12 Bayesian Methods

So far, we have considered statistical methods which select a single "best" model given the data (i.e., a so-called point estimate of model parameters). This approach can have problems, such as over-fitting when there is not enough data to fully constrain the model fit. In contrast, in the "pure" Bayesian approach, as much as possible we only compute distributions over unknowns; we never maximize anything. For example, consider a model parameterized by some weight vector $\mathbf{w}$, and some training data $\mathcal{D}$ that comprises input-output pairs x_i, y_i, for $i = 1...N$. The posterior probability distribution over the parameters, conditioned on the data is, using Bayes' rule, given by

$$p(\mathbf{w} \,|\, \mathcal{D}) \;=\; \frac{p(\mathcal{D} \,|\, \mathbf{w}) p(\mathbf{w})}{p(\mathcal{D})} \,. \tag{12.1}$$

The reason we want to fit the model in the first place is to allow us to make predictions with future test data. That is, given some future input x_{new}, we want to use the model to predict y_{new}. To accomplish this task through estimation in previous chapters, we used optimization to find maximum likelihood (ML) or maximum a posteriori (MAP) estimates of $\mathbf{w}$, e.g., by maximizing (12.1).

In a Bayesian approach, rather than estimation a single best value for $\mathbf{w}$, we compute (or approximate) the entire posterior distribution $p(\mathbf{w}|\mathcal{D})$. Given the entire distribution, we can still make predictions with the following integral:

$$\begin{aligned} p(y_{new} \,|\, \mathcal{D}, x_{new}) \;&=\; \int p(y_{new}, \mathbf{w} \,|\, \mathcal{D}, x_{new}) \; d\mathbf{w} \\ &=\; \int p(y_{new} \,|\, \mathbf{w}, \mathcal{D}, x_{new}) \, p(\mathbf{w} \,|\, \mathcal{D}, x_{new}) \; d\mathbf{w} \end{aligned} \tag{12.2}$$

The first step in this equality follows from the Sum Rule. The second follows from the Product Rule. Additionally, the outputs y_{new} and training data $\mathcal{D}$ are independent conditioned on $\mathbf{w}$, so $p(y_{new} \,|\, \mathbf{w}, \mathcal{D}) = p(y_{new} \,|\, \mathbf{w})$. That is, given $\mathbf{w}$, we have all available information about making predictions that we could possibly get from the training data $\mathcal{D}$ (according to the model). Finally, given $\mathcal{D}$, it is safe to assume that x_{new}, in itself, provides no information about $\mathbf{W}$. With these assumptions we have the following expression for our predictions:

$$p(y_{new} \,|\, \mathcal{D}, x_{new}) \;=\; \int p(y_{new} \,|\, \mathbf{w}, x_{new}) \, p(\mathbf{w} \,|\, \mathcal{D}) \; d\mathbf{w} \tag{12.3}$$

In the case of discrete parameters $\mathbf{w}$, the integral becomes a summation.

The posterior distribution $p(y_{new} \,|\, \mathcal{D}, x_{new})$ tells us everything there is to know about our beliefs about the prediction y_{new}. It does so by combining two sources of information. One is the distribution over the target y given the weights $\mathbf{w}$; e.g., where the distribution $p(y|\mathbf{w}, x)$, and its intrinsic uncertainty, is often derived from a model of the form $y = f(\mathbf{x}, \mathbf{w}) + \eta$, where η is a random variable used to model measurement noise. The second source of information (and uncertainty) is what we know about the model parameters $\mathbf{w}$, characterized by the posterior distribution $p(\mathbf{w} \,|\, \mathcal{D})$. In essence, the posterior $p(y_{new} \,|\, \mathcal{D}, x_{new})$ is a weighted average of distributions, $p(y|\mathbf{w}, x)$, over all possible values of the model parameters $\mathbf{w}$, weighted by how well they fit the data, i.e., $p(\mathbf{w}|\mathcal{D})$.

And there are many things we can do with the prediction distribution in (12.3). For example, we could pick the most likely prediction, i.e., $\arg\max_y p(y_{new} \mid \mathcal{D}, x_{new})$, or we could compute the variance of this distribution to get a sense of how much confidence we have in the prediction. We could sample from this distribution in order to visualize the range of models that are plausible for this data.

