Assignment 1 - Report

Pietro Alovisi

11-17-2018

Todo

- finire risposta 4
- check e finire risposta 5
- question 6 mettere una immagine della marginale
- domanda 9, che varianza scegliere nel prior?
- domanda 9 rifare i calcoli per fgare unn sample delle rette

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 random error ε that distributes as $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. This can be written as:

$$\mathbf{t} = f(\mathbf{x}) + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$$

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 σ^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.

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:

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

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.

$$p(\mathbf{T}|\mathbf{X},\mathbf{W}) = \prod_{i=1}^{N} \frac{1}{\sigma^D(2\pi)^{\frac{D}{2}}} \cdot e^{-\frac{1}{2\sigma^2}(\mathbf{t_i}-\mathbf{W}\mathbf{x_i})^T(\mathbf{t_i}-\mathbf{W}\mathbf{x_i})} =$$

$$= \frac{1}{\sigma^{ND}(2\pi)^{\frac{ND}{2}}} \cdot e^{-\frac{1}{2\sigma^2}\sum_i^N(\mathbf{W}\mathbf{x_i}-\mathbf{t_i})^T(\mathbf{W}\mathbf{x_i}-\mathbf{t_i})} =$$

$$= \frac{1}{\sigma^{ND}(2\pi)^{\frac{ND}{2}}} \cdot e^{-\frac{1}{2\sigma^2}Tr\left((\mathbf{X}\mathbf{W^T}-\mathbf{T})(\mathbf{X}\mathbf{W^T}-\mathbf{T})^T\right)} =$$

$$= \mathcal{N}(\mathbf{X}\mathbf{W}^T,\mathbf{I},\sigma^2\mathbf{I})$$

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 $Tr\left((\mathbf{X}\mathbf{W^T}-\mathbf{T})(\mathbf{X}\mathbf{W^T}-\mathbf{T})^T\right)$ by noting that the summation is just the sum of the diagonal of the matrix $(\mathbf{X}\mathbf{W^T}-\mathbf{T})(\mathbf{X}\mathbf{W^T}-\mathbf{T})^T$.

Question 4

The two penalization terms can be obtained from the prior. 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:

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

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

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

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

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

Where we can easily see the penalizing factor:

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

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

$$\frac{1}{2\tau^2}\sum_i^N |w_i| = \frac{1}{2\tau^2}\sum_i^N \|w_i\|_{L_1}$$

Using L_1 norm will perform some kind of dimensionality reduction by setting some variables to 0, while the quadratic term will try to balanced the parameter. We can see this effect by inspecting the derivative of the penalizing term: for the $L1$ norm it is always constant, while for the $L2$ it decreases as we get closer to zero, this means that in $L2$ optimizing values that are close to the origin does not get me any 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.

We can also see visually by looking at the iso-contours of these functions in figure 1.

We can see that the corners of the square lie on the axis, so where one of the two variables is zero.

$w^T w$: for the L_2 norm which is just the Froebenius norm $|w|_F$

These two priors will introduce an additive term depending on the model parameter $|w|$, which would be of second order for the L_2 metric, and a first order L_1 for the other one.

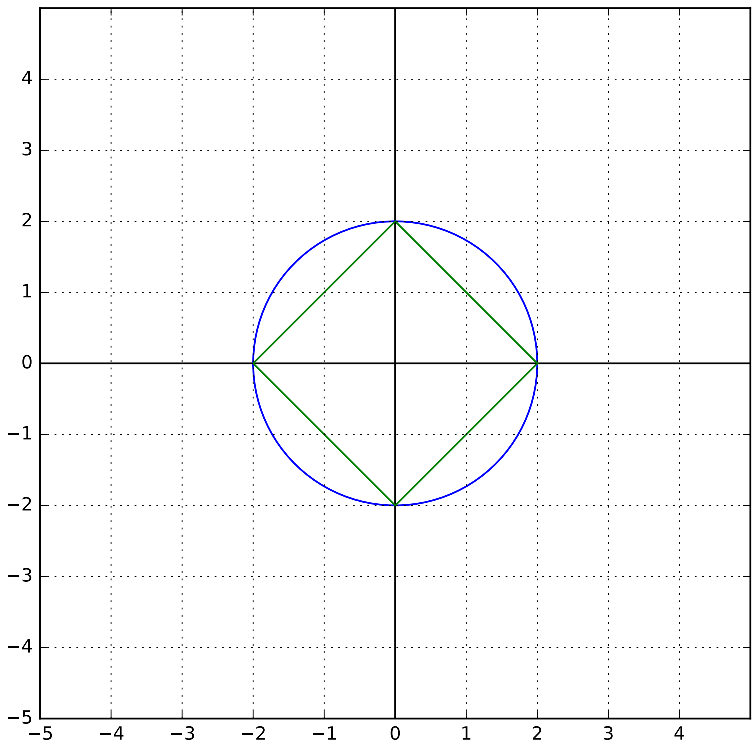

Figure 1: Showing L1 and L2 differences.

Question 5

We will use the square completion to perform this task. So we assume that the output is normal with the following parametrs:

$$p(W) = \frac{1}{\xi} \cdot e^{-\frac{1}{2}tr(V^{-1}(W-W_0)\Sigma^{-1}(W-W_0)^T)} =$$

$$= \frac{1}{\xi} \cdot e^{-\frac{1}{2}tr(V^{-1}W\Sigma^{-1}W^T)} e^{\cdot tr(V^{-1}W\Sigma^{-1}W_0^T)} e^{-\frac{1}{2}tr(V^{-1}W_0\Sigma^{-1}W_0^T)}$$

Where ξ 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 $p(\mathbf{t_i}|\mathbf{x_i}, \mathbf{W})$.

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

Since they have the same dimensions, we can do 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 t_i we can get to the full posterior.

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

Where we substituted $\sum_i \mathbf{x_i}\mathbf{t_i}^T$ with $\mathbf{X}^T\mathbf{T}$. This can be demostrated to be true, and so I do in appendix A.

Now we can retrieve the variance and the mean of our prior.

$$\mathbb{E}(\mathbf{W}|\mathbf{T}, \mathbf{X}) =$$
$$\text{Var}(\mathbf{W}|\mathbf{T}, \mathbf{X}) =$$

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.

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). Basically the covariance function defines a transfer of information between one point and the other based on their distance.

Put figure(s) here

Question 7

If we also assume that $\mathbf{X}$ and θ 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 θ 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:

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

Where 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.

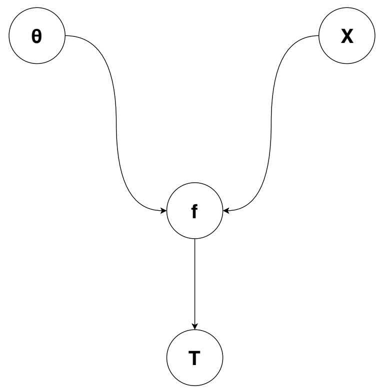

Figure 2: Graphical model of the assumption I made about the variables

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 ε, and the other comes from the uncertainty we have on the data that is can be seen as uncertainty on the shape of the functions of the gaussian process that we marginalize out. We still condition in θ 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 θ, which is useful for performing hyperparameter optimization.

Question 9

There are 2 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 fades away as we get more and more points.

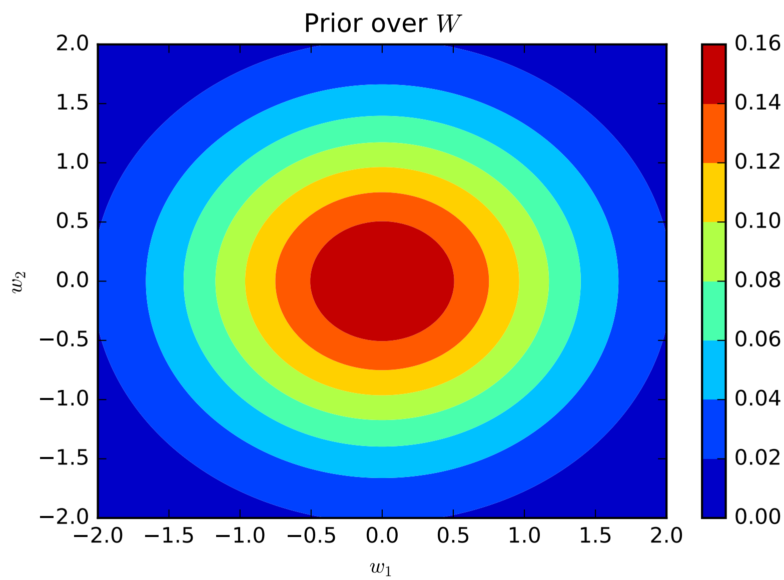

Figure 3: Prior over the parameters

Question 10

The lenghtscale defines a "unit of measure" between the two points. 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.

Question 11

Question 12

The preference is that our variable X is a normal distribution whose elemetns are independent, and distribute around zero.

Question 13

Since the model is:

$$\mathbf{y_i} = \mathbf{W}\mathbf{x_i} + \varepsilon$$

The mean of each $\mathbf{y_i}$ is :

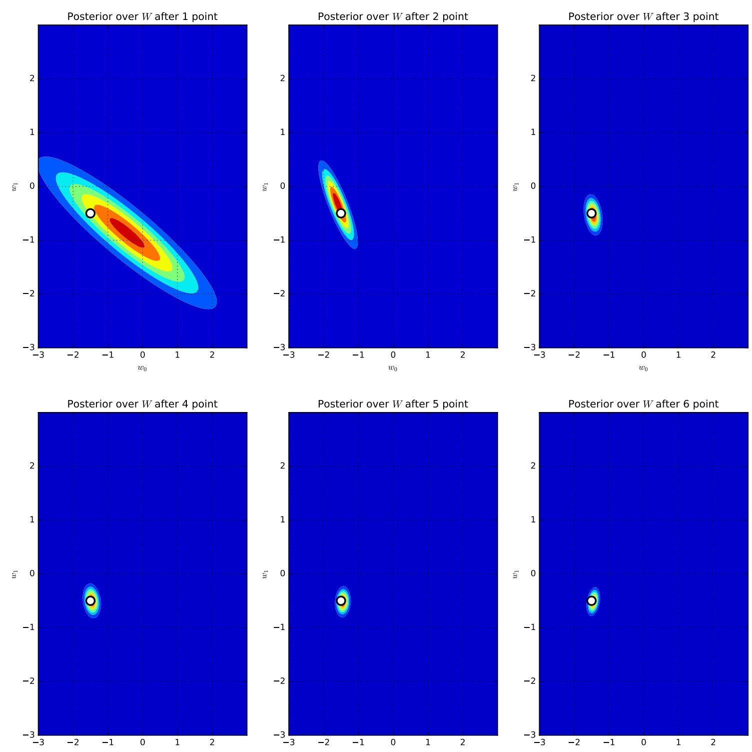

Figure 4: Posterior over the parameters after observing one point with $\sigma = 0.1$

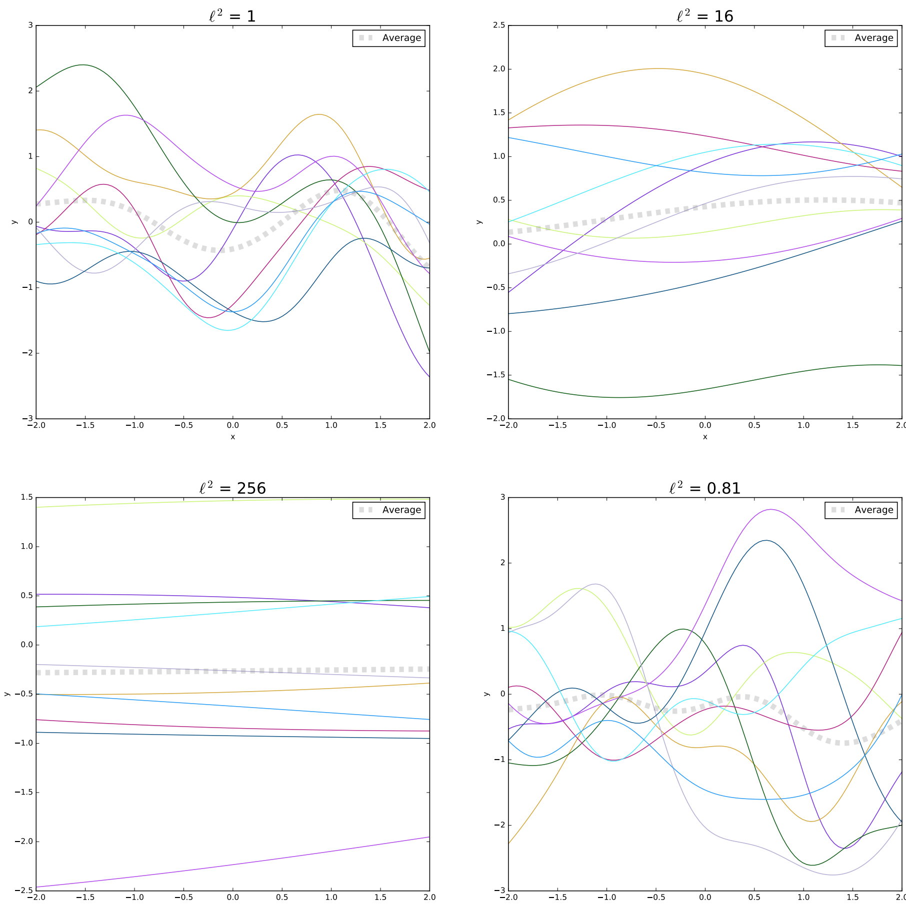

Figure 5: Samples from a gaussian process

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

We can describe the probability by only the first 2 moments of the random variable:

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

Since they are uncorrelated, we can write:

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

Since each y_i is independent with each other, we can combine the results we got into the distribution:

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

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

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:

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

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:

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

Where the second term acts as a reguralizing term.

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.

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

The two expression in equation 25 are equal because the denominator (the evidence) is constant for any choice of the model parameter W. 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.

Question 15

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

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

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

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

We can remove the constant terms, and remove 1/2 by multiplying by 2.

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

Moving on to the derivative:

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

Let's expand one term at a time:

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

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

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

Where the derivation $\frac{\partial(\mathbf{W})}{\partial W_{ij}}$ give rise to the single-entry element J_{ij}[1].

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

The next term to derive is:

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

Using the identity:

$$\partial\mathbf{X}^{-1} = \mathbf{X}^{-1} \cdot \partial\mathbf{X} \cdot \mathbf{X}^{-1}$$

Mi sono dimneticato un meno

We get to:

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

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

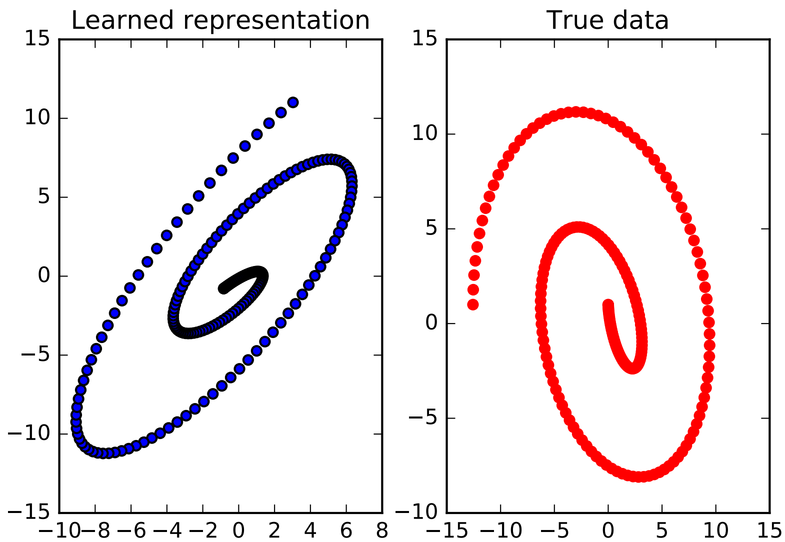

Figure 6: Representation learned

Question 16

Here is the representation I have learned, and beside the true data.

There is an invariance in the parameter matrx $\mathbf{W}$ with respect to the dot procut of any Orthogonal matrx, that is all possible rotation. We can see matematically, if

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

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 $\mathbf{W'_{opt}}$:

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

We can conlude that $\mathcal{L}(\mathbf{W}_{opt}) = \mathcal{L}(\mathbf{W'_{opt}})$, and so both are valid optimal solutions.

[1]This notation is taken from the wikipedia page.

Question 17

This is the simples model because it is uninformative, each data set is equally likely. This means that

Question 18

Question 19

Question 20

Question 21

Question 22

It is also possible to understand how the evidence discourages overcomplex models and therefore embodies Occam's Razor by using the following interpretation. The evidence is the probability that if you randomly selected parameter values from your model class, you would generate data set Y

Question 23

Question 24

Appenddix A

Demonstration 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:

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

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

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

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: