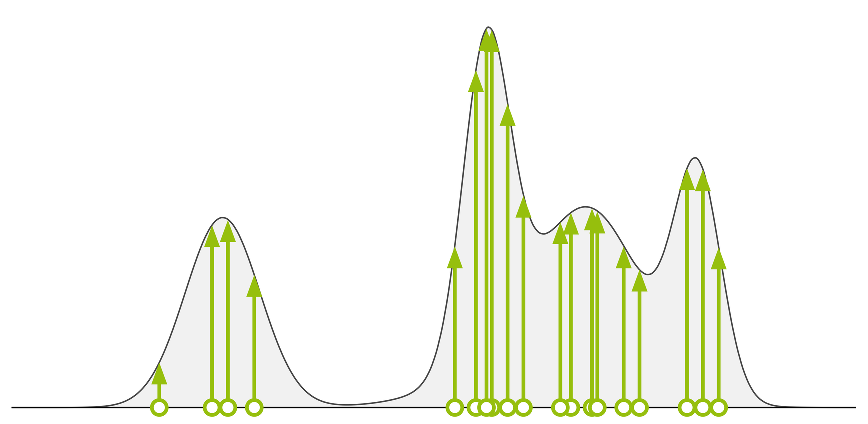
Jakob Kruse[1], Gianluca Detommaso[23], Robert Scheichl[12], Ullrich Köthe[1]

[1]Heidelberg University [2]University of Bath [3] amazon.com Contact: jakob.kruse@iwr.uni-heidelberg.de

Density Estimation

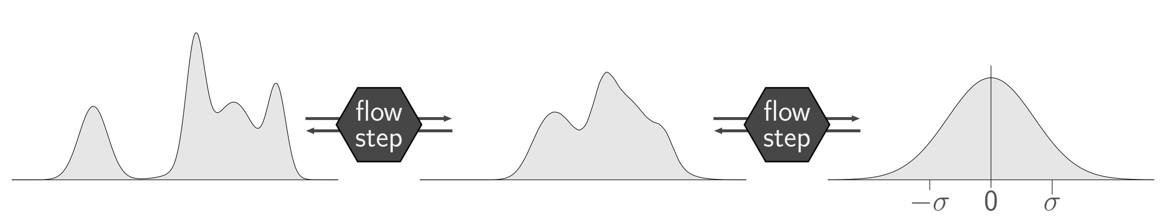
Given a sample of data points $\mathbf{x}$ that all stem from the same, unknown source distribution $p^*(\mathbf{x})$,

model the underlying density $p(\mathbf{x})$ to gain insights into the data $\mathbf{x}$ or generate new samples:



Transport and Normalizing Flow

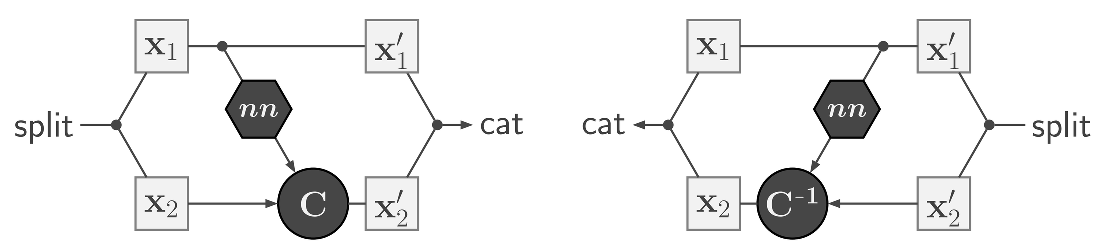
- **Transport** bijective mapping between probability densities



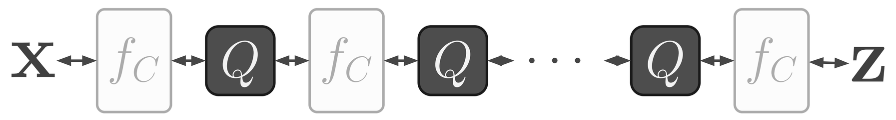
- **Normalizing Flow** transport from $p(\mathbf{x})$ to standard normal latent distribution $\mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$ as a sequence of neural steps [1]

$$\mathcal{L}(\mathbf{x}) = \tfrac{1}{2}\|T(\mathbf{x})\|_2^2 - \log|\mathbf{J}_T(\mathbf{x})|$$

