Question 1:
In Figure 1, by increasing the model order, the root mean square of both training and validation set is decreasing until model order (M) reaches 5. However, with larger value of M the root mean square of the training is decreasing while the root mean square error of validation dataset is decreasing. This fact is an indicator that from M=5 the model is starting to memorize the model which yields to overtraining. Although M=6 yields the same amount root mean square on both tests, the best model order would be M=5. I think in any polynomial fitting regression problem we need to balance the model analytical ability with its complexity to avoid complex models which might cause overtraining in our question. I tested the different number of training and validation size and different variances for noise distribution. The small values of variation for noise distribution made the two plots be almost identical and stacked on top of each other. Therefore, I decided to keep the noise variance around 0.2 to make sure we can see each plot separately, so we can make better interpretations on which model order is the best. According to Bishop [1], the model order smaller than 5 gives us high testing RMS which could be related to the fact that Sinc(x) is not a very simple or flexible function.
Another experiment I have done was that when I changed the number of training and/or testing size, I needed to change the range of model order to see any pick on the plots. This fact was also approved by Bishop textbook which is mentioned "the larger the data set, the more complex (in other words more flexible) the model that we can afford to fit to the data." Therefore, I choose training size between 50 to 70 to make sure the function does not become too complex and at the same time the overtraining is not happening.

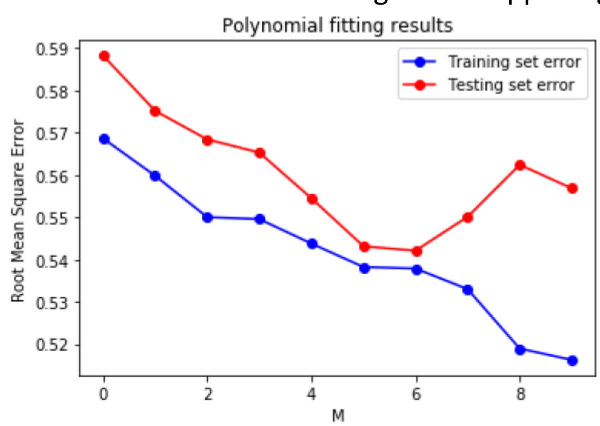

Figure 1:Root Mean Square of training and validation set

Question 2:
For small value of N, there is a significant difference between posterior mean and maximum likelihood mean. However, as N goes large enough, the mean of posterior converges to the mean of maximum likelihood. On the other hand, the variance of posterior distribution goes to zero and we have a peak on the value of maximum likelihood. We plotted the prior and posterior normal distribution in the Figure 2. In this problem the posterior of the previous draw would be the prior of the next draw which results in the peak of the prior. I experimented different number of draws to investigate what is its effect on the prior and true distribution in this problem. As you could see in Figure 2, once number of samples are increasing the prior mu is decreasing and ultimately the maximum posterior would be equal with the maximum likelihood.

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML}$$

Equation 1: Mean of Gaussian Distribution

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma^2} + \frac{N}{\sigma^2}$$

Equation 2: Variance of Multivariate Gaussian

1. What happens when the prior mean is initialized to the wrong value? To the correct value? According to the Equation 1, if we enter the prior mean equal to the true mean then the posterior distribution equals with the maximum likelihood. Therefore, we do not include our bias which may affect the data generation process. On the other hand, if we assign the wrong value to the prior mean, then there might be a huge difference between the maximum likelihood mean and maximum posterior mean. When we collect data, we expect that the sample mean is close to the underlying true distribution. However, this is not the case when we pick a wrong mean value.
2. What happens as you vary the prior variance from small to large? According to equation 2, if the prior variance is small the posterior variance would be equal with the prior Nvariance. However, for large values of prior variance the posterior variance would be equal to $\sigma 2$ /N. The values of prior variance would affect the posterior mean as well. According to equation 4, for small value of prior variance the posterior mean leans more toward prior mean as the coefficient of maximum likelihood would become small.
   On the other hand, for the large values of prior variance the posterior mean would become equal with maximum likelihood which means that our bias is not affecting posterior mean.
3. What happens when the likelihood variance is varied from small to large?
   If the likelihood variance grows from small to large, the posterior variance would be equal with prior variance. This means that our bias is affecting the data significantly.
4. How do the initial values of the prior mean, prior variance, and likelihood variance interact to affect the final estimate of the mean? For small values of N, the initial values of the prior mean, prior variance and likelihood variance have significant effect on the final estimate of the mean. However, as N goes large the posterior mean would converge to the maximum likelihood mean or in another MAP result would be equal to ML results.

