Bayesian Learning: Homework 1

Christian Steinmetz
Machine Learning - Winter 2019

November 13, 2019

Problem 2.8. Assuming a uniform prior on f_H, $P(f_H) = 1$, solve the problem posed in example 2.7 (p.30). Sketch the posterior distribution of f_H and compute the probability that the $N + 1$th outcome will be a head, for

(a) $N = 3$ and $n_H = 0$;

(b) $N = 3$ and $n_H = 2$;

(c) $N = 10$ and $n_H = 3$;

(d) $N = 300$ and $n_H = 29$.

You will find the beta integral useful:

$$\int_0^1 dp_a p_a^{F_a}(1 - p_a)^{F_b} = \frac{F_a!F_b!}{(F_a + F_b + 1)!} \tag{1}$$

You may also find it instructive to look back at example 2.6 (p.27) and equation (2.31).

Solution First, we note that we are asked to find f_H given an observation of N trials, where n_H is the number of heads observed. Implicitly, the number of tails, n_T, is given by $n_T = N - n_H$. Using Bayes' theorem we can find the desired posterior,

$$P(f_H|N, n_H) = \frac{P(n_H|f_H, N)P(f_H)}{P(n_H|N)} \tag{2}$$

which is of the form

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}. \tag{3}$$

In the problem statement, we are told to assume $P(f_H) = 1$ as the prior, which simplifies the problem. We can solve for the likelihood using the data for each case, which is a function of our parameter f_H, as follows, since the coin flipping process follows a binomial distribution.

$$P(n_H|f_H, N) = (f_H)^{n_H}(1 - f_H)^{N-n_H} \tag{4}$$

To find the normalized posterior, we also solve for the evidence, which requires calculating the marginal probabilities with this integral.

$$p(n_H|N) = \int_0^1 f_H^{n_H}(1-f_H)^{N-n_H} df_H \tag{5}$$

We notice that this follows the same form as the beta integral in Equation 1 and we can substitute our values of n_H and N as follows

$$p(n_H|N) = \frac{(n_H)!(N-n_H)!}{(n_H+N-n_H+1)!} = \frac{(n_H)!(N-n_H)!}{(N+1)!}. \tag{6}$$

Finally, this yields the following relationship for the posterior

$$\begin{align} P(f_H|N,n_H) &= \frac{P(n_H|f_H,N)P(f_H)}{P(n_H|N)} \tag{7}\\ &= \frac{(f_H)^{n_H}(1-f_H)^{N-n_H}}{\frac{(n_H)!(N-n_H)!}{(N+1)!}} \tag{8}\\ &= \frac{(f_H)^{n_H}(1-f_H)^{N-n_H}(N+1)!}{(n_H)!(N-n_H)!} \tag{9}\end{align}$$

Below we plot the posterior over values of f_H. We note that as N increases, the peak within the plot becomes sharper. This indicates that we are more certain of the true value of the unknown parameter f_H.

To find our prediction for the probability of a heads in the $N+1$ coin toss, which we will call H_{N+1}, we apply the sum rule as follows

