# Assignment 1 - Report

Pietro Alovisi

11-31-2018

## Part I : The Prior

### Question 1

Choosing the gaussian distribution means that the vaules $t_i$ is distributed simmetrically around the true determinsitic function. because the gaussian distribution is a unimodal distribution, which means that has only one mode and for this particular distribution it coincides with the mean. This can be rephrased as assuming a determinsitic model $f(\mathbf{x})$ that generates realizations with a white noise $\varepsilon$ that distributes as $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, which is a sensible assumption when using real data. This can be written as:

$$\mathbf{t_i} = f(\mathbf{x_i}) + \varepsilon \quad \varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$$

A prior oservation about the covariance matrix is that it is constant, it does not dependent on the input vector $\mathbf{x}$. Then the spherical covariance matrix implies two facts:

- All the scalar random variables $t_{ij}$ of the vector $\mathbf{t_i}$ have the same variance $\sigma^2$ (called homoscedasticity).
- The fact that the covariance matrix is diagonal means that all the output scalar component $t_{ij}$ of the vector $\mathbf{t_i}$ are independent one another.

Moreover the normal distribution has a lot of properties that makes it easy to work with, and also is ubiquitous in practice as an approximation because of the central limit theorem.

### Question 2

If we do not assume independence of the samples, we must turn to the joint probability distribution

$$p(\mathbf{T}|f, \mathbf{X}) = p(\mathbf{t_1}, \mathbf{t_2}, \ldots, \mathbf{t_N}|f, \mathbf{X})$$

**Question 3**

Equation 5 is a linear transformation of a normal distribution which, from its properties, is again a normal distribution equal to:

$$p(\mathbf{t_i}) \sim \mathcal{N}(\mathbf{W}\,\mathbf{x_i}, \sigma^2\mathbf{I})$$

Still assuming conditionally independent samples, from Eq. 3 the likelihood is just:

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

Having defined the two matrix $\mathbf{T}$ and $\mathbf{X}$ as:

$$\mathbf{T} = \begin{bmatrix} \mathbf{t_1}^T \\ \mathbf{t_2}^T \\ \dots \\ \mathbf{t_N}^T \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} \mathbf{x_1}^T \\ \mathbf{x_2}^T \\ \dots \\ \mathbf{x_N}^T \end{bmatrix}$$

Which we can also write by expanding the whole product, by noting that since all the $\mathbf{t_i}$ have the same variance, the exponents in the probability density function sum up.

$$\begin{aligned} p(\mathbf{T}|\mathbf{X}, \mathbf{W}) &= \prod_{i=1}^{N} \frac{1}{\sigma^D (2\pi)^{\frac{D}{2}}} \cdot e^{-\frac{1}{2\sigma^2}(\mathbf{t_i}-\mathbf{W}\mathbf{x_i})^T(\mathbf{t_i}-\mathbf{W}\mathbf{x_i})} = \\ &= \frac{1}{\sigma^{ND} (2\pi)^{\frac{ND}{2}}} \cdot e^{-\frac{1}{2\sigma^2}\sum_i^N (\mathbf{W}\mathbf{x_i}-\mathbf{t_i})^T(\mathbf{W}\mathbf{x_i}-\mathbf{t_i})} = \\ &= \frac{1}{\sigma^{ND} (2\pi)^{\frac{ND}{2}}} \cdot e^{-\frac{1}{2\sigma^2} Tr\left((\mathbf{X}\mathbf{W}^\mathbf{T}-\mathbf{T})(\mathbf{X}\mathbf{W}^\mathbf{T}-\mathbf{T})^T\right)} = \\ &= \mathcal{N}(\mathbf{X}\mathbf{W}^T, \mathbf{I}, \sigma^2\mathbf{I}) \end{aligned}$$

Where we substituted the expression at the exponent $\sum_i^N (\mathbf{W}\mathbf{x_i} - \mathbf{t_i})^T(\mathbf{W}\mathbf{x_i} - \mathbf{t_i})$ with $Tr\big((\mathbf{X}\mathbf{W}^\mathbf{T} - \mathbf{T})(\mathbf{X}\mathbf{W}^\mathbf{T} - \mathbf{T})^T\big)$ by noting that the summation is just the sum of the diagonal of the matrix $(\mathbf{X}\mathbf{W}^\mathbf{T} - \mathbf{T})(\mathbf{X}\mathbf{W}^\mathbf{T} - \mathbf{T})^T$.

**Question 4**

The two penalization terms comes from the prior, and can be obtained through the posterior. First let's do the one for the $L_2$ norm, and then we will generalize to the $L_1$. We can write the prior on $W$ as:

$$p(W) = \frac{1}{\tau^2(2\pi)^{\frac{D}{2}}} \cdot e^{-\frac{tr((W-W_0)(W-W_0)^T)}{2\tau^2}} = \frac{1}{\tau^2(2\pi)^{\frac{D}{2}}} \cdot e^{-\frac{\sum_i^D (w_i-w_i^0)^T \cdot (w_i-w_i^0)}{2\tau^2}}$$

If we multiply with the expression computed above for the likelihood $p(\mathbf{T}|\mathbf{X}, \mathbf{W})$ we get:

$$p(\mathbf{W}|\mathbf{X}, \mathbf{T}) \propto e^{-\frac{1}{2\sigma^2}\sum_i^D (\mathbf{W}\mathbf{x_i}-\mathbf{t_i})^T(\mathbf{W}\mathbf{x_i}-\mathbf{t_i}) - \frac{1}{2\tau^2}\sum_i^D (w_i-w_i^0)^T \cdot (w_i-w_i^0)}$$

Where we have disregarded the multiplicative factor in front of the exponential, because by taking the log will lead to a costant factor. Now we take the negative logarithm:

$$-log(p(\mathbf{W}|\mathbf{X}, \mathbf{T})) \propto \frac{1}{2\sigma^2}\sum_i^D (\mathbf{W}\mathbf{x_i} - \mathbf{t_i})^T(\mathbf{W}\mathbf{x_i} - \mathbf{t_i}) + \frac{1}{2\tau^2}\sum_i^D (w_i - w_i^0)^T \cdot (w_i - w_i^0)$$

If we assume the mean $W_0$ to be 0 we can write the penalizing factor:

$$\frac{1}{2\tau^2}\sum_i^D w_i^T \cdot w_i = \frac{1}{2\tau^2}\sum_i^D \|w_i\|_{L_2}^2$$

By considering a non zero mean the penalizing factor is just the Frobenius norm of the difference of the matrix with its mean:

$$\frac{1}{2\tau^2}\sum_i^D w_i^T \cdot w_i = \frac{1}{2\tau^2}\|\mathbf{W} - \mathbf{W_0}\|_F^2$$

course the proper extension to the $L_1$ norm will lead to the penalizing term:

$$\frac{1}{2\tau^2}\sum_i^D |w_i - w_i^0| = \frac{1}{2\tau^2}\sum_i^D \left\|w_i - w_i^0\right\|_{L_1}$$

Which correspond to a Laplace distribution. From now on I will assume $W_0 = 0$ for simplicity. Using $L_1$ norm will perform some kind of dimensionality reduction

by setting some variables to 0, while the quadratic term in $L_2$ will try to balanced the parameter. We can see this effect by inspecting the derivative of the penalizing term: for the $L_1$ norm it is always constant, while for the $L_2$ it decreases as we get closer to zero:

$$\partial \|w_i\|_{L_1} = \pm 1 \qquad \partial \|w_i\|_{L_2} = 2w_i$$

this means that in $L_2$ optimizing values that are already close to the origin does not get me any relevant decrease in the penalazing term, while if I take a value far away from the origin then this will decrease a lot my penalizing term. For $L_1$ any optimization has the same effect in reducing the penalizing term because it does not depend on the position of the weight vector I'm optimizing.

We can take another perspective by looking at the distribution, in fact the Laplace one has a peak in the origin, so it will spread most of its mass there, making weights zero more likely. The gaussian still has a peak in the origin but it is more smooth. Since the optimization of the log likelihood can be seen as a constrained optimization problem then transformed using Lagrange multipliers, we can also see visually by looking at the iso-contours of the constraining functions in figure 1.

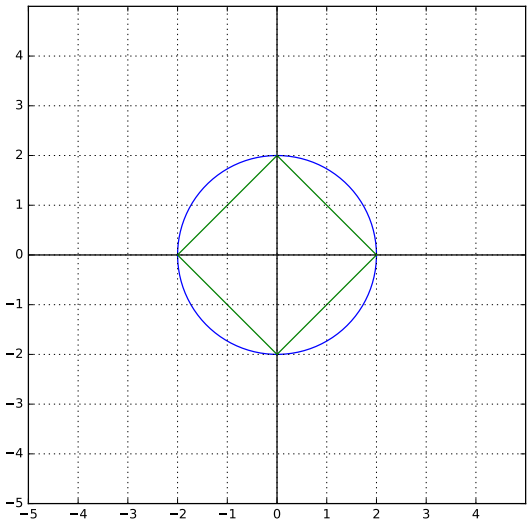


Figure 1: Showing iso-contours of value 1 for the $L_1$ norm (green), and for $L_2$ (blue).

We can see that the corners of the square lie on the axis, so where one of the

two variables is zero.

**Question 5**

We will use the square completion to perform this task. Since the product of gaussian is stilla gaussian we assume the posterior to be normal with the following parametrs:

$$p(\mathbf{W}|\mathbf{T},\mathbf{X}) = \mathcal{N}(\mathbf{M},\mathbf{\Sigma},\mathbf{V}) =$$
$$= \frac{1}{\xi}\cdot e^{-\frac{1}{2}tr(V^{-1}(W-M)\Sigma^{-1}(W-M)^T)} =$$

$$= \frac{1}{\xi}\cdot e^{-\frac{1}{2}tr(V^{-1}W\Sigma^{-1}W^T)}e^{\cdot tr(V^{-1}W\Sigma^{-1}M^T)}e^{-\frac{1}{2}tr(V^{-1}M\Sigma^{-1}M^T)}$$

Where $\xi$ is just the normlaizing factor to make the integral of the function 1.

