Daiki Shirafuji
daikish@kth.se

Assignment 1

DD2434 MLadv 2019
January 6, 2020

Grade-E Level

Question 1

First, assuming the distribution of samples as a *Gaussian form* is a sensible choice as the following reason:

- According to Central Limit Theorem, the features, which are assumed i.i.d, converge into normal distributions.
- Therefore, the noise also tends to be a normal distribution.
- The conditions on *the likelihood* are giving f and x_i, so that it just depends on the noise, whose distribution may be a normal distribution.

Therefore, it is natural to assume that the distribution of *the likelihood* is a normal distribution.

Second, *the assumptions we make by this choice (Gaussian form)* is that all of features x_i are i.i.d.

Finally, by choosing a spherical covariance matrix for the likelihood, the assumption is that a spherical covariance matrix for the likelihood means that the off-diagonal correlations are 0, which means features are independent to all other features.

Question 2

According to $P(X, Y) = P(Y|X)P(X)$, the answer is following:

$$\begin{aligned} p\left(\mathbf{T}|f, X\right) &= p(t_1|f, X)p(t_2, \ldots, t_N|t_1, f, X) \\ &= p\left(t_1|f, X\right) * \prod_{i=2}^{N} p\left(t_i|t_1, \ldots, t_{i-1}, f, X\right). \end{aligned} \tag{1}$$

Question 3

According to $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, $p(\mathbf{T}|\mathbf{X}, \mathbf{W})$ is following:

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

where $p(t_i|x_i, \mathbf{W}) = N(\mathbf{W}x_i, \sigma^2\mathbf{I})$.

Question 4

- What would be the effect of encoding the preferred model with L_1 norm (for model parameters)?

When encoding with L_1 norm instead of L_2 norm, the prior become ***less*** smooth because of the variable selection in L_1 norm.

- Discuss how these two types of priors affect the posterior from the regularisation perspective. Write down the penalization term, i.e. the negative log-prior, and illustrate for a two-dimensional problem (in the two-dimensional parameter space).

First, according to Bayesian Rule, we can say that:

$$posterior \propto likelihood * prior \tag{3}$$

and this is equivalent to:

$$ln(posterior) \sim ln(likelihood) + ln(penalty). \tag{4}$$

This is how these priors affect the posterior as penalties.
Now, start to write down the penalisation term (i.e. $ln(penalty)$ or $ln(prior)$):

$$-ln[p(\mathbf{W}|\mathbf{D})] = \frac{1}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T f(\mathbf{x}_n))^2 + \frac{\lambda}{2}\sum_{i=1}^{K}|w_i|^q + const \tag{5}$$

where $\mathbf{D}$ is the data, K is the number of elements in one of $\mathbf{W}$ row, q is an integer value (0, 1, ...), and $const$ is a constant value.
In the L_1 case, q is 1, and in the L_2 case, q is 2.

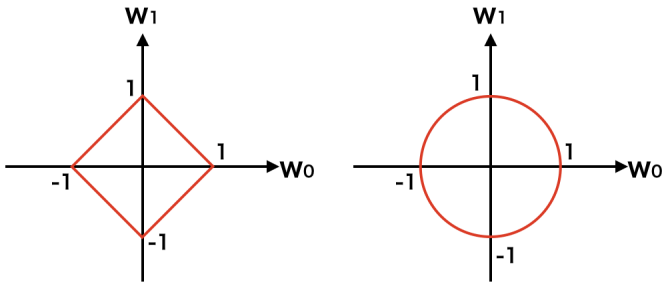

Figure 1: Regularisation on L1 (left) and L2 (right)

Finally, the regularisation in the two-dimensional parameter space are shown in Figure 1.
As Figure 1 is shown, L1 regularisation tend to yield more sparse solution than L2.

