Advanced Machine Learning, Assignment 1

Youssef Taoudi, yousseft@kth.se

December 24, 2019

1 The Prior

1.1 Theory

1.1.1 Question 1

Insert statement about Gaussian form of likelihood
The assumption we make about the data by choosing a spherical covariance matrix is that the covariance becomes a diagonal matrix, meaning all data points are assumed to be independent. The assumption also means that all features will share the same variance.

1.1.2 Question 2

The likelihood will be a multivariate normal distribution with a mean vector as a collection of the function of all data points of the distribution. If the data points are not assumed to be independent, then the covariance will be used instead of the variance in a joint distribution.

$$P(\mathbf{T}|f,\mathbf{X}) \sim \mathcal{N}\left(\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \dots & \Sigma_{1n} \\ \vdots & & \\ \Sigma_{n1} & \dots & \Sigma_{nn} \end{bmatrix}\right) \tag{1}$$

1.1.3 Question 3

$$p(\mathbf{T}|\mathbf{X},\mathbf{W}) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{T}|f(\mathbf{x}_i,\mathbf{W}),\beta^{-1}) \tag{2}$$

Equation 2 shows the probability $p(\mathbf{T}|\mathbf{X},\mathbf{W})$ as a normal distribution with the mean $\mathbf{T}|f(\mathbf{x}_i,\mathbf{W})$ and variance β. Where t is the target value that is conditioned on the deterministic function $f(\mathbf{x}_i,\mathbf{W}) = \mathbf{W}\mathbf{x}_i$ and β is the precision (inverse of the variance) for the error function ϵ [1, p. 142]. In our case, the error function has precision $\beta = (\sigma^2\mathbf{I})^{-1}$. The inverse precision β^{-1} can thus written as $(\sigma^2\mathbf{I})$ which gives us the Gaussian in equation 3. Note that the product

of the probabilities hold true only because the data samples are assumed to be conditionally independent.

$$p(\mathbf{T}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{T}|\mathbf{W}\mathbf{x}_i, \sigma^2\mathbf{I}) \tag{3}$$

1.1.4 Question 4

Regularizing with L1 norm is known as the lasso. One property of the lasso is that as the shrinkage term λ increases, an increasing number of coefficients w_i are driven to zero which leads to a sparse model where some features will not be used for predictions [1, p. 145].Regularization terms are used to reduce complexity on models that are trained on limited data [1, p. 145] which will lead to a smoother distribution.

The distribution of the prior is shown in equation 4 as a normal distribution [3] where the denominator is a constant value for all vectors $\mathbf{w}$.

$$p(\mathbf{w}) = \frac{\exp -\frac{1}{2\tau^2}(\mathbf{w} - \mathbf{w}_0)(\mathbf{w} - \mathbf{w}_0)^T}{C} \tag{4}$$

The negative logarithmic prior of one row takes the form of equation 5 derived from the logarithm of the normal distribution of the prior in equation 4.

$$\ln p(\mathbf{w}) = \frac{1}{2\tau^2}(\mathbf{w} - \mathbf{w}_0)(\mathbf{w} - \mathbf{w}_0)^T + \ln C \tag{5}$$

Equation 6 and 7 derived by Bishop [1, p. 153] represent the penalization terms for L_1 and L_2 regularization respectively. In our case, $\lambda = \frac{1}{\tau^2}$ and $\mathbf{w}$ is one row in our $\mathbf{W}$ matrix. Naturally, as L_2 takes a very similar form as our log prior, an L_2 regularization would be more suitable for the problem.

$$\frac{\lambda}{2}\sum_{j=1}^{D}|w_j| \tag{6}$$

$$\frac{\lambda}{2}\sum_{j=1}^{D}|w_j|^2 = \frac{\lambda}{2}\mathbf{w}\mathbf{w}^T \tag{7}$$

Figure 1 illustrates the general contour forms in two dimensions when using L_1 and L_2 regularization respectively.

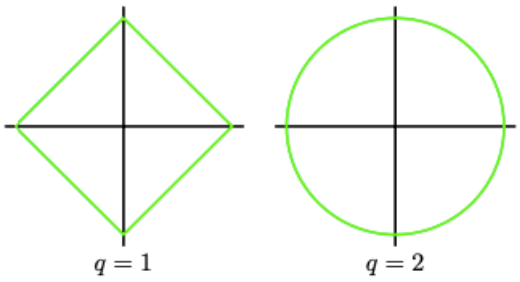


Figure 1: The figure illustrates the contours of the regularization terms where q=1 (L_1) on the left and q=2 (L_2) on the right. Image taken from Bishop[1, p. 145].

1.1.5 Question 5

Our posterior for each row in matrix W_i will be a Gaussian distribution in the form 8. Because output points are assumed independent, we can formulate a normal distribution over each output feature by dividing the W matrix in rows and the T matrix in columns where $\boldsymbol{W_i}$ is one row in $\mathbf{W}$ and $\boldsymbol{T_i}$ is one column in $\mathbf{T}$.

$$p(\boldsymbol{W_i}|\mathbf{X}, \boldsymbol{T_i}) = \mathcal{N}(\boldsymbol{\mu_w}, \boldsymbol{\Sigma_w}) \tag{8}$$

The posterior 8 is made up of three factors: the prior, the likelihood and the evidence shown in expression 10. The first exponent is quadratic in terms of $\boldsymbol{W_i}$, the second exponent is linear and the third exponent is constant.

$$p(\boldsymbol{W_i}|\boldsymbol{X}, \boldsymbol{T_i}) \propto P(\boldsymbol{T_i}|\boldsymbol{W_i}, \boldsymbol{X}) \times P(\boldsymbol{W_i}) = \exp\left(-\frac{1}{2}\boldsymbol{W_i}\boldsymbol{\Sigma_w^{-1}}\boldsymbol{W_i^T}\right) \times \exp\left(\boldsymbol{W_i}\boldsymbol{\Sigma_w^{-1}}\boldsymbol{\mu_w}\right) \tag{9}$$

