MADE

by Mathieu Germain, Karol Gregor, Iain Murray, Hugo Larochelle

Masked Autoencoder for Distribution Estimation

July 9th 2015

UNIVERSITÉ DE SHERBROOKE

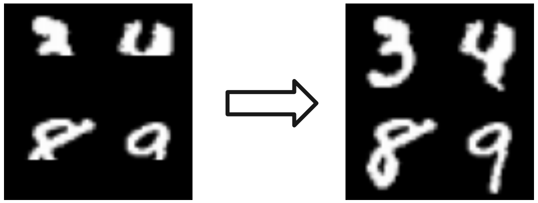
Some Perspective

Why Generative models

- Probabilistic reasoning
  - Denoising

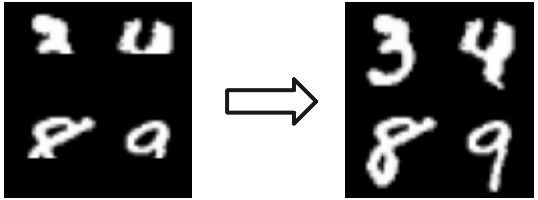
Why Generative models

- Probabilistic reasoning
  - Denoising
  - Missing-data imputation



Why Generative models

- Probabilistic reasoning
  - Denoising
  - Missing-data imputation



- Simulation-based
  - Planning and model-based reinforcement learning
  - Robots learning!

Previous work

<u>Binary Distribution Estimators</u>

- RBM (Smolensky 1986)
- NADE (Larochelle & Murray 2011)
- Deep NADE (Uria & al. 2014)
- DARN (Gregor & al. 2014)

Previous work

Binary Distribution Estimators

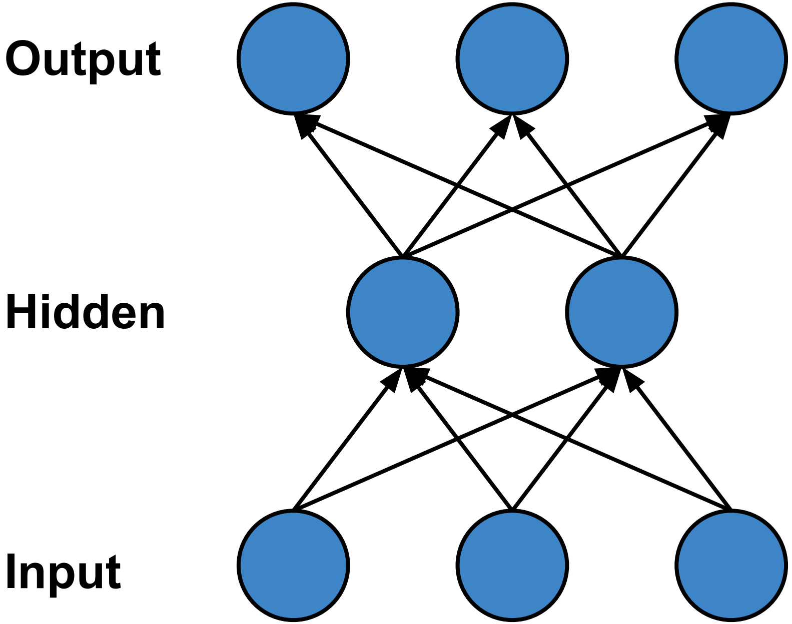
- RBM (Smolensky 1986)
- NADE (Larochelle & Murray 2011)
- Deep NADE (Uria & al. 2014)
- DARN (Gregor & al. 2014)

Problems

- Slow
- Intractable

MADE

Autoencoder

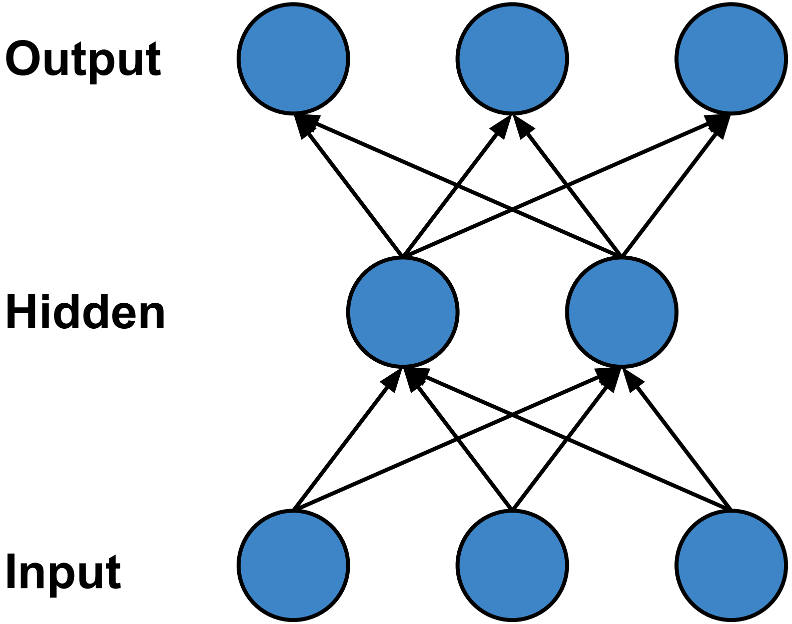


$$\widehat{\mathbf{x}} = \mathrm{sigm}(\mathbf{b}_1 + \mathbf{W}_1\mathbf{h})$$

$$\mathbf{h} = \sigma(\mathbf{b}_0 + \mathbf{W}_0\mathbf{x})$$

$$\mathbf{x}$$

Autoencoder



$$\widehat{\mathbf{x}} = \mathrm{sigm}(\mathbf{b}_1 + \mathbf{W}_1 \mathbf{h})$$

$$\mathbf{h} = \sigma(\mathbf{b}_0 + \mathbf{W}_0 \mathbf{x})$$

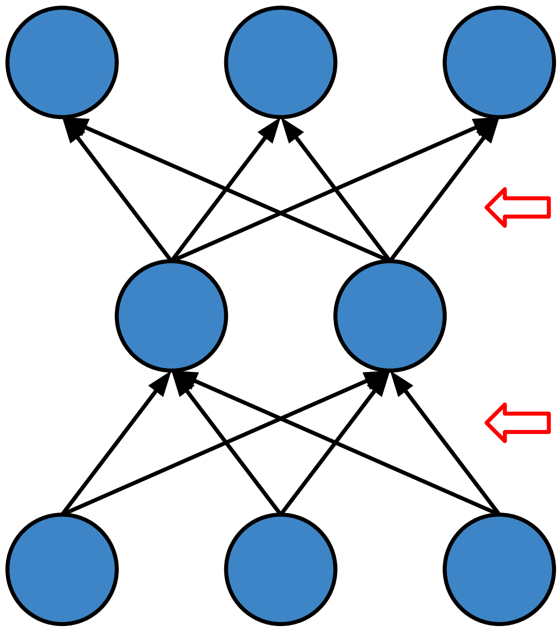
$$\mathbf{x}$$

Autoencoder

Output

Hidden

Input

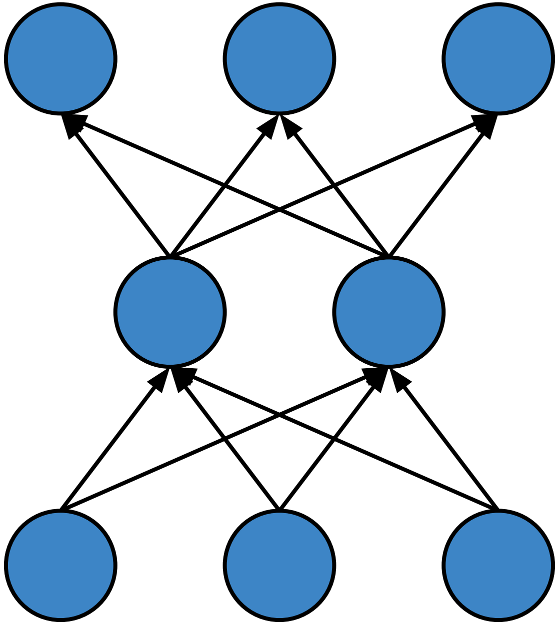


$$\hat{\mathbf{x}} = \mathrm{sigm}(\mathbf{b}_1 + \mathbf{W}_1\mathbf{h})$$

$$\mathbf{h} = \sigma(\mathbf{b}_0 + \mathbf{W}_0\mathbf{x})$$

$$\mathbf{x}$$

Autoencoder



Output

$$\hat{\mathbf{x}} = \mathrm{sigm}(\mathbf{b}_1 + \mathbf{W}_1 \mathbf{h})$$

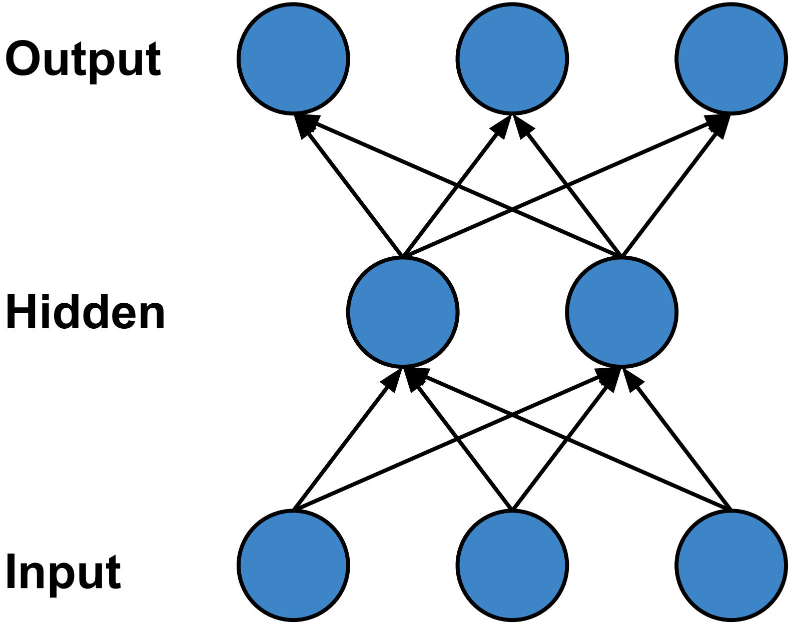
Hidden

$$\mathbf{h} = \sigma(\mathbf{b}_0 + \mathbf{W}_0 \mathbf{x})$$

Input

$$\mathbf{x}$$

Autoencoder

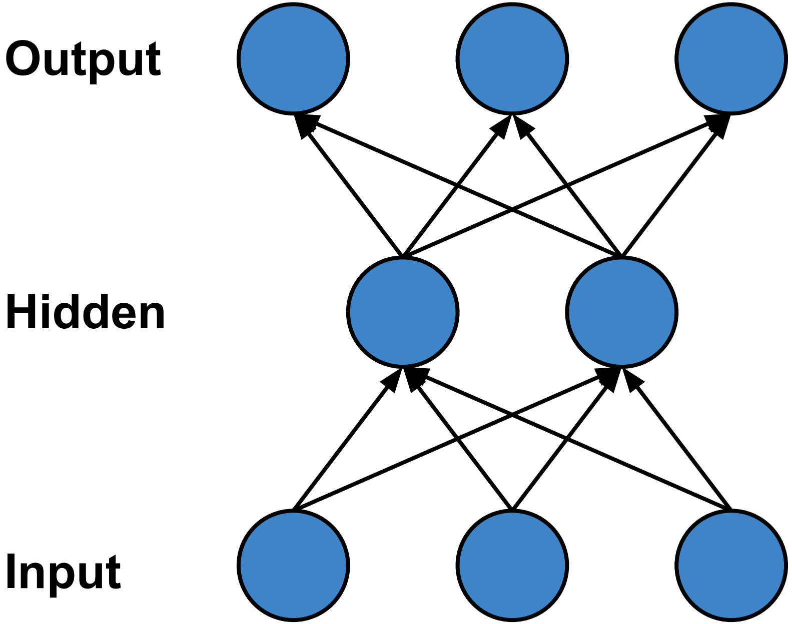


$$\hat{\mathbf{x}} = \mathrm{sigm}(\mathbf{b}_1 + \mathbf{W}_1 \mathbf{h})$$

$$\mathbf{h} = \sigma(\mathbf{b}_0 + \mathbf{W}_0 \mathbf{x})$$

$$\mathbf{x}$$

Autoencoder

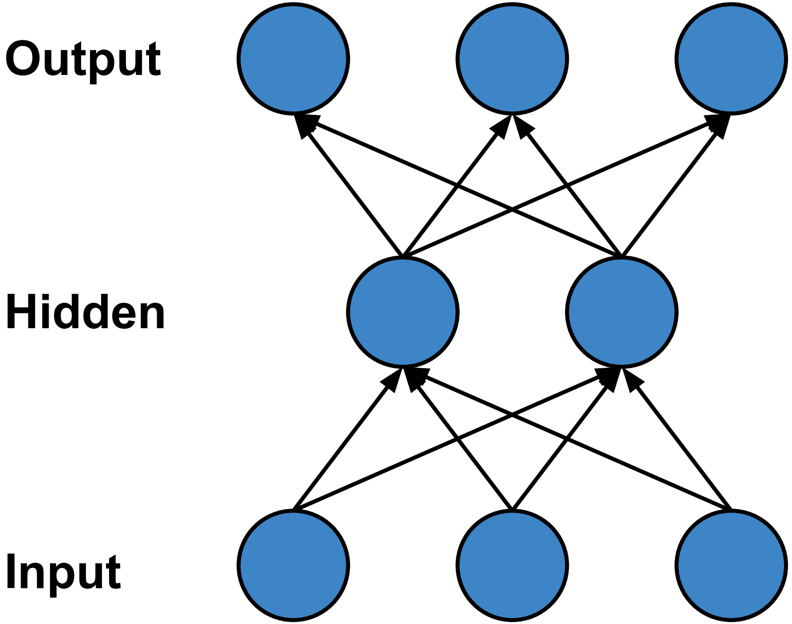


$$\widehat{\mathbf{x}} = \operatorname{sigm}(\mathbf{b}_1 + \mathbf{W}_1 \mathbf{h})$$

$$\mathbf{h} = \sigma(\mathbf{b}_0 + \mathbf{W}_0 \mathbf{x})$$

$$\mathbf{x}$$

Autoencoder



$$\hat{\mathbf{x}} = \mathrm{sigm}(\mathbf{b}_1 + \mathbf{W}_1 \mathbf{h})$$

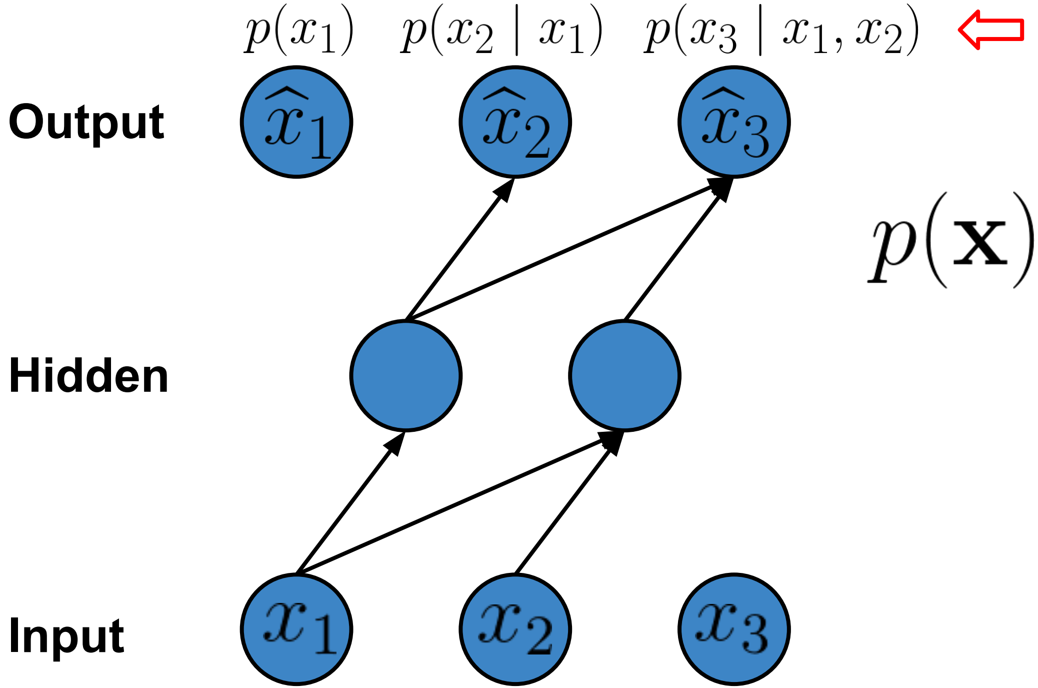
$$\mathbf{h} = \sigma(\mathbf{b}_0 + \mathbf{W}_0 \mathbf{x})$$

$$\mathbf{x}$$

Key contribution :