Question 5

According to Question 3, $p(\mathbf{T}|\mathbf{X},\mathbf{W}) = \prod_{i=1}^{N-1} p(t_i|x_i,\mathbf{W})$ and $p(t_i|x_i,\mathbf{W}) = N(\mathbf{W}x_i, \sigma^2\mathbf{I})$, where $i = 1, ..., N$. The distribution which is produced by Gaussian forms Gaussian distribution, so that it can be said that:

$$p(\mathbf{T}|\mathbf{X},\mathbf{W}) = N(\mathbf{W}\mathbf{X}, \sigma^2\mathbf{I}). \tag{6}$$

The definition of a Gaussian distribution is represented with following:

$$N(\mu,\Sigma) = \frac{1}{\sqrt{2\pi\Sigma}} exp(\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)) \tag{7}$$

Now, we set Σ as τ^2. Then, according with $p(\mathbf{W}) = MN(\mathbf{W}_0, \mathbf{I}, \tau^2\mathbf{I})$,

$$\begin{aligned} p(\mathbf{W}) &= \frac{1}{\sqrt{2\pi\Sigma}} exp(\frac{-1}{2}(W-W_0)^T\Sigma^{-1}(W-W_0)) \\ p(\mathbf{T}|\mathbf{X},\mathbf{W}) &= \frac{1}{\sqrt{2\pi\sigma^2}} exp(\frac{-1}{2\sigma^2}(T-XW)^T(T-XW)). \end{aligned} \tag{8}$$

In addition, according to Bayesian rule, $p(\mathbf{W}|\mathbf{T},\mathbf{X}) \propto p(\mathbf{W}) * p(\mathbf{T}|\mathbf{X},\mathbf{W})$. Therefore, the posterior can be derived with the following:

$$\begin{aligned} p(\mathbf{W}|\mathbf{T},\mathbf{X}) &\propto \frac{1}{\sqrt{2\pi\Sigma}} exp(\frac{-1}{2}W^T\Sigma^{-1}W) \frac{1}{\sqrt{2\pi\sigma^2}} exp(\frac{-1}{2\sigma^2}(T-XW)^T(T-XW)) \\ &\propto exp(\frac{-1}{2}W^T\Sigma^{-1}W + \frac{-1}{2\sigma^2}(T-XW)^T(T-XW)) \\ &= exp(-\frac{1}{2}W^T\Sigma^{-1}W - \frac{1}{2\sigma^2}T^TT + \frac{1}{\sigma^2}T^T(XW) - \frac{1}{2\sigma^2}(XW)^T(XW)). \end{aligned} \tag{9}$$

Finally, the posterior is formulated with Equation (9). Next, we explain this posterior in terms of the mean μ_W and the covariance Σ_W. $p(\mathbf{W}|\mathbf{T},\mathbf{X})$ can be formulated with μ_W and Σ_W as following:

$$p(\mathbf{W}|\mathbf{T},\mathbf{X}) = exp(-\frac{1}{2}W^T\Sigma_W^{-1}W + W^T\Sigma_W^{-1}\mu_W - \frac{1}{2}\mu_W^T\Sigma_W^{-1}\mu_W). \tag{10}$$

Therefore, the mean and the covariance in the posterior is represented with following:

$$
\begin{aligned}
u_W &= (\Sigma^{-1} + \sigma^{-2}\mathbf{X}^T\mathbf{X})^{-1}(\sigma^{-2}\mathbf{X}^T\mathbf{T} + \Sigma^{-1}W_0) \\
\Sigma_W &= (\Sigma^{-1} + \sigma^{-2}\mathbf{X}^T\mathbf{X})^{-1}.
\end{aligned} \tag{11}
$$

As we can see from Equation (10) and (11), the posterior also form Gaussian distribution in Bayesian linear regression. Therefore, it can be updated after the relevant information has been taken into account. This is how the posterior form relates to the estimator.

In terms of Z, there is no affect on the solution. Z is the constant value here, so that Z will not affect the posterior mean and the covariance. Z only represents the evidence, which we are interested in during the model selection.

Question 6

In Gaussian Process, the prior represents a probability distribution over the functions, not over the parameters.

The choice of this prior is nice because the function, which is used in Gaussian Process, is not clear usually.

Now, the prior can be represented with $p(f|\mathbf{X}, \theta) = N(\mathbf{0}, k(\mathbf{X}, \mathbf{X}))$ according to Equation (11), which is given from the questions. And $k(\mathbf{X}, \mathbf{X})$ is given from the following:

$$
k(\mathbf{x}_n, \mathbf{x}_m) = \theta_0 exp(-\frac{\theta_1}{2}||\mathbf{x}_n - \mathbf{x}_m||^2) + \theta_2 + \theta_3\mathbf{x}_n^T\mathbf{x}_m. \tag{12}
$$

Samples from the prior is shown in Figure 2, which is quoted from Bishop pp 308.

According with this Figure, the parameters θ represents how θ control the shape of the function. Therefore, we can say the choice of this prior is nice.

Question 7

According to Bayesian Rule, $p(T, X, f, \theta)$ can be represented with the following equation.

This is because $\mathbf{T}$ is conditionally dependent on $\mathbf{X}$ and θ as the Figure 3 shows.

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

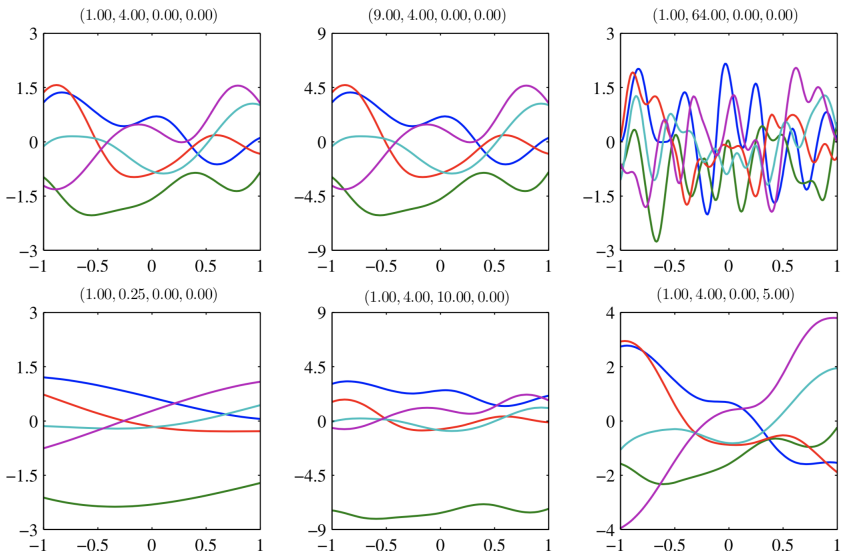

Figure 2: Samples from a Gaussian process prior defined by the covariance function. The title above each plot denotes (θ_0, θ_1, θ_2, θ_3).

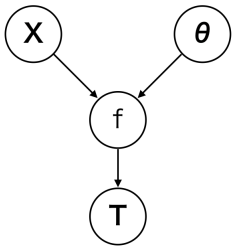

Figure 3: Graphical Model

Question 8

Because $\mathbf{T}$ is conditionally dependent on θ and $\mathbf{X}$, $p(\mathbf{T}|\mathbf{X}, \theta)$ is generally formulated with the following:

$$p(\mathbf{T}|\mathbf{X}, \theta) = \int p(\mathbf{T}|f)p(f|\mathbf{X}, \theta)df \tag{14}$$

where $p(f|\mathbf{X}, \theta)$ is the prior over the data $\mathbf{X}$. Now, the prior of all $\mathbf{X}$ over all f is averaged. The **uncertainty** of the value from f and the true value is represented with $p(\mathbf{T}|f)$. In addition, $p(f|\mathbf{X}, \theta)$ represents the uncertainty of $\mathbf{X}$ and f. Now, we know that these two uncertainties, which are represented from $p(\mathbf{T}|f)$ and $p(f|\mathbf{X}, \theta)$, do "filter" the uncertainty of the inputs $\mathbf{X}$ and the values $\mathbf{T}$.

In Gaussian Process, we get the average over all functions f. Now, θ is a constant value in the process of averaging over f. Therefore, θ remains after averaging, so that the final result become a function of θ.

Question 9

1. Set the prior distribution over $\mathbf{W}$ and visualise it.

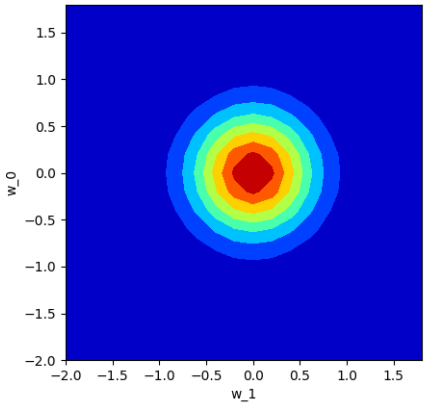

Figure 4: The Prior Distribution over $\mathbf{W}$

2. Pick a single data point (x, t) and visualise the posterior distribution over $\mathbf{W}$.
3. Draw 5 samples from the posterior and plot the resulting functions.

