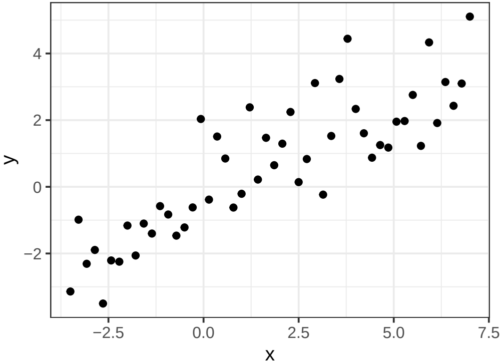
REVIEW: THE BAYESIAN LINEAR MODEL

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



Let $\mathbf{y} = (y^{(1)}, \ldots, y^{(n)})^\top$ 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

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

or on the data:

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

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 $f\left(\mathbf{x}^{(i)}\right)$ 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 $\boldsymbol{\theta}$.

REVIEW: THE BAYESIAN LINEAR MODEL / 3

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

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

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

REVIEW: THE BAYESIAN LINEAR MODEL / 4

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

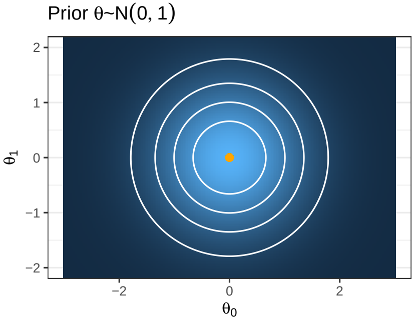
$$\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y} \sim \mathcal{N}(\sigma^{-2}\boldsymbol{A}^{-1}\mathbf{X}^\top \mathbf{y}, \boldsymbol{A}^{-1})$$

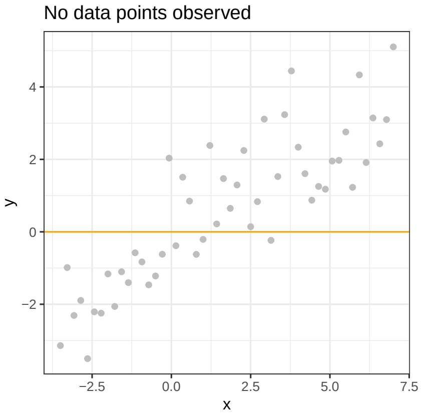
with $\boldsymbol{A} := \sigma^{-2}\mathbf{X}^\top\mathbf{X} + \frac{1}{\tau^2}\boldsymbol{I}_p$.

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

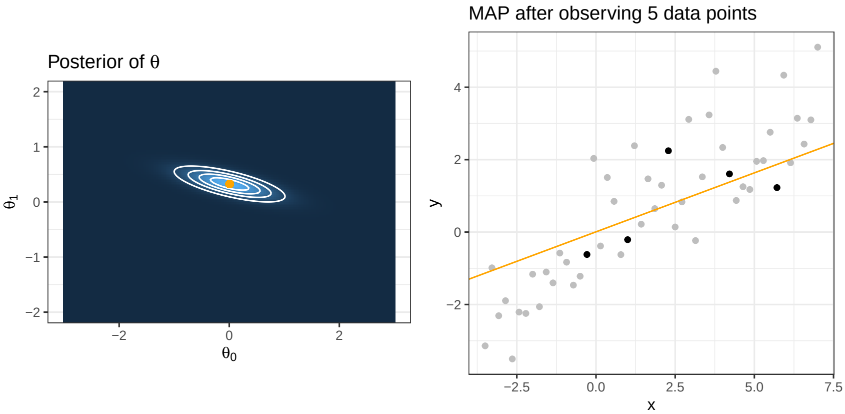
Note: If the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y})$ are in the same probability distribution family as the prior $q(\boldsymbol{\theta})$ w.r.t. a specific likelihood function $p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta})$, they are called **conjugate distributions**. 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 / 5

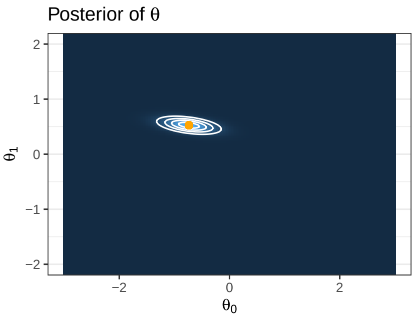


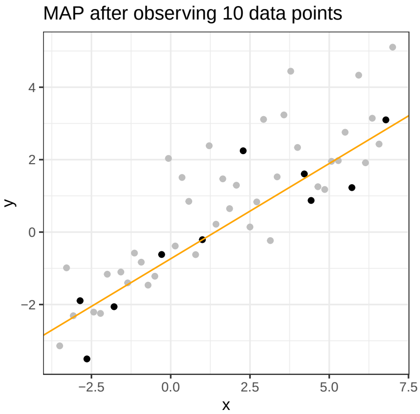


REVIEW: THE BAYESIAN LINEAR MODEL

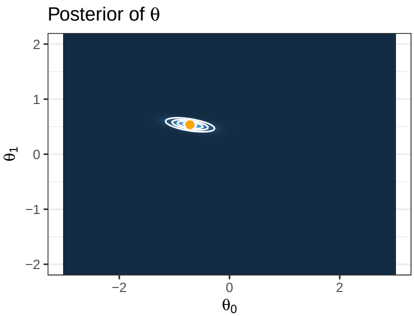


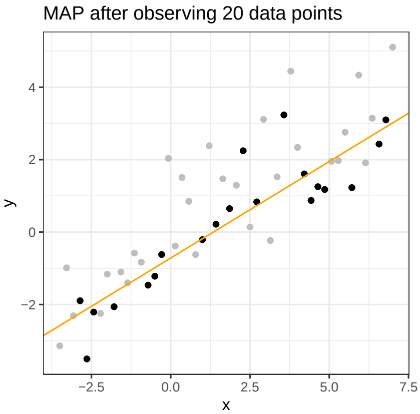
REVIEW: THE BAYESIAN LINEAR MODEL





REVIEW: THE BAYESIAN LINEAR MODEL





REVIEW: THE BAYESIAN LINEAR MODEL

Based on the posterior distribution

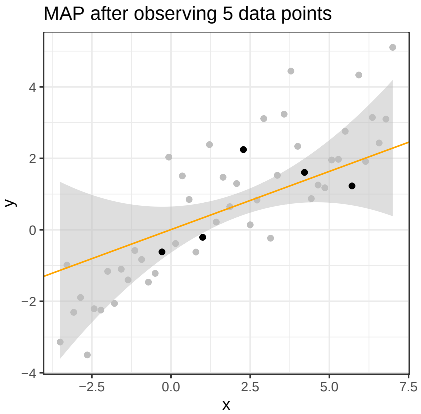
$$\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y} \sim \mathcal{N}(\sigma^{-2}\boldsymbol{A}^{-1}\mathbf{X}^\top\mathbf{y}, \boldsymbol{A}^{-1})$$

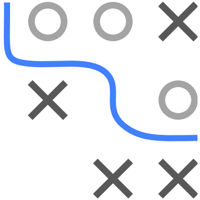
we can derive the predictive distribution for a new observation $\mathbf{x}_*$. The predictive distribution for the Bayesian linear model, i.e. the distribution of $\boldsymbol{\theta}^\top\mathbf{x}_*$, is

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

Please see the Deep Dive part for more details.

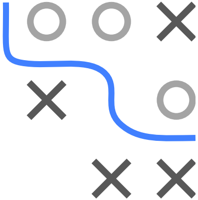
REVIEW: THE BAYESIAN LINEAR MODEL

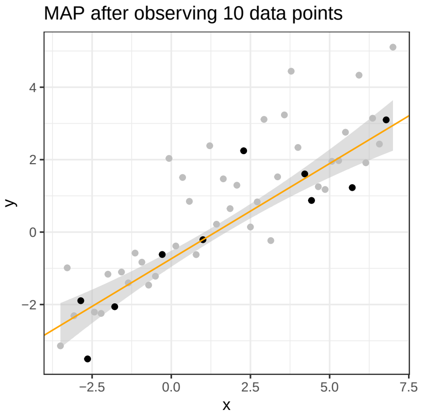




For every test input $\mathbf{x}_*$, we get a distribution over the prediction y_*. In particular, we get a posterior mean (orange) and a posterior variance (grey region equals $+/-$ two times standard deviation).

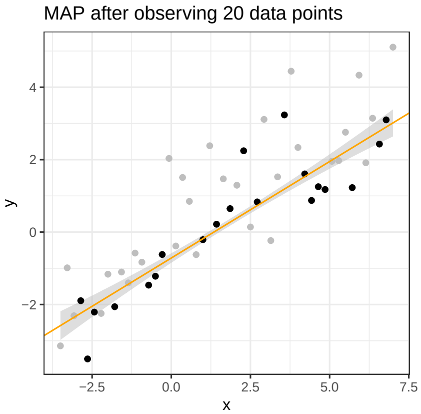
REVIEW: THE BAYESIAN LINEAR MODEL





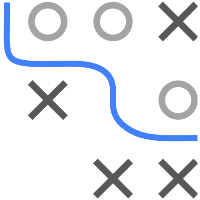
For every test input $\mathbf{x}_*$, we get a distribution over the prediction y_*. In particular, we get a posterior mean (orange) and a posterior variance (grey region equals $+/-$ two times standard deviation).

REVIEW: THE BAYESIAN LINEAR MODEL



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 $+/-$ 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 $\boldsymbol{\theta}$, but whole **distributions**
- From the posterior distribution of $\boldsymbol{\theta}$, we can derive a predictive distribution for $y_* = \boldsymbol{\theta}^\top \mathbf{x}_*$.
- We can perform online updates: Whenever datapoints are observed, we can update the **posterior distribution** of $\boldsymbol{\theta}$

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