Neural networks

Deep learning - deep belief network

DEEP BELIEF NETWORK

Topics: deep belief network

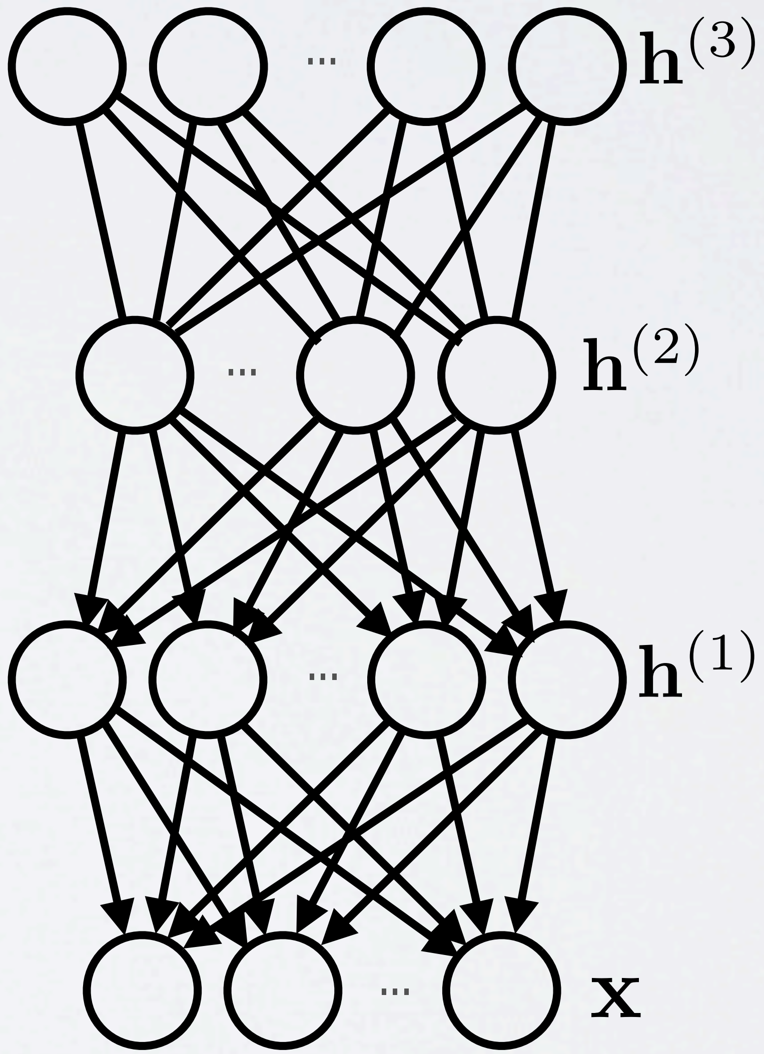
- The idea of pre-training came from work on deep belief networks (DBNs)
  - it is a generative model that mixes undirected and directed connections between variables
  - top 2 layers' distribution $p(\mathbf{h}^{(2)}, \mathbf{h}^{(3)})$ is an RBM
  - other layers form a Bayesian network:
    - the conditional distributions of a layers given the one above it are

$$p(h_j^{(1)} = 1|\mathbf{h}^{(2)}) = \mathrm{sigm}(\mathbf{b}^{(1)} + \mathbf{W}^{(2)}{}^\top \mathbf{h}^{(2)})$$

$$p(x_i = 1|\mathbf{h}^{(1)}) = \mathrm{sigm}(\mathbf{b}^{(0)} + \mathbf{W}^{(1)}{}^\top \mathbf{h}^{(1)})$$

    - this is referred to as a **sigmoid belief network** (SBN)
  - a DBN **is not** a feed-forward network

