CMPUT 365: Introduction to Reinforcement Learning, Winter 2023
Worksheet #1: Bandits

Manuscript version: *#9dd9e9-dirty* - 2023-04-05 12:37:31-06:00

Question 1. Suppose a game where you choose to flip one of two (possibly unfair) coins.[1] You win \$1 if your chosen coin shows heads and lose \$1 if it shows tails. Let the bias (probability that the toss comes up as head) of the first coin be p_1 and the bias of the second coin be p_2.

(a) Model this as a k-armed stochastic bandit problem: Specify the action set and the reward distributions for each of the actions. Choose the rewards so that the total reward achieved gives how much you earned in dollars.

(b) You do not know the coin flip probabilities. Instead, you are able to view 6 sample flips for each coin respectively: (T, H, H, T, T, T) and (H, T, H, H, H, T). Use the sample average formula (Equation 2.1 in the book) to compute the estimates of the value of each action.

(c) Decide on which coin to flip next! Assume it's an exploit step.

Note An informal language that people use when talking about that the reward is a function of the action chosen is that the reward is a **stochastic function** of the action. This is quite descriptive, but a mathematician would not know what you mean by it! Intuitively, a stochastic function f is one that it maps one set to another one in such a way that in the computation a random number that is freshly drawn from some distribution can also be used. Formally, it is mapping the set of inputs to distributions over the outputs. In probability theory, one would call f a probability kernel (mathematicians love the word kernel, whose meaning is the unhelpful "the central, most important part of everything" according to the Oxford Dictionary of English). For f to qualify as a stochastic kernel a little more is needed if the set of inputs is very large (more than countably infinite), but in all examples we will work with the extra conditions are met, so we will just ignore this.

---

[1] Funny note; even if you load a dice and you catch it in the air, after a vigorous toss, and you are not a magician, the outcome will be close to fair! You need to let the dice come down to the ground for the loading to matter (and this is not obvious), or you need to spin it (it will tend to land heavy side up). And of course there is a paper about this important discovery. However entertaining this paper is, a more closer look reveals that coin tosses caught in hand with even an *unloaded*, standard coin, tend to have a tiny bias: For "natural flips", the probability of the coin coming up as started is around 0.51 (see here). There you go gamblers, first useful real-life lesson of the class!? But how long do you need to play with your naive fellows if you were so mean to take advantage of them? Next exercise!?

Question 2. (*Exercise 2.2 S&B*) Consider a k-armed bandit problem with $k = 4$ actions, denoted $1, 2, 3$, and 4. Consider applying to this problem a bandit algorithm using ϵ-greedy action selection, using sample-average action-value estimates and initial estimates of $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1, R_1 = -1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = -2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$. On some of these time steps the ϵ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

Question 3. Imagine a poor Edmontonian student who lives along a bus line. Every morning of weekdays, the student shows up at 8am in the bus-stop, when the bus is supposed to arrive. Walking to the school from the bus stop takes 10 minutes (hard but doable in $-40°C$), while by bus it only takes 5 minutes to get to the school (lots of red light). Sadly, the bus arrival times are effectively random, but at least they come from a fixed distribution because in Edmonton the traffic is the same every weekday. The student wants to find out the optimal amount of time to wait for the bus before they start to walk (after walking, the bus cannot be caught; it may not even stop!) so that the total expected travel time (from 8am until the student arrives at the school) is the smallest possible. [2]

(a) Assume that the distribution of the bus arrival time follows the power law: Let the pdf p underlying the random arrival time $X \geq 0$ of the bus be $p(x) = \mathbb{I}(x \geq 1)/x^2$, $x \geq 0$.[3] What is the expected travel time assuming the student waits until time y?

(b) What is the optimal expected waiting time if the random arrival time follows the power law distribution as in the previous part? What is the total expected travel time with the optimal choice?

(c) How do the answers to the previous two questions change if the bus arrival's time pdf is exponential with parameter $\lambda > 0$? That is, $p(x) = \lambda e^{-\lambda x}$ for $x \geq 0$?

(d) Edmonton being Edmonton, perhaps neither the power distribution or the exponential distribution are the correct distribution for the bus arrival times. Hence, we better use a learning algorithm. Formulate the problem as a bandit problem with an infinite action set[4] so that a good bandit algorithm that maximizes reward in the problem you set up can be used by the student to minimize their total travel time.

Hint: When formulating this bandit problem, feel free to specify the reward distributions in an implicit fashion (in particular, as a function of the bus's random arrival time). Indeed, for a full specification of a bandit problem we do not need to write down the distributions underlying each arm, it suffices if we know that there is a specific reward distribution that generates the reward every time the arm is pulled.

Note: The bandit problem formulated here belongs to the so-called structured bandit problems. In a structured bandit problem, an unknown "parameter" influences the reward distributions of *all* the arms. Here, the unknown parameter is the distribution of the bus arrival time. The point about structured bandit problems is that pulling one arm can give information about the other arms. As such, often better algorithms exist to address structured bandit problems; algorithms that, unsurprisingly, exploit the structure of the problem. This is the case, here as well. In this problem the unknown parameter comes from a very big space (all distributions), but the structure is still very helpful in playing well. In fact, in the above problem, some algorithms achieve a regret (see Worksheet 1!) of at most $\tilde{O}(\sqrt{n})$. And note that this is for a bandit problem with an *infinite* action space: The action space cardinality does not matter if the problem has sufficiently strong extra structure!

---

[2] This exercise is (vaguely) based on, not a true story, but a paper that goes and "solves" this super important bandit problem. The interested student can check the paper out by clicking here.

[3] Here, $\mathbb{I}(\text{predicate}) = 1$ if predicate evaluates to true, otherwise $\mathbb{I}(\text{predicate}) = 0$. The function $\mathbb{I}(\cdot)$ is called the indicator function.

[4] A bandit problem with infinite action sets is just like a bandit problem with finitely many actions, except that there are infinitely many actions.

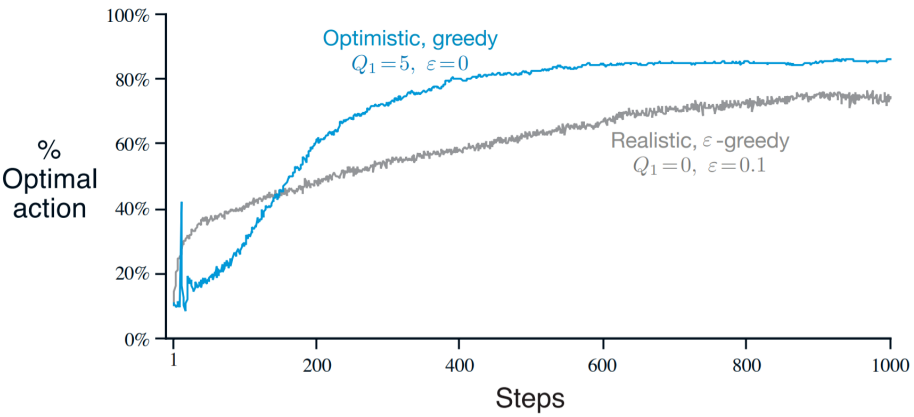

Figure 2.3: The effect of optimistic initial action-value estimates on the 10-armed testbed. Both methods used a constant step-size parameter, $\alpha = 0.1$.

Question 4. (****) (*Exercise 2.6 S&B*) The results shown in Figure 2.3 should be quite reliable because they are averages over 2000 individual, randomly chosen 10-armed bandit tasks. Why, then, are there oscillations and spikes in the early part of the curve for the optimistic method? In other words, what might make this method perform particularly better or worse, on average, on particular early steps?

Question 5. (***)[5]

Fix a positive integer n. Let $v(\mathcal{A}, \nu)$ denote the total expected reward an algorithm $\mathcal{A}$ generates when interacting with a k-armed stochastic bandit environment $\nu = (P_a)_{a\in[k]}$ for n steps. Assume that the algorithms under consideration are *broad-sense invariant* (BSI)[6] to shifting the rewards, which is defined as follows: For $\lambda \in \mathbb{R}$, let $P_a + \lambda$ denote the distribution of $R + \lambda$, where $R \sim P_a$ (the distribution P_a is shifted by λ). When $\nu = (P_a)_{a\in[k]}$, let $\nu + \lambda$ denote the bandit environment $(P_a + \lambda)_{a\in[k]}$. We say that algorithm $\mathcal{A}$ is *broad-sense invariant* to shifting the rewards (in short, BSI) if for any $\lambda \in \mathbb{R}$ and any bandit environment ν, $v(\mathcal{A}, \nu + \lambda) = v(\mathcal{A}, \nu) + \lambda n$, assuming the values here are well-defined.[7]

Let $\mathcal{E} = (\nu_1, \ldots, \nu_m)$ be a finite list of environments. Extending the previous notation, for $\lambda \in \mathbb{R}^m$, let $\mathcal{E} + \lambda$ denote $(\nu_1 + \lambda_1, \ldots, \nu_m + \lambda_m)$. Let $p : [m] \to [0, 1]$ be a probability mass function on $[m]$. Let $v(\mathcal{A}, p, \mathcal{E}) = \sum_{i=1}^m p(i) v(\mathcal{A}, \nu_i)$.

(a) Is the algorithm that always chooses the first arm BSI?

(b) Is the algorithm that plays the arms uniformly at random BSI?

(c) (***) Is ϵ-greedy BSI? Assume that when taking the greedy action and there is a tie, the arm with the smallest index is preferred. Further, assume that in time steps $t \in [k]$, arm t is pulled, so that the means are properly initialized.

(d) Is UCB BSI?[8]

(e) (****) Is Gradient Bandit (Section 2.8) BSI?

(f) Is Optimistic Initialization BSI? That is, $Q_1(a) = +5$ for each action $a \in [k]$. The choice is greedy, with ties broken with preference towards smaller indices. The algorithm updates the action-values using $Q_{n+1}(A_t) = Q_n(A_t) + \frac{1}{N_t(A_t)}(R_t - Q_n(A_t))$ where $N_t(a)$ is the number of times action a was chosen during time steps $1, \ldots, t$.

(g) Take any BSI algorithm $\mathcal{A}$. Show that for any $\lambda \in \mathbb{R}^m$, $v(\mathcal{A}, p, \mathcal{E} + \lambda) = v(\mathcal{A}, p, \mathcal{E}) + n \sum_{i=1}^m p(i)\lambda_i$.

(h) Let $\mathcal{A}^*(p, \mathcal{E}) = \arg\max_{\mathcal{A}\in\mathrm{BSI}} v(\mathcal{A}, p, \mathcal{E})$. (Here, arg max returns all the maximizing algorithms and the argument ranges over all broad-sense reward-shift invariant algorithms.)[9] Show that $\mathcal{A}^*(p, \mathcal{E} + \lambda) = \mathcal{A}^*(p, \mathcal{E})$ holds for any $\lambda \in \mathbb{R}^m$, that is, the "choice map" $\mathcal{E} \mapsto \mathcal{A}^*(p, \mathcal{E})$ is invariant to reward shifts. (What a choice map tells us is how should we choose the algorithm given a set of possible environments.)

(i) Fix an arbitrary BSI bandit algorithm $\mathcal{A}_0$. Show that the choice map

$$\mathcal{E} \mapsto \arg\max_{\mathcal{A}\in\mathrm{BSI}} \min_{\nu\in\mathcal{E}} \Big[v(\mathcal{A}, \nu) - v(\mathcal{A}_0, \nu) \Big]$$

is invariant to reward shifts, i.e. for any $\lambda \in \mathbb{R}^m$, show that

$$\arg\max_{\mathcal{A}\in\mathrm{BSI}} \min_{\nu\in\mathcal{E}} \Big[v(\mathcal{A}, \nu + \lambda) - v(\mathcal{A}_0, \nu + \lambda) \Big] = \arg\max_{\mathcal{A}\in\mathrm{BSI}} \min_{\nu\in\mathcal{E}} \Big[v(\mathcal{A}, \nu) - v(\mathcal{A}_0, \nu) \Big] .$$

Argue for and against using this choice map as a way of choosing a good algorithm.

(j) Let $v^*(\nu) = \max_{\mathcal{A}\in\mathrm{BSI}} v(\mathcal{A}, \nu)$. Show that if the means of the arms in ν are $\mu_1, \ldots, \mu_k$ then $v^*(\nu) = n \max(\mu_1, \ldots, \mu_k)$.

[5] This is not necessarily difficult in a technical sense, but it requires quite a bit of development of tools, hence the many stars.

[6] In class, we just used invariant here. I added the qualifier "broad-sense" in this exercise because I want to use invariant for the case when the conditional distribution of choosing any particular action given any particular history is independent of whether the rewards are shifted by a constant, whereas the definition here is milder.

[7] Read the note at the end for the explanation of the terminology used.

[8] This does not get stars, assuming you can solve the previous problem.

[9] In this exercise we assume that this set is nonempty. We also use the same assumption for argmin.

(k) Show that the choice map

$$\mathcal{E} \mapsto \operatorname*{argmin}_{\mathcal{A}\in \mathrm{BSI}} \max_{\nu\in\mathcal{E}} \Big[v^*(\nu) - v(\mathcal{A},\nu)\Big]$$

is invariant to reward shifts.

Note on the word invariant and the terminology used: Invariant means "staying the same in the face of changes". What we see here though is that the value generated by the algorithm changes, in a linear fashion. In other words, the value varies together (in a systematic fashion) with the change introduced. This is usually called *covariance*. Using this word, we could say that a broadly-sense invariant algorithm is one whose value is covariant to shifts of reward by a constant. Since there is nothing that is required to stay the same, there is nothing really invariant here. Should we then call $\mathcal{A}$ covariant rather than invariant? Not really. It is not really $\mathcal{A}$ that is covariant (an algorithm is a way of assigning probability distributions over actions given histories), but the map $\nu \mapsto v(\mathcal{A},\nu)$, which is the interconnection (composition) of $\mathcal{A}$ and v. Hence, the terminology. Defining an algorithm to be invariant to reward-shifts if the probability it assigns to actions given any history stays the same, you should be able to verify that such an algorithm is BSI. Is the reverse also true? Is it true that every BSI algorithm is invariant.

Question 6. Let X, Y be discrete-valued random variables such that the expectations of X, Y and XY are all well-defined.

1. Show that if X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ holds.

2. Show that if $\mathbb{P}(X = c) = 1$ for some real c then X is independent of Y (actually, this holds no matter whether the expectations of X, Y or XY are well-defined).

3. Show that if X and Y are independent then

$$\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = 0\,.$$

   Note: Recall that for nonconstant random variables, the correlation is defined as the above expectation divided by $\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}$. Hence, for such random variables it holds that if they are independent, then their correlation is zero.

4. Provide an example of X and Y that are uncorrelated but not independent. **Hint**: There is an example when the sample space has 4 elements. (Challenge: Does there exist an example when the sample space has fewer elements?)

Note: The results of this exercise remain true regardless of whether X and Y are discrete.
