

IIC3675: Tarea 3

Bruno Cerda Mardini

a)

Q-learning es *off-policy* ya que en su función de actualización, el valor futuro de la state-action value function ($Q(S', A')$) se elige de manera *greedy*. En específico, se elige la acción que maximiza dicho valor, como se ve en la fórmula:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t) \right]$$

SARSA es *on-policy* ya que, a diferencia de Q-learning, el valor futuro de la state-action value function ($Q(S', A')$) se determina eligiendo la siguiente acción según la política que se está optimizando actualmente:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha [R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)]$$

b)

Si ambos algoritmos empiezan a elegir acciones de manera greedy, entonces se vuelven muy similares, pero de todas formas **no van a ser equivalentes**.

Esto debido al orden en que se toman las acciones. En **SARSA**, se elige una acción (A') para el siguiente estado (S'), luego se hace la actualización, y después esta misma acción (A') se utiliza en el siguiente paso. En cambio, **Q-learning** NO elige una acción (A') y se hace la actualización basado en la acción que simplemente maximice el valor de Q ($\max_a Q(S', a)$), de esta manera Q-learning no se compromete con ninguna acción, luego en el siguiente paso Q-learning es capaz de elegir una acción pero con la tabla Q ya actualizada.

Un ejemplo claro en donde ambos algoritmos podrían tomar decisiones distintas es en caso de empate de valores, debido a que la política greedy no podría elegir un claro ganador, es una situación en donde ambos algoritmos podrían diverger en resultados y la toma de acciones, el ejemplo es el siguiente:

Supongamos que un agente se encuentra en el estado S , toma la acción A y debido a esto ahora se encuentra en el estado S' . Supongamos también que en la tabla Q, el valor para el estado S' tiene un empate:

- $Q(S', A_1) = -1$

- $Q(S', A_2) = -1$

Debido a que hay un empate, la política greedy no puede elegir una acción que sea claramente mejor, por lo que lo tiene que hacer de manera uniforme. **SARSA** va a actualizar $Q(S, A)$, para esto va a elegir la siguiente acción A' , pero como hay un empate, supongamos que va a terminar eligiendo A_1 , por lo que en el siguiente paso va a estar obligado a ejecutar A_1 . **Q-Learning** va a actualizar $Q(S, A)$ usando la acción que maximice el valor de retorno, pero como hay un empate, supongamos que va a tomar la acción A_2 como la acción que maximice el valor de $Q(S', a)$ (**Pero no la guarda para el siguiente paso**). La actualización para $Q(S, A)$ de ambos algoritmos va a ser la misma, pero **eligieron acciones distintas**, por lo que no son equivalentes.

c)

hola