

described as a mixture of Dirac distributions. Because Dirac distributions are not described by a single probability distribution function, no Dirac or mixture of Dirac distribution corresponds to a single specific point in function space. These distributions are thus invisible to our method of solving for a specific point where the functional derivatives are zero. This is a limitation of the method. Distributions such as the Dirac must be found by other methods, such as guessing the solution and then proving that it is correct.

19.4.3 Continuous Latent Variables

When our graphical model contains continuous latent variables, we may still perform variational inference and learning by maximizing \mathcal{L} . However, we must now use calculus of variations when maximizing \mathcal{L} with respect to $q(\mathbf{h} \mid \mathbf{v})$.

In most cases, practitioners need not solve any calculus of variations problems themselves. Instead, there is a general equation for the mean field fixed point updates. If we make the mean field approximation

$$q(\mathbf{h} \mid \mathbf{v}) = \prod_i q(h_i \mid \mathbf{v}), \quad (19.55)$$

and fix $q(h_j \mid \mathbf{v})$ for all $j \neq i$, then the optimal $q(h_i \mid \mathbf{v})$ may be obtained by normalizing the unnormalized distribution

$$\tilde{q}(h_i \mid \mathbf{v}) = \exp(\mathbb{E}_{\mathbf{h}_{-i} \sim q(\mathbf{h}_{-i} \mid \mathbf{v})} \log \tilde{p}(\mathbf{v}, \mathbf{h})) \quad (19.56)$$

so long as p does not assign 0 probability to any joint configuration of variables. Carrying out the expectation inside the equation will yield the correct functional form of $q(h_i \mid \mathbf{v})$. It is only necessary to derive functional forms of q directly using calculus of variations if one wishes to develop a new form of variational learning; equation 19.56 yields the mean field approximation for any probabilistic model.

Equation 19.56 is a fixed point equation, designed to be iteratively applied for each value of i repeatedly until convergence. However, it also tells us more than that. It tells us the functional form that the optimal solution will take, whether we arrive there by fixed point equations or not. This means we can take the functional form from that equation but regard some of the values that appear in it as parameters, that we can optimize with any optimization algorithm we like.

As an example, consider a very simple probabilistic model, with latent variables $\mathbf{h} \in \mathbb{R}^2$ and just one visible variable, v . Suppose that $p(\mathbf{h}) = \mathcal{N}(\mathbf{h}; 0, \mathbf{I})$ and $p(v \mid \mathbf{h}) = \mathcal{N}(v; \mathbf{w}^\top \mathbf{h}; 1)$. We could actually simplify this model by integrating out \mathbf{h} ; the result is just a Gaussian distribution over v . The model itself is not