Lecture Notes: Bayesian Linear Regression

Ted Meeds[1,2]

[1] Informatics Institute, University of Amsterdam
[2] The Centre for Integrative Bioinformatics, Vrije University
tmeeds@gmail.com

Abstract In this note we examine Bayesian linear regression using a Gaussian likelihood for the targets and a Gaussian prior over the weights. A Bayesian approach involves 1) computing a posterior distribution over parameters $p(\mathbf{w}|\mathcal{D})$ and *integrating* over the parameters $\mathbf{w}$ from $p(\mathbf{w}|\mathcal{D})$ when making predictions. The posterior distribution encodes the uncertainty we have about parameters after observing the data and by integrating over this uncertainty for new data, we mostly avoid overfitting problems, and results in the posterior predictive distribution. In this note we will first derive the posterior $p(\mathbf{w}|\mathcal{D})$, then derive the posterior predictive distribution $p(t_\star|\boldsymbol{\phi}_\star)$. The intuition behind the Bayesian approach is illustrated using an example of sequential inference (data arriving one at a time). Finally, Bayesian model averaging and prediction are described.

1 Motivation for a Bayesian approach

Previously we used a probabilistic view of linear regression to find the maximum likelihood and maximum a posterior estimators (MLE and MAP). Although probabilistic, these solutions are *point estimates*, that is a single parameter setting that either maximize the likelihood or posterior. A Bayesian treatment of linear regression provides a complete posterior distribution of parameters. Further, the posterior distribution is used at prediction time to integrate over the uncertainty we have for each possible parameter setting in the posterior. This is a very powerful concept because it greatly mitigates the problem of overfitting in machine learning, even when the models are complex and the number of training points is small. However, there are a limited number of models where we can *analytically* solve for the posterior, let alone the posterior predictive distribution; Gaussian-based linear regression is one of these models.

Model selection still remains for Bayesian linear regression. Although we integrate over our uncertainty in $\mathbf{w}$, we still have hyperparameters α and β, for example, and also the choice of basis functions, etc. Some of these we can treat in a Bayesian way, but for others we may use cross-validation to select them.

2 The Bayesian Set-up

A Bayesian modeling approach requires two ingredients: a **prior** distribution over parameters $p(\boldsymbol{\theta})$, a **likelihood** function (aka a conditional probability density function of data given parameters) $p(\mathcal{D}|\boldsymbol{\theta})$. Both the prior and likelihood may have additional hyperparameters $\boldsymbol{\gamma}$ (e.g. $\boldsymbol{\gamma} = \{\alpha, \beta\}$), which we can write $p(\boldsymbol{\theta}|\boldsymbol{\gamma})$ and $p(\mathcal{D}|\boldsymbol{\theta}, \boldsymbol{\gamma})$.

Based on these ingredients, a Bayesian approach then used the rules of probability—the sum and product rule—to compute the **posterior** distribution $p(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{\gamma})$. This is often non-trivial, as it requires computing the **evidence** which acts as a partition function (reciprocal of the normalizing constant) for the posterior to be properly normalized.

2.1 Some definitions

Prior distribution distribution over parameters *before seeing the data*; has hyperparameters that may be fixed, or themselves part of a Bayesian hierarchy. NB: we often use a **conjugate prior** which allows us to analytically integrate over **w** and get an analytic solution for the posterior (hint: a Gaussian prior for Gaussian likelihoods gives a Gaussian posterior).

Likelihood function probability density function of data conditioned on a single setting of parameters. We can integrate over posterior uncertainty during prediction, this is called the posterior predictive distribution.

Posterior distribution The distribution over parameters *after seeing the data*; still conditioned on fixed hyperparameters. At an abstract level, the Bayesian approach involves the following computation:

$$\text{posterior} = \left.\frac{\text{prior} \times \text{likelihood}}{\text{evidence}}\right\} \text{Bayes rule} \tag{1}$$

A general formula is:

$$p(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{\gamma}) = \left.\frac{p(\boldsymbol{\theta}|\boldsymbol{\gamma}) \times p(\mathcal{D}|\boldsymbol{\theta}, \boldsymbol{\gamma})}{p(\mathcal{D}|\boldsymbol{\gamma})}\right\} \text{Bayes rule} \tag{2}$$

We have made explicit the conditioning on hyperparameters $\boldsymbol{\gamma}$; for simplicity, below, we may drop this dependence.

Evidence aka **marginal likelihood** (some parameters have marginalized); aka **model evidence**: e.g. $p(\mathcal{D}|\text{model})$, i.e. the evidence for this particular model – can be used for model comparison/averaging. Computing the evidence requires computing the following integral:

$$p(\mathcal{D}|\boldsymbol{\gamma}) = \int p(\boldsymbol{\theta}|\boldsymbol{\gamma}) p(\mathcal{D}|\boldsymbol{\theta}, \boldsymbol{\gamma}) d\boldsymbol{\theta} \tag{3}$$

Posterior predictive distribution Keep in mind our goal is prediction, in particular prediction of real-valued targets given new, unseen test inputs vectors. Given a posterior distribution $p(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{\gamma})$, we can compute the posterior *predictive* distribution:

$$p(t_\star|\boldsymbol{\phi}_\star, \mathcal{D}, \boldsymbol{\gamma}) = \int p(t_\star|\boldsymbol{\phi}_\star, \boldsymbol{\theta}, \boldsymbol{\gamma}) p(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{\gamma}) d\boldsymbol{\theta} \tag{4}$$

Notice that we are not using Bayes rule; instead just marginalizing over $\boldsymbol{\theta}$ from the posterior. Also note we condition of $\mathcal{D}$, the training data.

