Andrew Li-Yang Liu (ALL2209)
Professor Verma
COMS 4771 Machine Learning
Columbia University
November 21, 2021

Machine learning homework 3

Problem 1:

1i.

We are given a finite class F. Given $\epsilon > 0$.

$$\begin{aligned}
&P_{(x_i,y_i)_{i=1}^m}[f(x_i) = y_i | f \text{ is } \epsilon - \text{bad}] \\
&= P[f(x_1) = y_1, f(x_2) = y_2 \dots, f(x_m) = y_m | f \text{ is } \epsilon - \text{bad}] \\
&= P[f(x_1) = y_1 | f \ is \ \epsilon - bad] \dots P[f(x_m) = y_m | f \ is \ \epsilon - bad] \\
&< (1-\epsilon)^m
\end{aligned}$$

The last step is due to independence of draws.

1ii.

For notation, let $|F|$ denote the class size (number of elements) of hypothesis class F.

Let's first rephrase the expression inside the probability to give ourselves clarity:

$$\begin{aligned}
&P_{(x_i,y_i)_{i=1}^m}[\exists \epsilon \text{ bad } f \in F \text{ such that } f(x_i) = y_i] \\
&= P_{(x_i,y_i)_{i=1}^m}[\exists f \in F \text{ such that } f \text{ is bad } AND\ f(x_i) = y_i]
\end{aligned}$$

Let's call the condition of simultaneously being epsilon-bad and getting all training data correctly labelled as condition C. i.e.:

$$= 1 - P_{(x_i,y_i)_{i=1}^m}[\text{No } f \text{ satisfies } C]$$
$$= 1 - P_{(x_i,y_i)_{i=1}^m}[f_1 \text{ doesn't satisfy } C] P_{(x_i,y_i)_{i=1}^m}[f_2 \text{ doesn't satisfy } C] \dots P_{(x_i,y_i)_{i=1}^m}\left[f_{|F|} \text{ doesn't satisfy } C\right]$$

Without loss of generality, let the perfect predictor f* be $f_{|F|}$, so $P_{(x_i,y_i)_{i=1}^m}\left[f_{|F|} \text{ doesn't satisfy } C\right] = 1$. So the above equation contains $|F| - 1$ Bernoulli probabilities we need to compute. Denote the random variable Z_j as:

$$Z_j = \begin{cases} 1, & \text{if } f_j \text{ does not satisfy condition C} \\ 0, & \text{if } f_j \text{ does satisfy condition C} \end{cases}$$

$$P(Z_j = 0) = P_{(x_i,y_i)_{i=1}^m}\left[f_j \text{ is bad}, f_j(x_i) = y_i\right] = P_{(x_i,y_i)_{i=1}^m}\left[f_j(x_i) = y_i | f_j \text{ is bad}\right] P[f_j \text{ is bad}]$$

Here, let's assume that $P[f_j \text{ is bad}]$ varies somewhat uniformly with ϵ, i.e. $P\left[f_j \text{ is bad}\right] = 1 - \epsilon$. Then, using the result in part 1i, we get:

$$P\left(Z_j = 0\right) < (1-\epsilon)^{m+1}$$
$$P\left(Z_j = 1\right) > 1-(1-\epsilon)^{m+1}$$

Hence:

$$P_{(x_i,y_i)_{i=1}^m}[\exists\epsilon \text{ bad } f \in F \text{ such that } f(x_i) = y_i] \leq 1-[1-(1-\epsilon)^{m+1}]^{|F|-1}$$

We see that this satisfies what we would expect in limiting conditions. For example, when $|F| \to \infty$, the upper bound of the probability of existing an epsilon-bad F that also matches all the training data becomes 1. This makes sense because increasing the number of functions in the class increases the probability that any one function could satisfy condition C. When |F|=1, the above probability is zero, which makes sense since the one predictor that exists is a perfect predictor. Also, as $m \to \infty$, the probability goes to zero, since increasing m makes it harder to match all of the training labels.

1iii.

Want:

$$P_{(x_i,y_i)_{i=1}^m}[\exists\epsilon - \text{bad } f \in F \text{ such that } f(x_i) = y_i] \leq \delta$$
$$1-[1-(1-\epsilon)^{m+1}]^{|F|-1} \leq \delta$$
$$[1-(1-\epsilon)^{m+1}]^{|F|-1} \geq 1-\delta$$
$$1-(1-\epsilon)^{m+1} \geq \sqrt[|F|-1]{1-\delta}$$

$$(1-\epsilon)^{m+1} \leq 1-\sqrt[|F|-1]{1-\delta}$$
$$m+1 \geq \frac{\log\left(1-\sqrt[|F|-1]{1-\delta}\right)}{\log(1-\epsilon)}$$
$$m \geq \frac{\log\left(1-\sqrt[|F|-1]{1-\delta}\right)}{\log(1-\epsilon)} - 1$$

Problem 2:

$$p(x) = (2\pi)^{-\frac{d}{2}} \exp\left\{-\frac{||x-\mu||^2}{2}\right\}$$

2i:

Since the dependence on x behaves as a negative squared exponential, p(x) is maximized when the exponent is zero, hence when $x = \mu$.