DBN's graphical model



DEEP BELIEF NETWORK

Topics: deep belief network

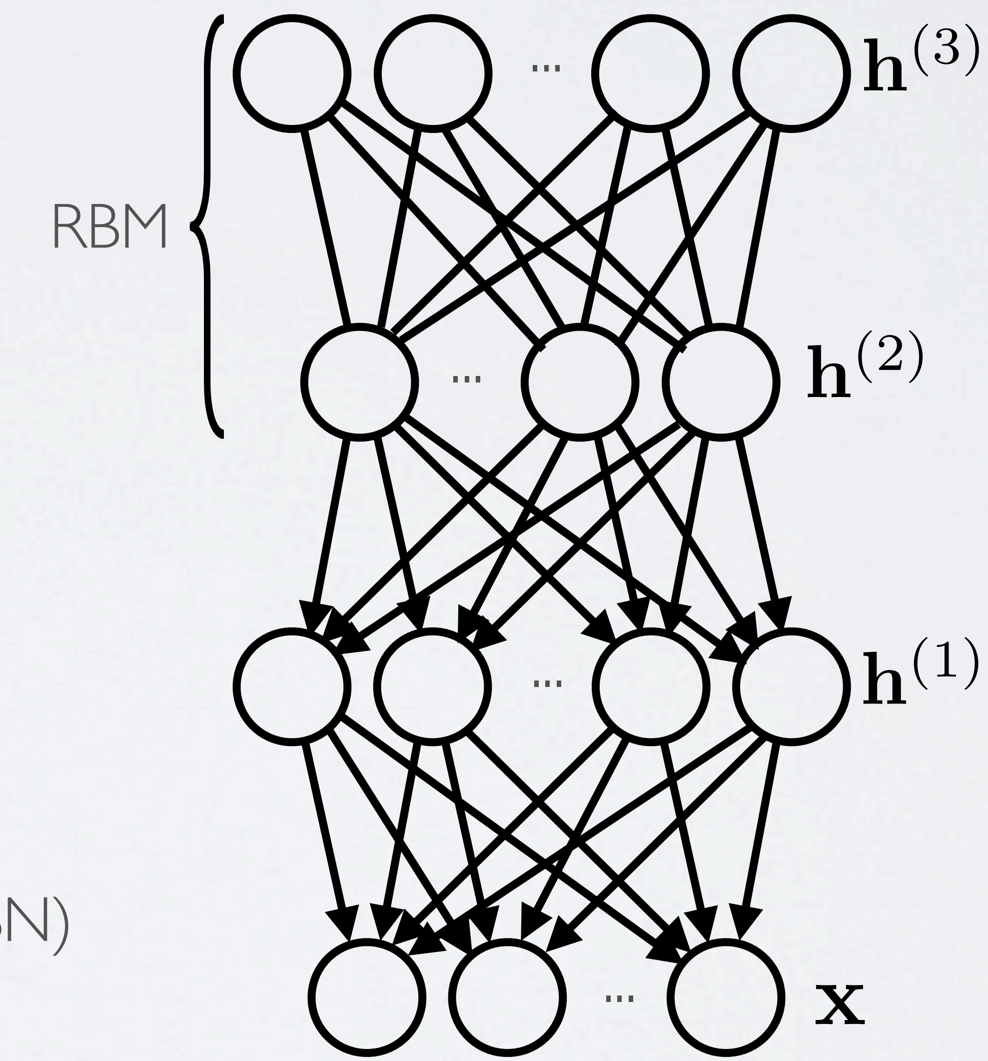
- The idea of pre-training came from work on deep belief networks (DBNs)
  - it is a generative model that mixes undirected and directed connections between variables
  - top 2 layers' distribution $p(\mathbf{h}^{(2)}, \mathbf{h}^{(3)})$ is an RBM
  - other layers form a Bayesian network:
    - the conditional distributions of a layers given the one above it are

$$p(h_j^{(1)} = 1|\mathbf{h}^{(2)}) = \mathrm{sigm}(\mathbf{b}^{(1)} + \mathbf{W}^{(2)^\top}\mathbf{h}^{(2)})$$

$$p(x_i = 1|\mathbf{h}^{(1)}) = \mathrm{sigm}(\mathbf{b}^{(0)} + \mathbf{W}^{(1)^\top}\mathbf{h}^{(1)})$$

    - this is referred to as a **sigmoid belief network** (SBN)
  - a DBN **is not** a feed-forward network

DBN's graphical model



DEEP BELIEF NETWORK

Topics: deep belief network

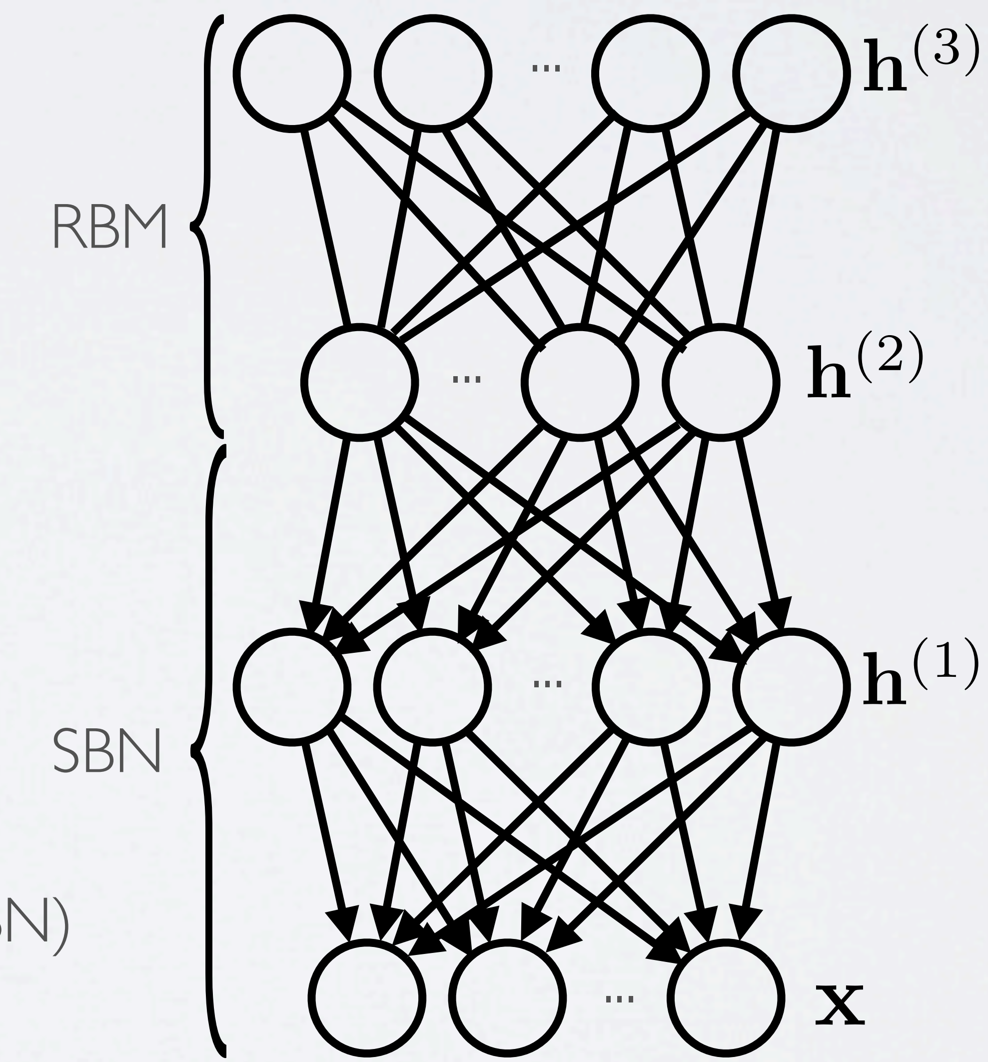
- The idea of pre-training came from work on deep belief networks (DBNs)
  - it is a generative model that mixes undirected and directed connections between variables
  - top 2 layers' distribution $p(\mathbf{h}^{(2)}, \mathbf{h}^{(3)})$ is an RBM
  - other layers form a Bayesian network:
    - the conditional distributions of a layers given the one above it are

$$p(h_j^{(1)} = 1|\mathbf{h}^{(2)}) = \mathrm{sigm}(\mathbf{b}^{(1)} + \mathbf{W}^{(2)}{}^\top \mathbf{h}^{(2)})$$

$$p(x_i = 1|\mathbf{h}^{(1)}) = \mathrm{sigm}(\mathbf{b}^{(0)} + \mathbf{W}^{(1)}{}^\top \mathbf{h}^{(1)})$$

    - this is referred to as a **sigmoid belief network** (SBN)
  - a DBN **is not** a feed-forward network

DBN's graphical model



DEEP BELIEF NETWORK

Topics: deep belief network

- The full distribution of a DBN is as follows

$$p(\mathbf{x}, \mathbf{h}^{(1)}, \mathbf{h}^{(2)}, \mathbf{h}^{(3)}) = p(\mathbf{h}^{(2)}, \mathbf{h}^{(3)})\, p(\mathbf{h}^{(1)}|\mathbf{h}^{(2)})\, p(\mathbf{x}|\mathbf{h}^{(1)})$$

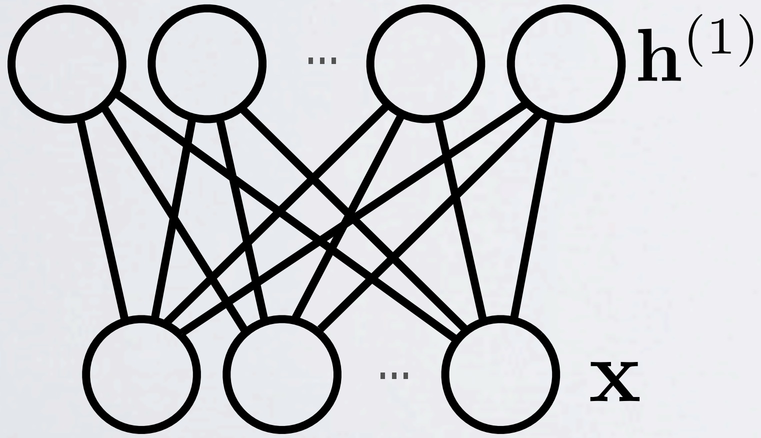
  - where:
    - $p(\mathbf{h}^{(2)}, \mathbf{h}^{(3)}) = \exp\left(\mathbf{h}^{(2)\top}\mathbf{W}^{(3)}\mathbf{h}^{(3)} + \mathbf{b}^{(2)\top}\mathbf{h}^{(2)} + \mathbf{b}^{(3)\top}\mathbf{h}^{(3)}\right)/Z$
    - $p(\mathbf{h}^{(1)}|\mathbf{h}^{(2)}) = \prod_j p(h_j^{(1)}|\mathbf{h}^{(2)})$
    - $p(\mathbf{x}|\mathbf{h}^{(1)}) = \prod_i p(x_i|\mathbf{h}^{(1)})$
- To observe a DBN trained on MNIST in action:
  - http://www.cs.toronto.edu/~hinton/adi/index.htm
- As in a deep feed-forward network, training a DBN is hard
  - initialization will play a crucial role on the results

DEEP BELIEF NETWORK

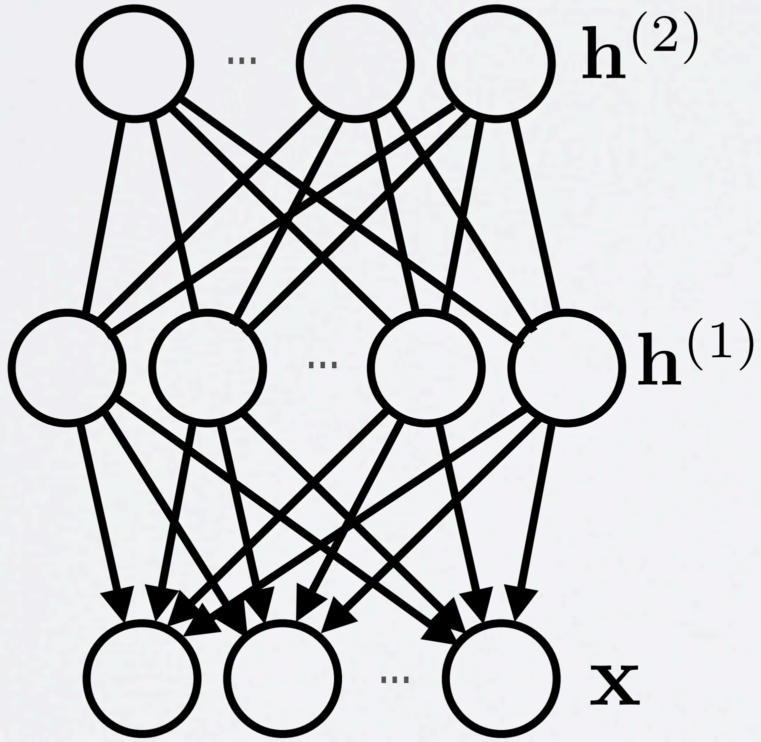
Topics: deep belief network

- This is where the RBM stacking procedure comes from
  - idea: improve prior on last layer by adding another hidden layer
  - how do we train these additional layers?

