Machine Learning, Advanced Course (DD2434)
Assignment 1

Martin Hwasser, hwasser@kth.se

November 30, 2016

Question 1. Why is the Gaussian form of the likelihood a sensible choice? What does it mean that we have chosen a spherical covariance matrix for the likelihood?

The Gaussian form of the likelihood is sensible for many reasons. The normal distribution is common in nature as well as in statistics. If we assume that our data will contain noise (random variables), the Central Limit Theorem states that as the input grows sufficiently large, the sum of these random variables will have an approximately normal distribution. In other words, if we are unsure how or why something happens, at least we know that if it happens often enough it will behave Gaussian.

A spherical covariance matrix means that the matrix is proportional to the identity matrix, such that the random variables are independent.

Question 2. If we do not assume that the data points are independent how would the likelihood look then? Remember that $Y = [y_1, \ldots, y_N]$

If the data points are not independent, we need to use the product rule:

$$p(\boldsymbol{Y} \mid f, \boldsymbol{X}) = p(y_1 \mid f, \boldsymbol{X}) p(y_2 \mid y_1, f, \boldsymbol{X}) p(y_3 \mid y_1, y_2, f, \boldsymbol{X}) \ldots p(y_n \mid y_1, \ldots y_{n-1}, f, \boldsymbol{X}), \quad (1)$$

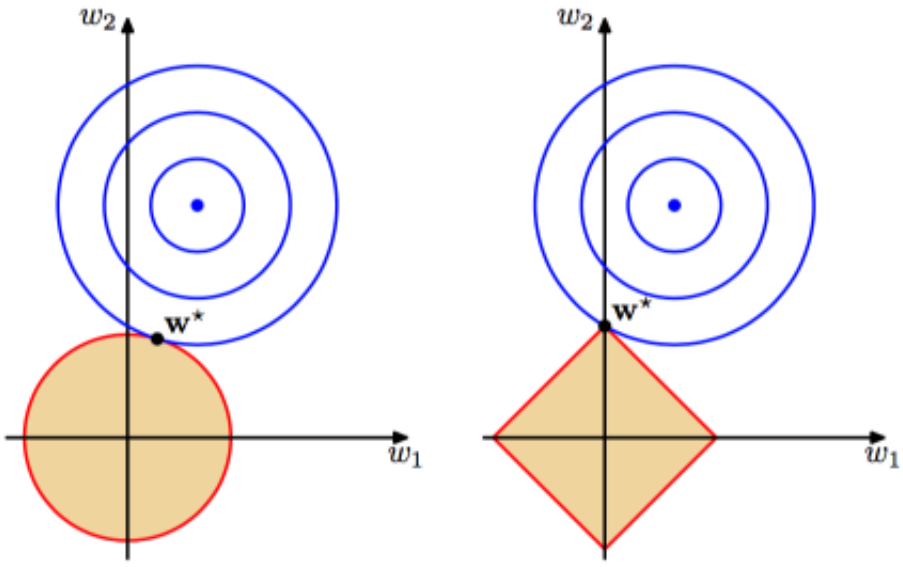
which becomes very unwieldy with many data points.

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

From Bishop (3.10):

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{W}) = \prod_{n=1}^{N} \mathcal{N}(y_n \mid \boldsymbol{W}\boldsymbol{x}_n, \sigma^2 \boldsymbol{I}) \quad (2)$$

Figure 1: L2 norm and L1 norm respectively



Question 4. Explain the concept of conjugate distributions. Why is this a motivated choice?

If the prior and the posterior distribution are in the same family of distributions, the prior and posterior are called conjugate distributions, and the prior is a conjugate prior for the likelihood function. So for example, if the likelihood function is a Gaussian, choosing a prior that is also a Gaussian distribution will result in a posterior distribution that is Gaussian. This is convenient and leads to a simplified expression for the posterior since we can determine what functional form it will have.

Question 5. The prior in Eq.8 is a spherical Gaussian. This means that the preference is encoded in terms of a L_2 distance in the space of the parameters. With this view, how would the preference change if the preference was rather encoded using a L_1 norm? Compare and discuss the different type of solutions these two priors would encode.

The L_1 leads to a sparse solution where parameters drop toward zero, since it has a preference for values near the median. This is also known as a Lasso. On the other hand, the L_2 norm can have many weights with low values, but rarely exactly zero.

The classic illustration of these two norms can be seen in figure 1.

Question 6. Derive the posterior over the parameters. Please, do these calculations by hand as it is very good practice. However, in order to pass the assignment you only need

to outline the calculation and highlight the important steps. For simplicity, please make derivations for a single output variable y. Otherwise, you would have to apply vectorization techniques.

- Why does it have the form that it does?
- What is the effect of the constant Z, are we interested in this?

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto exp\Big\{ -\frac{1}{2}\boldsymbol{X^T}\Sigma^{-1}\boldsymbol{X}\Big\}exp\Big\{\boldsymbol{X^T}\Sigma^{-1}\mu\Big\}exp\Big\{ -\frac{1}{2}\mu^T\Sigma^{-1}\mu\Big\}$$

The prior $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{\mu}_w, \tau^2\boldsymbol{I})$

$$p(\boldsymbol{w}) \propto \frac{1}{\sqrt{2\tau^2\pi}} \exp\Big\{ -\frac{1}{2\tau^2}(\boldsymbol{w} - \boldsymbol{\mu}_w)^T(\boldsymbol{w} - \boldsymbol{\mu}_w)\Big\} \tag{3}$$

The likelihood $p(y \mid \boldsymbol{w}, \boldsymbol{x}) = \mathcal{N}(\boldsymbol{xw}, \sigma^2\boldsymbol{I})$

$$p(y \mid \boldsymbol{w}, \boldsymbol{x}) \propto \frac{1}{\sqrt{2\sigma^2\pi}} \exp\Big\{ -\frac{1}{2\sigma^2}(y - \boldsymbol{xw})^T(y - \boldsymbol{xw})\Big\} \tag{4}$$

The posterior $p(\boldsymbol{w} \mid \boldsymbol{x}, y) = \mathcal{N}(\mu \mid \Sigma)$

$$p(\boldsymbol{w} \mid \boldsymbol{x}, y) \propto \frac{1}{\sqrt{2\sigma^2\pi}} \exp\Big\{ -\frac{1}{2}(\boldsymbol{w} - \mu)^T\Sigma^{-1}(\boldsymbol{w} - \mu)\Big\} \tag{5}$$

We can now write the posterior:

$$\begin{aligned} p(\boldsymbol{w} \mid \boldsymbol{x}, y) &\propto p(y \mid \boldsymbol{w}, \boldsymbol{x})p(\boldsymbol{w}) \\ &= \frac{1}{\sqrt{2\sigma^2\pi}} \exp\{-\frac{1}{2\sigma^2}(y - \boldsymbol{xw})^T(y - \boldsymbol{xw})\}\frac{1}{\sqrt{2\sigma^2\pi}} \exp\{-\frac{1}{2\tau^2}(\boldsymbol{w} - \boldsymbol{\mu}_w)^T(\boldsymbol{w} - \boldsymbol{\mu}_w)\} \\ &= \Big\{\text{drop the constants}\Big\} \\ &= \exp\{-\frac{1}{2\sigma^2}(y - \boldsymbol{xw})^T(y - \boldsymbol{xw}) - \frac{1}{2\tau^2}(\boldsymbol{w} - \boldsymbol{\mu}_w)^T(\boldsymbol{w} - \boldsymbol{\mu}_w)\} \end{aligned} \tag{6}$$

