

mating a particular target distribution—in our case, the posterior distribution over the hidden units given the visible units—by some reasonably simple family of distributions. In the case of the mean field approximation, the approximating family is the set of distributions where the hidden units are conditionally independent.

We now develop the mean field approach for the example with two hidden layers. Let  $Q(\mathbf{h}^{(1)}, \mathbf{h}^{(2)} | \mathbf{v})$  be the approximation of  $P(\mathbf{h}^{(1)}, \mathbf{h}^{(2)} | \mathbf{v})$ . The mean field assumption implies that

$$Q(\mathbf{h}^{(1)}, \mathbf{h}^{(2)} | \mathbf{v}) = \prod_j Q(h_j^{(1)} | \mathbf{v}) \prod_k Q(h_k^{(2)} | \mathbf{v}). \quad (20.29)$$

The mean field approximation attempts to find a member of this family of distributions that best fits the true posterior  $P(\mathbf{h}^{(1)}, \mathbf{h}^{(2)} | \mathbf{v})$ . Importantly, the inference process must be run again to find a different distribution  $Q$  every time we use a new value of  $\mathbf{v}$ .

One can conceive of many ways of measuring how well  $Q(\mathbf{h} | \mathbf{v})$  fits  $P(\mathbf{h} | \mathbf{v})$ . The mean field approach is to minimize

$$\text{KL}(Q \| P) = \sum_{\mathbf{h}} Q(\mathbf{h}^{(1)}, \mathbf{h}^{(2)} | \mathbf{v}) \log \left( \frac{Q(\mathbf{h}^{(1)}, \mathbf{h}^{(2)} | \mathbf{v})}{P(\mathbf{h}^{(1)}, \mathbf{h}^{(2)} | \mathbf{v})} \right). \quad (20.30)$$

In general, we do not have to provide a parametric form of the approximating distribution beyond enforcing the independence assumptions. The variational approximation procedure is generally able to recover a functional form of the approximate distribution. However, in the case of a mean field assumption on binary hidden units (the case we are developing here) there is no loss of generality resulting from fixing a parametrization of the model in advance.

We parametrize  $Q$  as a product of Bernoulli distributions, that is we associate the probability of each element of  $\mathbf{h}^{(1)}$  with a parameter. Specifically, for each  $j$ ,  $\hat{h}_j^{(1)} = Q(h_j^{(1)} = 1 | \mathbf{v})$ , where  $\hat{h}_j^{(1)} \in [0, 1]$  and for each  $k$ ,  $\hat{h}_k^{(2)} = Q(h_k^{(2)} = 1 | \mathbf{v})$ , where  $\hat{h}_k^{(2)} \in [0, 1]$ . Thus we have the following approximation to the posterior:

$$Q(\mathbf{h}^{(1)}, \mathbf{h}^{(2)} | \mathbf{v}) = \prod_j Q(h_j^{(1)} | \mathbf{v}) \prod_k Q(h_k^{(2)} | \mathbf{v}) \quad (20.31)$$

$$= \prod_j (\hat{h}_j^{(1)})^{h_j^{(1)}} (1 - \hat{h}_j^{(1)})^{(1-h_j^{(1)})} \times \prod_k (\hat{h}_k^{(2)})^{h_k^{(2)}} (1 - \hat{h}_k^{(2)})^{(1-h_k^{(2)})}. \quad (20.32)$$

Of course, for DBMs with more layers the approximate posterior parametrization can be extended in the obvious way, exploiting the bipartite structure of the graph