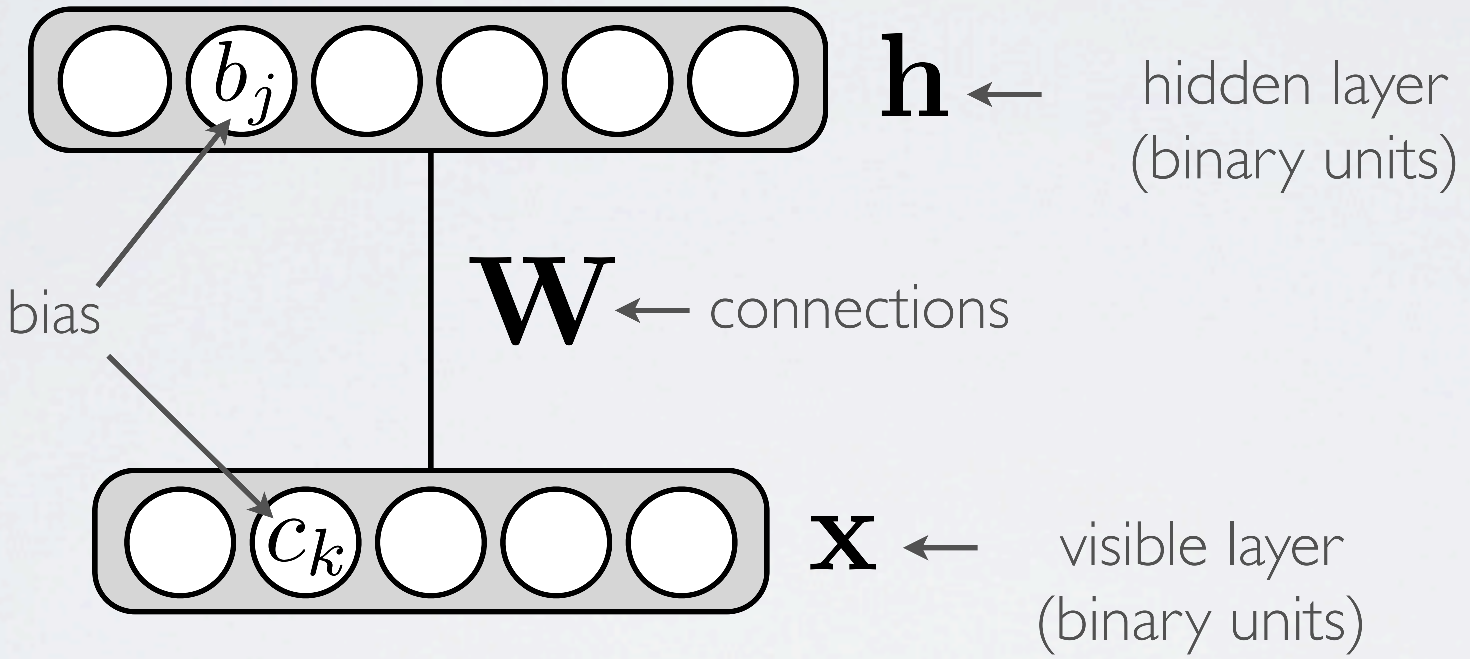
Neural networks

Restricted Boltzmann machine - contrastive divergence

RESTRICTED BOLTZMANN MACHINE

Topics: RBM, visible layer, hidden layer, energy function



Energy function:

$$\begin{aligned} E(\mathbf{x}, \mathbf{h}) &= -\mathbf{h}^\top \mathbf{W} \mathbf{x} - \mathbf{c}^\top \mathbf{x} - \mathbf{b}^\top \mathbf{h} \\ &= -\sum_j \sum_k W_{j,k} h_j x_k - \sum_k c_k x_k - \sum_j b_j h_j \end{aligned}$$

Distribution: $p(\mathbf{x}, \mathbf{h}) = \exp(-E(\mathbf{x}, \mathbf{h}))/Z$

partition function
(intractable)

