Learning to Play Blackjack using Deep Reinforcement Learning

## Chris Durbin

## Research Draft Summary

I have achieved my original research project goal of being able to learn optimal play for blackjack using deep reinforcement learning with a Deep Q-Network (DQN) agent. In the remaining weeks I have decided to attempt a more difficult challenge of counting cards to achieve a positive expected reward and learn optimal bet sizes with another DQN agent. My primary goal with the project is to continue to learn and gain more experience with deep reinforcement learning so that I can apply it to other problems in the future after graduation. I fleshed out the paper some so that you will have an idea of what I am attempting to achieve by the end.

## Introduction

The purpose of the project is to attempt to use deep reinforcement learning to learn optimal play in the game of Blackjack. By counting cards and properly sized bets players have been able to gain an advantage over casinos and have a positive expected reward. In the project I will attempt to entirely use reinforcement learning algorithms to achieve the same.

My hypothesis is that it is possible for an agent to learn to play Blackjack with positive expected reward using deep reinforcement learning without any programmed knowledge of optimal play nor using the Kelly criterion calculation (Kelly, 1956) to determine bet size.

• Student clearly defines the research topic and project goals. Describes the rough research and/or development approach, hypotheses

• What ML problem solved? What method analysis applied?

## Related Work

Prior research projects have …

1. Original Atari DQN paper (Minh, 2013) – shows how to replace the Q table with a neural network and calculate a target to use for error calculations for backprop. Shows it can learn many Atari games – no game specific knowledge in the DQN agent.
2. Deadly Triad from Sutton (Sutton, 2018) – convergence is not guaranteed when using function approximation, off-policy learning, and bootstrapping (temporal difference methods)
3. DDQN (Wang, 2016) – shows how to use two networks freezing the target network during the experience replay learning

• Correct approach or convincing novel approach to ML problem

• Theoretical research has merit and complete

## Research Project Problem

TODO – add in blackjack rules used. Move hypothesis to this section – delete the comment below.

NOTE: I do not plan to include this section in my final paper. I noticed the submission asked for this section, but I think I cover everything related to the problem in the introduction and methods sections. Please let me know if I should move anything into this section.

## Methods

The project includes several components to implement the reinforcement learning environment:

* Blackjack environment that implements the mechanics and enforces the blackjack rules.
* DQN agent to make the decisions for selecting actions for playing hands.
* DQN agent to make the decision on the bet size to use for each hand (not fully implemented yet).
* Driver to put all the components together and run the experiments.

TODO – Add an illustration of the component interaction

In addition to the deep reinforcement learning components two other agents are included as a baseline comparison for selecting actions for playing hands. One agent implements the first-visit Monte Carlo algorithm, and the other agent implements the Q-learning algorithm.

To train my agents I repeat the same experiment for TBD episodes. The constraints of an episode are as follows:

* Agent starts with a balance of $2,500
* Minimum bet for a hand is $10
* Maximum bet for a hand is $50
* Play continues until the agent no longer has enough to place a minimum bet or has completed 1,000 hands.

For my tests after the initial training has completed, I repeat the same experiment twenty times and track the results of the individual experiments as well as the average results across all twenty experiments.

### Metrics

The primary metrics I am tracking are the final average balance for each of the experiments, and if the balance is less than the minimum bet, I track the number of hands played for that experiment.

In addition, I track a few other metrics to better understand how quickly the agent learns, both in terms of number of hands played as well as the wall clock time to train.

## Results

The following…

Show a chart with expected return.

Show learning rate over number of episodes.

• Sound solution with metrics, correctly doing the ML method

• Proper runs and statistics collections

• Proper association between problem and solution shown

## Conclusions

I would recommend further…

• Relevant conclusion with support from the method sections

• Answers "why we have conducted this work "

## References

1. Kelly, John L. "A new interpretation of information rate." *the bell system technical journal* 35.4 (1956): 917-926.
2. Riedmiller, Martin. "Neural fitted Q iteration–first experiences with a data efficient neural reinforcement learning method." *Machine Learning: ECML 2005: 16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005. Proceedings 16*. Springer Berlin Heidelberg, 2005.
3. Mnih, Volodymyr, et al. "Playing atari with deep reinforcement learning." *arXiv preprint arXiv:1312.5602* (2013).
4. Schaul, Tom, et al. "Prioritized experience replay." *arXiv preprint arXiv:1511.05952* (2015).
5. Wang, Ziyu, et al. "Dueling network architectures for deep reinforcement learning." *International conference on machine learning*. PMLR, 2016.
6. Sutton, Richard S., and Andrew G. Barto. *Reinforcement Learning, second edition: An Introduction*. MIT Press, 2018.