11 Bayesian Methods

So far, we have considered statistical methods which select a single "best" model given the data. 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

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \tag{1}$$

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 ML or MAP estimates of $\mathbf{w}$, e.g., by maximizing (1).

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

$$\begin{aligned} p(y_{new}|\mathcal{D}, x_{new}) &= \int p(y_{new}, \mathbf{w}|\mathcal{D}, x_{new}) d\mathbf{w} \\ &= \int p(y_{new}|\mathbf{w}, \mathcal{D}, x_{new})\, p(\mathbf{w}|\mathcal{D}, x_{new}) d\mathbf{w} \end{aligned} \tag{2}$$

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:

$$p(y_{new}|\mathcal{D}, x_{new}) = \int p(y_{new}|\mathbf{w}, x_{new})\, p(\mathbf{w}|\mathcal{D}) d\mathbf{w} \tag{3}$$

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 new value y_{new}. There are many things we can do with this distribution. For example, we could pick the most likely prediction, i.e., $\arg\max_y p(y_{new}|\mathcal{D}, x_{new})$, 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 (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. However, in a few cases, the Bayesian computations can be done exactly, as in the regression case discussed below.

11.1 Bayesian Regression

Recall the statistical model used in basis-function regression:

$$y \;=\; \mathbf{b}(x)^T\mathbf{w} + n, \quad n \sim \mathcal{N}(0, \sigma^2) \tag{4}$$

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 $p(\mathbf{w})$) 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:

$$\mathbf{w} \sim N(0, \alpha^{-1}\mathbf{I}) \tag{5}$$

which expresses a prior belief that smooth functions are more likely. 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 "hyperparameter" 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 the input and output components of the training dataset denoted $x_{1:N}$ and $y_{1:N}$) is

$$p(y_{1:N}|x_{1:N}, \mathbf{w}) = \prod_{i=1}^{N} p(y_i|x_i, \mathbf{w}) \tag{6}$$

and so the posterior is:

$$p(\mathbf{w}|x_{1:N}, y_{1:N}) = \frac{\left(\prod_{i=1}^{N} p(y_i|x_i, \mathbf{w})\right) p(\mathbf{w})}{p(y_{1:N}|x_{1:N})} \tag{7}$$

In the negative log-domain, using Equations (4) and (5), the model is given by:

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

As above in the regression notes, 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 log posterior as follows:

$$\begin{aligned} -\ln p(\mathbf{w}|x_{1:N}, y_{1:N}) &= \frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{B}\mathbf{w}||^2 + \frac{\alpha}{2}||\mathbf{w}||^2 + \text{constants} \\ &= \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{B}\mathbf{w})^T(\mathbf{y} - \mathbf{B}\mathbf{w}) + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} + \text{constants} \\ &= \frac{1}{2}\mathbf{w}^T(\mathbf{B}^T\mathbf{B}/\sigma^2 + \alpha\mathbf{I})\mathbf{w} - \frac{1}{2}\mathbf{y}^T\mathbf{B}\mathbf{w}/\sigma^2 - \frac{1}{2}\mathbf{w}^T\mathbf{B}^T\mathbf{y}/\sigma^2 + \text{constants} \\ &= \frac{1}{2}(\mathbf{w} - \bar{\mathbf{w}})^T\mathbf{K}^{-1}(\mathbf{w} - \bar{\mathbf{w}}) + \text{constants} \qquad (8) \end{aligned}$$

where

$$\begin{aligned} \mathbf{K} &= \left(\mathbf{B}^T\mathbf{B}/\sigma^2 + \alpha\mathbf{I}\right)^{-1} \qquad (9) \\ \bar{\mathbf{w}} &= \mathbf{K}\mathbf{B}^T\mathbf{y}/\sigma^2 \qquad (10) \end{aligned}$$

(The last step of the derivation uses the methods of *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 multidimensional Gaussian with mean $\bar{\mathbf{w}}$ and covariance matrix $\mathbf{K}$, i.e.,

$$p(\mathbf{w}|x_{1:N}, y_{1:N}) = G(\mathbf{w}; \bar{\mathbf{w}}, \mathbf{K}) \qquad (11)$$

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 1. Note that $\bar{\mathbf{w}}$ is the MAP estimate for regression, since it maximizes the posterior.

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

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

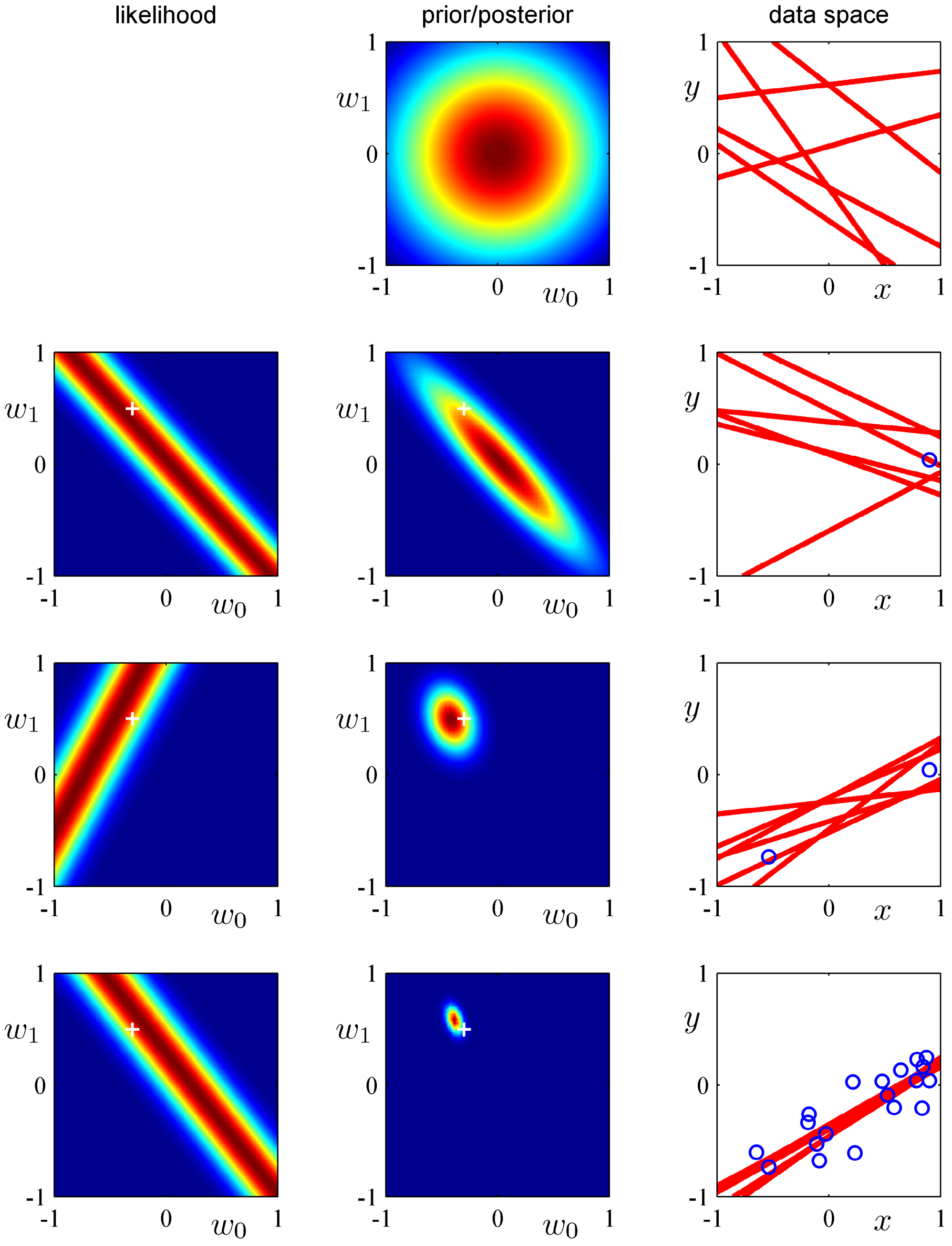

Figure 1: Iterative posterior computation for a linear regression model: $y = w_0 x + w_1$. The top row shows the prior distribution, and several fair samples from the prior distribution. The second row shows the likelihood over w after observing a single data point (i.e., an x, y pair), along with the 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.

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 then 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 given in Figure 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.

11.2 Hyperparameters

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 "hyperparameters." For example, in the RBF model, the hyperparameters 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 hyperparameters 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 hyperparameters). 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:

$$\alpha^* = \arg\max \ln p(\alpha|x_{1:N}, y_{1:N}) \tag{12}$$

where

$$p(\alpha|x_{1:N}, y_{1:N}) = \frac{p(y_{1:N}|x_{1:N}, \alpha)p(\alpha)}{p(y_{1:N}|x_{1:N})} \tag{13}$$

and

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

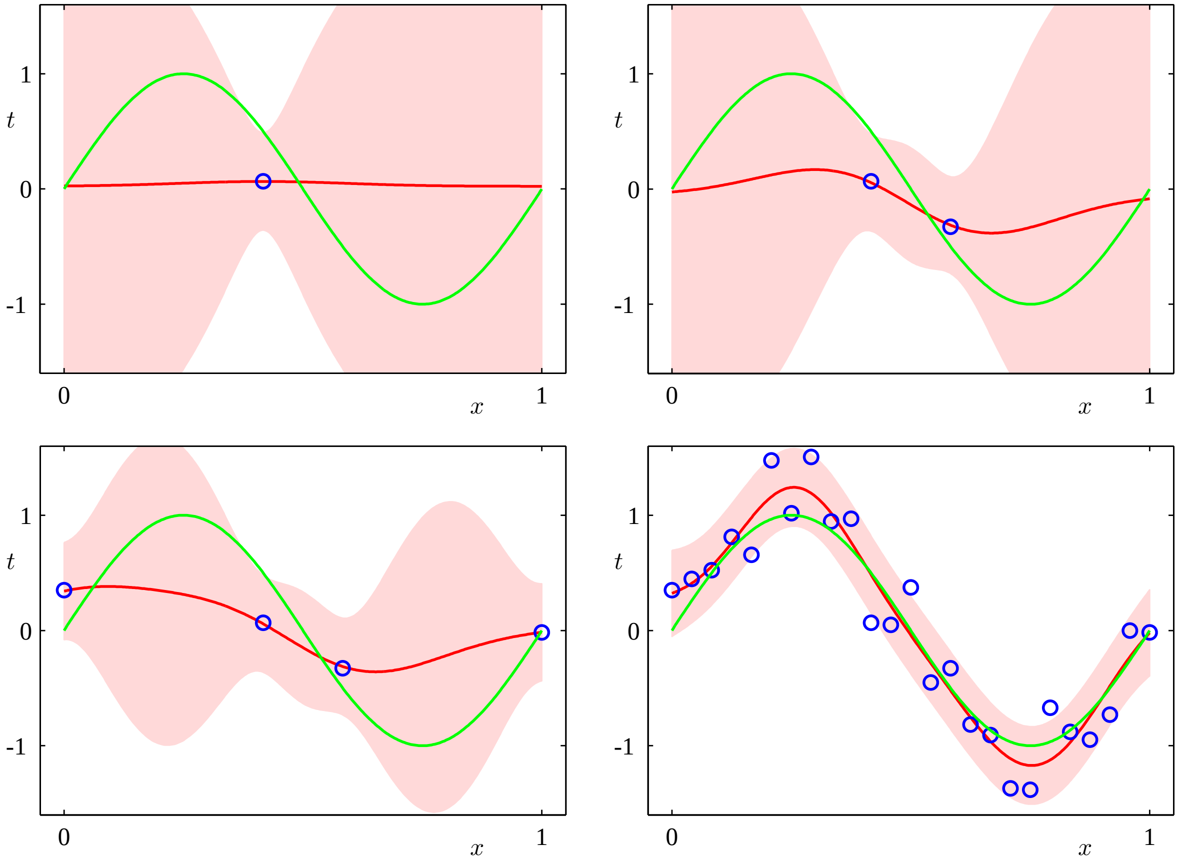

Figure 2: Predictive distribution for an RBF model (with 9 basis functions), trained on noisy sinusoidal data. The green curve is the true underlying sinusoidal function. The blue circles are data points. The red curve is the mean prediction as a function of the input. The pink region represents 1 standard 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, this objective function can be computed in closed-form. However, depending on the form of the prior over the hyperparameters, it is often necessary to use some form of numerical optimization, such as gradient descent.

11.3 Bayesian Model Selection

How do we choose which model to use? For example, we might like to automatically choose the form of the basis functions or the number of basis functions. Cross-validation is one approach, but it can be expensive, and, more importantly, inaccurate if small amounts of data are available. In general one intuition is that we want to choose simple models over complex models to avoid overfitting,insofar as they provide equivalent fits to the data. Below we consider a Bayesian approach to model selection which provides just such a bias to simple models.

The goal of model selection is to choose the best model from some set of candidate models $\{\mathcal{M}_i\}_{i=1}^{L}$ based on some observed data $\mathcal{D}$. This may be done either with a maximum likelihood approach (picking the model that assigns the largest likelihood to the data) or a MAP approach (picking the model with the highest posterior probability). If we take a uniform prior over models (i.e. $p(\mathcal{M}_i)$ is a constant for all $i = 1...L$) then these approaches can be seen to be equivalent since:

$$\begin{aligned} p(\mathcal{M}_i|\mathcal{D}) &= \frac{p(\mathcal{D}|\mathcal{M}_i)p(\mathcal{M}_i)}{p(\mathcal{D})} \\ &\propto p(\mathcal{D}|\mathcal{M}_i) \end{aligned}$$

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 here we will assume a uniform prior over models and focus on $p(\mathcal{D}|\mathcal{M}_i)$.

In some sense, whenever we estimate a parameter in a model we are doing model selection where the family of models is indexed by the different values of that parameter. However the term "model selection" can also mean choosing the best model from some set of parametric models that are parameterized differently. A good example of this would be choosing the number of basis functions to use in an RBF regression model. Another simple example is choosing the polynomial degree for polynomial regression.

The key quantity for Bayesian model selection is $p(\mathcal{D}|\mathcal{M}_i)$, often called the *marginal data likelihood*. Given two models, $\mathcal{M}_1$ and $\mathcal{M}_2$, we will choose the model $\mathcal{M}_1$ when $p(\mathcal{D}|\mathcal{M}_1) > p(\mathcal{D}|\mathcal{M}_1)$. To specify these quantities in more detail we need to take the model parameters into account. Different models may have different numbers of parameters (e.g., polynomials of different degrees), or entirely different parameterizations (e.g., RBFs and neural networks). In what follows, let $\mathbf{w}_i$ be the vector of parameters for model $\mathcal{M}_i$. In the case of regression, for example, $\mathbf{w}_i$ might comprise the regression weights and hyper-parameters like the weight on the regularizer.

The extent to which a model explains (or fits) the data depends on the choice of the right parameters. Using the sum rule and Bayes' rule it follows we can write the marginal data likelihood as

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}, \mathbf{w}_i|\mathcal{M}_i)d\mathbf{w}_i = \int p(\mathcal{D}|\mathbf{w}_i, \mathcal{M}_i)p(\mathbf{w}_i|\mathcal{M}_i)d\mathbf{w}_i \tag{14}$$

This tells us that the ideal model is one that assigns high prior probability $p(\mathbf{w}_i|\mathcal{M}_i)$ to every weight vector that also yields a high value of the likelihood $p(\mathcal{D}|\mathbf{w}_i, \mathcal{M}_i)$ (i.e., to parameter vectors that fit the data well). One can also recognize that the product of the data likelihood and the prior in the integrand is proportional to the posterior over the parameters that we previously maximized to find MAP estimates of the model parameters. [1]

Typically, a "complex model" that assigns a significant posterior probability mass to complex data will be able to assign significantly less mass to simpler data than a simpler model would. This is because the integral of the probability mass must sum to 1 and so a complex model will have less mass to spend on simpler data. Also, since a complex model will require higher-dimensional parameterizations, mass must be spread over a higher-dimensional space and hence more thinly. This phenomenon is visualized in Figure 4.

As an aid to intuition to explain why this marginal data likelihood helps us choose good models, we consider a simple approximation to the marginal data likelihood $p(\mathcal{D}|\mathcal{M}_i)$ (depicted in Figure 3 for a scalar parameter w). First, as is common in many problems of interest, the posterior distribution over the model parameters $p(\mathbf{w}_i|\mathcal{D}, \mathcal{M}_i) \propto p(\mathcal{D}|\mathbf{w}_i, \mathcal{M}_i)p(\mathbf{w}_i|\mathcal{M}_i)$ to have a strong peak at the MAP parameter estimate $\mathbf{w}_i^{MAP}$. Accordingly we can approximate the integral in Equation (14) 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 its width $\Delta\mathbf{w}_i^{posterior}$.

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

We then assume that the prior distribution over parameters $p(\mathbf{w}_i|\mathcal{M}_i)$ 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:

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

Taking the logarithm, this becomes

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

Intuitively, this approximation tells us that models with wider prior distributions on the parameters will tend to assign less likelihood to the data because the wider prior captures a larger variety of data (so the density is spread thinner over the data-space). Similarly, models that have a very narrow peak around their modes are generally less preferable because they assign a lower probability mass to the surrounding area (and so a slightly perturbed setting of the parameters would provide a poor fit to the data, suggesting that over-fitting has occurred).

[1] This is the same quantity we compute when optimizing hyper-parameters (which is a type of model selection) and also corresponds to the denominator "$p(\mathcal{D})$" in Bayes' rule for finding the posterior probability of a particular setting of the parameters $\mathbf{w}_i$. Note that above we generally wrote $p(\mathcal{D})$ and not $p(\mathcal{D}|\mathcal{M}_i)$ because we were only considering a single model, and so it was not necessary to condition on it.

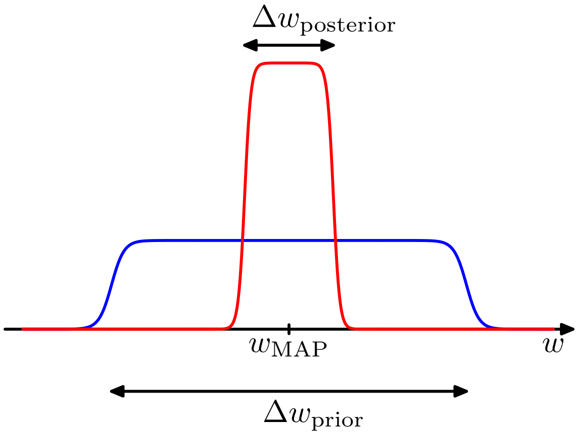

Figure 3: A visualization of the width-based evidence approximation. (Figure from *Pattern Recognition and Machine Learning* by Chris Bishop.)

From another perspective, note that in most cases of interest we can assume that $\Delta \mathbf{w}_i^{posterior} < \Delta \mathbf{w}_i^{prior}$. I.e., the posterior width will be less than the width of the prior. The log ratio is maximal when the prior and posterior widths are equal. For example, a complex model with many parameters, or a a very broad prior over the parameters will necessarily assign a small probability to any single value (including those under the posterior peak). A simpler model will assign a higher prior probability to the useful parameter values (ie those under the posterior peak). When the model is too simple, then the likelihood term in the integrand will be particularly high and therefore lowers the marginal data likelihood. So, 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 that assigns the highest posterior probability to the data 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 should instead combine estimates from all models according to their respective posterior probabilities:

$$p(y_{new}|\mathcal{D}, x_{new}) = \sum_i p(y_{new}|\mathcal{M}_i, \mathcal{D}, x_{new})\, p(\mathcal{M}_i|\mathcal{D}) \tag{15}$$

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

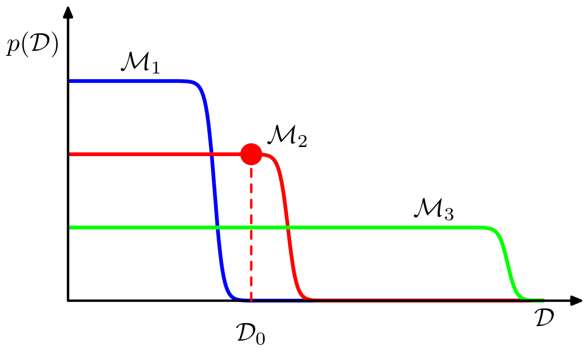

Figure 4: The x-axis is data complexity from simplest to most complex, and models $\mathcal{M}_i$ are indexed in order of increasing complexity. Note that in this example M_2 is the best model choice for data $\mathcal{D}_0$ since it simultaneously is complex enough to assign mass to $\mathcal{D}_0$ but not so complex that it must spread its mass too thinly. (Figure from *Pattern Recognition and Machine Learning* by Chris Bishop.)