TRAINING

Topics: training objective

- To train an RBM, we'd like to minimize the average negative log-likelihood (NLL)

$$\frac{1}{T}\sum_t l(f(\mathbf{x}^{(t)})) = \frac{1}{T}\sum_t -\log p(\mathbf{x}^{(t)})$$

- We'd like to proceed by stochastic gradient descent

$$\frac{\partial -\log p(\mathbf{x}^{(t)})}{\partial\theta} = \underbrace{\mathrm{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x}^{(t)},\mathbf{h})}{\partial\theta}\,\middle|\,\mathbf{x}^{(t)}\right]}_{\text{positive phase}} - \underbrace{\mathrm{E}_{\mathbf{x},\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial\theta}\right]}_{\text{negative phase}}$$

TRAINING

Topics: training objective

- To train an RBM, we'd like to minimize the average negative log-likelihood (NLL)

$$\frac{1}{T}\sum_t l(f(\mathbf{x}^{(t)})) = \frac{1}{T}\sum_t -\log p(\mathbf{x}^{(t)})$$

- We'd like to proceed by stochastic gradient descent

hard to compute

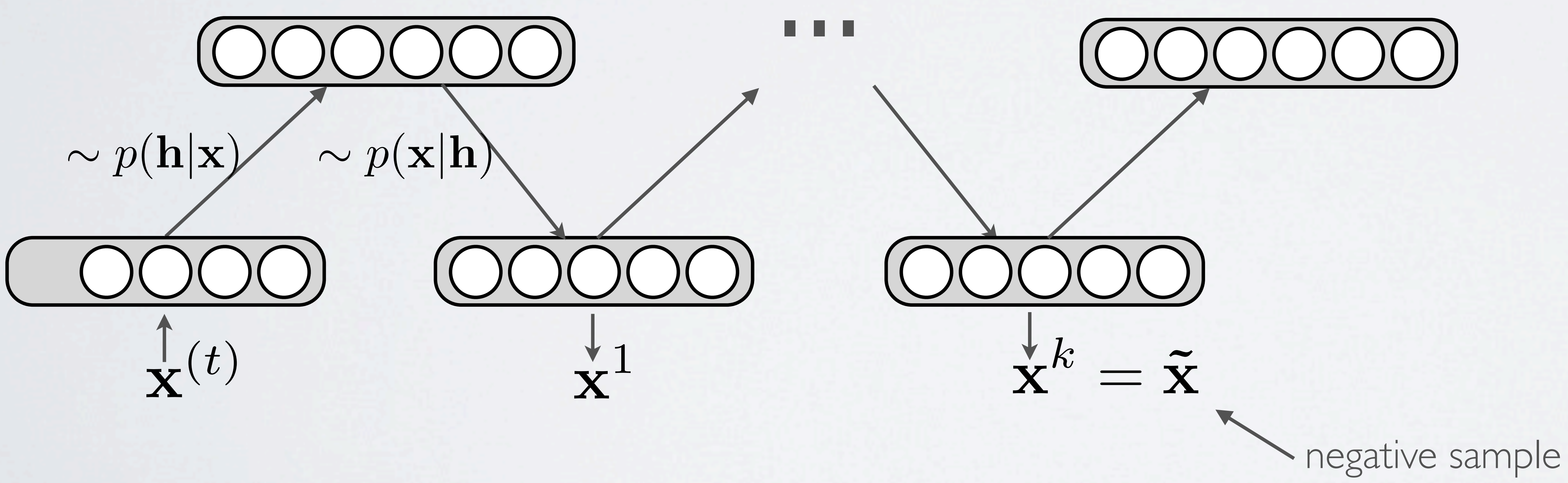
$$\frac{\partial -\log p(\mathbf{x}^{(t)})}{\partial \theta} = \underbrace{\mathrm{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x}^{(t)},\mathbf{h})}{\partial \theta}\,\middle|\,\mathbf{x}^{(t)}\right]}_{\text{positive phase}} - \underbrace{\boxed{\mathrm{E}_{\mathbf{x},\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial \theta}\right]}}_{\text{negative phase}}$$

