Blackjack using Q-Learning

Malhar Bhoite, Jeffrey Kim, Zhonglin Cao
Carnegie Mellon University

Abstract

What is Blackjack?

Blackjack is a famous and poker game which is widely played in almost every casino in the world. In a single blackjack game, players will compare the points of cards in their hands with the dealer's. The objective of this game is to obtain points as close as possible to 21 without exceeding it (called "bust" in the game). Whoever gets the higher points without busting will be the winner and can collect all bets at the end of game.

Why Reinforcement Learning?

The core of blackjack is to decide the action with the knowledge of current hand, which makes it a perfect case to be applied reinforcement learning on.

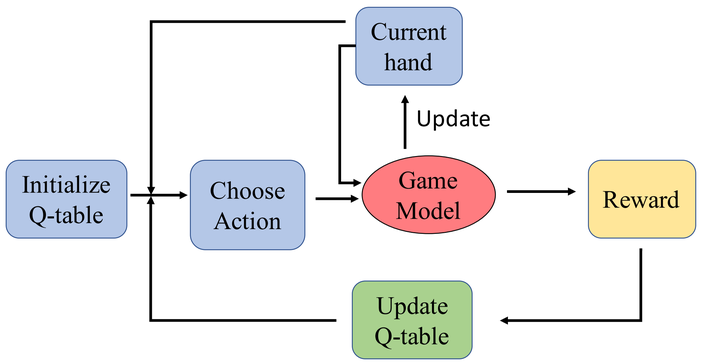
Problem:

Objective

The goal of this project include:

1. Apply Q-learning algorithm on blackjack game.
2. Observe the changes in total payout with the parameter variation.
3. Try apply other algorithms such as: deep Q-learning, exploration function etc, and compare their performance
4. Compare with traditional strategy and random choice

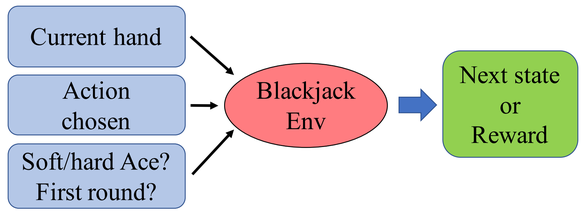
Reinforcement Learning & Game Model:



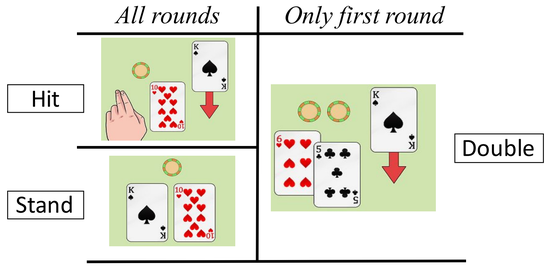
Method

Gym toolkit:

Gym Blackjack-v0 environment is used as the game model.



The original version of blackjack environment in Gym can only take two actions: hit or stand. To simulate the real game in casino as close as possible, we modified the environment two have four actions.



Q-learning:

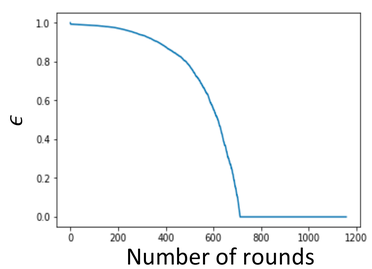
The Q-table will be updated after each game. The model will generate 1000 samples and each model will run 1000 rounds such that the total epoch trained is 1 million.

Greedy policy:

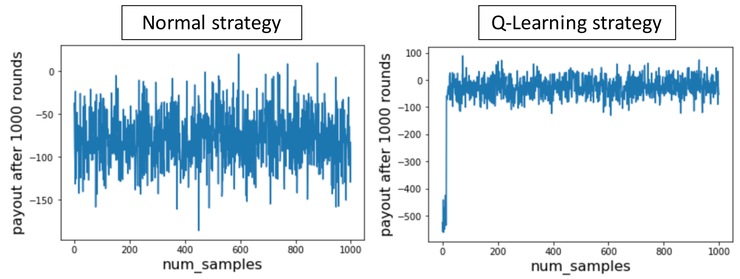
We used annealing ϵ in ϵ -greedy exploration. This means that the chance of exploration will be very high at the beginning of training and it will decrease exponentially with the number of rounds. The mathematical function of ϵ is:

$$\epsilon = \epsilon_s + (\epsilon_f - \epsilon_s)e^{\frac{t-f}{k}}$$

where s and f are the starting and final rounds number and k is the scaling scaling factor. The ϵ change with rounds is shown below.



Results



Average payout after 1000 rounds is -77.805

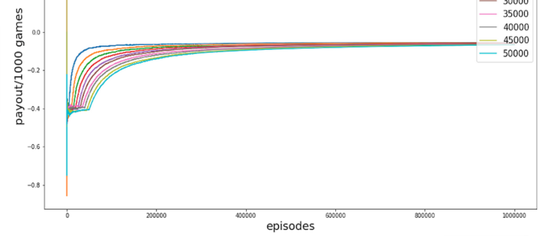
Average payout after 1000 rounds is -33.067

Payout changes with γ

Exploration round:

	5000	10000	15000	20000	25000	30000	35000
Gamma: 0.1	-58.948	-55.982	-56.582	-56.623	-56.392	-60.998	-61.161
Gamma: 0.2	-56.381	-54.547	-54.513	-58.045	-59.555	-58.411	-61.831
Gamma: 0.3	-59.112	-67.561	-64.773	-63.437	-66.711	-68.754	-69.178
Gamma: 0.4	-73.888	-73.909	-68.386	-62.856	-69.112	-73.227	-73.593
Gamma: 0.5	-76.281	-76.058	-79.853	-80.155	-77.436	-76.272	-77.179
Gamma: 0.6	-84.501	-90.217	-109.587	-90.858	-84.382	-88.708	-112.884
Gamma: 0.7	-95.089	-114.089	-111.951	-115.196	-116.857	-126.648	-124.964
Gamma: 0.8	-110.568	-128.097	-125.721	-121.632	-132.862	-129.471	-137.443
Gamma: 0.9	-135.531	-128.020	-128.392	-140.325	-133.489	-165.942	-132.290
Gamma: 1.0	-162.537	-118.475	-134.864	-129.170	-172.335	-150.248	-144.133

Payout changes with Exploration time



Decision table comparison

Black is Q-learning, hard A
Blue is Q-learning, soft A
Red is Normal strategy

Player's Hand	Dealer's upcard									
	A	2	3	4	5	6	7	8	9	10
2	H	H	H	H	H	H	H	H	H	H
3	H	H	H	H	H	H	H	H	H	H
4	H	H	H	H	H	H	H	H	H	H
5	H	H	H	H	H	H	H	H	H	H
6	H	H	H	H	H	H	H	H	H	H
7	H	H	H	H	H	H	H	H	H	H
8	H	H	H	H	H	H	H	H	H	H
9	H	H/D	H/D	H/D	H/D	H	H	H	H	H
10	H	H/D	D	D	D	D	D	D	H/D	H
11	H	D	D	D	D	D	D	H/D	D	H/D
12	H	S/H/H	S/H/H	S/H	S/H	S/H	H	H	H	H
13	H	H/S	H/S	H/S	H/S	S/H/H	H	H	H	H
14	H	S	S	S	S	S/D	H	H	H	H
15	H	S/H	S/H	S/H	S/H	S/D	H	H	H	H
16	H	H/S	S/H	S/H	S/H	S/D	S/H	S/H	H	H
17	H/S	S/H	S/H	S/H	S/H	S/D	S/H	S/H	S/H	S/H
18	S/H	S	S	S/H	S/H	S/D	S/H	S/H	S/H	S/H
19	S/H	S/H	S	S	S/D	S/D	S	S	S	S/H
20	S	S	S/D	S	S	S	S	S	S	S
21	S	S	S	S	S	S	S	S	S	S

Future Work

- Testing different exploration function algorithms
- Find explanation for the decrease of payout with higher γ