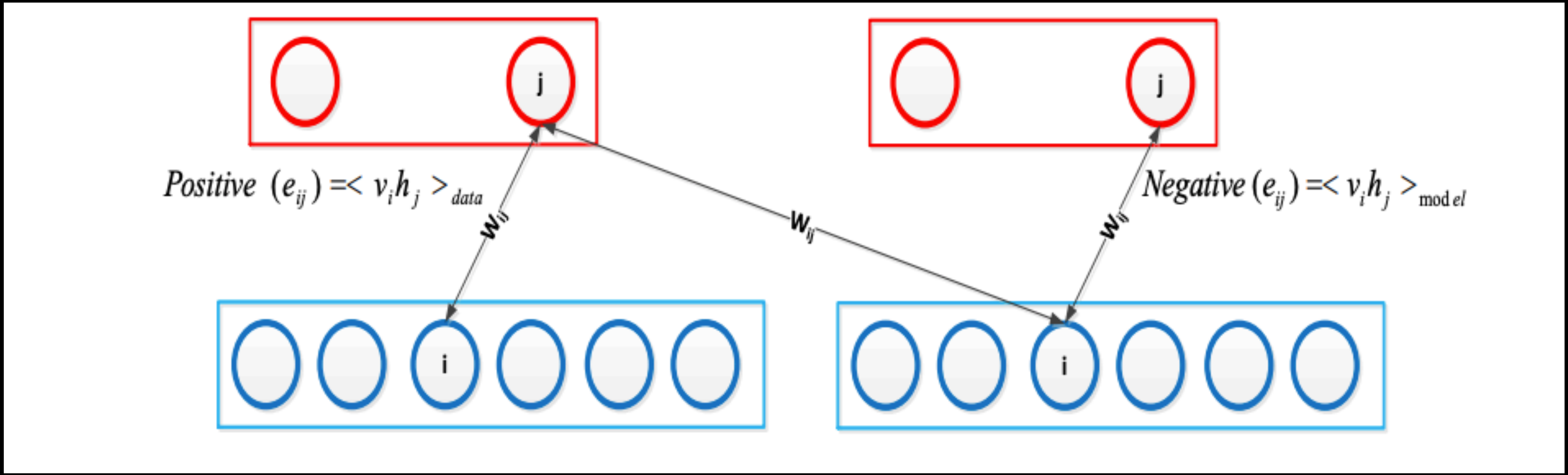

E9: 309 ADL 23-11-2020
http://leap.ee.iisc.ac.in/sriram/teaching/ADL2020/

Housekeeping

- Mid-term exam
  - December 5th (Saturday) [Topics covered up to Dec 2nd]
  - Mode of exam
    - Time to respond - 3 hours
      - Exam paper uploaded in Teams Channel and response (photo-scanned and uploaded in your folder).
      - Open book, open notes
        - Strictly no online communication or help sought.
          - Academic integrity and ethics strongly followed.

