ECE 8110 Machine Learning Homework Assignment 2

Febuary 7, 2014

Salvatore Giorgi
salvatore.giorgi@temple.edu

Exercise 1. *Generate two 2D GRVs with a mean of* $(1, 1)$ *and* $(-1, -1)$*. Plot the theoretical probability of error for an ML classifier as a function of the prior probabilities and the covariance matrices. Since there are a number of degrees of freedom, determine the best way to visualize the results.*

Solution: To calculate the probability of error for this classifier we first investigate the classifier decision surface as a function of the prior probabilities and covariance matrix. For the first case, we consider two GRVs with each with identity covariance matrices. With identity covariance matrices, each RV's distribution will be symmetric about its mean. Thus, we can expect the decision surface to be a straight line. With one RV centered at $(1, 1)$ and another centered at $(-1, -1)$, we expect the decision line to be of the form $y = -x + b$ with b depending on the prior probabilities. If the prior probabilities are equal, then we expect this straight line to be equi-distant from the means of the two distributions, i.e, $b = 0$. As the priors of one distribution grows, we expect the line to move towards the opposite distribution, since it is more likely that the data observed originates from the former distribution, rather than the later. Note that we have the following constraint for the prior probabilities: $P_1 + P_2 = 1$. We plot the support region for both RVs and the decision line for $P_1 \in \{0.1, \cdots, 0.9\}$. Figure **??** shows the results of our simulation. As we can see, the results match our intuitive guesses above.

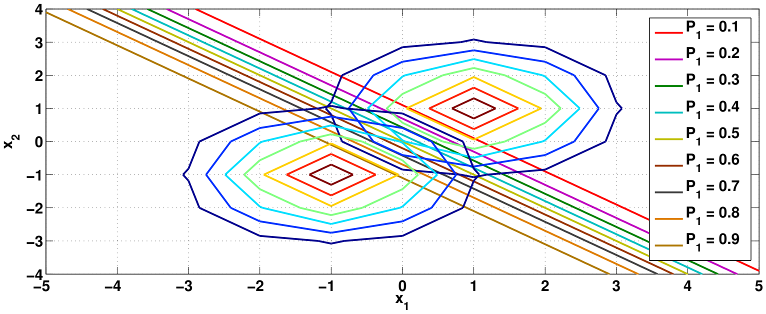


Figure 1: Support Region and Decision Surface for $\Sigma_1 = \Sigma_2 = I^{2\times 2}$

These results could also be deduced from the following equations, found in [**?**]

$$g_i = x'W_i x + w_i' x + w_{i0}, \tag{1}$$

where

$$W_i = -\frac{\Sigma_i^{-1}}{2}, \quad w_i = \mu_i \Sigma_i^{-1}, \tag{2}$$

$$w_{i0} = -\frac{\mu_i' \Sigma_i^{-1} \mu_i}{2} - \frac{\log(\det \Sigma_i)}{2} + \log(P_i) \tag{3}$$

for $i = 1, 2$ and $x = [x_1, x_2]'$. For $g = g_1 - g_2$ and setting $g = 0$, we have

$$g = g_1 - g_2 = 2x_1 + 2x_2 + \log\big(\frac{P_1}{P_2}\big) = 0 \Rightarrow x_2 = -x_1 - \frac{1}{2}\log\big(\frac{P_1}{P_2}\big).$$

Thus, we see that the quadratic terms will cancel out, leaving a linear equation, whose intercept depends on the prior probabilities. Not only does this show us that we will have a straight line in equal covariance case, but it shows us that if we do not have equal covariances, our decision surface will be some sort of parabola.

To find the theoretical error for this classifier, we integrate the pdf of each RV on the incorrect side of the decision surface. With the constraint that the total probability must add up to 1, we multiply each integral by its prior probability. First, we consider the case when $\Sigma_1 = \Sigma_2 = I^{2\times 2}$ and $P_1 \in \{0.1, \cdots, 0.9\}$. Since we are consider distributions equi-distant from the origin, with equal covariance, and the constraint that $P_1 + P_2 = 1$, we expect the theoretical error to exhibit some symmetry as we vary the priors. The theoretical errors are listed in Table **??** and plotted in Fig. **??**.

