Ejemplo Sencillo: Un MDP de Corredor Simple

Este documento proporciona un ejemplo concreto y a pequeña escala que ilustra todos los conceptos y definiciones clave de Reinforcement Learning (RL) introducidos anteriormente.

---

1. Planteamiento del Problema

```
[L] -- [M] -- [R] -- [T]
```

Representación ASCII del corredor: L (Izquierda), M (Medio, inicio), R (Derecha), y T (Meta terminal).

Un agente se encuentra en un corredor de 3 celdas: Izquierda (L), Medio (M), Derecha (R). El objetivo es llegar al extremo derecho (R) partiendo desde M. Cada paso de tiempo incurre en un costo (recompensa negativa) hasta que se alcanza la meta.

- **Estados**: $\mathcal{S} = {L, M, R, T}$, donde T es el estado terminal o meta.
- **Acciones**: $\mathcal{A} = {\text{Left}, \text{Right}}$ (Izquierda, Derecha).
- **Estado inicial**: $S_0 = M$.

---

2. Formulación del MDP

Definimos el MDP $(\mathcal{S},\mathcal{A},P,R,\gamma)$:

1. **Probabilidades de transición**:
   $
   P(s'|s,a) = \begin{cases}
   1 & \text{si } s'=\begin{cases}L & \text{y }(s,a)=(L,\cdot)\\
   M & \text{y }(s,a)=(L,\text{Right})\\
   R & \text{y }(s,a)=(M,\text{Right})\\
   M & \text{y }(s,a)=(M,\text{Left})\\
   T & \text{y }(s,a)=(R,\text{Right})\\
   R & \text{y }(s,a)=(R,\text{Left})
   \end{cases}\\
   0 & \text{en otro caso}
   \end{cases}
   $
   (es decir, los intentos de moverse más allá de un borde dejan al agente en la misma celda).

2. **Función de recompensa**:
   $
   R(s,a,s') = \begin{cases}
   -1 & \text{para cualquier transición que no lleve a }T,\\
   0 & \text{para transiciones hacia }T.
   $

PROF

\end{cases}
$

3. **Factor de descuento**: $\gamma = 0.9$.

---

3. Política π

Consideremos una **política determinista** simple:

$
\pi(\text{Right}|s) = 1, \quad \forall s \in {L,M,R},
$
lo que significa que el agente siempre se mueve a la derecha.

---

4. Retorno y Objetivo de Desempeño

Comenzando desde M, el agente se mueve a la derecha dos veces para alcanzar el estado terminal:

- Secuencia: $M \xrightarrow{\text{Right}} R \xrightarrow{\text{Right}} T$.
- Recompensas: $-1, -1, 0$.

El **retorno**:
$
G_0 = -1 + \gamma\cdot(-1) + \gamma^2\cdot 0 = -1 - 0.9 = -1.9.
$

El **objetivo de desempeño** para esta política:
$
J = \mathbb{E}[G_0] = -1.9.
$

PROF

5. Función de Valor de Estado $V^\pi(s)$

Resolvemos las ecuaciones de Bellman para V^π:

$
\begin{cases}
V(L) = -1 + 0.9,V(M),\
V(M) = -1 + 0.9,V(R),\
V(R) = -1 + 0.9,V(T),\
V(T) = 0.
\end{cases}
$

Sustituimos $V(T)=0$:

1. $V(R) = -1 + 0.9\cdot0 = -1.$

2. $V(M) = -1 + 0.9(-1) = -1 -0.9 = -1.9.$
3. $V(L) = -1 + 0.9(-1.9) = -1 -1.71 = -2.71.$

Por lo tanto,
$
V^\pi = {-2.71, -1.9, -1, 0}\quad\text{para }{L,M,R,T}.
$

5.1 Visualización de V^π(s)

```
[L: -2.71] -- [M: -1.9] -- [R: -1] -- [T: 0]
```

Una representación visual de la función de valor de estado anotada en el corredor: valores menos negativos están más cerca de la meta.

---

6. Función de Valor-Acción $Q^\pi(s,a)$

Por definición:
$
Q^\pi(s,a) = R(s,a,s') + \gamma V^\pi(s').
$

Por ejemplo:

- $Q^\pi(M,\text{Right}) = -1 + 0.9,V(R) = -1 + 0.9(-1) = -1.9.$
- $Q^\pi(M,\text{Left}) = -1 + 0.9,V(M) = -1 + 0.9(-1.9) = -2.71.$

Esto muestra que moverse a la derecha es estrictamente mejor en M.

PROF

6.1 Visualización de Q^π(s,a)

```
        Left    Right
L       -2.71   -2.71
M       -2.71   -1.9
R       -1.0    -1.0
T        0       0
```

Una tabla ASCII que muestra los valores de acción para cada par estado-acción. En M, elegir Right (-1.9) produce un mejor retorno esperado que Left (-2.71).

Cálculo Explícito de Q^π en M

Usando la definición de Bellman $Q^\pi(s,a) = R(s,a,s') + \gamma V^\pi(s')$, en el estado M obtenemos $Q^\pi(M,\text{Left}) = -1 + 0.9 V(M) = -2.71$ y $Q^\pi(M,\text{Right}) = -1 + 0.9 V(R) = -1.9$. Esto claramente muestra por qué se prefiere 'Right' en M.

7. Intuición

- Los **estados** son las celdas del corredor.
- Las **acciones** mueven izquierda/derecha; chocar contra una pared te deja en el mismo lugar.
- Las **recompensas** penalizan cada paso hasta la meta, incentivando caminos más cortos.
- La **política** siempre se mueve a la derecha.
- El **valor** de cada estado es igual al costo total esperado comenzando desde ahí.
- El **valor-acción** indica que moverse a la derecha es mejor que a la izquierda en cada estado no terminal.

Este pequeño ejemplo concretiza cómo interactúan los MDPs, los retornos, las funciones de valor y las políticas en RL.

Conexión con la Teoría: Este ejemplo del corredor instancia las definiciones del documento introductorio:

- **Sección 2 (MDP)**: espacio de estados {L, M, R, T}, espacio de acciones {Left, Right}, probabilidades de transición de los movimientos en el corredor, y función de recompensa que penaliza cada paso.
- **Sección 3 (Política)**: política determinista pi(s) = Right para todos los estados no terminales.
- **Sección 4 (Retorno y Objetivo de Desempeño)**: retorno $G_0 = -1 + 0.9(-1) = -1.9$, por lo que $J = E[G_0] = -1.9$.
- **Sección 5 (Funciones de Valor)**: V^pi resuelto mediante ecuaciones de Bellman, dando valores {-2.71, -1.9, -1, 0}.
- **Sección 6 (Valor-Acción y Ecuaciones de Bellman)**: Q^pi(s,a) = R + gamma V^pi(s'), con derivación explícita en M en la sección 6.1.

---

8. Comparación Resumen de Conceptos de RL

PROF

Concepto	Símbolo	Ejemplo del Corredor
Espacio de estados	$\mathcal{S}$	{L, M, R, T}
Espacio de acciones	$\mathcal{A}$	{Left, Right}
Política	$\pi(a	s)$
Retorno	G_t	$-1 + 0.9(-1) = -1.9$
Objetivo de desempeño	J	$\mathbb{E}[G_0] = -1.9$
Función de valor	$V^\pi(s)$	$V(M) = -1.9$, etc.
Función valor-acción	$Q^\pi(s,a)$	$Q(M,Right) = -1.9$, etc.
Ecuación de Bellman	Relación recursiva para V y Q	Ver secciones 5 y 6

9. Comparación Numérica de V y Q

Estado	V^pi(s)	Q^pi(s, Left)	Q^pi(s, Right)
L	-2.71	-2.71	-2.71
M	-1.9	-2.71	-1.9
R	-1.0	-1.0	-1.0
T	0	0	0