Bayesian Linear Regression 6

In this chapter we will introduce the simplest Bayesian machine learning approach, namely Bayesian linear regression. Under a Gaussian likelihood model for observations, the solution can be computer analytically, requiring a matrix inversion from a computational perspective. We will start by an introduction to Gaussian distributions, Bayesian inference and linear regression (to fix notation). Then we will dig into Bayesian regression, discuss the role of model evidence or marginal likelihood and briefly touch upon model comparison.

6.1 Gaussian Distribution

We are going to view in detail a number of useful properties of Multivariate Gaussian Distribution, which will be very useful in the following sections and chapters.

6.1.1 Definition

Let's start by defining the probability density of the d-dimensional Multivariate Gaussian, denoted by $N(x|\mu, \Sigma)$, where μ is a d-dimensional vector and represents the mean of the Gaussian and Σ is a $d \times d$ positive definite matrix, called **Covariance matrix**

$$\mathcal{N}(x|\mu, \Sigma) = ((2\pi)^d \det(\Sigma))^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

Sometimes $\Sigma^{-1} = A$, called the **Precision matrix**, is used instead of the covariance matrix in the definition of a Gaussian. We also refer to $(x-\mu)^T\Sigma^{-1}(x-\mu)$ as the **Mahalanobis distance**. Notice that having $\Sigma = I$ one obtains the **euclidean distance**.

6.1.2 Principal components

Since Σ is positive definite, we can diagonalize it and decompose it in $\Sigma = E\Lambda E^T$ where $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix composed of the eigenvalues of Σ and E is an orthogonal matrix (i.e. such that $EE^T = I$ holds) whose rows are eigenvectors of Σ.

We can do a change of coordinate in this way:

$$y = \Lambda^{-\frac{1}{2}} E^T (x - \mu)$$

Then we have that

$$(x-\mu)^T\Sigma^{-1}(x-\mu) = (x-\mu)^T E\Lambda^{-\frac{1}{2}}\Lambda^{-\frac{1}{2}}E^T(x-\mu) = y^T y$$

Practically, we obtain a Gaussian distribution with mean zero and Covariance distribution equal to the identity, $\mathcal{N}(x|0, I)$. Geometrically we are roto-translating the ellipsoids describing the level sets of a Gaussian Distribution into a sphere centered in the axis, such that points at one standard from the mean have distance 1 from the origin. This linear change of basis can always be performed.

6.1.3 Completing the square

Suppose that we have a probability density like

$$p(x) = c \cdot \exp\left(-\frac{1}{2}x^T Ax - b^T x\right)$$

This is actually a Gaussian distribution; to show it we need to do some algebra on $\log p(x)$. The following identity can be proved:

$$-\frac{1}{2}x^T Ax - b^T x = \frac{1}{2}(x - A^{-1}b)^T A(x - A^{-1}b) - \frac{1}{2}b^T A^{-1}b$$

Therefore we can use the properties of the exponential to get

$$c\cdot\exp(-\frac{1}{2}x^T Ax - b^T x) = \mathcal{N}(x|A^{-1}b, A^{-1}) \underbrace{\sqrt{(2\pi)^d \det A^{-1}} \cdot \exp(-\frac{1}{2}b^T A^{-1}b) \cdot c}_{=1}$$

which means that

$$p(x) = \mathcal{N}(x|A^{-1}b, A^{-1}).$$

Staten otherwise: **every distribution which is an exponential of a quadratic form is a Gaussian distribution**.

6.1.4 Further closure properties

The following properties will not be proved. You can find more details in the Bishop book

- **Linear transformation** Suppose to have $y = Mx + \eta$ where $x \sim \mathcal{N}(\mu_x, \Sigma_x)$, $\eta \sim \mathcal{N}(\mu, \Sigma)$, $x \perp\!\!\!\perp \eta$. Then y is Gaussian, $y \sim \mathcal{N}(M\mu_x + \mu, M\Sigma_x M^T + \Sigma)$. This means that Gaussians are closed under linear transformations.
- **Marginals and conditionals** Assume that

$$z = \begin{bmatrix} x \\ y \end{bmatrix} \qquad z \sim \mathcal{N}(\mu, \Sigma)$$
$$\mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

  The marginal distribution is pretty easy to obtain:

$$x \sim \mathcal{N}(\mu_x, \Sigma_{xx})$$

While the conditional distribution is just a little more complicated

$$p(x|y) = \mathcal{N}(x|\mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y), \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx})$$

- **Product of Gaussians** The product of two Gaussians densities is still a Gaussian

$$\mathcal{N}(x|\mu_1, \Sigma_1)\mathcal{N}(x|\mu_2, \Sigma_2) = \mathcal{N}(x|\mu, \Sigma) \cdot K$$

where

$$\begin{aligned} S &= \Sigma_1 + \Sigma_2 \\ \mu &= \Sigma_1 S^{-1}\mu_2 + \Sigma_2 S^{-1}\mu_1 \\ \Sigma &= \Sigma_1 S^{-1}\Sigma_2 \\ K &= \frac{\exp\left(-\frac{1}{2}(\mu_1 - \mu_2)^T S^{-1}(\mu_1 - \mu_2)\right)}{\sqrt{\det(2\pi S)}} \end{aligned}$$

- **Bayesian Theorem** Supposing to have

$$x \sim \mathcal{N}(x|\mu, A^{-1}), \qquad p(y|x) = \mathcal{N}(y|Mx + b, L^{-1})$$

Then the joint distribution of x and y is still a Gaussian

$$z = \begin{bmatrix} x \\ y \end{bmatrix}$$

In fact

$$\ln p(z) = \ln p(x) + \ln p(y|x) =$$
$$\text{const } -\frac{1}{2}(x - \mu)^T A(x - \mu) - \frac{1}{2}(y - Mx - b)^T L(y - Mx - b)$$

By completing the square we obtain a Gaussian with the following mean and covariance

$$z \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ M\mu + b \end{bmatrix}, R^{-1}\right)$$

Where

$$R = \begin{bmatrix} A + M^T LM & -M^T L \\ -LM & L \end{bmatrix}$$

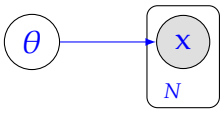
And

$$R^{-1} = \begin{bmatrix} A^{-1} & A^{-1}M^T \\ MA^{-1} & L^{-1} + MA^{-1}M^T \end{bmatrix}$$

And then we can apply the marginalization and the conditional distributions formula that we have seen before to compute $p(y)$ and $p(x|y)$.

6.2 Bayesian Estimation

Consider n observations of i.i.d. random variables $\underline{x} = x_1, \ldots, x_n$ and a family of models $p(x|\theta)$, corresponding to our likelihood distribution, in which we are looking for the model that best fits our data.



In the Bayesian context, we also have a **prior** distribution $p(\theta)$ and we can compute the posterior distribution

$$p(\theta|\underline{x}) = \frac{p(\underline{x}|\theta)p(\theta)}{p(\underline{x})}$$

The problem in this scenario, as usual, is computing the **marginal likelihood**, because $p(\underline{x}) = \int p(\underline{x}|\theta)p(\theta)d\theta$ is a hard integral to approximate in a high dimensional setting.

Now, we could compute the maximum a-posteriori, i.e. $\theta_{\text{map}} = \max_\theta p(\theta|\underline{x})$, but in this way, from the estimated parameters, we would obtain only point estimates as output. Instead, we want to get the entire distribution in order to have a complete representation of uncertainty. So, it's more convenient to compute the **predictive distribution** of x by using all the information contained in the posterior

$$p(x|\underline{x}) = \int p(x|\theta)p(\theta|\underline{x})d\theta$$

Which is obtained by the usual factorization

$$\int p(x,\theta|\underline{x})d\theta = \int p(x|\theta,\underline{x})p(\theta|\underline{x})d\theta = \int p(x|\theta)p(\theta|\underline{x})d\theta$$

Let's study an example: suppose to have $x_1,\dots,x_n \in \{0,1\}$ so that $p(\underline{x}|\theta) = \text{Bernoulli}(\theta)$, and that we observed (1) k times and (0) $n-k$ times. A good choice for our prior in this scenario is

$$p(\theta) = \text{Beta}(\theta|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Where B is the **Beta function**. It represents a family of distributions having domain in $[0,1]$, which can be skewed more toward 0 or 1 by playing with the parameters. It can be showed that

$$\mathbb{E}_{\text{Beta}(\alpha,\beta)}[\theta] = \frac{\alpha}{\alpha+\beta}$$

By using the prior and the likelihood we can compute the logarithm of the posterior

$$\begin{aligned}\log p(\theta|\underline{x}) &= L(\theta) + \log B(\theta|\alpha,\beta) - \log p(\underline{x}) \\ &= k\log\theta + (n-k)\log(1-\theta) + (\alpha-1)\log\theta + (\beta-1)\log(1-\theta) + C \\ &= C + (k+\alpha-1)\log\theta + (N-k+\beta-1)\log(1-\theta)\end{aligned}$$

where the term C incorporates all the terms that are independent of θ. By recognising the functional form of a Beta distribution we arrive at the

last equality

$$\log p(\theta|\underline{x}) = \log \text{Beta}(\theta|k+\alpha, N-k+\beta)$$

The predictive distribution is then defined through the following expectation:

$$p(x=1|\underline{x}) = \int p(x=1|\theta)p(\theta|\underline{x})d\theta = \int_0^1 \theta \text{Beta}(\theta|k+\alpha, N-k+\beta)d\theta = \frac{k+\alpha}{N+\alpha+\beta}$$

The Bernoulli and the Beta distribution are an example of **conjugate priors**.

Definition 6.2.1 *We say that the prior distribution and the likelihood distribution are **conjugate priors** if the corresponding posterior distribution has the the same functional form of the prior.*

6.3 Introduction to Linear Regression

We will start with an introduction to linear regression. This will serve as a recap of notions that will be expanded in a Bayesian setting later on.

We are moving from the problem of describing probabilistic models and performing inference on them, to the problem of **supervised learning**.

We have data, in the form of $\underline{x}, \underline{y} : (x_i, y_i)$ where $i = 1, \dots N$, i.e we have pairs of inputs and outputs. Assume that $p(y|x) = p(y|x, \theta)$ is a parametric model of our random variables. At first, we are going to choose θ_{ML} with a maximum likelihood approach

$$\theta_{ML} = \text{argmax } p(\underline{y}|\underline{x}, \theta)$$

Therefore what we need to do is to identify our parametric model, which in linear regression is just

$$p(y|x, \theta) = \mathcal{N}(y|f(x, w), \beta^{-1})$$

Notice that in this case $\theta = (w, \beta)$. We can equivalently rewrite this with the (perhaps more familiar) notation

$$y = f(x, w) + \epsilon \qquad \epsilon \sim \mathcal{N}(0, \beta^{-1})$$

In particular, in linear regression, we will have that our function f is linear with respect to our weights w, that is

$$f(x, w) = w_0\phi_0(x) + \dots + w_{M-1}\phi_{M-1}(x)$$

Notice that ϕ can be, and usually are, non-linear functions of the input data. They are the **basis function** for our regression model. They can be monomials, Gaussian RBF, sigmoids, ...

This means that the log likelihood of our model is, in fact

$$\log p(y|x,\theta) \propto -E_D(w) = -\frac{1}{2}\sum_{i=1}^{N}\left(y_i - w^T\phi(x_i)\right)^2$$

Minimizing the sum of squares E_D means maximizing the likelihood, hence, taking the gradient of the function we get

$$\nabla_w E_D(w) = \sum_{i=1}^{N}\left(y_i - w^T\phi(x_i)\right)\phi^T(x_i) = 0$$

Which yields the following close form for our weights

$$w_M = \left(\Phi^T\Phi\right)^{-1}\Phi^T\underline{y}$$

Definition 6.3.1 *$\Phi_{ij} = \phi_j(x_i)$ is called the **design matrix**, it is the $j-th$ feature (basis function) evaluated on the i-th datum.*

If M is large, solving the direct problem might be computationally difficult, but we can rely on optimization algorithms, such as gradient descent, to actually find the minimum of our negative log-likelihood, exploiting the fact that we are dealing with a quadratic form here.

In order to avoid overfitting, especially if the chosen basis functions are enough complex and expressive, we seldom introduce further regularization terms in the loss function, such as

$$E_W(w) = \begin{cases} \frac{1}{2}||w||_2^2 & \text{Ridge} \\ \frac{1}{2}||w||_1 & \text{Lasso} \end{cases}$$

Therefore we will minimize the quadratic loss plus one of the two penalty terms above:

$$E_D(w) + \lambda E_W(w)$$

where λ is the regularization coefficient.

6.3.1 Example

As an example, we can generate synthetic datasets by adding Gaussian noise to a set of points belonging to the curve $y = sin(2\pi x)$

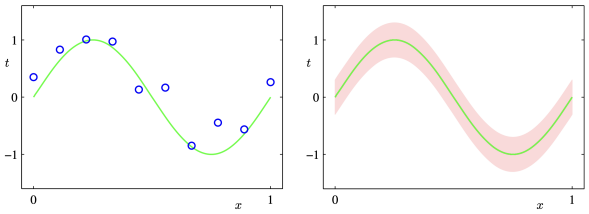


Figure 6.1: On the left, the sinusoidal function and the generated data points. On the right, the true conditional distribution $p(t|x)$ in which the green curve denotes the mean and the shaded region spans one standard deviation on each side of the mean (Bishop)

We build 100 data sets, each having 25 data points, and we perform Linear Regression using 24 Gaussian basis functions on the datasets varying the regularization coefficient λ.

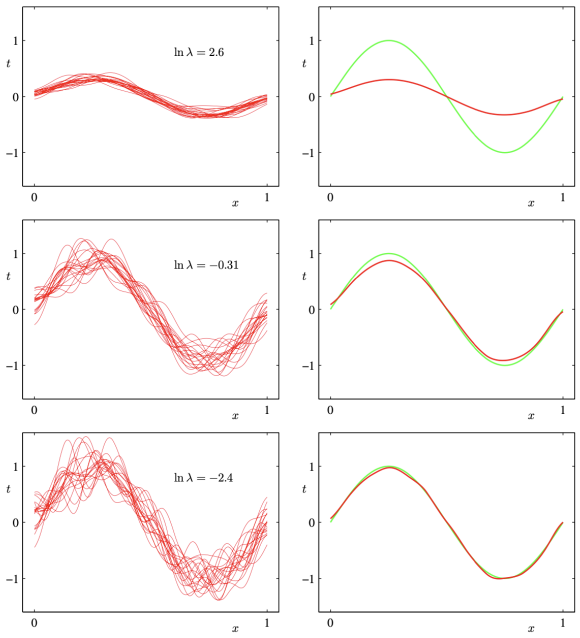


Figure 6.2: On the left, the result of fitting the model to the different data sets varying $ln(\lambda)$. On the right, the average of the fits on the 100 datasets (Bishop)

6.4 Bayesian Linear Regression

By adding the regularization term, we are modifying the loss function in order to obtain better results in terms of overfitting, but one of the drawbacks is that we lose a nice probabilistic interpretation of our model: the object we are minimizing is not a negative likelihood anymore. How can we go back towards a more rigorous probabilistic setting?

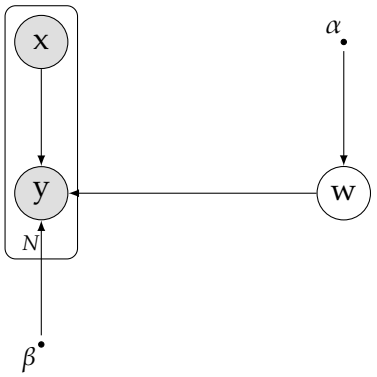
The key point to observe in order to do this step is that the regularization term is, in fact, a **bias** that we introduce in our data. What we can do instead of adding a penalty term in the loss, is to encode the bias in a **prior** distribution of our parameters, $p(w|\alpha)$ and then treat our problem in a Bayesian way. For example

$$p(w|\alpha) = \mathcal{N}(w|0, \alpha^{-1}I)$$

where α is a **hyper-parameter** of our distribution (we will see some methods to choose it). This is a typical choice for our bias since then our weights will be forced to be small (which is the goal of regularization).

For the moment let's suppose we have fixed our α. Since we have a likelihood of our observation, we can compute the posterior. Let's also introduce a Gaussian noise in the observations with precision β.

We can visualize the model in graphical terms as:



Applying Bayes theorem, the posterior is

$$p(w|\underline{x}, \underline{y}, \alpha, \beta) = \frac{p(\underline{y}|\underline{x}, w, \beta)p(w|\alpha)}{p(\underline{y}|\underline{x}, \alpha, \beta)}$$

In this scenario, choosing a Gaussian prior and having the Gaussian likelihood of the linear regression problem, we have an analytical form for our posterior distribution, as we are about to see.

Let's consider the logarithm of the posterior first

$$\log p(w|\underline{x}, \underline{y}, \alpha, \beta) = -\frac{\beta}{2}\sum_{j=1}^{N}\left(y_i - w^T\phi(x_i)\right)^2 - \alpha w^T w + \text{const}$$

The logarithm of the marginal likelihood does not depend on w and is treated as a constant. The trick is to notice that we have a quadratic function of w, and, as we have seen in the first paragraph, if the logarithm of a distribution is a quadratic function then that distribution is a Gaussian.

$$p(w|\underline{x}, \underline{y}, \alpha, \beta) = \mathcal{N}(w|m_N, S_N)$$

where

$$m_N = \beta S_N \Phi^T \underline{y}$$
$$S_N^{-1} = \alpha I + \beta \Phi^T \Phi$$

which is very similar (but not equal) to what we get as a solution in Ridge Regression.

If we use a general Gaussian prior instead, of the form $p(w|m_0, S_0) = \mathcal{N}(w|m_0, S_0)$, then our posterior becomes

$$p(w|\underline{x}, \underline{y}, \alpha, \beta) = \mathcal{N}(w|m_N, S_N)$$

with

$$m_N = S_N\left[S_0^{-1}m_0 + \beta\Phi^T\underline{y}\right]$$
$$S_N^{-1} = S_0^{-1} + \beta\Phi^T\Phi$$

So, in Bayesian regression, we treat our parameters probabilistically, placing a prior distribution over them, computing the posterior given the observation that we have and we using it to make predictions. We have seen that for a Gaussian prior we have a Gaussian posterior, and we also have an analytically computable posterior, given that we know how to invert matrices.

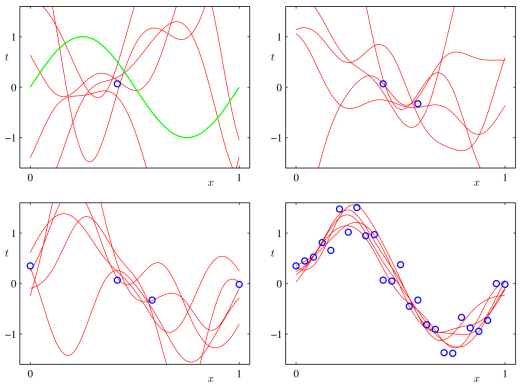


Figure 6.3: Samples from the posterior distribution of a Bayesian regression model on the dataset shown in section 6.3.1. Increasing the size of the dataset, the predictive distribution approximates better the true data distribution (Bishop)

6.4.1 Online Learning

There is an extra feature which is typical of Bayesian learning. When we first pick a prior distribution we are basically having random weights, corresponding to random lines in the data space (remember what the weights actually represent!). Once we compute the posterior, we are reshaping the Gaussian distribution from which we sample our weights, up until it becomes highly centered towards a point once we get a lot of observations. In this process, nothing forbids us to start from a prior that already takes into account some observations! This is a general principle of Bayesian learning: when we observe new data points, we use as new prior the posterior relative to the previous observations. Hence, Bayesian linear regression, is, *naturally*, an **online method**, which means that every time we observe new data we can easily incorporate them into our model without retraining the model from the start!

Example

In order to visualize how the posterior distribution is updated when including new training data, we consider the simple example reported in the Bishop's book. We want to fit a linear model of the form $f(x, \mathbf{w}) = w_0 + w_1 \cdot x$, assuming α and β known. The columns of Figure 7.1 show:

- First column: the likelihood of the last data point $(\underline{x}, \underline{y})$ added to the training set as a function of $\mathbf{w}$, i.e. $p(\underline{y}|\underline{x}, \mathbf{w})$
- Second column: the posterior distribution obtained multiplying the prior (which is the posterior of the previous row) by the likelihood reported in the same row
- Third column: some samples of the regression function obtained by drawing samples of $\mathbf{w}$ from the posterior distribution

The first row corresponds to the situation before any data points are observed: the prior distribution of w_0, w_1 is a multivariate standard

normal distribution. In the next rows this distribution is reshaped by the information contained in the dataset and the posterior distribution becomes sharper and centred on the true parameter values.

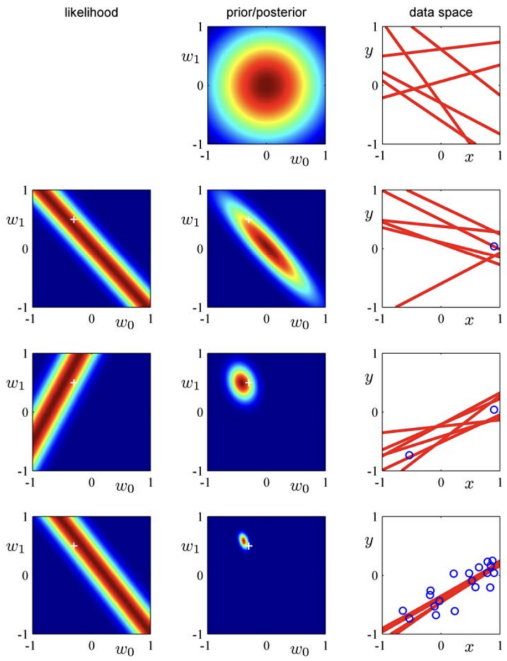


Figure 6.4: Illustration of sequential Bayesian learning for a simple linear model of the form $f(x, \mathbf{w}) = w_0 + w_1 \cdot x$ (Bishop)

6.5 Predictive distribution

Remember that the probability distribution is just the prediction of a new point given our observations.

$$p(y|x, \underline{x}, \underline{y}, \alpha, \beta)$$

In Bayesian learning we average over all possible models

$$p(y|x, \underline{x}, \underline{y}, \alpha, \beta) = \int p(y|x, w, \alpha, \beta) p(w|\underline{x}.\underline{y}, \alpha, \beta) dw$$

Since in the linear regression setting we have that both these distributions are Gaussians, we know that the product of Gaussians densities is a Gaussian, and also the marginal of a Gaussian is a Gaussian. Therefore it can be shown that the probability above is

$$p(y|x, \underline{x}, \underline{y}, \alpha, \beta) = \mathcal{N}\left(y|m_N^T \phi(x), \sigma_N^2(x)\right)$$

$$\sigma_N^2(x) = \frac{1}{\beta} + \phi^T(x) S_N \phi(x)$$

$$\sigma_N^2(x) \geq \sigma_{N+1}^2(x), \qquad \sigma_N^2 \to \frac{1}{\beta}, N \to \infty$$

The prediction is a Gaussian centered on an average prediction and having a variance which has two terms: one takes into account the

noise of observations, while the second one takes into account the epistemic uncertainty of our model. As we increase our knowledge, the epistemic uncertainty goes to zero and we are left with just the aleatoric uncertainty.

As such, we can see, graphically, that the credibility interval of our model shrinks the more we add observations.

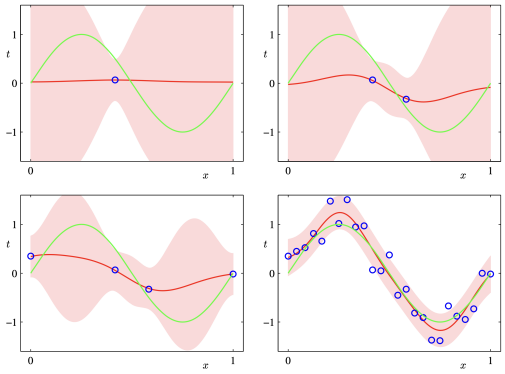
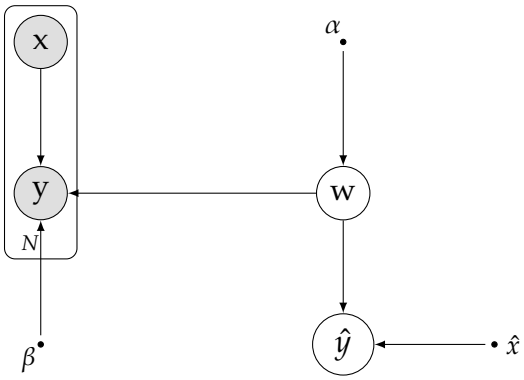


Figure 6.5: Example of the predictive distribution for a Bayesian linear regression model on the dataset shown in section 6.3.1. Increasing the number of data points, the red curve (which represents the mean of the predictive distribution) approximates better the sinusoidal function and the standard deviation (the shaded region) decreases. Note that the predictive uncertainty is smallest in the neighbourhood of the data points (Bishop)

We can represent the Bayesian linear regression model including the predictive with the following probabilistic graphical model:



Potentially we could treat this problem also as an inference in a PGM (even though this does not make much sense in practice since we have an analytical solution for the problem).

6.6 Model Evidence

How can we deal with the hyper-parameters α and β? Remember that α^{-1} represents the variance of the prior distribution whereas β^{-1} is the noise of the observations. The tool to estimate them is to use the marginal likelihood

$$p(\underline{y}|\underline{x}, \alpha, \beta) = \int p(\underline{y}|\underline{x}, w, \alpha, \beta)p(w|\alpha)dw$$

as the posterior is

$$p(w|\underline{y}, \underline{x}, \alpha, \beta) = \frac{p(\underline{y}|\underline{x}, w, \alpha, \beta)p(w|\alpha)}{p(\underline{y}|\underline{x}, \alpha, \beta)}$$

If we would treat the hyper-parameters in a Bayesian way then we would need to place a prior over them, and this would mean having a hyperprior $p(\alpha, \beta)$ and then computing the posterior distribution $p(\alpha, \beta|\underline{y}, \underline{x}) \propto p(\underline{y}|\underline{x}, \alpha, \beta)p(\alpha, \beta)$.

This is doable, but then we would need to introduce other hyper-hyper parameters and this would lead us into a hierarchy of hyper parameters.

An alternative instead is to make an approximation at this level. We need two approximations in fact:

1. First of all, we ask for an **uninformative prior** $p(\alpha, \beta)$, it can be a uniform distribution over an interval, or a Gaussian with a very broad variance. Let's suppose then that $p(\alpha, \beta) = \text{const}$
2. The posterior should be sharply peaked around the Maximum a Posteriori (MAP) of α and β. Hence $p(\alpha, \beta|\underline{y}, \underline{x}) \approx \delta_{MAP(\alpha,\beta)}$

In fact, these two approximations means that we can fix α and β with Maximum Likelihood, which means that we can find our best hyper-parameters by maximizing the Marginal Likelihood.

How to compute the marginal likelihood? Again, we rely on the closure properties of the Gaussian distribution. Since we have computed the posterior, and we already know the likelihood and the prior, we can just take the logarithm of the left and right term of Bayes' Theorem and solve the equation, else we can compute it directly (it is an integral of a Gaussian).

Therefore it can be proved that the marginal likelihood has the form

$$\log p(\underline{y}|\underline{x}, \alpha, \beta) = \frac{M}{2}\log\alpha + \frac{N}{2}\log\beta - E(m_N) - \frac{1}{2}\log|S_N^{-1}| - \frac{N}{2}\log 2\pi$$
$$E(m_N) = \frac{\beta}{2}||y - \Phi m_N||^2 + \frac{\alpha}{2}m_N^T m_n$$
$$m_N = \beta S_N \Phi \cdot \underline{y}$$
$$S_N^{-1} = \alpha I + \beta\Phi^T\Phi$$

where N is the size of the dataset and M is the number of parameters. In order to maximize this we can of course compute the gradient with respect to α and β and do gradient ascent on the expression, for example.

6.6.1 Fixed-point algorithm (*)

Here we will consider an alternative approach via a fixed-point algorithm. The general idea is to take $\nabla \log p(\underline{y}|\underline{x}, \alpha, \beta) = 0$ and to derive fix point equations

$$\alpha = g_\alpha(\alpha, \beta)$$
$$\beta = g_\beta(\alpha, \beta)$$

The algorithm works like this:

1. Fix α_0 and β_0
2. Compute

$$\alpha_{n+1} = g_\alpha(\alpha_n, \beta_n)$$
$$\beta_{n+1} = g_\beta(\alpha_n, \beta_n)$$

3. Iterate step 2 until convergence, i.e. up until

$$||\alpha_{n+1} - \alpha_n|| + ||\beta_{n+1} - \beta n|| < \epsilon$$

Therefore we just need to compute

$$\nabla \log p(\underline{y}|\underline{x}, \alpha, \beta) = \frac{M}{2}\log\alpha + \frac{N}{2}\log\beta - E(m_N) - \frac{1}{2}\log|S_N^{-1}| - \frac{N}{2}\log 2\pi$$

Let's start from the term

$$log|S_N^{-1}|$$
$$S_N^{-1} = \alpha I + \beta\Phi^T\Phi$$

In order to compute this determinant we first need to compute the eigenvalues λ_i of $\beta\Phi^T\Phi$. Then

$$|S_N^{-1}| = \prod_{i=0}^{m-1}(\alpha + \lambda_i)$$

Notice that λ_i does not depend on α. Which means that

$$\frac{\partial}{\partial\alpha}\log|S_N^{-1}| = \frac{\partial}{\partial\alpha}\sum\log(\alpha + \lambda_i) = \sum_{i=1}^{m-1}\frac{1}{\alpha + \lambda_i}$$

Moreover

$$\frac{\partial}{\partial\beta}\lambda_i = \frac{\lambda_i}{\beta}$$

In the end, by also deriving all the other terms we get

$$\alpha = \frac{\gamma}{m_N^T m_N} = g_\alpha(\alpha, \beta), \qquad \gamma = \sum_{i=0}^{m-1}\frac{\lambda_i}{\alpha + \lambda_i}$$
$$\frac{1}{\beta} = \frac{1}{N - \gamma}\sum_{n=1}^{N}\left(y_n - m_N^T\phi(x_n)\right)^2 = \frac{1}{g_\beta(\alpha, \beta)}$$

6.6.2 Effective number of parameters

Let's focus on the parameter

$$\gamma = \sum_{i=0}^{M-1}\frac{\lambda_i}{\alpha + \lambda_i}$$

where λ_i are the eigenvalues of $\beta\Phi^T\Phi$ and they give us information about the maximum Likelihood solution for w. In fact, they give us the

curvature of the likelihood function (they represent the Hessian of the likelihood). Small λ_i means a large curvature of the likelihood function which implies a large uncertainty on w_i and vice versa. When we have a large uncertainty on w_i, it means that taking the Maximum Likelihood solution of that particular weight is not very sensible. That's because introducing a prior on the parameter would likely change this value a lot (the Bayesian estimation would be different than the maximum likelihood estimation). The effective number of parameters gives us the effective number of parameters which Maximum Likelihood estimation is close to their Maximum a Posteriori estimation.

In fact, we have that

$$\lambda_i << \alpha \rightarrow \frac{\lambda_i}{\lambda_i + \alpha} \approx 0$$
$$\lambda_i >> \alpha \rightarrow \frac{\lambda_i}{\lambda_i + \alpha} \approx 1$$

and by definition of γ we have the meaning that we have been introducing before.

Notice also that in the regime of large data $N >> M$, then $\gamma \approx M$. Here the equation for α and β are also simplified.

In this scenario, the theorem of Bernstein-von Mises implies that the prior has no asymptotic influence on the posterior and that posterior inference is consistent with the frequentist approach (i.e. Maximum Likelihood estimation). Of course, there are some assumptions for this theorem to hold: the key assumption is that the "true" value of the parameter is interior to the parameters space.

Thus, the effective number of parameters in Bayesian estimation is adaptive: parameters will be "included" (in the sense that their uncertainty is low) in the model only if there is enough evidence in the data to justify their use. In a certain sense, a Bayesian model can say "I don't know" when needed. This has the effect of avoiding overfitting and giving a more correct estimation of the error when doing predictions.

6.7 Model Comparison

Imagine that we are doing linear regression and we pick two different sets of basis functions. Which of the two models should we choose?

To answer this question, we can leverage the marginal likelihood.

Suppose that we have two models $\mathbb{M}_1$ and $\mathbb{M}_2$ which are different (in the linear regression context, this means having two different sets of basis functions). Which one is the best to explain the data $D = \{\underline{x}, \underline{y}\}$?

Since we want to be Bayesian, let's place a prior distribution on the models, $p(\mathbb{M}_j)$. The posterior distribution, by Bayes' theorem, is

$$p(\mathbb{M}_j|D) = \frac{p(D|\mathbb{M}_j)p(\mathbb{M}_j)}{\sum_j p(D|\mathbb{M}_j)p(\mathbb{M}_j)}$$

Notice that $p(D|\mathbb{M}_j) = \int p_{\mathbb{M}_j}(D, \theta_{\mathbb{M}_j}|\mathbb{M}_j)d\theta_{\mathbb{M}_j}$ is the **marginal likelihood** with respect to the parameters of $\mathbb{M}_j$. In fact, since we are not looking at a specific configuration of the parameters of the model $\mathbb{M}_j$, we have to marginalize them, obtaining the marginal likelihood. We also assume that hyper-parameters are fixed in this scenario.

How to perform model selection? We have two choices

1. Model averaging (a fully Bayesian approach): instead of choosing one model we consider both of them, weighted according to the posterior distribution. The predictive distribution then is

$$p(y|x, D) = \sum_j p(y|x, D, \mathbb{M}_j) \cdot p(\mathbb{M}_j|D)$$

2. Choose the best model by computing

$$\frac{p(D|\mathbb{M}_1)}{p(D|\mathbb{M}_2)}$$

which is known as the **Bayes Factor** (of $\mathbb{M}_1$ versus $\mathbb{M}_2$). It is basically a ratio of the evidences of the two models. The model $\mathbb{M}_j$ to choose is the one with the largest Bayes factor.
Given that

$$\int p(D|\mathbb{M}_1) \log \frac{p(D|\mathbb{M}_1)}{p(D|\mathbb{M}_2)} dD > 0$$

since this is a Kullback-Leibler divergence, we observe that if $\mathbb{M}_1$ is the true model (i.e. $D \sim p(D|M_1)$), the expectation of the logarithm of the Bayes Factor $\log \frac{p(D|\mathbb{M}_1)}{p(D|\mathbb{M}_2)}$ will be positive. Hence, on average, the correct model will have the largest Bayes factor.

Example. Let $\mathbb{M}_1$ and $\mathbb{M}_2$ be two models s.t. $\mathbb{M}_1$ is nested in $\mathbb{M}_2$, i.e, the set of parameters of $\mathbb{M}_1$ is a subset of the parameters of $\mathbb{M}_2$ (for example, linear models where the set of basis functions of $\mathbb{M}_1$ is contained in the one of $\mathbb{M}_2$). This implies $\mathbb{M}_2$ is a more complex model than $\mathbb{M}_1$, and that the distribution $p(D|\mathbb{M}_2)$ is more spread than $p(D|\mathbb{M}_1)$ since the model can explain more data instances. Nevertheless, if $\mathbb{M}_1$ generated the data, then the Bayes factor will be in favor of $\mathbb{M}_1$, since $p(D|\mathbb{M}_1)$ is more concentrated on the few data instances that it can explain. Hence we can see the Occam's Razor principle emerging from the use of the Bayes factor, since the simpler model will be favored in absence of enough evidence to accept the more complex one.