16.2.2 Parametric Approximation of the Posterior Distribution

Instead of replacing $p(f|Z)$ by its mode, we may want to resort to slightly more sophisticated approximations. A first improvement is to use a normal distribution $\mathcal{N}(\mu, \sigma)$, with a mean μ coincides with the mode of $p(f|Z)$, and to use the second derivative of $-\ln p(f_{\text{MAP}}|Z)$ for the variance σ. This is often referred to as the *Gaussian Approximation*. In practice, we set (see for instance [338])

Gaussian Approximation

$$f|Z \sim \mathcal{N}(E[f|Z], \Sigma^{-1}) \text{ where } \Sigma = -\partial_f^2 \left[\ln p(f|Z)\right]\big|_{E[f|Z]}. \tag{16.28}$$

The advantage of such a procedure is that the integrals remain tractable. This is also one of the reasons why normal distributions enjoy a high degree of popularity in Bayesian methods. Besides, the normal distribution is the least informative distribution (largest entropy) among all distributions with bounded variance.

As Figure 16.2 indicates, a single Gaussian may not always be sufficient to capture the important properties of $p(y|X, Y, x)$. A more elaborate *parametric* model $q_\theta(f)$ of $p(f|X, Y)$, such as a mixture of Gaussian densities, can then be used to improve the approximation of (16.15). A common strategy is to resort to variational methods. The details are rather technical and go beyond the scope of this section. The interested reader is referred to [274] for an overview, and to [53] for an application to the Relevance Vector Machine of Section 16.6. The following theorem describes the basic idea.

Variational Approximation

Theorem 16.2 (Variational Approximation of Densities) *Denote by f, y random variables with corresponding densities $p(f, y), p(f|y)$, and $p(f)$. Then for any density $q(f)$, the following bound holds;*

$$\ln p(y) = \int_f \ln \frac{p(f,y)}{q(f)} q(f) df - \int_f \ln \frac{p(f|y)}{q(f)} q(f) df \leq \int_f \ln \frac{p(f,y)}{q(f)} q(f) df. \tag{16.29}$$

Proof We begin with the first equality of (16.29). Since $p(f, y) = p(f|y)p(y)$, we may decompose

$$\ln \frac{p(f,y)}{q(f)} = \ln p(y) + \ln \frac{p(f|y)}{q(f)}. \tag{16.30}$$

Additionally, $\int_f \ln \frac{p(f|y)}{q(f)} q(f) df = \mathrm{KL}(p(f|y) \| q(f))$ is the Kullback-Leibler divergence between $p(f|y)$ and $q(f)$ [114]. The latter is a nonnegative quantity which proves the second part of (16.29). ■

The true posterior distribution is usually $p(f|y)$, and $q(f)$ an approximation of it. The practical advantage of (16.29) is that $L := \ln \frac{p(f,y)}{q(f)} q(f) df$ can often be computed more easily, at least for simple enough $q(f)$. Furthermore, by maximizing L via a suitable choice of q, we maximize a lower bound on $\ln p(y)$.