The integral in (12.3) is rarely easy to compute, often involving intractable integrals or exponentially large summations. Thus, Bayesian methods often rely on numerical approximations, such as Monte Carlo sampling. MAP estimation can also be viewed as an approximation, for which the posterior distribution is approximated by a delta function. In some cases, however, the Bayesian computations can be done exactly, an example of which is the Bayesian approach to regression below.

12.1 Bayesian Regression

Recall the statistical model used in basis-function regression:

$$y \;=\; \mathbf{b}(x)^T\mathbf{w} + \eta\,, \quad \eta \sim \mathcal{N}(0, \sigma^2) \tag{12.4}$$

for a fixed set of basis functions $\mathbf{b}(x) = [b_1(x), ...b_M(x)]^T$.

To complete the model, we also need to define a "prior" distribution over the weights $\mathbf{w}$ (denoted $p(\mathbf{w})$). The prior which expresses what we believe about $\mathbf{w}$ in absence of any training data. One might be tempted to assign a constant density over all possible weights. There are several problems with this. First, the result cannot be a valid probability distribution since no choice of the constant will give the density a finite integral. We could, instead, choose a uniform distribution with finite bounds, however, this will make the resulting computations more complex.

More importantly, a uniform prior is often inappropriate. We often find that smoother functions are more likely in practice (at least for functions that we have any hope in learning), and so we should employ a prior that prefers smooth functions. A choice of prior that does so is a Gaussian prior:

$$\mathbf{w} \sim \mathcal{N}(0, \alpha^{-1}\mathbf{I}) \tag{12.5}$$

which expresses a prior belief that smooth functions are more likely.[1] This prior also has the additional benefit that it will lead to tractable integrals later on. Note that this prior depends on a parameter α. We will see later in this chapter how this "hyper-parameter" can be determined automatically as well.

As developed in previous chapters on regression, the data likelihood function that follows from the above model definition (with input and output components of the training data, denoted $x_{1:N}$ and $y_{1:N}$) is

$$p(y_{1:N} \mid x_{1:N}, \mathbf{w}) \;=\; \prod_{i=1}^{N} p(y_i \mid x_i, \mathbf{w})\,, \tag{12.6}$$

[1] Why such a prior encourages smoothness is not always obvious. But in many cases of regression, for example, the gradient of the model is linear in the weights, so smaller weights means smaller gradients.

and so the posterior is:

$$p(\mathbf{w} \,|\, x_{1:N}, y_{1:N}) \;=\; \frac{\left(\prod_{i=1}^{N} p(y_i \,|\, x_i, \mathbf{w})\right) p(\mathbf{w})}{p(y_{1:N} \,|, x_{1:N})} \,. \tag{12.7}$$

In the negative log-domain, using Equations (12.4) and (12.5), the model is given by

$$\begin{aligned} -\ln p(\mathbf{w} \,|\, x_{1:N}, y_{1:N}) &= -\sum_{i=1}^{N} \ln p(y_i \,|\, x_i, \mathbf{w}) \;-\; \ln p(\mathbf{w}) \;+\; \ln p(y_{1:N} \,|\, x_{1:N}) \\ &= \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - f(x_i))^2 \;+\; \frac{\alpha}{2} ||\mathbf{w}||^2 \;+\; \text{constants} \end{aligned}$$

As in the regression chapters above, it is useful if we collect the training outputs into a single vector, i.e., $\mathbf{y} = [y_1, ..., y_N]^T$, and we collect the all basis functions evaluated at each of the inputs into a matrix $\mathbf{B}$, with elements $\mathbf{B}_{i,j} = b_j(x_i)$. In doing so we can simplify the expression of the log posterior as follows