Table 1: Theoretical Error for $\Sigma_1 = \Sigma_2 = I^{2\times 2}$ and Varying Priors

P_1	P_2	Theoretical Error
0.1	0.9	0.0390
0.2	0.8	0.0583
0.3	0.7	0.0700
0.4	0.6	0.0766
0.5	0.5	0.0786
0.6	0.4	0.0766
0.7	0.3	0.0700
0.8	0.2	0.0583
0.9	0.1	0.0390

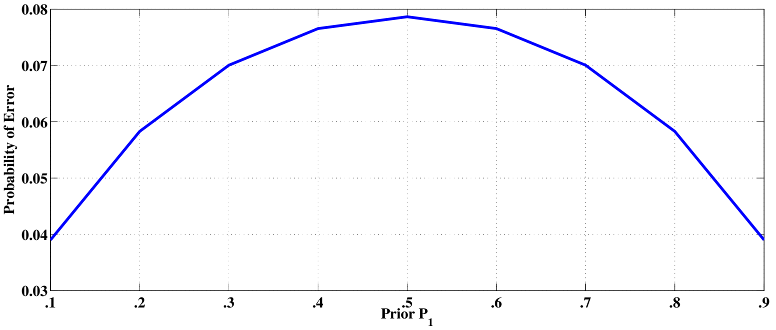


Figure 2: Probability of Error for GRVs with $\Sigma = I$ and Varying Priors

As stated above, to calculate the theoretical probability of error, one must integrate each pdf which lies on the incorrect side of the decision surface. For equal covariance matrices, our decision surface is a line, so our integration is over a rectangular area. For parabolic decision surfaces our integration area becomes much more complicated. I was not able to implement this integration in my software. While we cannot show the theoretical error, we are able to show how the decision surface changes when the two distributions have unequal covariance matrices. For simplicity, we will change only the covariance of the first distribution and keep the second distribution as $\mathcal{N}((-1,-1), I^{2\times 2})$. We consider three cases with the following covariance matrices

$$\Sigma_{21} = \begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix}, \quad \Sigma_{22} = \begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix}, \quad \Sigma_{23} = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}.$$

In each of the above cases we consider $P_1 \in \{0.1, \cdots, 0.9\}$. We make the following observations:

1) We saw in the previous homework assignment that changing the upper left term of the covariance matrix in turn makes the support grow horizontally, changing the lower right term in the covariance matrix makes the support grow vertically, and changing the diagonal terms makes the support grow along its diagonal axis.
2) In the case of equal covariances, changing the prior translated the decision surface.
3) In Σ_{21}, Σ_{22} we now have a value of 10, which is larger than in the case when we considered $\Sigma_2 = I^{2\times 2}$. Also, for Σ_{23} we now have a value of 0.6 in the off diagonal terms, which is smaller than when we considered $\Sigma_2 = I^{2\times 2}$.

From these observations, we expect the following, respectively,

1) The axis of symmetry of the decision surface (parabola) will be perpendicular to the axis of the support region which is changing.
2) Any change in the priors will translate the parabola along the axis of symmetry.
3) With the support region of the random variable corresponding to Σ_{21} and Σ_{22} growing, the decision surface will bend away from this distribution. With the support region of the random variable corresponding to Σ_{23} shrinking, the decision surface should bend towards this distribution.

Figures **??**, **??**, and **??** show the decision surfaces plotted with the support regions of the RVs for Σ_{21}, Σ_{22}, and Σ_{23}, respectively.