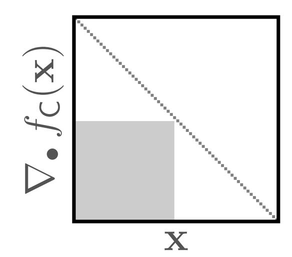
- **Split Coupling** fast inversion and Jacobian determinant [2]



- **Affine Coupling** take $\mathbf{C} = \mathbf{x}_2 \odot \exp\big(\mathrm{nn}_s(\mathbf{x}_1)\big) + \mathrm{nn}_t(\mathbf{x}_1)$ [3]
- **Varying Splits** ensured by orthogonal matrices, $\mathbf{Q}^{-1} = \mathbf{Q}^\top$



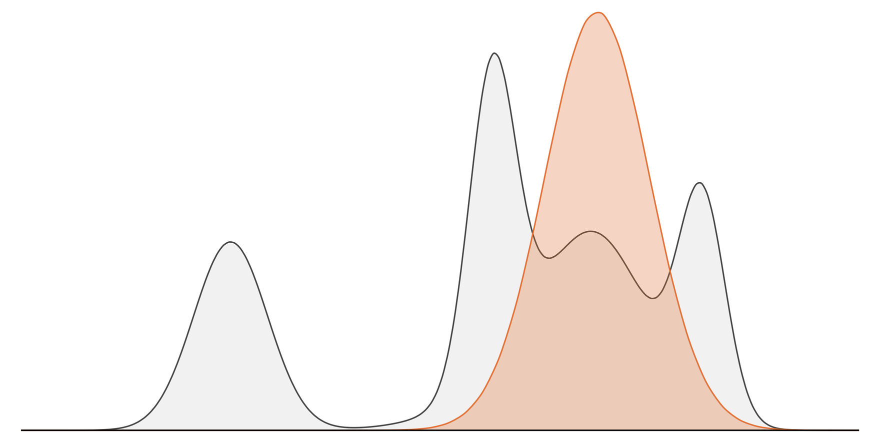
- **Jacobian Matrix** lower triangular [3], but high sparsity means reduced expressive power because many variable interactions are not modelled



Bayesian Inference

Given a sample of data points $\mathbf{x}$ with potentially ambiguous and uncertain attributes $\mathbf{y}$, in this example their color,

model the posterior density $p(\mathbf{x}\,|\,\mathbf{y})$ of data $\mathbf{x}$ given a partial observation $\mathbf{y}$, in this example $p(\mathbf{x}\,|\,\mathbf{y} = \text{orange})$:

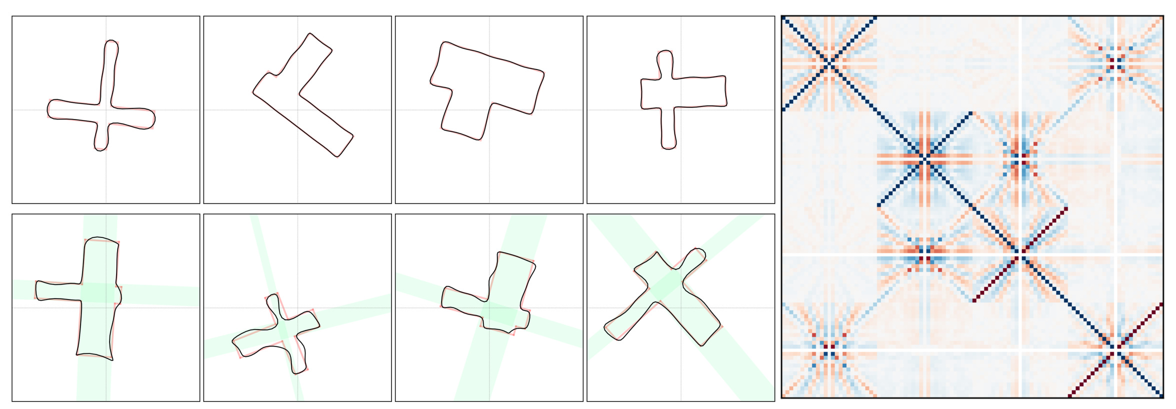


Fourier Curve Data Set

We can parameterize a closed 2d curve in terms of $2M+1$ complex, 2d Fourier coefficients $\mathbf{a}_m$, based on a set of points $\mathbf{p}_n$:

$$\mathbf{g}(t) = \sum_{m=-M}^{M} \mathbf{a}_m \cdot e^{2\pi \cdot i \cdot m \cdot t} \quad [5]$$

