Universidad de los Andes
Facultad de Ingeniería
Escuela de Sistemas
Departamento de Computación
Nombres: Diego Alejandro Martínez Rívas y José Gregorio Contreras Suárez
Profesor: Jesús Peréz
Asignatura: Sistemas Computacionales A2023

# Lectura 2: Markov Decision Process

Comparación de los métodos de policy iteration y value iteration en un entorno Frozen Lake.

| Método | γ | ϵ | t | π* | Nº pruebas | $R_T$ promedio |
|---|---|---|---|---|---|---|
| Policy Iterarion | 0.9 | 1e-10 | | [0. 3. 0. 3. 0. 0. 0. 0. 3. 1. 0. 0. 0. 2. 1. 0.] | 100 | 0.8 |
| Value Iteration | 0.9 | | 100 | [0. 3. 0. 3. 0. 0. 0. 0. 3. 1. 0. 0. 0. 2. 1. 0.] | 100 | 0.82 |
| Policy Iterarion | 0.8 | 1e-10 | | [2. 3. 2. 3. 0. 0. 0. 0. 3. 1. 0. 0. 0. 2. 1. 0.] | 100 | 0.46 |
| Value Iteration | 0.8 | | 100 | [2. 3. 2. 3. 0. 0. 0. 0. 3. 1. 0. 0. 0. 2. 1. 0.] | 100 | 0.45 |
| Policy Iteration | 0.7 | 1e-10 | | [1. 3. 2. 3. 0. 0. 0. 0. 3. 1. 0. 0. 0. 2. 1. 0.] | | |
| Value Iteration | 0.7 | | 100 | [1. 3. 2. 3. 0. 0. 0. 0. 3. 1. 0. 0. 0. 2. 1. 0.] | | |
| Policy Iteration | 0.6 | 1e-10 | | [1. 3. 2. 3. 0. 0. 0. 0. 3. 1. 0. 0. 0. 2. 1. 0.] | | |
| Value Iteration | 0.6 | | 100 | [1. 3. 2. 3. 0. 0. 0. 0. 3. 1. 0. 0. 0. 2. 1. 0.] | | |

Se realizaron 100 pruebas para los casos con gamma de 0.9 y 0.8, las pruebas de value iteration se realizaron con un número de 100 iteraciones, mientras que las pruebas de policy iteration con un valor de 1e-10 para epsilon hasta que convergiera.

Ambos metodos son muy similares realmente, la diferencia se encuentra en que para encontrar la politica optima con en el metodo de policy iteration es necesario hacer un procedimiento de *policy improvement* luego de encontrar los valores de los estados para un tiempo t $V_{\pi}^{t}(s)$ , mientras que en value iteration ese procedimiento ya está incluido cuando se calculan los valores optimos de la politica $V_{opt}^{t}(s)$ y la politica optima $\pi_{opt}(s)$. De las politicas obtenidas se puede observar que los dos metodos convergen a la misma cuando el valor de gamma es el mismo, y a una recompensa total promedio $R_T$ similar.

Dado que se usó la $R_T$ promedio como criterio de comparación, se puede llegar a una conclusión que tanto con policy iteration como con value iteration se puede alcanzar la convergencia a la misma politica optima, y que tan buena sea dependerá directamente del valor de gamma. Los resultados de la $R_T$ promedio dan a entender que mientras gamma sea más cercana a 1 mejor será la politica.