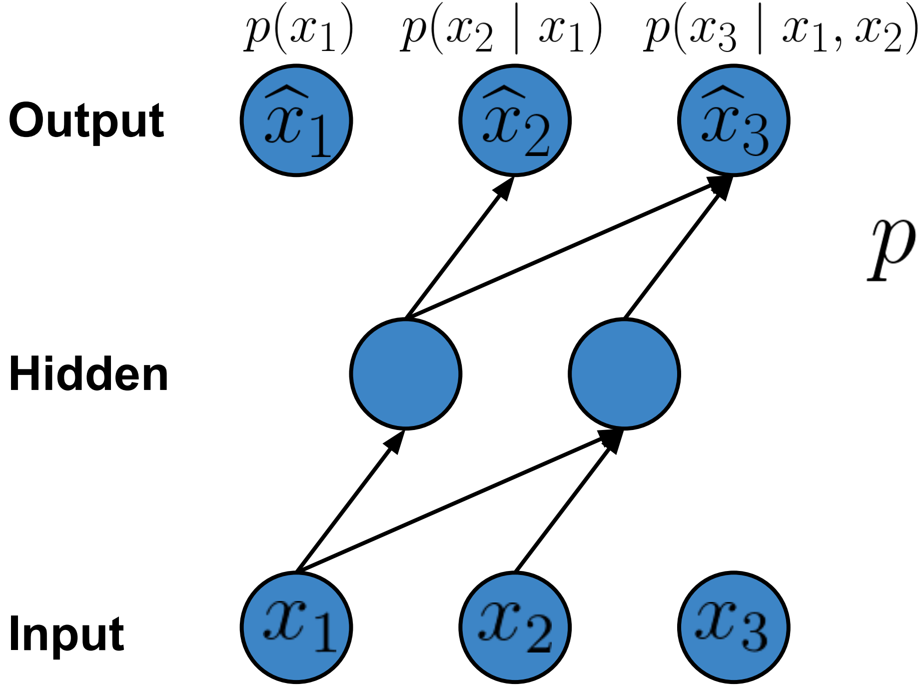
How to simply adapt an **autoencoder** into a **generative model**

Autoregression



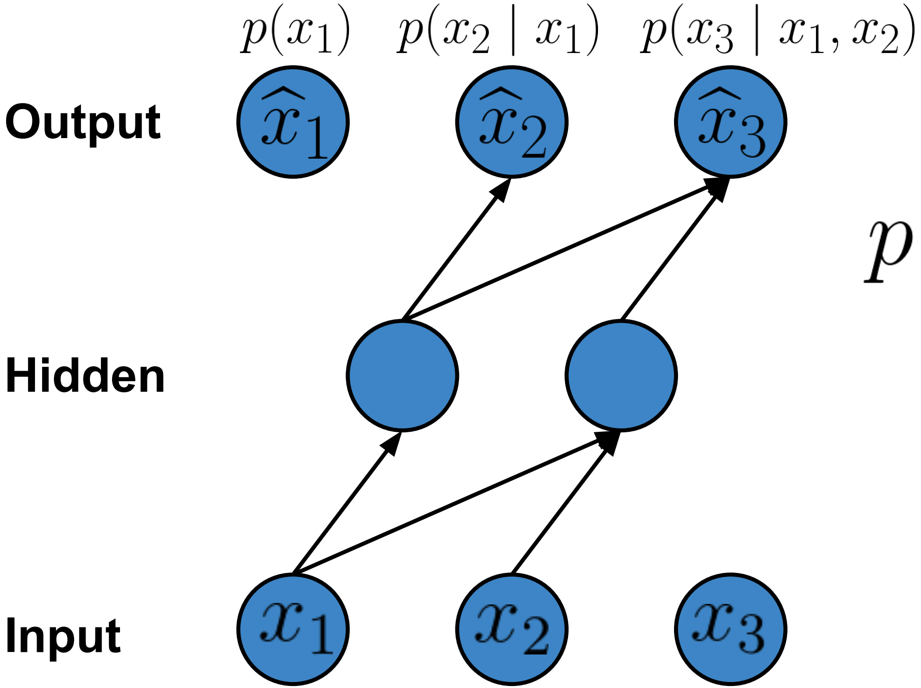
$$p(\mathbf{x}) = p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_1, x_2)$$

Autoregression



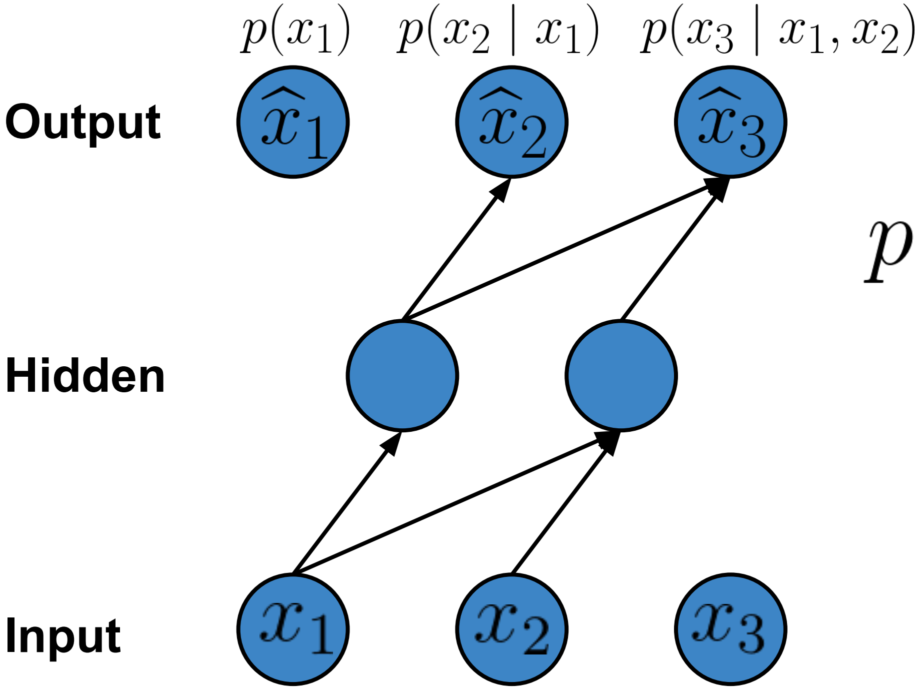
$$p(\mathbf{x}) = p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_1, x_2)$$

Autoregression



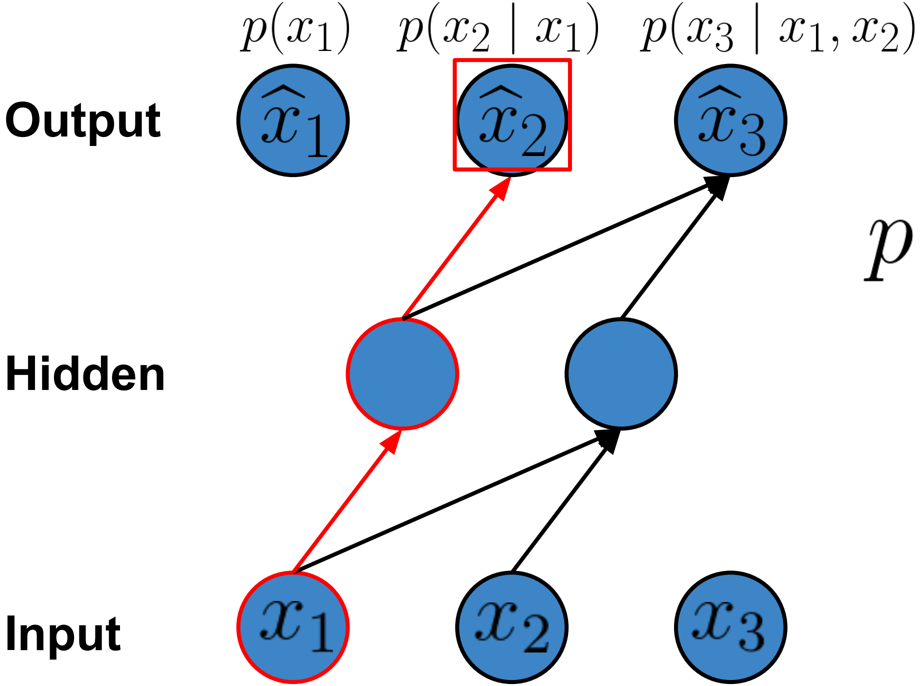
$$p(\mathbf{x}) = p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_1, x_2)$$

Autoregression



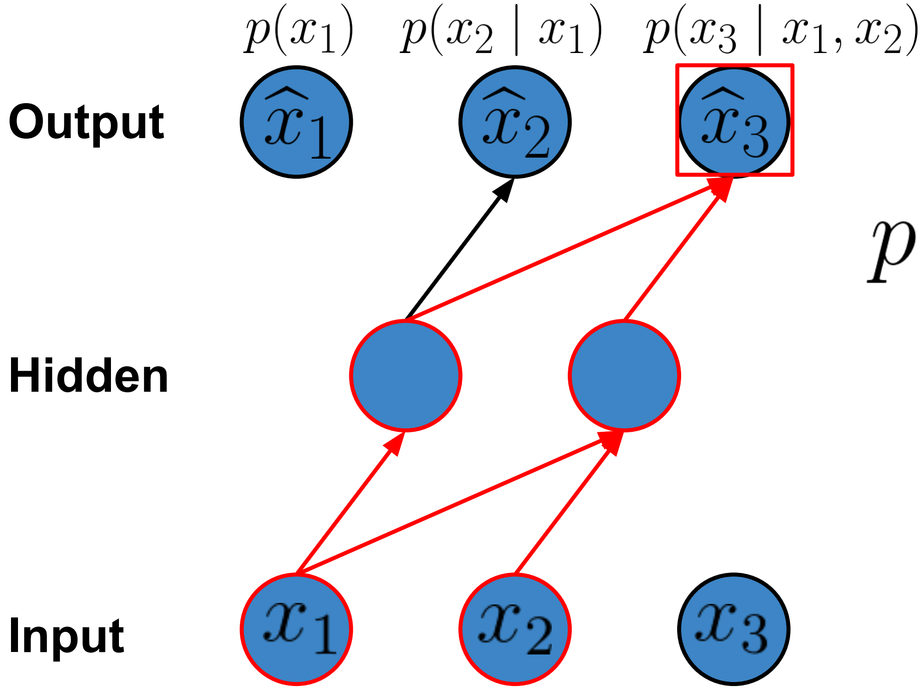
$$p(\mathbf{x}) = p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_1, x_2)$$

Autoregression



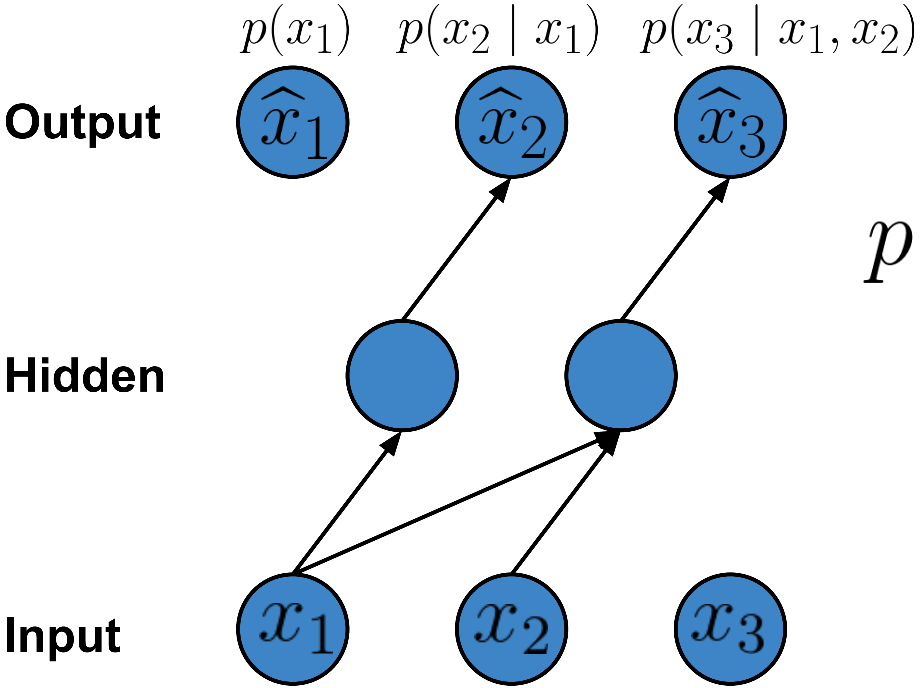
$$p(\mathbf{x}) = p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_1, x_2)$$

Autoregression



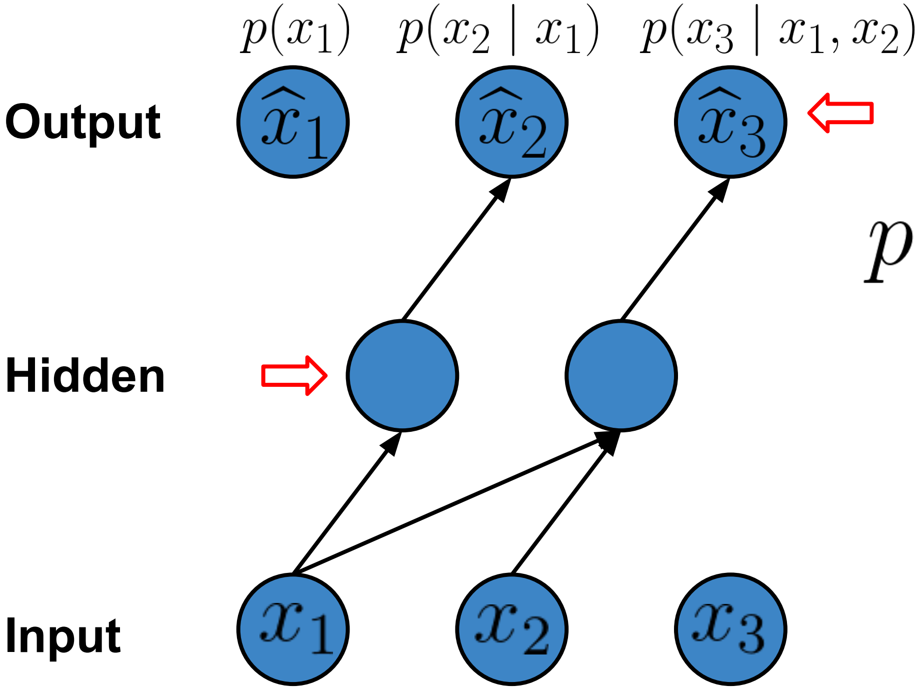
$$
\begin{aligned}
p(\mathbf{x}) = p(x_1) \\
\cdot\, p(x_2 \mid x_1) \\
\cdot\, p(x_3 \mid x_1, x_2)
\end{aligned}
$$

Autoregression



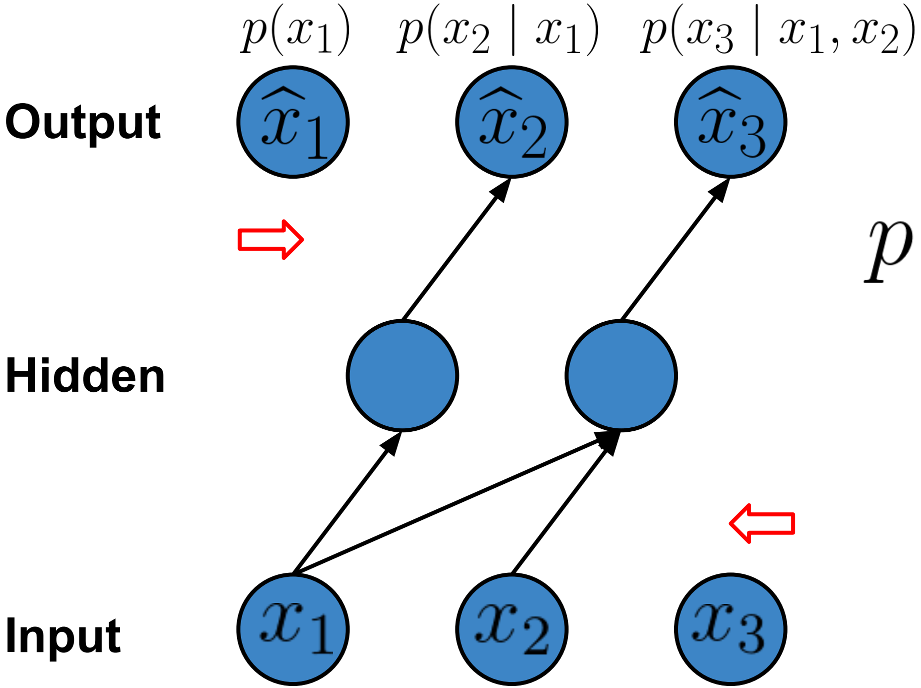
$$
\begin{aligned}
p(\mathbf{x}) = p(x_1)& \\
\cdot\, p(x_2 \mid x_1)& \\
\cdot\, p(x_3 \mid x_1, x_2)&
\end{aligned}
$$

Autoregression



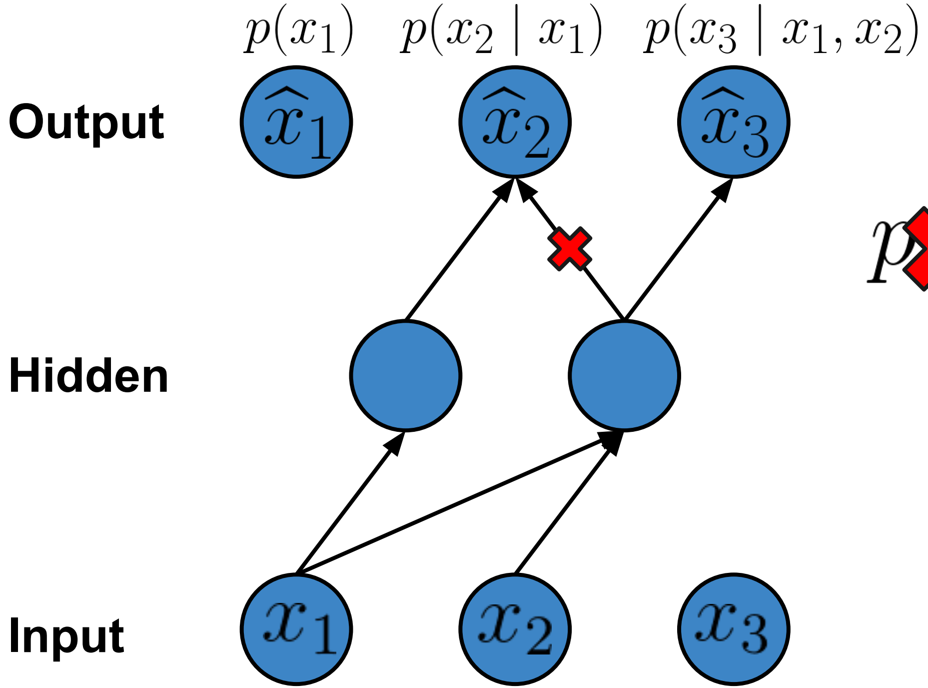
$$
\begin{aligned}
p(\mathbf{x}) = p(x_1)& \\
&\cdot p(x_2 \mid x_1) \\
&\cdot p(x_3 \mid x_1, x_2)
\end{aligned}
$$

Autoregression



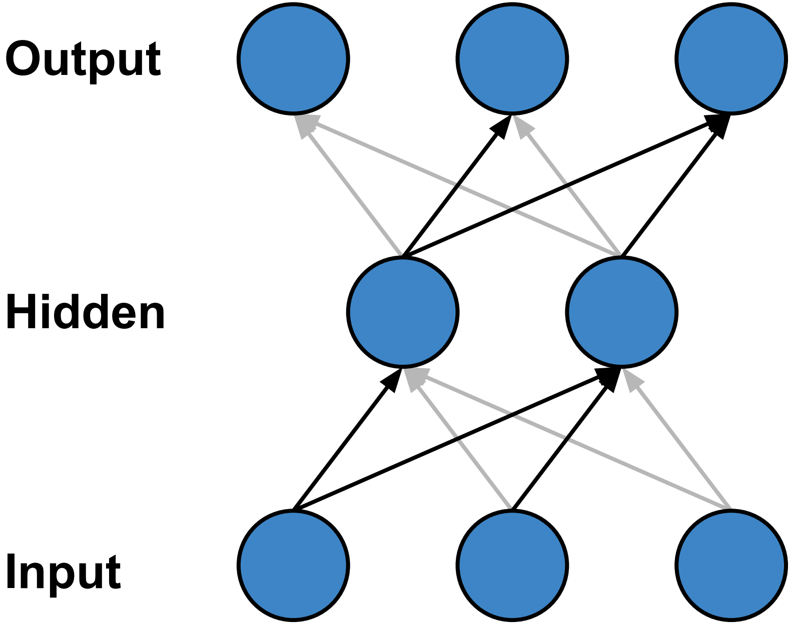
$$p(\mathbf{x}) = p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_1, x_2)$$

Autoregression



$$p(\mathbf{x}) = p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_1, x_2)$$

Masks



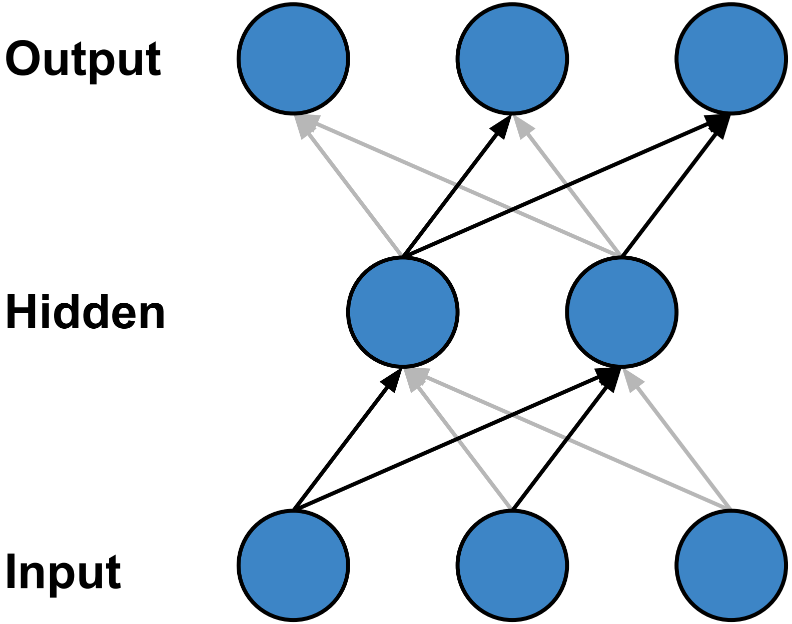
$$\hat{\mathbf{x}} = \mathrm{sigm}(\mathbf{b_1} + (\mathbf{W_1} \odot \mathbf{M_1})\mathbf{h})$$

$$\mathbf{h} = \sigma(\mathbf{b_0} + (\mathbf{W_0} \odot \mathbf{M_0})\mathbf{x})$$

$\mathbf{x}$

$$\mathrm{M_0} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \mathrm{M_1} = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

Masks



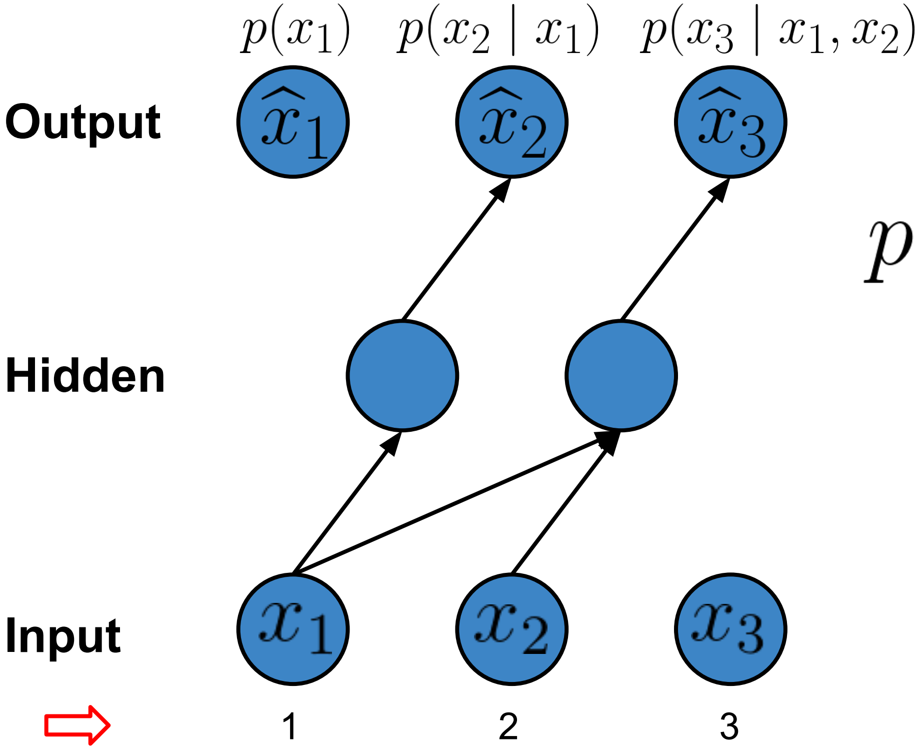
$$\hat{\mathbf{x}} = \mathrm{sigm}(\mathbf{b}_1 + (\mathbf{W}_1 \odot \mathbf{M}_1)\mathbf{h})$$

$$\mathbf{h} = \sigma(\mathbf{b}_0 + (\mathbf{W}_0 \odot \mathbf{M}_0)\mathbf{x})$$

$\mathbf{x}$

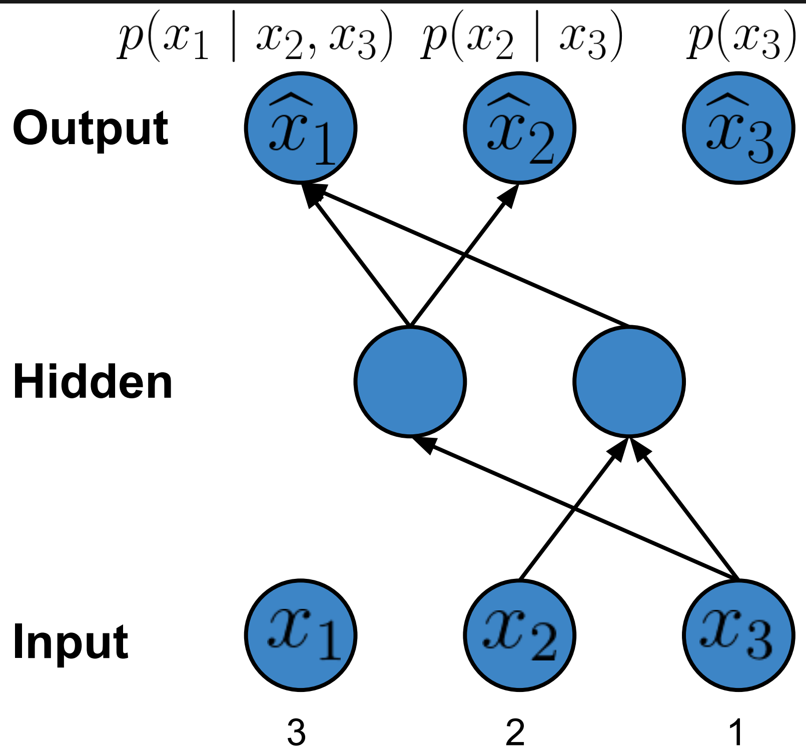
$$\mathbf{M}_0 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \mathbf{M}_1 = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

Orderings



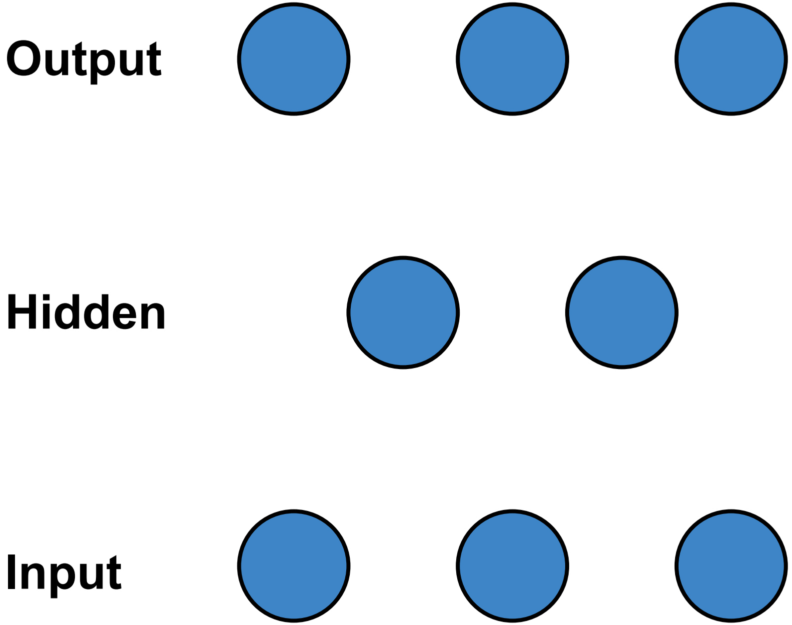
$$p(\mathbf{x}) = p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_1, x_2)$$

