11767 - Robots Intel·ligents Autònoms
Master Universitari en Sistemes Intel·ligents
Pràctica 1

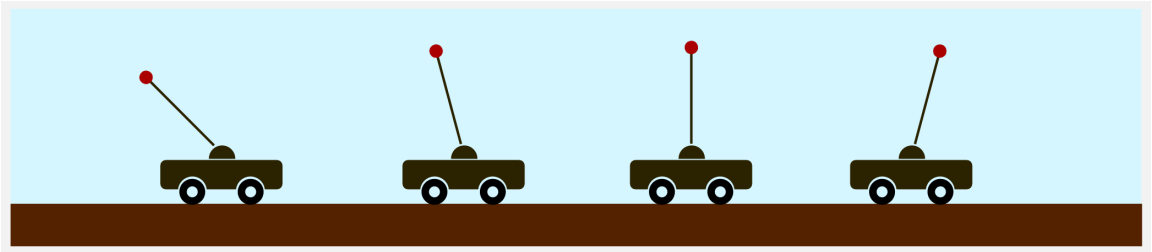
Xisco Bonnín - Alberto Ortiz

Universitat de les Illes Balears
4 de maig de 2023

Pèndul invertit

El pèndul invertit és un problema classic de control. La primera solució a aquest problema l'aportà James Roberge l'any 1960. El pèndul està format per un pal situat en posició vertical al damunt d'un carretó, al qual està fixat mitjançant una articulació o frontissa. **El problema consisteix en desplaçar el carretó per tal d'aconseguir que el pal es mantingui en posició vertical**.

L'estat està format per l'angle $\phi \in \left[\frac{-\pi}{2}, \frac{\pi}{2}\right]$ (rad), que és zero quan el pal es troba completament vertical, i la velocitat angular $\dot{\phi} \in [-\pi, \pi]$ (rad/seg).



El sistema conté diferents paràmetres que afecten a la seva dinàmica: la massa del pal *m*, la massa del carretó *M*, la longitud del pal *d* i l'increment de temps Δt, tots expressats en el Sistema Internacional d'Unitats. Donats aquests paràmetres, l'angle ϕ i la velocitat angular $\dot{\phi}$ s'actualitzen per l'instant $t+1$ segons:

$$\phi_{t+1} = \phi_t + \dot{\phi}_{t+1}\Delta t,$$

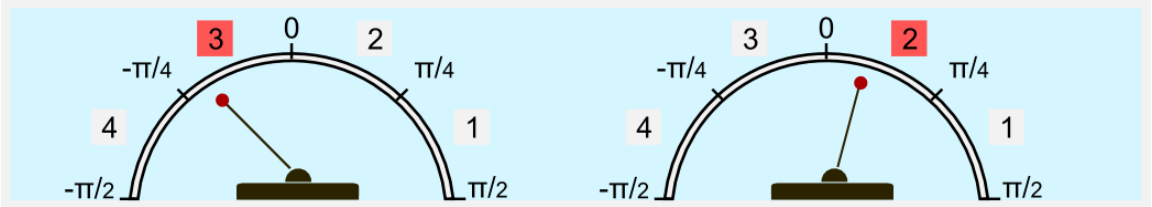
$$\dot{\phi}_{t+1} = \dot{\phi}_t + \frac{g\sin(\phi_t) - \alpha\, m\, d\, (\dot{\phi}_t)^2 \sin(2\phi_t)/2 + \alpha\cos(\phi_t)a_t}{4d/3 - \alpha\, m\, d\, \sin^2(\phi_t)}\Delta t$$

on $\alpha = 1/(M+m)$, i a_t és l'acció presa a l'instant t. Aquesta acció és una força expressada en Newtons que s'aplica al carretó, per exemple deguda a un motor que l'empeny endavant o enrere segons calgui.