- **flexible** number of dimensions and **easy to scale up**
- **complicated correlation** structure between dimensions
- exact ground truth and **quantitative metrics** for samples
- natural **representation in 2d** enables visual inspection

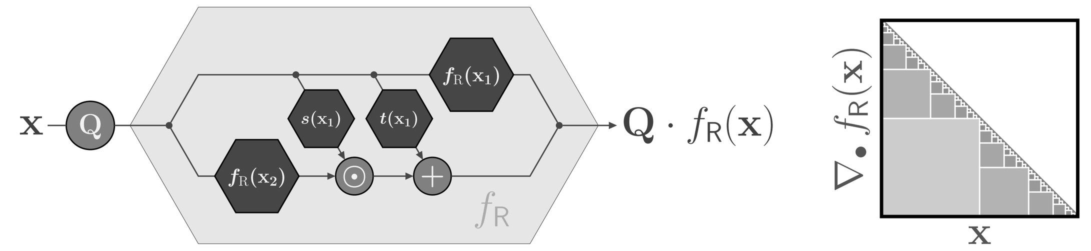


Cross shapes with random angles, bar lengths/widths/shifts and positions; parameterized by vectors $\mathbf{a} \in \mathbb{C}^{25\times 2} \to \mathbf{x} \in \mathbb{R}^{100}$

Recursive Coupling

We extend the split coupling design by making it recursive:

- **Recursion** one sub-block for each lane in a block



- **Stopping Criterion** when no further splits are possible

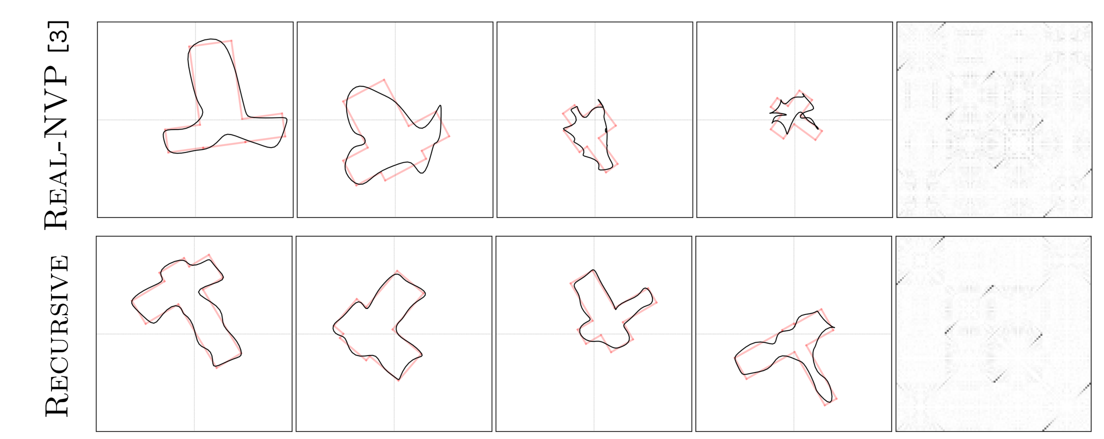
$$f_{\mathrm{R},k}(\mathbf{x}) = \begin{cases} f_{\mathrm{C}}(\mathbf{x}), & \text{if } N_k \le 3, \\ \begin{bmatrix} f_{\mathrm{R},k\text{-}1}(\mathbf{x}_1) \\ C_k\left(f_{\mathrm{R},k\text{-}1}(\mathbf{x}_2)\,\middle|\,\mathbf{x}_1\right) \end{bmatrix} & \text{otherwise} \end{cases}$$

- **Jacobian Matrix** becomes dense lower triangular
- **Train Loss** no change, $\log|\mathbf{J}_T(\mathbf{x})| = \sum$ all diagonal entries
- **Complexity** number of OPs decreases with recursion depth, all subnetworks within one block can be evaluated in parallel
- **Results On UCI Benchmarks** [4] normal REAL-NVP [3] versus our RECURSIVE model at equal parameter budget, in terms of *log-likelihood* on test set (higher is better)