Orderings



$$p(\mathbf{x}) = p(x_1 \mid x_2, x_3) \cdot p(x_2 \mid x_3) \cdot p(x_3)$$

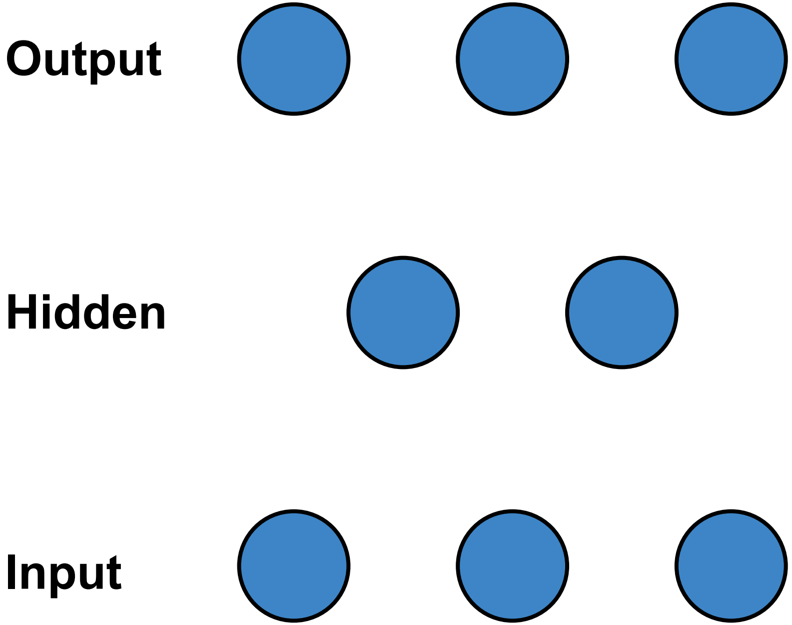
Mask Sampling



$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

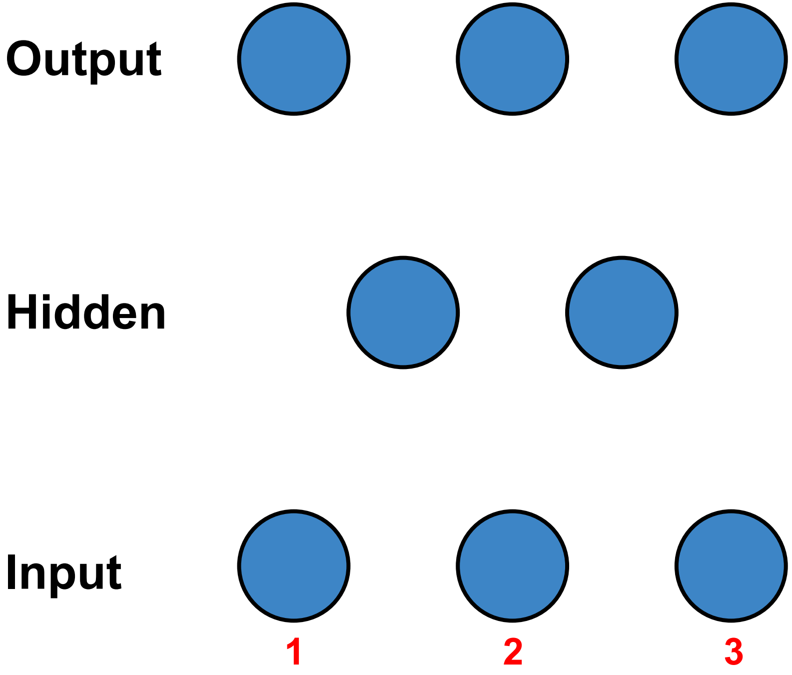
Mask Sampling



$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

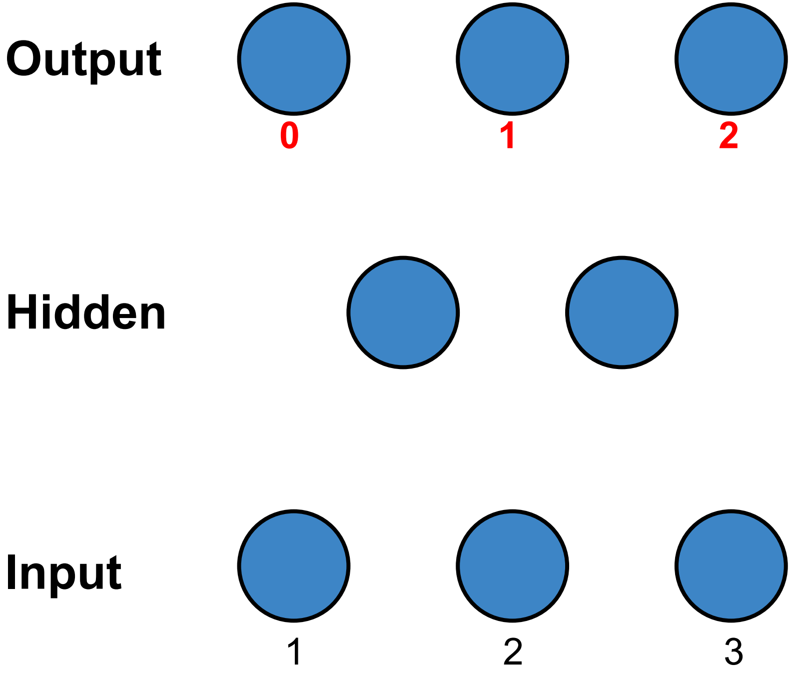
Mask Sampling



$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

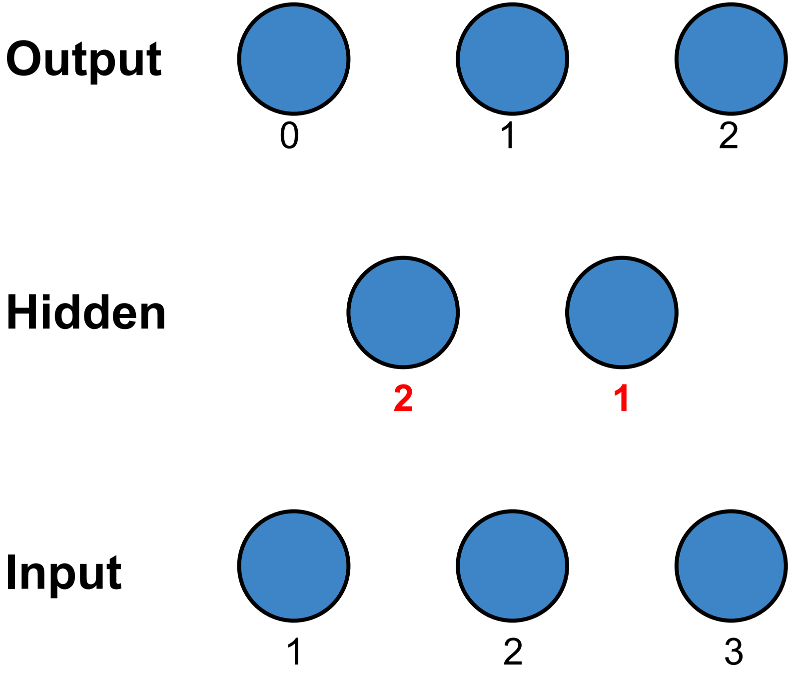
Mask Sampling



$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

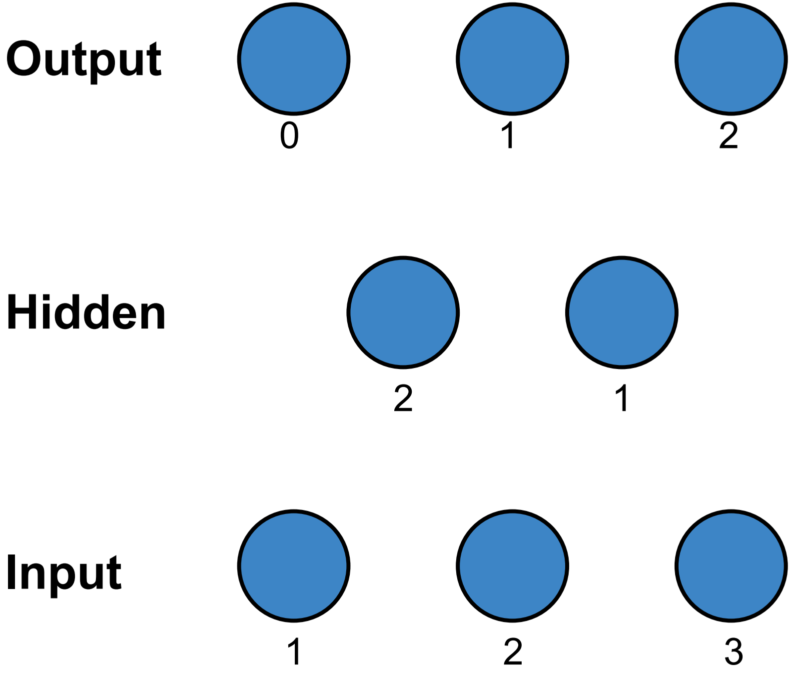
Mask Sampling



$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

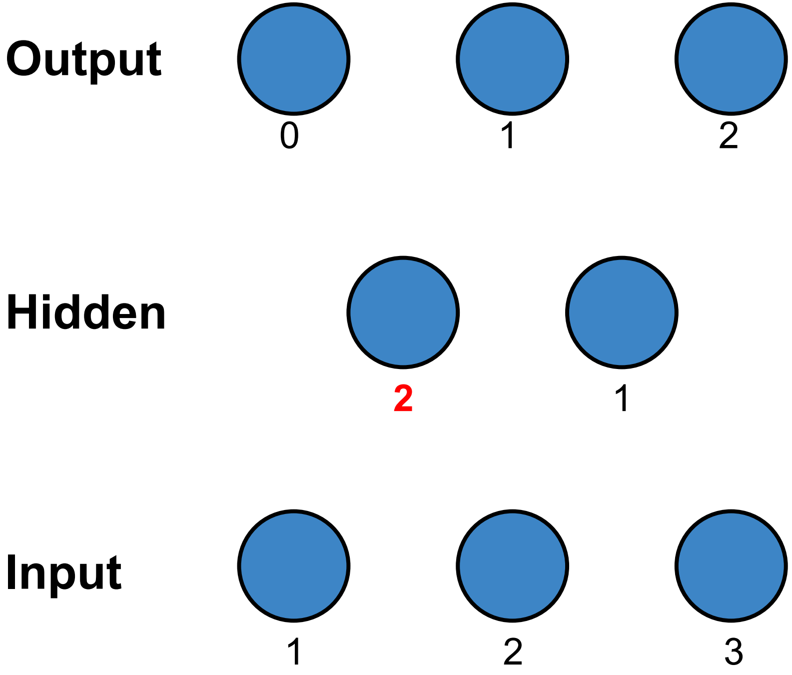
Mask Sampling



$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

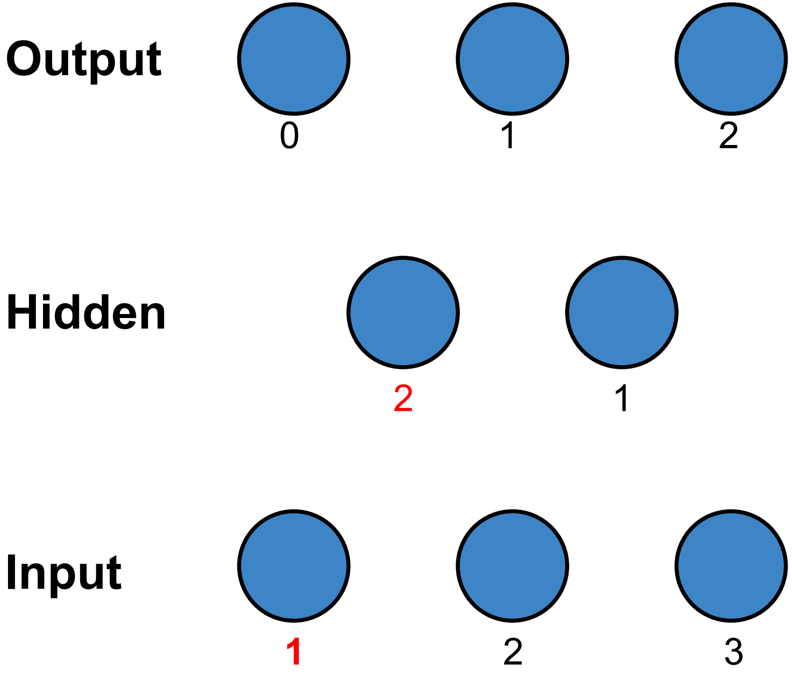
Mask Sampling



$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

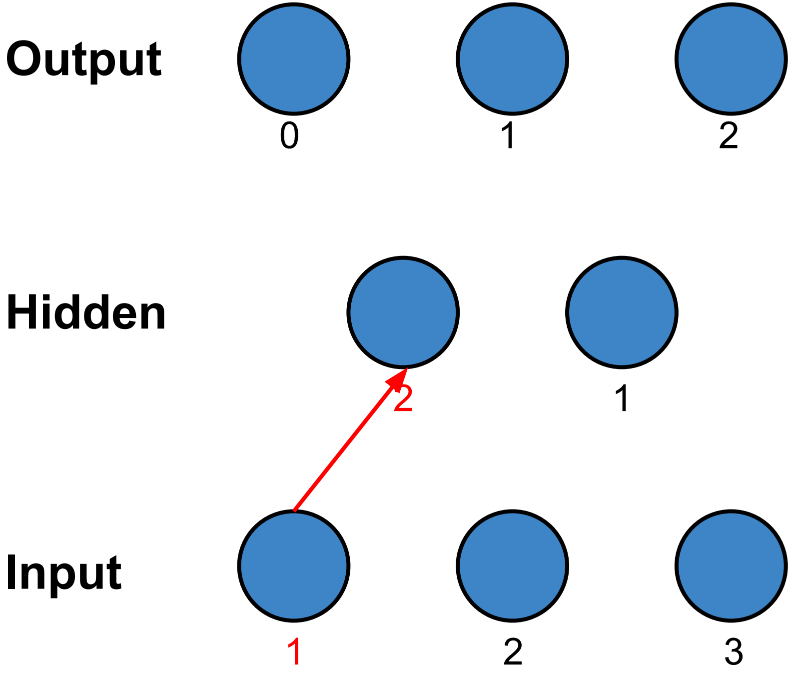
Mask Sampling



$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

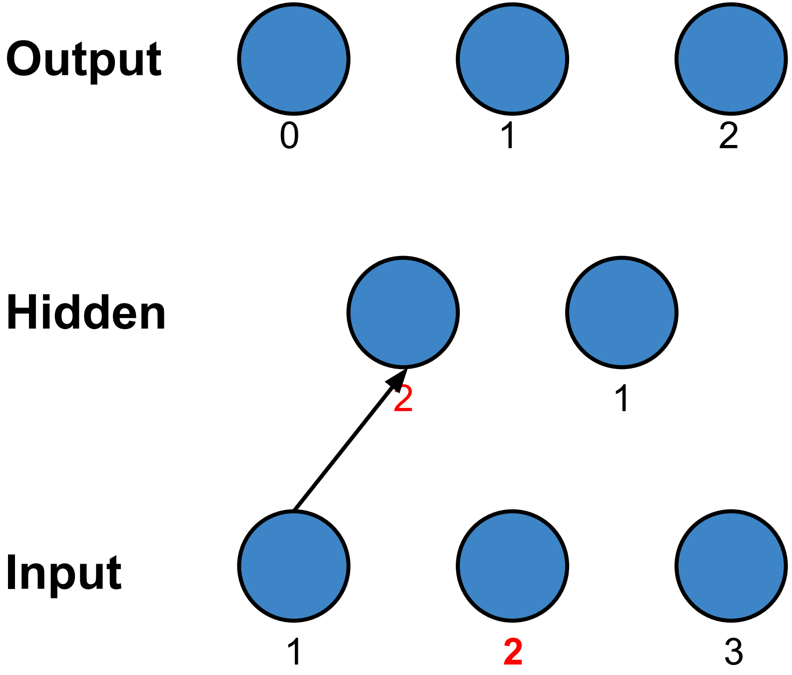
Mask Sampling



$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

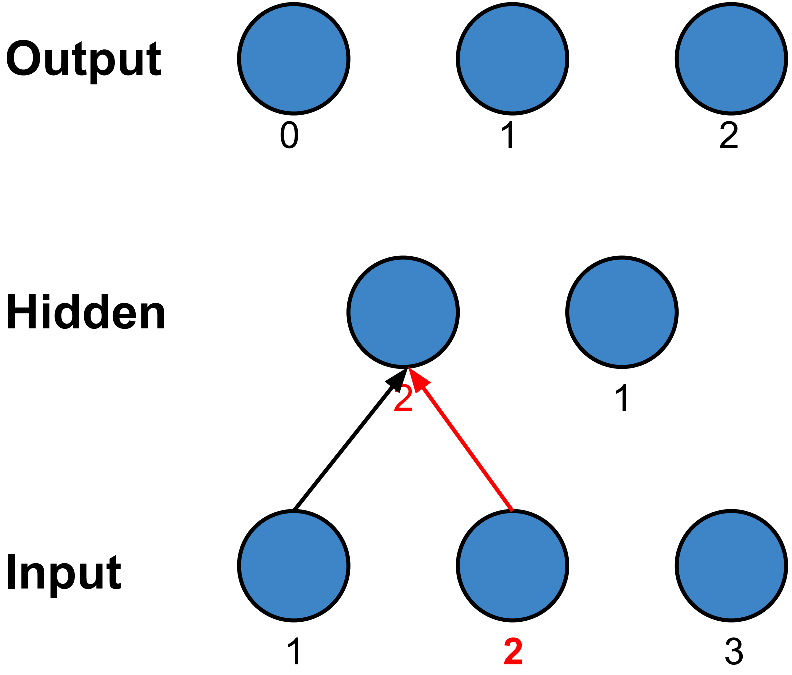
Mask Sampling



$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

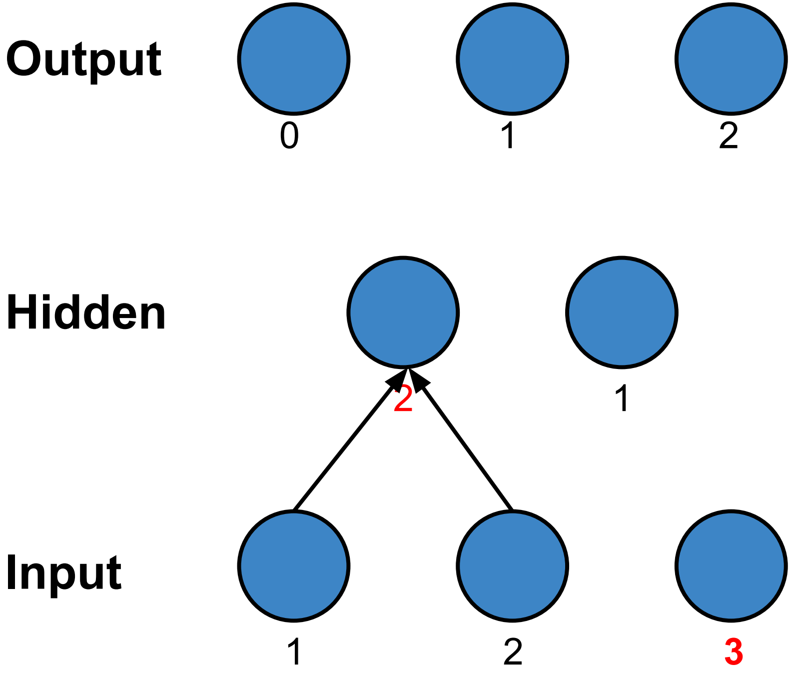
Mask Sampling



$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}$$

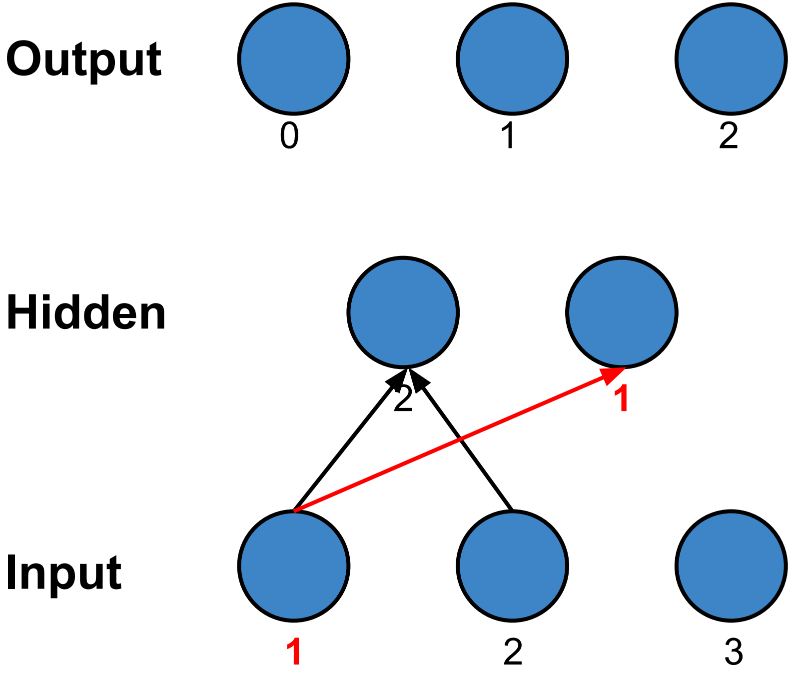
Mask Sampling



$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}$$

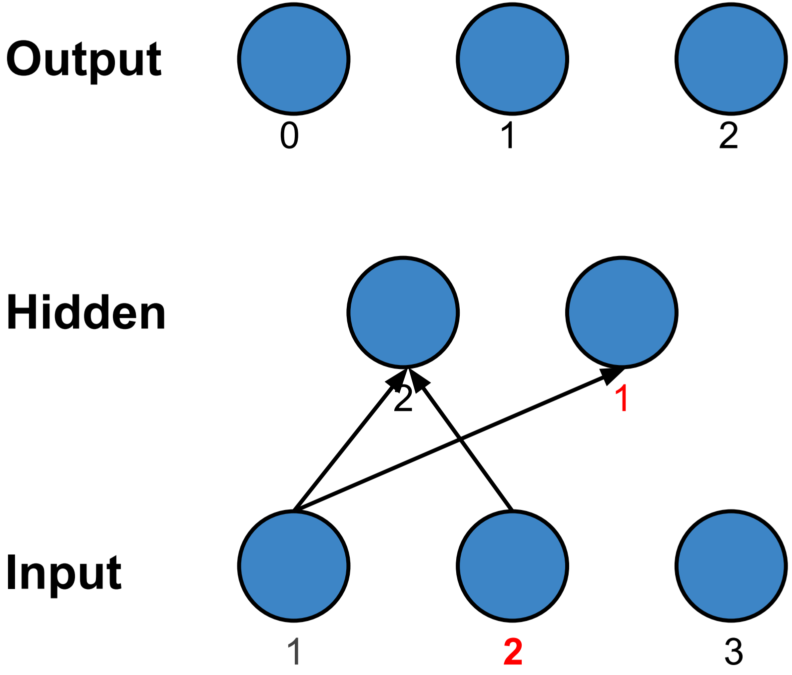
Mask Sampling



$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 1 & \color{red}{1} \\ 1 & 0 \\ 0 & 0 \end{bmatrix}$$

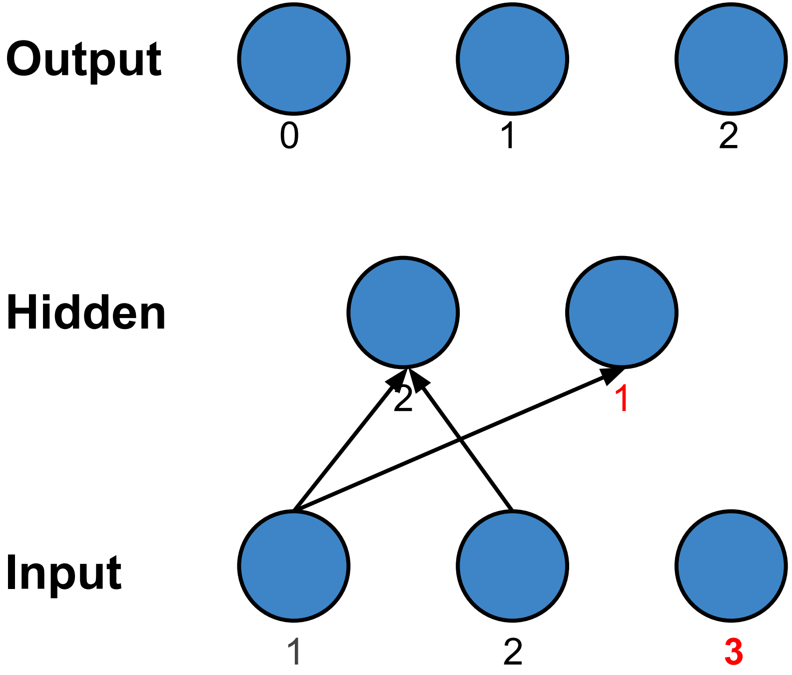
Mask Sampling



$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}$$

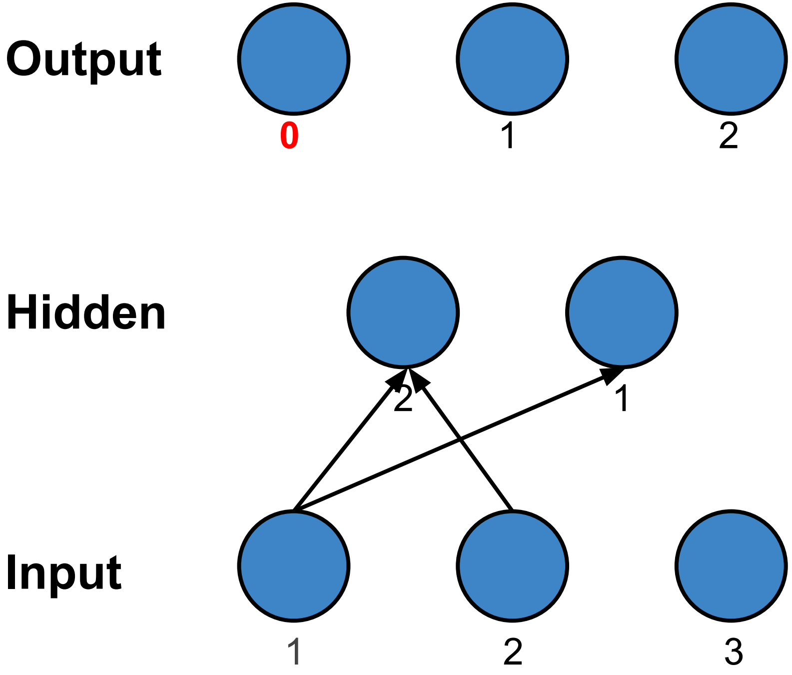
Mask Sampling



$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}$$

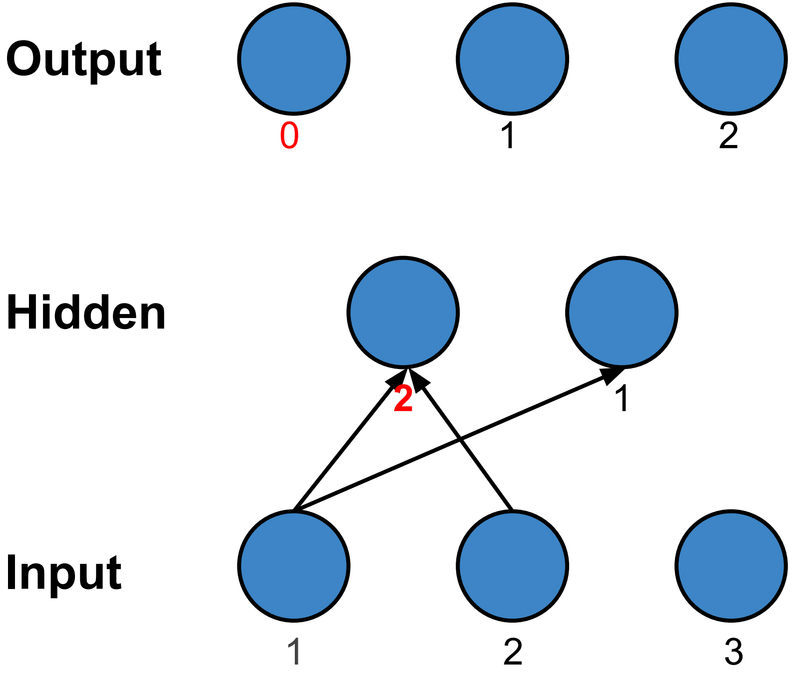
Mask Sampling



$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}$$

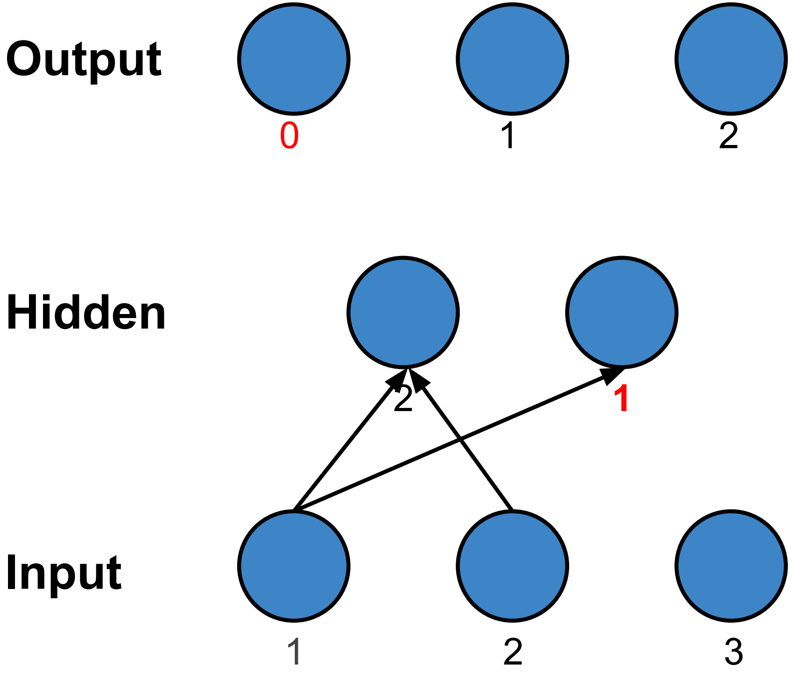
Mask Sampling



$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}$$

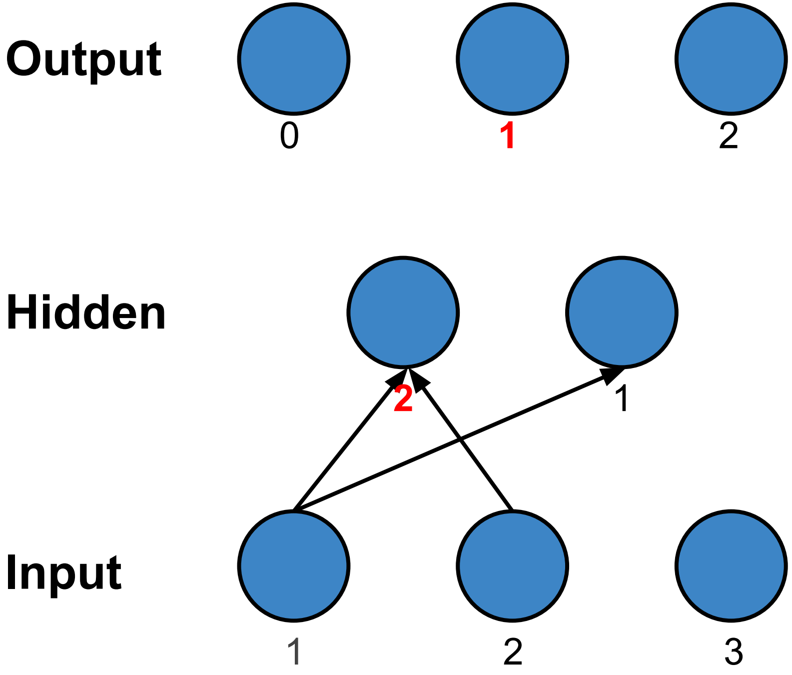
Mask Sampling



$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}$$

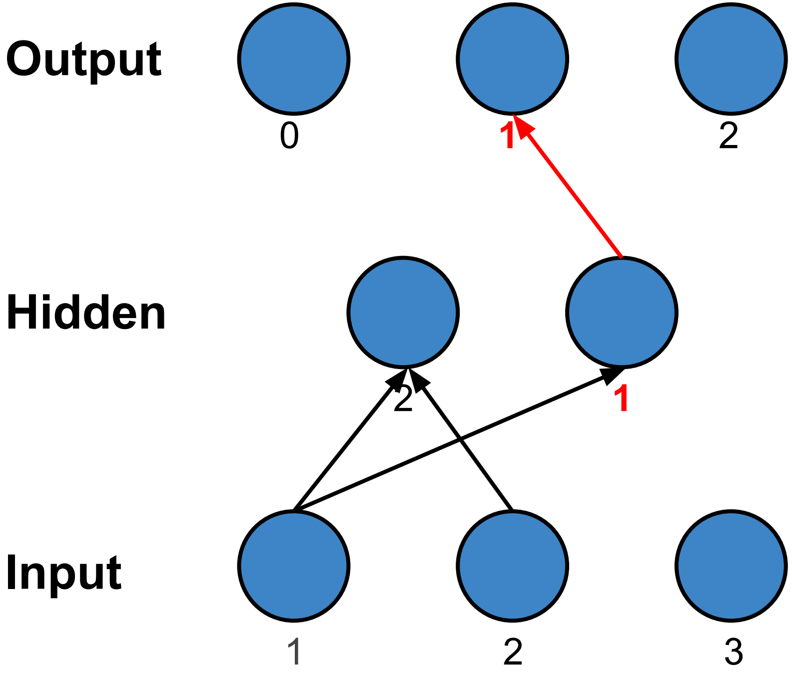
Mask Sampling



$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}$$

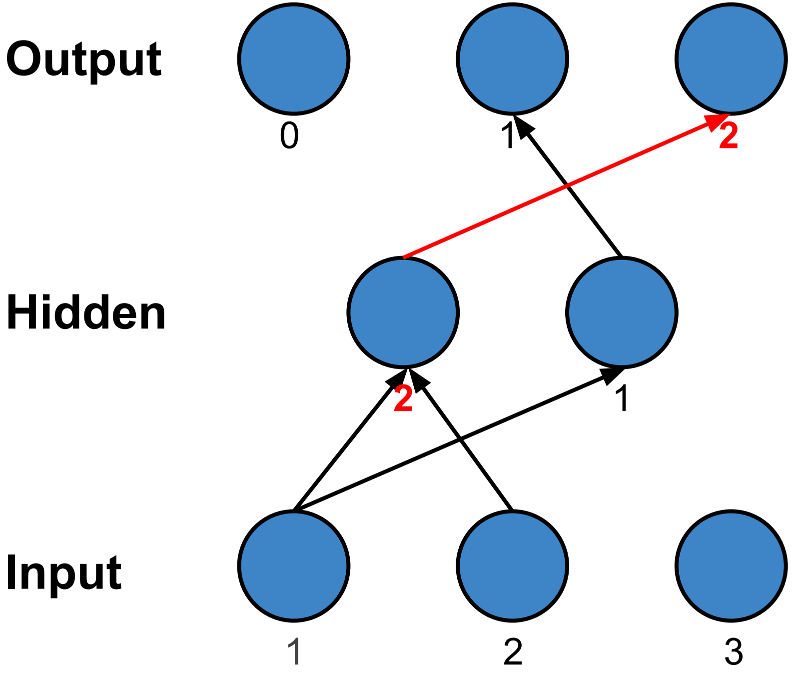
Mask Sampling



$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \color{red}{1} & 0 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}$$

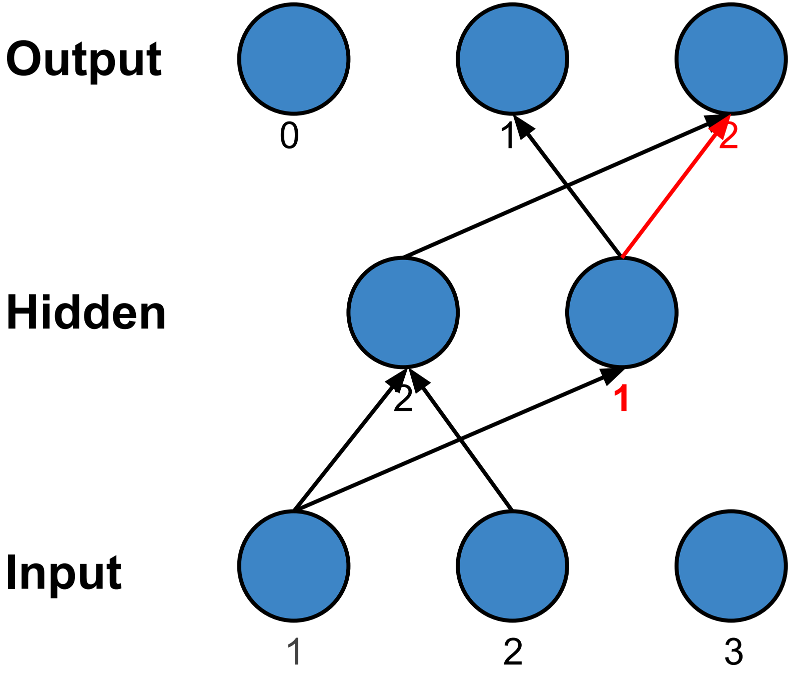
Mask Sampling



$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & {\color{red}1} \\ 0 & 1 & 0 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}$$

