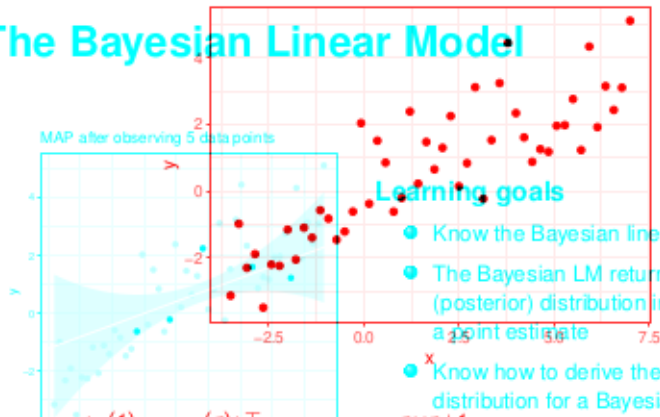


# REVIEW: THE BAYESIAN LINEAR MODEL

Let  $\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), \dots, (\mathbf{x}^{(n)}, y^{(n)})\}$  be a training set of i.i.d. observations from some unknown distribution.



## The Bayesian Linear Model



### Learning goals

- Know the Bayesian linear model
- The Bayesian LM returns a (posterior) distribution instead of a point estimate
- Know how to derive the posterior distribution for a Bayesian LM

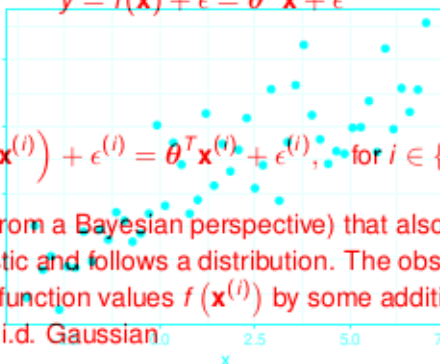
Let  $\mathbf{y} = (y^{(1)}, \dots, y^{(n)})^T$  and  $\mathbf{X} \in \mathbb{R}^{n \times p+1}$  be the design matrix where the  $i$ -th row contains vector  $\mathbf{x}^{(i)}$ .

## REVIEW: THE BAYESIAN LINEAR MODEL / 2

The linear (regression) model is defined as a training set of i.i.d. observations from some unknown distribution.



or on the data:

$$y = f(\mathbf{x}) + \epsilon = \boldsymbol{\theta}^T \mathbf{x} + \epsilon$$

$$y^{(i)} = f(\mathbf{x}^{(i)}) + \epsilon^{(i)} = \boldsymbol{\theta}^T \mathbf{x}^{(i)} + \epsilon^{(i)}, \text{ for } i \in \{1, \dots, n\}$$

We now assume (from a Bayesian perspective) that also our parameter vector  $\boldsymbol{\theta}$  is stochastic and follows a distribution. The observed values  $y^{(i)}$  differ from the function values  $f(\mathbf{x}^{(i)})$  by some additive noise, which is assumed to be i.i.d. Gaussian

Let  $\mathbf{y} = (y^{(1)}, \dots, y^{(n)})^T$  and  $\mathbf{X} \in \mathbb{R}^{n \times p}$  be the design matrix where the  $i$ -th row contains vector  $\mathbf{x}^{(i)}$ .

$$\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$$

## REVIEW: THE BAYESIAN LINEAR MODEL / 3

Let us assume we have **prior beliefs** about the parameter  $\theta$  that are represented in a prior distribution  $\theta \sim \mathcal{N}(\mathbf{0}, \tau^2 I_p)$ .

$$y = f(\mathbf{x}) + \epsilon = \theta^T \mathbf{x} + \epsilon$$

Whenever data points are observed, we update the parameters' prior distribution according to Bayes' rule **or on the data:**

$$y^{(i)} = f(\mathbf{x}^{(i)}) + \epsilon^{(i)} = \theta^T \mathbf{x}^{(i)} + \epsilon^{(i)} \quad \text{for } i \in \{1, \dots, n\}$$
$$p(\theta | \mathbf{X}, \mathbf{y}) = \frac{\overbrace{p(\mathbf{y} | \mathbf{X}, \theta)}^{\text{likelihood}} \overbrace{q(\theta)}^{\text{prior}}}{\underbrace{p(\mathbf{y} | \mathbf{X})}_{\text{marginal}}}$$

We now assume (from a Bayesian perspective) that also our parameter vector  $\theta$  is stochastic and follows a distribution. The observed values  $y^{(i)}$  differ from the function values  $f(\mathbf{x}^{(i)})$  by some additive noise, which is assumed to be i.i.d. Gaussian

$$\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$$

and independent of  $\mathbf{x}$  and  $\theta$ .



