The Austalian National University Semester 2, 2021
School of Computing Assignment 3 Theory Questions

COMP3670/6670: Introduction to Machine Learning

Errata: All corrections are in red.

Note: For the purposes of this assignment, if X is a random variable we let p_X denote the probability density function (pdf) of X, F_X to denote it's cumulative distribution function, and P to denote probabilities. These can all be related as follows:

$$P(X \leq x) = F_X(x) = \int_{-\infty}^{x} p_X(z)dz$$

$$P(a \leq X \leq b) = F_X(b) - F_X(a) = \int_{a}^{b} p_X(z)dz$$

Often, we will simply write p_X as p, where it's clear what random variable the distribution refers to. You should show your derivations, but **you may use a computer algebra system (CAS)** to assist with integration or differentiation. We are not assessing your ability to integrate/differentiate here.[1].

Question 1 **Continuous Bayesian Inference** 5+5+2+4+4+6+6+5=37 credits

Let X be a random variable representing the outcome of a biased coin with possible outcomes $\mathcal{X} = \{0, 1\}$, $x \in \mathcal{X}$. The bias of the coin is itself controlled by a random variable Θ, with outcomes[2] $\theta \in \boldsymbol{\theta}$, where

$$\boldsymbol{\theta} = \{\theta \in \mathbb{R} : 0 \leq \theta \leq 1\}$$

The two random variables are related by the following conditional probability distribution function of X given Θ.

$$p(X = 1 \mid \Theta = \theta) = \theta$$

$$p(X = 0 \mid \Theta = \theta) = 1 - \theta$$

We can use $p(X = 1 \mid \theta)$ as a shorthand for $p(X = 1 \mid \Theta = \theta)$.

We wish to learn what θ is, based on experiments by flipping the coin.

We flip the coin a number of times.[3] After each coin flip, we update the probability distribution for θ to reflect our new belief of the distribution on θ, based on evidence.

Suppose we flip the coin n times, and obtain the sequence of coin flips [4] $x_{1:n}$.

a) Compute the new PDF for θ after having observed n consecutive **ones** (that is, $x_{1:n}$ is a sequence where $\forall i.x_i = 1$), for an arbitrary prior pdf $p(\theta)$. Simplify your answer as much as possible.

Solution. We are given the prior PDF before any evidence as $p(\theta) = 1$.

$$\begin{aligned} &p(\theta \mid x_{1:n} = 1^n) \\ &= \frac{p(x_{1:n} = 1^n \mid \theta)p(\theta)}{p(x_{1:n} = 1^n)} \\ &= \frac{p(x_{1:n} = 1^n \mid \theta)p(\theta)}{\int_0^1 p(x_{1:n} = 1^n \mid \theta)p(\theta)d\theta} \end{aligned}$$

[1] For example, asserting that $\int_0^1 x^2 \left(x^3 + 2x\right) dx = 2/3$ with no working out is adequate, as you could just plug the integral into Wolfram Alpha using the command `Integrate[x^2(x^3 + 2x),{x,0,1}]`

[2] For example, a value of $\theta = 1$ represents a coin with 1 on both sides. A value of $\theta = 0$ represnts a coin with 0 on both sides, and $\theta = 1/2$ represents a fair, unbaised coin.

[3] The coin flips are independent and identically distributed (i.i.d).

[4] We write $x_{1:n}$ as shorthand for the sequence $x_1x_2\ldots x_n$.

Now, by definition of X, $p(x_1 = 1 \mid \theta) = \theta$, and since the coins are sampled i.i.d, $p(x_{1:n} = 1^n) = p(X = 1)^n = \theta^n$.

$$= \frac{\theta^n p(\theta)}{\int_0^1 \theta^n p(\theta) d\theta}$$

b) Compute the new PDF for θ after having observed n consecutive **zeros**, (that is, $x_{1:n}$ is a sequence where $\forall i. x_i = 0$) for an arbitrary prior pdf $p(\theta)$. Simplify your answer as much as possible.

Solution. We are given the prior PDF before any evidence as $p(\theta) = 1$.

$$\begin{aligned} & p(\theta \mid x_{1:n} = 0^n) \\ &= \frac{p(x_{1:n} = 0^n \mid \theta) p(\theta)}{p(x_{1:n} = 0^n)} \\ &= \frac{p(x_{1:n} = 0^n \mid \theta) p(\theta)}{\int_0^1 p(x_{1:n} = 0^n \mid \theta) p(\theta) d\theta} \end{aligned}$$

Now, by definition of X, $p(x_1 = 0 \mid \theta) = 1 - \theta$, and since the coins are sampled i.i.d, $p(x_{1:n} = 0^n) = p(X = 0)^n = (1 - \theta)^n$.

$$= \frac{(1-\theta)^n p(\theta)}{\int_0^1 (1-\theta)^n p(\theta) d\theta}$$

c) Compute $p(\theta | x_{1:n} = 1^n)$ for the uniform prior $p(\theta) = 1$.

Solution. Using the previous answer,

$$p(\theta \mid x_{1:n} = 1^n) = \frac{\theta^n p(\theta)}{\int_0^1 \theta^n p(\theta) d\theta} = \frac{\theta^n}{\int_0^1 \theta^n d\theta} = (n+1)\theta^n$$

d) Compute the expected value μ_n of θ after observing n consecutive ones, with a uniform prior $p(\theta) = 1$. Provide intuition explaining the behaviour of μ_n as $n \to \infty$.

Solution. The expectation value of a continuous probability distribution $f(\theta)$ on the domain $[0, 1]$ is given by

$$\mu = \int_0^1 \theta f(\theta) d\theta$$

Hence,

$$\mu_n = \int_0^1 \theta p(\theta | X_{1:n} = 1^n) d\theta = (n+1) \int_0^1 \theta^{n+1} d\theta = \frac{n+1}{n+2} \to 1$$

As we observe an increasingly large number of consecutive ones, the model becomes more and more biased towards 1, in the limit placing all the probability mass on 1 and 0 everywhere else. Hence, the mean should also converge to 1.

e) Compute the variance σ_n^2 of the distribution of θ after observing n consecutive ones, with a uniform prior $p(\theta) = 1$. Provide intuition explaining the behaviour of σ_n^2 as $n \to \infty$.

Solution. The variance of a continuous probability distribution $f(\theta)$ on the domain $[0, 1]$ is given by

$$\sigma^2 = \int_0^1 (\theta - \mu)^2 f(\theta) d\theta$$

Hence,

$$\sigma_n^2 = \int_0^1 (\theta - \mu_n)^2 p(\theta | X_{1:n} = 1^n) d\theta$$
$$= \int_0^1 \left(\theta - \frac{n+1}{n+2}\right)^2 (n+1)\theta^n d\theta = \frac{n+1}{(n+2)^2(n+3)} \to 0$$

As we observe more and more consecutive ones in a row, the model becomes more and more confident that the coin can only generate ones. So, the posterior becomes more narrow, and the variance goes to zero.

f) Compute the *maximum a posteriori* estimation θ_{MAP_n} of the distribution on θ after observing n consecutive ones, with a uniform prior $p(\theta) = 1$. Provide intuition explaining how θ_{MAP_n} varies with n.

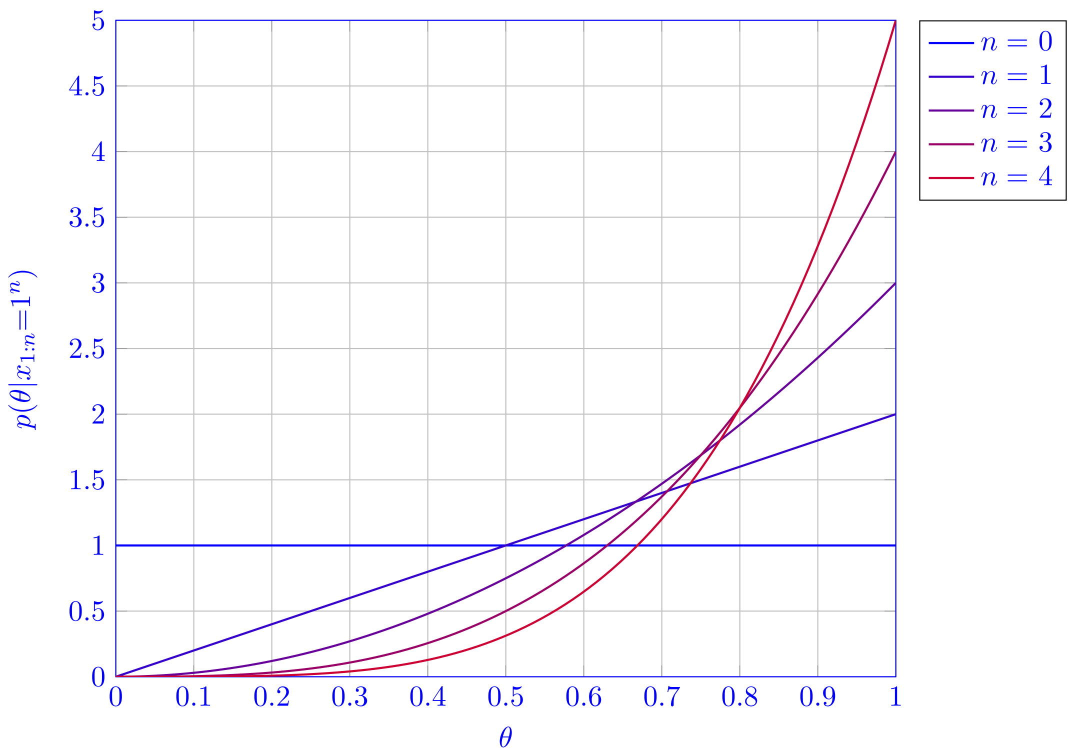
Solution. $p(\theta \mid x_{1:n} = 1^n) = (n+1)\theta^n$ is a monotonically increasing function, so $\theta_{MAP_n} = 1$, so θ_{MAP_n} in fact does not depend on n. This is because since we have only observed nothing but consecutive ones, the coin most likely to generate this sequence is a coin with 1 on both sides (i.e. a coin with $\theta = 1$.)

g) Given we have observed n consecutive coin flips of ones in a row, what do you think would be a better choice for the best guess of the true value of θ? μ_n or θ_{MAP}? Justify your answer. (Assume $p(\theta) = 1$.)

Solution. Even after a single coinflip, θ_{MAP} always declares that the coin is deterministic, with $\theta = 1$. The mean μ_n takes a more conservative approach, converging closer to 1 as the number of observed ones tends to infinity.

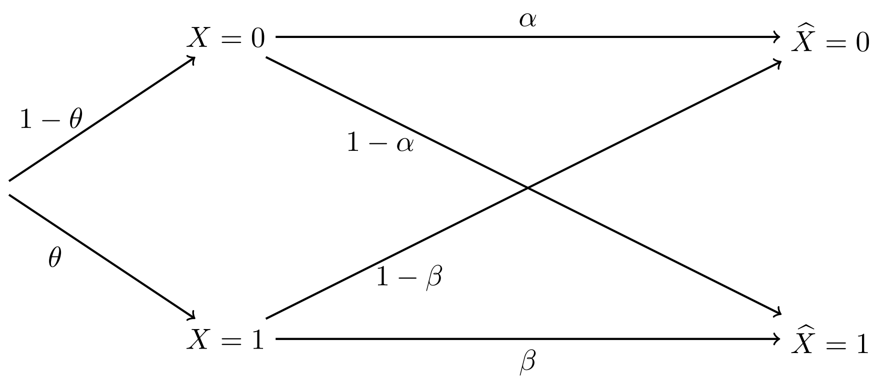
h) Plot the probability distributions $p(\theta | x_{1:n} = 1)$ over the interval $0 \le \theta \le 1$ for $n \in \{0, 1, 2, 3, 4\}$ to compare them. Assume $p(\theta) = 1$.

Solution.