1 Plantejament usant aprenentatge per reforç

Per tal de resoldre el problema del pèndul invertit mitjançant aprenentatge per reforç, cal definir cóm calculem la recompensa ("reward" en anglés), i cóm discretitzem l'estat i el conjunt d'accions. Podem considerar que la recompensa asociada a un estat ve donada pel cosinus de l'angle ϕ, de manera que com més gran sigui l'angle més baix resulti la recompensa. Dit d'una altra manera, la recompensa és 0.0 quan el pal es troba completament horitzontal, i 1.0 quan la seva posició és totalment vertical.

L'angle i la velocitat angular les podem discretitzar en regions d'iguals dimensions. Per exemple, la figura següent mostra el cas de discretitzar l'angle en quatre regions. Quan la inclinació del pal és de $\frac{-\pi}{5}$ rad. (cas de l'esquerra), podem dir que es troba en la tercera regió, mentre que si la seva inclinació és de $\frac{\pi}{6}$ rad. (cas de la dreta), podem dir que es troba en la segona regió. Quant al conjunt d'accions, per simplificar, l'hem fixat al conjunt discret de valors $[-50, 0, 50]$ (Newtons).



L'arxiu *inverted_pendulum.py* conté un script Python que defineix la classe *InvertedPendulum*. Aquesta classe conté els mètodes *reset()*, *step()* i *render()* que ens permeten, respectivament, preparar un nou episodi de simulació, desplaçar el carretó per tal d'executar un pas de simulació, i generar un arxiu gif amb el resultat de la simulació. Aquesta animació es realitza usant la llibreria *Matplotlib* i s'ha d'interpretar com la vista des d'una càmera centrada en la unió entre el carretó i el pal, i que es mou de manera solidària a aquesta unió. Per tal de crear una nova simulació, és necessari crear una nova instància de *InvertedPendulum*, definint els seus paràmetres (masses, longitud del pal i increment de temps). A continuació en teniu un exemple:

```
from inverted_pendulum import InvertedPendulum

# Defining a new environment with pre-defined parameters
my_pole = InvertedPendulum( pole_mass=2.0,
                            cart_mass=8.0,
                            pole_lenght=0.5,
                            delta_t=0.1)
```

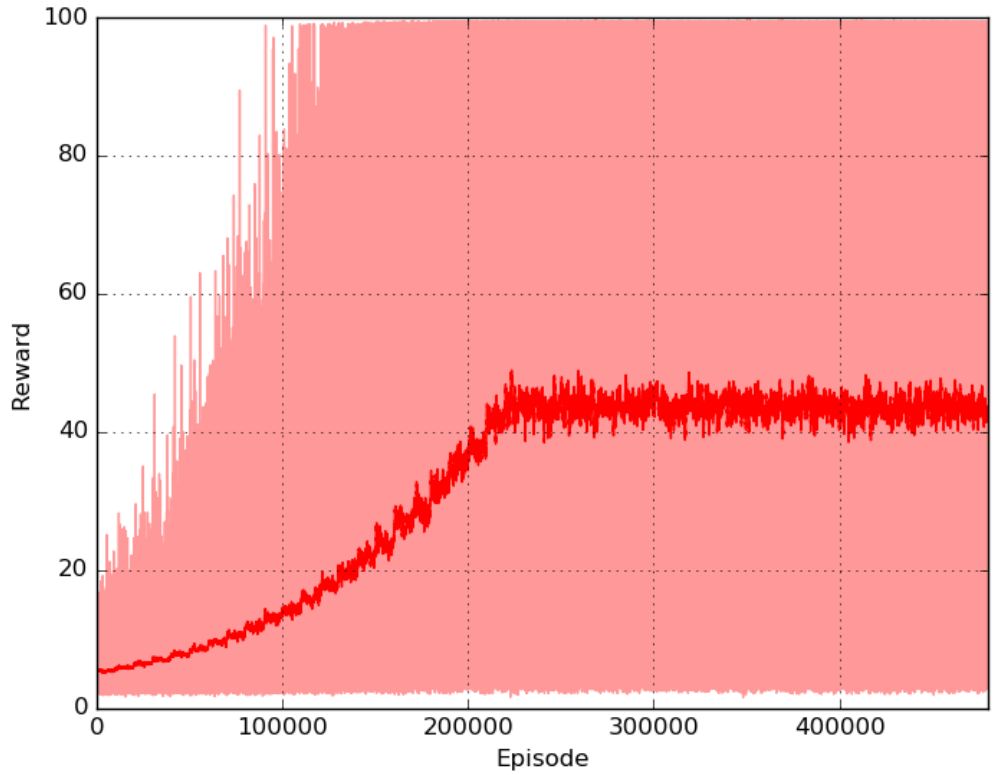
L'objectiu que ens fixem és el de poder mantenir el pèndul sense que caigui durant **10 segons** de simulació (100 passos a un increment Δt de 0.1s). Si el pal cau, l'episodi de simulació es dona per acabat.