$$\begin{aligned} -\ln p(\mathbf{w} \,|\, x_{1:N}, y_{1:N}) &= \frac{1}{2\sigma^2} ||\mathbf{y} - \mathbf{B}\mathbf{w}||^2 + \frac{\alpha}{2} ||\mathbf{w}||^2 + \text{constants} \\ &= \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{B}\mathbf{w})^T (\mathbf{y} - \mathbf{B}\mathbf{w}) + \frac{\alpha}{2} \mathbf{w}^T \mathbf{w} + \text{constants} \\ &= \frac{1}{2} \mathbf{w}^T \left(\frac{\mathbf{B}^T \mathbf{B}}{\sigma^2} + \alpha \mathbf{I} \right) \mathbf{w} - \frac{1}{2} \frac{\mathbf{y}^T \mathbf{B} \mathbf{w}}{\sigma^2} - \frac{1}{2} \frac{\mathbf{w}^T \mathbf{B}^T \mathbf{y}}{\sigma^2} + \text{constants} \\ &= \frac{1}{2} (\mathbf{w} - \bar{\mathbf{w}})^T \mathbf{K}^{-1} (\mathbf{w} - \bar{\mathbf{w}}) + \text{constants} \end{aligned} \tag{12.8}$$

where

$$\mathbf{K} \;=\; \left(\frac{\mathbf{B}^T \mathbf{B}}{\sigma^2} + \alpha \mathbf{I} \right)^{-1} \qquad \text{and} \qquad \bar{\mathbf{w}} \;=\; \frac{\mathbf{K} \mathbf{B}^T \mathbf{y}}{\sigma^2} \tag{12.9}$$

(The last step of the derivation uses a technique referred to as *completing the square*. It is easiest to verify the last step by going backwards, that is by multiplying out $(\mathbf{w} - \bar{\mathbf{w}})^T \mathbf{K}^{-1} (\mathbf{w} - \bar{\mathbf{w}})$.)

The derivation above tells us that the posterior distribution over the weight vector is a multi-dimensional Gaussian with mean $\bar{\mathbf{w}}$ and covariance matrix $\mathbf{K}$, i.e.,

$$p(\mathbf{w} \,|\, x_{1:N}, y_{1:N}) \;=\; G(\mathbf{w}; \bar{\mathbf{w}}, \mathbf{K}) \tag{12.10}$$

In other words, our belief about $\mathbf{w}$ once we have seen the data is specified by a Gaussian density. We believe that $\bar{\mathbf{w}}$ is the most probable value for $\mathbf{w}$, but we have uncertainty about this estimate, as determined by the covariance $\mathbf{K}$. The covariance expresses our uncertainty about these parameters. If the covariance is very small, then we have a lot of confidence in the MAP estimate. The nature of the posterior distribution is illustrated visually in Figure 12.1. Note that $\bar{\mathbf{w}}$ is the MAP estimate for regression, since it maximizes the posterior.