Equation 9 can be written in terms of one quadratic, one linear and one constant exponent in terms of W as shown in equation 10.

$$p(\boldsymbol{W_i}|\boldsymbol{X}, \boldsymbol{T_i}) \propto \exp\left(-\frac{1}{2}\boldsymbol{W_i}\boldsymbol{\Sigma_w^{-1}}\boldsymbol{W_i^T}\right) \times \exp\left(\boldsymbol{W_i}\boldsymbol{\Sigma_w^{-1}}\boldsymbol{\mu_w}\right) \times \exp\left(-\frac{1}{2}\boldsymbol{\mu_w}\boldsymbol{\Sigma_w^{-1}}\boldsymbol{\mu_w}\right) \tag{10}$$

Our task is thus to match posterior 10 with our prior 4 and likelihood 1. The evidence can be ignored for now as it is constant and does not need to be used to find the parameters $\boldsymbol{\mu_w}$ and $\boldsymbol{\Sigma_w}$. If we only focus on the exponents, we can write a tidy expression as shown in 11

$$-\frac{1}{2\sigma^2}(\boldsymbol{T_i} - \boldsymbol{X}\boldsymbol{W_i^T})^T(\boldsymbol{T_i} - \boldsymbol{X}\boldsymbol{W_i^T}) - \frac{1}{2}\boldsymbol{W_i}\Sigma^{-1}\boldsymbol{W_i^T} \tag{11}$$

Expression 11 can be written into constant, linear and quadratic terms in the perspective of $\boldsymbol{W_i}$ as shown in 12 where the first term is constant, the second term is linear and the last two terms are quadratic.

$$-\frac{1}{2\sigma^2}\boldsymbol{T_i^T T_i} + \frac{1}{\sigma^2}\boldsymbol{T_i^T X W_i^T} - \frac{1}{2\sigma^2}(\boldsymbol{X W_i^T})^T(\boldsymbol{X W_i^T}) - \frac{1}{2}\boldsymbol{W_i}\Sigma^{-1}\boldsymbol{W_i^T} \tag{12}$$

The quadratic terms in 12 can be used to complete the square by setting them equal to the quadratic exponent in 10.

$$-\frac{1}{2\sigma^2}(\boldsymbol{X W_i^T})^T(\boldsymbol{X W_i^T}) - \frac{1}{2}\boldsymbol{W_i}\Sigma^{-1}\boldsymbol{W_i^T} = -\frac{1}{2}\boldsymbol{W_i}\boldsymbol{\Sigma_w^{-1}}\boldsymbol{W_i^T} \tag{13}$$

The expression on the left can be re-written to the same form as the right side expression.

$$-\frac{1}{2}\boldsymbol{W_i}(\frac{1}{\sigma^2}\boldsymbol{X^T X} + \boldsymbol{\Sigma^{-1}})\boldsymbol{W_i^T} = -\frac{1}{2}\boldsymbol{W_i}\boldsymbol{\Sigma_w^{-1}}\boldsymbol{W_i^T} \tag{14}$$

The covariance $\boldsymbol{\Sigma_w}$ can be identified easily through 14.

$$\boldsymbol{\Sigma_w} = (\frac{1}{\sigma^2}\boldsymbol{X^T X} + \boldsymbol{\Sigma^{-1}})^{-1} \tag{15}$$

To find the mean $\boldsymbol{\mu_w}$, we have to look at the linear terms of equation 12 and the exponents in equation 10.

$$\frac{1}{\sigma^2}\boldsymbol{T_i^T X W_i^T} = (\boldsymbol{W_i}\boldsymbol{\Sigma_w^{-1}}\boldsymbol{\mu_w}) \tag{16}$$

Once again, the left side expression can be tweaked to make the calculation easier.

$$\frac{1}{\sigma^2}\boldsymbol{W_i^T X^T T_i} = (\boldsymbol{W_i}\boldsymbol{\Sigma_w^{-1}}\boldsymbol{\mu_w}) \tag{17}$$

Since we know $\boldsymbol{\Sigma_w}$, it can be substituted into equation 17.

$$\frac{1}{\sigma^2}\boldsymbol{W_i^T X^T T_i} = (\boldsymbol{W_i}(\frac{\boldsymbol{1}}{\boldsymbol{\sigma^2}}\boldsymbol{X^T X} + \boldsymbol{\Sigma^{-1}})\boldsymbol{\mu_w}) \tag{18}$$

The mean $\boldsymbol{\mu_w}$ can thus be rewritten as shown in expression 19.

$$\boldsymbol{\mu_w} = \frac{1}{\sigma^2}(\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{X}^T\boldsymbol{T_i} \tag{19}$$

Since we now know the parameters for the distribution, we can now express the posterior for each output feature i as a multivariate gaussian as shown in 20. Note that the covariance $\boldsymbol{\Sigma} = \tau^2\boldsymbol{I}$

$$p(\boldsymbol{W_i}|\boldsymbol{X}, \boldsymbol{T_i}) = \mathcal{N}(\frac{1}{\sigma^2}(\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X} + \frac{1}{\tau^2}\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{T_i}, (\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X} + \frac{1}{\tau^2}\boldsymbol{I})^{-1}) \tag{20}$$

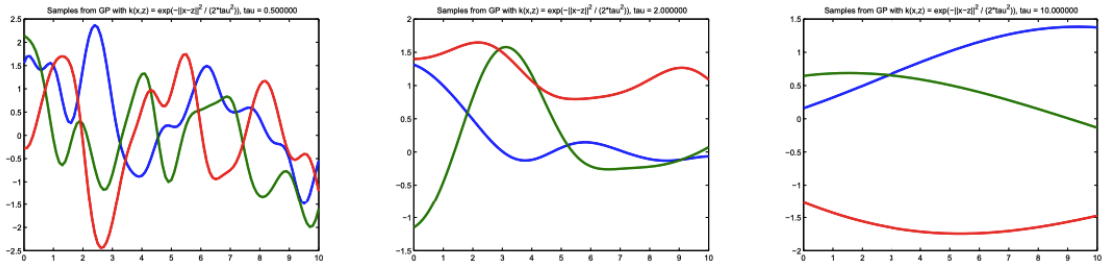
Note that the constant term corresponds to the evidence (Z) and is not interesting for finding the gaussian distributions for the posteriors.

