Ejemplo Sencillo: Un MDP de Corredor Simple con Diferencias On-policy y Off-policy

Este documento proporciona un ejemplo concreto que ilustra conceptos clave de Reinforcement Learning (RL), destacando la diferencia entre aprendizaje on-policy y off-policy.

---

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 (T) 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 $P(s'|s,a)$**: Son deterministas. Si el agente intenta moverse a una celda válida, lo logra; si intenta moverse contra un límite (izquierda desde L, derecha desde R), se queda en la misma celda. El estado T es terminal. Específicamente:
   - $P(L | L, \text{Left}) = 1$
   - $P(M | L, \text{Right}) = 1$
   - $P(L | M, \text{Left}) = 1$
   - $P(R | M, \text{Right}) = 1$
   - $P(M | R, \text{Left}) = 1$
   - $P(T | R, \text{Right}) = 1$
   - Para todas las demás combinaciones de (s, a, s'), $P(s'|s,a) = 0$.
   - Una vez en T, el episodio termina. $V(T) = 0$.

2. **Función de recompensa $R(s,a,s')$**: Se recibe una recompensa de -1 por cada paso, excepto cuando se llega al estado terminal T, donde la recompensa es 0.
   $
   R(s,a,s') = \begin{cases}
   -1 & \text{si } s' \neq T \\
   0 & \text{si } s' = T

\end{cases}
$
(Nota: En la práctica, la recompensa se asocia a menudo solo con el estado de llegada s', o con la tupla (s, a, s') como aquí).

3. **Factor de descuento**: $\gamma = 0.9$. Un valor menor que 1 asegura que las recompensas futuras valen menos que las inmediatas y que los valores convergen.

---

3. Comparación On-Policy vs Off-Policy

Para ilustrar la diferencia, consideremos dos políticas diferentes:

Política óptima (π)

Una política determinista que siempre elige Right:
$\pi(a|s) = \begin{cases} 1 & \text{si } a = \text{Right} \\ 0 & \text{si } a = \text{Left} \end{cases}$ para todo estado $s \in {L,M,R}$.

Política exploratoria (μ)

Una política estocástica que explora ocasionalmente:
$\mu(a|s) = \begin{cases} 0.8 & \text{si } a = \text{Right} \\ 0.2 & \text{si } a = \text{Left} \end{cases}$ para todo estado $s \in {L,M,R}$.

---

4. Diferencia Conceptual: On-Policy vs Off-Policy

Aprendizaje On-Policy

- **Definición**: Aprendemos sobre la misma política que estamos siguiendo.
- **Ejemplo**: Si seguimos μ (política exploratoria), aprendemos los valores de μ.
- **Característica**: Lo que aprendo es exactamente sobre lo que hago.
- **Función de valor**: $V^{\mu}(s)$ - valor del estado s bajo la política μ.
- **Función de valor-acción**: $Q^{\mu}(s,a)$ - valor de tomar la acción a en el estado s y luego seguir μ.

PROF

Aprendizaje Off-Policy

- **Definición**: Aprendemos sobre una política mientras seguimos otra diferente.
- **Ejemplo**: Seguimos μ (exploramos), pero aprendemos los valores de π (política óptima).
- **Característica**: Puedo explorar con una política, pero aprender sobre otra.
- **Función de valor**: $V^{\pi}(s)$ - valor del estado s bajo la política π (que no estamos siguiendo).
- **Función de valor-acción**: $Q^{\pi}(s,a)$ - valor de tomar la acción a en el estado s y luego seguir π.

---

5. Ecuaciones de Bellman para Ambas Políticas

Las ecuaciones de Bellman relacionan el valor de un estado (o par estado-acción) con los valores de los estados sucesores. Son la base para calcular las funciones de valor V y Q.