Now we will take the product of the prior over $W$, and the likelihood $p(\mathbf{t_i}|\mathbf{x_i},\mathbf{W})$.

$$p(\mathbf{W}|\mathbf{t_i},\mathbf{x_i}) = e^{-\frac{1}{2\sigma^2}(\mathbf{t_i}-W\mathbf{x_i})^T(\mathbf{t_i}-W\mathbf{x_i})}\cdot e^{-\frac{1}{2\tau^2}tr((W-W_0)(W-W_0)^T)}$$
$$= e^{-\frac{1}{2\sigma^2}tr((\mathbf{t_i}-W\mathbf{x_i})(\mathbf{t_i}-W\mathbf{x_i})^T)}\cdot e^{-\frac{1}{2\tau^2}tr((W-W_0)(W-W_0)^T)}$$

Since $(\mathbf{t_i}-W\mathbf{x_i})(\mathbf{t_i}-W\mathbf{x_i})^T)$ and $(W-W_0)(W-W_0)^T$ $(D\prod D)$, we can merge the two traces:

$$p(\mathbf{W}|\mathbf{t_i},\mathbf{x_i}) = e^{-\frac{1}{2\sigma^2}(\mathbf{t_i}-W\mathbf{x_i})^T(\mathbf{t_i}-W\mathbf{x_i})}\cdot e^{-\frac{1}{2\tau^2}tr((W-W_0)(W-W_0)^T)} =$$
$$= e^{-\frac{1}{2\sigma^2}tr((\mathbf{t_i}-W\mathbf{x_i})(\mathbf{t_i}-W\mathbf{x_i})^T)}\cdot e^{-\frac{1}{2\tau^2}tr((W-W_0)(W-W_0)^T)} =$$
$$= e^{-\frac{1}{2\sigma^2}tr(\mathbf{t_i}\mathbf{t_i}^T)}e^{\frac{1}{\sigma^2}tr(W\mathbf{x_i}\mathbf{t_i}^T)}e^{-\frac{1}{2\sigma^2}tr(W\mathbf{x_i}\mathbf{x_i}^TW^T)}e^{-\frac{1}{2\tau^2}tr(WW^T)}e^{\frac{1}{\tau^2}tr(WW_0^T)}e^{-\frac{1}{2\tau^2}tr(W_0W_0^T)} =$$
$$= e^{-tr(W(\frac{1}{2\tau^2}\mathbf{I}+\frac{1}{2\sigma^2}\mathbf{x_i}\mathbf{x_i}^T)W^T)}e^{tr(W(\frac{1}{\sigma^2}\mathbf{x_i}\mathbf{t_i}^T+\frac{1}{\tau^2}W_0^T))}e^{-tr(\frac{1}{2\sigma^2}\mathbf{t_i}\mathbf{t_i}^T+\frac{1}{2\tau^2}W_0W_0^T)}$$

Then assuming the independence of the $\mathbf{t_i}$ we can get to the full posterior, noting that the only thing that changes is the likelyhood (the product of each $p(\mathbf{t_i}|\mathbf{x_i},\mathbf{W})$), which turns into summation in the exponent.

$$p(\mathbf{W}|\mathbf{T},\mathbf{X}) = e^{-tr(W(\frac{1}{2\tau^2}\mathbf{I}+\frac{1}{2\sigma^2}\sum_i^N\mathbf{x_i}\mathbf{x_i}^T)W^T)}e^{tr(W(\frac{1}{\sigma^2}\mathbf{x_i}\mathbf{t_i}^T+\frac{1}{\tau^2}W_0^T))}e^{-tr(\frac{1}{2\sigma^2}\mathbf{t_i}\mathbf{t_i}^T+\frac{1}{2\tau^2}W_0W_0^T)}$$

$$p(\mathbf{W}|\mathbf{T},\mathbf{X}) = e^{-tr(W(\frac{1}{2\tau^2}\mathbf{I}+\frac{1}{2\sigma^2}\mathbf{X}^T\mathbf{X})W^T)}e^{tr(W(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{T}+\frac{1}{\tau^2}W_0^T))}e^{-tr(\frac{1}{2\sigma^2}\mathbf{T}^T\mathbf{T}+\frac{1}{2\tau^2}W_0W_0^T)}$$

Where we substituted $\sum_i^N \mathbf{x_i}\mathbf{t_i}^T$ with $\mathbf{X}^T\mathbf{T}$. This can be demostrated to be true, and so I did in appendix A. Now we can retrieve the variance and the mean

of our prior by comparing the first expression with the derived one, and then match the elements. For the second order elements we have:

$$\frac{1}{2}V^{-1}W\Sigma^{-1}W^T = W(\frac{1}{2\tau^2}\mathbf{I} + \frac{1}{2\sigma^2}\mathbf{X}^T\mathbf{X})W^T$$

We can derive that $\mathbf{V} = \mathbf{I}$, and $\mathbf{\Sigma} = (\frac{1}{\tau^2}\mathbf{I} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X})^{-1}$. Now we can compute the mean:

$$V^{-1}W\Sigma^{-1}M^T = W(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{T} + \frac{1}{\tau^2}W_0^T)$$

$$\Sigma^{-1}M^T = (\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{T} + \frac{1}{\tau^2}W_0^T)$$

$$M^T = \Sigma \cdot (\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{T} + \frac{1}{\tau^2}W_0^T) = \frac{1}{\sigma^2}\Sigma \cdot \mathbf{X}^T\mathbf{T} + \frac{1}{\tau^2}\Sigma \cdot W_0^T) =$$

$$M^T = \frac{1}{\sigma^2}(\frac{1}{\tau^2}\mathbf{I} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X})^{-1} \cdot \mathbf{X}^T\mathbf{T} + \frac{1}{\tau^2}(\frac{1}{\tau^2}\mathbf{I} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X})^{-1} \cdot W_0^T)$$

So in the end the posterior is:

$$\mathcal{N}(\mathbf{M}, \mathbf{\Sigma}, \mathbf{I})$$

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

$$M^T = \Sigma \cdot (\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{T} + \frac{1}{\tau^2}W_0^T)$$

Z represents the regularizing term for our posterior distribution and, from Bayes rule, it must be equal to the evidence. But we are not intrested in it for the computation of the posterior, and it does not affect our derivation since it does not depend on $\mathbf{W}$. The maximum likelihood approach gives a result that is the mean of the posterior when the prior is an uninformative one uniform distribution (we can think of this as having the $\tau^2 \to \infty$). So the resulting mean is then:

$$M^T = \frac{1}{\sigma^2}(\frac{1}{\infty}\mathbf{I} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X})^{-1} \cdot \mathbf{X}^T\mathbf{T} + \frac{1}{\infty}(\frac{1}{\infty}\mathbf{I} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X})^{-1} \cdot W_0^T) =$$

$$= \frac{1}{\sigma^2}(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X})^{-1} \cdot \mathbf{X}^T\mathbf{T} = (\mathbf{X}^T\mathbf{X})^{-1} \cdot \mathbf{X}^T\mathbf{T}$$

Which is just the maximum likelihood solution.

**Question 6**

This is a prior on functions, where a function is seen as a collection of infinite random variables, and for any subset of it the joint probability is a multivariate gaussian. To comment the prior we will analize its two components. The least important is the mean, which is set arbitrarely to 0, which means that the functions we'll have zero mean. The most important component is the covariance, that is computed as a kernel function. The kernel function should implement some kind of "closeness measure" between two points $x_i$ and $x_j$, with the kernel having high values id $x_i$ is similar to $x_j$, low otherwise. This function sets the correlation between two points, so if $x_i$ and $x_j$ are close, their values will be high correlated, on the opposite side if $k(x_i, x_j) = 0$, then the two values $y_i$ and $y_j$ are independent (works only assuming the distribution normal). The effect of the covariance is shown in figure 2 where we can see how the mean on the marginal variable depends on the value of the first one. Basically the covariance function defines a transfer of information between one point and the other based on their distance.

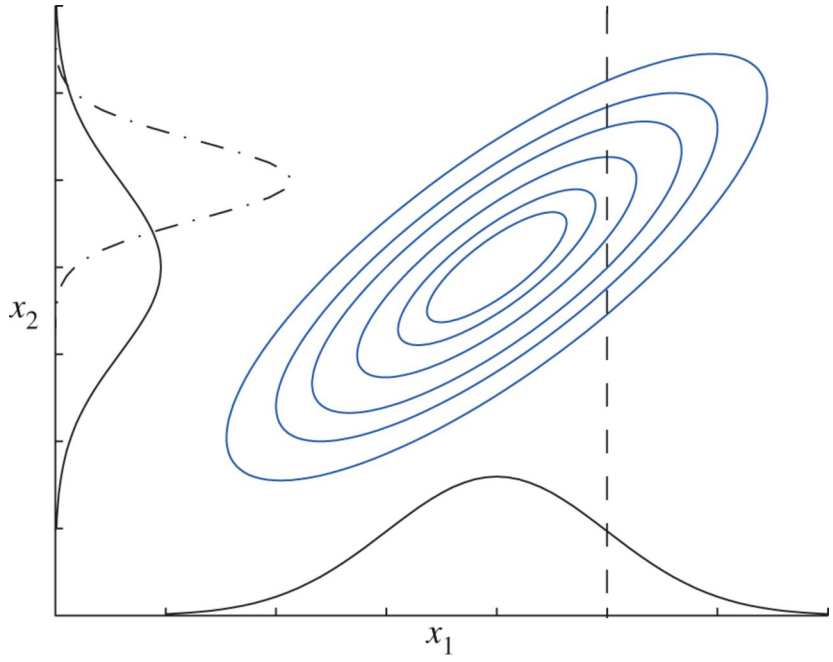


Figure 2: Effect of conditioning one variable $x_2$ by another one $x_1$ when they are correlated.

**Question 7**

If we also assume that $\mathbf{X}$ and $\theta$ are random variables, we can apply the chain rule and easily decompose the formula into:

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

Moreover it's safe to assume that $\mathbf{X}$ and $\theta$ are independent, which lets me factor even more the formula into:

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

I can use the chain rule one again on the first term to get:

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

A graphical model of these assumption can be found in figure 3. From the last formula we know that $p(\mathbf{f}|\mathbf{X}, \theta)$ is a multivariate normal distribution for the definition of the gaussian processes. While we can get some insights in the term $p(\mathbf{T}|\mathbf{f}, \mathbf{X}, \theta)$ by looking at the relation between $t_i$ and $f_i$. Since $t_i$ depends on $f_i$ and the latter, being conditioned, is known.

$$p(\mathbf{t_i} = t^*|\mathbf{f} = f^*, \mathbf{X}, \theta) = p(f^* + \varepsilon = t^*|\mathbf{f} = f^*, \mathbf{X}, \theta) = p(\varepsilon = t^* - f^*|\mathbf{f} = f^*, \mathbf{X}, \theta)$$

So this term is also gaussian.

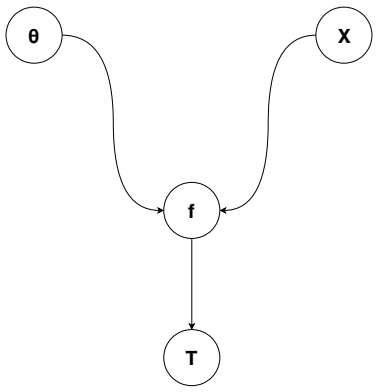


Figure 3: Graphical model of the joint likelihood in Question 7.

**Question 8**

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

The integral has the meaning of a weighted average of the likelyhood of the data over all possible function, where the weight is given by the prior on the functions. The uncertainty is reflected in the covariance matrix of the marginalized distribution, and it has 2 independent components: one comes from the noise $\varepsilon$, and the other comes from the uncertainty we have on the data that can be seen as uncertainty on the shape of the functions resulting from the gaussian process,, this last term is the result of the marginalization of the functions. We still condition in $\theta$ because we assumed it as a constant, it could be marginalized if we have had assumed it was a random variable. In this form the marginal distribution is a function of $\theta$, which is useful for performing hyperparameter optimization.

**Question 9**

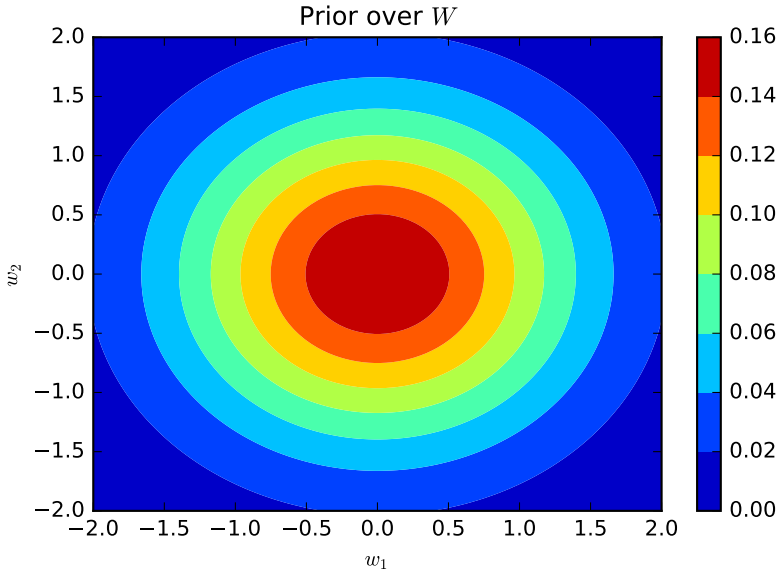


Figure 4: Prior over the parameters, it is a $\mathcal{N}(\mathbf{0},\mathbf{I})$.

The prior is shown in figure 4. Then the evolution of the posterior is shown in figure 5, where le left column is the posterior contour plot, and on the right are the samples taken from it.

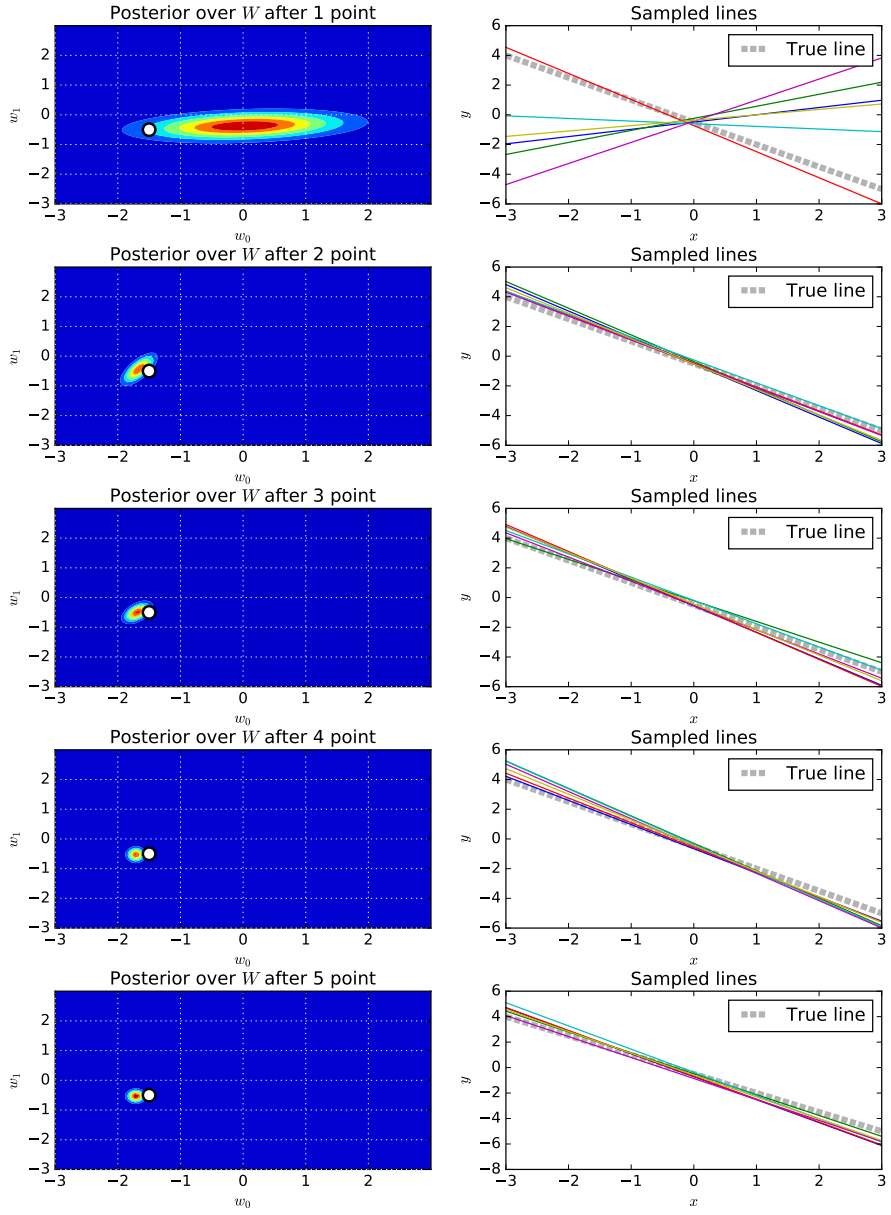


Figure 5: In the left column the posterior over the parameters, the true value of the parameters is marked by a white dot. The left column represent samples from the posterior, and the gray line represents the line with the true parameters.This is the case having the error with $\sigma = 0.3$

There are two main effects in adding data to my model :

- The first one is the fact that the posterior moves its center towards the true value of my weights pair
- The variance of the posterior shrinks as I add more points

These two effects can be explained easily. The latter occurs because as we get more data we are more certain about the model, our belief increases, and so our variance reduces. This rate of change depends on the noise in our data. The first effect is determined by the fact that our belief changes as we see more points. Starting from the prior at the origin, we move towards the pair that better fits our data. The prior encodes some bias that gets less relavant(that we forget) as we get more and more points. We can think of this as a iterative process, we start with our belief that is the prior, then we obser our data point, and now our belief is the posterior. Then the posterior is used as a prior in the next iteration where we observe the new data point. This process goes on and on until we observe all data.

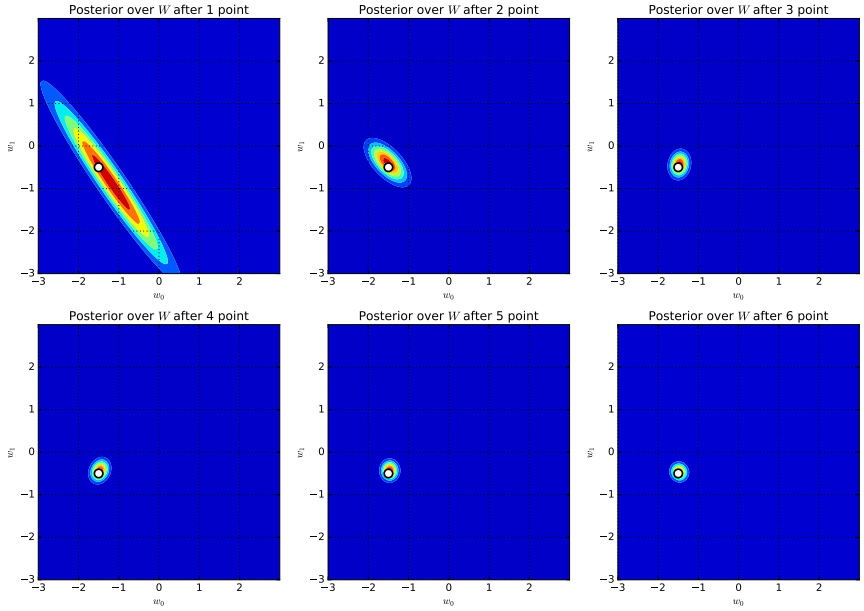


Figure 6: Posterior evolution when observing 6 points with $\sigma = 0.1$

The effect of changing the error standar deviation $\sigma$ is shown in figure 6 and 7. We can see that a big uncertainty in the data reflects to uncertainty in the posterior, and in the end uncertainty in the parameter. We can also see that the convergence of the posterior distribution over the right parameter value is slower for the model with high $\sigma$.

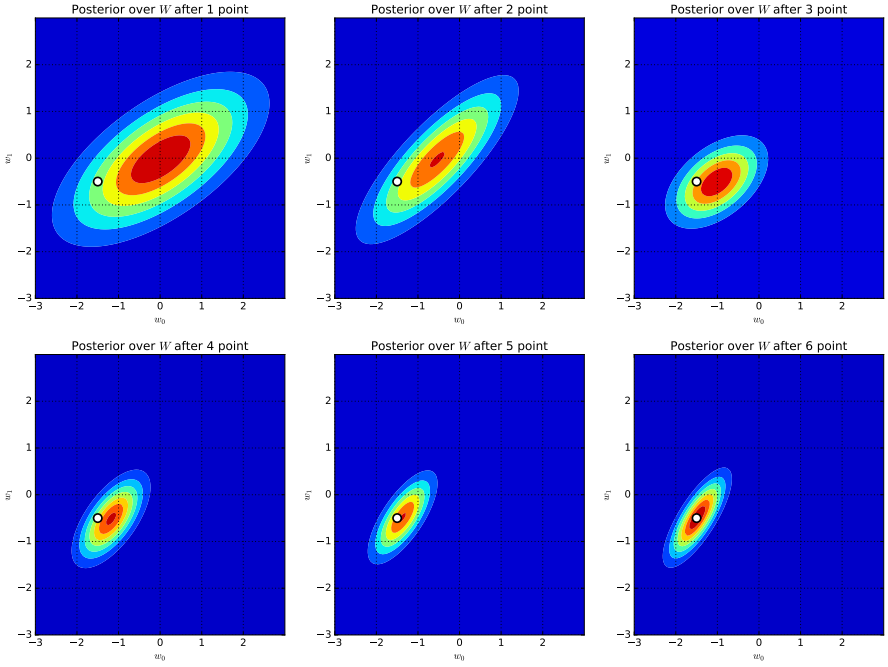


Figure 7: Posterior evolution when observing 6 points with $\sigma = 0.7$

### Question 10

The result are shown in figure 8. The lenghtscale defines a "scale" for measuring the closeness of two points. We can relate the value of the lenghtscale to the numerical value of the squared exponential. Since it is the denominator of a (negative) exponent, if the value of the exponent is high then the squared exponential will be low (close to zero) otherwise if the exponenti is small, then the value of the squared exponential will be high.

Since it divides the difference between two points, if the lenghtscale is low the two points will be less correlated, if the value of the lenghtscale is high, then they will be highly correlated.

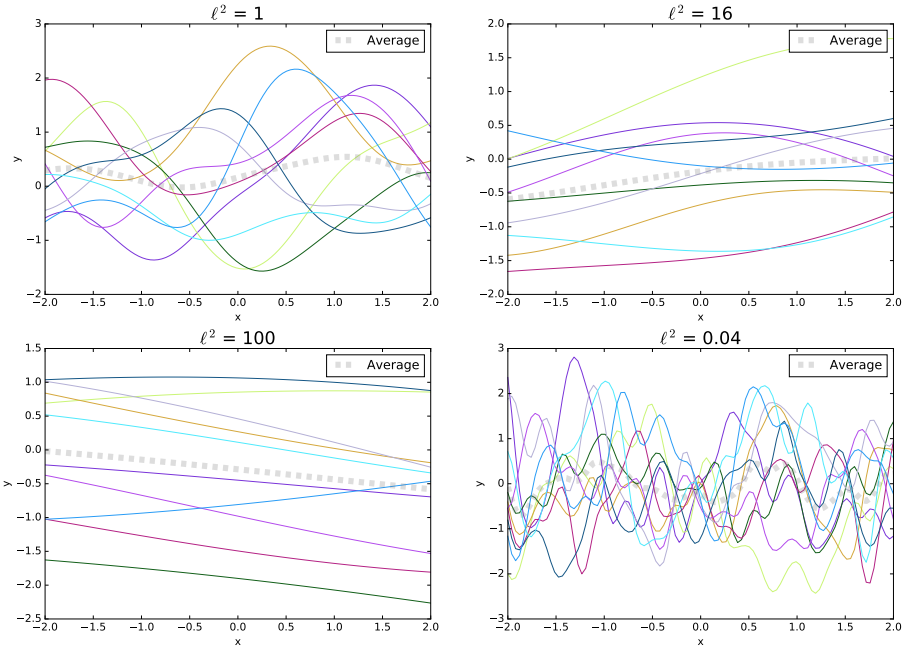


Figure 8: Samples from a gaussian process using different lenghtscales. For all of the picture I have used $\sigma_f^2 = 1$.

### Question 11

If we don't have any data then the posterior is just the prior. In figure 9 we sample some points in some fixed location $x_i$ from the predictive distribution $p(t_i|D)$. Here we see that the further away we are from the data points, the higher the variance in my sampled points. This is because we are less certain

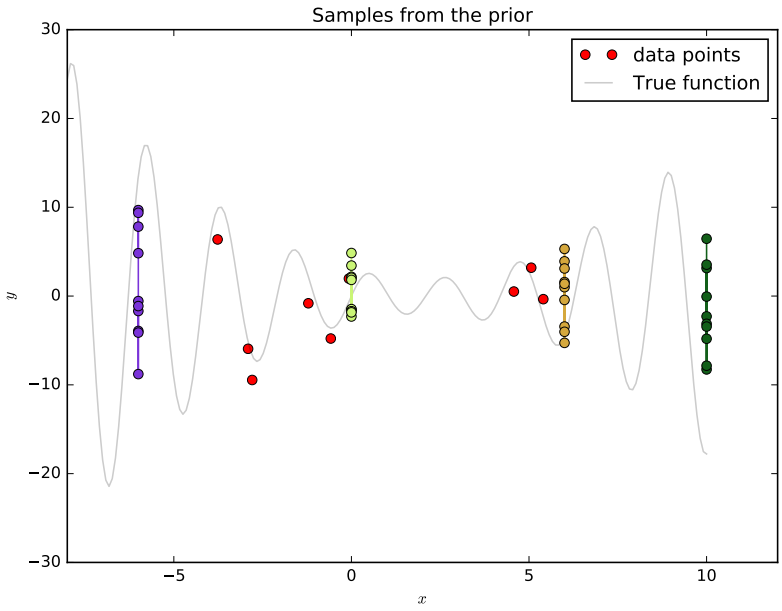


Figure 9: Samples from the predictive posterior for some $x$. In red are the known data points,

about the value of the function there. While near the data points, the variance is quite low.

We then plotted the predictive mean and variance in figure 10 and 11. Again the variance behaves as we described before. These function are better than the one drawn from the prior because they use the information of the data. These functions start behaving like the one drawn from the prior when they are outside the neighborhood of the data points.

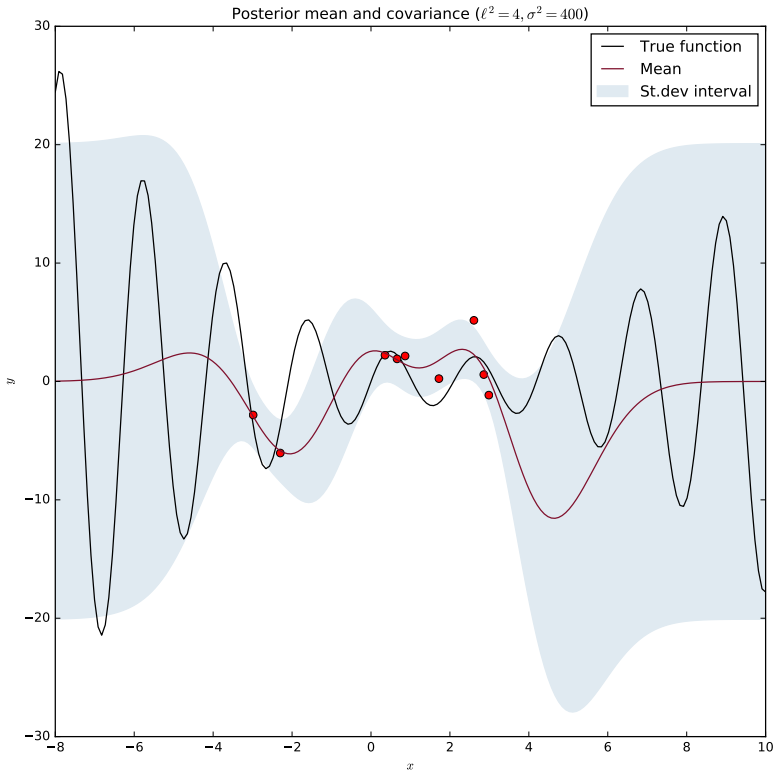


Figure 10: Mean for the predictive posterior, with shaded area of $\pm\sigma$, the standard deviation.

We can also sample some function from the posterior $p(f_i|\mathbf{D})$ which is easily obtained from the predictive posterior by removing the diagonal term due to noise. The results are in image 12. We can see that the function follow a common trend near the data points, while far away they behave randomly.

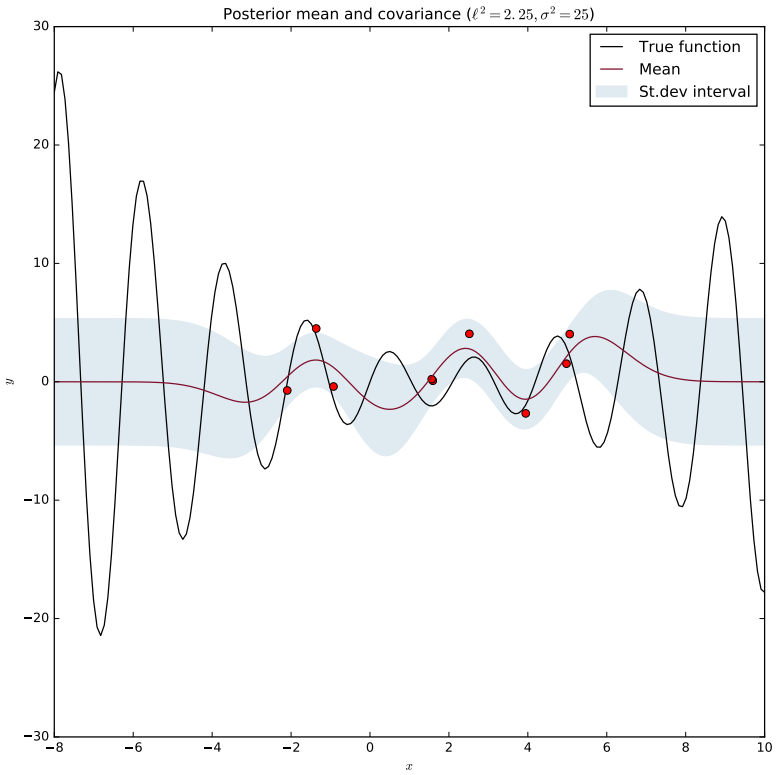


Figure 11: Mean for the predictive posterior, with shaded area of $\pm\sigma$, the standard deviation.

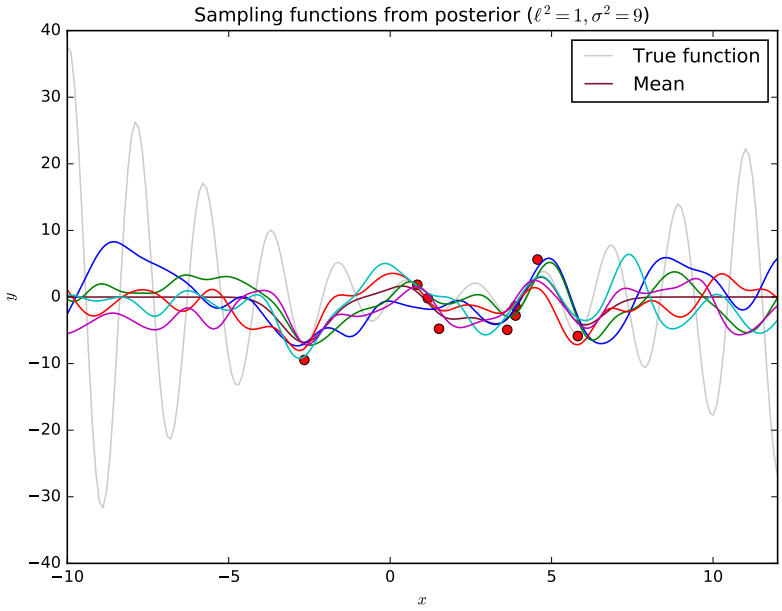


Figure 12: Sampled function from the posterior.

We can investigate the effect of the diagonal term in the kernel function in figure 13. Since it is a diagonal term it only adds uncertainty to each prediction variable $t_i$, increasing its variance, as we can see from the plot.

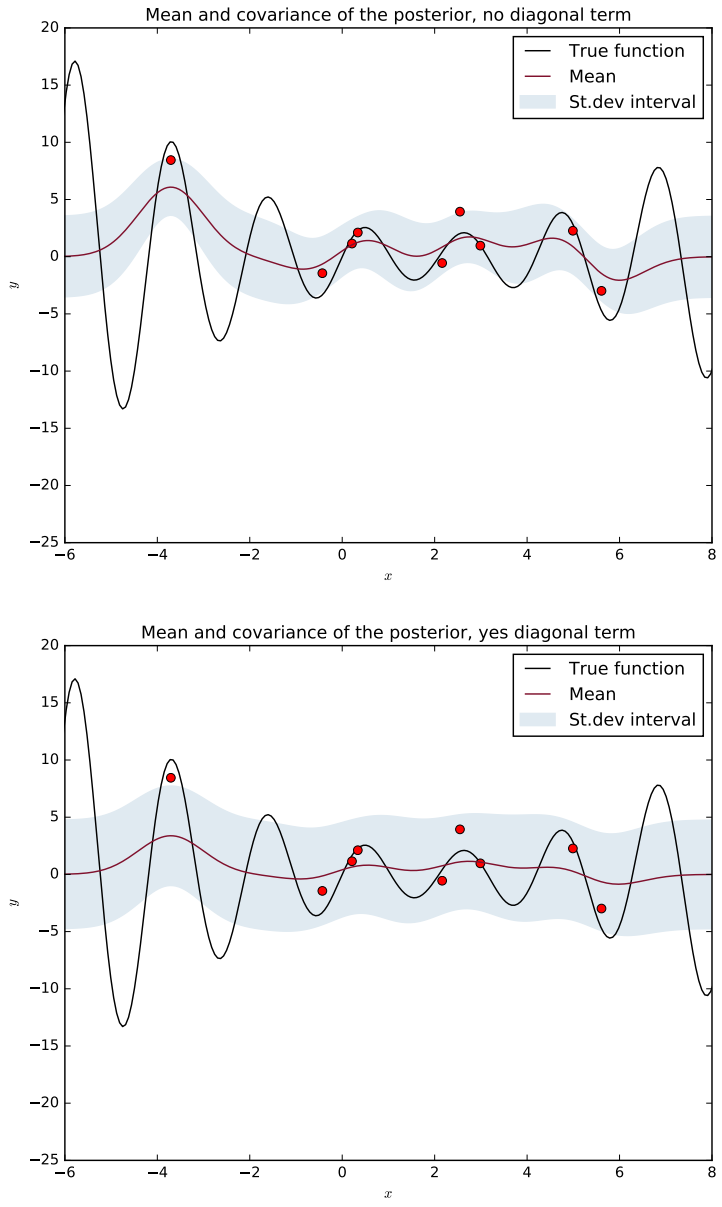


Figure 13: The top plot does not add a diagonal term to the kernel, the bottom one does.The parameter used are $\ell^2 \stackrel{!}{=} 1, \sigma^2 = 9$ and the diagonal term was scaled by a factor of 10.

## Part II : The Posterior

### Question 12

The preference is that our latent variable $X$ is a normal distribution whose elements are independent because of the diagonal covariance matrix. Moreover the values of the latent variable distribute around zero.

### Question 13

Since the marginalization of a gaussian by a gaussian prior is still a gaussan(by the *Gaussian Algebra*), we only need to compute its mean and covariance to describe fully the marginalized distribution. Given the independence of each row of the matrix $\mathbf{Y}$, we can write the following relationship:

$$\mathbf{y_i} = \mathbf{W}\mathbf{x_i} + \varepsilon \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

We can then compute the first and second order statistic, the expected value and the variance, of each $\mathbf{y_i}$ in the random variable $\mathbf{x_i}$. We start from the expected value:

$$\begin{aligned}\mathbb{E}_X(\mathbf{y_i}) &= \mathbb{E}_X(\mathbf{W}\mathbf{x_i} + \varepsilon) \\ \mathbb{E}(\mathbf{y_i}) &= \mathbf{W}\,\mathbb{E}(\mathbf{x_i}) + \mathbb{E}(\varepsilon) \\ \mathbb{E}(\mathbf{y_i}) &= 0 + 0\end{aligned}$$

Here I have dropped the subscript $_X$ for convenience. I have also used the linearity of the expectation operator. Now let's move on to the variance:

$$\mathrm{Var}(\mathbf{y_i}) = \mathrm{Var}(\mathbf{W}\mathbf{x_i} + \varepsilon)$$

Since the white noise $\varepsilon$ is uncorrelated to the variable $\mathbf{W}\mathbf{x_i}$ , we can split the variance in two and get the following expression:

$$\begin{aligned}\mathrm{Var}(\mathbf{y_i}) &= \mathrm{Var}(\mathbf{W}\mathbf{x_i}) + \mathrm{Var}(\varepsilon) \\ \mathrm{Var}(\mathbf{y_i}) &= \mathbf{W}\mathrm{Var}(\mathbf{x_i})\mathbf{W}^T + \sigma^2\mathbf{I} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}\end{aligned}$$

Where I have used the properties of the variance to move $W$ outside the argument of the variance, since it as a constant. Each $y_i$ is independent with each other thus we can combine the results we got into the distribution:

$$p(\mathbf{Y}|\mathbf{W}) = \prod_i^N \mathcal{N}(\mathbf{y_i}|\mathbf{0}, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})$$

$$p(\mathbf{Y}|\mathbf{W}) \sim \mathcal{N}(\mathbf{Y}|\mathbf{0}, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}, \mathbf{I})$$

**Question 14**

**MLE**

From the derived distribution in question 3 we can compute the log likelihood:

$$log(p(\mathbf{Y}|\mathbf{X}, \mathbf{W})) = log\left(\frac{1}{\sigma^2(2\pi)^{\frac{D}{2}}} \cdot e^{-\frac{1}{2\sigma^2}\sum_i^N (\mathbf{W}\mathbf{x_i}-\mathbf{t_i})^T(\mathbf{W}\mathbf{x_i}-\mathbf{t_i})}\right) =$$

$$= -log\left(\sigma^2(2\pi)^{\frac{D}{2}}\right) - \frac{1}{2\sigma^2}\sum_i^N (\mathbf{W}\mathbf{x_i} - \mathbf{t_i})^T(\mathbf{W}\mathbf{x_i} - \mathbf{t_i})$$

In the maximization we disregard the constant factor $-log\left(\sigma^2(2\pi)^{\frac{D}{2}}\right)$ ,and then remove also the multiplicative constant in the second term $\frac{1}{2\sigma^2}$. So we are left with the maximization of:

$$\arg\max_W -\sum_i^N (\mathbf{W}\mathbf{x_i} - \mathbf{t_i})^T(\mathbf{W}\mathbf{x_i} - \mathbf{t_i})$$

Which is clealy the generalization of the sum of residual square for vectorial outputs.

**MAP**

We can derive the expression starting from the previous part of the question.

$$log\left(p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) \cdot p(\mathbf{W})\right) = log(p(\mathbf{Y}|\mathbf{X}, \mathbf{W})) + log(p(\mathbf{W}))$$

The first term of the summation is the MLE term from before, while the second one I have already computed in qustion 4 as:

$$log(p(\mathbf{W})) = -log\left(\tau^2(2\pi)^{\frac{D}{2}}\right) - \frac{1}{2\tau^2}\sum_i^D (\mathbf{w_i})^T(\mathbf{w_i})$$

Again we can disregard the constant term at the begining, but we need to keep the multiplicative terms both for the prior and for the least square. Putting everything together we get:

$$\arg\max_{W} \left\{ -\frac{1}{2\sigma^2} \sum_{i}^{N} (\mathbf{W}\mathbf{x_i} - \mathbf{t_i})^T (\mathbf{W}\mathbf{x_i} - \mathbf{t_i}) - \frac{1}{2\tau^2} \sum_{i}^{D} (\mathbf{w_i})^T (\mathbf{w_i}) \right\}$$

Where the second term acts as a reguralizing term, the $L_2$ norm we already talked about.

**Type II ML**

$$log\left(\int p(\mathbf{Y}|\mathbf{X},\mathbf{W})p(\mathbf{X})d\mathbf{X}\right) = log\left(p(\mathbf{Y}|\mathbf{W})\right) =$$

$$log\left(\prod_{i=0}^{N} p(\mathbf{y_i}|\mathbf{x_i},\mathbf{W})\right) =$$

$$= \sum_{i=0}^{N} log\left(p(\mathbf{y_i}|\mathbf{x_i},\mathbf{W})\right) =$$

If we substitute with the expression for $p(\mathbf{y_i}|\mathbf{x_i},\mathbf{W})$ we still have a log of a normal distribution.

$$= -\sum_{i=0}^{N} log\left((det(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})2\pi^D)^{\frac{1}{2}}\right) - \frac{1}{2}\sum_{i=0}^{N} \mathbf{y_i}^T(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}\mathbf{y_i}$$

This terms seek the $W$ that maximizes the covariance of the observed data points $\mathbf{y_i}$.

---

As we get more data MLE and MAP change their estimate for the optimal $W$, MAP does that slowly because it also has a belief the, expecially when we have not so much data, infulences a lot the estimate for $W$. In type-II Maximum-Likelihood we sum up the contributions of each new data point in the second term of its derived formula.

The two expression in equation 25 are equal because the denominator (the evidence) is constant for any choice of the model parameter $W$, and a multiplicative (positive) term does not chabge the optimization proble. The evidence only changes if we choose another model.

Type-II Maximum-Likelihood is a sensible way of learning the parameters because we first use the bayesian approach to avoid the overfit on data, and then we maximize the hyperparameter, which cannot overfit, because it is not backed by data. Also because here we do not have the data for the $X$, and so the only thing we can do is marginalize them out.

**Question 15**

$$\mathcal{L}(\mathbf{W}) = -log\left(\prod_{i=0}^{N}\mathcal{N}(\mathbf{y_i}|\mathbf{0}, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I} =)\right) =$$

$$= -log\left(\prod_{i=0}^{N}\frac{1}{(2\pi^D \cdot \det(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}))^{1/2}}e^{-\frac{1}{2}\mathbf{y_i}^T(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})^{-1}\mathbf{y_i}}\right) =$$

Then by using the properties of the logarithm we can obtain:

$$= \sum_{i=0}^{N}\frac{1}{2}log\left(2\pi^D \cdot \det(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})\right) + \sum_{i=0}^{N}\frac{1}{2}\mathbf{y_i}^T(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}\mathbf{y_i} =$$

$$= \frac{ND}{2}log\left(2\pi\right) + \frac{N}{2}log\left(\det(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})\right) + \frac{1}{2}\sum_{i=0}^{N}\mathbf{y_i}^T(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}\mathbf{y_i} =$$

$$= \frac{ND}{2}log\left(2\pi\right) + \frac{N}{2}log\left(\det(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})\right) + \frac{1}{2}Tr\left(\mathbf{Y}(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}\mathbf{Y}^T\right)$$

We can remove the first constant term, and remove 1/2 by multiplying by 2 which only scales the function. Removing these parameter will not change the maxima and minima of the log likelyhood. The final expression is:

$$\mathcal{L}(\mathbf{W}) = Nlog\left(\det(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})\right) + Tr\left(\mathbf{Y}(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}\mathbf{Y}^T\right)$$

Moving on to the derivative, since we are deriving a scalar by a matrix it's convinient to derive by an element of the matrix $W_{ij}$:

$$\frac{\partial\mathcal{L}(\mathbf{W})}{\partial W_{ij}} = N\frac{\partial log\left(\det(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})\right)}{\partial W_{ij}} + \frac{\partial Tr\left((\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}\mathbf{Y}^T\mathbf{Y}\right)}{\partial W_{ij}}$$

Here we have two terms to derive. Let's start from the first one:

$$\frac{\partial log\left(\det(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})\right)}{\partial W_{ij}} = Tr\left(\left(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I}\right)^{-1}\cdot\frac{\partial\left(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I}\right)}{\partial W_{ij}}\right) =$$

By using the property[1] $\partial\left(log\left(det\left(\mathbf{X}\right)\right)\right) = Tr\left(\mathbf{X}^{-1}\partial\mathbf{X}\right)$ we get:

$$\frac{\partial log\left(\det(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})\right)}{\partial W_{ij}} = Tr\left(\left(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I}\right)^{-1}\cdot\frac{\partial\left(\mathbf{W}\mathbf{W}^T\right)}{\partial W_{ij}}\right)$$

We only need to develop the right factor of the multiplication, and by applying the simple product rule for matrix derivation we obtain:

$$\frac{\partial\left(\mathbf{W}\mathbf{W}^T\right)}{\partial W_{ij}} = \frac{\partial\left(\mathbf{W}\right)}{\partial W_{ij}}\cdot\mathbf{W}^T+\mathbf{W}\cdot\frac{\partial\left(\mathbf{W}^T\right)}{\partial W_{ij}}$$

Where the derivation $\frac{\partial(\mathbf{W})}{\partial W_{ij}}$ give rise to the single-entry element $J_{ij}$[2].

$$\frac{\partial\left(\mathbf{W}\mathbf{W}^T\right)}{\partial W_{ij}} = \mathbf{J_{ij}}\cdot\mathbf{W}^T+\mathbf{W}\cdot\mathbf{J_{ij}}^T = \mathbf{J_{ij}}\cdot\mathbf{W}^T+(\mathbf{J_{ij}}\cdot\mathbf{W}^T)^T$$

Before putting everythin together let's derive the next term:

$$\frac{\partial Tr\left(\mathbf{Y}(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})^{-1}\mathbf{Y}^T\right)}{\partial W_{ij}} = Tr\left(\frac{\partial\mathbf{Y}(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})^{-1}\mathbf{Y}^T}{\partial W_{ij}}\right)$$

Where I have used the linearity of the the derivation of the trace (afterall the trace is just a sum). Using the identity $\partial\mathbf{X}^{-1} = -\mathbf{X}^{-1}\cdot\partial\mathbf{X}\cdot\mathbf{X}^{-1}$ we obtain:

$$Tr\left(\frac{\partial\left[\mathbf{Y}(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})^{-1}\mathbf{Y}^T\right]}{\partial W_{ij}}\right) =$$

$$= Tr\left(-\mathbf{Y}(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})^{-1}\cdot\frac{\partial(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})}{\partial W_{ij}}\cdot(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})^{-1}\cdot\mathbf{Y}^T\right) =$$

$$Tr\left(-\mathbf{Y}\frac{\partial\left[\mathbf{Y}(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})^{-1}\mathbf{Y}^T\right]}{\partial W_{ij}}\right) =$$

[1] Taken from *The Matrix Cookbook* available here.
[2] This notation is taken from the wikipedia page.

$$= Tr\left(-\mathbf{Y}(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1} \cdot \left[\mathbf{J_{ij}} \cdot \mathbf{W}^T + \mathbf{W} \cdot \mathbf{J_{ij}}^T\right] \cdot (\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1} \cdot \mathbf{Y}^T\right)$$

Now we will put everything together.If we make the following substitution for a cleaner formula:

$$\mathbf{\Sigma} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$$

$$\mathbf{H_{ij}} = \frac{\partial\left(\mathbf{W}\mathbf{W}^T\right)}{\partial W_{ij}} = \mathbf{J_{ij}} \cdot \mathbf{W}^T + \mathbf{W} \cdot \mathbf{J_{ij}}^T$$

The final formula is then:

$$\frac{\partial \mathcal{L}(\mathbf{W})}{\partial W_{ij}} = NTr\left(\mathbf{\Sigma}^{-1} \cdot \mathbf{H_{ij}}\right) + Tr\left(-\mathbf{Y} \cdot \mathbf{\Sigma}^{-1} \cdot \mathbf{H_{ij}} \cdot \mathbf{\Sigma}^{-1} \cdot \mathbf{Y}^T\right)$$

**Question 16**

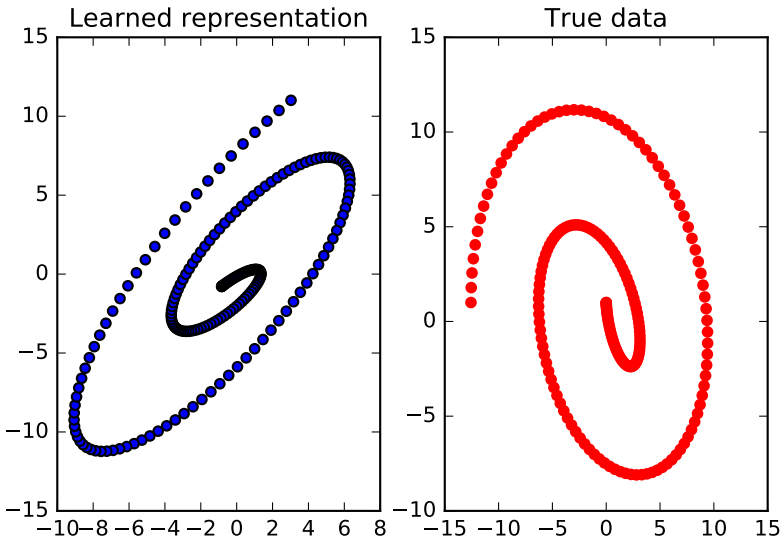


Figure 14: On the left the latent variable representation of the data, while on the right the true representation that generated the data.

The result of the algorithm are presented in figure 14. There is shown the learned latent representation of the data and the starting representation that then we used to generate the data. As we can se the learned latent representation is a rotated version of the true representation. This is because there is an invariance in the parameter matrx $\mathbf{W}$ with respect to the dot product (function composition) with any orthogonal matrx. Since any orthogonal matrix represent a rotation in its appropriate space, it means that the latent representation is invariant to rotation. We can see it matematically, if we assume that there is a matrix $W_{opt}$ which is a minimum of $\partial\mathcal{L}$ and we create a new matrix:

$$\mathbf{W'_{opt}} = \mathbf{W_{opt}}\mathbf{R}$$

For any $\mathbf{R}$ orthogonal. In the likelihood formula the parameter $\mathbf{W}$ is present only in the form $\mathbf{W} \cdot \mathbf{W}^T$. So substituting both $\mathbf{W_{opt}}$ and $\mathbf{W'_{opt}}$:

$$\mathbf{W'}_{opt} \cdot (\mathbf{W'}_{opt})^T = \mathbf{W}\mathbf{R} \cdot \mathbf{R}^T\mathbf{W}^T = \mathbf{W} \cdot \mathbf{W}^T$$

Where we used the fact that $\mathbf{R} \cdot \mathbf{R}^T = \mathbf{I}$ for any orthogonal matrix. We can conlude that $\mathcal{L}(\mathbf{W}_{opt}) = \mathcal{L}(\mathbf{W'_{opt}})$, and so both are valid optimal solutions. If we plug $\mathbf{W'}_{opt}$ into the model formula:

$$\mathbf{Y} = \mathbf{X} \cdot (\mathbf{W'}_{opt})^T = \mathbf{X} \cdot \mathbf{R}^T(\mathbf{W}_{opt})^T = \mathbf{X'} \cdot (\mathbf{W}_{opt})^T$$

Where $\mathbf{X'}$ it's the other latent representation, and we can see that it only differs from the $\mathbf{X}$ by a rotation, therefore proving that we may learn any rotated version of the true representation.

## Part III : The Evidence

### Question 17

This model is the simples because it does not have any parameter and so its probability density function is fixed. In particular this model spreads all its probability equaly over all dataset which means that the model does not have a more likely dataset, one that it can "explain" the most. Basically this is a model that "explains" all dataset, but badly. The lack of parameter means that we cannot also tune the model, we cannot make it learn or adapt to our data. The class of models $M_0$ is composed only by one model.

### Question 18

Each of the next model resembles the logistic regression model, in fact we have in all setting a logistic sigmoid of a linear function in $x$ and $y$.

$$\frac{1}{1 + \exp\left(-y_i \cdot (\theta^T \cdot \mathbf{x} + \theta_0)\right)}$$

So each model gives more probability mass to the dataset for which the quantity $y_i \cdot (\theta^T \cdot \mathbf{x} + \theta_0)$ is positive and (possibly) large. This quantity has a geometric meaning: if the bias is 0 ($\theta_0 = 0$) the quantity is the projection of $\mathbf{x}$ on $\theta$, which is just perpendicular distance with sign between the line given by $\theta^T \cdot \mathbf{x} = 0$ and the point $\mathbf{x}$ scaled by the norm of $\theta$. Then by multipling by $y_i$ we might change the sign. if $\mathbf{x}$ lies on the side of the line "pointed" by $\theta$ then the sign of $y_i$ is preserved. If we include the bias, we can think of it as a threshold on this "distance", that graphically changes the intercept of the boundary $\theta^T \cdot \mathbf{x} + \theta_0 = 0$ This boundary defines an are where $y_i = 1$ are more probable (the one pointed by $\theta$) and another one where $y_i = -1$ are. This difference is more enhanced if we have a high value of $\theta$ that "sharpens" the boundary.

#### Model $M1$

Here we only focus the value of $x_1$ and we don't consider $x_2$, the boundary induced by $\theta$ is orthogonal to the $x_1$ axis through the origin. Then $\theta$, as we said, controls the "strictness" on which the model tollerate point on the wrong side. This puts it's probability mass over datasets that can be split by such vertical boundary.

#### Model $M2$

Here we also care about the value of $x_2$, and the resulting line can have any orientation, but it still passes throught the origin. Same line of reasoning as

before for the $\theta$ parameter.Again this puts it's probability mass over datasets that can be split by such boundary.

**Model** $M3$

Here we also have a bias term $\theta_3$, and the resulting line can have any orientation and itercept. This is the most general linear model, we cannot have more degrees of freedom. This model spread its probability mass over all linearly separable datasets.

Each of these datasets cannot "explain" any non-linear dataset, while $M_0$ can. On the other hand these models are flexible because they have parameter that can be tuned to a particular subset of datasets.

**Question 19**

$M_3$ is the most flexible one because it can move its decision boundary however it wants, it can express any linear boundary. In particular it can explain all the datasets of $M_1$ and $M_2$ (because both linear) plus some more, which means that it must allocate less probability mass over those datasets spanned also by $M_1$ and $M_2$ which is a somehow constraing. So the model pays this flexibility in less probability per each single dataset it "explains". The same consideration works for $M_2$ which is has more flexibility than $M_1$. Then $M_0$ spread its probability over all dataset equally, which is the best a model can do to explain all dataset, but by doing so it does not leave any degree of freedom to move its probability on other datasets.

**Question 20**

Marginalization is the process through which we can obtain a probability distribution of a subset of random variable, from a joint probability distribution by marginalizing the other random variables away. This process effectively "removes" the dependency of the other variables, the marginalzied ones, by a process that looks like a weighted average. This integration conveys the effect of the dependencies of the marginalized variable into the output distribution. In our case this is done by:

$$\int_\theta p(D|M_i,\theta)p(\theta)d\theta$$

Which expresses an average of models $p(D|M_i,\theta)$, by the probability of the model parameters $p(\theta)$. So we are averaging models taking into account the probability density function of $\theta$. We are mixing the possible models, but giving more weight to the ones that are more probable, given the prior distribution.

**Question 21**

By choosing that distribution for the prior we again assume that all the parameter are independent, because of the diagonal covariance matrix. And then we can relate the choosen parameters $\mu$ and $\sigma^2$ to decision boundary defined by the expression $\theta^T \cdot \mathbf{x}$ that we talked about in Question 18. The mean is zero, which means that the value of the parameters $\theta_i$ will distribute around zero, but their variance is very high, which means that is not unlikely the event of having a parameter far from zero. This also relates to what we said about the value of the parameters, if they are high I get a more "strict" boundary. Since parameter are independent and with mean 0, we don't restrict the orientation of the lines in $M_2$ and $M_3$. If we had choosen a non diagonal covariance matrix we would have gotten a bias in the orientation of the lines, because the ratio of the parametrs of the line would be biased. In the extreme case where there is a relation between two parameters of $\theta$, the resulting boundaries form a pencil of lines (either parallel or incident). For the models, the fact the the mean is 0 means that the normal vector $\theta$ is not biased. If the mean had not been non zero, that would have ment a biased in the direction of $\theta$. We can see this effect in $M_1$, where if the mean is positive, then the model assigns more porobability to the datasets having positive $y_i$ on the right than the other one, because the parameter theta is more likely positive. For $M_2$ the effect is that, if we sample a lot of paramerts, we have an average decision boundary with parameters given by the mean. See figure 21 to see the effect of changing mean, and figure 22 for the effect of the covariance.

**Question 22**

If we sum up the evidence over all dataset, for each model we get 1 (obviously in practice there is a small error, in my case on the order of $10^5$). That is beacause this is a probabilistic distribution, and so must sum up to 1. We can interpret this distribution from a generative prospective: if we would sample the model parameters at random, the probability of generating dataset $D_i$ is $p(D_i|M_i)$.

There there are few comment about figure 15, where we can see that the left plot is symmetric, and given how the datasets are indexed in that figure, that means that the probability of a dataset doesn't change if we filp the sign to all the $y_i$ in a dataset. This of course makes sense because the only thing that changes is the orientation of the normal to the decision boundary $\theta^T\mathbf{x}$, and since the distribution of the parameter is symmetric the probability of obtainig $\theta$ or $-\theta$ is the same.

Another comment is the shape of the plot on the right side that I have also plotted in figure 16 using a log scale to better highlight its feature.

We can see how there is a range over which almost all models from $M_1$ to $M_3$ have a high probability. These are the linearly separable datasets (or almost-linearly).

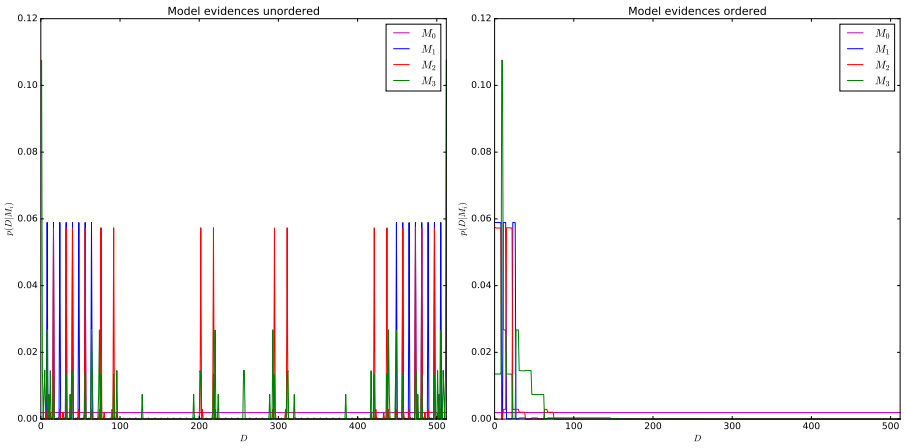


Figure 15: Shown here are the plots of the evidence of each dataset per each model. On the left using the order of the datasets as they were generated, on the right using the same ordering procedure proposed in the paper.

While in all the other dataset $M_0$ puts more probability than the other models.

By looking at the range of linear datasets we see that often the probability of $M_1$ is greater than the one of $M_2$ which is greater than the one of $M_3$. These are the common datasets explained by all the models, that is all the dataset explained by $M_1$. And since $M_1$ is the simplest it will have more probability mass to spread in this range. Then there are the datasets explained only by $M_2$, in this range $p(M_1) < p(M_3) < p(M_2)$, and then the datasets explained only by $M_3$ where it's true $p(M_1) < p(M_2) < p(M_3)$.

Here lies the automatic Occam's razor, we choose the model that puts the most probability on the datasets that we are intrested in, which as we have seen, is usually the simplest among the one that can explain it.

### Question 23

We can comment the most likely and the least likely for each models, represented in figure 17. Of course we do not have a maximum and minimum in the $M_0$ case, since all the models are equally probable, so they are not shown in the figure. For $M_1$ the most likely is one that can be separated by a vertical line, that is the model that it encodes. For $M_2$ the most likely is one that can be separated by a line that passes through the origin, again because the model it represents divide the space with a line crossing the origin. For $M_3$ the most likely is the one with all $+1$ (or $-1$). This makes because we have also the bias term, and having a big variance this can move the line very far away from the square $[-1, 1] \times [-1, 1]$ where the data lies. If the boundary don't cut in half

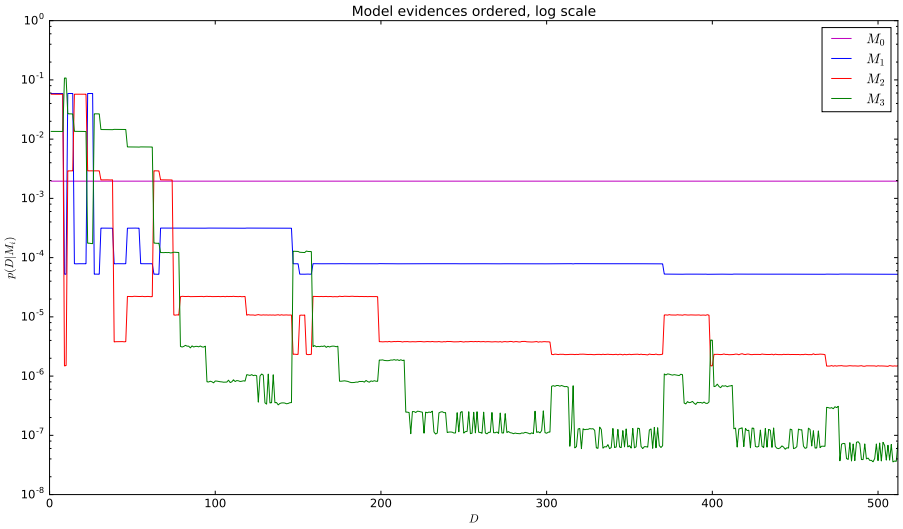


Figure 16: Same as the right plot in the previous figure but using log scale on the y axis.

the data it "classifies" all points as $-1$ or $+1$ intuitively the average decision boundary should be $y = \pm x \pm 1$[3], which does not cut the square $[-1, 1] \times [-1, 1]$.

While for the least probable dataset we see that it represent non-linerly separable configuration, which makes sense, because none of the model is likely to generate it, since they represent linear boundaries.

### Question 24

The prior encodes the preferences about the parameter of the linear boundaries that each model represents. So changing the mean, will make the "average" boundary the one having as parameter, the one specified by the mean. If we change the covariance, we introduce some dependance in the parameters of the decision boundary, which can cause the line parameters to have some ratios between them, which can reflects into a preferred direction or a preferred intercept. These effects are explained by figure 21 and 22.

### Changing covariance

I experimented by changing the covariance matrix. Obviously it makes no sense

[3]This can be shown by taking the expected value of the line $\mathbb{E}(\theta_1 x + \theta_2 y + \theta_3) = \mu x + \mu y + \mu = 0$, then we should take the limit $\mu \to 0$, but we can divide by $\mu$ prior and obtain $\pm x + \pm y + \pm = 0$.

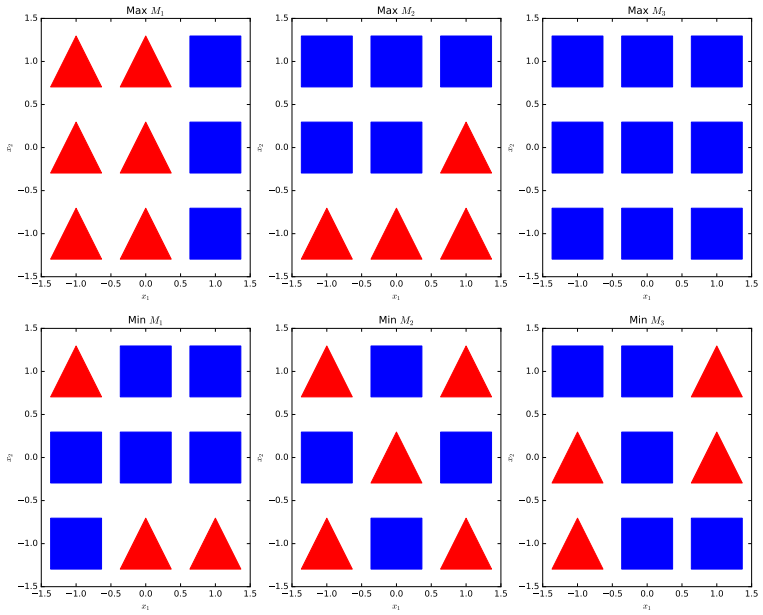


Figure 17: Representation of the most and least likely datasets, as said before, because of simmetricity we don't care if triangles represent $+1$ or $-1$. in the first row are the most probable datasets, and in the second one the least.

for $M_0$ and neither for $M_1$ because it has only one parameter. The covariance I used was:

$$\mathbf{C} = \sigma^2 \begin{pmatrix} 1 & -0.9 & -0.5 \\ -0.9 & 1 & 0.5 \\ -0.5 & 0.5 & 1 \end{pmatrix}$$

For $M_2$ I restricted $\mathbf{C}$ to the top left $2 \prod 2$ square submatrix. We can see a comparison between the model having independent parameters, and one having as covariance $\mathbf{C}$ in figure 18.

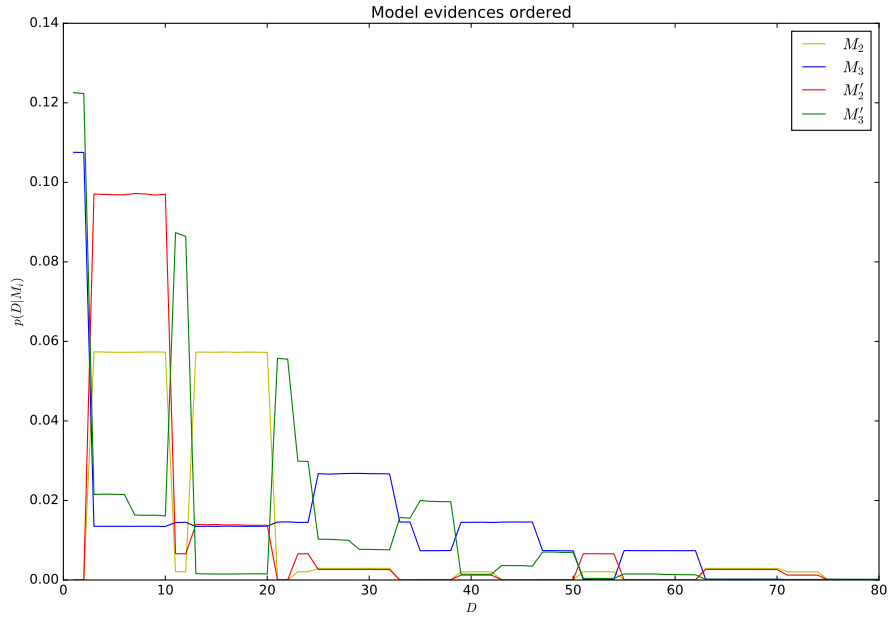


Figure 18: Comparison, in a restricted portion of the domain, for model 2 and 3 between when parameters are drawn independently ($M_2$ and $M_3$), and the one with a non diagonal covariance matrix ($M_2'$ and $M_3'$). Mean is zero in both cases.

If we compare $M_2$ with $M_2'$ we can see that they do not differ a lot in terms of dataset spanned, but puts a different probability mass on each of these. This means that the covarianc we choosed encodes some preferences for those models we have increased the probability. In comparing $M_3$ with $M_3'$ we can make the same considerations as before, but here we also have some sharp differences in the shape of the distribution. It might be that the bias enhance th probability of some datasets that were not considered as likely before.

**Changing mean**

The result of changing the mean to 5, are depicted in figure 19. First of all we can see that we have lost the symmetricity in the left plot, because having changed the mean, we encoded preference of the direction of the normal of the decision boundary towards the first quadrant, so we will prefer datasets for which negative $y_i$ lies "below" the line. One difference is therefore the distribution of probability over the linerly separable models. Moreover the right plot is a little bit squahsed toward the y axis and more irregular compared to the one of figure 15. We can see a scaled version in figure 20. This effect is due to the fact the the magnitude of $\theta$ has increased because of the mean. As already discussed the magnitude of $\theta$ determines how strict the bounday is, in particular it will assign even less probability to non-linearly separable datasets compared to the original case, because the magnitude of $\theta$ acts as a multiplicative factor in the exponent of the sigmoid.

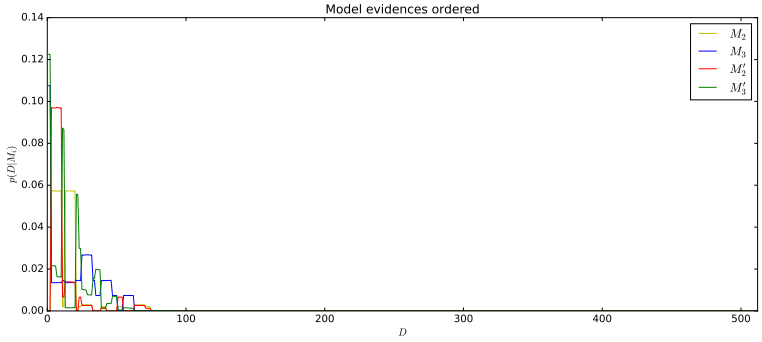


Figure 19: Evidence for each model in the unordered and ordered datasets

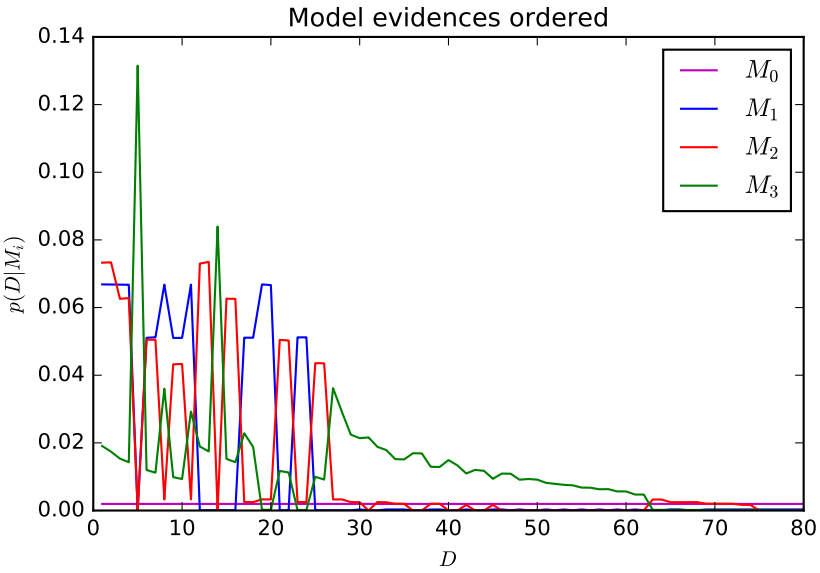


Figure 20: Close up of the model evidence until the first 80 most probable datasets

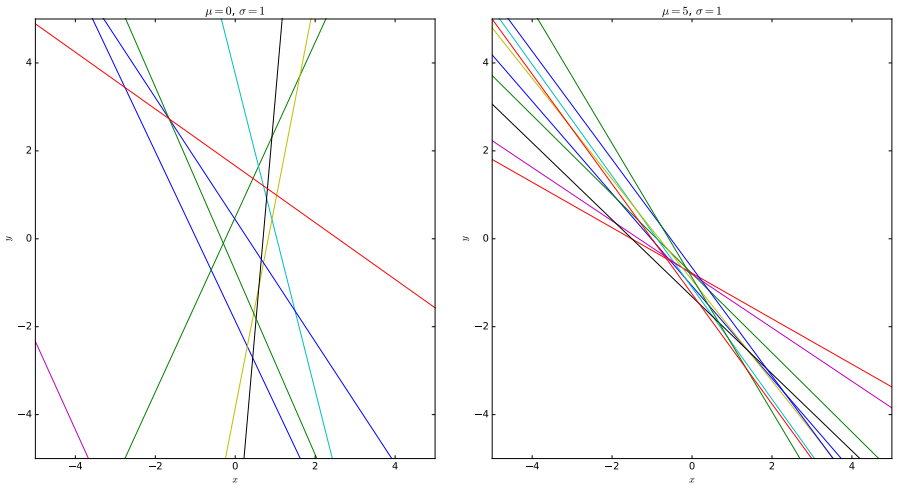


Figure 21: Samples of lines where the parameters are drawn from a normal distribution with the parameters specified in the title. Here all the parameters are assumed independent. We can see how changing the mean affects the resulting lines.

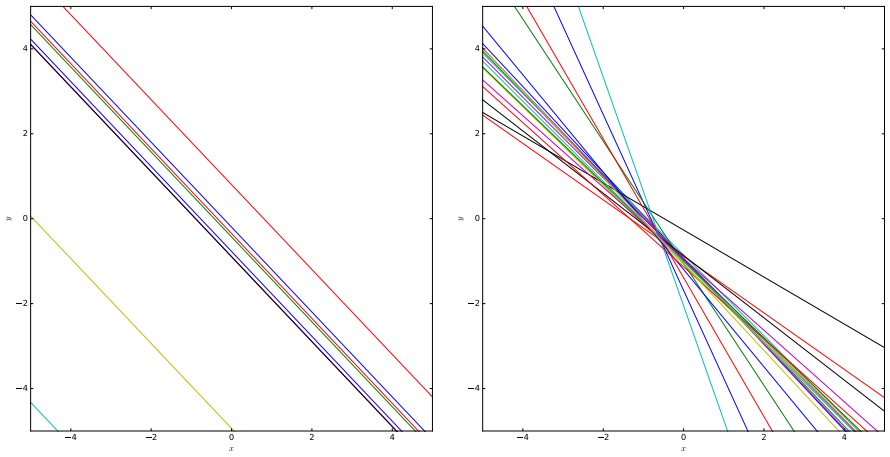

Figure 22: Samples of lines where the parameters are drawn from a normal distribution with mean zero. In the left plot $\theta_1$ and $\theta_2$ are equal while $\theta_3$ is independent.On the right the covariance matrix has 0.9 in all the entries, except on the diagonal where it is 1.

# Appenddix A

## Proof of $\sum_i \mathbf{x_i}\mathbf{x_i}^T = \mathbf{X}^T\mathbf{X}$

Suppose we have a matrix $X$:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x_1}^T \\ \mathbf{x_2}^T \\ \dots \\ \mathbf{x_N}^T \end{bmatrix}$$

We can decompose this matrix as:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x_1}^T \\ \mathbf{0}^T \\ \dots \\ \mathbf{0}^T \end{bmatrix} + \begin{bmatrix} \mathbf{0}^T \\ \mathbf{x_2}^T \\ \dots \\ \mathbf{0}^T \end{bmatrix} + \cdots + \begin{bmatrix} \mathbf{0}^T \\ \mathbf{0}^T \\ \dots \\ \mathbf{x_n}^T \end{bmatrix}$$

From the previous it immediatly follows that its transpose can be expressed as:

$$\mathbf{X}^T = \begin{bmatrix} \mathbf{x_1} & \mathbf{0} & \dots & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{x_2} & \dots & \mathbf{0} \end{bmatrix} + \cdots + \begin{bmatrix} \mathbf{0} & \mathbf{0} & \dots & \mathbf{x_n} \end{bmatrix}$$

Now if we multiply the decomposed version of the matrix $\mathbf{X}^T$ together with $\mathbf{X}$ and apply the distributive property we get:

$$\mathbf{X}^T\mathbf{X} = \left(\begin{bmatrix}\mathbf{x_1} & \mathbf{0} & \dots & \mathbf{0}\end{bmatrix} + \dots + \begin{bmatrix}\mathbf{0} & \mathbf{0} & \dots & \mathbf{x_n}\end{bmatrix}\right) \cdot \left(\begin{bmatrix}\mathbf{x_1}^T \\ \mathbf{0}^T \\ \dots \\ \mathbf{0}^T\end{bmatrix} + \begin{bmatrix}\mathbf{0}^T \\ \mathbf{x_2}^T \\ \dots \\ \mathbf{0}^T\end{bmatrix} + \dots + \begin{bmatrix}\mathbf{0}^T \\ \mathbf{0}^T \\ \dots \\ \mathbf{x_n}^T\end{bmatrix}\right)$$

From which we can see that the multiplication of any corresponding matrices containing the same vector $\mathbf{x_i}$ we get:

$$\begin{bmatrix}\mathbf{0} & \dots & \mathbf{x_i} & \dots & \mathbf{0}\end{bmatrix} \cdot \begin{bmatrix}\mathbf{0}^T \\ \dots \\ \mathbf{x_i}^T \\ \dots \\ \mathbf{0}^T\end{bmatrix} = \mathbf{x_i} \cdot \mathbf{x_i}^T$$

While by multipling two matrices containing different vectors we get:

$$\begin{bmatrix}\mathbf{0} & \dots & \mathbf{x_j} & \dots & \mathbf{0}\end{bmatrix} \cdot \begin{bmatrix}\mathbf{0}^T \\ \dots \\ \mathbf{x_i}^T \\ \dots \\ \mathbf{0}^T\end{bmatrix} = \mathbf{0}$$

We can therfore conclude that:

$$\mathbf{X}^T\mathbf{X} = \sum_i \mathbf{x_i}\mathbf{x_i}^T$$

If we have 2 different matrices $\mathbf{X}$ and $\mathbf{Y}$ we can repeat the procedure and conclude that: