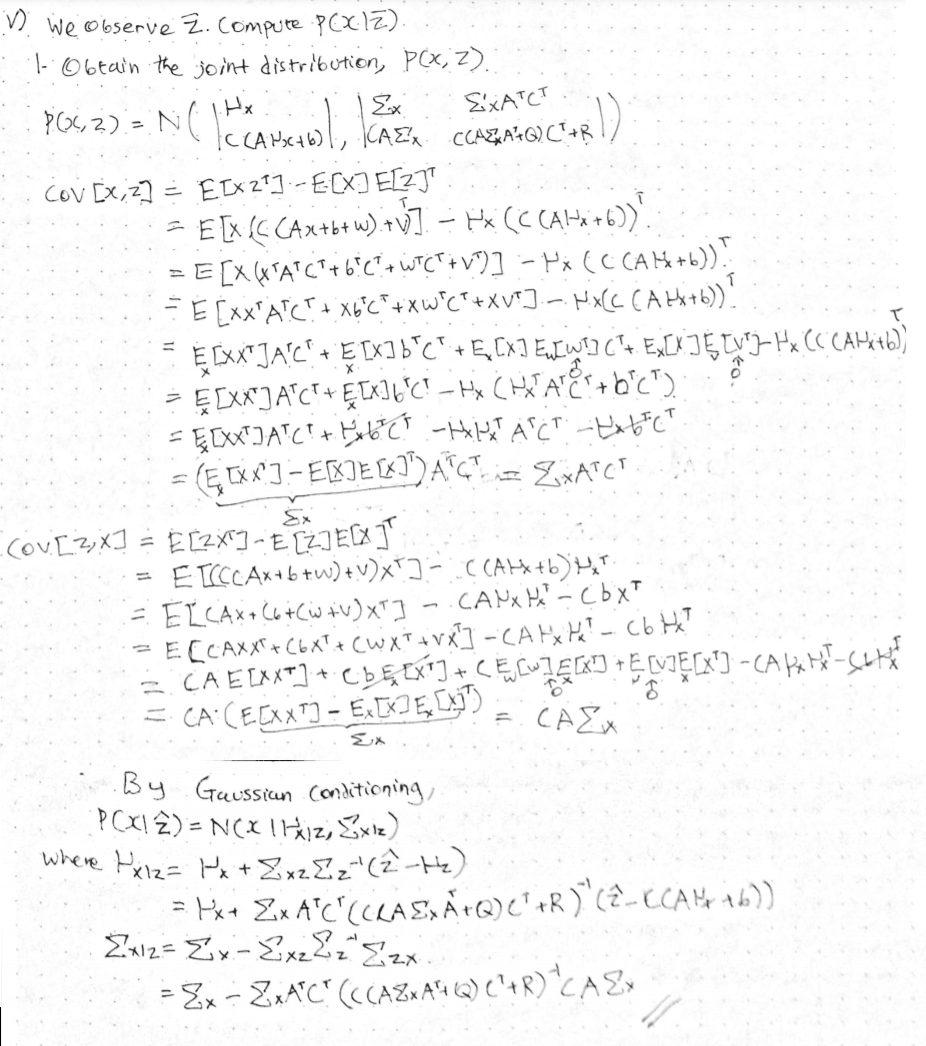
1 a)

1. Not assessed anymore
2. Ax is gaussian (multiplier), b is constant (so still gaussian) and w is gaussian (sum of gaussians is gaussians), therefore p(y) is Gaussian Distribution
3. ,
5. We have to find join distribution of x, z then uses that result from lectures to give us p(x|z). See [https://piazza.com/class/ kf7uh9z2iryy6?cid=252](https://piazza.com/class/kf7uh9z2iryy6?cid=252).

where is as in (iv)

and

Detailed working:



Alternatively, you can obtain the covariance values by writing [x, z] as a linear transformation of [x, w, z].

b) i) Very overfit model because the degree is large (9) and the data points are very spread out

ii) Purpose of the penalty is to prevent overfitting by penalizing large parameters, forcing the solution to balance fitting the points as well as possible with the smallest parameters possible

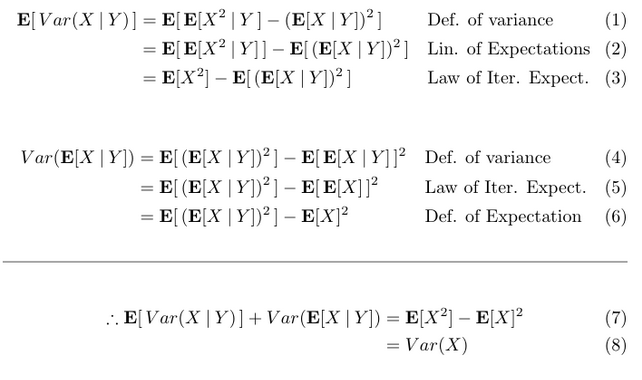
iii) Could pick different lambda and then use cross validation to find the average error

iv) Ridge regression is equivalent to introducing some prior knowledge about what we want the distribution for the parameters to be, this makes it very unlikely to have parameter values that stray too far from the prior distribution given.