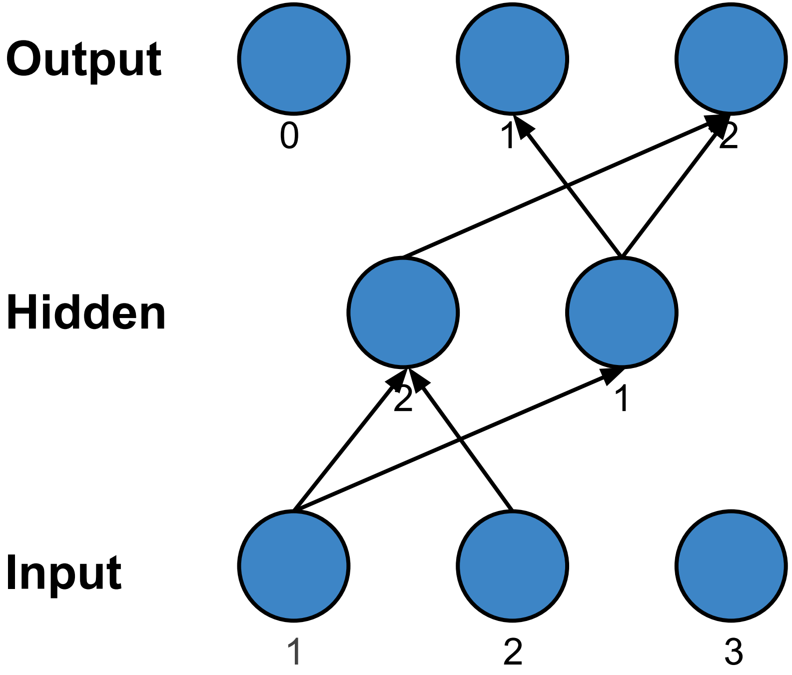
Mask Sampling



$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}$$

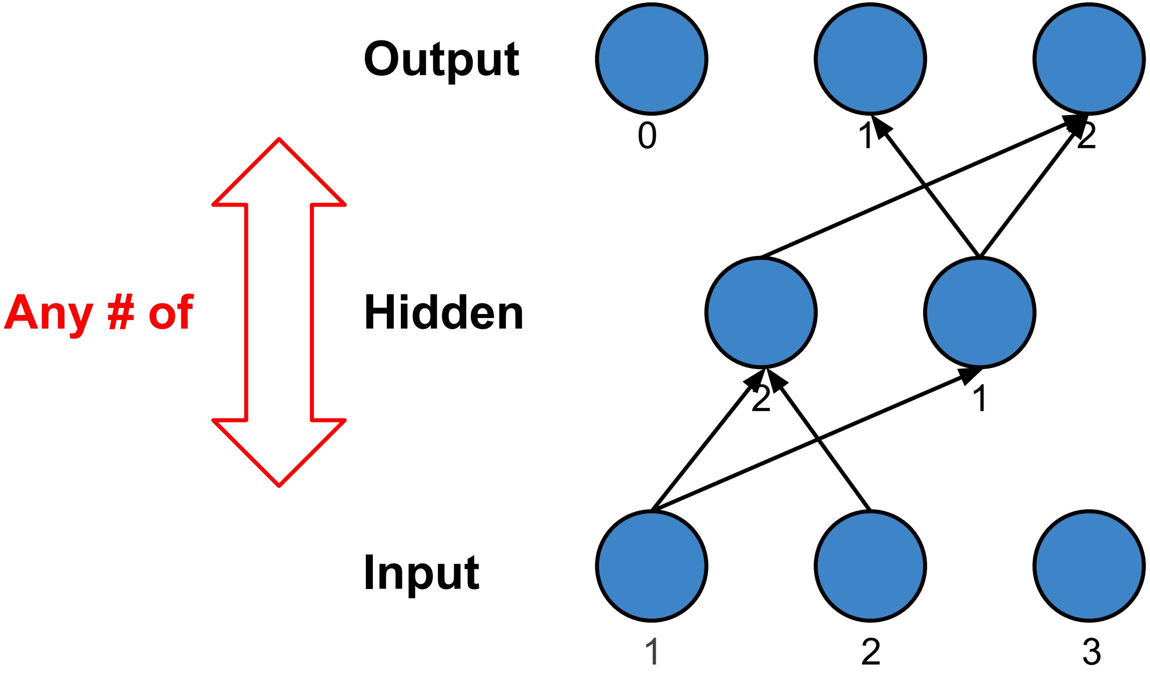
Mask Sampling



$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}$$

Mask Sampling



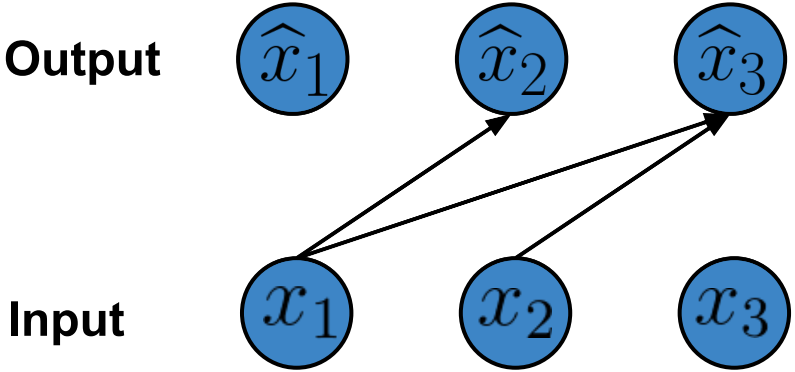
$$\mathbf{M_1} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

$$\mathbf{M_0} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}$$

Shallow MADE aka FVSBN

- No Hidden Layers
- No Mask Sampling
- No Reordering

FVSBN*



* *Frey, 1998; Bengio and Bengio, 2000*

Training

Training

- Binary cross entropy
- AdaDelta or AdaGrad
- Mini-Batches: 100
- GPU

Training

Training

- Binary cross entropy
- AdaDelta or AdaGrad
- Mini-Batches: 100
- GPU

Mask & Testing

- Mask sampling
  - Infinite set
  - Finite set
- Test:Ensemble of MADE *

** Uria & al. 2014*

Sampling

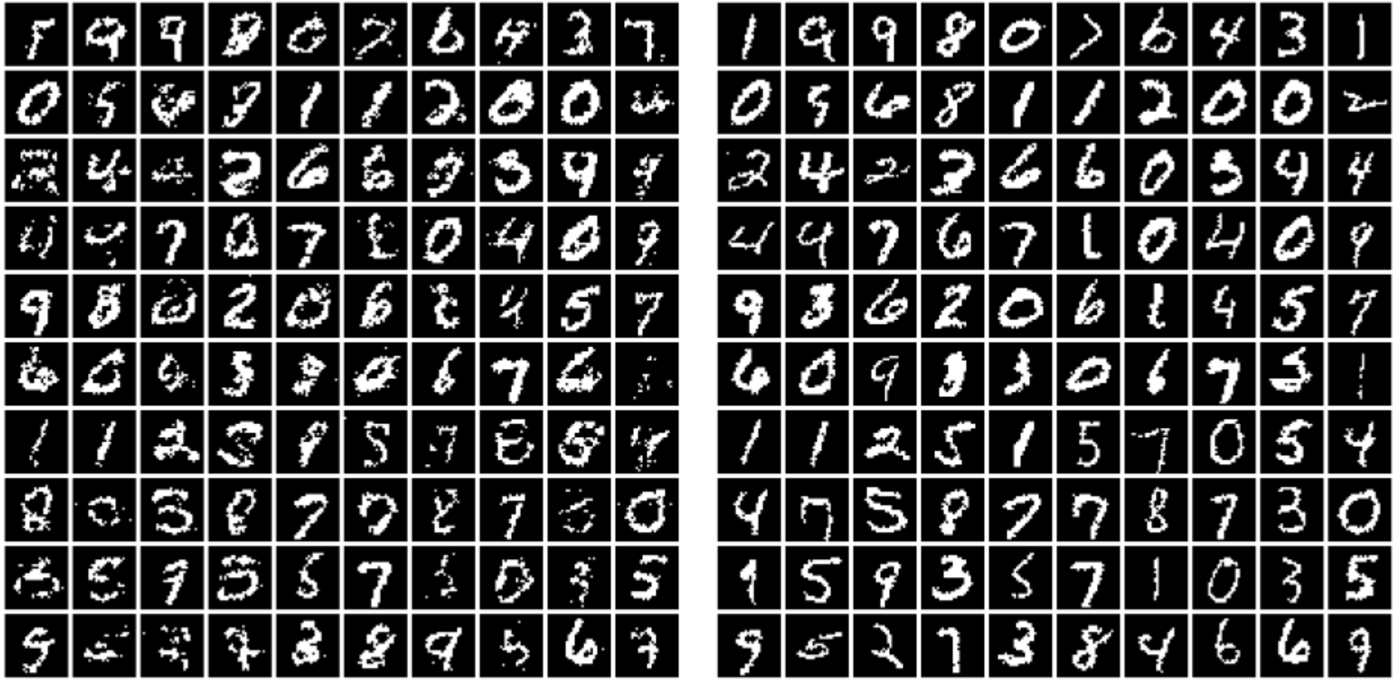


Figure 3. Left: Samples from a 2 hidden layer MADE. Right: Nearest neighbour in binarized MNIST.

