Problem 1 (Bayesian Methods)

This question helps to build your understanding of making predictions with a maximum-likelihood estimation (MLE), a maximum a posterior estimator (MAP), and a full posterior predictive.

Consider a scalar variable x with the following generative process: First, the mean μ is sampled from a prior $N(0, \tau^2)$. Next, each x_n is generated as $x_n = \mu + \epsilon_n$, where $\epsilon_n \sim N(0, \sigma^2)$. All ϵ_n's are independent of each other and of μ.

For this problem, use $\sigma^2 = 1$ and $\tau^2 = 5$.

Now, we see 14 independent samples of x to yield data

$$D = 3.3, 3.5, 3.1, 1.8, 3.0, 0.74, 2.5, 2.4, 1.6, 2.1, 2.4, 1.3, 1.7, 0.19$$

Make sure to include all required plots in your PDF.

1. Derive the expression for $p(\mu|D)$. Do *not* plug in numbers yet!

   Hint: Use properties of normal-normal conjugacy to simplify your derivation. You can also refer to this paper.

2. Now we get to our core interest: the predictive distribution of a new datapoint x^* given our observed data D, $p(x^*|D)$. Write down the expression for the full posterior predictive distribution:

   $$p(x^*|D) = \int p(x^*|\mu)p(\mu|D)d\mu$$

   Interpret your expression in a few words. Do *not* plug in numbers yet!

   Hint: To simplify your derivation, use the fact that $x = \mu + \epsilon$, and $\mu|D$ and ϵ are independent Gaussians whose distributions you know from above.

3. The full posterior predictive distribution had a nice analytic form in this case, but in many problems, it is often difficult to calculate because we need to marginalize out the parameters (here, the parameter is μ). We can mitigate this problem by plugging in a point estimate of μ^* rather than a distribution. Derive the estimates of $p(x^*|D) \approx p(x^*|\mu^*)$ for $\mu^* = \mu_{MLE}$ and $\mu^* = \mu_{MAP}$. How do these expressions compare to the expression for the full posterior above? Do *not* plug in numbers yet!

4. Plot how the above 3 distributions change after each data point is gathered. You will have a total of 15 plots, starting with the plot for the case with no data (they can be small, e.g. in a 3×5 grid). The x-axis of each plot will be the x value and the y-axis the density. You can make one plot for each estimator, or combine all three estimators onto one plot with a different colored line for each estimator.

5. How do the means of the predictive distributions vary with more data? How do the variances vary? Interpret the differences you see between the three different estimators.

6. Does the ordering of the data matter for the final predictive distributions?

7. Derive an expression for and then compute the marginal likelihood of the training data $p(D)$.

   Hint: You can rearrange the required integral such that it looks like an un-normalized Gaussian distribution in terms of μ, and take advantage of the fact that integrating over a normalized Gaussian distribution is equal to 1. You will need to complete the square.

8. Now consider an alternate model in which we were much more sure about the mean: $\mu \sim N(0, \tau^2)$, where $\tau^2 = 0.1$. Compute the marginal likelihood $p(D)$ for this model. Which of the two models has a higher marginal likelihood? Interpret this result.

Solution:

1. Let n be the number of observations in our data. Via Bayes rule, we have that the posterior distribution of μ is proportional to the probability of the data given μ, times the prior distribution of μ:

$$p(\mu|D) \propto p(D|\mu,\sigma)p(\mu|\tau)$$

We have that the probability of the data given μ as $p(D|\mu) \sim N(\mu,\sigma^2)$, so we may write the likelihood as:

$$p(D|\mu,\sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{D_i-\mu}{\sigma})^2}$$

$$p(D|\mu,\sigma) = (\sigma\sqrt{2\pi})^{-n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(D_i-\mu)^2}$$

Let $\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$ be the empirical mean, such that we may write

$$\sum_{i=1}^{n}(D_i-\mu)^2 = \sum_{i=1}^{n}[(D_i-\bar{D})-(\mu-\bar{D})]^2$$

$$\sum_{i=1}^{n}(D_i-\mu)^2 = \sum_{i=1}^{n}(D_i-\bar{D})^2 + \sum_{i=1}^{n}(\bar{D}-\mu)^2 - 2\sum_{i=1}^{n}(D_i-\bar{D})(\mu-\bar{D})$$

Notice that

$$\sum_{i=1}^{n}(D_i-\bar{D})(\mu-\bar{D}) = (\mu-\bar{D})\big((\sum_{i=1}^{n} D_i) - n\bar{D}\big) = (\mu-\bar{D})(n\bar{D}-n\bar{D}) = 0$$

Therefore we have that

$$\sum_{i=1}^{n}(D_i-\mu)^2 = \sum_{i=1}^{n}(D_i-\bar{D})^2 + \sum_{i=1}^{n}(\bar{D}-\mu)^2$$