Looking at 11, it is clear that we can identify the log-posterior as a least square likelihood with L2-regularization as prior.

1.1.6 Question 6

The prior must be a gaussian distribution in order for it to be conjugate to the posterior gaussian process. The gaussian prior $p(f|\boldsymbol{X}, \boldsymbol{\theta})$ with a covariance function is chosen to express the property that the values for similar points will be more strongly correlated than for dissimilar points [1, p. 306].

The reason for using a covariance function (kernel) is because it can be used to adjust smoothing in the distribution. Figure 2 illustrates how smoothness can be adjusted using a squared exponential kernel $k(x, x') = \exp\left(-\frac{1}{2\tau^2}\left\|x - x'\right\|^2\right)$ for different values of τ[2].



(a) (b) (c)

Figure 2: The figure illustrates samples from three different gaussian process priors with squared exponential covariance functions. a) has $\tau = 0.5$, b) has $\tau = 2$ and c) has $\tau = 10$. Image taken from Chuong B. Do[2].

1.1.7 Question 7

Figure 21 is the formula for the joint probability of the four variables. The graphical model 3 displays how the variables are related. f is conditioned on $\boldsymbol{X}$ and θ and $\boldsymbol{T}$ is in turn conditioned on f.

$$p(\mathbf{t}, \boldsymbol{f}, \mathbf{X}, \theta) = p(\mathbf{t}|\boldsymbol{f})p(\boldsymbol{f}|\mathbf{X}, \theta)p(\mathbf{X})p(\theta) \tag{21}$$

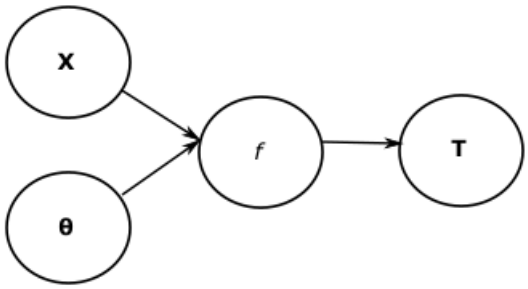


Figure 3: Graphical model displaying the relationships between the variables

1.1.8 Question 8

$$p(\mathbf{T}|\mathbf{X}, \theta) = \int p(\mathbf{T}|\boldsymbol{f})p(\boldsymbol{f}|\mathbf{X}, \theta)df \tag{22}$$

The prior gives all possible functions $\boldsymbol{f}$ generated by data points from the input space $\boldsymbol{X}$ and hyperparameters θ. The resulting functions from the prior ($\boldsymbol{f}$) are then conditioned on by the output points $\boldsymbol{T}$ in the likelihood to find the posterior.

There are uncertainties in our data because of the noise (ϵ) and in our functions ($\boldsymbol{f}$). Because they are Gaussian sources of uncertainty, their covariances will simply be added and thus pass through the marginalisation [1, p. 307].

The θ is still needed because it is a constant in our covariance function which is needed to calculate the probabilities inside the integral.

1.2 Practical

1.2.1 Question 9

The prior over W is displayed in figure 4.

It is clear from plots 5 and 6 that the more data points that are used for the model, the more certain it will be in predictions. For example, in the plots 5 the uncertainty of the posterior becomes smaller the more datapoints are used in training (smaller circles in the graph). The same can be said for the lineplots 6, the more datapoints that are used the closer (on average) we get to the true values of $w_0 = 0.5$ and $w_1 = -1.5$. Varying the noise

levels did not affect the uncertainty (variance) of the posterior but it did change the average values (center) of w_0 and w_1 especially for lower amounts of data points.

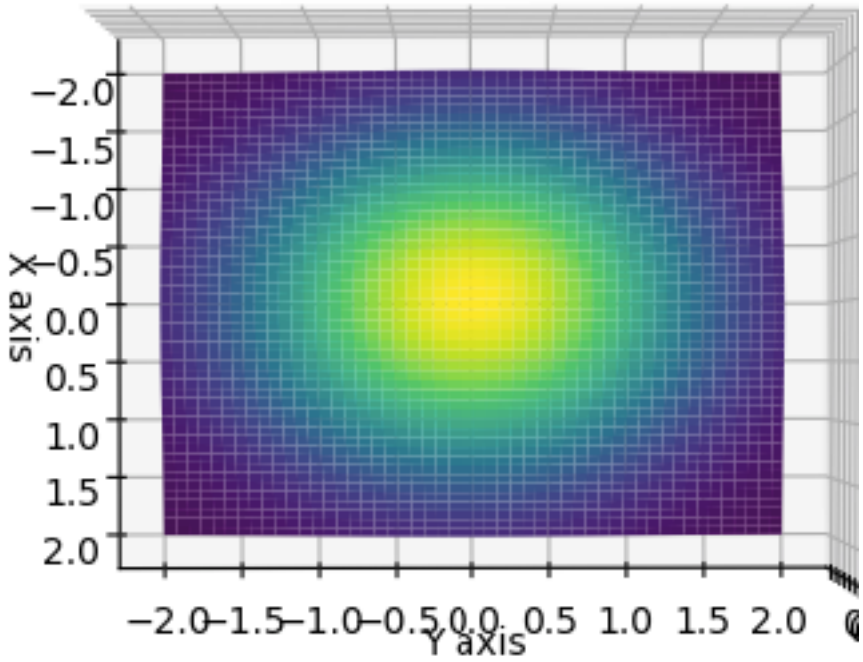


Figure 4: Vizualization of the prior over the $\boldsymbol{W}$. The mean is assumed to be 0 and the variance is assumed to be 1.