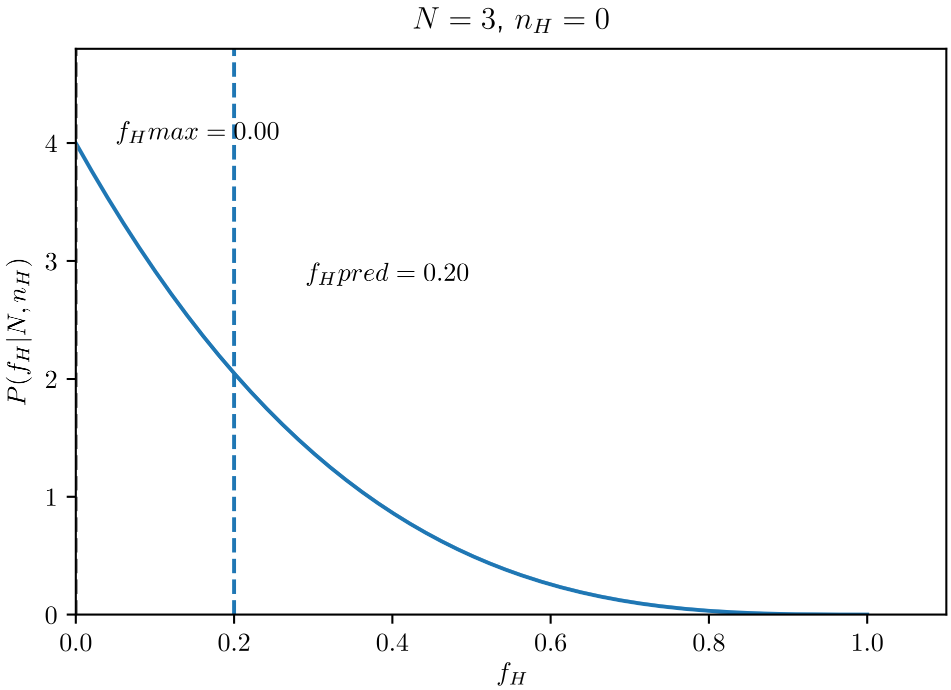
$$\begin{align} P(H_{N+1}|n_H,H) &= \int df_H P(H_{N+1}|f_H)P(f_H|n_H,N) \tag{10}\\ &= \int_0^1 df_H P(f_H|n_H,N) \tag{11}\\ &= \int_0^1 df_H \frac{(f_H)^{n_H+1}(1-f_H)^{N-n_H}}{\frac{(n_H)!(N-n_H)!}{(N+1)!}} \tag{12}\\ &= \frac{1}{\frac{(n_H)!(N-n_H)!}{(N+1)!}} \int_0^1 df_H (f_H)^{n_H+1}(1-f_H)^{N-n_H} \tag{13}\\ & \tag{14}\end{align}$$

We note that the denominator is not dependent on f_H and there we can once again apply the beta integral relationship to solve the remaining integral.

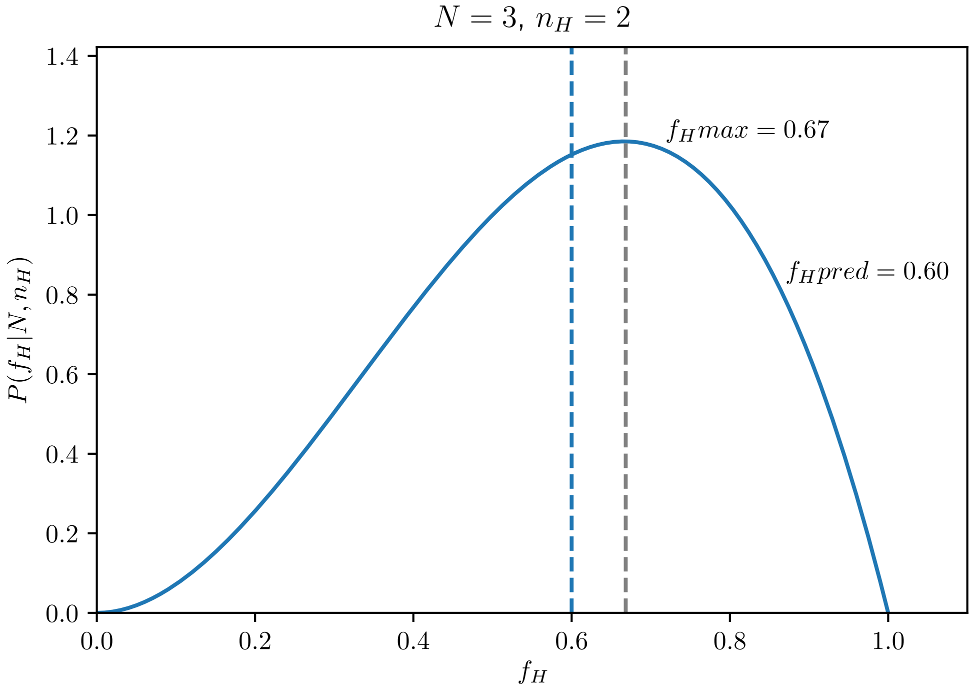
On the plots below we also show the maximum of f_H and we notice that our prediction is different when N is small. This is an advantage to using this Bayesian method, since our predictions with few examples tend to be closer to the underlying data generation distribution.