Let $s^2 = \frac{1}{n}\sum_{i=1}^{n}(D_i-\bar{D})$ be the empirical variance, such that we may write

$$\sum_{i=1}^{n}(D_i-\mu)^2 = ns^2 + n(\bar{D}-\mu)^2$$

Therefore we may substitute back into our likelihood:

$$p(D|\mu,\sigma) = (\sigma\sqrt{2\pi})^{-n} e^{-\frac{1}{2\sigma^2}\left(ns^2+n(\bar{D}-\mu)^2\right)}$$

$$p(D|\mu,\sigma) \propto \sigma^{-n}\exp\Big(-\frac{n}{2\sigma^2}(\bar{D}-\mu)^2\Big)\exp\Big(-\frac{ns^2}{2\sigma^2}\Big)$$

From the problem statement, we may assume that σ is constant, therefore we may rewrite the likelihood as:

$$p(D|\mu,\sigma) \propto \exp\Big(-\frac{n}{2\sigma^2}(\bar{D}-\mu)^2\Big) \propto N(\bar{D}|\mu,\frac{\sigma^2}{n})$$

We are given the prior distribution of μ as $p(\mu|\tau) \sim N(0,\tau^2)$, so we can write its probability as:

$$p(\mu|\tau) = \frac{1}{\tau\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\mu}{\tau})^2}$$

$$p(\mu|\tau) \propto \tau^{-1}\exp\Big(-\frac{\mu^2}{2\tau^2}\Big)$$

Once again, from the problem statement, we may assume that τ is constant, therefore we may rewrite the prior probability as

$$p(\mu|\tau) \propto \exp\Big(-\frac{\mu^2}{2\tau^2}\Big) \propto N(\mu|0,\tau^2)$$

Therefore we can combine our results to write the posterior distribution of μ be:

$$p(\mu|D) \propto p(D|\mu,\sigma)p(\mu|\tau)$$

$$p(\mu|D) \propto \exp\Big(-\frac{n}{2\sigma^2}(\bar{D}-\mu)^2\Big)\exp\Big(-\frac{\mu^2}{2\tau^2}\Big)$$

$$p(\mu|D) \propto \exp\Big(-\frac{n}{2\sigma^2}(\bar{D}^2-2\mu\bar{D}+\mu^2)-\frac{\mu^2}{2\tau^2}\Big)$$

$$p(\mu|D) \propto \exp\Big(\mu^2(-\frac{n}{2\sigma^2}-\frac{1}{2\tau^2})+\mu(\frac{2n\bar{D}}{2\sigma^2})-\frac{n\bar{D}^2}{2\sigma^2}\Big)$$

$$p(\mu|D) \propto \exp\Big(-\frac{\mu^2}{2}(\frac{n}{\sigma^2}+\frac{1}{\tau^2})+\mu(\frac{n\bar{D}}{\sigma^2})-\frac{n\bar{D}^2}{2\sigma^2}\Big)$$

By normal-normal conjugacy, we know that the product of two Gaussians is also Gaussian. Therefore, define some a and b such that the following equality holds:

$$\exp\Big(-\frac{1}{2a}(\mu-b)^2\Big) = \exp\Big(-\frac{\mu^2}{2}(\frac{n}{\sigma^2}+\frac{1}{\tau^2})+\mu(\frac{n\bar{D}}{\sigma^2})-\frac{n\bar{D}^2}{2\sigma^2}\Big)$$

$$\exp\Big(-\frac{1}{2a}(\mu^2-2b\mu+b^2)\Big) = \exp\Big(-\frac{\mu^2}{2}(\frac{n}{\sigma^2}+\frac{1}{\tau^2})+\mu(\frac{n\bar{D}}{\sigma^2})-\frac{n\bar{D}^2}{2\sigma^2}\Big)$$

To find our desired values for a and b to satisfy the above equality, we can "complete the square" by matching the coefficients for μ^2, μ, and the constant term.
We can solve for a by matching the coefficients of μ^2:

$$-\frac{1}{2a}\mu^2 = -\frac{\mu^2}{2}(\frac{n}{\sigma^2}+\frac{1}{\tau^2})$$

$$\frac{1}{a} = \frac{n}{\sigma^2}+\frac{1}{\tau^2} = \frac{n\tau^2+\sigma^2}{\sigma^2\tau^2}$$

$$a = \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}} = \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}$$

Next, we can solve for b by matching the coefficients of μ:

$$-\frac{-2b\mu}{2a} = \mu\frac{n\bar{D}}{\sigma^2}$$

$$\frac{b}{a} = \frac{n\bar{D}}{\sigma^2}$$

$$b = a\frac{n\bar{D}}{\sigma^2} = (\frac{\sigma^2\tau^2}{n\tau^2+\sigma^2})(\frac{n\bar{D}}{\sigma^2}) = \frac{n\bar{D}\tau^2}{n\tau^2+\sigma^2}$$

Therefore, returning to our equality we have that

$$p(\mu|D) \propto \exp\Big(-\frac{1}{2a}(\mu-b)^2\Big) = \exp\Big(-\frac{\mu^2}{2}(\frac{n}{\sigma^2}+\frac{1}{\tau^2})+\mu(\frac{n\bar{D}}{\sigma^2})-\frac{n\bar{D}^2}{2\sigma^2}\Big)$$

