SID Primavera 2025
Práctica 2 (Laboratorio)

Sergio Álvarez
Javier Vázquez

Abril/Mayo 2025

1. Introducción

Durante las sesiones de laboratorio sobre aprendizaje por refuerzo, hemos visto posibles maneras de abordar el entorno FrozenLake-v1[1]: Iteración de valor, Estimación directa (*model-based*), Q-Learning y REINFORCE en un entorno con un MDP accesible y con espacios discretos de observaciones y acciones. En esta práctica exploraréis cómo entrenar un agente en un entorno de características similares.

2. Objetivo de la práctica

En esta práctica se os pide que diseñéis y ejecutéis experimentos para realizar un estudio del rendimiento en el entorno CliffWalking-v0[2], en su modo *slippery* (`is_slippery=True`), de los siguientes algoritmos:

- Iteración de valor
- Estimación directa
- Q-Learning

Se dará una puntuación extra a los grupos que también añadan un método de gradiente de política al estudio, como por ejemplo REINFORCE.

Los parámetros e hiperparámetros que podéis tener en cuenta para realizar el análisis pueden incluir (pero no tienen por qué limitarse a):

- El número de episodios a entrenar
- El factor de descuento

[1] https://gymnasium.farama.org/environments/toy_text/frozen_lake/
[2] https://gymnasium.farama.org/environments/toy_text/cliff_walking/

- La señal de recompensa[3]
- El coeficiente de exploración y su descuento (en Q-Learning)
- La tasa de aprendizaje y su descuento (en Q-Learning)

Para evaluar la calidad de los entrenamientos, se pueden usar los siguientes elementos, aunque no tiene por qué ser exactamente esta lista (podéis añadir o reemplazar elementos si lo justificáis):

- Tiempo de entrenamiento por episodio
- Número de episodios
- Tiempo de entrenamiento total
- Recompensa obtenida
- Optimalidad de la política resultante

El objetivo de la práctica es que redactéis un documento que muestre las ventajas y los inconvenientes de cada método en este entorno particular y que, respecto a cada método, permita entender el efecto de los diferentes parámetros en el rendimiento del agente entrenado.

Podéis utilizar como base el código de los notebooks de la sesión de laboratorio. También podéis modificar, optimizar o reimplementar los algoritmos siempre y cuando documentéis el proceso de diseño e implementación en la entrega. Prestad especial atención a la eficiencia del algoritmo porque los espacios de observaciones y acciones son más grandes que en el entorno FrozenLake-v1.

3. Plazos y evaluación

La entrega tiene que incluir el código de los algoritmos, así como también el código para lanzar los experimentos y no ha de incluir notebooks[4]. Las únicas dependencias que podéis utilizar son las mismas que ya se utilizan en los notebooks.

La fecha de entrega será el día **11 de mayo (11/05/2025)**, y consistirá en:

- El código de experimentación y los algoritmos.
- Una documentación en `.pdf` incluyendo:
    - Las decisiones de diseño de los experimentos, incluyendo hipótesis y una justificación de parámetros a probar y con qué rangos.

[3] Podéis probar alternativas a la señal de recompensa por defecto pero, si lo hacéis, tenéis que comparar entre vuestras propuestas y la versión por defecto.

[4] Podéis utilizar los notebooks para desarrollar algoritmos o probar experimentos, pero no para la versión final del código que entreguéis.

    - Un resumen de los resultados empíricos.
    - Análisis comparativo de los algoritmos y sus parametrizaciones, contextualizando cada resultado con respecto a las propiedades de los algoritmos tal como se vieron en las sesiones de teoría.

- Un fichero README con las instrucciones para parametrizar y ejecutar los experimentos.

Criterio	Peso	Expectativa
Ejecución	40 %	El código no depende de notebooks y funciona correctamente a partir de las instrucciones del README, permitiendo fácilmente parametrizar nuevos experimentos.
Implementación	20 %	Los algoritmos que se utilizan son consistentes con su funcionamiento teórico, por ejemplo en lo que concierne a los parámetros γ, ϵ, α y los respectivos decay cuando corresponda.
Diseño de experimentos	20 %	Los experimentos están planteados a partir de hipótesis y están justificados en base al comportamiento teórico de los diferentes algoritmos y a las características del entorno.
Análisis de resultados	20 %	El resultado de los experimentos se analiza de manera correctamente contextualizada con respecto a las hipótesis de partida y a los algoritmos en cuestión y a sus propiedades y se concluye con posibles recomendaciones de cómo entrenar agentes para entornos similares.

Figura 1: Rúbrica de evaluación de la práctica

La nota base de la práctica se evaluará según la rúbrica descrita en la Figura 1. Esta nota tendrá un valor máximo de 10. Aparte de esta nota base, existe la posibilidad de sumar puntos extra que permitirían tener una nota mayor de 10:

- 2 puntos extra a los grupos que añadan, aparte de los tres algoritmos propuestos, uno de gradiente de política (e.g. REINFORCE).

La nota de esta práctica contará un tercio de la nota global de laboratorio. Si la nota es superior al 10, para hacer media no se reducirá a 10 sino que se mantendrá superior.