

19.4 Variational Inference and Learning

We have seen how the evidence lower bound $\mathcal{L}(\mathbf{v}, \boldsymbol{\theta}, q)$ is a lower bound on $\log p(\mathbf{v}; \boldsymbol{\theta})$, how inference can be viewed as maximizing \mathcal{L} with respect to q , and how learning can be viewed as maximizing \mathcal{L} with respect to $\boldsymbol{\theta}$. We have seen that the EM algorithm allows us to make large learning steps with a fixed q and that learning algorithms based on MAP inference allow us to learn using a point estimate of $p(\mathbf{h} \mid \mathbf{v})$ rather than inferring the entire distribution. Now we develop the more general approach to variational learning.

The core idea behind variational learning is that we can maximize \mathcal{L} over a restricted family of distributions q . This family should be chosen so that it is easy to compute $\mathbb{E}_q \log p(\mathbf{h}, \mathbf{v})$. A typical way to do this is to introduce assumptions about how q factorizes.

A common approach to variational learning is to impose the restriction that q is a factorial distribution:

$$q(\mathbf{h} \mid \mathbf{v}) = \prod_i q(h_i \mid \mathbf{v}). \quad (19.17)$$

This is called the **mean field** approach. More generally, we can impose any graphical model structure we choose on q , to flexibly determine how many interactions we want our approximation to capture. This fully general graphical model approach is called **structured variational inference** (Saul and Jordan, 1996).

The beauty of the variational approach is that we do not need to specify a specific parametric form for q . We specify how it should factorize, but then the optimization problem determines the optimal probability distribution within those factorization constraints. For discrete latent variables, this just means that we use traditional optimization techniques to optimize a finite number of variables describing the q distribution. For continuous latent variables, this means that we use a branch of mathematics called calculus of variations to perform optimization over a space of functions, and actually determine which function should be used to represent q . Calculus of variations is the origin of the names “variational learning” and “variational inference,” though these names apply even when the latent variables are discrete and calculus of variations is not needed. In the case of continuous latent variables, calculus of variations is a powerful technique that removes much of the responsibility from the human designer of the model, who now must specify only how q factorizes, rather than needing to guess how to design a specific q that can accurately approximate the posterior.

Because $\mathcal{L}(\mathbf{v}, \boldsymbol{\theta}, q)$ is defined to be $\log p(\mathbf{v}; \boldsymbol{\theta}) - D_{\text{KL}}(q(\mathbf{h} \mid \mathbf{v}) \parallel p(\mathbf{h} \mid \mathbf{v}; \boldsymbol{\theta}))$, we can think of maximizing \mathcal{L} with respect to q as minimizing $D_{\text{KL}}(q(\mathbf{h} \mid \mathbf{v}) \parallel p(\mathbf{h} \mid \mathbf{v}))$.