TD(0)

Supongamos que tienes un agente que se mueve en una cuadrícula de 4x4. El objetivo del agente es llegar a la esquina inferior derecha (4,4) partiendo de la esquina superior izquierda (1,1). Las recompensas son -1 por cada movimiento. Los movimientos posibles en cada estado son abajo y derecha.

Asuma que con una política π se generaron los siguientes episodios:

- **Episodio 1**: Derecha (D), Derecha (D), Derecha (D), Abajo (A), Abajo (A), Abajo (A)
- **Episodio 2**: Derecha (D), Abajo (A), Abajo (A), Derecha (D), Derecha (D), Abajo (A)
- **Episodio 3**: Abajo (A), Abajo (A), Derecha (D), Derecha (D), Derecha (D), Abajo (A)

Utilize el enfoque de SARSA para estimar $V_{\pi}(s)$ y $Q_{\pi}(s,a)$

> Notas:
>
> - Tasa de aprendizaje (α): 0.1
> - Factor de descuento (γ): 0.9

Ejercicio 2: Q-Learning

Utilizando los mismos episodios anteriores genere una política π_{2} utilizando Q-Learning.

- ¿Se puede decir que π_2 es mejor que π?
- Si se generaran miles de episodios con una política random, ¿qué ocurre con π_{2} utilizando Q-Learning?

Ejercicio 3: Evitando el Abismo

Una vez jugando los episodios anteriores, de repente se añade un "abismo" en la celda (2, 3) (utilizando la notación (fila, columna) con base en 1) y la recompensa de caer en el abismo es -10.

Se genera el siguiente episodio 💀 : Derecha (D), Abajo (A), Derecha (D)

- Actualice la estimación π utilizando SARSA.
- Actualice la política π_2 usando Q-Learning.

Reflexión sobre el Abismo

- **Comparación de Estrategias**: Después de completar ambas partes, reflexiona sobre cómo el abismo afecta las decisiones de política en SARSA y Q-Learning. Considera cómo cada algoritmo se adapta a las recompensas negativas severas y qué esto puede decir sobre su uso en entornos con penalizaciones significativas.

- **Riesgo vs. Seguridad**: Piensa en cómo la presencia del abismo puede cambiar la estrategia de exploración del agente (teniendo $\epsilon>0$). ¿El agente se vuelve más cauteloso con SARSA en comparación con Q-Learning, o viceversa?

Para más info:

- Example 6.6: Cliff Walking (cap 6.5, Reinforcement Learning. An Introduction", R.S. Sutton & A.G. Barto (2018))
- Temporal Difference Learning (including Q-Learning) | Reinforcement Learning Part 4