Alternatively, can take the derivative and set it to zero:

$$\frac{dp}{dx} = (2\pi)^{-\frac{d}{2}} \exp\left\{-\frac{||x-\mu||^2}{2}\right\}\{-(x-\mu)\} = 0$$

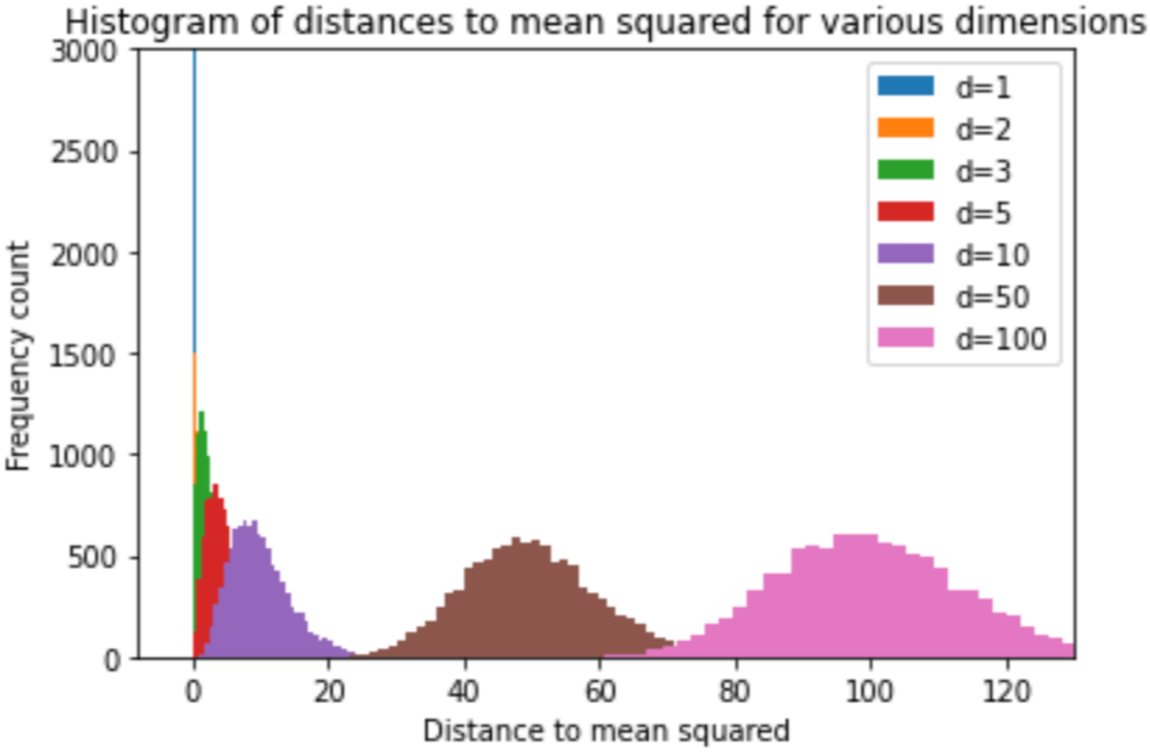
$$x = \mu$$

We can take the second derivative test to show this is a maximum, but we know that Gaussians don't have local minima, so $x = \mu$ is the maximum point.

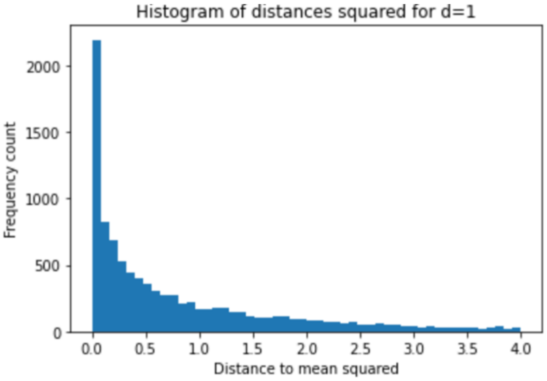
2ii:

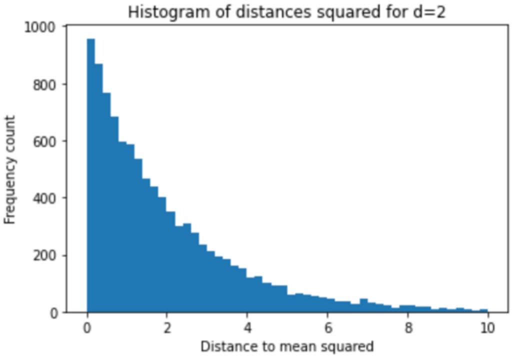
We see that the center of the samples dramatically shifts towards the right as we increase the number of dimensions. In other words, as we increase the number of dimensions, the majority of the data lie increasingly farther from the mean.

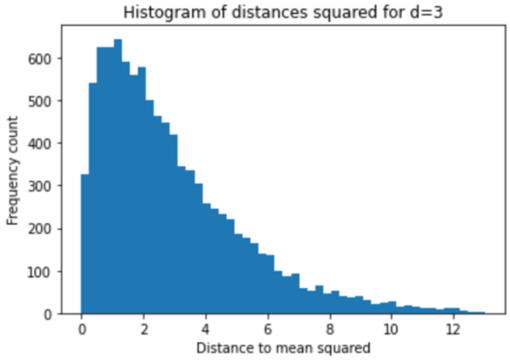
Plotting together:

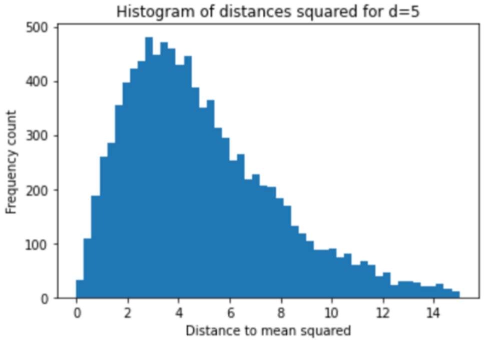


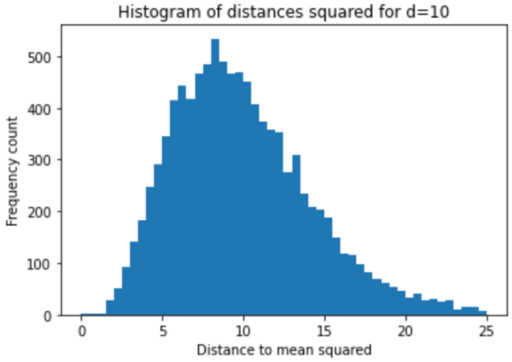
Plotting separately:

