UNIVERSITY
of
GLASGOW

Machine Learning

Lecture. 6.

Mark Girolami

girolami@dcs.gla.ac.uk

Department of Computing Science
University of Glasgow

Bayesian Regression

UNIVERSITY of GLASGOW

- Likelihood methods interested in how likely the data is given our model and associated parameters $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma)$

Bayesian Regression

UNIVERSITY
of
GLASGOW

- Likelihood methods interested in how likely the data is given our model and associated parameters $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma)$
- But we really want to know about the model parameters given the data then the quantity we should concerned with is $p(\mathbf{w}, \sigma|\mathbf{X}, \mathbf{t})$.

Bayesian Regression

UNIVERSITY of GLASGOW

- Likelihood methods interested in how likely the data is given our model and associated parameters $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma)$
- But we really want to know about the model parameters given the data then the quantity we should concerned with is $p(\mathbf{w}, \sigma|\mathbf{X}, \mathbf{t})$.
- From our likelihood $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma)$ we can obtain $p(\mathbf{w}, \sigma|\mathbf{X}, \mathbf{t})$ via Bayes rule

Bayesian Regression

UNIVERSITY of GLASGOW

- Likelihood methods interested in how likely the data is given our model and associated parameters $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma)$
- But we really want to know about the model parameters given the data then the quantity we should concerned with is $p(\mathbf{w}, \sigma|\mathbf{X}, \mathbf{t})$.
- From our likelihood $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma)$ we can obtain $p(\mathbf{w}, \sigma|\mathbf{X}, \mathbf{t})$ via Bayes rule
- Let's then look at our linear regression model within the Bayesian formalism

Bayesian Regression

UNIVERSITY
of
GLASGOW

- For simplicity assume that we know the noise level σ

Bayesian Regression

UNIVERSITY *of* GLASGOW

- For simplicity assume that we know the noise level σ
- We know the value of σ and the input data $\mathbf{X}$ is given, so no uncertainty in these, we only reason about the target values $\mathbf{t}$ and the parameters $\mathbf{w}$ so the joint probability of everything associated with our model can be written as...

Bayesian Regression

- For simplicity assume that we know the noise level σ
- We know the value of σ and the input data $\mathbf{X}$ is given, so no uncertainty in these, we only reason about the target values $\mathbf{t}$ and the parameters $\mathbf{w}$ so the joint probability of everything associated with our model can be written as...

$$p(\mathbf{t}, \mathbf{w}|\mathbf{X}, \sigma) = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma)p(\mathbf{w}) = p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma)p(\mathbf{t}|\mathbf{X}, \sigma)$$

Bayesian Regression

UNIVERSITY *of* GLASGOW

- For simplicity assume that we know the noise level σ
- We know the value of σ and the input data $\mathbf{X}$ is given, so no uncertainty in these, we only reason about the target values $\mathbf{t}$ and the parameters $\mathbf{w}$ so the joint probability of everything associated with our model can be written as...

$$p(\mathbf{t}, \mathbf{w}|\mathbf{X}, \sigma) = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma)p(\mathbf{w}) = p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma)p(\mathbf{t}|\mathbf{X}, \sigma)$$

- Using the expressions above using Bayes theorem we can invert our probabilities to obtain

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma) = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma)\frac{p(\mathbf{w})}{p(\mathbf{t}|\mathbf{X}, \sigma)}$$

Bayesian Regression

UNIVERSITY of GLASGOW

- The probability $p(\mathbf{w})$ is the probability distribution of the parameters prior to observing any data - refered to as the prior

Bayesian Regression

UNIVERSITY of GLASGOW

- The probability $p(\mathbf{w})$ is the probability distribution of the parameters prior to observing any data - refered to as the prior
- The conditional probability $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma)$ we have met previously and is the data likelihood

Bayesian Regression

UNIVERSITY *of* GLASGOW

- The probability $p(\mathbf{w})$ is the probability distribution of the parameters prior to observing any data - refered to as the prior
- The conditional probability $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma)$ we have met previously and is the data likelihood
- The distribution $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma)$ is obtained after observing data (post observation) and is called the posterior distribution

Bayesian Regression

UNIVERSITY of GLASGOW

- The probability $p(\mathbf{w})$ is the probability distribution of the parameters prior to observing any data - refered to as the prior
- The conditional probability $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma)$ we have met previously and is the data likelihood
- The distribution $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma)$ is obtained after observing data (post observation) and is called the posterior distribution
- Note for $p(\mathbf{w})$ there is no conditioning on $\mathbf{X}$ or σ as we set the prior before seeing any data

