A. ResMADE

Residual connections (He et al., 2016a) are widely used in deep neural networks, and have demonstrated favourable performance relative to standard networks. Residual networks typically consist of many stacked transformations of the form

$$\mathbf{h}_{l+1} = \mathbf{h}_l + \mathbf{f}(\mathbf{h}_l), \tag{13}$$

where $\mathbf{f}$ is a residual block. In this work, we observe that it is possible to equip a MADE (Germain et al., 2015) with residual connections when certain conditions on its masking structure are met.

At initialization, each unit in each hidden layer of a MADE is assigned a positive integer, termed its degree. The degree specifies the number of input dimensions to which that particular unit is connected. For example, writing d_k^l for unit k of layer l, a value $d_k^l = m$ means that the unit depends only on the first m dimensions of the input, when the input is prescribed a particular ordering (we always assume the ordering given by the data). In successive layers, this means that a unit may only be connected to units in the previous layer whose degrees strictly do not exceed its own. In other words, the degree assignment defines a binary mask matrix which multiplies the weight matrix of each layer in a MADE elementwise, maintaining autoregressive structure. Indeed, the mask M for layer l is given in terms of the degrees for layer $l-1$ and layer l:

$$M_{ij}^l = \begin{cases} 1 & \text{if } d_i^l \geq d_j^{l-1} \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

Through this masking process, it can be guaranteed that output units associated with data dimension d only depend on the previous inputs $\mathbf{x}_{<d}$. Though the original MADE paper considered a number of ways in which to assign degrees, we focus here on fixed sequential degree assignment, used also by MAF (Papamakarios et al., 2017). For an input of dimension D, we define the degree

$$d_k^l = (k-1) \pmod{(D-1)} + 1. \tag{15}$$

We also assume the dimension H of each hidden layer is at least D, so that no input information is lost. The result of sequential degree assignment for $D = 3$ and $H = 4$ is illustrated in fig. 7.

If all hidden layers in the MADE are of the same dimensionality H, sequential degree assignment means that each layer will also have the same degree structure. In this way, two vectors of hidden units in the MADE may be combined using any binary element-wise operation, also shown in Figure fig. 7. In particular, a vector computed from a fully-connected block with masked layers can be added to the input to that block while maintaining the same autoregressive structure, allowing for the traditional residual connection to be added to the MADE architecture.

Not only do residual connections enhance a MADE in its own right, but the resulting ResMADE architecture can be used as a drop-in replacement wherever a MADE is used as a building block, such as IAF, MAF, or NAF (Kingma et al., 2016; Papamakarios et al., 2017; Huang et al., 2018).

B. Experimental settings

In all experiments, we use the same ResMADE and ENN architectures, each with 4 pre-activation residual blocks (He et al., 2016b). The number of hidden units and the dimensionality of the context vector for the ENN are fixed across all tasks at 128 and 64 respectively. The number of hidden units in the ResMADE is tuned per experiment. For the exact experimental settings used in each experiment, see Tables 3 and 4.

For the proposal distributions, we use a mixture of Gaussians in all cases except for the checkerboard experiment, where we use a fixed uniform distribution. We use 10 mixture components for the synthetic experiments, and 20 components for all other experiments. We use a minimum scale of 10^{-3} for the Gaussian distributions in order to prevent numerical issues.

For optimization we use Adam (Kingma & Ba, 2014) with a cosine annealing schedule (Loshchilov & Hutter, 2016) and an initial learning rate of 5×10^{-4}. The number of training steps is adjusted per task. For Miniboone, we use only 6000 training updates, as we found overfitting to be a significant problem for this dataset. Early-stopping is used to select models, although in most cases the best models are obtained at the end of training. Normalizing constants are estimated using 20 importance samples, and dropout is applied to counter overfitting, with a rate that is tuned per task. Dropout is applied between the two layers of each residual block in both the ENN and ResMADE.