c) i) <https://en.wikipedia.org/wiki/Law_of_total_probability>

(okay this is jumping a lot of steps, buts its justified by tonnelis theorem, and we don’t even need to know what that is, so I think it's fine to just assume this is all ok)

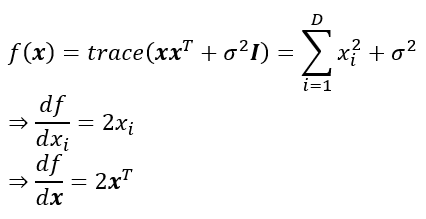
ii) <https://en.wikipedia.org/wiki/Law_of_total_variance>



2)

a)

b)



c)

b)

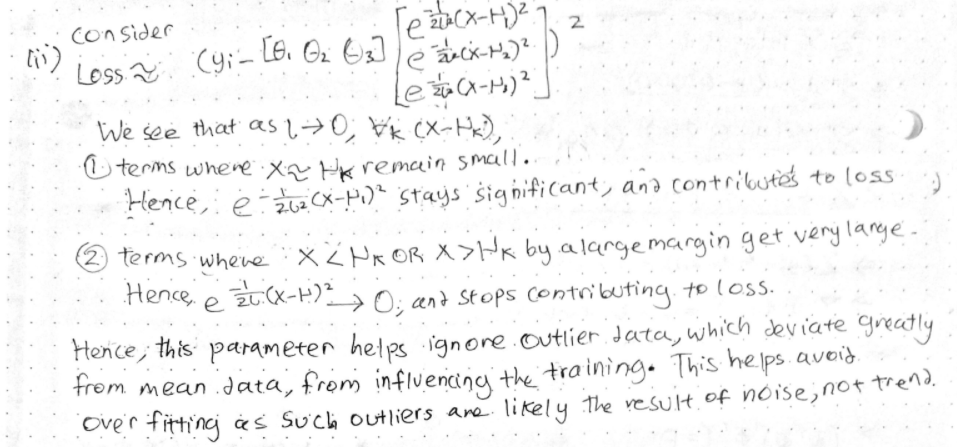
and expand this.... gets messy (do we need to? now that's a fantastic question)

1. Differentiate above wrt alpha and = 0. Probably should check second derivative also  
   is it:  
    alpha = (y – theta^T \* phi(x))^T \* (y – theta^T \* phi(x))

1. Each basis function will be a (unnormalised) gaussian, so either we have exactly as many points as basis functions (then we get overfitting, but doesn’t really matter). Or we have less points, in which case the length scale needs to related to (some of) the distances between points. Some, as they can be very spread out altogether, but as long as there are at least clusters of points which are together, these can be modelled by one gaussian. If this is not the case: overfitting occurs.

Limiting case is spikes of points (v. thin gaussians), which get very large if they want to match up to multiple points.

Alternative (not sure if correct!):



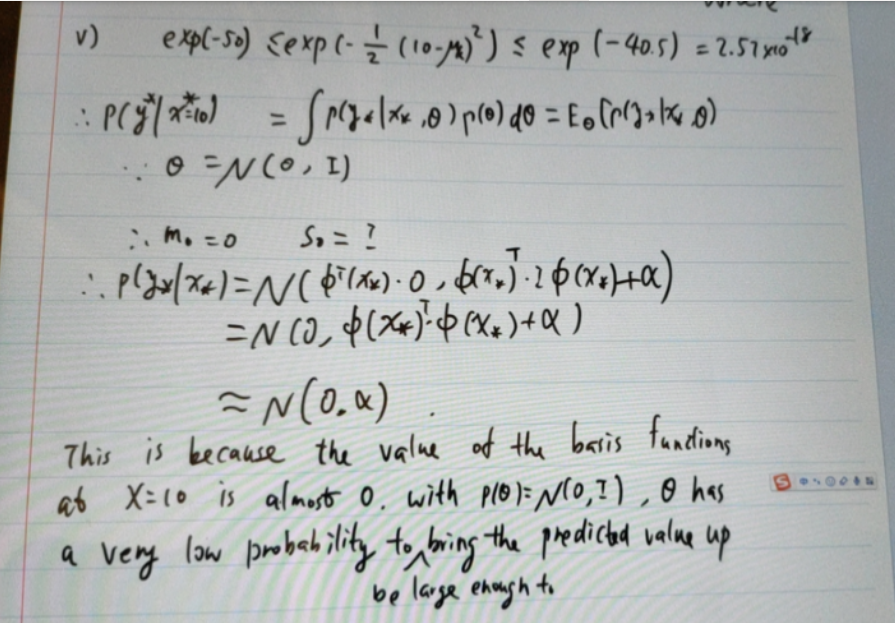
1. First eigen decomposition of .

Then generate vector sample from MV standard normal, put this into .