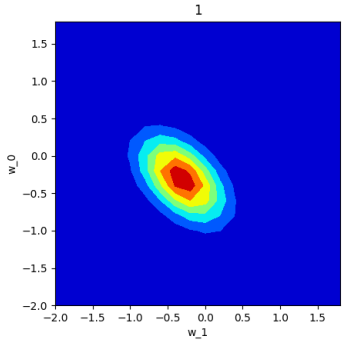

Figure 5: Posterior (1 Observation)

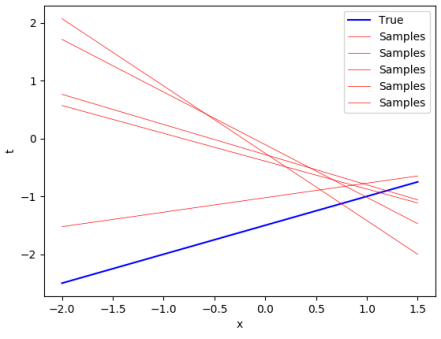

Figure 6: Five Samples from Posterior

4. Repeat 2 to 3 by adding additional data points up to 7.

Answers are shown in the following figures (Figure 7 to Figure 12). The numbers of adding additional data points are 3, 5, and 7.

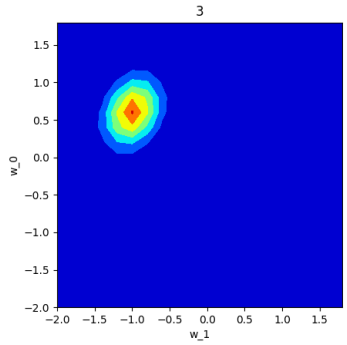

Figure 7: Posterior (3 Observations)

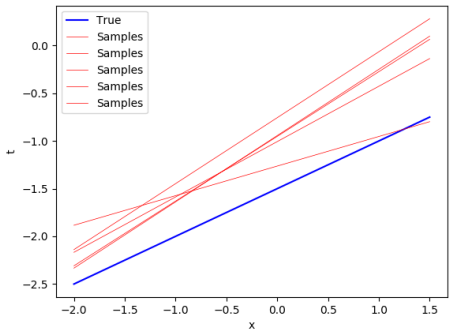

Figure 8: Five Samples from Posterior

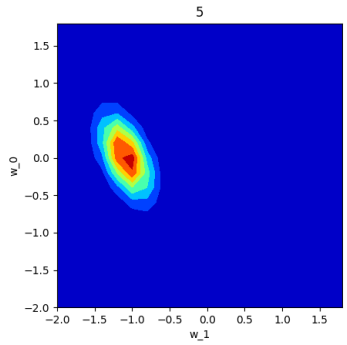

Figure 9: Posterior (5 Observations)

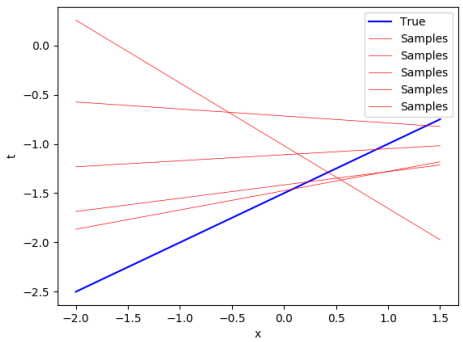

Figure 10: Five Samples from Posterior

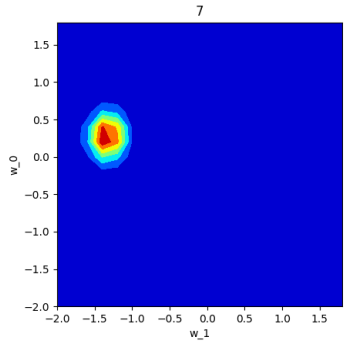

Figure 11: Posterior (7 Observations)

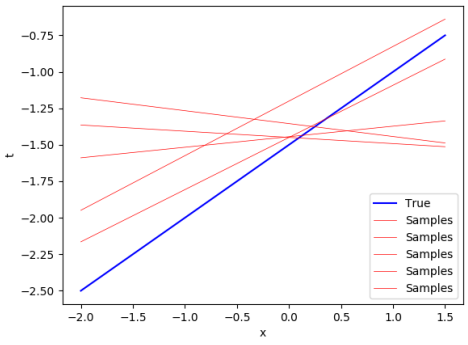

Figure 12: Five Samples from Posterior

5. Given the plots explain the effect of adding more data on the posterior as well as the functions. How would you interpret this effect?

The more the data is, the samples can represent closer function to the True Function.

This can be explained with the reason that the more data points we have in advance, the data points converge into True points.

6. Finally, test the exercise for different values of σ e.g. 0.1, 0.4 and 0.8. How does your model account for data with varying noise levels? What is the effect on the posterior?

The results when $\sigma = 0.1$, 0.4, and 0.8 (under the condition of 7 Observations) are shown in the following figures (Figure 13 to Figure 18).

If the noise levels on data are high (i.e. sigma is a large value), the posterior tends to take a large range in the 2-dimensional space. This means that it becomes unpredictable in the models.

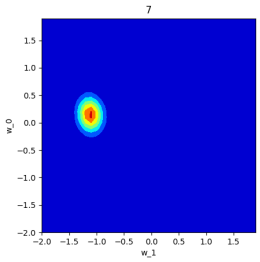

Figure 13: Posterior (7 Observations) on $sigma = 0.1$.

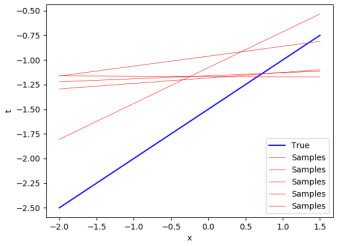

Figure 14: The Prior Distribution over $\mathbf{W}$ on $sigma = 0.1$.

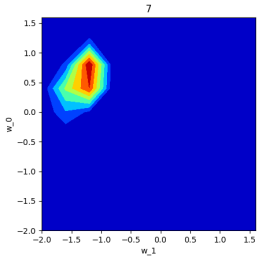

Figure 15: Posterior (7 Observations) on $sigma = 0.4$.

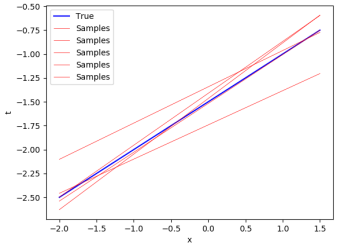

Figure 16: The Prior Distribution over $\mathbf{W}$ on $sigma = 0.4$.

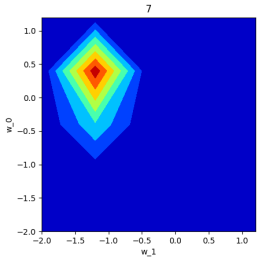

Figure 17: Posterior (7 Observations) on $sigma = 0.8$.

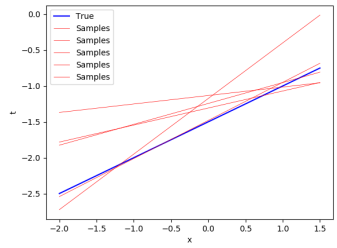

Figure 18: The Prior Distribution over $\mathbf{W}$ on $sigma = 0.8$.

Question 10

For each 4 different length scales (0.1, 1, 10, and 100), 10 samples from the prior are shown in the following figures (Figure 19 to Figure 22).

From these figures, we can know that the samples, which are gotten from the results, show more smooth carve when the length scales become bigger.

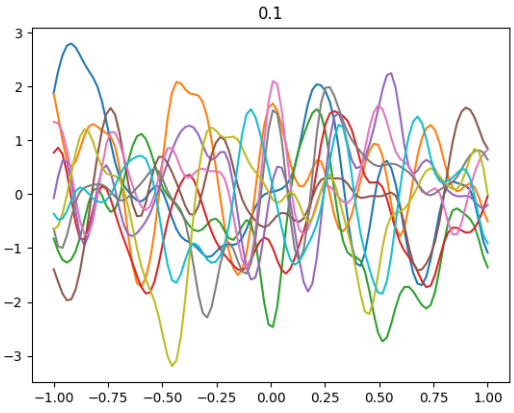

Figure 19: 10 Samples from GP when length-scale = 0.1.

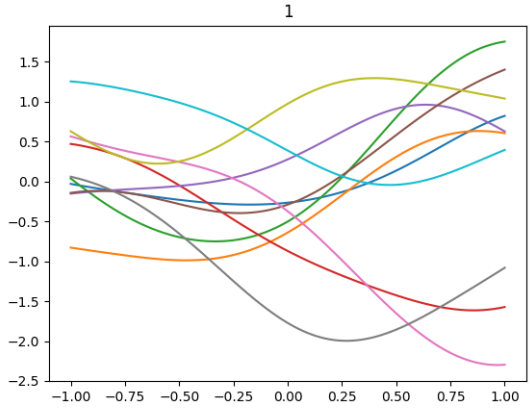

Figure 20: 10 Samples from GP when length-scale = 1.

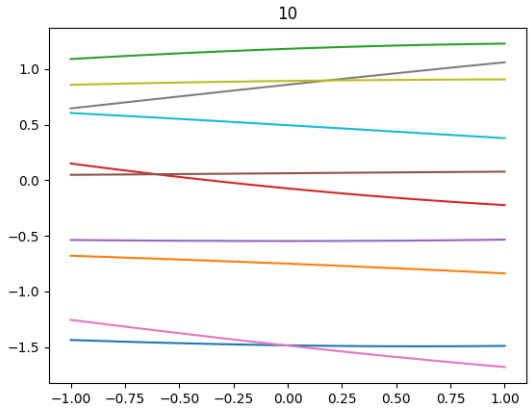

Figure 21: 10 Samples from GP when length-scale = 10.

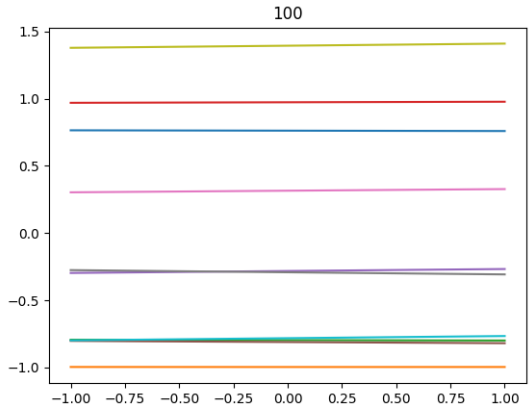

Figure 22: 10 Samples from GP when length-scale = 100.

Question 11

1. What is the posterior before we observe any data?

Before we observe any data, we have no information for the posterior in advance, so that we set the posterior as the value of the prior. Therefore, the posterior is represented with the following formula:

$$p(\mathbf{f}|\mathbf{X}, \theta) = N(0, k(\mathbf{X}, \mathbf{X})). \tag{15}$$

2. Compute the predictive posterior distribution of the model.

We now compute the predictive posterior distribution in the general case.

First of all, let me set that t_{N+1} is a target value to predict under the condition of x_{N+1}. In addition, from Bishop p. 308, $m(x_{N+1}) = \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{t}$ and $\sigma^2(x_{N+1}) = c - \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{k}$ are defined, where $c = k(x_{N+1}, x_{N+1}) + \beta^{-1}$ and

$$\mathbf{C}_{N+1} = \begin{pmatrix} \mathbf{C}_N & \mathbf{k} \\ \mathbf{k}^T & c \end{pmatrix}. \tag{16}$$

The predictive posterior distribution can be represented with $p(t_{N+1}|\mathbf{t}_N)$, and the joint distribution $p(\mathbf{t}_{N+1})$ can be represented with $N(t_{N+1}|0, \mathbf{C}_{N+1})$.

Therefore, the predictive posterior distribution is represented with following:

$$p(t_{N+1}|\mathbf{t}) = N(t_{N+1}|m(x_{N+1}), \sigma^2(x_{N+1})) \tag{17}$$

3. Sample from this posterior with points both close to and far away from the observed data. Explain the observed effects.

As shown in the following figure (Figure 23), we can say that as the samples is closer to the observed data, the samples are converged into it. On the other hand, we can also say that when the samples go away from the observed data, the samples tend to be not able to be controllable.

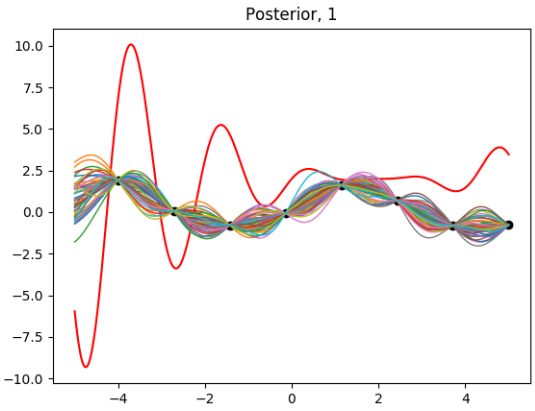

Figure 23: Observations with length-scale = 1 (Noise included).

4. Plot the data, the predictive mean and the predictive variance of the posterior from the data.

The required figure is shown in the following figure (Figure 24).

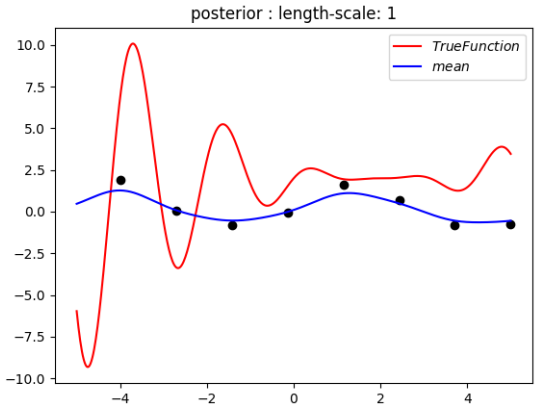

Figure 24: With noises in a diagonal covariance matrix to the squared exponential.

5. Compare the samples of the posterior with the ones from the prior. Is the observed behavior desirable?

Comparing the samples of the posterior with the ones from the prior, we can know that the difference between the prior and the posterior is that the samples are collected more or not. This behaviour is desirable because it relate to the model information.

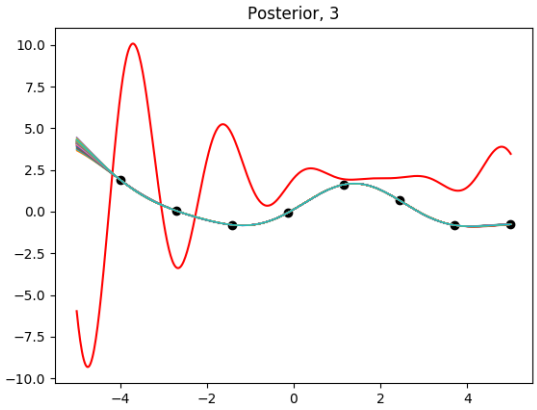

Figure 25: Length-scale=3.

6. What would happen if you added a diagonal covariance matrix to the squared exponential?

We know from comparing Figure 25 with the following figure (Figure 26) that when we added the noise to a diagonal covariance matrix to the squared exponential, it become more difficult to predict the True Function than without noises.

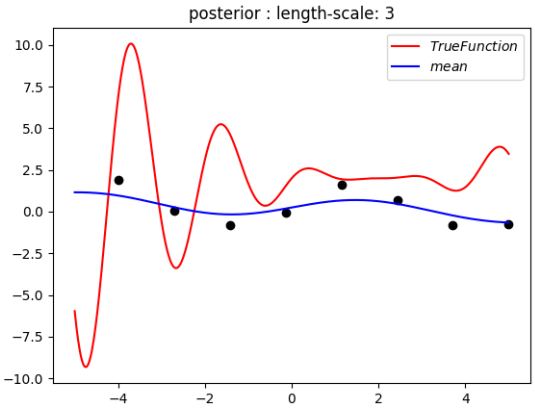

Figure 26: Adding noises in a diagonal covariance matrix to the squared exponential (Length-scale=3).