

# COMS30035, Machine learning:

From regression to classification  
and neural networks:

Sequential Bayesian regression

Rui Ponte Costa

Department of Computer Science, SCEEM  
University of Bristol

September 29, 2022

# Textbooks

We will follow parts of the Chapter 3 of the Bishop book closely:

- ▶ Bishop, C. M., Pattern recognition and machine learning (2006). Available for free [here](#).

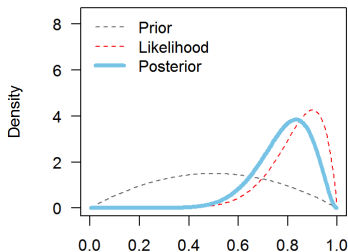
# Agenda

- ▶ Revising linear and nonlinear regression [see extra slides; Chapter 3, Bishop]
  - ▶ Linear regression
  - ▶ Nonlinear regression
  - ▶ Probabilistic models
  - ▶ Maximum likelihood estimation
- ▶ **Sequential Bayesian regression** [Chapter 3, Bishop]
  - ▶ Bayesian formulation
  - ▶ Conjugate priors
  - ▶ Example
- ▶ Classification and neural networks [Chapter 5, Bishop]
  - ▶ Architectures (Parametric model)
  - ▶ The supervised case
  - ▶ Optimising nnets using backprop
  - ▶ Highly flexible model → overfitting: early stopping/drop-out.

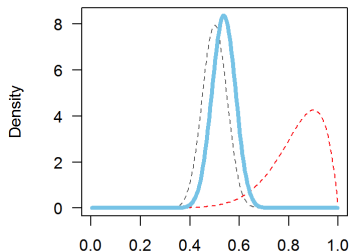
# Why Bayesian?

- ▶ In the previous lecture we considered likelihood methods
- ▶ But these ignore any prior knowledge we often have about  $\theta$
- ▶ We should use Bayesian inference, which combines prior and likelihood probabilities as
- ▶  $p(\theta|D) = p(D|\theta)p(\theta)/p(D)$ , where  $p(\theta|D)$  is the posterior,  $p(D|\theta)$  the likelihood of the data,  $p(\theta)$  the prior over the parameters and  $p(D)$  is the normalization term also called evidence.

Weak Prior



Strong Prior



# Example: Sequential Bayesian Linear Regression

We are going to go through an example of a Bayesian model that nicely illustrates its advantages.

- ▶ Goal: We want a model that iteratively adjusts as new data comes in
- ▶ Model: The posterior is given by

$$p(\theta|D) = p(D|\theta)p(\theta)/Z$$

where the likelihood is similar the previous lecture <sup>1</sup>:

$$p(D|\theta) = \prod_{i=1}^N p(y_i|x_i, \theta)$$

and we use a conjugate prior  $p(\theta) = \mathcal{N}(\theta|m_0, S_0)$ , where  $m_0$  and  $S_0$  are the mean and precision (inverse variance), respectively.

---

<sup>1</sup> To be consistent with Bishop, instead of using  $\sigma$  directly we use the precision parameter  $\beta = 1/\sigma^2$  which we assume to be given.

# Why conjugate priors?

A conjugate prior is one such that the posterior is in the probability distribution family. For example, given a Gaussian likelihood if we choose a Gaussian prior, then the posterior is guaranteed to be a Gaussian – this makes this prior a conjugate prior for the posterior.

**Why are they useful?** The Bayes theorem has a nasty normalising term  $p(D)$ , which is given by  $\int p(D|\theta)p(\theta)d\theta$ . For models with more than a few  $\theta$  it is intractable to compute this integral. Conjugate priors save us from this, as they lead to exact posterior for which we do not need to compute integrals.

# Example: Sequential Bayesian Linear Regression

Given the conjugate prior  $p(\theta) = \mathcal{N}(\theta|m_0, S_0)$ , then our posterior is

$$p(\theta|D) = \mathcal{N}(\theta|m_N, S_N)$$

where <sup>2</sup>

$$m_N = S_N(S_0^{-1}m_0 + \beta\Phi^T D)$$

$$S_N^{-1} = S_0^{-1} + \beta\Phi^T\Phi$$

$m_0$  and  $m_N$  are the mean of the prior and posterior, respectively.

$S_0$  and  $S_N$  are the precision of the prior and posterior (i.e.  $1/\sigma^2$ ), respectively.

$\beta$  is the noise precision parameter <sup>3</sup> (which we assume to be given).