Com a primera prova, podem avaluar el rendiment usant una política aleatòria. Això ho podeu fer executant l'arxiu *random_agent_inverted_pendulum.py*. Com podreu comprovar si executeu aquest codi, el rendiment d'una política aleatòria no és gens satisfactori, aconseguint realitzar episodis de molt curta durada.

2 Solució usant Monte Carlo per control

Com podeu imaginar, la política òptima seria la que aconsegueix compensar l'angle i les variacions de velocitat per tal que el pal estigui vertical el màxim temps possible. A l'arxiu *montecarlo_control_inverted_pendulum.py* trobareu una solució fent servir **Monte Carlo per control usant primera visita**. Aquest codi realitza un entrenament de la política durant 5×10^5 episodis, fent servir un *factor de depreciació* $\gamma = 0.999$ i discretitzant l'angle i la velocitat angular en 12 regions iguals cadascun (paràmetre *tot_bins*).

Per tal d'afavorir l'exploració, s'ha fet servir una estratègia ϵ-voraç amb un decaïment lineal de 0.99 fins 0.1. A més, s'ha fet servir la tècnica d'*inicis exploratoris*, per tal d'iniciar cada episodi amb una inclinació del pal aleatòria. Cada episodi es simula durant 10 segons (100 passos d'increment $\Delta t = 0.1$s). Fixeu-vos que la màxima recompensa que es pot obtenir és 100, que es donaria en el cas que el pal estigués completament vertical durant els 100 passos. A la imatge següent es pot veure com l'algoritme proporciona bons resultats, proporcionant una recompensa mitjana de 45 a partir de, més o manco, l'episodi 220000 (es promitja la recompensa obtinguda a cada pas de tot l'episodi = 100 passos). És a dir, el pèndol està vertical o quasi vertical més o manco la meitat de l'episodi, o tot l'episodi es manté inclinat entre $\pm 63^{\text{o}}$ ($\cos^{-1} 0.45 = 63$).



La política resultant és bastant satisfactoria i, amb una inclinació inicial favorable, és capaç de mantenir el pal a dalt durant l'episodi complet (10 segons).

Ara és un bon moment per que intenteu realitzar un entrenament executant l'arxiu *montecarlo_control_inverted_pendulum.py*. Veureu que, durant l'entrenament, es generen i actualitzen periòdicament tres arxius:

- *inverted_pendulum.gif*: una animació amb la darrera política que s'ha provat.
- *reward.png*: una gràfica amb la mitjana de les recompenses per tots els episodis simulats.
- *step.png*: una gràfica amb la mitjana de nombre de passos simulats, per a tots els episodis.

Arribat a aquest punt, comencem amb les tasques que heu de realitzar.

Tasca 1: Mireu detingudament els diferents codis Python esmentats: *inverted_pendulum.py*, *random_agent_inverted_pendulum.py* i *montecarlo_control_inverted_pendulum.py*. Intenteu entendre tot el codi i compareu els resultats obtinguts amb la política aleatòria amb els obtinguts amb el mètode de Monte Carlo per control.

Tasca 2: Modifiqueu el codi *montecarlo_control_inverted_pendulum.py* configurant els paràmetres del pal i del carretó amb els valors que teniu assignats segons el vostre nombre de grup de pràctiques (podeu trobar la taula al final del document). Aquests valors són els que haureu de fer servir en la resta de la pràctica. Realitzeu un nou entrenament usant els vostres paràmetres i guardeu els arxius resultants *inverted_pendulum.gif*, *reward.png* i *step.png* per tal de comparar amb resultats futurs.

3 Solució usant SARSA

Tasca 3: Usant els apunts de l'assignatura com a guia, i partint del codi de Monte Carlo, implementeu la versió **SARSA(0)**, creant un nou fitxer anomenat *sarsa_0_inverted_pendulum.py*. Tal com es fa en el codi de Monte Carlo, feu servir l'estratègia ϵ-voraç amb un decaïment lineal de 0.99 fins 0.1, i la tècnica d'inicis exploratoris. Un cop implementat el codi, realitzeu un entrenament per un nombre suficient d'episodis i guardeu els arxius resultants *inverted_pendulum.gif*, *reward.png* i *step.png* per comparar amb resultats futurs.

Tasca 4: Implementeu la versió **SARSA(λ)** dins d'un nou fitxer anomenat *sarsa_lambda_inverted_pendulum.py*. Podeu partir de la versió SARSA(0) a la qual li heu d'incorporar el mecanisme de traces d'elegibilitat. Concretament, heu d'implementar les traces de reemplaçament amb neteja de traces d'altres accions. Per altra banda, manteniu l'estratègia ϵ-voraç amb un decaïment lineal i l'ús d'inicis exploratoris per tal d'afavorir l'exploració. Un cop implementat, realitzeu un entrenament per un nombre suficient d'episodis i guardeu els arxius resultants.

Tasca 5: Afegiu al codi *sarsa_lambda_inverted_pendulum.py* la possibilitat de provar altres estratègies ϵ-voraces. Concretament, volem provar les següents estratègies:

- ús de ϵ amb decaïment lineal de 0.99 fins 0.1 (el que ja tenim implementat),
- ús de ϵ fixe, configurant ϵ al valor que creieu adequat (realitzeu les proves necessaries) i

- ús de ϵ amb decaïment segons la fórmula

$$\epsilon(ep) = max\left\{\mu_\epsilon,\, \epsilon \cdot d^{ep}\right\},$$

on ep és el nombre d'episodi i ϵ, d i μ_ϵ s'han de configurar al valor que creieu adequat.

Teniu en compte que hem de poder escollir una de les tres estratègies esmentades mitjançant un paràmetre en la funció *main()*. Un cop implementades i configurades les diferents estratègies, realitzeu entrenaments per a cada una d'elles i guardeu els arxius resultants.

4 Anàlisi comparatiu i documentació

Tasca 6: Escriviu un informe de la pràctica repassant totes i cada una de les tasques que heu desenvolupat, punt per punt. Heu de descriure quines són les modificacions que heu introduït a cada punt, i quines són les millores que esperau aconseguir amb aquestes modificacions. En especial, ens interessa que compareu els diferents mètodes o estratègies, demostrant que enteneu el que heu fet. Aquesta comparació l'heu de fer a nivell de mètodes/idees i a nivell dels resultats que heu obtingut amb la implementació d'aquestes idees. Incloeu les gràfiques de recompenses i de passos que heu obtingut per a les diferents versions/configuracions.

Lliurament de la pràctica

La pràctica s'ha de lliurar via UIBDigital abans del dimecres 24 de maig a les 23:55. Trobareu un enllaç a l'espai de l'assignatura. El lliurament consistirá en un arxiu comprimit (zip) que contingui tots els codis Python que hagueu desenvolupat, així com l'arxiu pdf corresponent a l'informe.

Taula de paràmetres

# de grup	M	m	d	Δt
1	7	1.5	0.4	0.1
2	8	1.5	0.4	0.1
3	9	1.5	0.4	0.1
4	7	1.5	0.5	0.1
5	8	1.5	0.5	0.1
6	9	1.5	0.5	0.1
7	7	2.0	0.4	0.1
8	8	2.0	0.4	0.1
9	9	2.0	0.4	0.1
10	7	2.0	0.5	0.1
11	8	2.0	0.5	0.1
12	9	2.0	0.5	0.1
13	7	2.5	0.4	0.1
14	8	2.5	0.4	0.1