$$\begin{aligned} P(H_{N+1}|n_H, H) &= \frac{(N+1)!}{(n_H)!(N-n_H)!} \int_0^1 df_H (f_H)^{n_H+1}(1-f_H)^{N-n_H} && (15) \\ &= \frac{(N+1)!}{(n_H)!(N-n_H)!} \frac{(n_H+1)!(N-n_H)!}{(n_H+1+1+N-n_H)!} && (16) \\ &= \frac{(N+1)!}{(n_H)!(N-n_H)!} \frac{(n_H+1)!(N-n_H)!}{(N+2)!} && (17) \\ &= \frac{n_H+1}{N+2} && (18) \end{aligned}$$

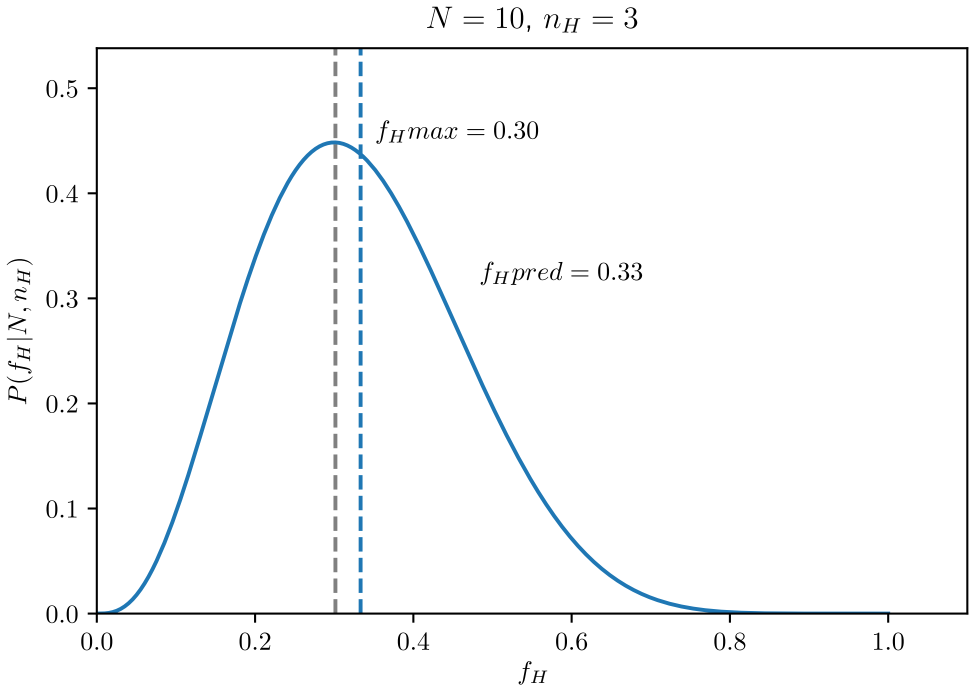
(a) $N = 3$ and $n_H = 0$;



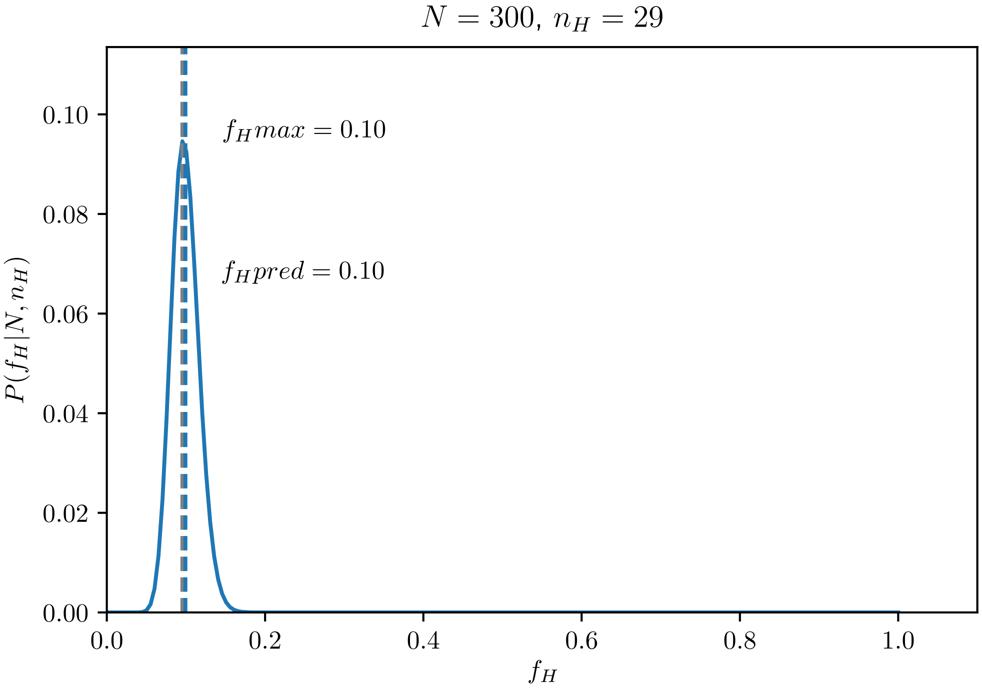
(b) $N = 3$ and $n_H = 2$;



