

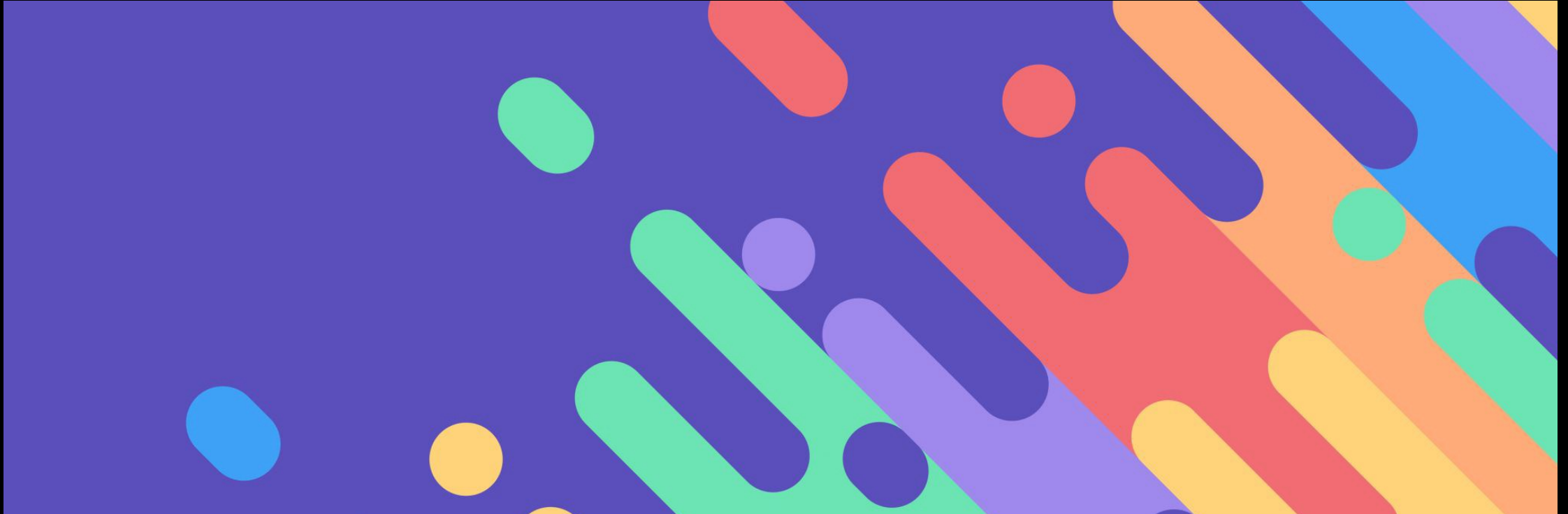
APRENDIZAJE POR REFUERZO



Inteligencia Artificial

CEIA - FIUBA

Dr. Ing. Facundo Adrián
Lucianna

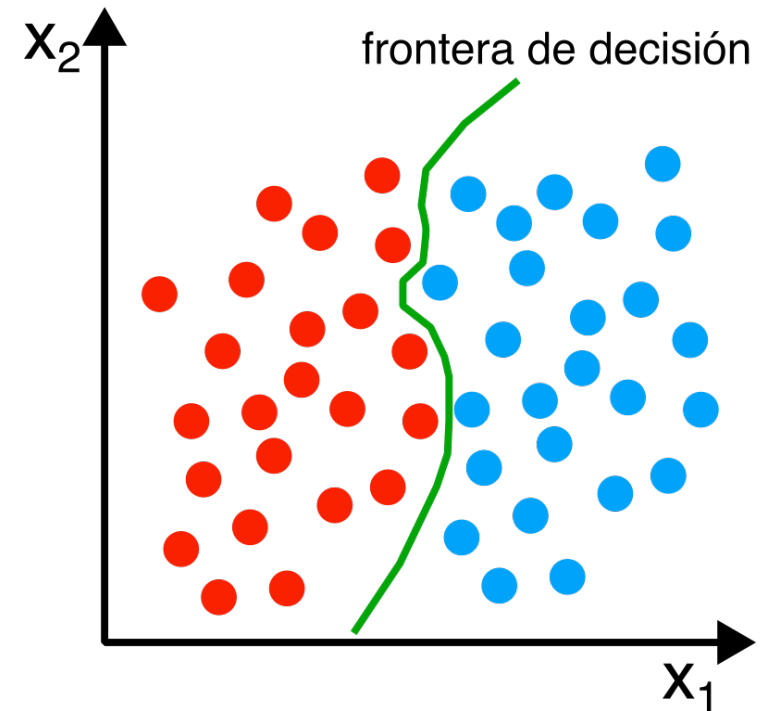


LO QUE VIMOS LA CLASE ANTERIOR...

CLASIFICACIÓN

Es más común encontrarnos con problema de clasificación que de regresión:

- Una persona llega a una guardia con un set de síntomas atribuidos a una de tres condiciones médicas.
- Un servicio de banca online debe determinar si una transacción en el sitio es fraudulenta o no, usando como base la dirección IP, historia de transacciones, etc.
- En base a la secuencia de ADN de un número de pacientes con y sin una enfermedad dada, un genetista debe determinar que mutaciones de ADN genera un efecto nocivo relacionado a la enfermedad o no.



REGRESIÓN LOGÍSTICA

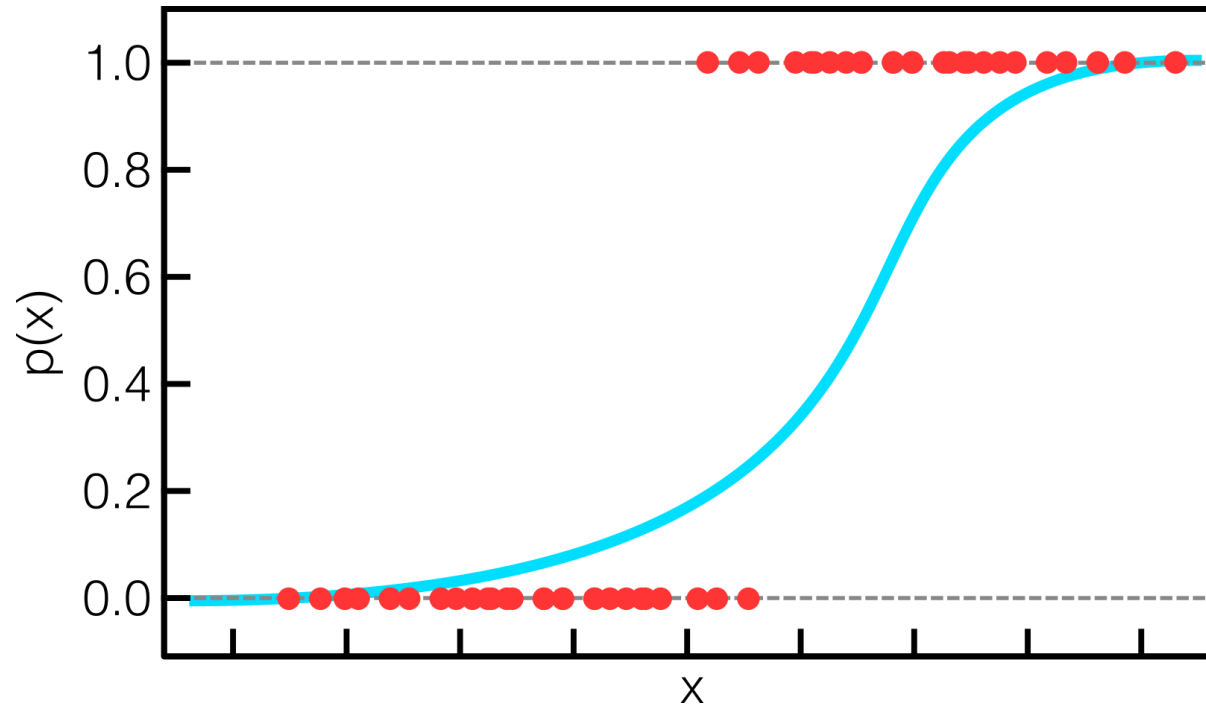
Para clasificar, en vez de modelar la salida, modelamos la probabilidad de que una observación es de una clase u otra. Para ello, podemos modelar a la probabilidad usando una función que nos asegure que siempre tendremos valores entre 0 y 1.

En regresión logística, esto lo resolvemos usando una función sigmoide:

$$p(x) = \frac{e^{b+w_0x}}{1 + e^{b+w_0x}} = \frac{1}{1 + e^{-(b+w_0x)}}$$

REGRESIÓN LOGÍSTICA

Lo que visualmente se observa:



CLASIFICADOR BAYESIANO INGENUO

Teorema de Bayes

Este teorema es uno de los teoremas más importantes de probabilidad, y uno que hasta el día de hoy genera divisiones en el plano filosófico por su implicancia

Este describe la probabilidad de un evento, basado en conocimiento previo de condiciones que pueden estar relacionados con el evento.

Por ejemplo, si se sabe que el riesgo de desarrollar problemas de salud aumenta con la edad, el teorema de Bayes permite evaluar con mayor precisión el riesgo para un individuo de una edad conocida condicionándolo en relación con su edad, en lugar de asumir que el individuo es típico de la población en su conjunto.

CLASIFICADOR BAYESIANO INGENUO

Una de las aplicaciones de este teorema es el denominado clasificador bayesiano ingenuo.

Este clasificador utiliza la probabilidad de observar atributos, dado un resultado, para estimar la probabilidad de observar el resultado y_j , dado un conjunto de atributos.



APRENDIZAJE POR REFUERZO

APRENDIZAJE POR REFUERZO

En las últimas clases vimos dos áreas conocidas de aprendizaje supervisados. Dejaremos a Aprendizaje no supervisado para AMq1, y vayamos al **Aprendizaje por Refuerzo**.

El aprendizaje por refuerzo es un enfoque de aprendizaje automático donde un agente aprende a tomar decisiones óptimas al interactuar con un entorno, maximizando una señal de recompensa a lo largo del tiempo.

Para este tipo de aprendizaje es más fácil de pensar usando el concepto de agente.

APRENDIZAJE POR REFUERZO

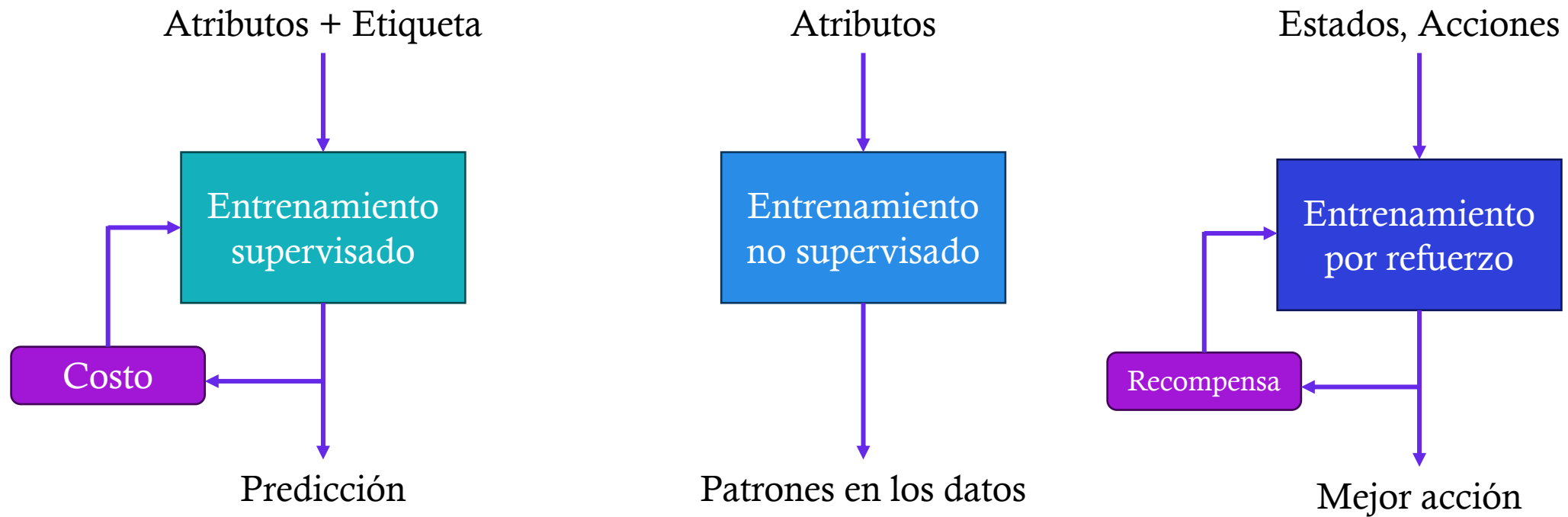
Un agente puede aprender a jugar al ajedrez con aprendizaje supervisado, proporcionándole ejemplos de situaciones de juego con los mejores movimientos para dichas situaciones. Pero si no hay un profesor que proporcione ejemplos, ¿qué puede hacer el agente?

Intentando movimientos aleatorios, el agente puede eventualmente **construir un modelo de predicción** de su entorno: cómo estará el tablero después de que haga un movimiento e incluso cómo es probable que el oponente responda en una situación dada.

El problema es el siguiente: sin cierta realimentación de lo que es bueno y de lo que es malo, el agente no tendrá razones para decidir qué movimiento hacer.

El agente necesita saber que algo bueno ha ocurrido cuando gana y que algo malo ha ocurrido cuando pierde. Esta clase de realimentación se denomina **recompensa**, o **refuerzo**.

APRENDIZAJE POR REFUERZO



APRENDIZAJE POR REFUERZO

Aprendizaje por refuerzo se utiliza más comúnmente para resolver una clase diferente de problemas del mundo real, como una tarea de control o una tarea de decisión, en las que se opera un sistema que interactúa con el mundo real.

Es útil para una variedad de aplicaciones como:

- Operar un dron o un vehículo autónomo
- Manipular un robot para navegar por el entorno y realizar diversas tareas.
- Gestionar una cartera de inversiones y tomar decisiones comerciales.
- Jugar juegos como Go, Ajedrez, videojuegos.



PROCESO DE DECISIÓN DE MÁRKOV

PROCESO DE DECISIÓN DE MÁRKOV

Un proceso de decisión de Márkov (MDP) es un proceso de control estocástico en tiempo discreto. Proporciona un marco matemático para modelar la toma de decisiones en situaciones en las que los resultados son en **parte aleatorios** y en parte están bajo el **control del agente**.

En cada paso temporal, el proceso se encuentra en el estado s , y agente puede elegir cualquier acción a que esté disponible en el estado s . El proceso responde en el siguiente paso temporal pasando aleatoriamente a un nuevo estado s' , y ofreciendo al responsable de la toma de decisiones la recompensa correspondiente $R(s, a, s')$.

La probabilidad de que el proceso pase a su nuevo estado s' está influida por la acción elegida. En concreto, viene dada por la función de transición de estado $P(s, a, s')$. Así, el siguiente estado s' depende del estado actual s y de la acción del decisor a . Pero dado s y a , es condicionalmente independiente de todos los **estados y acciones anteriores**.

Los procesos de decisión de Márkov son una extensión de las cadenas de Márkov; la diferencia es la adición de acciones y recompensas.

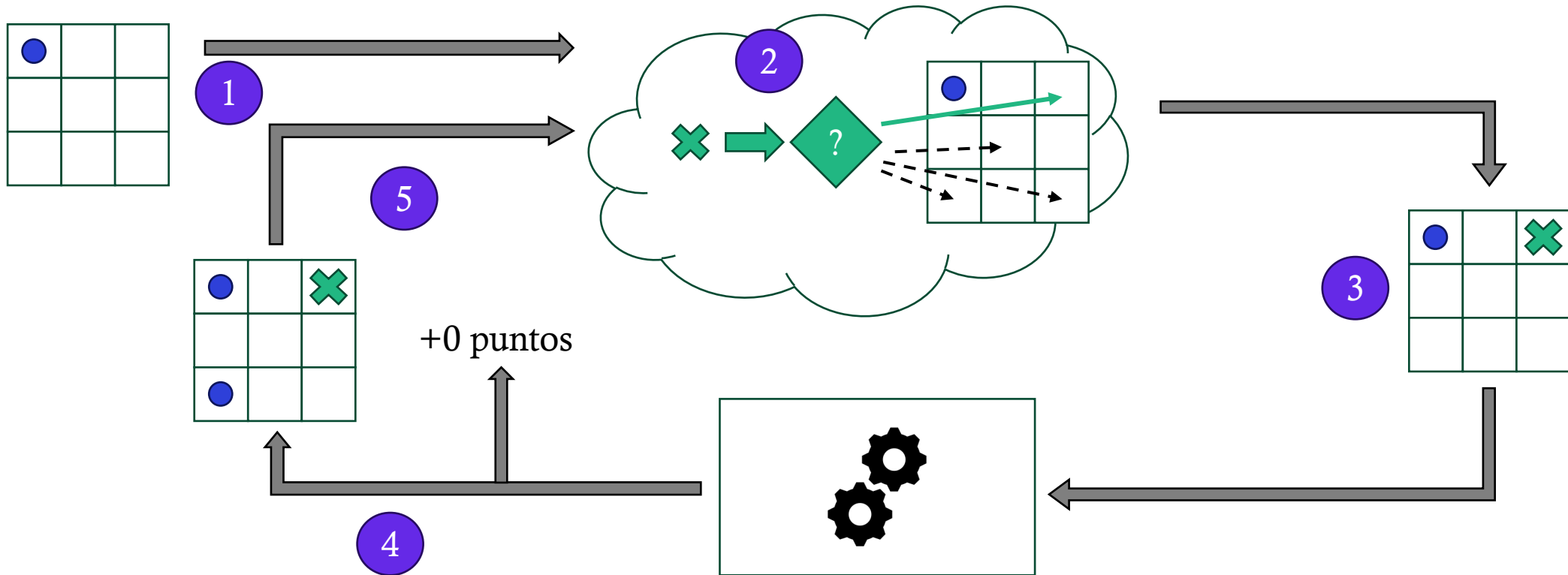
PROCESO DE DECISIÓN DE MÁRKOV

MDP tiene 5 componentes:

- **Agente**
- **Ambiente**
- **Estado**
- **Acción**
- **Recompensa**: es el refuerzo positivo o negativo que el agente recibe del entorno como resultado de sus acciones. Es una forma de evaluar la "bondad" o la "maldad" de una acción en particular.

PROCESO DE DECISIÓN DE MÁRKOV

Veamos un ejemplo usando tatetí en ejemplo de funcionamiento de MDP:



PROCESO DE DECISIÓN DE MÁRKOV

La ejecución de un MDP puede ser descrita como una secuencia de ocurrencias en términos de estado, acción y recompensa sobre una secuencia de pasos temporales.

En tareas que tienen un final, la secuencia es episódica. En un juego de tatetí, un episodio es cada partida que se juega. **Cada episodio es independiente del siguiente**. Por lo que el funcionamiento de Aprendizaje por refuerzo (AR) se repite en episodios, donde cada episodio, se repite varios pasos de tiempo.

Por otro lado, las tareas de AR no tienen fin se conocen como *Tareas Continuas* y pueden continuar para siempre.

PROCESO DE DECISIÓN DE MÁRKOV

El **agente** y el **entorno** controlan las transiciones estado-acción:

El MDP opera alternando entre el **agente** realizando una acción y luego el **entorno** haciendo algo. En cada paso de tiempo:

- Dado el estado actual, la siguiente acción la decide el **agente** de un grupo de posibles acciones.
- Dado el estado actual y la siguiente acción elegida por el agente, la transición al siguiente estado y la recompensa están controladas por el **entorno**. Esta parte del entorno tomando una decisión es lo que el agente ve como que sus acciones son probabilísticas.

PROCESO DE DECISIÓN DE MÁRKOV

¿Cómo el agente toma una decisión?

Este es precisamente el problema de aprendizaje por refuerzo que queremos resolver. Para ello se utiliza tres conceptos:

- **Retorno**: A medida que el agente ejecuta pasos de tiempo, acumula recompensas en cada paso de tiempo. La recompensa acumulada es lo que se llama retorno.
- **Política**: La política es la estrategia que se sigue para elegir una acción.
- **Valor**: Indica el Retorno esperado siguiendo alguna Política

PROCESO DE DECISIÓN DE MÁRKOV

Retorno

Cuando se calcula el retorno, en lugar de sumar todas las recompensas, se aplica un factor de descuento γ para ponderar las recompensas posteriores a lo largo del tiempo (γ está entre 0 y 1). Estos se conocen como recompensas con descuento:

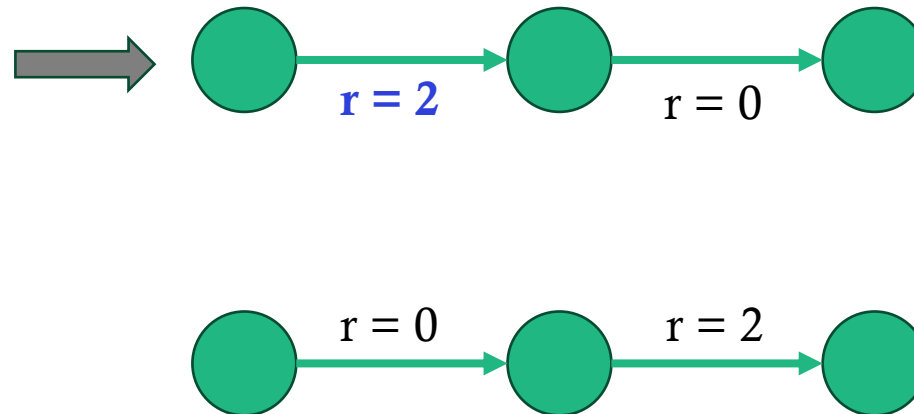
$$Return = r_0 + \gamma r_1 + \gamma^2 r_2 + \cdots + \gamma^n r_n$$

De esta manera, las recompensas acumulativas no crecen infinitamente a medida que el número de pasos de tiempo se vuelve muy grande.

PROCESO DE DECISIÓN DE MÁRKOV

Retorno

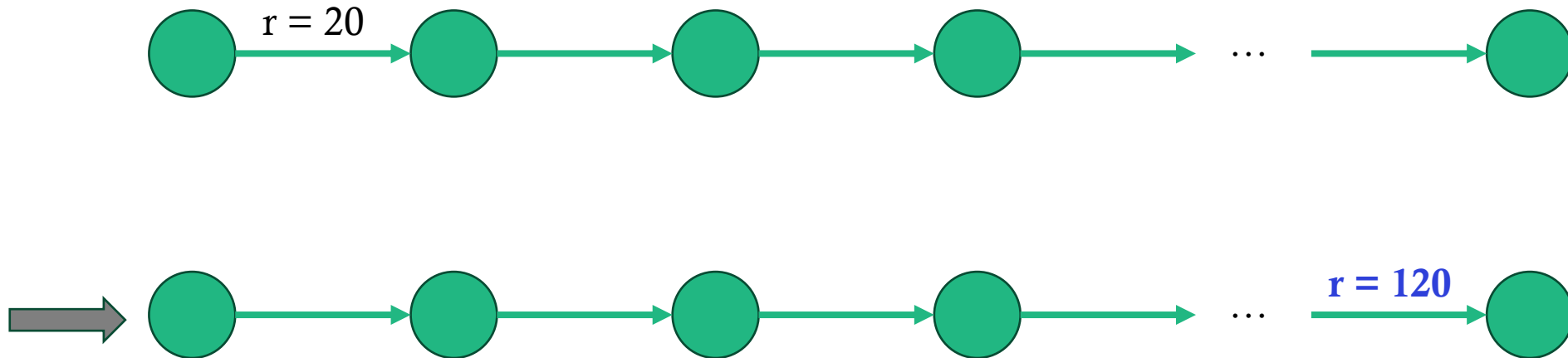
La recompensa inmediata es más valiosa que la recompensa posterior:



PROCESO DE DECISIÓN DE MÁRKOV

Retorno

Si el agente tiene que elegir entre obtener alguna recompensa ahora u obtener una recompensa mucho mayor más adelante, lo más probable es que la recompensa mayor sea preferible. Esto se debe a que queremos que el agente mire los rendimientos totales en lugar de las recompensas individuales.



PROCESO DE DECISIÓN DE MÁRKOV

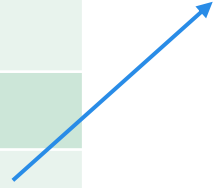
Política

Esto es básicamente que define el modelo, es la acción que va a tomar dado un estado que se encuentre.

La forma más trivial es volver a la tabla de estado-acción que vimos en la clase 2.

	a1	a2	a3
s1	$\pi(a_1 s_1)$
s2
s3
s4	$\pi(a_3 s_4)$

Probabilidad
de realizar la
acción a_3
cuando se
encuentra en
el estado s_4



PROCESO DE DECISIÓN DE MÁRKOV

Política

Las políticas pueden ser:

- **Deterministas**: Una política determinista es una política en la que el agente siempre elige la misma acción fija cuando alcanza un estado particular.
- **Estocásticas**: La política estocástica es una política en la que el agente varía las acciones que elige para un estado, en función de cierta probabilidad para cada acción. Podría hacer esto mientras juega, por ejemplo, para que no se vuelva completamente predecible.

PROCESO DE DECISIÓN DE MÁRKOV

Política

El agente realmente no tiene una política útil cuando comienza y no tiene idea de qué acción debe tomar en un estado determinado. Luego, al utilizar el algoritmo de aprendizaje por refuerzo, aprende lentamente una política útil que puede utilizar.

Hay tantas políticas posibles, ¿cuál debería utilizar el Agente?

El objetivo del agente es seguir una Política que maximice su Retorno. Entonces, de todas las políticas que el agente podría seguir, quiere elegir la mejor, es decir. el que le da mayor retorno.

PROCESO DE DECISIÓN DE MÁRKOV

Valor

Digamos que el agente se encuentra en un estado particular. Además, digamos que al agente se le ha dado de alguna manera una política, π .

Si parte de ese estado y siempre elige acciones basadas en esa política, ¿cuál es el rendimiento que podría esperar obtener?

Este rendimiento promedio a largo plazo, o rendimiento esperado, se conoce como **valor** de ese estado en particular según la política π .

PROCESO DE DECISIÓN DE MÁRKOV

Valor

Tenemos dos tipos de Valores:

- **Valor de Estado:** el rendimiento esperado de un estado determinado, mediante la ejecución de acciones basadas en una política determinada π desde ese estado en adelante.
- **Valor de Estado-Acción (valor Q):** el rendimiento esperado al realizar una acción determinada desde un estado determinado y luego ejecutar acciones basadas en una política determinada π después de eso.

PROCESO DE DECISIÓN DE MÁRKOV

Relación entre los tres:

- La **Recompensa** es la recompensa inmediata obtenida por una sola acción.
- El **Retorno** es el total de todas las recompensas con descuento obtenidas hasta el final de ese episodio.
- El **Valor** es el rendimiento medio (también conocido como rendimiento esperado) de muchos episodios.

Se puede pensar en el **Valor** de la siguiente manera. El agente aprende de la experiencia. A medida que interactúa con el entorno y completa episodios, obtiene los retornos de cada episodio.

A medida que acumula más experiencia (es decir, obtiene retornos por más y más episodios), tiene una idea de qué estados y qué acciones en esos estados producen el mayor retorno.

Almacena esta *experiencia* como **Valor**

PROCESO DE DECISIÓN DE MÁRKOV

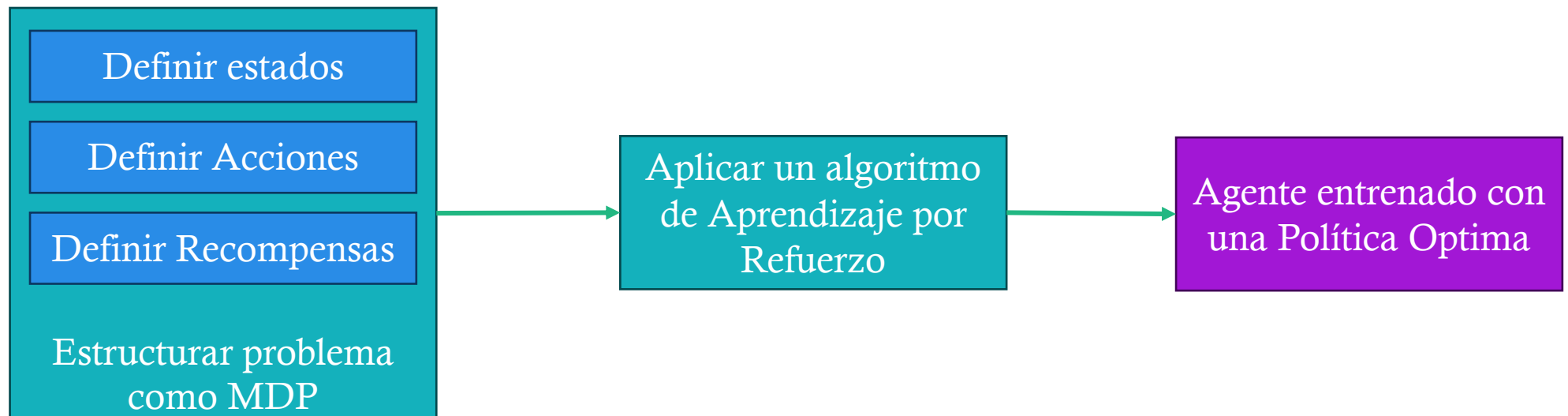
Utilizando la función de Valor para comparar políticas

Podemos comparar políticas para determinar cuáles son *buenas* y cuáles son *malas*, también podemos utilizar eso para encontrar la *mejor* política. Esto se conoce como **Política Óptima**.

PROCESO DE DECISIÓN DE MÁRKOV

Resolviendo el problema de Aprendizaje por Refuerzo encontrando la política óptima

Entonces, podemos entender como entrenar un agente usando Aprendizaje por Refuerzo, usando MDP. Lo que se busca encontrar la **Política Óptima** para el agente. Una vez que tiene la Política Óptima, simplemente usa esa política para elegir acciones de cualquier estado.





CATEGORÍAS DE SOLUCIONES

CATEGORÍAS DE SOLUCIONES

Predicción vs Control

Los problemas se pueden dividir en dos tipos:

- **Problemas de predicción:** Se proporciona una política como entrada y el objetivo es generar la función de Valor. No es necesario que sea la Política Óptima.
- **Problemas de control:** No se proporciona información y el objetivo es explorar el espacio de políticas y encontrar la Política Óptima.

CATEGORÍAS DE SOLUCIONES

Hay muchos algoritmos que buscan encontrar la **Política Óptima** para el agente.
Hay dos grandes categorías:

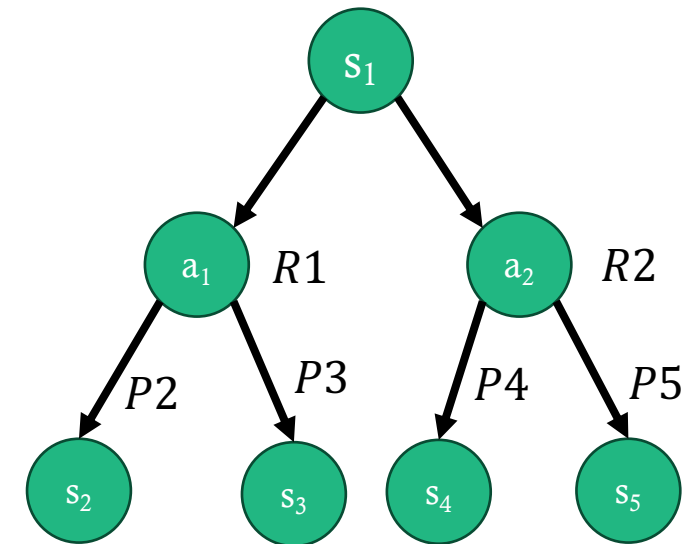
- Basado en modelo
- Sin modelo

CATEGORÍAS DE SOLUCIONES

Basado en modelo

Los enfoques basados en modelos se utilizan cuando se conoce el funcionamiento interno del entorno. Podemos decir de manera confiable qué próximo estado y recompensa generará el entorno cuando se realice alguna acción desde algún estado actual.

Estos modelos son los que se conocen como planificación. Son aquellos que ya vimos en clase 2 y 3 (y que hay más avanzados). No es AR.

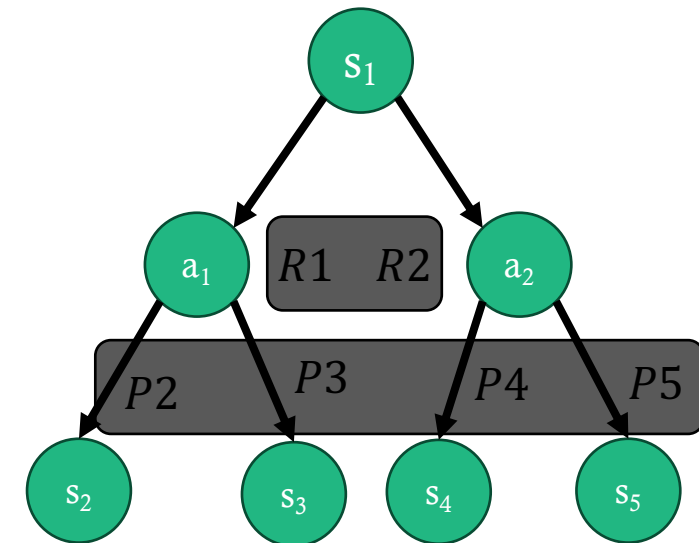


CATEGORÍAS DE SOLUCIONES

Libre de modelos

Dado que el funcionamiento interno del entorno es invisible para el agente,

¿Cómo observa el algoritmo el comportamiento del entorno?



CATEGORÍAS DE SOLUCIONES

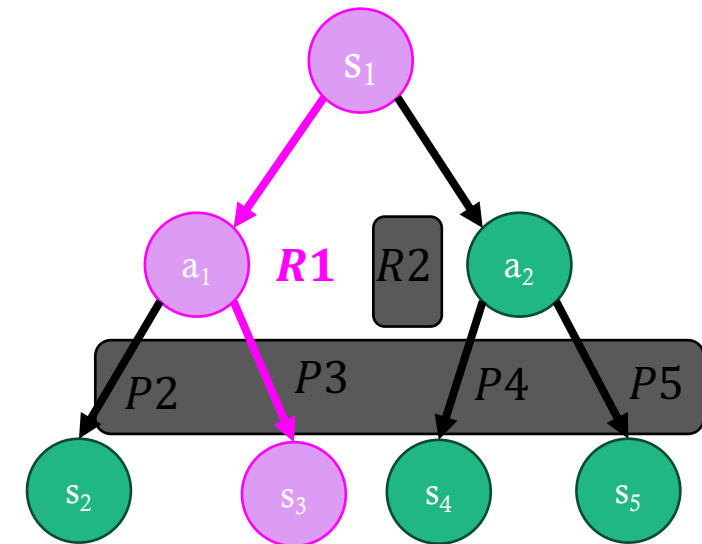
Libre de modelos

Dado que el funcionamiento interno del entorno es invisible para el agente,

¿Cómo observa el algoritmo el comportamiento del entorno? De la interacción con el mismo.

El famoso **prueba y error**.

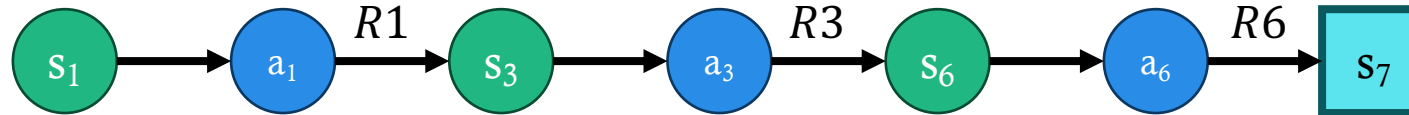
Intenta pasos y recibe comentarios positivos o negativos. Esto es muy parecido al aprendizaje de seres biológicos.



CATEGORÍAS DE SOLUCIONES

Libre de modelos

A medida que el agente da cada paso, sigue un camino. La trayectoria del agente se convierte en los *datos de entrenamiento* del algoritmo.





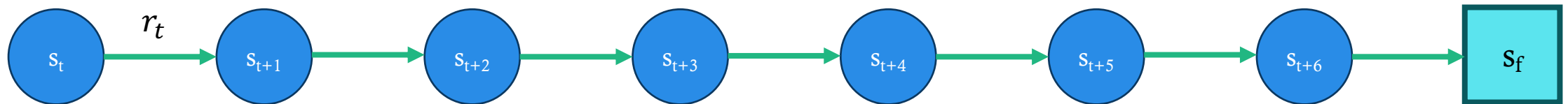
ECUACIÓN DE BELLMAN

ECUACIÓN DE BELLMAN

La ecuación de Bellman es un concepto fundamental en el campo del Aprendizaje por Refuerzo.

Esencialmente, describe cómo el valor de estar en un estado particular bajo una política específica se relaciona con el valor de estar en el próximo estado y las recompensas esperadas.

Vayamos paso a paso para crear una intuición de esto. Para calcular el valor de estado s_t , $V(s_t)$ necesitamos calcular la suma ponderada de las recompensas de los pasos futuros (el retorno).



$$V(s_t) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots + \gamma^6 r_6$$

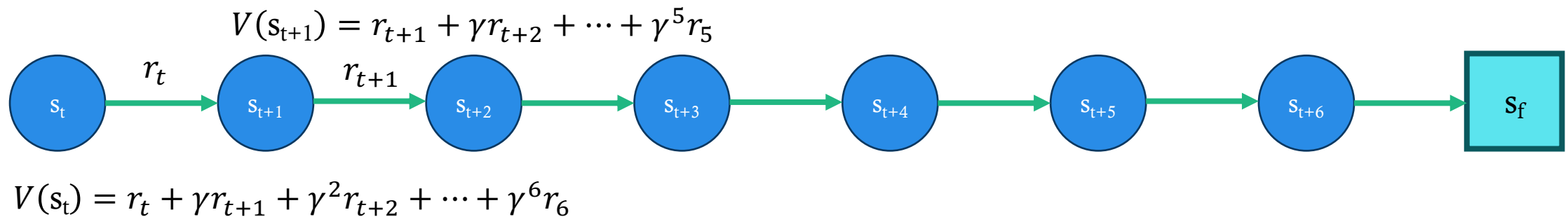
ECUACIÓN DE BELLMAN

La ecuación de Bellman es un concepto fundamental en el campo del Aprendizaje por Refuerzo.

Esencialmente, describe cómo el valor de estar en un estado particular bajo una política específica se relaciona con el valor de estar en el próximo estado y las recompensas esperadas.

Vayamos paso a paso para crear una intuición de esto. Para calcular el valor de estado s_t , $V(s_t)$ necesitamos calcular la suma ponderada de las recompensas de los pasos futuros (el retorno).

Para calcular $V(s_{t+1})$ tenemos que calcular el retorno de s_{t+1}



ECUACIÓN DE BELLMAN

Lo que podemos ver que hay una recursividad, esto lo podemos aprovechar usando la ecuación de Bellman, el cual se calcula para un valor actual se calcula con la recompensa inmediata r_t más el valor descontado del siguiente estado $\gamma V(s_{t+1})$:

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

ECUACIÓN DE BELLMAN

Lo que podemos ver que hay una recursividad, esto lo podemos aprovechar usando la ecuación de Bellman, el cual se calcula para un valor actual se calcula con la recompensa inmediata r_t más el valor descontado del siguiente estado $\gamma V(s_{t+1})$:

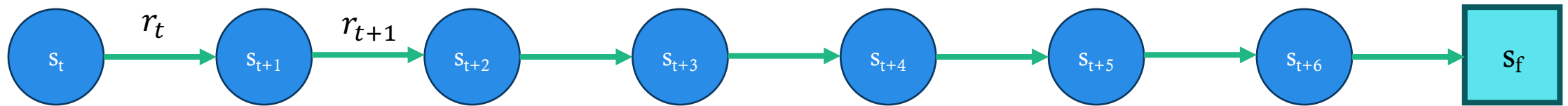
$$V_{\pi}(s_t) = E_{\pi}[r_t + \gamma V_{\pi}(s_{t+1})]$$

The diagram illustrates the components of the Bellman equation $V_{\pi}(s_t) = E_{\pi}[r_t + \gamma V_{\pi}(s_{t+1})]$ using colored arrows and labels:

- A red arrow points from the $V_{\pi}(s_t)$ term to the label "Bajo la política π ".
- A green arrow points from the $V_{\pi}(s_t)$ term to the label "Valor del estado s_t ".
- A purple arrow points from the E_{π} term to the label "Valor esperado de la recompensa".
- A blue arrow points from the $\gamma V_{\pi}(s_{t+1})$ term to the label "El valor descontado del siguiente estado".

ECUACIÓN DE BELLMAN

Volviendo al caso anterior:



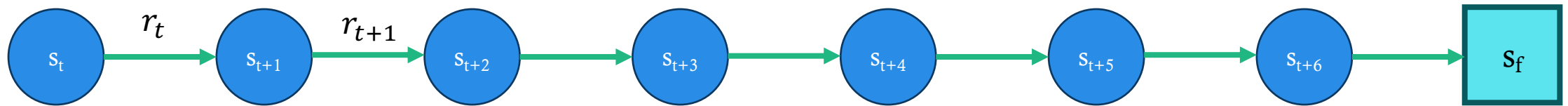
$$V(s_t) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots + \gamma^6 r_6$$

Si el agente arranca en el estado t , y sigue la misma política para todos los estados:

$$V(s_t) = r_t + \gamma V(s_{t+1})$$

ECUACIÓN DE BELLMAN

Volviendo al caso anterior:



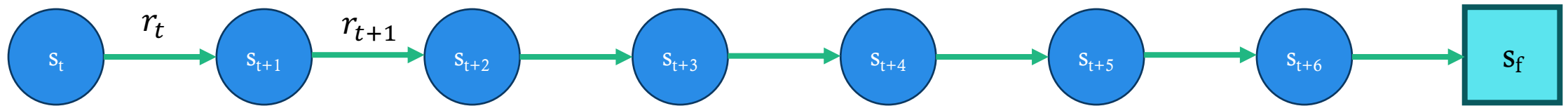
$$V(s_t) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots + \gamma^6 r_6$$

Si el agente arranca en el estado t , y sigue la misma política para todos los estados:

$$\begin{aligned} V(s_t) &= r_t + \gamma V(s_{t+1}) \\ V(s_{t+1}) &= r_{t+1} + \gamma V(s_{t+2}) \end{aligned}$$

ECUACIÓN DE BELLMAN

Volviendo al caso anterior:



$$V(s_t) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots + \gamma^6 r_6$$

Si el agente arranca en el estado t , y sigue la misma política para todos los estados:

$$\begin{aligned} V(s_t) &= r_t + \gamma V(s_{t+1}) \\ V(s_{t+1}) &= r_{t+1} + \gamma V(s_{t+2}) \end{aligned}$$

La idea es que en lugar de calcular cada valor como la suma del rendimiento esperado, que es un proceso largo, se calcula el valor como la suma de la recompensa inmediata + el valor descontado del estado siguiente.



ESTRATEGIAS DE APRENDIZAJE

ESTRATEGIAS DE APRENDIZAJE

Recordemos, el agente aprende interactuando con el entorno. La idea es que, dada la experiencia y la recompensa recibida, el agente actualice su función de valor o política.

El tema es como hacemos para que el agente actualice su función de valor o política, es decir aprenda:

- **Montecarlo**: Utiliza un episodio completo de experiencia para antes de aprender.
- **Aprendizaje por diferencia temporal**: Aprende en un paso.

ESTRATEGIAS DE APRENDIZAJE

Montecarlo

En esta técnica, se espera que el agente termine el episodio, calcula el retorno G_t y usa como objetivo para adaptar $V(s_t)$:

$$V(s_t) \leftarrow V(s_t) + \alpha[G_t - V(s_t)]$$

ESTRATEGIAS DE APRENDIZAJE

Montecarlo

En esta técnica, se espera que el agente termine el episodio, calcula el retorno G_t y usa como objetivo para adaptar $V(s_t)$:

$$\underline{V(s_t)} \leftarrow \underline{V(s_t)} + \underline{\alpha} [\underline{G_t} - \underline{V(s_t)}]$$

Nuevo valor del estado t

Constante de aprendizaje

Retorno en el paso

Estimación del valor del estado t
(retorno esperado de ese estado)

ESTRATEGIAS DE APRENDIZAJE

Montecarlo

Al final del episodio se tiene una lista de estados, acciones, recompensas y nuevos estados.

El agente suma todas las recompensas G_t para ver que tan bien le fue.

Luego actualiza $V(s_t)$ con la fórmula que vimos.

Luego arranca un nuevo juego con este nuevo conocimiento, a medida que avanza los episodios, el agente va a aprender a jugar mejor y mejor.

ESTRATEGIAS DE APRENDIZAJE

Aprendizaje por diferencia temporal

En el otro extremo, aprendizaje por diferencia temporal (DT), espera una interacción $t+1$ para actualizar $V(s_t)$ usando r_t y $\gamma V(s_{t+1})$.

DT basa las actualizaciones en parte de una estimación de $V(s_{t+1})$ y no del retorno total G_t

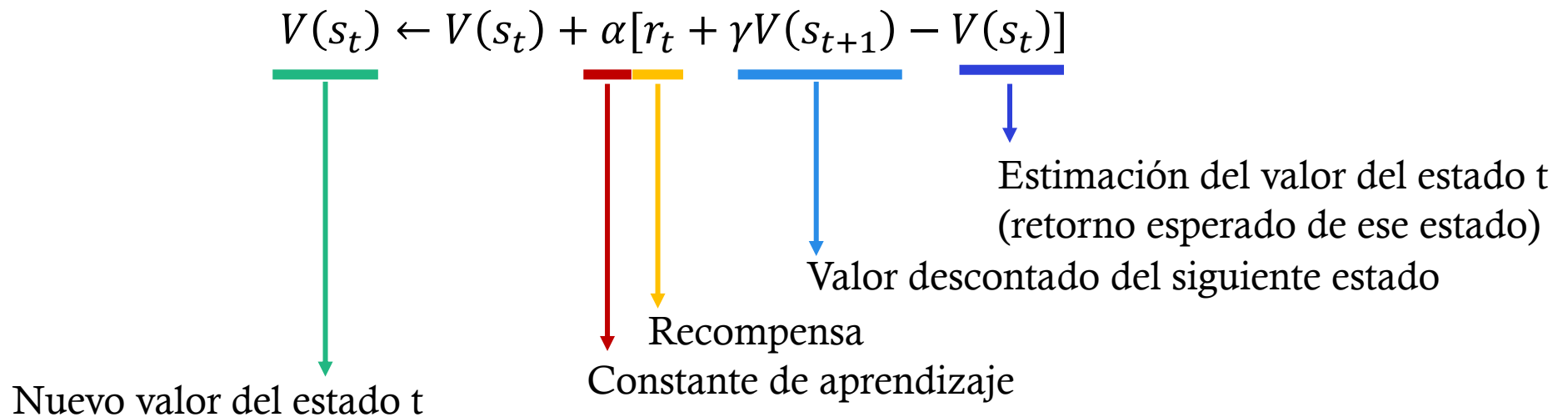
$$V(s_t) \leftarrow V(s_t) + \alpha[r_t + \gamma V(s_{t+1}) - V(s_t)]$$

ESTRATEGIAS DE APRENDIZAJE

Aprendizaje por diferencia temporal

En el otro extremo, aprendizaje por diferencia temporal (DT), espera una interacción $t+1$ para actualizar $V(s_t)$ usando r_t y $\gamma V(s_{t+1})$.

DT basa las actualizaciones en parte de una estimación de $V(s_{t+1})$ y no del retorno total G_t

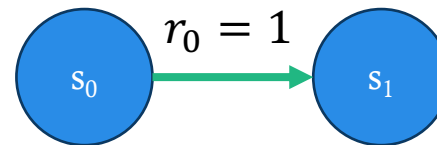


ESTRATEGIAS DE APRENDIZAJE

Aprendizaje por diferencia temporal

Por ejemplo,

1. Inicializamos la función de valor con ceros, la constante de aprendizaje igual a 0.1.
2. El agente realiza una acción y pasa del estado 0 al estado 1



3. El agente tiene una recompensa r_0 igual a 1.

ESTRATEGIAS DE APRENDIZAJE

Aprendizaje por diferencia temporal

Por ejemplo,

4. Ahora actualizamos el valor $V(s_0)$

$$V(s_0) = V(s_0) + \alpha[r_0 + \gamma V(s_1) - V(s_0)]$$

$$V(s_0) = 0 + 0.1[1 + 0.9 * 0 - 0]$$

$$V(s_0) = 0.1$$

5. Con esta fórmula, actualizamos la función de valor en el estado 0.
6. Para los siguientes pasos, actualizamos de igual forma.



Q-LEARNING

Q-LEARNING

Existen varios algoritmos basados en valores y basados en políticas. Aunque podemos reducir a unos pocos principios esenciales que todos emplean.

En un nivel alto, todos los algoritmos, tanto los basados en valores como los basados en políticas, arrancan con una estimación inicial que van ajustando en iteraciones posteriores. Para ello, realizan cuatro operaciones básicas:

1. Inician estimaciones
2. Toma una acción o más acciones
3. Obtiene retroalimentación de ambiente
4. Mejora la estimación

Q-LEARNING

Q-Learning es un método basado en valores que utiliza un enfoque de aprendizaje por diferencia temporal:

- Método basado en valores: encuentra la política óptima de forma indirecta entrenando una función de valor o acción-valor que nos dirá el valor de cada estado o de cada par estado-acción.
- Enfoque de aprendizaje por DT: actualiza su función de valor de acción en cada paso en lugar de al final del episodio.
- Este algoritmo aprende el Valor de **Estado-Acción (valor Q)**

Q-LEARNING

Internamente, la función de valor Q es estructurada con una tabla Q , en donde cada celda corresponde a un par de estado-acción.

	a1	a2
s ₁	12	0.3
s ₂	43	12
s ₃	10	93
s ₄	0	5

Q-LEARNING

Este algoritmo, arranca **inicializando las estimaciones** de la tabla, en general, con todos valores iguales a cero:

	a1	a2
s ₁	0	0
s ₂	0	0
s ₃	0	0
s ₄	0	0

Q-LEARNING

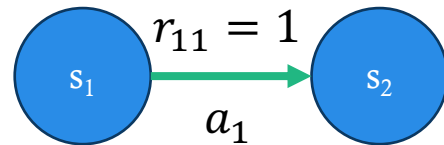
A medida que el agente explora, en un paso, recibe una recompensa y actualiza la tabla usando aprendizaje de DT con una pequeña variación:

$$Q(s_i, a_j) = Q(s_i, a_j) + \alpha [(r_{ij} + \gamma \max_k(Q(s_k, :))) - Q(s_i, a_j)]$$

Q-LEARNING

A medida que el agente explora, en un paso, recibe una recompensa y actualiza la tabla usando aprendizaje de DT con una pequeña variación:

$$Q(s_1, a_1) = Q(s_1, a_1) + \alpha \left[(r_{11} + \gamma \max(Q(s_2, :))) - Q(s_1, a_1) \right]$$

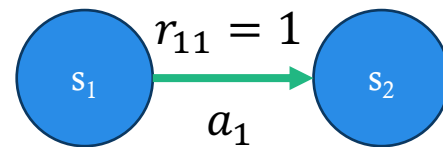


	a1	a2
s ₁	1.2	0
s ₂	1	0.3
s ₃	0	0
s ₄	0	0

Q-LEARNING

A medida que el agente explora, en un paso, recibe una recompensa y actualiza la tabla usando aprendizaje de DT con una pequeña variación:

$$Q(s_1, a_1) = Q(s_1, a_1) + \alpha \left[(r_{11} + \gamma \max(Q(s_2, :))) - Q(s_1, a_1) \right]$$

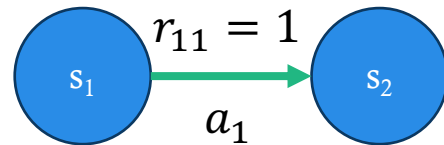


	a1	a2
s1	1.2	0
s2	1	0.3
s3	0	0
s4	0	0

Q-LEARNING

A medida que el agente explora, en un paso, recibe una recompensa y actualiza la tabla usando aprendizaje de DT con una pequeña variación:

$$Q(s_1, a_1) = 1.2 + 0.1 [(1 + 0.9 * 1) - 1.2] = 1.27$$

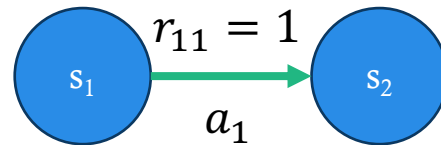


	a1	a2
s ₁	1.2	0
s ₂	1	0.3
s ₃	0	0
s ₄	0	0

Q-LEARNING

A medida que el agente explora, en un paso, recibe una recompensa y actualiza la tabla usando aprendizaje de DT con una pequeña variación:

$$Q(s_1, a_1) = 1.2 + 0.1 [(1 + 0.9 * 1) - 1.2] = 1.27$$



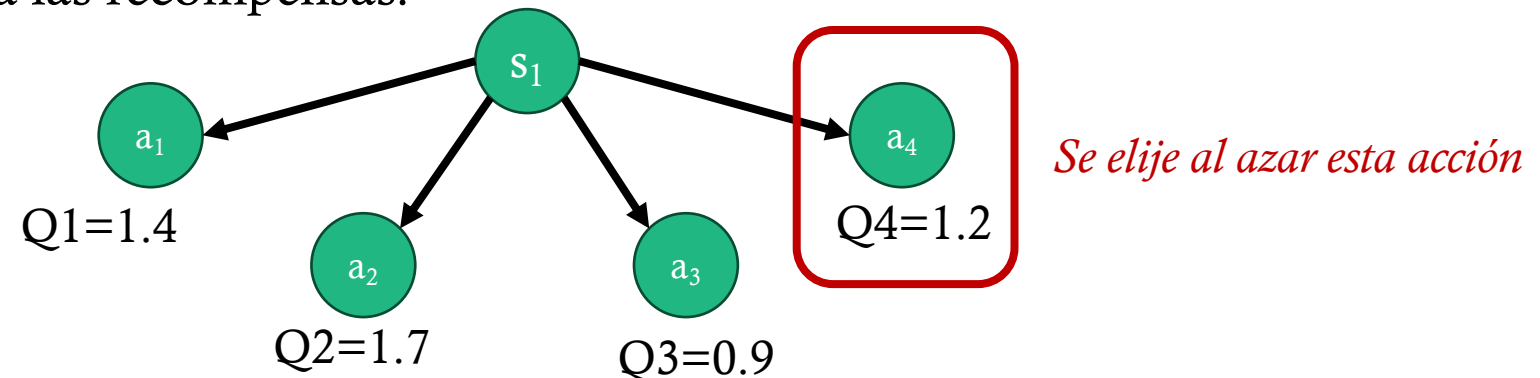
	a1	a2
s ₁	1.27	0
s ₂	1	0.3
s ₃	0	0
s ₄	0	0

Q-LEARNING

Toma una acción

Ahora, nos queda la pregunta de cómo hace el agente para tomar una decisión. El agente necesita encontrar el equilibrio adecuado entre **Exploración** y **Explotación**, para tomar una acción.

Exploración: cuando se comienza a aprender, no se tiene idea de qué acciones son *buenas* y cuáles son *malas*. Entonces se pasa por un proceso de descubrimiento en el que se prueba diferentes acciones al azar y se observa las recompensas.

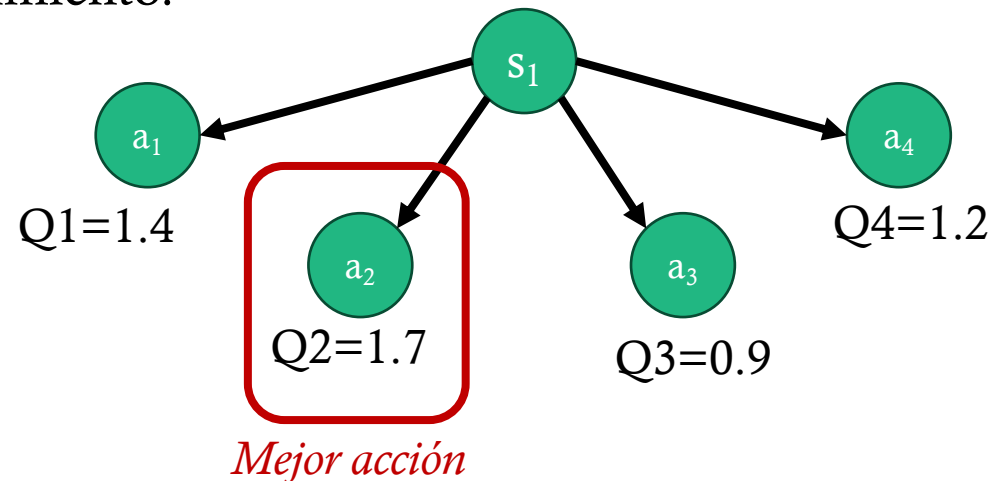


Q-LEARNING

Toma una acción

Ahora, nos queda la pregunta de cómo hace el agente para tomar una decisión. El agente necesita encontrar el equilibrio adecuado entre **Exploración** y **Explotación**, para tomar una acción.

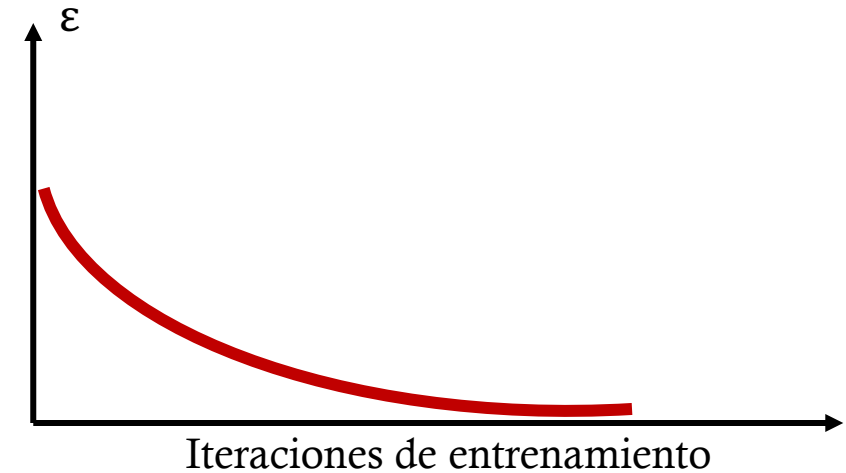
Explotación: en el otro extremo, cuando el modelo está completamente entrenado, ya se ha explorado todas las acciones posibles, por lo que se puede elegir las mejores acciones que producirán el máximo rendimiento.



Q-LEARNING

Toma una acción

Un agente basado en valor adopta una estrategia dinámica conocida como **ϵ -Greedy**. Utiliza una tasa de exploración ϵ que se ajusta a medida que avanza el entrenamiento para garantizar una mayor exploración en las primeras etapas del entrenamiento y cambia hacia una mayor explotación en las etapas posteriores.



$$\epsilon = \epsilon_0 e^{-\lambda t}$$

Si $U[0,1] < \epsilon$, **exploramos**

Si no, **explotación**

Q-LEARNING

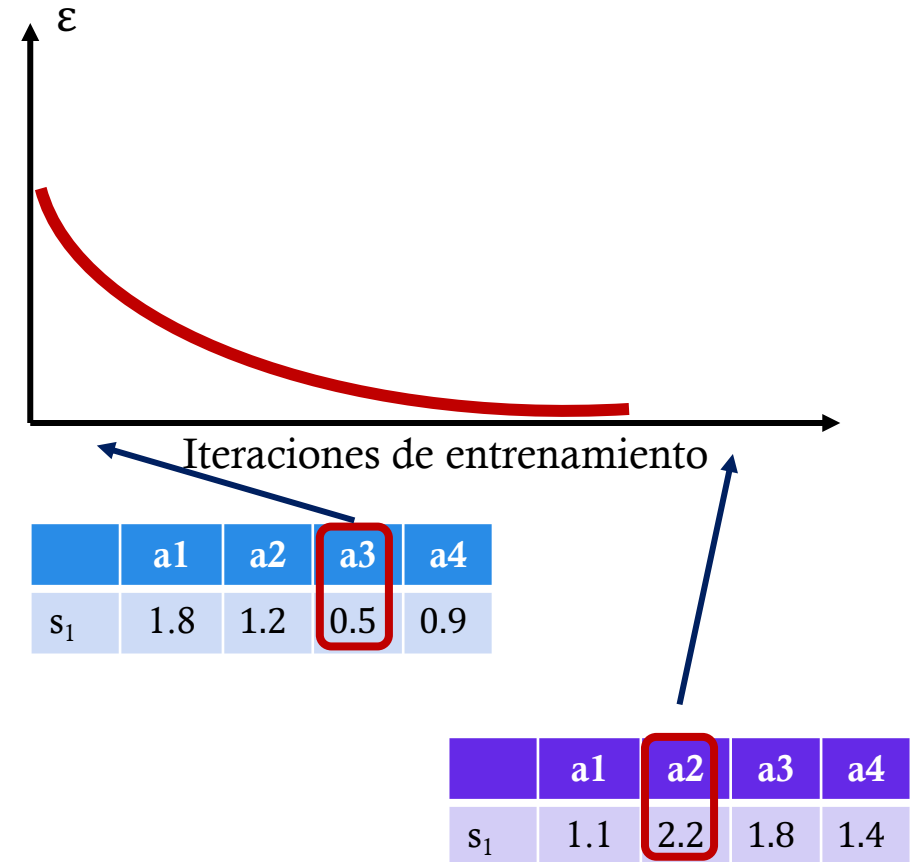
Toma una acción

Se establecemos ϵ inicialmente en 1.

Luego, al comienzo de cada episodio, se reduce ϵ a cierta tasa.

Cada vez que elige una acción en cada estado, se **explora** con una probabilidad ϵ . Se **explota** con una probabilidad $1-\epsilon$.

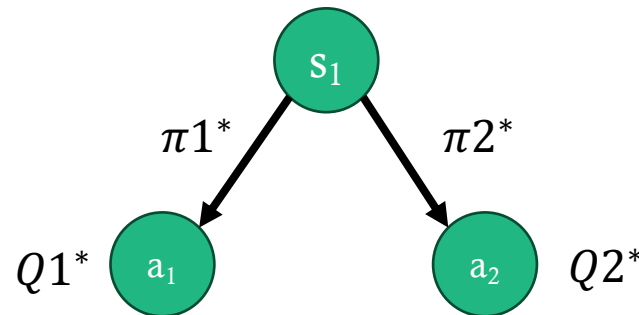
Dado que ϵ es mayor en las primeras etapas, es más probable que el agente **explore**. A medida que ϵ disminuye, la probabilidad de exploración disminuye y el agente se vuelve **codicioso** al explotar cada vez más el entorno.



Q-LEARNING

Una vez que termina el entrenamiento, asumimos que el modelo aprende. La política encontrada es la óptima, y se verifica que:

$$\pi^*(s) = \operatorname{argmax}(Q(s, a))$$



$\pi^* = 1$ para la acción que corresponda al $\max(Q1^*, Q2^*)$

$\pi^* = 0$ para las restantes

Q-LEARNING

¿Pero, como hace para seleccionarse la política a raíz del valor?

Generalmente, la Política Óptima es en este proceso **determinista** ya que siempre elige la mejor acción.

Sin embargo, la Política Óptima puede ser estocástica si hay un empate entre dos valores Q. En ese caso, la Política Óptima elige cualquiera de las dos acciones correspondientes con igual probabilidad.

En juego con oponentes, una política óptima estocástica es necesaria porque una política determinista daría como resultado que el agente realice movimientos predecibles que su oponente podría derrotar fácilmente.

Q-LEARNING

Otros algoritmos se basan en procedimientos similares, aunque algunos se basan directamente en política en vez de valores. Las diferencias se pueden resumir en:

- **Frecuencia:** El número de pasos hacia adelante dados antes de una actualización.
 - Por episodio.
 - Por cada paso
 - n-pasos
- **Profundidad:** El número de pasos hacia atrás para propagar una actualización.
 - Por episodio.
 - Por cada paso
 - n-pasos
- **Formula:** Fórmula que se utiliza para calcular la estimación actualizada.
 - Diferentes algoritmos usan diferentes variantes de la ecuación de Bellman.
 - Si se basa en políticas, se aumenta o disminuye la probabilidad en base a la recompensa recibida.

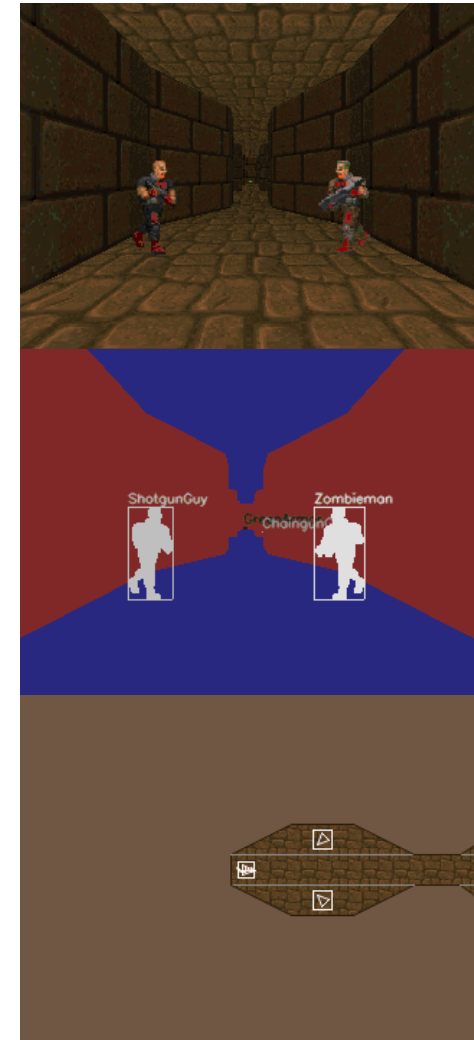
CAN IT RUN DOOM?

PLAY

CAN IT PLAY DOOM? KINDA...

Algo que tiene interesante el AR es que, dado que se usa en problemas de control, se necesita crear escenarios simulados. Hay varias implementaciones disponibles, entre ellas:

- **Gymnasium**: Un fork de Gym de OpenAI con múltiples entornos de simulación en 2D y 3D, y videos juegos de Atari. Además de miles de entornos desarrollados por third parties o herramientas para desarrollar uno propio.
- **VizDoom**: Librería para desarrollar agentes que juegan Doom usando información visual. Tiene un wrapper de compatibilidad con Gymnasium, pero al ser un desarrollo previo, tiene su propia API en C++, Python y Julia.
 - Está diseñado para ofrecer altas opciones de personalización
 - Admite modos para un jugador y multijugador
 - Creación de escenarios personalizados basado en el editor de mapa de Doom.
 - Basado en **ZDoom**.



CAN IT PLAY DOOM? KINDA...

Armemos una implementación de un algoritmo de AR usando **VizDoom** como entorno de simulación. Dado que vamos a implementar un modelo sencillo, vamos a usar un escenario simple.

Tenemos el siguiente escenario (muy simplificado):

Objetivo: Matar al demonio con la menos cantidad de balas posibles.

El demonio siempre está en el mismo lugar y fijo, se muere de una sola bala.

El jugador solo se puede mover de forma lateral, y solo en 11 posiciones discretas.

Cada episodio puede tener 15 movimientos, y termina si el jugador hace 15 movimientos o el demonio muere.



CAN IT PLAY DOOM? KINDA...

Armemos una implementación de un algoritmo de AR usando **VizDoom** como entorno de simulación. Dado que vamos a implementar un modelo sencillo, vamos a usar un escenario simple.

El agente tiene las siguientes acciones:

- Moverse un bloque a la derecha o a la izquierda de donde esta.
- Disparar.

El agente no puede percibir nada.



CAN IT PLAY DOOM? KINDA...

Armemos una implementación de un algoritmo de AR usando **VizDoom** como entorno de simulación. Dado que vamos a implementar un modelo sencillo, vamos a usar un escenario simple.

El escenario da los siguientes castigos y recompensas:

- **Castigo:**
 - La distancia horizontal en bloques entre el jugador y el demonio. -1 puntos entre bloque de distancia.
 - Si dispara y no le pega al demonio, -20 puntos
- **Recompensa:**
 - 1000 puntos cada vez que le pega al demonio.



CAN IT PLAY DOOM? KINDA...

Q-Learning

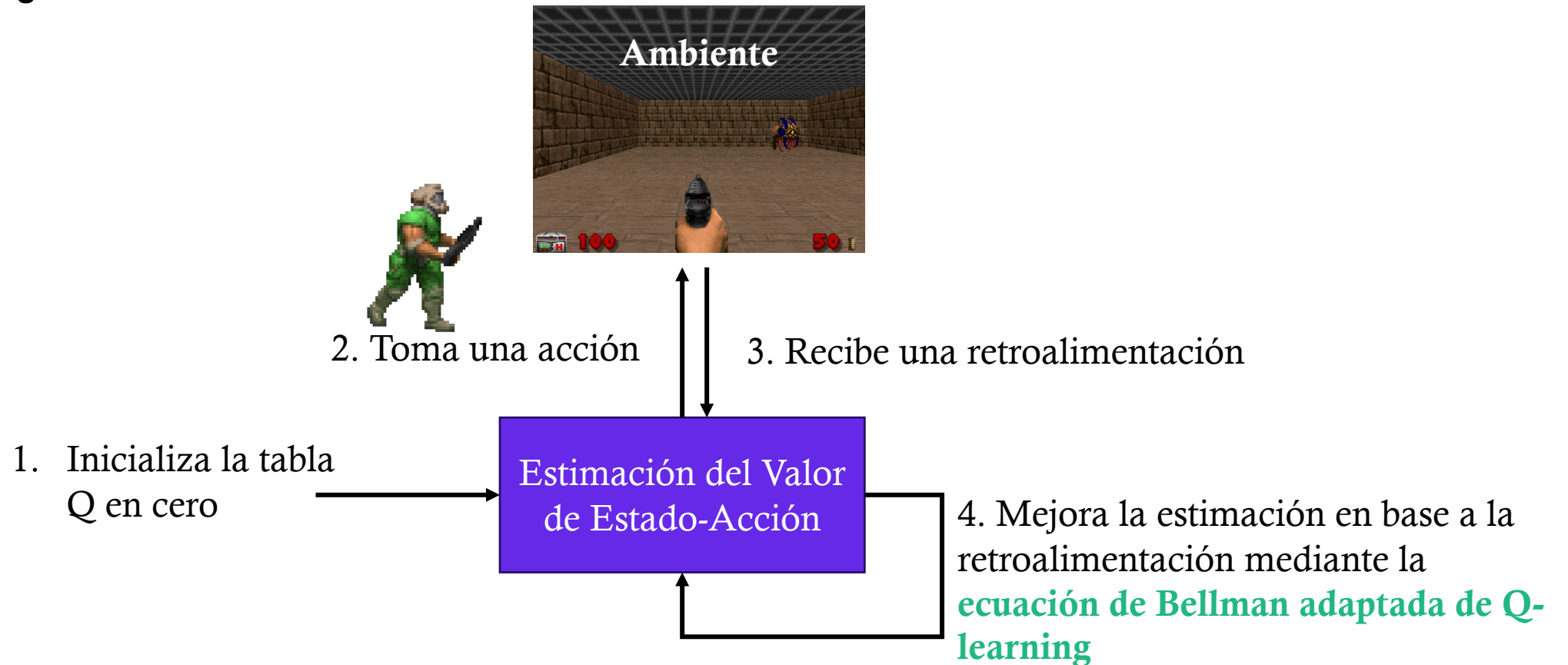
El algoritmo que vamos a implementar es Q-Learning. Este es muy sencillo y por consiguiente no útil en aplicaciones más complejas, por eso este escenario esta simplificado.

Para este problema tenemos 11 estados (uno por posición que el agente puede ubicarse) y 3 acciones (moverse a la derecha, a la izquierda y disparar).

Estado	Derecha	Izquierda	Disparo
0			
1			
...
10			

CAN IT PLAY DOOM? KINDA...

Q-Learning

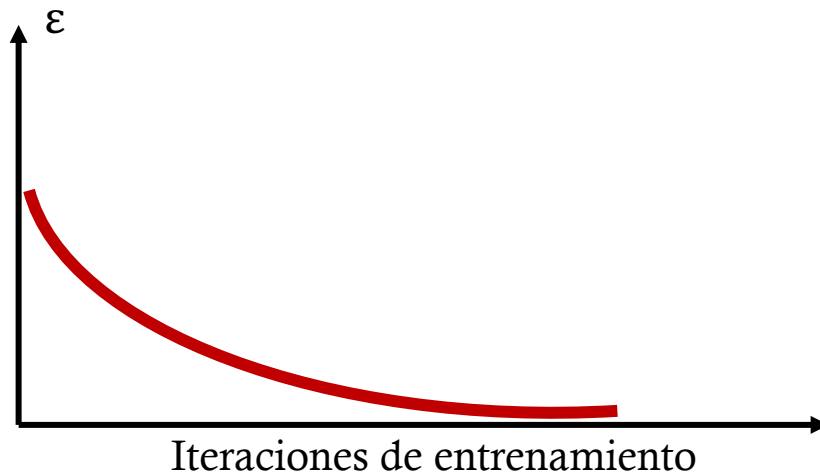


CAN IT PLAY DOOM? KINDA...

Q-Learning

Vimos que los algoritmos basados en valor realizan **exploración** y **explotación** en cada episodio. De la tabla Q, para un estado, se elige la mejor acción (mayor valor) o un valor al azar mediante la estrategia ϵ -Greedy.

Para este problema en particular vamos a usar 100 episodios y $\lambda = 0.03$



$$\epsilon = \epsilon_0 e^{-\lambda t}$$

Si $U[0,1] < \epsilon$, **exploramos**

Si no, **explotación**

CAN IT PLAY DOOM? KINDA...

Q-Learning

Para este caso vamos a actualizar los valores mediante aprendizaje por diferencia temporal:

$$Q(s_i, a_j) = Q(s_i, a_j) + \alpha [(r_{ij} + \gamma \max_k (Q(s_k, :))) - Q(s_i, a_j)]$$

$$Q(s_i, a_j) = (1 - \alpha)Q(s_i, a_j) + \alpha (R_{ij} + \gamma \max_k (Q(s_k, :)))$$

Para este ejercicio usamos $\alpha=0.1$ y $\gamma=0.99$

CAN IT PLAY DOOM? KINDA...

Q-Learning

Por ejemplo, si realizamos un entrenamiento, llegamos a la siguiente tabla:

	Estado	Derecha	Izquierda	Disparo
	0	0	0	0
	1	-0.1	48.8	-0.2
	2	0.7	432.8	-0.2
Acá termina	3	114.2	277.5	1000
	4	989.8	415.9	-1.2
	5	974.8	296.3	-1.3
Acá empieza el agente	6	952.8	6.4	-1.8
	7	312.9	-0.5	-1.7
	8	-0.4	-0.4	-1.2
	9	-0.3	-0.3	-0.6
	10	0	0	0

**La recompensa promedio
para este agente es: 5790**

*La recompensa promedio para
un agente aleatorio es: -5113*