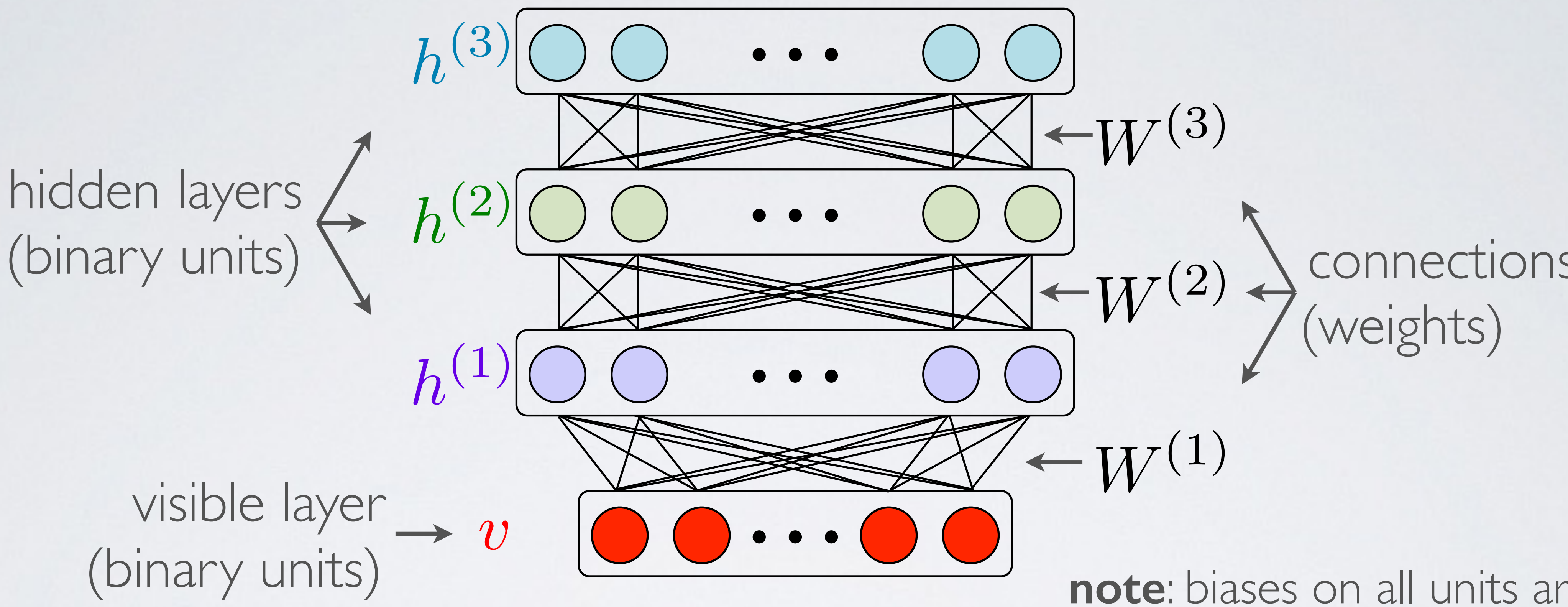
Deep Boltzmann Machines

DEEP BOLTZMANN MACHINES



note: biases on all units are present but suppressed to simplify notation

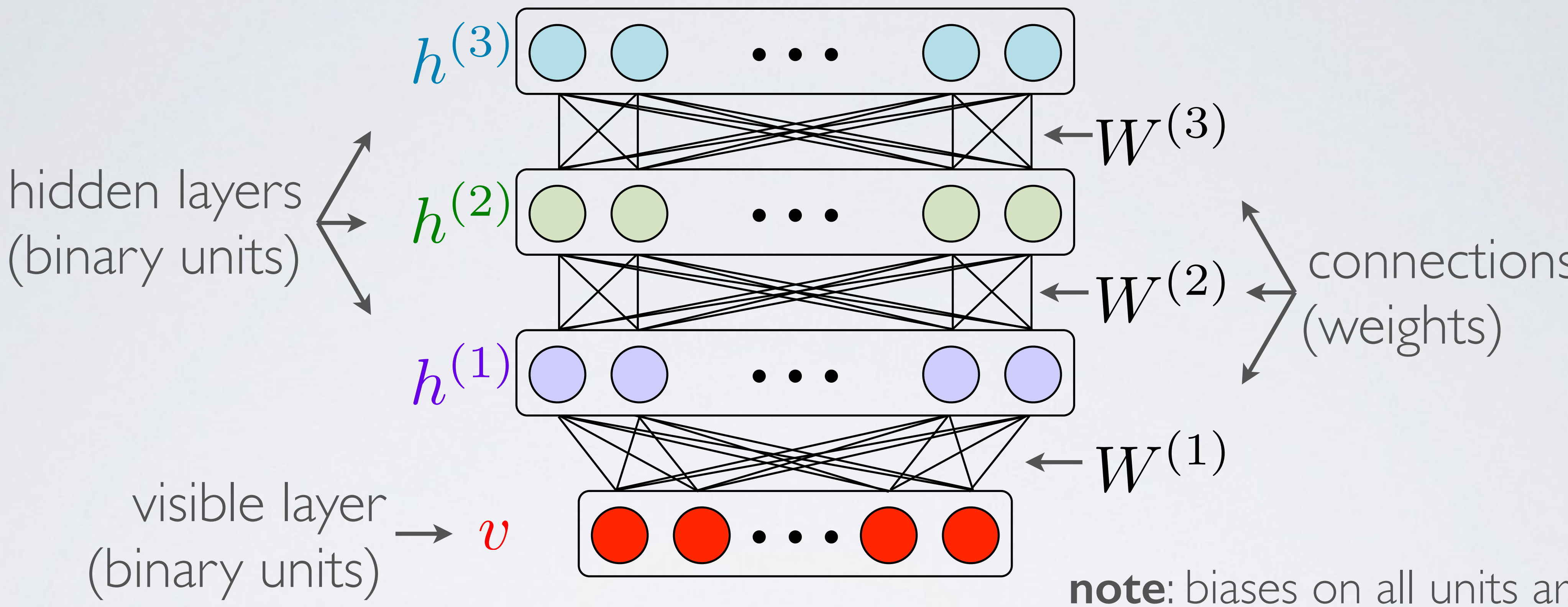
Energy function:

$$E(v, h^{(1)}, h^{(2)}, h^{(3)}; \theta) = -v^T W^{(1)} h^{(1)} - h^{(1)T} W^{(2)} h^{(2)} - h^{(2)T} W^{(3)} h^{(3)}$$

Joint distribution:

$$p\left(v, h^{(1)}, h^{(2)}, h^{(3)}\right) = \frac{1}{Z(\theta)} \exp\left(-E(v, h^{(1)}, h^{(2)}, h^{(3)}; \theta)\right)$$

DEEP BOLTZMANN MACHINES



note: biases on all units are present but suppressed to simplify notation

Bipartite structure:

- Undirected connections between neighbouring layers.
  - eg. $h^{(2)}$ is connected only to $h^{(1)}$ and $h^{(3)}$
- No connections between the nodes in the same layer.

DBM: CONDITIONAL DISTRIBUTION

- DBM joint distribution:

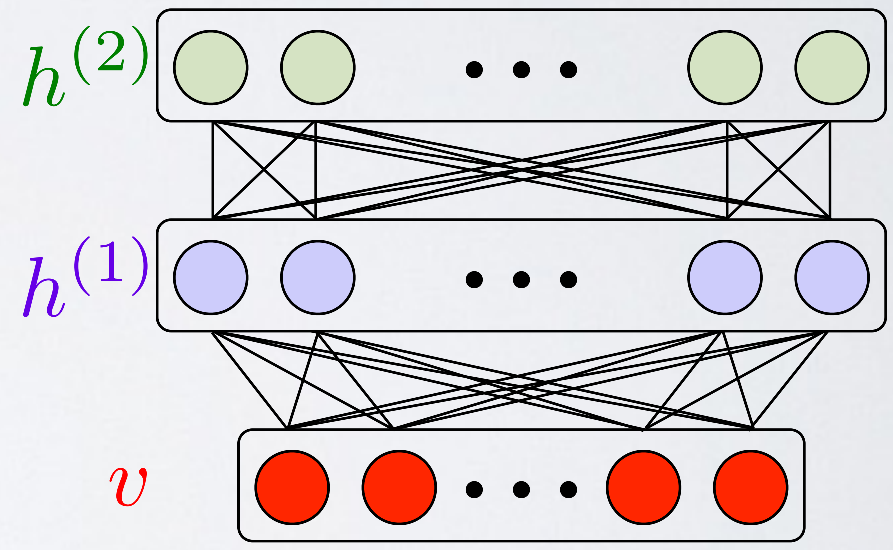
$$p(v, h^{(1)}, h^{(2)}) = \frac{1}{Z} \exp\left\{ v^T W^{(1)} h^{(1)} + h^{(1)T} W^{(2)} h^{(2)} \right\}$$

- DBM Property - Conditional distribution factorize (like RBMs)

$$p(v_i = 1 \mid h^{(1)}) = \mathrm{sigm}\left(\sum_j W_{ij}^{(1)} h_j^{(1)} \right)$$

$$p(h_k^{(2)} = 1 \mid h^{(1)}) = \mathrm{sigm}\left(\sum_j W_{jk}^{(2)} h_j^{(1)} \right)$$

$$p(h_j^{(1)} = 1 \mid v, h^{(2)}) = \mathrm{sigm}\left(\sum_i W_{ij}^{(1)} v_i + \sum_k W_{jk}^{(2)} h_k^{(2)} \right)$$



DBM: INFERENCE

- Unlike the RBM, inference in the DBM is intractable.
    - That is, computing the posterior $p(h^{(1)}, h^{(2)} \mid v)$ is intractable. Why?

DBM: INFERENCE

- Unlike the RBM, inference in the DBM is intractable.
  - That is, computing the posterior $p(h^{(1)}, h^{(2)} \mid v)$ is intractable. Why?
  - Because the latent variables, h, are not independent given an observed v.

$$p(h^{(1)}, h^{(2)} \mid v) = \frac{1}{Z'} \exp\left\{ v^T W^{(1)} h^{(1)} + h^{(1)T} W^{(2)} h^{(2)} \right\}$$

$W^{(2)}$ induces conditional dependencies between $h^{(1)}$ and $h^{(2)}$

DBM: INFERENCE

- Unlike the RBM, inference in the DBM is intractable.
  - That is, computing the posterior $p(h^{(1)}, h^{(2)} \mid v)$ is intractable. Why?
  - Because the latent variables, h, are not independent given an observed v.

$$p(h^{(1)}, h^{(2)} \mid v) = \frac{1}{Z'} \exp\left\{ v^T W^{(1)} h^{(1)} + h^{(1)T} W^{(2)} h^{(2)} \right\}$$

$W^{(2)}$ induces conditional dependencies between $h^{(1)}$ and $h^{(2)}$

- **Strategy**: use a (variational) mean-field approximation to the posterior distribution $p(h^{(1)}, h^{(2)} \mid v)$.

DBM: APPROXIMATE INFERENCE

- Mean-field approximate inference:

DBM: APPROXIMATE INFERENCE

- Mean-field approximate inference:
  - Find the approximating dist. $q_v(h^{(1)}, h^{(2)})$ that best fits $p(h^{(1)}, h^{(2)} \mid v)$

DBM: APPROXIMATE INFERENCE

- Mean-field approximate inference:
  - Find the approximating dist. $q_v(h^{(1)}, h^{(2)})$ that best fits $p(h^{(1)}, h^{(2)} \mid v)$
  - **Mean-field assumption:** $q_v(h^{(1)}, h^{(2)}) = \prod_j q_v(h_j^{(1)}) \prod_k q_v(h_k^{(2)})$
    - approximating distribution has only independent elements.

DBM: APPROXIMATE INFERENCE

- Mean-field approximate inference:
  - Find the approximating dist. $q_v(h^{(1)}, h^{(2)})$ that best fits $p(h^{(1)}, h^{(2)} \mid v)$
  - **Mean-field assumption:** $q_v(h^{(1)}, h^{(2)}) = \prod_j q_v(h_j^{(1)}) \prod_k q_v(h_k^{(2)})$
    - approximating distribution has only independent elements.
  - Choose $q_v(h^{(1)}, h^{(2)})$ that minimizes the KL divergence:

$$\mathrm{KL}(q \| p) = - \sum_{h^{(1)}, h^{(2)}} q_v(h^{(1)}, h^{(2)}) \ln \left(\frac{p(h^{(1)}, h^{(2)} \mid v)}{q_v(h^{(1)}, h^{(2)})} \right)$$

DBM: APPROXIMATE INFERENCE

- Parameterization of $q_v(h^{(1)}, h^{(2)})$:

DBM: APPROXIMATE INFERENCE

- Parameterization of $q_v(h^{(1)}, h^{(2)})$:

$$\hat{h}_j^{(1)} \equiv q_v(h_j^{(1)} = 1), \quad \hat{h}_k^{(2)} \equiv q_v(h_k^{(2)} = 1)$$

DBM: APPROXIMATE INFERENCE

- Parameterization of $q_v(h^{(1)}, h^{(2)})$:

$$\hat{h}_j^{(1)} \equiv q_v(h_j^{(1)} = 1), \quad \hat{h}_k^{(2)} \equiv q_v(h_k^{(2)} = 1)$$

- Combine with MF assumption: $q_v(h^{(1)}, h^{(2)}) = \prod_j q_v(h_j^{(1)}) \prod_k q_v(h_k^{(2)})$

DBM: APPROXIMATE INFERENCE

- Parameterization of $q_v(h^{(1)}, h^{(2)})$:

$$\hat{h}_j^{(1)} \equiv q_v(h_j^{(1)} = 1), \quad \hat{h}_k^{(2)} \equiv q_v(h_k^{(2)} = 1)$$

  - Combine with MF assumption: $q_v(h^{(1)}, h^{(2)}) = \prod_j q_v(h_j^{(1)}) \prod_k q_v(h_k^{(2)})$

... and we get:

$$\boxed{\begin{aligned} q_v(h^{(1)}, h^{(2)}) = & \prod_j (\hat{h}_j^{(1)})^{h_j^{(1)}} (1 - \hat{h}_j^{(1)})^{(1-h_j^{(1)})} \\ & \times \prod_k (\hat{h}_k^{(2)})^{h_k^{(2)}} (1 - \hat{h}_k^{(2)})^{(1-h_k^{(2)})} \end{aligned}}$$

DBM: APPROXIMATE INFERENCE

- Plugging in this $q_v(h^{(1)}, h^{(2)})$ into the KL divergence:

$$\mathrm{KL}(q\|p) = -\sum_{h^{(1)},h^{(2)}} q_v(h^{(1)}, h^{(2)}) \ln\left(\frac{p(h^{(1)}, h^{(2)} \mid v)}{q_v(h^{(1)}, h^{(2)})}\right)$$

- and optimize it w.r.t. the parameters of $q_v(h^{(1)}, h^{(2)})$, i.e. solve the system of equations:

$$\frac{\partial}{\partial \hat{h}^{(1)}} \mathrm{KL}(q\|p) = 0 \quad \text{and} \quad \frac{\partial}{\partial \hat{h}^{(2)}} \mathrm{KL}(q\|p) = 0$$

- Defines iterative update equations (convergence to local fixed point):

$$\hat{h}_j^{(1)} = \mathrm{sigm}\left(\sum_i W_{ij}^{(1)} v_i + \sum_k W_{jk}^{(2)} \hat{h}_k^{(2)}\right), \quad \hat{h}_k^{(2)} = \mathrm{sigm}\left(\sum_j W_{jk}^{(2)} \hat{h}_j^{(1)}\right)$$

DBM: LEARNING

DBM: LEARNING

Energy function: $E(v, h^{(1)}, h^{(2)}; \theta) = -v^T W^{(1)} h^{(1)} - h^{(1)T} W^{(2)} h^{(2)}$

Joint distribution: $p\left(v, h^{(1)}, h^{(2)}\right) = \frac{1}{Z(\theta)} \exp\left(-E(v, h^{(1)}, h^{(2)}; \theta)\right)$

DBM: LEARNING

Energy function: $E(v, h^{(1)}, h^{(2)}; \theta) = -v^T W^{(1)} h^{(1)} - h^{(1)T} W^{(2)} h^{(2)}$

Joint distribution: $p\left(v, h^{(1)}, h^{(2)}\right) = \frac{1}{Z(\theta)} \exp\left(-E(v, h^{(1)}, h^{(2)}; \theta)\right)$

Marginal distribution: $p\left(v\right) = \sum_{h^{(1)}, h^{(2)}} \frac{1}{Z(\theta)} \exp\left(-E(v, h^{(1)}, h^{(2)}; \theta)\right)$

Partition function: $Z(\theta) = \sum_{v, h^{(1)}, h^{(2)}} \exp\left(-E(v, h^{(1)}, h^{(2)}; \theta)\right)$

DBM: LEARNING

Marginal distribution: $$p(v) = \sum_{h^{(1)}, h^{(2)}} \frac{1}{Z(\theta)} \exp\left(-E(v, h^{(1)}, h^{(2)}; \theta)\right)$$

Partition function: $$Z(\theta) = \sum_{v, h^{(1)}, h^{(2)}} \exp\left(-E(v, h^{(1)}, h^{(2)}; \theta)\right)$$

DBM: LEARNING

Marginal distribution: $p(v) = \sum_{h^{(1)}, h^{(2)}} \frac{1}{Z(\theta)} \exp\left(-E(v, h^{(1)}, h^{(2)}; \theta)\right)$

Partition function: $Z(\theta) = \sum_{v, h^{(1)}, h^{(2)}} \exp\left(-E(v, h^{(1)}, h^{(2)}; \theta)\right)$

- Maximum likelihood estimation via gradient descent (as in RBMs):

$$\frac{\partial \ln p(v)}{\partial \theta} = \frac{\partial}{\partial \theta} \ln \left(\frac{1}{Z(\theta)} \sum_{h^{(1)}, h^{(2)}} \exp\left\{-E(v, h^{(1)}, h^{(2)}, \theta)\right\} \right)$$

$$= \frac{\partial}{\partial \theta} \ln \left(\sum_{h^{(1)}, h^{(2)}} \exp\left\{-E(v, h^{(1)}, h^{(2)}, \theta)\right\} \right) - \frac{\partial}{\partial \theta} \ln Z(\theta)$$

DBM: LEARNING

- Maximum likelihood estimation via gradient descent (as in RBMs):

$$\frac{\partial \ln p(v)}{\partial \theta} = \frac{\partial}{\partial \theta} \ln \left(\frac{1}{Z(\theta)} \sum_{h^{(1)},h^{(2)}} \exp\left\{ -E(v, h^{(1)}, h^{(2)}, \theta) \right\} \right)$$

$$= \frac{\partial}{\partial \theta} \ln \left(\sum_{h^{(1)},h^{(2)}} \exp\left\{ -E(v, h^{(1)}, h^{(2)}, \theta) \right\} \right) - \frac{\partial}{\partial \theta} \ln Z(\theta)$$

DBM: LEARNING

- Maximum likelihood estimation via gradient descent (as in RBMs):

$$\frac{\partial \ln p(v)}{\partial \theta} = \frac{\partial}{\partial \theta} \ln \left(\frac{1}{Z(\theta)} \sum_{h^{(1)},h^{(2)}} \exp\left\{-E(v, h^{(1)}, h^{(2)}, \theta)\right\} \right)$$

$$= \underbrace{\frac{\partial}{\partial \theta} \ln \left(\sum_{h^{(1)},h^{(2)}} \exp\left\{-E(v, h^{(1)}, h^{(2)}, \theta)\right\} \right)} - \underbrace{\frac{\partial}{\partial \theta} \ln Z(\theta)}$$

$$\frac{\partial \ln p(v)}{\partial \theta} = \boxed{-\mathbb{E}_{p(h^{(1)},h^{(2)}|v)} \frac{\partial}{\partial \theta} E(v, h^{(1)}, h^{(2)}, \theta)} + \boxed{\mathbb{E}_{p(v,h^{(1)},h^{(2)})} \frac{\partial}{\partial \theta} E(v, h^{(1)}, h^{(2)}, \theta)}$$

Data term, a.k.a. positive phase

Model term, a.k.a. negative phase

DBM: LEARNING

- Maximum likelihood estimation via gradient descent (as in RBMs):

$$\frac{\partial \ln p(v)}{\partial \theta} = \boxed{-\mathbb{E}_{p(h^{(1)},h^{(2)}|v)} \frac{\partial}{\partial \theta} E(v, h^{(1)}, h^{(2)}, \theta)} + \boxed{\mathbb{E}_{p(v,h^{(1)},h^{(2)})} \frac{\partial}{\partial \theta} E(v, h^{(1)}, h^{(2)}, \theta)}$$

Data term, a.k.a. positive phase

Model term, a.k.a. negative phase

DBM: LEARNING

- Maximum likelihood estimation via gradient descent (as in RBMs):

$$\frac{\partial \ln p(v)}{\partial \theta} = \boxed{-\mathbb{E}_{p(h^{(1)},h^{(2)}|v)} \frac{\partial}{\partial \theta} E(v, h^{(1)}, h^{(2)}, \theta)} + \boxed{\mathbb{E}_{p(v,h^{(1)},h^{(2)})} \frac{\partial}{\partial \theta} E(v, h^{(1)}, h^{(2)}, \theta)}$$

Data term, a.k.a. positive phase

Model term, a.k.a. negative phase

- For the DBM (unlike the RBM), the expectation in **both** the data term and the model term are intractable.

- How are we going to approximate these expectations?

DBM: LEARNING

- Maximum likelihood estimation via gradient descent (as in RBMs):

$$\frac{\partial \ln p(v)}{\partial \theta} = \boxed{-\mathbb{E}_{p(h^{(1)},h^{(2)}|v)} \frac{\partial}{\partial \theta} E(v, h^{(1)}, h^{(2)}, \theta)} + \boxed{\mathbb{E}_{p(v,h^{(1)},h^{(2)})} \frac{\partial}{\partial \theta} E(v, h^{(1)}, h^{(2)}, \theta)}$$

Data term, a.k.a. positive phase

Model term, a.k.a. negative phase

Approach of Salakhutdinov & Hinton (2009):

Mean-field approximation: we assume posterior dist. $p(h^{(1)},\ h^{(2)} \mid v)$ is relatively simple, i.e. unimodal.

Monte Carlo approximation: we assume the joint dist. $p(v,\ h^{(1)},\ h^{(2)})$ is much more complex, i.e. multimodal.

➡ Exactly as in Persistent-CD when training an RBM

VARIATIONAL APPROACH

- How can we justify this combination of variational inference and maximum likelihood?
- Variational methods are based on the relationship:

$$\ln p(v) = \ln p(v) + \sum_h q(h \mid v) \ln \left(\frac{p(v,h)}{q(h \mid v)} \right) - \sum_h q(h \mid v) \ln \left(\frac{p(v,h)}{q(h \mid v)} \right)$$

VARIATIONAL APPROACH

- How can we justify this combination of variational inference and maximum likelihood?
- Variational methods are based on the relationship:

$$
\begin{aligned}
\ln p(v) &= \ln p(v) + \sum_h q(h \mid v) \ln \left(\frac{p(v,h)}{q(h \mid v)} \right) - \sum_h q(h \mid v) \ln \left(\frac{p(v,h)}{q(h \mid v)} \right) \\
&= \sum_h q(h \mid v) \ln \left(\frac{p(v,h)}{q(h \mid v)} \right) - \sum_h q(h \mid v) \ln \left(\frac{p(v,h)}{q(h \mid v)} \right) + \sum_h q(h \mid v) \log p(v)
\end{aligned}
$$

VARIATIONAL APPROACH

- How can we justify this combination of variational inference and maximum likelihood?
- Variational methods are based on the relationship:

$$
\begin{aligned}
\ln p(v) &= \ln p(v) + \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) - \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) \\
&= \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) - \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) + \sum_h q(h \mid v) \log p(v) \\
&= \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) - \sum_h q(h \mid v) \left[\ln\left(\frac{p(v,h)}{q(h \mid v)}\right) - \ln p(v)\right]
\end{aligned}
$$

VARIATIONAL APPROACH

- How can we justify this combination of variational inference and maximum likelihood?
- Variational methods are based on the relationship:

$$\begin{aligned}
\ln p(v) &= \ln p(v) + \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) - \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) \\
&= \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) - \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) + \sum_h q(h \mid v) \log p(v) \\
&= \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) - \sum_h q(h \mid v) \left[\ln\left(\frac{p(v,h)}{q(h \mid v)}\right) - \ln p(v)\right] \\
&= \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) - \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{p(v)q(h \mid v)}\right)
\end{aligned}$$

VARIATIONAL APPROACH

- How can we justify this combination of variational inference and maximum likelihood?
- Variational methods are based on the relationship:

$$
\begin{aligned}
\ln p(v) &= \ln p(v) + \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) - \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) \\
&= \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) - \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) + \sum_h q(h \mid v) \log p(v) \\
&= \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) - \sum_h q(h \mid v) \left[\ln\left(\frac{p(v,h)}{q(h \mid v)}\right) - \ln p(v)\right] \\
&= \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) - \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{p(v)q(h \mid v)}\right) \\
&= \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) - \sum_h q(h \mid v) \ln\left(\frac{p(h \mid v)}{q(h \mid v)}\right)
\end{aligned}
$$

VARIATIONAL APPROACH

- How can we justify this combination of variational inference and maximum likelihood?
- Variational methods are based on the relationship:

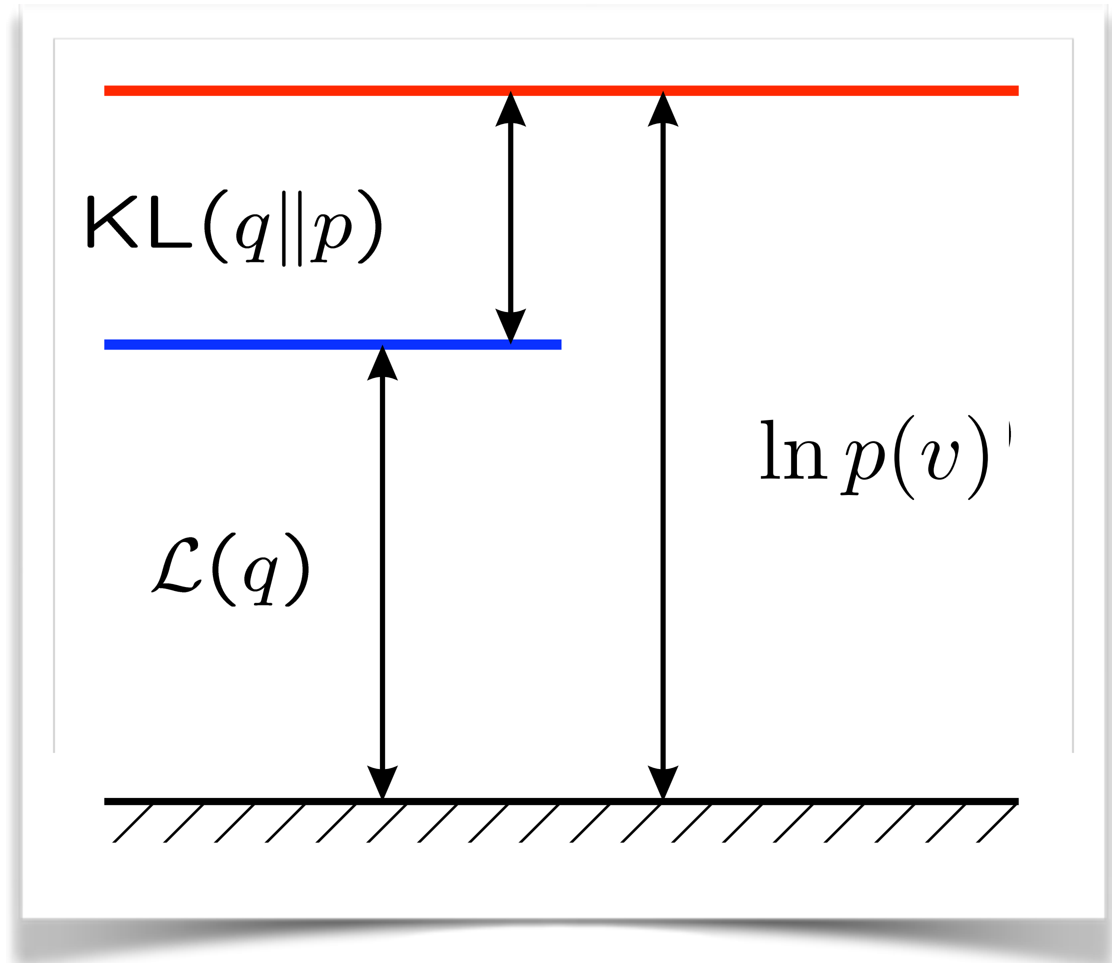
$$
\begin{aligned}
\ln p(v) &= \ln p(v) + \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) - \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) \\
&= \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) - \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) + \sum_h q(h \mid v) \log p(v) \\
&= \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) - \sum_h q(h \mid v) \left[\ln\left(\frac{p(v,h)}{q(h \mid v)}\right) - \ln p(v)\right] \\
&= \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) - \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{p(v)q(h \mid v)}\right) \\
&= \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right) - \sum_h q(h \mid v) \ln\left(\frac{p(h \mid v)}{q(h \mid v)}\right) \\
&= \mathcal{L}(q) + \mathrm{KL}(q \| p)
\end{aligned}
$$

UNDERSTANDING VARIATIONAL LOWER BOUND

Lower bound

$$\ln p(v) = \mathcal{L}(q) + \mathrm{KL}(q\|p)$$

Kullback-Leibler divergence



- We have a lower bound on the data likelihood

$$\therefore \quad \ln p(v) \geq \mathcal{L}(q) \qquad \text{where} \quad \mathcal{L}(q) = \sum_h q(h \mid v) \ln\left(\frac{p(v,h)}{q(h \mid v)}\right)$$

VARIATIONAL EXPECTATION MAXIMIZATION

- We can approximately maximize the likelihood by maximizing the lower bound.

$$\mathcal{L}(q) = \sum_h q(h \mid v) \ln \left(\frac{p(v, h; \theta)}{q(h \mid v)} \right)$$
$$= \sum_h q(h \mid v) \ln p^*(v, h; \theta) - \ln Z(\theta) - \sum_h q(h \mid v) \ln q(h \mid v)$$
$$= -\sum_h q(h \mid v) E(v, h; \theta) - \ln Z(\theta) + \mathcal{H}(q)$$

- We can do this in 2 steps:

1. Variation expectation: Maximize the lower bound w.r.t. the variational distributions: $q(h \mid v)$.

2. Variational maximization: Maximizing the lower bound w.r.t the model parameters via gradient ascent.

STEP 1: VARIATIONAL E-STEP

- For the DBM, we want to maximize w.r.t $q(h^{(1)}, h^{(2)} \mid v)$:

$$\mathcal{L}(q) = \sum_{h^{(1)}, h^{(2)}} q(h^{(1)}, h^{(2)} \mid v) \ln \left(\frac{p(v, h^{(1)}, h^{(2)}; \theta)}{q(h^{(1)}, h^{(2)} \mid v)} \right)$$

$$= - \sum_{h^{(1)}, h^{(2)}} q(h^{(1)}, h^{(2)} \mid v) E(v, h^{(1)}, h^{(2)}; \theta) - \ln Z(\theta) + \mathcal{H}(q)$$

- Mean-field assumption: $q(h^{(1)}, h^{(2)} \mid v) = \prod_j q_v(h_j^{(1)}) \prod_k q_v(h_k^{(2)})$
  - Posterior has only independent elements.

PARAMETERIZING THE APPROXIMATE POSTERIOR

- Parametrization of q : $\hat{h}_j^{(1)} \equiv q_v(h_j^{(1)} = 1), \quad \hat{h}_k^{(2)} \equiv q_v(h_k^{(2)} = 1)$

$$q_v(h_j^{(1)}) = (\hat{h}_j^{(1)})^{h_j^{(1)}} (1 - \hat{h}_j^{(1)})^{(1-h_j^{(1)})}$$

$$q_v(h_k^{(2)}) = (\hat{h}_k^{(2)})^{h_k^{(2)}} (1 - \hat{h}_k^{(2)})^{(1-h_k^{(2)})}$$

- With the mean field assumption:

$$q(h^{(1)}, h^{(2)} \mid v) = \prod_j (\hat{h}_j^{(1)})^{h_j^{(1)}} (1 - \hat{h}_j^{(1)})^{(1-h_j^{(1)})} \times \prod_k (\hat{h}_k^{(2)})^{h_k^{(2)}} (1 - \hat{h}_k^{(2)})^{(1-h_k^{(2)})}$$

MAXIMIZING THE LOWER BOUND

- Putting the DBM energy function, mean field q into $\mathcal{L}(q)$:

$$\mathcal{L}(q) = -\sum_{h^{(1)},h^{(2)}} q(h^{(1)}, h^{(2)} \mid v) E(v, h^{(1)}, h^{(2)}; \theta) - \ln Z(\theta) + \mathcal{H}(q)$$

$$= \sum_i \sum_{j'} v_i W^{(1)}_{ij'} \hat{h}^{(1)}_{j'} + \sum_{j'} \sum_{k'} \hat{h}^{(1)}_{j'} W^{(2)}_{j'k'} \hat{h}^{(2)}_{k'} - \ln Z(\theta) + \mathcal{H}(q)$$

- We want to maximize $\mathcal{L}(q)$ w.r.t. $q(h^{(1)},h^{(2)} \mid v)$
  - Solve system of eqns: $\frac{\partial}{\partial \hat{h}^{(1)}_j} \mathcal{L}(q) = 0, \quad \text{and} \quad \frac{\partial}{\partial \hat{h}^{(2)}_k} \mathcal{L}(q) = 0$

MAXIMIZING THE LOWER BOUND

$$\frac{\partial}{\partial \hat{h}_j^{(1)}} \mathcal{L}(q) = \frac{\partial}{\partial \hat{h}_j^{(1)}} \left[\sum_i \sum_{j'} v_i W_{ij'}^{(1)} \hat{h}_{j'}^{(1)} + \sum_{j'} \sum_{k'} \hat{h}_{j'}^{(1)} W_{j'k'}^{(2)} \hat{h}_{k'}^{(2)} - \ln Z(\theta) + \mathcal{H}(q) \right]$$

MAXIMIZING THE LOWER BOUND

$$\frac{\partial}{\partial \hat{h}_j^{(1)}} \mathcal{L}(q) = \frac{\partial}{\partial \hat{h}_j^{(1)}} \left[\sum_i \sum_{j'} v_i W_{ij'}^{(1)} \hat{h}_{j'}^{(1)} + \sum_{j'} \sum_{k'} \hat{h}_{j'}^{(1)} W_{j'k'}^{(2)} \hat{h}_{k'}^{(2)} - \ln Z(\theta) + \mathcal{H}(q) \right]$$

$$\begin{aligned} &= \frac{\partial}{\partial \hat{h}_j^{(1)}} \left[\sum_i \sum_{j'} v_i W_{ij'}^{(1)} \hat{h}_{j'}^{(1)} + \sum_{j'} \sum_{k'} \hat{h}_{j'}^{(1)} W_{j'k'}^{(2)} \hat{h}_{k'}^{(2)} - \ln Z(\theta) \right. \\ &\qquad - \sum_{j'} \left(\hat{h}_{j'}^{(1)} \ln \hat{h}_{j'}^{(1)} + (1 - \hat{h}_{j'}^{(1)}) \ln(1 - \hat{h}_{j'}^{(1)}) \right) \\ &\qquad \left. - \sum_{k'} \left(\hat{h}_{k'}^{(2)} \ln \hat{h}_{k'}^{(2)} + (1 - \hat{h}_{k'}^{(2)}) \ln(1 - \hat{h}_{k'}^{(2)}) \right) \right] \end{aligned}$$

MAXIMIZING THE LOWER BOUND

$$
\begin{aligned}
\frac{\partial}{\partial \hat{h}_j^{(1)}} \mathcal{L}(q) &= \frac{\partial}{\partial \hat{h}_j^{(1)}} \left[\sum_i \sum_{j'} v_i W_{ij'}^{(1)} \hat{h}_{j'}^{(1)} + \sum_{j'} \sum_{k'} \hat{h}_{j'}^{(1)} W_{j'k'}^{(2)} \hat{h}_{k'}^{(2)} - \ln Z(\theta) + \mathcal{H}(q) \right] \\
&= \frac{\partial}{\partial \hat{h}_j^{(1)}} \left[\sum_i \sum_{j'} v_i W_{ij'}^{(1)} \hat{h}_{j'}^{(1)} + \sum_{j'} \sum_{k'} \hat{h}_{j'}^{(1)} W_{j'k'}^{(2)} \hat{h}_{k'}^{(2)} - \ln Z(\theta) \right. \\
&\qquad - \sum_{j'} \left(\hat{h}_{j'}^{(1)} \ln \hat{h}_{j'}^{(1)} + (1 - \hat{h}_{j'}^{(1)}) \ln(1 - \hat{h}_{j'}^{(1)}) \right) \\
&\qquad \left. - \sum_{k'} \left(\hat{h}_{k'}^{(2)} \ln \hat{h}_{k'}^{(2)} + (1 - \hat{h}_{k'}^{(2)}) \ln(1 - \hat{h}_{k'}^{(2)}) \right) \right] \\
&= \sum_i v_i W_{ij}^{(1)} + \sum_{k'} W_{jk'}^{(2)} \hat{h}_{k'}^{(2)} - \ln \left(\frac{\hat{h}_{j'}^{(1)}}{1 - \hat{h}_{j'}^{(1)}} \right)
\end{aligned}
$$

MAXIMIZING THE LOWER BOUND

$$\frac{\partial}{\partial \hat{h}_j^{(1)}} \mathcal{L}(q) = 0 = \sum_i v_i W_{ij}^{(1)} + \sum_{k'} W_{jk'}^{(2)} \hat{h}_{k'}^{(2)} - \ln\left(\frac{\hat{h}_j^{(1)}}{1 - \hat{h}_j^{(1)}}\right)$$

MAXIMIZING THE LOWER BOUND

$$\frac{\partial}{\partial \hat{h}_j^{(1)}} \mathcal{L}(q) = 0 = \sum_i v_i W_{ij}^{(1)} + \sum_{k'} W_{jk'}^{(2)} \hat{h}_{k'}^{(2)} - \ln\left(\frac{\hat{h}_j^{(1)}}{1 - \hat{h}_j^{(1)}}\right)$$

$$\hat{h}_j^{(1)} = \mathrm{sigmoid}\left(\sum_i v_i W_{ij}^{(1)} + \sum_{k'} W_{jk'}^{(2)} \hat{h}_{k'}^{(2)}\right)$$

MAXIMIZING THE LOWER BOUND

$$\frac{\partial}{\partial \hat{h}_j^{(1)}} \mathcal{L}(q) = 0 = \sum_i v_i W_{ij}^{(1)} + \sum_{k'} W_{jk'}^{(2)} \hat{h}_{k'}^{(2)} - \ln\left(\frac{\hat{h}_j^{(1)}}{1 - \hat{h}_j^{(1)}}\right)$$

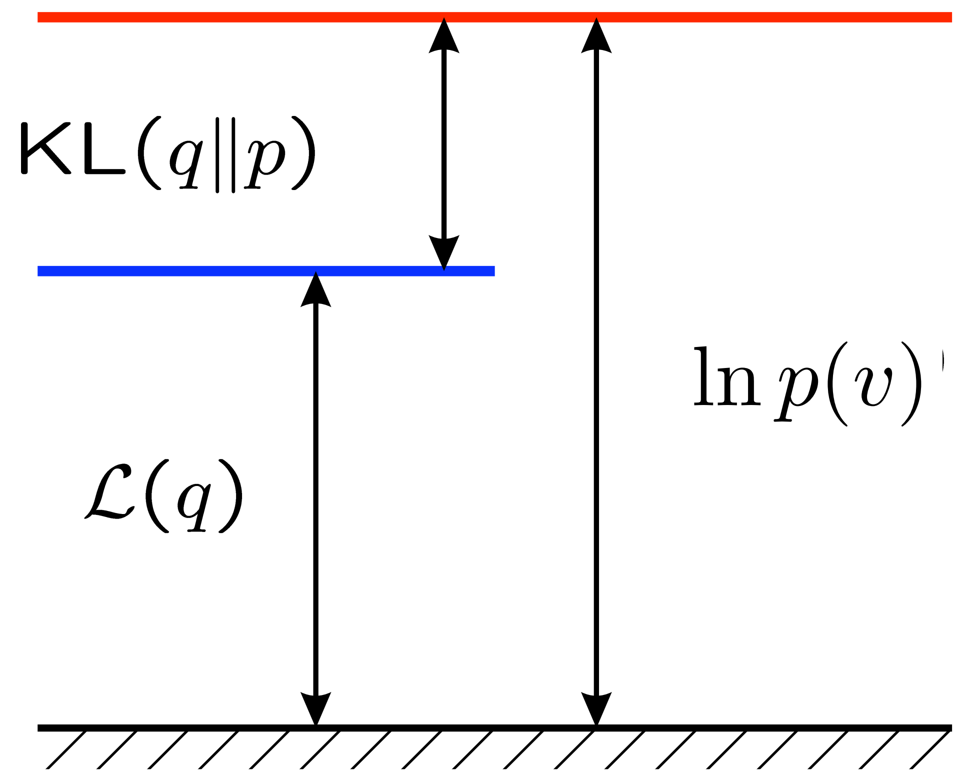
$$\hat{h}_j^{(1)} = \text{sigmoid}\left(\sum_i v_i W_{ij}^{(1)} + \sum_{k'} W_{jk'}^{(2)} \hat{h}_{k'}^{(2)}\right)$$

- So at the max of $\mathcal{L}(q)$ w.r.t. q, we have:

$$\hat{h}_j^{(1)} = \text{sigmoid}\left(\sum_i v_i W_{ij}^{(1)} + \sum_{k'} W_{jk'}^{(2)} \hat{h}_{k'}^{(2)}\right), \quad \forall j$$

$$\hat{h}_k^{(2)} = \text{sigmoid}\left(\sum_{j'} W_{j'k}^{(2)} \hat{h}_{j'}^{(1)}\right), \quad \forall k$$

MAXIMIZING THE LOWER BOUND



- Iterate until convergence:

$$\hat{h}_j^{(1)} = \mathrm{sigmoid}\left(\sum_i v_i W_{ij}^{(1)} + \sum_{k'} W_{jk'}^{(2)} \hat{h}_{k'}^{(2)}\right), \quad \forall j$$

$$\hat{h}_k^{(2)} = \mathrm{sigmoid}\left(\sum_{j'} W_{j'k}^{(2)} \hat{h}_{j'}^{(1)}\right), \quad \forall k$$

STEP 2: VARIATIONAL M-STEP

- Maximize $\mathcal{L}(q)$ with respect to the model parameters:
  - we will use the stochastic gradient descent:

$$\frac{\partial}{\partial\theta}\mathcal{L}(q) = \frac{\partial}{\partial\theta}\left(\sum_i \sum_{j'} v_i W^{(1)}_{ij'} \hat{h}^{(1)}_{j'} + \sum_{j'} \sum_{k'} \hat{h}^{(1)}_{j'} W^{(2)}_{j'k'} \hat{h}^{(2)}_{k'} - \ln Z(\theta) + \mathcal{H}(q)\right)$$

$$= \frac{\partial}{\partial\theta}\left(\sum_i \sum_{j'} v_i W^{(1)}_{ij'} \hat{h}^{(1)}_{j'} + \sum_{j'} \sum_{k'} \hat{h}^{(1)}_{j'} W^{(2)}_{j'k'} \hat{h}^{(2)}_{k'}\right) - \frac{\partial}{\partial\theta}\ln Z(\theta)$$

STEP 2: VARIATIONAL M-STEP

- Maximize $\mathcal{L}(q)$ with respect to the model parameters:
  - we will use stochastic gradient descent:

$$\frac{\partial}{\partial\theta}\mathcal{L}(q) = \frac{\partial}{\partial\theta}\left(\sum_i\sum_{j'} v_i W^{(1)}_{ij'}\hat{h}^{(1)}_{j'} + \sum_{j'}\sum_{k'}\hat{h}^{(1)}_{j'}W^{(2)}_{j'k'}\hat{h}^{(2)}_{k'} - \ln Z(\theta) + \mathcal{H}(q)\right)$$

$$= \underbrace{\boxed{\frac{\partial}{\partial\theta}\left(\sum_i\sum_{j'} v_i W^{(1)}_{ij'}\hat{h}^{(1)}_{j'} + \sum_{j'}\sum_{k'}\hat{h}^{(1)}_{j'}W^{(2)}_{j'k'}\hat{h}^{(2)}_{k'}\right)}}_{\text{Easy.}} - \underbrace{\boxed{\frac{\partial}{\partial\theta}\ln Z(\theta)}}_{\text{Hard!}}$$

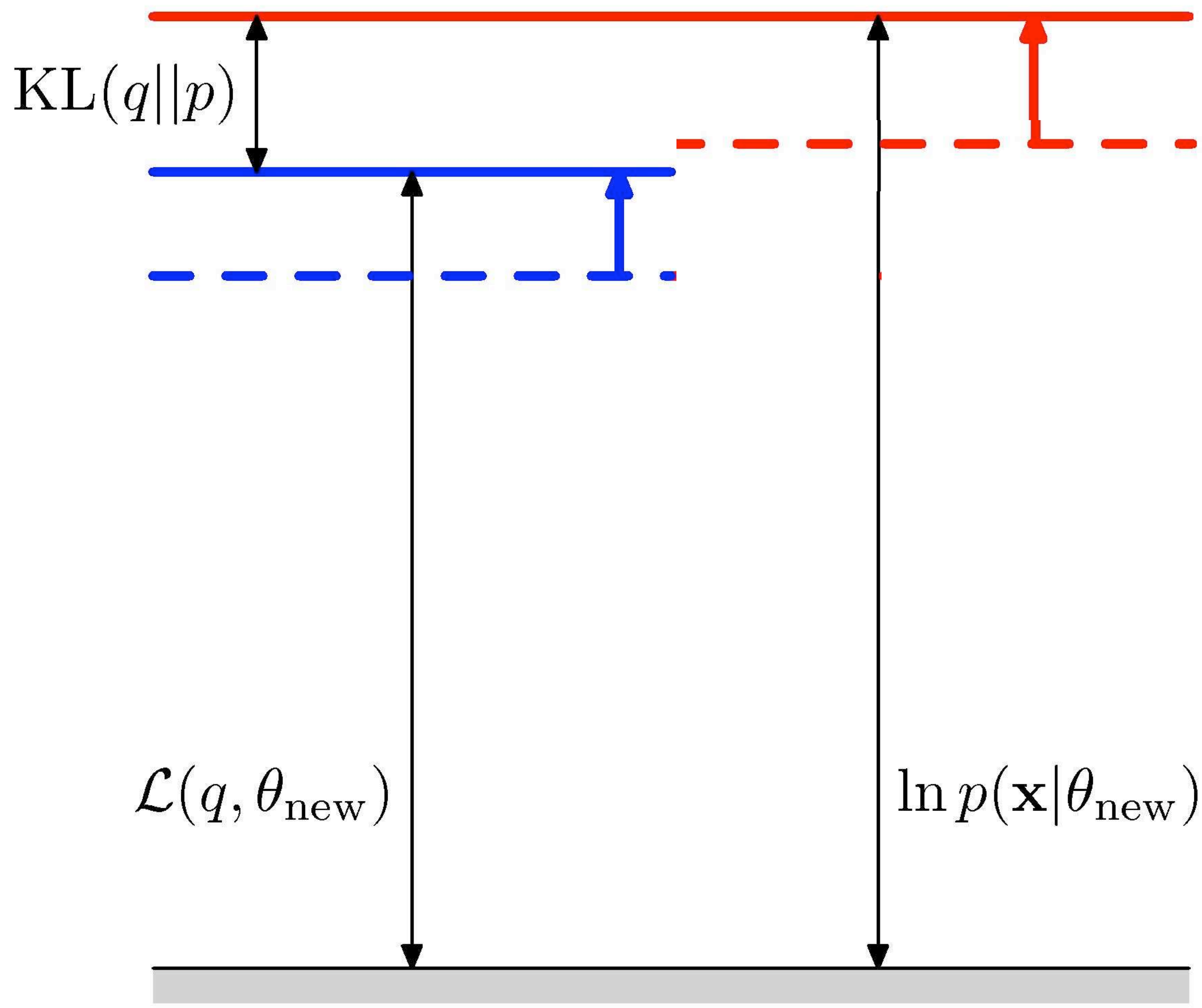
- As in PCD for the RBM, we will make use of a persistent Gibbs chain to approximate this term.

MAXIMIZING LOWER BOUND W.R.T. PARAMETERS

- Using Gibbs samples, we estimate $\frac{\partial}{\partial \theta} \mathcal{L}(q)$

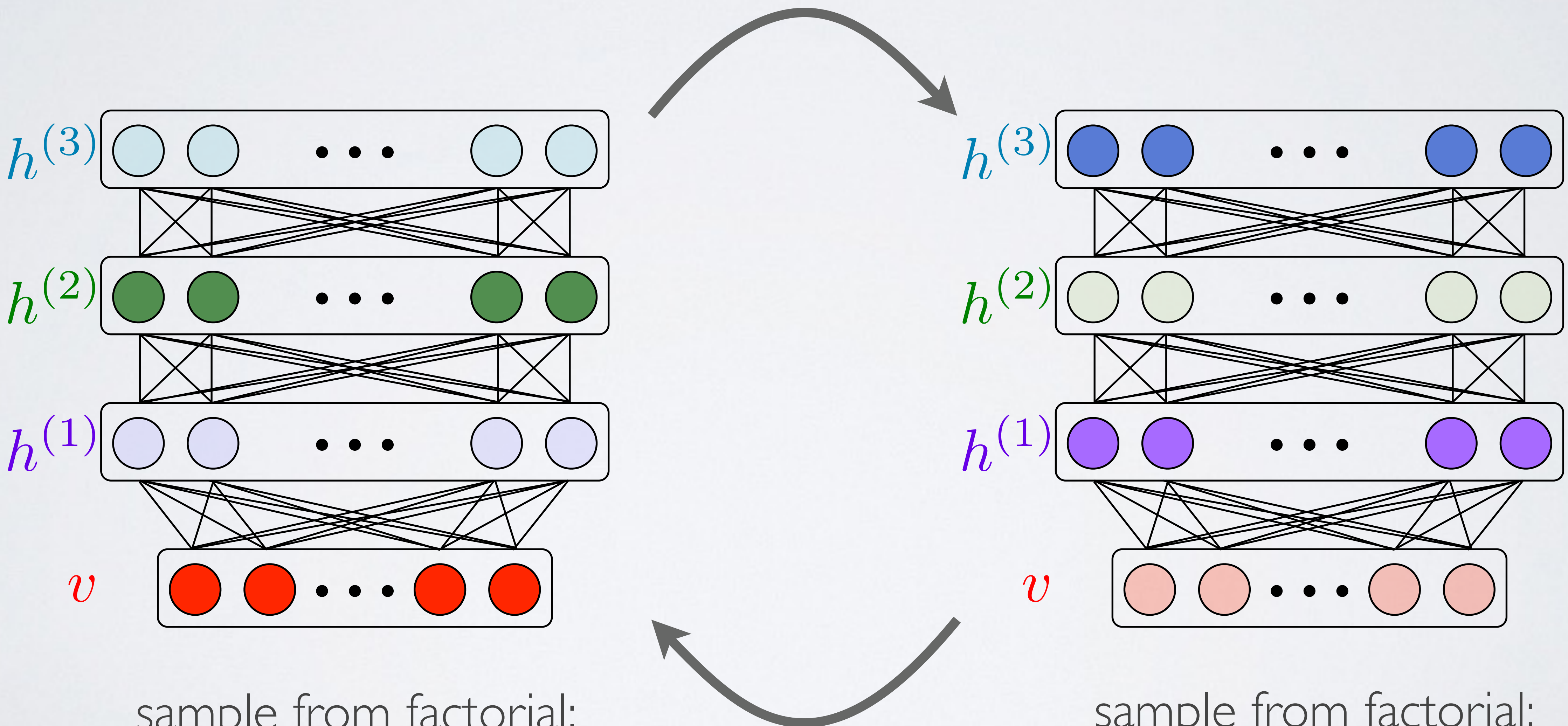
- Apply SGD:

$$\theta_{t+1} = \theta_t + \epsilon \frac{\partial}{\partial \theta} \mathcal{L}(q)$$



DBM: GIBBS SAMPLING

- Gibbs sampling in DBMs is similar to Gibbs sampling in RBMs.
- Iterate (exploiting the **factorization of conditionals**)

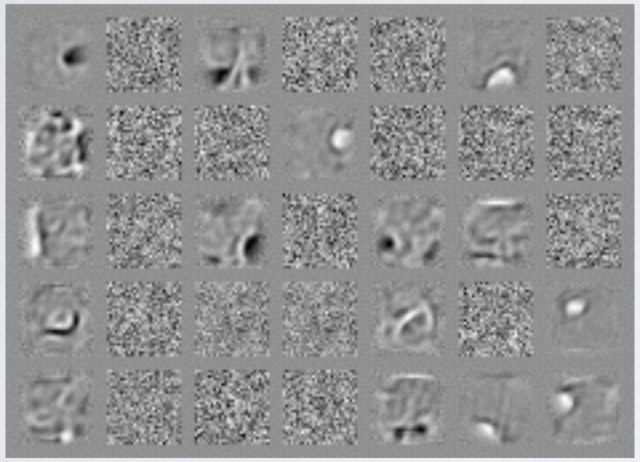


sample from factorial: $p(h^{(1)}, h^{(3)} \mid v, h^{(2)})$

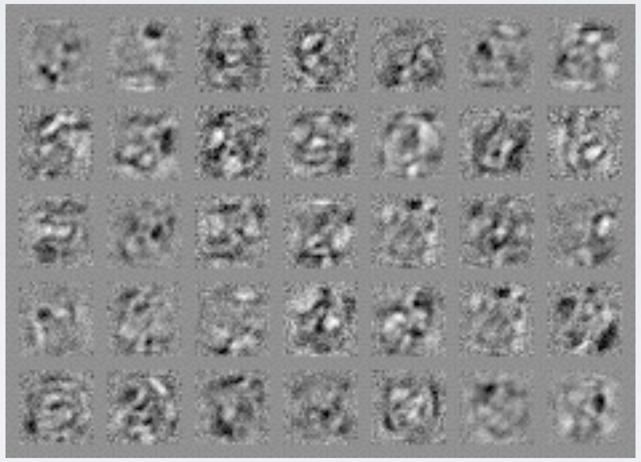
sample from factorial: $p(v, h^{(2)} \mid h^{(1)}, h^{(3)})$

DBM: LEARNING

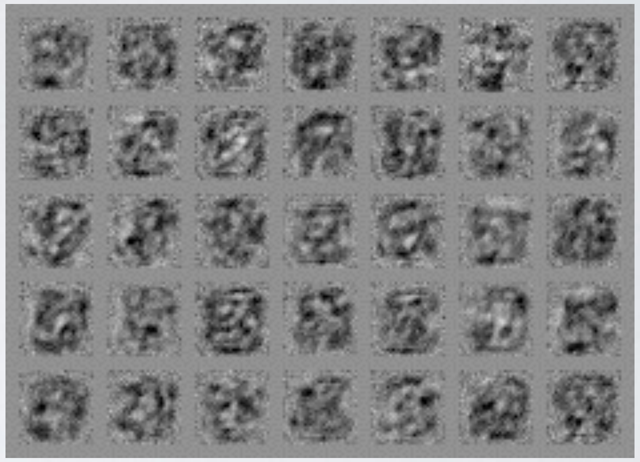
- Training a DBM from random initial weights is difficult:



1st layer



2nd layer



3rd layer

- Two strategies proposed:

1. Greedy layer-wise pretraining with RBMs (Salakhutdinov & Hinton, 2009).
2. Centering the DBM energy function (Montavon and Müller, 2012).

GREEDY LAYER-WISE PRETRAINING

- Salakhutdinov & Hinton (2009) propose to greedily pretrain the model as a stack of RBMs

➡ Important note: **not quite the same as in the DBN case**.

➡ Eg. doubling up W^1 and W^2 representations while pretraining can account for h^1 connecting to both v and h^2.

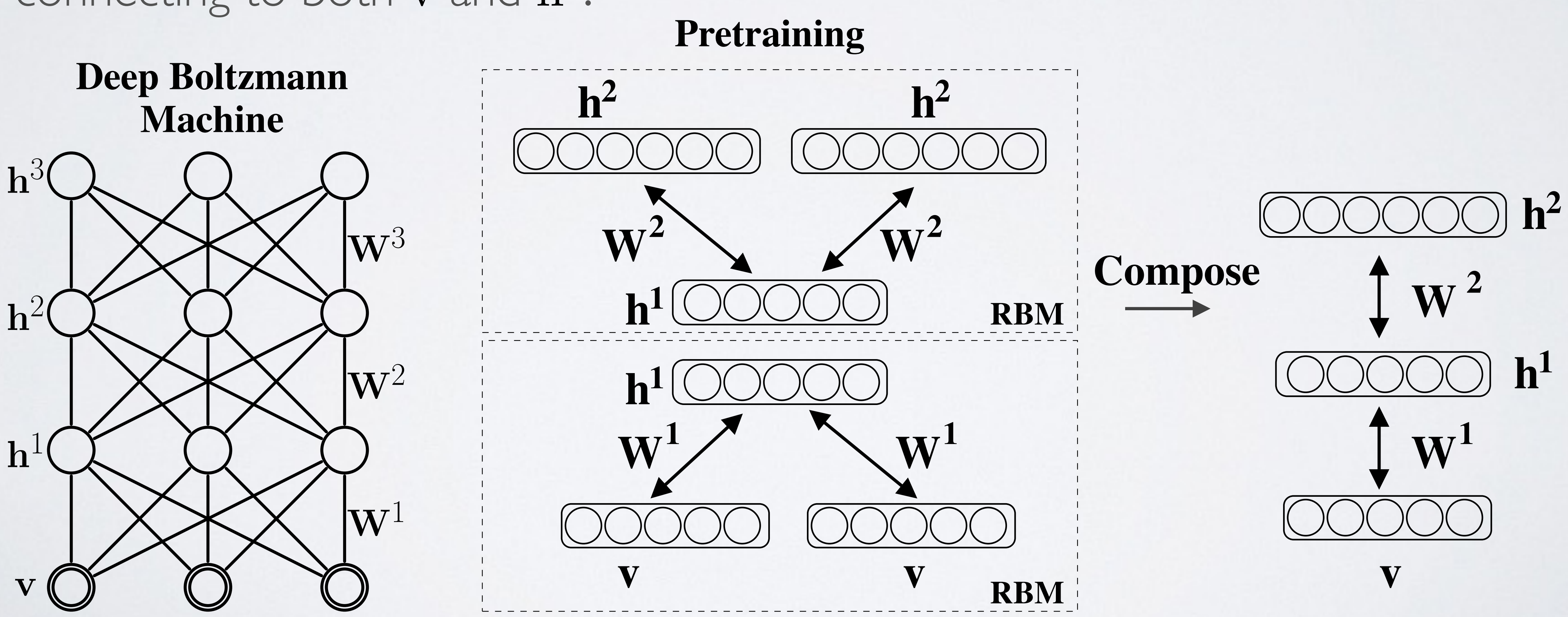


image from Salakhutdinov & Hinton (2009)

CENTERING DBMS

(Montavon and Müller, 2012)

- Promote learning by reparameterizing the DBM energy function:

Original:

$$E(v, h^{(1)}, h^{(2)}, \theta) = -v^T W^{(1)} h^{(1)} - h^{(1)T} W^{(2)} h^{(2)}$$

Centered DBM energy function:

$$E(v, h^{(1)}, h^{(2)}, \theta) = -(v - \alpha)^T W^{(1)} (h^{(1)} - \beta) - (h^{(1)} - \beta)^T W^{(2)} (h^{(2)} - \gamma)$$

- A few possible choices for α, β, γ
  - Montavon and Müller (2012) advocate:

$$\alpha = \langle v \rangle_D , \beta = \left\langle h^{(1)} \right\rangle_D , \gamma = \left\langle h^{(2)} \right\rangle_D$$

DBM APPLICATIONS

- Used to pretrain the NN to achieve state-of-the-art performance for permutation invariant MNIST.

Method	Unit Type	Error %
2 layer NN [19]	Logistic	1.60
SVM gaussian kernel	-	1.4
Dropout	ReLU	1.25
Dropout + weight norm constraint	ReLU	1.05
DBN + finetuning	Logistic	1.18
DBN + dropout fine-tuning	Logistic	0.92
DBM + finetuning	Logistic	0.96
DBM + dropout finetuning	Logistic	**0.79**

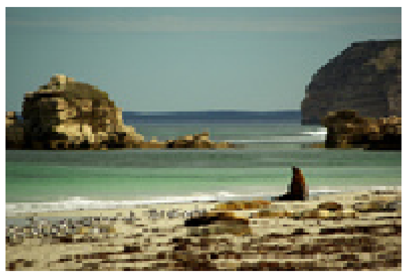
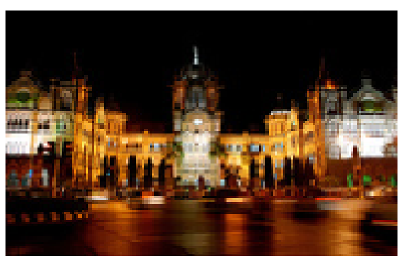
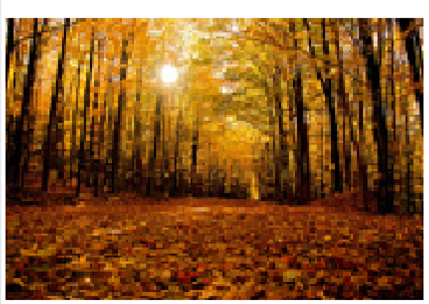
- State-of-the-art MNIST likelihood.
- State-of-the-art joint model of images and text. Using the flickr-1M dataset of mostly unlabeled data. (Srivastava and Salakhutdinov, 2012)
    - ➡ Task: image-topic classification

Flickr-1M dataset:

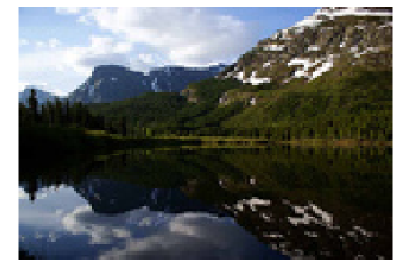
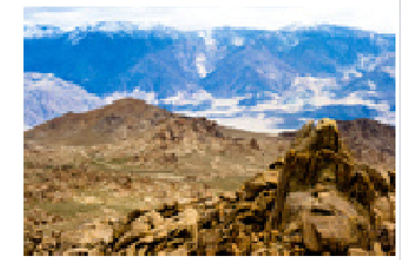
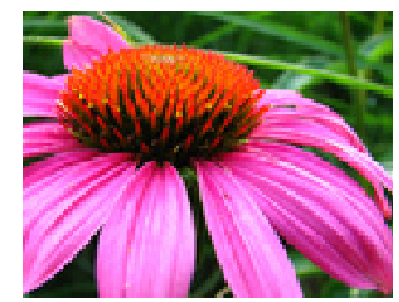
Method	Mean Average Precision %	Precision at 50
LDA [8]	0.492	0.754
SVM [8]	0.475	0.758
DBN [22]	0.599	0.867
Autoencoder (based on [15])	0.600	0.875
DBM [22]	0.609	0.873
Multiple Kernel Learning SVMs [4]	0.623	-
DBN with dropout finetuning	0.628	0.891
DBM with dropout finetuning	**0.632**	**0.895**

DBM ON FLICKR

Generated tags:

Image	Given Tags	Generated Tags
	pentax, k10d, kangarooisland, southaustralia, sa, australia, australiansealion, 300mm	beach, sea, surf, strand, shore, wave, seascape, sand, ocean, waves
	<no text>	night, lights, christmas, nightshot, nacht, nuit,notte, longexposure, noche, nocturna
	aheram, 0505 sarahc, moo	portrait, bw, blackandwhite, woman, people, faces, girl,blackwhite, person, man
	unseulpixel, naturey crap	fall, autumn, trees, leaves, foliage, forest, woods, branches, path

Images retrieved from tags:

Input Text	2 nearest neighbours to generated image features	
nature, hill scenery, green clouds		
flower, nature, green, flowers, petal, petals, bud		
blue, red, art, artwork, painted, paint, artistic surreal, gallery bleu		
bw, blackandwhite, noiretblanc, biancoenero blancoynegro		

Images from Srivastava and Salakhutdinov (2012)

DBM APPLICATIONS

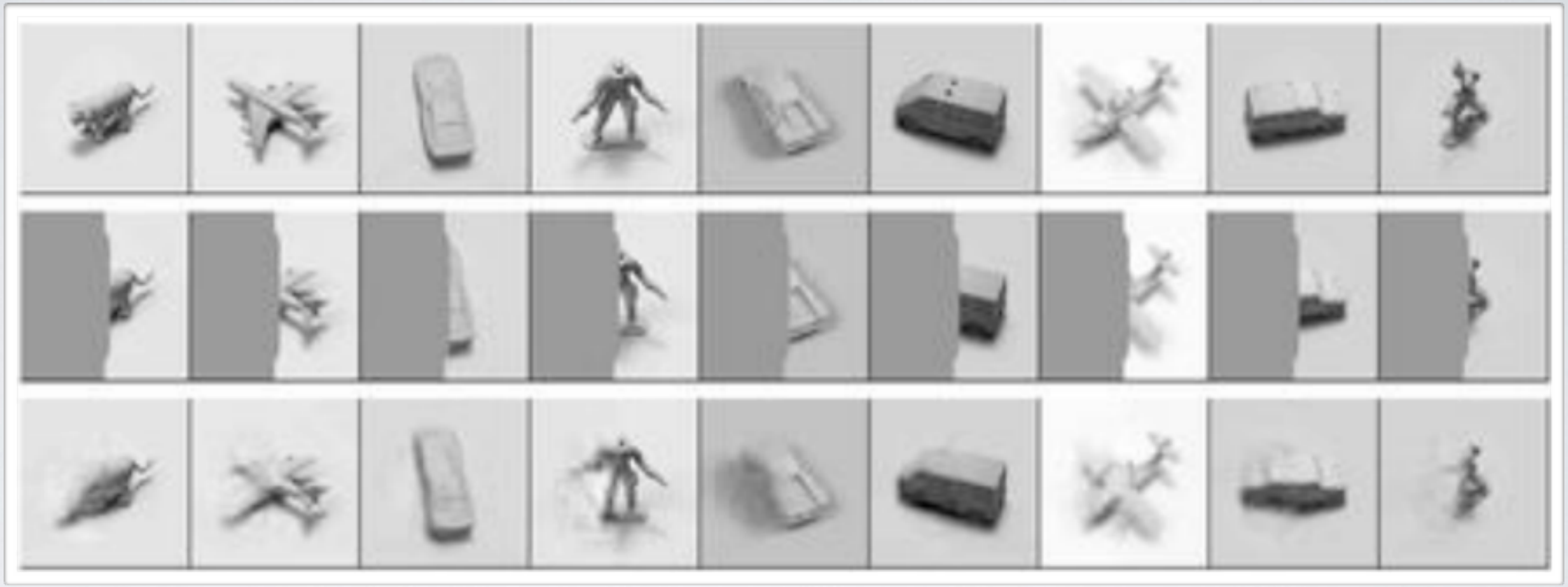
- NORB inpainting task:

images from Salakhutdinov & Hinton (2009)

original image

masked image

inpainted image



Capable of fantasizing plausible alternatives: