

Peeking Blackjack

Stanford CS221 Spring 2018-2019



Owner CA: Jaebum Lee

Version: 1

General Instructions

This (and every) assignment has a written part and a programming part.

The full assignment with our supporting code and scripts can be downloaded as [blackjack.zip](#).

-  This icon means a written answer is expected in [blackjack.pdf](#).
-  This icon means you should write code in [submission.py](#).

All written answers must be **in order** and **clearly and correctly labeled** to receive credit.

You should modify the code in [submission.py](#) between

```
# BEGIN_YOUR_CODE
```

and

```
# END_YOUR_CODE
```

but you can add other helper functions outside this block if you want. Do not make changes to files other than [submission.py](#).

Your code will be evaluated on two types of test cases, **basic** and **hidden**, which you can see in [grader.py](#). Basic tests, which are fully provided to you, do not stress your code with large inputs or tricky corner cases. Hidden tests are more complex and do stress your code. The inputs of hidden tests are provided in [grader.py](#), but the correct outputs are not. To run the tests, you will need to have [graderUtil.py](#) in the same directory as your code and [grader.py](#). Then, you can run all the tests by typing

```
python grader.py
```

This will tell you only whether you passed the basic tests. On the hidden tests, the script will alert you if your code takes too long or crashes, but does not say whether you got the correct output. You can also run a single test (e.g., [3a-0-basic](#)) by typing

```
python grader.py 3a-0-basic
```

We strongly encourage you to read and understand the test cases, create your own test cases, and not just blindly run [grader.py](#).

The search algorithms explored in the previous assignment work great when you know exactly the results of your actions. Unfortunately, the real world is not so predictable. One of the key aspects of an effective AI is the ability to reason in the face of uncertainty.

Markov decision processes (MDPs) can be used to formalize uncertain situations. In this homework, you will implement algorithms to find the optimal policy in these situations. You will then formalize a modified version of Blackjack as an MDP, and apply your algorithm to find the optimal policy.

Problem 1: Value Iteration

In this problem, you will perform the value iteration updates manually on a very basic game just to solidify your intuitions about solving MDPs. The set of possible states in this game is $\{-2, -1, 0, 1, 2\}$. You start at state 0, and if you reach either -2 or 2, the game ends. At each state, you can take one of two actions: $\{-1, +1\}$.

If you're in state s and choose -1:


- You have an 80% chance of reaching the state $s - 1$.
- You have a 20% chance of reaching the state $s + 1$.

If you're in state s and choose +1:



- You have a 70% chance of reaching the state $s + 1$.
- You have a 30% chance of reaching the state $s - 1$.

If your action results in transitioning to state -2, then you receive a reward of 20. If your action results in transitioning to state 2, then your reward is 100. Otherwise, your reward is -5. Assume the discount factor γ is 1.

- a.  [3 points] Give the value of $V_{\text{opt}}(s)$ for each state s after 0, 1, and 2 iterations of value iteration. Iteration 0 just initializes all the values of V to 0. Terminal states do not have any optimal policies and take on a value of 0.

Iteration 0:

$$V_{\text{opt}}(-2) = 0,$$

$$V_{\text{opt}}(-1) = 0,$$

$$V_{\text{opt}}(0) = 0,$$

$$V_{\text{opt}}(1) = 0,$$

$$V_{\text{opt}}(2) = 0$$

Iteration 1:

$$V_{\text{opt}}(-2) = 0,$$

$$V_{\text{opt}}(-1) = \max\{0.8(20 + 0) + 0.2(-5 + 0), 0.7(-5 + 0) + 0.3(20 + 0)\} = 15,$$

$$V_{\text{opt}}(0) = \max\{0.8(-5 + 0) + 0.2(-5 + 0), 0.7(-5 + 0) + 0.3(-5 + 0)\} = -5,$$

$$V_{\text{opt}}(1) = \max\{0.8(-5 + 0) + 0.2(100 + 0), 0.7(100 + 0) + 0.3(-5 + 0)\} = 68.5,$$

$$V_{\text{opt}}(2) = 0$$

Iteration 2:


$$V_{\text{opt}}(-2) = 0,$$

$$V_{\text{opt}}(-1) = \max\{0.8(20 + 0) + 0.2(-5 - 5), 0.7(-5 - 5) + 0.3(20 + 0)\} = 14,$$

$$V_{\text{opt}}(0) = \max\{0.8(-5 + 15) + 0.2(-5 + 68.5), 0.7(-5 + 68.5) + 0.3(-5 + 15)\} = 47.45,$$

$$V_{\text{opt}}(1) = \max\{0.8(-5 - 5) + 0.2(100 + 0), 0.7(100 + 0) + 0.3(-5 - 5)\} = 67,$$

$$V_{\text{opt}}(2) = 0$$

- b.  [3 points] What is the resulting optimal policy π_{opt} for all non-terminal states?

$$\pi_{\text{opt}}(-2) = \text{no action (end state)}$$

$$\pi_{\text{opt}}(-1) = -1$$


$$\pi_{\text{opt}}(0) = +1$$

$$\pi_{\text{opt}}(1) = +1$$

$$\pi_{\text{opt}}(2) = \text{no action (end state)}$$

Problem 2: Transforming MDPs

Let's implement value iteration to compute the optimal policy on an arbitrary MDP. Later, we'll create the specific MDP for Blackjack.

- a.  [3 points] If we add noise to the transitions of an MDP, does the optimal value always get worse? Specifically, consider an MDP with reward function $\text{Reward}(s, a, s')$, states States , and transition function $T(s, a, s')$. Let's define a new MDP which is identical to the original, except that on each action, with probability $\frac{1}{2}$, we randomly jump to one of the states that we could have reached before with positive probability. Formally, this modified transition function is:

$$T'(s, a, s') = \frac{1}{2}T(s, a, s') + \frac{1}{2} \cdot \frac{1}{|\{s'' : T(s, a, s'') > 0\}|}.$$

Let V_1 be the optimal value function for the original MDP, and V_2 the optimal value function for the modified MDP. Is it always the case that $V_1(s_{\text{start}}) \geq V_2(s_{\text{start}})$? If so, prove it on the written portion and put **return None** for each of the code blocks. Otherwise, construct a counterexample by filling out **CounterexampleMDP** in [submission.py](#).

For one of many possible counterexamples, consider an MDP with three states:

$B \leftarrow A \rightarrow C$

Let A be the start state and both B and C be terminal states. Assume there is only one action (say, 'travel') to perform from state A, which takes you to state B with reward 0 (probability 0.9) and state C with reward 10 (probability 0.1). Intuitively, if we add noise, it makes it more likely you'll end up in C and receive reward 10, so the optimal value goes up. This counterexample is coded up in [CounterexampleMDP](#).

- b. [3 points] Suppose we have an acyclic MDP for which we want to find the optimal value at each node. We could run value iteration, which would require multiple iterations -- but it would be nice to be more efficient for MDPs with this acyclic property. Briefly explain an algorithm that will allow us to compute V_{opt} for each node with only a single pass over all the (s, a, s') triples.

Since the MDP is acyclic, we can simply compute $V(s)$ by using the dynamic programming recurrence, which goes over each (s, a, s') triple once. The reason that value iteration requires multiple passes is that we don't have an ordering over the states.

- c. [3 points] Suppose we have an MDP with states States and a discount factor $\gamma < 1$, but we have an MDP solver that can only solve MDPs with discount factor of 1. How can we leverage the MDP solver to solve the original MDP?

Let us define a new MDP with states $\text{States}' = \text{States} \cup \{o\}$, where o is a new state. Let's use the same actions ($\text{Actions}'(s) = \text{Actions}(s)$), but we need to keep the discount $\gamma' = 1$. Your job is to define new transition probabilities $T'(s, a, s')$ and rewards $\text{Reward}'(s, a, s')$ in terms of the old MDP such that the optimal values $V_{\text{opt}}(s)$ for all $s \in \text{States}$ are equal under the original MDP and the new MDP.

Hint: If you're not sure how to approach this problem, go back to the notes from the first MDP lecture and read closely the slides on convergence, toward the end of the deck.

The idea is to interpret the discount γ as the probability of not transitioning into o . Let o be a terminal state. This admits two solutions:

1. (correct)

- $T'(s, a, s') \doteq \gamma T(s, a, s')$ for $s' \in \text{States}$
- $T'(s, a, o) \doteq 1 - \gamma$
- $\text{Reward}'(s, a, s') \doteq \text{Reward}(s, a, s')$ for all $s' \in \text{States}$
- $\text{Reward}'(s, a, o) \doteq \sum_{s' \in \text{States}} T(s, a, s') \text{Reward}(s, a, s')$

2. (correct)

- $T'(s, a, s') \doteq \gamma T(s, a, s')$ for $s' \in \text{States}$
- $T'(s, a, o) \doteq 1 - \gamma$
- $\text{Reward}'(s, a, s') \doteq \frac{1}{\gamma} \text{Reward}(s, a, s')$ for all $s' \in \text{States}$
- $\text{Reward}'(s, a, o) \doteq 0$

3. (incorrect)

- $T'(s, a, s') \doteq \gamma T(s, a, s')$ for $s' \in \text{States}$
- $T'(s, a, o) \doteq 1 - \gamma$
- $\text{Reward}'(s, a, s') \doteq \text{Reward}(s, a, s')$ for all $s' \in \text{States}$
- $\text{Reward}'(s, a, o) \doteq 0$

Recall the recurrence for the new optimal value:

$$\begin{aligned}
V'_{\text{opt}}(s) &= \max_{a \in \text{Actions}(s)} \sum_{s' \in \text{States}'} T'(s, a, s') [\text{Reward}'(s, a, s') + V'_{\text{opt}}(s')]. \\
&= \max_{a \in \text{Actions}(s) \setminus o} \sum_{s' \in \text{States}'} T'(s, a, s') [\text{Reward}'(s, a, s') + V'_{\text{opt}}(s')] \\
&\quad + T'(s, a, o) [\text{Reward}'(s, a, o) + V'_{\text{opt}}(o)]
\end{aligned}$$

Plugging in the definitions of the new transitions and rewards, we get that, for each solution:

$$\begin{aligned}
1. \quad V'_{\text{opt}}(s) &= \max_{a \in \text{Actions}(s)} \sum_{s' \in \text{States}} \gamma T(s, a, s') [\text{Reward}(s, a, s') + V'_{\text{opt}}(s')] + \\
&\quad (1 - \gamma) \sum_{s' \in \text{States}} T(s, a, s') \text{Reward}(s, a, s'), \\
&= \max_{a \in \text{Actions}(s)} \sum_{s' \in \text{States}} T(s, a, s') [\text{Reward}(s, a, s') + \gamma V'_{\text{opt}}(s')]. \\
2. \quad V'_{\text{opt}}(s) &= \max_{a \in \text{Actions}(s)} \sum_{s' \in \text{States}} \gamma T(s, a, s') \left[\frac{1}{\gamma} \text{Reward}(s, a, s') + V'_{\text{opt}}(s') \right] + \\
&\quad (1 - \gamma) * 0 \\
&= \max_{a \in \text{Actions}(s)} \sum_{s' \in \text{States}} T(s, a, s') [\text{Reward}(s, a, s') + \gamma V'_{\text{opt}}(s')] \\
3. \quad V'_{\text{opt}}(s) &= \max_{a \in \text{Actions}(s)} \sum_{s' \in \text{States}} \gamma T(s, a, s') [\text{Reward}(s, a, s') + V'_{\text{opt}}(s')] + \\
&\quad (1 - \gamma) * 0 \\
&= \max_{a \in \text{Actions}(s)} \sum_{s' \in \text{States}} T(s, a, s') [\gamma \text{Reward}(s, a, s') + \gamma V'_{\text{opt}}(s')]
\end{aligned}$$

In all cases except for the third, we have recovered the original recurrence:

$$V'_{\text{opt}}(s) = \max_{a \in \text{Actions}(s)} \sum_{s' \in \text{States}} T(s, a, s') [\text{Reward}(s, a, s') + \gamma V'_{\text{opt}}(s')].$$

Therefore, the new MDP and the old MDP have the same optimal values.

Problem 3: Peeking Blackjack

Now that we have gotten a bit of practice with general-purpose MDP algorithms, let's use them to play (a modified version of) Blackjack. For this problem, you will be creating an MDP to describe states, actions, and rewards in this game.

For our version of Blackjack, the deck can contain an arbitrary collection of cards with different face values. At the start of the game, the deck contains the same number of each card of each face value; we call this number the 'multiplicity'. For example, a standard deck of 52 cards would have face values $[1, 2, \dots, 13]$ and multiplicity 4. You could also have a deck with face values $[1, 5, 20]$; if we used multiplicity 10 in this case, there would be 30 cards in total (10 each of 1s, 5s, and 20s). The deck is shuffled, meaning that each permutation of the cards is equally likely.

The game occurs in a sequence of rounds. Each round, the player either (i) takes the next card from the top of the deck (costing nothing), (ii) peeks at the top card (costing `peekCost`, in which case the next round, that card will be drawn), or (iii) quits the game. (Note: it is not possible to peek twice in a row; if the player peeks twice in a row, then `succAndProbReward()` should return `[]`.)

The game continues until one of the following conditions becomes true:

- The player quits, in which case her reward is the sum of the face values of the cards in her hand.
- The player takes a card and "goes bust". This means that the sum of the face values of the cards in her hand is strictly greater than the threshold specified at the start of the game. If this happens, her reward is 0.
- The deck runs out of cards, in which case it is as if she quits, and she gets a reward which is the sum of the cards in her hand. *Make sure that if you take the last card and go bust, then the reward becomes 0. (Not the sum of values of cards)*

In this problem, your state s will be represented as a 3-element tuple:

`(totalCardValueInHand, nextCardIndexIfPeeked, deckCardCounts)`

As an example, assume the deck has card values `[1, 2, 3]` with multiplicity 1, and the threshold is 4. Initially, the player has no cards, so her total is 0; this corresponds to state `(0, None, (1, 1, 1))`. At this point, she can take, peek, or quit.

- If she takes a card, then the three possible successor states (each of which has equal probability of $1/3$) are:

`(1, None, (0, 1, 1))`
`(2, None, (1, 0, 1))`
`(3, None, (1, 1, 0))`

She will receive a reward of 0 for reaching any of these states. (Remember, even though she now has a card in her hand for which she may receive a reward at the end of the game, the reward is not actually granted until the game ends.)

- If she peeks, the three possible successor states are:


`(0, 0, (1, 1, 1))`
`(0, 1, (1, 1, 1))`
`(0, 2, (1, 1, 1))`


She will receive (immediate) reward `-peekCost` for reaching any of these states. Things to remember about the states after a peek action:

- From `(0, 0, (1, 1, 1))`, taking a card will lead to the state `(1, None, (0, 1, 1))` deterministically.
- The second element of the state tuple is not the face value of the card that will be drawn next, but the index into the deck (the third element of the state tuple) of the card that will be drawn next. In other words, the second element will always be between 0 and `len(deckCardCounts)-1`, inclusive.
- If she quits, then the resulting state will be `(0, None, None)`. (Remember that setting the deck to `None` signifies the end of the game.)

As another example, let's say the player's current state is `(3, None, (1, 1, 0))`, and the threshold remains 4.


- If she quits, the successor state will be `(3, None, None)`.
- If she takes, the successor states are `(3 + 1, None, (0, 1, 0))` or `(3 + 2, None, None)`. Note that in the second successor state, the deck is set to `None` to signify the game ended with a bust. You should also set the deck to `None` if the deck runs out of cards.

a.  [10 points] Implement the game of Blackjack as an MDP by filling out the `succAndProbReward()` function of class `BlackjackMDP`.

b.  [4 points] Let's say you're running a casino, and you're trying to design a deck to make people peek a lot. Assuming a fixed threshold of 20, and a peek cost of 1, design a deck where for at least 10% of states, the optimal policy is to peek. Fill out the function `peekingMDP()` to return an instance of `BlackjackMDP` where the optimal action is to peek in at least 10% of states. *Hint: Before randomly assigning values, think of the case when you really want to peek instead of blindly taking a card.*


Problem 4: Learning to Play Blackjack

So far, we've seen how MDP algorithms can take an MDP which describes the full dynamics of the game and return an optimal policy. But suppose you go into a casino, and no one tells you the rewards nor the transitions. We will see how reinforcement learning can allow you to play the game and learn its rules & strategy at the same time!

a.  [8 points] You will first implement a generic Q-learning algorithm `QLearningAlgorithm`, which is an instance of an `RLAlgorithm`. As discussed in class, reinforcement learning algorithms are capable of executing a policy while simultaneously improving that policy. Look in `simulate()`, in `util.py` to see how the `RLAlgorithm` will be used. In short, your `QLearningAlgorithm` will be run in a simulation of the MDP, and will alternately be asked for an action to perform in a given state (`QLearningAlgorithm.getAction()`), and then be informed of the result of that action (`QLearningAlgorithm.incorporateFeedback()`), so that it may learn better actions to perform in the future.

We are using Q-learning with function approximation, which means $\hat{Q}_{\text{opt}}(s, a) = \mathbf{w} \cdot \phi(s, a)$, where in code, `w` is `self.weights`, `phi` is the `featureExtractor` function, and `Qopt` is `self.getQ`.

We have implemented `QLearningAlgorithm.getAction` as a simple ϵ -greedy policy. Your job is to implement `QLearningAlgorithm.incorporateFeedback()`, which should take an (s, a, r, s') tuple and update `self.weights` according to the standard Q-learning update.

- b.  [4 points] Now let's apply Q-learning to an MDP and see how well it performs in comparison with value iteration. First, call `simulate` using your Q-learning code and the `identityFeatureExtractor()` on the MDP `smallMDP` (defined for you in `submission.py`), with 30000 trials and default `explorationProb`.

How does the Q-learning policy compare with a policy learned by value iteration (i.e., for how many states do they produce a different action)? (Don't forget to set the `explorationProb` of your Q-learning algorithm to 0 after learning the policy.) Now run `simulate()` on `largeMDP`, again with 30,000 trials. How does the policy learned in this case compare to the policy learned by value iteration? What went wrong?


The proportion of states where Q-learning's actions differ from the actions chosen by value iteration will vary slightly due to differences in implementation, as well as whether the simulation was run just once for each MDP, or multiple times. (On each run, there is some randomness in the results due to the random exploration behavior of the Q-learning algorithm.) In general, though, the expected ranges for this problem are 0-10% of states for `smallMDP`, and 30-33% of states for `largeMDP`.


In providing an explanation for Q-learning's worse performance on `largeMDP` vs. `smallMDP`, the most important factor to mention for full credit is the fact that the state space for `largeMDP` is much larger than the state space for `smallMDP`. Random exploration of (state, action) pairs over a large state space is often not sufficient to allow the Q-learning algorithm to learn accurate Q-values for each such pair, even with a large number of iterations. Q-learning does better on `smallMDP` because the state space is relatively small, so the algorithm is likely to encounter each possible (state, action) pair many times, and can thus learn more accurate Q-values.

A secondary factor worth noting is that our `identityFeatureExtractor` uses features that are unique to each (state, action) pair, so our function approximation weights cannot be used to generalize learned Q-values to (state, action) pairs that haven't been seen.

Some common mistakes:

- o Mentioning the problem of feature generalization without mentioning the large state space for `largeMDP`. (The former problem wouldn't matter as much if not for the latter problem, as we can see in the case of `smallMDP`, where our feature extractor still isn't great but Q-learning nevertheless performs pretty well.)
- o Highlighting the specific behavior of the Q-learning algorithm in terms of how and why specific actions differed. A number of students gave an explanation about how Q-learning on either MDP was more "aggressive" or more "conservative" based on the number of times it chose to take/peek/quit. Remember that the Q-learning algorithm has no notion of what it means to be "aggressive" or "conservative" in this or any other game; it simply tries to estimate rewards over a sequence of states/actions, but can't do that very well if it hasn't explored enough of those states/actions.
- o Giving an explanation of how Q-learning performs worse than value iteration because it doesn't know the rewards or transition probabilities in advance. While this is true, this fact doesn't offer us any insight into why Q-learning does a good job approximating VI on `smallMDP`, but does a relatively poor job approximating the optimal solution on `largeMDP`.

- c.  [5 points] To address the problems explored in the previous exercise, let's incorporate some domain knowledge to improve generalization. This way, the algorithm can use what it has learned about some states to improve its prediction performance on other states. Implement `blackjackFeatureExtractor` as described in the code comments. Using this feature extractor, you should be able to get pretty close to the optimum on the `largeMDP`.

- d.  [4 points] Sometimes, we might reasonably wonder how an optimal policy learned for one MDP might perform if applied to another MDP with similar structure but slightly different characteristics. For example, imagine that you created an MDP to choose an optimal strategy for playing "traditional" blackjack, with a standard card deck and a threshold of 21. You're living it up in Vegas every weekend, but the casinos get wise to your approach and decide to make a change to the game to disrupt your strategy: going forward, the threshold for the blackjack tables is 17 instead of 21. If you continued playing the modified game with your original policy, how well would you do? (This is just a hypothetical example; we won't look specifically at the blackjack game in this problem.)

To explore this scenario, let's take a brief look at how a policy learned using value iteration responds to a change in the rules of the MDP.

- First, run value iteration on the `originalMDP` (defined for you in `submission.py`) to compute an optimal policy for that MDP.
- Next, simulate your policy on `newThresholdMDP` (also defined for you in `submission.py`) by calling `simulate` with an instance of `FixedRLAlgorithm` that has been instantiated using the policy you computed with value iteration. What is the expected reward from this simulation? *Hint: read the documentation (comments) for the `simulate` function in `util.py`, and look specifically at the format of the function's return value.*
- Now try simulating Q-learning on `originalMDP` (30,000 trials). Then, using the learned parameters, run Q-learning again on `newThresholdMDP` (again, 30000 trials). What is your expected reward under the new Q-learning policy? Provide some explanation for how the rewards compare with when `FixedRLAlgorithm` is used. Why they are different?

(Below is the previous version of this problem. If you have followed this one, it is okay. You don't have to change your solution. But the correct way to observe the difference is following the above procedure;

Now try simulating Q-learning directly on `newThresholdMDP` instead. What is your expected reward under the new Q-learning policy? Provide some explanation for how the rewards compare, and why they are different.)

For the updated version; The average total reward will approximately be 8.82. Still, it performs much better than the fixed policy version as Q learning can adapt.

You get relatively low rewards (approximately 6.84) for `FixedRLAlgorithm` because you are passing in the policy learned for `originalMDP`. Because `FixedRLAlgorithm` doesn't adapt, the actions taken are not optimal actions for `newThresholdMDP`.

Running Q-learning directly on the `newThresholdMDP` produces a policy with higher rewards (approximately 9.57) because it is able to adapt to the specific rules of `newThresholdMDP` -- in this case, the higher threshold value of 15 (vs. 10 in the original MDP).

Common mistakes:

- Forgot to mention that Q-learning can adapt.
- Got incorrect average rewards.
- Reported rewards did not make sense. (Some people reported that the reward for Q-learning was always 12, or they didn't specify that it could be a different value from 12. This is incorrect, as the rewards returned by `simulate(newThresholdMdp, QLearningAlgorithm, ...)` are often less than 12.)