(c) $N = 10$ and $n_H = 3$;



(d) $N = 300$ and $n_H = 29$.



Problem 2.10. Urn A contains three balls: one black, and two white; urn B contains three balls: two black, and one white. One of the urns is selected at random and one ball is drawn. The ball is black. What is the probability that the selected urn is urn A?

Solution We apply Bayes' theorem

$$P(A|n_B, N) = \frac{P(n_B, N|A)P(A)}{P(n_B, N)} \tag{19}$$

The likelihood asks what is the probability that we chose a single black ball, n_B in a single draw, N, assuming that A was in fact the urn that was drawn from. Since urn A has one black ball and two black balls, the probability of choosing a single black ball is $P(n_B, N|A) = \frac{1}{3}$.

The prior asks what is the probability of selecting urn A among the possible choices. Since there are only two urns we know that $P(A) = \frac{1}{/}2$.

Finally, the evidence asks what is the probability of the outcome we observed given all of the possible options, which we find using the sum rule over both urns.

$$\begin{aligned} P(n_B, N) &= P(n_B|A) \times P(A) + P(n_B|B) \times P(B) && (20) \\ &= \frac{1}{3} \times \frac{1}{2} + \frac{2}{3} \times \frac{1}{2} && (21) \\ &= \frac{1}{2} && (22) \end{aligned}$$

And we then use all of these results to compute our probability

$$\begin{aligned} P(A|n_B, N) &= \frac{P(n_B, N|A)P(A)}{P(n_B, N)} && (23) \\ &= \frac{\frac{1}{3} \times \frac{1}{2}}{\frac{1}{2} \times \frac{1}{3} + \frac{2}{3} \times \frac{1}{2}} && (24) \\ &= \frac{\frac{1}{3} \times \frac{1}{2}}{\frac{1}{2}} && (25) \\ &= \frac{1}{3} && (26) \end{aligned}$$

Problem 2.11. Urn A contains five balls: one black, two white, one green and one pink; urn B contains five hundred balls: two hundred black, one hundred white, 50 yellow, 40 cyan, 30 sienna, 25 green, 25 silver, 20 gold, and 10 purple. [One fifth of A's balls are black; two-fifths of B's are black.] One of the urns is selected at random and one ball is drawn. The ball is black. What is the probability that the urn is urn A?

Solution This is fairly similar to the above problem, just with different distribution of the balls within the two urns. The question we want to answer is still the same though; what is the probability that the urn is urn A?

We again apply Bayes' theorem

$$P(A|n_B, N) = \frac{P(n_B, N|A)P(A)}{P(n_B, N)} \tag{27}$$

Since urn A has one black ball and four other coloured balls, the probability of choosing a single black ball given we draw from urn A $P(n_B, N|A) = \frac{1}{5}$. Since there are still only two urns we know that $P(A) = \frac{1}{2}$. Again, the evidence asks what is the probability of the outcome we observed given all of the possible options, which we find using the sum rule over both urns.

$$\begin{align} P(n_B, N) &= P(n_B|A) \times P(A) + P(n_B|B) \times P(B) \tag{28} \\ &= \frac{1}{5} \times \frac{1}{2} + \frac{2}{5} \times \frac{1}{2} \tag{29} \\ &= \frac{3}{10} \tag{30} \end{align}$$

And we then use all of these results to compute our probability

$$\begin{align} P(A|n_B, N) &= \frac{P(n_B, N|A)P(A)}{P(n_B, N)} \tag{31} \\ &= \frac{\frac{1}{5} \times \frac{1}{2}}{\frac{3}{10}} \tag{32} \\ &= \frac{1}{3} \tag{33} \end{align}$$

Interestingly enough we arrive at the same answer for both scenarios. This is the result of the *likelihood principle*, wherein the detailed contents of each urn is not important for our answer, we only care about the probability that our outcome actually occurred. This is simply a function of the probability of a black ball being drawn from either urn.

