Universidad del Valle de Guatemala
Reinforcement Learning

Paola De León 20360
Diego Cordova 20212
Gabriela Contreras 20213

Laboratorio No. 4

Task no.1

1. ¿Cómo afecta la elección de la estrategia de exploración (exploring starts vs soft policy) a la precisión de la evaluación de políticas en los métodos de Monte Carlo?
    a. Considere la posibilidad de comparar el desempeño de las políticas evaluadas con y sin explorar los inicios o con diferentes niveles de exploración en políticas blandas.

	Exploring starts	**Soft policy**
Cobertura	Esta es alta ya que todos los estados y acciones tienen la oportunidad de ser explorados	Esta es media ya que depende mucho del valor epsilon que se coloque
Precisión	Debido a que se tiene una exploración más amplia la precisión otorgada es más alta.	Esta depende del balance que se realice a epsilon pero por lo general es más consistente
Convergencia	Esta es variable y más lenta debido a al a alateoridad	Esta es más rápida y cuenta con un comportamiento más estable, no obstante a largo plazo puede ser menor debido a que no se cuenta con una exploración suficiente
Impacto	Debido a que cada episodio inicia desde un estado y acción diferente, se asegura una exploración más exhaustiva generando una evaluación de políticas más precisa evitando así sesgos hacia ciertos estados o acciones	Este depende del epsilon, ya que si este es pequeño la política no explora lo suficiente por lo que se obtiene una evaluación menos precisa. Ahora bien si este es muy grande puede existir demasiada exploración por lo que se reduce la precisión a corto plazo

2. En el contexto del aprendizaje de Monte Carlo fuera de la póliza, ¿cómo afecta la razón de muestreo de importancia a la convergencia de la evaluación de políticas? Explore cómo la razón de muestreo de importancia afecta la estabilidad y la convergencia.

    La razón de muestreo afecta la estabilidad del aprendizaje y la velocidad en la que va a converger hacia una política estable. Esto sucede ya que los valores que se están aprendiendo pueden cambiar drásticamente con cada nueva muestra, dificultando que el algoritmo converja a un resultado estable.

3. ¿Cómo puede el uso de una soft policy influir en la eficacia del aprendizaje de políticas óptimas en comparación con las políticas deterministas en los métodos de Monte Carlo? Compare el desempeño y los resultados de aprendizaje de las políticas derivadas de estrategias épsilon-greedy con las derivadas de políticas deterministas.

	Políticas derivadas de estrategias épsilon-greedy	**Derivadas de políticas deterministas**
Ventajas	Esta explora diferentes acciones por lo que el agente tiene menos probabilidad de quedar atrapado facilitando así encontrar la política global Esta puede adaptarse a entornos dinámicos, es decir donde las recompensas y transiciones pueden cambiar con el tiempo. Esta puede ser más robusta y se adapta a los diferentes cambios que puedan existir en los entornos	Esta es más simple de implementar y fácil de entender, ya que siempre se sigue una estrategia fija basada en la información inicial o actual.
Desempeño	El agente es puede llegar a obtener una idea del entorno de manera rápida	En entornos complejos el agente puede quedar atrapado en políticas subóptimas, dado a que no se realiza una exploración
Exploración & Explotación	Esta es capaz de balancear entre la exploración y la explotación, logrando así mejorar la capacidad de descubrir y aprender nuevas estrategias. Debido a la componente aleatoria el agente es capaz de descubrir nuevas estrategias y estados que una política determinista podría pasar por alto	Este se enfoca en la explotación de la estrategia ya conocida, por lo que cuenta con una capacidad menor para descubrir nuevas estrategias. Las políticas deterministas pueden ser menos flexibles o adaptables a cambios en el entorno ya que no se exploran nuevas estrategias
Convergencia	A pesar de que la convergencia puede ser más lenta dada la exploración continua, la calidad de la política es mejor	Dado que en esta no se exploran acciones subóptimas el agente es capaz de converger de manera rápida generando una política, la cual

		puede ser óptima si la información inicial es adecuada.

4. ¿Cuáles son los posibles beneficios y desventajas de utilizar métodos de Monte Carlo off-policy en comparación con los on-policy en términos de eficiencia de la muestra, costo computacional. y velocidad de aprendizaje?

	Métodos off-policy	**Métodos on-policy**
Ventajas	• Se pueden reutilizar experiencias previas. No requiere que las trayectorias sean generadas por la política objetivo. • Puede evaluar y mejorar múltiples políticas a la vez.	• Estimaciones con menor varianza para que las trayectorias son generadas por política objetivo. • Mayor estabilidad y convergencia más rápida.
Desventajas	• Alta varianza en estimaciones, afectando estabilidad. • Alto costo computacional, especialmente cuando se deben procesar grandes cantidades de datos o ajustes de pesos.	• Poca eficiencia ya que las trayectoria son generadas por política objetivo • Exploraciones limitadas lo cual afecta el aprendizaje de políticas óptimas.