Entrega tarea 5: Offline Reinforcement Learning

Código: EL7021-1

Nombre: José Luis Cádiz Sejas

1. **Parte I**

   1.1. **Clase PIDController**: Código adjunto.
   1.2. **Ajuste de parámetros**: Código adjunto.
   1.3. Reporte de parámetros.

Parámetros (kp,ki,kd)	Average Reward	Reward Deviation	Average Steps	Success Rate
0.1,0.3,0.7	143.7	30.54	143.7	0.1
0.5,0.3,0.1	200	0	200	1
1,1.5,3	155.17	22.34	155.17	0.07
3,2,1.5	200	0	200	1
2,3,4	174.4	25.61	174.4	0.33
4,2,3	200	0	200	1
5,5,5	200	0	200	1

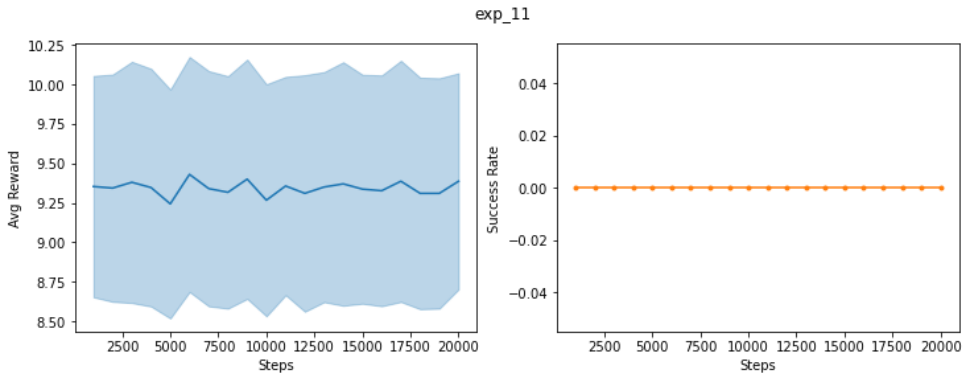
Finalmente se seleccionan los valores 5,5,5.

2. **Parte II**

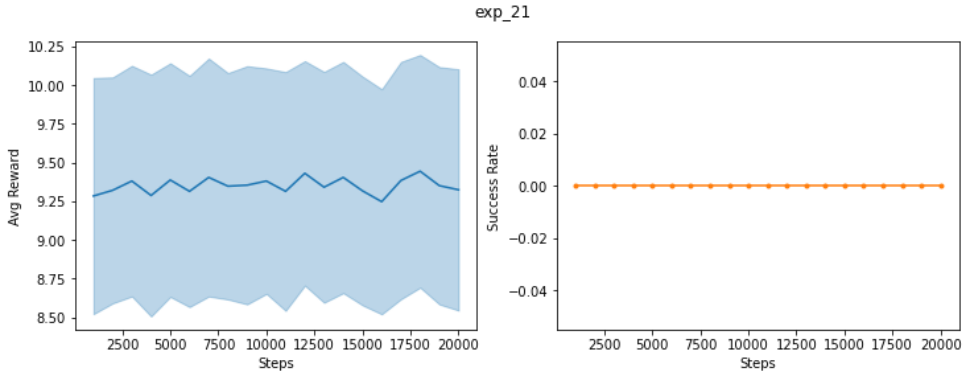
   2.1. **Función update**: Código adjunto.

   2.2. **Experimentos**:
   2.2.1. Ambiente CartPole.

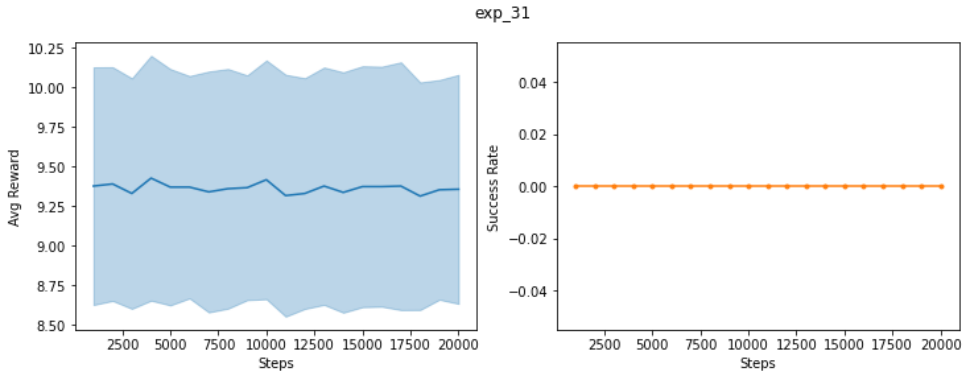
```
exp_11={"name":"exp_11","nb_rollouts":10, "alpha":0}
```