Substituting for our values of a and b gives:

$$\boxed{p(\mu|D) \propto \exp\Big(-\frac{n\tau^2+\sigma^2}{2\sigma^2\tau^2}(\mu-\frac{n\bar{D}\tau^2}{n\tau^2+\sigma^2})^2\Big)}$$

or equivalently

$$\boxed{p(\mu|D) \propto N\Big(\mu \,\Big|\, \frac{n\bar{D}\tau^2}{n\tau^2+\sigma^2}, \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}\Big)}$$

2. The posterior predictive for a new datapoint x^* given our observed data D is:

$$p(x^*|D) = \int p(x^*|\mu)p(\mu|D)d\mu$$

Recall that we had derived the likelihood in part 1 to be:

$$p(\mu|D) \propto \exp\Big(-\frac{n\tau^2+\sigma^2}{2\sigma^2\tau^2}(\mu - \frac{n\bar{D}\tau^2}{n\tau^2+\sigma^2})^2\Big)$$

From the problem statement we had defined $x_n = \mu + \epsilon_n$. Therefore

$$p(x^*|D) = p(\mu + \epsilon_n|D)$$

We are given that ϵ_n and μ are independent and both normally distributed, so we know that

$$p(x^*|D) = p(\mu|D) + p(\epsilon_n|D)$$

In our generative model, our noise is independent of our data since it is meant to be unobservable noise. Therefore $p(\epsilon_n|D) = p(\epsilon_n)$. We are given that $\epsilon_n \sim N(0, \sigma^2)$ such that

$$p(\epsilon_n) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{1}{2}(\frac{\epsilon_n}{\sigma})^2\Big)$$

To compute $p(x^*|D) = p(\mu|D) + p(\epsilon_n)$ we can employ the properties of Gaussian variables, since both $p(\mu|D)$ and $p(\epsilon_n)$ independent and normally distributed. The sum of two independent random variables (let them be X and Y) that are normally distributed is also normally distributed, with it's mean being the sum of the two means, and its variance being the sum of the two variances. That is

$$\text{let } X \sim N(\mu_X, \sigma_X^2) \text{ and } Y \sim N(\mu_Y, \sigma_Y^2). \text{ If } X \perp\!\!\!\perp Y, \text{ then } X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Applying this to the sum of $\mu|D$ and ϵ_n, which are also independent and normally distributed, we find:

$$\boxed{p(x^*|D) \sim N\Big(\frac{n\bar{D}\tau^2}{n\tau^2+\sigma^2}, \sigma^2 + \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}\Big)}$$

or equivalently

$$\boxed{p(x^*|D) = \frac{1}{\sqrt{2\pi\big(\sigma^2 + \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}\big)}} \exp\left(-\frac{1}{2\big(\sigma^2 + \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}\big)}(x^* - \frac{n\bar{D}\tau^2}{n\tau^2+\sigma^2})^2\right)}$$

- The variance of the posterior predictive $p(x^*|D)$ is independent of the data D, implying that the variance of $p(x^*|D)$ is strictly controlled by the variance of the mean parameter μ and the variance of the noise ϵ_n in our generative process. Notice the n that appears in the denominator of the variance for $p(x^*|D)$, this makes intuitive sense since more data points (larger n) will reduce the variance of our estimation for a new datapoint x^*.

- τ^2 is our uncertainty on the mean parameter μ. Larger τ^2 will increase our variance, which implies that our data is more important for larger τ^2, since more data yields a larger n and therefore a lower variance on the posterior predictive. This makes intuitive sense, since if we are more uncertain about the prior distribution of the mean parameter μ, then having more data will allow us to have a more accurate estimation of a new datapoint x^*. The converse is also true where our data is less important for smaller τ^2 since that implies that we are more certain about our prior estimation of μ.

- σ^2 has a similar interpretation, where a larger σ^2 will increase our variance and mean we place greater importance on having more data. This again makes intuitive sense, since greater noise in our generative process will increase the variance of the data D that we condition on in the posterior predictive $p(x^*|D)$. Once again, the converse is also true where a smaller σ^2 implies that it is less important to have lots of data (or larger n).

3. We had that the noise $\epsilon_n \sim N(0, \sigma^2)$ is normally distributed and that $x_n = \mu + \epsilon_n$, therefore

$$x^*|\mu^* \sim N(\mu^*, \sigma^2) \qquad p(x^*|\mu^*) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{1}{2}(\frac{x^* - \mu^*}{\sigma})^2\Big)$$

We are finding point estimates $\mu^* = \mu_{MLE}$ and $\mu^* = \mu_{MAP}$ to substitute into the above. Begin by finding μ_{MLE} which is μ that maximizes the likelihood of our data $p(D|\mu)$ (letting n again be the number of observations in our data):

$$p(D|\mu) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{1}{2}(\frac{D_i - \mu}{\sigma})^2\Big)$$

since each observation $D_i \sim N(\mu, \sigma^2)$. Converting the likelihood into the log-likelihood and dropping constants (since we assume σ is constant):

$$\log(p(D|\mu)) = -\frac{1}{2}\sum_{i=1}^{n}(\frac{D_i - \mu}{\sigma})^2$$

Solving the First Order Condition of our log-likelihood gives us the maximum likelihood estimation μ_{MLE}:

$$\frac{\partial \log(p(D|\mu))}{\partial \mu} = \sum_{i=1}^{n} \frac{D_i - \mu_{MLE}}{\sigma^2} = 0$$

$$\sum_{i=1}^{n}(D_i - \mu_{MLE}) = \sum_{i=1}^{n} D_i - \sum_{i=1}^{n} \mu_{MLE} = -n\mu_{MLE} + \sum_{i=1}^{n} D_i = 0$$

$$\mu_{MLE} = \frac{1}{n}\sum_{i=1}^{n} D_i$$

So our estimate for $\mu^* = \mu_{MLE}$ is

$$\boxed{p(x^*|D) \approx p(x^*|\mu_{MLE}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{1}{2}(\frac{x^* - \frac{1}{n}\sum_{i=1}^{n} D_i}{\sigma})^2\Big)}$$

or equivalently

$$\boxed{p(x^*|\mu_{MLE}) \sim N\Big(\frac{1}{n}\sum_{i=1}^{n} D_i, \sigma^2\Big)}$$

μ_{MAP} maximizes the posterior distribution $p(\mu|D)$. Using Bayes' Rule we can write the posterior as

$$p(\mu|D) \propto p(D|\mu)p(\mu)$$

Maximizing the logged posterior $\log(p(\mu|D))$ is equivalent to maximizing the posterior $p(\mu|D)$, because the logarithm is a monotonic function. Therefore:

$$\log(p(\mu|D)) \propto \log(p(D|\mu)) + \log(p(\mu))$$

$\mu^* = \mu_{MAP}$ is the value of μ which maximizes $\log(p(\mu|D))$, that is the value of μ which solves the first order condition of $\log(p(\mu|D))$:

$$\frac{\partial \log(p(\mu_{MAP}|D))}{\partial \mu_{MAP}} \propto \frac{\partial \log(p(D|\mu_{MAP}))}{\partial \mu_{MAP}} + \frac{\partial \log(p(\mu_{MAP}))}{\partial \mu_{MAP}} = 0$$

Recall that earlier we had found that

$$\log(p(D|\mu)) = -\frac{1}{2}\sum_{i=1}^{n}\big(\frac{D_i - \mu}{\sigma}\big)^2 \quad \text{and} \quad \frac{\partial \log(p(D|\mu))}{\partial \mu} = \sum_{i=1}^{n}\frac{D_i - \mu}{\sigma^2}$$

We are given the prior distribution $p(\mu) \sim N(0, \tau^2)$, so therefore

$$p(\mu) = \frac{1}{\tau\sqrt{2\pi}} \exp\Big(-\frac{1}{2}\big(\frac{\mu}{\tau}\big)^2\Big)$$

Converting into the logged prior distribution and dropping constants (since we assume τ is constant) and deriving with respect to μ:

$$\log(p(\mu)) = -\frac{1}{2}\big(\frac{\mu}{\tau}\big)^2 \quad \text{and} \quad \frac{\partial \log(p(\mu))}{\partial \mu} = -\frac{\mu}{\tau^2}$$

Therefore, we have

$$\frac{\partial \log(p(\mu_{MAP}|D))}{\partial \mu_{MAP}} \propto \sum_{i=1}^{n}\frac{D_i - \mu_{MAP}}{\sigma^2} - \frac{\mu_{MAP}}{\tau^2} = 0$$

$$\sum_{i=1}^{n}\frac{D_i - \mu_{MAP}}{\sigma^2} = \frac{\mu_{MAP}}{\tau^2}$$

$$\sum_{i=1}^{n}(D_i - \mu_{MAP}) = \frac{\mu_{MAP}\sigma^2}{\tau^2}$$

$$-n\mu_{MAP} + \sum_{i=1}^{n} D_i = \frac{\mu_{MAP}\sigma^2}{\tau^2}$$

$$\sum_{i=1}^{n} D_i = \frac{\mu_{MAP}\sigma^2}{\tau^2} + n\mu_{MAP} = \mu_{MAP}\big(\frac{\sigma^2}{\tau^2} + n\big) = \mu_{MAP}\big(\frac{\sigma^2 + n\tau^2}{\tau^2}\big)$$

$$\mu_{MAP} = \frac{\tau^2}{\sigma^2 + n\tau^2}\sum_{i=1}^{n} D_i$$

So our estimate for $\mu^* = \mu_{MAP}$ is

$$\boxed{p(x^*|D) \approx p(x^*|\mu_{MAP}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{1}{2}\big(\frac{x^* - \frac{\tau^2}{\sigma^2+n\tau^2}\sum_{i=1}^{n} D_i}{\sigma}\big)^2\Big)}$$

