APHY/ICSI/IINF 451/551 **BAYESIAN DATA ANALYSIS MIDTERM** 22 Oct 2020

NAME______________

This is a Take Home Exam
It is due to be emailed to knuthclass@gmail.com by Tuesday Oct 27, 2020 at 11:59 pm.

Do your own work.
You are not allowed to discuss the exam or problems herein with others.

Show all of your work!
Written solutions and all code are to be submitted. Just as with the homework, all figures are to be embedded in either a MSWord or a pdf document and described.

Copying and cheating will not be tolerated and will result in failure of the exam.
The exam (130 points total) consists of five problems, with sub-parts.

1. (15 points)
   Write Bayes' Theorem.
   Name each term in the equation and write a description (at least one sentence each) describing the meaning and/or use of each term. (15 points)

SOLUTION #1

$$p(m\,|d,I) = p(m\mid I)\,\frac{p(d\,|m,I)}{p(d\,|I)}$$

Posterior probability, $p(m\,|d,I)$, denotes the probability of the specific model m given both the data d and the prior information I.

The prior probability, $p(m\mid I)$, quantifies the probability of the specific model m given only the prior information.

The likelihood, $p(d\,|m,I)$, describes the likelihood that the data d could have been recorded given the model m and prior information I.

The evidence, $p(d\,|I)$, is the probability of obtaining the data based only on the prior information. It is found by marginalizing (integrating) the product of the prior and the likelihood over all of the model parameter values, m. The evidence often acts as a normalization factor.

2. (20 points)
   You are playing a role-playing adventure game in which a player determines what happens in the game by rolling a die. You cannot see the die being rolled and have to trust the numbers that the person claims to roll.

   The person is supposed to be rolling a fair 20-sided die, but after rolling a 3 and a 9, you begin to suspect that the person might be rolling a 12-sided die by mistake. If there is an equal chance that the person is rolling either die, what is the probability that they are rolling the 12-sided die by mistake?

SOLUTION #2

We have two hypotheses: the person is rolling a:

D12: 12-sided die
or
D20: 20-sided die

We have some data in the form of a 3 having been rolled (R3) and a 9 having been rolled (R9).

The desired posterior probability is: $P(D12 \mid R3, R9, I)$.

$$P(D12 \mid R3, R9, I) = P(D12 \mid I)\,\frac{P(R3, R9 \mid D12, I)}{P(R3, R9 \mid I)}$$

The rolls are independent, so the likelihood can be factored

$$P(R3, R9 \mid D12, I) = P(R3 \mid D12, I)\,P(R9 \mid D12, I) = \frac{1}{12} \cdot \frac{1}{12} = \frac{1}{144}$$

For the evidence we will need the likelihood conditional on the other hypothesis $D20$, too:

$$P(R3, R9 \mid D20, I) = P(R3 \mid D20, I)\,P(R9 \mid D20, I) = \frac{1}{20} \cdot \frac{1}{20} = \frac{1}{400}$$

Note that had the person rolled a number greater than 12, the likelihood for the hypothesis $D12$ would have been zero, and there would be no question as to which die was rolled.

The evidence is given by

$$P(R3, R9 \mid I) = P(D12 \mid I)\,P(R3, R9 \mid D12, I) + P(D20 \mid I)\,P(R3, R9 \mid D20, I)$$

$$= \frac{1}{2} \cdot \frac{1}{144} + \frac{1}{2} \cdot \frac{1}{400} = \frac{1}{2} \cdot \frac{144 + 400}{144 \cdot 400} = \frac{272}{144 \cdot 400}$$

The posterior for $D12$ is then

$$P(D12 \mid R3, R9, I) = P(D12 \mid I)\,\frac{P(R3, R9 \mid D12, I)}{P(R3, R9 \mid I)}$$

$$= \frac{1}{2} \cdot \frac{\frac{1}{144}}{\frac{272}{144 \cdot 400}} = \frac{1}{2} \cdot \frac{400}{272} = \frac{200}{272} = \mathbf{0.735}$$

The posterior for $D20$ is

$$P(D20 \mid R3, R9, I) = P(D20 \mid I)\, \frac{P(R3, R9 \mid D20, I)}{P(R3, R9 \mid I)}$$

$$= \frac{1}{2} \cdot \frac{\frac{1}{400}}{\frac{272}{144 \cdot 400}} = \frac{1}{2} \cdot \frac{144}{272} = \frac{72}{272} = 0.265$$

And they sum to one, as they should.

