Control continuo con aprendizaje por refuerzo profundo

Emmanuel Peto Gutiérrez

IIMAS
UNAM

8 de diciembre de 2023

Aprendizaje por refuerzo

Agente y ambiente

Resultados

Conclusiones

La interfaz agente-ambiente

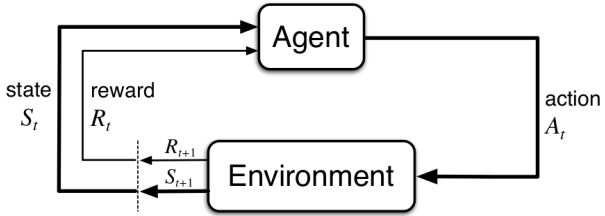


Figure 3.1: The agent–environment interaction in a Markov decision process.

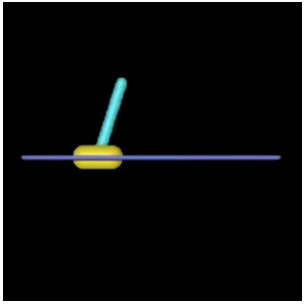
Actor-critic

Un método actor-crítico aprende las funciones de aproximación tanto para la política como para la función de valor.

- Actor: la función relacionada con la política ($\pi(a|s)$ o $\mu(s)$).
- Crítico: la función relacionada con el valor ($q(s,a)$ o $v(s)$).

El problema

Problema: encontrar una política donde las variables acción (a) y (estado) s son continuas, y probar resultados en problemas de control físico (como balancear un péndulo o manejar un carro).



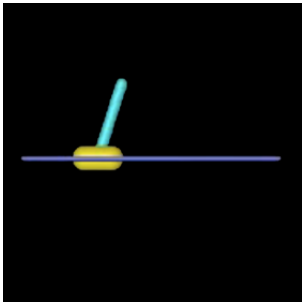
Los agentes

Para el agente se utilizaron los siguientes algoritmos:

- Deep Q-Network (DQN)
- Deep Deterministic Policy Gradient (DDPG)

El ambiente

Ambos algoritmos se probaron en el problema de cartpole, el cual consiste en balancear un péndulo moviendo el carro de manera horizontal.



DQN vs DDPG

DQN

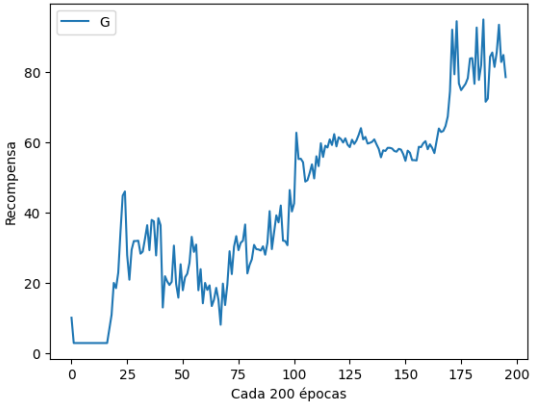
- Aproxima directamente la función Q
- Opera con acciones discretas
- Solo tiene una red que aproxima Q

DDPG

- Aproxima una política determinista que maximice la esperanza
- Opera con acciones continuas
- Tiene dos redes, la que aproxima Q y la que aproxima μ

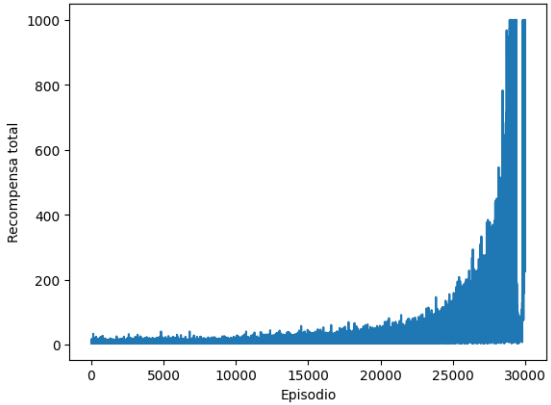
Recompensa en DDPG

Número de pasos: 40,000



Recompensa en DQN

Número de episodios: 30,000



Conclusiones

- La discretización de las acciones para usar DQN puede resultar mejor que DDPG si la dimensión de la acción es baja.
- El tiempo de cómputo de DQN puede ser mayor debido a que tiene que iterar sobre el espacio de estados discretizados.
- El agente DDPG puede no apreder correctamente si el ruido es muy alto.