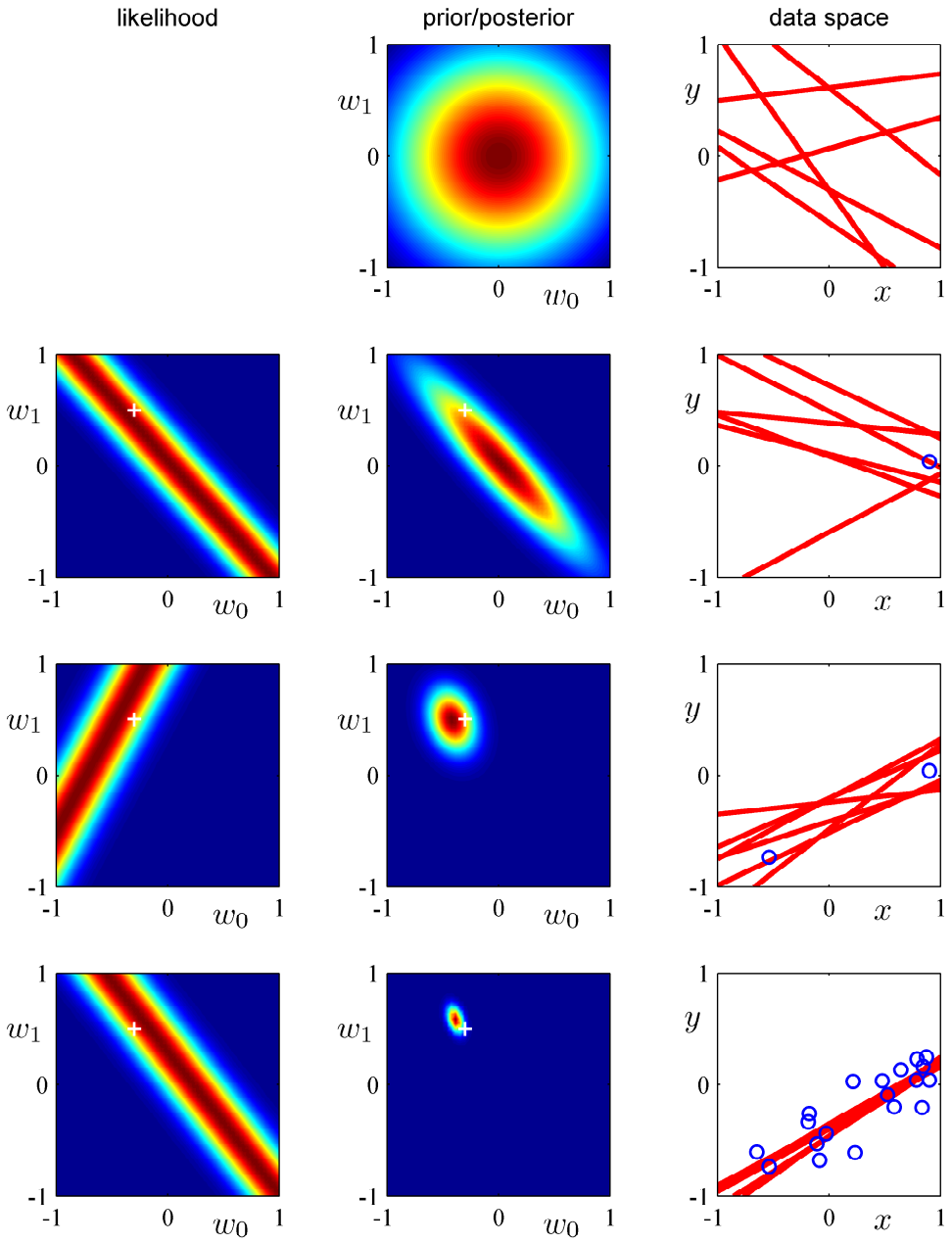

Figure 12.1: Iterative posterior computation for a linear regression model: $y = w_0 x + w_1$. The top row shows the prior distribution, and several fair samples from the prior distribution. The second row shows the likelihood over w after observing a single data point (i.e., an x, y pair), along with the resulting posterior (the normalized product of the likelihood and the prior), and then several fair samples from the posterior. The third row shows the liklihood when a new observation is added to the previous observation, followed by the corresponding posterior and random samples from the posterior. The final row shows the result of 20 observations.

Prediction. For a new data point x_{new}, the predictive distribution for y_{new} is given by:

$$\begin{aligned} p(y_{new} \,|\, x_{new}, \mathcal{D}) &= \int p(y_{new} \,|\, x_{new}, \mathcal{D}, \mathbf{w})\; p(\mathbf{w} \,|\, \mathcal{D})\, d\mathbf{w} \\ &= G(y_{new}; \mathbf{b}(x_{new})^T \bar{\mathbf{w}},\; \sigma^2 + \mathbf{b}(x_{new})^T \mathbf{K} \mathbf{b}(x_{new})) \;. \end{aligned}$$

To show that the prediction distribution has this form requires some work, and the use of some identities associated with Gaussian distributions. In particular, as explained at the end of Chapter 6 on the Gaussian Probability Density Function, the product of two Gaussians is Gaussian (albeit unnormalized), and the marginalization of a multi-dimensional Gaussian is Gaussian. Further, if $\mathbf{x}$ is a Gaussian random vector with mean μ_x and covariance Σ_x, and a new random vector is equal to $\mathbf{y} = A\mathbf{x}$ for some matrix A, then one can show that $\mathbf{y}$ is Gaussian with mean $A\mu_x$, and covariance matrix $A\Sigma_x A^T$. If $\mathbf{y} = A\mathbf{x} + \eta$ where η is mean-zero Gaussian with covariance Σ_η then $\mathbf{y}$ has mean $A\mu_x$ and covariance matrix $\Sigma_\eta + A\Sigma_x A^T$.