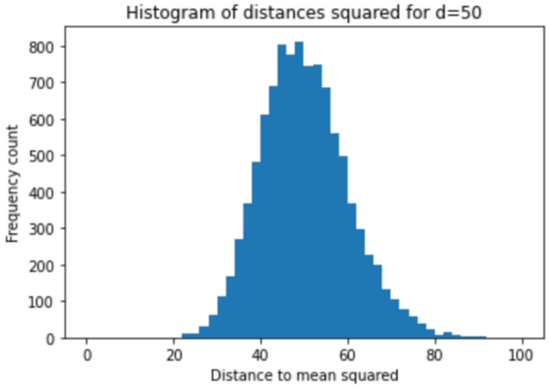


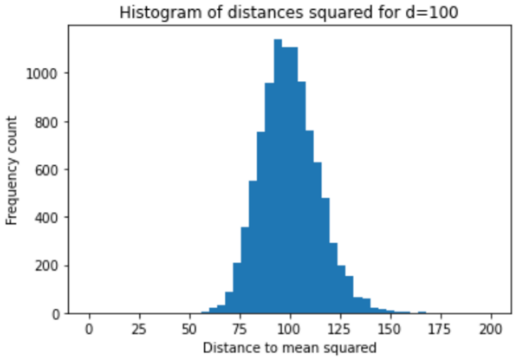












2iii:

$$E_{x\sim N(0,I_d)}\left[||x||^2\right]$$

$$= (2\pi)^{-\frac{d}{2}} \int_{R^d} \left||\vec{x}|\right|^2 \exp\left(-\frac{\left||\vec{x}|\right|^2}{2}\right) d^d\vec{x}$$

$$= (2\pi)^{-\frac{d}{2}} \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 \dots \int_{-\infty}^{\infty} dx_d \, (x_1^2 + x_2^2 + \dots + x_d^2) \exp\left(-\frac{x_1^2+x_2^2+\dots+x_d^2}{2}\right) \quad (\ * \ )$$

Consider just one term in this integral:

$$(2\pi)^{-\frac{d}{2}} \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 \dots \int_{-\infty}^{\infty} dx_d \, x_1^2 \exp\left(-\frac{x_1^2+x_2^2+\dots+x_d^2}{2}\right)$$

The integrals involving $dx_2 \dots dx_d$ treat x_1 as a constant, and hence each of these integrals integrate to $\sqrt{2\pi}$. So:

$$(2\pi)^{-\frac{d}{2}} \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 \dots \int_{-\infty}^{\infty} dx_d \, x_1^2 \exp\left(-\frac{x_1^2+x_2^2+\dots+x_d^2}{2}\right)$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dx_1 \, x_1^2 \exp\left(-\frac{x_1^2+x_2^2+\dots+x_d^2}{2}\right)$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dx_1 \, x_1^2 \exp\left(-\frac{x_1^2}{2}\right)$$

(shifting the center of the Gaussian does not change integration value with respect to x_1)

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x_1^2 \exp\left(-\frac{1}{2}x_1^2\right) dx_1$$
$$= -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x_1 \frac{d}{dx_1} \exp\left(-\frac{1}{2}x_1^2\right) dx_1$$
$$= -\frac{1}{\sqrt{2\pi}} \left[x_1 e^{-\frac{1}{2}x_1^2} |_{-\infty}^{\infty} - \int_{-\infty}^{\infty} e^{-\frac{1}{2}x_1^2} dx_1 \right]$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}x_1^2} dx_1 = 1$$

So each term in the integral (*) integrates to 1. There are d such terms and the integral in (*) is symmetric with respect to each feature, so:

$$(2\pi)^{-\frac{d}{2}} \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 \dots \int_{-\infty}^{\infty} dx_d \, (x_1^2 + x_2^2 + \dots + x_d^2) \exp\left(-\frac{x_1^2+x_2^2+\dots+x_d^2}{2}\right) = 1 * d$$
$$= d$$

$$E_{x \sim N(0, I_d)}\left[\left||x|\right|^2\right] = d$$

In hindsight, alternatively, we could've just used the fact that:

$$E\left[\left||x|\right|^2\right] = E[x_1^2 + \dots + x_d^2] = E[x_1^2] + \dots + E[x_d^2]$$

Note that:

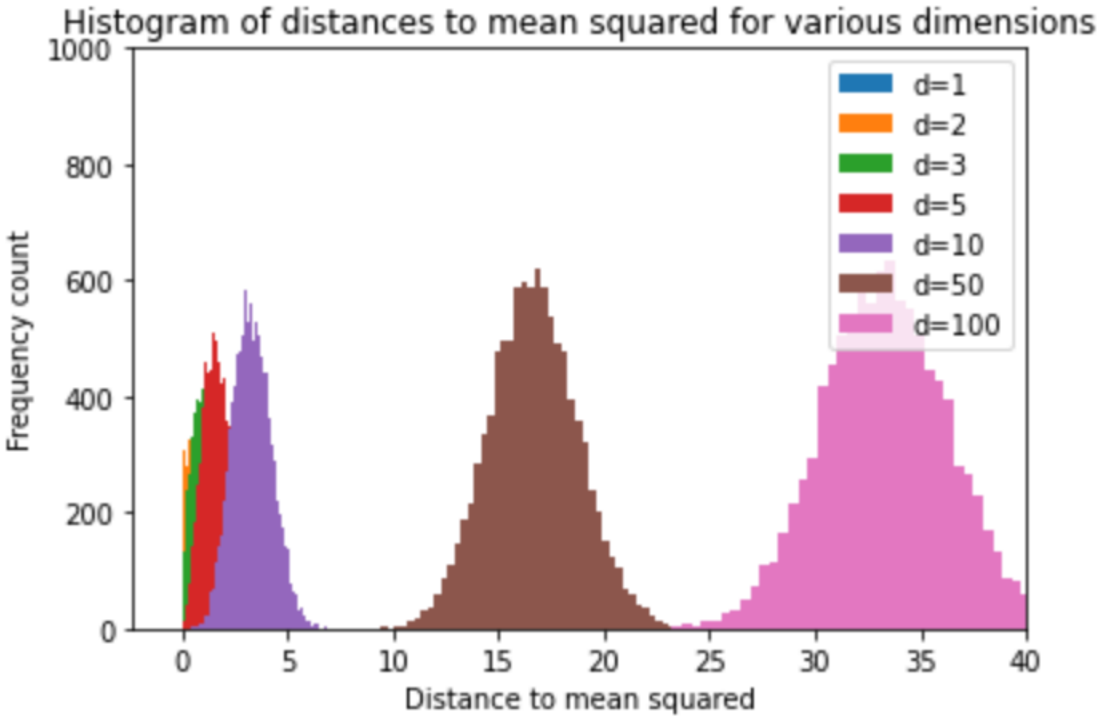
$$E[x_i^2] = Var(x_i) + E(x_i)^2 = 1 + 0 = 1$$