Ecuaciones de Bellman para la política π (Siempre Derecha)

La ecuación general para $V^{\pi}(s)$ es:
$V^{\pi}(s) = \sum_{a} \pi(a|s) \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma V^{\pi}(s')]$

Dado que $\pi(\text{Right}|s) = 1$ y $\pi(\text{Left}|s) = 0$ para $s \in {L, M, R}$:
$V^{\pi}(s) = \sum_{s'} P(s'|s,\text{Right})[R(s,\text{Right},s') + \gamma V^{\pi}(s')]$

Aplicando a nuestros estados (recordando que $V^{\pi}(T)=0$):

- **Estado L**: $V^{\pi}(L) = P(M|L,\text{Right})[R(L,\text{Right},M) + \gamma V^{\pi}(M)] = 1 \times [-1 + 0.9 V^{\pi}(M)]$
- **Estado M**: $V^{\pi}(M) = P(R|M,\text{Right})[R(M,\text{Right},R) + \gamma V^{\pi}(R)] = 1 \times [-1 + 0.9 V^{\pi}(R)]$
- **Estado R**: $V^{\pi}(R) = P(T|R,\text{Right})[R(R,\text{Right},T) + \gamma V^{\pi}(T)] = 1 \times [0 + 0.9 \times 0] = 0$
    - *Corrección*: La recompensa al llegar a T es 0, pero la transición *desde* R ocurre *antes* de llegar. La recompensa asociada a la *acción* de ir de R a T es -1 (el último paso), aunque el valor de T sea 0. Así, la ecuación correcta es:
    - **Estado R (corregido)**: $V^{\pi}(R) = P(T|R,\text{Right})[R(R,\text{Right},T) + \gamma V^{\pi}(T)] = 1 \times [-1 + 0.9 \times 0] = -1.0$

Resolviendo el sistema (de abajo hacia arriba):

1. $V^{\pi}(R) = -1.0$
2. $V^{\pi}(M) = -1 + 0.9 V^{\pi}(R) = -1 + 0.9 (-1.0) = -1 - 0.9 = -1.9$
3. $V^{\pi}(L) = -1 + 0.9 V^{\pi}(M) = -1 + 0.9 (-1.9) = -1 - 1.71 = -2.71$

Ecuaciones de Bellman para la política μ (Exploratoria: 80% Derecha, 20% Izquierda)

PROF

La ecuación general para $V^{\mu}(s)$ es:
$V^{\mu}(s) = \sum_{a} \mu(a|s) \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma V^{\mu}(s')]$

Aplicando a nuestros estados ($\mu(\text{Right}|s)=0.8$, $\mu(\text{Left}|s)=0.2$):
$V^{\mu}(s) = 0.8 \sum_{s'} P(s'|s,\text{Right})[R(s,\text{Right},s') + \gamma V^{\mu}(s')] + 0.2 \sum_{s'} P(s'|s,\text{Left})[R(s,\text{Left},s') + \gamma V^{\mu}(s')]$

- **Estado L**:
  $V^{\mu}(L) = 0.8 [P(M|L,\text{Right})(R(L,\text{Right},M) + \gamma V^{\mu}(M))] + 0.2 [P(L|L,\text{Left})(R(L,\text{Left},L) + \gamma V^{\mu}(L))]$
  $V^{\mu}(L) = 0.8 [-1 + \gamma V^{\mu}(M)] + 0.2 [-1 + \gamma V^{\mu}(L)]$
- **Estado M**:
  $V^{\mu}(M) = 0.8 [P(R|M,\text{Right})(R(M,\text{Right},R) + \gamma V^{\mu}(R))] + 0.2 [P(L|M,\text{Left})(R(M,\text{Left},L) + \gamma V^{\mu}(L))]$
  $V^{\mu}(M) = 0.8 [-1 + \gamma V^{\mu}(R)] + 0.2 [-1 + \gamma V^{\mu}(L)]$

- **Estado R**:
  $V^{\mu}(R) = 0.8 [P(T|R,\text{Right})(R(R,\text{Right},T) + \gamma V^{\mu}(T))] + 0.2 [P(M|R,\text{Left})(R(R,\text{Left},M) + \gamma V^{\mu}(M))]$
  $V^{\mu}(R) = 0.8 [-1 + \gamma \times 0] + 0.2 [-1 + \gamma V^{\mu}(M)]$
  $V^{\mu}(R) = -0.8 + 0.2 [-1 + 0.9 V^{\mu}(M)]$