Recap of previous class

Contrastive Divergence



One-step contrastive divergence

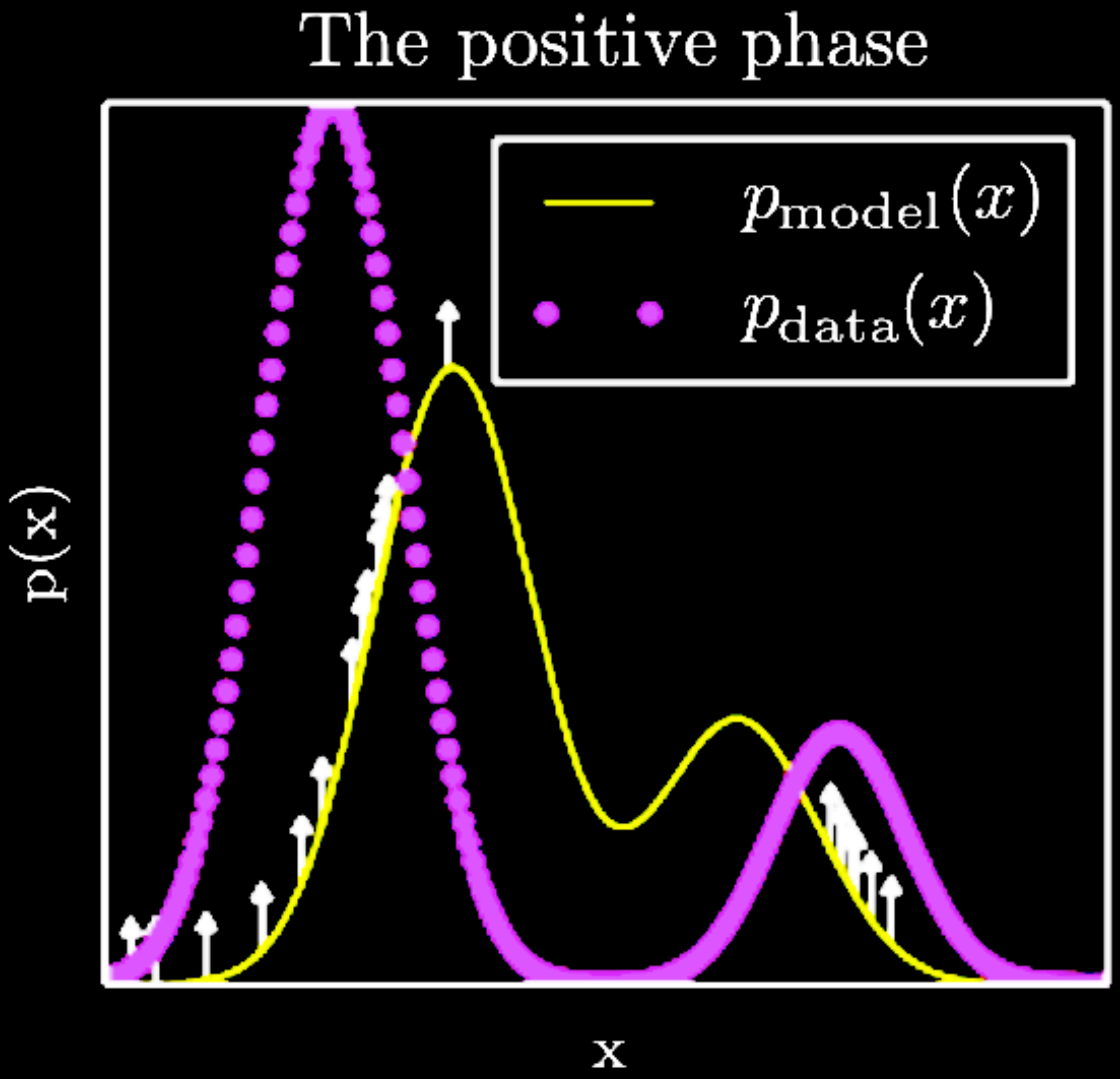
- Computing the gradient

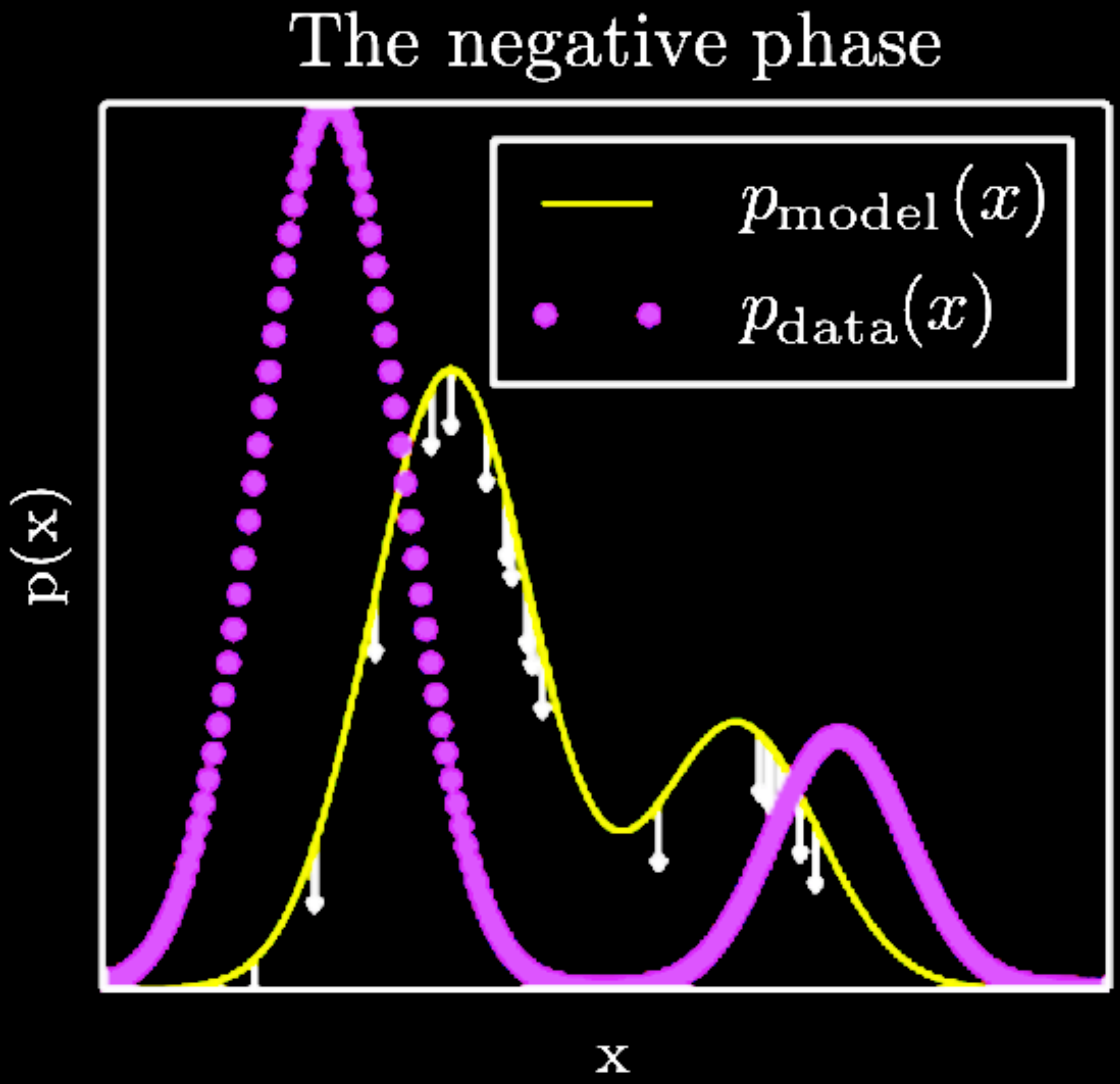
$$\frac{\partial(p([\mathbf{v}\ \ \mathbf{h}], \Theta)}{\partial \mathbf{W}} \approx \frac{1}{N}\sum_{q=1}^{N} \mathbf{v}_q \mathbf{h}_q^T - \frac{1}{N}\sum_{q=1}^{N} \tilde{\mathbf{v}}_q \tilde{\mathbf{h}}_q^T$$

- Performing gradient ascent using the approximate gradient

$$\boldsymbol{\Theta}^{k+1} = \boldsymbol{\Theta}^{k} + \eta \frac{\partial log(p(\mathbf{x}, \Theta))}{\partial \boldsymbol{\Theta}}\Bigg|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}^k}$$

Positive phase and negative phase





Gaussian Bernoulli RBM

- For modeling real observations $\mathbf{v} \in \mathcal{R}^D$
- Define the energy function

$$E[\mathbf{v},\mathbf{h}] = \frac{1}{2}(\mathbf{v}-\mathbf{a})^T(\mathbf{v}-\mathbf{a}) - \mathbf{v}^T\mathbf{W}\mathbf{h} - \mathbf{b}^T\mathbf{h}$$

$$p([\mathbf{v},\mathbf{h}]) = \frac{e^{-E[\mathbf{v},\mathbf{h}]}}{Z}$$

- The conditional distributions

$$p(\mathbf{v}|\mathbf{h}) = \mathcal{N}(\mathbf{W}\mathbf{h}+\mathbf{a},\mathbf{I})$$

$$p(h_j = 1|\mathbf{v}) = \sigma(\mathbf{v}^T\mathbf{W}^k_{:,j} + b_j)$$

Properties of GRBM

- $d = 0$

$$E(\mathbf{v}) = \frac{1}{2}(\mathbf{v} - \mathbf{a})^T(\mathbf{v} - \mathbf{a})$$

$$p(\mathbf{v}) = \frac{e^{-E(\mathbf{v})}}{Z}$$

- The marginal distribution is a Gaussian.