Then then:

$$E\left[||x||^2\right] = 1 * d = d$$

This is a general approach that does not depend on the distribution at hand. We see that this calculation agrees with our histograms.

2iv.



2v.

Again:

$$E\left[||x||^2\right] = E[x_1^2 + \cdots + x_d^2] = E[x_1^2] + \cdots + E[x_d^2]$$

This time, note that:

$$Var(x_i) = \frac{\left(1-(-1)\right)^2}{12} = \frac{1}{3}$$

So:

$$E[x_i^2] = Var(x_i) + E(x_i)^2 = \frac{1}{3} + 0 = \frac{1}{3}$$

Hence:

$$E\left[||x||^2\right] = \frac{1}{3} * d = \frac{d}{3}$$

We see that this calculation agrees with our histograms.

Problem 3:

i.

We use the marginalization property of multivariate Gaussians, which says that to find the marginal distribution we simply marginalize out the other variables outside of our interest.

Alternatively, use the results in part iii:

$$p(x_1, x_2) = N(x_1; \mu_1, \Sigma_{11}) N(x_2; b, A)$$

If we integrate out x2, we integrate over a normal distribution about x2, which gives 1, so:

$$p(x_1) = N(x_1; \mu_1, \Sigma_{11}) \int_{-\infty}^{\infty} dx_2 \, N(x_2; b, A) = N(x_1; \mu_1, \Sigma_{11})$$

Hence, $X_1 \sim N(\mu_1, \Sigma_{11})$.

$$p(x_1) = \frac{1}{(2\pi)^{\frac{d}{4}} |\Sigma_{11}|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1))$$

ii.

We simply need to verify that, in the exponent:

$$(x - \mu)^T \Sigma^{-1} (x - \mu) = (x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1) + (x_2 - b)^T A^{-1} (x_2 - b)$$

To achieve this, start with:

$$\begin{aligned}
&(x - \mu)^T \Sigma^{-1} (x - \mu) = ((x_1 - \mu_1)^T, (x_2 - \mu_2)^T) \begin{pmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{pmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \\
&= (x_1 - \mu_1)^T \Sigma^{11} (x_1 - \mu_1) + (x_1 - \mu_1)^T \Sigma^{12} (x_2 - \mu_2) + (x_2 - \mu_2)^T \Sigma^{21} (x_1 - \mu_1) \\
&\qquad + (x_2 - \mu_2)^T \Sigma^{22} (x_2 - \mu_2) \\
&= (x_1 - \mu_1)^T (\Sigma_{11}^{-1} + \Sigma_{11}^{-1} \Sigma_{12} A^{-1} \Sigma_{12}^T \Sigma_{11}^{-1})(x_1 - \mu_1) + (x_1 - \mu_1)^T (-\Sigma_{11}^{-1} \Sigma_{12} A^{-1})(x_2 - \mu_2) \\
&\qquad + (x_2 - \mu_2)^T (-\Sigma_{11}^{-1} \Sigma_{12} A^{-1})^T (x_1 - \mu_1) + (x_2 - \mu_2)^T A^{-1} (x_2 - \mu_2)
\end{aligned}$$

Note that $(x_1 - \mu_1)^T (-\Sigma_{11}^{-1} \Sigma_{12} A^{-1})(x_2 - \mu_2)$ is a scalar and $(x_2 - \mu_2)^T (-\Sigma_{11}^{-1} \Sigma_{12} A^{-1})^T (x_1 - \mu_1) = \{(x_1 - \mu_1)^T (-\Sigma_{11}^{-1} \Sigma_{12} A^{-1})(x_2 - \mu_2)\}^T$. Hence $(x_1 - \mu_1)^T (-\Sigma_{11}^{-1} \Sigma_{12} A^{-1})(x_2 - \mu_2) = (x_1 - \mu_1)^T (-\Sigma_{11}^{-1} \Sigma_{12} A^{-1})(x_2 - \mu_2)$. (Transpose of a scalar is itself).

Hence, so far we have:

$$\begin{aligned}(x-\mu)^T\Sigma^{-1}(x-\mu) \\ &= (x_1-\mu_1)^T(\Sigma_{11}^{-1}+\Sigma_{11}^{-1}\Sigma_{12}A^{-1}\Sigma_{12}^T\Sigma_{11}^{-1})(x_1-\mu_1) \\ &+ 2(x_1-\mu_1)^T(-\Sigma_{11}^{-1}\Sigma_{12}A^{-1})(x_2-\mu_2)+(x_2-\mu_2)^TA^{-1}(x_2-\mu_2)\end{aligned}$$

We can regroup some of the terms:

$$\begin{aligned}&= (x_1-\mu_1)^T\Sigma_{11}^{-1}(x_1-\mu_1)-(x_1-\mu_1)^T(\Sigma_{11}^{-1}\Sigma_{12}A^{-1})(x_2-b) \\ &\quad + \{(x_2-\mu_2)^T-(x_1-\mu_1)^T\Sigma_{11}^{-1}\Sigma_{12}\}A^{-1}(x_2-\mu_2)\end{aligned}$$

If we use the fact that the transpose of a sum is the sum of the transpose, the fact that covariance matrices are symmetric, and the fact that the transpose of the inverse is the inverse of the transpose, we can group the last two terms to get:

$$= (x_1-\mu_1)^T\Sigma_{11}^{-1}(x_1-\mu_1)+(x_2-b)^TA^{-1}(x_2-b)$$

As desired. To conclude, we must have:

$$p(x_1,x_2)=\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}[(x_1-\mu_1)^T\Sigma_{11}^{-1}(x_1-\mu_1)+(x_2-b)^TA^{-1}(x_2-b)]\right)$$

iii.