## REVIEW: THE BAYESIAN LINEAR MODEL / 4

The posterior distribution of the parameter  $\theta$  is again normal distributed (the Gaussian family is self-conjugate):  $\mathcal{N}(\mathbf{0}, \tau^2 I_p)$ .



Whenever data points  $\theta | \mathbf{X}, \mathbf{y} \sim \mathcal{N}(\sigma^{-2} \mathbf{A}^{-1} \mathbf{X}^T \mathbf{y}, \mathbf{A}^{-1})$  parameters' prior distribution according to Bayes' rule with  $\mathbf{A} := \sigma^{-2} \mathbf{X}^T \mathbf{X} + \frac{1}{\tau^2} I_p$ .

Remarks: (1) Please see the Deep Dive part for the detailed derivation.  
(2) The expectation of  $\theta | \mathbf{X}, \mathbf{y}$  is exactly the solution of ridge regression.

$$\underbrace{p(\theta | \mathbf{X}, \mathbf{y})}_{\text{posterior}} = \frac{\underbrace{p(\mathbf{y} | \mathbf{X}, \theta)}_{\text{likelihood}} \underbrace{q(\theta)}_{\text{prior}}}{\underbrace{p(\mathbf{y} | \mathbf{X})}_{\text{marginal}}}$$

**Note:** If the posterior distribution  $p(\theta | \mathbf{X}, \mathbf{y})$  are in the same probability distribution family as the prior  $q(\theta)$  w.r.t. a specific likelihood function  $p(\mathbf{y} | \mathbf{X}, \theta)$ , they are called **conjugate distributions**. The prior is then called a **conjugate prior** for the likelihood.

The Gaussian family is self-conjugate: Choosing a Gaussian

Likelihood ensures that the posterior is Gaussian.

## REVIEW: THE BAYESIAN LINEAR MODEL / 5

The posterior distribution of the parameter  $\theta$  is again normal distributed (the Gaussian family is self-conjugate):

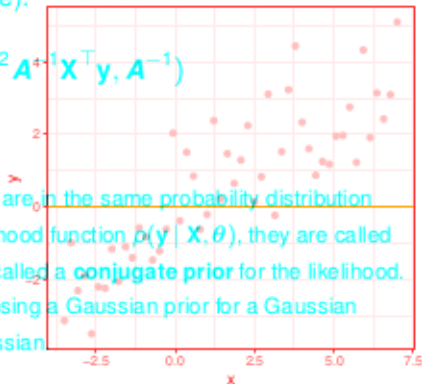
No data points observed

Prior  $\theta \sim \mathcal{N}(0, 1)$

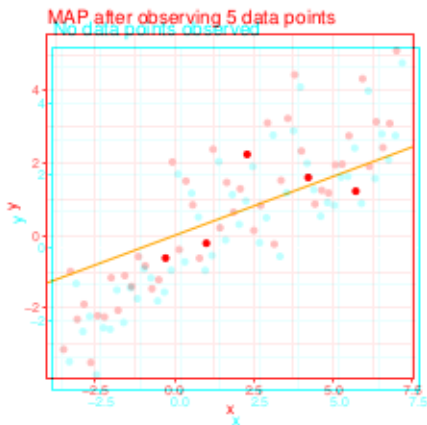
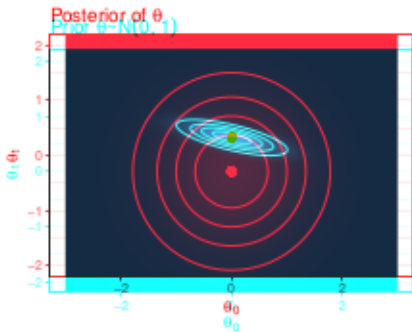
$$\theta | \mathbf{X}, \mathbf{y} \sim \mathcal{N}(\sigma^{-2} \mathbf{A}^{-1} \mathbf{X}^T \mathbf{y}, \mathbf{A}^{-1})$$

with  $\mathbf{A} := \sigma^{-2} \mathbf{X}^T \mathbf{X} + \frac{1}{\lambda} \mathbf{I}_p$ .