This is the Bayesian way to do regression. The predictive distribution may be viewed as a function from x_{new} to a distribution over values of y_{new}. An example of this for an RBF model is depicted in Figure 12.2. To predict a new value y_{new} for an input x_{new}, we don't estimate a single model $\mathbf{w}$. Instead we average over all possible models, weighting the different models according to their posterior probability.

12.2 Hyper-Parameters

There are often implicit parameters in our model that we hold fixed, such as the covariance constants in linear regression, or the parameters that govern the prior distribution over the weights. These are usually called "hyper-parameters." For example, in the RBF model, the hyper-parameters constitute the parameters α, σ^2, and the parameters of the basis functions (e.g., the width of the basis functions). Thus far we have assumed that the hyper-parameters were "known" (which means that someone must set them by hand), or estimated by cross-validation (which has a number of pitfalls, including long computation times, especially for large numbers of hyper-parameters). Instead of either of these approaches, we may apply the Bayesian approach in order to directly estimate these values as well.

To find a MAP estimate for the α parameter in the above linear regression example we compute:

$$\alpha^* \;=\; \arg\max \ln p(\alpha \,|\, x_{1:N}, y_{1:N}) \;, \tag{12.11}$$

where

$$p(\alpha \,|\, x_{1:N}, y_{1:N}) \;=\; \frac{p(y_{1:N} \,|\, x_{1:N}, \alpha)\; p(\alpha)}{p(y_{1:N} \,|\, x_{1:N})} \;, \tag{12.12}$$

and

$$\begin{aligned} p(y_{1:N} \,|\, x_{1:N}, \alpha) &= \int p(y_{1:N}, \mathbf{w} \,|\, x_{1:N}, \alpha) d\mathbf{w} \\ &= \int p(y_{1:N} \,|\, x_{1:N}, \mathbf{w}, \alpha) p(\mathbf{w} \,|\, \alpha) d\mathbf{w} \\ &= \int \left(\prod_i p(y_i \,|\, x_i, \mathbf{w}, \alpha) \right)\; p(\mathbf{w} \,|\, \alpha) d\mathbf{w} \end{aligned}$$

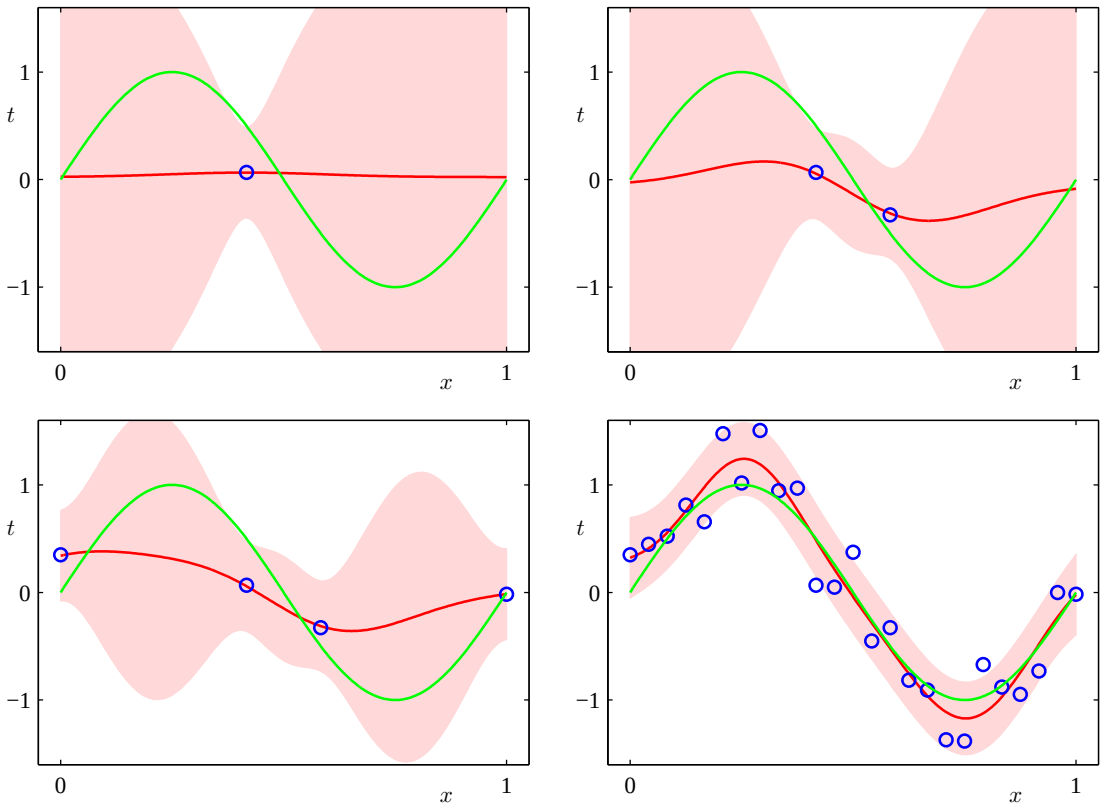

Figure 12.2: Predictive distribution for an RBF model (with 9 basis functions), trained on noisy sinusoidal data. The green curve is the true underlying sinusoidal function. The blue circles are data points. The red curve is the mean prediction as a function of the input. The pink region represents 1 standard deviation. Note how this region shrinks close to where more data points are observed. (Figure from *Pattern Recognition and Machine Learning* by Chris Bishop.)

For RBF regression, in the Gaussian case, this objective can be solved in closed-form. However, depending on the form of the prior over the hyper-parameters, it is often necessary to use some form of numerical optimization, such as gradient descent to find α.

12.3 Bayesian Model Selection

Cross-validation is one approach to model selection, where we measure the relative performance of different models on data held out from training. But it can be extremely expensive, as it require fitting multiple models, often many times to different subsets of data. Unless one has has a reasonably large dataset it can be unreliable.

Bayesian approaches offer a principled alternative, often capturing the general rule of thumb that one should choose simple models over complex models. That is, other things being equal (like how well they fit the data), it is hoped that simpler models are less prone to over-fitting.

By way of formulation, given data $\mathcal{D}$, assume there are L candidate models $\{\mathcal{M}_i\}_{i=1}^L$, each with a corresponding parameter vector $\mathbf{w}_i$. One might choose to use the model that maximizes the model posterior, $P(\mathcal{M}_i \,|\, \mathcal{D})$, or perhaps the one with the best data likelihood $p(\mathcal{D} \,|\, \mathcal{M}_i)$. Of course these two distributions are related through Bayes' rule:

$$P(\mathcal{M}_i \,|\, \mathcal{D}) \;=\; \frac{p(\mathcal{D} \,|\, \mathcal{M}_i)\, P(\mathcal{M}_i)}{p(\mathcal{D})} \;.$$

If we assume all models are equally likely *a priori*, then $P(\mathcal{M}_i)$ is constant, and the two approaches are essentially the same. In practice, a uniform prior over models may not be appropriate, but the design of suitable priors in these cases will depend significantly on one's knowledge of the application domain. So, for our purposes in this chapter we'll assume a uniform prior over models and focus on $p(\mathcal{D} \,|\, \mathcal{M}_i)$, often called the *marginal data likelihood*.

Given two models, $\mathcal{M}_1$ and $\mathcal{M}_2$, we will choose model $\mathcal{M}_1$ when $p(\mathcal{D} \,|\, \mathcal{M}_1) > p(\mathcal{D} \,|\, \mathcal{M}_2)$. But to understand these quantities we need to take the model parameters into account, i.e., $\mathbf{w}_i$ for model $\mathcal{M}_i$. Again, it is important to remember that different models will often have different numbers of parameters (e.g., polynomials of different degrees, differents numbers of basis function etc). We include model parameters by rewriting the marginal data likelihood as follows:

$$\begin{aligned} p(\mathcal{D} \,|\, \mathcal{M}_i) \;&=\; \int p(\mathcal{D}, \mathbf{w}_i \,|\, \mathcal{M}_i)\, d\mathbf{w}_i \\ &=\; \int p(\mathcal{D} \,|\, \mathbf{w}_i, \mathcal{M}_i)\, p(\mathbf{w}_i \,|\, \mathcal{M}_i)\, d\mathbf{w}_i \;. \end{aligned} \tag{12.13}$$