Given the results derived in part ii and the fact that:

$$|\Sigma|=|\Sigma_{11}||\Sigma_{22}-\Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12}|=|\Sigma_{11}||A|$$

We can factorize the joint distribution into:

$$\begin{aligned}p(x_1,x_2) &= \left\{\frac{1}{(2\pi)^{\frac{d}{4}}|\Sigma_{11}|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x_1-\mu_1)^T\Sigma_{11}^{-1}(x_1 - \mu_1)\right)\right\}\left\{\frac{1}{(2\pi)^{\frac{d}{4}}|\mathrm{A}|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x_2-b)^TA^{-1}(x_2-b)\right)\right\} \\ &= N(x_1;\mu_1,\Sigma_{11})N(x_2;b,A)\end{aligned}$$

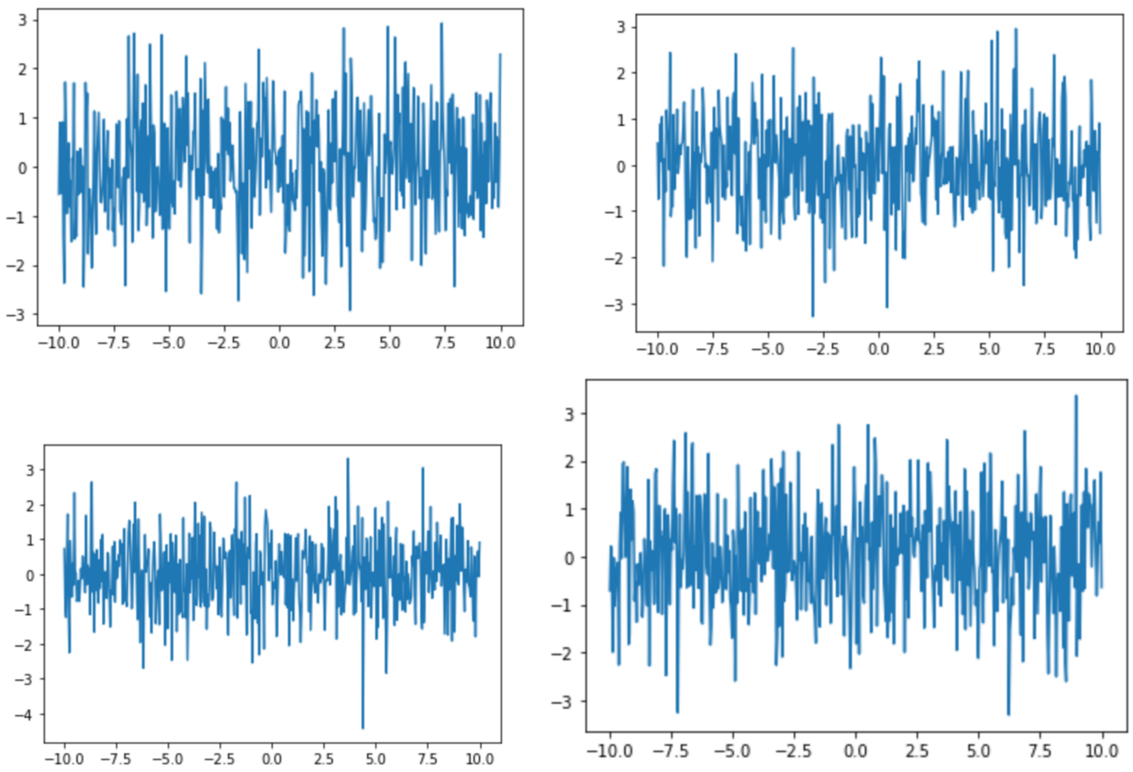
As desired.

iv.

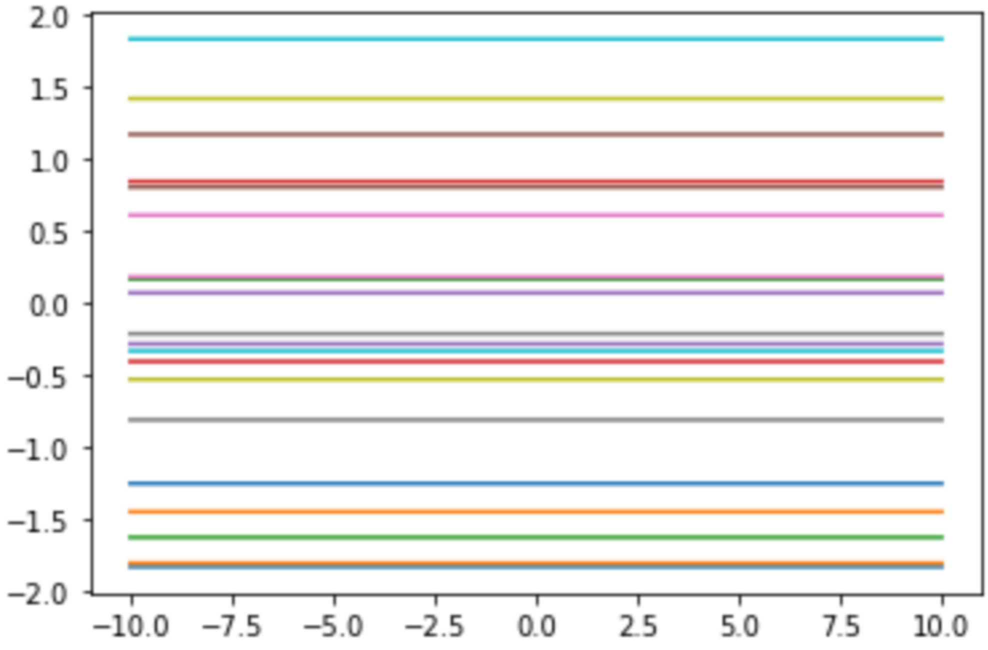
$$p(x_2|x_1)=\frac{p(x_1,x_2)}{p(x_1)}=\frac{1}{(2\pi)^{\frac{d}{4}}|\mathrm{A}|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x_2-b)^TA^{-1}(x_2-b)\right)$$