CONTRASTIVE DIVERGENCE (CD)

(HINTON, NEURAL COMPUTATION, 2002)

Topics: contrastive divergence, negative sample

- Idea:
  1. replace the expectation by a point estimate at $\tilde{\mathbf{x}}$
  2. obtain the point $\tilde{\mathbf{x}}$ by Gibbs sampling
  3. start sampling chain at $\mathbf{x}^{(t)}$

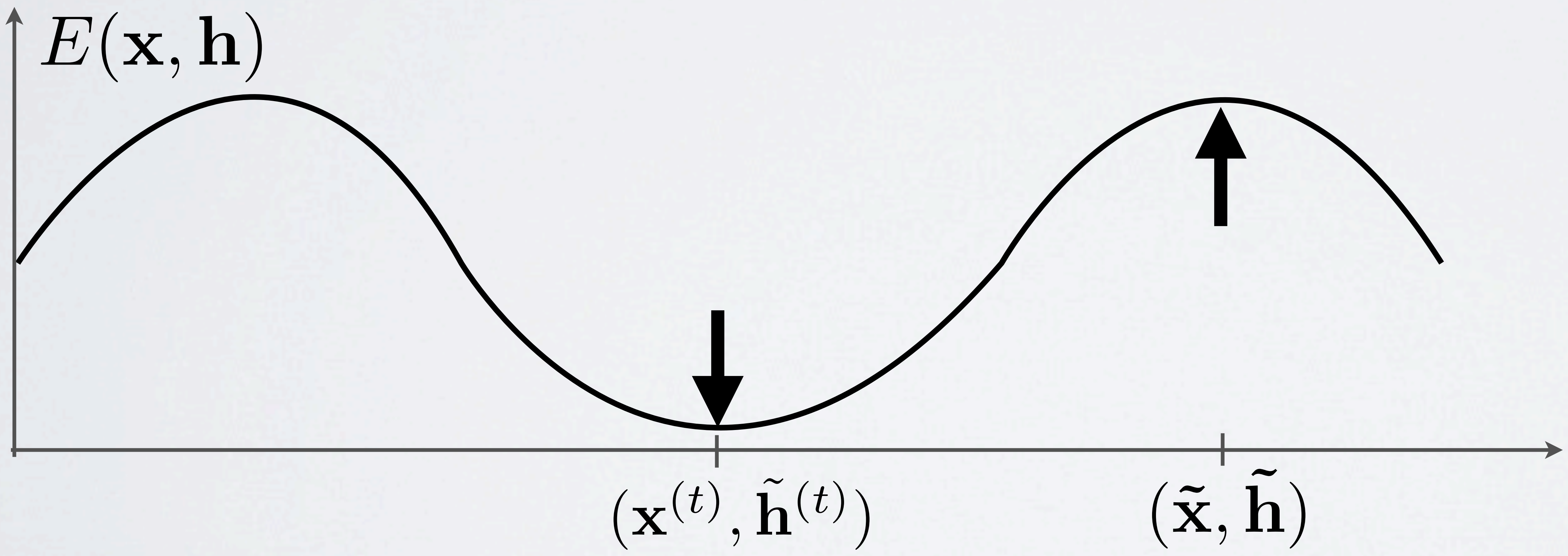


CONTRASTIVE DIVERGENCE (CD)

(HINTON, NEURAL COMPUTATION, 2002)

Topics: contrastive divergence, negative sample

$$\mathrm{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x}^{(t)},\mathbf{h})}{\partial\theta}\,\middle|\,\mathbf{x}^{(t)}\right] \approx \frac{\partial E(\mathbf{x}^{(t)},\tilde{\mathbf{h}}^{(t)})}{\partial\theta} \qquad \mathrm{E}_{\mathbf{x},\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial\theta}\right] \approx \frac{\partial E(\tilde{\mathbf{x}},\tilde{\mathbf{h}})}{\partial\theta}$$