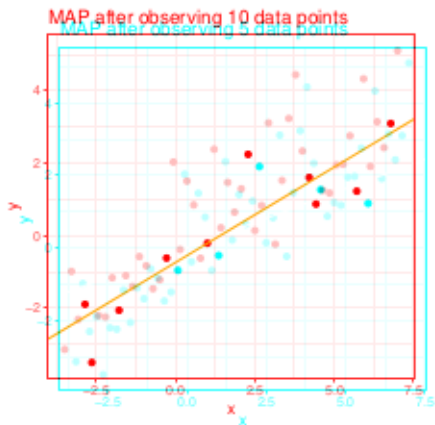
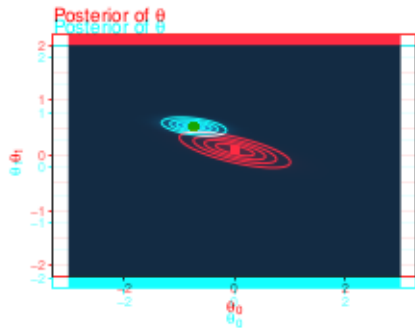
**Note:** If the posterior distribution  $p(\theta | \mathbf{X}, \mathbf{y})$  are in the same probability distribution family as the prior  $q(\theta)$  w.r.t. a specific likelihood function  $p(\mathbf{y} | \mathbf{X}, \theta)$ , they are called **conjugate distributions**. The prior is then called a **conjugate prior** for the likelihood. The Gaussian family is self-conjugate. Choosing a Gaussian prior for a Gaussian Likelihood ensures that the posterior is Gaussian.



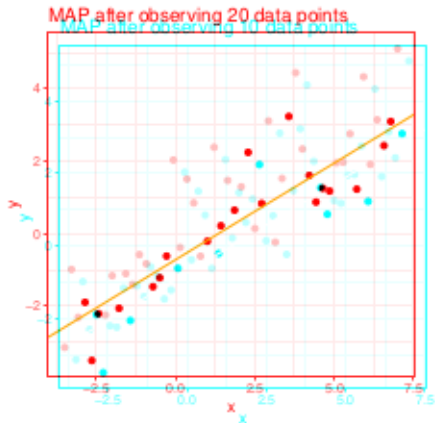
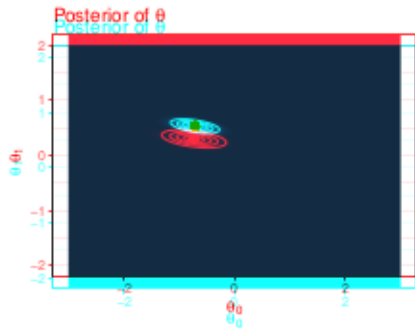
# REVIEW: THE BAYESIAN LINEAR MODEL



# REVIEW: THE BAYESIAN LINEAR MODEL



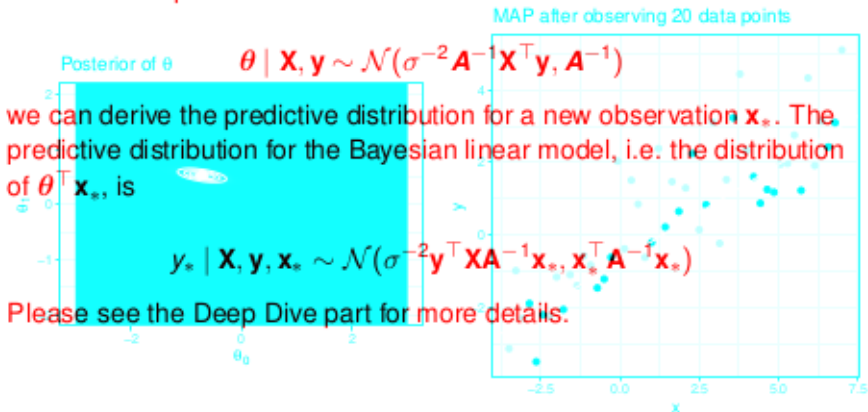
# REVIEW: THE BAYESIAN LINEAR MODEL





# REVIEW: THE BAYESIAN LINEAR MODEL

Based on the posterior distribution



# REVIEW: THE BAYESIAN LINEAR MODEL

## Proof:

We want to show that

- for a Gaussian prior on  $\theta \sim \mathcal{N}(0, \tau^2 I_p)$
- for a Gaussian Likelihood  $y | \mathbf{X}, \theta \sim \mathcal{N}(\mathbf{X}^T \theta, \sigma^2 I_n)$