or equivalently

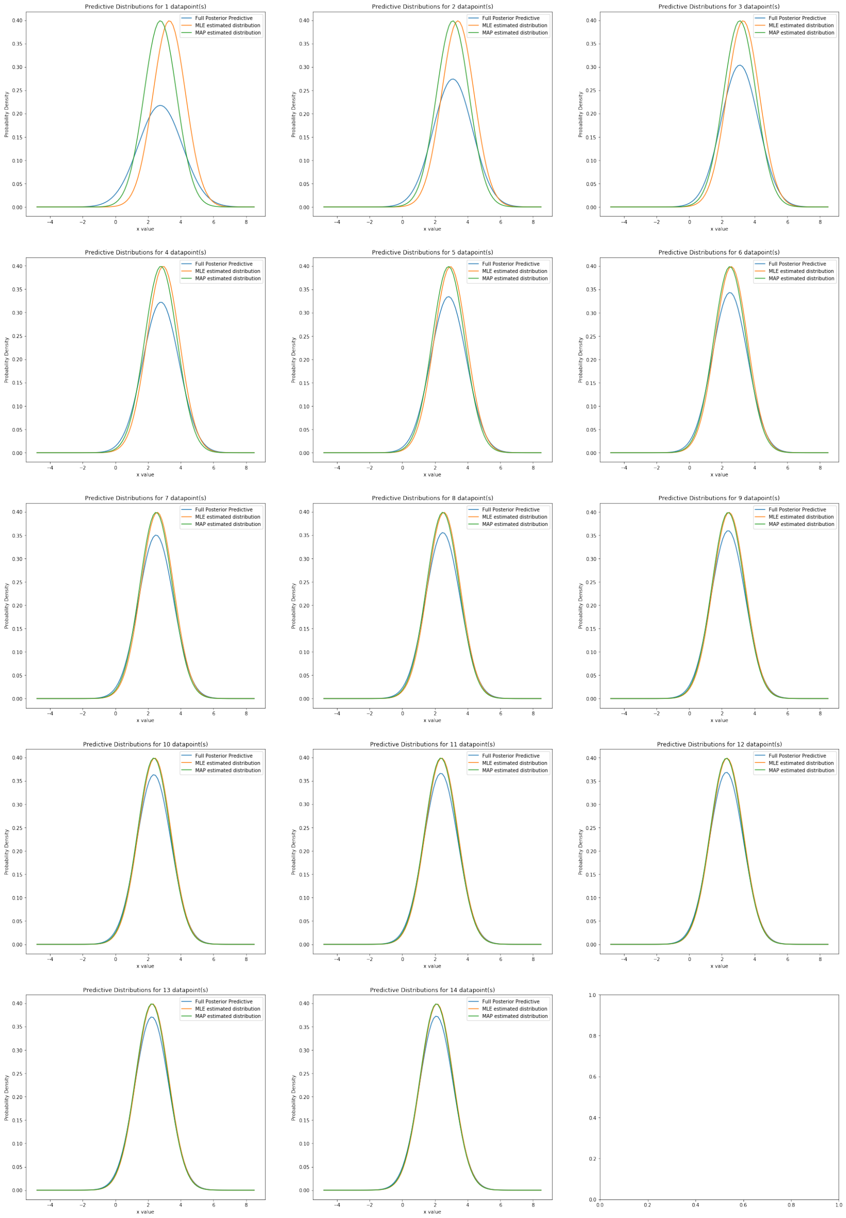
$$\boxed{p(x^*|\mu_{MAP}) \sim N\big(\frac{\tau^2}{\sigma^2 + n\tau^2}\sum_{i=1}^{n} D_i, \sigma^2\big)}$$

Recall the posterior predictive distribution was

$$p(x^*|D) \sim N\Big(\frac{n\bar{D}\tau^2}{n\tau^2 + \sigma^2}, \sigma^2 + \frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}\Big)$$

Unlike the variance for the posterior predictive, the variance for both of the estimated distributions is σ^2, which is the same variance as the noise ϵ_n in our generative process. This makes intuitive sense, since all we are doing is plugging in an estimation of the mean μ^*, so in our estimated distributions only the means should change, but not the variances. Therefore $p(x|\mu)$ has a mean of σ^2. Moreover, notice that the estimate for μ_{MLE} is simply the empirical average (sample mean) of our data D, just like in HW 1! The estimate for μ_{MAP} is simply the mode of the posterior distribution $p(\mu|D)$, but since μ is distributed normally, the mean and the mode of the posterior distribution are equivalent and therefore μ_{MAP} is also the mean of the posterior distribution. Moreover, the means for $p(x^*|D)$ and $p(x^*|\mu_{MAP})$ are identical, which makes intuitive sense because μ_{MAP} is the mean of the posterior distribution, and because both the posterior predictive and μ_{MAP} estimated distributions set a prior on μ.

4. The 14 plots for 3 distributions are shown below:



5. The means for the posterior distribution and the μ_{MAP} estimated distributions are equal by definition, so their means are the same for all 14 plots. Moreover, their means when there are few data points are near zero because both set a prior on μ and the prior had a mean of 0. In contrast, the MLE makes no assumptions on the mean parameter μ and is simply the empirical average of the datapoints. Notice that eventually, when we have more datapoints, the means of all three distributions converge to approximately 2. Moreover, observe that for many datapoints the distributions for the μ_{MLE} estimated distribution and the μ_{MAP} distribution appear to be almost identical, which makes sense since the both μ_{MLE} and μ_{MAP} estimate the means (since the mode of a normal distribution is also its mean), and since both methods employ the use of a point estimate. As we get more data, it matters less whether we use μ_{MLE} or μ_{MAP} as our point estimate, since for many datapoints their distributions will be approximately the same.

