Solution 1: Bayesian Linear Model

The posterior distribution is obtained by Bayes' rule

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

In the Bayesian linear model we have a Gaussian likelihood: $\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\theta}, \sigma^2 \mathbf{I}_n)$, i.e.,

$$\begin{split} p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) &\propto \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^{\top}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})\right] \\ &= \exp\left[-\frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2}{2\sigma^2}\right] \\ &= \exp\left[-\frac{\sum_{i=1}^n(y^{(i)} - \boldsymbol{\theta}^{\top}\mathbf{x}^{(i)})^2}{2\sigma^2}\right]. \end{split}$$

Moreover, note that the maximum a posteriori estimate of θ , which is defined by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})$$

can also be defined by

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} \log (p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})),$$

since log is a monotonically increasing function, so the maximizer is the same.

(a) If the prior distribution is a uniform distribution over the parameter vectors $\boldsymbol{\theta}$, i.e.,

$$q(\boldsymbol{\theta}) \propto 1$$
,

then

$$\begin{split} p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y}) & \propto & p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})q(\boldsymbol{\theta}) \\ & \propto & \exp\left[-\frac{\sum_{i=1}^{n}(y^{(i)}-\boldsymbol{\theta}^{\top}\mathbf{x}^{(i)})^{2}}{2\sigma^{2}}\right]. \end{split}$$

With this,

$$\begin{split} \hat{\boldsymbol{\theta}} &= \arg \max_{\boldsymbol{\theta}} \log \left(p(\boldsymbol{\theta} | \mathbf{X}, \mathbf{y}) \right) \\ &= \arg \max_{\boldsymbol{\theta}} - \frac{\sum_{i=1}^{n} (y^{(i)} - \boldsymbol{\theta}^{\top} \mathbf{x}^{(i)})^{2}}{2\sigma^{2}} \\ &= \arg \min_{\boldsymbol{\theta}} \frac{\sum_{i=1}^{n} (y^{(i)} - \boldsymbol{\theta}^{\top} \mathbf{x}^{(i)})^{2}}{2\sigma^{2}} \\ &= \arg \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (y^{(i)} - \boldsymbol{\theta}^{\top} \mathbf{x}^{(i)})^{2}, \end{split} \tag{$2\sigma^{2}$ is just a constant scaling)}$$

so the maximum a posteriori estimate coincides with the empirical risk minimizer for the L2-loss (over the linear models).

(b) If we choose a Gaussian distribution over the parameter vectors $\boldsymbol{\theta}$ as the prior belief, i.e.,

$$q(\boldsymbol{\theta}) \propto \exp \left[-\frac{1}{2\tau^2} \boldsymbol{\theta}^\top \boldsymbol{\theta} \right], \qquad \tau > 0,$$

then

$$p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) \propto p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})q(\boldsymbol{\theta})$$

$$\propto \exp\left[-\frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^{\top}\mathbf{x}^{(i)})^{2}}{2\sigma^{2}} - \frac{1}{2\tau^{2}}\boldsymbol{\theta}^{\top}\boldsymbol{\theta}\right]$$

$$= \exp\left[-\frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^{\top}\mathbf{x}^{(i)})^{2}}{2\sigma^{2}} - \frac{\|\boldsymbol{\theta}\|_{2}^{2}}{2\tau^{2}}\right]$$

With this,

$$\begin{split} \hat{\boldsymbol{\theta}} &= \arg \max_{\boldsymbol{\theta}} \log \left(p(\boldsymbol{\theta} | \mathbf{X}, \mathbf{y}) \right) \\ &= \arg \max_{\boldsymbol{\theta}} - \frac{\sum_{i=1}^{n} (y^{(i)} - \boldsymbol{\theta}^{\top} \mathbf{x}^{(i)})^{2}}{2\sigma^{2}} - \frac{\|\boldsymbol{\theta}\|_{2}^{2}}{2\tau^{2}} \\ &= \arg \min_{\boldsymbol{\theta}} \frac{\sum_{i=1}^{n} (y^{(i)} - \boldsymbol{\theta}^{\top} \mathbf{x}^{(i)})^{2}}{2\sigma^{2}} + \frac{\|\boldsymbol{\theta}\|_{2}^{2}}{2\tau^{2}} \\ &= \arg \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (y^{(i)} - \boldsymbol{\theta}^{\top} \mathbf{x}^{(i)})^{2} + \frac{\sigma^{2}}{\tau^{2}} \|\boldsymbol{\theta}\|_{2}^{2}, \end{split}$$

so the maximum a posteriori estimate coincides for the choice of $\lambda = \frac{\sigma^2}{\tau^2} > 0$ with the regularized empirical risk minimizer for the L2-loss with L2 penalty (over the linear models), i.e., the Ridge regression.

(c) If we choose a Laplace distribution over the parameter vectors $\boldsymbol{\theta}$ as the prior belief, i.e.,

$$q(\boldsymbol{\theta}) \propto \exp\left[-\frac{\sum_{i=1}^{p} |\boldsymbol{\theta}_i|}{\tau}\right], \quad \tau > 0,$$

then

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) & \propto & p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) q(\boldsymbol{\theta}) \\ & \propto & \exp\left[-\frac{\sum_{i=1}^{n} (y^{(i)} - \boldsymbol{\theta}^{\top} \mathbf{x}^{(i)})^{2}}{2\sigma^{2}} - \frac{\sum_{i=1}^{p} |\boldsymbol{\theta}_{i}|}{\tau}\right] \\ & = & \exp\left[-\frac{\sum_{i=1}^{n} (y^{(i)} - \boldsymbol{\theta}^{\top} \mathbf{x}^{(i)})^{2}}{2\sigma^{2}} - \frac{\|\boldsymbol{\theta}\|_{1}}{\tau}\right] \end{aligned}$$

With this,

$$\begin{split} \hat{\boldsymbol{\theta}} &= \arg \max_{\boldsymbol{\theta}} \log \left(p(\boldsymbol{\theta} | \mathbf{X}, \mathbf{y}) \right) \\ &= \arg \max_{\boldsymbol{\theta}} - \frac{\sum_{i=1}^{n} (y^{(i)} - \boldsymbol{\theta}^{\top} \mathbf{x}^{(i)})^{2}}{2\sigma^{2}} - \frac{\|\boldsymbol{\theta}\|_{1}}{\tau} \\ &= \arg \min_{\boldsymbol{\theta}} \frac{\sum_{i=1}^{n} (y^{(i)} - \boldsymbol{\theta}^{\top} \mathbf{x}^{(i)})^{2}}{2\sigma^{2}} + \frac{\|\boldsymbol{\theta}\|_{1}}{\tau} \\ &= \arg \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (y^{(i)} - \boldsymbol{\theta}^{\top} \mathbf{x}^{(i)})^{2} + \frac{2\sigma^{2}}{\tau} \|\boldsymbol{\theta}\|_{1}, \end{split}$$

so the maximum a posteriori estimate coincides for the specific choice of $\lambda = \frac{2\sigma^2}{\tau}$ with the regularized empirical risk minimizer for the L2-loss with L1 penalty (over the linear models), i.e., the Lasso regression.

Solution 2: Gaussian Posterior Process

(a) Prior distribution (assuming the same notation as in the lecture):

$$f \sim \mathcal{N}(m, K)$$

with $\mathbf{m} = m(\mathbf{x})$ and \mathbf{K} defined by the entries $\mathbf{K}_{ij} = k(x_i, x_j)$. NB: Note the (in-)finite Gaussian property of a GP.

(b) Note that the posterior distribution f|y,x in this case is different from the one of $f_*|x_*,x,y$ and also from the marginal distribution of $y \sim \mathcal{N}(m, K + \sigma^2 I)$! We have:

$$p(\boldsymbol{f}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{f}) \cdot p(\boldsymbol{f})$$

$$\propto \exp(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{f})^{\top}(\sigma^{2}\boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{f})) \cdot \exp(-\frac{1}{2}(\boldsymbol{f} - \boldsymbol{m})^{\top}\boldsymbol{K}^{-1}(\boldsymbol{f} - \boldsymbol{m}))$$

$$\propto \exp(-\frac{1}{2}\{\boldsymbol{f}^{\top}\underbrace{((\sigma^{2}\boldsymbol{I})^{-1} + \boldsymbol{K}^{-1})}_{=:\boldsymbol{K}_{post}^{-1}}\boldsymbol{f} - 2\boldsymbol{f}^{\top}\underbrace{((\sigma^{2}\boldsymbol{I})^{-1}\boldsymbol{y} + \boldsymbol{K}^{-1}\boldsymbol{m})}_{=:\tilde{\boldsymbol{f}}}\})$$

$$\propto \exp(-\frac{1}{2}\{\boldsymbol{f}^{\top}\boldsymbol{K}_{post}^{-1}\boldsymbol{f} - 2\boldsymbol{f}^{\top}\tilde{\boldsymbol{f}}\})$$

$$(1)$$

by removing all constant factors that do not depend on f as we only need to know the density up to a constant of proportionality. By extending the proportionality, we can get a quadratic form in f:

$$p(\boldsymbol{f}|\boldsymbol{y}) \propto \exp(-\frac{1}{2} \{ \boldsymbol{f}^{\top} \boldsymbol{K}_{post}^{-1} \boldsymbol{f} - 2 \boldsymbol{f}^{\top} \tilde{\boldsymbol{f}} \})$$

$$\propto \exp(-\frac{1}{2} \{ \boldsymbol{f}^{\top} \boldsymbol{K}_{post}^{-1} \boldsymbol{f} - 2 \boldsymbol{f}^{\top} \boldsymbol{K}_{post}^{-1} \underbrace{\boldsymbol{K}_{post} \tilde{\boldsymbol{f}}}_{post} \})$$

$$\approx \exp(-\frac{1}{2} (\boldsymbol{f} - \boldsymbol{f}_{post})^{\top} \boldsymbol{K}_{post}^{-1} (\boldsymbol{f} - \boldsymbol{f}_{post}))$$

$$(2)$$

which is the so-called kernel of a multivariate normal distribution $\mathcal{N}(f_{post}, K_{post})$, i.e., $f|y \sim \mathcal{N}(f_{post}, K_{post})$.

(c) In order to get the posterior predictive distribution for a new sample x_* from the same data-generating process, we could derive

$$p(y_*|x_*, \boldsymbol{y}, \boldsymbol{x}) = \int p(y_*|x_*, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{f}) \cdot p(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{x}) d\boldsymbol{f}.$$

This is feasible but cumbersome. Alternatively, we can make use of the fact that the joint distribution of y and y_* is known (cf. slides on noisy GP):

$$\begin{pmatrix} \boldsymbol{y} \\ y_* \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \boldsymbol{m} \\ m_* \end{pmatrix}, \begin{pmatrix} \boldsymbol{K} + \sigma^2 \boldsymbol{I} & \boldsymbol{K}_* \\ \boldsymbol{K}_*^\top & K_{**} \end{pmatrix} \right),$$

with $m_* = m(x_*)$, $K_* = k(x_*, x)$ and $K_{**} = k(x_*, x_*)$. The conditional distribution can then be derived using the rule of conditioning for Gaussian distributions:

$$y_*|x_*, x, y \sim \mathcal{N}(m_* + K_*^{\top}(K + \sigma^2 I)^{-1}(y - m), K_{**} - K_*^{\top}(K + \sigma^2 I)^{-1}K_*).$$

(d) To implement a GP with squared exponential kernel and $\ell = 1$, we need the inverse of K. x being a vector implies that we have only one feature and thus the entries of our matrix K are

$$K = \begin{pmatrix} 1 & \exp(-0.5(x^{(1)} - x^{(2)})^2) \\ \exp(-0.5(x^{(2)} - x^{(1)})^2) & 1 \end{pmatrix}.$$

The inverse of \boldsymbol{K} is then given by

$$\frac{1}{1-\exp(-(x^{(1)}-x^{(2)})^2)}\begin{pmatrix} 1 & -\exp(-0.5(x^{(1)}-x^{(2)})^2) \\ -\exp(-0.5(x^{(2)}-x^{(1)})^2) & 1 \end{pmatrix}.$$

If we have a noisy GP, we would have to add $\sigma^2 I_2$ to K with resulting inverse

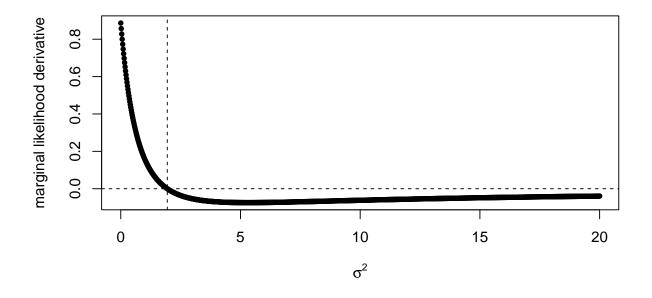
$$\boldsymbol{K}_y^{-1} = \frac{1}{(1+\sigma^2)^2 - \exp(-(x^{(1)}-x^{(2)})^2)} \begin{pmatrix} 1+\sigma^2 & -\exp(-0.5(x^{(1)}-x^{(2)})^2) \\ -\exp(-0.5(x^{(2)}-x^{(1)})^2) & 1+\sigma^2 \end{pmatrix}.$$

Assuming a zero mean GP, we can derive $\frac{\partial \mathbf{K}_y}{\partial \theta}$ with $\theta = \sigma^2$, which gives us the identity matrix. We can thus maximize the marginal likelihood (slide on *Gaussian Process Training*), by finding σ^2 that yields

$$\operatorname{tr}\left(\boldsymbol{K}_{y}^{-1}\boldsymbol{y}\boldsymbol{y}^{\top}\boldsymbol{K}_{y}^{-1}-\boldsymbol{K}_{y}^{-1}\right)=0.$$

This can be solved analytically (though quite tedious). We will use a root-finding function for this. For the posterior predictive distribution we can make use of the results from the previous exercise.

```
library(kernlab)
# set seed, define n, true (unknown) sigma
set.seed(4212)
n <- 2
sigma <- 1
# define kernel with l = 1
kernel_fun <- function(x)</pre>
 kernelMatrix(kernel = rbfdot(sigma = 1/2),
               x = x
kernel_fun_pred <- function(x,y)</pre>
 kernelMatrix(kernel = rbfdot(sigma = 1/2),
               x = x, y = y
# draw data according to the generating process:
x <- rnorm(n)
K <- kernel_fun(x)</pre>
K_y <- K + diag(rep(sigma^2,2))</pre>
(y <- t(mvtnorm::rmvnorm(1, sigma = K_y)))</pre>
            [,1]
## [1,] 2.012317
## [2,] 1.866819
# function to find the best sigma^2
root_fun <- function(sigmaSq){</pre>
 K_y_inv <- solve(K + diag(rep(sigmaSq,2)))</pre>
 0.5*sum(diag(K_y_inv%*%y%*%t(y)%*%K_y_inv - K_y_inv))
# get the best sigma
(bestSigmaSq \leftarrow uniroot(f = root_fun, interval = c(0,20)))root
## [1] 1.943684
# plot the optimization problem and best sigma
possible_sigvals <- seq(0.001,20,l=1000)
plot(possible_sigvals, sapply(possible_sigvals, root_fun),
     xlab = expression(sigma^2), ylab = "marginal likelihood derivative",
     pch = 20)
abline(h=0, lty=2)
abline(v=bestSigmaSq$root, lty=2)
```



```
# function to draw samples from the predictive posterior
draw_from_pred_posterior <- function(number_samples, y, x, xstar, sigmaSq = 1)</pre>
  # invert noisy K
 K_y_inv <- solve(kernel_fun(x) + diag(rep(sigmaSq,2)))</pre>
  # get the other K's for new data
 Kstar <- kernel_fun_pred(x,xstar)</pre>
 Kstarstar <- kernel_fun(xstar)</pre>
  # draw samples according to Ex. (d)
  rnorm(number_samples,
        mean = as.numeric(t(Kstar) %*% K_y_inv %*% y),
        sd = sqrt(as.numeric(Kstarstar - t(Kstar) %*% K_y_inv %*% Kstar))
  )
# draw enough samples to get a feeling for the distribution
samples_posterior <-</pre>
       draw_from_pred_posterior(number_samples = 1000, sigmaSq = bestSigmaSq$root,
                                 y = y, x = x, xstar = 0)
# plot the distribution
hist(samples_posterior, breaks=50, xlab=expression(y["*"]^b))
```

Histogram of samples_posterior