Sampling

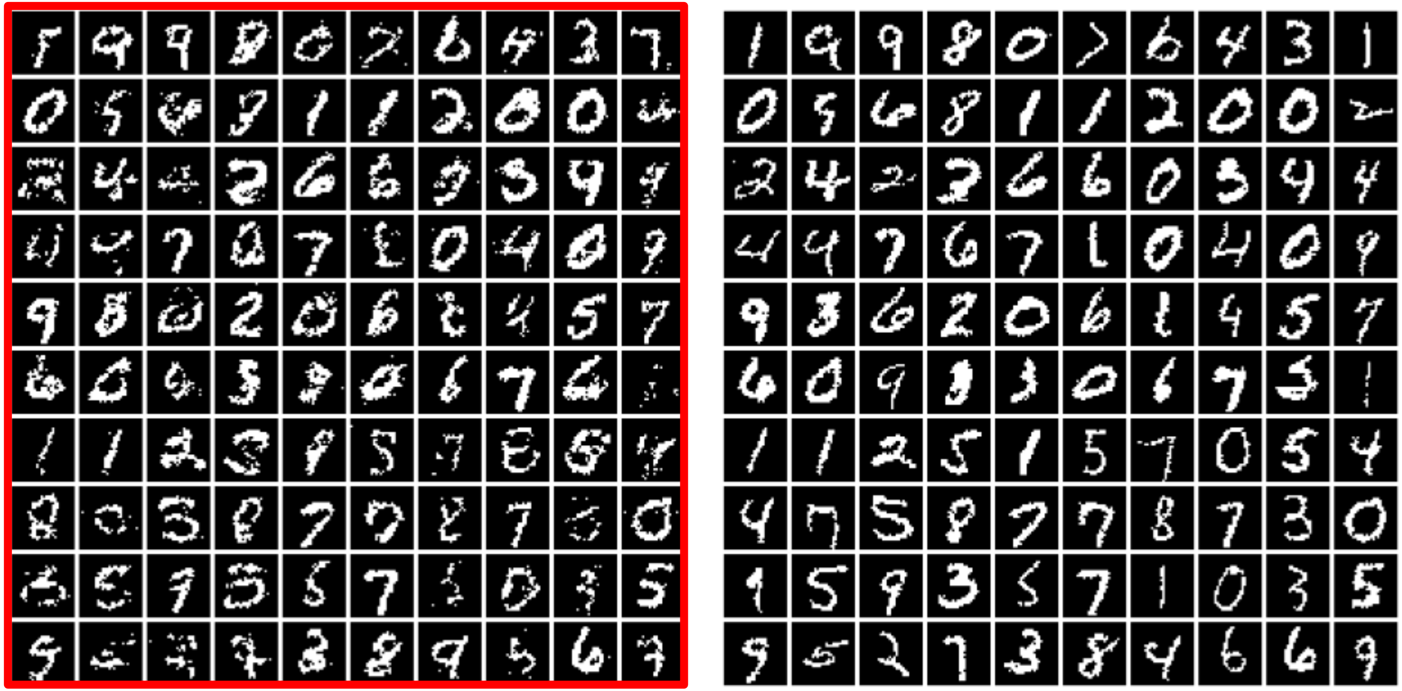


Figure 3. Left: Samples from a 2 hidden layer MADE. Right: Nearest neighbour in binarized MNIST.

Sampling

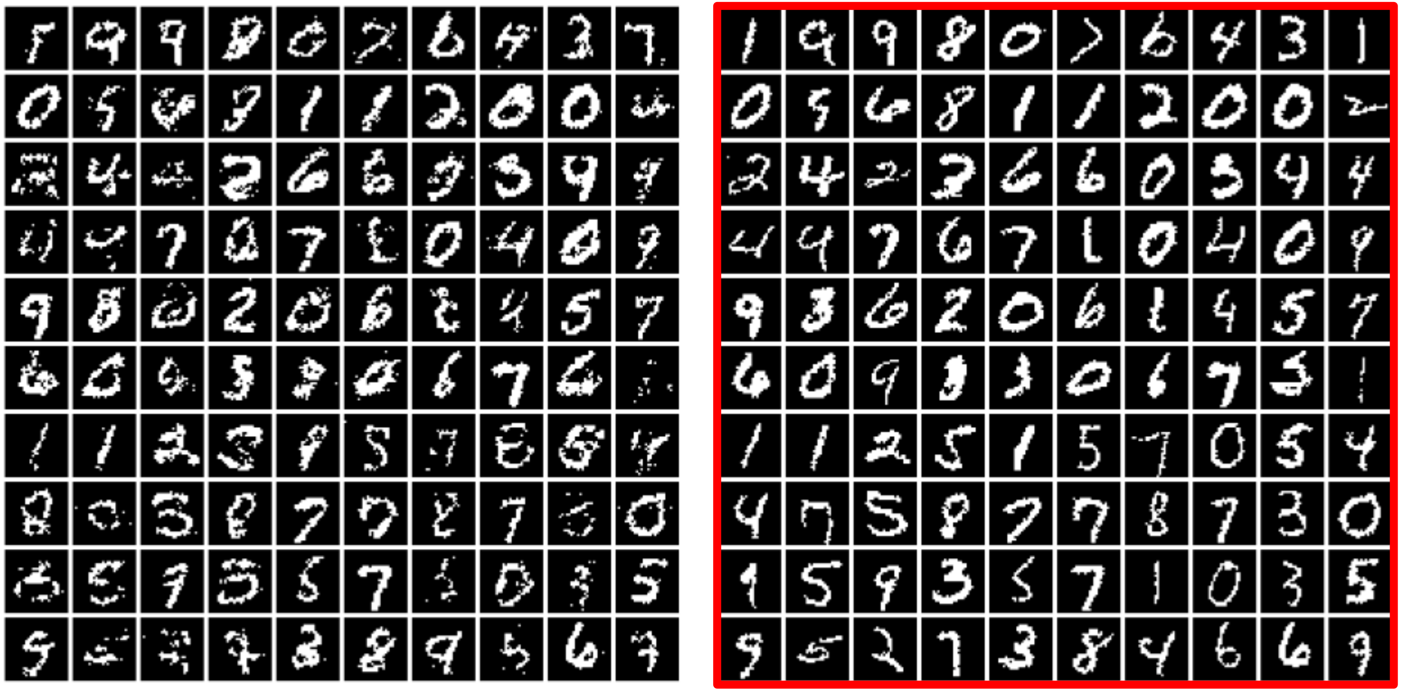
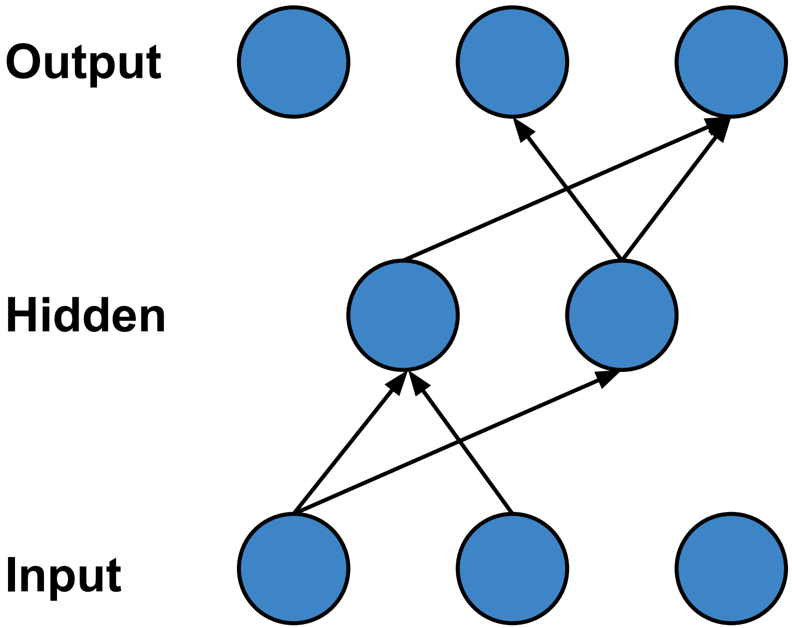


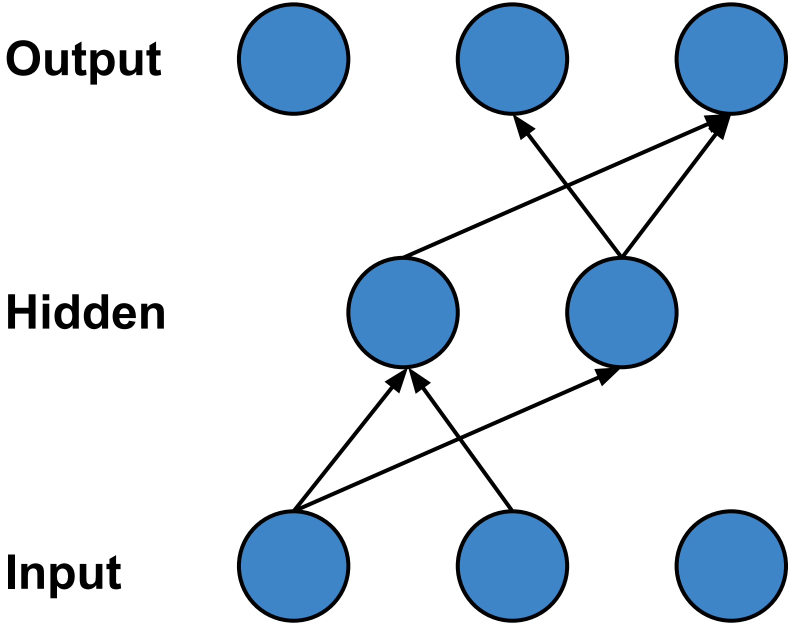
Figure 3. Left: Samples from a 2 hidden layer MADE. Right: Nearest neighbour in binarized MNIST.

Ancestral Sampling



Ancestral Sampling

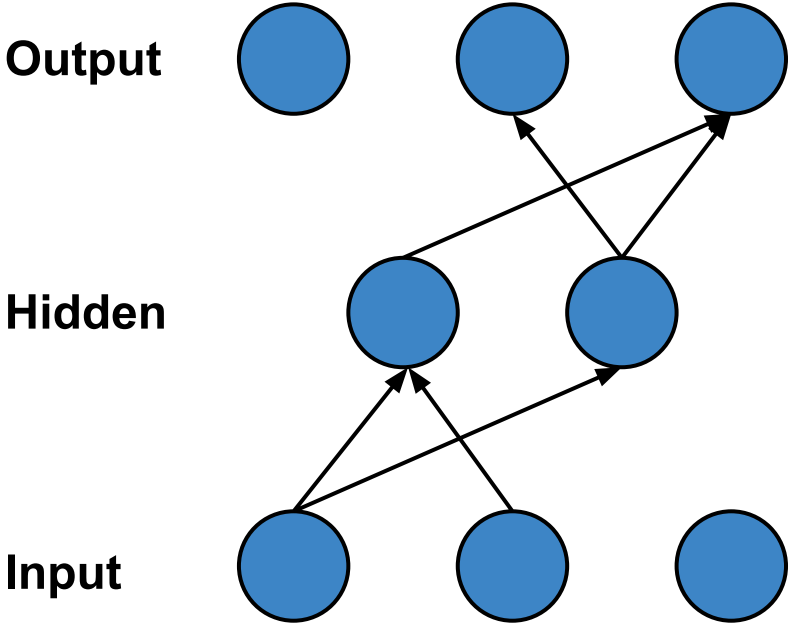
$$\mathbf{x} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$



Ancestral Sampling

$$position = 1$$

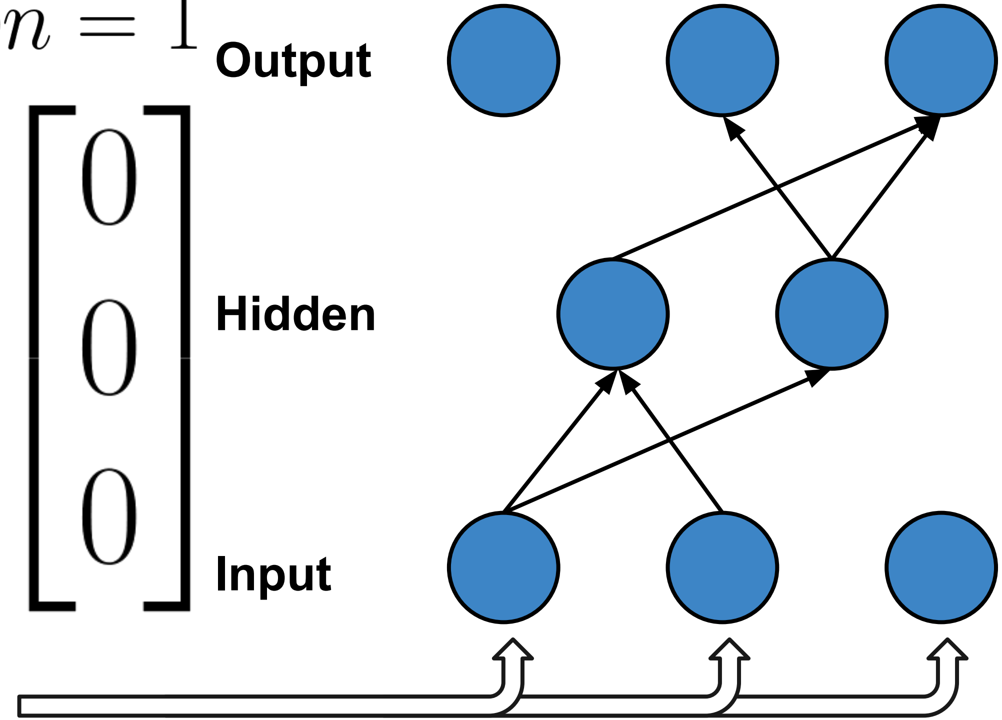
$$\mathbf{x} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$



Ancestral Sampling

$$position = 1$$

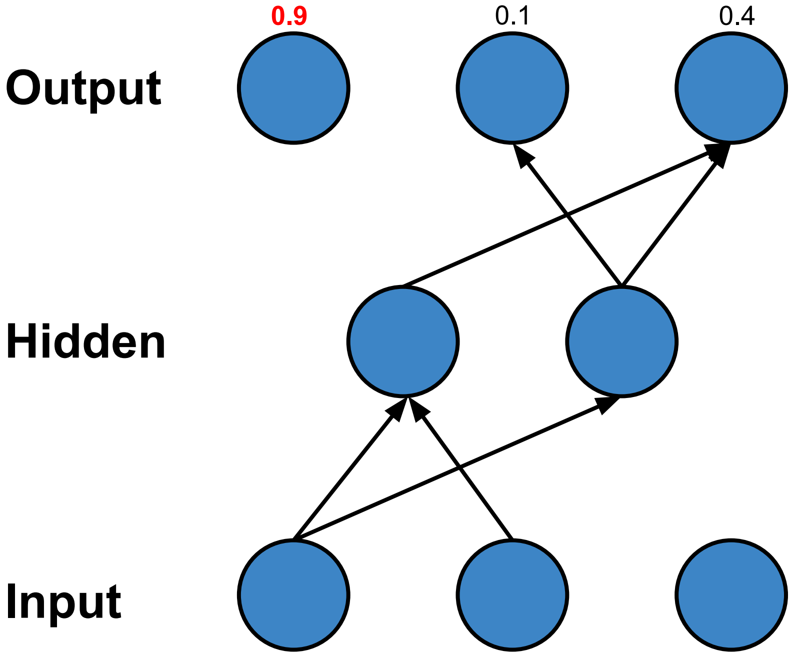
$$\mathbf{x} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$



Ancestral Sampling

$$position = 1$$

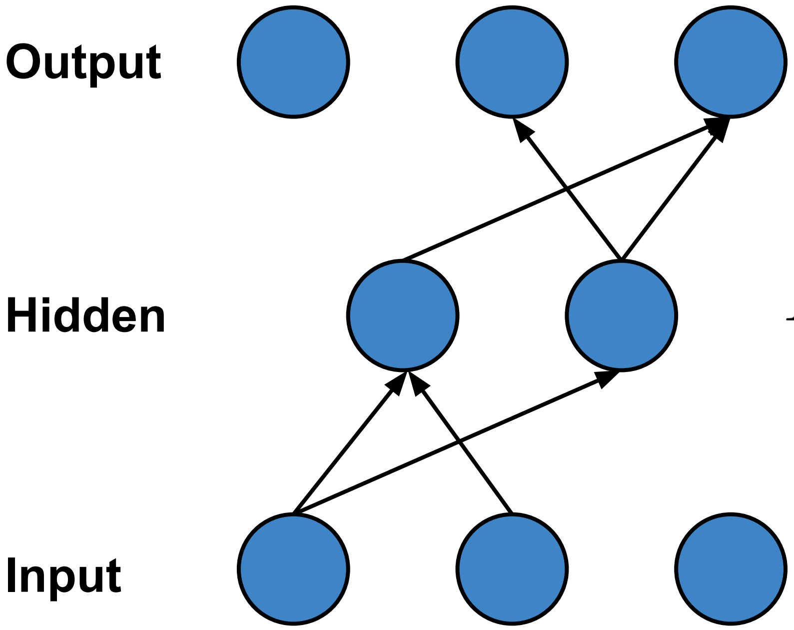
$$\mathbf{x} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$



