### **Comparación de algoritmos Deep Q-Learning en el entorno LunarLander-v2**

#### Características del environment

El entorno a resolver es LunarLander-v2, un entorno clásico de control de refuerzo donde un módulo lunar debe aterrizar suavemente en una zona designada. Las características del environment son las siguientes:

* **Tipos de estados**:
  + El espacio de estados es continuo y multidimensional, compuesto por:
    - Posición horizontal (x) y vertical (y) del módulo.
    - Velocidad horizontal (vx) y vertical (vy).
    - Ángulo de orientación (θ).
    - Velocidad angular (ω).
    - Dos booleanos indican si las patas izquierda y derecha han tocado el suelo.
* **Tipos de acciones**:
  + El espacio de acciones es discreto, con cuatro posibles acciones:
    - 0: No hacer nada.
    - 1: Disparar el propulsor principal.
    - 2: Disparar el propulsor izquierdo.
    - 3: Disparar el propulsor derecho.
* **Recompensas**:
  + El agente recibe una recompensa basada en la distancia al objetivo de aterrizaje y en la velocidad.
  + Se obtiene una recompensa adicional por aterrizar suavemente y una penalización por chocar.
  + El episodio termina al aterrizar correctamente o al chocar.

#### Recompensa promedio esperada

En LunarLander-v2, el objetivo es que el agente aprenda a aterrizar suavemente en la zona designada, minimizando las penalizaciones y maximizando las recompensas acumuladas. La recompensa promedio esperada depende del rendimiento del agente, pero un agente bien entrenado debería aprender a aterrizar con suavidad y precisión, logrando una recompensa promedio positiva o cercana a cero con el tiempo.

### Comparación de agentes y explicación del entrenamiento

#### Elección de los agentes

En este análisis se han comparado tres variantes del algoritmo DQN para el entorno LunarLander-v2:

1. **Deep Q-Learning (DQN)**: Un método de aprendizaje por refuerzo que utiliza redes neuronales para aproximar la función de valores Q (acción-estado). Es adecuado para entornos con espacios de estado continuos y complejos.
2. **Double DQN**: Una variante del DQN que mitiga la sobreestimación de los valores Q al utilizar una red separada para seleccionar las acciones en el estado siguiente.
3. **Dueling DQN**: Otra variante del DQN que separa la estimación del valor de estado y las ventajas de cada acción, lo que puede proporcionar una mejor representación de la función Q.

#### Pasos realizados durante el entrenamiento

1. **Inicialización**:
   * Se crean dos redes neuronales para cada variante: una para estimar la función Q (red local) y otra para mantener objetivos estables (red objetivo).
   * Se inicializa una memoria de repetición para almacenar las experiencias pasadas y se establecen los parámetros de aprendizaje: tasa de aprendizaje, factor de descuento, epsilon inicial, y su tasa de decaimiento.
2. **Política de exploración**:
   * Se utiliza una política epsilon-greedy para balancear la exploración y la explotación:
     + Al principio del entrenamiento, se usa un valor alto de epsilon (1.0) para fomentar la exploración.
     + Epsilon decae gradualmente en cada episodio para favorecer la explotación de las acciones conocidas como buenas a medida que el agente aprende.
3. **Bucle de entrenamiento**:
   * Para cada episodio:
     + El entorno se reinicia y se obtiene el estado inicial.
     + En cada paso del episodio, el agente selecciona una acción basada en la política epsilon-greedy.
     + Se ejecuta la acción en el entorno, obteniendo el nuevo estado y recompensa.
     + Se almacena la transición (estado, acción, recompensa, nuevo estado, done) en la memoria de repetición.
     + Se actualiza la red local mediante el muestreo de mini-lotes de experiencias de la memoria de repetición.
     + Se actualiza la red objetivo periódicamente para mejorar la estabilidad del entrenamiento.
     + Se ajusta epsilon para reducir la exploración a medida que avanza el entrenamiento.
4. **Grabación de videos para comparar gráficamente**:
   * Se graban videos de los episodios cada 100 episodios para analizar visualmente el comportamiento del agente.

#### Resultados de cada variante

Para evaluar el rendimiento de cada variante, se graficaron las recompensas totales por episodio a lo largo de los 2000 episodios de entrenamiento. Los resultados se presentan a continuación:

1. **DQN**:
   * **Inicio del entrenamiento**: Las recompensas por episodio son altamente variables debido a la exploración intensiva y la falta de conocimiento del entorno.
   * **Medio del entrenamiento**: La recompensa promedio comienza a estabilizarse, reflejando un incremento en la frecuencia con la que el agente aterriza suavemente.
   * **Final del entrenamiento**: La recompensa por episodio se estabiliza en valores positivos, indicando que el agente ha aprendido una política efectiva.

Episodio 100 Average Score: -161.58

Episodio 200 Average Score: -81.91

Episodio 300 Average Score: -29.94

Episodio 400 Average Score: 23.15

Episodio 500 Average Score: 194.54

Episodio 510 Average Score: 200.38

Environment resuelto en 510 episodios! Average Score: 200.38

1. **Double DQN**:
   * **Inicio del entrenamiento**: Similar al DQN, las recompensas son variables, pero tienden a ser ligeramente mejores debido a la mitigación de la sobreestimación.
   * **Medio del entrenamiento**: La recompensa promedio mejora más rápidamente en comparación con el DQN.
   * **Final del entrenamiento**: La estabilidad en las recompensas es más pronunciada, con valores ligeramente más altos que el DQN.

Episode 100 Average Score: -175.34

Episode 200 Average Score: -122.47

Episode 300 Average Score: -71.18

Episode 400 Average Score: 20.58

Episode 500 Average Score: 88.49

Episode 598 Average Score: 200.04

Environment solved in 598 episodes! Average Score: 200.04

1. **Dueling DQN**:
   * **Inicio del entrenamiento**: Las recompensas son variables pero muestran una mejor tendencia desde el principio gracias a la separación del valor de estado y las ventajas de las acciones.
   * **Medio del entrenamiento**: La recompensa promedio mejora de manera más consistente y rápida.
   * **Final del entrenamiento**: Las recompensas son las más altas y estables entre las tres variantes, indicando una política muy efectiva.

Episode 100 Average Score: -181.79

Episode 200 Average Score: -129.62

Episode 300 Average Score: -90.43

Episode 400 Average Score: -4.07

Episode 500 Average Score: -8.86

Episode 600 Average Score: 6.29

Episode 700 Average Score: 57.39

Episode 800 Average Score: 192.95

Episode 815 Average Score: 200.37

Environment solved in 815 episodes! Average Score: 200.37

### Análisis comparativo

* **DQN**: Buen desempeño general, pero con variabilidad en las recompensas debido a la sobreestimación de valores Q.
* **Double DQN**: Mejor rendimiento que el DQN estándar al mitigar la sobreestimación, proporcionando recompensas más consistentes.
* **Dueling DQN**: El mejor rendimiento entre las tres variantes, con una mejora significativa en la estabilidad y la magnitud de las recompensas debido a la estructura de la red que separa el valor de estado y las ventajas de las acciones.