Equation (12.13) tells us that more probable models will have both $p(\mathbf{w}_i \,|\, \mathcal{M}_i)$ and $p(\mathcal{D} \,|\, \mathbf{w}_i,\, \mathcal{M}_i)$ large for the same weight vectors. In other words, a model is considered good if it assign high prior probability to the same weight vectors that also fit the data well (ie yield high likelihoods).

As an aside, remember than when we first introduced regression from a probabilistic point of view, we determined the MAP estimate of the model parameters, $\mathbf{w}$, by maximizing the posterior; i.e.,

$$\max_{\mathbf{w}} p(\mathbf{w}|\mathcal{D}) \;=\; \max_{\mathbf{w}} \frac{p(\mathcal{D} \,|\, \mathbf{w})\, p(\mathbf{w})}{p(\mathcal{D})} \;. \tag{12.14}$$

There are three important things to note here. First, although we didn't mention it at the time, all the quantities above in Equation (12.14) are implicitly conditioned on a particlar class of model (eg K-th order polynomials). We didn't mention it explicitly early on since we weren't comparing different classes of model. We were only lookinng for the best parameters given the model class. Second, we ignored the denominator in this regression optimization since to find the MAP estimate we only need to consider terms that explicitly depend on the parameters $\mathbf{w}$. And third, we called that denominator the 'evidence' without much explanation at the time. Note that the evidence, i.e., $p(\mathcal{D})$ or to explicitly condition on the model, $p(\mathcal{D} \,|\, \mathcal{M})$, is precisely the quantity of interest above. That is, it is the evidence with which we select one model over another.

Returning to the issue at hand, i.e., Equation (12.13), we note that the prior $p(\mathbf{w}_i \,|\, \mathcal{M}_i)$ in plays a particularly important role in model selection. One way to measure model complexity is the effective size of the parameter space, or perhaps the entropy of the prior. For discrete parameters we might count the number of possible parameter values. For continuous parameters we might count the number of parameters and their ranges (up to uncertainty in parameter estimation). And a model with a larger parameter space will assign lower prior probability to any one parameter value. That is, the prior must integrate to one, so a complex model spreads its prior probability mass more thinly over its parameter space. So if two models fit the data equally well for some range of parameters, then the more complex model will have a lower marginal data likelihood since the prior density in (12.13) will be smaller. In this way, the marginal data likelihood captures a very natural bias toward simpler models, whose range of parameters in the prior is well matched to the range of parameters for which the data likelihood $p(\mathcal{D} \,|\, \mathbf{w}_i, \mathcal{M}_i)$ is large.

As an aid to intuition on model selection, consider a simple approximation to the marginal data likelihood $p(\mathcal{D} \,|\, \mathcal{M}_i)$ (depicted in Fig. 12.3 for a single scalar parameter w). First, as is common for many problems of interest, let's assume the posterior distribution over the model parameters, $p(\mathbf{w}_i \,|\, \mathcal{D}, \mathcal{M}_i) \propto p(\mathcal{D} \,|\, \mathbf{w}_i, \mathcal{M}_i) p(\mathbf{w}_i \,|\, \mathcal{M}_i)$, has a strong peak at the MAP estimate $\mathbf{w}_i^{MAP}$. Accordingly, we can approximate the integral in Eqn. (12.13) as the height of the peak, i.e., $p(\mathcal{D} \,|\, \mathbf{w}_i^{MAP}, \mathcal{M}_i)\, p(\mathbf{w}_i^{MAP} \,|\, \mathcal{M}_i)$, multiplied by width of the peak, $\Delta \mathbf{w}_i^{posterior}$:

$$\int p(\mathcal{D} \,|\, \mathbf{w}_i, \mathcal{M}_i) p(\mathbf{w}_i \,|\, \mathcal{M}_i) d\mathbf{w}_i \;\; \approx \;\; p(\mathcal{D} \,|\, \mathbf{w}_i^{MAP}, \mathcal{M}_i)\, p(\mathbf{w}_i^{MAP} \,|, \mathcal{M}_i)\, \Delta \mathbf{w}_i^{posterior}$$