Ancestral Sampling

$$position = 1$$

$$\mathbf{x} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

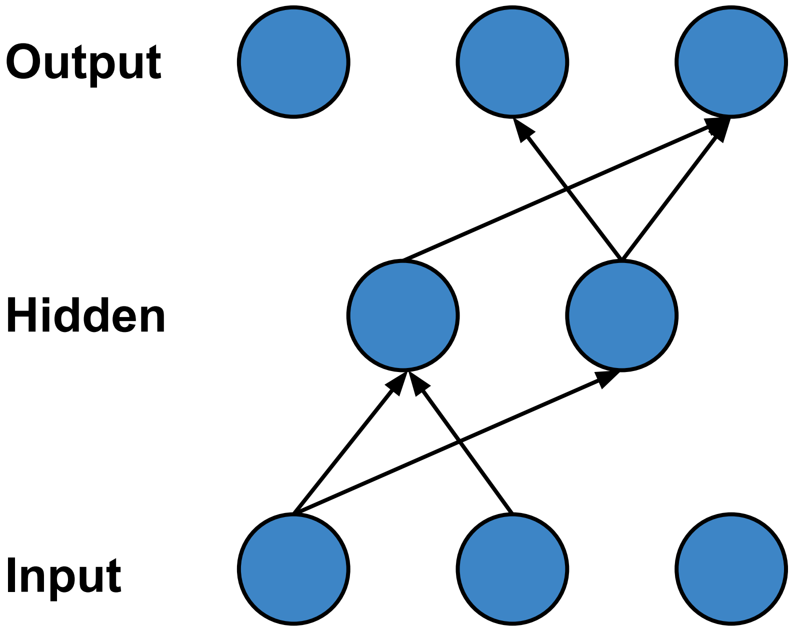


$$Bernoulli(0.9) = 1$$

Ancestral Sampling

$$position = 1$$

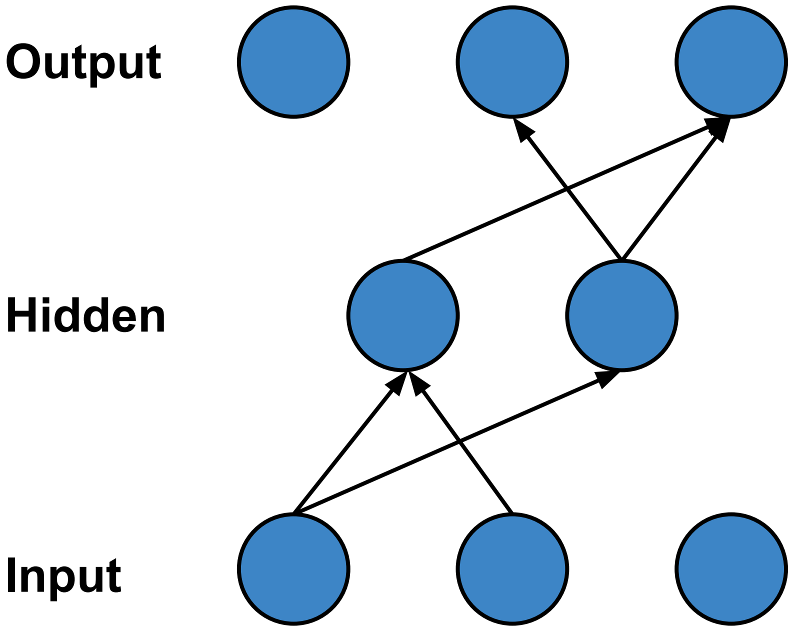
$$\mathbf{x} = \begin{bmatrix} \color{red}{1} \\ 0 \\ 0 \end{bmatrix}$$



Ancestral Sampling

$$position = \color{red}{2}$$

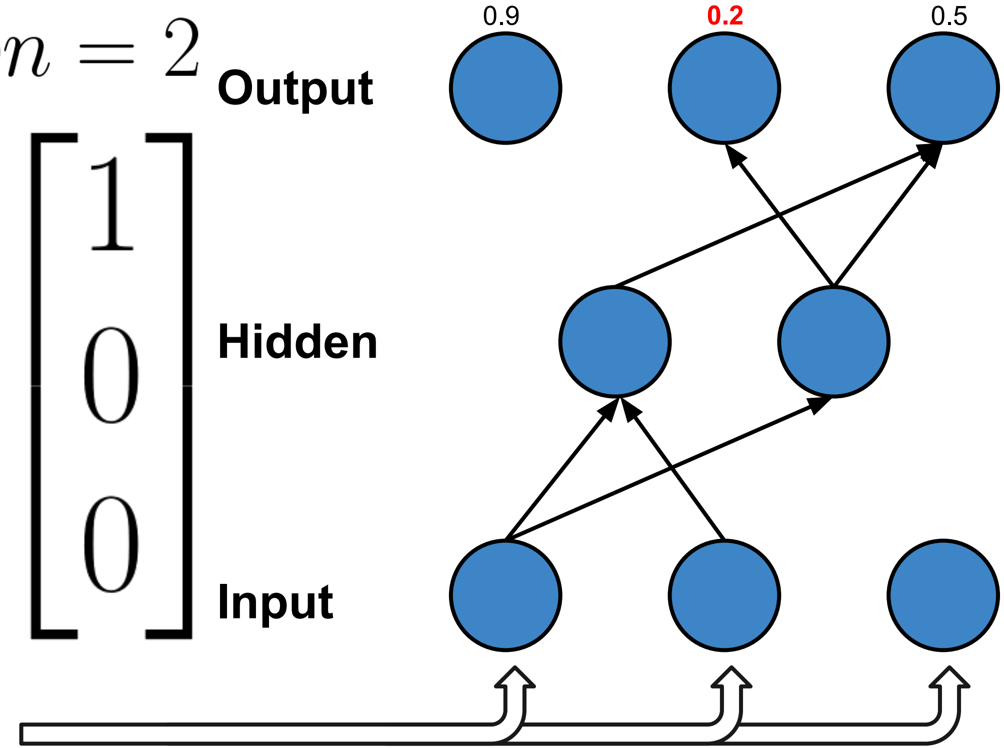
$$\mathbf{x} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$



Ancestral Sampling

$$position = 2$$

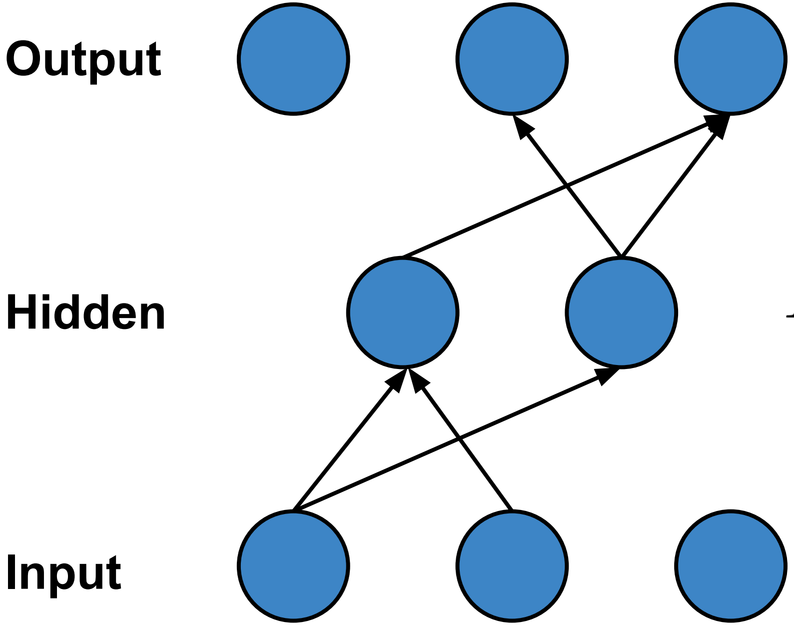
$$\mathbf{x} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$



Ancestral Sampling

$$position = 2$$

$$\mathbf{x} = \begin{bmatrix} 1 \\ \color{red}{0} \\ 0 \end{bmatrix}$$

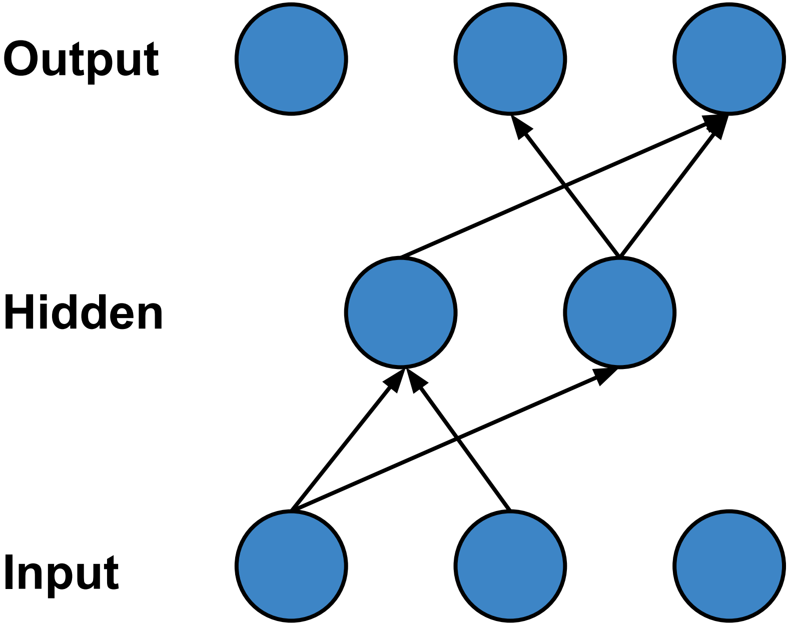


$$Bernoulli(0.2) = 0$$

Ancestral Sampling

$$position = \color{red}{3}$$

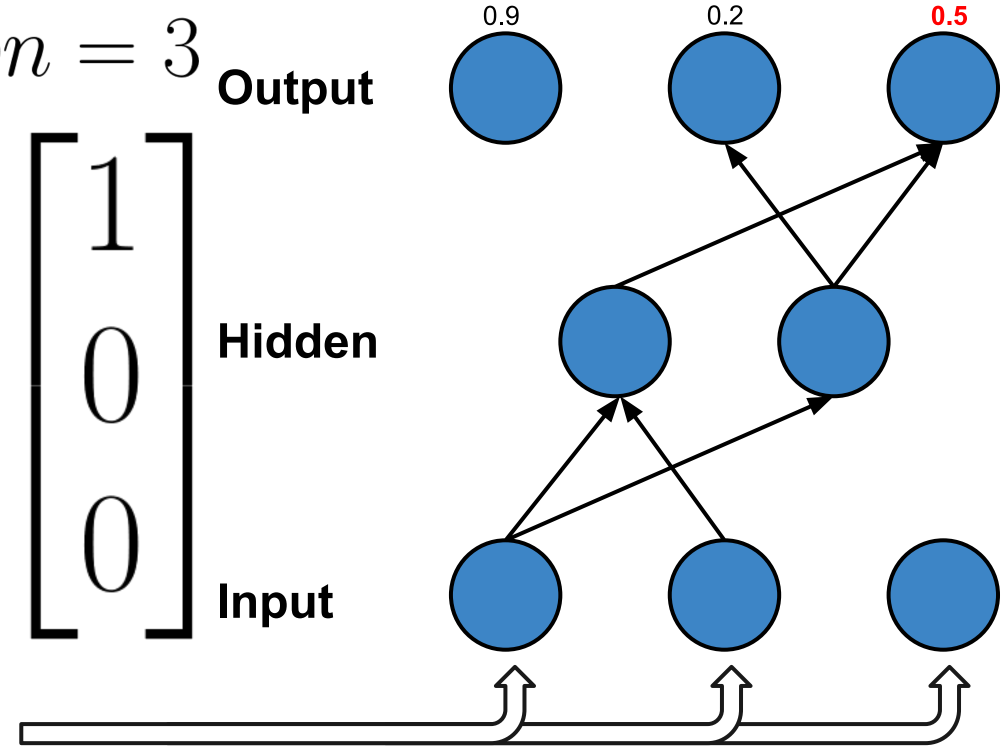
$$\mathbf{x} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$



Ancestral Sampling

$$position = 3$$

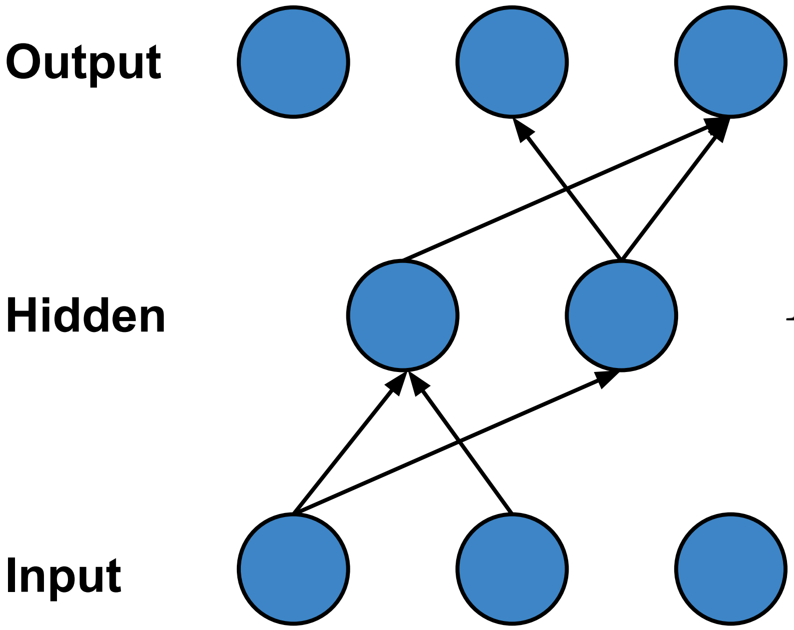
$$\mathbf{x} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$



Ancestral Sampling

$$position = 3$$

$$\mathbf{x} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$



$$Bernoulli(0.5) = 1$$

Experiments

Datasets

- Adult
- Connect 4
- DNA
- Mushrooms
- NIPS
- OCR Letters
- RCV1
- Web

- MNIST

Results : UCI

Negative log-likelihood test results of different models on multiple datasets.
The best result as well as any other result with an overlapping confidence interval is shown in bold.

Model	Adult	Connect4	DNA	Mushrooms	NIPS-0-12	OCR-letters	RCV1	Web
MoBernoullis	20.44	23.41	98.19	14.46	290.02	40.56	47.59	30.16
RBM	16.26	22.66	96.74	15.15	277.37	43.05	48.88	29.38
FVSBN	**13.17**	12.39	83.64	10.27	276.88	39.30	49.84	29.35
NADE (fixed order)	**13.19**	11.99	84.81	9.81	**273.08**	**27.22**	46.66	28.39
EoNADE 1hl (16 ord.)	**13.19**	12.58	82.31	9.69	**272.39**	**27.32**	**46.12**	**27.87**
DARN	13.19	11.91	81.04	**9.55**	274.68	≈28.17	≈**46.10**	≈28.83
MADE	**13.12**	**11.90**	83.63	9.68	280.25	28.34	47.10	28.53
MADE mask sampling	**13.13**	**11.90**	**79.66**	9.69	277.28	30.04	46.74	**28.25**

Results : MNIST

Model	$-\log p$	
RBM (500 h, 25 CD steps)	≈ 86.34	Intractable
DBM 2hl	≈ 84.62	
DBN 2hl	≈ 84.55	
DARN n_h=500	≈ 84.71	
DARN n_h=500, adaNoise	≈ 84.13	
MoBernoullis K=10	168.95	Tractable
MoBernoullis K=500	137.64	
NADE 1hl (fixed order)	88.33	
EoNADE 1hl (128 orderings)	87.71	
EoNADE 2hl (128 orderings)	85.10	
MADE 1hl (1 mask)	88.40	
MADE 2hl (1 mask)	89.59	
MADE 1hl (32 masks)	88.04	
MADE 2hl (32 masks)	86.64	

Conclusion

Generative Autoencoder $\Rightarrow$ MADE!

- Fast
- Tractable
- State of the art

The End!