Looking at the exponent, we have:

$$\frac{1}{2\sigma^2}[\boldsymbol{w}^T\boldsymbol{w} - 2\boldsymbol{w}^T\boldsymbol{\mu}_w + \boldsymbol{\mu}_w^T\boldsymbol{\mu}_w] - \frac{1}{2\tau^2}[\boldsymbol{x}^T\boldsymbol{x}\boldsymbol{w}^T\boldsymbol{w} - 2y(\boldsymbol{xw}) + y^2] \tag{7}$$

We complete the squares:

$$\overbrace{\frac{\boldsymbol{w}^T\boldsymbol{w}}{2\tau^2} + \frac{\boldsymbol{x}^T\boldsymbol{x}\boldsymbol{w}^T\boldsymbol{w}}{2\sigma^2}}^{\text{quadratic in } \boldsymbol{w}} = \frac{\boldsymbol{w}^T\Sigma^{-1}\boldsymbol{w}}{2} \tag{8}$$

$$\overbrace{\frac{\boldsymbol{w}^T\boldsymbol{\mu}_w}{\tau^2} + \frac{y\boldsymbol{x}^T\boldsymbol{w}}{\sigma^2}}^{\text{linear in } \boldsymbol{w}} = \boldsymbol{w}^T\Sigma^{-1}\mu \tag{9}$$

$$\overbrace{\frac{\boldsymbol{\mu}_w^T\boldsymbol{\mu}_w}{2\tau^2} + \frac{y^2}{2\sigma^2}}^{\text{constant in } \boldsymbol{w}} = \frac{\mu^T\Sigma^{-1}\mu}{2} \tag{10}$$

From the quadratic term we can find the Σ^{-1} of the posterior:

$$\Sigma^{-1} = \frac{\boldsymbol{x}^T\boldsymbol{x}}{\sigma^2} + \frac{1}{\tau^2} \tag{11}$$

From the linear term, by replacing Σ^{-1}, we find μ:

$$\mu = (\frac{\boldsymbol{x}^T\boldsymbol{x}}{\sigma^2} + \frac{1}{\tau^2})^{-1}(\frac{y\boldsymbol{x}}{\sigma^2} + \frac{\boldsymbol{\mu}_w}{\tau^2}) \tag{12}$$

The effect of the constant Z is to normalize the probabilities so that they sum to 1.

Question 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:

- Representability?
- Interpretability?

The name non-parametric model is somewhat of an unfortunate misnomer. A non-parametric model can, just like a parametric model, have parameters, but these will concern the model complexity rather than the distribution. For a non-parametric model, predicting unseen data is not only based on given parameters but also the data that has been observed so far. In contrast to the parametric model, the non-parametric model does not assume a specific form of distribution.

Parametric models are easier to interpret, but might lead to a model that does a poor job of representing the data. For example, with a parametric model we can never represent data whose complexity is larger than we presume. Non-parametric models can usually represent the data better since the models can be arbitrarily complex. Non-parametric models can also use the kernel trick to represent data in higher dimensions. This also makes non-parametric models much harder to interpret.

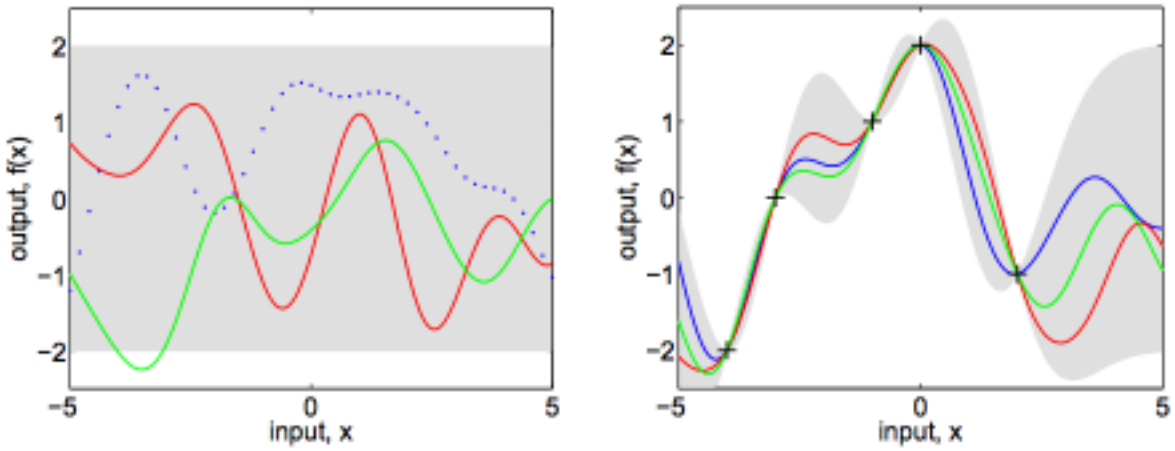


Figure 2: (a) shows the prior, where the dotted line is the true function y we are looking for. (b) reflects our posterior, which is the prior conditioned by the five points. The shaded area is 2 standard deviations from the mean. The colored lines are sampled from the prior and posterior respectively.

Question 8. Explain what this prior does? Why is it a sensible choice? Use images to show your reasoning. Clue: use the marginal distribution to explain the prior

---

$$p(f \mid \boldsymbol{X}, \theta) = \mathcal{N}(0, k(\boldsymbol{X}, \boldsymbol{X}))$$

The prior allows us to express uncertainty in f. The covariance (kernel function) says that for similar values of $\boldsymbol{x}$, the output f will be similar. For the assignment, we used the squared exponential kernel. The mean assumes that the values of f will distributed around 0. The prior is sensible choice because it means that we can define points nearby each other to be highly correlated, which will strengthen our confidence as we see more data. Similarly, we can express that points that are far away from each other will have low correlation. In a sense it's telling us that we are confident in areas with a lot of data, and unsure in areas where we have little or no data. This is illustrated in figure 2. Moreover, it's flexible since θ allows us to specify how much or how little the variables should correlate.

Question 9. Formulate the joint likelihood of the full model that you have defined above,

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

Try to draw a very simple graphical model to clearly show the assumptions that you have made.

First we use the product rule:

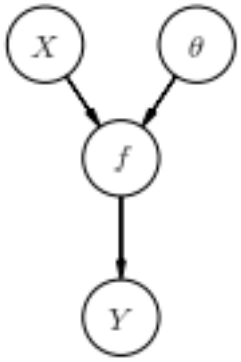


Figure 3: Graphical model illustrating our assumptions.

$$p(\boldsymbol{Y}, \boldsymbol{X}, f, \theta) = \underbrace{p(\boldsymbol{Y} \mid \boldsymbol{X}, f, \theta)p(\boldsymbol{X} \mid f, \theta)p(f \mid \theta)p(\theta)}_{\text{product rule}} \tag{13}$$

We make the assumption that $\boldsymbol{Y}$ is independent of $\boldsymbol{X}$ and θ given f.

$$p(\boldsymbol{Y} \mid f)p(f \mid \boldsymbol{X}, \theta)p(\boldsymbol{X})p(\theta) \implies \underbrace{p(\boldsymbol{Y} \mid \boldsymbol{X}, f)}_{\text{model assumption}} \tag{14}$$

The figure 3 illustrates the assumptions.

Question 10. Explain the marginalisation in Eq.12,

- Explain how this connects the prior and the data?
- How does the uncertainty filter through this?
- What does it imply that is left on the left-hand side of the expression after marginalisation?

