# Models

**Anonymous Author(s)**
Affiliation
Address
email

**Machine Learning Coursework**

Ben Norris - bn15932
Bilal Kazi - bk15841
Kyle Welch - kw15469
Greg Sims - gs15687

### 1

a) The uncertainty in $y_i$ can be modelled with a single variate $\epsilon$ drawn from a zero mean Gaussian with variance $\sigma^2$. So

$$y_i = f(\boldsymbol{x_i}) + \epsilon$$

If we assume $\boldsymbol{y_i} \in \boldsymbol{Y}$ is independent of one another, we know due to the central limit theorem that their normalized sum tends towards a Gaussian distribution. From this, we can derive the likelihood to be

$$p(y_i|f,\boldsymbol{x_i}) \sim \mathcal{N}(f(\boldsymbol{x_i}),\sigma^2 I)$$

b) A spherical (or isotropic) covariance matrix is when the covariance matrix is proportional to the identity matrix, which means it is diagonal, and all diagonal elements are exactly the same. In context this means that there is assumed to be no correlation between each dimension.

### 2

If we don't assume independence then we can't express the probability as a product of composite probabilities. We therefore have to express the probability in terms of each element.

$$p(\mathbf{Y}|f,\mathbf{X}) = p(\bigcap_{i}^{N}(y_i|f,x_i))$$

### 3

$$p(\mathbf{Y}|\mathbf{X},\mathbf{W}) = \prod_{n=1}^{N} N(\boldsymbol{y_i}|\boldsymbol{W}^T\boldsymbol{\phi}(\boldsymbol{x_n}),\sigma^2 I)$$

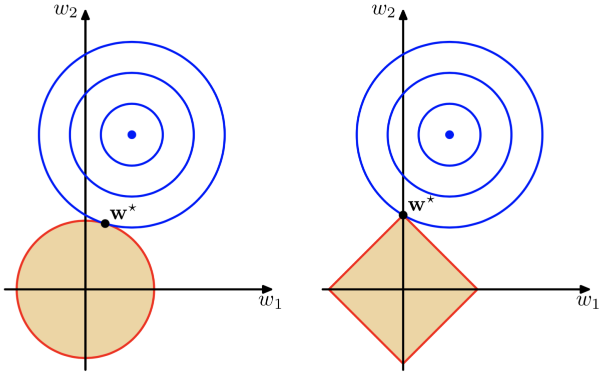
### 4

Conjugate distributions occur when a prior and a posterior distribution are of the same family, in this case the prior is called a conjugate prior.

A conjugate prior is helpful, as if we choose to use a Gaussian conjugate prior, it ensures that our posterior is also a Gaussian, since Gaussian distributions are conjugate to themselves. A conjugate prior gives a closed-form expression for the posterior, which prevents the need for integration which may be otherwise necessary. Additionally, conjugate priors can more easily show how a likelihood function updates the prior.

## 5

A Gaussian distribution is parameterised on the L2 distance of a point from the mean of the distribution. If we encode the preference using L1 norm it will change the shape of out prior (Laplace distribution). The change in shape of prior would also change the value of learned paramaters since different priors mean different things.



On the left is a prior encoded using L2 norm and right is prior encoded using L1 norm

As you can see in the image the shape of the prior determines the parameters that are learned. The L2 norm places an equal importance in every direction whereas the L1 norm tends to bias parameters more toward the axes. This results in the L1 norm preferring certain dimension more than others.

## 6

a) We can derive our posterior from Bayes Theorem. From this we know

$$p(\boldsymbol{W}|\boldsymbol{X},\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{W})p(\boldsymbol{W})$$

We assume that each observation $\boldsymbol{y_i} \in Y$ is independent from one another, giving

$$p(Y|X,W) = \prod_{i=1}^{N} p(y_i|\boldsymbol{x_i},\boldsymbol{W})$$

We model an observation $y_i$ by the equation $y_i = \boldsymbol{W}\boldsymbol{x_i} + \epsilon$ where $\epsilon$ represents Gaussian noise with 0 mean and variance $\sigma^2$. Therefore $p(Y|X,W)$ can be modelled:

$$\prod_{i=1}^{N} \mathcal{N}(y_i|\boldsymbol{W}\boldsymbol{\phi}(\boldsymbol{x_i}),\sigma^2 I)$$

To derive the posterior, we shall substitute in the formula for the multivariate Gaussian, then we will focus on just the exponent of the $e$ since

$$p(\boldsymbol{W}|\boldsymbol{X},\boldsymbol{Y}) \propto \prod_{i=1}^{N} e^{-\frac{1}{2}(y_i-\boldsymbol{W}\boldsymbol{x_i})^T(\sigma^2 I)^{-1}(y_i-\boldsymbol{W}^T\boldsymbol{x_i})} e^{-\frac{1}{2}(\boldsymbol{W}-\boldsymbol{W_0})^T(\tau^2 I)^{-1}(\boldsymbol{W}-\boldsymbol{W_0})}$$

$$= e^{-\frac{1}{2}\sum_{i=1}^{N}((y_i-\boldsymbol{W}\boldsymbol{x_i})^T(\sigma^2 I)^{-1}(y_i-\boldsymbol{W}\boldsymbol{x_i}))} e^{-\frac{1}{2}(\boldsymbol{W}-\boldsymbol{W_0})^T(\tau^2 I)^{-1}(\boldsymbol{W}-\boldsymbol{W_0})}$$

This expands to

$$e^{\frac{-1}{2\sigma^2}\boldsymbol{Y}^T\boldsymbol{Y}+\frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{X}\boldsymbol{W})-\frac{1}{2\sigma^2}(\boldsymbol{X}\boldsymbol{W})^T(\boldsymbol{X}\boldsymbol{W})} e^{-\frac{1}{2\tau^2}\boldsymbol{W}^T\boldsymbol{W}+\frac{1}{\tau^2}\boldsymbol{W}^T\boldsymbol{W_0}-\frac{1}{2\tau^2}\boldsymbol{W_0}^T\boldsymbol{W_0}}$$

Since the posterior is Gaussian, we can use the general form to derive the new parametres of the Gaussian. The exponent of the general form contains 3 key terms, a constant term (A), a mixed term (B) and a term quadratic in the parameters (C). To get the updated variance $\boldsymbol{S}^{-1}$ we set C equal to our term that is quadratic in the parameters:

$$-\frac{1}{2\tau^2}\boldsymbol{W}^T\boldsymbol{W} + \frac{1}{\tau^2}\boldsymbol{W}^T\boldsymbol{W_0} - \frac{1}{2\tau^2}\boldsymbol{W_0}^T\boldsymbol{W_0} - \frac{1}{2\sigma^2}(\boldsymbol{XW})^T(\boldsymbol{XW}) = \boldsymbol{W}^T\boldsymbol{S}^{-1}\boldsymbol{W}$$
$$\implies \boldsymbol{S}^{-1} = \frac{1}{\tau^2}\boldsymbol{I} + \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X} - \frac{2}{\tau^2}\boldsymbol{W_0}\boldsymbol{W}^{-1} + \frac{1}{\tau^2}\boldsymbol{W}^{-1}\boldsymbol{W_0}^T\boldsymbol{W}^{-1}$$

With this, we can calculate $\boldsymbol{\mu}$ by setting the general mixed term (B), equal to the mixed term in our exponent.

$$\boldsymbol{W}^T\boldsymbol{S}^{-1}\boldsymbol{\mu} = \frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{XW})$$
$$\implies \boldsymbol{\mu} = \frac{1}{\sigma^2}\boldsymbol{S}^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

This gives us the parameters for our posterior disribution so we can state

$$p(\boldsymbol{W}|\boldsymbol{Y},\boldsymbol{X}) \propto \mathcal{N}(\boldsymbol{\mu},\boldsymbol{S}^{-1})$$

It's form is that of a Gaussian, which is expected since we have a conjugate prior, suggesting the posterior should be Gaussian.

c) Z is used for normalisation, since the prior and posterior are conjugate, we know they are proportional, so Z can be seen as nothing more than a constant and shouldn't affect the actual model.

**7**

A non-parametric model is one that assumes the data distribution cannot be defined in terms of some finite parameters, and instead defines them by some infinite dimensional function.