Sustituyendo $\gamma=0.9$:

1. $V^{\mu}(L) = 0.8 [-1 + 0.9 V^{\mu}(M)] + 0.2 [-1 + 0.9 V^{\mu}(L)]$
2. $V^{\mu}(M) = 0.8 [-1 + 0.9 V^{\mu}(R)] + 0.2 [-1 + 0.9 V^{\mu}(L)]$
3. $V^{\mu}(R) = -0.8 + 0.2 [-1 + 0.9 V^{\mu}(M)]$

Este es un sistema de 3 ecuaciones lineales con 3 incógnitas ($V^{\mu}(L), V^{\mu}(M), V^{\mu}(R)$). Resolviéndolo (por sustitución o métodos matriciales) obtenemos los valores aproximados mencionados más adelante.

Ecuación de Bellman de Optimalidad (Base para Q-learning)

Q-learning no calcula V^{π} o V^{μ} directamente, sino que busca la función de valor-acción óptima $Q^*(s, a)$, que satisface la ecuación de Bellman de optimalidad:

$Q^*(s, a) = \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma \max_{a'} Q^*(s', a')]$

Esta ecuación establece que el valor óptimo de tomar la acción a en el estado s es la recompensa inmediata más el valor descontado del *mejor* valor Q posible desde el siguiente estado s'. Q-learning utiliza esta relación para actualizar sus estimaciones de Q^*.

---

6. Comparación de Valores y Cálculo Detallado

Aquí calcularemos explícitamente los valores V y Q para ambas políticas.

Valores bajo la política óptima π (Siempre Derecha)

PROF

Ya resolvimos el sistema para V^{π} en la sección anterior:

- $V^{\pi}(R) = -1.0$
- $V^{\pi}(M) = -1.9$
- $V^{\pi}(L) = -2.71$

Ahora calculemos los valores $Q^{\pi}(s,a)$ usando la definición:
$Q^{\pi}(s,a) = \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma V^{\pi}(s')]$

- **Estado L**:
  - $Q^{\pi}(L, \text{Left}) = P(L|L,\text{Left})[R(L,\text{Left},L) + \gamma V^{\pi}(L)] = 1 \times [-1 + 0.9(-2.71)] = -1 - 2.439 = -3.439$
  - $Q^{\pi}(L, \text{Right}) = P(M|L,\text{Right})[R(L,\text{Right},M) + \gamma V^{\pi}(M)] = 1 \times [-1 + 0.9(-1.9)] = -1 - 1.71 = -2.71$
- **Estado M**:

    - $Q^{\pi}(M, \text{Left}) = P(L|M,\text{Left})[R(M,\text{Left},L) + \gamma V^{\pi}(L)] = 1 \times [-1 + 0.9(-2.71)] = -1 - 2.439 = -3.439$
    - $Q^{\pi}(M, \text{Right}) = P(R|M,\text{Right})[R(M,\text{Right},R) + \gamma V^{\pi}(R)] = 1 \times [-1 + 0.9(-1.0)] = -1 - 0.9 = -1.9$
- **Estado R**:
    - $Q^{\pi}(R, \text{Left}) = P(M|R,\text{Left})[R(R,\text{Left},M) + \gamma V^{\pi}(M)] = 1 \times [-1 + 0.9(-1.9)] = -1 - 1.71 = -2.71$
    - $Q^{\pi}(R, \text{Right}) = P(T|R,\text{Right})[R(R,\text{Right},T) + \gamma V^{\pi}(T)] = 1 \times [-1 + 0.9(0)] = -1.0$
- **Estado T**: $Q^{\pi}(T, \text{Left}) = 0$, $Q^{\pi}(T, \text{Right}) = 0$ (Estado terminal)

Observa que $V^{\pi}(s) = Q^{\pi}(s, \text{Right})$ porque la política π *siempre* elige Right. El valor $Q^{\pi}(s, \text{Left})$ representa el valor si *forzáramos* la acción Left una vez y *luego* siguiéramos la política óptima π (siempre Right).

Valores bajo la política exploratoria μ (80% Derecha, 20% Izquierda)

Resolver el sistema de ecuaciones para V^{μ} de la sección anterior:

1. $V^{\mu}(L) = 0.8 [-1 + 0.9 V^{\mu}(M)] + 0.2 [-1 + 0.9 V^{\mu}(L)]$
2. $V^{\mu}(M) = 0.8 [-1 + 0.9 V^{\mu}(R)] + 0.2 [-1 + 0.9 V^{\mu}(L)]$
3. $V^{\mu}(R) = -1.0 + 0.18 V^{\mu}(M)$

Sustituyendo (3) en (2):
$V^{\mu}(M) = 0.8 [-1 + 0.9 (-1.0 + 0.18 V^{\mu}(M))] + 0.2 [-1 + 0.9 V^{\mu}(L)]$
$V^{\mu}(M) = 0.8 [-1 - 0.9 + 0.162 V^{\mu}(M)] + 0.2 [-1 + 0.9 V^{\mu}(L)]$
$V^{\mu}(M) = 0.8 [-1.9 + 0.162 V^{\mu}(M)] - 0.2 + 0.18 V^{\mu}(L)$
$V^{\mu}(M) = -1.52 + 0.1296 V^{\mu}(M) - 0.2 + 0.18 V^{\mu}(L)$
$V^{\mu}(M) (1 - 0.1296) = -1.72 + 0.18 V^{\mu}(L)$
$0.8704 V^{\mu}(M) = -1.72 + 0.18 V^{\mu}(L)$ (Ecuación 4)

De la ecuación (1):
$V^{\mu}(L) = -0.8 + 0.72 V^{\mu}(M) - 0.2 + 0.18 V^{\mu}(L)$
$V^{\mu}(L) (1 - 0.18) = -1.0 + 0.72 V^{\mu}(M)$
$0.82 V^{\mu}(L) = -1.0 + 0.72 V^{\mu}(M)$
$V^{\mu}(L) = \frac{-1.0 + 0.72 V^{\mu}(M)}{0.82}$ (Ecuación 5)

Sustituyendo (5) en (4):
$0.8704 V^{\mu}(M) = -1.72 + 0.18 \left(\frac{-1.0 + 0.72 V^{\mu}(M)}{0.82} \right)$
$0.8704 V^{\mu}(M) = -1.72 + \frac{-0.18 + 0.1296 V^{\mu}(M)}{0.82}$
Multiplicando por 0.82:
$0.713728 V^{\mu}(M) = -1.4104 - 0.18 + 0.1296 V^{\mu}(M)$
$0.713728 V^{\mu}(M) - 0.1296 V^{\mu}(M) = -1.5904$
$0.584128 V^{\mu}(M) = -1.5904$
$V^{\mu}(M) = \frac{-1.5904}{0.584128} \approx -2.7227$

Ahora encontramos $V^{\mu}(L)$ y $V^{\mu}(R)$:
$V^{\mu}(L) = \frac{-1.0 + 0.72 (-2.7227)}{0.82} \approx \frac{-1.0 - 1.9603}{0.82} = \frac{-2.9603}{0.82}$

PROF

\approx -3.6101$
$V^{\mu}(R) = -1.0 + 0.18 (-2.7227) \approx -1.0 - 0.4901 = -1.4901$

Resumen de valores V (aproximados):

- $V^{\mu}(L) \approx -3.61$
- $V^{\mu}(M) \approx -2.72$
- $V^{\mu}(R) \approx -1.49$

(Nota: Estos valores difieren ligeramente de los -3.0, -2.3, -1.2 estimados previamente. La diferencia puede deberse a la precisión de los cálculos o a una posible simplificación en la estimación inicial. Usaremos estos valores calculados más precisos).

Ahora calculemos los valores $Q^{\mu}(s,a)$:
$Q^{\mu}(s,a) = \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma V^{\mu}(s')]$

- **Estado L**:
  - $Q^{\mu}(L, \text{Left}) = P(L|L,\text{Left})[R(L,\text{Left},L) + \gamma V^{\mu}(L)] \approx 1 \times [-1 + 0.9(-3.61)] = -1 - 3.249 = -4.249$
  - $Q^{\mu}(L, \text{Right}) = P(M|L,\text{Right})[R(L,\text{Right},M) + \gamma V^{\mu}(M)] \approx 1 \times [-1 + 0.9(-2.72)] = -1 - 2.448 = -3.448$
- **Estado M**:
  - $Q^{\mu}(M, \text{Left}) = P(L|M,\text{Left})[R(M,\text{Left},L) + \gamma V^{\mu}(L)] \approx 1 \times [-1 + 0.9(-3.61)] = -1 - 3.249 = -4.249$
  - $Q^{\mu}(M, \text{Right}) = P(R|M,\text{Right})[R(M,\text{Right},R) + \gamma V^{\mu}(R)] \approx 1 \times [-1 + 0.9(-1.49)] = -1 - 1.341 = -2.341$
- **Estado R**:
  - $Q^{\mu}(R, \text{Left}) = P(M|R,\text{Left})[R(R,\text{Left},M) + \gamma V^{\mu}(M)] \approx 1 \times [-1 + 0.9(-2.72)] = -1 - 2.448 = -3.448$
  - $Q^{\mu}(R, \text{Right}) = P(T|R,\text{Right})[R(R,\text{Right},T) + \gamma V^{\mu}(T)] = 1 \times [-1 + 0.9(0)] = -1.0$
- **Estado T**: $Q^{\mu}(T, \text{Left}) = 0$, $Q^{\mu}(T, \text{Right}) = 0$

PROF Verificación: Podemos comprobar si $V^{\mu}(s)$ es igual a la esperanza de $Q^{\mu}(s,a)$ bajo μ:
$V^{\mu}(s) \approx 0.8 Q^{\mu}(s,\text{Right}) + 0.2 Q^{\mu}(s,\text{Left})$

- $V^{\mu}(L) \approx 0.8(-3.448) + 0.2(-4.249) = -2.7584 - 0.8498 = -3.6082 \approx -3.61$ (Coincide)
- $V^{\mu}(M) \approx 0.8(-2.341) + 0.2(-4.249) = -1.8728 - 0.8498 = -2.7226 \approx -2.72$ (Coincide)
- $V^{\mu}(R) \approx 0.8(-1.0) + 0.2(-3.448) = -0.8 - 0.6896 = -1.4896 \approx -1.49$ (Coincide)

Los cálculos son consistentes. La política exploratoria μ resulta en valores de estado y acción peores (más negativos) que la política óptima π, ya que a veces toma la acción Left, lo que retrasa la llegada a la meta T.

---

7. Ejemplos Detallados de Actualizaciones On-Policy y Off-Policy

Para ilustrar concretamente la diferencia entre métodos on-policy y off-policy, vamos a ver paso a paso cómo funcionarían SARSA (on-policy) y Q-learning (off-policy) en nuestro ejemplo del corredor.

Ejemplo On-Policy: SARSA

SARSA es un algoritmo on-policy que actualiza valores Q basados en las acciones realmente tomadas según la política actual.

Fórmula de actualización SARSA:
$Q^{\mu}(s_t, a_t) \leftarrow Q^{\mu}(s_t, a_t) + \alpha [r_{t+1} + \gamma Q^{\mu}(s_{t+1}, a_{t+1}) - Q^{\mu}(s_t, a_t)]$

Ejemplo concreto:
Supongamos que seguimos la política exploratoria μ y estamos en el estado M. Inicializamos todos los valores Q^{μ} en -1.0.

1. **Estado actual**: $s_t = M$
2. **Elegimos acción** según μ: $a_t = \text{Right}$ (con probabilidad 0.8)
3. **Observamos**: $r_{t+1} = -1$, $s_{t+1} = R$
4. **Elegimos siguiente acción** según μ: $a_{t+1} = \text{Right}$ (con probabilidad 0.8)
5. **Actualización SARSA** (con $\alpha = 0.1$):
   $Q^{\mu}(M, \text{Right}) \leftarrow -1.0 + 0.1 [-1 + 0.9 \times Q^{\mu}(R,\text{Right}) - Q^{\mu}(M, \text{Right})]$
   $Q^{\mu}(M, \text{Right}) \leftarrow -1.0 + 0.1 [-1 + 0.9 \times (-1.0) - (-1.0)]$
   $Q^{\mu}(M, \text{Right}) \leftarrow -1.0 + 0.1 [-1 - 0.9 + 1.0]$
   $Q^{\mu}(M, \text{Right}) \leftarrow -1.0 + 0.1 [-0.9]$
   $Q^{\mu}(M, \text{Right}) \leftarrow -1.0 - 0.09 = -1.09$

Después de muchas iteraciones, $Q^{\mu}(M, \text{Right})$ convergerá a $Q^{\mu}(M, \text{Right}) \approx -2.34$, el valor real bajo la política μ calculado en la sección 6.

Importante: En SARSA, usamos la acción a_{t+1} que realmente tomaremos según μ, incluso si no es la óptima. Esto hace que SARSA aprenda el valor real de seguir μ, incluyendo sus movimientos exploratorios.

PROF

Ejemplo Off-Policy: Q-learning

Q-learning es un algoritmo off-policy que actualiza valores Q basándose en la acción óptima para el siguiente estado, independientemente de qué política estamos siguiendo.

Fórmula de actualización Q-learning:
$Q^(s_t, a_t) \leftarrow Q^(s_t, a_t) + \alpha [r_{t+1} + \gamma \max_a Q^(s_{t+1}, a) - Q^(s_t, a_t)]$

Ejemplo concreto:
Misma situación que antes, seguimos μ pero aprenderemos valores Q óptimos Q^* (que corresponden a la política óptima π).

1. **Estado actual**: $s_t = M$
2. **Elegimos acción** según μ: $a_t = \text{Right}$ (con probabilidad 0.8)

3. **Observamos**: $r_{t+1} = -1$, $s_{t+1} = R$
4. **Actualización Q-learning** (con $\alpha = 0.1$):
   $Q^*(M, \text{Right}) \leftarrow Q^*(M, \text{Right}) + \alpha [r_{t+1} + \gamma \max_a Q^*(R, a) - Q^*(M, \text{Right})]$
   $Q^*(M, \text{Right}) \leftarrow -1.0 + 0.1 [-1 + 0.9 \times \max(Q^*(R,\text{Left}), Q^*(R,\text{Right})) - (-1.0)]$
   $Q^*(M, \text{Right}) \leftarrow -1.0 + 0.1 [-1 + 0.9 \times \max(-1.0, -1.0) - (-1.0)]$
   $Q^*(M, \text{Right}) \leftarrow -1.0 + 0.1 [-1 - 0.9 + 1.0]$
   $Q^*(M, \text{Right}) \leftarrow -1.0 + 0.1 [-0.9]$
   $Q^*(M, \text{Right}) \leftarrow -1.0 - 0.09 = -1.09$

Después de muchas iteraciones, $Q^*(M, \text{Right})$ convergerá a $Q^*(M, \text{Right}) = Q^{\pi}(M, \text{Right}) = -1.9$, el valor óptimo calculado en la sección 6.

Diferencia clave: En Q-learning, usamos $\max_a Q^*(s_{t+1}, a)$ para la actualización, no la acción que realmente tomaremos según μ. Esto permite que Q-learning aprenda los valores Q óptimos Q^* (que corresponden a la política óptima π) independientemente de qué política estamos siguiendo.

---

8. El Desafío Off-Policy: ¿Cómo aprender π mientras seguimos μ?

Problema: Si seguimos μ (explorando), ¿cómo podemos estimar los valores de π?

Este es el desafío central del aprendizaje off-policy: necesitamos una manera de "corregir" las experiencias obtenidas con μ para estimar los valores bajo π.

Solución 1: Q-learning

El Q-learning es un método off-policy que aprende directamente los valores Q^* óptimos:

$Q^_{t+1}(s_t, a_t) \leftarrow Q^t(s_t, a_t) + \alpha [r_{t+1} + \gamma \max_a Q^t(s_{t+1}, a) - Q^_t(s_t, a_t)]$

PROF

Con Q-learning:

- Exploramos usando μ
- Actualizamos Q hacia los valores óptimos Q^* usando $\max_a Q^*$
- No necesitamos conocer explícitamente π

Solución 2: Importance Sampling

Si queremos estimar directamente V^{π} mientras seguimos μ:

1. **Concepto**: Reponderar las experiencias según la relación entre π y μ
2. **Idea básica**: Las trayectorias que son más probables bajo π que bajo μ deben tener más peso

El peso de importance sampling para una trayectoria $\tau = (s_0, a_0, s_1, a_1, ..., s_T)$ es:

$\rho_{\tau} = \prod_{t=0}^{T-1} \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}$

Por ejemplo, para la trayectoria M → R → T:

- Probabilidad bajo π: $\pi(\text{Right}|M) \times \pi(\text{Right}|R) = 1.0 \times 1.0 = 1.0$
- Probabilidad bajo μ: $\mu(\text{Right}|M) \times \mu(\text{Right}|R) = 0.8 \times 0.8 = 0.64$
- Ratio = $\frac{1.0}{0.64} = 1.5625$

Este ratio nos permite ajustar los retornos observados siguiendo μ para estimar los retornos esperados bajo π.

Visualizacion

9. Intuición: ¿Cuándo Usar On-Policy vs Off-Policy?

Entender cuándo usar cada enfoque es crucial para aplicar RL de manera efectiva:

Ventajas del Aprendizaje On-Policy (SARSA)

1. **Más seguro para aprender en entornos reales**:
    - Como aprende sobre lo que realmente hace, considera los riesgos de exploración
    - Ejemplo: Un robot que aprende a caminar evitará acciones que puedan provocar caídas frecuentes
2. **Más estable durante el entrenamiento**:
    - Menor varianza en las actualizaciones
    - Convergencia más suave pero potencialmente más lenta
3. **Mejor cuando no existe una política óptima clara**:
    - En algunos problemas con aleatoriedad o adversarios, no hay una política "perfecta"
    - On-policy aprende a optimizar la política que realmente estamos usando
4. **Ideal para problemas donde exploración y explotación están balanceadas**:
    - En nuestro ejemplo del corredor: SARSA aprendería que, bajo μ, $Q^{\mu}(M, \text{Right}) \approx -2.34$, reconociendo que a veces tomaremos Left

PROF

Ventajas del Aprendizaje Off-Policy (Q-learning)

1. **Mayor eficiencia en el uso de datos**:
    - Puede aprender de cualquier experiencia, incluso de datos recolectados por políticas antiguas
    - Permite reutilizar experiencias pasadas (experience replay)
2. **Puede aprender la política óptima mientras explora**:
    - Aprende Q^* directamente, sin importar qué política estamos siguiendo

- En nuestro ejemplo: Q-learning aprenderá que $Q^*(M, \text{Right}) = Q^{\pi}(M, \text{Right}) = -1.9$ (el valor óptimo), ignorando las exploraciones de μ

3. **Permite separar exploración de aprendizaje**:
   - Podemos usar una política muy exploratoria sin comprometer el aprendizaje
   - Útil cuando la exploración puede ser costosa o peligrosa

4. **Necesario para aprendizaje desde observación**:
   - Permite aprender de demostraciones humanas o de otros agentes
   - No necesitamos seguir la misma política que queremos aprender

Consideraciones Prácticas

- **Complejidad computacional**: Off-policy suele requerir más cómputo
- **Estabilidad**: On-policy es generalmente más estable, off-policy puede diverger
- **Tipo de problema**:
  - Problemas seguros, donde la exploración no es costosa → Off-policy
  - Problemas de seguridad crítica → On-policy
  - Problemas donde los datos son escasos → Off-policy (reutiliza datos)

En nuestro ejemplo del corredor, la diferencia de rendimiento no es dramática ($V^{\pi}(M) = -1.9$ vs $V^{\mu}(M) \approx -2.3$), pero en problemas más complejos, elegir el enfoque correcto puede marcar una gran diferencia en rendimiento, estabilidad y seguridad.

---

10. Visualización Comparativa de Funciones de Valor

Aquí presentamos los valores V y Q calculados con mayor precisión en la Sección 6.

PROF

```
Valores de Estado V:
[L] --- [M] --- [R] --- [T]
-2.71  -1.90  -1.00    0     (V^π - Política Óptima)
-3.61  -2.72  -1.49    0     (V^μ - Política Exploratoria)
```

Nota: $V(s)$ representa el retorno esperado total (suma de recompensas descontadas) comenzando en s y siguiendo la política correspondiente.

```
Valores Q para la política óptima π:
Estado | Q^π(s,Left) | Q^π(s,Right)
-------|-------------|-------------
  L    |   -3.439    |   -2.710
  M    |   -3.439    |   -1.900
  R    |   -2.710    |   -1.000
  T    |     0       |     0
```

Nota: $Q^{\pi}(s,a)$ es el retorno esperado si se toma la acción a en el estado s y luego se sigue la política π. Como π siempre elige 'Right', $V^{\pi}(s) = Q^{\pi}(s, \text{Right})$.

```
Valores Q para la política exploratoria μ:
Estado | Q^μ(s,Left) | Q^μ(s,Right)
-------|-------------|-------------
  L    |   -4.249    |   -3.448
  M    |   -4.249    |   -2.341
  R    |   -3.448    |   -1.000
  T    |     0       |     0
```

Nota: $Q^{\mu}(s,a)$ es el retorno esperado si se toma la acción a en el estado s y luego se sigue la política μ. Aquí, $V^{\mu}(s) = 0.8 Q^{\mu}(s, \text{Right}) + 0.2 Q^{\mu}(s, \text{Left})$. Los valores son peores que con π debido a la exploración (acción 'Left').

---

11. Métodos Típicos

Tipo	Método	Característica	Actualización
On-Policy	SARSA	Aprende Q^{μ} para la política actual	$Q^{\mu}(s_t,a_t) \leftarrow Q^{\mu}(s_t,a_t) + \alpha[r_{t+1} + \gamma Q^{\mu}(s_{t+1},a_{t+1}) - Q^{\mu}(s_t,a_t)]$
	Actor-Critic	Actualiza política y función de valor en paralelo	Actualiza V^{μ} y μ simultáneamente
Off-Policy	Q-learning	Aprende valores Q^* óptimos	$Q^(s_t,a_t) \leftarrow Q^(s_t,a_t) + \alpha[r_{t+1} + \gamma \max_a Q^(s_{t+1},a) - Q^(s_t,a_t)]$
	DQN	Versión de Q-learning con redes neuronales	Aprende Q^* mientras sigue μ (ε-greedy)

PROF

12. Ventajas y Desventajas

Aspecto	On-Policy	Off-Policy
Exploración	La política μ debe balancear exploración y explotación	Puede usar una política μ altamente exploratoria mientras se aprende π
Estabilidad	Más estable, menor varianza	Puede tener alta varianza (especialmente con importance sampling)
Eficiencia de datos	Menos eficiente (μ cambia, los datos viejos no aplican)	Más eficiente (puede reutilizar datos aunque π cambie)

Aspecto	On-Policy	Off-Policy
Complejidad	Generalmente más simple	Suele ser más complejo debido a las correcciones necesarias

13. Conclusión

La distinción entre on-policy y off-policy es fundamental en RL:

- **On-policy**: "aprendo sobre lo que hago" (μ → V^{μ}, Q^{μ})
- **Off-policy**: "hago una cosa, aprendo sobre otra" (μ → V^{π}, Q^{π} o Q^*)

En nuestro ejemplo del corredor:

- On-policy: Seguimos μ y aprendemos V^{μ} y Q^{μ}
- Off-policy: Seguimos μ (exploramos) pero aprendemos V^{π}, Q^{π} o directamente Q^* (valores óptimos)

Esta distinción es crucial para entender algoritmos modernos de RL y el balance entre exploración y explotación.