UNIVERSIDAD DEL VALLE DE GUATEMALA

CC3104 – Aprendizaje por Refuerzo

Ing, Javier Josué Fong Guzmán

Laboratorio 8
Policy Based Methods

José Pablo Orellana 21970

Diego Alberto Leiva 21752

Guatemala, 12 de septiembre de 2025

En este laboratorio se implementó un agente de aprendizaje por refuerzo con arquitectura Actor–Critic, utilizando el entorno CartPole-v1 de Gymnasium.

El estado del entorno está representado por un vector de 4 variables continuas: posición y velocidad del carro, ángulo de la varilla y velocidad angular.

El agente combina dos aproximadores:

- **Actor:** red neuronal encargada de representar la política $\pi(a|s)$π(a|s), la cual genera una distribución de probabilidad sobre las acciones. Durante el entrenamiento se utiliza dist.sample() para fomentar exploración, mientras que en la evaluación se aplica dist.probs.argmax() para una política greedy.
- **Crítico:** red neuronal que estima el valor de estado $V(s)$V(s), usado como referencia para evaluar la calidad de las acciones.

El entrenamiento se realizó durante 100,000 frames, reutilizando el mejor modelo del laboratorio previo como punto de partida para el crítico. La función de pérdida total combina los tres términos:

$L=Lactor+cv\cdot Lcritic-\beta\cdot H$L=Lactor +cv ·Lcritic −β·H

donde $Lactor$Lactor maximiza la probabilidad de buenas acciones en función de la ventaja $At=Gt-V(st)$At =Gt −V(st), $Lcritic$Lcritic es un error cuadrático medio sobre los retornos, y el término βHβH corresponde a la entropía, que promueve la exploración.

Para mejorar la estabilidad, la ventaja fue normalizada en cada batch. Además, se usó advantage.detach() para evitar retropropagar el gradiente desde el actor hacia el crítico.

Evolución del Desempeño

Las gráficas muestran el comportamiento del agente durante el entrenamiento:

- Recompensa por episodio: se observa un aprendizaje rápido al inicio, llegando a episodios cercanos al máximo de 500 pasos. Sin embargo, después de la mitad del entrenamiento se producen caídas bruscas y oscilaciones, evidenciando problemas de estabilidad.
- Pérdidas del Actor y del Crítico: la pérdida del actor se mantuvo estable en valores bajos, mientras que el crítico presentó picos pronunciados, especialmente entre los 60k y 90k frames. Esto indica dificultades del crítico para aproximar el valor de los estados.
- Recompensa en evaluación: refleja la misma tendencia: fases de alto rendimiento intercaladas con colapsos donde el agente no logra sostener el equilibrio.
- Entropía del actor: decreció desde ~0.7 hasta ~0.4, lo que significa que la política pasó de ser exploratoria a más determinista. Esta reducción de la exploración puede explicar por qué, tras buenos resultados iniciales, el agente dejó de adaptarse a nuevas situaciones.
- Error del crítico (MAE): mostró una disminución progresiva, pero con incrementos esporádicos que afectaron directamente la estabilidad del aprendizaje.

Conclusiones

1. Eficiencia de los aproximadores:
   El actor mantuvo un comportamiento estable en su pérdida, lo cual indica que la política fue consistente. En contraste, el crítico presentó mayor inestabilidad y contribuyó a los colapsos en el rendimiento global.
2. Sensibilidad a hiperparámetros:
   El desempeño del agente se vio altamente influenciado por la selección de tasas de aprendizaje, coeficientes de entropía y peso del crítico en la función de pérdida. Pequeños cambios en estos parámetros modifican la estabilidad y el rendimiento.
3. Desempeño general:
   Aunque el agente alcanzó periodos de desempeño óptimo (500 pasos), no logró mantenerlos consistentemente a lo largo del entrenamiento. Esto refleja la naturaleza inestable de los métodos Actor–Critic.