TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

VIỆN TOÁN ỨNG DỤNG & TIN HỌC

_____ * _____

BÁO CÁO CUỐI KÌ MÔN MÔ HÌNH NGẪU NHIÊU VÀ ỨNG DỤNG

BÀI TOÁN QUYẾT ĐỊNH MARKOV VÀ MÔ HÌNH HỌC TĂNG CƯỜNG

Giảng viên hướng dẫn: TS. Nguyễn Thị Ngọc Anh

Sinh viên thực hiện, mã số sinh viên, lớp:

Trương Tiến Dũng - 20150722 - Toán tin 02-K60

Nguyễn Thành Đức - 20151043 - Toán tin 02-K60

Nguyễn Trung Đức - 20151047 - Toán tin 02-K60

HÀ NỘI - 2019

Mục lục

Lời mở đầu

Trong những năm gần đây, khoa học máy tính phát triển như vũ bão, các mô hình trí tuệ nhân tạo được ứng dụng vào nhiều lĩnh vực khác nhau trong cuộc sống, đem lại nhiều lợi ích cho con người và xã hội. Trong đó, các mô hình học máy đã đạt được những thành tựu vô cùng đáng kể trong những năm gần đây. Học tăng cường (Reinforcement Learning) là một phương pháp phổ biến để giải bài toán quyết định Markov. Nó có thể rất nhiều ứng dụng trong các lĩnh vực kỹ thuật như quy hoạch toán học, điều khiển tối ưu, ...

Trong báo cáo này, chúng em sẽ trình bày về bài toán quyết định Markov (Markov decision process - MDP) và thuật toán học tăng cường Q-learning áp dụng cho bài toán "Nhà thám hiểm và đường hầm", bố cục của bài bào cáo của chúng em bao gồm các nội dung sau

- **Cơ sở lý thuyết:** Trình bày về một số khái niệm và một số tính chất của MDP, giới thiệu thuật toán Q-learning.
- **Áp dụng:** Trình bày khái quát về bài toán "Nhà thám hiểm và đường hầm", thuật toán áp dụng và kết quả đạt được.

Chúng em xin tỏ lòng biết ơn sâu sắc tới TS. Nguyễn Thị Ngọc Anh, cô đã hướng dẫn rất tận tình và giúp đỡ chúng em rất nhiều trong học phần "Các mô hình ngẫu nhiêu và ứng dụng", giúp chúng em hoàn thành báo cáo này.

Chương 1

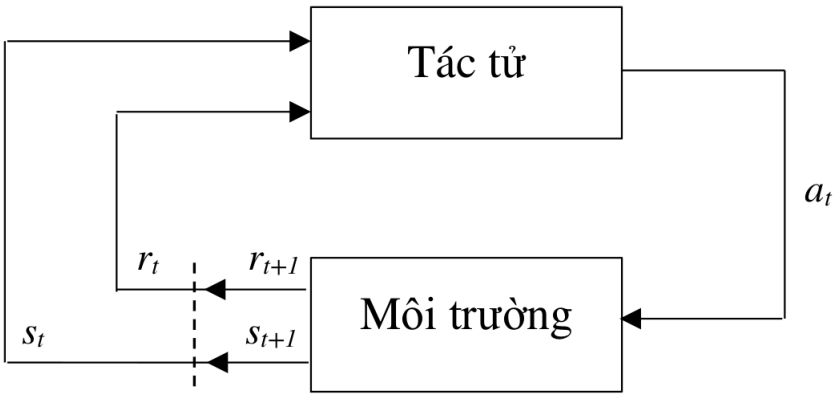
Cơ sở lý thuyết

Trong phần này, chúng em sẽ trình bày sơ lược về một số tính chất liên quan đến bài toán quyết định Markov và thuật toán Q-learning, các khái niệm được chúng em tham khảo tại các công trình khoa học đã có từ trước [3] [5].

1.1 Bài toán quyết định Markov

1.1.1 Phát biểu bài toán

Bài toán quyết định Markov là bài toán học từ các tác động để đạt mục đích, trong đó, người học và người ra quyết định được gọi là tác tử; bên ngoài tác tử được gọi là môi trường. Tác động thực hiện một cách liên tục và tác tử lựa chọn hành động và môi trường để đáp ứng lại tác động và chuyển từ trạng thái hiện tại sang trạng thái mới. Các đặc tính quan trọng của bài toàn quyết định Markov gồm (1) Tác tử tương tác với môi trường và cặp "Tác tử + Môi trường" tạo thành một hệ thống động; và (2) Tín hiệu tăng cường, được nhận biết dựa vào mục tiêu, cho phép tác tử thay đổi hành vi của nó.



Hình 1.1: Mô hình tương tác giữa tác tử và môi trường

Như minh họa trong Hình 1.1 , tác tử và môi trường tác động lẫn nhau trong chuỗi các bước thời gian rời rạc $t = 0, 1, 2, ..., T$. Tại mỗi bước, tác tử nhận một số biểu diễn về trạng thái môi trường $s_t \in S$ (S là tập các trạng thái có thể) và thực hiện ánh xạ đến hành động có thể lựa chọn $a_t \in A(s_t)$ ($A(s_t)$ là tập các hành vi có hiệu lực trong trạng thái s_t). Phép ánh xạ được gọi là chiến lược của tác tử, ký hiệu $\pi_t(\pi_t = S \longrightarrow A)$. Mục tiêu của nó là xác định giá trị tăng cường $r_{t+1} \in R$ cực đại và tự động tìm kiếm trạng thái mới s_{t+1}.

Hàm $P_{a_t}(s + t, s_{t+1})$ được gọi là hàm chuyển trạng thái, thể hiện xác xuất chuyển từ s_t sang s_{t+1} thông qua hành động a_t. Do a_t và s_t độc lập có điều kiện với toàn bộ trạng thái và hành động trước đó, các trạng thái chuyển tiếp của một quá trình MDP thỏa mãn thuộc tính Markov.

Quá trình quyết định Markov là một phần mở rộng của chuỗi Markov, khác biệt là ở sự bổ sung của các hành động và phần thưởng. Ngược lại, nếu chỉ có một hành động tồn tại cho mỗi trạng thái và tất cả các phần thưởng là giống nhau, một quá trình quyết định Markov làm giảm một chuỗi Markov.

1.1.2 Các phần tử của bài toán quyết định Markov

Một quá trình quyết định Markov $M = (S, A, P_0)$, với S là tập các trạng thái đếm được khác rỗng, $\mathcal{A}$ tập các hành động (A cũng là tập đếm được và khác rỗng). P_0 là tập các xác suất chuyển ứng với mỗi cặp trạng thái và hành động tương ứng $(s, a) \in S \times A$ một độ đo xác suất trên $S \times \mathbb{R}$ mà chúng ta định nghĩa bởi $= P_0(s, a)$.

P_0 có thể hiểu: cho $U \subset S \times \mathbb{R}$, $P_0(U|, a)$ là xác suất mà trạng thái kế tiếp và phần thưởng của U với trạng thái hiện tại là s, và hành động được thực hiện là a. Hệ số chiết khấu γ với $0 \leq \gamma \leq 1$.

Dựa vào tác tử và môi trường, bài toán Markov đưa ra bốn phần tử con gồm chiến lược (policy), hàm phản hồi (reward function), hàm giá trị (value function) và (không bắt buộc) một mô hình về môi trường.

1.1.2.1 Chiến lược

Chiến lược định nghĩa cách thức tác tử học từ hành động tại thời điểm đưa ra. Chiến lược là một ánh xạ từ tập các trạng thái của môi trường đến tập các hành động được thực hiện khi môi trường ở trong các trạng thái đó. Nó tương ứng với tập luật nhân quả trong lĩnh vực tâm lý học. Với một số trường hợp, chiến lược có thể là một hàm đơn giản hoặc một bảng tra cứu. Với những trường hợp khác, nó có thể liên quan đến tính toán mở rộng ví dụ như một tiến trình tìm kiếm. Chiến lược là nhân của một tác tử với nhận thức rằng một mình nó đủ quyết định hành động.

Nếu tác tử tuân theo chiến lược π tại thời điểm t, thì $\pi(a|s)$ là xác xuất có điều kiện để $A_t = a$ nếu $S_t = s$. Có thể tấy hàm $\pi(a|s)$ thể hiện một phân phối xác xuất của $a \in A(s)$ với $s \in S$.

1.1.2.2 Hàm phản hồi

Hàm phản hồi định nghĩa mục tiêu trong bài toán quyết định Markov. Nó ánh xạ mỗi trạng thái quan sát được (hoặc một cặp hành động-trạng thái) của môi trường với một giá trị phản hồi R để chỉ ra mong muốn thực chất về trạng thái đó. Mục đích duy nhất của tác tử là cực đại hoá tổng giá trị phản hồi nó nhận được trong suốt thời gian chạy. Nó định nghĩa sự kiện tốt/xấu, là những đặc trưng có tính tức thời và là vấn đề mà tác tử phải đối mặt. Như vậy, hàm phản hồi cần phải có khả năng thay đổi bởi tác tử. Tuy nhiên, nó có thể phục vụ dưới dạng một yếu tố cơ bản để thay đổi chiến lược. Ví dụ, nếu hành động lựa chọn bởi chiến lược được theo sau bởi một hàm phản hồi thấp, thì chiến lược có thể được thay đổi để lựa chọn hành động khác thay thế trong tương lai.

Giá trị G_t được coi là tổng thưởng tại thời điểm t, được tính bằng công thức:

$$G_t = R_{t+1} + R_{t+2} + ... + R_T = \sum_{k=t+1}^{T} R_k \tag{1.1}$$

Trong trường hợp có chiếu khấu γ, hàm G được tính thông qua biểu thức

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... = \sum_{t=0}^{\infty} \gamma^t R_{t+1} \tag{1.2}$$

Nếu $\gamma < 1$ thì phần thưởng ở tương lai sẽ nhỏ hơn phần thưởng ở trạng thái ban đầu. Một quá trình quyết định Markov khi mà lợi nhuận được định nghĩa theo công thức thì được gọi là chiết khấu phần thưởng MDP. Khi $\gamma = 1$ thì quá trình quyết định Markov được gọi là không chiết khấu.

1.1.2.3 Hàm giá trị

Hàm giá trị chỉ ra mong muốn trong cả quá trình về các trạng thái sau khi đưa vào bản miêu tả các trạng thái tiếp theo, và các mục tiêu hiệu quả trong các trạng thái đó. Ví dụ, một trạng thái có thể thường xuyên mang lại một hàm phản hồi tức thì thấp, nhưng vẫn có một hàm giá trị cao, vì nó thường được theo sau bởi các trạng thái khác mà mang lại các giá trị phản hồi cao, hoặc ngược lại. Để tạo ra các mô hình tương tự con người, các giá trị phản hồi giống như là sự hài lòng (khi hàm phản hồi có giá trị lớn) và hình phạt (khi hàm phản hồi có giá trị thấp), trong khi các hàm giá trị tương ứng với một sự phán đoán tinh tế hơn và nhìn xa trông rộng hơn về việc hài lòng hoặc không hài lòng với môi trường ở trong một trạng thái riêng biệt. Biểu diễn theo cách này, hàm giá trị rõ ràng là một ý tưởng khuôn mẫu thân thiện và căn bản. Hàm phản hồi biểu diễn ngữ cảnh chính, trong khi hàm giá trị đưa ra giá trị dự đoán của giá trị phản hồi. Không có các giá trị phản hồi thì sẽ không có các hàm giá trị. Mục đích duy nhất của việc ước lượng các hàm giá trị là để đạt được các giá trị phản hồi lớn hơn. Tuy nhiên, chính hàm giá trị mới là đối tượng được đề cập đến nhiều nhất khi ra quyết định và đánh giá quyết định. Việc lựa chọn quyết định dựa trên sự phán đoán về hàm giá trị. Ta tìm kiếm các hành động mà đem lại các trạng thái với giá trị lớn nhất, chứ không phải là các phản hồi lớn nhất, bởi vì các hành động này chứa số lượng phản hồi lớn nhất trong cả giai đoạn. Thật không may, việc xác định giá trị khó hơn nhiều so với xác định giá trị phản hồi. Các giá trị phản hồi về cơ bản được đưa ra trực tiếp bởi môi trường, nhưng các hàm giá trị cần phải được ước lượng hoặc ước lượng lại từ chuỗi các quan sát tác tử có được qua toàn bộ thời gian sống của nó. Thực tế, thành phần quan trọng nhất của tất cả các thuật toán học tăng

cường là phương pháp để ước lượng các hàm giá trị một cách hiệu quả nhất. Vai trò trung tâm của phép ước lượng hàm giá trị có thể xem là điều quan trọng nhất về phương pháp học tăng cường trong suốt các thập kỷ gần đây. Mặc dù hầu hết các phương pháp học tăng cường được xem xét tuân theo cấu trúc xung quanh việc ước lượng các hàm giá trị, đây cũng không phải là nhân tố bắt buộc để giải quyết được các bài toán quyết định Markov. Trong thực tế, có thể sử dụng các phương pháp tìm kiếm như các thuật toán phát sinh, lập trình phát sinh, huấn luyện tái tạo và các phương pháp tối ưu hoá chức năng khác để giải quyết các bài toán quyết định Markov. Các phương pháp này tìm kiếm trực tiếp trong không gian các chiến lược mà không phải sử dụng các hàm giá trị. Chúng được gọi đây là phương pháp "tiến hoá" bởi vì hoạt động tương tự như cách mà phép tiến hoá sinh vật học tạo ra các sinh vật với các hành động có kỹ năng thậm chí khi chúng không học trong suốt chu kỳ sống. Phương pháp "tiến hóa có ưu thế trong những bài toán ở đó tác tử học không thể phán đoán chính xác trạng thái của môi trường. Tuy nhiên, nó bỏ qua rất nhiều cấu trúc có ích của bài toán quyết định Markov. Bên cạnh đó, phương pháp "tiến hóa" cũng không tận dụng thực tế rằng chiến lược mà nó đang tìm kiếm là một hàm từ các trạng thái đến hành động. Mặc dù việc "học" và "tiến hoá" chia sẻ nhiều đặc tính và có thể kết hợp cùng với nhau, như chúng thực hiện trong tự nhiên, ta sẽ không xem xét các phương pháp tiến hoá đặc biệt là trong các bài toán quyết định Markov.

Hàm giá trị v của trạng thái s theo chiến lược π, kí hiệu $v_\pi(s)$, là kỳ vọng các giá trị của hàm phản hồi khi đi từ s theo chiến lược π, công thức cụ thể được mô tả như sau:

$$V^\pi(s) = \mathbb{E}\Big[G_t|S_t = s\Big] = \mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t R_{t+1}|S_0 = s\Big], s \in S \qquad (1.3)$$

Công thức 1.3 được gọi là hàm trạng thái - giá trị cho chiến lược π.Tương tự, chúng ta định nghĩa hàm hành động - giá trị theo chiến lược π như sau

$$q^{\pi}(s,a) = \mathbb{E}\Big[G_t|S_t = s, A_t = a\Big] = \mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t R_{t+1}|S_0 = s\Big|S_t = s, A_t = a\Big] \tag{1.4}$$

1.1.2.4 Mô hình của môi trường

Phần tử cuối cùng của bài toán quyết định Markov đó là mô hình của môi trường. Đây là đối tượng để bắt chước hành vi của môi trường. Ví dụ, khi đưa ra một trạng thái và hành động, mô hình có thể dự đoán tổng hợp trạng thái tiếp theo và giá trị phản hồi tiếp theo. Các mô hình được sử dụng để lập kế hoạch để đưa ra dự đoán về quyết định trên một tiến trình của hành động bằng cách xem xét các tình huống trong tương lai có thể xảy ra trước khi chúng có kinh nghiệm thực sự. Sự hợp nhất giữa các mô hình và kế hoạch trong các hệ thống học tăng cường là một phát triển mới. Các hệ thống học tăng cường ban đầu là những người học "thử và lỗi", với cách tiếp cận này những gì chúng thực hiện được xem như là đối lập với kế hoạch. Tuy nhiên, ngày càng rõ ràng rằng các phương pháp học tăng cường có liên quan gần gũi với các phương pháp quy hoạch động, trong đó cũng sử dụng các mô hình và chúng cũng lần lượt có liên quan gần gũi với các phương pháp lập kế hoạch không gian trạng thái. Các phương pháp học tăng cường hiện đại mở rộng sự phân bố từ học thử và lỗi mức thấp sang việc lập kế hoạch có tính thảo luận mức cao.

1.2 Thuật toán Q-learning trong bài toán MDP

Q-learning [6] là một kỹ thuật đánh giá hành động nào sẽ được chấp nhận dựa trên một hàm giá trị-hành động xác định giá trị của một trạng thái nào đó và thực hiện một hành động nào đó ở trạng thái đó.

Cho quá trình quyết đinh Markov hữu hạn $MDP(\mathcal{X}, \mathcal{A}, \mathcal{P}, r)$. Trong đó hàm thưởng r:

$$r : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}$$

là hàm xác định và bị chặn.

Giá trị đạt được khi thực hiện chuỗi hành động A_t ở trạng thái x được định nghĩa:

$$J(x, A_t) = E\big[\sum_{t=0}^{\infty} \gamma^t R(X_t, A_t) | X_0 = x\big]$$

Hàm giá trị tối ưu được định nghĩa:

$$V^*(x) = max_{\mathcal{A}} J(x, A_t)$$

và thỏa mãn:

$$V^*(x) = max_{\mathcal{A}} \sum_{y \in \mathcal{X}} P(x, a, y)[r(x, a, y) + \gamma V^*(y)]$$

Từ đây, ta định nghĩa 1 hàm Q tối ưu như sau:

$$Q^*(x, a) = \sum_{y \in \mathcal{X}} P(x, a, y)[r(x, a, y) + \gamma V^*(y)]$$

Năm 1992, Watkin và Dayan đã chứng minh thuật toán Q-learning với công thức lặp:

$$Q_{t+1}(s_t, a_t) = Q_t(x_t, a_t) + \alpha \left[r_t + \gamma max_{b \in \mathcal{A}} Q_t(x_{t+1}, b) - Q_t(x_t, a_t) \right] \quad (1.5)$$

cho chúng ta hàm Q hội tụ tới Q^* với xác suất bằng 1 khi $0 \leq \alpha \leq 1$ và tát cả các cặp trạng thái - hành động (x, a) được lặp lại vô hạn cho công thức trên.

Chúng ta có một hàm Q với đầu vào là một trạng thái và một hành động,đầu ra là phần thưởng mong đợi của hành động đó (và tất cả các hành động tiếp theo) ở trạng thái đó. Trước khi chúng ta khám phá môi trường, Q cho cùng một giá trị cố định (tùy ý). Nhưng khi chúng ta khám phá môi trường nhiều hơn, Q cho chúng ta một sự ước lượng ngày một tốt hơn về giá trị của một hành động ở trạng thái nào đó. Giá trị của Q được cập nhật liên tục trong quá trình tác tử hành động. Phương trình của Q-learning giải thích tất cả một cách rất độc đáo. Nó cho thấy cách chúng ta cập nhật giá trị của Q dựa trên phần thưởng mà chúng ta nhận được từ môi trường:

$$Q(x_t, a_t) \leftarrow Q(x_t, a_t) + \alpha \left[r_t + \gamma max_a Q(x_{t+1}, a) - Q(x_t, a_t) \right] \quad (1.6)$$

Trong đó $Q : S \times A \rightarrow \mathbb{R}$ là hàm đánh giá giá trị chất lương của hành động ứng với trạng thái hiện tại. r_t là phần thưởng đã nhận được ở thời trạng thái hiện tại, α là hệ số học, γ là hệ số chiết khấu, và $maxQ(s_{t+1}, a_t)$ là giá trị ước tính nhận được trong tương lai. Các biến trên có ảnh hưởng tới thuật toán như sau:

- **Hệ số học**: thể hiện cường độ cập nhật giá trị Q. α càng gần 0 thì giá trị Q cập nhật càng chậm, $\alpha = 0$ thì tác tử gần như không học được gì,

$\alpha = 1$ thì có nghĩa là ta thay thế giá trị Q cũ bằng giá trị mới cập nhật, không lưu giữ lại thông tin giá trị cũ. Trong môi trường tất định, $\alpha = 1$ là tối ưu. Còn trong môi trường ngẫu nhiên, thuật toán chỉ hội tụ khi hệ số học tiến dần tới 0. Trong thực tiễn α thường được chọn là 0.1 tại mọi thời điểm t.

- **Hệ số chiết khấu**: xác định tầm ảnh hưởng của phần thưởng trong tương lai. Hệ số chiết khấu càng nhỏ (gần 0) thì tác tử sẽ hầu như chỉ quan tâm tới phần thưởng hiện tại, bỏ qua phần thưởng trong tương lai, ngược lại, hệ số càng gần 1 thì các phần thưởng trong tương lai sẽ được đánh giá cao hơn. Để giá trị của Q hội tụ thì γ phải nhỏ hơn 1. Nếu γ quá gần 1 cũng sẽ dẫn tới giá trị của Q không ổn định và việc học mất nhiều thời gian. Khi đó việc giảm giá trị γ sẽ làm tăng tốc việc học.

- **Điều kiện ban đầu** Q_0 : Do Q-learning là một thuật toán lặp nên việc chọ giá trị khởi tạo có vai trò quan trọng và đòi hỏi kinh nghiệm khi thực hiện. Với giá trị khởi tạo cao, thuật toán có tác dụng khám phá nhiều hơn.

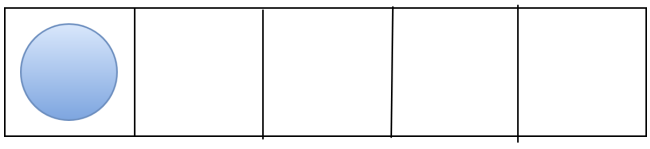
Chương 2

Bài toán nhà thám hiểm hang động

Trong chương này chúng em sẽ áp dụng phương pháp học tăng cường cho một bài toán cụ thể. Sau quá trình cài đặt thuật toán, chúng em đã thu được những kết quả khả quan, qua đó hiểu rõ hơn về mô hình quyết đinh Markov và phương pháp học tăng cường.

2.1 Mô tả bài toán

Một nhà thám hiểm khai phá một hang tối hang tối, tìm những vật giá trị chỉ có thể di chuyển theo các phương thức đã quy định [4].



Hình 2.1: Mô tả bài toán "hang động"

Một số mô tả chi tiết về bản đồ như sau:

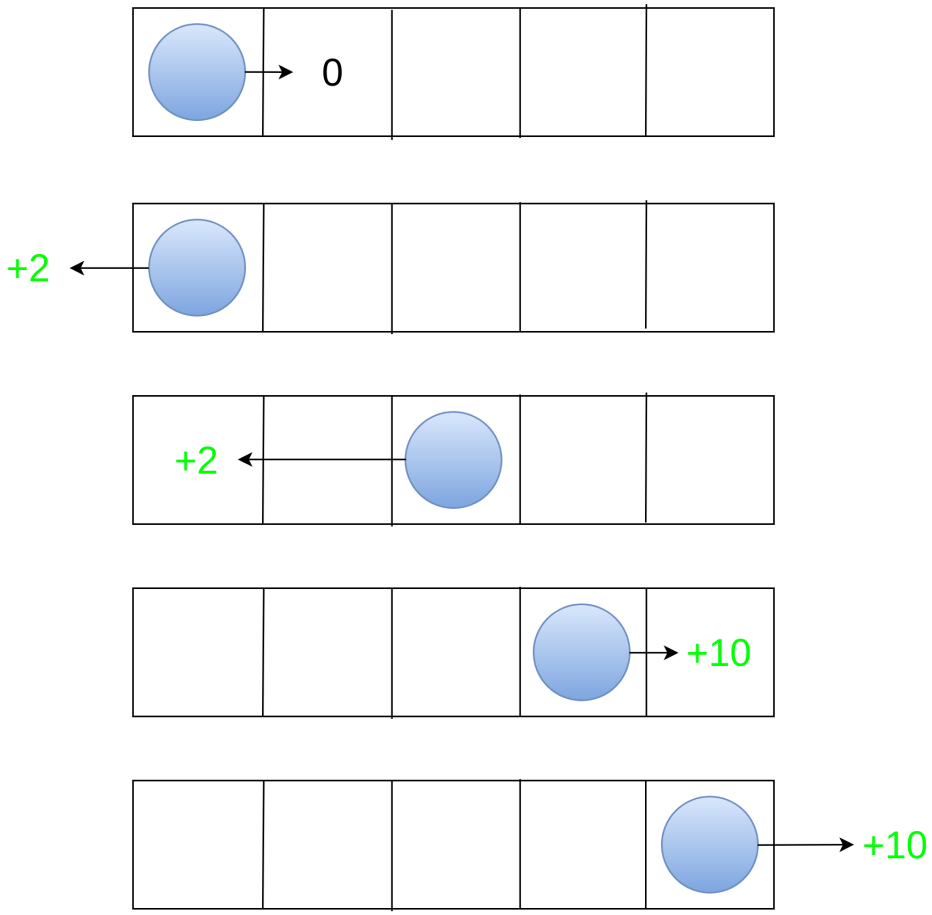
- Hang tối là một hàng ngang có N vị trí.
- Hai phương thức di chuyển duy nhất trong hang là FORWARD (tiến) và BACKWARD (lùi).
- FORWARD được quy định là tiến 1 bước theo chiều thuận.
- BACKWARD được quy định là trở lại điểm ban đầu.
- Thi thoảng có một cơn gió làm đảo lộn hành động của nhà thám hiểm.
- Nhà thám hiểm sẽ nhận được nhận giải thưởng ki đi vào một số số nhất định.
- Nhà thám hiểm ghi chép được các thông tin liên quan về các ô mình đã đi qua.

Trong giới hạn của môn học, chúng em quy định một số tham số như sau:

- Độ dài đường hầm N = 0.
- Mỗi lần quay lại ô đầu tiên sẽ nhận được +2 giá trị thưởng.
- Mỗi lần đi đến cuối đường hầm sẽ nhận được +10 giá trị thưởng.
- Các ô khác không được tính điểm.

Một trong những chiến lược đơn giản mà nhà giảm hiểm có thể lựa chọn chiến lược Accountant, được mô tả như sau là:

- Luôn ưu tiên chọn đi vào các ô cho lợi điểm thưởng cao
- Lựa chọn hành động ngẫu nhiên tại các ô cho giá trị 0



Hình 2.2: Quy chế điểm thưởng trong "hang động"

Có thể dễ dàng nhận ra nhà thám hiểm luôn ưu tiên lựa chọn BACKWARD dù cho lựa chọn tối ưu bài toán là đi về phía FORWARD, từ điểm thưởng của ô cuối cùng. Kể các khi đã chạm được vào ô cuối, nhà thám hiểm vẫn lưỡng lựa khi lựa chọn chiến thuận FORWARD do anh ấy đã chọn một chiến lược tham lam. Khi đã chọn BACKWARD và được điểm thưởng bé, anh ta quyến định sẽ luôn chọn điều đó.

Trong phần tiếp theo, chúng em sẽ sử dụng các phương pháp học tăng cường nhằm tối ưu lại chiến thuật, nhằm gặt hái giải thưởng lớn hơn.

2.2 Áp dụng học tăng cường cho bài toán

2.2.1 Q-learning

Dựa vào công thức đã có ở mục 1.3, chúng em áp dụng Q-table để ghi lại giá trị của các ô và đề ra chiến thuật "Đánh bạc" cho bài toán, chi tiết như sau:

- Lựa chọn phương pháp cho điểm thưởng cao nhất.
- Đôi lúc đánh bạc và lựa chọn nhẫu nhiên.
- Khi bảng ghi chưa có giá trị, lựa cọn giải pháp đánh bạc.
- Bắt đầu trò chơi với việc đánh bạc, kết thúc khi đã có đủ thông tin và chiến thuật cố định.

Các lựa chọn đánh bạc (ngẫu nhiên) được chúng em thêm vào do bản chất thuật toán của chúng em vẫn là một thuận toán tham lam. Các lựa chọn ngẫu nhiên được thêm vào để đảm bảo mô hình học được tất cả các cặp <Trạng thái, hành động> có thể thực hiện được, ngưỡng ngẫu nhiệu được giảm đến và bằng 0 khi mô hình đã học được cách để tối ưu khi khám phá hầm tối.

Chi tiết thuận toán cập nhật Q-table được mô tả như sau với ngưỡng ngẫu nhiên $exploration\_rate = 1$:

1: $Q(x_t, a_t) \leftarrow q\_table[action][x_t]$;
2: **if** $random < exploration_rate$ **then**
3: $\quad action \leftarrow random\_action$
4: **else**
5: $\quad action \leftarrow max(q\_table[x_t])$
6: **end if**

7: $\alpha \leftarrow q_table[action][x_{t+1}]$

8: $Q(x_t, a_t) \leftarrow Q(x_t, a_t) + \alpha \left[r_t + \gamma max_a Q(x_{t+1}, a) - Q(x_t, a_t)\right]$;

9: $q\_table[action][x_t]] \leftarrow Q(x_t, a_t)$

10: **if** $exploration\_rate > 0$ **then**

11: $\quad exploration\_rate- = exploration\_delta$

12: **end if**

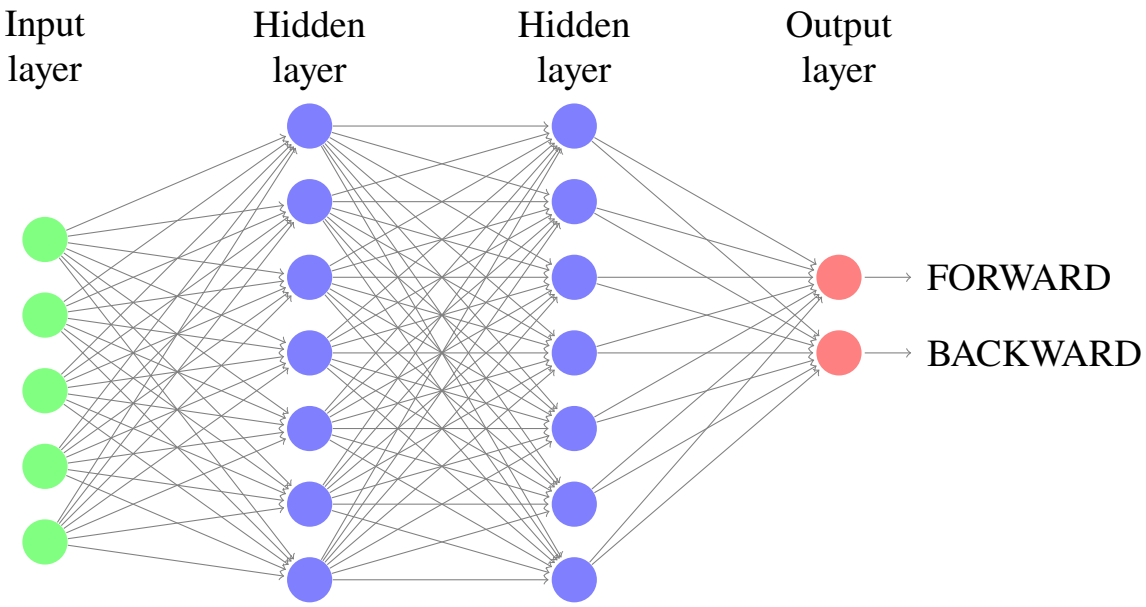
2.2.2 Deep Q-learning

Việc kết hợp các phương pháp học sâu và phương pháp học tăng cường đang ngày càng được áp dụng phổ biến [1] [2]. Với sự khó khăn trong tính toán ma trận lớn của Q_table, các mạng nơ-rơn được sử dụng để thay thế nhằm giảm khối lượng tính toán. Thay thế cho việc sử dụng giá trị của bảng Q_table, việc huấn luyện cho mạng ước lượng được các giá trị thông qua các trọng số của mạng, cấu trúc của mạng được chúng em sử dụng gồm 2 lớp mạng đầy đủ sử dụng hàm kích hoạt tuyến tính. Mô hình mạng được minh họa như Hình 2.3, với đầu vào là trạng thái dưới dạng onehot vector.

Khác với công thức học tăng cường, **Hệ số học** không còn cần thiết, thay thế vào đó là quá trình lan truyền ngược của mạng, khi đó, công thức cập nhật lại hệ số của mạng là:

$$Q(x_t) \leftarrow r_t + \gamma max(Q(x_{t+1})$$

Khi kết hợp với phương pháp học máy, dữ liệu học của mạng được thu thập lần lượt trong quá trình khám phá hang động và sử dụng trực tiếp, được gọi là các "kinh nghiệm đơn". Đầu vào của một lần học là trạng thái tương ứng, đầu ra là 2 hệ số xác xuất cho quá trình tiến hoặc lùi của nhà thám hiểm.



Hình 2.3: Mô hình mạng nơ-rơn cho thuật toán Q-learning

Từ đó, thuật toán học và khám phá hang động được mô tả như sau:

1: $Q(x_t) \leftarrow network[x_t]$;
2: **if** $random < exploration_rate$ **then**
3: $\quad action \leftarrow random\_action$
4: **else**
5: $\quad action \leftarrow max(network[x_t])$
6: **end if**
7: $Q'(x_t) \leftarrow r_t + \gamma max_{network(x_{t+1})}$;
8: $Training(Q(x_t), Q'(x_t))$
9: **if** $exploration\_rate > 0$ **then**
10: $\quad exploration\_rate- = exploration\_delta$
11: **end if**

Cách tiếp cận mà chúng em đang sử dụng và triển khai được gọi là học trực tuyến (Online learning), phương thức này khá đơn giản và phù hợp với bài toán học tăng cường

2.3 Thực nghiệm và đánh giá

Bảng 2.1: Kết quả thực nghiệm của bài toán nhà thám hiểm và đường hầm

	Phương pháp	Hệ số học	Hệ số chiết khấu	Hiệu năng	Tổng thưởng
1	**Accountant**	**x**	**x**	**1.61**	**32884**
2	Gambler	0.1	0.1	1.8	30028
2	**Gambler**	**0.1**	**0.9**	**6.6**	**56232**
4	Gambler	0.5	0.5	2.5	32972
5	Gambler	0.5	0.9	5.99	52788
6	Gambler	0.9	0.9	1.86	37612
7	Deep Gambler	x	0.1	1.84	28914
8	Deep Gambler	x	0.5	5.08	52952
9	**Deep Gambler**	**x**	**0.9**	**5.78**	**54612**

Bảng 2.1 là tổng hợp các kết quả mà chúng em đã thực hiện được. Trong đó, số vòng lặp được thực hiện là 20000 nghìn, tổng thưởng là tổng điểm thưởng trong quá trình khai phá đường hầm, hiệu năng được tính bằng trung bình điểm thưởng tính được sau mỗi 250 vòng lặp. Chúng em đã cho chạy thực nghiệm 1 phương pháp 5 lần và lấy trung bình để cho ra được kết quả như trên.

Từ kết quả trên, chúng em đã thu được các tri thức:

- Các phương pháp đề xuất Gambler và Deep Gambler của chúng em thu được kết qua vượt trôi cho với phương pháp Accountant đơn giản.
- Hệ số chiết khấu ảnh hưởng lớn đến kết quả thu được, hệ số càng cao, mô hình càng có khả năng nhìn được dài hạn giúp kết quả tốt hơn.
- Hệ số học ảnh hưởng tương đối với phương pháp Gambler.
- Hai phương pháp Gambler và Deep Gambler cho kết quả gần tương đương. Tuy nhiên đây chỉ là bài toán đơn giản, trong các bài toán phức

tạp, với Q_table là ma trận lớp, chúng em tự tin rằng học tăng cường kết hợp mạng nơ-rơn sẽ cho hiệu quả vượt trội.

Kết luận

Tóm lại, qua báo cáo cuối kì môn học "Các mô hình ngẫu nhiên", chúng em đã đạt được các kết quả như sau:

- Tổng hợp và khái quát chung phương pháp học tăng cường cho bài toán quyết định Markov
- Tìm hiểu về thuật toán Q-learning và ứng dụng trong một bài toán thực tế, mã nguồn có thể tìm thấy tại **https://github.com/ttdung997/Neural-network-and-Reinforcement-learning**
- Đề xuất sử dụng mạng nơ-rơn nhân tạo kết hợp thuật toán Q-learning coo bài toán "nhà thám hiểm và đường hầm" và thu được các kết quả khả quan trong thực nghiệm

Trong tương lai, hướng nghiên cứu tiếp theo của chúng em sẽ là nghiên cứu, áp dụng phương pháp đề xuất cho nhiều bài toán phức tạp hơn.

Tài liệu tham khảo

[1] COULOM, R. *Reinforcement learning using neural networks, with applications to motor control*. PhD thesis, Institut National Polytechnique de Grenoble-INPG, 2002.

[2] RIBEIRO, C. H. C. A tutorial on reinforcement learning techniques. In *Supervised Learning Track Tutorials of the 1999 International Joint Conference on Neuronal Networks* (1999).

[3] SUTTON, R. S., AND BARTO, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

[4] VALOHAI. Reinforcement learning tutorial, 2019.

[5] VAN OTTERLO, M., AND WIERING, M. Reinforcement learning and markov decision processes. In *Reinforcement Learning*. Springer, 2012, pp. 3–42.

[6] WATKINS, C. J., AND DAYAN, P. Q-learning. *Machine learning 8*, 3-4 (1992), 279–292.