In our example  $\Phi = X$  as we do not use any basis functions (see Bishop).

**Note:** The posterior becomes the prior (i.e.  $S_0 = S_N$  and  $m_0 = m_N$ ) for new data (see example next).

---

<sup>2</sup>More details on how to derive this equations in Bishop p.153 and [here](#) (w/ Python code).

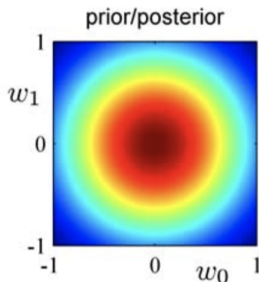
<sup>3</sup>As in the previous lecture  $y = a_1x + \mathcal{N}(0, \beta = 1/\sigma^2)$

# Example: Sequential Bayesian Linear Regression

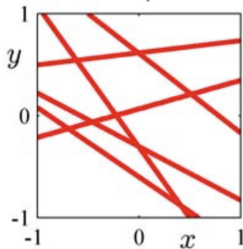
Lets run this Bayesian model using the linear model  $y = w_0 + w_1 x$

Before data arrives:

likelihood



data space

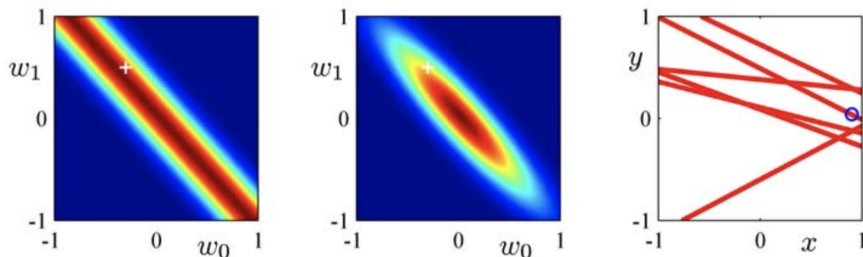


Notes: No data yet = no likelihood; prior is broad consequently so is the posterior; the model generates 'totally' random samples (red lines).



# Example: Sequential Bayesian Linear Regression

After 1 datapoint:



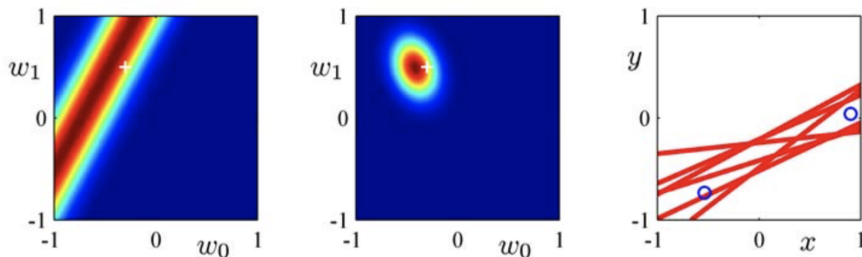
Notes: Data <sup>4</sup> = we have a likelihood; posterior is more precise <sup>5</sup>, and is used as prior for next iteration. This is the most important point of this lecture, **a Bayesian framework enables you to automatically consider previous estimates as priors for future model fittings!**

<sup>4</sup>Data samples are represented by blue scatter plot on the right.

<sup>5</sup>Parameter set used to generate the data is indicated by the white cross.

# Example: Sequential Bayesian Linear Regression

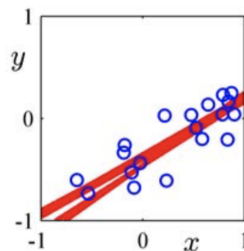
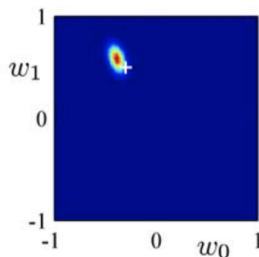
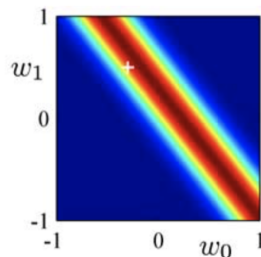
After 2 datapoints:



Notes: The posterior is even more precise, and the lines sampled from the model are better defined.

# Example: Sequential Bayesian Linear Regression

After 20 datapoints:



Notes: Posterior is very close to the original parameter set (white cross).

Quiz time!



Go to Blackboard unit page » Quizzes » Lecture 3.2

[Should take you less than 5 minutes]