For the VAE, we use an architecture similar to those used in IAF and NAF (Kingma et al., 2016; Huang et al., 2018). For the encoder, we alternate between residual blocks that preserve spatial resolution, and downsampling residual blocks with strided convolutions. The input is projected to 16 channels using a 1×1 convolution, and the number of channels is doubled at each downsampling layer. After three downsampling layers we flatten the feature maps and use a linear layer to regress the means and log-variances of the approximate posterior with 32 latent units. The decoder mirrors the encoder, with the input latents being linearly projected and reshaped to a $[4, 4, 128]$ spatial block, which is then upsampled using transpose-convolutions in order to output pixel-wise Bernoulli logits.

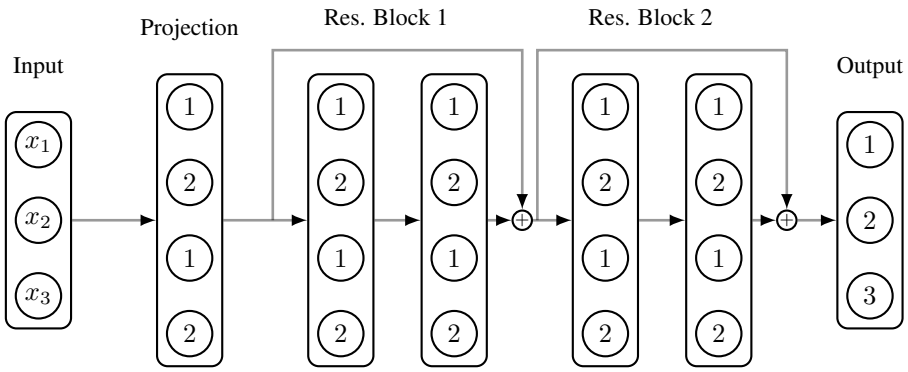

Figure 7: ResMADE architecture with $D = 3$ input data dimensions and $H = 4$ hidden units. The degree of each hidden unit and output is indicated with an integer label. Sequential degree assignment results in each hidden layer having the same masking structure, here alternating between dependence on the first input, or the first two inputs. These layers can be combined using any binary elementwise operation, while preserving autoregressive structure. In particular, residual connections can be added in a straightforward manner.The ResMADE architecture consists of an initial masked projection to the target hidden dimensionality, a sequence of masked residual blocks, and finally a masked linear layer to the output units.

Table 3: Experimental setting for synthetic data and VAE experiments.

HYPERPARAMETER	SPIRALS	CHECKERBOARD	DIAMOND	EINSTEIN	VAE
BATCH SIZE	256	256	256	256	256
RESMADE HIDDEN DIM.	256	256	256	256	512
RESMADE ACTIVATION	RELU	RELU	RELU	RELU	RELU
RESMADE DROPOUT	0	0	0	0	0.5
CONTEXT DIM.	64	64	64	64	64
ENN HIDDEN DIM.	128	128	128	128	128
ENN ACTIVATION	RELU	RELU	RELU	RELU	RELU
ENN DROPOUT	0	0	0	0	0.5
MIXTURE COMPS.	10	-	10	10	20
MIXTURE COMP. SCALE MIN.	1E-3	-	1E-3	1E-3	1E-3
LEARNING RATE	5E-4	5E-4	5E-4	5E-4	5E-4
TOTAL STEPS	400000	400000	400000	3000000	100000
WARM-UP STEPS	5000	0	0	0	0

Table 4: Experimental settings for UCI and BSDS300 datasets.

HYPERPARAMETER	POWER	GAS	HEPMASS	MINIBOONE	BSDS300
BATCH SIZE	512	512	512	512	512
RESMADE HIDDEN DIM.	512	512	512	512	1024
RESMADE ACTIVATION	RELU	RELU	RELU	RELU	RELU
RESMADE DROPOUT	0.1	0	0.2	0.5	0.2
CONTEXT DIM.	64	64	64	64	64
ENN HIDDEN DIM.	128	128	128	128	128
ENN ACTIVATION	RELU	TANH	RELU	RELU	RELU
ENN DROPOUT	0.1	0	0.2	0.5	0.2
MIXTURE COMPS.	20	20	20	20	20
MIXTURE COMP. SCALE MIN.	1E-3	1E-3	1E-3	1E-3	1E-3
LEARNING RATE	5E-4	5E-4	5E-4	5E-4	5E-4
TOTAL STEPS	800000	400000	400000	6000	400000
WARM-UP STEPS	5000	5000	5000	0	5000