   For the variance, the μ_{MLE} estimated distribution and the μ_{MAP} distribution have the same variance by definition, so their distributions have the same shape but are situated at different means for all of the 14 plots (except for the plots of 5 or more datapoints where their means are approxiamtely the same). For these two distributions, the shape of their distribution stays the same across all 14 plots since their variance is σ^2 which is independent of our data. In contrast, the variance of the posterior predictive decreases for larger datasets, as evidenced by the posterior predictive distribution becoming narrower and taller for more datapoints. This makes intuitive sense, since the variance of the posterior predictive distribution was $\sigma^2 + \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}$ which decreases for larger datasets (which have larger n). Eventually, the posterior predictive begins to approach the same shape as the point estimated distributions, implying that if we had a sufficiently large dataset then all three distributions would look identical. This is implied by the fact that

   $$\lim_{n\to\infty} \sigma^2 + \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2} = \sigma^2$$

   showing that for an infinitely large dataset the posterior predictive distribution actually has the same analytical form for its variance as the point estimated distributions.

6. Given the same dataset, the ordering of the datapoints do not matter for any of the three final predictive distributions. This is evidenced by the fact that the the data only appears as sums or arithmetic averages in the expressions for the mean and variances of the three predictive distributions. Because we are only concerned with the sum and average of our datapoints, the actual order of the datapoints don't matter since the ordering will not change the calculation of the sum or average.

7. Using the Law of Total Probability, we can write

   $$p(D) = \int p(D|\mu)p(\mu)d\mu$$

   Recall that we had found that

   $$p(D|\mu) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{1}{2}(\frac{D_i-\mu}{\sigma})^2\Big) \quad \text{and} \quad p(\mu) = \frac{1}{\tau\sqrt{2\pi}} \exp\Big(-\frac{1}{2}(\frac{\mu}{\tau})^2\Big)$$

   Plugging these into our expression for $p(D)$ gives:

   $$p(D) = \int \Bigg(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{1}{2}(\frac{D_i-\mu}{\sigma})^2\Big)\Bigg)\Bigg(\frac{1}{\tau\sqrt{2\pi}} \exp\Big(-\frac{1}{2}(\frac{\mu}{\tau})^2\Big)\Bigg)d\mu$$

   $$p(D) = (\frac{1}{\sigma\sqrt{2\pi}})^n(\frac{1}{\tau\sqrt{2\pi}}) \int \Bigg(\prod_{i=1}^{n} \exp\Big(-\frac{1}{2}(\frac{D_i-\mu}{\sigma})^2\Big)\Bigg) \exp\Big(-\frac{1}{2}(\frac{\mu}{\tau})^2\Big)d\mu$$

   $$p(D) = (\frac{1}{\sigma\sqrt{2\pi}})^n(\frac{1}{\tau\sqrt{2\pi}}) \int \exp\Big(-\frac{1}{2}\sum_{i=1}^{n}(\frac{D_i-\mu}{\sigma})^2\Big) \exp\Big(-\frac{1}{2}(\frac{\mu}{\tau})^2\Big)d\mu$$

$$p(D) = (\frac{1}{\sigma\sqrt{2\pi}})^n(\frac{1}{\tau\sqrt{2\pi}})\int \exp\Big(-\frac{1}{2}(\frac{\mu}{\tau})^2 - \frac{1}{2}\sum_{i=1}^{n}(\frac{D_i-\mu}{\sigma})^2\Big)d\mu$$

$$p(D) = (\frac{1}{\sigma\sqrt{2\pi}})^n(\frac{1}{\tau\sqrt{2\pi}})\int \exp\Big(-\frac{1}{2}\frac{\mu^2}{\tau^2} - \frac{1}{2}\sum_{i=1}^{n}\frac{D_i^2-2\mu D_i+\mu^2}{\sigma^2}\Big)d\mu$$

$$p(D) = (\frac{1}{\sigma\sqrt{2\pi}})^n(\frac{1}{\tau\sqrt{2\pi}})\int \exp\Big(-\frac{1}{2}(\frac{\mu^2}{\tau^2} + \frac{1}{\sigma^2}\sum_{i=1}^{n}(D_i^2-2\mu D_i+\mu^2))\Big)d\mu$$

Let $\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$ (the empirical average defined previously) and let $\bar{D}^2 = \frac{1}{n}\sum_{i=1}^{n} D_i^2$. Therefore we may rewrite our expression for $p(D)$ as:

$$p(D) = (\frac{1}{\sigma\sqrt{2\pi}})^n(\frac{1}{\tau\sqrt{2\pi}})\int \exp\Big(-\frac{1}{2}(\frac{\mu^2}{\tau^2} + \frac{1}{\sigma^2}(n\bar{D}^2 - 2\mu n\bar{D} + n\mu^2))\Big)d\mu$$

Combining the terms for μ^2, μ, and the constant term gives:

$$p(D) = (\frac{1}{\sigma\sqrt{2\pi}})^n(\frac{1}{\tau\sqrt{2\pi}})\int \exp\Big(-\frac{1}{2}(\mu^2(\frac{1}{\tau^2} + \frac{n}{\sigma^2}) - \frac{2\mu n\bar{D}}{\sigma^2} + \frac{n\bar{D}^2}{\sigma^2})\Big)d\mu$$

$$p(D) = (\frac{1}{\sigma\sqrt{2\pi}})^n(\frac{1}{\tau\sqrt{2\pi}})\int \exp\Big(-\frac{1}{2}(\mu^2(\frac{\sigma^2+n\tau^2}{\sigma^2\tau^2}) - \frac{2\mu n\bar{D}}{\sigma^2} + \frac{n\bar{D}^2}{\sigma^2})\Big)d\mu$$

$$p(D) = (\frac{1}{\sigma\sqrt{2\pi}})^n(\frac{1}{\tau\sqrt{2\pi}})\int \exp\Big(-\frac{1}{2\sigma^2}(\mu^2(\frac{\sigma^2+n\tau^2}{\tau^2}) - 2\mu n\bar{D} + n\bar{D}^2)\Big)d\mu$$

Observe that this appears as an un-normalized Gaussian distribution as described in the hint. To rewrite the integral as a normalized Gaussian distribution, we employ completing the square as suggested in the hint. Using the resource linked in the EdStem Addendums for HW 3 we are given that

$$ax^2 + bx + c = a(x+d)^2 + e$$

such that

$$d = \frac{b}{2a} \quad \text{and} \quad e = c - \frac{b^2}{4a}$$

Consider the polynomial in the exponential:

$$\mu^2(\frac{\sigma^2+n\tau^2}{\tau^2}) - 2\mu n\bar{D} + n\bar{D}^2$$

Employing our given formula for completing the square, we have that:

$$a = \frac{\sigma^2+n\tau^2}{\tau^2} \quad b = -2n\bar{D} \quad \text{and} \quad c = n\bar{D}^2 \quad \text{such that} \quad d = -\frac{n\bar{D}\tau^2}{\sigma^2+n\tau^2} \quad \text{and} \quad e = n\bar{D}^2 - \frac{n^2(\bar{D})^2\tau^2}{\sigma^2+n\tau^2}$$

Therefore we have that:

$$\mu^2(\frac{\sigma^2+n\tau^2}{\tau^2}) - 2\mu n\bar{D} + n\bar{D}^2 = \frac{\sigma^2+n\tau^2}{\tau^2}(\mu - \frac{n\bar{D}\tau^2}{\sigma^2+n\tau^2})^2 + n\bar{D}^2 - \frac{n^2(\bar{D})^2\tau^2}{\sigma^2+n\tau^2}$$

Substituting into the marginal distribution of D gives:

$$p(D) = (\frac{1}{\sigma\sqrt{2\pi}})^n(\frac{1}{\tau\sqrt{2\pi}})\int \exp\Big(-\frac{1}{2\sigma^2}(\frac{\sigma^2+n\tau^2}{\tau^2}(\mu - \frac{n\bar{D}\tau^2}{\sigma^2+n\tau^2})^2 + n\bar{D}^2 - \frac{n^2(\bar{D})^2\tau^2}{\sigma^2+n\tau^2})\Big)d\mu$$

$$p(D) = (\frac{1}{\sigma\sqrt{2\pi}})^n(\frac{1}{\tau\sqrt{2\pi}})\int \exp\Big(-\frac{\sigma^2+n\tau^2}{2\sigma^2\tau^2}(\mu - \frac{n\bar{D}\tau^2}{\sigma^2+n\tau^2})^2 - \frac{1}{2\sigma^2}(n\bar{D}^2 - \frac{n^2(\bar{D})^2\tau^2}{\sigma^2+n\tau^2})\Big)d\mu$$

$$p(D) = (\frac{1}{\sigma\sqrt{2\pi}})^n (\frac{1}{\tau\sqrt{2\pi}}) \exp\Big(-\frac{1}{2\sigma^2}(n\bar{D}^2 - \frac{n^2(\bar{D})^2\tau^2}{\sigma^2+n\tau^2})\Big) \int \exp\Big(-\frac{\sigma^2+n\tau^2}{2\sigma^2\tau^2}(\mu - \frac{n\bar{D}\tau^2}{\sigma^2+n\tau^2})^2\Big) d\mu$$