v.

We see that the functions are not smooth at all.

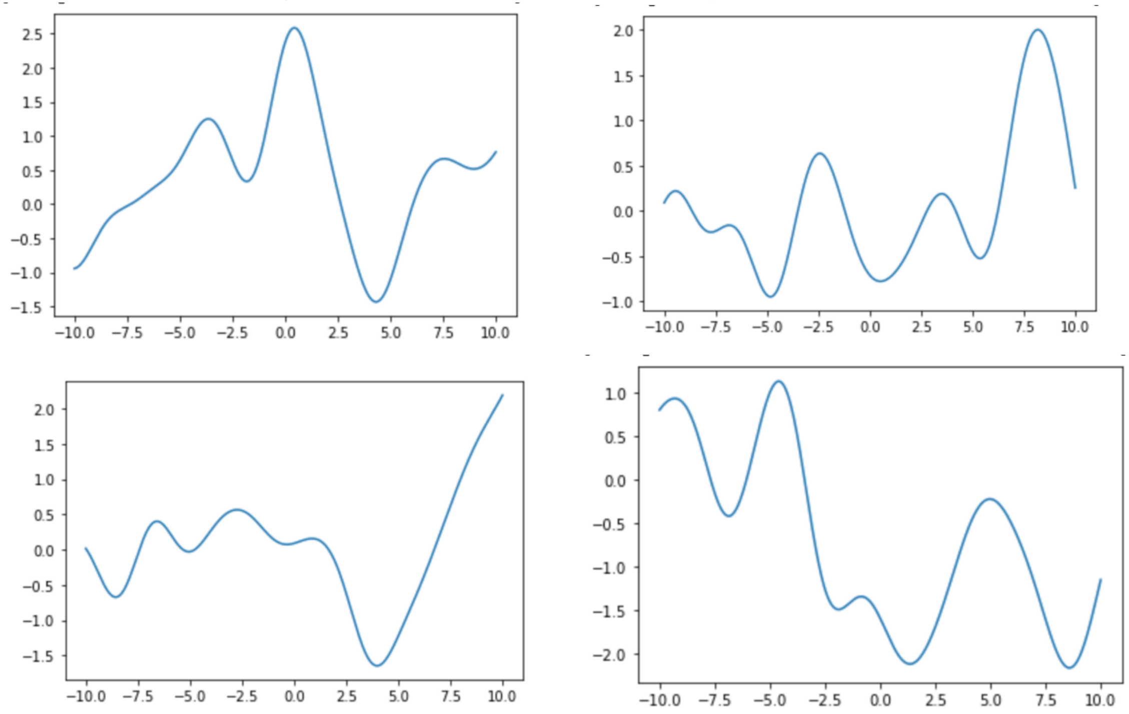


If we set the covariance matrix to all ones, we see that the oscillations are much smaller. In fact, if we plot multiple of such functions on the same graph, we observe the following:



We see that we get horizontal functions. This occurs because a covariance matrix filled with ones is singular and thus gives us a degenerate distribution with a definite constant function value.

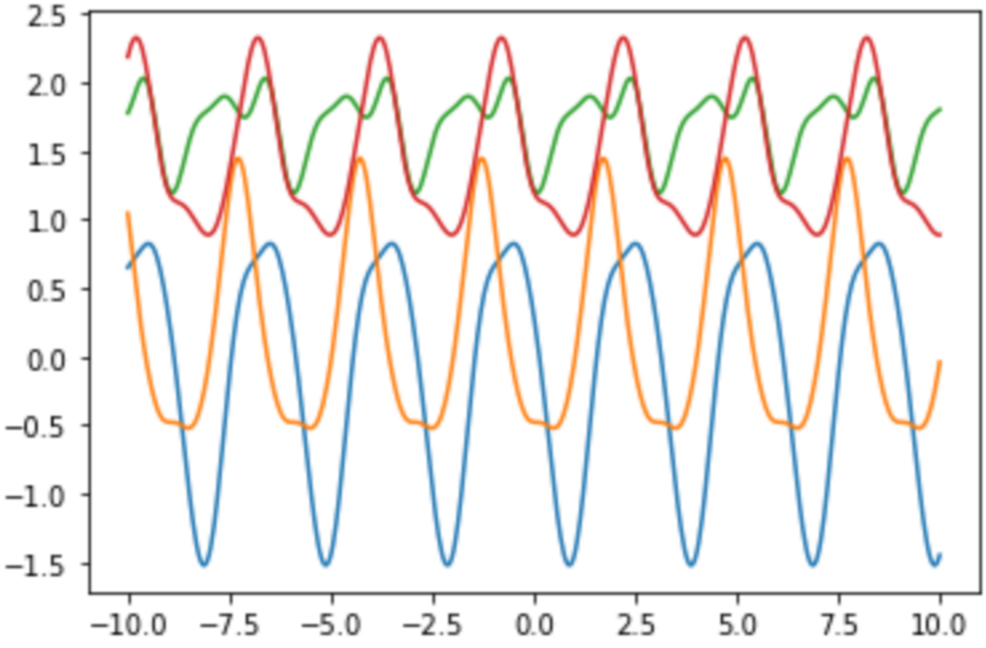
vi.
We see that this setting results in functions that are very smooth.



vii.
If we want random periodic functions, we may try to use a periodic kernel. Our setting for μ can be arbitrary, as this parameter only serves to shift the function. The period/frequency is controlled by the covariance matrix Σ.