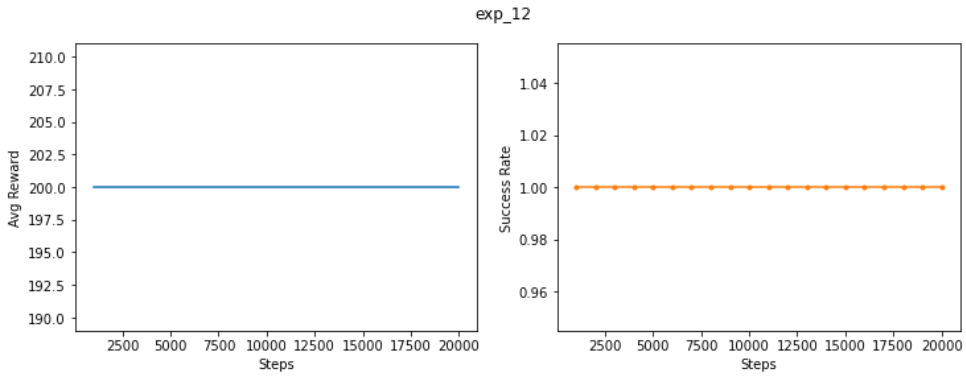
```
exp_21={"name":"exp_21","nb_rollouts":100, "alpha":0}
```



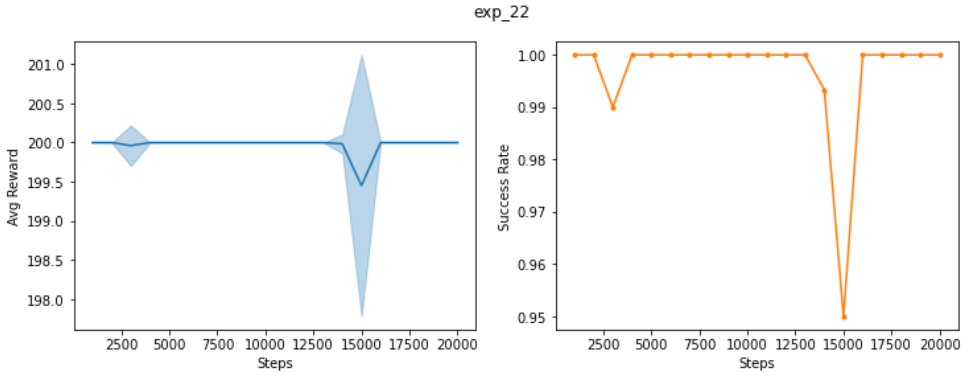

```
exp_31={"name":"exp_31","nb_rollouts":1000, "alpha":0}
```



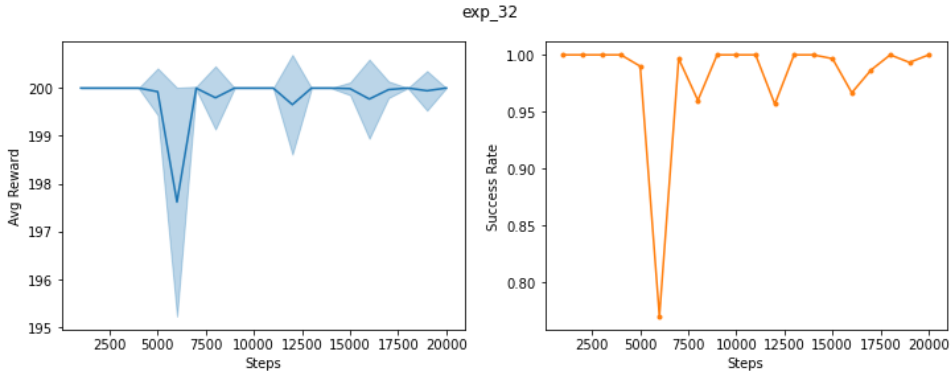

```
exp_12={"name":"exp_12", "nb_rollouts":10, "alpha":5}
```