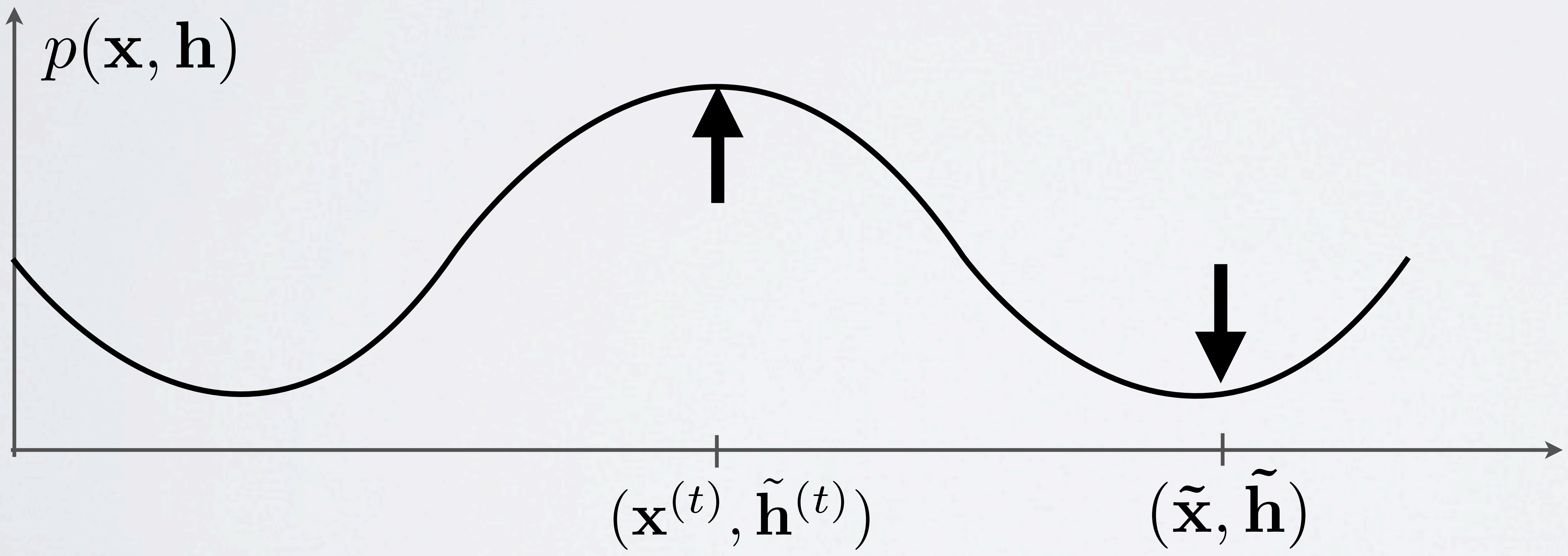


CONTRASTIVE DIVERGENCE (CD)

(HINTON, NEURAL COMPUTATION, 2002)

Topics: contrastive divergence, negative sample

$$\mathrm{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x}^{(t)},\mathbf{h})}{\partial\theta}\,\middle|\mathbf{x}^{(t)}\right]\approx\frac{\partial E(\mathbf{x}^{(t)},\tilde{\mathbf{h}}^{(t)})}{\partial\theta}\qquad\mathrm{E}_{\mathbf{x},\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial\theta}\right]\approx\frac{\partial E(\tilde{\mathbf{x}},\tilde{\mathbf{h}})}{\partial\theta}$$