If we want random periodic functions with period 3, we can try:

$$K(x_i, x_j) = \exp\left(-\frac{4}{5}\sin^2\left(\frac{\pi|x_i - x_j|}{3}\right)\right).$$



We see that the above 4 functions all have the same period of 3, as expected. Moreover, the kernel form of an exponential of a squared trigonometric function ensures that the covariance matrix is positive definite.

viii.

Using the result from part iv, we see that the posterior distribution is: (simply plug in $x_2 = Y, x_1 = \bar{Y}$

$$p(Y|\bar{Y}) = \frac{1}{(2\pi)^{\frac{d}{4}}|\mathrm{A}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(Y-b)^T A^{-1}(Y-b)\right)$$

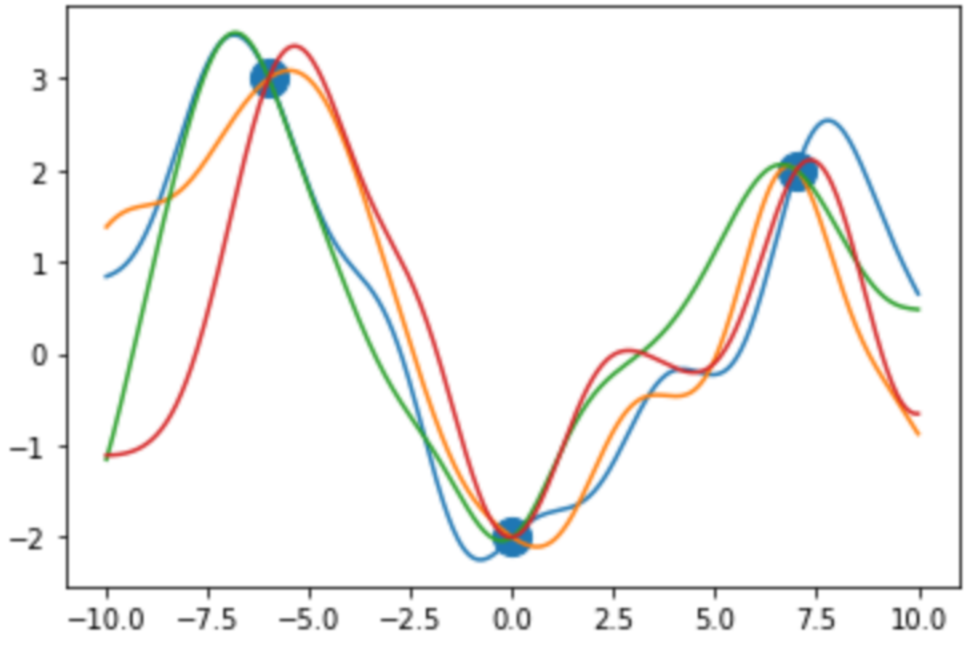
Where:

$b = \mu_n + \Sigma_{12}^T \Sigma_{11}^{-1}(\bar{Y} - \mu_m) = \mu_n + \Sigma_{mn}^T \Sigma_{mm}^{-1}(\bar{Y} - \mu_m) = \mu_n + [K(\bar{X}, X)]^T [K(\bar{X}, \bar{X})]^{-1}(\bar{Y} - \mu_m)$

And:

$$A = \Sigma_{nn} - \Sigma_{mn}^T \Sigma_{mm}^{-1} \Sigma_{mn} = K(X,X) - [K(\bar{X}, X)]^T [K(\bar{X}, \bar{X})]^{-1} K(\bar{X}, X)$$

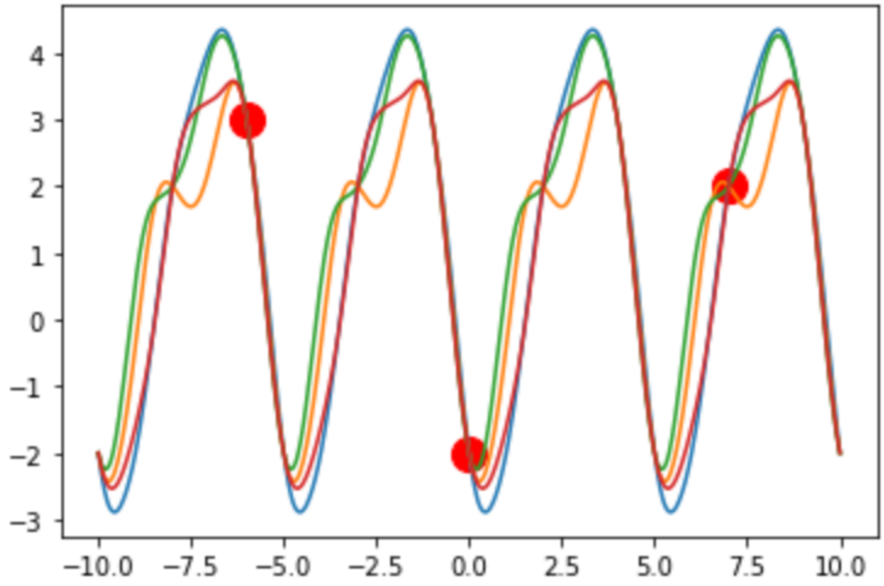
ix.



We see that the random functions drawn from the posterior distribution perfectly interpolate our training data, as expected.

x.

I adjusted the periodicity to be 5. We see that we get periodic functions that interpolate the training data.



xi.