Properties of GRBM

* d = 1

$$E[\mathbf{v}, h] = \frac{1}{2}(\mathbf{v} - \mathbf{a})^T(\mathbf{v} - \mathbf{a}) - h\mathbf{v}^T\mathbf{w} - hb$$

$$p([\mathbf{v}, h]) = \frac{e^{-E[\mathbf{v},h]}}{Z}$$

$$p([\mathbf{v}, h = 0]) = \alpha\mathcal{N}(\mathbf{a}, \mathbf{I})$$

$$p([\mathbf{v}, h = 1]) = (1 - \alpha)\mathcal{N}(\mathbf{a} + \mathbf{w}, \mathbf{I})$$

* The marginal distribution is then

$$p(\mathbf{v}) = p([\mathbf{v}, h = 0]) + p([\mathbf{v}, h = 1])$$

  - ✓ 2-mixture Gaussian

Properties of GRBM

- For any general d dimensions

$$E[\mathbf{v}, \mathbf{h}_d] = E[\mathbf{v}, \mathbf{h}_{d-1}] + h_d\ \mathbf{v}^T \mathbf{W}_{:,d} + b_d h_d$$

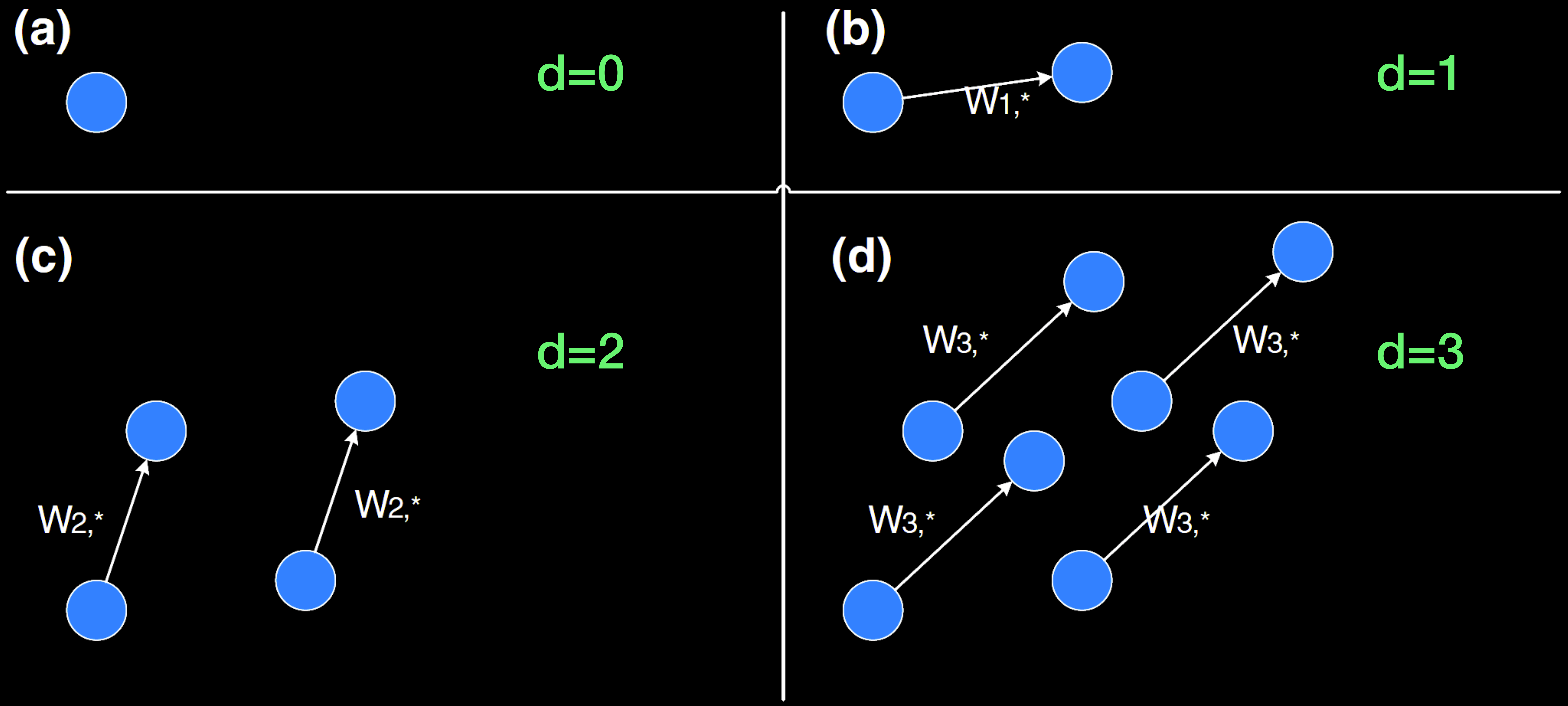
$$p([\mathbf{v}, [\mathbf{h}_{d-1}, h_d = 0]]) = \alpha\ p([\mathbf{v}, \mathbf{h}_{d-1}])$$

$$p([\mathbf{v}, [\mathbf{h}_{d-1}, h_d = 1]]) = (1 - \alpha)\ p([(\mathbf{v} + \mathbf{W}_{:,d}), \mathbf{h}_{d-1}])$$

- For d=0, 1 Gaussian, d=1, 2-mix Gaussian, …

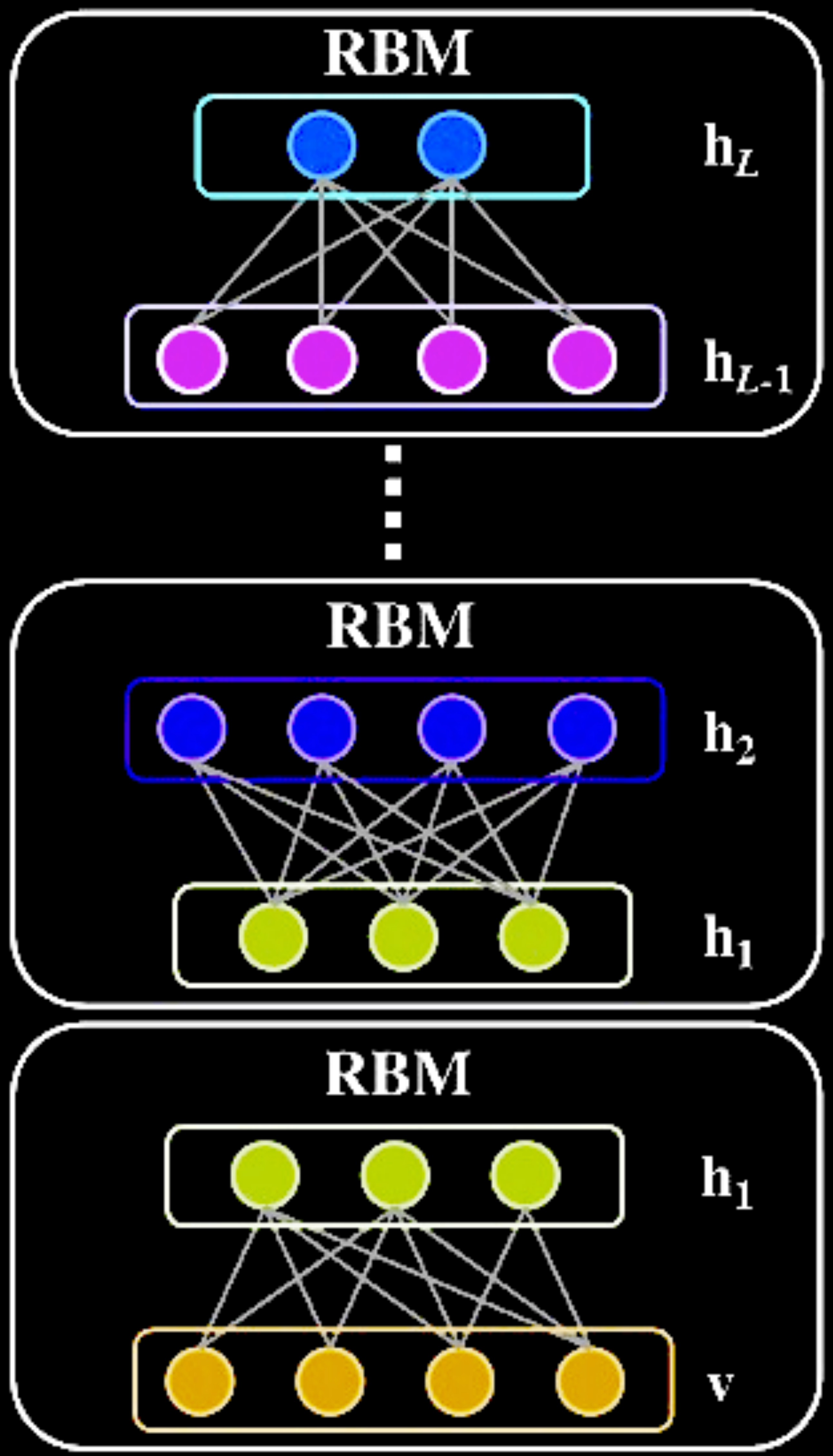
  ➡ 2^d mixture Gaussian for any arbitrary d.