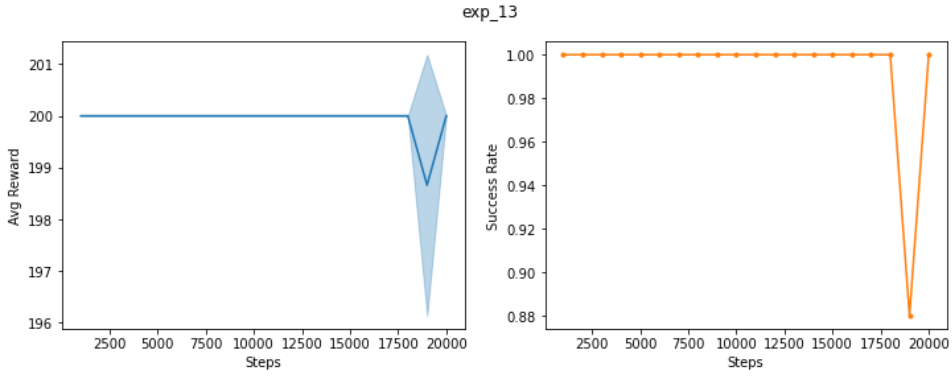
```
exp_22={"name":"exp_22","nb_rollouts":100, "alpha":5}
```



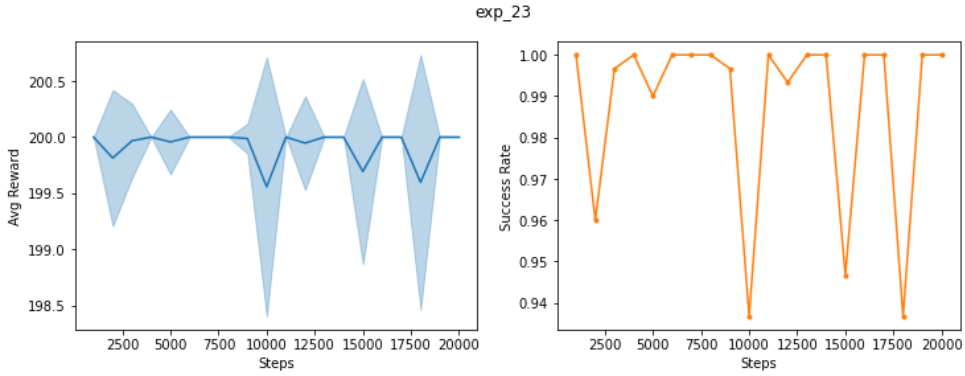

```
exp_32={"name":"exp_32", "nb_rollouts":1000, "alpha":5}
```



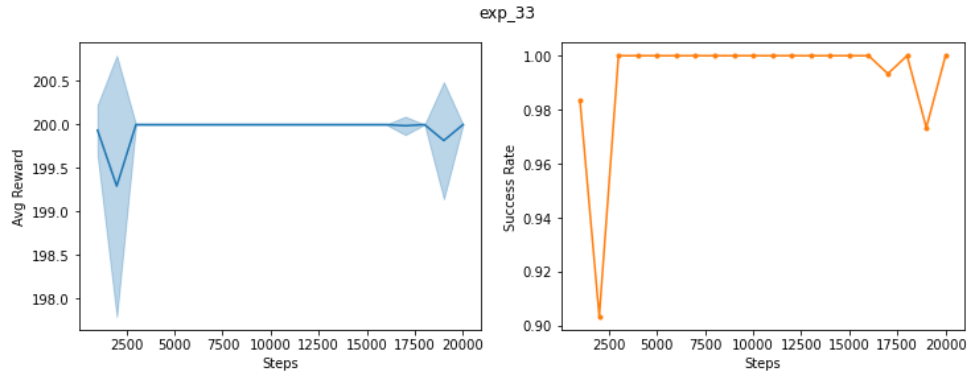

```
exp_13={"name":"exp_13","nb_rollouts":10, "alpha":10}
```



