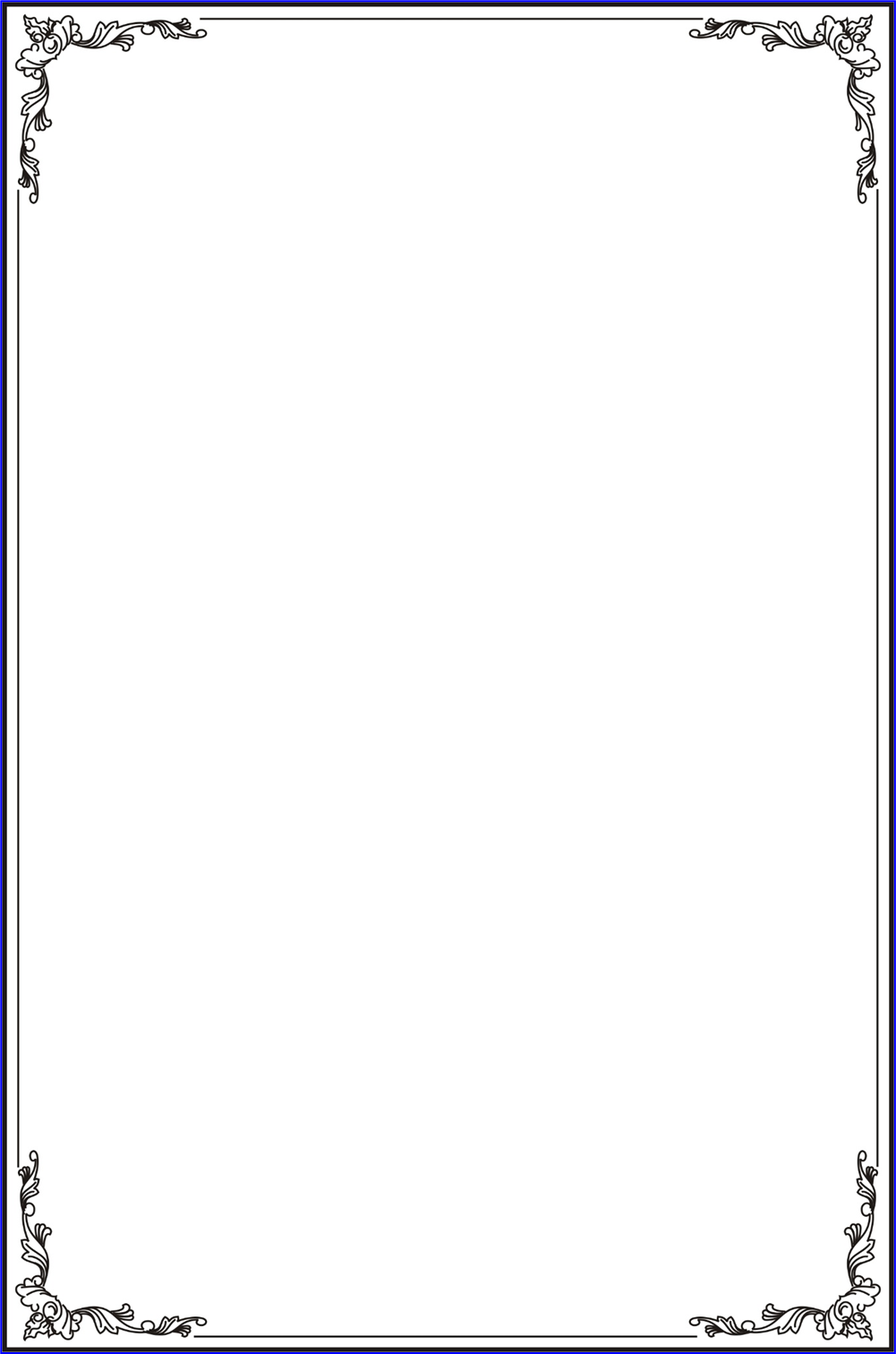
****

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

**🕯✡🕮🕮✡🕯**

****

**Đề tài**

**REINFORCEMENT LEARNING**

**BASIC CONCEPTS, FARAMA GYMNASIUM PACKAGE**

**MÔN HỌC: TRÍ TUỆ NHÂN TẠO**

**GVHD: ThS. Lê Minh Tân**

|  |  |
| --- | --- |
| ***Nhóm sinh viên thực hiện :*** | |
| Hoàng Mai Hiếu | 21110882 |
| Trương Nguyễn Thùy Trang | 21110691 |
| Đặng Công Tuấn | 21110709 |
| Nguyễn Văn Tín | 21110931 |
| Nguyễn Thành Lộc | 21110897 |
| Trần Hoàng Phúc  Nguyễn Hoàng Phương  Mai Đình Hồng Sơn | 21110606  21110609  21110465 |

**TP Hồ Chí Minh , tháng 4 năm 2023**

# Mục Lục

[I. KHÁI QUÁT CƠ BẢN VỀ REINFORCEMENT LEARNING VÀ LEARNING FROM REWARD 1](#_Toc133683640)

[1.Giới thiệu về học tăng cường và học từ phần thưởng 1](#_Toc133683641)

[2. Các thành phần chính của một bài toán Learning from Rewards 1](#_Toc133683642)

[3. Các phương pháp học từ phần thưởng 2](#_Toc133683643)

[4. Các thuật toán trong học tăng cường 2](#_Toc133683644)

[II. GYMNASIUM PACKAGE 3](#_Toc133683645)

[1. Cơ bản về Gymnasium Package 3](#_Toc133683646)

[2. Bắt đầu với Gymnasium 3](#_Toc133683647)

[3. Các môi trường cung cấp bởi Gymnasium Package 4](#_Toc133683648)

[4. Các thành phần chính của Gymnasium Package 6](#_Toc133683649)

[III.  THỰC NGHIỆM VỚI MÔI TRƯỜNG BLACKJACK 7](#_Toc133683650)

[1. Các thành phần của môi trường Blackjack trong gymnasium 7](#_Toc133683651)

[2. Giải quyết môi trường Blackjack với chiến lược Epsilon-greedy và thuật toán Q-learning 8](#_Toc133683652)

[IV. ỨNG DỤNG CỦA HỌC TĂNG CƯỜNG TRONG LĨNH VỰC CỤ THỂ, ƯU VÀ NHƯỢC ĐIỂM 9](#_Toc133683653)

[1. Ứng dụng 9](#_Toc133683654)

[2. Ưu điểm và nhược điểm của học tăng cường 9](#_Toc133683655)

[V. KẾT LUẬN 10](#_Toc133683656)

# I. KHÁI QUÁT CƠ BẢN VỀ REINFORCEMENT LEARNING VÀ LEARNING FROM REWARD

## **1.Giới thiệu về học tăng cường và học từ phần thưởng**

Reinforcement Learning hay còn gọi là “học tăng cường” là một nhánh quan trọng của Machine Learning, đóng vai trò quan trọng trong nhiều ứng dụng của trí tuệ nhân tạo, bao gồm cả trò chơi, robot, học tập và các tác vụ tự động hóa khác

Reinforcement Learning tập trung vào việc huấn luyện các mô hình máy tính để tương tác với một môi trường để đạt được một mục tiêu cụ thể thông qua phương thức phản hồi thích nghi. Trong Reinforcement Learning, mô hình máy tính cố gắng tìm ra các hành động tối ưu dựa trên các phản hồi của môi trường.

Learning from Rewards là một phương pháp học tăng cường (reinforcement learning) trong đó tác tử (agent) học cách lựa chọn hành động trong một môi trường (environment) để tối đa hóa tổng phần thưởng (reward) nhận được từ môi trường.

Khi tác tử thực hiện một hành động trong một trạng thái, môi trường sẽ cung cấp phản hồi dưới dạng phần thưởng, cho biết giá trị của hành động đó trong trạng thái hiện tại. Tác tử sử dụng thông tin phần thưởng này để cập nhật chính sách (policy), tăng cường khả năng lựa chọn hành động trong tương lai.

Quá trình này được lặp lại nhiều lần, cho đến khi tác tử học được một chính sách tối ưu để tối đa hóa phần thưởng trong môi trường. Trong quá trình này, hàm giá trị (value function) được sử dụng để đánh giá chất lượng của các hành động trong một trạng thái, giúp tác tử cập nhật chính sách một cách hiệu quả.

## **2. Các thành phần chính của một bài toán Learning from Rewards**

Một bài toán Learning from Rewards bao gồm các thành phần chính sau:

1. Môi trường (Environment): Môi trường là nơi mà hành động xảy ra và cung cấp thông tin về trạng thái hiện tại của hệ thống và các tác động của các hành động.
2. Trạng thái (State): Trạng thái là một tập hợp các đặc trưng của môi trường được sử dụng để mô tả tình trạng của hệ thống.
3. Hành động (Action): Hành động là các quyết định mà hệ thống có thể thực hiện để tương tác với môi trường.
4. Phần thưởng (Reward): Phần thưởng là một giá trị số được cung cấp cho hệ thống bởi môi trường để đánh giá hiệu quả của các hành động.
5. Chính sách (Policy): Chính sách là một quy tắc quyết định được sử dụng để lựa chọn hành động dựa trên trạng thái hiện tại của hệ thống.
6. Giá trị (Value): Giá trị là một hàm số được sử dụng để đánh giá sự hiệu quả của một chính sách hoặc một hành động trong một trạng thái cụ thể.
7. Mục tiêu (Objective): Mục tiêu là một tiêu chí được sử dụng để đánh giá hiệu quả của hệ thống và định hướng quá trình học.

## **3. Các phương pháp học từ phần thưởng**

- MODEL-BASED REINFORCEMENT LEARNING:là một phương pháp học tăng cường trong đó tác tử cố gắng xây dựng một mô hình của môi trường mà nó đang tương tác. Mô hình này được sử dụng để dự đoán trạng thái tiếp theo của môi trường và tính toán giá trị của các hành động.

- MODEL-FREE REINFORCEMENT LEARNING: là một phương pháp học tăng cường không sử dụng một mô hình của môi trường để tính toán giá trị của hành động. Thay vào đó, nó tập trung vào việc cập nhật hàm giá trị (value function) hoặc chính sách (policy) mà không cần biết một mô hình rõ ràng của môi trường:

+ ACTION-UTILITY LEARNING: là một phương pháp học tăng cường cho phép tác tử học cách chọn hành động tối ưu dựa trên độ ưu tiên của các mục tiêu cần đạt được. Điều này được thực hiện bằng cách ước lượng giá trị của từng hành động, dựa trên mức độ đóng góp của hành động đó đối với các mục tiêu cần đạt được.

+ POLICY SEARCH: là một phương pháp học tăng cường trong đó mục tiêu là tìm ra chính sách tối ưu trực tiếp, thay vì ước lượng hàm giá trị của một chính sách đã cho. Trong policy search, chúng ta cố gắng tìm kiếm chính sách mà cho phép tác tử đạt được hiệu suất tốt nhất trong một nhiệm vụ cụ thể. Chính sách có thể được biểu diễn dưới dạng một bản đồ từ trạng thái sang hành động, hoặc bằng các phương pháp khác như mạng nơ-ron.

## **4. Các thuật toán trong học tăng cường**

Thuật toán học tăng cường là một lĩnh vực quan trọng trong machine learning, nó tập trung vào việc giúp tác tử tự động học cách tương tác với một môi trường bằng cách đưa ra các hành động để đạt được các mục tiêu nhất định.

Một trong những thuật toán MODEL-BASED REINFORCEMENT LEARNING đáng chú ý là Dyna-Q, thuật toán này kết hợp MODEL-BASED và MODEL-FREE REINFORCEMENT LEARNING bằng cách sử dụng mô hình của môi trường để dự đoán trạng thái tiếp theo và tính toán giá trị của các hành động. Monte Carlo Tree Search (MCTS) là một thuật toán khác của MODEL-BASED REINFORCEMENT LEARNING, được sử dụng trong các trò chơi như cờ vua và go.

Trong MODEL-FREE REINFORCEMENT LEARNING, có nhiều thuật toán đáng chú ý được sử dụng như Q-Learning và SARSA. Cả hai đều tập trung vào cập nhật hàm giá trị của từng hành động để tìm kiếm hành động tối ưu. Deep Q-Networks (DQN) là một trong những thuật toán phổ biến khác của MODEL-FREE REINFORCEMENT LEARNING, nó sử dụng mạng nơ-ron sâu để học các hàm giá trị và tối ưu hành động.

# II. GYMNASIUM PACKAGE

## **1. Cơ bản về Gymnasium Package**

Thư viện Gymnasium là một thư viện mã nguồn mở cho phép tạo ra các môi trường tương tác để huấn luyện các mô hình học tăng cường (reinforcement learning). Nó cung cấp một bộ các môi trường phổ biến và đa dạng cho các bài toán học tăng cường, bao gồm các môi trường về trò chơi, robot và điều khiển tối ưu.

Gymnasium cung cấp một bộ sưu tập các môi trường hoặc "nhiệm vụ" có thể được sử dụng để kiểm tra và phát triển các thuật toán học tăng cường. Các môi trường này thường giống như các trò chơi, với các quy tắc được định nghĩa rõ ràng và một cấu trúc phần thưởng, làm cho chúng hữu ích để đánh giá và so sánh các thuật toán học tăng cường khác nhau.

Các môi trường trong Gymnasium Package đều cung cấp cho các tác tử một trạng thái (state) của môi trường, các hành động (actions) có thể thực hiện, và các phần thưởng (rewards) nhận được sau mỗi hành động. Tác tử sử dụng các thuật toán học tăng cường để tìm hiểu cách chọn hành động tối ưu dựa trên trạng thái hiện tại của môi trường để đạt được phần thưởng cao nhất.

Vì vậy, các môi trường trong Gymnasium Package rất phù hợp cho việc nghiên cứu và phát triển các thuật toán học tăng cường.

## **2. Bắt đầu với Gymnasium**

Quy trình cơ bản để sử dụng Gymnasium để huấn luyện mô hình Học tăng cường như sau:

Bước 1: Xác định môi trường bạn muốn làm việc.

Bước 2: Khởi tạo môi trường.

Bước 3: Xác định chính sách của tác tử (tức là cách nó quyết định hành động nào sẽ được thực hiện).

Bước 4: Tương tác với môi trường, thực hiện các hành động và nhận phần thưởng.

Bước 5: Cập nhật chính sách của tác tử dựa trên các phần thưởng mà nó nhận được.

Bước 6: Dừng việc huấn luyện nếu đã huấn luyện đủ số lần lặp hoặc tác tử đã đạt hiệu suất hài lòng. Nếu không, lặp lại bước 4 và 5 cho đến  khi trò chơi kết thúc và reset môi trường.

## **3. Các môi trường cung cấp bởi Gymnasium Package**

**3.1. Classic Control**

**Ảnh có chứa biểu đồ

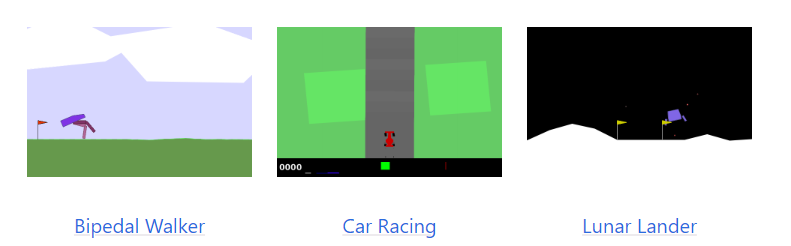
Mô tả được tạo tự động**

Có năm môi trường điều khiển cổ điển: Acrobot, CartPole, Mountain Car, Continuous Mountain Car, and Pendulum. Tất cả các môi trường này đều ngẫu nhiên về trạng thái ban đầu của chúng, trong một phạm vi nhất định. Ngoài ra, Acrobot có tiếng ồn được áp dụng cho hành động được thực hiện. Ngoài ra, đối với cả hai môi trường mountain car, những chiếc xe này đều không đủ sức mạnh để leo lên núi, vì vậy bạn phải nỗ lực một chút để lên đến đỉnh.

Trong số các môi trường Gymnasium, tập hợp các môi trường này có thể được coi là những môi trường dễ giải quyết hơn bằng chính sách.

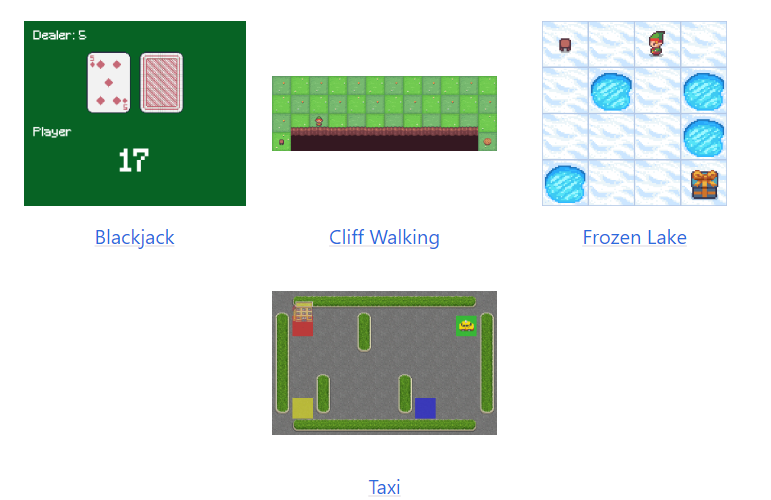
Tất cả các môi trường đều có thể định cấu hình cao thông qua các đối số được chỉ định trong tài liệu của từng môi trường.

**3.2. Box2D**

****

Tất cả các môi trường này đều liên quan đến các trò chơi đồ chơi dựa trên điều khiển vật lý, sử dụng vật lý dựa trên box2d và kết xuất dựa trên PyGame. Những môi trường này đã được Oleg Klimov đóng góp vào những ngày đầu của OpenAI Gym và đã trở thành tiêu chuẩn đồ chơi phổ biến kể từ đó. Tất cả các môi trường đều có thể định cấu hình cao thông qua các đối số được chỉ định trong tài liệu của từng môi trường.

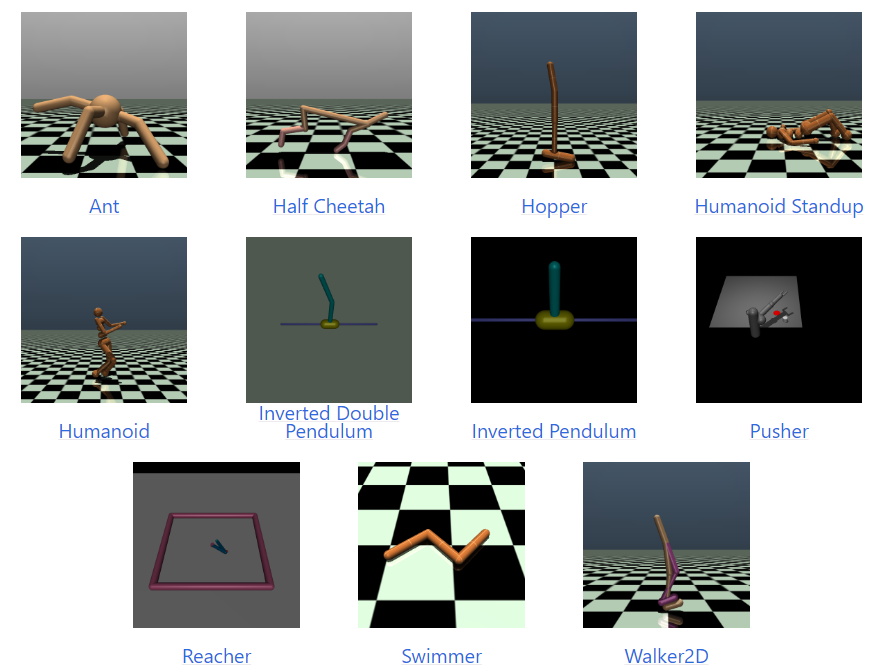
**3.3. Toy Text**

****

Môi trường văn bản đồ chơi được thiết kế cực kỳ đơn giản, với các không gian hành động và trạng thái riêng biệt nhỏ, do đó dễ học. Do đó, chúng phù hợp để gỡ lỗi triển khai các thuật toán học tăng cường.

Các môi trường có thể định cấu hình thông qua các đối số được chỉ định trong tài liệu của từng môi trường.

**3.4 MuJoCo**

****

MuJoCo là viết tắt của Multi-Joint dynamics with Contact. Nó là một công cụ vật lý để hỗ trợ nghiên cứu và phát triển về rô-bốt, cơ chế sinh học, đồ họa và hoạt hình cũng như các lĩnh vực khác cần mô phỏng nhanh và chính xác.

Có 11 môi trường Mujoco: Ant, HalfCheetah, Hopper, Humanoid, HumanoidStandup, InvertedDoublePendulum, InvertedPendulum, Pusher, Reacher, Swimmer và Walker2d.

Xem thêm tại: https://gymnasium.farama.org/environments/mujoco/

**3.5. Atari**

Một tập hợp các môi trường Atari 2600 được mô phỏng thông qua Stella và Môi trường học tập Arcade.

Ảnh có chứa văn bản

Mô tả được tạo tự động Xem thêm tại: https://gymnasium.farama.org/environments/atari/

## **4. Các thành phần chính của Gymnasium Package**

Observation and Action Spaces: Không gian quan sát (observation space) là tập hợp các trạng thái có thể mà một tác tử có thể quan sát được trong môi trường. Còn không gian hành động (action space) là tập hợp các hành động có thể mà một tác tử có thể thực hiện trong môi trường.

Episode: Một Episode là một vòng chơi hoàn chỉnh của một môi trường, bắt đầu từ trạng thái ban đầu và tiếp tục cho đến khi đạt được trạng thái kết thúc. Mỗi tập được tạo thành bởi một chuỗi các trạng thái, hành động và phần thưởng.

Wrapper: Một wrapper là một công cụ trong Gymnasium cho phép bạn thay đổi hành vi của một môi trường mà không cần thay đổi mã nguồn của nó. Các wrapper có thể được sử dụng để thêm các tính năng như giới hạn thời gian, hình dạng phần thưởng và chặn hành động.

Benchmark: Gymnasium cung cấp một bộ các môi trường đánh giá chuẩn để đánh giá và so sánh các thuật toán học tăng cường. Các môi trường đánh giá này bao gồm các nhiệm vụ điều khiển cổ điển, các trò chơi Atari và các nhiệm vụ robot.

# III.  THỰC NGHIỆM VỚI MÔI TRƯỜNG BLACKJACK

## **1. Các thành phần của môi trường Blackjack trong gymnasium**

**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 tổng giá trị bài của họ, hoặc "stick" để giữ bài hiện tại.

5. Người chia bài phải "hit" cho đến khi tổng giá trị bài của họ bằng 17 hoặc hơn.

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

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

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

**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)

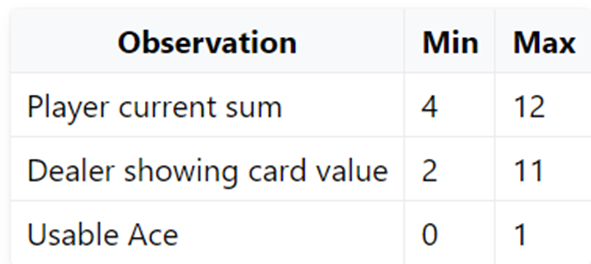
·         1: hit (rút thêm một lá bài)

**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 ngửa của nhà cái (từ 2 đến 11, trong đó 11 là quân Át), một giá trị cho biết rằng liệu người chơi giữ quân Át có sử dụng được nó với giá trị là 11 hay không (0, 1).

**1.4. Starting State**

Trạng thái bắt đầu được khởi tạo trong phạm vi sau:

****

**1.5. Rewards**

· Thắng trận: +1

· Thua trận: -1

· Hòa trận: 0

**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.

## **2. Giải quyết môi trường Blackjack với chiến lược Epsilon-greedy và thuật toán Q-learning**

**2.1. Chiến lược Epsilon-greedy**

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, tác tử (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, tác tử sẽ học được chính sách tối ưu nhằm tối đa hóa phần thưởng trung bình. Ban đầu, tác tử 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, tác tử sẽ bắt đầu tận dụng chính sách tốt nhất đã biết, nhằm tối đa hóa phần thưởng trung bình trong thời gian dài.

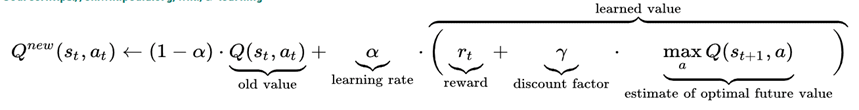
**2.2. Huấn luyện tác tử chơi Blackjack với thuật toán Q-learning**

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. Các bước xây dựng Q-table:

1. Khởi tạo bảng Q với các giá trị bất kỳ.

2. Dựa vào chiến lược epsilon-greedy để lựa chọn hành động.

3. 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.

4. Sử dụng công thức Q-learning để cập nhật giá trị trong Q-table theo công thức :  
  
****

5.  Lặp lại các bước từ 2 đến 4 cho đến khi đạt được một điều kiện dừng nhất định, ví dụ như đạt đến số lần lặp tối đa hoặc bảng Q được lấp đầy.

# IV. ỨNG DỤNG CỦA HỌC TĂNG CƯỜNG TRONG LĨNH VỰC CỤ THỂ, ƯU VÀ NHƯỢC ĐIỂM

## **1. Ứng dụng**

Các ứng dụng của Reinforcement Learning rất đa dạng và phong phú. Chúng ta có thể kể đến một số ứng dụng như: quản lý xe ô tô tự lái, giải quyết vấn đề tiêu thụ năng lượng, điều khiển tín hiệu giao thông, chăm sóc sức khỏe, phát triển người máy, marketing, game, hỗ trợ điều khiển nhập học trong viễn thông và hỗ trợ phi công trực thăng, và cả trong khoa học để giúp hiểu các phản ứng hóa học. Các ứng dụng này sử dụng các thuật toán Reinforcement Learning để học từ các tương tác với môi trường, giúp tối ưu hóa các hành động và đưa ra quyết định chính xác. Thuật toán Deep Learning như AlphaGo, Alpha Zero cũng được áp dụng để tạo ra các máy chơi game và phát hiện lỗi trong môi trường trò chơi.

## **2. Ưu điểm và nhược điểm của học tăng cường**

**Ưu điểm:**

- Nó có thể giải quyết các vấn đề phức tạp và bậc cao hơn. Ngoài ra, các giải pháp thu được sẽ rất chính xác.

- Có thể sửa các lỗi xảy ra trong quá trình đào tạo.

- Kể cả khi không có dữ liệu đào tạo, nó sẽ học được thông qua kinh nghiệm mà nó có được từ việc xử lý dữ liệu đào tạo

- Đối với các vấn đề khác nhau, có vẻ phức tạp đối với chúng tôi, nó cung cấp các mô hình hoàn hảo để giải quyết chúng.

**Nhược điểm:**

- Không thích hợp để giải quyết các vấn đề đơn giản hơn. Lý do là, các mô hình thường giải quyết các vấn đề phức tạp. Và cũng không thích hợp vì sẽ lãng phí sức mạnh xử lý và không gian không cần thiết

- Reinforcement Learning cần nhiều dữ liệu để cung cấp cho mô hình tính toán. Điều này tiêu tốn thời gian

- Chi phí bảo trì rất cao.

- Việc đào tạo quá nhiều có thể dẫn đến quá tải các trạng thái của mô hình, có thể làm giảm kết quả.

# V. KẾT LUẬN

Tóm lại, Learning from Rewards là một phương pháp trong Học máy, trong đó tác tử học cách tương tác với môi trường và tối ưu hóa hành động của mình dựa trên phần thưởng hoặc hình phạt. Phương pháp này có ứng dụng rộng rãi trong nhiều lĩnh vực như robotica, trò chơi điện tử, tối ưu hóa tự động và hệ thống điều khiển.

Gymnasium Package là một thư viện mã nguồn mở cung cấp một môi trường mô phỏng các vấn đề tối ưu hóa cho phép các nhà nghiên cứu và lập trình viên có thể thử nghiệm các thuật toán học tăng cường và tạo ra các mô hình học máy.

Sử dụng Gymnasium Package, các nhà nghiên cứu và lập trình viên có thể tạo và kiểm tra các thuật toán học tăng cường trên một loạt các môi trường khác nhau, từ đơn giản đến phức tạp, để giải quyết các vấn đề từ học tăng cường đơn giản đến các vấn đề phức tạp như tự động hóa sản xuất và các bài toán tối ưu hóa trong các lĩnh vực kinh tế và tài chính.

Hướng phát triển tiềm năng của Learning from Rewards và Gymnasium Package có thể bao gồm mở rộng các mô hình học tăng cường để giải quyết các vấn đề khó khăn hơn trong các môi trường mô phỏng, tập trung vào các ứng dụng thực tế và tích hợp các công nghệ mới vào thư viện.