Exercise D.1. Berkeley Pacman Exercises Q1-6

Question 2

As we were supposed to change the value of only one of the variables, we changed only the value of the noise from 0.2 to 0. This was because with 0.2 the agent was not able to cross the bridge and so reducing the value would mean that the agent is less likely to end up in an unintended successor state when performing an action. The agent has a number of actions that it can take but there are a subset of actions that might be invalid at certain states, and with lower noise we reduce the number of a failure successor state.

Question 3

For part a the optimal policy the goal was to prefer the close exit and risk the cliff. So, in this case the answer discount was kept from the previous question at 0.9 and the noise was 0.1. However, what is important here is that the living reward was set to -5 because the agent is risking the cliff and at each cliff region the payoff is -10, which means the living reward has to decrease by a lot in order to obtain the optimal policy. As for part b, the case is different in the sense that the agent avoids the cliff. Therefore, the living reward is just 0 as it is less likely to incur in negative payoff. Nonetheless, the paths are longer and so the answer discount has to be lowered and in this case we chose it to be 0.3 with a higher noise of 0.2. For part c, the problem changes since now the agent prefers the distant exit while risking the cliff. At each cliff region the payoff is -10 but its choosing a distant exit and so if it follows an optimal policy the payoff will be $+10$, which means that the noise and the living reward don't need to be considered, they are set to 0. For part d, the agent still prefers the distant exit, but it avoids the cliff and so the noise and discount values could be set to the default 0.9 discount and 0.2 noise. Finally, for part e, for the episode to never terminate, where it avoids both the exits and the cliff, the agent should attempt to keep a similar discount and noise that are no less than 0.5, here what worked for us was having the same values for both the discount and noise of 0.5 and again as the aforementioned, keeping the living reward at 0.

Question 6

Because 50 iterations are too few iterations, there is not an epsilon and learning rate for which it is highly likely that the optimal policy will be learned. For this reason both the epsilon and learning rate were kept with the none values and in turn would return "NOT POSSIBLE".

Exercise D.2. This is a modeling-only exercise, no need to code anything here, just to think (!)

(a) Provide the formalization of the above problem as an infinite-horizon discounted reward MDP

$$\mathcal{M} = \left\langle S, A, (P_a)_{a \in A}, r, s_0, \gamma \right\rangle$$

- The set of States S is defined as the number of times I can roll a dice in the context provided by the problem, that is, I can not get more than 6 rolls because there is a probability of 100% I will repeat a number, so we define $N = 6 = \{roll_1, roll_2, roll_3, roll_4, roll_5, roll_6\}$ and $M = 6 = \{dice_value_1, dice_value_2, dice_value_3, dice_value_4, dice_value_5, dice_value_6\}$. Additionally I have the State $LOST$ of being out of the game when a repeated number appears or because I decided to finish the episode. Formally, I have 2 variables:

  $X_1 = N + 1 = \{roll_1, roll_2, roll_3, roll_4, roll_5, roll_6, LOST\}$

  $X_2 = M = 6 = \{dice\_value_1, dice\_value_2, dice\_value_3, dice\_value_4, dice\_value_5, dice\_value_6\}$

$$S = M * N + 1 = \{roll_{11}, roll_{12}, roll_{13}, roll_{14}, roll_{15}, roll_{16},$$
$$roll_{21}, roll_{22}, roll_{23}, roll_{24}, roll_{25}, roll_{26},$$
$$roll_{31}, roll_{32}, roll_{33}, roll_{34}, roll_{35}, roll_{36},$$
$$roll_{41}, roll_{42}, roll_{43}, roll_{44}, roll_{45}, roll_{46},$$
$$roll_{51}, roll_{52}, roll_{53}, roll_{54}, roll_{55}, roll_{56},$$
$$roll_{61}, roll_{62}, roll_{63}, roll_{64}, roll_{65}, roll_{66}, LOST\}$$

- The set of Actions A includes the action $a_0 = roll$ modelling the possibility of keep rolling a dice and the action $a_1 = finish$ for the option of finalizing the game. Formally:

  $$A = \{a_0 = roll, a_1 = finish\}$$

- The transition probabilities $(P_a)_{a \in A}(s|s')$ of going from state s to s' when taking action $a \in A$ in state s is given by Figure 1. Notice that these probabilities only show $a_0 = roll$ because the $a_1 = finish$ will have $P_{a_1} = 1.0$ in every state S. The probabilities are calculated following Table 1 where the probabilities of repeating a number or not are defined, as well as taking into consideration that you will be able to roll the dice up to 6 times without repetition, where the probability or repeating is with full certainty equal to 1:

$Number of Roll$	$P(repeating)$	$P(not\_repeating)$
$roll_1$	0	1
$roll_2$	1/6	5/6
$roll_3$	2/6	4/6
$roll_4$	3/6	3/6
$roll_5$	4/6	5/6
$roll_6$	5/6	1/6
$roll_7$	6/6	0

Table 1: Probabilities of repeating and not repeating a number

For example, if it is my first time rolling a dice $roll = r = 1$ I will have the same probability for all the possible numbers $I = \{1, 2, 3, 4, 5, 6\}$ in the dice such that $(P_{r=1,i})_{i \in I} = (1/6 * P_{r=1}(not\_repeating))$. If it is my second rolling, it will be $(P_{r=2,i})_{i \in I} = (1/6 * P_{r=2}(not\_repeating))$, and so on. All in all, my final transition probabilities will be represented as P_r, i (see Figure 1) defined as the probability of rolling number r and obtaining dice number i.

	Roll11	Roll12	Roll13	Roll14	Roll15	Roll16	Roll21	Roll22	Roll23	Roll24	Roll25	Roll26	Roll31	Roll32	Roll33	Roll34	Roll35	Roll36	Roll41	Roll42	Roll43	Roll44	Roll45	Roll46	Roll51	Roll52	Roll53	Roll54	Roll55	Roll56	Roll61	Roll62	Roll63	Roll64	Roll65	Roll66	LOST	TOTAL
Roll11	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
Roll12	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
Roll13	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
Roll14	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
Roll15	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
Roll16	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
Roll21	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.86	1.00
Roll22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.86	1.00
Roll23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.86	1.00
Roll24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.86	1.00
Roll25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.86	1.00
Roll26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.89	1.00
Roll31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.89	1.00
Roll32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.89	1.00
Roll33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.89	1.00
Roll34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.89	1.00
Roll35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.89	1.00
Roll36	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.89	1.00
Roll41	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.92	1.00
Roll42	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.92	1.00
Roll43	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.92	1.00
Roll44	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.92	1.00
Roll45	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.92	1.00
Roll46	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.92	1.00
Roll51	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	1.00
Roll52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	1.00
Roll53	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	1.00
Roll54	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	1.00
Roll55	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	1.00
Roll56	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.94	1.00
Roll61	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.97	1.00
Roll62	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.97	1.00
Roll63	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.97	1.00
Roll64	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.97	1.00
Roll65	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.97	1.00
Roll66	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.97	1.00
LOST	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00

Figure 1: Transition Probabilities $(P_a)_{a \in A}(s|s')$

- The reward function $r(a, s)$ obtained from doing action a in state s which provides the immediate reward value is basically given by the dice number obtained in the particular state s for action a. For example:
  - $r(a_0, s_r = 1_i = 1) = 1$
  - $r(a_0, s_r = 1_i = 2) = 2$
  - $r(a_0, s_r = 4_i = 5) = 5$
  - $r(a_1, s_r = 2_i = 3) = 3$

  Bare in mind that these are immediate rewards, meaning, that accumulated rewards would depend on the previous values given by the successive dice values obtained; and given that there is a probability of repeating a dice value after the first roll, this would affect the accumulated reward by multiplying all the values obtained by 0 because the rules of this particular game established it this way.
- The initial state s_0 is given by the first rolling state $roll_1$ as I am starting to play and can be in any of these states $s_0 = \{roll_{11}, roll_{12}, roll_{13}, roll_{14}, roll_{15}, roll_{16}\}$. It can also be seen as the initial state the state $LOST$ (or OUT of the game) and you are about to start playing again.
- The factor γ for the proposed model can be any value between $0 < \gamma < 1$. The intention of using this factor is to inform the agent how important is the rewards in terms of time, that is, if it cares more about rewards now compared to rewards that can be received in the future. If I use a $\gamma = 0$ it means that our agent would care about the first reward only (in other words, is short-sighted). On the other hand, using a $\gamma = 1$ our agent would care more about future rewards. We are just stating this explanation here in order to clarify the intention of using a γ factor on discounted MDPs.

(b) Discuss briefly what should be changed in your model in order to work for games with dice having an arbitrary number N of sides.

As our number of states depends directly on the number of sides that the dice have (6 in the original case) and the possible rolling number before there is a repetition (same number as dice's sides) plus one state to model the LOST state, the only part of our model that would be affected is the calculations of the transition probabilities $(P_a)_{a \in A}(s|s')$. Nevertheless, it would still comply with the same size of $(N * M + 1)$ probabilities as well as the number of the set of States S.

(c) Discuss briefly what modifications could be done to model the following variation of the game still as a discounted-reward MDP (or justify why it is not possible to model that variation). In this variation, the player only plays one single episode, that is, either she rolls the die a few times without getting any repeated outcome, collects the reward, and "goes home" (no further rewards), or a repeated outcome shows up before she has decided to stop, in which case she goes home with empty pockets.

In the answer **a)** for the γ factor, we established the importance of the different values that γ can take and what would that mean: I care more about the future rewards or the immediate ones? So, by following that logic and taking into consideration the variation of the problem proposed, if I will only play for a single episode, I would try to get the best out of the situation, meaning that I would use a γ closer to 0 than to 1. Normally, I will play several episodes, where the agent is not certain about what would happen in the future, so γ defines a kind of finite-horizon for what to care about. In conclusion, if I keep the factor γ I would still be able to model the problem as a discounted-reward MDP, although I may not get the maximum accumulated reward but I still can take advantage of using a γ closer to 0.

Exercise D.3. Berkeley Pacman — Discussion

When trying to relate the "noise" and "living reward" to what a classic MDP terminology uses, we believe that "noise" intervenes directly on the transition probability function given that the description they provide is "Noise refers to how often an agent ends up in an unintended successor state when they perform an action", where it is known that successor states (that is exactly what the transition probability function does) are calculated exactly there. And for the "living reward", using for instance the negative living reward example in which will affect agent behaviour by giving it an incentive to finish the game sooner than later, one can see that affects rewards in terms of cumulative rewards calculation (similar to what γ factor does when is closer to 0, favoring the recent events than the ones occurring in the future). On that same note, we can see from the *gridworld.py* file that if the north or south actions are chosen, the the transition probabilities with the added

noise will go into states and similarly, with the probability minus 1 divided by 2 it will go to east and west. The same would happen the other way around if the actions chosen are east and west, then the probabilities would go into north and south. Reward functions describe how the agent should behave, this means they stipulate what we want the agent to accomplish. In relation to the code, it is related to the reward function in the sense that it serves as an incentive to finish the game as fast as possible because the living reward is how much reward if received on non-exit states.