3 Gaussian Linear Regression

By keeping the prior and likelihood function Gaussian, we will be able to solve analytically for the evidence, posterior, and posterior predictive distribution. In modeling situations when this is not the case, we may resort to *approximate inference* or *sampling* techniques. These approaches are tackled in the next course.

Precision or variance? Bishop and others likes to parameterize Gaussian distributions in terms of *precision* which is exactly the reciprocal of the variance, ie $\sigma^2 = 1/\beta$. Using precision just makes the derivations easier, but it is exactly equivalent to using variances instead.

3.1 A Gaussian prior over weights

For the linear regression model we assume a Gaussian prior:

$$\begin{aligned} p(\mathbf{w}|\mathbf{m_0}, \mathbf{S}_0) &= \mathcal{N}(\mathbf{w}|\mathbf{m_0}, \mathbf{S}_0) \\ &= \underbrace{\frac{|\mathbf{S}_0|^{-1/2}}{(2\pi)^{D/2}}}_{C_W} \exp\left(-\frac{1}{2}(\mathbf{w} - \mathbf{m_0})^T \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m_0})\right) \\ &= C_W \exp\left(-\frac{1}{2}(\mathbf{w} - \mathbf{m_0})^T \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m_0})\right) \end{aligned}$$

i.e. the prior is a D-dimensional Gaussian, and the density is proportional to a quadratic term (the parts that depend on $\mathbf{w}$). We can simplify the expression for the prior further by assuming the prior mean $\mathbf{m_0} = \mathbf{0}$ and $\mathbf{S}_0 = \frac{1}{\alpha}\mathbf{I}$:

$$\begin{aligned} p(\mathbf{w}|\alpha) &= \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}\right) \\ &= \underbrace{\frac{\alpha^{D/2}}{(2\pi)^{D/2}}}_{C_W} \exp\left(-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right) \end{aligned}$$

3.2 Gaussian likelihood function

For the linear regression model we assume a Gaussian likelihood:

$$\begin{aligned} p(\mathbf{t}|\boldsymbol{\Phi},\mathbf{w},\beta) &= \mathcal{N}\left(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w},1/\beta\mathbf{I}\right) \\ &= \underbrace{\frac{\beta^{N/2}}{(2\pi)^{N/2}}}_{C_L} \exp\left(-\frac{\beta}{2}\left(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}\right)^T\left(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}\right)\right) \\ & C_L \exp\left(-\frac{\beta}{2}\left(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}\right)^T\left(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}\right)\right) \end{aligned}$$

where $\beta = 1/\sigma^2$, the likelihood is an N-dimensional Gaussian, it has a **diagonal** covariance (though in principle we could use a full-covariance). Recall that this is *linear* regression because the prediction $\boldsymbol{\Phi}\mathbf{w}$ is linear in $\mathbf{w}$.

4 Derivation of Posterior via Evidence derivation

The derivation of the posterior involves a lot of linear algebra, but do not get confused along the way. Keep your eye on the derivation of the evidence. To perform the integral over weights we will rearrange the joint $p(\mathbf{w})p(\mathcal{D}|\mathbf{w})$ into an expression proportional to $p(\mathbf{w}|\mathcal{D})p(\mathcal{D})$. This makes the integral simple because we only need to know the normalizing constant for $p(\mathbf{w}|\mathcal{D})$ to normalize $p(\mathcal{D})$ (which is a constant inside the integral). We get a bonus result doing it this way, the rearrangement gives us both the posterior $p(\mathbf{w}|\mathcal{D})$ and evidence $p(\mathcal{D})$.

Keep in mind the posterior computation we are doing:

$$p\left(\mathbf{w}|\underbrace{\mathbf{t},\boldsymbol{\Phi}}_{\mathcal{D}},\underbrace{\beta,\mathbf{m_0},\mathbf{S}_0}_{\text{fixed}}\right) = \frac{p\left(\mathbf{w}|\mathbf{m_0},\mathbf{S}_0\right)p\left(\mathbf{t}|\mathbf{w},\boldsymbol{\Phi},\beta\right)}{Z}$$

$$\begin{aligned} \underbrace{Z}_{\text{partition function / evidence}} &= \left.\begin{aligned} &\int \underbrace{p\left(\mathbf{w}|\mathbf{m_0},\mathbf{S}_0\right)p\left(\mathbf{t}|\mathbf{w},\boldsymbol{\Phi},\beta\right)}_{\text{product rule}} d\mathbf{w}\ \}\ \text{sum rule} \end{aligned}\right\}\ \text{marginalization} \\ &= p\left(\mathbf{t}|\boldsymbol{\Phi},\beta,\mathbf{m_0},\mathbf{S}_0\right) \end{aligned}$$

We now turn to the derivation of the evidence, which, along the way, will give us the posterior distribution. Our first calculation will be the integral over $\mathbf{w}$. Since we have a conjugate prior (the same form as likelihood), we can do this analytically. Our strategy will be

Step 1: find the joint of $p(\mathbf{w},\mathbf{t}|\text{rest})$ and separate the $\mathbf{w}$ parts from the rest
Step 2: get Gaussian form by completing the square for the parts that depend on $\mathbf{w}$
Step 3: find posterior by inspection
Step 4: integrate over $\mathbf{w}$ to get evidence
Step 5: simplify into Gaussian form