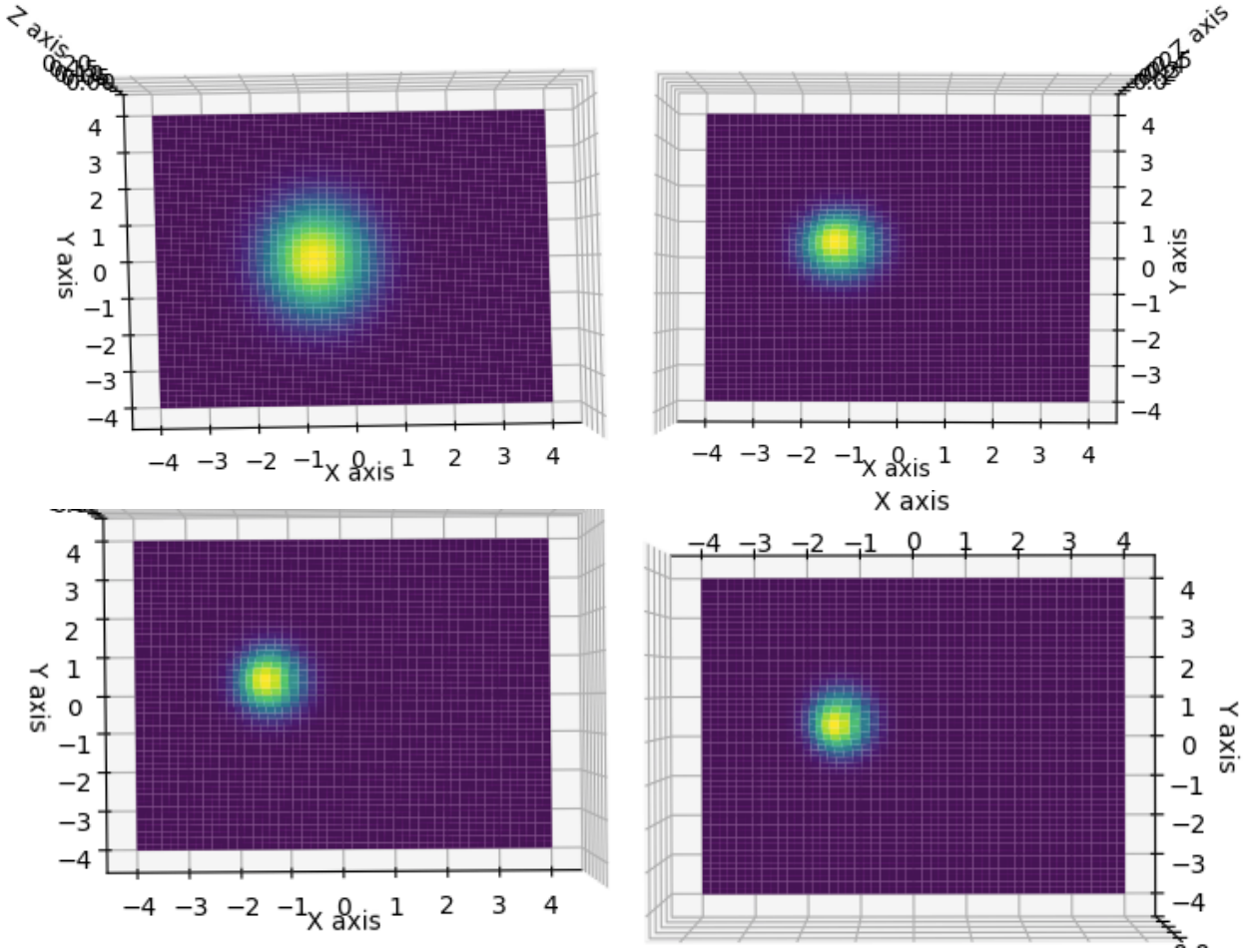


Figure 5: Posterior distributions over W for up to 7 different datapoints. The top-left is for 1 datapoint, top-right for 3 datapoints, bottom-left for 5 datapoints and bottom-right for 7 datapoints. Note that the Y-axis corresponds to w_0 and the X-axis to w_1

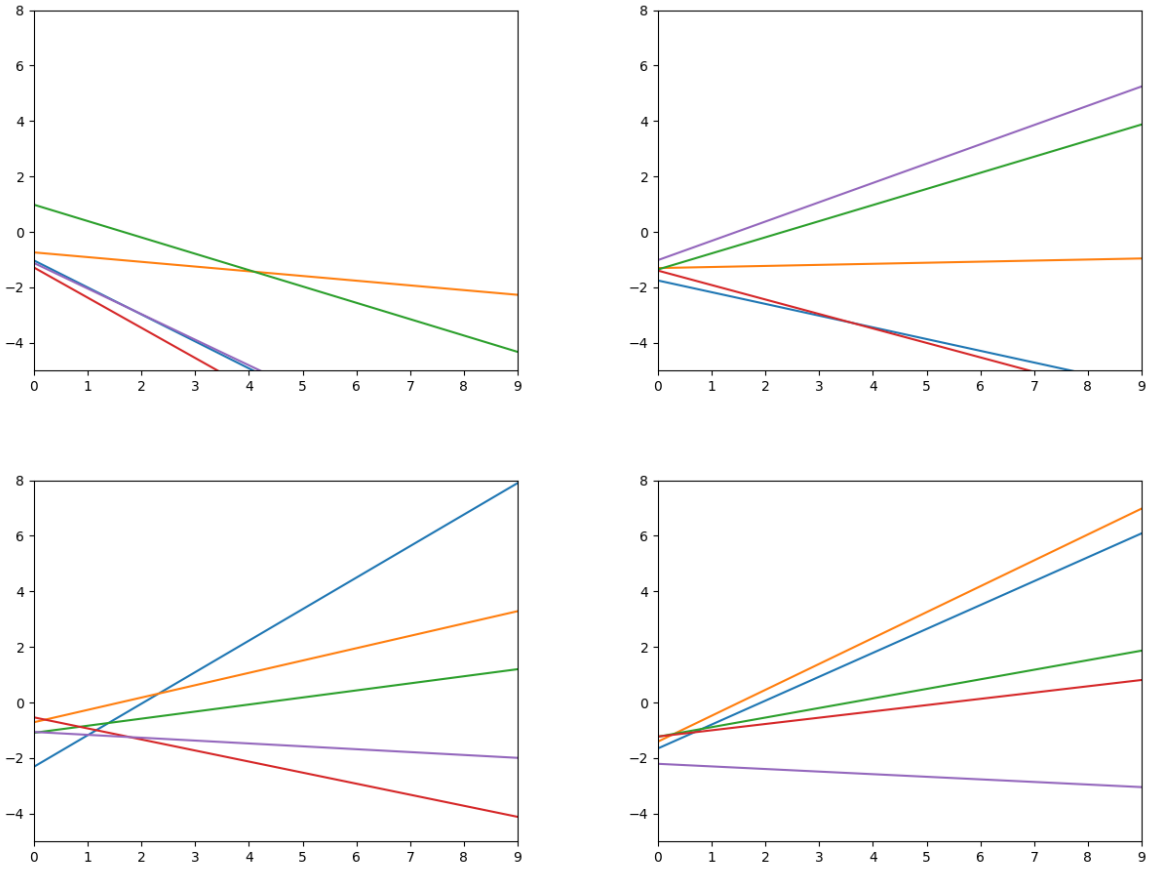


Figure 6: 5 Sample lines drawn from different amount datapoints. Top left is sampled from a distribution with 1 datapoint, top-right from 3 datapoints, botom-left from 5 datapoints and bottom-right from 7 datapoints. Note that the intersection in the vertical axis when the horisontal axis equals 0 is the predicted value of w_1 while the slope of the line is the predicted value of w_0

1.2.2 Question 10

It is clear from the plots in figure 7 that the length scale adjusts the smoothness of the function curve. For a high value (for example top-left image in 7), the lines will be fairly flat. For lower values on the length scale on the other hand, the plots will be more complex.

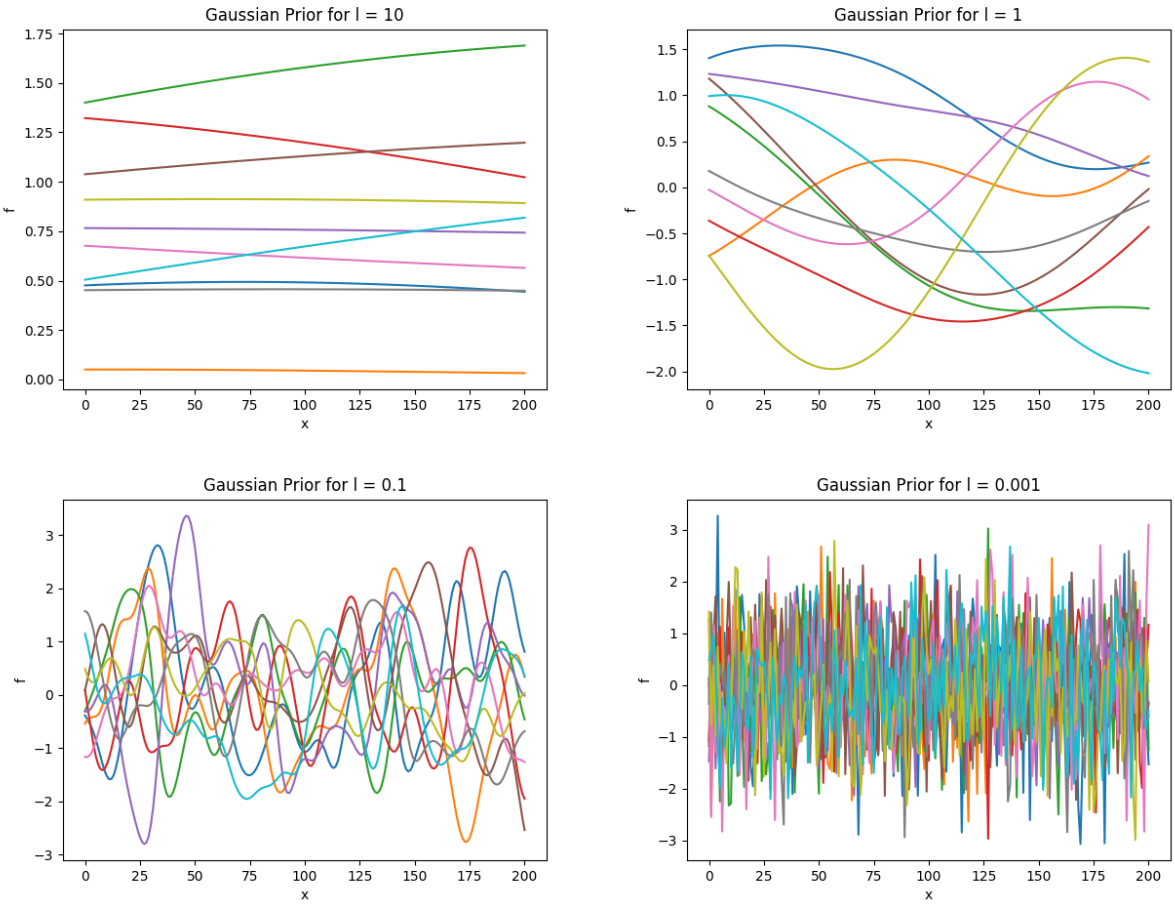


Figure 7: The prior from 10 samples for 4 different values of the length scale l.

1.2.3 Question 11

$$p(\mathbf{f}^*|\boldsymbol{X}^*, \boldsymbol{X}, \mathbf{f}) \sim \mathcal{N}\left(k\left(\mathbf{X}^*, \mathbf{X}\right)^{\mathrm{T}} K(\mathbf{X}, \mathbf{X})^{-1} \boldsymbol{f}, k\left(\mathbf{X}^*, \mathbf{X}^*\right) - k\left(\mathbf{X}^*, \mathbf{X}\right)^{\mathrm{T}} K(\mathbf{X}, \mathbf{X})^{-1} K\left(\mathbf{X}, \mathbf{X}^*\right)\right) \tag{23}$$

The predictive posterior is a gaussian distribution as shown in equation 23 [5, p. 16]. X is the training data, X^* is the data that is to be predicted and f are the functions of the training data. If we were to have no training data available, the only term that would remain would be $k\left(\mathbf{X}^*, \mathbf{X}^*\right)$. Thus the equation can be re-written as 24 which is the same as the gaussian prior (as in the assignment).

$$p(\mathbf{f}^*|\boldsymbol{X}^*, \boldsymbol{X}, \mathbf{f}) \sim \mathcal{N}\left(0, k\left(\mathbf{X}^*, \mathbf{X}^*\right)\right) \tag{24}$$

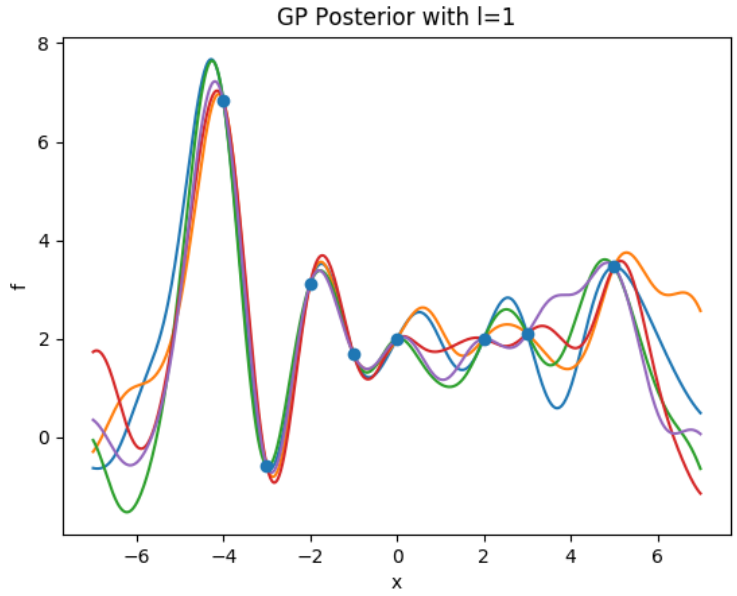


Figure 8: GP Posterior over 1400 observations of x^* where $\{x^* \in \mathbf{Q} | \ -7 \leq x \leq 7\}$. The blue dots are the true values of $\boldsymbol{f}$ for the training data $\boldsymbol{x}$.

In figure 8 you can see the predictive distribution of the posterior where each line is a sample of f. Note that all lines at the true values of the observed data $\boldsymbol{x}$.

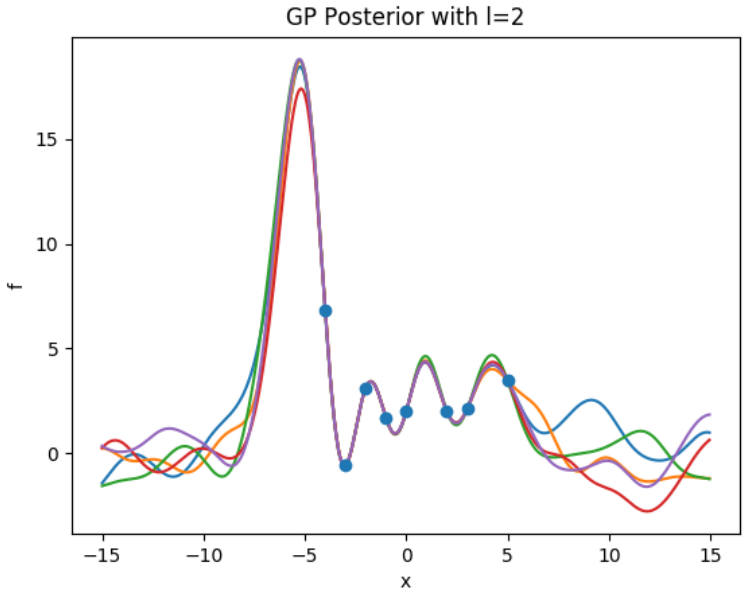


Figure 9: GP Posterior over the x-axis where $\{x^* \in \mathbf{Q} | \ -15 \leq x \leq 15\}$. The blue dots are the true values of $\boldsymbol{f}$ for the training data $\boldsymbol{x}$.

In figure 9 it is clear that the uncertainty grows the further away from the observed data while points close to the observed data generally has lower variance.

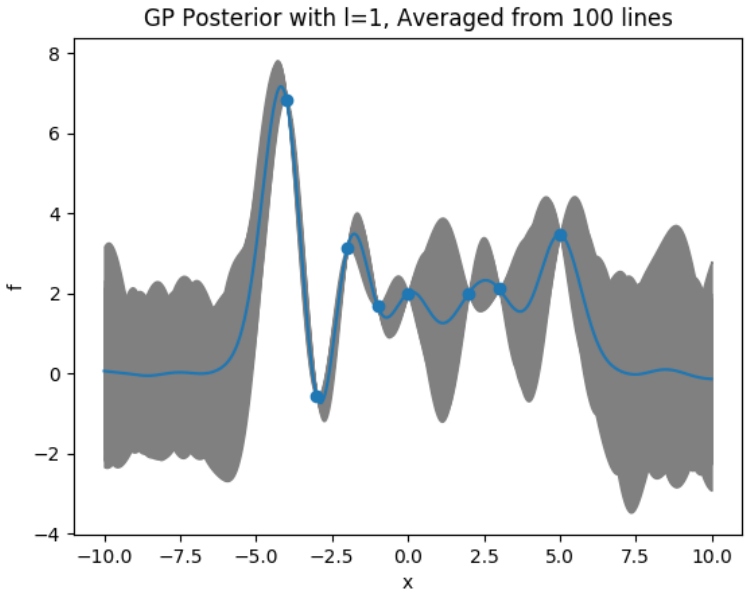


Figure 10: The blue line is the averaged posterior over 100 different samples from many observations x^* where $\{x^* \in \mathbf{Q} | \ -10 \leq x \leq 10\}$. The shaded areas is the predictive variance of the posterior over the 100 lines.

Comparing all posterior plots to the prior plot 7 the main difference is that the posterior is fit after our labeled data $\boldsymbol{t}$ whereas the prior has no knowledge of the data. This is a desirable effect as it leads to better predictions, especially for points close to the training data.

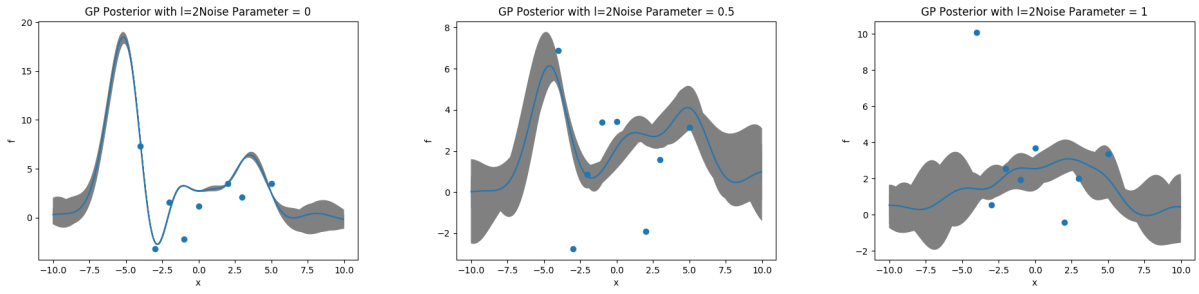


Figure 11: The average predictive posterior distribution with increasing values for the noise variance (σ_n^2) going from left to right.

Figure 11 shows how noisy data is predicted using different values for the noise variance[5] on the diagonal covariance matrix. Higher values for the noise variance lead to higher

uncertainties (variance) for f and it also leads to a smoother function of the predictive mean.

2 The Posterior

2.1 Theory

2.1.1 Question 12

The preferred prior is defined as a Gaussian because of conjunction, with independence over all latent variables $\boldsymbol{X}$ which are properties that simplifies our derivations and calculations.

2.1.2 Question 13

The prior as given in the assignment 25.

$$p(\boldsymbol{X}) = \mathcal{N}(0, \boldsymbol{I}) \tag{25}$$

We want to calculate the marginal likelihood 26.

$$p(\boldsymbol{Y}|\boldsymbol{W}) = \int p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{W}) p(\boldsymbol{X}) \boldsymbol{d}\mathrm{X} \tag{26}$$

Due to the independence in $\mathbf{X}$ and $\mathbf{Y}$, the marginal posterior for the entire $\mathbf{Y}$ can be written as a product through Naive Bayes:

$$p(\boldsymbol{Y}|\boldsymbol{W}) = \prod p(\boldsymbol{Y_i}|\boldsymbol{W}) \tag{27}$$

Thus, we can simply calculate the posterior over each row of $\mathbf{Y}$ by using the likelihood 28 as given from Bishop[1, p. 571] where $\mathbf{Y_i}$ is a single output vector and $\mathbf{X_i}$ is corresponding $\mathbf{X}$.

$$p(\boldsymbol{Y_i}|\boldsymbol{X_i}) = \mathcal{N}\left(\boldsymbol{Y_i}|\boldsymbol{W}\,\boldsymbol{X_i} + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}\right) \tag{28}$$

Since we know that the marginal is Gaussian, we can derive the mean and covariance from every instance of $p(\boldsymbol{Y_i}|\boldsymbol{X_i})$ defined in equation 28[1, p. 573].

$$\begin{aligned} E[\mathbf{Y_i}] &= E[\mathbf{W}\mathbf{X_i} + \boldsymbol{\mu} + \boldsymbol{\epsilon}] = \boldsymbol{\mu} \\ \mathrm{cov}[\mathbf{Y_i}] &= E\left[(\mathbf{W}\mathbf{X_i} + \boldsymbol{\epsilon})(\mathbf{W}\mathbf{X_i} + \boldsymbol{\epsilon})^{\mathrm{T}}\right] \\ &= E\left[\mathbf{W}\mathbf{X_i}^T\mathbf{X_i}\mathbf{W}^{\mathrm{T}}\right] + E\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}\right] = \mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I} \end{aligned} \tag{29}$$

Following 29, we may express our Gaussian as:

$$p(\boldsymbol{Y_i}|\boldsymbol{W}) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}) \tag{30}$$

2.1.3 Question 14

We know that maximizing the likelihood is the same as minimizing the sum of square error in the log probability [1, p. 23]:

$$\ln p\left(\mathbf{Y}|\mathbf{W},\mathbf{X}\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(\mathbf{Y_n} - \mathbf{W}\mathbf{X_n}\right)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi) \tag{31}$$

MAP on the other hand is a means of finding the parameters that mazimise the posterior. In this case, the log-prior is added onto the maximum likelihood as a constraint[1, p. 30]:

$$\ln p\left(\mathbf{Y}|\mathbf{W},\mathbf{X}\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(\mathbf{Y_n} - \mathbf{W}\mathbf{X_n}\right)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi) - \frac{1}{2\sigma_w}\mathbf{W}\mathbf{W}^{\mathrm{T}} \tag{32}$$

Note that for ML, there is no prior knowledge about the observations, so far smaller number of observations, there can occur overfitting. This in means that ML is more useful when a high amount of data has been observed and less useful for low amount of observed data. MAP on the other hand has a regularization term that counteracts the problem of overfitting making it useful even for low amounts of observed data [1, p. 29]. Also worth noting is that MAP will share the same behaviour as ML when sufficient data has been observed (i.e when the observed data is very large). For type 2 maximum likelihood, the observed variables are marginilised out, meaning there is no need for observed data as long as there is some belief of the prior distribution of the data. Thus there is no need for actual data when using Type 2 ML.

$$\hat{\mathbf{W}} = \mathrm{argmax}_{\mathbf{W}}\frac{p(\mathbf{Y}|\mathbf{X},\mathbf{W})p(\mathbf{W})}{\int p(\mathbf{Y}|\mathbf{X},\mathbf{W})p(\mathbf{W})\mathrm{d}\mathbf{W}} = \mathrm{argmax}_{\mathbf{W}}\, p(\mathbf{Y}|\mathbf{X},\mathbf{W})p(\mathbf{W}) \tag{33}$$

The integral can be re-written in terms of only X and Y.

$$\hat{\mathbf{W}} = \mathrm{argmax}_{\mathbf{W}}\frac{p(\mathbf{Y}|\mathbf{X},\mathbf{W})p(\mathbf{W})}{p(\mathbf{Y}|\mathbf{X})} = \mathrm{argmax}_{\mathbf{W}}\, p(\mathbf{Y}|\mathbf{X},\mathbf{W})p(\mathbf{W}) \tag{34}$$

Type 2 ML is the best suitable choice in our case because we have no observed data to use, but we have an assumption of how the data is distributed. Thus we can simply marginalize out the observed data.

2.1.4 Question 15

For convinience sake, we will call the covariance $\mathbf{C}$

$$p(\boldsymbol{Y}) = \mathcal{N}(\mu, \boldsymbol{C}) \tag{35}$$

This is of course the same covariance that we calculated in 30.

$$\mathbf{C} = \mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I} \tag{36}$$

The inverse of $\mathbf{C}$ is needed for the next step:

$$\mathbf{C}^{-1} = \sigma^{-1}\mathbf{I} - \sigma^{-2}\mathbf{W}(\mathbf{W}^{\mathrm{T}}\mathbf{W} + \sigma^{\mathbf{2}}\mathbf{T})^{-1}\mathbf{W}^{\mathrm{T}} \tag{37}$$

Now on to our log-posterior which of course is made up of the sum of all individual probabilities for $\mathbf{Y_i}$. Note that N is the amount of observations and D is the amount of features.

$$\ln p\left(\mathbf{Y}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2\right) = \sum_{n=1}^{N} \ln p\left(\mathbf{Y}_n|\mathbf{W}, \boldsymbol{\mu}, \sigma^2\right) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\mathbf{C}| - \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{Y}_n - \boldsymbol{\mu}\right)^{\mathrm{T}}\mathbf{C}^{-1}\left(\mathbf{Y}_n - \boldsymbol{\mu}\right) \tag{38}$$

$$\mathcal{L}(\mathbf{W}) = \frac{N}{2}\left\{D\ln(2\pi) + \ln|\mathbf{C}| + \mathrm{Tr}\left(\mathbf{C}^{-1}\frac{\mathbf{1}}{\mathbf{N}}\sum_{\mathbf{n=1}}^{\mathbf{N}}\left(\mathbf{Y_n} - \mu\right)\left(\mathbf{Y_n} - \mu\right)^{\mathrm{T}}\right)\right\} \tag{39}$$

For the gradient we follow rules derived from The Matrix Cookbook[4].

$$\begin{aligned} \partial(\det(X)) &= \mathrm{Tr}\left(X^{-1}\partial X\right) \\ \partial\,\mathrm{Tr}(X) &= \mathrm{Tr}(\partial X) \\ \partial X^{-1} &= -X^{-1}(\partial X)X^{-1} \end{aligned} \tag{40}$$

We can write the derivative of $\mathbf{C}$ in terms of $\mathbf{W}$:

$$\frac{\partial \mathbf{C}}{\partial W_{ij}} = \frac{\partial \mathbf{W}\mathbf{W}^T}{\partial W_{ij}} \tag{41}$$

We choose to solve the determinant and the trace separately. Starting with the determinant.

$$\frac{\partial(\det(\mathbf{C}))}{\partial W_{ij}} = Tr(\mathbf{C}^{-1}\frac{\partial \mathbf{C}}{\partial W_{ij}}) \tag{42}$$

Now for the trace, firstly the partial can be moved into the trace because of the rules in 40 meaning we want to solve the derivative inside the trace, note that constants with respect to W can be removed from the partial:

$$\frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{Y}_n - \mu\right)\left(\mathbf{Y}_n - \mu\right)^{\mathrm{T}}\frac{\partial(\mathbf{C}^{-1})}{\partial W_{ij}} \tag{43}$$

$$\frac{\partial(\mathbf{C}^{-1})}{\partial W_{ij}} = -\mathbf{C}^{-1}\frac{\partial \mathbf{C}}{\partial W_{ij}}\mathbf{C}^{-1} \tag{44}$$

Thus our final derivative with respects to W is:

$$\frac{\mathcal{L}(\boldsymbol{W})}{\partial W_{ij}} = \frac{N}{2}Tr(\mathbf{C}^{-1}\frac{\partial \mathbf{W}\mathbf{W}^{\mathbf{T}}}{\partial W_{ij}}) - Tr((-\mathbf{C}^{-1}\frac{\partial \mathbf{W}\mathbf{W}^{\mathbf{T}}}{\partial W_{ij}}\mathbf{C}^{-1})\frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{Y}_n - \mu\right)\left(\mathbf{Y}_n - \mu\right)^{\mathrm{T}}) \tag{45}$$

2.1.5 Question 16

References

[1] Christopher M. Bishop. *Pattern recognition and machine learning*. en. Information science and statistics. New York: Springer, 2006. ISBN: 978-0-387-31073-2.

[2] Chuong B. Do. *Gaussian Processes*. en. 2008. URL: http://cs229.stanford.edu/section/cs229-gaussian_processes (visited on 12/04/2019).

[3] *Multivariate normal distribution*. en. Page Version ID: 922344737. Oct. 2019. URL: https://en.wikipedia.org/w/index.php?title=Multivariate_normal_distribution&oldid=922344737 (visited on 11/18/2019).

[4] Kaare Brandt Petersen and Michael Syskind Pedersen. "[http://matrixcookbook.com]". en. In: (), p. 72.

[5] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. en. Adaptive computation and machine learning. OCLC: ocm61285753. Cambridge, Mass: MIT Press, 2006. ISBN: 978-0-262-18253-9.