GRBMs and GMMs

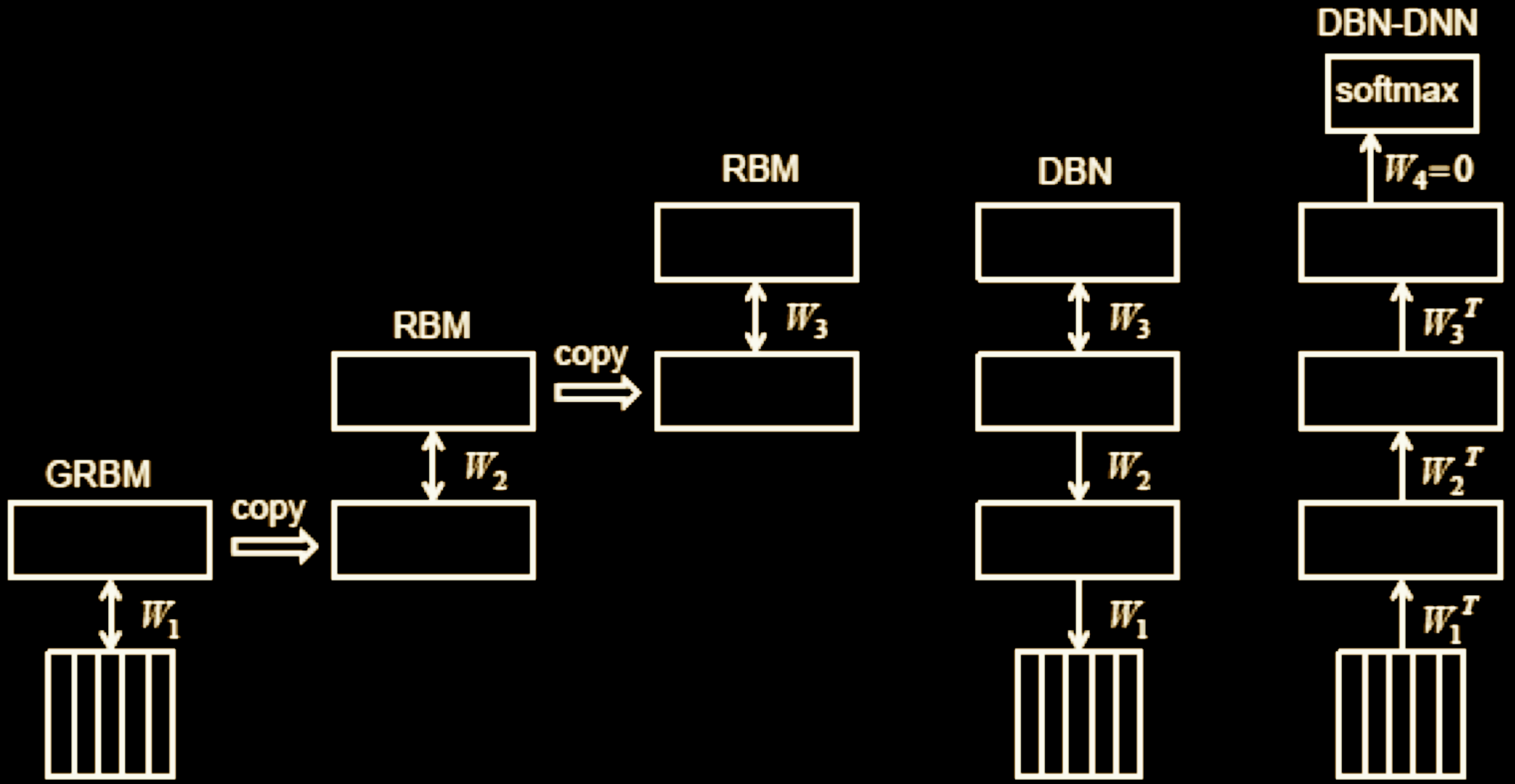


Deep Belief Networks (DBN)