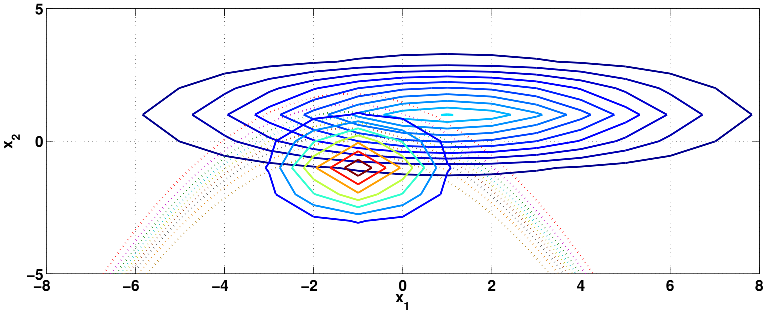


Figure 3: Support Region and Decision Surface for Σ_{21}

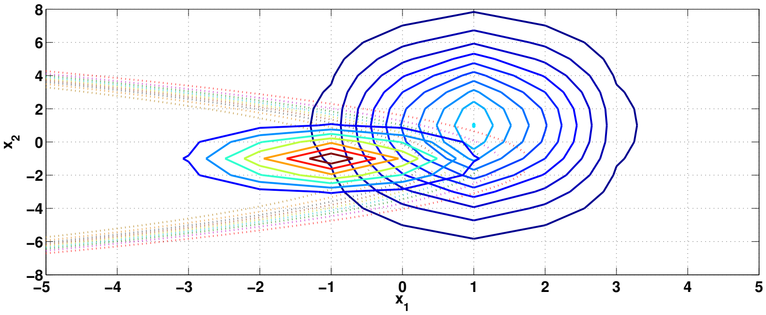


Figure 4: Support Region and Decision Surface for Σ_{22}

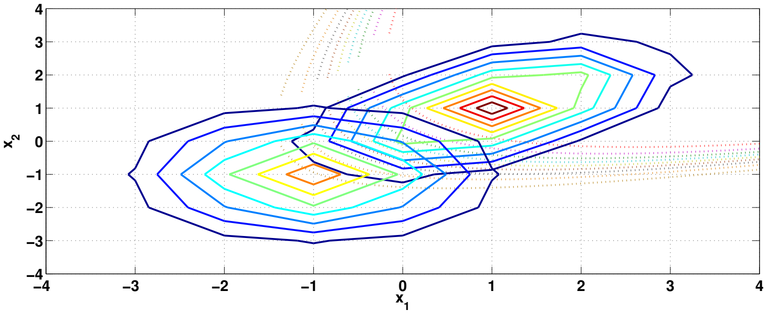


Figure 5: Support Region and Decision Surface for Σ_{23}

Exercise 2. *Compare the theoretical results in Exercise* **??** *to those obtained when you construct an ML classifier by generating* $[100, 1000, 10000, 100000]$ *random variables for each cass. Estimate the means and variances from the data. Only consider the equal priors case for this example, and focus on a small representative set of covariances.*

Solution: For the first part, we set $\Sigma_1 = \Sigma_2 = I^{2\times 2}$ and vary both the number of data points N and the prior P_1. To find the experimental error we count the number of data points on the wrong side of the decision surface, divide by the total number of points, and multiply this by the prior probabilities. We expect the experimental errors to converge to the theoretical errors as $N \to \infty$. The results of our experiments can be seen in Figure **??** and Table **??**. As expected, as N increases, we see that the experimental errors converge to the theoretical errors.

Table 2: Experimental Error for Varying Priors and Number of Data Points

	$P_1 = 0.1$	$P_1 = 0.2$	$P_1 = 0.3$	$P_1 = 0.4$	$P_1 = 0.5$	$P_1 = 0.6$	$P_1 = 0.7$	$P_1 = 0.8$	$P_1 = 0.9$
$N = 10$	0.0300	0.0600	0.1600	0.1800	0.1500	0.1600	0.0300	0.0200	0.0200
$N = 100$	0.0470	0.0800	0.0900	0.1000	0.1050	0.0900	0.0960	0.0769	0.0540
$N = 1000$	0.0401	0.608	0.0685	0.0728	0.0725	0.0736	0.0749	0.0620	0.0429
$N = 10000$	0.0443	0.0612	0.0737	0.0799	0.0811	0.0764	0.0702	0.0591	0.0426
$N = 100000$	0.0428	0.0613	0.0713	0.0761	0.0781	0.0765	0.0713	0.0612	0.0431

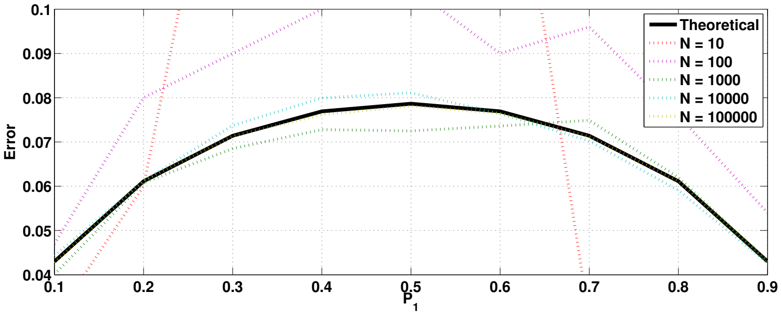


Figure 6: Theoretical and Experimental Error Rates

We now consider the case when we have the following covariance matrices

$$\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix}.$$

The support regions and decision surface are seen in Figure **??** from Exercise 1. Again, we vary both the number of data points N and the prior P_1. While we don't have a theoretical error to compare these to, we do expect the error curve to smooth and for the distance between the curves to shrink as N increases. Figure **??** shows the results of this simulation.

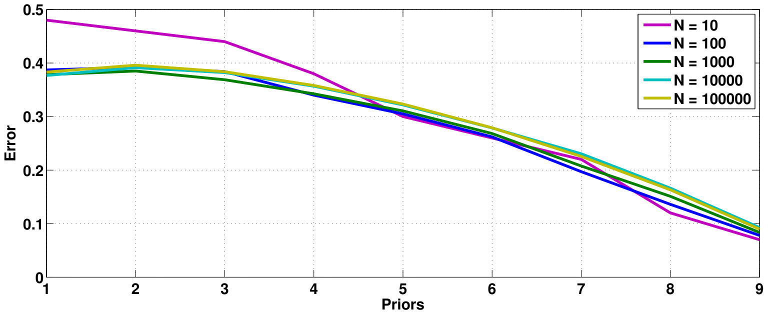


Figure 7: Experimental Error Rates for Unequal Covariance Matrices

Exercise 3. *Generate two 2D GRVs with means $[-2,-5]$ and $[3,6]$ an covariance matrices*

$$\begin{bmatrix} 2 & 1.5 \\ 1.5 & 2 \end{bmatrix} \quad \begin{bmatrix} 3 & -2 \\ -2 & 3 \end{bmatrix} \tag{4}$$

respectively.

