

Optimización mediante algoritmos por refuerzo

Robótica



Alberto Díaz y Raúl Lara

Curso 2022/2023

Departamento de Sistemas Informáticos

License CC BY-NC-SA 4.0

Introducción

Paradigmas de aprendizaje en *Machine Learning*

Supervisado : Se aprende de ejemplos con sus correspondientes respuestas.

- Problemas de regresión y clasificación.

No supervisado : Búsqueda de patrones en datos no etiquetados.

- Problemas de *clustering*, reducción de la dimensionalidad, recodificación, ...
-

Por refuerzo : Se aprende a través de la experiencia a base de recompensas.

- Problemas de aprendizaje de políticas de decisión.
- No se le presentan ejemplos-respuestas
- La evaluación del sistema es concurrente con el aprendizaje.

"Las respuestas que producen un efecto positivo en una situación concreta aumentan la probabilidad de repetirse en dicha situación, mientras que las que producen un efecto negativo la reducen."

- Edward Thorndike - Law of Effect (1898) -

Caja de Skinner

Experimento desarrollado en 1938 por Burrus F. Skinner.

- También **cámara del condicionamiento operante**.
- ¿Animal realiza acción deseada? Recompensa
- ¿No? Penalización

Se vio que algunos comportamientos de aprendizaje son bucles observación-acción-recompensa



Aprendizaje por refuerzo (RL)

Área del *machine learning* donde **los agentes aprenden interactuando** :

- **Imita** de manera fundamental el **aprendizaje** de muchos **seres vivos** .
- Esa interacción produce tanto resultados deseados como no deseados.
- Se entrena con la **recompensa o castigo** determinados para dicho resultado.
- El agente tratará de maximizar la recompensa a largo plazo.

Se utiliza principalmente en dos áreas hoy en día:

- **Juegos** : Los agentes aprenden las reglas y las jugadas jugando¹.
- **Control** : Los agentes aprenden en entornos de simulación las mejores políticas de control para un problema determinado.

¹ Un ejemplo curioso es el publicado en <https://www.nature.com/articles/nature14236>, donde describen cómo un agente aprende a jugar a 49 juegos de Atari 2600 llegando a un nivel de destreza comparable al humano.

Terminología

Agente inteligente (agente, robot): Entidad que interactúa con el **entorno**.

Espacio de **estados** S y de **observaciones** O : Información obtenida del entorno:

- **Estado** $s_t \in S$: Descripción **completa** del estado del entorno en un instante t .
- **Observación** $o_t \in O$: Descripción **parcial** del estado del entorno en un instante t .

Espacio de acciones A : Conjunto de acciones que puede realizar el agente:

- **Discreto**: El conjunto es finito (e.g. juego del Go).
- **Continuo**: El conjunto es infinito (e.g. vehículo autónomo).

Conjunto de recompensas R : Todas las recompensas que puede recibir un agente.

- $r_t \in R$: La recompensa recibida por el agente en un instante t .



Ejemplo #1: Juego del Go

- Agente: Robot que juega al Go.
- Entorno/mundo: El tablero en el que se juega.
- Estado: Colocacion concreta de las piedras.
- Observación: Estado (sin información oculta).
- Espacio de acciones (finito): Poner piedra en una casilla vacía.



Ejemplo #2: Warcraft II

- Agente: Robot que juega al Warcraft II.
- Entorno/mundo: Pantalla en la que se juega.
- Estado: Situación de la pantalla en un momento determinado.
- Observación: Lo que el agente ve en un instante determinado (sin la niebla de guerra).
- Espacio de acciones (finito): Mover unidades, construir edificios, ...



Ejemplo #3: Coche autónomo

- Agente: Robot que conduce el vehículo.
- Entorno/mundo: El continente en el que se encuentra el vehículo.
- Estado: Estado del continente en un momento determinado.
- Observación: Lo que el agente ve por sus sensores en un instante determinado.
- Espacio de acciones (infinito): Girar el volante un determinado ángulo, aumentar y disminuir aceleración, ...

Modelo de interacción agente-entorno

El proceso de aprendizaje por refuerzo es el siguiente:



1. El agente lee un estado s_0 del entorno.
2. De acuerdo a s_0 , realiza la acción a_0 .
3. El entorno pasa al nuevo estado s_1 .
4. El agente recibe una recompensa r_1 .
5. Iterar hasta encontrar estrategia óptima

Este bucle produce una secuencia de estados, acciones y recompensas:

$$s_0, a_0, r_1, s_1, a_1, \dots$$

Markov Decision Processes (MDP)

Propiedad de Márkov

El estado futuro de un proceso depende del estado actual, y no de los anteriores.

- Es un estado que cumplen ciertos procesos estocásticos.
- Definida por Andréi Markov en 1906 en su Teoría de Cadenas de Márkov².

Al proceso que satisface esta propiedad se denomina **Proceso de Márkov**.

- Concretamente se denominan Procesos de Márkov de **primer orden**.
- La definición se puede extender a n estados anteriores (proceso de orden n).

Si hay que quedarse con algo, nos dice que nuestro agente sólo necesita el estado actual para decidir qué acción tomar.

² Más información en https://en.wikipedia.org/wiki/Markov_chain.

Procesos de decisión de Márkov (MDP)

Proceso **estocástico** de **tiempo discreto** que satisface la **propiedad de Márkov**.



Matemáticamente se define como una 4-tupla (S, A, P_a, R_a) donde:

- S y A : Espacios de estados y de acciones del proceso respectivamente.
- $P_a(s, s')$: Probabilidad de que la acción a nos lleve de s a s' .
- $R_a(s, s')$: Recompensa inmediata por pasar del estado s al estado s' con la acción a .

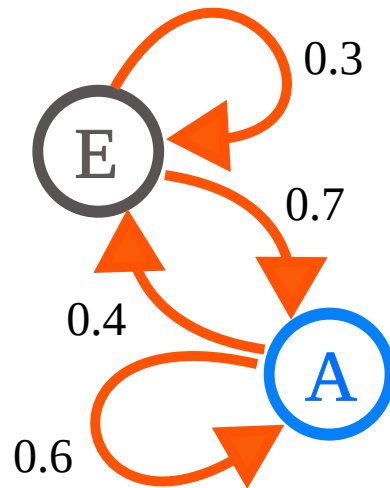
A la función $\pi : S \rightarrow A$ que define las políticas de decisión se le denomina **policy**.

Diferencia entre un MDP y Cadena de Márkov

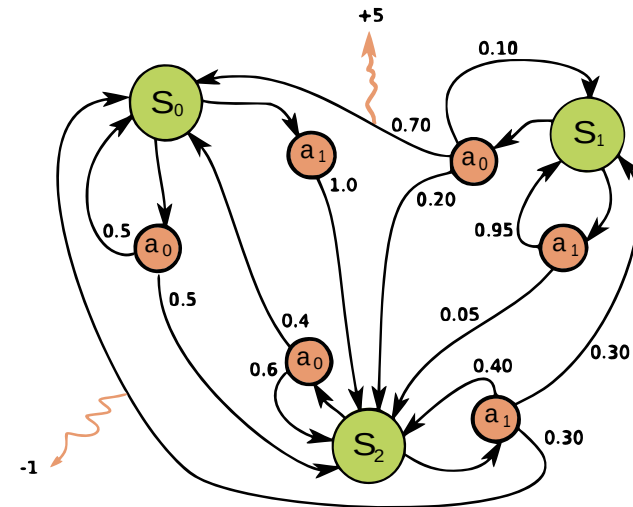
Los MDP extienden a las cadenas de Márkov en dos aspectos:

- Permiten elegir **acciones** para realizar transiciones entre estados.
- Incluyen **recompensas** a una o más de esas transiciones.

Cadenas de Márkov



MDP



Recompensas y tomas de decisiones

Hipótesis de la recompensa

El agente quiere **maximizar la recompensa acumulada** (rendimiento esperado).

- Recompensa: *Feedback* que recibe el agente para saber si la acción es buena o no.

Recompensa acumulada: Suma de todas las recompensas de la secuencia.

$$R(\tau) = \sum_{i=0}^{\infty} \gamma^i r_{t+i+1} = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots$$

Sin embargo, las recompensas no tienen por qué tener todo su valor siempre.

- De ahí el **factor de ajuste** $\gamma \in [0, 1]$ aunque se le aplica a la recompensa.
- Las recompensas a corto plazo tienen más probabilidades de suceder.
- γ indica si interesan más recompensas a **corto** ($\gamma \approx 0$) o a **largo** ($\gamma \approx 1$) **plazo**.

Función de políticas de decisión

La función de *policy* (π) es la que **asigna** una **acción** $a \in A$ a cada **estado** $s \in S$.

- Realiza el mapeo entre el espacio de estados y el de acciones.
- Define completamente el comportamiento de un agente.

Buscamos π que **maximice el rendimiento esperado** ; existen dos métodos:

- **Directo** : ¿Qué acción debe realizar en el estado actual?
- **Indirecto** : ¿Qué estados son mejores para tomar la acción que lleva a esos estados?

Métodos directos (basados en políticas)

En estos métodos **aprendemos directamente la función π** . Existen dos tipos:

Determinista

Devuelve **siempre la misma acción** para un estado determinado.

$$\pi(S) = A$$

Por ejemplo:

$$\pi(s_t) = \{\blacktriangleright\}$$

No determinista

Devuelve una **distribución de probabilidad** sobre las acciones.

$$\pi(S) = P[A|S]$$

Por ejemplo:

$$\pi(s_i) = \{(\blacktriangleleft, 0.3), (\blacktriangleright, 0.5), (\blacktriangledown, 0.1), (\blacktriangle, 0.1)\}$$

Métodos indirectos (basados en valores)

Aprendemos una función v_π (o q_π) que **relaciona un estado con su valor estimado**.

- Valor: Recompensa acumulada si empieza en ese estado y se mueve al mejor estado.
- El agente selecciona la acción de mayor valor.

Valor estado

$$v_\pi(s_t) = E_\pi[r_{t+1} + \gamma v_\pi(s_{t+1})]$$

Valor par estado-acción

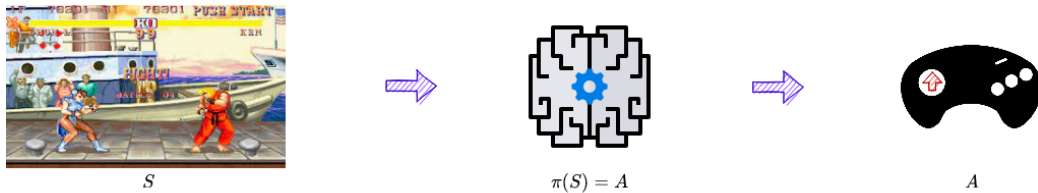
$$q_\pi(s_t, a_t) = E_\pi[r_{t+1} + \gamma q_\pi(s_{t+1}, a_{t+1})]$$

Independientemente de la función elegida, el resultado será la recompensa esperada.

- Ojo: **Para** calcular **cada valor de un estado** (o par estado-acción), hay que **sumar todas las recompensas** que puede obtener un agente si empieza en ese estado.

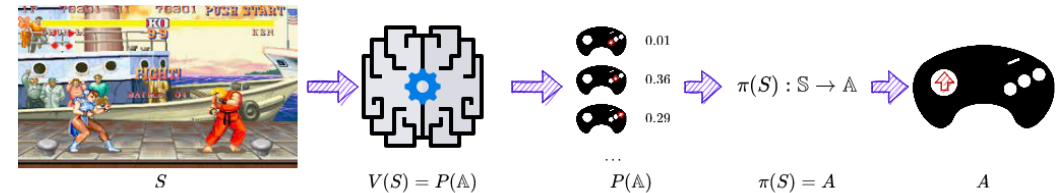
Comparativa entre métodos directos e indirectos

Métodos directos



La **política óptima** se encuentra **entrenando** la política **directamente**.

Métodos indirectos



Encontrar una **función de valor óptima** lleva a tener una **política óptima**.

Por lo tanto Independientemente del método, tendremos una política.

- Pero en el caso de los métodos basados en valores no la entrenamos.
- Será una "simple" función que usará los valores dados por la función v_π o q_π .

Estrategia epsilon-greedy

Política sencilla para elegir acción que mantiene el equilibrio exploración/explotación.

- El agente elige una acción de forma aleatoria con probabilidad ϵ
- La mejor acción conocida con probabilidad $1 - \epsilon$.

Por lo general, se empieza con un **épsilon alto** (much **exploración**).

- Según el agente aprende más, **epsilon disminuye**, aumentando la **explotación**.

Tareas y problemas

Se entiende por tarea a una instancia de un problema. Existen dos tipos diferenciados:

Tenemos dos tipos bien diferenciados de tareas:

- **Episódicas** : Poseen estado inicial y terminal o final (e.g. Sonic the Hedgehog).
- **Continuas** : Tarea que no posee estado terminal (e.g. vehículo autónomo).

Ecuación de Bellman

Simplifica el cálculo del **valor** del estado o del par estado-acción.

ME HE QUEDADO AQUÍ

Gymnasium

OpenAI

Empresa dedicada al I+D de sistemas de Inteligencia Artificial

- Fundada en 2015 por Elon Musk¹ y Sam Altman², y financiada por muchas empresas³.
- Misión: Garantizar que la Inteligencia Artificial General (AGI)⁴ beneficia al ser humano.

Han desarrollado múltiples soluciones basadas en IA, entre las que destacan:

- Dall-e: Generador de imágenes de alta calidad a partir de texto.
- GPT-3: Generador de texto que simula la redacción humana.
- *OpenAI Gym*: Entorno de pruebas para IA.

¹ Cofundador, entre otros, de [Tesla Inc.](https://www.tesla.com/) (<https://www.tesla.com/>) y [SpaceX](https://www.spacex.com/) (<https://www.spacex.com/>)

² Cofundador, entre otros, de [Y Combinator](https://www.ycombinator.com) (<https://www.ycombinator.com>)

³ Micro\$oft invirtió más de 1000 millones de dólares para colaborar con Azure. Más información en: <https://learn.microsoft.com/azure/cognitive-services/openai/concepts/models>

⁴ [Wikipedia](https://es.wikipedia.org/wiki/Inteligencia_artificial_general) (https://es.wikipedia.org/wiki/Inteligencia_artificial_general)

El entorno *Gymnasium*⁵

Biblioteca la que investigar algoritmos de aprendizaje por refuerzo.

- Bueno, y más areas como algorítmica, teoría de juegos, investigación operativa, ...
- Es un *fork* de *OpenAI Gym* que mantiene por un equipo externo a *OpenAI*.
- Puede considerarse (de momento) un *drop-in replacement* de la original.

Proporciona una API estándar para comunicar algoritmos de RL y entornos

- Ofrece entornos ya preparados y mecanismos para crear entornos personalizados.

Se encarga de proporcionar toda la información que el agente necesitaría:

- Entorno, posibles acciones y recompensas, estado actual, ...
- Sólo tenemos que preocuparnos de la lógica del agente.

⁵ Información sobre la biblioteca disponible en <https://gymnasium.farama.org/>.

Instalación

La biblioteca está disponible a través de Pypi como `gymnasium` :

- En algunos entornos será necesario instalar otras bibliotecas (e.g. `box2d`).
- Solo está soportado en GNU/Linux y macOS. Para Windows hay que usar el WSL.
- Y para sacar representaciones gráficas, el Xming Server.

El "Hola, mundo!" de rigor

```
import gymnasium as gym

env = gym.make('CartPole-v1', render_mode='human')
obs = env.reset()

print(f'Training in environment "{env}" (shape: {env.observation_space.shape})')
print(f'- {epochs} epochs ({iterations} iterations per epoch)')
print(f'- Actions: {env.unwrapped.get_action_meanings()}')

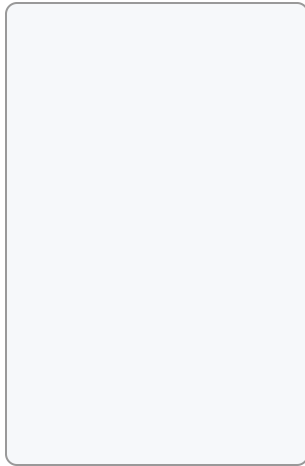
for ep in range(epochs):
    total_reward = 0
    for _ in range(iterations):
        env.render()
        action = env.action_space.sample() # Aquí la lógica del agente
        obs, reward, done, info = env.step(action)
        total_reward += reward
        if done:
            break
    print(f'Epoch {ep}: Total reward: {total_reward}')
```


- **Probabilidad de transición**: $P(S_{t+1} | (S_t, A_t), (S_{t-1}, A_{t-1}), \dots, (S_1, A_1))$
- **Probabilidad de recompensa**: $P(R_{t+1} | (S_t, A_t), (S_{t-1}, A_{t-1}), \dots, (S_1, A_1))$

Q-Learning

Es un método indirecto (método basado en valores)

Ajuste en la recompensa



- El agente se mueve una celda cada t .
- Los zombies también.
- Objetivo: Conseguir el mayor número de puntos de supervivencia...
- Si nos arañan perdemos.

Propiedad de Márkov

El estado futuro del proceso depende del estado actual, y no de los anteriores.

- Es un estado que cumplen ciertos procesos estocásticos.
- Definida por Andréi Markov en 1906 en su Teoría de Cadenas de Márkov.

Al proceso que satisface esta propiedad se denomina **Proceso de Márkov**.

- Concretamente se denominan Procesos de Márkov de **primer orden**.
- La definición se puede extender a n estados anteriores (proceso de orden n).
- Si el espacio de estados es finito, equivale a una **cadena de Márkov**.

Si hay que quedarse con algo, nos dice que nuestro agente sólo necesita el estado actual para decidir qué acción tomar.

Se asume que el proceso de decisión de un agente es un MDP:

- $P(s_{t+1} | (s_t, a_t), (s_{t-1}, a_{t-1}), \dots, (s_1, a_1)) = P(s_{t+1} | (s_t, a_t))$
- $P(r_{t+1} | (s_t, a_t), (s_{t-1}, a_{t-1}), \dots, (s_1, a_1)) = P(r_{t+1} | (s_t, a_t))$

Los procesos de decisión de un

Sistemas de toma de decisiones basados en procesos de Márkov. Incluyen:

- S : Conjunto finito de estados.
- A : Conjunto finito de acciones.
- $P(s_i | (s_j, a))$: Probabilidad de transición de s_i a s_j con la acción a .
- $\pi : S \rightarrow A$: Función que define las políticas de decisión.
-
- **Transiciones** entre estados.
- **Recompensas** por transición. Pueden ser positivas o negativas.
- Factor de descuento $\gamma \in [0, 1]$: Importancia entre recompensas inmediatas o futuras 41 / 60
(generalmente γ^t)

MDP en nuestro ejemplo

El objetivo del superviviente es intentar maximizar la suma de las recompensas futuras tomando la mejor acción para cada estado:

$$\sum_{t=0}^{\infty} r_{e_t, a_t} \cdot \gamma^t$$

Explicado:

1. Estamos sumando para cada paso de tiempo t , de ahí el sumatorio.
2. Cada paso de tiempo tiene una recompensa r_{e_t, a_t} asociada la acción tomada.
3. γ^t es el factor de descuento en 1 por ahora y olvidémonos de ello.

Una vez formalizado el problema, vamos a explorar algunas soluciones.

Solución #1: Q-learning

Se apoya en una función denominada acción-valor (*action-value*) o función Q :

- Entrada: Estado y acción a realizar.
- Salida: Recompensa esperada de esa acción (y de todas las posteriores).

La función Q se actualiza de forma iterativa:

1. Antes de explorar el entorno, Q da el mismo valor fijo (arbitrario).
2. Según se explora, aproxima mejor el valor de la acción a en un estado s .
3. Según se avanza, la función Q se actualiza.

Representa suma de las recompensas de elegir la acción Q y todas las acciones óptimas posteriores.

$$Q(e_t, a_t) = Q(e_t, a_t) + \alpha \cdot (r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t))$$

Realizar a_t en el estado e_t actualiza su valor con un término que contiene:

- α : Lo "agresivo" que estamos haciendo el entrenamiento.
- r_t : Estimación que obtuvimos al actuar en el estado e_t anteriormente.
- $\max_a Q(s_{t+1}, a)$: Recompensa futura estimada.
- Se resta además el valor antiguo para incrementar o disminuir la diferencia en la estimación.

Ahora tenemos una estimación de valor para cada par estado-acción.

- Con el podemos elegir la acción que nos interesa (e.g. usando epsilon-greedy)

Solución #2: *Policy learning*

Trata de determinar una función π asigna la mejor acción a un estado dado:

$$a = \pi(e)$$

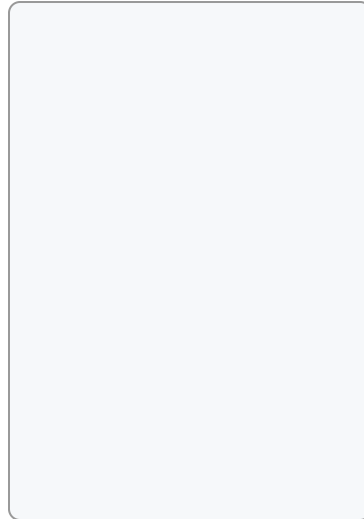
"Cuando observo el estado e , lo mejor que puedo hacer es tomar la acción a "

Esta función es una función compleja que tratamos de aproximar.

- Y lo más "sencillo" y rápido es usar redes neuronales para ello.

Ejemplo: Hambre y zombies

Objetivo: Utilizar técnicas de RL para que el superviviente llegue a su destino.



Hay que comenzar considerando los estados, las acciones y las recompensas.

Estados

El agente se encuentra en un estado y toma una acción de acuerdo a este.

Espacio de estados: Todas las situaciones posibles en las que se puede encontrar el agente.

- Debe contener información suficiente para tomar una decisión correcta.

En el ejemplo, son todas las posiciones que podría ocupar el agente (35).

- Podríamos complicarlo más, por ejemplo, obligando a llevar comida.
- Esto implicaría los 35 estados con y sin comida encima ($35 + 35 = 70$).
- Pero nos quedaremos con el ejemplo simple.

Acciones

El agente se encuentra con uno de los 35 estados y realiza una acción.

- 5 acciones posibles: arriba, abajo, izquierda, derecha y coger comida.





Espacio de acciones : Conjunto de todas las acciones posibles para un estado.

Recompensas

El superviviente está motivado por la recompensa, así que aprenderá a:

- Encontrar la comida y el objetivo.
- Evitar las zonas infestadas de zombies.

Algunos puntos a tener en cuenta para el agente:

- Alta recompensa por llegar a las montañas  (+1000); es el objetivo.
- Ligera recompensa por encontrar comida   (+10) porque está bien.
- Penalización si llega a un zombie  (-50) porque no interesa en absoluto.

Es importante tener en cuenta que la recompensa no siempre es inmediata:

- Puede haber tramos sin nada hasta llegar a un estado muy bueno.

Otras soluciones

Deep Q-networks (DQN)

Son aproximaciones de funciones Q utilizando redes neuronales profundas².

Asynchronous Advantage Actor-Critic (A3C)

Es una combinación de las dos técnicas anteriores³, combinando:

- Un actor: Red de políticas de actuación que deciden qué acción tomar.
- Un crítico: DQN que decide el valor de cada acción a tomar.

² <https://www.nature.com/articles/nature14236>

³ <https://proceedings.mlr.press/v48/mniha16.html>

▶ 0:00 / 1:02



Relevancia del aprendizaje por refuerzo

"El Go es un juego estudiado por los humanos durante más de 2500 años. AlphaZero, en un tiempo insignificante (3 días), pasó de conocer sólo las reglas del juego a vencer a los mejores jugadores del mundo, superando todo nuestro conocimiento acumulado durante milenios. Ningún campo del aprendizaje automático ha permitido avanzar tanto en este tipo de problemas como el aprendizaje por refuerzo."

Relevancia del aprendizaje por refuerzo hoy en día

Podemos decir que es prácticamente el único paradigma de aprendizaje:

- Capaz de aprender comportamientos complejos en entornos complejos.
- Que ha podido hacerlo prácticamente sin supervisión humana.

Ofrece a la robótica forma abordar cómo diseñar comportamientos difíciles.

- Que por otro lado, son prácticamente todos.
- Las cosas fáciles para un humano suelen ser las más complejas de diseñar.

Permite a robots descubrir de forma autónoma comportamientos óptimos:

- No se detalla la solución al problema, sino que se interacciona con el entorno.
- La retroalimentación de el efecto sobre el entorno permite aprender.

La utilidad de los modelos aproximados

Los datos del mundo real pueden usarse para aprender modelos aproximados.

- Mejor, porque el proceso de aprendizaje por ensayo y error es muy lento.
- Sobre todo en un sistema que tiene que hacerlo en un entorno físico.
- Las simulaciones suelen ser mucho más rápidas que el tiempo real.
- Y también también mucho más seguras para el robot y el entorno
- ***Mental rehearsal***: Describe el proceso de aprendizaje en simulación.

Suele ocurrir que un modelo aprende en simulación pero falla en la realidad:

- Esto se conoce como **sesgo de simulación**.
- Es análogo al sobreajuste en el aprendizaje supervisado.
- Se ha demostrado que puede abordarse introduciendo modelos estocásticos.

Impacto del conocimiento o información previa

El conocimiento previo puede ayudar a guiar el proceso de aprendizaje:

- Este enfoque reduce significativamente el espacio de búsqueda.
- Esto produce una **aceleración** dramática **en el proceso de aprendizaje**.
- También **reduce la posibilidad de encontrar mejores óptimos**¹.

Existen dos técnicas principales para introducir conocimiento previo:

- A través de la **demostración**: Se da una política inicial semi-exitosa.
- A través de la **estructuración de la tarea**: Se da la tarea dividida.

¹ Alpha Go fue entrenado con un conocimiento previo de Go, pero Alpha Go Zero no sabía nada del juego. El resultado fue que Alpha Go Zero jugó y ganó a Alpha Go en 100 partidas.

Desafíos del aprendizaje por refuerzo

La maldición de la dimensionalidad : El espacio de búsqueda crece exponencialmente con el número de estados.

La maldición del mundo real : El mundo real es muy complejo y no se puede simular.

- Desgaste, estocasticidad, cambios de dinámica, intensidad de la luz, ...

La maldición de la incertidumbre del modelo : El modelo no es perfecto y no se puede simular.

- Cada pequeño error se acumula, haciendo que conseguir un modelo suficientemente preciso del robot y su entorno sea un reto

La pregunta central de la filosofía moral es: ¿qué debemos hacer?

- ¿Cómo debemos vivir? ¿Qué acciones son correctas o incorrectas?
- Una posible respuesta es que, claramente, depende de los valores de cada uno.

A medida que vayamos creando una IA cada vez más avanzada, ésta empezará a salir de los problemas donde la recompensa se define mediante un número de puntos ganados en el juego, y requerirá recompensas más complejas.

- Los vehículos autónomos, por ejemplo, son agentes que tienen que tomar decisiones con una definición de recompensa algo más compleja
- Al principio, la recompensa podría estar ligada a algo como "llegar a salvo al destino".
- Pero ¿y si se ve obligado a elegir entre mantener el rumbo y atropellar a cinco peatones o desviarse y atropellar a uno? ¿debe desviarse o incluso dañar al conductor con una maniobra peligrosa? ¿Y si el único peatón es un niño, o un anciano, o el próximo Einstein o Hitler? ¿Cambia eso la decisión? ¿por qué? ¿Y si al dar un volantazo también se destruimos una escultura extremadamente valiosa e irremplazable?
- De repente tenemos un problema mucho más complejo cuando intentamos definir la función objetivo, y las respuestas no son tan sencillas.

¡GRACIAS!