Bayesian Regression

UNIVERSITY *of* GLASGOW

- The probability $p(\mathbf{w})$ is the probability distribution of the parameters prior to observing any data - refered to as the prior
- The conditional probability $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma)$ we have met previously and is the data likelihood
- The distribution $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma)$ is obtained after observing data (post observation) and is called the posterior distribution
- Note for $p(\mathbf{w})$ there is no conditioning on $\mathbf{X}$ or σ as we set the prior before seeing any data
- The term $p(\mathbf{t}|\mathbf{X}, \sigma)$ which is called the marginal likelihood as $p(\mathbf{t}|\mathbf{X}, \sigma) = \int p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma)p(\mathbf{w})d\mathbf{w}$ where we integrate out or *marginalise* the model parameters

Bayesian Regression

UNIVERSITY *of* GLASGOW

- So our posterior distribution for the parameters can be seen as the prior belief being updated after we observe our data so in other words.

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$$

Bayesian Regression

UNIVERSITY of GLASGOW

- So our posterior distribution for the parameters can be seen as the prior belief being updated after we observe our data so in other words.

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$$

- The form of the likelihood has been previously defined now we have to consider the form of the prior distribution over the parameter values

Setting the Prior

- We are perfectly free to make whatever assumptions are most appropriate at this point

Setting the Prior

UNIVERSITY of GLASGOW

- We are perfectly free to make whatever assumptions are most appropriate at this point
- These assumptions can be encoded in a prior distribution

Setting the Prior

- We are perfectly free to make whatever assumptions are most appropriate at this point
- These assumptions can be encoded in a prior distribution
- Let's say that before seeing any data we would prefer some parameter values to be small, this is a sensible strategy especially when there are many possibly redundant parameter values.

Setting the Prior

- We are perfectly free to make whatever assumptions are most appropriate at this point
- These assumptions can be encoded in a prior distribution
- Let's say that before seeing any data we would prefer some parameter values to be small, this is a sensible strategy especially when there are many possibly redundant parameter values.
- So assume that all our parameter values will follow a Gaussian distribution with a mean of zero and a standard deviation of α.

Setting the Prior

UNIVERSITY
of
GLASGOW

- We also assume that the parameters are *a priori* independent of each other so $w_0 \sim \mathcal{N}(0, \alpha)$ and likewise $w_1 \sim \mathcal{N}(0, \alpha)$.

Setting the Prior

UNIVERSITY *of* GLASGOW

- We also assume that the parameters are *a priori* independent of each other so $w_0 \sim \mathcal{N}(0, \alpha)$ and likewise $w_1 \sim \mathcal{N}(0, \alpha)$.
- So $p(\mathbf{w}|\alpha) = \mathcal{N}_{w_0}(0, \alpha)\mathcal{N}_{w_1}(0, \alpha) = \mathcal{N}_{\mathbf{w}}(\mathbf{0}, \mathbf{\Lambda})$ where $\mathbf{\Lambda} = \alpha\mathbf{I}$

Setting the Prior

UNIVERSITY *of* GLASGOW

- We also assume that the parameters are *a priori* independent of each other so $w_0 \sim \mathcal{N}(0, \alpha)$ and likewise $w_1 \sim \mathcal{N}(0, \alpha)$.
- So $p(\mathbf{w}|\alpha) = \mathcal{N}_{w_0}(0, \alpha)\mathcal{N}_{w_1}(0, \alpha) = \mathcal{N}_{\mathbf{w}}(\mathbf{0}, \mathbf{\Lambda})$ where $\mathbf{\Lambda} = \alpha\mathbf{I}$
- The most probable values that the parameters will take on are small centered around zero, once observations are made this prior will be updated in the light of the data : prior $\longrightarrow$ posterior via data likelihood

Setting the Prior

UNIVERSITY *of* GLASGOW

- We also assume that the parameters are *a priori* independent of each other so $w_0 \sim \mathcal{N}(0, \alpha)$ and likewise $w_1 \sim \mathcal{N}(0, \alpha)$.
- So $p(\mathbf{w}|\alpha) = \mathcal{N}_{w_0}(0, \alpha)\mathcal{N}_{w_1}(0, \alpha) = \mathcal{N}_{\mathbf{w}}(\mathbf{0}, \mathbf{\Lambda})$ where $\mathbf{\Lambda} = \alpha\mathbf{I}$
- The most probable values that the parameters will take on are small centered around zero, once observations are made this prior will be updated in the light of the data : prior $\longrightarrow$ posterior via data likelihood

To the Posterior

- The likelihood is an N-dimensional multivariate Gaussian $\prod_{n=1}^{N} \mathcal{N}_{t_n}(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n, \sigma) = \mathcal{N}_{\mathbf{t}}(\mathbf{X}\mathbf{w}, \sigma\mathbf{I})$ and so we can write the posterior as

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma, \alpha) = \frac{\mathcal{N}_{\mathbf{t}}(\mathbf{X}\mathbf{w}, \sigma\mathbf{I})\mathcal{N}_{\mathbf{w}}(\mathbf{0}, \mathbf{\Lambda})}{\int \mathcal{N}_{\mathbf{t}}(\mathbf{X}\mathbf{w}, \sigma\mathbf{I})\mathcal{N}_{\mathbf{w}}(\mathbf{0}, \mathbf{\Lambda}) d\mathbf{w}}$$

To the Posterior

UNIVERSITY *of* GLASGOW

- The likelihood is an N-dimensional multivariate Gaussian $\prod_{n=1}^{N} \mathcal{N}_{t_n}(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n, \sigma) = \mathcal{N}_{\mathbf{t}}(\mathbf{X}\mathbf{w}, \sigma\mathbf{I})$ and so we can write the posterior as

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma, \alpha) = \frac{\mathcal{N}_{\mathbf{t}}(\mathbf{X}\mathbf{w}, \sigma\mathbf{I})\mathcal{N}_{\mathbf{w}}(\mathbf{0}, \mathbf{\Lambda})}{\int \mathcal{N}_{\mathbf{t}}(\mathbf{X}\mathbf{w}, \sigma\mathbf{I})\mathcal{N}_{\mathbf{w}}(\mathbf{0}, \mathbf{\Lambda}) d\mathbf{w}}$$

- Miraculously the posterior is also a Normal distribution

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma, \alpha) = \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{\Sigma}\right)$$

where

$$\boldsymbol{\mu} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \frac{\sigma^2}{\alpha}\mathbf{I}\right)^{-1} \mathbf{X}^{\mathsf{T}}\mathbf{t} \text{ and } \mathbf{\Sigma} = \sigma^2\left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \frac{\sigma^2}{\alpha}\mathbf{I}\right)^{-1}$$

To the Posterior

UNIVERSITY *of* GLASGOW

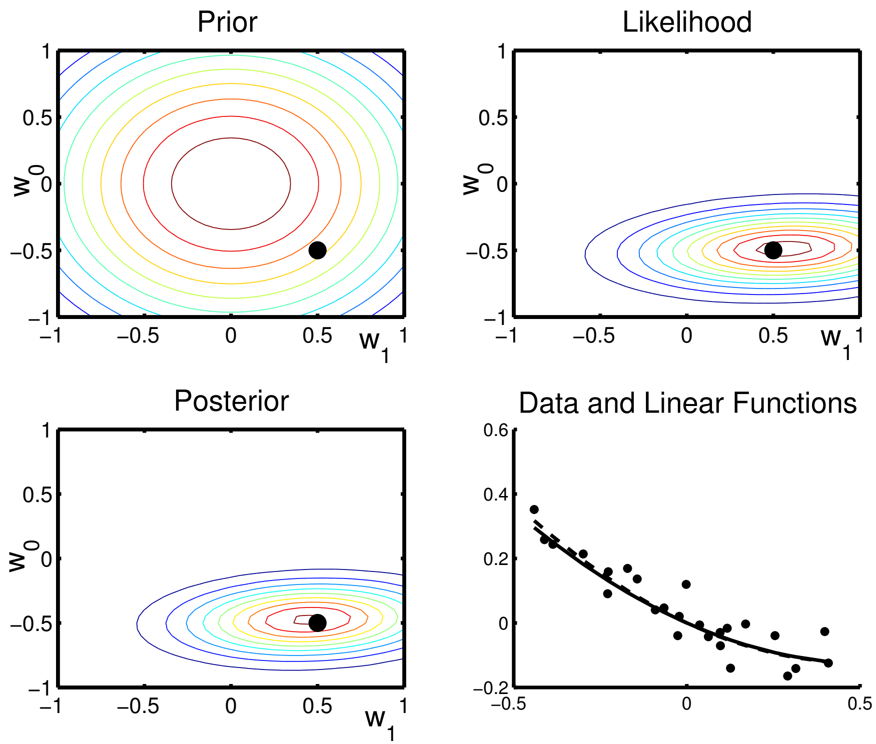

Figure 1: Top Left shows the prior distribution with the black-spot highlighting the *true* parameter values. The top right plot shows the likelihood and we can see that it is concentrated around the true values. The bottom left shows the corresponding posterior and finally the bottom right shows the data the true function and the estimated one when σ is known and α, the prior variance, is set to unity.

To the Posterior

UNIVERSITY of GLASGOW

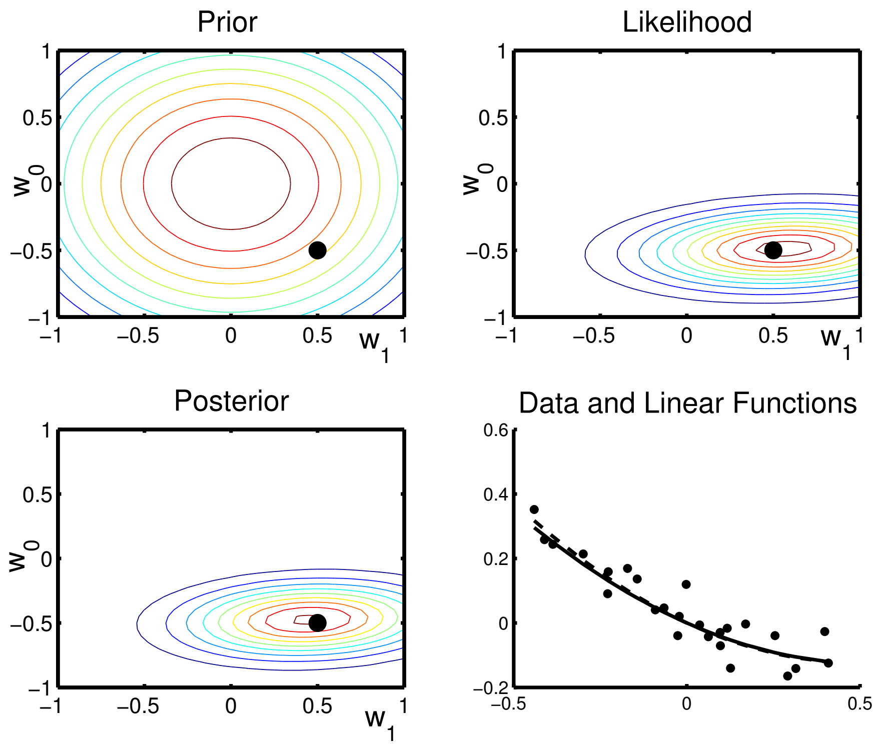

Figure 2: Top Left shows the prior distribution with the black-spot highlighting the *true* parameter values. The top right plot shows the likelihood and we can see that it is concentrated around the true values. The bottom left shows the corresponding posterior and finally the bottom right shows the data the true function and the estimated one when σ is known and α, the prior variance, is set to unity.

To the Posterior

UNIVERSITY *of* GLASGOW

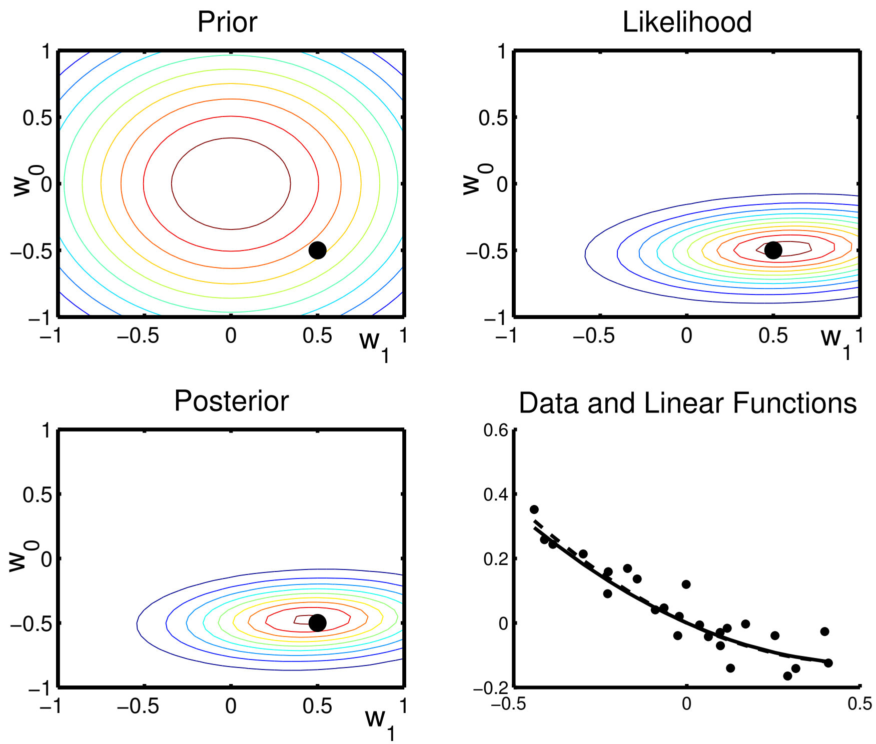

Figure 3: Top Left shows the prior distribution with the black-spot highlighting the *true* parameter values. The top right plot shows the likelihood and we can see that it is concentrated around the true values. The bottom left shows the corresponding posterior and finally the bottom right shows the data the true function and the estimated one when σ is known and α, the prior variance, is set to unity.

To the Posterior

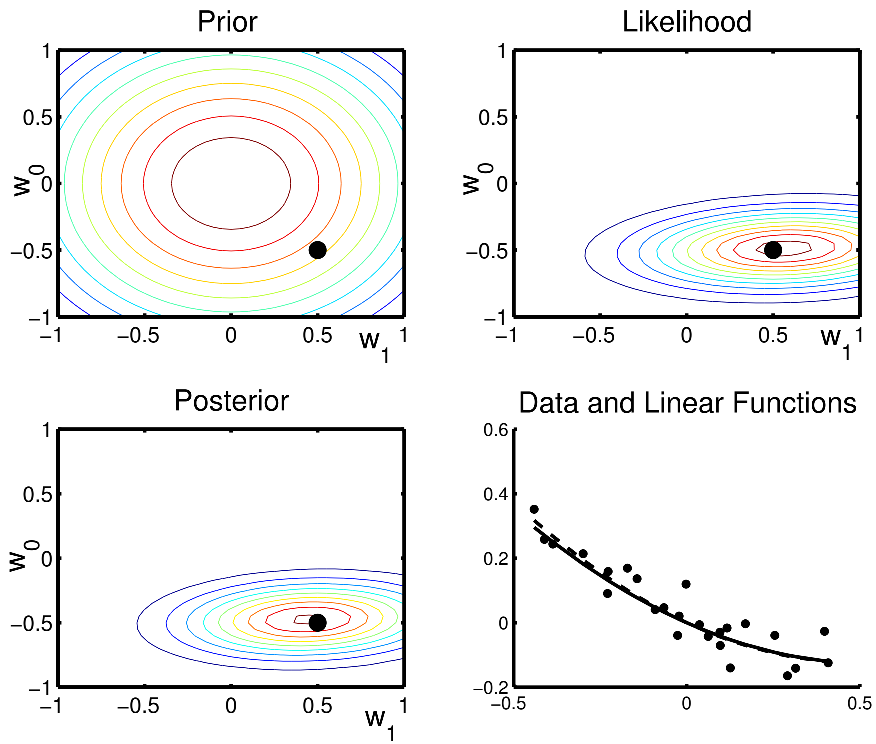

Figure 4: Top Left shows the prior distribution with the black-spot highlighting the *true* parameter values. The top right plot shows the likelihood and we can see that it is concentrated around the true values. The bottom left shows the corresponding posterior and finally the bottom right shows the data the true function and the estimated one when σ is known and α, the prior variance, is set to unity.

Posterior Inference

- Maximum Likelihood framework the MLE is plugged in to obtain predicted target values for a new data point

Posterior Inference

UNIVERSITY
of
GLASGOW

- Maximum Likelihood framework the MLE is plugged in to obtain predicted target values for a new data point
- Bayesian framework we can use our posterior distribution to average (or integrate) over our uncertainty in the possible parameter values

Posterior Inference

UNIVERSITY of GLASGOW

$$
\begin{aligned}
E_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma,\alpha)}\left\{t_{new}|\mathbf{x}_{new}\right\} &= E_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma,\alpha)}\left\{\mathbf{x}_{new}^{\mathsf{T}}\mathbf{w}\right\} \\
&= \mathbf{x}_{new}^{\mathsf{T}}\int \mathbf{w}p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma,\alpha)d\mathbf{w} \\
&= \mathbf{x}_{new}^{\mathsf{T}}\boldsymbol{\mu} = \mathbf{x}_{new}^{\mathsf{T}}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}+\frac{\sigma^2}{\alpha}\mathbf{I}\right)^{-1}\mathbf{X}
\end{aligned}
$$

Posterior Inference

UNIVERSITY of GLASGOW

$$
\begin{aligned}
E_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma,\alpha)}\left\{t_{new}|\mathbf{x}_{new}\right\} &= E_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma,\alpha)}\left\{\mathbf{x}_{new}^{\mathsf{T}}\mathbf{w}\right\} \\
&= \mathbf{x}_{new}^{\mathsf{T}}\int \mathbf{w}p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma,\alpha)d\mathbf{w} \\
&= \mathbf{x}_{new}^{\mathsf{T}}\boldsymbol{\mu} = \mathbf{x}_{new}^{\mathsf{T}}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}+\frac{\sigma^2}{\alpha}\mathbf{I}\right)^{-1}\mathbf{X}
\end{aligned}
$$

and var$(t_{new}|\mathbf{x}_{new})$

$$
\begin{aligned}
&= E_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma,\alpha)}\left\{t_{new}^2|\mathbf{x}_{new}\right\} - E^2_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma,\alpha)}\left\{t_{new}|\mathbf{x}_{new}\right\} \\
&= \mathbf{x}_{new}^{\mathsf{T}}E_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma)}\left\{\mathbf{w}\mathbf{w}^{\mathsf{T}}\right\}\mathbf{x}_{new} - \left(\mathbf{x}_{new}^{\mathsf{T}}E_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma)}\left\{\mathbf{w}\right\}\right)^2 \\
&= \mathbf{x}_{new}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{x}_{new} = \sigma^2\mathbf{x}_{new}^{\mathsf{T}}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}+\frac{\sigma^2}{\alpha}\mathbf{I}\right)^{-1}\mathbf{x}_{new}
\end{aligned}
$$

Effect of Prior

UNIVERSITY of GLASGOW

- Now as $\alpha \to \infty$ then we will recover the MLE prediction and this makes sense because the width of our Gaussian prior $p(\mathbf{w}|\alpha)$ will increase as α increases which means that we will become less precise about the prior values which the parameters should take and in the limit they will all become equally likely *a priori*.

Effect of Prior

- Now as $\alpha \rightarrow \infty$ then we will recover the MLE prediction and this makes sense because the width of our Gaussian prior $p(\mathbf{w}|\alpha)$ will increase as α increases which means that we will become less precise about the prior values which the parameters should take and in the limit they will all become equally likely *a priori*.
- Effect of prior on solution introduces bias what effect does this have on predictive power?

Effect of Prior

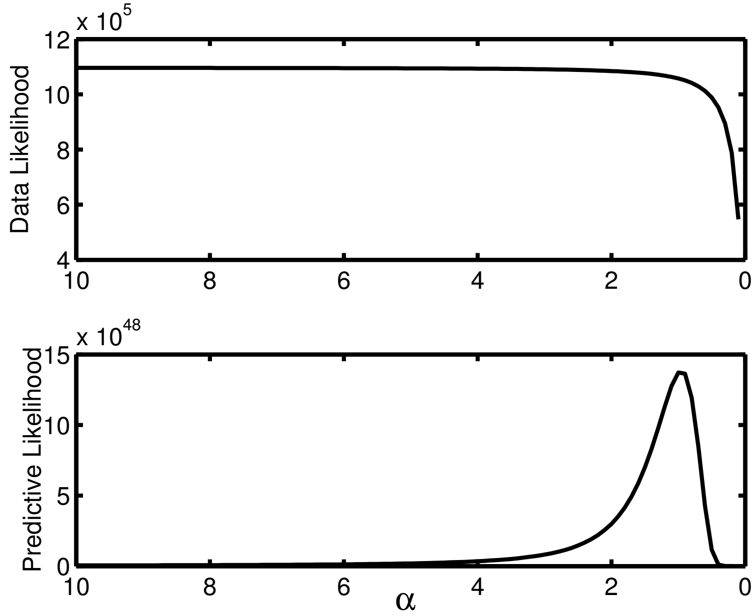

Figure 5: The top chart shows the in-sample likelihood as a function of the prior variance and we can see a drop in likelihood as the regularising effect of the prior becomes significant. The bottom chart shows how the out-of-sample predictive likelihood varies with α with a significant increase in performance at a specific α value. This is a nice example of the effect that bias & variance has on a predictive model.