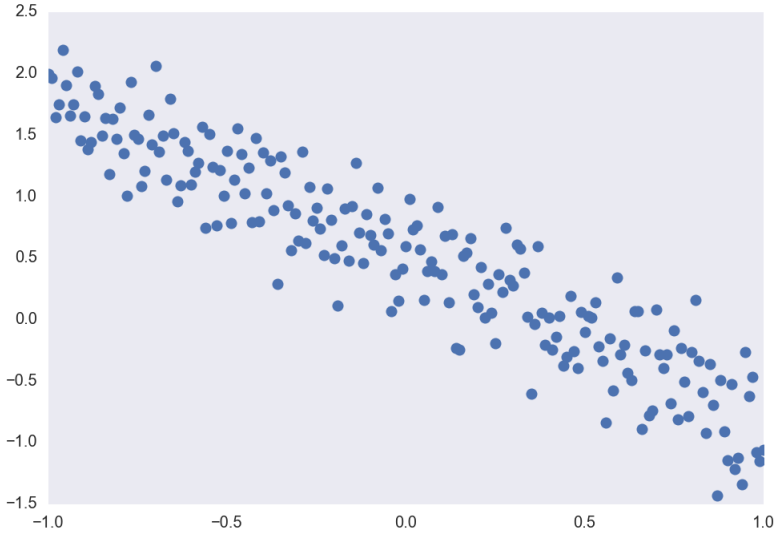
$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \theta) = \int p(\boldsymbol{Y} \mid f)p(f \mid \boldsymbol{X}, \theta)df$$

By marginalizing, we are dropping f. The prior is connected to the data since f is dependent on $\boldsymbol{X}$ and θ, and after marginalizing, $\boldsymbol{Y}$ is directly dependent on $\boldsymbol{X}$ and θ.

The uncertainty is "filtered" or "propagated" through the prior by integrating all values of the possible functions f. The uncertainty of $p(\boldsymbol{Y} \mid f)$ is now expressed by $p(\boldsymbol{Y} \mid \boldsymbol{X}, \theta)$. The more sure we are of the prior, the more data we will need to "correct" the posterior.

We keep θ on the left-hand side, since the distribution will still depend on it. We keep it as a hyperparameter to be able to tweak our model.

Figure 4: Noisy data from **Question 11**



Question 11.

- Visualize the prior distribution over W.
- Pick a single data-point from the data and visualize the posterior distribution over W.
- Sample from the posterior and show a couple of functions.
- Repeat 2-3 by adding additional data points.

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

The prior distribution over $\boldsymbol{W}$ is:

$$p(\boldsymbol{W}) = \mathcal{N}(\boldsymbol{\mu}_w, \tau^2 \boldsymbol{I})$$

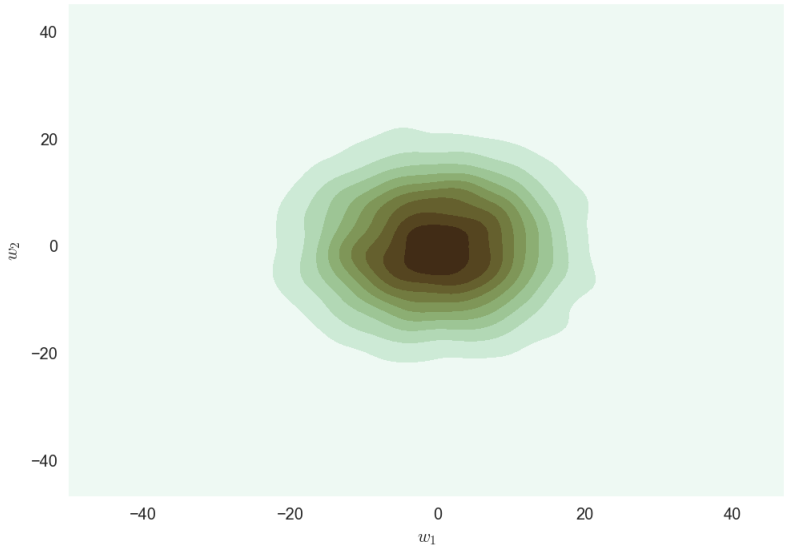
We have no prior information about the mean and covariance, so we can set $\boldsymbol{\mu}_w = (0, 0)$ as it is convenient and symmetrical. Being Bayesian, we can express uncertainty by assuming a large covariance matrix:

$$\Sigma = \begin{bmatrix} 10^2 & 0 \\ 0 & 10^2 \end{bmatrix}$$

which means high variance and implies that we have no strong beliefs. By letting Σ be a diagonal matrix, we are also assuming independence between the variables. In the particular case of this assignment, a large covariance like this leads to slower convergence, but we can be sure that the data has "spoken for itself". Figure 5 illustrates this prior.

As we can see in figure 6, by adding more samples from the data, the parameters converge toward $(-1.3, 0.5)$.

Figure 5: Prior



Question 12.

- Create a GP-prior with a squared exponential covariance function.
- Sample from this prior and visualize the samples.
- Show samples using different length-scale for the squared exponential.

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

A high l-value means the curves are smoother. The l-value is a constraint between points that define how closely they correlate. A large l-value means that the distance between the points matters less. Changing the l-value does not affect the diagonal, but with a higher l-value, variables are highly correlated, and conversely, with a lower l-value, the correlation between variables fall off so that they depend less on each other, and the graph becomes wiggly. As the l-value approaches zero, the covariance matrix becomes closer to the identity matrix. A low l-value will lead to overfitting of the data, and a high value will lead to underfitting.

Figure 7 shows varying l-values.

Question 13. The posterior and the prior are the same object if we do not have any observed data.

Explain the above statement, why is this?

The posterior is the probability of a certain to occur after data has been observed. The prior reflects the subjective beliefs about some phenomena. By combining the prior with