Problem 3.5. Given a string of length F of which F_a are **a**s and F_b are **b**s, we are interested in (a) inferring what p_a might be; (b) predicting whether the next character is an **a** or a **b**.

Sketch the posterior probability $P(p_a|\mathrm{s} = \mathbf{aba}, F = 3)$. What is the most probable value of p_a (i.e., the value that maximizes the posterior probability density)? What is the mean value of p_a under this distribution?

Answer the same questions for the posterior probability $P(p_a|s = bbb, F = 3)$.

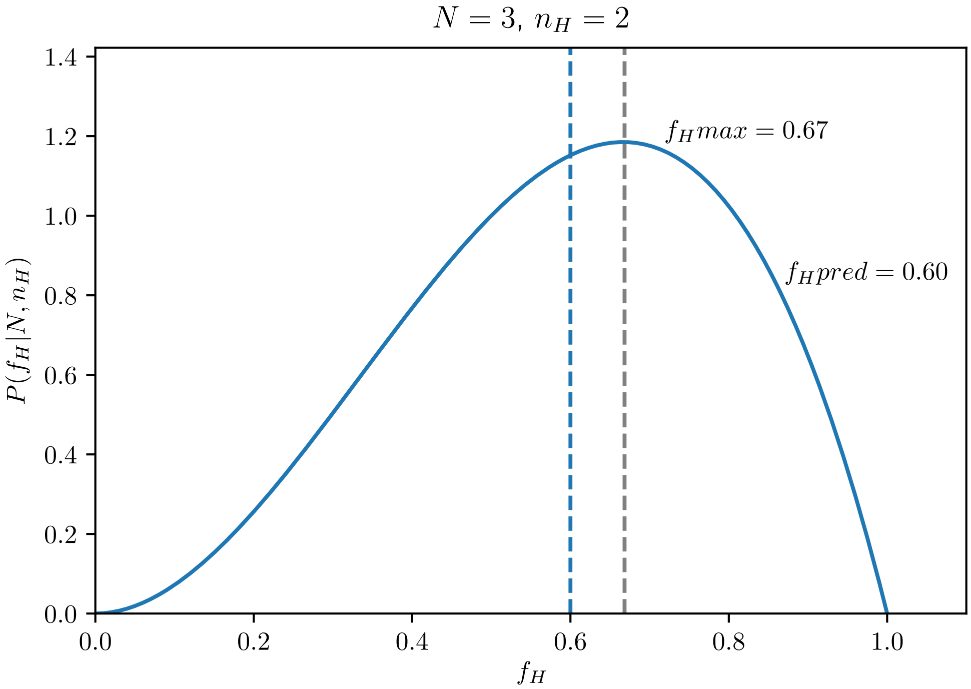
Solution We will continue with notation and assumptions provided in the text. As usual, we start by applying Bayes' theorem as follows, where we let s = **aba**, $F = 3$, and notate our assumptions as $\mathcal{H}_1$.

$$P(p_a|\mathrm{s}, F, \mathcal{H}_1) = \frac{P(\mathrm{s}|p_a, F, \mathcal{H}_1)P(p_a|\mathcal{H}_1)}{P(\mathrm{s}, F)} \tag{34}$$

Interestingly, we notice that this is equivalent to our results in Problem 2.8.

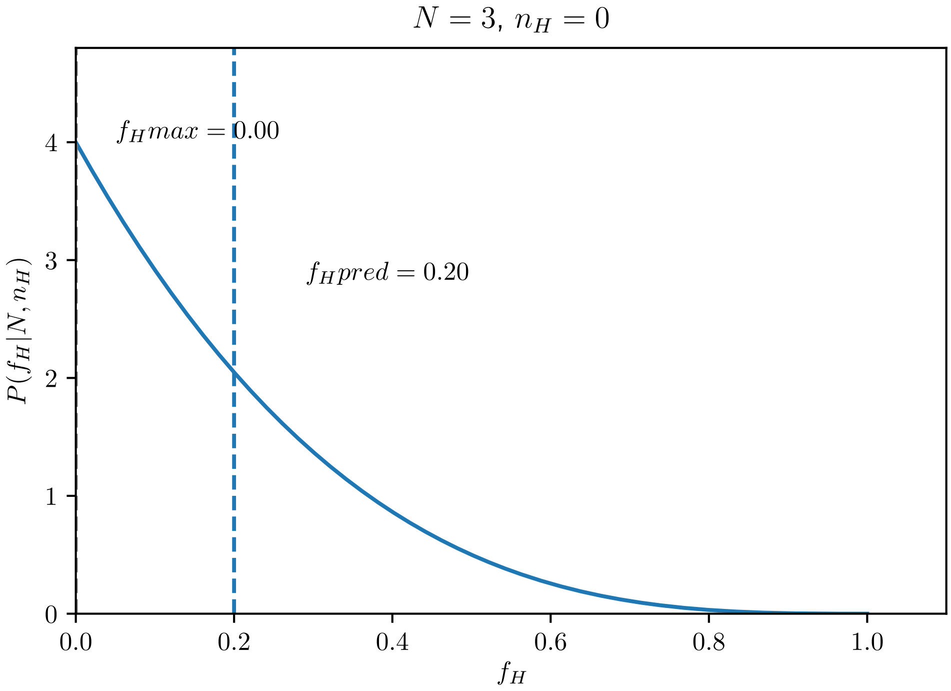
(a) s = **aba** and $F = 3$;

The maximum is given by $p_a = \frac{2}{3}$ and the mean value is given by $p_a = \frac{3}{5}$. (Excuse the labeling from the previous problem in the plot.)



(b) s = **bbb** and $F = 3$;

The maximum is given by $p_a = 0$ and the mean value is given by $p_a = \frac{1}{5}$. (Excuse the labeling from the previous problem in the plot.)



Problem 3.12. A bag contains one counter, known to be either white or black. A white counter is put in, the bag is shaken, and a counter is drawn out, which proves to be white. What is now the chance of drawing a white counter? [Notice that the state of the bag, after the operations, is exactly identical to its state before.]

Solution Again we start with Bayes' theorem and we let C_W be the even that the original counter was white and we let P_W be the event that we pull a white counter on the first try. We then want to find the chance of drawing a white counter after this has occurred, or $P(C_W|P_W)$.

$$P(C_W|P_W) = \frac{P(P_W|C_W)P(C_W)}{P(P_W)} \tag{35}$$

We know that $P(P_W|C_W) = 1$ since if the original counter is white we were ensured to have selected a white counter on the first attempt. Furthermore, we know that the probability of the original counter being white is $P(C_W) = \frac{1}{2}$ since we assume there was an equal chance of it being either white or black. We also know that the probability of selecting a white counter on our first try has two possible routes, giving us

$$P(P_W) = P(C_W)P(P_W|C_W) + P(C_B)P(P_W|C_B) \tag{36}$$

where C_B is the event that the original counter was black.

$$\begin{align} P(P_W) &= P(C_W)P(P_W|C_W) + P(C_B)P(P_W|C_B) \tag{37} \\ &= \frac{1}{2} \times 1 + \frac{1}{2} \times \frac{1}{2} \tag{38} \\ &= \frac{1}{2} + \frac{1}{4} \tag{39} \\ &= \frac{3}{4} \tag{40} \end{align}$$

And putting all of these pieces together yields

$$\begin{align} P(C_W|P_W) &= \frac{P(P_W|C_W)P(C_W)}{P(P_W)} \tag{41} \\ &= \frac{1 \times \frac{1}{2}}{\frac{3}{4}} \tag{42} \\ &= \frac{2}{3} \tag{43} \end{align}$$

Problem 3.14. In a game, two coins are tossed. If either of the coins comes up heads, you have won a prize. To claim the prize, you must point to one of your coins that is a head and say 'look, that coin's a head, I've won'. You watch Fred play the game. He tosses the two coins, and he points to a coin and says 'look, that coin's a head, I've won'. What is the probability that the other coin is a head?

Solution Intuitively we know that each coin flip is independent of the other, therefore we would assume that the probability of the second coin landing on heads is $P(H) = \frac{1}{2}$, but let's investigate.

At the start we determine that there are four possible outcomes for the game: HH, HT, TH, TT. The information provided to us tells us that one of the coins that was flipped landed on heads and therefore there are only 3 possible outcomes: HH, HT, TH. Therefore, we can see that $P(HH|H) = \frac{1}{3}$, and this is in fact a dependent probability.