(a) Construct an ML esitmator and measure the error rate.

(b) Convert each GRV to another GRV with identity covarance matrix by performing Principal Component Analysis (PCA).

(c) Classify each data point y transforming it to the PCA space using a whitening transformation and computing the distance from the mean. Select the class assignment by choosing the class that has the smallest distance. This is essentualy an ML classifier, but implemented in a slightly different way. Do your results match part (a)?

(d) Examine the eigenvectors of each covariance matrix and relate those to the support region for each GRV. It is preferable to visualize this with a graph of the vectors overlaid on the support region.

Solution of (a):

As discussed above, a non-diagonal covariance matrix will give us a curved decision region. Hence, to find the probability of error, we must integrate the GRV over regions defined by a polynomial, which is not easy to do analytically. So we generate N data points and count how many data points are on the incorrect side of the decision surface. As $N \to \infty$ we expect the experimental probability of error to converge to the theoretical probability of error. Table **??** shows the result of this experiment.

Table 3: Experimental Error for Varying Priors and Number of Data Points

	$P_1 = 0.1$	$P_1 = 0.2$	$P_1 = 0.3$	$P_1 = 0.4$	$P_1 = 0.5$	$P_1 = 0.6$	$P_1 = 0.7$	$P_1 = 0.8$	$P_1 = 0.9$
$N = 100$	0.0470	0.0800	0.0900	0.1000	0.1050	0.0900	0.0960	0.0769	0.0540
$N = 1000$	0.0401	0.608	0.0685	0.0728	0.0725	0.0736	0.0749	0.0620	0.0429
$N = 10000$	0.0443	0.0612	0.0737	0.0799	0.0811	0.0764	0.0702	0.0591	0.0426

Solution of (b): Let

$$\Sigma_1 = \begin{bmatrix} 2 & 1.5 \\ 1.5 & 2 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 3 & -2 \\ -2 & 3 \end{bmatrix}$$

and define D_i to be a diagonal matrix whose diagonal elements correspond to the eigenvalues of Σ_i. Also, define V_i to be a matrix whose columns are right eigenvectors of Σ_i with the added constraint that the eigenvectors are orthonormal. This gives us

$$\Sigma_i V_i = V_i D_i.$$

From this, we get the following:

$$\Sigma_i V_i = V_i D_i \Leftrightarrow V_i^{-1} \Sigma_i V_i = D_i = D_i^{\frac{1}{2}} D_i^{\frac{1}{2}} \Leftrightarrow D_i^{-\frac{1}{2}} V_i^{-1} \Sigma_i V_i D_i^{-\frac{1}{2}} = I.$$

With the values of Σ_1, Σ_2 above, we have

$$V_1 = \begin{bmatrix} -0.7071 & 0.7071 \\ 0.7071 & 0.7071 \end{bmatrix} \quad D_1 = \begin{bmatrix} 0.5 & 0 \\ 0 & 3.5 \end{bmatrix}$$

$$V_2 = \begin{bmatrix} -0.7071 & -0.7071 \\ -0.7071 & 0.7071 \end{bmatrix} \qquad D_2 = \begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix}.$$

Solution of (c):

Denote our original data sets by X_1, X_2. From our solution to Part (b), we get the following transformed data sets

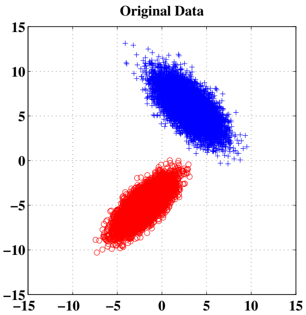
$$Y_1 = V_1^{-1} D_1^{-\frac{1}{2}} V_1^T X_1 \quad \text{and} \quad Y_2 = V_1^{-1} D_2^{-\frac{1}{2}} V_2^T X_2.$$

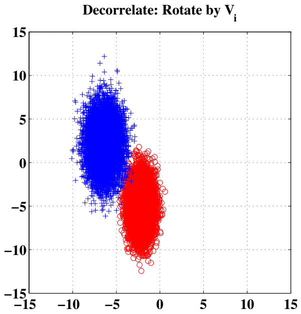
We see that the transfomation $Y_i = V_i^T X_i$ decorelates the data, $Y_i = D_i^{-\frac{1}{2}} V_i^T X_i$ decorelates and whitens the data, and $Y_i = V_i^{-1} D_i^{-\frac{1}{2}} V_i^T X_i$ decorelates, whitens, and then rotates the data back into the original space, as seen in Fig. **??**. Note that the transformation $Y_i = V_i^T X_i$ orients the data such that the principal axes of the data are aligned with the axes along which the data has the largest (orthogonal) variance.

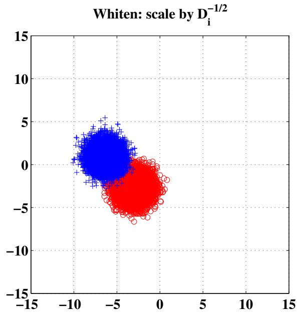
Table 4: Theoretical Error for PCA Space and Varying Priors

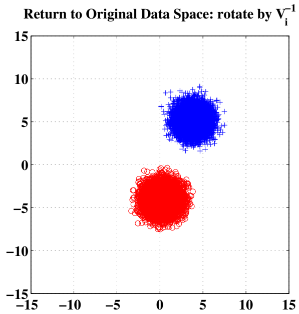
P_1	P_2	Theoretical Error
0.1	0.9	0.0038
0.2	0.8	0.0054
0.3	0.7	0.0063
0.4	0.6	0.0068
0.5	0.5	0.0069
0.6	0.4	0.0068
0.7	0.3	0.0063
0.8	0.2	0.0054
0.9	0.1	0.0038

Solution of (d):


Original Data


Decorrelate: Rotate by V_i


Whiten: scale by $D_i^{-1/2}$


Return to Original Data Space: rotate by V_i^{-1}

Exercise 4. *Following the example presented in the notes, assume you have a 1D GRV with mean and variance. Demonstrate estimation of the mean and variance using theoretical results derived in class for a Bayesian estimation. Compare this to an ML estimation. Show convergence as the number of data points increased from 100 to 100000.*

Solution: We consider a random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, with unknown mean μ and known variance σ^2, and generate N data points from this distribution. From the data points, we use both a Bayesian and Maximum Likelihood estimation to estimate the mean μ of the distribution of X.

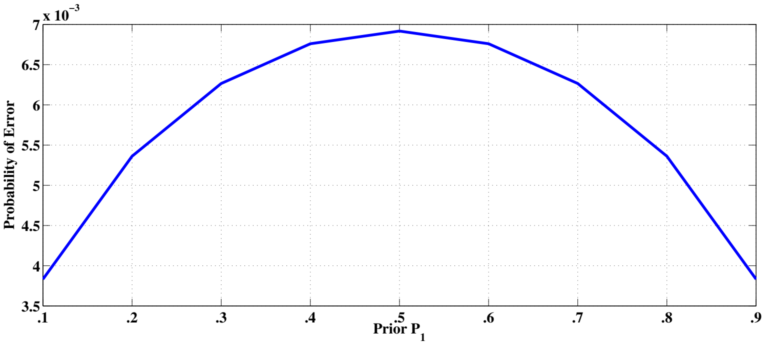


Figure 8: Probability of Error in PCA Space and Varying Priors

For the Maximum Likelihood case, we consider all

For the Bayesian case, we assume μ has a known prior distribution $p(\mu) \sim \mathcal{N}(\mu_0, \sigma_0^2)$. Here μ_0 is our best prior guess for μ and σ_0^2 measures our uncertainty about this guess. Our problem then reduces to estimating the true value of μ from the data $\mathcal{D}$, or in other words, determining the distribution of $p(\mu|\mathcal{D})$. Since we are conditioning on the data $\mathcal{D}$, one would expect the distribution $p(\mu|\mathcal{D})$ to depend on the number of samples N, which is indeed the case. It can be shown [**?**] that, if we assume $p(\mu|\mathcal{D}) \sim \mathcal{N}(\mu_N, \sigma_N^2)$, for the univariate Gaussian X, we have

$$\mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\hat{\mu}_N + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 \tag{5}$$

$$\sigma_N^2 = \frac{\sigma_0^2\sigma^2}{N\sigma_0^2 + \sigma^2} \tag{6}$$

where $\hat{\mu}_N$ is the sample mean, computed as

$$\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{7}$$

These equations show us that to find μ_N we use both our prior information and the information derived from the samples. Also, we see that as $N \to \infty$, our measure of uncertainty about μ_N, i.e., σ_N^2 approaches 0.

For this example we consider $X \sim \mathcal{N}(-10, 5)$ and assume our prior $p(\mu) \sim \mathcal{N}(1, 1)$. In the notation above we have $\mu = -10$, $\sigma^2 = 5$, $\mu_0 = 1$, and $\sigma_0^2 = 1$. Note, that our ”best guess” for μ is not too far from the correct value ($\mu_0 = 1$ and $\mu = -10$) and our uncertainty is relatively small ($\sigma_0^2 = 1$). We consider $N \in \{10, 20, 30, \cdots, 100000\}$. Figure **??** shows the values of the Bayesian and Maximum Likelihood estimates of the means. Figure **??** shows our uncertainty about our estimate of the mean σ_N^2. Figure **??** shows $p(\mu|\mathcal{D})$.

As another example, we consider $X \sim \mathcal{N}(-10, 5)$ and our prior $p(\mu) \sim \mathcal{N}(10000, 50)$. Here we note that both our initial guess for μ and our uncertainty about this guess are much greater than above. Again, $N \in \{10, 20, 30, \cdots, 100000\}$. Figure **??** shows the values of the Bayesian and Maximum Likelihood estimates of the means. Figure **??** shows our uncertainty about our estimate of the mean σ_N^2. Figure **??** shows $p(\mu|\mathcal{D})$.

Finally, we consider $X \sim \mathcal{N}(-10, 5)$ and our prior $p(\mu) \sim \mathcal{N}(-10, 1)$, which is saying we have a good initial guess for our value of μ. Again, $N \in \{10, 20, 30, \cdots, 100000\}$. Figure **??**

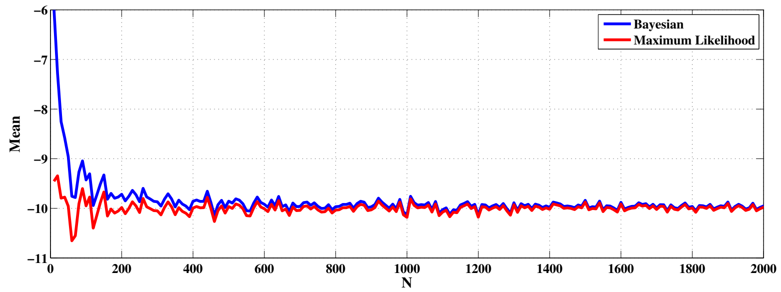


Figure 9: Values of μ_N and $\hat{\mu}_N$ for varying N

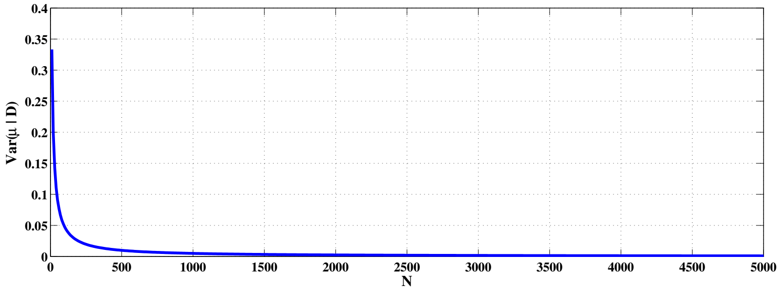


Figure 10: Values of σ_N^2 for varying N

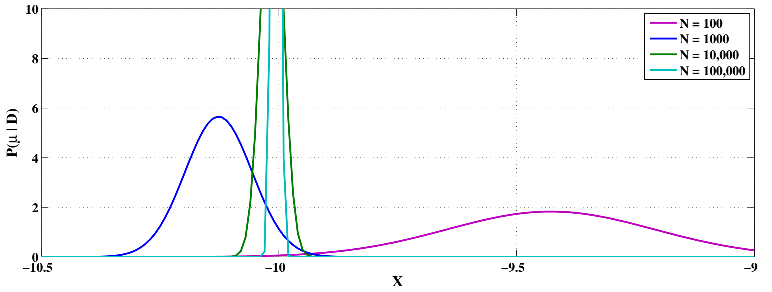


Figure 11: Probability distribution $p(\mu|\mathcal{D})$

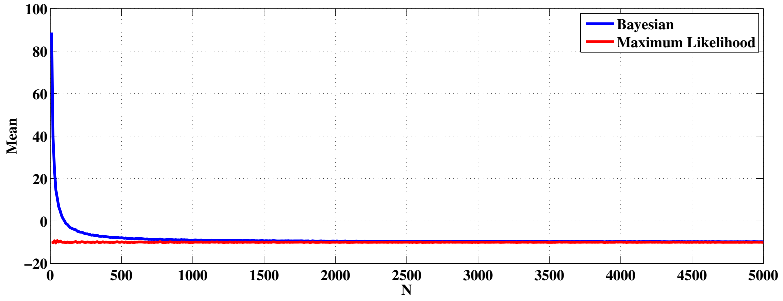


Figure 12: Values of μ_N and $\hat{\mu}_N$ for varying N

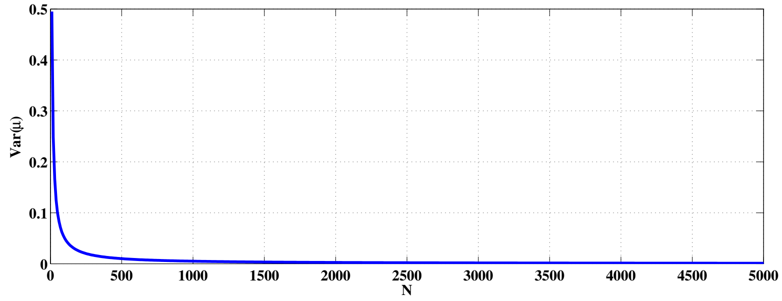


Figure 13: Values of σ_N^2 for varying N

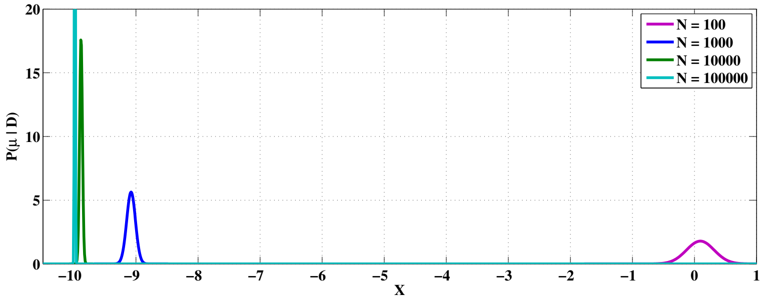


Figure 14: Probability distribution $p(\mu|\mathcal{D})$

shows the values of the Bayesian and Maximum Likelihood estimates of the means. Figure **??** shows our uncertainty about our estimate of the mean σ_N^2. Figure **??** shows $p(\mu|\mathcal{D})$.

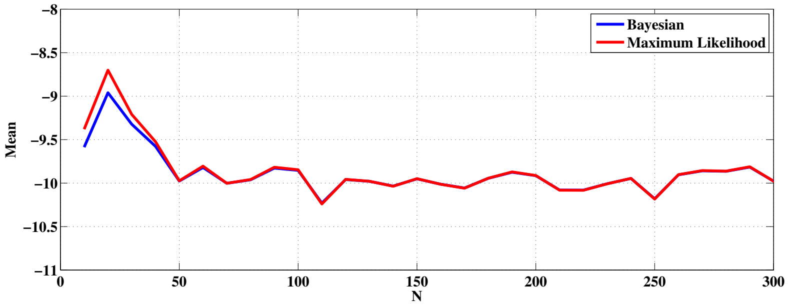


Figure 15: Values of μ_N and $\hat{\mu}_N$ for varying N

In the first two cases, we see that since our inital guess for μ was not correct, it took additional data points to get the correct estimate. Also, in both cases, we see that after 1000 data points both techniques approximately give the same estimate. In the last example, since our inital guess was correct, we see that with fewer sample points, we have a better estimate of the true value of the mean. In all three cases, we see our uncertainty σ_N^2 decrease to 0 as N increases. Additionally, as N increases, our distribution $p(\mu|\mathcal{D})$ is centered closer to the true mean and the variance of each distribution decreases.

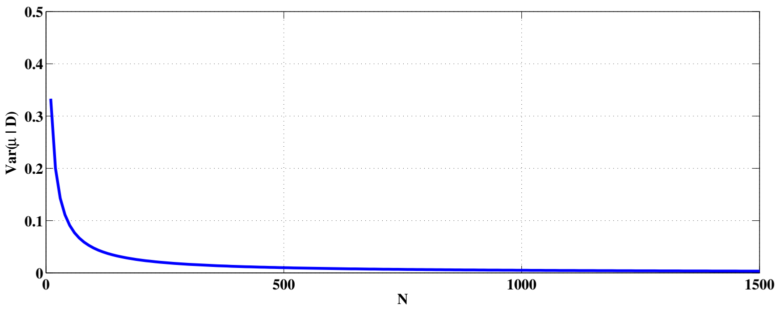


Figure 16: Values of σ_N^2 for varying N

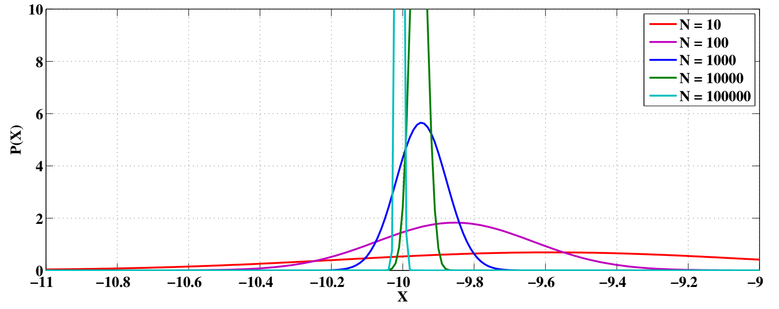


Figure 17: Probability distribution $p(\mu|\mathcal{D})$

Exercise 5. *Consider the Bayesian estimation of the mean of a 1D GRV. Suppose you are given the prior for the mean as $p(\mu)$ $\mathcal{N}(\mu_0, \sigma_0)$. Generate 1000 data points for $\mathcal{N}(1, 1)$. Assume $\mu_0 = 0$ and $\sigma_0 = 1$ (assume the variance is known). Plot the Bayesian estimate of $p(x|D)$ and μ as a function of the number of data points, n. Do this for 100, 1000, 10000, and 100000 points. Explain your results.*

Solution: We consider a random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, with given prior $p(\mu) \sim \mathcal{N}(\mu_0, \sigma_0^2)$. Let $\mu = 1$, $\sigma^2 = 1$, $\mu_0 = 0$, and $\sigma_0^2 = 1$. We note that the conditional distribution $p(x|\mathcal{D})$, with data $\mathcal{D}$ and N data points, follows a normal distribution $\mathcal{N}(\mu_N, \sigma^2 + \sigma_N^2)$, where the parameters are defined as follows:

$$\mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\hat{\mu}_N + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 \tag{8}$$

$$\sigma_N^2 = \frac{\sigma_0^2\sigma^2}{N\sigma_0^2 + \sigma^2} \tag{9}$$

where $\hat{\mu}_N$ is the sample mean, computed as

$$\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{10}$$

The parameter μ_N is our estimate of the true mean given the data and our prior knowledge, and σ_N^2 is our incertainty of this estimate. From the equations, we can see that as $N \to \infty$ both $\mu_N \to \hat{\mu}_N$ and $\sigma_N^2 \to 0$. Also, we have that $\hat{\mu}_N \to \mu$ as N approaches infinity. One expects that as we aquire more and more data, i.e. N increases, our estimate of the mean should approach the true value of the mean. Hence, at the same time our uncertainty of our estimate should decrease at the same time. We see both of these intuitive ideas reflected in the above equations. Therefore, we expect that as $N \to \infty$ we will have $p(x|\mathcal{D}) \to p(x)$.

To see this, we let $N \in \{10, 20, 30, \cdots, 100000\}$ and plot both $p(x|\mathcal{D})$ and μ_N. Figure **??** shows the plots of $p(x|\mathcal{D})$ for $N \in \{10, 20, 100, 10000, 100000\}$, as well as a plot of $\mathcal{N}(1, 1)$. We see from this figure that as N grows, the distributions approach $\mathcal{N}(1, 1)$. For N greater than or equal to 100 it is difficult to distinguish the plots from $\mathcal{N}(1, 1)$.

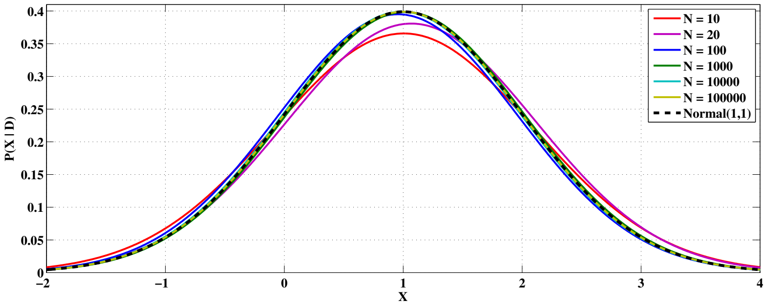


Figure 18: Probability distribution $p(x|\mathcal{D})$

Figures **??** and **??** show plots of our estimate of the mean μ_N and our uncertainty of this estimate σ_N^2. As expected, we see the value of our estimated mean converging to the value of our true mean. Additionally, we see our uncertainty of this estimates tends to 0 as N increases. After 1000 data points, we do not see much difference in our estimated value and our uncertainty is approximately 0.

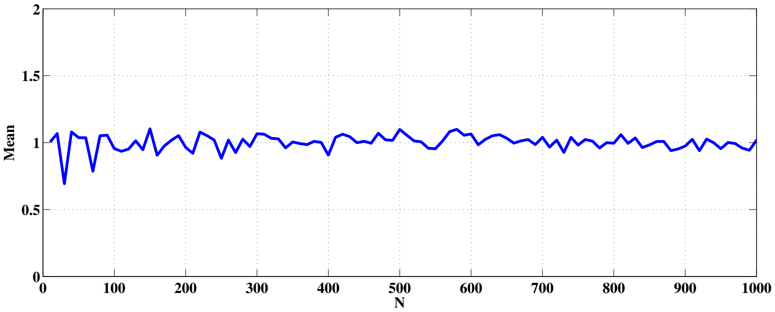


Figure 19: Estimate of mean μ_N with increasing N

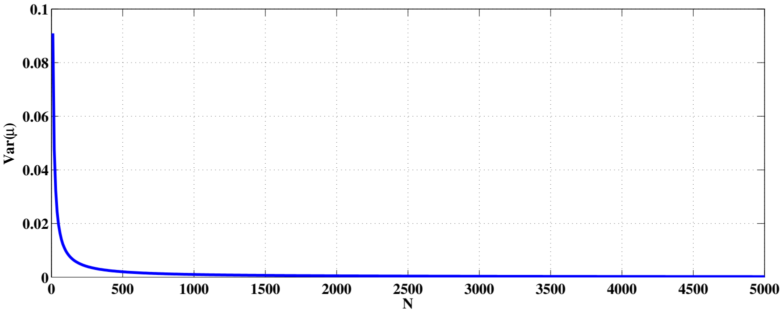


Figure 20: Uncertainty of estimate σ_N^2 with increasing N