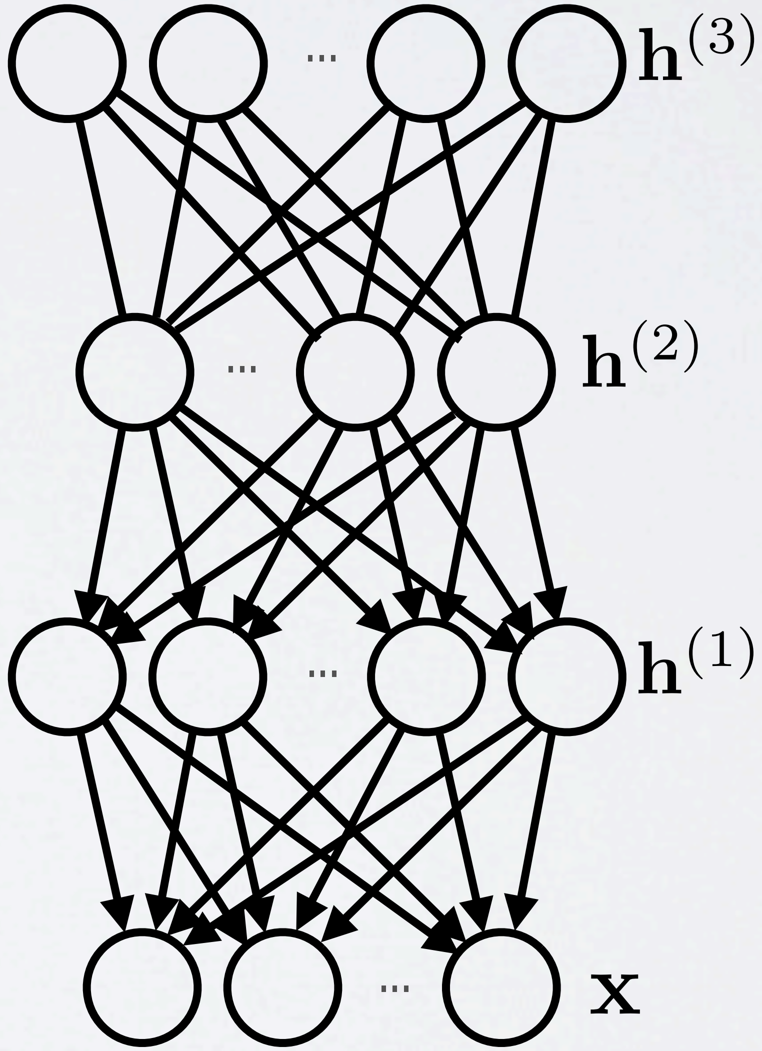
$$p(\mathbf{x}) = \sum_{\mathbf{h}^{(1)}} p(\mathbf{x}, \mathbf{h}^{(1)})$$



$$p(\mathbf{x}, \mathbf{h}^{(1)}) = p(\mathbf{x}|\mathbf{h}^{(1)}) \sum_{\mathbf{h}^{(2)}} p(\mathbf{h}^{(1)}, \mathbf{h}^{(2)})$$



$$p(\mathbf{h}^{(1)}, \mathbf{h}^{(2)}) = p(\mathbf{h}^{(1)}|\mathbf{h}^{(2)}) \sum_{\mathbf{h}^{(3)}} p(\mathbf{h}^{(2)}, \mathbf{h}^{(3)})$$