Step 1: Compute joint and separate terms During the derivation we ignore the normalizing constants for the prior and likelihood, and instead focus on the parts that depend on $\mathbf{w}$ and $\mathbf{t}$; we can just call them $C_P = 1/Z_P$ and $C_L = 1/Z_L$ for prior and likelihood normalizing constants:

$$
\begin{aligned}
&p(\mathbf{w}|\mathbf{m_0}, \mathbf{S}_0)p(\mathbf{t}|\boldsymbol{\Phi}, \mathbf{w}, \beta) \\
&\quad= C_P C_L \exp\left(-\frac{1}{2}\left(\mathbf{w}-\mathbf{m_0}\right)^T \mathbf{S}_0^{-1}\left(\mathbf{w}-\mathbf{m_0}\right)\right)\exp\left(-\frac{\beta}{2}\left(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}\right)^T\left(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}\right)\right) \\
&\quad= C_P C_L \exp\left(-\frac{1}{2}\left[\mathbf{w}^T \underbrace{\left(\mathbf{S}_0^{-1}+\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)}_{\mathbf{S}_N^{-1}}\mathbf{w} - 2\mathbf{w}^T\underbrace{\left(\mathbf{S}_0^{-1}\mathbf{m_0}+\beta\boldsymbol{\Phi}^T\mathbf{t}\right)}_{\mathbf{S}_N^{-1}\mathbf{m}_N}\right]\right) \\
&\qquad\qquad \cdot \exp\left(-\frac{1}{2}\left[\mathbf{m_0}^T\mathbf{S}_0^{-1}\mathbf{m_0}+\beta\mathbf{t}^T\mathbf{t}\right]\right)
\end{aligned}
$$

Step 2: complete the square If we let $\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}$ and $\mathbf{S}_N^{-1}\mathbf{m}_N = \mathbf{S}_0^{-1}\mathbf{m_0} + \beta\boldsymbol{\Phi}^T\mathbf{t}$, we can 1) solve for $\mathbf{m}_N$, then 2) complete the square to get a Gaussian quadratic expression for $\mathbf{w}$. To complete the square we need to add $\mathbf{m}_N^T\mathbf{S}_N^{-1}\mathbf{m}_N$ to the exponential with $\mathbf{w}$ terms and subtract it from the other (multiplying to two exponentials the adding and subtracting cancel, so we haven't changed the value of the expression).

$$
\begin{aligned}
\mathbf{m}_N &= \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m_0}+\beta\boldsymbol{\Phi}^T\mathbf{t}\right) \\
&= \left(\mathbf{S}_0^{-1}+\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\left(\mathbf{S}_0^{-1}\mathbf{m_0}+\beta\boldsymbol{\Phi}^T\mathbf{t}\right)
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{m}_N^T\mathbf{S}_N^{-1}\mathbf{m}_N &= \left(\mathbf{S}_0^{-1}\mathbf{m_0}+\beta\boldsymbol{\Phi}^T\mathbf{t}\right)^T\mathbf{S}_N\mathbf{S}_N^{-1}\mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m_0}+\beta\boldsymbol{\Phi}^T\mathbf{t}\right) \\
&= \left(\mathbf{S}_0^{-1}\mathbf{m_0}+\beta\boldsymbol{\Phi}^T\mathbf{t}\right)^T\mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m_0}+\beta\boldsymbol{\Phi}^T\mathbf{t}\right)
\end{aligned}
$$

where $\mathbf{S}_N = \left(\mathbf{S}_0^{-1} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi}\right)^{-1}$. Subbing back in for the joint:

$$\begin{aligned}
&\overbrace{p(\mathbf{w}|\mathbf{m_0}, \mathbf{S}_0)}^{\text{prior}}\overbrace{p(\mathbf{t}|\boldsymbol{\Phi}, \mathbf{w}, \beta)}^{\text{likelihood}} \\
&\quad = C_P C_L \exp\left(-\frac{1}{2}\left[\mathbf{w}^T\mathbf{S}_N^{-1}\mathbf{w} - 2\mathbf{w}^T\mathbf{S}_N^{-1}\mathbf{m}_N + \mathbf{m}_N^T\mathbf{S}_N^{-1}\mathbf{m}_N\right]\right) \\
&\qquad\qquad \cdot \exp\left(-\frac{1}{2}\left[\mathbf{m_0}^T\mathbf{S}_0^{-1}\mathbf{m_0} + \beta\mathbf{t}^T\mathbf{t} - \mathbf{m}_N^T\mathbf{S}_N^{-1}\mathbf{m}_N\right]\right) \\
&\quad = C_P C_L \exp\left(-\frac{1}{2}\left[\left(\mathbf{w} - \mathbf{m}_N\right)^T\mathbf{S}_N^{-1}\left(\mathbf{w} - \mathbf{m}_N\right)\right]\right) \\
&\qquad\qquad \cdot \exp\left(-\frac{1}{2}\left[\mathbf{m_0}^T\mathbf{S}_0^{-1}\mathbf{m_0} + \beta\mathbf{t}^T\mathbf{t} - \mathbf{m}_N^T\mathbf{S}_N^{-1}\mathbf{m}_N\right]\right) \\
&\qquad\qquad = \underbrace{p(\mathbf{w}|\mathbf{m_0}, \mathbf{S}_0, \mathbf{t}, \boldsymbol{\Phi}, \beta)}_{\text{posterior}}\underbrace{p(\mathbf{t}|\boldsymbol{\Phi}, \mathbf{m_0}, \mathbf{S}_0, \beta)}_{\text{evidence}}
\end{aligned}$$

we now have an expression for $\mathbf{w}$ in the form of a quadratic; we can integrate over $\mathbf{w}$ and add $1/C_W$ its reciprocal normalizing constant. Next we need to rearrange terms for another quadratic for $\mathbf{t}$. But note we have simply rearrange the joint in terms of $p(\mathbf{w}|\mathcal{D})$ and $p(\mathcal{D})$, thus we have simultaneously solved for the posterior and the evidence! I.e. we started with $p(\mathbf{w})$ and $p(\mathcal{D}|\mathbf{w})$ and reversed the conditioning in the joint.