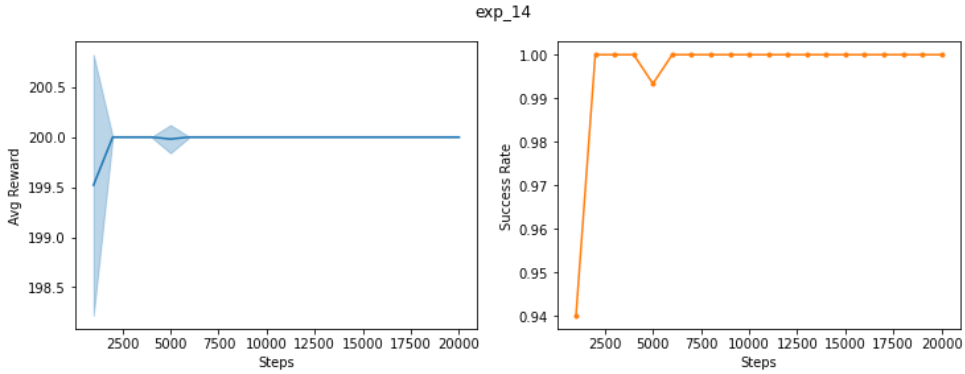
```
exp_23={"name":"exp_23","nb_rollouts":100, "alpha":10}
```



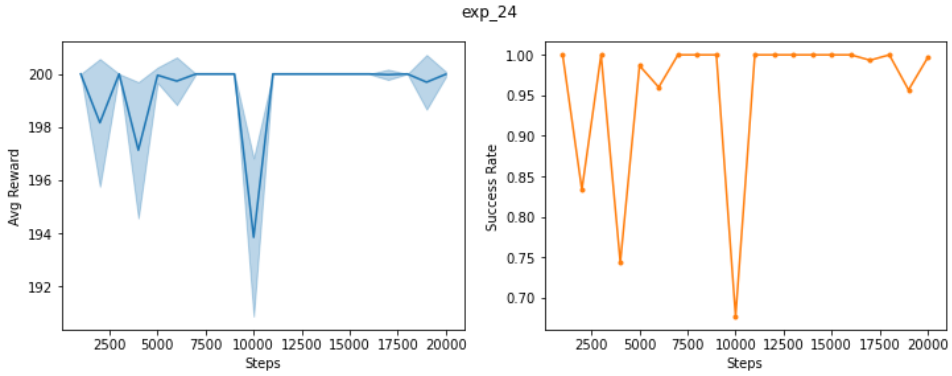

```
exp_33={"name":"exp_33","nb_rollouts":1000, "alpha":10}
```



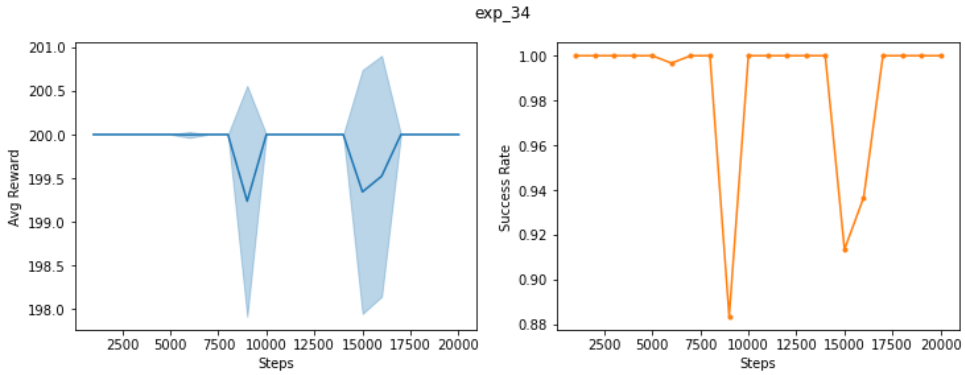

```
exp_14={"name":"exp_14", "nb_rollouts":10, "alpha":20}
```



```
exp_24={"name":"exp_24","nb_rollouts":100, "alpha":20}
```

