

$$= b_i - \log \hat{h}_i + \log(1 - \hat{h}_i) + \mathbf{v}^\top \beta \mathbf{W}_{:,i} - \frac{1}{2} \mathbf{W}_{:,i}^\top \beta \mathbf{W}_{:,i} - \sum_{j \neq i} \mathbf{W}_{:,j}^\top \beta \mathbf{W}_{:,i} \hat{h}_j. \quad (19.43)$$

To apply the fixed point update inference rule, we solve for the \hat{h}_i that sets equation 19.43 to 0:

$$\hat{h}_i = \sigma \left(b_i + \mathbf{v}^\top \beta \mathbf{W}_{:,i} - \frac{1}{2} \mathbf{W}_{:,i}^\top \beta \mathbf{W}_{:,i} - \sum_{j \neq i} \mathbf{W}_{:,j}^\top \beta \mathbf{W}_{:,i} \hat{h}_j \right). \quad (19.44)$$

At this point, we can see that there is a close connection between recurrent neural networks and inference in graphical models. Specifically, the mean field fixed point equations defined a recurrent neural network. The task of this network is to perform inference. We have described how to derive this network from a model description, but it is also possible to train the inference network directly. Several ideas based on this theme are described in chapter 20.

In the case of binary sparse coding, we can see that the recurrent network connection specified by equation 19.44 consists of repeatedly updating the hidden units based on the changing values of the neighboring hidden units. The input always sends a fixed message of $\mathbf{v}^\top \beta \mathbf{W}$ to the hidden units, but the hidden units constantly update the message they send to each other. Specifically, two units \hat{h}_i and \hat{h}_j inhibit each other when their weight vectors are aligned. This is a form of competition—between two hidden units that both explain the input, only the one that explains the input best will be allowed to remain active. This competition is the mean field approximation’s attempt to capture the explaining away interactions in the binary sparse coding posterior. The explaining away effect actually should cause a multi-modal posterior, so that if we draw samples from the posterior, some samples will have one unit active, other samples will have the other unit active, but very few samples have both active. Unfortunately, explaining away interactions cannot be modeled by the factorial q used for mean field, so the mean field approximation is forced to choose one mode to model. This is an instance of the behavior illustrated in figure 3.6.

We can rewrite equation 19.44 into an equivalent form that reveals some further insights:

$$\hat{h}_i = \sigma \left(b_i + \left(\mathbf{v} - \sum_{j \neq i} \mathbf{W}_{:,j} \hat{h}_j \right)^\top \beta \mathbf{W}_{:,i} - \frac{1}{2} \mathbf{W}_{:,i}^\top \beta \mathbf{W}_{:,i} \right). \quad (19.45)$$

In this reformulation, we see the input at each step as consisting of $\mathbf{v} - \sum_{j \neq i} \mathbf{W}_{:,j} \hat{h}_j$ rather than \mathbf{v} . We can thus think of unit i as attempting to encode the residual