By inspection, find posterior By inspection of the form of the joint, rearranged so that only one term depends on $\mathbf{w}$, we can infer the posterior distribution:

$$\boxed{\begin{aligned}
p\left(\mathbf{w}|\mathbf{t}, \boldsymbol{\Phi}, \sigma^2, \mathbf{m_0}, \mathbf{S}_0\right) &= \mathcal{N}\left(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N\right) && \text{(5a)} \\
\mathbf{m}_N &= \left(\mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\left(\mathbf{S}_0^{-1}\mathbf{m_0} + \beta\boldsymbol{\Phi}^T\mathbf{t}\right) && \text{(5b)} \\
\mathbf{S}_N &= \left(\mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1} && \text{(5c)}
\end{aligned}}$$

furthermore is we assume $\mathbf{m_0} = \mathbf{0}$ and $\mathbf{S}_0 = \frac{1}{\alpha}\mathbf{I}$ then

$$\begin{aligned}
\mathbf{m}_N &= \left(\alpha\mathbf{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\beta\boldsymbol{\Phi}^T\mathbf{t} \\
&= \left(\beta\left(\frac{\alpha}{\beta}\mathbf{I} + \boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\right)^{-1}\beta\boldsymbol{\Phi}^T\mathbf{t} \\
&= \frac{1}{\beta}\left(\lambda\mathbf{I} + \boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\beta\boldsymbol{\Phi}^T\mathbf{t} \\
&= \left(\lambda\mathbf{I} + \boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\mathbf{t}
\end{aligned}$$

$$
\begin{aligned}
\mathbf{S}_N &= \left(\alpha \mathbf{I} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi}\right)^{-1} \\
&= \frac{1}{\beta}\left(\lambda \mathbf{I} + \boldsymbol{\Phi}^T \boldsymbol{\Phi}\right)^{-1}
\end{aligned}
$$

where $\lambda = \frac{\alpha}{\beta}$.

Compute evidence by integrating over posterior

$$
\begin{aligned}
p\left(\mathbf{t}|\boldsymbol{\Phi}, \beta, \mathbf{m_0}, \mathbf{S}_0\right) &= \int p\left(\mathbf{w}|\mathbf{m_0}, \mathbf{S}_0\right) p\left(\mathbf{t}|\mathbf{w}, \boldsymbol{\Phi}, \sigma^2\right) d\mathbf{w} \\
&= \int p(\mathbf{w}|\mathbf{m_0}, \mathbf{S}_0, \mathbf{t}, \boldsymbol{\Phi}, \beta) p(\mathbf{t}|\boldsymbol{\Phi}, \mathbf{m_0}, \mathbf{S}_0, \beta) d\mathbf{w} \\
&= C_P C_L \int \exp\left(-\frac{1}{2}\left[\left(\mathbf{w} - \mathbf{m}_N\right)^T \mathbf{S}_N^{-1}\left(\mathbf{w} - \mathbf{m}_N\right)\right]\right) \\
&\cdot \exp\left(-\frac{1}{2}\left[\mathbf{m_0}^T \mathbf{S}_0^{-1} \mathbf{m_0} + \beta \mathbf{t}^T \mathbf{t} - \mathbf{m}_N^T \mathbf{S}_N^{-1} \mathbf{m}_N\right]\right) d\mathbf{w} \\
&= C_P C_L \exp\left(-\frac{1}{2}\left[\mathbf{m_0}^T \mathbf{S}_0^{-1} \mathbf{m_0} + \beta \mathbf{t}^T \mathbf{t} - \mathbf{m}_N^T \mathbf{S}_N^{-1} \mathbf{m}_N\right]\right) \\
&\int \exp\left(-\frac{1}{2}\left[\left(\mathbf{w} - \mathbf{m}_N\right)^T \mathbf{S}_N^{-1}\left(\mathbf{w} - \mathbf{m}_N\right)\right]\right) d\mathbf{w} \\
&= \frac{C_P C_L}{C_W} \exp\left(-\frac{1}{2}\left[\mathbf{m_0}^T \mathbf{S}_0^{-1} \mathbf{m_0} + \beta \mathbf{t}^T \mathbf{t} - \mathbf{m}_N^T \mathbf{S}_N^{-1} \mathbf{m}_N\right]\right)
\end{aligned}
$$

where $1/C_W = \int \exp\left(-\frac{1}{2}\left[\left(\mathbf{w} - \mathbf{m}_N\right)^T \mathbf{S}_N^{-1}\left(\mathbf{w} - \mathbf{m}_N\right)\right]\right) d\mathbf{w}$. By inspection we can see that the exponential terms have Gaussian forms, so we do not to explicitly solve for the normalizing constants. We still need to put the evidence into a proper Gaussian form:

$$
\begin{aligned}
&\mathbf{m_0}^T \mathbf{S}_0^{-1} \mathbf{m_0} + \beta \mathbf{t}^T \mathbf{t} - \mathbf{m}_N^T \mathbf{S}_N^{-1} \mathbf{m}_N \\
&\quad = \beta \mathbf{t}^T \mathbf{t} - \left(\mathbf{S}_0^{-1} \mathbf{m_0} + \beta \boldsymbol{\Phi}^T \mathbf{t}\right)^T \mathbf{S}_N \left(\mathbf{S}_0^{-1} \mathbf{m_0} + \beta \boldsymbol{\Phi}^T \mathbf{t}\right) + \mathbf{m_0}^T \mathbf{S}_0^{-1} \mathbf{m_0} \\
&\quad = \mathbf{t}^T \left(\underbrace{\beta \mathbf{I} - \beta^2 \boldsymbol{\Phi} \mathbf{S}_N \boldsymbol{\Phi}^T}_{\boldsymbol{\Sigma}_N^{-1}}\right) \mathbf{t} - 2\mathbf{t}^T \left(\underbrace{\beta \boldsymbol{\Phi} \mathbf{S}_N \mathbf{S}_0^{-1} \mathbf{m_0}}_{\boldsymbol{\Sigma}_N^{-1} \boldsymbol{\mu}_N}\right) + \underbrace{\mathbf{m_0}^T \mathbf{S}_0^{-1} \mathbf{m_0} - \mathbf{m_0}^T \mathbf{S}_0^{-1} \mathbf{S}_N \mathbf{S}_0^{-1} \mathbf{m_0}}_{\boldsymbol{\mu}_N^T \boldsymbol{\Sigma}_N^{-1} \boldsymbol{\mu}_N}
\end{aligned}
$$

To solve for $\boldsymbol{\Sigma}_N^{-1}$ we use the **Woodbury identity** (found in Matrix cookbook):

$$
\left(\mathbf{A} + \mathbf{C}\mathbf{B}\mathbf{C}^T\right)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{C}\left(\mathbf{B}^{-1} + \mathbf{C}^T \mathbf{A}^{-1} \mathbf{C}\right)^{-1} \mathbf{C}^T \mathbf{A}^{-1} \qquad (6)
$$

Expanding the current form of $\boldsymbol{\Sigma}_N^{-1}$ we can map it to the Woodbury identity and solve directly. Note we do not need Woodbury to solve it, we can do it may matrix manipulation alone.

$$\begin{aligned}
\boldsymbol{\Sigma}_N^{-1} &= \beta\mathbf{I} - \beta^2\boldsymbol{\Phi}\mathbf{S}_N\boldsymbol{\Phi}^T \\
&= \beta\mathbf{I} - \beta\mathbf{I}\boldsymbol{\Phi}\left(\mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\beta\mathbf{I} \\
&= \beta\mathbf{I} - \beta\mathbf{I}\boldsymbol{\Phi}\left(\mathbf{S}_0^{-1} + \boldsymbol{\Phi}^T(\beta\mathbf{I})\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\beta\mathbf{I} \\
&= \left(\frac{1}{\beta}\mathbf{I} + \boldsymbol{\Phi}\mathbf{S}_0\boldsymbol{\Phi}^T\right)^{-1}
\end{aligned}$$

Solving for $\boldsymbol{\mu}_N$ is just an exercise in matrix manipulation:

$$\begin{aligned}
\boldsymbol{\mu}_N &= \boldsymbol{\Sigma}_N\beta\boldsymbol{\Phi}\left(\mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\mathbf{S}_0^{-1}\mathbf{m_0} \\
&= \left(\frac{1}{\beta}\mathbf{I} + \boldsymbol{\Phi}\mathbf{S}_0\boldsymbol{\Phi}^T\right)\beta\boldsymbol{\Phi}\left(\mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\mathbf{S}_0^{-1}\mathbf{m_0} \\
&= \boldsymbol{\Phi}\left[\left(\mathbf{I} + \beta\mathbf{S}_0\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\left(\mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\mathbf{S}_0^{-1}\right]\mathbf{m_0} \\
&= \boldsymbol{\Phi}\left[\mathbf{S}_0\mathbf{S}_0^{-1}\left(\mathbf{I} + \beta\mathbf{S}_0\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\left(\mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\mathbf{S}_0^{-1}\right]\mathbf{m_0} \\
&= \boldsymbol{\Phi}\left[\mathbf{S}_0\left(\mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\left(\mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\mathbf{S}_0^{-1}\right]\mathbf{m_0} \\
&= \boldsymbol{\Phi}\left[\mathbf{S}_0\mathbf{S}_0^{-1}\right]\mathbf{m_0} \\
&= \boldsymbol{\Phi}\mathbf{m_0}
\end{aligned}$$

Therefore:

$$\begin{aligned}
p\left(\mathbf{t}|\boldsymbol{\Phi}, \beta, \mathbf{m_0}, \mathbf{S}_0\right) &= \mathcal{N}\left(\mathbf{t}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N\right) && (7) \\
\boldsymbol{\mu}_N &= \boldsymbol{\Phi}\mathbf{m_0} && (8) \\
\boldsymbol{\Sigma}_N &= \frac{1}{\beta}\mathbf{I} + \boldsymbol{\Phi}\mathbf{S}_0\boldsymbol{\Phi}^T && (9)
\end{aligned}$$

furthermore is we assume $\mathbf{m_0} = \mathbf{0}$ and $\mathbf{S}_0 = \frac{1}{\alpha}\mathbf{I}$, then

$$\boxed{p\left(\mathbf{t}|\boldsymbol{\Phi}, \beta, \mathbf{m_0}, \mathbf{S}_0\right) = \mathcal{N}\left(\mathbf{t}|\mathbf{0}, \frac{1}{\beta}\mathbf{I} + \frac{1}{\alpha}\boldsymbol{\Phi}\boldsymbol{\Phi}^T\right)} \quad (10a)$$

5 Posterior Predictive Distribution

Let's recap what we have done so far for the Bayesian linear regression model. Using a (conjugate) Gaussian prior over $\mathbf{w}$ we have solved for the exact analytic posterior distribution for $\mathbf{w}$. We have also solved for the model evidence, which is also Gaussian. The model evidence could be used for model comparison, to for example selection different hyperparameters α and β. We now turn to the task of **prediction** using a Bayesian model.

