UNIVERSIDAD DEL VALLE DE GUATEMALA

CC3104 – Aprendizaje por Refuerzo

Ing, Javier Josué Fong Guzmán

Laboratorio 5 - Monte Carlo

Diego Alberto Leiva 21572

José Pablo Orellana 21970

Guatemala, 12 de agosto de 2025

Descripción general de implementación

Se implemento un agente de control por Monte Carlo sobre el entorno BlackJack V1. La solución sigue un ciclo clásico de generacion de episodios, estimacion de Q(s,a) por retornos promedio y mejora de política con esquema ε-greedy:

- Política inicial y exploración. Se parte de una política ε-greedy sobre Q si el estado no existe aún en Q, se selecciona una acción aleatoria si existe, se elige al azar entre las acciones empatadas con el máximo Q(s,·), con exploración con probabilidad ε. La ruptura de empates aleatoria evita sesgos tempranos.
- Generación de episodios. Para cada episodio se registra la secuencia *(st,at,rt)*(st ,at ,rt) hasta terminar la mano.
- Actualización First-Visit. Se recorre el episodio en reversa y se acumula el retorno *G*G. Para cada par *(s,a)*(s,a) se actualiza solo en su primera visita dentro del episodio, promediando retornos:

$$\mathbf{Q(s,a) \leftarrow (\sum G(s,a)) / (\#visitas(s,a))}$$

- Mejora de política. Al finalizar el entrenamiento, la política determinista final asigna a cada estado alguna de las acciones con máximo Q(s,·) (si hay empates, se elige aleatoriamente entre los mejores).
- Semillas y reproducibilidad. Se fijan semillas para numpy y random y se controla el seed en env.reset() por episodio.
- Evaluación. Se incluyen dos evaluadores:
    - Desempeño promedio (recompensa media y error estándar) siguiendo la política entrenada.
    - Tsas de victoria/empate/derrota sobre 100 000 episodios; además, un baseline aleatorio para comparar.

Hiperparámetros principales de entrenamiento:

- Episodios de entrenamiento: 300 000.
- $\Gamma = 1.0$ $\gamma = 1.0$.
- Annealing de ϵ: de $\epsilon start = 1.0$ ϵstart = 1.0 a $\epsilon end = 0.1$ ϵend = 0.1 con decaimiento lineal controlado por epsilon_decay = 0.98 (define la fracción inicial de episodios en la fase de descenso).
- Variante del entorno con sab = True (reglas del dealer estándar).

Evolución de la política a lo largo de los episodios

Durante la fase inicial (ε alto) la exploración amplia permite cubrir el espacio de estados y reducir el sesgo por inicialización. Conforme ε desciende linealmente hacia 0.1:

- Estabilización de Q(s,a). Aumenta el conteo de visitas y los promedios de retorno se vuelven menos volátiles, se observa convergencia cualitativa hacia una política tipo “stick alto/hit bajo”.

- Reglas emergentes típicas.
    - Con suma del jugador ≥ 20, la política converge a stick (acción 0).
    - Entre 12 y 16, la acción tiende a hit (acción 1), especialmente cuando la carta visible del dealer es fuerte (A, 7–10).
    - Con As utilizable, la política es más agresiva (más hits con sumas aparentes altas) por el "colchón" del As.

Estas tendencias se visualizaron con mapas de calor separados para usable ace = True/False (0 = stick, 1 = hit), donde el patrón muestra claramente el umbral de detención alto y mayor agresividad con As utilizable.

Resultados de la política óptima

Desempeño de la política entrenada (100 000 episodios):

- Recompensa media: −0.06658
- Error estándar: 0.00304

Tasas de resultado (100 000 episodios):

- Victoria / Empate / Derrota: 0.431 / 0.072 / 0.497

Comparativo con baseline aleatorio:

- Política aleatoria (100 000 episodios): media = −0.39678, error estándar ≈ 0.00283.
  En Blackjack con reglas estándar del entorno, la expectativa del jugador suele ser negativaincluso con políticas razonables, debido a la ventaja de la banca. No obstante, la política aprendida mejora sustancialmente sobre el azar (de −0.397 a ∼−0.067), reduciendo la desventaja y aumentado la tasa de victorias en ~43.1%. El empate ~7.2% es consistente con la dinámica del juego y el dealer fijo.

Conclusiones

- MC First-Visit con ε-greedy y decaimiento lineal logró una política estable que captura heurísticas óptimas conocidas (stick con 20–21; mayor hit con manos duras 12–16 y dealer fuerte; mayor agresividad con As utilizable).
- Aunque la expectativa sigue siendo negativa (propio del juego con ventaja de la casa), la mejora frente a la política aleatoria es notable (~0.33 puntos de recompensa media).
- El enfoque es simple y sin modelo, pero efectivo: no requiere dinámicas de transición ni estimaciones bootstrap, lo que facilita la implementación y la explicación.