DD2434 Machine Learning, Advanced Course - Assignment 1

Hannes Kindbom

1 The prior

1.1 Theory

Question 1

We often model the relationship between two single points $\boldsymbol{x_i} \in R^q$ and $\boldsymbol{t_i} \in R^D$ as $\boldsymbol{t_i} = f(\boldsymbol{x_i}) + \boldsymbol{\epsilon}$, where the noise $\boldsymbol{\epsilon}$ follows a Gaussian distribution.

One argument for assuming a Gaussian distribution on the noise term is that the Gaussian distribution has certain properties which makes it easy to work with. The Gaussian choice must however of course be appropriate and not off.

Since we often do not have information about the noise distribution, we can assume that it consists of a large sum of independent and identically distributed random variables (measurement errors) with finite variance. According to the central limit theorem, this large sum will converge in distribution to Gaussian. We do also assume that all $\boldsymbol{\epsilon}$'s and ϵ_i's are independent. The likelihood is a linear combination $(f(\boldsymbol{x_i}) + \boldsymbol{\epsilon})$ of the noise random variables conditioned on f and $\boldsymbol{x_i}$, which is thus also (multivariate) Gaussian. This is the rational argument behind the Gaussian assumption on the likelihood.
The components in $\boldsymbol{t_i}$ of this multivariate Gaussian likelihood are independent (since the noise terms are independent), which implies that the covariances between them are 0. Thus, the covariance matrix is proportional to the identity matrix. To be specific, this spherical covariance matrix comes from the independence assumption on the noise terms.

Question 2

If we assume that each output point is conditionally dependent given the input and the mapping, the likelihood of the data takes the following form:

$$p(\boldsymbol{T}|f, \boldsymbol{X}) = p(\boldsymbol{t_1}|\boldsymbol{t_{1+1}}, ..., \boldsymbol{t_N}, f, \boldsymbol{X}) \prod_{i=2}^{N} p(\boldsymbol{t_i}|f, \boldsymbol{X}) \tag{1}$$

where we have applied the product rule iteratively.

1.1.1 Linear Regression

Question 3

Suppose we assume $\boldsymbol{t_i} = \boldsymbol{W}\boldsymbol{x_i} + \boldsymbol{\epsilon}$, i = 1, .., N, where $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Then the likelihood of the

data becomes:

$$p(\boldsymbol{T}|\boldsymbol{X},\boldsymbol{W}) = \prod_{i=1}^{N} p(\boldsymbol{t_i}|\boldsymbol{x_i},\boldsymbol{W}) \tag{2}$$

where $\boldsymbol{t_i}|\boldsymbol{x_i},\boldsymbol{W} \sim N(\boldsymbol{W}\boldsymbol{x_i},\sigma^2\boldsymbol{I})$. Note that we get a product on the right hand side of equation 2, since the random variables $\boldsymbol{t_i}$ in the joint distribution $\boldsymbol{T}|\boldsymbol{X},\boldsymbol{W}$ are conditionally independent given $\boldsymbol{x_i},\boldsymbol{W}$.

Question 4

The conjugate prior for the Gaussian likelihood is a Gaussian prior. A Gaussian prior over each row of $\boldsymbol{W}$ on the form $p(\boldsymbol{w}) = N(\boldsymbol{w_0},\tau^2 I)$ will thus lead to a Gaussian posterior with the term $-\frac{1}{2\tau^2}(\boldsymbol{w}-\boldsymbol{w_0})^T(\boldsymbol{w}-\boldsymbol{w_0})$ being part of the exponent, since *posterior* $\propto$ *likelihood * prior*. When searching for the MAP (maximum aposteriori estimator), the posterior can be logarithmised. Thereafter maximizing with respect to $\boldsymbol{w}$ is equivalent to minimization of a sum of squares error function with a quadratic (L_2 distance) penalization term: $\frac{1}{2\tau^2}(\boldsymbol{w}-\boldsymbol{w_0})^T(\boldsymbol{w}-\boldsymbol{w_0})$ (this is the negative log-prior), where $\frac{1}{2\tau^2}$ can be seen as the regularization coefficient. Regularization is used to reduce variance and overfitting. If $\boldsymbol{w_0} = 0$, then the estimated weights $\boldsymbol{w_{MAP}}$ will be encouraged to tend towards 0, if not supported by the data.

If we instead of a Gaussian prior, choose a Laplace prior, the preferred model would be encoded with L_1 norm for model parameters. The one dimensional version of a Laplace distribution may be written as $p(w) = \frac{1}{2b}\exp(-\frac{|w-w_0|}{b})$, where $w_0 \in R$ is the mean and $b > 0$ is a scaling coefficient. The penalization term (part of the log posterior) in this case thus becomes $\frac{|w-w_0|}{b}$.

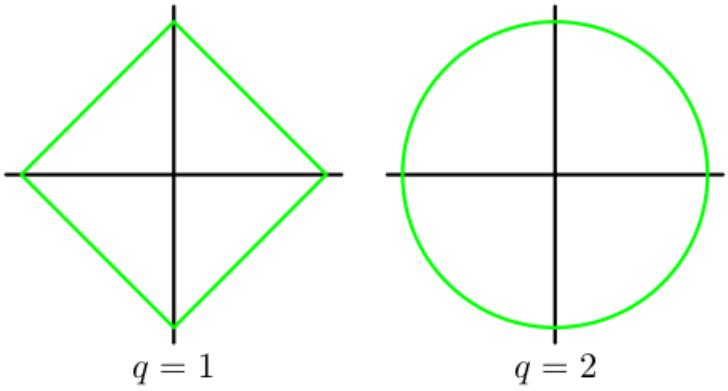

Figure 1: Contours of two different regularization terms when $\boldsymbol{w}$ is two dimensional. The left plot is for L_1 and the right is for L_2 regularization.

