Universidad del Valle de Guatemala
Reinforcement Learning

Paola De León 20360
Diego Cordova 20212
Gabriela Contreras 20213

**Laboratorio No. 5**

*Task no.1*

1. Defina y explique qué “expected sarsa”
    a. Definición
        i. Es una versión de Sarsa que usa el promedio de todos los posibles valores de acción para un estado, en lugar de solo el valor de la acción que realmente tomamos. Esto ayuda a que el aprendizaje sea más estable y menos sensible a las decisiones específicas que tomamos. Es un método de control basado en políticas, lo que significa que mejora la política mientras sigue aprendiendo.

    b. ¿Cómo se diferencia de “sarsa”?
        i. La principal diferencia entre estas es la forma en que estiman el valor Q. Sarsa estima el valor Q utilizando la regla de actualización de aprendizaje Q, seleccionando el valor Q máximo del siguiente par de estado y acción. Expected sarsa estima el valor Q tomando un promedio ponderado de los valores Q de todas las acciones posibles en el siguiente estado.

    c. ¿Para qué sirven las modificaciones que se hacen sobre “sarsa”?
        i. Las modificaciones que se hacen sobre sarsa sirven para para hacer que el aprendizaje sea más estable, logre manejar de mejor manera la aleatoriedad en las políticas, reducir la variabilidad y optimizar la convergencia.

2. Defina y explique qué es “n-step TD”
    a. Definición
        i. Este es un método para actualizar los valores de los estados o acciones. Hace uso de la información de los próximos n pasos para hacer la actualización mediante el promedio de las recompensas y estimaciones de valor durante esos n pasos.

    b. ¿Cómo se diferencia de TD(0)?
        i. Estos se diferencian principalmente en la cantidad de datos que utilizan para actualizar los valores, ya que TD(0) actualiza el valor de un estado o acción basándose únicamente en la recompensa inmediata y el valor estimado del siguiente estado mientras que n-step TD considera la recompensa acumulada y las estimaciones de valor a lo largo de los próximos n pasos.

    c. ¿Cuál es la utilidad de esta modificación?
        i. Estas modificaciones son de utilidad ya que estas permiten mejorar la precisión y estabilidad del aprendizaje puesto a que

utilizan información de pasos futuros. También estas permiten obtener una estimación más precisa del valor de un estado o acción, reducen la variabilidad en las actualizaciones y logra generar un aprendizaje más estable.

   d. ¿Qué usa como objetivo?
      i. Este utiliza el retorno acumulado de los próximos n pasos como objetivo para actualizar el valor del estado o acción proporcionado así una estimación más completa y más precisa del valor del estado o acción.

3. ¿Cuál es la diferencia entre SARSA y Q-learning?
   a. La principal diferencia entre estos es en la forma en la que actualiza el valor Q ya que SARSA actualiza dicho valor utilizando el valor Q del próximo estado y próxima acción de la política. Mientras que Q-Learning lo actualiza utilizando el valor Q del siguiente estado y la acción codiciosa posterior.