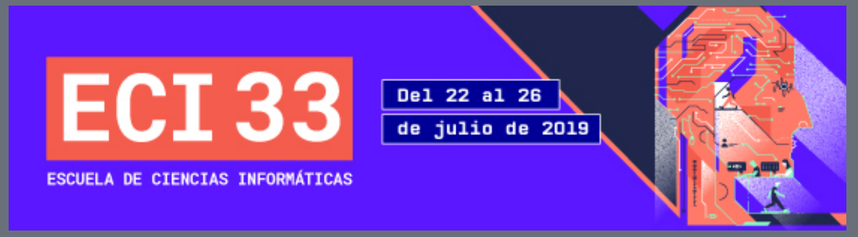
Aprendizaje Profundo por Refuerzo
01. Introducción

Dr. Juan Gómez Romero
Investigador Senior
Departamento de Ciencias de la Computación e Inteligencia Artificial
Universidad de Granada

UBA
Universidad de Buenos Aires
Argentina virtus robur et studium

UNIVERSIDAD DE GRANADA



Introducción

Aprendizaje profundo por refuerzo

UNIVERSIDAD DE GRANADA

AlphaGo (2018). Dirigida por Greg Kohs. Más: https://www.alphagomovie.com

Introducción

Curso

UNIVERSIDAD DE GRANADA

Aprendizaje profundo por refuerzo

(*deep reinforcement learning*)

Profesor: Juan Gómez Romero (PhD)

jgomez@ugr.es http://decsai.ugr.es/~jgomez

Material **https://github.com/jgromero/eci2019-DRL**

Organización del curso 3 horas / día (30’ Q&A, 2h teoría + prácticas, 30’ lab) x 5 días

1. Introducción al aprendizaje profundo (días 1, 2)
2. Aprendizaje por refuerzo (día 3)
3. Aprendizaje profundo por refuerzo (días 4, 5)

Introducción

Aprendizaje profundo por refuerzo

UNIVERSIDAD DE GRANADA

Prerrequisitos

Se recomienda contar con conocimientos sobre Álgebra, Inteligencia Artificial y Redes Neuronales.

El lenguaje de programación que se utilizará en el curso es Python.

Conceptos general de AI, ML y ANN

- Del curso "AI for Everyone", de Coursera, Week 1:
  - *Vídeo*: Machine Learning
  - *Vídeo*: The Terminology of AI
  - *Vídeo*: Non-technical explanation of Deep Learning (Part 1, Part 2)
- Del libro de Chollet “Deep Learning with Python”:
  - Capítulo 1: What is Deep Learning

Formalización de modelos de ANN

- Del libro de Russell & Norvig “Artificial Intelligence: A Modern Approach”:
  - Sección 18.7: Artificial Neural Networks (20.5 en la versión en español)

Primeros pasos en Deep Learning & PyTorch

- Del libro de Stevens & Antiga “Deep Learning with PyTorch”:
  - Capítulo 1: PyTorch from 1 Mile Away

UNIVERSIDAD DE GRANADA

Evaluación

La evaluación del curso se hará mediante:

- Examen de tipo test, a celebrar el último día del curso (viernes 26) [hasta **8.5 puntos** sobre 10]
- La entrega de un trabajo teórico-práctico, que consistirá en la resolución de un problema de optimización en un entorno virtual aplicando algoritmos de aprendizaje profundo por refuerzo [hasta **5 puntos** sobre 10].

  Se entregará el software junto a una breve memoria explicando la solución desarrollada.

La nota final del curso será la suma de ambas calificaciones. Siguiendo la normativa de la UBA, para superar el curso será necesario obtener una **nota final >= 4**.

UNIVERSIDAD DE GRANADA

Actividades prácticas

El curso tiene una orientación teórico-práctica y se desarrollará con trabajos de ejercitación y experimentación en computadoras en el laboratorio:

- **Introducción práctica a *Deep Learning***: definición, entrenamiento y validación de redes neuronales *feed-forward* utilizando PyTorch y computación en la nube, para conocer sus elementos principales, las etapas del proceso de aprendizaje y los algoritmos de optimización. (2 días)
- **Métodos de aprendizaje por refuerzo**: implementación de algoritmos fundamentales (Monte Carlo y Q-learning), para comprender los conceptos de recompensa acumulada y política de actuación. (1 día)
- **Métodos de aprendizaje profundo por refuerzo**: introducción a la implementación de algoritmos fundamentales (DQN, DDPG), para comprender el concepto de estimación de recompensa y optimización de acción, así como su cálculo utilizando redes neuronales profundas. (1.5 días)

UNIVERSIDAD DE GRANADA

Bibliografía

Deep Learning

- Y. LeCun, Y. Bengio, G. Hinton (2015) Deep Learning. Nature 521, 436-444.
- ★ I. Goodfellow, Y. Bengio, A. Courville (2016) Deep Learning. MIT Press. http://www.deeplearningbook.org
- F. Berzal (2019) Redes Neuronales & Deep Learning. [link]
- E. Stevens, L. Antiga (2019) Deep Learning with PyTorch. Manning.

Reinforcement Learning

- ★ R.S. Sutton, A.G. Barto (2018) Reinforcement Learning. MIT Press. https://mitpress.mit.edu/books/reinforcement-learning-second-edition

Deep Reinforcement Learning

- ★ A. Zai, B. Brown (2018) Deep Reinforcement Learning in Action. Manning.
- M. Morales (2018) Grokking Deep Reinforcement Learning. Manning.
- M. Pumperla, K. Ferguson (2019). Deep Learning and the Game of Go. Manning.
- OpenAI (2018) Spinning Up in Deep RL. https://spinningup.openai.com/

UNIVERSIDAD DE GRANADA

Bibliografía

DeepMind

Control (juegos Atari)

- V. Mnih et al. (2015) Human-Level Control through Deep Reinforcement Learning. Nature 518, 529-533.

Juegos con adversario e información perfecta (Go)

- D. Silver et al. (2016) Mastering the game of Go with Deep Neural Networks and Tree Search. Nature 529, 484-489.
- D. Silver et al. (2017) Mastering the game of Go without human knowledge. Nature 550, 354-359.
- D. Silver et al. (2018) A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science 362(6419), 1140-1144.

Juegos con adversario e información imperfecta (StarCraft II, Quake III)

- O. Vinyals et al. (2019) AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/
- M. Jaderberg et al. (2019) Human-level performance in 3D multiplayer games with population-based reinforcement learning. Science 364, 859-865.

Introducción

APRENDIZAJE POR REFUERZO, APRENDIZAJE PROFUNDO

UNIVERSIDAD DE GRANADA

Aprendizaje por refuerzo

Planteamiento: Un agente, situado en un entorno, debe resolver una tarea mediante experimentación repetida. El entorno proporciona al agente una recompensa positiva cuando realiza una acción “buena” y una penalización cuando realiza una acción “mala”.

Objetivo: Seleccionar acciones que maximicen la recompensa acumulada por el agente durante su vida

Dificultad: Seleccionar secuencias de acciones óptimas a largo plazo + Diversificar entre acciones que se sabe que son (en general) positivas y otras novedosas

Aprendizaje profundo por refuerzo *(deep reinforcement learning)*

Utilizar **redes neuronales profundas** para optimizar la recompensa acumulada

Introducción

Aprendizaje por refuerzo

UNIVERSIDAD DE GRANADA

Objetivo

Maximizar ganancia acumulada

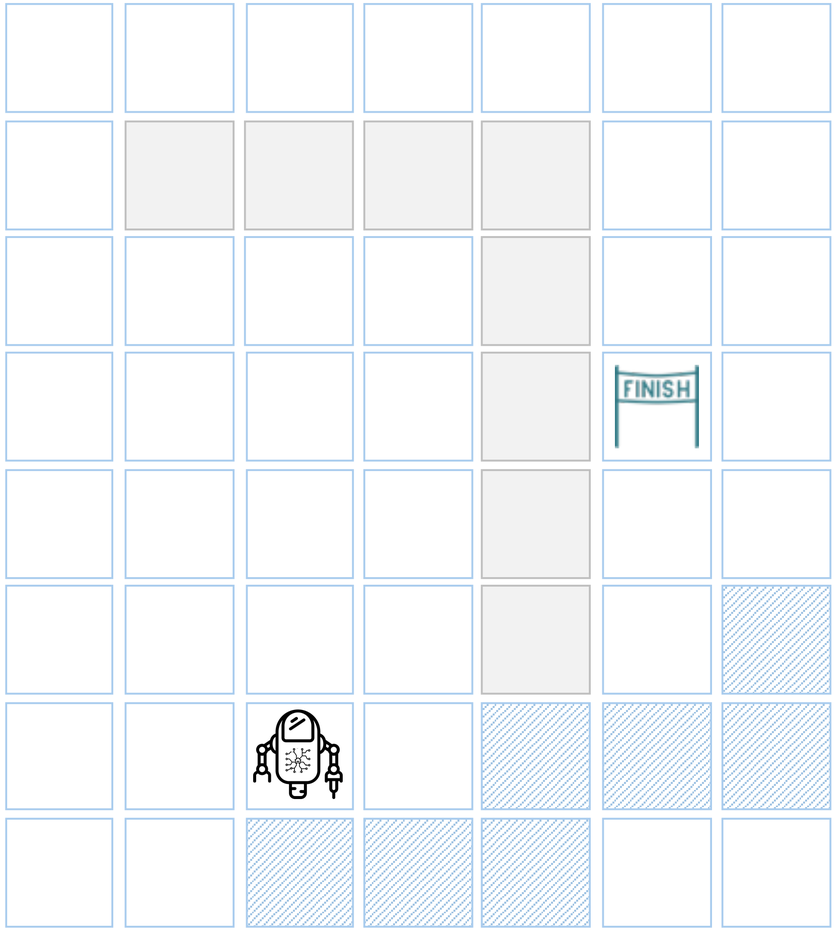
Introducción

Aprendizaje por refuerzo

UNIVERSIDAD DE GRANADA

Objetivo

Encontrar la meta en el laberinto en el menor tiempo posible



Introducción

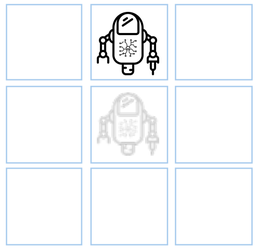
Aprendizaje por refuerzo

UNIVERSIDAD DE GRANADA

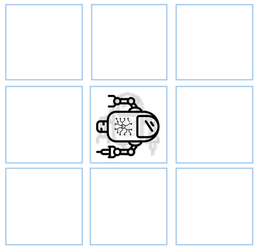
Objetivo

Encontrar la meta en el laberinto en el menor tiempo posible

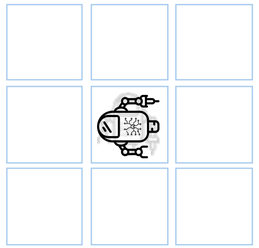
Acciones



avanzar



giro derecha



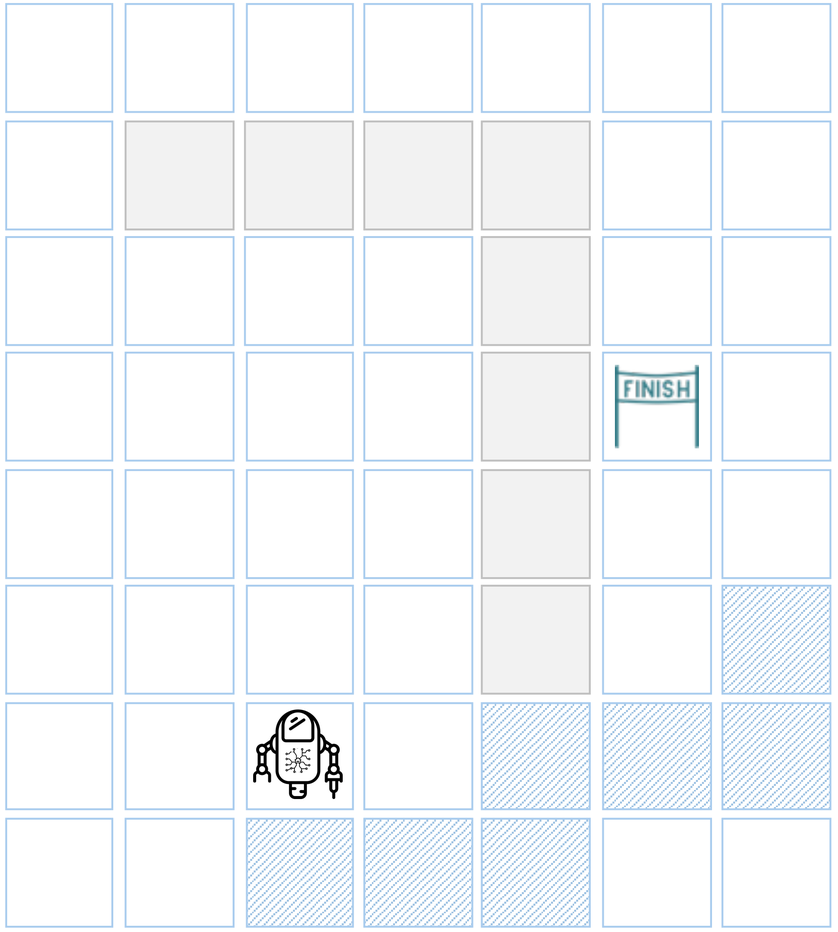
giro izquierda

Tiempo

+1

+5

no autorizado



Introducción

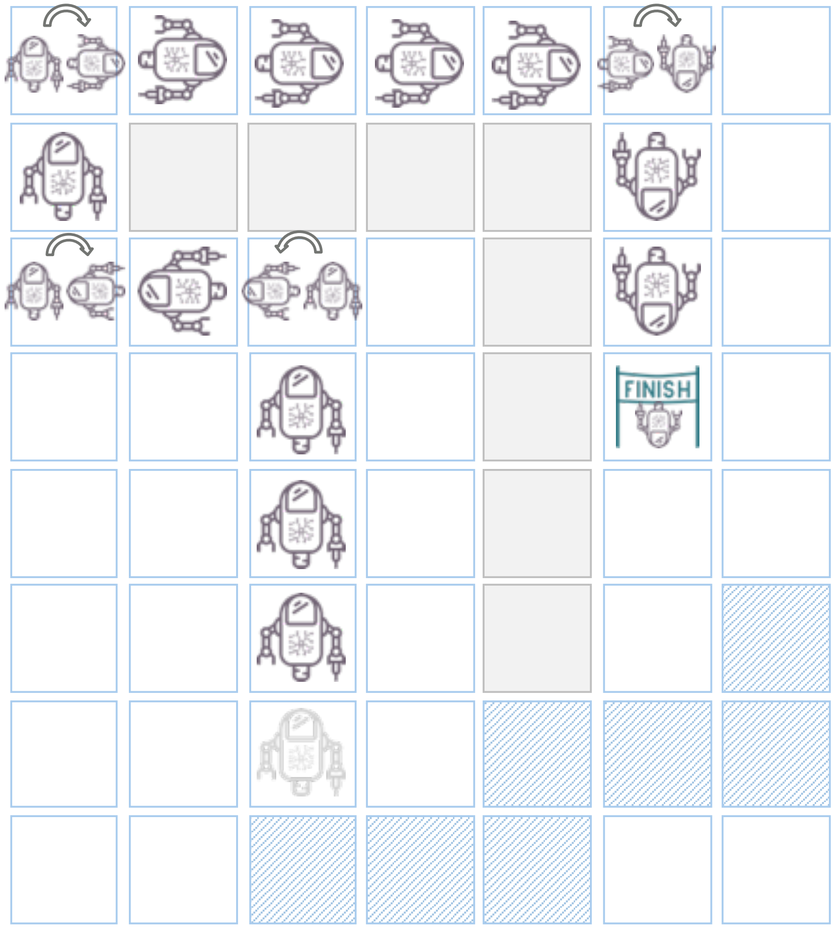
Aprendizaje por refuerzo

UNIVERSIDAD DE GRANADA

Objetivo

Encontrar la meta en el laberinto en el menor tiempo posible

Camino óptimo: 20s



Algoritmos de búsqueda óptimos

+ *Búsqueda con coste*

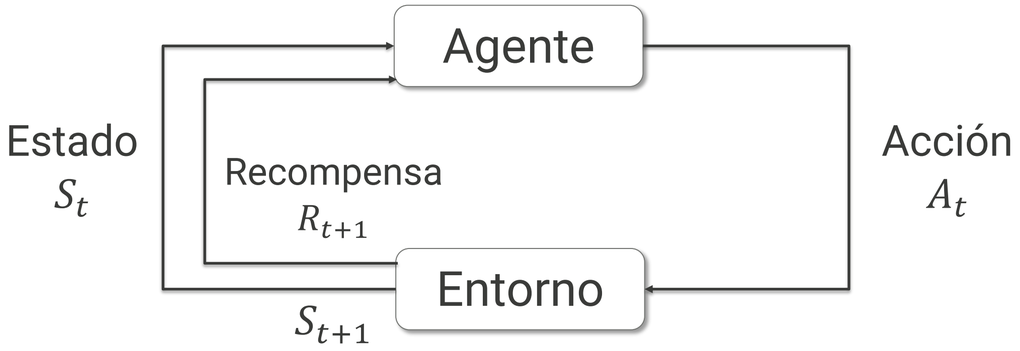
Explora todos los posibles caminos, asignando un coste según el tiempo que se tarda en realizar una acción

+ *Búsqueda con función de potencial*
+ *Búsqueda con función heurística (e.g A*)*

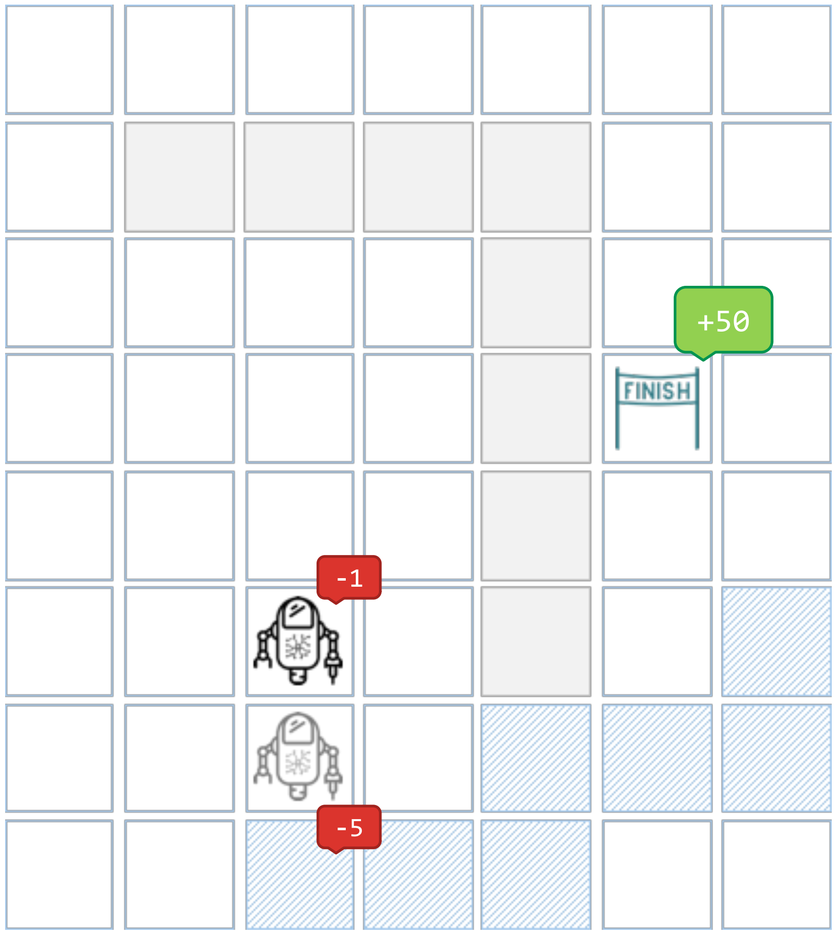
Necesitan conocer el mapa y la posición objetivo
No son viables en problemas de tamaño moderado

UNIVERSIDAD DE GRANADA

Aprendizaje por refuerzo



R.S. Sutton, A.G. Barto (2018) **Reinforcement Learning**. MIT Press.



S_t = { (6, 2), ↑ } ⟶ S_{t+1} = { (5, 2), ↑ }

A_t = **AVANZAR**

R_{t+1} = −1

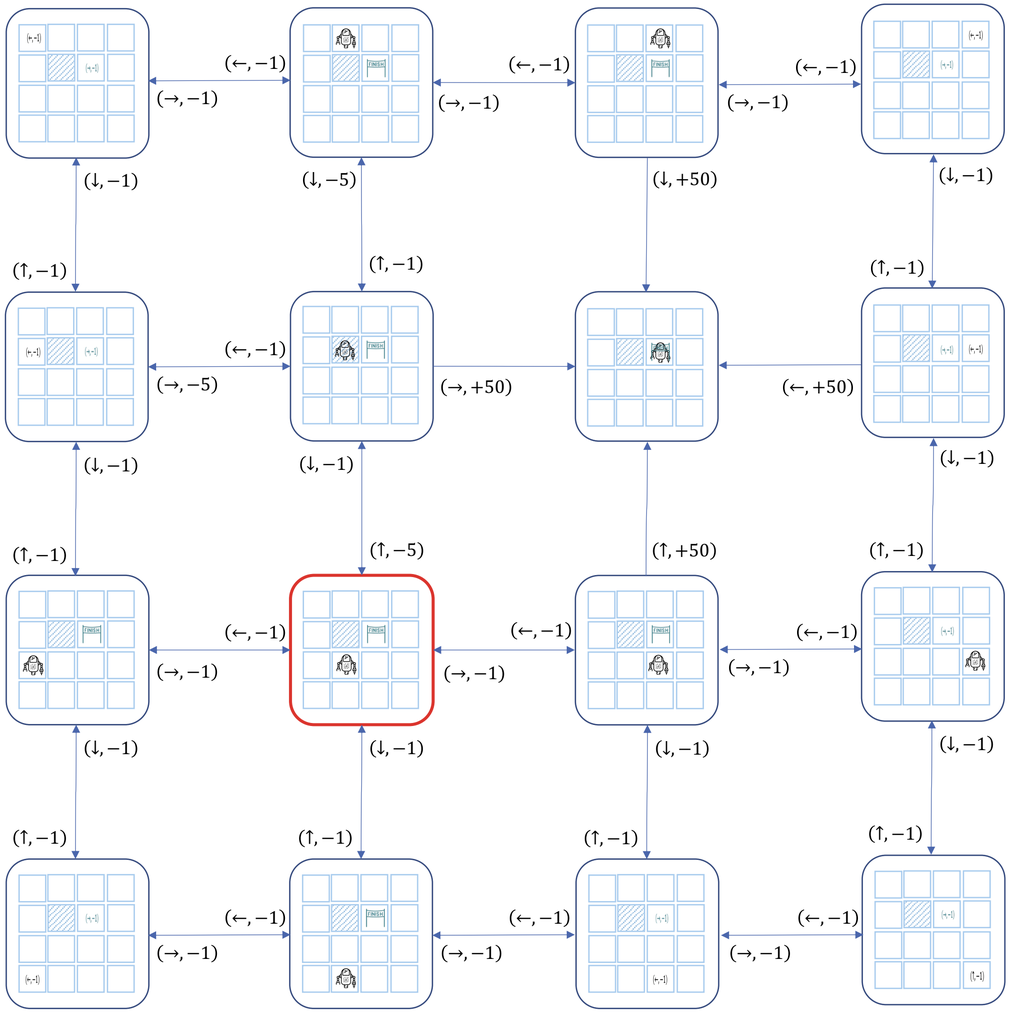
Introducción

Aprendizaje por refuerzo

UNIVERSIDAD DE GRANADA

Estados y transiciones

(robot simplificado)



Introducción

Aprendizaje por refuerzo

UNIVERSIDAD DE GRANADA

Política de actuación: π

Si normal: mover arriba
Si rayas: mover derecha

Secuencia o episodio:

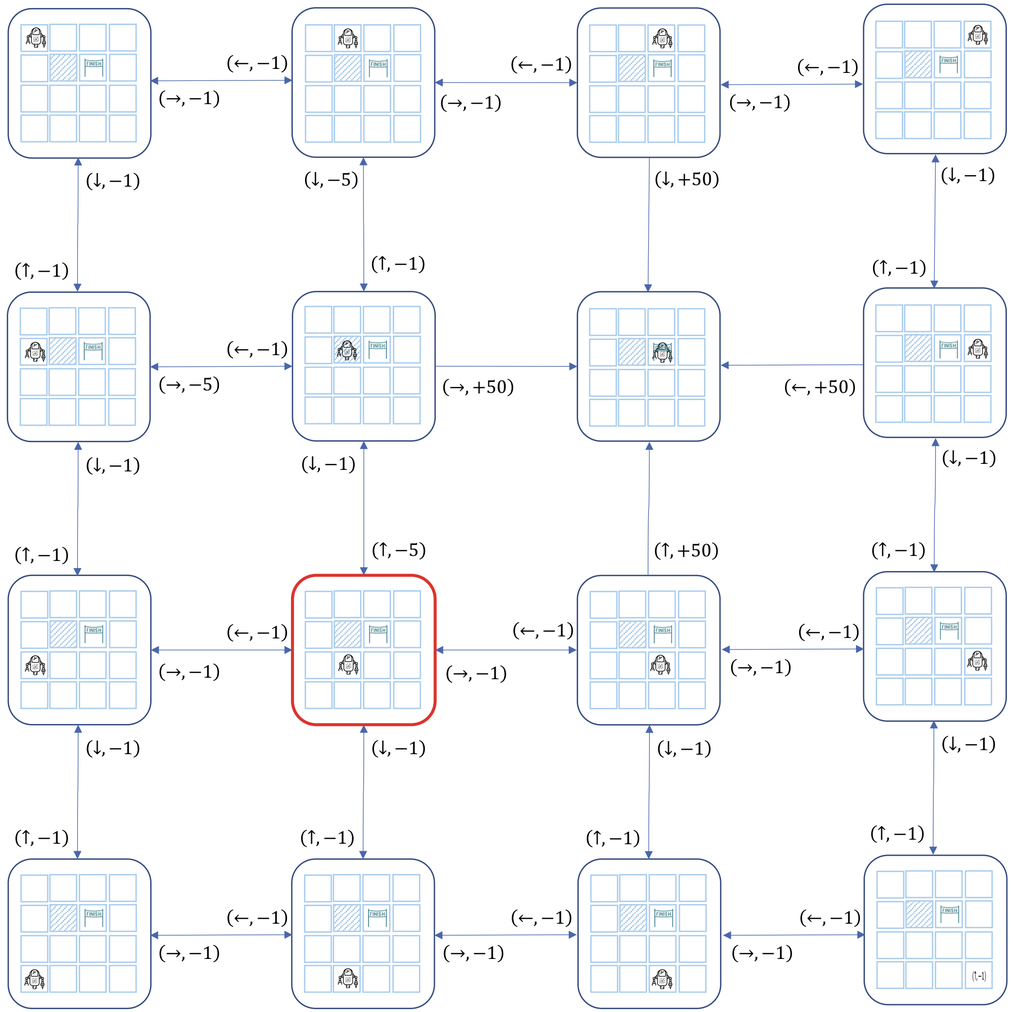
$(2,1) \uparrow -5\ (1,1)$
$(1,1) \rightarrow +50\ (1,2)$

Recompensa acumulada:

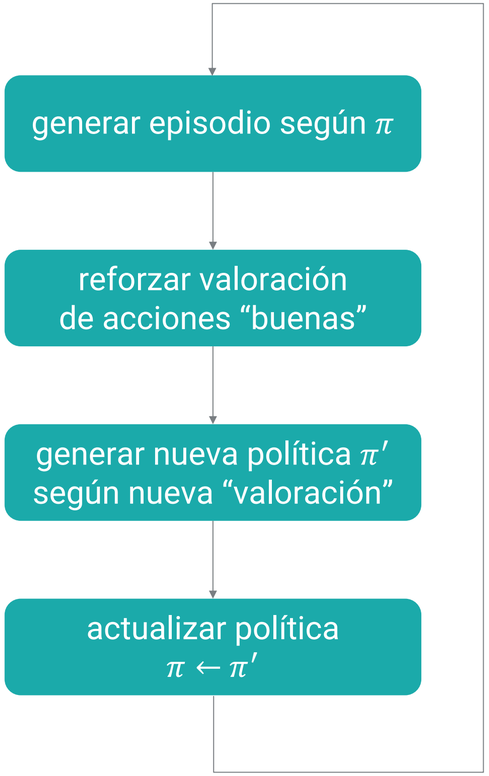
$-5 + 50 = -45$

Política óptima π_* :

Estado	Acción
$(2,1)$	$\rightarrow$
$(2,2)$	$\uparrow$
$(2,3)$	$\leftarrow$
…	



Aprendizaje



Introducción

Aprendizaje profundo por refuerzo

UNIVERSIDAD DE GRANADA

Aprendizaje profundo por refuerzo (deep reinforcement learning)

Utilizar **redes neuronales profundas** para optimizar la recompensa acumulada

↓

La red neuronal:

- devuelve la próxima acción que se debe realizar…
- … en función de una estimación de cómo de buena es un acción en el estado actual

Entrenar la red para encontrar π_*:

- A partir de ejemplos de secuencias de acciones y recompensas asociadas