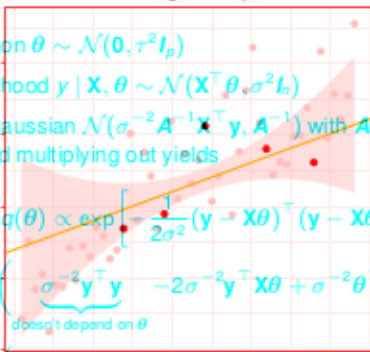
the resulting posterior is Gaussian  $\mathcal{N}(\sigma^{-2} \mathbf{A}^{-1} \mathbf{X}^T \mathbf{y}, \mathbf{A}^{-1})$  with  $\mathbf{A} := \sigma^{-2} \mathbf{X}^T \mathbf{X} + \frac{1}{\tau^2} I_p$ .  
 Plugging in Bayes' rule and multiplying out yields

$$\begin{aligned}
 p(\theta | \mathbf{X}, \mathbf{y}) &\propto p(\mathbf{y} | \mathbf{X}, \theta) p(\theta) \propto \exp \left[ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\theta)^T (\mathbf{y} - \mathbf{X}\theta) - \frac{1}{2\tau^2} \theta^T \theta \right] \\
 &= \exp \left[ -\frac{1}{2} \left( \underbrace{\sigma^{-2} \mathbf{y}^T \mathbf{y}}_{\text{doesn't depend on } \theta} - 2\sigma^{-2} \mathbf{y}^T \mathbf{X}\theta + \sigma^{-2} \theta^T \mathbf{X}^T \mathbf{X}\theta + \tau^{-2} \theta^T \theta \right) \right] \\
 &\propto \exp \left[ -\frac{1}{2} \left( \sigma^{-2} \theta^T \mathbf{X}^T \mathbf{X}\theta + \tau^{-2} \theta^T \theta - 2\sigma^{-2} \mathbf{y}^T \mathbf{X}\theta \right) \right]
 \end{aligned}$$

For every test input  $\mathbf{x}_*$ , we get a distribution over the prediction  $y_*$ . In particular, we get a posterior mean (orange) and a posterior variance (grey region equals  $\pm$  two times standard deviation).

This expression resembles a normal density - except for the term in red!

MAP after observing 5 data points



# REVIEW: THE BAYESIAN LINEAR MODEL

**Note:** We need not worry about the normalizing constant since its mere role is to convert probability functions to density functions with a total probability of one.

We subtract a (not yet defined) constant  $c$  while compensating for this change by adding the respective terms ("adding 0"), emphasized in green:

$$p(\theta | \mathbf{X}, \mathbf{y}) \propto \exp \left[ -\frac{1}{2} (\theta - c)^T \mathbf{A} (\theta - c) - c^T \mathbf{A} \theta + \frac{1}{2} c^T \mathbf{A} c + \sigma^{-2} \mathbf{y}^T \mathbf{X} \theta \right]$$

doesn't depend on  $\theta$

$$\propto \exp \left[ -\frac{1}{2} (\theta - c)^T \mathbf{A} (\theta - c) - c^T \mathbf{A} \theta + \sigma^{-2} \mathbf{y}^T \mathbf{X} \theta \right]$$

If we choose  $c$  such that  $-c^T \mathbf{A} \theta + \sigma^{-2} \mathbf{y}^T \mathbf{X} \theta = 0$ , the posterior is normal with mean  $c$  and covariance matrix  $\mathbf{A}^{-1}$ . Taking into account that  $\mathbf{A}$  is symmetric, this is if we choose

$$\begin{aligned} \sigma^{-2} \mathbf{y}^T \mathbf{X} &= c^T \mathbf{A} \\ \Leftrightarrow \sigma^{-2} \mathbf{y}^T \mathbf{X} \mathbf{A}^{-1} &= c^T \\ \Leftrightarrow c &= \sigma^{-2} \mathbf{A}^{-1} \mathbf{X}^T \mathbf{y} \end{aligned}$$

For every test input  $\mathbf{x}_*$ , we get a distribution over the prediction  $y_*$ . In particular, we get a posterior mean (orange) and a posterior variance (grey region equals  $\pm$  two times standard deviation).  
as claimed.



# REVIEW: THE BAYESIAN LINEAR MODEL

Based on the posterior distribution

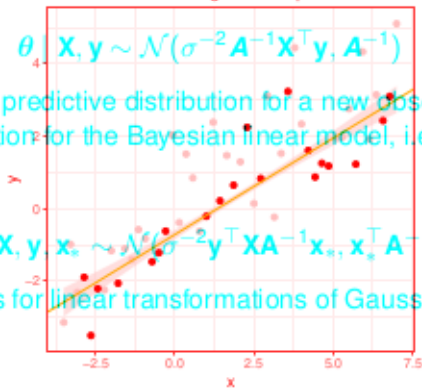
MAP after observing 20 data points

$$\theta | \mathbf{X}, \mathbf{y} \sim \mathcal{N}(\sigma^{-2} \mathbf{A}^{-1} \mathbf{X}^T \mathbf{y}, \mathbf{A}^{-1})$$

we can derive the predictive distribution for a new observations  $\mathbf{x}_*$ . The predictive distribution for the Bayesian linear model, i.e. the distribution of  $\theta^T \mathbf{x}_*$ , is

$$y_* | \mathbf{X}, \mathbf{y}, \mathbf{x}_* \sim \mathcal{N}(\sigma^{-2} \mathbf{y}^T \mathbf{X} \mathbf{A}^{-1} \mathbf{x}_*, \mathbf{x}_*^T \mathbf{A}^{-1} \mathbf{x}_*)$$

(applying the rules for linear transformations of Gaussians).

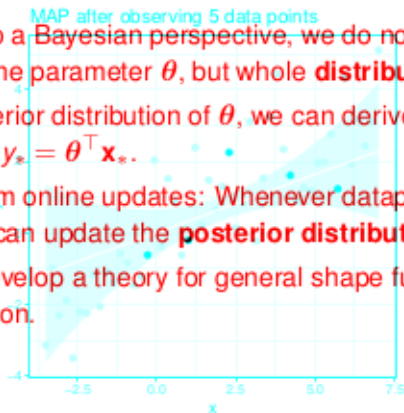


For every test input  $\mathbf{x}_*$ , we get a distribution over the prediction  $y_*$ . In particular, we get a posterior mean (orange) and a posterior variance (grey region equals  $\pm$  two times standard deviation).

## SUMMARY: THE BAYESIAN LINEAR MODEL

- By switching to a Bayesian perspective, we do not only have point estimates for the parameter  $\theta$ , but whole **distributions**
- From the posterior distribution of  $\theta$ , we can derive a predictive distribution for  $y_* = \theta^T \mathbf{x}_*$ .
- We can perform online updates: Whenever datapoints are observed, we can update the **posterior distribution** of  $\theta$

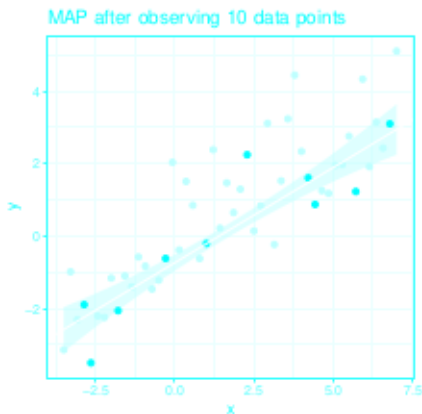
Next, we want to develop a theory for general shape functions, and not only for linear function.



For every test input  $\mathbf{x}_*$ , we get a distribution over the prediction  $y_*$ . In particular, we get a posterior mean (orange) and a posterior variance (grey region equals  $\pm$  two times standard deviation).

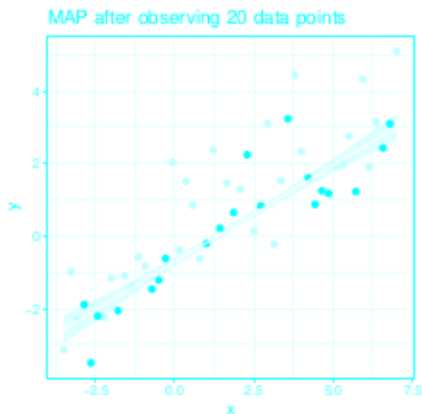


## REVIEW: THE BAYESIAN LINEAR MODEL



For every test input  $x_*$ , we get a distribution over the prediction  $y_*$ . In particular, we get a posterior mean (orange) and a posterior variance (grey region equals  $\pm$  two times standard deviation).

## REVIEW: THE BAYESIAN LINEAR MODEL



For every test input  $x_*$ , we get a distribution over the prediction  $y_*$ . In particular, we get a posterior mean (orange) and a posterior variance (grey region equals  $\pm$  two times standard deviation).

## SUMMARY: THE BAYESIAN LINEAR MODEL

- By switching to a Bayesian perspective, we do not only have point estimates for the parameter  $\theta$ , but whole **distributions**
- From the posterior distribution of  $\theta$ , we can derive a predictive distribution for  $y_* = \theta^\top \mathbf{x}_*$ .
- We can perform online updates: Whenever datapoints are observed, we can update the **posterior distribution** of  $\theta$

Next, we want to develop a theory for general shape functions, and not only for linear function.