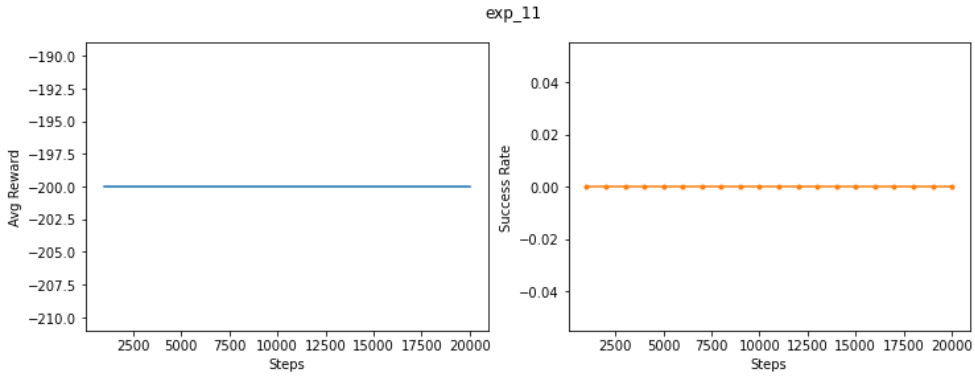



```
exp_34={"name":"exp_34", "nb_rollouts":1000, "alpha":20}
```

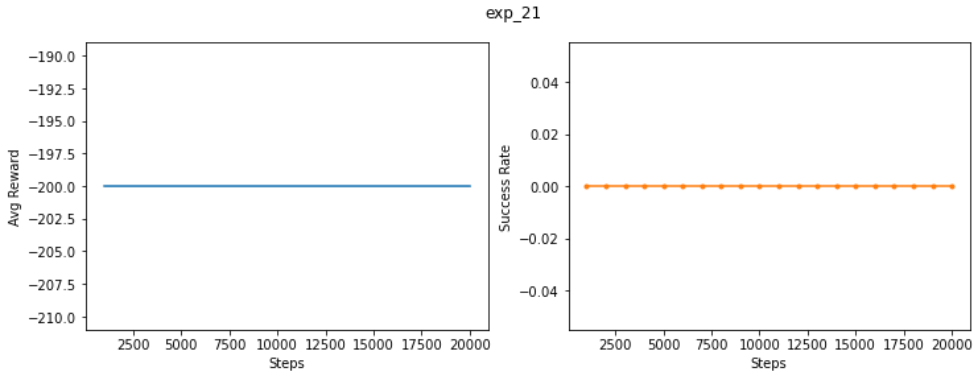


2.2.2. Ambiente MountainCar.

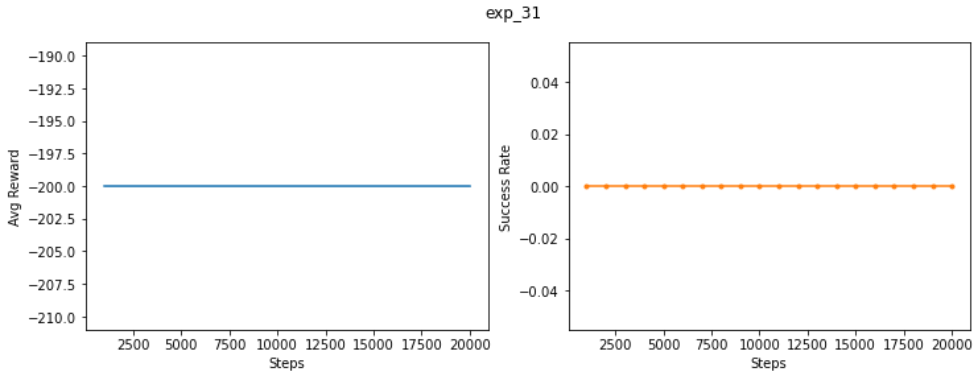
```
exp_11={"name":"exp_11","nb_rollouts":10, "alpha":0}
```



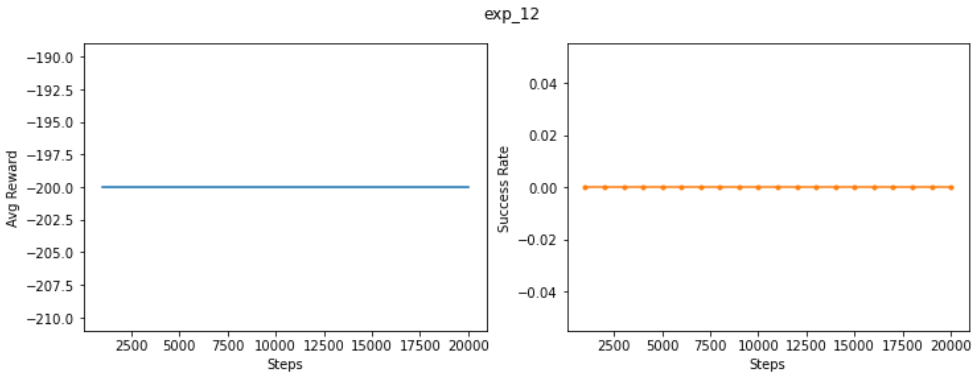

```
exp_21={"name":"exp_21","nb_rollouts":100, "alpha":0}
```



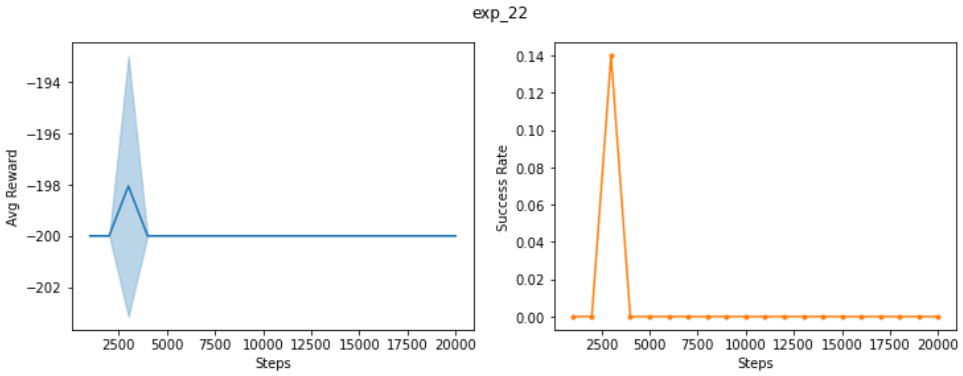

```
exp_31={"name":"exp_31","nb_rollouts":1000, "alpha":0}
```



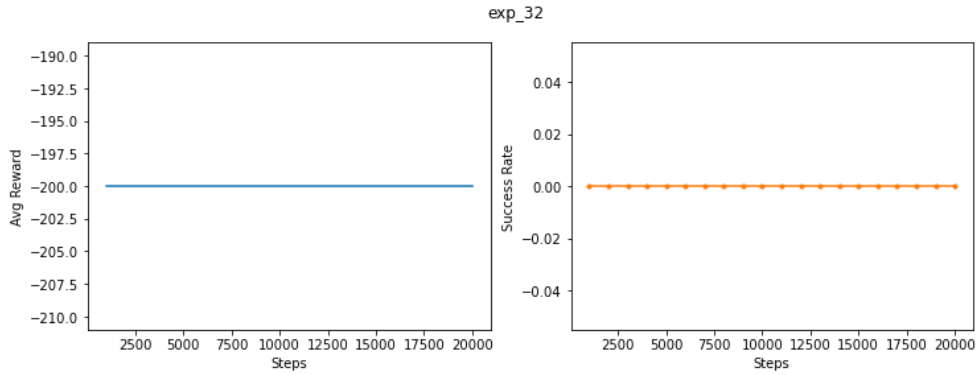

```
exp_12={"name":"exp_12", "nb_rollouts":10, "alpha":5}
```



