Policy Iteration

Họ tên: Phan Anh Lộc

MSSV: 19521766

Lớp: CS106.M11.KHCL

I. Value Iteration

Value Iteration là một phương pháp tính toán tối ưu MDP policy và cho giá trị của nó.Value Iteration hoạt động từ cuối và hoạt động bằng cách tính ngược lại các giá trị ở đầu,tinh chỉnh giá trị của Q* hoặc V*. Vì bài toán trên không có điểm cuối thật sự, nên ta sẽ thành lập một điểm endpoint tùy chọn, Vk là một value function có k trạng thái để đi,QK có k trạng thái để đi. Chúng có thể được định nghĩa một cách đệ quy. Phép lặp giá trị bắt đầu với một hàm tùy ý V0 và sử dụng các phương trình sau để nhận các hàm cho k + 1 trạng thái đi từ các hàm cho k trạng thái:

$$Q_{k+1}(s,a) = \sum_{s'} P(s' \mid s,a)(R(s,a,s') + yV_k(s')) \quad \text{(k > 0)}$$

$$V_k(s) = \max_a Q_k(s,a) \quad \text{(k > 0)}$$

Nó có thể lưu mảng V [S] hoặc mảng Q [S, A]. Lưu mảng V dẫn đến việc lưu trữ ít hơn, nhưng khó xác định một optimal action hơn và cần thêm một lần lặp để xác định hành động nào dẫn đến max value.

II. Policy Iteration

Policy Iteration là một phương pháp tối ưu policy dựa trên các action và state trước đó. Giả sử chúng ta có một policy ($\boldsymbol{\pi}$: S → A) chỉ định một action cho state. Các action sẽ được chọn mỗi khi hệ thống ở state s.

- Idea: Đánh giá một policy nhất định (ví dụ: tự ý khởi tạo policy cho tất cả các state ∈ S) bằng cách tính toán value function cho tất cả các state ∈ S theo policy nhất định

$$V_{\pi}(s) = E[R(S, \pi(s), s') + \gamma V(s')]$$

Value function = kì vọng giá trị nhận được ở step đầu+ cộng với discount ở state kế tiếp

Improve policy : tìm một action tốt hơn với state s ∈ S

$$\pi_1(s) = \arg\max_{a \in A} E[R(s, a, s') + \gamma V(s')]$$

Lặp lại step 1,2 cho tới khi value function hội tụ tại optimal value function

III. Kết quả:

- **FrozenLake-v0**

Algorithm	Converged at	Successful time
Value Iteration	79	725
Policy Iteration	5	744

- **FrozenLake8x8-v0'**

Algorithm	Converged at	Successful time
Value Iteration	117	764
Policy Iteration	9	725

- **Taxi-v3**

Algorithm	Converged at	Successful time
Value Iteration	116	1000
Policy Iteration	16	1000

IV. Kết luận:

Từ kết quả chạy thục nghiệm ta có thể thấy rằng kết quả về số ván thắng (success) lẫn score của 3 map trên với 2 loại giải thuật Value Iteration và Policy Iteration là không quá chênh lệch. Bù lại thì Policy cho kết quả tốt hơn

và có thời gian hội tụ nhanh hơn so với Value Iteration, ngoài ra có một lưu ý rằng kể cả ta reset 1 map nào đó thì kết quả hội tụ vẫn không đổi, Và phân bố kết quả vẫn chưa phản ánh tốt hoàn toàn vì mỗi lần reset map giữa 2 map là độc lập(ta không đảm bảo rằng các loại map nhận được khi thực thi 2 phương pháp là giống nhau).