Figure 6: Posterior and corresponding samples

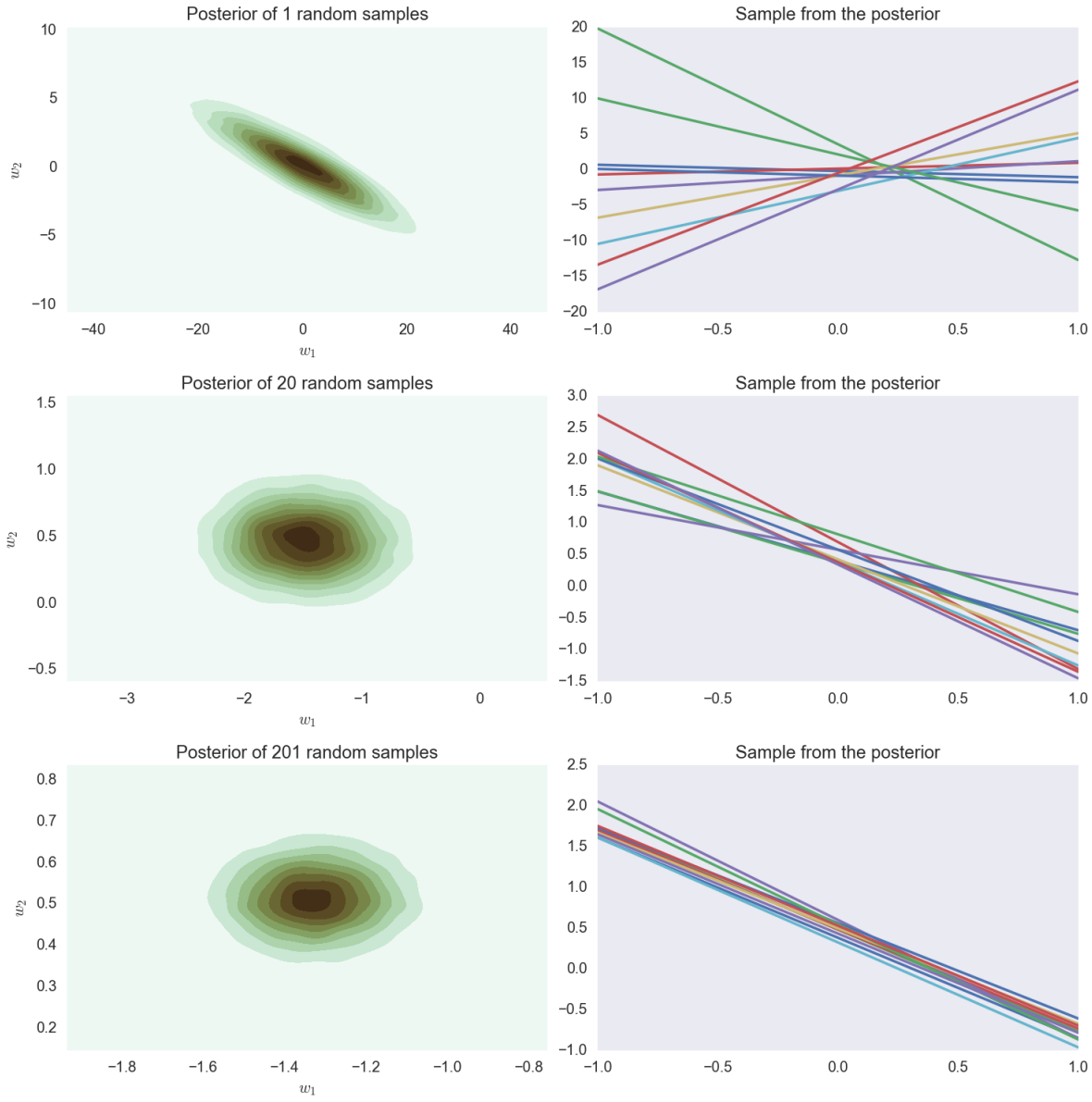


Figure 7: Samples from the prior with different l-values

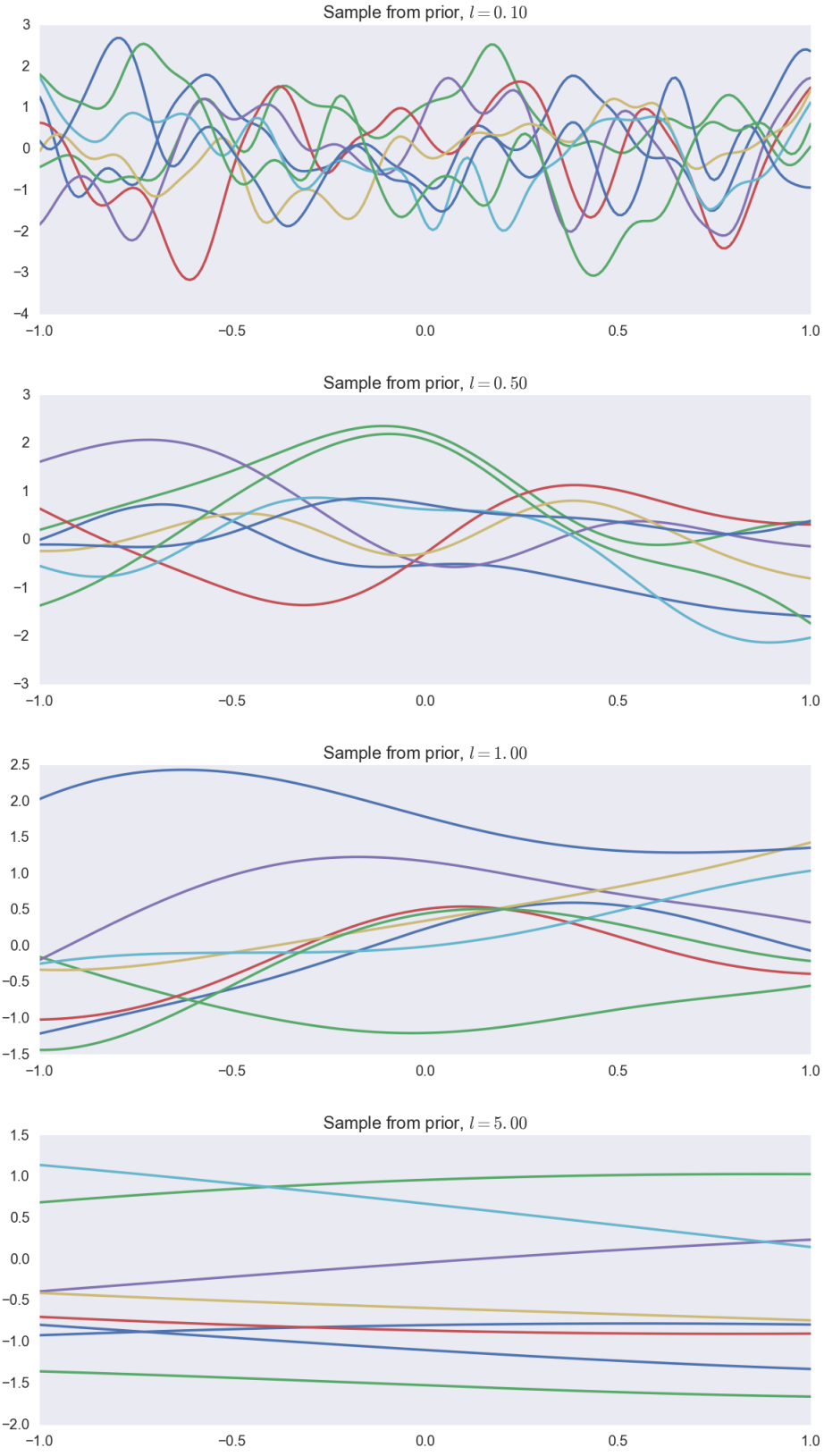
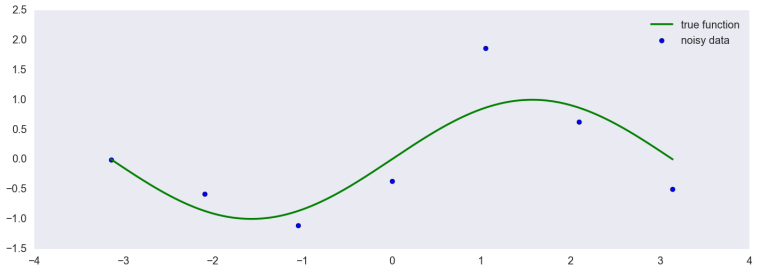


Figure 8: True function & noisy data



evidence of gathered data, we obtain the posterior using Bayes' theorem. If the prior is uninformative, the posterior will be strongly based on data. On the other hand, if the prior is informative, the posterior will be a mixture of the prior and the data. If we don't have any data at all, the prior and the posterior will thus be the same object.

Question 14.

- Compute the predictive posterior distribution of the model
- Sample from this posterior with points both close to the data and far away from the observed data.
- Plot the data, the predictive mean and the predictive variance of the posterior from the data.

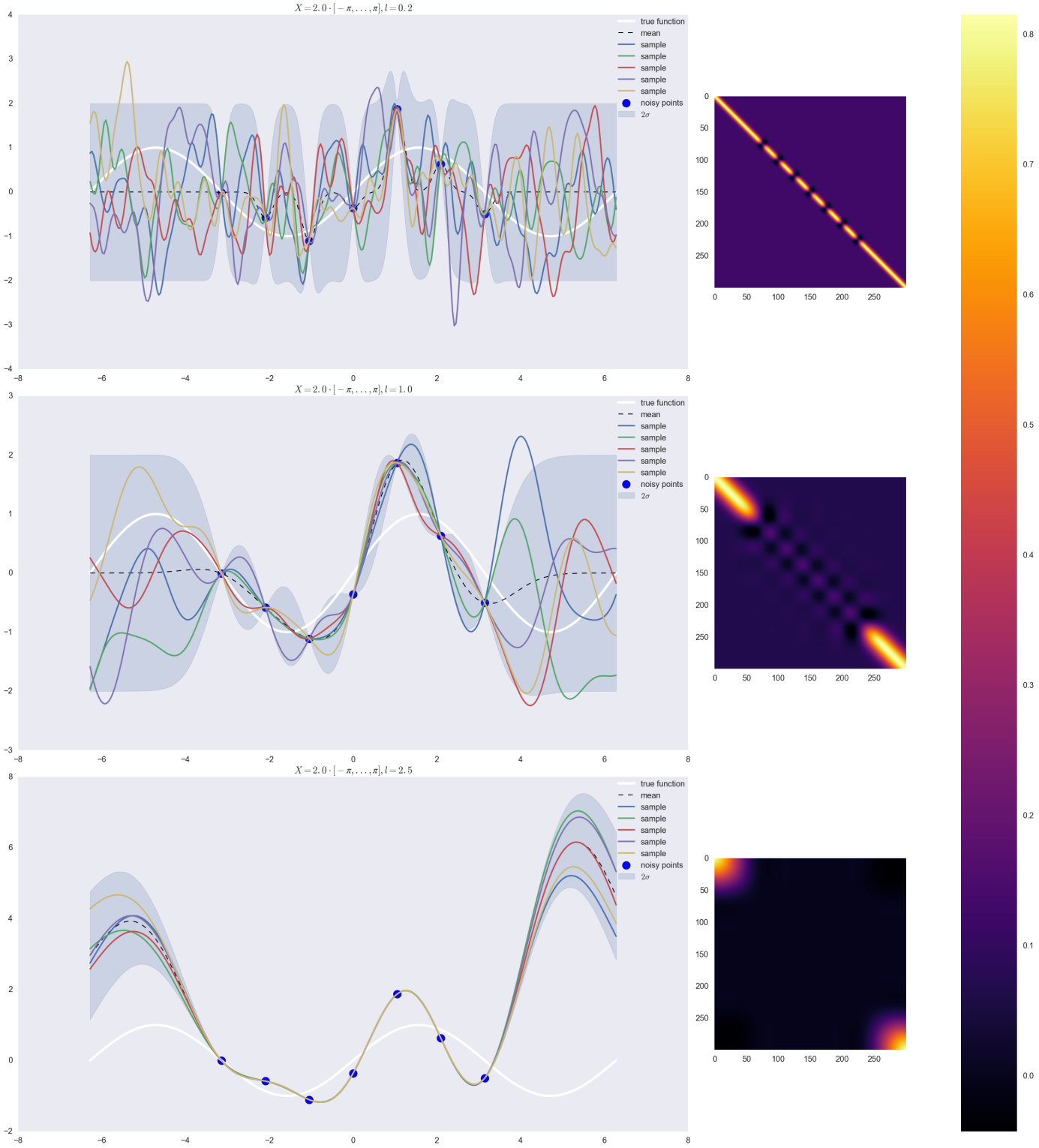
Explain the behavior of the samples and compare the samples of the posterior with the ones from the prior. Is this behavior desirable? What would happen if you would add a diagonal covariance matrix to the squared exponential?

The true function and the noisy data can be seen in figure 8. From figure 9, one can see that within the range $[-\pi, ..., \pi]$, given the "right" l-value, the model fits the data quite well as desired. The variance increases between each point, and outside the input range the variance is wild. This is especially visible in the covariance plots, where with a low l-value, it's easy to distinguish the points which appear as dark spaces between areas of larger variance. With a higher l-value, the variance drops off between the points as the points have less and less correlation between each other.

If we add a diagonal covariance matrix to the squared exponential, the posterior will have more variance, and the samples from the posterior will not necessarily go through the points.

Question 15. Elaborate on this, why can one view a prior as encoding a preference?

Figure 9: Samples from the posterior with different l-values, with corresponding illustrations of covariance.



A prior encodes a preference in the sense that the more informative the prior, the more data we need to change our beliefs since the posterior becomes more influenced by the prior. That is to say that we prefer a specific model more than other models. Being Bayesian, we are also allowed to marginalize a variable that we're not interested in.

Question 16.

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

What type of "preference" does this prior encode?

This prior encodes a preference for circular data centered around 0, possibly as an initial guess without actual information on where the latent variables X is centered. The prior also says that the variables are mutually independent because of the identity covariance matrix.

Question 17. Perform the marginalisation in Eq. 23 and write down the expression. As previously, it is recommended that you do this by hand even though you only need to outline the calculations and show the approach that you would take to pass the assignment. Hint: The marginal can be computed by integrating out X with the use of Gaussian algebra we exploited in the exercise derivations and, in particular, by completing the square. However it is much easier to derive the mean and covariance, knowing that the marginal is Gaussian, from the linear equation of Y(X).

$$p(\boldsymbol{Y} \mid \boldsymbol{W}) = \int p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{W}) p(\boldsymbol{X}) d\boldsymbol{X}$$

We could complete the square of the exponent, but we know that the $P(\boldsymbol{Y} \mid \boldsymbol{W})$ is Gaussian since the right hand side is Gaussian. So we only need to find the mean and covariance for $p(\boldsymbol{Y} \mid \boldsymbol{W})$. If we assume a model $y = \boldsymbol{W}\boldsymbol{x} + \mu + \epsilon$, we can find the mean and covariance from their definitions.

Expected value:

$$\begin{aligned} \mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{W}] &= \mathbb{E}[\boldsymbol{W}\boldsymbol{x} + \mu + \epsilon] \\ &= \mathbb{E}[\boldsymbol{W}\boldsymbol{x}] + \mathbb{E}[\mu] + \mathbb{E}[\epsilon] \\ &= \boldsymbol{W}\,\mathbb{E}[\boldsymbol{x}] + \mathbb{E}[\ \mu] + \mathbb{E}[\epsilon] \end{aligned} \tag{15}$$

By assuming the noise ϵ is uncorrelated with the underlying data:

$$\implies \boldsymbol{W}\,\mathbb{E}[\boldsymbol{x}] + \mu + 0 = \mu \tag{16}$$

Covariance:

$$\begin{aligned} cov(\boldsymbol{Y} \mid \boldsymbol{W}, \boldsymbol{Y}) &= cov(\boldsymbol{W}\boldsymbol{X} + \epsilon, \boldsymbol{W}\boldsymbol{X} + \epsilon) \\ &= \mathbb{E}[(\boldsymbol{W}\boldsymbol{X} + \epsilon - \mathbb{E}[\boldsymbol{W}\boldsymbol{X} + \epsilon])((\boldsymbol{W}\boldsymbol{X} + \epsilon)^T - \mathbb{E}[(\boldsymbol{W}\boldsymbol{X} + \epsilon)^T])] \\ &= \mathbb{E}[(\boldsymbol{W}\boldsymbol{X} + \epsilon)(\boldsymbol{W}\boldsymbol{X} + \epsilon)^T] \\ &= \mathbb{E}[\boldsymbol{W}\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}^{\boldsymbol{T}}] + \mathbb{E}[\epsilon^2] \\ &= \boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}} + \sigma^2\boldsymbol{I} \end{aligned} \tag{17}$$

After the marginalization, $\boldsymbol{Y}$ no longer depends on $\boldsymbol{X}$.

$$p(\boldsymbol{Y} \mid \boldsymbol{W}) = \mathcal{N}(\boldsymbol{Y} \mid \mu, \boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}} + \sigma^2\boldsymbol{I}) \tag{18}$$

Question 19. Compare these three estimation procedures above in log-space.

- How are they different?
- How are MAP and ML different when we observe more data?
- Why are the two last expressions of Eq. 25 equal?

ML in log-space:

$$\underset{\boldsymbol{W}}{argmax}_{\text{ML}} = -\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{W}x_n - y_n)^2 \tag{19}$$

MAP in log-space:

$$\underset{\boldsymbol{W}}{argmax}_{\text{MAP}} = -\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{W}x_n - y_n)^2 - \frac{1}{2}\boldsymbol{W}^T\boldsymbol{W} \tag{20}$$

ML Type-II in log-space:

$$\underset{\boldsymbol{W}}{argmax}_{\text{Type-II ML}} = -\frac{N}{2}(D \cdot \log 2\pi + \log|\boldsymbol{C}| + trace(\boldsymbol{C}^{-1}S)) \tag{21}$$

where

$$\boldsymbol{C} = \boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}} + (\sigma^2\mathcal{I}), S = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2$$

First of all, the ML approach finds the mode of the likelihood distribution, and MAP finds the mode of the posterior distribution. In constrast to ML, MAP estimation is Bayesian in the sense that it it looks at the prior of the parameters. One advantage with MAP is

that it has built-in regularization due to the uncertainty which decreases the variance and prevents overfitting. Lacking this term to penalize heavy variations, the ML estimation will tend to easily overfit when there's not so much data. However, with plenty of data, both ML and MAP will converge to the same point. The ML Type-II estimate is especially useful in some situations, since it does not depend on $\boldsymbol{X}$.

The two last expressions are equal since we can disregard the denominator when we are just looking for the $\boldsymbol{W}$ that maximizes the equation. The denominator is just a scaling factor.

Question 19.

- Write down the objective function $-log(p(Y|W)) = \mathcal{L}(W)$ for the marginal distribution in Eq. 23.
- Write down the gradients of the objective with respect to the parameters $\frac{\delta \mathcal{L}}{\delta \boldsymbol{W}}$

$$p(\boldsymbol{Y} \mid \boldsymbol{W}) = \int p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{W}) p(\boldsymbol{X}) d\boldsymbol{X} = \tag{22}$$

$$\frac{1}{(2\pi)^{\frac{ND}{2}}} \frac{1}{\mid \Sigma \mid^{\frac{N}{2}}} \exp\left\{ \sum_{i=1}^{N} - \frac{1}{2} y_i^T \Sigma^{-1} y_i \right\} \tag{23}$$

The cost function $\mathcal{L}(\boldsymbol{W})$:

$$\begin{aligned} \mathcal{L}(\boldsymbol{W}) &= -\log(p(\boldsymbol{Y} \mid \boldsymbol{W})) \\ &= \frac{ND}{2} \log 2\pi + \frac{N}{2} \log |\boldsymbol{C}| + \sum_{i=1}^{N} \frac{1}{2} y_i^T \boldsymbol{C}^{-1} y_i \\ &= \frac{N}{2} \Big(D \log 2\pi + \log |\boldsymbol{C}| + (\underbrace{Tr(\boldsymbol{C}^{-1} y y^T)}_{\text{Matrix Cookbook (16)}}) \Big) \end{aligned} \tag{24}$$

where

$$\boldsymbol{C} = \boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}} + \sigma^2 \boldsymbol{I} \tag{25}$$

The gradient is given by:

$$\frac{\delta \mathcal{L}}{\delta \boldsymbol{W}} = -N(\boldsymbol{C}^{-1}\boldsymbol{S}\boldsymbol{C}^{-1}\boldsymbol{W} - \boldsymbol{C}^{-1}\boldsymbol{W}) \tag{26}$$

Question 20. 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. Discuss any invariance you observe.

Reducing 10-dimensional data down to 2 dimensions has resulted in some information loss, as seen in figure 10. I was not surprised by this since we discarded the original data, added noise and projected the data into a higher dimension. I was actually surprised that it captured the basic structure, however the scale is somewhat different and the data appears rotated. This is because there's no one unique solution for $\boldsymbol{W}$, the marginal likelihood is invariant to rotation, since given an orthogonal rotation matrix $\boldsymbol{R}$, such that:

$$\boldsymbol{R}\boldsymbol{R}^{\boldsymbol{T}} = \boldsymbol{I}$$

we see that the covariance matrix $\boldsymbol{C}$ is independent of $\boldsymbol{R}$:

$$\begin{aligned}
\boldsymbol{C} &= \tilde{\boldsymbol{W}}\tilde{\boldsymbol{W}}^{\boldsymbol{T}} \\
&= \boldsymbol{W}\boldsymbol{R}\boldsymbol{R}^{\boldsymbol{T}}\boldsymbol{W}^{\boldsymbol{T}} \\
&= \boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}} \\
&\implies \tilde{\boldsymbol{W}} = \boldsymbol{W}\boldsymbol{R} \\
&\implies \mathcal{N}(\mu, \tau + \boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}}) = \mathcal{N}(\mu, \tau + \boldsymbol{W}\boldsymbol{R}\boldsymbol{R}^{\boldsymbol{T}}\boldsymbol{W}^{\boldsymbol{T}})
\end{aligned} \tag{27}$$

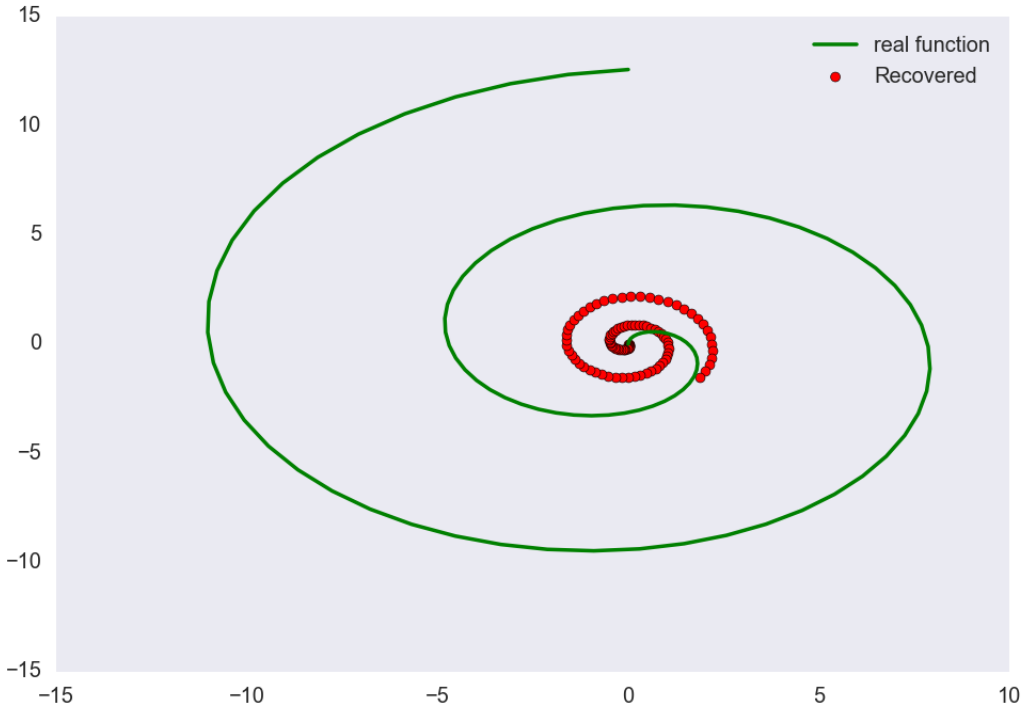
Question 21. Why is this the simplest model, and what does it actually imply? Discuss its implications, why is this a bad model and why is it a good model?

$\mathcal{M}_0$ is the simplest model in terms of form since it takes 0 parameters. It's a bad model since it actually knows nothing about $\mathcal{D}$ except its cardinality. It spreads the probability density uniformly over the entire dataset, which implies that it has no preference at all among the datasets. However, in terms of Occam's razor, it is quite complex since it considers all datasets. It's good at reflecting our lack of knowledge in the absence of data.

Question 22. Explain how each separate model works. In what way is this model more or less flexible compared to $\mathcal{M}_0$? How does this model spread its probability mass over $\mathcal{D}$?

The model $\mathcal{M}_1$ divides the probability distribution according to the data it is given, but only by looking at at one axis. It's more flexible than $\mathcal{M}_0$ since it takes one parameter. It can also realize model $\mathcal{M}_0$, and this is true for each subsequent model. For example, $\mathcal{M}_3$ can take the form of $\mathcal{M}_2$ by setting the bias term to 0, and $\mathcal{M}_2$ can take the form of $\mathcal{M}_1$ by setting θ_1 to 0. In terms of model complexity, $\mathcal{M}_1$ is simpler than $\mathcal{M}_0$ since it will consider less datasets.

Figure 10: Recovering latent variables



Question 23. How have the choices we made above restricted the distribution of the model? What datasets 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.

The models $\mathcal{M}_2$ and $\mathcal{M}_3$ are restricted to a more narrow subset of datasets. Having more parameters causes the models to give more probability distribution to complex datasets at the cost of giving lower probability to simpler datasets. Model $\mathcal{M}_2$ & $\mathcal{M}_3$ are more flexible since they accept more parameters.

Model $\mathcal{M}_2$ & $\mathcal{M}_3$ can only find linear decision boundaries, with the difference being that $\mathcal{M}_3$ can find boundaries that does not go through the origin. This makes it the only model that can actually look at the point in the origin.

Question 24. Explain the process of marginalization. Discuss its implications.

Since we don't know θ, we can marginalize it. By specifying a prior over θ, we are expressing our uncertainty of the parameter. The evidence for $p(\mathcal{D} \mid \mathcal{M}_i)$ is the weighted average of all possible θ in $p(\mathcal{D} \mid \mathcal{M}_i, \theta)$. The evidence is also called marginal likelihood, since the it can be interpreted as a likelihood function over the models where θ has been marginalized.

We can think of the evidence being the probability for a certain dataset, given a specific model whose parameters were sampled from the prior.

Question 25. What does this choice of prior imply? How does the choice of the parameters of the prior μ and Σ effect the model?

$$p(\theta \mid \mathcal{M}_i) = \mathcal{N}(\mu, \Sigma)$$
$$\mu = 0$$
$$\Sigma = \boldsymbol{I} \cdot 10^3$$

This prior says that the weights, θ, will be centered around 0, but that we are unsure since we assume a large variance. Moreover, we presume that the weights are mutually independent with an isotropic covariance matrix. The spread of weights will be pretty large, and there will be an even spread of positive and negative weights.

More importantly, the prior puts most of its mass on settings of the parameters that give sharp linear boundaries. This is easy to see since that when the weights are close to 0, the models will behave like $\mathcal{M}_0$ and the probability mass will be uniform.

Question 26. For each model sum the evidence for the whole of $\mathcal{D}$, what numbers do you get? Explain these numbers for all the models and relate them to each other.

The evidence for each model for the whole dataset sums up to 1. This is obvious for the first model, $\mathcal{M}_0$, since we have 512 datasets and each dataset has an evidence of $\frac{1}{512}$. For the three other models, $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$, we get a summed evidence of 1 due to the symmetry and completeness of the dataset. There are 9 points in each dataset, and given that the prior is uniform across all datasets, the expected value for a given point will be:

$$\frac{1}{1 + e^0} = 0.5$$

So for 512 datasets, and 9 point in each: $512 \cdot (\frac{1}{2})^9 = 1$

Question 27. Plot the evidence over the whole dataset for each model. The x-axis index the different instances in $\mathcal{D}$ and each models evidence is on the y-axis. How do you interpret this? Relate this to the parametrisation of each model.

The evidence over the whole dataset is shown in figure 11.

We see that $\mathcal{M}_0$ spreads its probability distribution evenly across all datasets, while $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ cover fewer. $\mathcal{M}_1$ and $\mathcal{M}_2$ can capture linear boundaries that intersects the origin, whereas $\mathcal{M}_3$ can capture linear boundaries that don't go through origin.

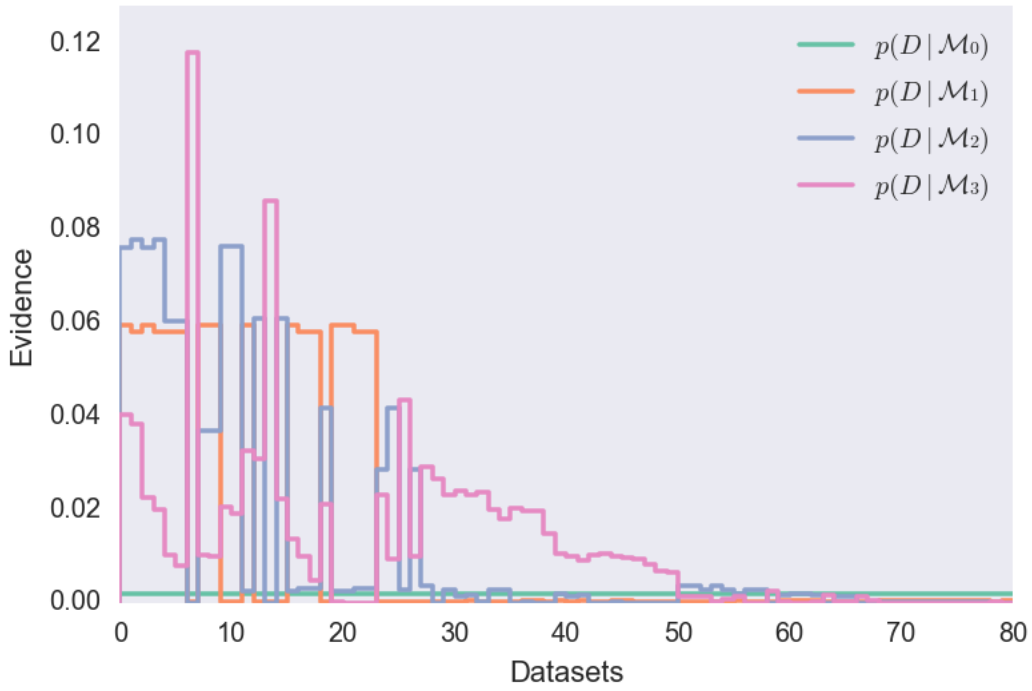


Figure 11: Evidence over the whole dataset $\mathcal{D}$

Question 28. Find using **np.argmax** and **np.argmin** which part of the $\mathcal{D}$ that is given most and least probability mass by each model. Plot the data-sets which are given the highest and lowest evidence for each model. Discuss these results, does it make sense?

The datasets giving the most and least probability mass by each model can be seen in figure 12.

In dataset (c) the decision boundary is a function of x_1 but not x_2. Model $\mathcal{M}_1$ is a simple model that only looks at one axis and captures such decision boundaries, so it makes sense that it gives (c) high probability. In dataset (e), the decision boundary is a function of both x_1 and x_2 so it makes sense that model $\mathcal{M}_2$ gives it high mass. The best dataset for $\mathcal{M}_3$ is one-sided, which is sensible since it's the only model to have a bias term θ so that it can have a decision boundary that does not go through the origin, and thus ignoring that point. The worst dataset for $\mathcal{M}_3$ is (h), which makes sense it doesn't have a linear decision boundary.

Question 29. What is the effect of the prior $p(\theta)$.

- What happens if we change its parameters?
- What happens if we use a non-diagonal covariance matrix for the prior?
- Alter the prior to have a non-zero mean, such that $\mu = (5, 5)$?
- Redo evidence plot for these and explain the changes compared to using zero-mean

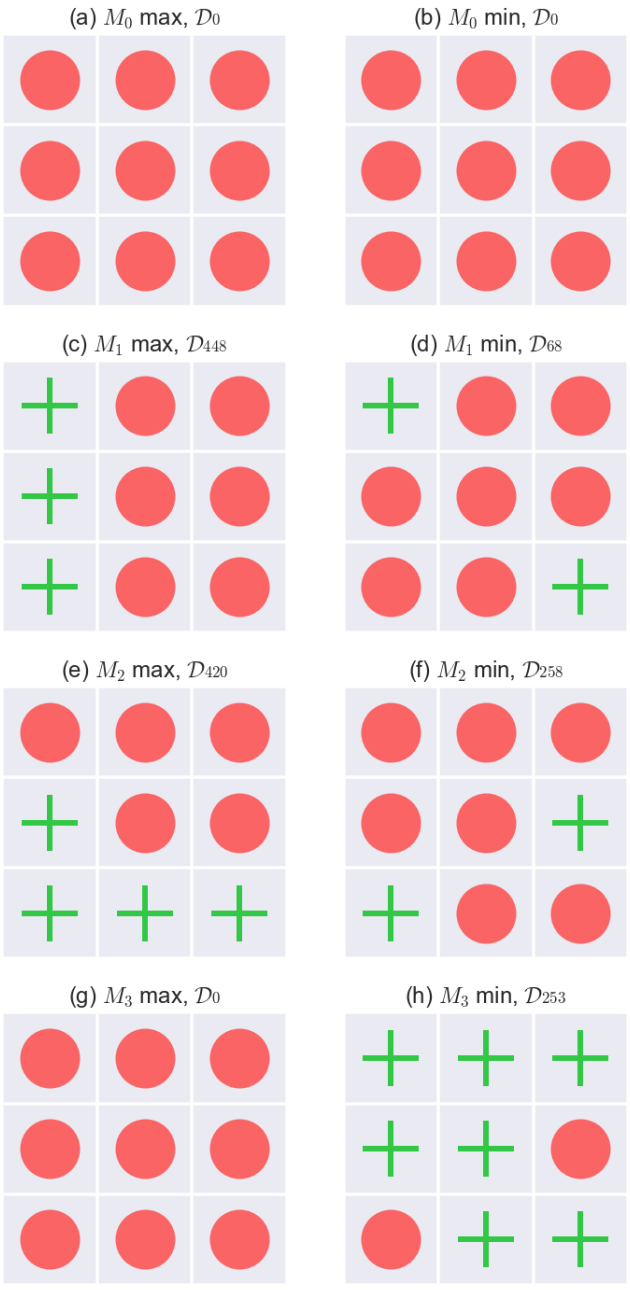


Figure 12: Probability mass by model

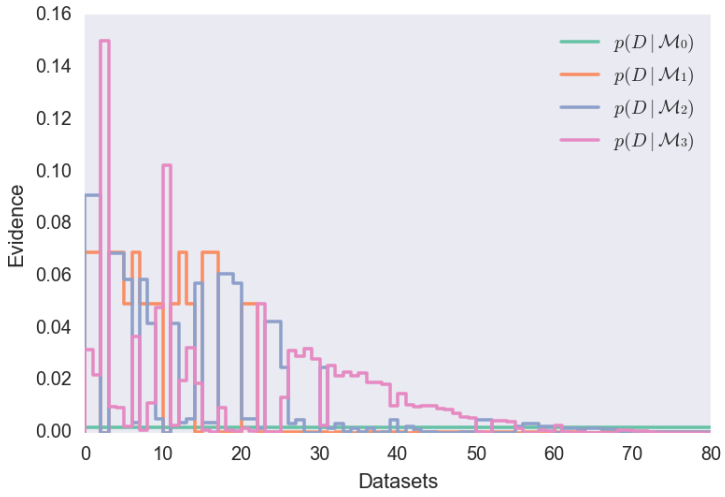


Figure 13: Probability mass by model, $\mu = (5, 5)$

When using a non-diagonal covariance matrix, the parameters θ_0, θ_1 and θ_2 will have some correlation. For example, if θ_0 correlates positively with θ_1, it will favor certain models over others. This is seen in the plots, where a few datasets have a high evidence. Figure 14 shows the evidence with a 0 mean and a covariance matrix:

$$\Sigma = \begin{bmatrix} 1000 & -355 & 444 \\ -355 & 1000 & -756 \\ 444 & -756 & 1000 \end{bmatrix}$$

Plotting the evidence for a prior with non-zero mean (figure 13), one can see that the majority of probability mass is distributed to a few datasets. This seems sensible because some datasets have decision boundaries for which the parameters correspond to zero weights. By moving the mean, these datasets will be given a lower probability mass.

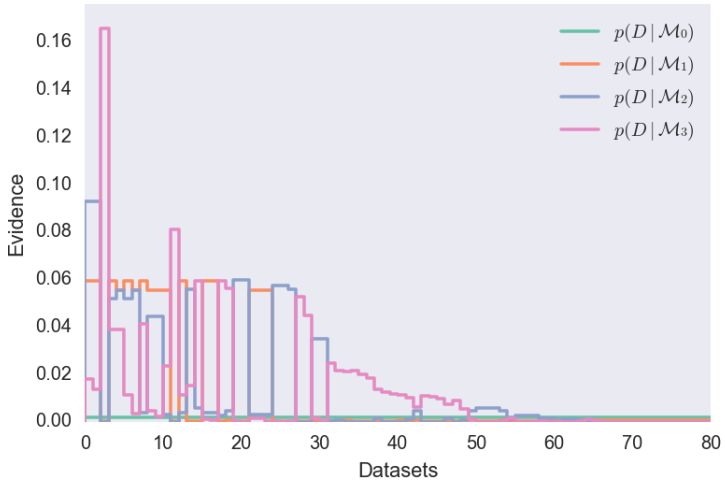


Figure 14: Probability mass by model, non-identity Σ