Consider the term:

$$\exp\Big(-\frac{\sigma^2+n\tau^2}{2\sigma^2\tau^2}(\mu - \frac{n\bar{D}\tau^2}{\sigma^2+n\tau^2})^2\Big)$$

which appears in the form of an un-normalized Gaussian with mean $\frac{n\bar{D}\tau^2}{\sigma^2+n\tau^2}$ and variance $\frac{\sigma^2\tau^2}{\sigma^2+n\tau^2}$. Therefore, to normalize this term, we can multiply and divide our expression for $p(D)$ by $\frac{1}{\sqrt{\frac{2\pi\sigma^2\tau^2}{\sigma^2+n\tau^2}}}$:

$$p(D) = (\frac{1}{\sigma\sqrt{2\pi}})^n (\frac{1}{\tau\sqrt{2\pi}}) \exp\Big(-\frac{1}{2\sigma^2}(n\bar{D}^2 - \frac{n^2(\bar{D})^2\tau^2}{\sigma^2+n\tau^2})\Big)(\sqrt{\frac{2\pi\sigma^2\tau^2}{\sigma^2+n\tau^2}})$$

$$\times \int \frac{1}{\sqrt{\frac{2\pi\sigma^2\tau^2}{\sigma^2+n\tau^2}}} \exp\Big(-\frac{\sigma^2+n\tau^2}{2\sigma^2\tau^2}(\mu - \frac{n\bar{D}\tau^2}{\sigma^2+n\tau^2})^2\Big) d\mu$$

Observe that $\frac{1}{\sqrt{\frac{2\pi\sigma^2\tau^2}{\sigma^2+n\tau^2}}} \exp\Big(-\frac{\sigma^2+n\tau^2}{2\sigma^2\tau^2}(\mu - \frac{n\bar{D}\tau^2}{\sigma^2+n\tau^2})^2\Big)$ is a normalized Gaussian, so the integral will evaluate to 1 as given in the hint and we can simplify to our expression for the marginal likelihood of the training data $p(D)$:

$$p(D) = (\frac{1}{\sigma\sqrt{2\pi}})^n (\frac{1}{\tau\sqrt{2\pi}}) \exp\Big(-\frac{1}{2\sigma^2}(n\bar{D}^2 - \frac{n^2(\bar{D})^2\tau^2}{\sigma^2+n\tau^2})\Big)(\sqrt{\frac{2\pi\sigma^2\tau^2}{\sigma^2+n\tau^2}})$$

$$p(D) = (\frac{1}{\sigma\sqrt{2\pi}})^n (\frac{1}{\tau\sqrt{2\pi}}) \exp\Big(-\frac{1}{2\sigma^2}(n\bar{D}^2 - \frac{n^2(\bar{D})^2\tau^2}{\sigma^2+n\tau^2})\Big)(\frac{\sigma\tau\sqrt{2\pi}}{\sqrt{\sigma^2+n\tau^2}})$$

$$\boxed{p(D) = (\frac{1}{\sigma\sqrt{2\pi}})^n (\frac{\sigma}{\sqrt{\sigma^2+n\tau^2}}) \exp\Big(-\frac{1}{2\sigma^2}(n\bar{D}^2 - \frac{n^2(\bar{D})^2\tau^2}{\sigma^2+n\tau^2})\Big)}$$

Substituting $\sigma^2 = 1$ and $\tau^2 = 5$ gives

$$p(D) = (\frac{1}{\sqrt{2\pi}})^n (\frac{1}{\sqrt{1+5n}}) \exp\Big(-\frac{1}{2}(n\bar{D}^2 - \frac{5n^2(\bar{D})^2}{1+5n})\Big)$$

We can plug in our data where $n = 14$ and

$$n\bar{D} = \sum_{i=1}^{14} D_i = 3.3 + 3.5 + 3.1 + 1.8 + 3.0 + 0.74 + 2.5 + 2.4 + 1.6 + 2.1 + 2.4 + 1.3 + 1.7 + 0.19 = 29.63$$

$$n\bar{D}^2 = \sum_{i=1}^{14} D_i^2 = 3.3^2+3.5^2+3.1^2+1.8^2+3.0^2+0.74^2+2.5^2+2.4^2+1.6^2+2.1^2+2.4^2+1.3^2+1.7^2+0.19^2 = 74.8937$$

Computing:

$$p(D) = (\frac{1}{\sqrt{2\pi}})^{14} (\frac{1}{\sqrt{1+5(14)}}) \exp\Big(-\frac{1}{2}(74.8937 - \frac{5(14^2)(\frac{29.63}{14})^2}{1+5(14)})\Big)$$

$$\boxed{p(D) = 4.46287 \times 10^{-10}}$$

8. Recall our expression for the marginal likelihood $p(D)$:

$$p(D) = (\frac{1}{\sigma\sqrt{2\pi}})^n (\frac{\sigma}{\sqrt{\sigma^2 + n\tau^2}}) \exp\Big(- \frac{1}{2\sigma^2}\big(n\bar{D}^2 - \frac{n^2(\bar{D})^2\tau^2}{\sigma^2 + n\tau^2}\big)\Big)$$

substituting $\tau^2 = 0.1$ instead of $\tau^2 = 5$ allows us to reuse most of our computation from above:

$$p(D) = (\frac{1}{\sqrt{2\pi}})^{14} (\frac{1}{\sqrt{1 + 0.1(14)}}) \exp\Big(- \frac{1}{2}\big(74.8937 - \frac{0.1(14^2)(\frac{29.63}{14})^2}{1 + 0.1(14)}\big)\Big)$$

$$\boxed{p(D) = 8 \times 10^{-15}}$$