UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

Reinforcement Learning - Laboratorio 6 -

Semestre II - 2024

Cayetano Elvira, Estefanía Elvira y Priscilla González

Task 1

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

1. **¿Qué es Prioritized sweeping para ambientes determinísticos?**

   En el ámbito de reinforcement learning, el prioritized sweeping es una estrategia empleada, en particular en entornos determinísticos, donde la transición entre estados es inequívoca y muy predecible. En este caso, se toman en cuenta más las actualizaciones de los estados que tienen un mayor impacto en el valor estimado de las acciones o en la política de decisión. Esta no actualiza de manera uniforme los valores de todos los estados, sino que le toma más importancia a aquellos estados que son más "importantes" en términos de su impacto en el valor de otros estados.

2. **¿Qué es Trajectory Sampling?**

   Se trata de una técnica que trata en crear y usar trayectorias o secuencias de estados, acciones y recompensas para así poder aprender de una mejor manera o mejorar una política. A diferencia de otros métodos que pueden actualizar políticas o valores basados en un solo paso de transición, este lo que hace es usar trayectorias desde un estado inicial hasta un estado terminal para poder evaluar la política.

3. **¿Qué es Upper Confidence Bounds para Árboles (UCT por sus siglas en inglés)?**

   Esta combina la estrategia de Upper Confidence Bounds (UCB) con la búsqueda en árbol para resolver problemas al momento de la toma de decisiones en entornos inciertos. También es utilizada en la teoría de bandits que realiza un balance entre la exploración y la explotación al elegir acciones que maximizan la recompensa esperada. Cuando se aplica UCB en la búsqueda en árboles, UCT realiza un árbol de decisiones donde se ven las ramas del árbol que maximizan una combinación del valor estimado y la incertidumbre asociada con ese valor.

Task 2

Preguntas

1. **Estrategias de exploración:**

   a. **¿Cómo influye la bonificación de exploración en Dyna-Q+ en la política en comparación con el equilibrio de exploración-explotación en MCTS? ¿Qué enfoque conduce a una convergencia más rápida en el entorno FrozenLake-v1?**

      Dyna-Q+: La bonificación de exploración en Dyna-Q+ incentiva al agente a explorar pares de estado-acción menos visitados, lo que puede ayudar a descubrir mejores políticas en entornos estocásticos. Esta bonificación se añade al valor Q, lo que hace que el agente valore más las acciones menos exploradas.

      MCTS (Monte Carlo Tree Search): MCTS utiliza un equilibrio de exploración-explotación basado en el criterio UCB1 (Upper Confidence Bound), que también incentiva la exploración de acciones menos visitadas pero de una manera diferente. MCTS construye un árbol de búsqueda y selecciona acciones basadas en una combinación de valor esperado y una bonificación de exploración.

      Convergencia: En general, Dyna-Q+ puede converger más rápido en entornos donde la exploración inicial es crucial para descubrir buenas políticas, ya que la bonificación de exploración puede guiar al agente de manera más efectiva. Sin embargo, MCTS puede ser más eficiente en entornos donde la planificación a corto plazo es más beneficiosa.

2. **Rendimiento del algoritmo:**

    a. **¿Qué algoritmo, MCTS o Dyna-Q+, ¿tuvo un mejor rendimiento en términos de tasa de éxito y recompensa promedio en el entorno FrozenLake-v1? Analice por qué uno podría superar al otro dada la naturaleza estocástica del entorno.**

    Tasa de éxito y recompensa promedio: En el entorno FrozenLake-v1, Dyna-Q+ puede tener un mejor rendimiento en términos de tasa de éxito y recompensa promedio debido a su capacidad para aprender y planificar simultáneamente. La bonificación de exploración ayuda a descubrir políticas óptimas más rápidamente.

    Naturaleza estocástica: La naturaleza estocástica del entorno puede hacer que MCTS sea menos eficiente, ya que depende de simulaciones precisas para la planificación. Dyna-Q+, al actualizar continuamente su modelo del entorno y usar planificación basada en experiencias simuladas, puede adaptarse mejor a la estocasticidad.

3. **Impacto de las transiciones estocásticas:**

    a. **¿Cómo afectan las transiciones probabilísticas en FrozenLake-v1 al proceso de planificación en MCTS en comparación con Dyna-Q+? ¿Qué algoritmo es más robusto a la aleatoriedad introducida por el entorno?**

    MCTS: Las transiciones estocásticas pueden dificultar la planificación en MCTS, ya que las simulaciones pueden no reflejar con precisión la verdadera dinámica del entorno. Esto puede llevar a decisiones subóptimas.

    Dyna-Q+: Dyna-Q+ es más robusto frente a la aleatoriedad, ya que actualiza su modelo del entorno basado en experiencias reales y utiliza planificación para mejorar continuamente su política. Esto le permite adaptarse mejor a las transiciones estocásticas.

4. **Sensibilidad de los parámetros:**

    a. **En la implementación de Dyna-Q+, ¿cómo afecta el cambio de la cantidad de pasos de planificación n y la bonificación de exploración a la curva de aprendizaje y al rendimiento final? ¿Se necesitarían diferentes configuraciones para una versión determinista del entorno?**

    Cantidad de pasos de planificación (n): Aumentar el número de pasos de planificación generalmente mejora el rendimiento, ya que el agente puede simular más experiencias y actualizar sus valores Q de manera más precisa. Sin embargo, hay un punto de rendimientos decrecientes donde más planificación no resulta en mejoras significativas.

    Bonificación de exploración: Una mayor bonificación de exploración puede incentivar al agente a explorar más, lo que es beneficioso en entornos estocásticos. Sin embargo, si la bonificación es demasiado alta, puede llevar a una exploración excesiva y ralentizar la convergencia.

    Entorno determinista vs. estocástico: En un entorno determinista, la necesidad de exploración es menor, por lo que una bonificación de exploración más baja y menos pasos de planificación pueden ser suficientes. En entornos estocásticos, se necesitan configuraciones que fomenten más exploración y planificación para manejar la incertidumbre.

UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

Reinforcement Learning - Laboratorio 6 -

Referencias bibliográficas

Papers with Code - Prioritized Sweeping Explained. (2022). Paperswithcode.com. https://paperswithcode.com/method/prioritized-sweeping

Trajectory Sampling. (2024). Incompleteideas.net. http://incompleteideas.net/book/first/ebook/node100.html#:~:text=In%20other%20words%2C%20one%20simulates,experience%20and%20backups%20trajectory%20sampling.

Yeshwanth, N. (2024). Upper Confidence Bounds for Tress. https://www.linkedin.com/pulse/upper-confidence-bounds-trees-yeshwanth-n-jnfwc/