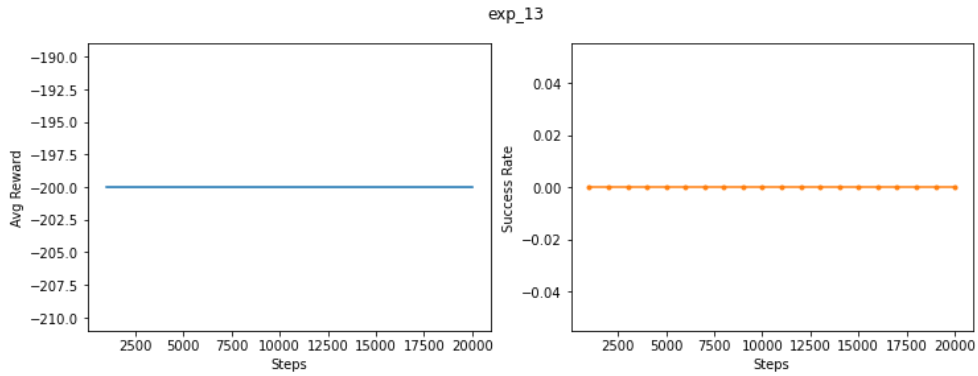
```
exp_22={"name":"exp_22","nb_rollouts":100, "alpha":5}
```



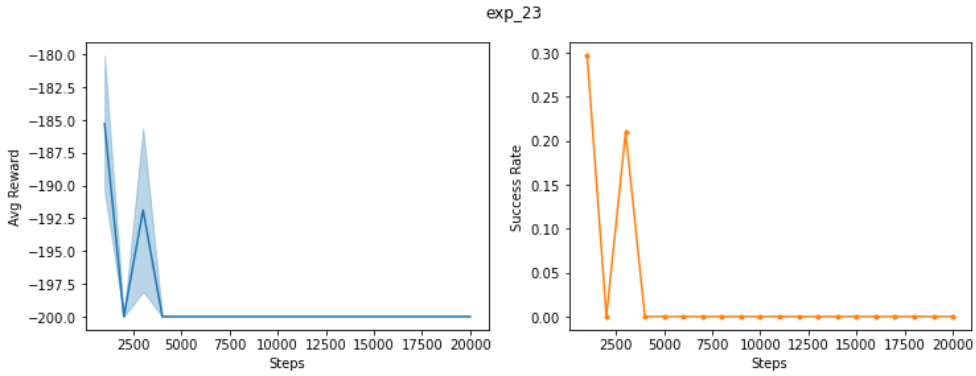

```
exp_32={"name":"exp_32", "nb_rollouts":1000, "alpha":5}
```



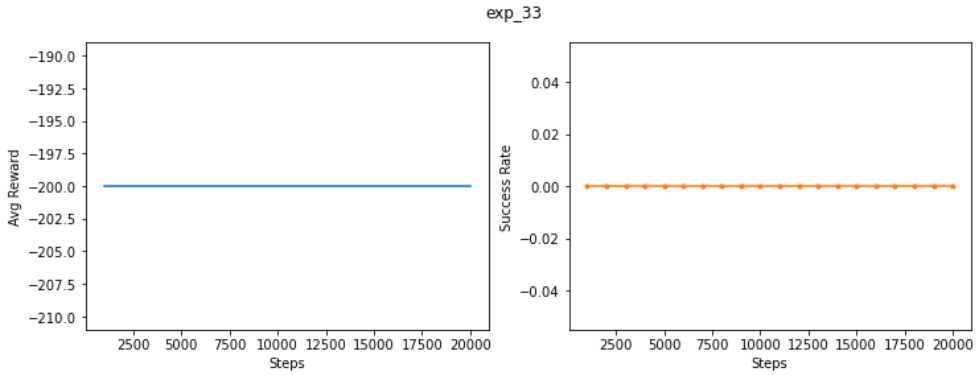

```
exp_13={"name":"exp_13","nb_rollouts":10, "alpha":10}
```



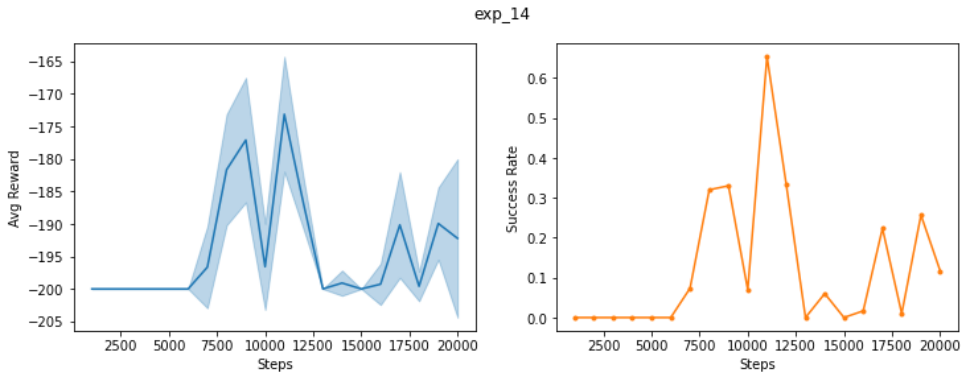
```
exp_23={"name":"exp_23","nb_rollouts":100, "alpha":10}
```



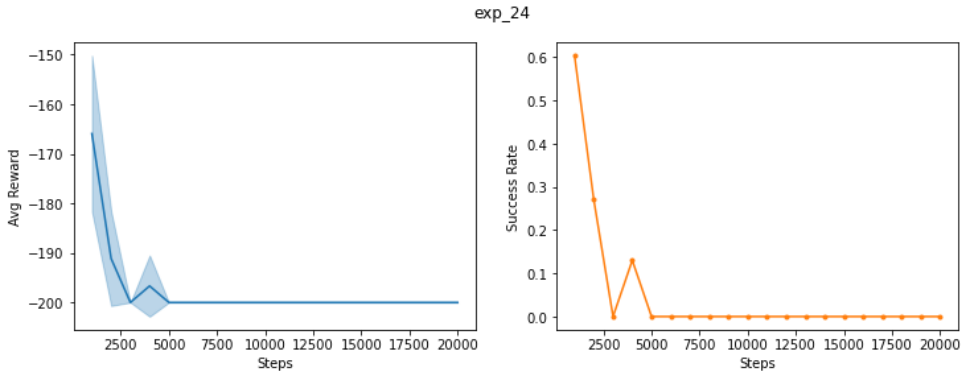

```
exp_33={"name":"exp_33","nb_rollouts":1000, "alpha":10}
```



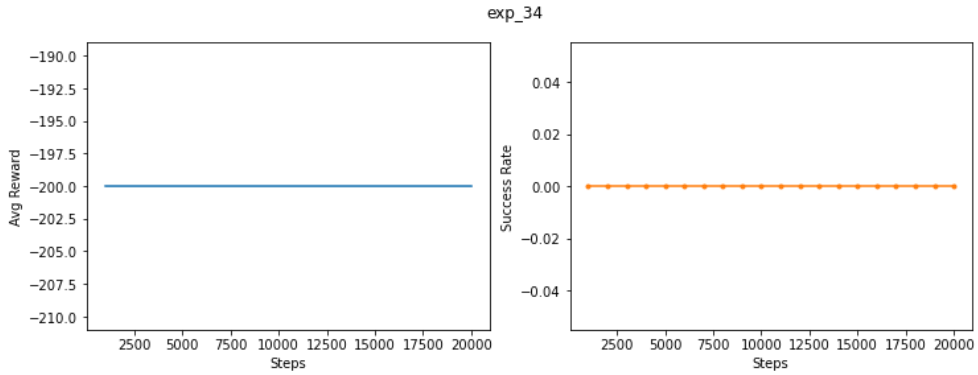

```
exp_14={"name":"exp_14", "nb_rollouts":10, "alpha":20}
```



```
exp_24={"name":"exp_24","nb_rollouts":100, "alpha":20}
```




```
exp_34={"name":"exp_34", "nb_rollouts":1000, "alpha":20}
```



2.2.3. Análisis de resultados.

A partir de los resultados de los experimentos en el ambiente CartPole y MountainCar, se observa que para un valor nulo de Alpha no se logra cumplir con la tarea de control, esto se asocia a que se están sobreestimando valores de Q para acciones que están fuera de distribución, esto ya que el factor de penalización en el Loss de la red nueronal deja de ser considerado. Por otro lado, se observa que para Cartpole se obtienen buenos rendimientos para cualquier valor distinto de 0 en Alpha, sin embargo, para MountainCar se observa que se logra un buen rendimiento solo para alpha=20.

Respecto al número de rolllouts, se espera que para un número mayor la recopilación de experiencias sea mayor y por tanto mejore el rendimiento, sin embargo, se observa que este es un parámetro que debe ser elegido acorde al ambiente.