Part 1 - Problems

Årne, Runar and Lars

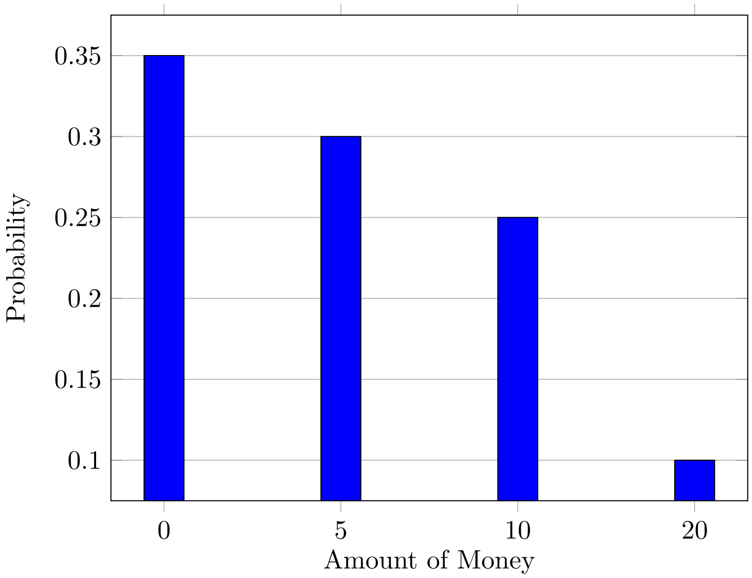
February 2, 2024

1 Foundations

Exercise 1

Consider interacting with a slot-machine which returns the following amount of money R: with probability 0.35 it returns nothing, with probability 0.3 it returns 5$; with probability 0.25 it returns 10$, with probability 0.1 it returns 20$.

1. Plot the distribution of P(R)?



2. Compute E[R] and Var[R]?

The expected value is calculated as:

$$E[X] = \sum_i x_i \cdot f(x_i)$$

Substituting the given values:

$$E[R] = 0.35 \cdot 0 + 0.3 \cdot 5 + 0.25 \cdot 10 + 0.1 \cdot 20 = 6$$

To calculate the variance we can use this shortcut formula:

$$Var[X] = E[X^2] - E[X]^2$$

Substituting the given values:

$$Var[R] = (0.35 \cdot 0^2 + 0.3 \cdot 5^2 + 0.25 \cdot 10^2 + 0.1 \cdot 20^2) - 6^2 = 36.5$$

3. What is the probability P($R \geq 10\$$)?

$$P(R \geq 10\$) = P(R = 10\$) + P(R = 20\$)$$
$$P(R \geq 10\$) = 0.25 + 0.1 = 0.35$$

4. What is the probability P($R > 10\$$)?

$$P(R \geq 10\$) = P(R = 20\$) = 0.1$$

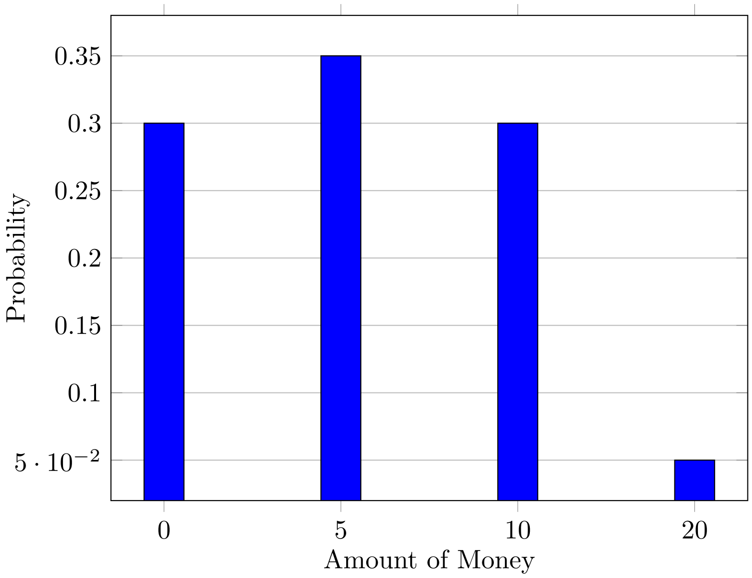
5. If I play the slot-machine twice, what is the probability that the total reward is greater than 15$?

Playing the slot-machine twice and getting a reward greater than 15 yields the following outcomes: (10, 10),(20, 20), (20, 0), (20, 5), (20, 10), (0, 20), (5, 20), (10, 20)

$$P = 0.25^2 + 0.1^2 + 2 \cdot (0.1 \cdot 0.35 + 0.1 \cdot 0.3 + 0.1 \cdot 0.25) = 0.2525$$

Consider now interacting also with a second slot-machine which returns the following amount of money S: with probability 0.3 it returns nothing; with probability 0.35 it returns 5$; with probability 0.3 it returns 10$; with probability 0.05 it returns 20$.

6. Plot the distribution of P(S)?



7. If you were to choose which slot-machine to play, which one would you choose and why?

If we were to play one of the two slot-machine we would like the highest expected value of money. We can calculate E[S] and compare this to the other machine.

$$E[S] = 0.3 \cdot 0 + 0.35 \cdot 5 + 0.3 \cdot 10 + 0.05 \cdot 20 = 5.75$$

Seeing as the first machine has a higher expected value than the newly introduced one, we would like to play the first machine.

8. What is the distribution of P(R, S)?

Table 1: Joint Probability Distribution $P(R, S)$

	R=0	**R=5**	**R=10**	**R=20**	**P(S)**
S=0	0.105	0.09	0.075	0.03	0.3
S=5	0.1225	0.105	0.0875	0.035	0.35
S=10	0.105	0.09	0.075	0.03	0.3
S=20	0.0175	0.015	0.0125	0.005	0.05
P(R)	0.35	0.3	0.25	0.1	

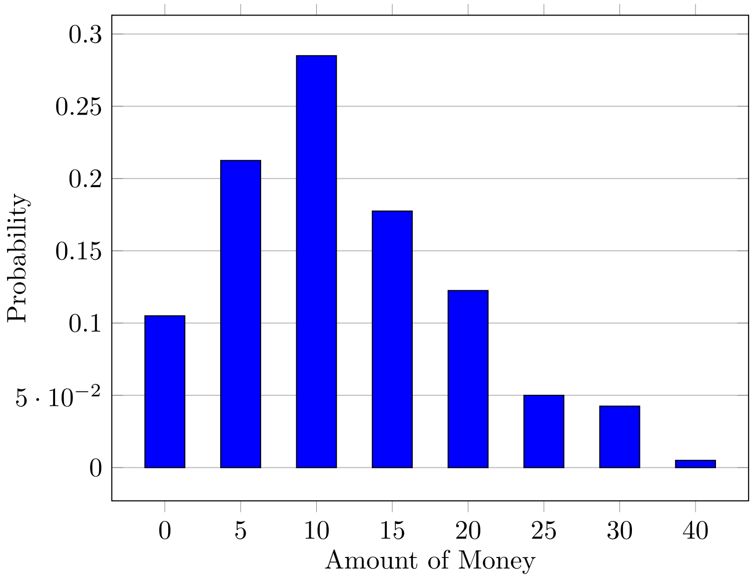
9. What is the distribution of P(S, R)? Does it differ from P(R, S)?

The joint probability distribution P(S,R) is identical to P(R,S), as the order in which the random variables are listed does not alter the assigned probabilities.

10. (*) If I play both slot machine, what is the distribution of the sum of the rewards? How do you express it formally, and how does it differ from P(R, S)?

To compute the distribution of the sum we take each possible sum of rewards and add up the probabilities for the events that add up to that sum. Ex.

$P(R+S = 20) = P(R = 10, S = 10)+P(R = 20, S = 0)+P(R = 0, S = 0) = 0.1225$

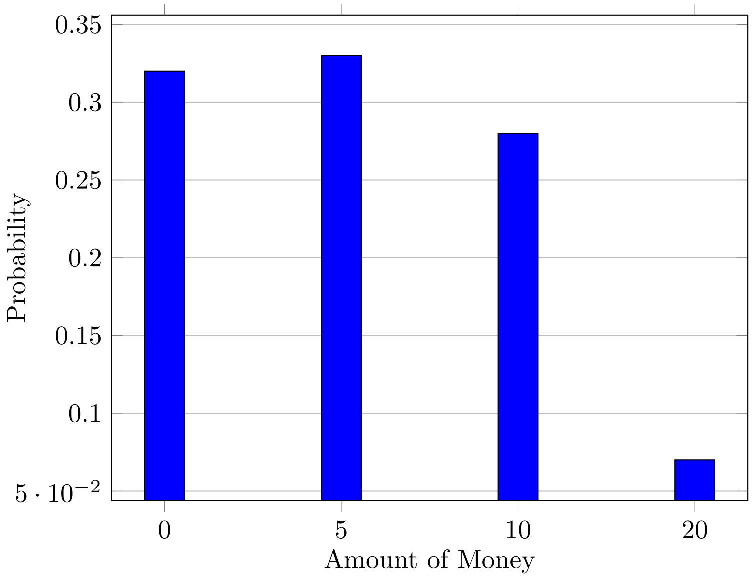


The main difference compared to P(R,S) is that P(R,S) treats P(R = 10, S = 0) and P(R = 0, S = 10) as separate outcomes with distinct probabilities while P(R + S) treats these events as the same and combines their probabilities.

11. If I play the first slot machine with probability 0.4 and the second slot machine with probability 0.6, what is the distribution of my reward? How do you express it formally, and how does it differ from P (R, S)?

To find the distribution of our reward Y we add up the probabilities for each machine weighted by the probability of selecting said machine. Ex.

$$P(Y = 0) = 0.4 \cdot P(R = 0) + 0.6 \cdot P(S = 0) = 0.32$$



12. If I play both slot machines, and I receive 10$ in outcome from the first one, what is the probability that after playing the second slot machine I win less than 10$? How do you express it formally?

This situation can be expressed as

$$P(R + S \leq 10|R = 10)$$

. Since R = 10 then

$$P(R + S \leq 10|R = 10) = 0$$

.

13. If I play both slot machines, and I receive 10$ in outcome from the first one, what is the probability that after playing the second slot machine I win at least 10$? How do you express it formally?

This situation can be expressed as

$$P(R + S \geq 10|R = 10$$

Since R = 10 then

$$P(R + S \geq 10|R = 10) = 0.3 + 0.35 + 0.3 + 0.05 = 1$$

14. (*) If I play both slot machines, and I receive 10$ in outcome from the first one, what is the probability that after playing the second slot machine I win more than 20$? How do you express it formally?

This can be expressed as

$$P(R+S>20|R=10)$$

Since the only possible outcome that adds up to be greater than 20 is S = 20 then the probability is

$$P(R+S>20|R=10)=P(S=20)=0.05$$

Exercise 2

Bayes' formula is an important identity in probability and statistics, stating:

$$P(X|Y)=\frac{P(Y|X)\cdot P(X)}{P(Y)}$$

1. (*) Using the two probability laws presented in the class, prove that Bayes' formula reduces to an identity.

Sum rule:

$$P(X)=\sum_{\Omega_Y}P(X,Y)$$

Product rule:

$$P(X,Y)=P(X|Y)\cdot P(Y)$$

From the product rule we can derive that

$$P(X|Y)=\frac{P(X,Y)}{P(Y)}$$

Since P(X,Y) = P(Y,X) then

$$P(X|Y)=\frac{P(Y,X)}{P(Y)}$$

Once again using the product rule

$$P(X|Y)=\frac{P(Y|X)\cdot P(X)}{P(Y)}$$

Exercise 3

Assume someone played one of the two slot machines in the first exercise above, and reported the following list of ten rewards Di = [10, 0, 5, 10, 5, 0, 0, 10, 0, 0].

1. Estimate Ê[D]?

We can calculate this using the following formula:

$$\hat{E}[R|a_i] = \frac{1}{N_i} \sum_{t|a^t=a_i} r^t$$

Substituting the given values:

$$\hat{E}[D|a_i] = \frac{1}{10} \cdot 10 + 0 + 5 + 10 + 5 + 0 + 0 + 10 + 0 + 0 = 4$$

2. Which of the two slot machines do you expect the person was playing?

Since the value 4 is closer to the expectation value of machine 2, we assume that the person was playing that machine.

The same person plays more the same machine, and provide you with additional ten samples Dj = 10, 20, 0, 5, 10, 5, 10, 0, 20, 10.

3. Update your estimate Eˆ[D] using the twenty sample.

$$\hat{E}[D|a_i] = \frac{1}{20} \cdot 10+0+5+10+5+0+0+10+0+0+10+20+0+5+10+5+10+0+20+10 = 6.5$$

4. (*) Does your opinion on which of the two slot machines the person was playing change?

Seeing as we now have gotten more samples, we can better estimate the true expected value. Now we assume that the person was rather playing machine 1, since 6.5 is closer to its expected value of 6.

5. (*) If the person plays once more (21st sample), do you have particular expectations on the reward that will be produced?

Since the expectation value is 6.5 for the 20 samples played, we expect the 21st sample to produce a reward of 6.5.

2 Basics of Reinforcement Learning

Exercise 4

Consider the following scenario. A company wants to develop an automatic floorsweeping robot to be able to effectively clean house rooms. The idea is to train such a robot with reinforcement learning within the premises of the company, and you have been put in charge of the project.

1. Do you consider the project sensible?

Yes, this does indeed seem like a sensible project.

2. (*) How would you define the main elements of your RL project (states, actions, rewards)?

States are the different situations or configurations the robot can find itself in within the environment. For example, a state might include information about the current location of the robot, the dirtiness level of the floor, and any obstacles in the environment.

Actions are the possible moves or decisions that the robot can take in each state. These actions might include moving forward, turning, stopping, or activating the cleaning mechanism.

Rewards are numerical values that the robot receives as feedback for its actions in a given state. The goal is to design the reward system in a way that encourages the robot to learn optimal behavior. This might include assigning positive rewards for cleaning dirty areas, negative rewards for collisions or inefficiencies, and possibly a higher reward for completing the cleaning task quickly.

3. Is this a planning, prediction, or control problem?

Since the robot in our projects objective is to make decisions in real-time to effectively clean rooms. The robot learns from the consequences of its actions and adjusts its behavior over time. This means that its a control problem because we are trying to find an optimal policy π^*.

After satisfactorily training your robot, the management gets its hand on the robot of another firm and it asks you to compare your prototype against the competitor.

4. Do you have now a planning, prediction, or control problem?

We now have a prediction problem, because we are given two different policies and we want to estimate which one is better.

5. What metric would you use to compare the robots?

This is really problem specific, but we could use the likes of speed, cleanliness or carefulness to compare the two robots.

6. (*) The competing robot seems to perform better, although it often hits tables and bookshelves, causing papers and mugs to crash on the floor. Do you think this might have something to do with the training of the robot? How would you explain its behaviour?

This could possibly happen if the robot have been trained in a room with no furniture or that its main objective is to clean the fastest, without regarding bumping into things while its moving fast enough to knock things down.

The management is concerned with your prototype being competitive enough, and so it considers hitting the less crowded market of factories instead of homes. Your trained robot will be deployed on factory floors instead of house floors.

7. (*) Do you think this choice sensible from a technical point of view? Is there any issue you think worth discussing with the management?

No, we believe that this may not be a sensible choice. The dynamic environment of a factory floor, with constant movements and changes, differ from the static nature of a house floor. Training the robot in an environment that is mostly static might make it difficult for the robot to achieve satisfactory results in a more dynamic environment.

Exercise 5

Consider the following scenario. A company is developing a software for streamlining the hiring process, and it decided to use RL to develop an artificial interviewing agent to optimize the whole process.

1. (*) Do you consider the choice of RL sensible? Discuss.

Whether the choice of RL is sensible or not really depends on the approach rather than as a whole. Its quality also hinges on the correctness of our initial assumptions.

The different properties of reinforcement learning include a delayed feedback signal, a time dimension and an environment. The delayed feedback signal is a property which could be beneficial in a hiring process (as we generally do not know how good a certain candidate is until after we've hired them).

The time dimension and environment's usefulness really instead depends on how good the specific representations are. This hold for the choice of algorithm as well, as some algorithms could fit the problem very well given the initial assumptions (while others may not).

In conclusion, given the diffuse choice of "using RL to develop an artificial interviewing agent" it will be hard to decide its sensibility.

3 Multi-armed Bandits

Exercise 6

Consider the following scenario. Facing a new epidemics, the government is considering the deployment of three different vaccines. However, the success rate of these vaccines is still uncertain. Officials of the government heard that multi-armed bandits are an effective tool to design the deployment of new medicines, and so they consider adopting the same model.

1. (*) Do you consider the choice of MABs sensible? Discuss.

In contrast to Exercise 5, this exercise's choice is clearly sensible! The multi-armed bandit model is an effective tool at balancing "exploration" and "exploitation" in such a way that given enough runtime, we (hopefully) eventually converge at the (or a) optimal policy.

If we represent each vaccine as an action, and the rewards as how effective a given vaccine is. Then the expected rewards estimated within the MAB models will represent an estimation on the effectiveness of each vaccine.

Selecting the action with the highest expected reward would then be equivalent to choosing the vaccine with the highest expected success rate.

In conclusion, this choice is very sensible and is a very typical problem solved by multi-armed bandit models.

Exercise 7

You are tossing a biased coin C, and collected 10 samples such that Eˆ[C] = 0.5.

1. (*) Use Hoeffding's Inequality to evaluate the probability you are underestimating the true expected value E[C] by 0.01, 0.1 and 0.5.

Hoeffding's Inequality is this:

$$P(E[X] > \hat{E}^{(t)}[X] + u) \leq e^{-2tu^2}$$

Tossing a biased coin C for 10 samples. We have that $\hat{E}[C] = 0.5$.

Using Hoeffding's Inequality we can evaluate the probability that we are underestimating the true expected value $E[C]$ by 0.01, 0.1 and 0.5 like this:

$$P(E[C] > \hat{E}^{(10)}[C] + 0.01) = P(E[C] > 0.51) \leq e^{-2 \cdot 10 \cdot (0.01)^2} \approx 0.9980$$

$$P(E[C] > \hat{E}^{(10)}[C] + 0.1) = P(E[C] > 0.6) \leq e^{-2 \cdot 10 \cdot (0.1)^2} \approx 0.8187$$

$$P(E[C] > \hat{E}^{(10)}[C] + 0.5) = P(E[C] > 1.0) \leq e^{-2 \cdot 10 \cdot (0.5)^2} \approx 0.0067$$

Therefore we can conclude that we are probably underestimating the true expected value $E[C]$ by somewhere between 0.1 and 0.5.