The difference between having L_1 and L_2 as penalizing terms is mainly that L_1 works more as a feature selector, forcing some components in $\boldsymbol{w}$ to be zero if the regularization coefficient is sufficiently large. The L_2 norm will generally not make any feature exactly zero, only small. This can

be seen from the contours of the regularization term in figure 1[1] or from the fact that the density in the Laplace distribution is highly concentrated and peaked at its mean, while the Gaussian density is more wide spread. If we set the mean of the prior to be zero, then the Laplace prior will force weights to zero, unless the likelihood indicates something else.

Question 5

The posterior when using a Gaussian prior and likelihood can be derived by using that $p(\boldsymbol{W}|\boldsymbol{X},\boldsymbol{T}) \propto p(\boldsymbol{T}|\boldsymbol{X},\boldsymbol{W})p(\boldsymbol{W})$. Gaussian is the conjugate prior to Gaussian, which means that the posterior will also be Gaussian. Let us firstly derive an expression for the likelihood, when assuming conditional independence of the target variables, as follows:

$$p(\boldsymbol{T}|\boldsymbol{X},\boldsymbol{W}) = \prod_{i=1}^{N} p(\boldsymbol{t_i}|\boldsymbol{x_i},\boldsymbol{W}) \quad \text{where} \quad \boldsymbol{t_i}|\boldsymbol{x_i},\boldsymbol{W} \sim N(\boldsymbol{W}\boldsymbol{x_i},\sigma^2\boldsymbol{I}) \tag{3}$$

Equation 3 can be written as:

$$p(\boldsymbol{T}|\boldsymbol{X},\boldsymbol{W}) = N(\boldsymbol{W}\boldsymbol{X},\sigma^2\boldsymbol{I}) \tag{4}$$

because a product of multiple Gaussian distributions is in turn Gaussian and each factor's covariance matrix is diagonal (spherical).
Let us now assume a Gaussian prior on the form $p(\boldsymbol{W}) = N(\boldsymbol{W_0},\tau^2\boldsymbol{I})$. Note thereafter that the probability density function of a multivariate Gaussian distribution with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{\Sigma}$ is proportional to $\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{m}))$. Hence, we can obtain an expression for the posterior:

$$p(\boldsymbol{W}|\boldsymbol{X},\boldsymbol{T}) \propto exp(-\frac{1}{2\sigma^2}(\boldsymbol{T}-\boldsymbol{W}\boldsymbol{X})^T(\boldsymbol{T}-\boldsymbol{W}\boldsymbol{X}))exp(-\frac{1}{2\tau^2}(\boldsymbol{W}-\boldsymbol{W_0})^T(\boldsymbol{W}-\boldsymbol{W_0})) \tag{5}$$

As mentioned above, this will result in a Gaussian posterior. Let us take a closer look on the exponent to figure out the mean and covariance of the posterior. Adding the exponents in equation 5 yields:

$$-\frac{1}{2\sigma^2}(\boldsymbol{T}-\boldsymbol{W}\boldsymbol{X})^T(\boldsymbol{T}-\boldsymbol{W}\boldsymbol{X}) - \frac{1}{2\tau^2}(\boldsymbol{W}-\boldsymbol{W_0})^T(\boldsymbol{W}-\boldsymbol{W_0}) \tag{6}$$

We often assume $\boldsymbol{W_0} = \boldsymbol{0}$. Then, expression 6 simplifies to:

$$-\frac{1}{2}[\frac{1}{\sigma^2}(\boldsymbol{T}^T\boldsymbol{T} - 2\boldsymbol{X}^T\boldsymbol{W}^T\boldsymbol{T}) + \boldsymbol{W}^T(\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X} + \frac{1}{\tau^2}\boldsymbol{I})\boldsymbol{W}] \tag{7}$$

We can now identify the covariance matrix Σ of the posterior immediately as:

$$\Sigma = (\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X} + \frac{1}{\tau^2}\boldsymbol{I})^{-1} \tag{8}$$

The term $-2\boldsymbol{X}^T\boldsymbol{W}^T\boldsymbol{T}\frac{1}{\sigma^2}$, can be identified as the term $-2\boldsymbol{W}^T\Sigma^{-1}\boldsymbol{m}$ obtained from expanding the general form of a multivariate Gaussian exponent $(\boldsymbol{W}-\boldsymbol{m})^T\Sigma^{-1}(\boldsymbol{W}-\boldsymbol{m})$. Rearranging a bit yields:

$$\boldsymbol{m} = \Sigma\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{T}\frac{1}{\sigma^2} = (\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X} + \frac{1}{\tau^2}\boldsymbol{I})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{T}\frac{1}{\sigma^2} \tag{9}$$

[1] C.M. Bishop. Pattern recognition and machine learning. 2nd ed. 2006. p.145

which is the desired mean of the posterior. We have thus found that the posterior is as follows:

$$p(\boldsymbol{W}|\boldsymbol{X},\boldsymbol{T}) \sim N(\boldsymbol{m},\Sigma) \tag{10}$$

Note that the least squares estimator $\boldsymbol{W_{LS}}$ of $\boldsymbol{W}$ is given by $\boldsymbol{W_{LS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{T}$ when $\boldsymbol{X}$ has full rank. When using L_2 regularization in ridge regression we get $\boldsymbol{W_{Ridge}} = (\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{T}$, where λ is a real constant. We can observe that the least squares estimate of $\boldsymbol{W}$ is somewhat similar to the mean $\boldsymbol{m}$ of the posterior but the ridge estimate is essentially on the same form. Since the Gaussian posterior is maximized at the mean value, $\boldsymbol{m}$ is the MAP of $\boldsymbol{W}$. Hence, we can say that Bayesian linear regression has incorporated regularization and deals with overfitting, which can otherwise be a problem in the frequentist's approach.

