

exactly in closed form. Some other deep models, such as the deep Boltzmann machine, combine both the difficulty of an intractable partition function and the difficulty of intractable inference.

20.3 Deep Belief Networks

Deep belief networks (DBNs) were one of the first non-convolutional models to successfully admit training of deep architectures (Hinton *et al.*, 2006; Hinton, 2007b). The introduction of deep belief networks in 2006 began the current deep learning renaissance. Prior to the introduction of deep belief networks, deep models were considered too difficult to optimize. Kernel machines with convex objective functions dominated the research landscape. Deep belief networks demonstrated that deep architectures can be successful, by outperforming kernelized support vector machines on the MNIST dataset (Hinton *et al.*, 2006). Today, deep belief networks have mostly fallen out of favor and are rarely used, even compared to other unsupervised or generative learning algorithms, but they are still deservedly recognized for their important role in deep learning history.

Deep belief networks are generative models with several layers of latent variables. The latent variables are typically binary, while the visible units may be binary or real. There are no intralayer connections. Usually, every unit in each layer is connected to every unit in each neighboring layer, though it is possible to construct more sparsely connected DBNs. The connections between the top two layers are undirected. The connections between all other layers are directed, with the arrows pointed toward the layer that is closest to the data. See figure 20.1b for an example.

A DBN with l hidden layers contains l weight matrices: $\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(l)}$. It also contains $l+1$ bias vectors: $\mathbf{b}^{(0)}, \dots, \mathbf{b}^{(l)}$, with $\mathbf{b}^{(0)}$ providing the biases for the visible layer. The probability distribution represented by the DBN is given by

$$P(\mathbf{h}^{(l)}, \mathbf{h}^{(l-1)}) \propto \exp \left(\mathbf{b}^{(l)\top} \mathbf{h}^{(l)} + \mathbf{b}^{(l-1)\top} \mathbf{h}^{(l-1)} + \mathbf{h}^{(l-1)\top} \mathbf{W}^{(l)} \mathbf{h}^{(l)} \right), \quad (20.17)$$

$$P(h_i^{(k)} = 1 \mid \mathbf{h}^{(k+1)}) = \sigma \left(b_i^{(k)} + \mathbf{W}_{:,i}^{(k+1)\top} \mathbf{h}^{(k+1)} \right) \forall i, \forall k \in 1, \dots, l-2, \quad (20.18)$$

$$P(v_i = 1 \mid \mathbf{h}^{(1)}) = \sigma \left(b_i^{(0)} + \mathbf{W}_{:,i}^{(1)\top} \mathbf{h}^{(1)} \right) \forall i. \quad (20.19)$$

In the case of real-valued visible units, substitute

$$\mathbf{v} \sim \mathcal{N} \left(\mathbf{v}; \mathbf{b}^{(0)} + \mathbf{W}^{(1)\top} \mathbf{h}^{(1)}, \beta^{-1} \right) \quad (20.20)$$