UNIVERSIDAD DEL VALLE DE GUATEMALA
CC3104 – Aprendizaje por Refuerzo
Ing, Javier Josué Fong Guzmán

Laboratorio 6
Temporal Difference Learning

José Pablo Orellana 21970

Diego Alberto Leiva 21572

Guatemala, 17 de agosto de 2025

Descripción general de implementación

Trabajamos en Blackjack-v1 con reglas sab y el estado por defecto: suma del jugador, carta visible del crupier y si el As es usable. Las acciones son quedarse (0) y pedir (1). Implementamos dos agentes de diferencias temporales:

- SARSA(0), on-policy: actualiza Q con la acción que realmente ejecuta la política.
- Q-Learning, off-policy: actualiza Q usando el mejor valor futuro estimado.

Usamos una política epsilon-greedy con epsilon que decae de 1.0 a 0.05, tasa de aprendizaje 0.02 y descuento 1.0. Entrenamos 300000 episodios y, cada cierto intervalo, evaluamos de forma greedy. Para comparar con el lab anterior, cargamos la tabla Q de Monte Carlo y la evaluamos con el mismo protocolo.

Evolución de la política a lo largo de los episodios

Vimos que al inicio el retorno sube rápido y luego se estabiliza. A medida que baja epsilon, exploramos menos y consolidamos decisiones más rentables. Q-Learning se mantuvo un poco por encima de SARSA durante casi todo el entrenamiento. Frente a Monte Carlo, los dos agentes TD tienden a jugar manos un poco más largas, lo que reduce derrotas y sube empates.

Resultados de la política óptima

Evaluamos con 100000 episodios, greedy:

- SARSA(0): retorno promedio −0.05323, victorias 0.42744, empates 0.09189, derrotas 0.48067, longitud 1.57102.
- Q-Learning: retorno promedio −0.05136, victorias 0.42872, empates 0.09120, derrotas 0.48008, longitud 1.58456.
- Monte Carlo: retorno promedio −0.07444, victorias 0.42598, empates 0.07360, derrotas 0.50042, longitud 1.34345.

Los dos métodos TD mejoran a Monte Carlo en retorno y en menos derrotas. Q-Learning queda primero por poco sobre SARSA. Recordamos que en Blackjack los retornos negativos son normales por la ventaja de la casa.

Conclusiones

Con el mismo número de episodios y la misma evaluación, Q-Learning logra el mejor retorno y la menor tasa de derrotas, SARSA(0) queda muy cerca, y Monte Carlo se queda atrás. Para este problema, nos conviene usar aprendizaje por diferencias temporales, de preferencia Q-Learning. Si tuviéramos más tiempo, haríamos un pequeño barrido de alfa y epsilon para verificar que estas diferencias se mantienen.