Note that throughout the derivation of the posterior, we have omitted the normalizing constant Z. I.e. the posterior is actually on the form $p(\boldsymbol{W}|\boldsymbol{X},\boldsymbol{T}) = \frac{1}{Z}p(\boldsymbol{T}|\boldsymbol{X},\boldsymbol{W})p(\boldsymbol{W})$. Z fills the purpose of making the posterior a probability distribution, which integral over the real line converges to 1. Z is furthermore often called the *evidence*, used for model selection.

1.1.2 Non-parametric Regression

Question 6
Assume we have the relationship $\boldsymbol{t_i} = \boldsymbol{f}(\boldsymbol{x_i}) + \boldsymbol{\epsilon}$, where $\boldsymbol{f}_i$ is considered as a random variable. When we take a non-parametric approach and assign $\boldsymbol{f}$ to be a Gaussian Process on the space of real valued functions on the input space, we put a prior on the space of functions instead of on some parameters. The problem of "manually" choosing an appropriate parametric form of the mapping is therefore avoided. Samples from a Gaussian process prior are illustrated in figure 8, where it can be observed that the curves do not follow a specified parametric mapping. We can formally define the Gaussian Process prior as:

$$p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{\theta}) = N(\boldsymbol{0}, k(\boldsymbol{X},\boldsymbol{X})) \tag{11}$$

where k is the covariance function and $\boldsymbol{\theta}$ the hyper-parameters. The joint distribution of any $\boldsymbol{f}_1(x_1),..,\boldsymbol{f}_n(x_n)$ is normally distributed per definition in a Gaussian Process. The marginal distribution $p(\boldsymbol{f})$ is also normally distributed, which motivates the choice of this prior; The mapping between X and T can results in any real value in the output space, which is possible with the normal distribution on $\boldsymbol{f}$.

Regarding the mean and covariance function, we often set the mean to zero in the prior due to lack of knowledge about $\boldsymbol{f}$. The covariance function can be chosen on the form, such that similar x_i's produce corresponding $\boldsymbol{f}_i$'s that are highly correlated and vice versa. This may be a desired behaviour. One such covariance function is the squared exponential covariance function. The hyper-parameters θ furthermore allow for customization of the covariance function to the specific dataset.

Question 7
The joint likelihood of the full model above can without making any assumptions (only applying the product rule) be formulated as follows:

$$p(\boldsymbol{T},\boldsymbol{X},\boldsymbol{f},\boldsymbol{\theta}) = p(\boldsymbol{T}|\boldsymbol{f},\boldsymbol{X},\boldsymbol{\theta})p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{\theta})p(\boldsymbol{X},\boldsymbol{\theta}) \tag{12}$$

Using the assumptions that $\boldsymbol{X}$ and $\boldsymbol{\theta}$ are independent and that $\boldsymbol{T}$ is conditionally independent of $\boldsymbol{X}$ and $\boldsymbol{\theta}$ given $\boldsymbol{f}$, equation 12 may be written as:

$$p(\boldsymbol{T}, \boldsymbol{X}, \boldsymbol{f}, \boldsymbol{\theta}) = p(\boldsymbol{T}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\theta})p(\boldsymbol{X})p(\boldsymbol{\theta}) \tag{13}$$

Figure 2 represents the right-hand side of (13) in terms of a simple graphical model (a so called Bayesian Network).

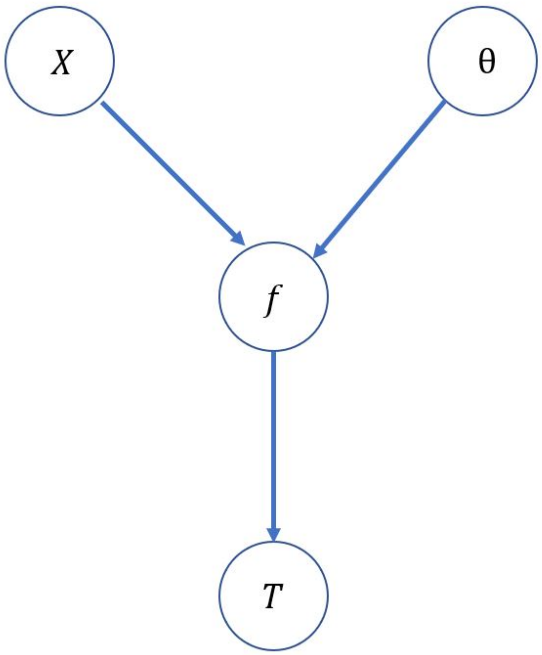

Figure 2: Bayesian Network for (13), where directed arrows are drawn from the nodes corresponding to the variables on which the distribution is conditioned.

Question 8

An expression for $p(\boldsymbol{T}|\boldsymbol{X}, \boldsymbol{\theta})$ is derived below, using the result from (13).

$$p(\boldsymbol{T}, \boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\theta}) = p(\boldsymbol{T}, \boldsymbol{X}, f, \boldsymbol{\theta})/(p(\boldsymbol{X}, \boldsymbol{\theta})) = p(\boldsymbol{T}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\theta}) \implies$$

$$p(\boldsymbol{T}|\boldsymbol{X}, \boldsymbol{\theta}) = \int p(\boldsymbol{T}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\theta})d\boldsymbol{f}$$

Here $\boldsymbol{f}$ is marginalized out since we are not interested in it directly. $p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\theta})$ in the integral above can be identified as the prior over instantiation of functions in (11). $p(\boldsymbol{T}|\boldsymbol{f})$ is the likelihood. The data in $\boldsymbol{T}$ and $\boldsymbol{X}$ is thus connected and the uncertainty filter through, since the uncertainty in $\boldsymbol{T}$ remains. $\boldsymbol{\theta}$ is held constant during the integration, wherefore we still condition on it after the marginalization.