The key is that the coin that is pointed at as having heads could be either coin 1 or coin 2, we do not know, and in order for pointing to occur either a single coin must be heads (HT and TH), or both coins must be heads (HH), giving us three outcomes, which are all equally likely, given that we know at least one of the coins is heads.

Problem 28.1. Random variables x come independently from a probability distribution $P(x)$. According to model $\mathcal{H}_0$, $P(x)$ is a uniform distribution

$$P(x|\mathcal{H}_0) = \frac{1}{2} \quad x \in (-1, 1). \tag{44}$$

According to model $\mathcal{H}_1$, $P(x)$ is a nonuniform distribution with an unknown parameter $m \in (-1, 1)$

$$P(x|m, \mathcal{H}_1) = \frac{1}{2}(1 + mx) \quad x \in (-1, 1). \tag{45}$$

Given the data $D = \{0.3, 0.5, 0.7, 0.8, 0.9\}$, what is the evidence for $\mathcal{H}_0$ and $\mathcal{H}_1$?

Solution In the text we find that the evidence for a model $\mathcal{H}_i$, given some data D, can be determined from the relationship

$$P(D|\mathcal{H}_i) = \int P(D|\mathbf{w}, \mathcal{H}_i)P(\mathbf{w}|\mathcal{H}_i)d\mathbf{w}. \tag{46}$$

For our first model $\mathcal{H}_0$, we have chosen $P(x|\mathcal{H}_0) = \frac{1}{2}$, for $x \in (-1, 1)$, and we note that there are no free parameters, $\mathbf{w}$, in this model. This means that there is no integral to calculate. Since all values of D are between -1 and 1, they are all equally likely under this model, and all have probability $P(x) = \frac{1}{2}$.

$$P(D|\mathcal{H}_0) = \prod_{i=1}^{N} P(x_i|\mathcal{H}_0) = \left(\frac{1}{2}\right)^5 = \frac{1}{32} = 0.03125 \tag{47}$$

Our calculation of the evidence for the second model, $\mathcal{H}_1$, is slightly more involved since this model includes a single parameter, $m \in (-1, 1)$. We assume that $P(m|\mathcal{H}_1) = \frac{1}{2}$, so that our prior for m is uniformly distributed over the range $(-1, 1)$. We can then solve for $P(D|m, \mathcal{H}_1)$ as a function of our single parameter, m, where x_i is the i-th observation in our dataset D, with N examples. We calculate the evidence of each observation and then take the product.

$$\begin{aligned}
P(D|\mathcal{H}_1) &= \int_{-1}^{1} P(D|m, \mathcal{H}_1)P(m|\mathcal{H}_1)dm && (48)\\
&= \prod_{i=1}^{5} \int_{-1}^{1} P(x_i|m, \mathcal{H}_1)P(m|\mathcal{H}_1)dm && (49)\\
&= \prod_{i=1}^{5} \int_{-1}^{1} \frac{1}{2}(1 + mx) \times \frac{1}{2}dm && (50)\\
&= \prod_{i=1}^{5} \frac{1}{4} \int_{-1}^{1} (1 + mx)dm && (51)\\
&= \prod_{i=1}^{5} \frac{1}{4} \times 2 = \frac{1}{32} = 0.03125 && (52)\\
& && (53)
\end{aligned}$$

So the posterior probability ratio, assuming equal priors on the models, is

$$\frac{P(D|\mathcal{H}_0)P(\mathcal{H}_0)}{P(D|\mathcal{H}_1)P(\mathcal{H}_1)} = \frac{(\frac{1}{2})^5}{(\frac{1}{2})^5} = 1, \tag{54}$$

which means that under the data D, both models provide the same evidence. This is interesting as when we examine the data we notice that our sample potentially exhibits a skewed distribution on (-1, 1), since all of the data points are > 0. From this observation we might assume that $\mathcal{H}_1$ would be able to better explain this data, since it has the ability to represent this kind of skewed distribution, while $\mathcal{H}_0$ does not. But, our results show that the added complexity of $\mathcal{H}_1$ hurts its likelihood of explaining the data. Therefore, in this case we would chose $\mathcal{H}_0$, since it offers a simpler explanation of the data (applying occam's razor).