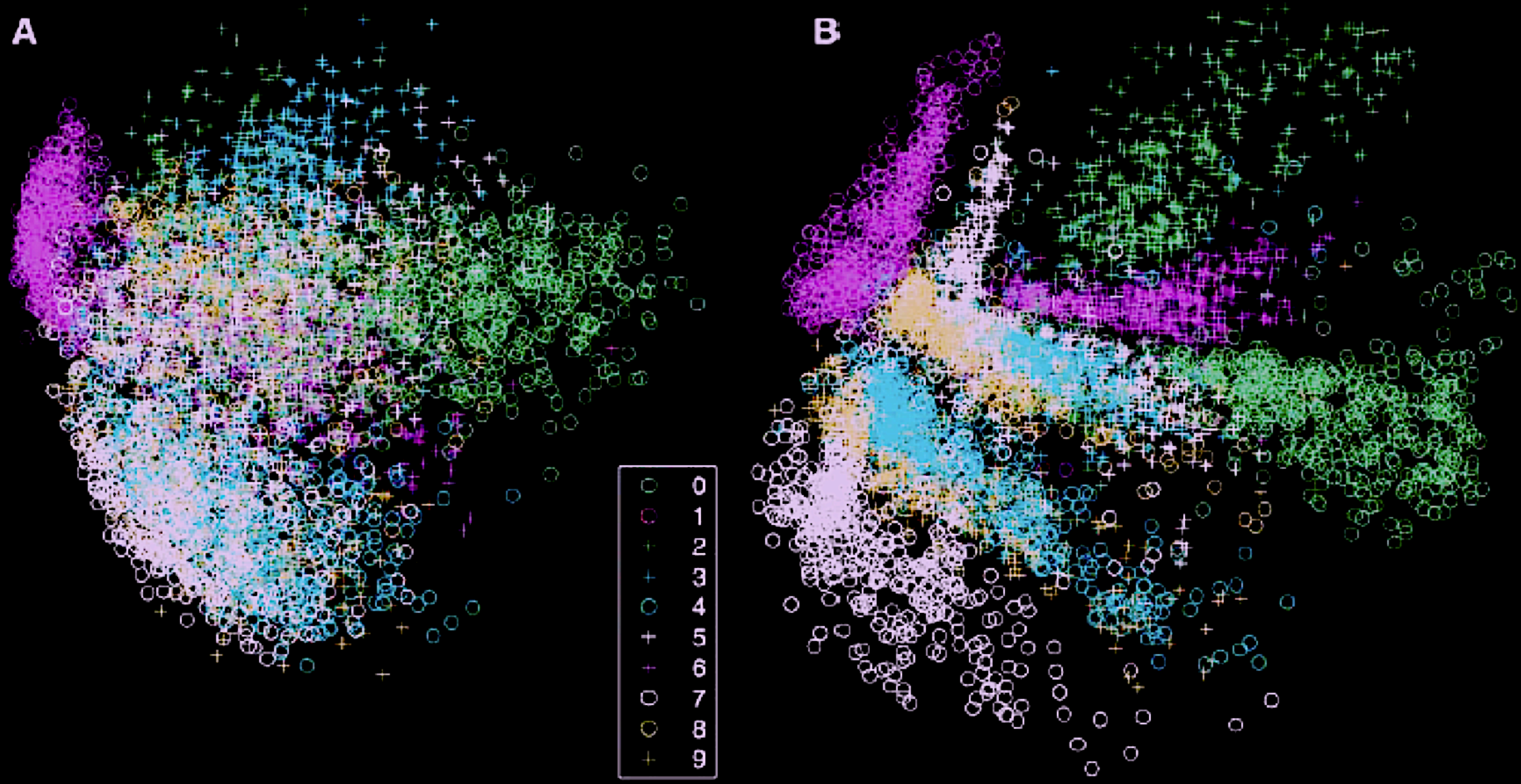
- Stacking RBMs in a disjoint fashion
  - ✓ Layer-wise training for each RBM.
    - ◉ Weights are frozen each layer before training the next layer.
  - ✓ Ancestral sampling can be performed for data generation
    - ★ Lossy sample generation due to accumulation of errors.
- Most common use - pre-training of DNNs.



DBNs for initialization



DBNs for visualization



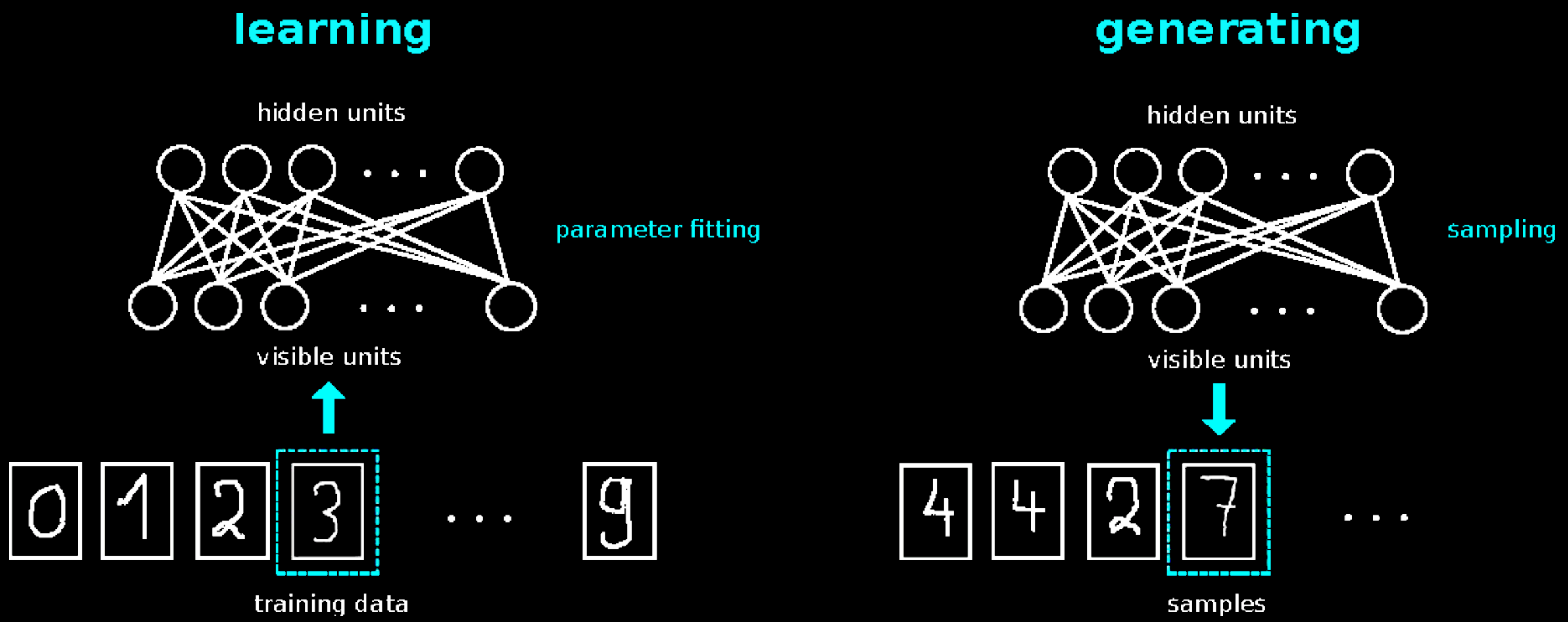
PCA

RBM

More reading

Salakhutdinov, Ruslan, Andriy Mnih, and Geoffrey Hinton. "Restricted Boltzmann machines for collaborative filtering." *Proceedings of the 24th international conference on Machine learning*. 2007.