BLOCKS		DIM	REAL-NVP [3]	RECURSIVE
4	POWER	6	-0.054 ± 0.017	$\mathbf{-0.027 \pm 0.018}$
	GAS	8	7.620 ± 0.136	$\mathbf{7.662 \pm 0.094}$
	MINIBOONE	42	-19.296 ± 0.395	$\mathbf{-14.547 \pm 0.164}$
8	POWER	6	$\mathbf{0.093 \pm 0.002}$	0.080 ± 0.007
	GAS	8	8.062 ± 0.177	$\mathbf{8.137 \pm 0.055}$
	MINIBOONE	42	-16.625 ± 0.119	$\mathbf{-14.117 \pm 0.163}$

→ positive effect of recursion increases with higher DIM

- **Results On Fourier Curve Data Set** generated samples and error in $\mathbf{x}$ correlation matrices (more white is better):



Quantitative results in terms of *log-likelihood*, *intersection-over-union* and average *Hausdorff distance*:

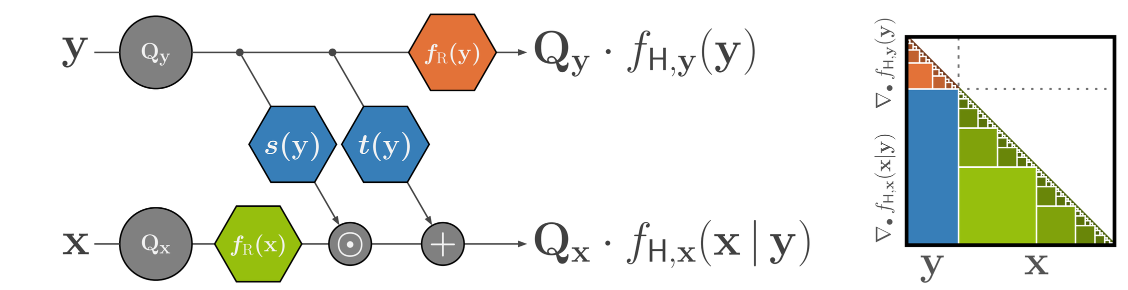
BLOCKS	4× RNVP [3]	4× REC	8× RNVP [3]	8× REC
LL ↑	3.419	**3.627**	3.329	**3.637**
IoU ↑	0.594	**0.823**	0.588	**0.823**
H-DIST ↓	0.134	**0.077**	0.138	**0.073**

→ **Recursive design works better in all 3 metrics**

Hierarchical Architecture – HINT

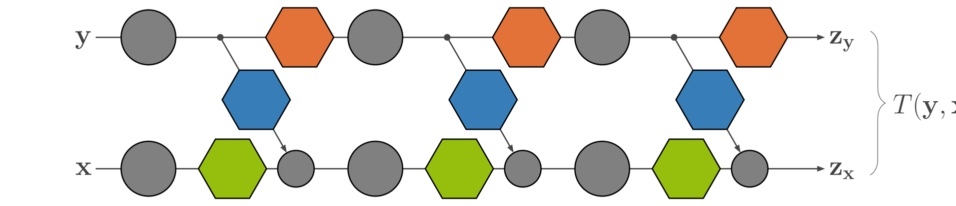
For paired data $\mathbf{x}$ and observations $\mathbf{y}$, the top level split in the recursion can represent their separation:

- **Dependency** of $\mathbf{x}$-lane on $\mathbf{y}$-lane, but not vice versa



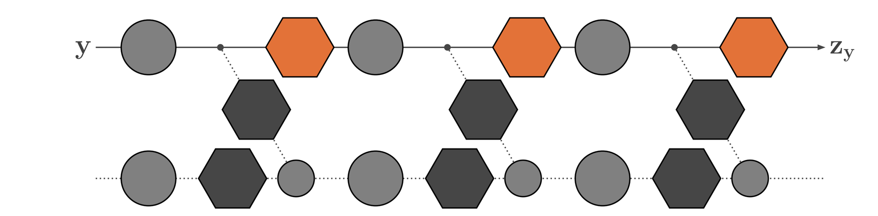
→ ensured by keeping shuffling matrices $\mathbf{Q}$ separate

- **Training** same as before, as one joint network $T(\mathbf{y}, \mathbf{x})$

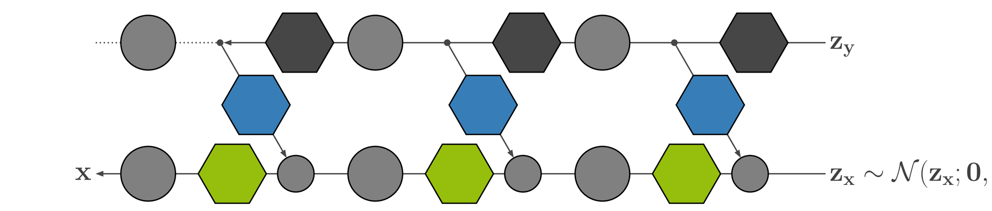


$$\mathcal{L}(\mathbf{y}, \mathbf{x}) = \tfrac{1}{2}\|T(\mathbf{y}, \mathbf{x})\|_2^2 - \log|\mathbf{J}_T(\mathbf{y}, \mathbf{x})|$$

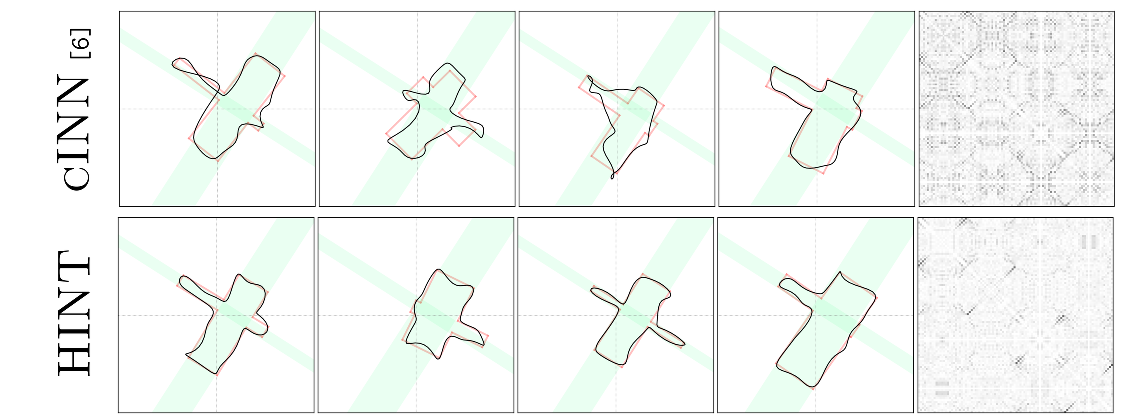
- **Sampling** by first running only $\mathbf{y}$-lane to get $\mathbf{z}_y$



then holding $\mathbf{z}_y$ fixed and running inverse with Gaussian $\mathbf{z}_x$



- **Results On Fourier Curve Data Set** samples $\mathbf{x}$ conditioned on $\mathbf{y}$ = (angle, center, thickness ratio); and corr. matrix errors:



Quantitative results in terms of *log-likelihood*, *intersection-over-union* and average *Hausdorff distance*:

BLOCKS	4× cINN [6]	4× HINT	8× cINN [6]	8× HINT
LL ↑	3.625	**3.724**	3.609	**3.766**
IoU ↑	0.654	**0.843**	0.590	**0.859**
H-DIST ↓	0.116	**0.073**	0.119	**0.066**

→ **Hierarchical transport better in all 3 metrics**

[1] Rezende and Mohamed. **Variational inference with normalizing flows.** ICML 2015.
[2] Dinh et al. **NICE: Non-linear Independent Components Estimation.** ICLR 2015.
[3] Dinh et al. **Density estimation using Real NVP.** ICLR 2017.
[4] Dua and Graff. **UCI Machine Learning Repository.** 2017. http://archive.ics.uci.edu/ml.
[5] McGarva and Mullineux. **Harmonic representation of closed curves.** Appl. Math. Model. 1993
[6] Ardizzone et al. **Conditional Invertible Neural Networks for Diverse Image-to-Image Translation.** GCPR 2020.

<CODE>
github.com/VLL-HD/HINT

<DATA SETS>
github.com/VLL-HD/inn_toy_data

This project has received funding from the **European Research Council** (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 647769). J. Kruse was supported by **Informatics for Life** funded by the Klaus Tschira Foundation. G. Detommaso was supported by the **EPSRC Centre** for Doctoral Training in Statistical Applied Mathematics at Bath (EP/L015684/1).