The posterior is a gaussian distribution with mean $\mu = b$, where:

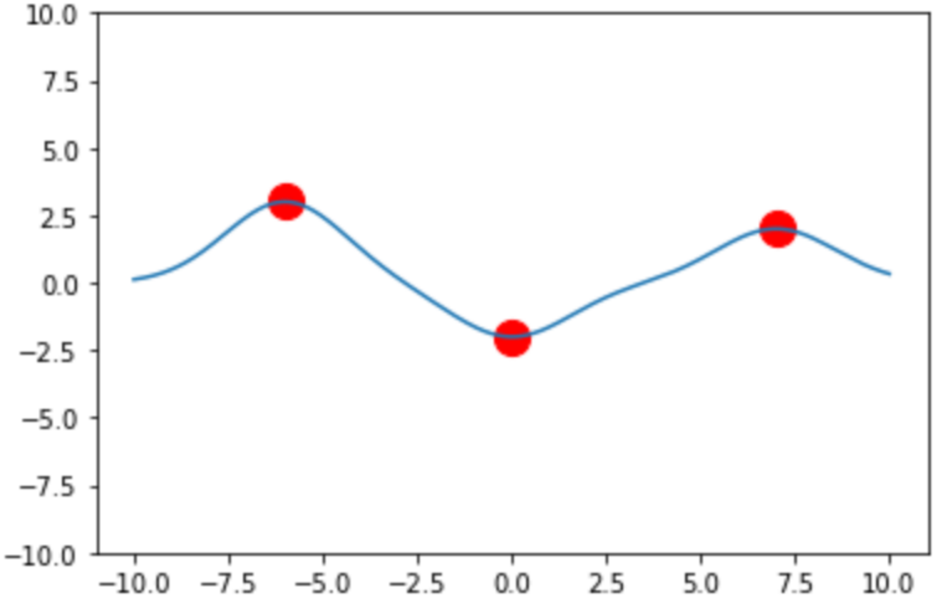
$$b = \mu_n + \Sigma_{12}^T \Sigma_{11}^{-1}(\bar{Y} - \mu_m) = \mu_n + \Sigma_{mn}^T \Sigma_{mm}^{-1}(\bar{Y} - \mu_m) = \mu_n + [K(\bar{X}, X)]^T [K(\bar{X}, \bar{X})]^{-1}(\bar{Y} - \mu_m)$$

xii.

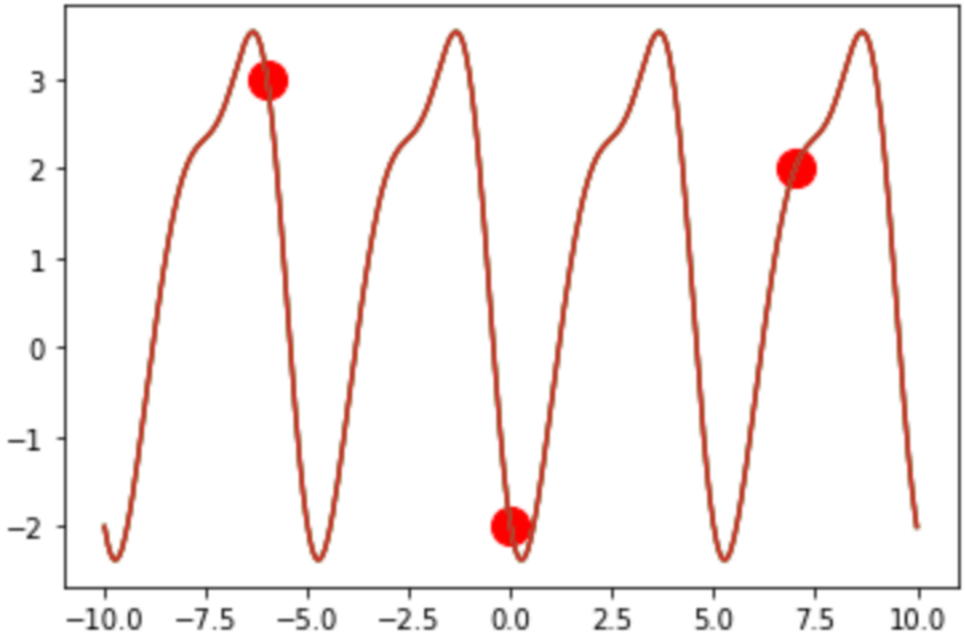
The “mean” functions were obtained by sampling from the posterior with zero covariance matrix.

“Mean” function for problem 3.ix:

```
(-10.0, 10.0)
```



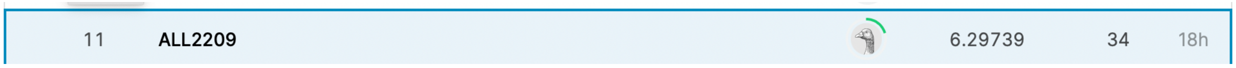
“Mean” function for problem 3.x:



Problem 4:

Neural network:
The best result achieved so far (MAE of 6.29739 on Kaggle test labels) was through a dense neural network with ReLU activation, MAE loss, and 2 hidden layers of size 90 and 20. I kept the depth of the network shallow so as to incentivize simpler representations, thereby preventing overfitting. I also found that normalizing the input data helped a lot in improving the convergence rate and test accuracy. In fact, the test error rises dramatically to nearly ~90 when normalization is not applied. Additionally, I used an Adam optimizer with weight decay (L2 regularization), which incentivizes the neural network to choose smaller weights and thus decreases chances of overfitting.

Random forest regressor:
I also tried using a random forest regressor since, from the previous homework, it is known that the random forest classifier is able to significantly decrease the bias via majority vote, hence yielding good generalizability. I chose max_leaves = 500 – which was determined via first varying the maximum number of leaves for single decision trees and selecting the number of leaves that yields lowest generalization error - and num_estimators = 100. This gave me a test error of around ~6.68. Since training a random forest regressor of this size takes a long time, I did not continue to experiment with other values for num_estimators.



11	ALL2209	6.29739	34	18h