**4.Thực nghiệm với môi trường BlackJack**

**4.1. Mô tả về môi trường BlackJack**

**4.1.1. Những quy tắc cơ bản của Blackjack:**

1.Trò chơi được chơi với một hoặc nhiều bộ bài tiêu chuẩn.

2.Mỗi người chơi được chia hai lá bài, và người chia bài cũng được chia hai lá, với một lá bài úp.

3.Giá trị của từng lá bài được xác định bởi thứ hạng của nó. Át được tính là 1 hoặc 11, các quân hình (kings, queens, và jacks) được tính là 10, và các lá bài khác được tính bằng giá trị của nó.

4.Người chơi có thể "hit" và lấy thêm lá bài để cải thiện tay của họ, hoặc "stand" và giữ tay hiện tại của họ.

5.Người chia bài phải "hit" cho đến khi tay của họ có giá trị là 17 hoặc hơn.

6.Nếu tay của một người chơi vượt quá 21, họ "bust" và thua cuộc.

7.Nếu tay của người chia bài vượt quá 21, người chơi thắng cuộc.

8.Nếu không người chơi hay người chia bài vượt quá 21, tay có giá trị tổng cộng cao nhất và nhỏ hơn hoặc bằng 21 sẽ thắng cuộc.

**4.1.2. Action Space**

Không gian hành động chỉ có một chiều và chỉ có hai hành động khả thi được biểu diễn bằng các số nguyên trong phạm vi {0,1}, trong đó:

* 0: stick ( dừng lại và giữ bài hiện tại)
* 1: Hit ( rút thêm một lá bài )

**4.1.3. Observation Space**

Không gian quan sát bao gồm một bộ 3 giá trị: tổng điểm hiện tại của người chơi, giá trị của lá bài mở của nhà cái (từ 1 đến 10, trong đó 1 là quân Át), và thông tin về việc người chơi có lá bài Át được tính là 1 hay không được tính là 0.

**4.1.4. Starting State**

The starting state is initialised in the following range.

Graphical user interface, application, table

Description automatically generated

**4.1.5. Rewards**

* Thắng trận: +1
* Thua trận: -1
* Hoà trận: 0

**4.1.6. Điều kiện kết thúc một Episode**

Một episode kết thúc ( một ván blackjack) kết thúc khi :

* Nếu người chơi rút thêm bài và tổng điểm vượt quá 21, ván chơi kết thúc.
* Nếu người chơi không rút thêm bài nữa, ván chơi kết thúc.
* Át là luôn được tính là 11 trừ khi nó làm cho tổng điểm vượt quá 21.

**4.2. Giải quyết môi trường Blackjack với chiến lượng epsilon-greedy và thuật toán Q-learning**

**4.2.1. Ứng dụng chiến lược Epsilon-greedy để giải quyết Blackjack**

1.Chiến lược epsilon-greedy là một phương pháp để giải quyết bài toán đưa ra hành động tốt nhất cho một tình huống dựa trên chính sách hiện tại. Theo chiến lược này, đôi khi tác tử sẽ chọn một hành động tốt nhất (với xác suất 1-epsilon) hoặc một hành động ngẫu nhiên (với xác suất epsilon) để tìm kiếm các chính sách mới, có thể đem lại phần thưởng tốt hơn trong tương lai.

2.Trong trò chơi Blackjack, chiến lược epsilon-greedy có thể được áp dụng để xác định liệu người chơi có nên rút thêm một lá bài hay dừng lại. Tại mỗi bước của trò chơi, đại lý (tức là người chơi) có thể chọn hành động được khuyến nghị bởi chính sách hiện tại hoặc một hành động ngẫu nhiên. Chính sách sẽ được học qua thời gian bằng cách cập nhật ước lượng giá trị hành động cho mỗi cặp trạng thái-hành động dựa trên các phần thưởng nhận được trong trò chơi.

3.Khi trò chơi được chơi nhiều lần, đại lý sẽ học được chính sách tối ưu nhằm tối đa hóa phần thưởng trung bình. Ban đầu, đại lý có thể khám phá bằng cách thực hiện các hành động ngẫu nhiên để tìm kiếm các chiến lược mới. Tuy nhiên, khi trò chơi diễn ra, đại lý sẽ bắt đầu tận dụng chính sách tốt nhất đã biết, nhằm tối đa hóa phần thưởng trung bình trong thời gian dài.

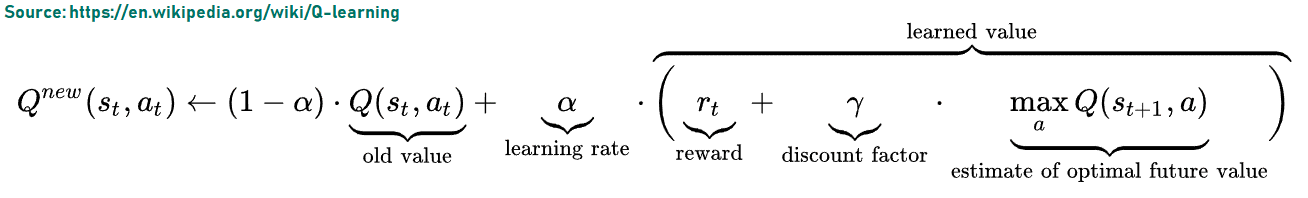
**4.2.2: Training agent chơi Blackjack với thuật toán Q-learning**

**4.2.2.1. Q-learing**

* Q-Learning là một thuật toán RL có thể áp dụng cho nhiều môi trường khác nhau.
* Q-learning thực hiện bằng cách lặp đi lặp lại các action và nhận được reward để xây dựng look-up table gọi là Q-table.

**4.2.2.2 Các bước thực hiện thuật toán Q-learning trong mô hình Blackjack:**

1. Khởi tạo bảng Q với các giá trị bất kỳ.
2. Chọn trạng thái ban đầu của mô hình. Trạng thái ban đầu này bao gồm tổng điểm hiện tại của người chơi, giá trị lá bài của thợ bàn hiện ra và trạng thái của lá bài Ace của người chơi.
3. Chọn hành động đầu tiên để thực hiện. Hành động này có thể được chọn bằng cách thực hiện một trong hai hành động có sẵn, là hit (rút thêm lá bài) hoặc stand (ngừng).
4. Thực hiện hành động và quan sát kết quả. Kết quả này sẽ bao gồm trạng thái mới, phần thưởng nhận được và trạng thái tiếp theo của người chơi.
5. Sử dụng công thức Q-learning để cập nhật giá trị trong bảng Q theo công thức :



1. Chọn hành động tiếp theo dựa trên giá trị trong bảng Q. Hành động này có thể được chọn dựa trên chiến lược epsilon-greedy để đảm bảo khả năng khám phá.
2. Lặp lại các bước từ bước 4 đến bước 6 cho đến khi đạt được một điều kiện dừng nhất định, ví dụ như đạt đến số lần lặp tối đa hoặc bảng Q đã hội tụ.
3. Sử dụng bảng Q để đưa ra quyết định về hành động tối ưu trong các trạng thái khác nhau trong mô hình Blackjack.