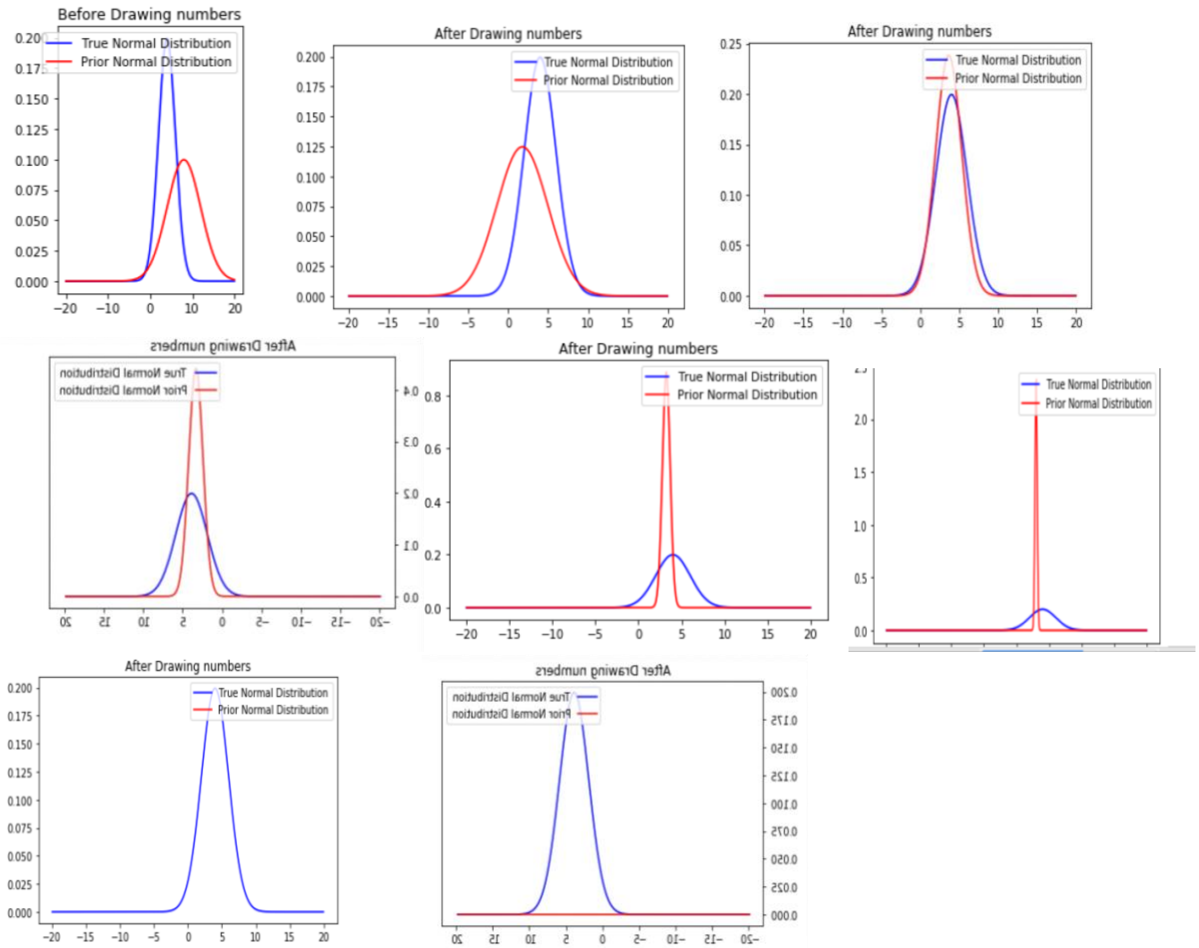

Figure 2-1 to 2-8 illustrate that how priors are changing through running the code and the maximum posterior would lean toward the maximum likelihood and the bias is eliminated