DEEP BELIEF NETWORK

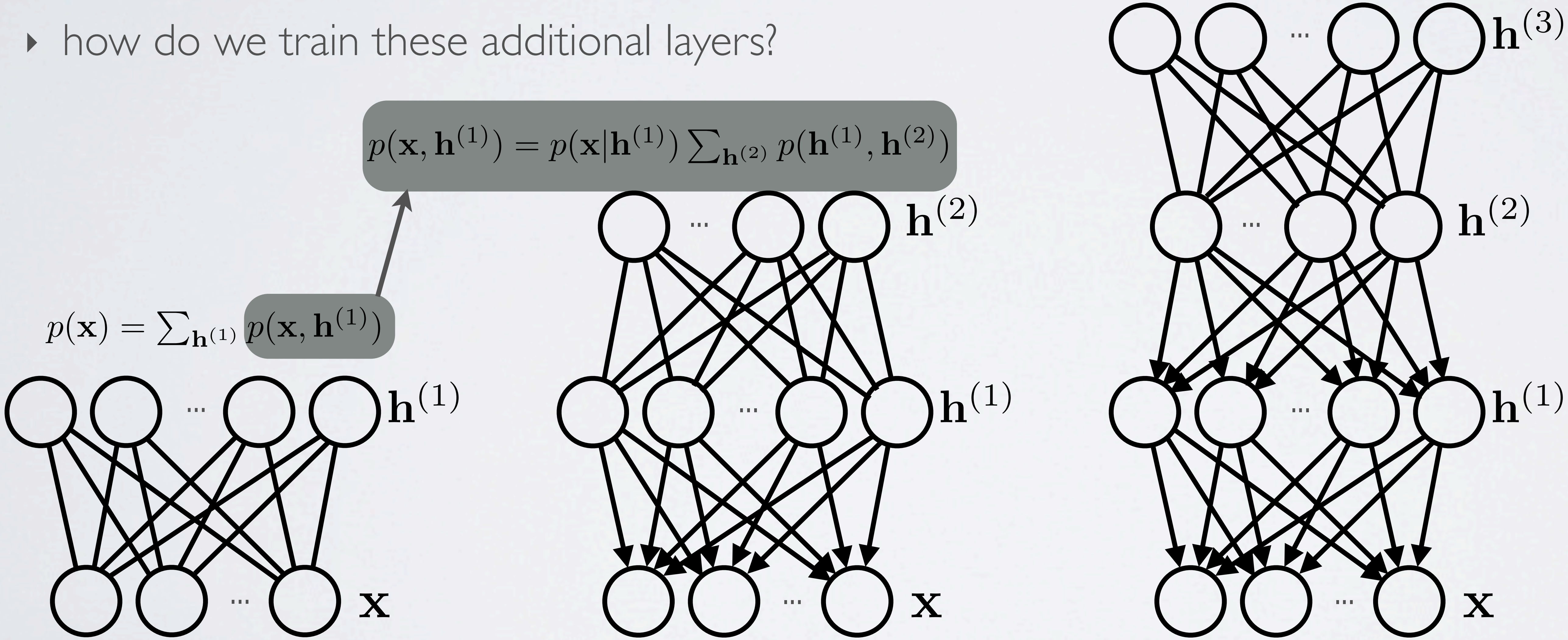
Topics: deep belief network

- This is where the RBM stacking procedure comes from
  - idea: improve prior on last layer by adding another hidden layer
  - how do we train these additional layers?

$$p(\mathbf{h}^{(1)}, \mathbf{h}^{(2)}) = p(\mathbf{h}^{(1)}|\mathbf{h}^{(2)}) \sum_{\mathbf{h}^{(3)}} p(\mathbf{h}^{(2)}, \mathbf{h}^{(3)})$$

$$p(\mathbf{x}, \mathbf{h}^{(1)}) = p(\mathbf{x}|\mathbf{h}^{(1)}) \sum_{\mathbf{h}^{(2)}} p(\mathbf{h}^{(1)}, \mathbf{h}^{(2)})$$

$$p(\mathbf{x}) = \sum_{\mathbf{h}^{(1)}} p(\mathbf{x}, \mathbf{h}^{(1)})$$



DEEP BELIEF NETWORK

Topics: deep belief network

- This is where the RBM stacking procedure comes from
  - idea: improve prior on last layer by adding another hidden layer
  - how do we train these additional layers?