The output vector will have the expectation and s.d. as required (Can be checked using standard results from lectures).

1. (We only touched on predictive distribution very briefly once, so this question might not apply anymore????)

]



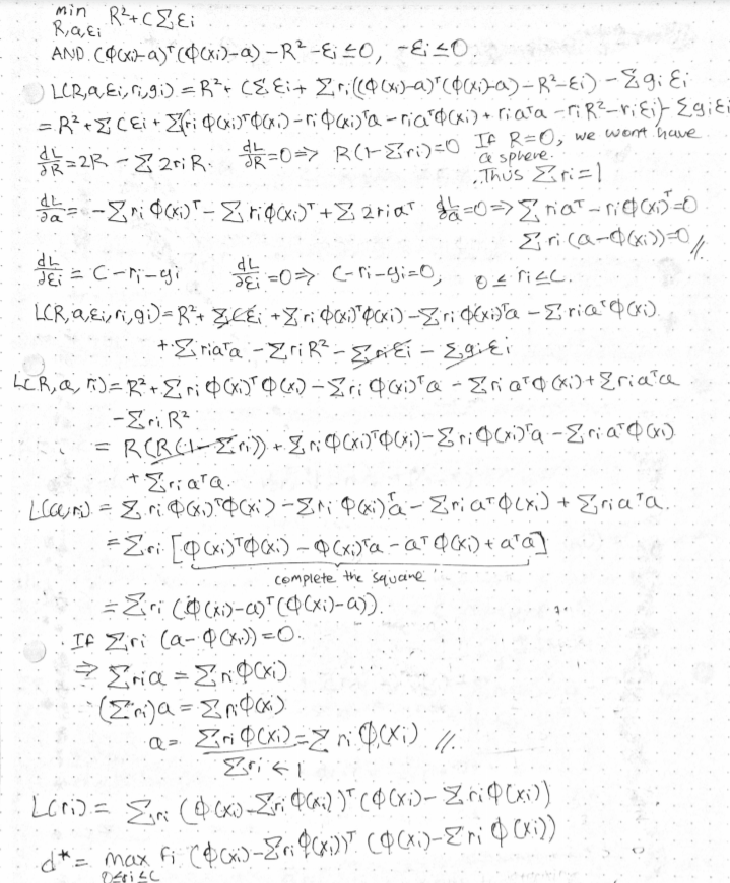
3) a) Exact repeat of 2018-2019 Q4

1. See 2018-2019 Q4i)
2. See 2018-2019 Q4 ii)

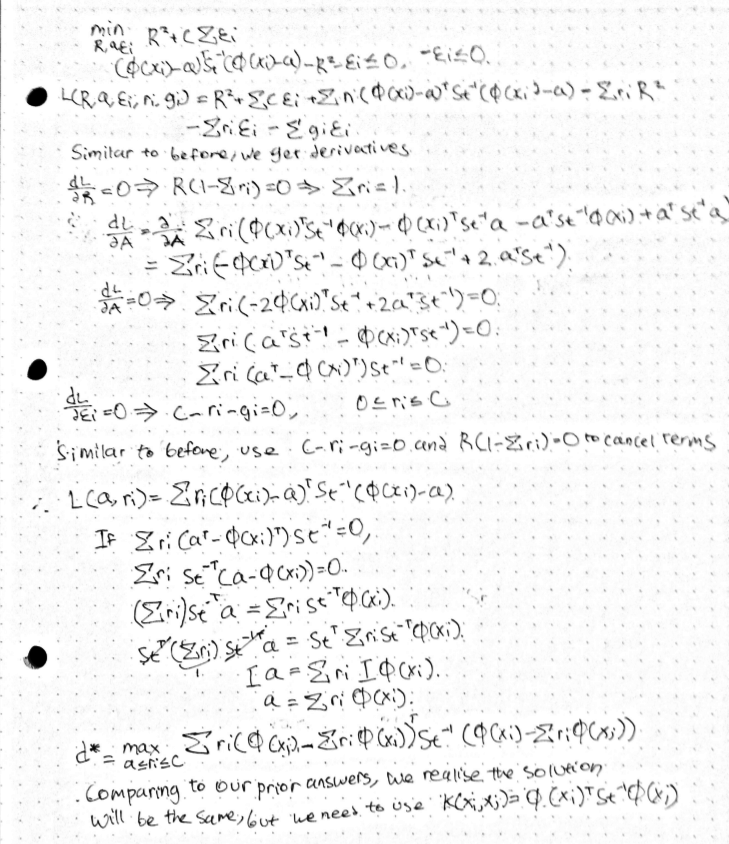
b) tutorial q

4) a)

this is from 2018-19, but its using a kernel, so just ignore that here (TODO: write up solutions without kernel here)



b)



Idk what to do when is singular though??????????? (This gets asked every year so may be important)

Choose kernel to be x\_i^TSt^(-1)x\_j and let a = UQ s.t. UStU^T = I. When you expand the quadratic term, St^-1 will be cancelled. See my answer for 2018-2019.