**Avance tarea 3:** Policy - Gradient

**Código:** EL7021-1

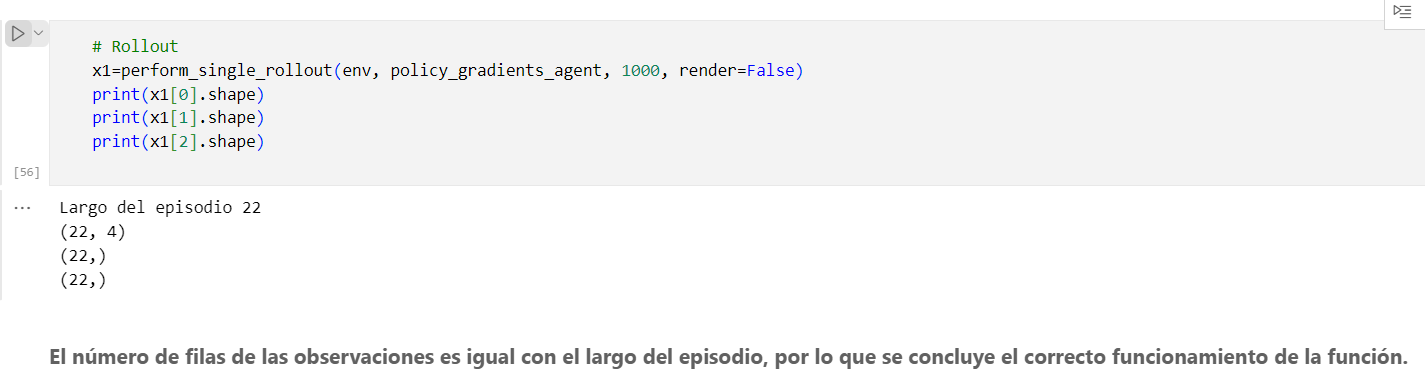
**Nombre:** José Luis Cádiz Sejas

1. **Parametrización de la política:** Código adjunto.
   1. Código adjunto.
   2. Código adjunto.

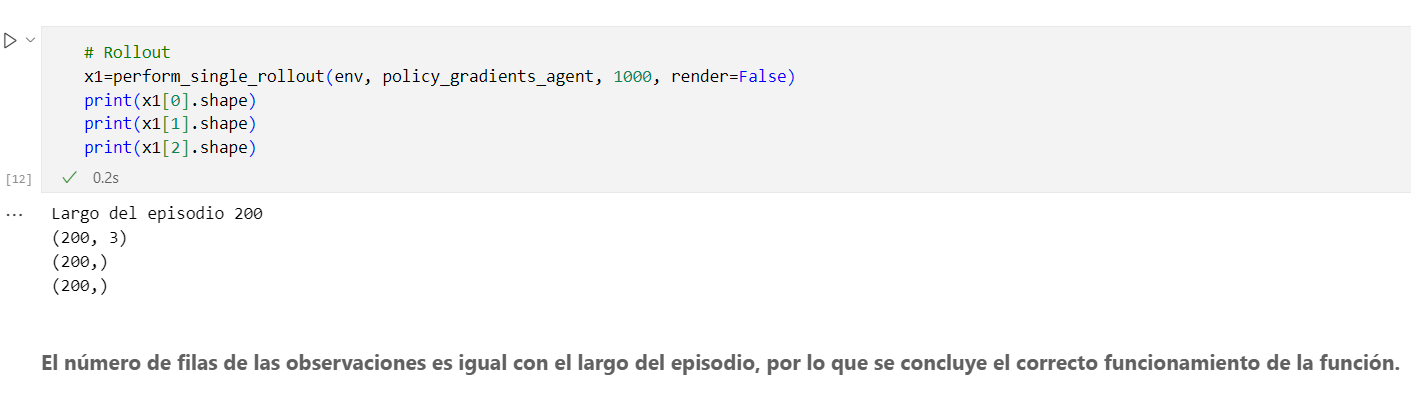
1. **Muestreo de trayectorias**.
   1. **Perform\_single\_rollout y verificación de funcionamiento**: Para verificar su funcionamiento se considera que el largo del episodio vine dado por

print(f"Largo del episodio {nb\_steps}") cuando la tarea está terminada.

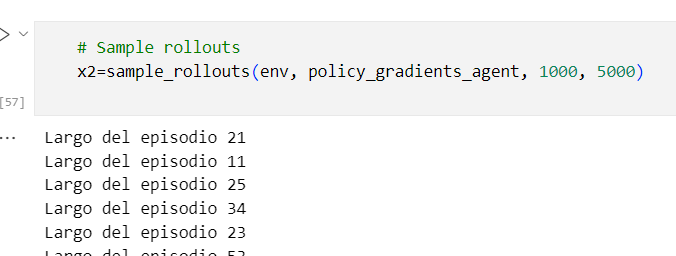
* + 1. **CartPole**:

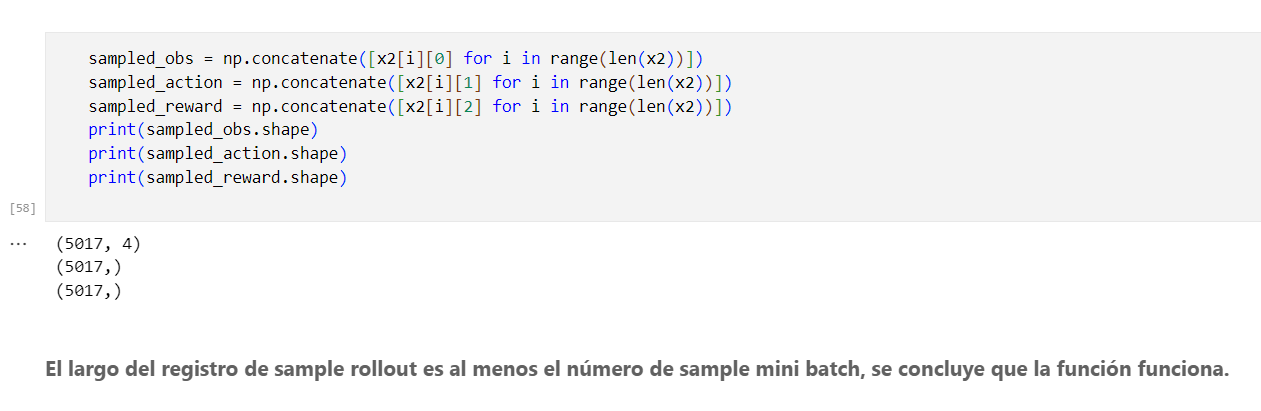


* + 1. **Pendulum**:

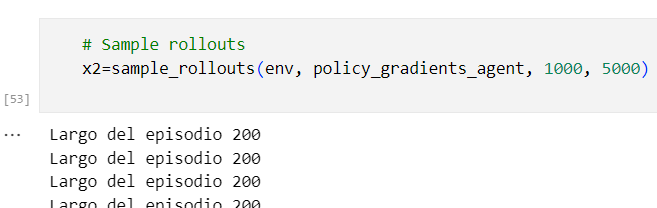


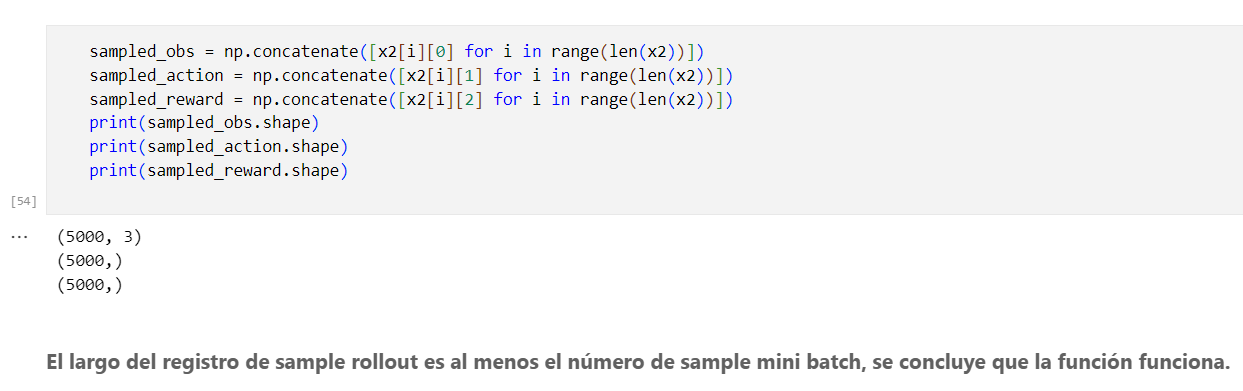
* 1. **Sample\_rollouts y verificación de funcionamiento**: Se verifica obteniendo un número de sample rollouts al menos igual o mayor que el tamaño del Batch.
     1. **CartPole**:





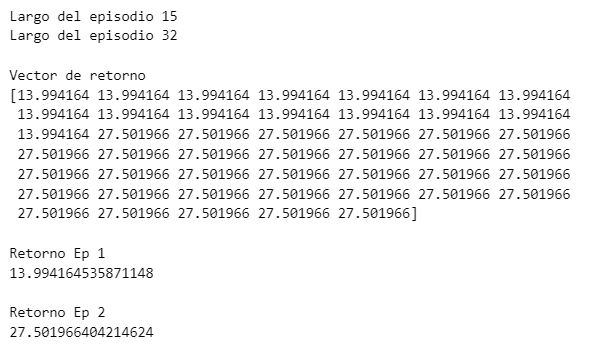
* + 1. **Pendulum**:





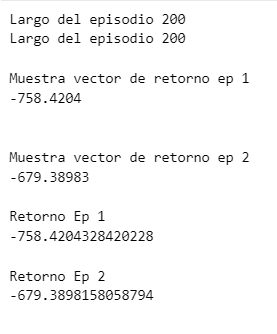
1. **Estimación de retornos**:
   1. Código adjunto.
   2. **Verificación de funcionamiento**: Se verifica mediante calculo manual del retorno descontado. Si este el mismo dentro de los steps de cada episodio, se verifica el correcto funcionamiento.
      1. **CartPole**:



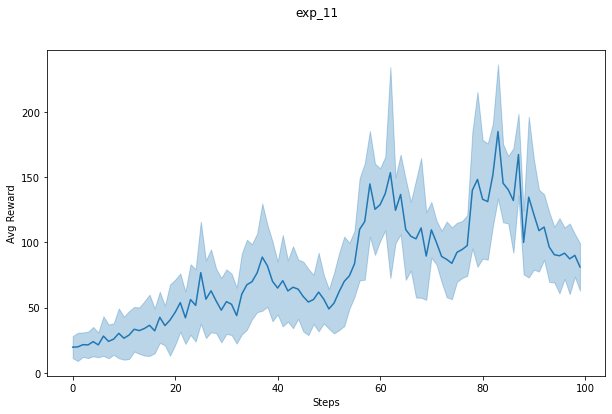


* + 1. **Pendulum**:

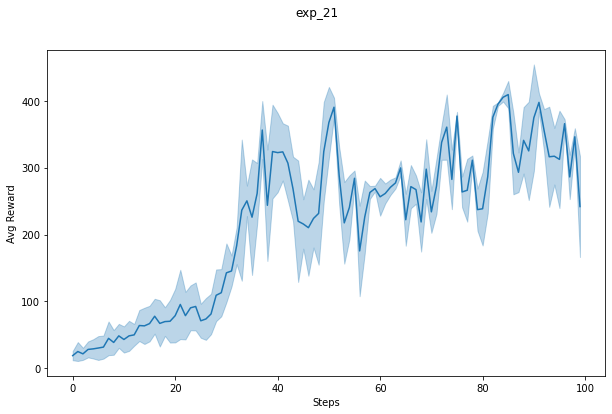




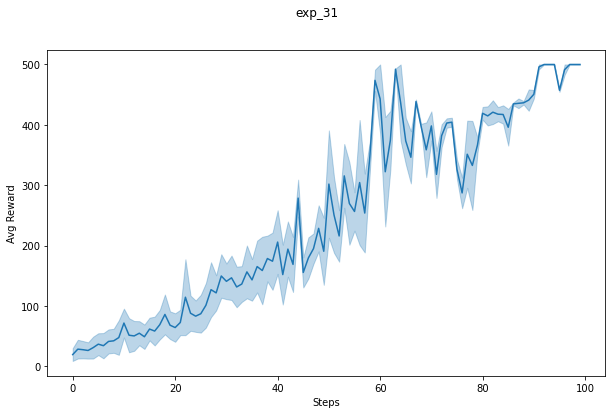
1. **Policy Gradients:**
   1. Código adjunto.
2. **Reducción de la varianza**.
   1. Código adjunto.
   2. Código adjunto.
3. **Evaluación del algoritmo:**
   1. **Entrenamiento y reporte de resultados Cartpole**.
      1. Exp 11: {"name":"exp\_11", "batch\_size":500, "use\_baseline":False,"reward\_to\_go":False}



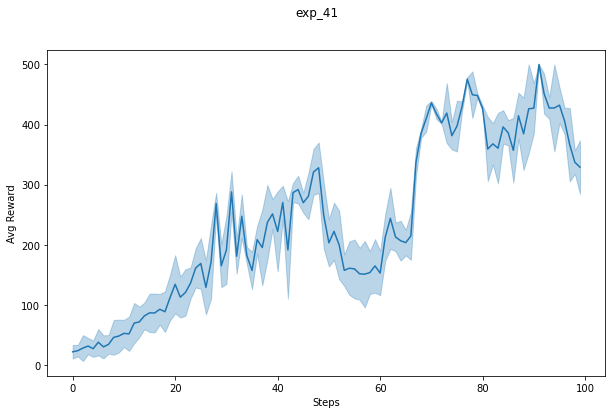
* + 1. Exp 21: {"name":"exp\_21", "batch\_size":500, "use\_baseline":False,"reward\_to\_go":True}



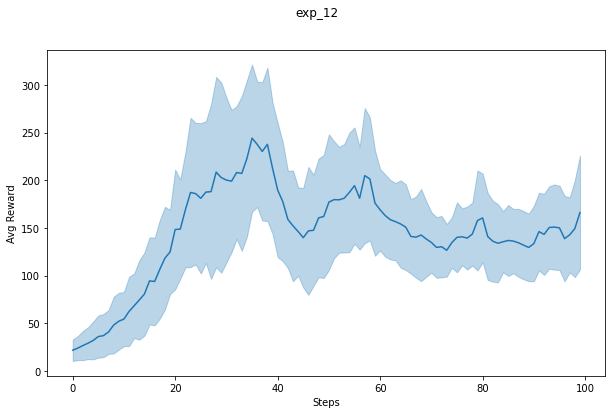
* + 1. Exp 31: {"name":"exp\_31", "batch\_size":500, "use\_baseline":True,"reward\_to\_go":False}



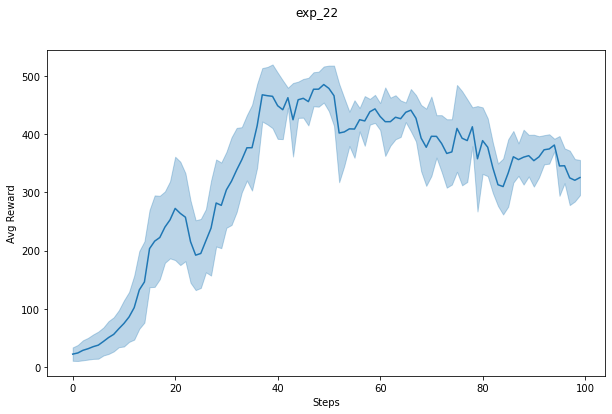
* + 1. Exp 41: {"name":"exp\_41", "batch\_size":500, "use\_baseline":True,"reward\_to\_go":True}



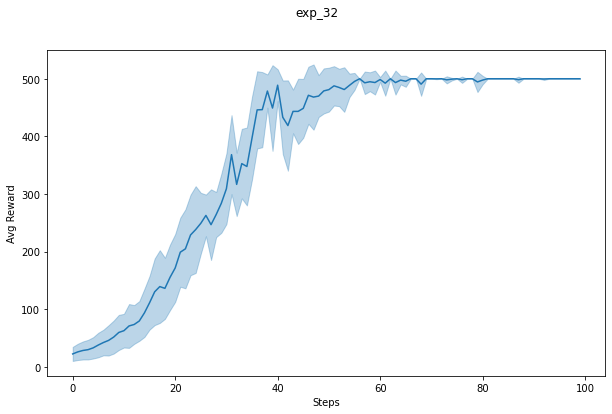
* + 1. Exp 12: {"name":"exp\_12", "batch\_size":5000, "use\_baseline":False,"reward\_to\_go":False}



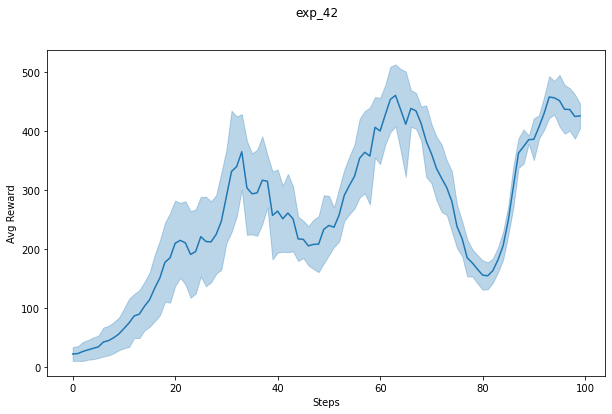
* + 1. Exp 22: {"name":"exp\_22", "batch\_size":5000, "use\_baseline":False,"reward\_to\_go":True}



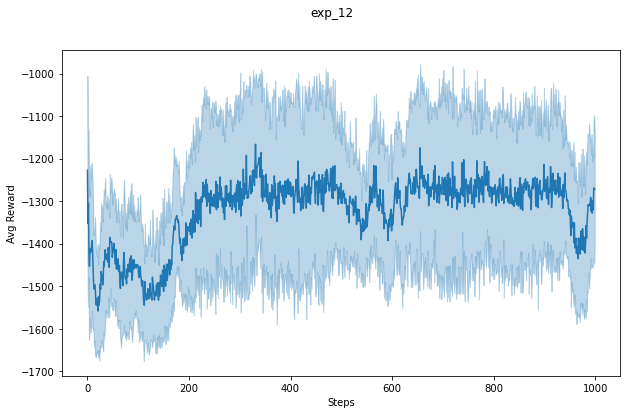
* + 1. Exp 32: {"name":"exp\_32", "batch\_size":5000, "use\_baseline":True,"reward\_to\_go":False}



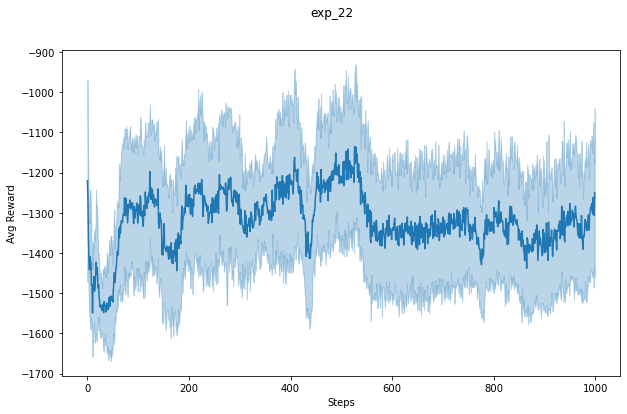
* + 1. Exp 42: {"name":"exp\_42", "batch\_size":5000, "use\_baseline":True,"reward\_to\_go":True}



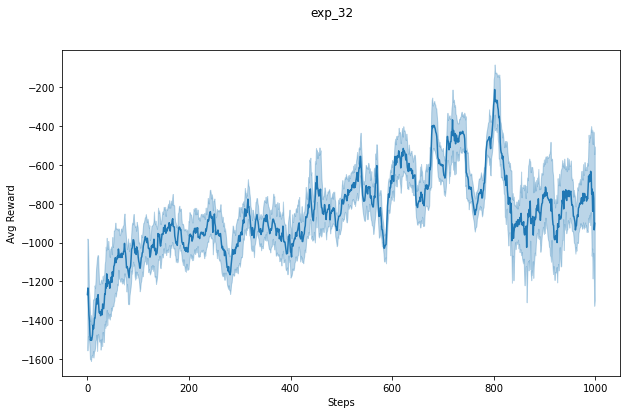
* 1. **Entrenamiento y reporte de resultados Pendulum**.
     1. Exp 12: {"name":"exp\_12", "batch\_size":5000, "use\_baseline":False,"reward\_to\_go":False}



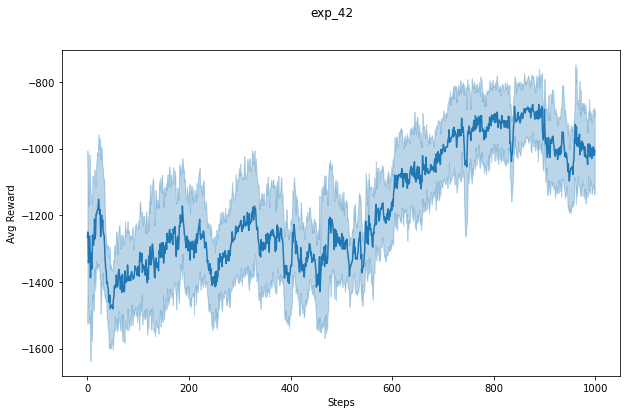
* + 1. Exp 22: {"name":"exp\_22", "batch\_size":5000, "use\_baseline":False,"reward\_to\_go":True}



* + 1. Exp 32: {"name":"exp\_32", "batch\_size":5000, "use\_baseline":True,"reward\_to\_go":False}



* + 1. Exp 42: {"name":"exp\_42", "batch\_size":5000, "use\_baseline":True,"reward\_to\_go":True}



* 1. **Análisis CartPole**: Respecto al tamaño de Batches, se observa que el desempeño de los modelos mejora con el aumento de Batches y disminuye la variabilidad del rendimiento de un entrenamiento a otro, sin embargo, aumenta la varianza dentro de un entrenamiento en particular.

Por otro lado, se observa que al incorporar uso de reward\_to\_go o baseline, aumenta el rendimiento del agente y disminuye la varianza de los entrenamientos, sin embargo, existe una anomalía en el experimento Exp\_42 al utilizar ambos enfoques.

* 1. **Análisis Pendulum**: A partir de los experimentos se observa que el uso de reward\_to\_go y baseline en conjunto aumentan considerablemente el rendimiento del agente y además disminuye en gran medida la varianza y variabilidad de los resultados.