How to get escorted out of the casino, with Reinforcement Learning

Kieran Rudd,[1,*] Michael Gamston,[1,†] and Scott Underdown[1,‡]

[1] *School of Physics and Astronomy, University of Nottingham, Nottingham, NG7 2RD, UK*

(Dated: January 25, 2025)

Note: use plain text.
Can prob just use AI to assist in this once done.

I. INTRODUCTION

* introduce the problem, its significance, how we intended to solve it

Blackjack, also known as twenty-one, is the most widely played casino game in the world, largely due to its simple game structure in which a player attempts to get the highest score by drawing cards from a deck. The probabilistic nature of Blackjack leads to

Whilst an optimum strategy for blackjack has been known by statisticians for decades by drawing on the expected value of future cards [1], training an intelligent agent to learn the optimum strategy is an approach not as often taken. Reinforcement learning is the subfield of machine learning which aims to train an agent in an environment based on a reward, so that an optimum 'policy' can be obtained. TALK ABOUT REINFORCEMENT LEARNING PLENTY (ref)

There are many variations to the game, but for the purposes of this project, the sequence of play was as follows:

1. A card is dealt to the player with value C_1.
2. For n iterations, or until a total score of 21 is exceeded, the player can make one of two choices;
   (a) Stick, and end the game.
   (b) Hit, and receive another card with value C_{n+1}.
3. The final score is calculated using

$$\text{Score} = \begin{cases} (\sum C_n)^2 & \text{if } \sum C_n \leq 21 \\ 0 & \text{if } \sum C_n > 21 \end{cases} \tag{1}$$

* describe the two situations

Optimal strategies have been developed for varieties of game play, with

To approach this challenge, two situations were considered, where 'infinite' describes the

This paper details the methodology taken for both problem situations, the result of each in context of an optimal result, and a conclusion on the ____.

II. METHODOLOGY

In training an agent to play Blackjack, an iterative Q-Learning approach has been taken. Watkins' Q-Learning aims to learn the optimal 'q-value' for given state-action pairs in an environment, i.e., the respective value of making a certain move in a certain environmental state. ... further description

This approach was selected because it does not require direct

The Bellman's equations for Q-Learning is defined as,

$$Q_{new}(s,a) = Q_{old}(s,a) + \alpha\underbrace{(R(s,a) + \gamma MaxQ(s',a') - Q_{old}(s,a))}_{\text{Temporal Difference}} \tag{2}$$

Where s, a are the current state and action, s', a' are the next state and action, α is the learning rate, γ is the discount factor, $R(s,a)$ is the reward received after taking action a in state s, and $Q(s,a)$ is the q-value for the state-action pair.

Temporal difference somewhere.

A. Infinite

B. Finite

more boobs

III. RESULTS

Here, one can display figures, such as in Figure 1.

IV. CONCLUSIONS

...

[*] efykr2@nottingham.ac.uk

[†] ppxmg5@nottingham.ac.uk

[‡] ppxsu1@nottingham.ac.uk

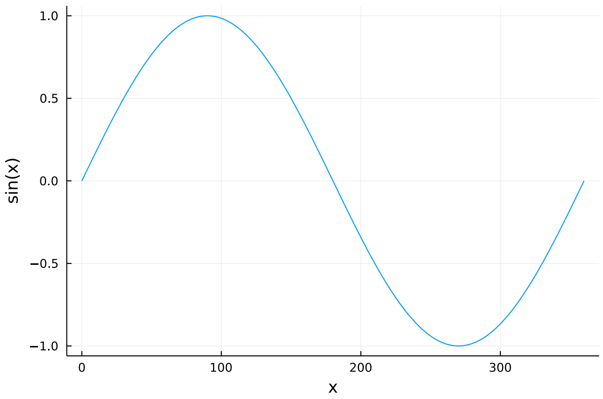

FIG. 1. Shows an example of a figure.

[1] H. M. Roger R. Baldwin, Wilbert E. Cantey and J. P. McDermott, Journal of the American Statistical Association **51**, 429 (1956), https://www.tandfonline.com/doi/pdf/10.1080/01621459.1956.105013