Question 2 **Bayesian Inference on Imperfect Information** (4+5+8+4+4=25 credits)

We have a Bayesian agent running on a computer, trying to learn information about what the parameter θ could be in the coin flip problem, based on observations through a noisy camera. The noisy camera takes a photo of each coin flip and reports back if the result was a 0 or a 1. Unfortunately, the camera is not perfect, and sometimes reports the wrong value.[5] The probability that the camera makes mistakes is controlled by two parameters α and β, that control the likelihood of correctly reporting a zero, and a one, respectively. Letting X denote the true outcome of the coin, and $\widehat{X}$ denoting what the camera reported back, we can draw the relationship between X and $\widehat{X}$ as shown.



So, we have

$$\begin{aligned} p(\widehat{X} = 0 \mid X = 0) &= \alpha \\ p(\widehat{X} = 0 \mid X = 1) &= 1 - \beta \\ p(\widehat{X} = 1 \mid X = 1) &= \beta \\ p(\widehat{X} = 1 \mid X = 0) &= 1 - \alpha \end{aligned}$$

We would now like to investigate what posterior distributions are obtained, as a function of the parameters α and β.

a) (5 credits) Briefly comment about how the camera behaves for $\alpha = \beta = 1$, for $\alpha = \beta = 1/2$, and for $\alpha = \beta = 0$. For each of these cases, how would you expect this would change how the agent updates it's prior to a posterior on θ, given an observation of $\widehat{X}$? (No equations required.) You shouldn't need any assumptions about $p(\theta)$ for this question.

Solution. For $\alpha = \beta = 1$, the camera always returns a perfect description of the coin, so we would expect the agent to update on observations of $\hat{X}$ in exactly the same way as they would for X. For $\alpha = \beta = 1/2$, the camera returns 0 half the time, regardless of what the actual value of X was. So the agent receives no information from the camera, and will not change their beliefs (the distribution will stay as the prior) based on camera observations. For $\alpha = \beta = 0$, the camera always returns the wrong result, so updating on $\hat{X} = 0$ is the same as updating on $X = 1$, and vice versa.

b) (10 credits) Compute $p(\widehat{X} = x|\theta)$ for all $x \in \{0, 1\}$.

Solution.

$$\begin{aligned} p(\hat{X} = 0|\theta) &= p(\hat{X} = 0|X = 0, \theta)p(X = 0|\theta) + p(\hat{X} = 0|X = 1, \theta)p(X = 1|\theta) \\ &= \alpha(1 - \theta) + (1 - \beta)\theta \end{aligned}$$

$$\begin{aligned} p(\hat{X} = 1|\theta) &= p(\hat{X} = 1|X = 0, \theta)p(X = 0|\theta) + p(\hat{X} = 1|X = 1, \theta)p(X = 1|\theta) \\ &= (1 - \alpha)(1 - \theta) + \beta\theta \end{aligned}$$

[5]The errors made by the camera are i.i.d, in that past camera outputs do not affect future camera outputs.

c) (15 credits) The coin is flipped, and the camera reports seeing a one. (i.e. that $\hat{X} = 1$.) Given an arbitrary prior $p(\theta)$, compute the posterior $p(\theta|\hat{X} = 1)$. What does $p(\theta|\hat{X} = 1)$ simplify to when $\alpha = \beta = 1$? When $\alpha = \beta = 1/2$? When $\alpha = \beta = 0$? Explain your observations.

Solution.

$$\begin{aligned} & p(\theta \mid \hat{X} = 1) \\ &= \frac{p(\hat{X} = 1 \mid \theta)p(\theta)}{p(\hat{X} = 1)} \\ &= \frac{p(\hat{X} = 1 \mid \theta)p(\theta)}{\int_0^1 p(\hat{X} = 1 \mid \theta)p(\theta)d\theta} \\ &= \frac{((1-\alpha)(1-\theta) + \beta\theta)p(\theta)}{\int_0^1 ((1-\alpha)(1-\theta) + \beta\theta)p(\theta)d\theta} \end{aligned}$$

(a) Now, if $\alpha = \beta = 1$, then the above collapses to

$$\frac{(0(1-\theta) + 1\theta)p(\theta)}{\int_0^1 (0(1-\theta) + 1\theta)p(\theta)d\theta} = \frac{\theta p(\theta)}{\int_0^1 \theta p(\theta)d\theta} = P(\theta|X = 0)$$

as the camera never makes any error, so we would expect that the true value of the coin provides the same evidence as what the camera claims the coin is, so this is why we get

$$p(\theta \mid X = 0) = p(\theta \mid \hat{X} = 0).$$

(b) If $\alpha = \beta = 1/2$, then the above collapses to

$$\frac{(1/2(1-\theta) + 1/2\theta)p(\theta)}{\int_0^1 (1/2(1-\theta) + 1/2\theta)p(\theta)d\theta} = \frac{p(\theta)}{\int_0^1 p(\theta)d\theta} = p(\theta)$$

Since $\alpha = \beta = 1/2$, then $p(\hat{X} = 0|X = 0) = p(\hat{X} = 0|X = 1)$, so the camera is just as likely to return a zero whether or not the coin actually was zero. This means that viewing the output of the camera provides no evidence at all for what the true value of the coin is, and so the agent does not change their belief at all based on camera data, so this is why we get

$$p(\theta \mid \hat{X} = 1) = p(\theta).$$

(c) If $\alpha = \beta = 0$, then the above collapses to

$$\frac{((1-0)(1-\theta) + 0\theta)p(\theta)}{\int_0^1 ((1-0)(1-\theta) + 0\theta)p(\theta)d\theta} = \frac{(1-\theta)p(\theta)}{\int_0^1 (1-\theta)p(\theta)d\theta} = p(\theta|X = 0)$$

Since $\alpha = \beta = 0$, the camera always reports back the opposite result with no error, so if the camera reports $\hat{X} = 1$, then it must have been the case that $X = 0$, hence we obtain

$$p(\theta|\hat{X} = 1) = p(\theta|X = 0)$$

d) Compute $p(\theta|\hat{X} = 1)$ for the uniform prior $p(\theta) = 1$. Simplify it under the assumption that $\beta := \alpha$.

Solution. From before, we obtain

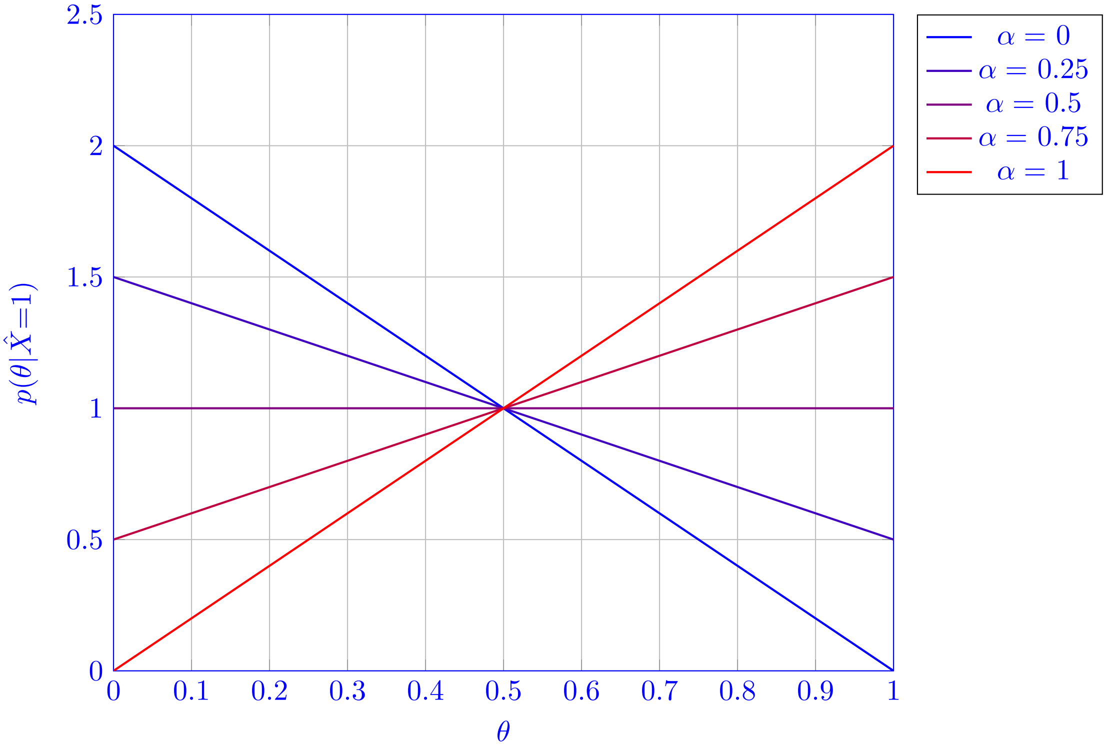
$$p(\theta|\hat{X} = 1) = \frac{((1-\alpha)(1-\theta) + \beta\theta)}{\int_0^1 ((1-\alpha)(1-\theta) + \beta\theta)d\theta} = 2\frac{(1-\alpha)(1-\theta) + \beta\theta}{1 - \alpha + \beta}$$

Then, replacing β with α, we obtain

$$p(\theta|\hat{X} = 1) = 2\left((1-\alpha) + (2\alpha - 1)\theta\right)$$

e) (10 credits) Let $\beta = \alpha$. Plot $p(\theta|\hat{X} = 1)$ as a function of θ, for all $\alpha \in \{0, \frac{1}{4}, \frac{2}{4}, \frac{3}{4}, 1\}$ on the same graph to compare them. Comment on how the shape of the distribution changes with α. Explain your observations. (Assume $p(\theta) = 1$.)

Solution.



For $\alpha = 1$, the posterior distribution updates in the same way as if the coin has been observed directly. As α decreases, the distribution begins to skew such that lower values of θ are more likely, as a decrease in α implies a camera more likely to misread a 1, which means it's less likely that if the camera reported a zero, the true value of the coin is zero. When $\alpha = 1/2$, the posterior matches that of the prior, as the camera provides no information. Beyond that, as α moves towards zero, the camera more than half the time gives the wrong answer, so having $\hat{X} = 1$ is actually evidence in favour of $X = 0$, so the posterior places more weight on lower values of θ. And when $\alpha = 0$, we recover the same posterior as for $\alpha = 1$, but flipped around (as if we had updated on $X = 0$.)

Question 3 **Relating Random Variables** (10+7+5+16=38 credits)

A casino offers a new game. Let $X \sim f_X$ be a random variable on $(0, 1]$ with pdf p_X. Let Y be a random variable on $[1, \infty)$ such that $Y = 1/X$. A random number c is sampled from Y, and the player guesses a number $m \in [1, \infty)$. If the player's guess m was lower than c, then the player wins $m - 1$ dollars from the casino (which means higher guesses pay out more money). But if the player guessed too high, $(m \geq c)$, they go bust, and have to pay the casino 1 dollar.

a) Show that the probability density function p_Y for Y is given by

$$p_Y(y) = \frac{1}{y^2} p_X(\frac{1}{y})$$

Solution. We compute the CDF function of Y,

$$
\begin{aligned}
&F_Y(y) \\
&= P(Y \leq y) \\
&= P(1/X \leq y) \\
&= P(X \geq 1/y) \\
&= 1 - P(X \leq 1/y) \\
&= 1 - F_X(1/y)
\end{aligned}
$$

Then, to obtain the pdf for Y, we can differentiate $F_Y(y)$.

$$p_Y(y) = F_Y'(y) = (1 - F_X(1/y))' = -F_X'(1/y) \times \frac{-1}{y^2} = \frac{1}{y^2} p_X(1/y)$$

b) Hence, or otherwise, compute the expected profit for the player under this game. Your answer will be in terms of m and p_X, and should be as simplified as possible.

Solution. The expected return is given by $P(\text{player win})(m-1) + P(\text{player loss})(-1)$.

$$
\begin{aligned}
&P(\text{player win})(m-1) + P(\text{player loss})(-1) \\
&= P(Y > m)(m-1) - P(Y \leq m) \\
&= (1 - P(Y \leq m))(m-1) - P(Y \leq m) \\
&= m - 1 - mP(Y \leq m) \\
&= m - 1 - m \int_1^m p_Y(y)\, dy \\
&= m - 1 - m \int_1^m \frac{1}{y^2} p_X(1/y)\, dy
\end{aligned}
$$

c) Suppose the casino chooses a uniform distribution over $(0, 1]$ for X, that is,

$$p_X(x) = \begin{cases} 1 & 0 < x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

What strategy should the player use to maximise their expected profit?

Solution. We use the previous question, using $p_X(1/y) = 1$, noting that for all $y \geq 1$, $0 < 1/y \leq 1$.

$$
\begin{aligned}
&m - 1 - m \int_1^m \frac{1}{y^2} p_X(1/y)\, dy \\
&= m - 1 - m \int_1^m \frac{1}{y^2}\, dy \\
&= m - 1 - m(1 - \frac{1}{m}) \\
&= 0
\end{aligned}
$$

So, regardless of the strategy (the players choice of m) the game will return on expectation zero profit, making it a fair game.

d) Find a pdf $p_X : (0, 1] \to \mathbb{R}$ such that for any $B > 0$, there exists a corresponding player guess m such that the expected profit for the player is at least B. (That is, prove that the expected profit for p_X, as a function of m, is unbounded.)

Make sure that your choice for p_X is a valid pdf, i.e. it should satisfy

$$\int_0^1 p_X(x)dx = 1 \text{ and } p_X(x) \geq 0$$

You should also briefly mention how you came up with your choice for p_X.

Hint: We want X to be extremely biased towards small values, so that Y is likely to be large, and the player can choose higher values of m without going bust.

Solution. We want the casino to use a p_X that is biased towards low numbers, so X is often small, so Y is often big, and we can make large guesses for m to make the most money. With that in mind, we choose $p_X(x) = \frac{1}{2\sqrt{x}}$. It is clear that $\frac{1}{2\sqrt{x}} \geq 0$ for all $x \in (0, 1]$, and one can verify that $\int_0^1 \frac{1}{2\sqrt{x}} dx = 1$. Then, the expected profit is

$$\begin{aligned}
&m - 1 - m \int_1^m \frac{1}{y^2} p_X(1/y)\, dy \\
&= m - 1 - m \int_1^m \frac{1}{y^2} \frac{1}{2} \sqrt{y}\, dy \\
&= m - 1 - m(1 - \frac{1}{\sqrt{m}})\, dy \\
&= \sqrt{m} - 1
\end{aligned}$$

Then, by choosing any $m \geq (B + 1)^2$ we have

$$\text{profit } = \sqrt{m} - 1 \geq \sqrt{(B + 1)^2} - 1 = B$$

which ensures an expected profit of at least B. Since B was arbitrary, the expected profit is unbounded.

Addendum: How did I know to chose $p_X(x) = \frac{1}{2\sqrt{x}}$? Well, one guess is to try x^n, and then work out an appropriate value for n, if one exists. First, we have to normalise it. We note that $\int_0^1 x^n dx = \frac{1}{n+1}$, for any $n > -1$. (For $n \leq -1$, the integral diverges.) So, we can choose $p_X(x) = (n + 1)x^n$, where $n > -1$. Then, we compute the expectation as before,

$$\begin{aligned}
&m - 1 - m \int_1^m \frac{1}{y^2} p_X(1/y)\, dy \\
&= m - 1 - m \int_1^m \frac{1}{y^2} (n + 1)(1/y)^n\, dy \\
&= m - 1 - m(n + 1) \int_1^m \frac{1}{y^{n+2}}\, dy \\
&= m - 1 - m(n + 1) \left(\frac{1 - m^{-n-1}}{n + 1} \right) \\
&= m - 1 - m(1 - m^{-n-1}) \\
&= \frac{1}{m^n} - 1
\end{aligned}$$

We now need to chose n such that $n > -1$ and $\frac{1}{m^n} - 1$ is unbounded. We can achieve this by choosing $n = -1/2$, and obtain $\sqrt{m} - 1$, as before.