Data generation using RBMs

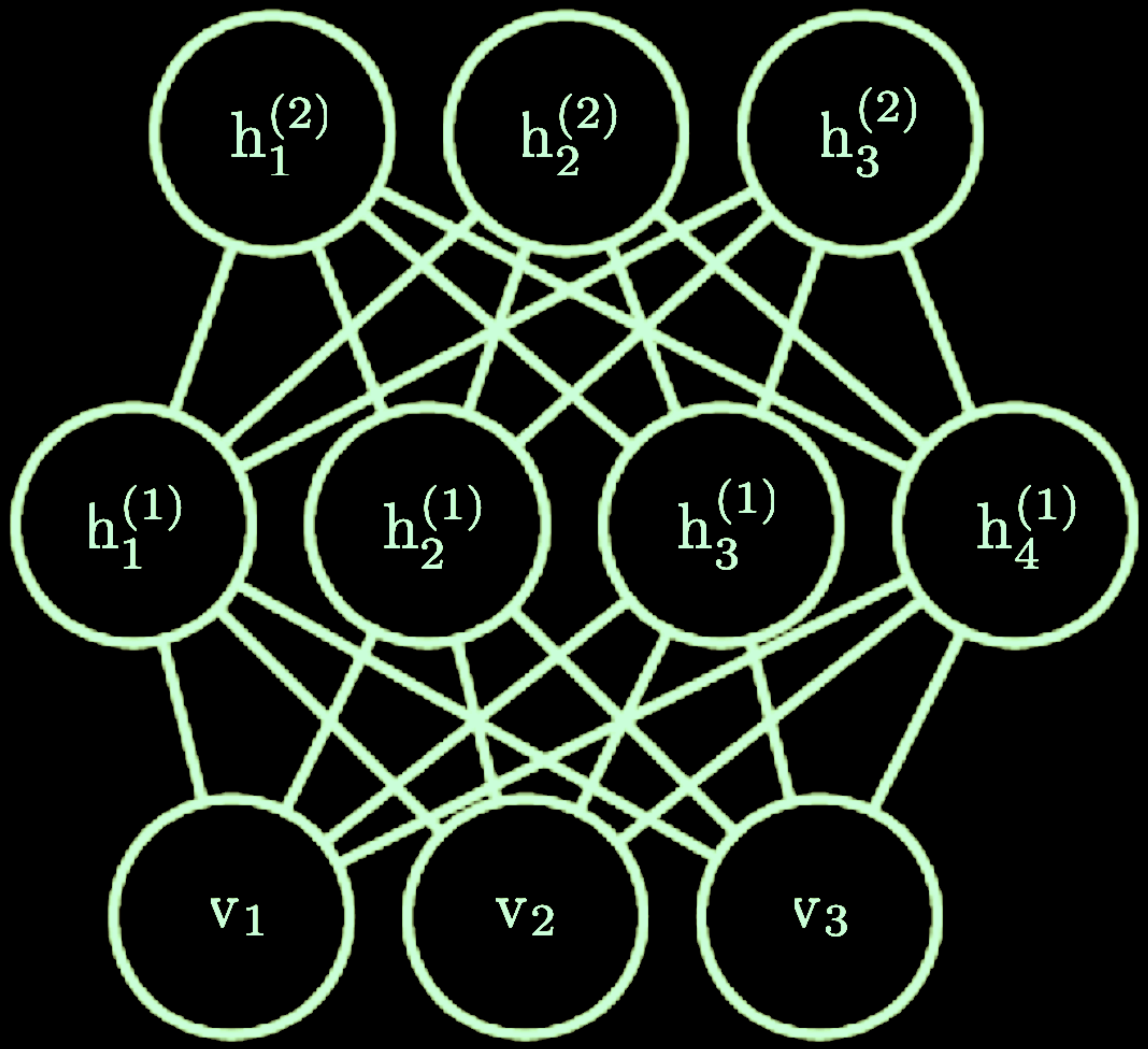


Data generation using RBMs

Source — https://lme.tf.fau.de/lecture-notes/lecture-notes-in-deep-learning-unsupervised-learning-part-1/

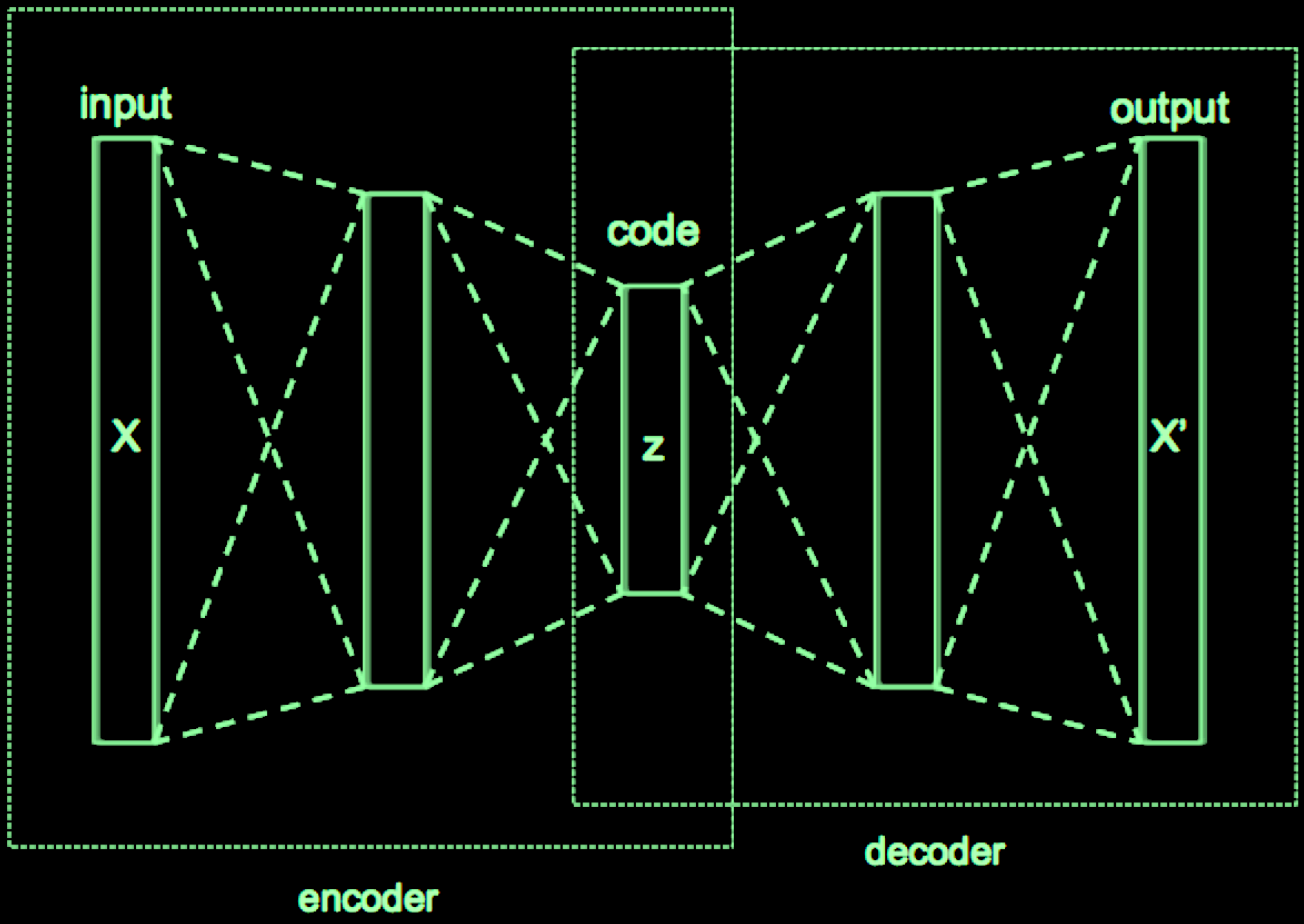
Deep Boltzmann machine

* Deep layers of connections with RBM structure.
  * ➡ Joint energy function.
  * ➡ Undirected graph
* Training and inference are more involved.



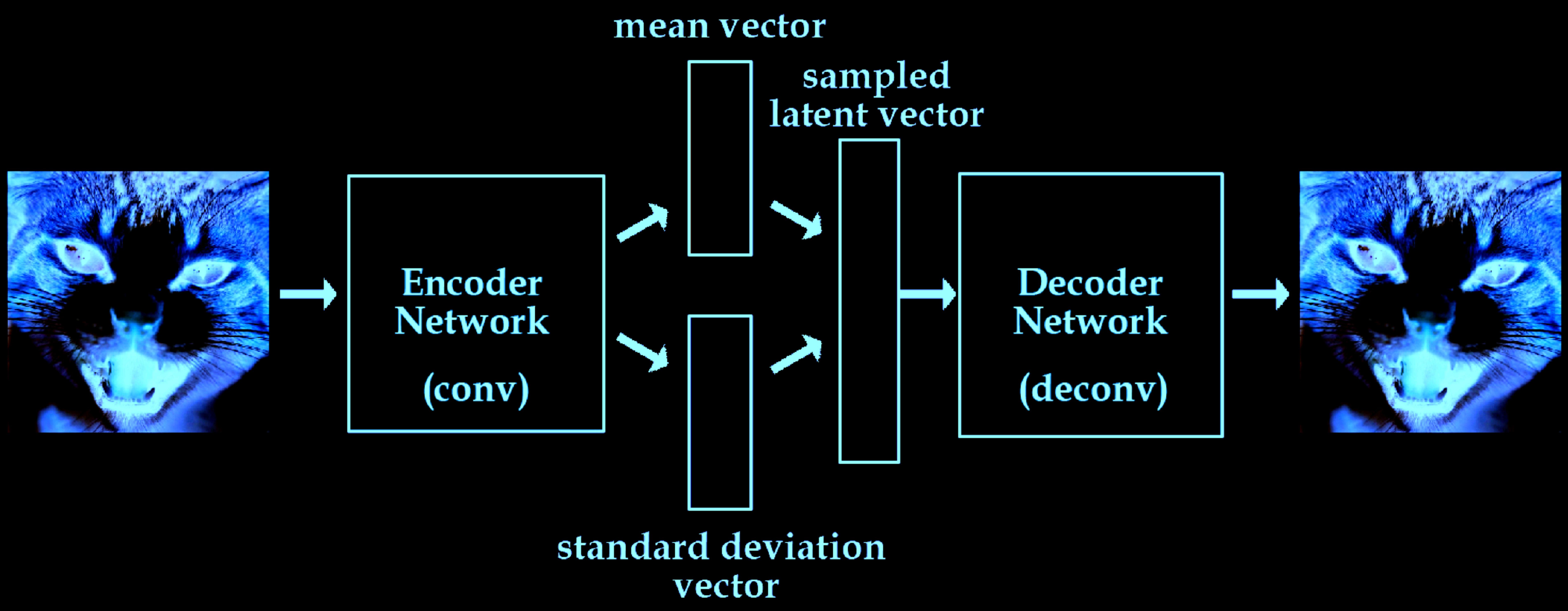
Autoencoders

- Encoder decoder
  - ➡ Latent representations capture a lower dimensional embedding of the data.
    - ✓ can be feedforward or convolutional layers.
- Model training
  - ➡ Using a reconstruction loss.



Variational auto encoders

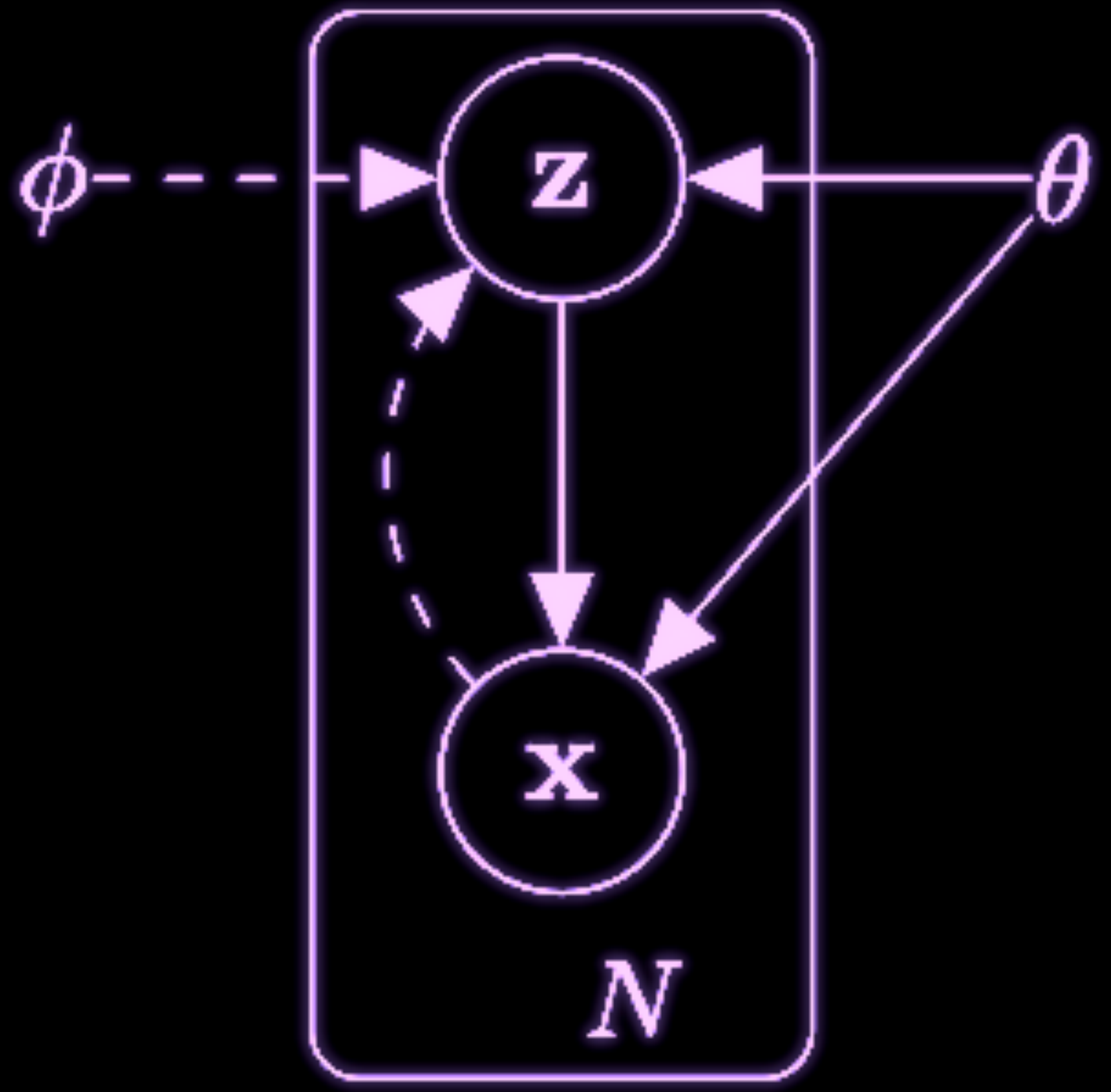
- Sampling from the latent space.
  - ➡ Making Gaussian assumptions at the latent layer