1.2 Practical

1.2.1 Linear Regression

Question 9
A synthetic dataset according to (14) was first generated.

$$t_i = w_0 x_i + w_1 + \epsilon = 0.5x_i - 1.5 + \epsilon \quad , \textit{ where } \epsilon \sim N(0, \sigma^2) \textit{ and } \boldsymbol{x} = [-1, -099, ..., 099, 1] \tag{14}$$

The generated dataset for standard deviation on the noise $\sigma = 0.2$ is shown in figure 3. A Gaussian prior on $W = [w_0, w_1]$ was thereafter set to $W \sim N(\begin{bmatrix}0\\0\end{bmatrix}, \tau^2 I)$, where $\tau = 0.5$. The prior is illustrated in figure 4.

The posterior was thereafter computed, using the theory described under question 5 above. The results are displayed in figure 5. As expected, the posterior gets more and more concentrated around the true value of W, as more data is added to the likelihood. This is reasonable, since it gets increasingly influenced by the dataset (generated with the true W).

Figure 6 shows the resulting functions, after drawing random samples from the seven different posteriors. The graphs in 6 are ordered in the same way as in figure 5. Again as expected, the function lines get more and more concentrated around the true blue fat line as more data is added to the likelihood. It is especially difficult to get a reasonable function line with only one data point, as can be observed in the first graph.

The same experiment was therafter performed with $\sigma = 0.4$ as well as $\sigma = 0.8$. The results are displayed in figure 7. It may be observed that the increased variance σ is reflected in the variance of the posterior as well as in the sampled function lines. This is consistent with the theory and a higher variance in the data is intuitively reflected in our parameter estimates.

1.2.2 Non-parametric Regression

Question 10
10 samples per length scale l were drawn from a Gaussian Process prior with four different length scales in the squared exponential kernel. The results are shown in figure 8. The Gaussian Process prior is defined as in (11). The squared exponential kernel is defined as in (15).

$$k(x_i, x_j) = \sigma_f^2 \exp(-(x_i - x_j)^T (x_i - x_j)/l^2) \ , \ \sigma_f^2 = 4 \tag{15}$$

As can be observed in figure 8, $f(x_i)$ becomes increasingly smooth/stable as l increases. The explanation for this is that, as l increases, $k(x_i, x_j)$ increases for fixed signal variance σ_f^2. A larger $k(x_i, x_j)$, means that $f(x_i)$ and $f(x_j)$ will have a stronger correlation, hence the smoother graphs. The opposite holds for small l.

Question 11
Suppose $\boldsymbol{x_*}$ is a collection of new inputs $\boldsymbol{x_1^*, ..., x_m^*}$. Also, let $\boldsymbol{T}$ be a collection of training target variables $\boldsymbol{t_1, ..., t_n}$ and $\boldsymbol{X}$ a collection of training inputs $\boldsymbol{x_1, ..., x_n}$. If the same relationship as stated under question 6 is assumed, then the joint distribution between $\boldsymbol{T}$ and $\boldsymbol{f_*}$ is Gaussian as

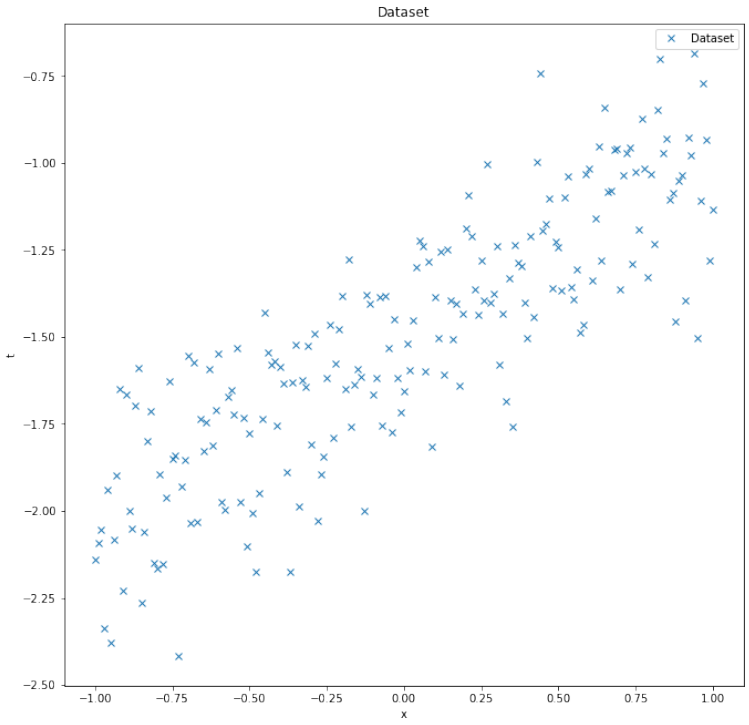

Figure 3: Generated dataset from equation 14 with $\sigma = 0.2$

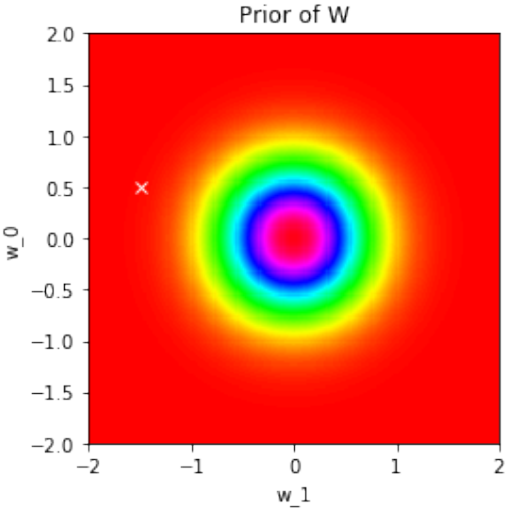

Figure 4: The prior on W, where the white X marks the true values of W.

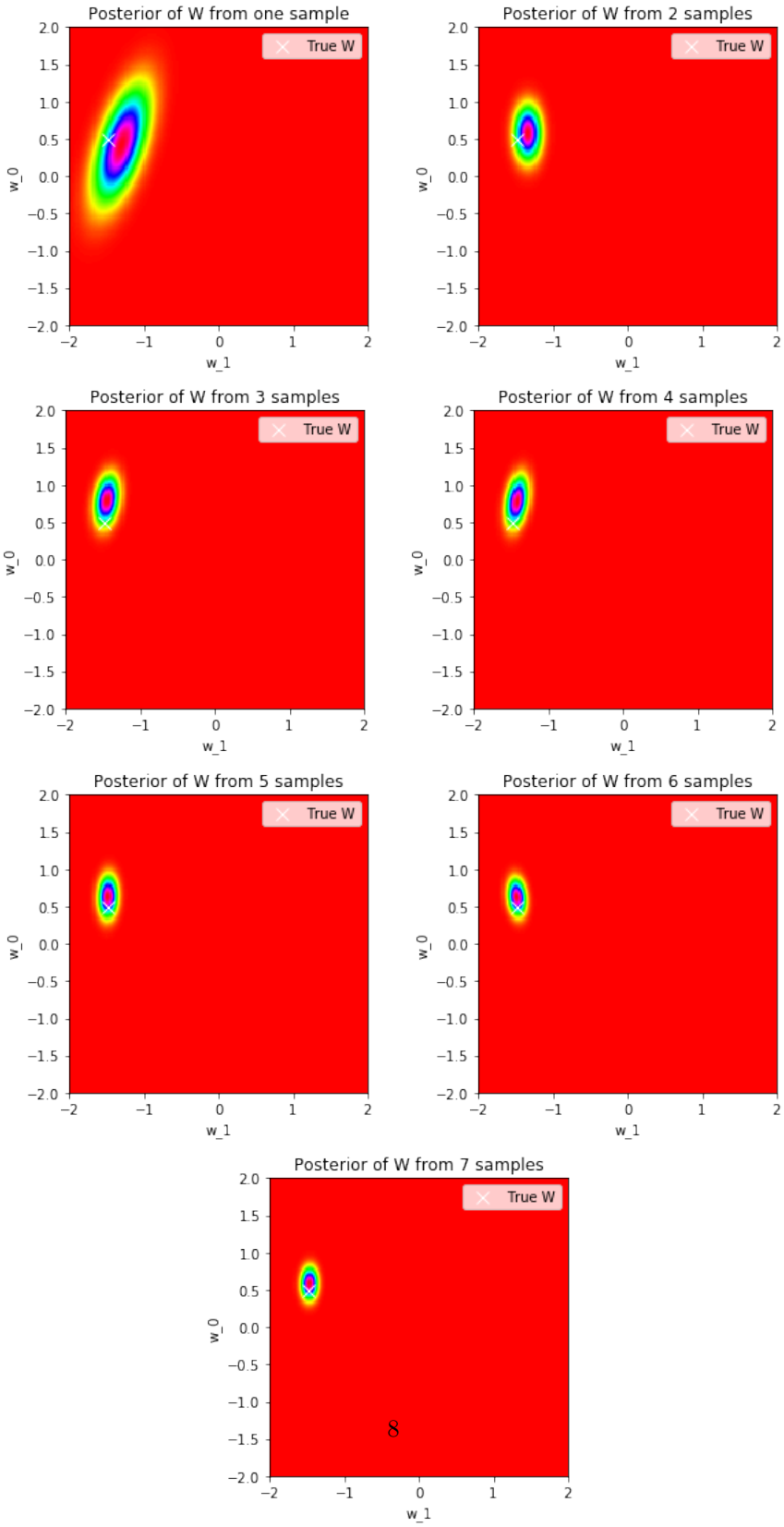

Figure 5: Posteriors for W, where random data points from the dataset have been added successively (starting at 1 ending at 7 data points). As before, the white cross marks the true values of W.

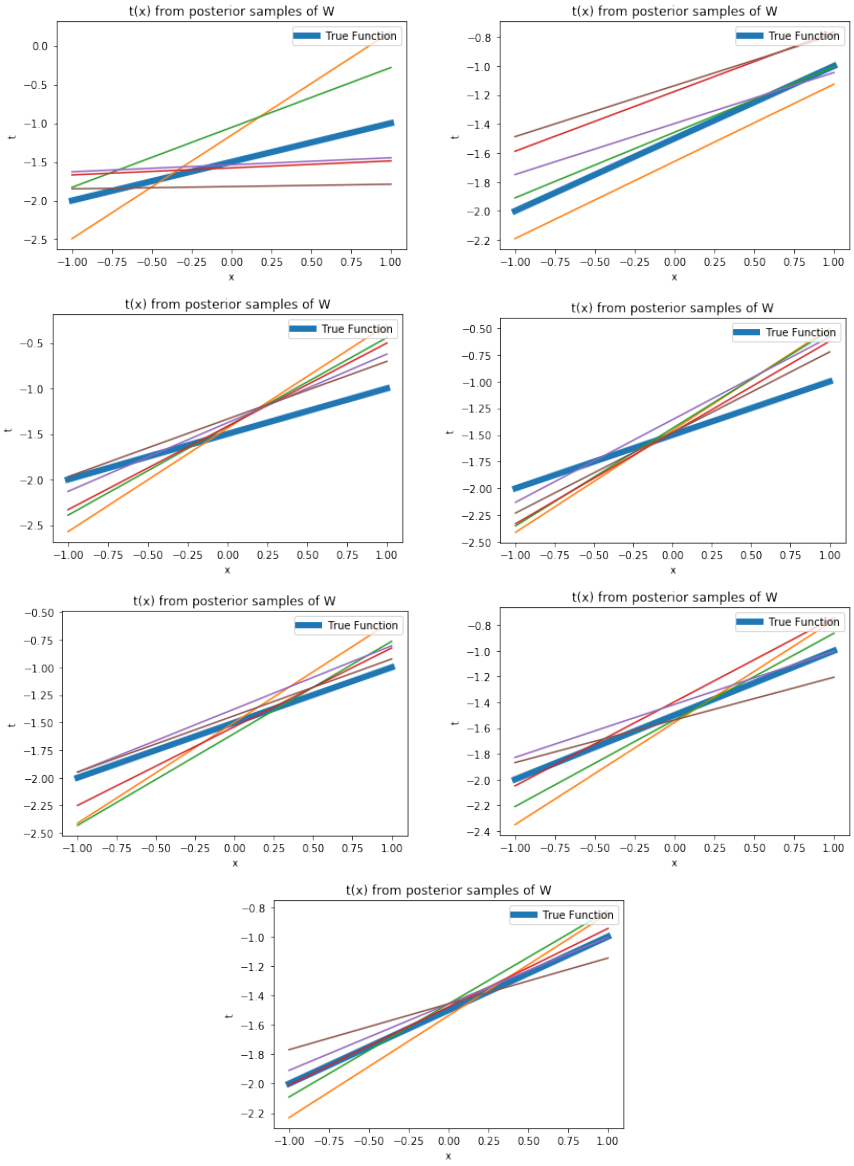

Figure 6: The function $t = w_0 x + w_1$ with five different W, sampled from the seven different posteriors in figure 5 (plots are ordered the same way). The "fat" blue line shows the true function.

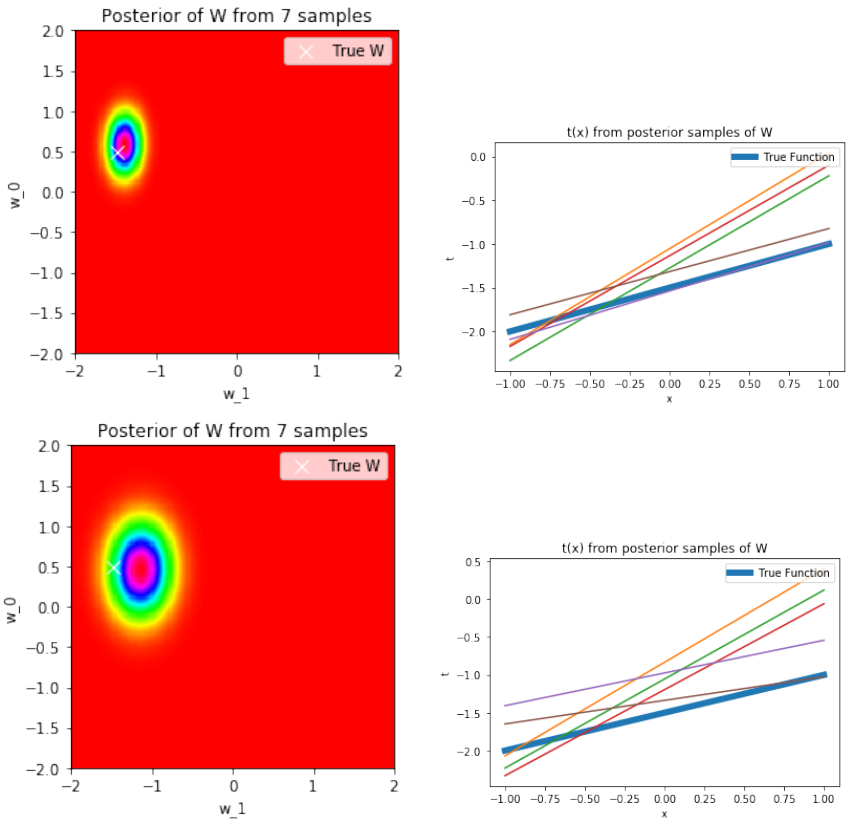

Figure 7: Posteriors for W and corresponding function lines, where 7 random data points from the data are included in the likelihood. As before, the white cross marks the true values of W and the blue line is the true one. The standard deviation on the noise term in the data is now $\sigma = 0.4$ in the first row and $\sigma = 0.8$ in the second row

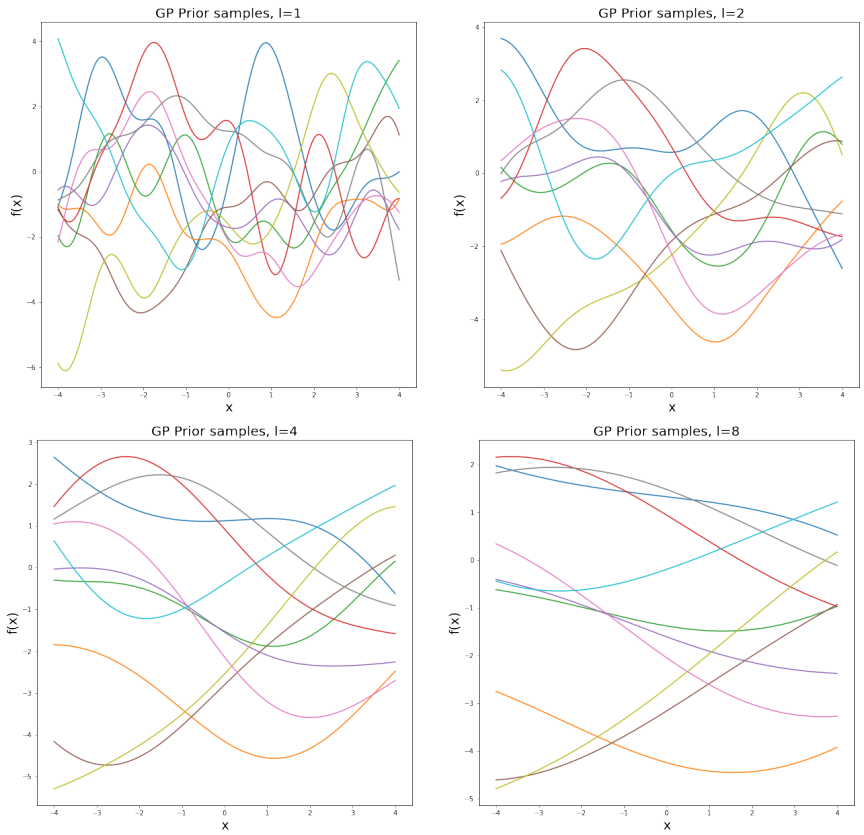

Figure 8: 10 samples from a Gaussian Process prior with varying length scale l in the squared exponential kernel.

follows:

$$\begin{bmatrix} \boldsymbol{T} \\ \boldsymbol{f}_* \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} k(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 I & k(\boldsymbol{X}, \boldsymbol{x}_*) \\ k(\boldsymbol{x}_*, \boldsymbol{X}) & k(\boldsymbol{x}_*, \boldsymbol{x}_*) \end{bmatrix} \right) \tag{16}$$

where $\sigma^2 I$ is the covariance matrix for the error term $\boldsymbol{\epsilon}$. The posterior $\boldsymbol{f}_* | \boldsymbol{T}, \boldsymbol{X}, \boldsymbol{x}_*$ is then also a Gaussian process with mean function $m_{post} = k(\boldsymbol{x}_*, \boldsymbol{X})(k(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 I)^{-1}\boldsymbol{T}$ and covariance function $cov_{post} = k(\boldsymbol{x}_*, \boldsymbol{x}_*) - k(\boldsymbol{x}_*, \boldsymbol{X})[k(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 I]^{-1} k(\boldsymbol{X}, \boldsymbol{x}_*)$. However if no data is observed and $\boldsymbol{T}$ and $\boldsymbol{X}$ do not exist, we have $\boldsymbol{f}_* | \boldsymbol{x}_* \sim N(\boldsymbol{0}, k(\boldsymbol{x}_*, \boldsymbol{x}_*))$, which is merely the Gaussian Process prior in (11) for $\boldsymbol{x}_*$.

A synthetic dataset was then generated according to the following:

$$t_i = (2 + (0.5x_i - 1)^2) * \sin(3x_i) + \epsilon_i \ \textit{where} \ \epsilon_i \sim N(0, 3) \ \textit{and} \ \boldsymbol{x} = [-4, -3, -2, -1, 0, 2, 3, 5]^T \tag{17}$$

Samples from the posterior using this dataset, where noise is included, is illustrated in figure 9. Figure 10 shows the result when ignoring the Gaussian noise in the model. A length scale of 2 and $\sigma_f^2 = 4$ was used throughout all experiments. It can be observed in both cases, that the posterior samples vary more the further away from the observed data. This is expected since $\boldsymbol{f}_*$ is conditioned on the observed data and the squared exponential kernel assigns larger covariance to nearby points. The higher uncertainty reflects our lack of knowledge in those regions.

Note how samples in figure 10 pass exactly through the observed data points, which is undesirable in this case, as we have noise in the observations and that uncertainty should be included in the model to avoid overfitting. Hence, the diagonal covariance matrix $\sigma^2 I$ should be added to the squared exponential as in (16). Such posterior is, as mentioned, shown in figure 9.

The predictive posterior mean with and without considering the noise in the data are furthermore shown in figure 11 and 12 respectively. The same behaviour as in figure 9 and 10 is illustrated here. Comparing the posterior to the prior samples, one may observe that the posterior mean converges towards the prior mean of 0 and the posterior standard deviation towards that of the prior (i.e $\sigma = 0.2$), as we move away from the observed data. This is desirable, since we do not want the observed data to affect our prior belief at a far distance from the observed data.

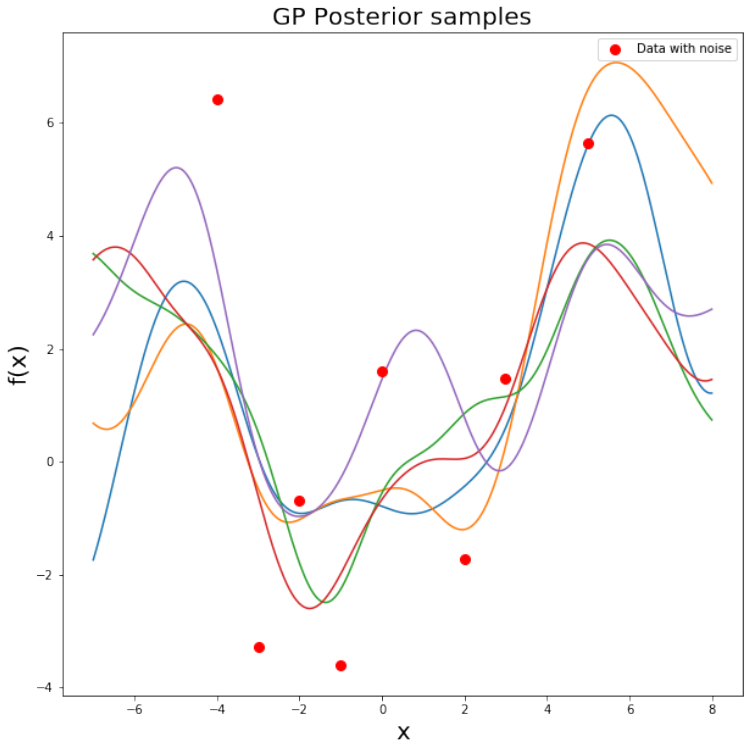

Figure 9: Samples from the Gaussian process posterior with noise term. The red dots mark the data used.

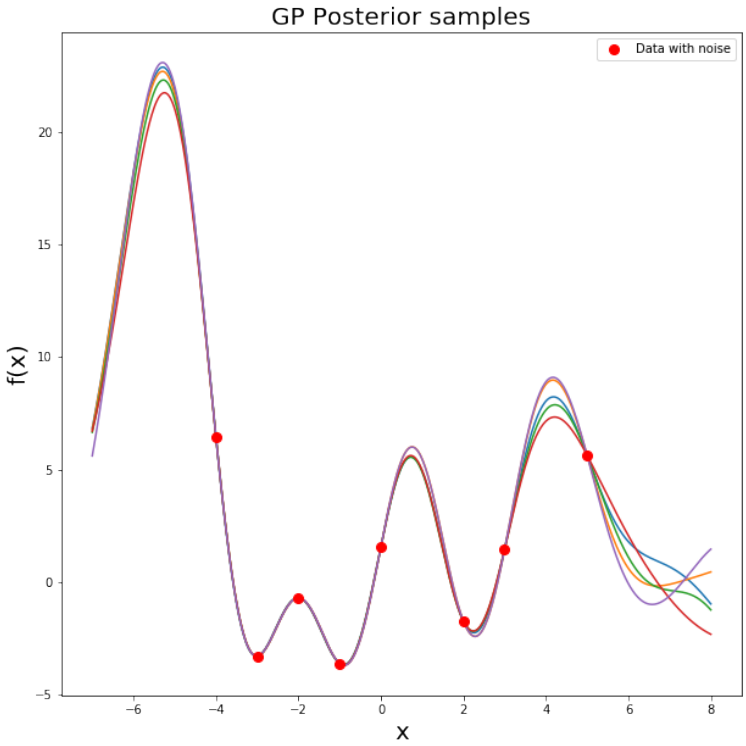

Figure 10: Samples from the Gaussian process posterior without noise term. The red dots mark the data used.

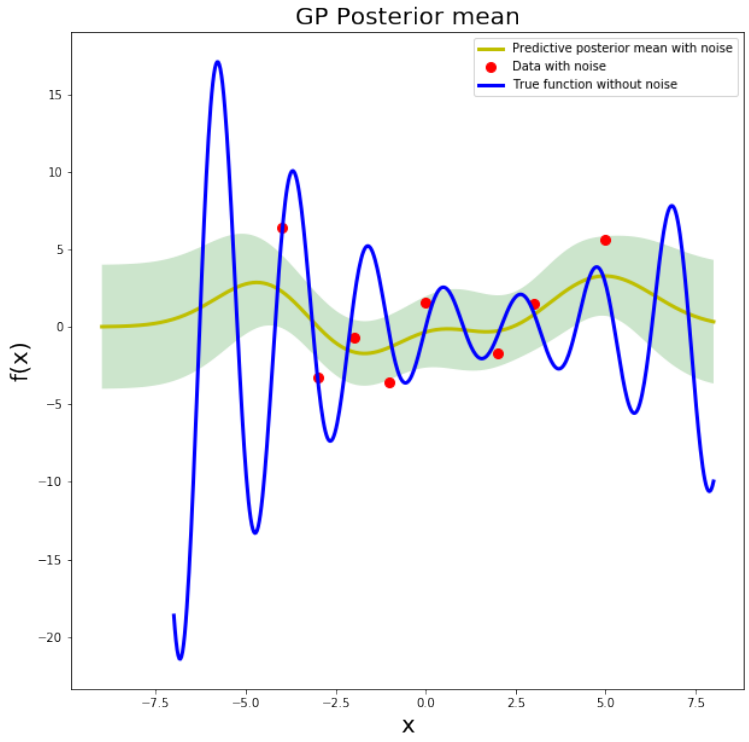

Figure 11: The Gaussian process posterior mean with noise term. The red dots mark the data used and the blue curve is the true function without noise term. The green shaded area illustrates mean plus minus two standard deviations.

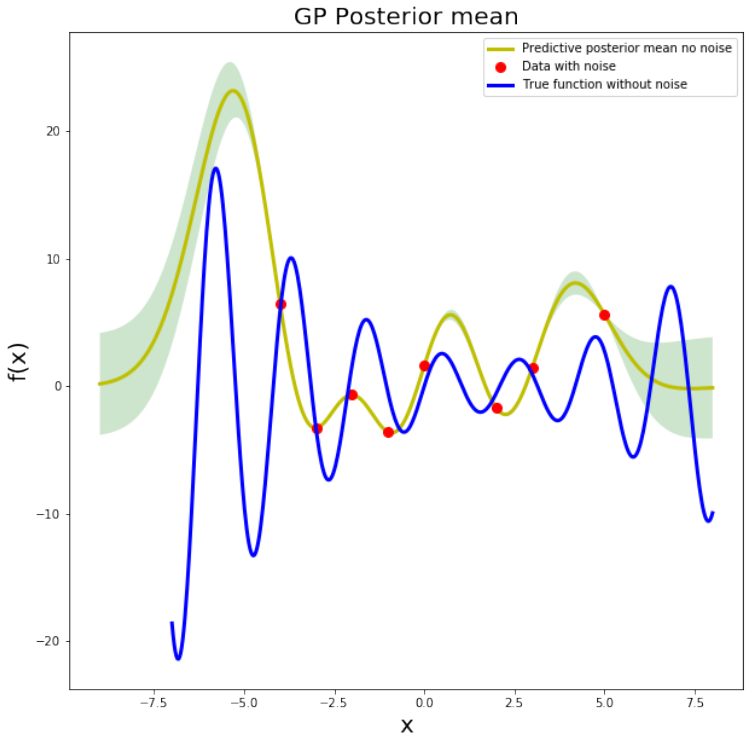

Figure 12: The Gaussian process posterior mean without noise term. The red dots mark the data used and the blue curve is the true function without noise term. The green shaded area illustrates mean plus minus two standard deviations.