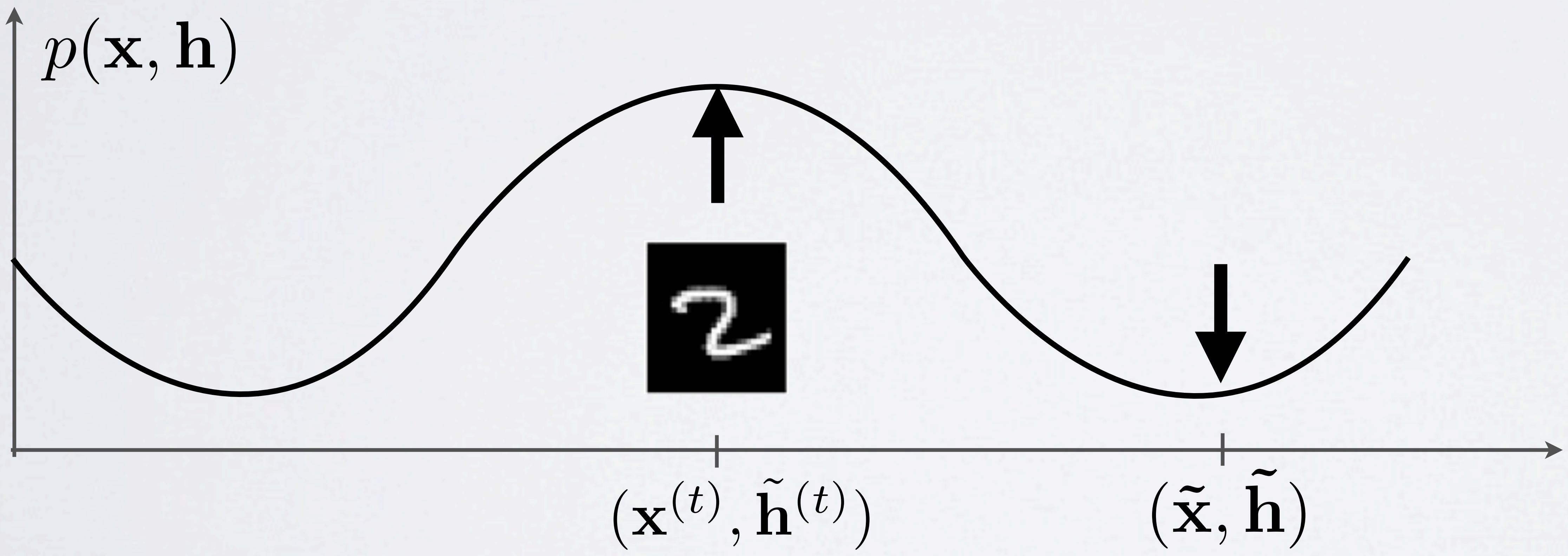


CONTRASTIVE DIVERGENCE (CD)

(HINTON, NEURAL COMPUTATION, 2002)

Topics: contrastive divergence, negative sample

$$\mathrm{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x}^{(t)},\mathbf{h})}{\partial\theta}\,\middle|\mathbf{x}^{(t)}\right] \approx \frac{\partial E(\mathbf{x}^{(t)},\tilde{\mathbf{h}}^{(t)})}{\partial\theta} \qquad \mathrm{E}_{\mathbf{x},\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial\theta}\right] \approx \frac{\partial E(\tilde{\mathbf{x}},\tilde{\mathbf{h}})}{\partial\theta}$$



CONTRASTIVE DIVERGENCE (CD)

(HINTON, NEURAL COMPUTATION, 2002)

Topics: contrastive divergence, negative sample

$$\mathrm{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x}^{(t)},\mathbf{h})}{\partial\theta}\,\middle|\,\mathbf{x}^{(t)}\right] \approx \frac{\partial E(\mathbf{x}^{(t)},\tilde{\mathbf{h}}^{(t)})}{\partial\theta} \qquad \mathrm{E}_{\mathbf{x},\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial\theta}\right] \approx \frac{\partial E(\tilde{\mathbf{x}},\tilde{\mathbf{h}})}{\partial\theta}$$