Z - 10, 11, 12, 18, 20, 26-30

L - 13, 14, 19, 22-25

1-1. What assumption does a Gaussian likelihood encode, i.e. what motivates the choice of this likelihood function?

The assumption is that each data point is independent of the others and identically distributed. That means that the likelihood is the product of the probabilities of each individual data point.

*Central limit theorem the average/sum of i.i.d. samples from any distribution will*

*follow a Gaussian distribution. This is one of these magical things when you see if*

*first and I do recommend testing it with some code. When you think a bit further*

*it makes perfect sense. It can be very useful to motivate the use of Gaussians as if I*

*don’t know which distribution something comes from, but I assume I get samples i.i.d.*

*then I’ll take an average and this will be Gaussian. (from SUMMARY)*

1-2. What does it mean that we have chosen a spherical covariance matrix for the likelihood, contrast with the non-spherical case?a

It means we assume each parameters in two dimensions is independent, and they have the same probability density for a Gaussian distribution that the covariance matrix is proportional to the identity matrix, i.e, .

2. If we do not assume that the data points are independent how would the likelihood look then?

*p(Y | f,X) = p(,...,| f,X)*

If we do not assume independence of data points we will have to deal with the joint *p(,...,| f,X),* instead of looking at individual terms *p(| f,X)*.

3. What is the specific form of the likelihood above, complete the right-hand side of the expression.

The likelihood is of Gaussian form.

Given N training examples, the likelihood assuming i.i.d. is

p(Y|X,W) = p(*|,W)*

4. Explain the concept of conjugate distributions, why do they help us compute the posterior distribu-

tion?

(Based on Baye’s rule, we know that the posterior is proportional to the likelihood times prior, if we assume that for a given probability distribution, we can seek a prior that is conjugate to the likelihood function, so that the posterior distribution has the same functional form as the prior, therefore we can avoid computing the evidence and just identify parameters.)

Conjugate distributions are prior and posterior distributions that are in the same probability distribution family. Choosing the conjugate prior (i.e. Gaussian) is a sensible choice as the posterior will also be Gaussian. A conjugate prior helps compute posterior distribution as it gives a closed-form expression for the posterior distribution, meaning the calculations can be done in a finite number of operations and typically involve no limit calculations. Knowing the form of posterior distribution helps to identify the parameters.

5. Reason about the Gaussian distribution in this context, which distance function does it encode with a spherical covariance matrix.

For a spherical covariance matrix, since two independent parameters have the same probability distribution, so we can think it as a single Gaussian distribution, simply use Euclidean distance function to estimate the probability density function of a random variable, then we can find the maximum likelihood.

6.

Write out the posterior over the parameters W. I recommend that you do these calculations by hand as it is very good practice and provides important intuitons. However, in order to pass the assignment you only need to outline the calculation and highlight the important steps.

• Justify the the posterior by providing an intuition of its form.

**We know that the posterior is proportional to the likelihood multiplied by prior due to conjugacy**

**Since we derived likelihood** p(Y|X,W) = p(*|,W)* and our prior is p(W) = *N*(, I), we can now compute the posterior as per our mapping function.

=

= p(*|,W) N*(, I)

**As the likelihood and prior are Gaussians, the intuition is that posterior will also be of Gaussian form.**

7.

What is a non-parametric model and what is the difference between non-parametrics and parametrics? In specific discuss these two aspects of non-parametrics,

• Representation/parametrisation of data?

• Interpretability of models?

A I’m not even sure if this is a word but hopefully you get what I mean. Write down what it means to you.

Parametrics models assume data can be defined by finite set of parameters. We encode relationship between variates using parameters w, which induced a corresponding prior distribution, to obtain the corresponding posterior distribution over parameters w. Given the parameters, predictions are independent of the observed data.

Therefore, the complexity of parametrics models are bounded even if the amount of data is unbounded, they only interpret the local data, they are not very flexible.

In contrast, non-parametric models assume data is defined by a function with infinite dimensional. We encode relationship between variates using variates themselves, define a prior probability distribution over functions directly, the corresponding posterior distribution may be various functions. Predictions are dependent of the observed data.

Therefore, non-parametric models about the data can grow as the amount of data grows, their complexity depends on data, they interpret the global data, they are more flexible.

The parametric model is more easy to interpret than a non-parametric one, as the complexity of the model does not change. The non-parametric model on the other hand allows unlimited number of parameters - it can create new parameters, as opposed to the parametric model.

8. Explain what this prior represents and how it places structure on the space of functions?

This is equivalent to choosing the mean of the prior over weight values p(w|α) to be zero in the basis function viewpoint. The specification of the Gaussian process is then completed by giving the covariance of y(x) evaluated at any two values of x, which is given by the kernel function and

P305

GP represents distribution over functions, and if we assume the mean function to be zero then the Gaussian process is fully defined by its covariance function k(X,X).

9. Does this prior encode all possible functions or only a subset?

**Checkthis**

This prior encodes all possible functions, because this prior can represent any function y(x) (?) **k(X,X)** which is evaluated at any two values of x.

10. Formulate the joint distribution of the full model that you have defined above,

p(Y, X, f, θ)

Draw the graphical model and clearly state the assumptions that has been made in bullet list.

The prior is



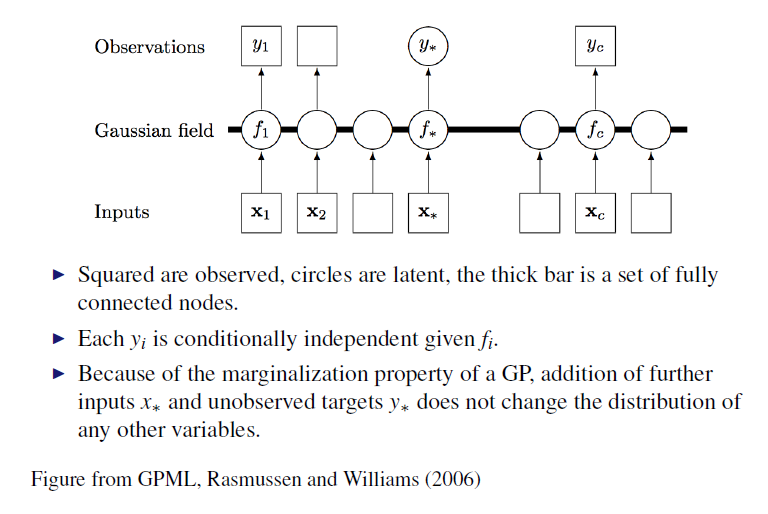
And therefore the joint distribution of training outputs f and test outputs

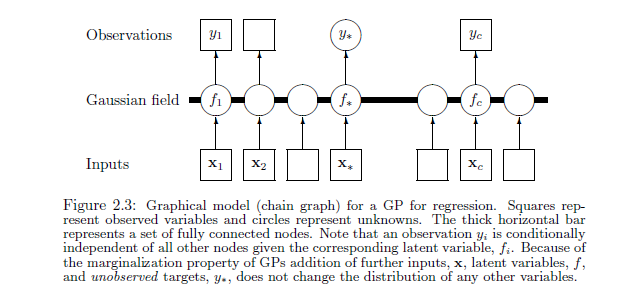


Where function f is evaluated at test points X\*. With n training points and n\* test points K(X,X\*) indicates a n×n\* matrix of the covariances assessed at all pairs of training points and test points. The same applies for K(X,X\*), K(X\*,X) and K(X\*,X\*).

Assumptions are

* No noise





Page 17

11.

Explain the marginalisation in Eq.2

• Explain how this connects the prior and the data?

• How does the uncertainty “filter” through this?

• What does it imply that θ is left on the left-hand side of the expression after marginalisation?

To perform marginalisation over the mapping f we calculate the integral of the likelihood multiplied by prior.

The Gaussian process marginalization allows for addition of more inputs x and targets y - it does not change the distribution of any other variables. This allows for uncertainty.

on the left means we calculate the marginal likelihood as a function of kernel hyperparameters .

12.

1. Visualise the prior distribution over W.

2. Pick a single data-point from the data and visualise the posterior distribution over W

3. Sample from the posterior and show a couple of functions

4. Repeat 2 − 3 by adding additional data points

5. Describe the plots, and the behavior when adding more data? Is this a desirable behavior?

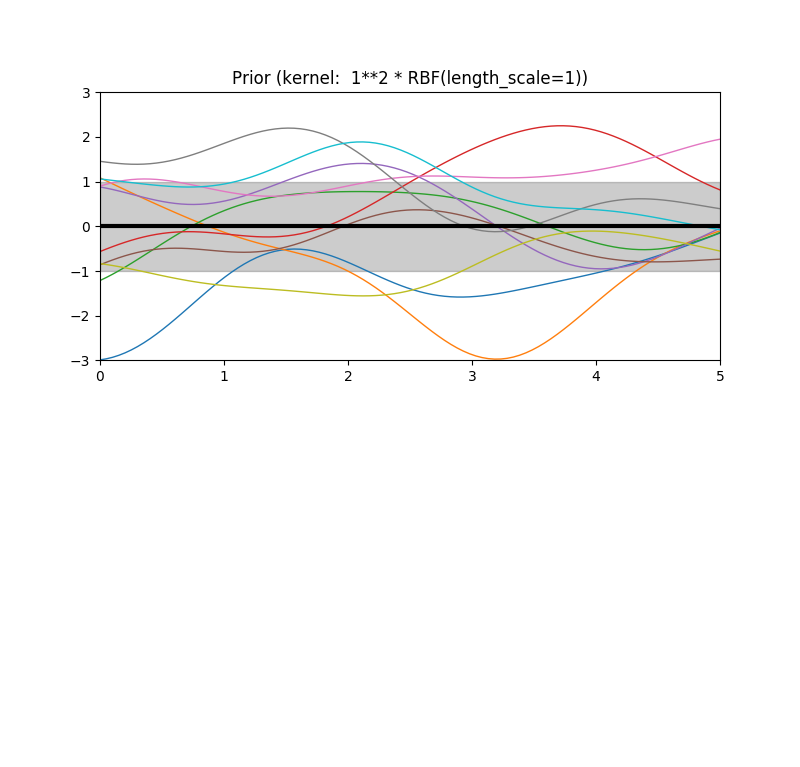
6. Relate to the expression of the posterior why you see the behaviour that you do when you add

more data

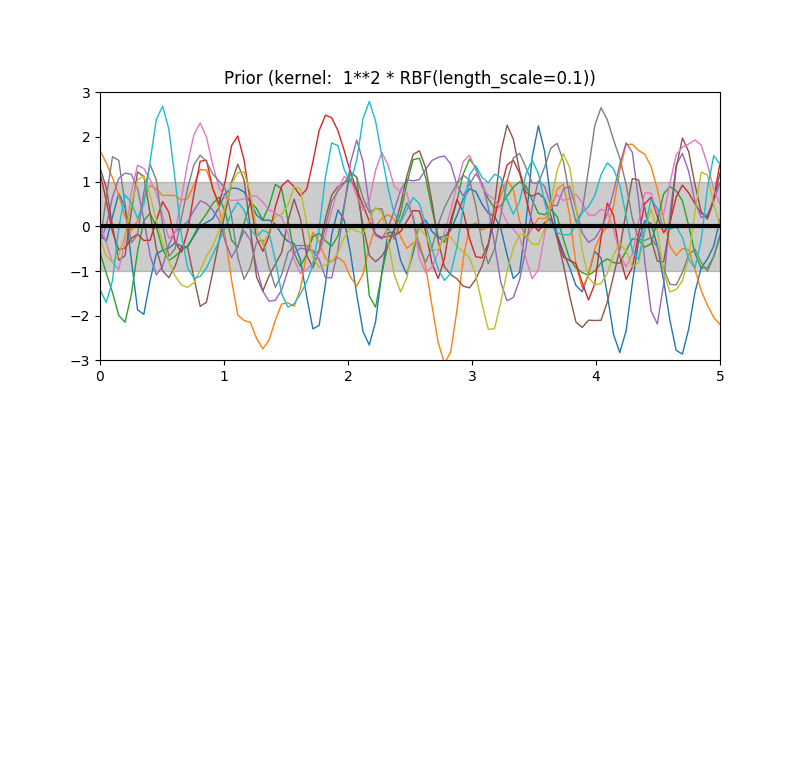
13.

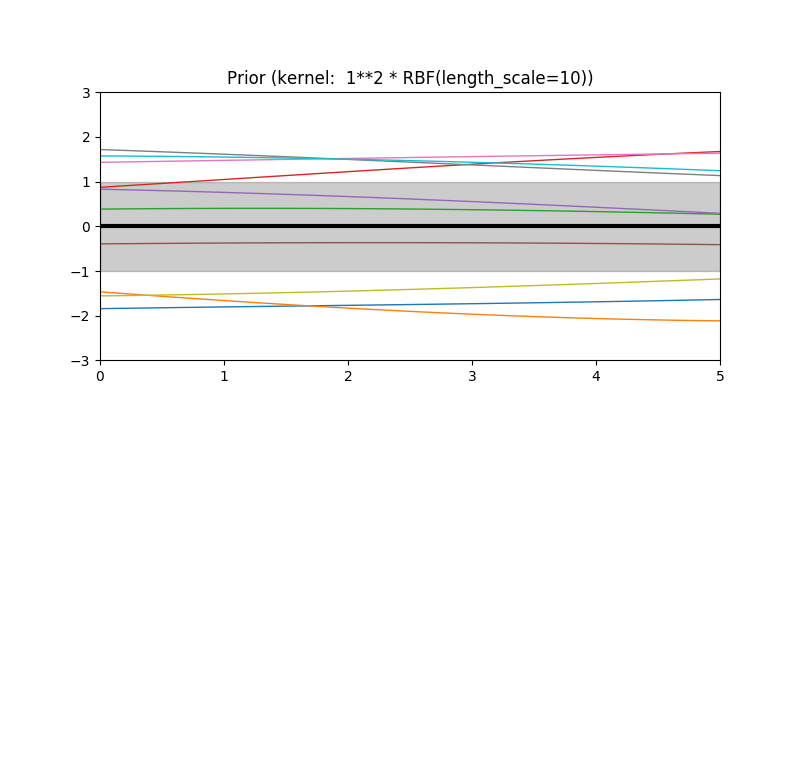
1. Create a GP-prior with a squared exponential covariance function.

2. Sample from this prior and visualise the samples



3. Show samples using different length-scale for the squared exponential





4. Explain the behavior of altering the length-scale of the covariance function.

We alter the length-scale to adjust how smooth the function is. Small length-scale value means that function values can change quickly, large length-scale values characterize functions that change very slowly.

5. What assumption does the length-scale encode?

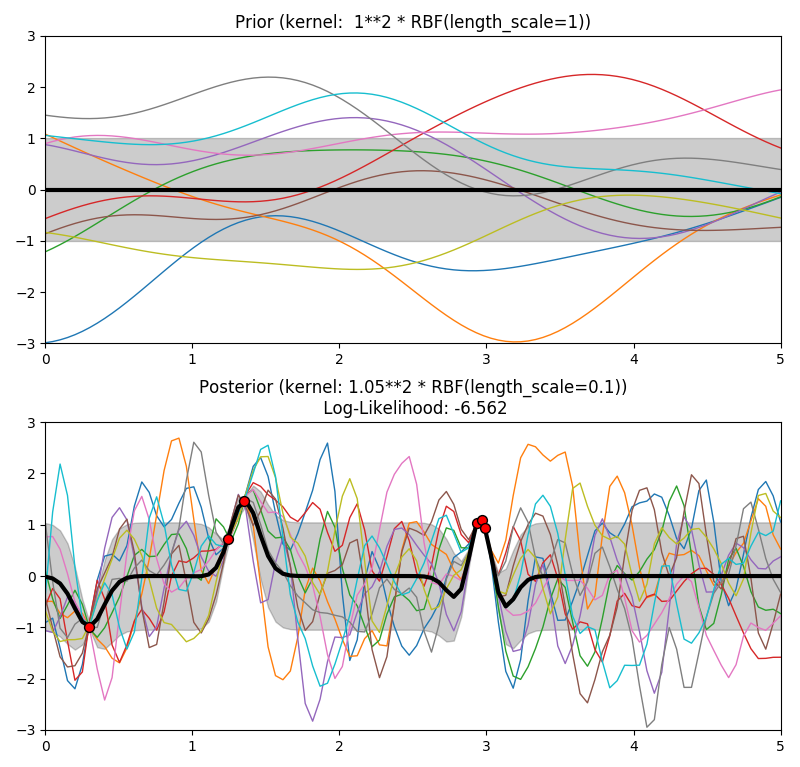
Length-scale assumes the how far we can reliably extrapolate from the training data. In general, we won't be able to extrapolate more than the length-scale away from the data.

14.

1. Compute the predictive posterior distribution of the model

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

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



Explain the behavior of the samples and compare the samples of the posterior with the ones from the prior. Is this behavior desirable? What would happen if you would add a diagonal covariance matrix to the squared exponential?

Using the non-parametric regression, the posterior has the similar shape with the observed data, it can go through the each observed data with small length-scale.

15.

Elaborate on the relationship between assumptions, belief and preference

There is no right or wrong answer here, I want to hear your thoughts on the what these words means to you in the context of what we have been doing.

Assumptions are the conditions added without proof. Those conditions are extremely sensitive and uncertain. For instance, if we assume some random variables can be categorized in linear regression model, so we can use a linear equations to represent the variables. Without this assumption, they maybe have any distribution we don’t know.

Belief is the possibility added on reality, the possibility can be changed within the range. For instance, if we believe all answers have the same possibility, so that the final answer is the mean of all answers. Without this belief, the final answer may be any one of the answers, but still in the range of all answers.

Preference is the conditions we added with expecting the data, usually they are existing options, we choose they to formulate the data more easily. For example, after observing the data, we can choose using different models to represent the data. They are different ways how we obtain predictions, but we can’t tell any of them will get the correct predictions.

16.

What is the assumption/preference we have encoded with this prior?

We make the assumption that , the is independent Gaussian distribution. And observed data shows a preference that in each dimension, is a standard Gaussian distribution with variance 1 and mean 0.

17.

Perform the marginalisation in Eq.2.1 and write down the expression. As previously, I do recommend that you do this by hand but to pass the assignment you only need to outline the calculations and show the approach that you would take.

as , so , +

18.

• How are the different?

• How are MAP and ML different when we observe more data?

• Why are the two expressions in Eq. 10 equal?

19.

• Write down the objective function −log(p(Y|W)) = L(W).

• Write down the gradients of the objective with respect to the parameters δLδWa

aLook in the appendix for help to compute the matrix derivatives. It is important that you follow how they are computed as this is something we often have to do.

Assume

The negative log likelihood is

The gradients

When

<http://www.robots.ox.ac.uk/~az/lectures/ml/2011/lect4.pdf>

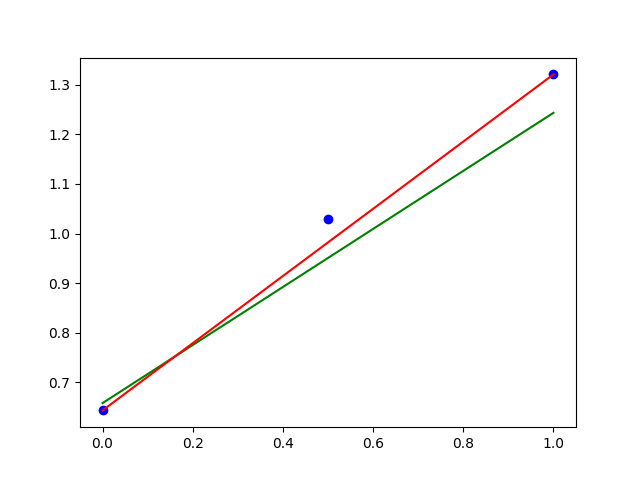
20.

Marginalisation of f is much simpler to do than marginalising out X, the latter is actually in most cases analytically intractable. Provide a simple reason why this is, the argument should be general about marginalisation and not about this model in specific. A good idea is to draw the graphical model and then follow the arrows (hint hint).

That is because X represents observed data.

21.

Plot the representation that you have learned. Explain why it looks the way it does. Was this the result that you expected? Hint: Plot X as a two-dimensional representation.

 Assume the model is , with 3 observed data points (blue point), draw a (green line), find the minimum sum of the gradient of the data at , so we can get the optimization function which represents the average of all observed data (red line).

22.

Draw a random two dimensional subspace (does not have to be an orthogonal basis) and plot the data. How is this result different compared to the subspace that you learnt? Provide a justification for the result.

23.

What does this assumption actually imply? Make an argument for why,

1. this is the simplest possible model

2. this is the most complex model

We can think of it as the M3 in Figure 1, it sets possibility distribution to which means it treats all data sets equally, so it’s the simplest possible model because it simply includes all given data sets. However, in this case, the prior , the parameterization becomes very important, and the quantities of observed data will affect the function a lot, since the mean of possibility will not be any longer, it will be most complex model.

24.

How have the choices we made above restricted the distribution of the model? What data sets are each model suited to model? What does this actually imply in terms of uncertainty? In what way are the different models more flexible and in what way are they more restrictive? Discuss and compare the models to each other?

model1 only restricts the first dimension of , so it can imply the uncertainty of , and it’s more flexible if the data sets have the decision boundary about function of but not . But it cannot model decision boundaries if the data sets has rotation invariance.

model2 restricts both dimension of , it can imply the uncertainty of , due to rotation invariance, it’s more flexible if the data sets have compared to model1.

model3 is a standard logistic regression with a bias weight which can account data sets with unequal distribution of +1 and −1.

If data set is not well modeled by any sharp linear boundary, none of models above is likely.

25.

What does this choice of prior imply? How does the choice of the parameters of the prior μ and Σ effect the model?

It implies that in each dimension, is a Gaussian distribution with mean 0 and variance, the models with usingcorrespond to a sharp linear boundary in x space, the models with a bias weight can account for data domain , the models which add times restrict the distribution of , but cover the most data sets. However, if the variance of is too small, adding restrict to will cause data loss, affect the data learning.

30.

Summarise the assignment in one paragraph, what have you learnt and what do you feel have been the prupose/message of performing this.