Then, assume that the prior over the parameter space, $p(\mathbf{w}_i | \mathcal{M}_i)$, is a relatively broad uniform with width $\Delta \mathbf{w}_i^{prior}$, so $p(\mathbf{w}_i) \approx \frac{1}{\Delta \mathbf{w}_i^{prior}}$. This yields a further approximation:

$$p(\mathcal{D} \,|\, \mathcal{M}_i) \;\; = \;\; \int p(\mathcal{D} \,|\, \mathbf{w}_i, \mathcal{M}_i)\, p(\mathbf{w}_i \,|\, \mathcal{M}_i) d\mathbf{w}_i \;\; \approx \;\; \frac{p(\mathcal{D} \,|\, \mathbf{w}_i^{MAP}, \mathcal{M}_i)\, \Delta \mathbf{w}_i^{posterior}}{\Delta \mathbf{w}_i^{prior}}$$

Taking the logarithm, this becomes

$$\ln p(\mathcal{D} \,|\, \mathcal{M}_i) \;\; \approx \;\; \ln p(\mathcal{D} | \mathbf{w}_i^{MAP}, \mathcal{M}_i) + \ln \frac{\Delta \mathbf{w}_i^{posterior}}{\Delta \mathbf{w}_i^{prior}}$$

In most cases of interest we can assume $\Delta \mathbf{w}_i^{posterior} < \Delta \mathbf{w}_i^{prior}$. That is, one can usually assume that the variance of the posterior is smaller than the variance of the prior. The log ratio is

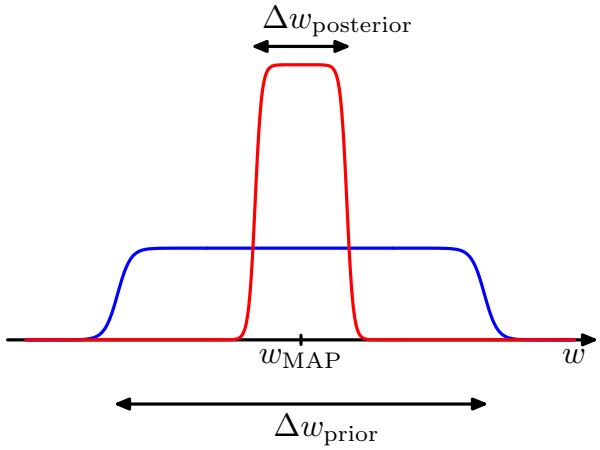

Figure 12.3: A visualization of the width-based evidence approximation. (Figure from *Pattern Recognition and Machine Learning* by Chris Bishop.)

maximal when the prior and posterior widths are equal. A complex model with many parameters, or a broad prior over the parameter space will necessarily assign small probability to any single parameter value (including those under the posterior peak). A simpler model will assign a higher prior probability to the useful parameter range under the posterior peak. When the model is too simple, the likelihood in the integrand tends to suffer, which lowers the marginal data likelihood. As models become more complex the data likelihood increasingly fits the data better. But as the models become more and more complex the log ratio $\ln \frac{\Delta \mathbf{w}_i^{posterior}}{\Delta \mathbf{w}_i^{prior}}$ acts as a penalty on unnecessarily complex models.

By selecting a model with the highest marginal data likelihood we are automatically balancing model complexity with the ability of the model to capture the data. This can be seen as the mathematical realization of Occam's Razor.

Model averaging. To be fully Bayesian, arguably, we shouldn't select a single "best" model, but instead combine estimates from all models according to their respective posterior probabilities:

$$p(y_{new} \,|\, \mathcal{D}, x_{new}) \;=\; \sum_i p(y_{new} \,|\, \mathcal{M}_i, \mathcal{D}, x_{new})\, P(\mathcal{M}_i \,|\, \mathcal{D}) \tag{12.15}$$

but this is often impractical and so we resort to model selection instead.