Variational auto encoders

* The data $\mathbf{x}$ and latent variable $\mathbf{z}$
* The forward model
  - ✓ Sample the latent variable
  - ✓ Sample the data given the latent $p_\theta(\mathbf{x}|\mathbf{z})$
* The marginal distribution $\int p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})d\mathbf{z}$
  - ✓ maybe intractable
* The posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$

➡ may also be intractable

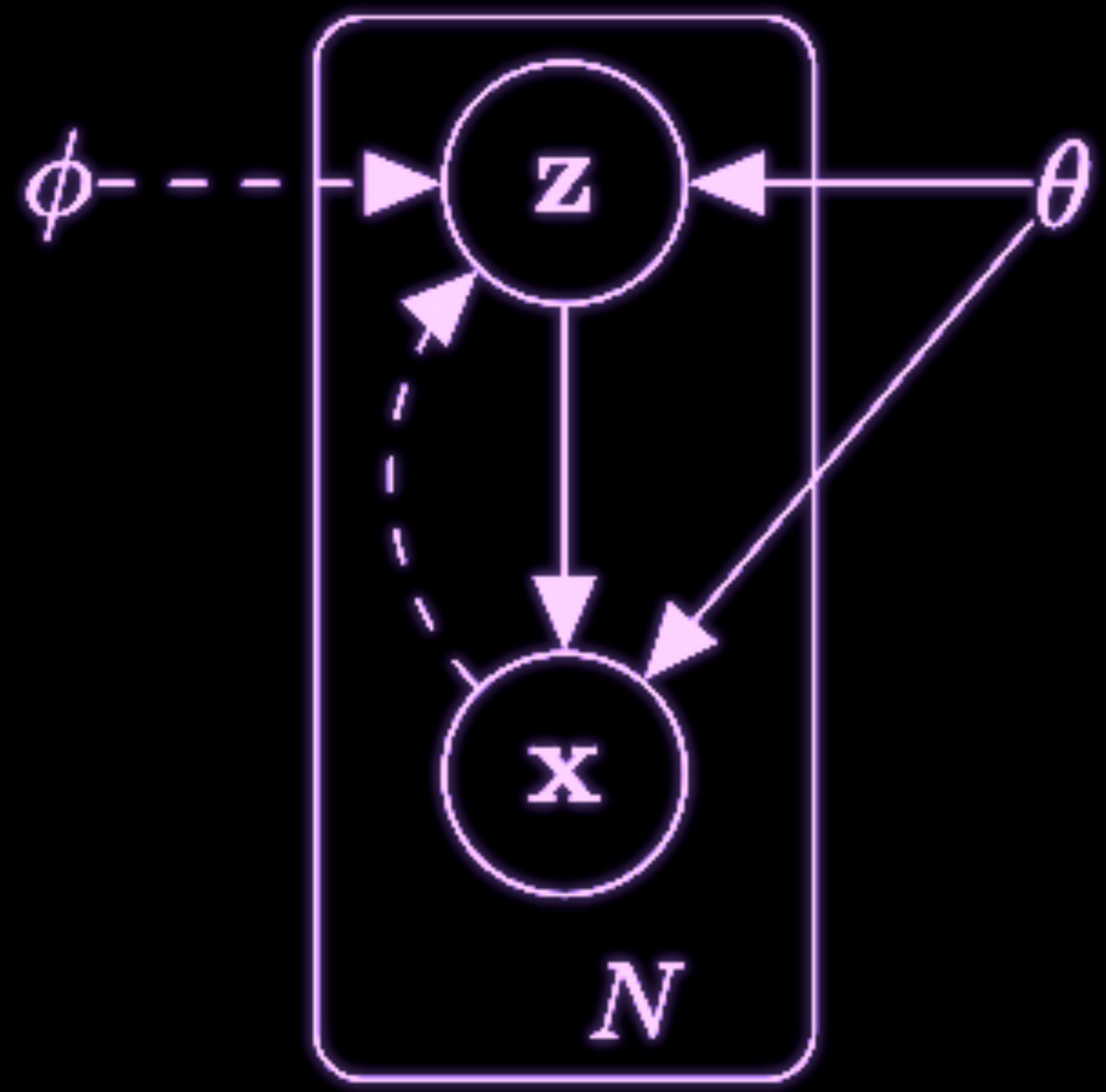


Variational auto encoders

* Approximate the posterior

$$q_\phi(\mathbf{z}|\mathbf{x}) \sim p_\theta(\mathbf{z}|\mathbf{x})$$

➡ Using variational lower bound.



Variational lower bound

$$
\begin{aligned}
log\ p_\theta(\mathbf{x}) &= log \int p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z} \\
&= log \int p_\theta(\mathbf{x}, \mathbf{z}) \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})} d\mathbf{z} \\
&= log \left(\mathbb{E} \left[\frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \right] \right) \\
&\geq \mathbb{E}_q \left(log \left[\frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \right] \right) \\
&= \mathbb{E}_q \left(log \left[\frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \right] \right) \\
&= \mathbb{E}_q \left(log \left[p_\theta(\mathbf{x}, \mathbf{z}) \right] \right) + H_q(\mathbf{z}|\mathbf{x}))
\end{aligned}
$$

Variational lower bound

$$
\begin{aligned}
KL\left[q(Z) \| p(Z|X)\right] &= \int_Z q(Z) \log \frac{q(Z)}{p(Z|X)} \\
&= -\int_Z q(Z) \log \frac{p(Z|X)}{q(Z)} \\
&= -\left(\int_Z q(Z) \log \frac{p(X,Z)}{q(Z)} - \int_Z q(Z) \log p(X) \right) \\
&= -\int_Z q(Z) \log \frac{p(X,Z)}{q(Z)} + \log p(X) \int_Z q(Z) \\
&= -L + \log p(X)
\end{aligned}
$$