For Bayesian models, the prediction for a new test point $\mathbf{x}$ should be based on the posterior distribution inferred from the training set $\mathcal{D}_{train}$. The likelihood function is the same as before, except now the likelihood becomes a predictive distribution, conditioned on $\mathbf{w}$, but the prior is now the posterior. Ideally we want to integrate over the uncertainty in the posterior when making predictions; we will see that this is possible with the Gaussian distributions we are using.

In general, we have

$$\underbrace{p(t_\star|\boldsymbol{\phi}_\star, \mathcal{D}_{\text{train}})}_{\text{predictive distribution}} = \underbrace{\int p(t_\star|\boldsymbol{\phi}_\star, \mathbf{w}) p(\mathbf{w}|\mathcal{D}_{train}) d\mathbf{w}}_{\text{integrate over posterior}}$$

For some models, like the Gaussian linear regression model, we can compute the integral analytically (and we will show two ways below) and use is directly to predict the target given any test vector. If we cannot compute the integral, we can use a Monte Carlo approximate to the integral by averaging over $\mathbf{w}^{(s)}$ where $\mathbf{w}^{(s)} \sim p(\mathbf{w}|\mathcal{D}_{train})$ (are draws from the posterior):

$$p(t_\star|\boldsymbol{\phi}_\star, \mathcal{D}_{train}) \approx \frac{1}{S}\sum_{s=1}^{S} p(t_\star|\boldsymbol{\phi}_\star, \mathbf{w}^{(s)}) \qquad \mathbf{w}^{(s)} \sim p(\mathbf{w}|\mathcal{D}_{train})$$

This is a good way of visualizing what the model is *thinking* by generating plausible weight vectors from the posterior. See for example Bishop Figure 3.9.

5.1 Solving the integral directly

The first approach rearranges the terms involving $\mathbf{w}$ from the others. We did this for solving the posterior distribution. The derivation is left as an exercise (HW2); the final result (for Gaussians) is the following:

$$\begin{aligned} p(t_\star|\boldsymbol{\phi}_\star, \mathcal{D}_{\text{train}}) &= \int p(t_\star|\boldsymbol{\phi}_\star, \mathbf{w}) p(\mathbf{w}|\mathcal{D}_{train}) d\mathbf{w} \\ &= \int \mathcal{N}\left(t_\star|\boldsymbol{\phi}_\star^T\mathbf{w}, 1/\beta\right) \mathcal{N}\left(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N\right) d\mathbf{w} \\ &= \mathcal{N}\left(t_\star|\boldsymbol{\phi}_\star^T\mathbf{m}_N, 1/\beta + \boldsymbol{\phi}_\star^T\mathbf{S}_N\boldsymbol{\phi}_\star\right) \end{aligned}$$

Solving the integral using evidence rations The second approach is to apply Bayes rule to the evidence distributions (that already have $\mathbf{w}$ integrated out). Consider Bayes rule for set of vectors:

$$p(\mathbf{x}_{N+1}|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) = \frac{p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1})}{p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)} \tag{11}$$

where we have implicitly integrated over $\mathbf{x}_{N+1}$ in the denominator.

Now consider the marginal likelihoods (evidences) for the training set and the training set *augmented* by the test vector:

$$p\left(t_\star|\mathbf{t},\boldsymbol{\Phi},\boldsymbol{\phi}_\star,\beta,\mathbf{m_0},\mathbf{S}_0\right)=\frac{p\left(t_\star,\mathbf{t}|\boldsymbol{\phi}_\star,\boldsymbol{\Phi},\beta,\mathbf{m_0},\mathbf{S}_0\right)}{p\left(\mathbf{t}|\boldsymbol{\Phi},\beta,\mathbf{m_0},\mathbf{S}_0\right)} \tag{12}$$

Remarks

1. The posterior predictive mean uses the posterior mean for $\mathbf{w}$.
2. The posterior predictive variance is a function of the input $\boldsymbol{\phi}$ (it was constant for maximum likelihood estimates of the noise).
3. The posterior predictive can be induced as a linear function of the posterior of $\mathbf{w}$ plus Gaussian noise $1/\beta$.
4. The variance decomposes into two parts: $1/\beta$ – the "data" uncertainty and $\boldsymbol{\phi}_\star^T\mathbf{S}_N\boldsymbol{\phi}_\star$ – the "model" uncertainty.
5. When $\boldsymbol{\phi}_\star$ is "close" to training data, then the model uncertainty is small (and vice versa) (for local basis functions).

5.2 Summary of Bayesian Linear Regression

posterior	model evidence	predictive
$\mathcal{N}\left(\mathbf{w}\|\mathbf{m}_N,\mathbf{S}_N\right)$	$\mathcal{N}\left(\mathbf{t}\|\boldsymbol{\mu}_N,\boldsymbol{\Sigma}_N\right)$	$\mathcal{N}\left(t_\star\|\boldsymbol{\phi}^T\mathbf{m}_N,1/\beta+\boldsymbol{\phi}_\star^T\mathbf{S}_N\boldsymbol{\phi}_\star\right)$
$\mathbf{m}_N=\left(\lambda\mathbf{I}+\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\mathbf{t}$	$\boldsymbol{\mu}_N=\mathbf{0}$	
$\mathbf{S}_N=\frac{1}{\beta}\left(\lambda\mathbf{I}+\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}$	$\boldsymbol{\Sigma}_N=\frac{1}{\beta}\mathbf{I}+\frac{1}{\alpha}\boldsymbol{\Phi}\boldsymbol{\Phi}^T$	

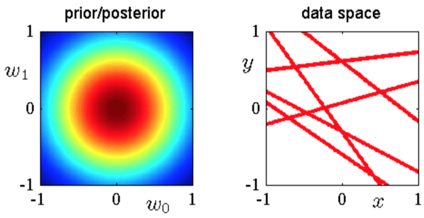
6 Understanding Bayesian Linear Regression

The derivations for the posterior and predictive distributions were a complicated exercise in linear algebra and application of the sum and product rules. To understand what is happening, it is useful to think of the Bayesian approach as a sequential learning problem. Recall that we said the prior describes our belief about $\mathbf{w}$ before seeing any data and the posterior our belief after seeing all the data. What makes our belief change?

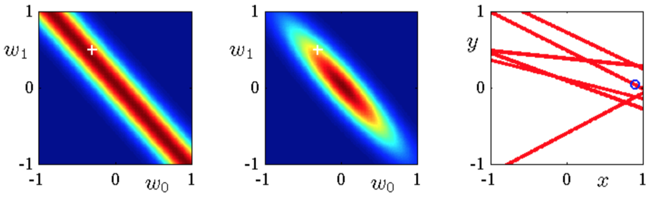
6.1 Example using Linear Basis Functions

Let's start with the example from Bishop Figure 3.7, where $y(x,\mathbf{w})=w_0+w_1x$ (ie the first dimension of $\mathbf{w}$ is the bias and the second is the slope). Consider data arriving sequentially and our task is to update the posterior and predictive distributions:

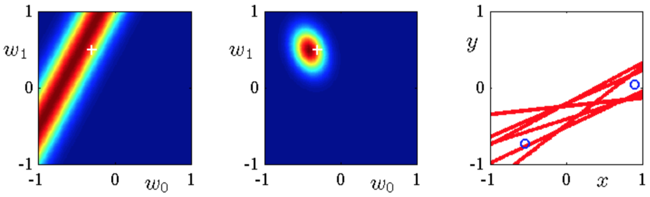
1. Before the arrival of the first data vector, our prior belief about $\mathbf{w}$ is a symmetric Gaussian, ie our posterior = prior, $p(\mathbf{w}|\mathcal{D})=p(\mathbf{w})$. Samples from the prior/posterior have biases that vary between -1 and 1, and slopes that vary between -1 and 1.



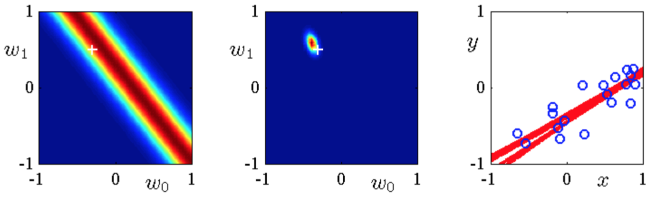
2. After the arrival of the first point (blue circle, right), the posterior is *proportional* to the product of the prior (a symmetric Gaussian) and one Gaussian likelihood term. On the left, notice how the likelihood for the data point cuts a region in weight space; this is adding a kind of soft constraint to the posterior. In the middle image we see the posterior is now the normalized product of the likelihood constraint and the prior. The solutions from the posterior are not much more restricted. For example, at one extreme a bias of -1 is ok, but it requires a slope of 1. The posterior is $p(\mathbf{w}|\mathcal{D}) \propto p(\mathbf{w})p(t_1|w_0 + x_1w_1)$.



3. With an additional data point comes an additional soft constraint (only the new one is shown left). This new constraint has a large effect of the posterior, as the two likelihood constraints intersect around where the posterior (middle) is shown. Now only negative biases and positive slopes have mass in the posterior. The posterior is $p(\mathbf{w}|\mathcal{D}) \propto p(\mathbf{w})p(t_1|w_0 + x_1w_1)p(t_2|w_0 + x_2w_1)$.



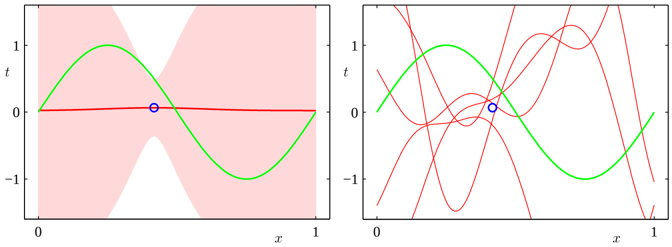
4. After the arrival of all the data points, the posterior is very small (though there is still some uncertainty). This is reflected in the low variability in the posterior samples and in turn by the predictions they make. The posterior is $p(\mathbf{w}|\mathcal{D}) \propto p(\mathbf{w}) \prod_{n=1}^{N=20} p(t_n|w_0 + x_nw_1)$.



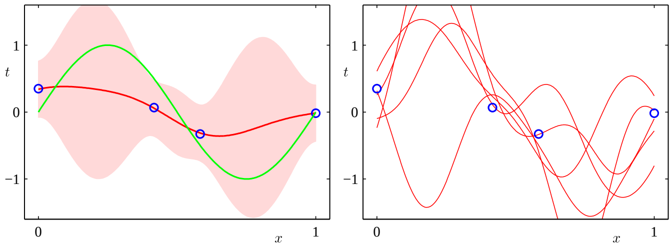
6.2 Example using Gaussian Basis Functions

The effect of adding data becomes more interesting if we change to local basis functions, in the example Gaussian basis functions in Bishop Figures 3.8-3.9.

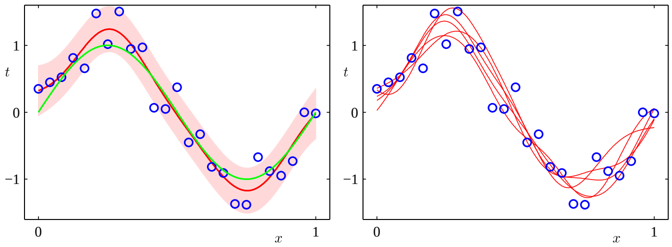
1. After observing a single data point, the posterior predictive distribution (left: expectation: red line, variance: shaded region) predicts 0, on average, since the prior is centered at zero, but has huge variance. Notice that the variance is smaller close to the observation, and grows away from it. We can gain an understanding of what is going on by looking at samples from the predictive distribution (by first sampling **w** from the posterior, then computing the prediction using these **w**). The likelihood constraint restricts weights so that the solutions pass close to the observations. Away from the observation, there is no such constraint, and therefore the variance increases again.



2. With three more data points, the predictive mean can still pass through all the data points, but the variance between points increases, accounting for the lack of observations in these regions. The predictive samples are still very wiggly between data, but all of the pass close to all the data points.



3. Finally, after observing all the data, the predictive distribution closely follows the true function, even though it is a very flexible model. At this point there is very little "model" uncertainty; instead the data uncertainty account for most of the uncertainty $(1/\beta)$



7 Bayesian Model Comparison

Recall: polynomial basis function regression implicitly made model choice $M = 0, 1, ...?$. We used cross-validation for selecting M and λ. Bayesian approaches integrates over $\mathbf{w}$, M avoiding cross-validation.

Recall: the model evidence for the linear regression model:

$$p\left(\underbrace{\mathbf{t}}_{\mathcal{D}} | \underbrace{\boldsymbol{\Phi}, \sigma^2, \mathbf{m_0}, \mathbf{S}_0}_{\text{model}\mathcal{M}}\right) = \int p\left(\mathbf{w}|\mathbf{m_0}, \mathbf{S}_0\right) p\left(\mathbf{t}|\mathbf{w}, \boldsymbol{\Phi}, \sigma^2\right) d\mathbf{w}$$

In general what we did was integrate over the model parameters:

$$p(\mathcal{D}|\mathcal{M}) = \int p(\boldsymbol{\theta}|\mathcal{M})p(\mathcal{D}|\boldsymbol{\theta})d\boldsymbol{\theta}$$

Thus the model evidence provides the likelihood of a model $\mathcal{M}$, so we can 1) compare models and 2) integrate over models.

Model Posterior Requires a prior over models $p(\mathcal{M}_i)$, and a model likelihood $p(\mathcal{D}|\mathcal{M}_i)$; the rest is just an application of Bayes rule:

$$p(\mathcal{M}_i|\mathcal{D}_i) = p(\mathcal{D}|\mathcal{M}_i)p(\mathcal{M}_i)/Z$$

Model Averaging
Average the posterior predictive distributions of different models:

$$p(t_\star|\mathbf{x}_\star, \mathcal{D}) = \sum_{i=1}^{L} p(t_\star|\mathbf{x}_\star, \mathcal{D}, \mathcal{M}_i)p(\mathcal{M}_i|\mathcal{D}_i)$$

Bayes Factors
Using marginal likelihood (aka model evidence). I.e. the ratio of marginal likelihoods:

$$p(\mathcal{D}|\mathcal{M}_i)/p(\mathcal{D}|\mathcal{M}_j)$$

Model Selection By maximizing the evidence:

$$\mathcal{M} = \arg\max_{\mathcal{M}_i} p(\mathcal{D}|\mathcal{M}_i)$$

8 Limitations of Fixed Basis Functions

Up to this point we have consider fixed basis functions, that is, transformations of raw input vectors $\mathbf{x}$ into a vector of feature $\boldsymbol{\phi}(\mathbf{x})$. Each dimension of $\boldsymbol{\phi}$ is the result of applying a single basis function to the input vector. Predictions are linear weightings of the feature functions. The form of the functions–their parameters–are fixed and pre-determined before learning. This has a couple advantages, in particular closed for solutions for **w** (in some cases). However, unless the dimension of the raw input space is small, it is rarely the case that we use fixed basis functions in practice. Not only does is scale poorly in dimension (ie the curse of dimensionality)—we must "cover" the input region with basis functions, but it also makes more sense to learn the locations or other parameters of the basis functions. This is the idea behind representation learning, and forms backbone of many more sophisticated models, such as neural networks and Gaussian processes.

The solution for these models is, for neural networks, is to use local basis functions (Gaussians, sigmoids) and *move* them to locations near data; the weights of the neural network do this (except for the last layer). In other models, the training data are used as templates for basis functions (Gaussian processes and Support Vector Machines).

Keep in mind, as we move to more complex models, that the much of the sophistication of complex models comes from representation learning. The "top" of the predictor remains the same for regression—a real-valued prediction that is minimized by sum-squared error, or by maximizing log-likelihood.