UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

Reinforcement Learning
- Laboratorio 3 -

Semestre II - 2025

Instrucciones:

- Esta es una actividad en grupos de 3 personas máximo
- No se permitirá ni se aceptará cualquier indicio de copia. De presentarse, se procederá según el reglamento correspondiente.
- Tendrán hasta el día indicado en Canvas.

Task 1

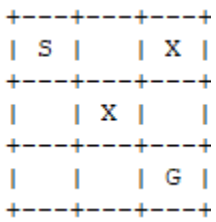
Responda a cada de las siguientes preguntas de forma clara y lo más completamente posible.

1. ¿Qué es Programación Dinámica y cómo se relaciona con RL?
2. Explique en sus propias palabras el algoritmo de Iteración de Póliza.
3. Explique en sus propias palabras el algoritmo de Iteración de Valor
4. En el laboratorio pasado, vimos que el valor de los premios obtenidos se mantienen constantes, ¿por qué?

Task 2

El objetivo principal de este ejercicio es que simule un MDP que represente un robot que navega por un laberinto de cuadrículas de 3x3 y evalúe una política determinada.
Por ello considere, a un robot navega por un laberinto de cuadrícula de 3x3. El robot puede moverse en cuatro direcciones: arriba, abajo, izquierda y derecha. El objetivo es navegar desde la posición inicial hasta la posición de meta evitando obstáculos. El robot recibe una recompensa cuando alcanza la meta y una penalización si choca con un obstáculo.

El laberinto es el siguiente



Donde:

- S = punto de inicio
- G = punto de meta
- X = son obstáculos

Instrucciones:

- Defina los componentes del MDP:
  - Estados: S = {0, 1, 2, 3, 4, 5, 6, 7, 8}, donde cada número representa una celda del laberinto.
  - Acciones: A = {arriba, abajo, izquierda, derecha}
  - Probabilidades de transición: P(s' | s, a)
  - Recompensas: R(s, a, s')
- Matriz de transición:
  - Defina las probabilidades de transición P como un diccionario donde P[s][a] asigna los siguientes estados s' a sus probabilidades.
- Función de recompensa:
  - Defina las recompensas R como un diccionario donde R[s][a][s'] da la recompensa por la transición del estado s al estado s' mediante la acción a.

- Inicializar función de valor:
  - Inicialice la función de valor V para todos los estados en 0.
- Algoritmo de iteración de valor:
  - Implemente el algoritmo de iteración de valores para actualizar la función de valor V y encontrar la política óptima.
  - Usa un factor de descuento γ de 0,9.
  - La iteración debe detenerse cuando el cambio máximo en la función de valor sea menor que un umbral (por ejemplo, 0,001).
- Extraiga la política óptima de la iteración de valor:
  - Después de converger, extraiga la política óptima de la función de valor.
- Algoritmo de iteración de políticas:
  - Implemente el algoritmo de iteración de políticas para encontrar la política óptima.
  - Inicialice una política aleatoria.
  - Evaluación de políticas: evalúe la política actual para encontrar la función de valor.
  - Mejora de la política: actualice la política en función de la función de valor.
  - La iteración debería detenerse cuando la política ya no cambie.

Asegúrese de mostrar la función de valor óptimo y la política óptima resultantes tanto del algoritmo de iteración de valor y del algoritmo de iteración de póliza.

Entregas en Canvas

1. Documento PDF con las respuestas a cada task
   a. Pueden exportar el JN como PDF si trabajan con esto.
2. Código de la implementación del Task 2
   a. Si trabaja con JN deje evidencia de la última ejecución
   b. Caso contrario, deje en comentarios el valor resultante

Evaluación

1. [1 pts] Task 1 (0.25 cada pregunta)
2. [4 pts] Task 2