3. (35 points)

In World War II, when the Allies (mainly Britain and America) were planning an invasion of France to take Europe back from the Germans, they needed to know how many German tanks they would be up against.

From 20 captured German tanks, the Allies found this set of serial numbers on the engine blocks:

038839	019401	012031	020106	004802	006570	046893	047594	028633
002976	011687	017580	040877	000767	002142	008412	032312	036423
032243	022446							

In summary, the serial numbers range from 000767 to 047594.

Assuming that the Germans assigned sequential numbers to the engine blocks, and that the twenty captured tanks represent a uniformly-distributed sample from the set of tanks that the German's possess, how many tanks do the Germans have?

We will approach this problem in steps.

The aim is to estimate the number N of tanks possessed by the German military.

A. Assign a prior probability, $P(N \mid I)$, for the number of tanks.
(You can safely assume that they cannot possibly have more than a million tanks.)

Using Bayes Theorem, we can write the posterior probability as

$$P(N \mid d_1, d_2, d_3, \cdots, d_{20}, I) = \frac{1}{Z}\ P(\, N \mid I\,)\ P(d_1, d_2, d_3, \cdots, d_{20} \mid N, I)$$

B. The posterior probability depends on the joint likelihood $P(d_1, d_2, d_3, \cdots, d_{20} \mid N,\ I)$. Assuming that the serial numbers are independent, use the product rule to write the joint likelihood in terms of the individual likelihoods $P(\, d_1 \mid N, I)$, $P(\, d_2 \mid N, I)$, etc.

C. Given that there are N tanks (where $N \gg 20$) and the assumption that the captured tanks represent a uniformly-distributed sample, assign a likelihood for $P(\, d_i \mid N, I)$ by considering the two cases below (and filling in the blanks):

$$P(\, d_i \mid N, I) = \begin{cases} \_\_ & \text{for } N \ \geq \ d_i \\ \_\_ & \text{for } N \ < \ d_i. \end{cases}$$

D. Write a MATLAB program to evaluate the posterior probability for values of N ranging from 1 to 100,000. You can store the posterior probability values in a 1 x 100000 array called p.
To start, ignore the normalization factor (evidence) Z in the posterior probability. This means that the probabilities will not sum to one, as they should. You can handle this numerically by summing the probabilities later and then renormalizing them:

```
sumprobs = sum(p);   % sum the probabilities

posteriors = p / sumprobs;    % renormalize the probabilities
```

You can now check that the probabilities are now normalized by evaluating `sum(posteriors)` and verifying that the sum is 1.

Produce a plot of the normalized posterior probabilities and show that there is a most probable solution to the problem.

E. Find the most probable value of N.

F. Use MATLAB code to compute the average value, or expected value of N.

It may help to create an array `n = 1:100000`; and use this with the array of normalized posterior probabilities, `posteriors`, to compute this expected value.

G. How many German tanks should the Allies expect? Explain why you would recommend this solution.

SOLUTION #3

This solution is only approximate.

A complete analytical solution can be found on page 41 in the book Bayesian Probability Theory: Applications in the Physical Sciences by von der Linden, Dose, and von Toussaint.

A. Assign a prior probability, $P(N \mid I)$, for the number of tanks.
(You can safely assume that they cannot possibly have more than a million tanks.)

When writing this problem, I choose 1 million as an arbitrary cut-off.

One could assign a uniform prior,

$$P(N \mid I) = \frac{1}{1000000} = 10^{-6}$$

One might acknowledge that making tanks is effortful, even for Germans, and assign Jeffrey's prior

$$P(N \mid I) \propto \frac{1}{N}$$

One really ought to make sure that this Jeffrey's prior is normalized

$$P(N \mid I) = \frac{C}{N}$$

with

$$\int_{1}^{1000000} \frac{C}{N}\, dN = 1$$

where $C = \frac{1}{\log 1000000} = \frac{1}{6 \log 10}$

so that

$$P(N \mid I) = \frac{N^{-1}}{6 \log 10}$$

I am going to work with the uniform prior.

B. The posterior probability depends on the joint likelihood $P(d_1, d_2, d_3, \cdots, d_{20} \,|N,\ I)$. Assuming that the serial numbers are independent, use the product rule to write the joint likelihood in terms of the individual likelihoods $P(\,d_1 \,|\, N, I)$, $P(\,d_2 \,|\, N, I)$, etc.

First, let us factor the joint likelihood.

$$P(d_1, d_2, d_3, \cdots, d_{20} \,|N, \qquad I) = P(d_1|\, N, I)\, P(d_2, d_3, \cdots, d_{20} \,|\, d_1, N, I)$$

$$= P(d_1|\, N, I)\, P(d_2 \,|\, d_1, N, I)\, P(d_3 \,|\, d_1, d_2, N, I) \ldots$$

We'll come back to this later in the second, more precise solution.

Here we will assume that $P(d_2 \,|\, d_1, N, I) = P(d_2 \,|\, N, I)$, which isn't exactly right because with d_1 known, there really are $N-1$ tanks left.

But let us use this approximation for the moment.

$$P(d_1, d_2, d_3, \cdots, d_{20} \,|N, \qquad I) \approx P(\,d_1 \,|\, N, I)\, P(\,d_2 \,|\, N, I) \ldots P(\,d_{20} \,|\, N, I)$$

In part C, we evaluate these terms

C. Given that there are N tanks (where $N \gg 20$) and the assumption that the captured tanks represent a uniformly-distributed sample, assign a likelihood for $P(\,d_i \,|\, N, I)$ by considering the two cases below (and filling in the blanks):

$$P(\,d_i \,|\, N, I) = \begin{cases} \_\_ & \text{for } N \ \geq \ d_i \\ \_\_ & \text{for } \ N \ < \ d_i. \end{cases}$$

Clearly, we cannot have $N \ < \ d_i$, so the likelihood must be zero in that case.

And if the tanks are uniformly-distributed, then for $N \ \geq \ d_i$, we have $P(\,d_i \,|\, N, I) = \frac{1}{N} = N^{-1}$.

So

$$P(\,d_i \,|\, N, I) = \begin{cases} N^{-1} & \text{for } N \ \geq \ d_i \\ 0 & \text{for } \ N \ < \ d_i \end{cases}$$

D. Write a MATLAB program to evaluate the posterior probability for values of N ranging from 1 to 100,000.

The posterior looks like

$$P(N \,|\, d_1, d_2, d_3, \cdots, d_{20}, I) = \frac{1}{Z}\ P(\,N \,|\, I\,)\ P(d_1, d_2, d_3, \cdots, d_{20} \,|N, I)$$

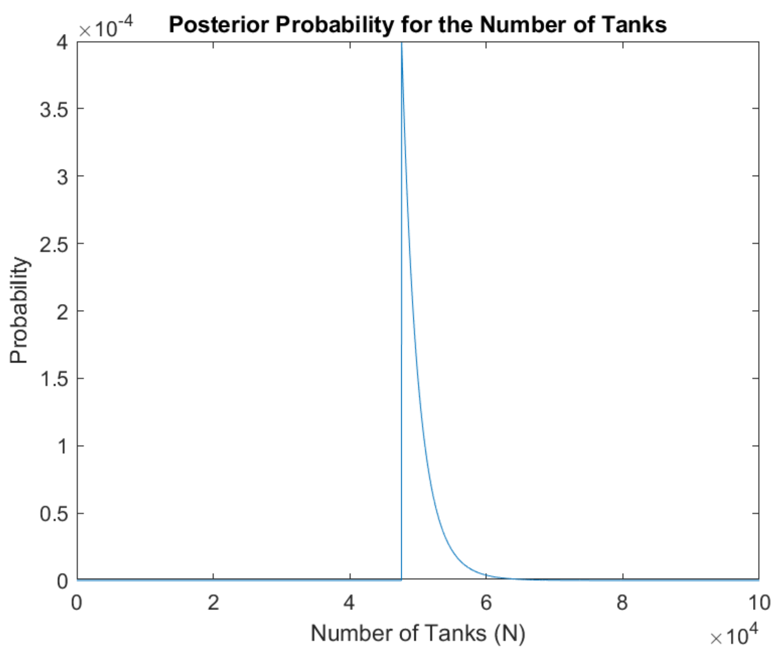
$$P(N \,|\, d_1, d_2, d_3, \cdots, d_{20}, I) = \frac{1}{Z}\ 10^{-6}\ P(\,d_1 \,|\, N, I)\, P(\,d_2 \,|\, N, I) \ldots P(\,d_{20} \,|\, N, I)$$

```
% tank serial numbers
x = [038839, 019401, 012031, 020106, 004802, 006570, 046893, 047594, 028633, 002976,...
    011687, 017580, 040877, 000767, 002142, 008412, 032312, 036423, 032243, 022446];
```

```
M = 100000;
for m = 1:M
    p(m) = 10^-6;    % start with the uniform prior

    for i = 1:length(x)      % multiply by each of the 20
likelihoods
        if (x(i) <= m)
            p(m) = p(m)*(1/m);
        else
            p(m) = p(m)*0;
        end
    end
end

% normalize the posterior p
sump = sum(p);
np = p/sump;
```



E. Find the most probable value of N.

```
% find the maximum
[v, n] = max(np);
```

One can see that the maximum is at $N = 47594$, which is approximately that largest serial number found. But the mean value will be higher.

F. Use MATLAB code to compute the average value, or expected value of N.

It may help to create an array `n = 1:100000`; and use this with the array of normalized posterior probabilities, `posteriors`, to compute this expected value.

```
% find mean
avg = sum((1:M).* np)
```

The mean number of tanks is $\overline{N} = 50238$.

G. How many German tanks should the Allies expect? Explain why you would recommend this solution.

I would recommend to prepare for 50238 tanks (the mean value), since the probability that the tank with the second largest serial number was captured would be very small $\approx \frac{20}{47594} = 0.042\%$. Also, the mean value is the larger of the two, and preparing for a larger number of tanks will be safer.

Incidentally, the serial numbers were randomly drawn from a set of 49777 tanks.
So, the correct number of tanks was 49777.

4. (35 points)
Finding the minimum of the Rosenbrock's Banana Function

$$f(x, y) = 100\,(y - x^2)^2 + \,(1 - x)^2$$

is notoriously difficult using numerical optimization due to the slow convergence rates.
(https://www.mathworks.com/help/optim/ug/banana-function-minimization.html)

We will work with this function as if it were a probability density function (I have multiplied $f(x, y)$ by -1 so that $p(x, y \mid I)$ has a maximum).

$$p(x, y \mid I) = \frac{-100\,(y - x^2)^2 - \,(1 - x)^2}{Z}$$

defined in the region $-c \leq x \leq c$ and $-d \leq y \leq d$ for some positive values of c and d.

A. Find the normalization constant Z.

B. Using calculus, find the (x, y) coordinates of the maximum of $p(x, y \mid I)$.

C. Find the covariance matrix about the peak of $p(x, y \mid I)$.

D. Find an expression for $p(x \mid I)$.

E. Find the peak of $p(x \mid I)$ and its associated uncertainty.

F. Is the peak of $p(x \mid I)$ at the same x-value as the peak of $p(x, y \mid I)$?
Discuss this.

G. Plot $p(x, y \mid I)$ in one figure and $p(x \mid I)$ in another.
You can concentrate on the range where $c = d = 5$.
You may want to look at the code included in the class notes from Oct 15.

SOLUTION #4

Given

$$p(x,y \mid I) = \frac{-100\,(y-x^2)^2 - \,(1-x)^2}{Z}$$

defined in the region $-c \le x \le c$ and $-d \le y \le d$ for some positive values of c and d.

A. Find the normalization constant Z.

To do this we use the fact that the probabilities must sum to one.

$$\int_{-c}^{c} dx \int_{-d}^{d} dy\, p(x,y \mid I) \;=1$$

Let's perform the integral over y

$$\int_{-d}^{d} dy\, \frac{-100\,(y-x^2)^2 - \,(1-x)^2}{Z} = \frac{1}{Z}\int_{-d}^{d} dy\,[-100\,(y-x^2)^2 - \,(1-x)^2]$$

$$= \frac{-100}{Z}\int_{-d}^{d} dy\,[\,(y-x^2)^2] - \frac{1}{Z}\int_{-d}^{d} dy\,(1-x)^2$$

$$= \frac{-100}{Z}\int_{-d}^{d} dy\,[\,y^2 - 2x^2y + \,x^4] - \frac{1}{Z}\int_{-d}^{d} dy\,(1-x)^2$$

$$= \frac{-100}{Z}\left[\frac{y^3}{3} - x^2y^2 + \,x^4y\right]_{-d}^{d} - \frac{1}{Z}\,[(1-x)^2\,y\,]_{-d}^{d}$$

$$= \frac{-200}{Z}\left[\frac{d^3}{3} - 0 + \,x^4d\right] - \frac{2}{Z}\,[(1-x)^2\,d\,]$$

$$= -\frac{2}{Z}\left[100\left(\frac{d^3}{3} + \,x^4d\right) + \,(1-x)^2\,d\right]$$

$$= -\frac{2}{Z}\,d\left[\,100\left(\frac{d^2}{3} + \,x^4\right) + \,(1-x)^2\,\right]$$

$$= -\frac{2}{Z}\,d\left[\,100\left(\frac{d^2}{3} + \,x^4\right) + \,x^2 - 2x + 1\,\right]$$

Now integrate over x

$$\int_{-c}^{c} dx \int_{-d}^{d} dy \, \frac{-100\,(y-x^2)^2 - (1-x)^2}{Z}$$

$$= \int_{-c}^{c} dx \left\{ -\frac{2}{Z}\, d \left[100 \left(\frac{d^2}{3} + x^4 \right) + x^2 - 2x + 1 \right] \right\}$$

$$= -\frac{2}{Z}\, d \int_{-c}^{c} dx \left[100 \left(\frac{d^2}{3} + x^4 \right) + x^2 - 2x + 1 \right]$$

$$= -\frac{2}{Z}\, d \left[100 \left(\frac{d^2}{3} x + \frac{x^5}{5} \right) + \frac{x^3}{3} - x^2 + x \right]_{-c}^{c}$$

$$= -\frac{4}{Z}\, d \left[100 \left(\frac{d^2}{3} c + \frac{c^5}{5} \right) + \frac{c^3}{3} - 0 + c \right]$$

$$= -\frac{4}{Z}\, c\, d \left[100 \left(\frac{d^2}{3} + \frac{c^4}{5} \right) + \frac{c^2}{3} + 1 \right]$$

$$1 = -\frac{4}{Z}\, c\, d \left[100 \left(\frac{d^2}{3} + \frac{c^4}{5} \right) + \frac{c^2}{3} + 1 \right]$$

$$Z = -4\, c\, d \left[100 \left(\frac{d^2}{3} + \frac{c^4}{5} \right) + \frac{c^2}{3} + 1 \right]$$

$$Z = -\frac{4}{3}\, c\, d \left[100 \left(d^2 + \frac{3c^4}{5} \right) + c^2 + 3 \right]$$

$$Z = -\frac{4}{3} \left[100\, cd^3 + 60\, c^5 d + c^3 d + 3cd \right]$$

Since c and d are positive, it is clear that Z is negative.

This means that the function $p(x, y \mid I)$ is negative within the domain $-c \leq x \leq c$ and $-d \leq y \leq d$. So clearly, this cannot be a probability density.

When designing this problem, I should have added a sufficiently large constant to the definition of $p(x, y \mid I)$ to ensure that the function was positive everywhere within the domain.

B. Using calculus, find the (x, y) coordinates of the maximum of $p(x, y \mid I)$.

Despite the fact that $p(x, y \mid I)$ is not a probability density function, we can find its maximum.

We need to look at

$$\partial_x\, p(x, y \mid I) = \frac{\partial}{\partial x}\, p(x, y \mid I)$$

$$\partial_y\, p(x, y \mid I) = \frac{\partial}{\partial y}\, p(x, y \mid I)$$

The derivatives are:

$$\frac{\partial}{\partial x}\, p(x, y \mid I) = \frac{\partial}{\partial x}\left[\frac{-100\,(y - x^2)^2 - \,(1 - x)^2}{Z}\right]$$

$$= \frac{1}{Z}\ \frac{\partial}{\partial x}\,[-100\,(y - x^2)^2 - \,(1 - x)^2\,]$$

$$= \frac{1}{Z}\ \,[-200\,(y - x^2)(-2x) - \,2(1 - x)(-1)\,]$$

$$= \boldsymbol{\frac{1}{Z}\ \,[400\,(y - x^2)x + \,2(1 - x)\,]}$$

$$\boldsymbol{\frac{\partial}{\partial x}\, p(x, y \mid I) = \frac{1}{Z}\ \,[400\,xy - 400\,x^3\, - \,2x + 2\,]}$$

and

$$\frac{\partial}{\partial y}\, p(x, y \mid I) = \frac{\partial}{\partial y}\left[\frac{-100\,(y - x^2)^2 - \,(1 - x)^2}{Z}\right]$$

$$= \frac{1}{Z}\ \frac{\partial}{\partial y}\,[-100\,(y - x^2)^2 - \,(1 - x)^2\,]$$

$$= \frac{1}{Z}\ \,[-200\,(y - x^2)\,]$$

$$\boldsymbol{\frac{\partial}{\partial y}\, p(x, y \mid I) = \frac{1}{Z}\ \,[-200\,(y - x^2)\,]}$$

Set both equations to zero and solve for x and y.

$$\frac{1}{Z}\ \left[400\,(y - x^2)x +\ 2(1 - x)\,\right] = 0$$

$$\frac{1}{Z}\ \left[-200\,(y - x^2)\,\right] = 0$$

The second equation gives

$$y =\ x^2$$

Substituting this into the first equation gives

$$[0 +\ 2(1 - x)\,] =\ 0$$

$$x = 1$$

The maximum is at $(\boldsymbol{x}, \boldsymbol{y}) =\ (\mathbf{1}, \mathbf{1})$.

One should really look at the second derivatives to verify that this is a maximum and not a minimum or inflection point.

To do this compute the three second derivatives: $p_{xx} = \frac{\partial^2}{\partial x^2}\, \boldsymbol{p}(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{I})$, $p_{yy} = \frac{\partial^2}{\partial y^2}\, \boldsymbol{p}(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{I})$, and $p_{xy} = \frac{\partial}{\partial x}\frac{\partial}{\partial y}\, \boldsymbol{p}(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{I})$. Verify that $p_{xx}\, p_{yy} -\ {p_{xy}}^2 > 0$, and that $p_{xx} < 0$ and $p_{yy} < 0$.

This will be problematic if one worries about the sign of Z, which is negative and will flip everything upside-down.

C. Find the covariance matrix about the peak of $p(x, y \mid I)$.

To do this, we need to compute the matrix of second derivatives of $\log[p(x, y \mid I)]$ evaluated at the maximum $(x, y) =\ (1,1)$. However, because of the logarithm, these derivatives are proportional to $\frac{1}{p(x,y \mid I)}$, which is not defined at the peak because $p(x = 1, y = 1 \mid I) = 0$.

Again, because $p(x, y \mid I)$ is not positive within the domain, it is not a probability density function.

D. Find an expression for $p(x \mid I)$.

Even though $p(x, y \mid I)$ is not a density function, we can go through the mechanics of integrating over y.

Fortunately, this integral was performed in Part A, when we found Z

$$p(x \mid I) = \int_{-d}^{d} dy \; p(x, y \mid I)$$

$$= \int_{-d}^{d} dy \; \frac{-100\,(y - x^2)^2 - (1-x)^2}{Z}$$

$$= -\frac{2}{Z}\, d \left[100 \left(\frac{d^2}{3} + x^4 \right) + x^2 - 2x + 1 \right]$$

E. Find the peak of $p(x \mid I)$ and its associated uncertainty.

Take the derivative with respect to x

$$\frac{\partial}{\partial x}\, p(x \mid I) = \frac{\partial}{\partial x} \left\{ -\frac{2}{Z}\, d \left[100 \left(\frac{d^2}{3} + x^4 \right) + x^2 - 2x + 1 \right] \right\}$$

$$= -\frac{2}{Z}\, d \,[\, 100\,(4\,x^3) + 2x - 2 \,]$$

$$= -\frac{2}{Z}\, d \,[\, 400\,x^3 + 2x - 2 \,]$$

$$= -\frac{4}{Z}\, d \,[\, 200\,x^3 + x - 1 \,]$$

Setting this equal to zero and solving for x, we have that

$$-\frac{4}{Z}\, d \,[\, 200\,x^3 + x - 1 \,] = 0$$

$$200\,x^3 + x - 1 = 0$$

The real solution to this equation is $\boldsymbol{x = 0.16126}$

Look at the second derivative to check that $x = 0.16126$ is the maximum.

$$\frac{\partial^2}{\partial x^2}\, p(x \mid I\,) = \frac{\partial}{\partial x}\left\{\frac{\partial}{\partial x}\, p(x \mid I\,)\right\}$$

$$= \frac{\partial}{\partial x}\left\{-\frac{4}{Z}\, d\,[\,200\, x^3 + x - 1\,]\right\}$$

$$= -\frac{4}{Z}\, d\,[\,600\, x^2 + 1\,]$$

which would be negative (and hence a peak), if Z were positive (like it ought to be).

F. Is the peak of $p(x \mid I)$ at the same x-value as the peak of $p(x, y \mid I)$? Discuss this.

No. The peak of $p(x, y \mid I)$ is at $(x, y) = (1, 1)$, whereas the peak of $p(x \mid I)$ is at $x = 0.16126$. Integrating over y can change the value of x at which the resulting function is maximum.

This is one reason why maxima are not an entirely reliable summary of the probability mass of the pdf.

5. (25 points)
An archeologist is studying drug abuse in the times of the late Roman Empire (circa 250–450 A.D.). This is accomplished by testing pottery for traces of opium. Samples from one such vessel were sent to five different labs to determine the age of the artifact.

Laboratory A returned a report stating that they dated the artifact at 327.1 ± 1.2 A.D.
Laboratory B returned a report stating that they dated the artifact at 332.5 ± 3.2 A.D.
Laboratory C returned a report stating that they dated the artifact at 321.2 ± 7.5 A.D.
Laboratory D returned a report stating that they dated the artifact at 318.0 ± 2.2 A.D.
Laboratory E returned a report stating that they dated the artifact at 325.3 ± 3.5 A.D.

Assuming that the listed uncertainties represent a 1-σ estimate from a Gaussian distribution:

A. Compute the most probable estimate for the age of the pottery and its uncertainty. You might consult with the Measuring Lengths lecture notes, but you will have to complete the notes by deriving the formula for estimating the uncertainty of the most probable estimate.

B. Plot the data with error bars along with lines illustrating the most probable age along with the uncertainties.

Given the data, which I will refer to as $x_1, x_2, \cdots, x_5$, for the data from labs A, B, …, E, respectively, we wish to infer the age, x, of the pottery. The data are independent. And will assign a Gaussian likelihood taking the reported error bars as the standard deviations of each Gaussian.

The likelihood for the ith data value is

$$p(x_i \mid x, I) = \frac{1}{\sqrt{2\pi}\,\sigma_i}\, exp\left[-\frac{1}{2\,\sigma_i^2}\,(x - x_i)^2\right]$$

And the joint likelihood is

$$p(\{x\} \mid x, I) = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi}\,\sigma_i}\, exp\left[-\frac{1}{2\,\sigma_i^2}\,(x - x_i)^2\right]$$

Which is simplified by considering the log likelihood

$$\ln p(\{x\} \mid x, I) = -\sum_{i=1}^{5} \log\sqrt{2\pi}\,\sigma_i - \frac{1}{2}\sum_{i=1}^{5}[\,\sigma_i^{-2}\,(x - x_i)^2]$$

Using Bayes' theorem, along with a uniform prior $p(x|I) = C$ over a reasonable range (one could use a range from 1 AD to 400 AD), we have that

$$\ln p(x \mid \{x\}, I) = \ln p(x|I) + \ln p(\{x\} \mid x, I) - \ln Z$$

We begin by finding the most probable value of x by taking the first derivative of the log posterior with respect to x.

$$\frac{d}{dx}[\ln p(x \mid \{x\}, I)] = \frac{d}{dx}[\ln p(x|I)] + \frac{d}{dx}[\ln p(\{x\} \mid x, I)] - \frac{d}{dx}[\ln Z]$$

$$\frac{d}{dx}[\ln p(x \mid \{x\}, I)] = 0 + \frac{d}{dx}[\ln p(\{x\} \mid x, I)] - 0$$

$$= \frac{d}{dx}[\ln p(\{x\} \mid x, I)]$$

$$= \frac{d}{dx}\left[-\sum_{i=1}^{5} \log \sqrt{2\pi}\,\sigma_i - \frac{1}{2}\sum_{i=1}^{5}[\,\sigma_i^{-2}\,(x - x_i)^2]\right]$$

$$= -\frac{1}{2}\,\frac{d}{dx}\left[\sum_{i=1}^{5}[\,\sigma_i^{-2}\,(x - x_i)^2]\right]$$

$$= -\frac{1}{2}\left[(2)\sum_{i=1}^{5}[\,\sigma_i^{-2}\,(x - x_i)]\right]$$

$$= -\sum_{i=1}^{5}[\,\sigma_i^{-2}\,(x - x_i)]$$

Setting this derivative equal to 0 and $x = \hat{x}$, solving for x:

$$\sum_{i=1}^{5}[\,\sigma_i^{-2}\,(\hat{x} - x_i)] = 0$$

$$\sum_{i=1}^{5}[\,\sigma_i^{-2}\,\hat{x}] - \sum_{i=1}^{5}[\,\sigma_i^{-2}\,x_i] = 0$$

$$\hat{x}\sum_{i=1}^{5}\sigma_i^{-2} - \sum_{i=1}^{5}[\,\sigma_i^{-2}\,x_i] = 0$$

$$\hat{x} = \frac{\sum_{i=1}^{5}[\,\sigma_i^{-2}\, x_i]}{\sum_{i=1}^{5}\sigma_i^{-2}}$$

The most probable estimate of x is found by taking a weighted average of the estimates provided by the five laboratories.

Taking the second derivative, we can estimate the uncertainty.

$$\frac{d^2}{dx^2}[\ln p(x \mid \{x\}, I)] = \frac{d}{dx}\left\{\frac{d}{dx}[\ln p(x \mid \{x\}, I)]\right\}$$

$$\frac{d^2}{dx^2}[\ln p(x \mid \{x\}, I)] = \frac{d}{dx}\left\{-\sum_{i=1}^{5}[\,\sigma_i^{-2}\,(x - x_i)]\right\}$$

$$= -\sum_{i=1}^{5}[\,\sigma_i^{-2}\,]$$

We then have that the uncertainty is given by

$$\sigma^2 = -\left[-\sum_{i=1}^{5}\sigma_i^{-2}\right]^{-1}$$

$$\sigma = \left[\sum_{i=1}^{5}\sigma_i^{-2}\right]^{-1/2}$$

We use the two equations above to evaluate the data.

$$\hat{\boldsymbol{x}} = \mathbf{325.64} \pm \mathbf{0.95}.$$