The difference between parametrics and non-parametrics is mainly that the first assumes the model is entirely defined in terms of some finite parameters, which means that the parameters contain all the information on the model, and due to finite parameters, the complexity of the model is bounded and therefore not flexible. However, for non-parametric models, the model is defined by a function relationship, and so the information the function contains grows as the amount of data grows, putting no bounds on the complexity and allowing the model to be more flexible.

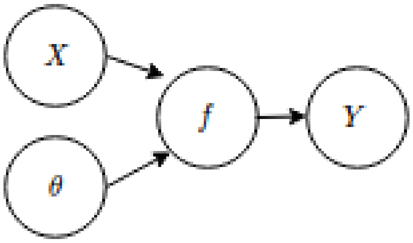
**8**

The prior is conditional on hyperparameter $\theta$ which tunes the kernel function which in turn adjusts the covariance of the Gaussian Process thus dictating the shape of the functions we prefer. This is useful as it allows us to control the types of functions from an infinite space of possible functions.

**9**

This prior encodes all possible functions, and this can be shown by looking at the spherical gaussian as a long line of "slices". Each of these "slices" is a regular gaussian, and therefore never touches zero, stretching from negative to positive infinity. In the same way, we can view the places we can take one of these slices as infinite, as the x-axis stretches from positive to negative infinity, and so we have an infinite function space.

**10**



$$p(\boldsymbol{Y}, \boldsymbol{X}, f, \boldsymbol{\theta}) = p(\boldsymbol{X})p(\boldsymbol{\theta})p(f|\boldsymbol{X}, \boldsymbol{\theta})p(\boldsymbol{Y}|f)$$

Assumptions: - All X's are independent, shown by the $k(\boldsymbol{X}, \boldsymbol{X})$

## 11

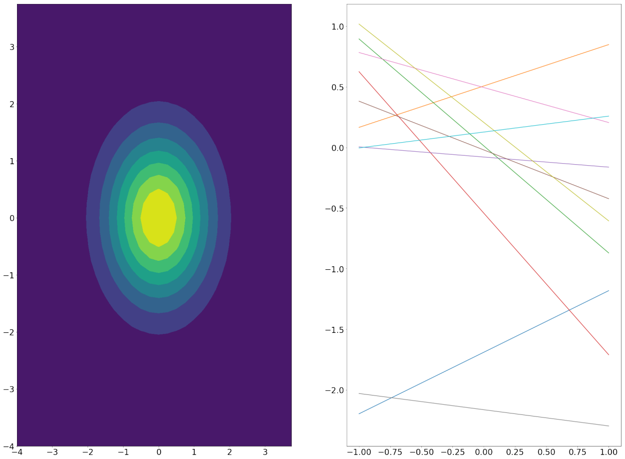
Marginalisation is a way to remove a variable that we aren't interested in, such that the information we need from them isn't lost. We take the liklihood function and multiply it by $p(f|\boldsymbol{X}, \boldsymbol{\theta})$ and integrate it with respect to $f$. This allows us to look at the set of all functions, and find an $f$ that maximises the marginal likelihood. We don't know what $f$ is chosen however this isn't important to us.

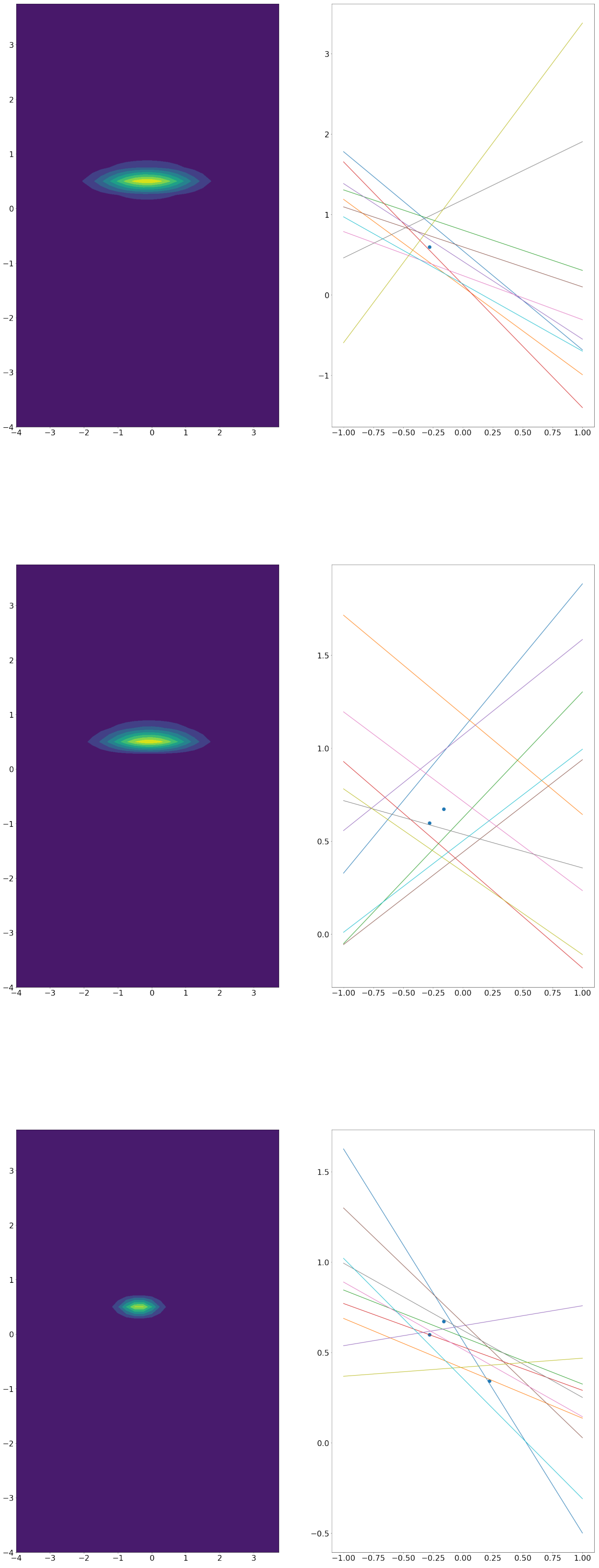
This connects the prior to the data by removing the intermediate paramater $f$ that was a function of the other 2 parameters. Instead of the data $(Y)$ relying on only some $f$, it now relies upon just the data $X$ and some parameter $\boldsymbol{\theta}$.

$\boldsymbol{\theta}$ remaining implies that $\boldsymbol{Y}$ is dependent on both the input data (X) as well as $\boldsymbol{\theta}$ and that different $\boldsymbol{\theta}$ values will result in a different $f$ that will give the maximum marginal liklihood.

## 12

On the left is the visualisation of the prior/posterior and on the right you can see a sample of functions from the prior/posterior.
Adding the first data point we see that the posterior shifts and becomes smaller and the functions seem to pass through the general area of the data point. On adding more data points the posterior starts to converge on the parameters and the functions move closer together and are very closely aligned to the direction of the data points.

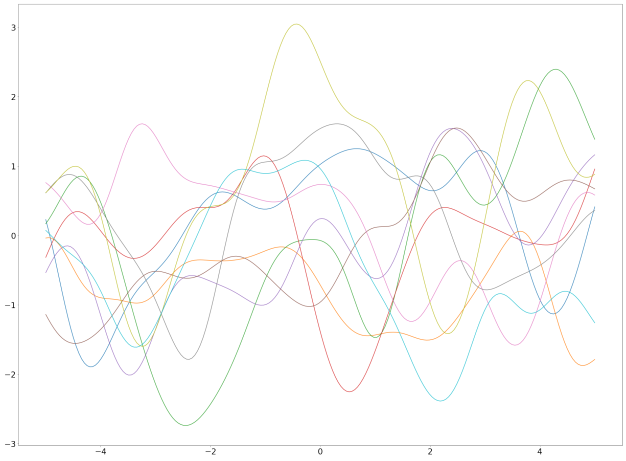


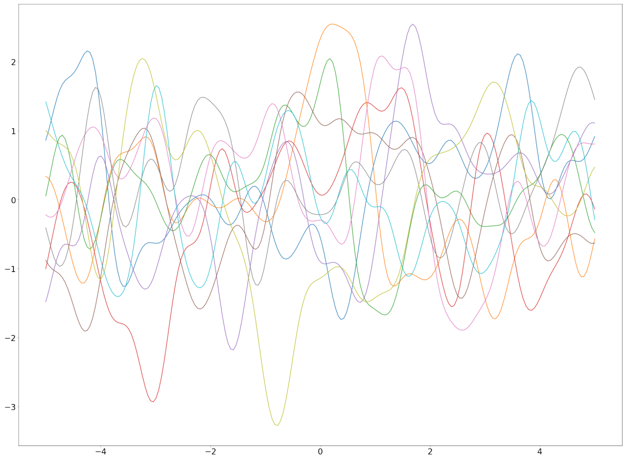

3
2
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
3
3
2
1
0
−1
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
1.5
1.0
0.5
0.0
1.5
1.0
0.5
0.0
−0.5

## 13

The first plot has value of length scale equal to 1 while the second one has length scale equal to 2. The length scale is directly proportional to the covariance of the Gaussian Process and we see that decreasing the value makes the functions noisier as the covariance decreases.

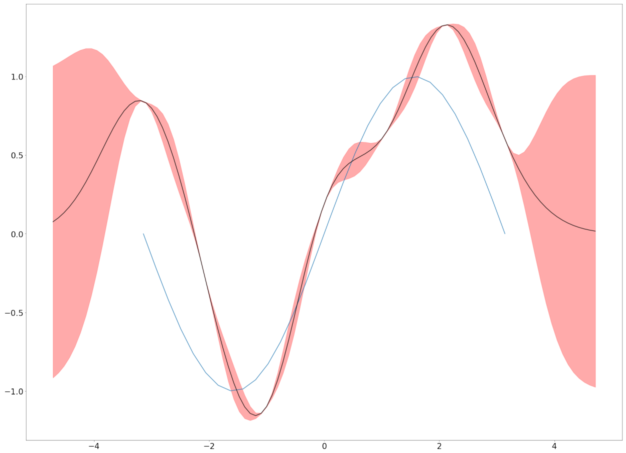
The length scale encodes the assumption of smoothness of the function.





## 14

We can see that the samples from the posterior are less varied near the data points and as we move away from the data points the uncertainity increases and the function starts to resemble the prior. This is desirable since it means that where we see data the model fits to the data and where no data is present it is represented by our prior belief but also has a high uncertainity.



### 15

The difference between a preference and an assumption is merely contextual. The overarching idea behind them is the same, however depending on the situation or context of the model you are creating, the decision is made as to whether preference or assumption is the preferred term. Both of these terms allow is to improve on our prior in some way, whether it be that we are "assuming" something about our input data, or we actually have some kind of "preference" to what form it takes.

### 16

The assumption we have encoded with this prior is that all the parameters are equally important, shown by the identity matrix as a covariance matrix. Since there is no covariance between the separate parameters, we are showing they're all equally important and independent of each other.

### 17

We know $y = Wx + \mu + \epsilon$

And the marginal distribution is a gaussian given by

$p(y) = \mathcal{N}(y \mid \mu, C)$

To derive the mean and covariance we get:

$\mathbb{E}[y] = \mathbb{E}[\boldsymbol{WX} + \mu + \epsilon] = \mu$

$cov[y] = \mathbb{E}[(\boldsymbol{WX} + \epsilon)(\boldsymbol{WX} + \epsilon)^T] = \mathbb{E}[\boldsymbol{WXX}^T\boldsymbol{W}^T] + \mathbb{E}[\epsilon.\epsilon^T] = \boldsymbol{WW}^T + \sigma^2\boldsymbol{I}$

This covariance shows that it's entirely in terms of W and no X, so we have removed X.

### 18

a) MAP and ML are different in that MAP maximises the posterior, whilst ML maximises the likelihood. However, if the prior of the distribution is a constant, MAP and ML are actually the same. Unlike MAP, ML does not weight it's data (shown by the exclusion of the prior) and therefore doesn't generate an estimate of the uncertainty of it's results. Type 2 ML marginalises out parts that we don't really need.

b) As we observe more data, you can update your posterior further and further, whilst throughout the likelihood does not particularly change. Visually, the posterior distribution homes in on the "correct answer", whilst the likelihood remains basically the same.

c) The two expressions are equal, as the denominator on the left hand side marginalises out W, and since we are maximising in terms of W, having a term without W is simply a constant, and so makes no different to our numerator.

### 19

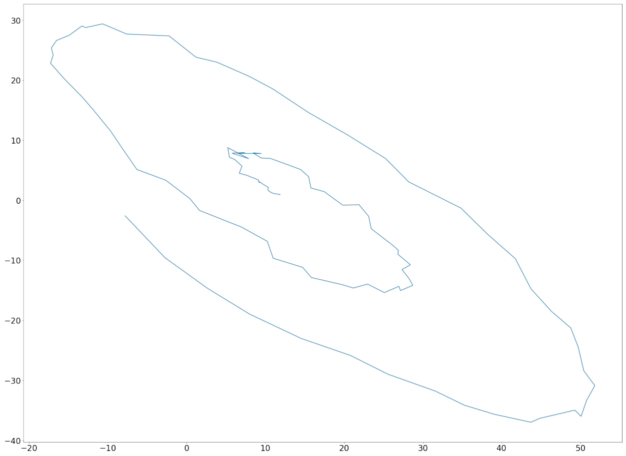
$$tr(\boldsymbol{YY}^T \tfrac{\delta \boldsymbol{C}^{-1}}{\delta \boldsymbol{W}_{ij}}) = tr(\boldsymbol{YY}^T(-\boldsymbol{C}^{-1}\tfrac{\delta \boldsymbol{C}}{\delta \boldsymbol{W}_{ij}}\boldsymbol{C}^{-1}))$$

### 20

Generally it's more simple to marginalise out $f$ since a function can easily be incorporated into the model itself, if we still have the variables that parametrize that function included in the model. Since $f$ only depends upon these parameters, it is therefore simple to marginalise. $\boldsymbol{X}$ however can be much larger since it's the input data itself and should have some effect on the output $\boldsymbol{Y}$. To marginalise $\boldsymbol{X}$ the model would have to change every time for a new input which makes the process a lot more complex. It also doesn't remove the uncertainty introduced by an intermediate function such as $f$.

### 21

Our data was generated by first passing the x through a non-linear function and then passing that output through a linear function. The linear function generates data that is 10 dimensional. This data cannot be visualised since it is in a higher dimension than we can visualise. Hence we need to learn a mapping from this higher dimensional space to one that we can visualise. Using our linear functions we can only learn the linear mapping of the data which means that what we are left with is the non-linear part of the Y which is a spiral since the non-linear part is described by [xsin(x), xcos(x)]



### 22

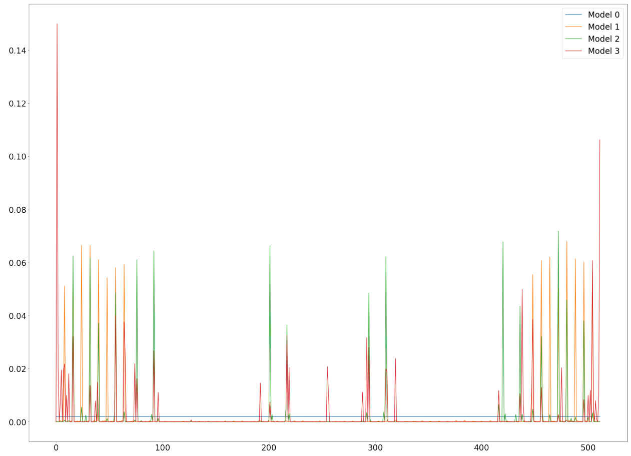
It is the simplest possible model since it places a uniform probability mass over the entire space which is mathematically very nice.

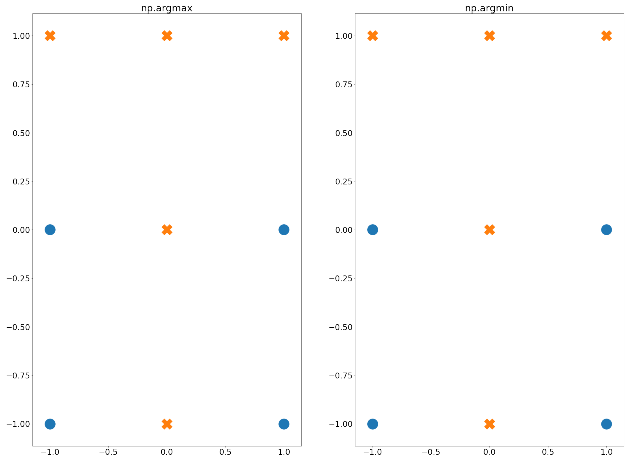
It is the most complex model since it places a uniform probability mass over the entire space and so tries to spread the probability mass over the entire dataset space which could lead to the model being penalised by Bayes' Rule.

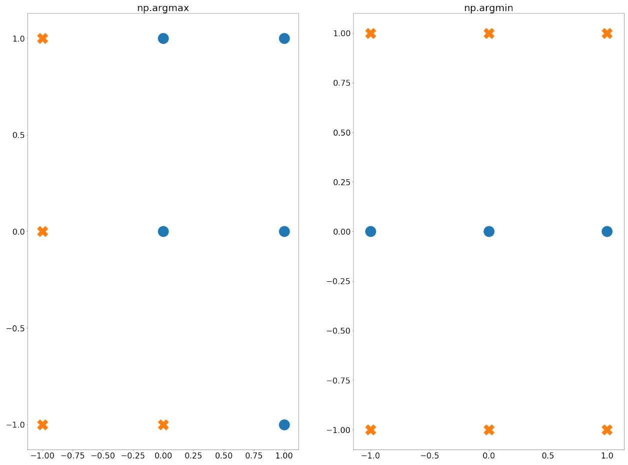
### 25

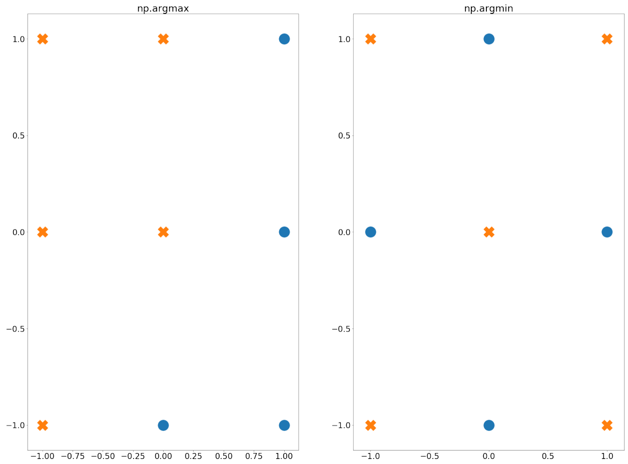
We see that the sum of the evidence over the entire dataspace for every model is about 1. This is the expected outcome since we are essentially integrating our probability distribution to get the area under the curve which should sum to one.

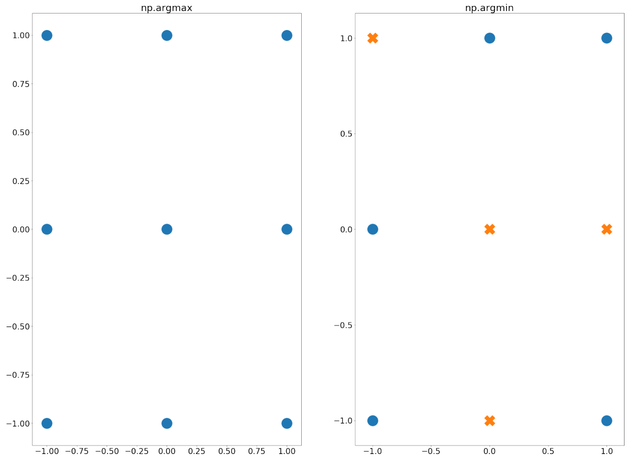
## 26 and 27


Model 0
Model 1
Model 2
Model 3
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
100
200
300
400
500


np.argmax
np.argmin
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.0
−0.5
0.0
0.5
1.0


np.argmax
np.argmin
1.0
0.5
0.0
−0.5
−1.0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.0
−0.5
0.0
0.5
1.0





## 30

Throughout this assignment, we have created and understood parametric and non-parametric models. We have learned the differences and uses of different model types and gaussian processes, as well as deriving and performing marginalisations amongst other things. We feel the purpose of performing this assignment has been to give as a "hands-on" approach to machine learning, and provide a clearer picture of what everything does.