**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC MÁY TÍNH**

**BÁO CÁO THỰC HÀNH**

**POLICY ITERATION**

**AND VALUE ITERATION**

**Giảng viên: TS. LƯƠNG NGỌC HOÀNG**

**Sinh viên thực hiện: PHẠM MINH LONG – 19521797**

**Lớp: CS106.L21.KHCL**

**TP. Hồ Chí Minh, tháng 6 năm 2021**

1. **MỤC LỤC**

[**A.** **MỤC LỤC** 1](#_Toc73822790)

[**B.** **NỘI DUNG** 2](#_Toc73822791)

[**I.** **VALUE ITERATION** 2](#_Toc73822792)

[**1.** **Value iteration** 2](#_Toc73822793)

[**2.** **Value extraction** 2](#_Toc73822794)

[**II.** **POLICY ITERATION** 2](#_Toc73822795)

[**1.** **Policy evaluation** 2](#_Toc73822796)

[**2.** **Policy improvement** 2](#_Toc73822797)

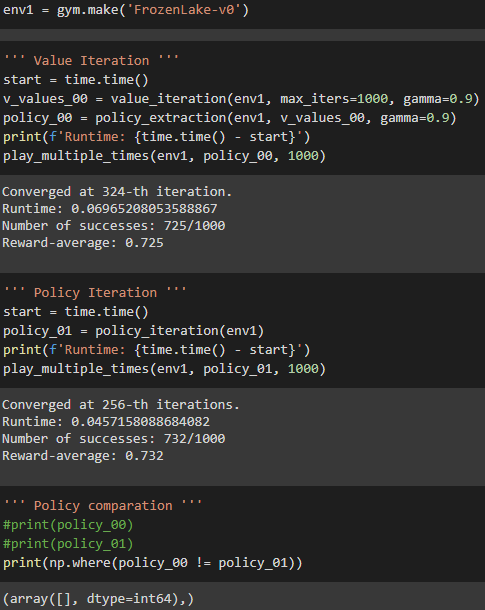
[**III.** **SO SÁNH** 3](#_Toc73822798)

[**1.** **FrozenLake-v0** 3](#_Toc73822799)

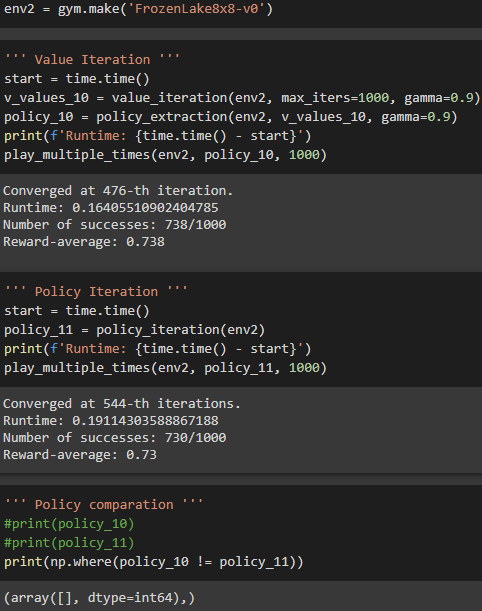
[**2.** **FrozenLake8x8-v0** 4](#_Toc73822800)

[**3.** **Taxi-v3** 5](#_Toc73822801)

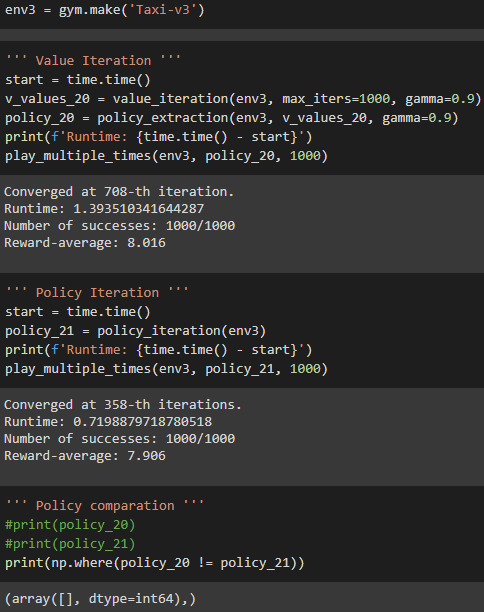
1. **NỘI DUNG**
   1. **VALUE ITERATION**
      1. **Value iteration**
      2. **Value extraction**
   2. **POLICY ITERATION**
      1. **Policy evaluation**
      2. **Policy improvement**
   3. **SO SÁNH**
      1. **FrozenLake-v0**



* + 1. **FrozenLake8x8-v0**



* + 1. **Taxi-v3**



**Nhận xét chung:**

* Policy Iteration đa số có vòng lặp tính theo độ phức tạp O(|S|) cho mỗi actions ít hơn Value Iteration càng nhiều khi số lượng trạng thái trong môi trường càng lớn (FrozenLake-v0: 256 vs 354 và Taxi-v3: 358 vs 708). Nhưng vẫn có trường hợp ở môi trường FrozenLake8x8-v0 có số lần lập của Policy Iteration cao hơn Value Iteration (544 vs 476)
* Về thời gian tìm ra policy, Policy Iteration có thời gian luôn tỉ lệ thuận cùng số vòng lập so với Value Iteration trong cả 3 môi trường.
* Số lần thành công ở 2 môi trường FrozenLake luôn do có xác suất ngẫu nhiên nên có sự giao động quanh một vài giá trị cho cả 2 giải thuật. Do 2 môi trường FrozenLake chỉ có 2 loại phần thưởng 0 khi thất lại và 1 khi thành công nên phần thưởng trung bình đạt được tương đương với số lần chơi thành công
* Ở môi trường Taxi-v3, do không bị ảnh hưởng bởi yếu tố xác suất trong các hành động nên luôn đạt tối đa số lần điểm thưởng lớn hơn 1. Taxi-v3 có giá trị phần thưởng khác nhau khai thành công do vị trí taxi và hành khách xuất hiện ngẫu nhiên, vì vậy có sự chênh lệch nhỏ giữa phần thưởng trung bình nhận được sau 1000 vòng chơi cho cả 2 giải thuật.
* Về policy, cả 2 phương pháp đều giống nhau cho cả 3 môi trường.
* Policy Iteration về tổng quan tối ưu hơn Value Iteration trên phương diện thời gian và độ phức tạp khi tính toán trong khi cả hai đều cho ra policy tối ưu như nhau.