# Control d'actitud de vehicles espacials mitjançant aprenentatge per reforç

Aquest document és un breu resum d'un dels dos problemes que vaig estudiar en el meu Treball de Final de Màster, "Solving classical Astrodynamics problems by means of Machine Learning approaches", per al Màster en Ciència de Dades de la Universitat de Girona (UdG), sota la tutela de la Dra. Esther Barrabés del grup d'Equacions Diferencials, Modelització i Aplicacions.

El problema que tractarem aquí és el que es coneix com el *problema del control d'actitud* de vehicles espacials, que consisteix en determinar els moviments que ha de fer un satèl·lit o una nau espacial per tal de controlar la seva orientació i velocitat angular. El propòsit d'aquest document és donar una idea general del problema i quins han estat els passos seguits per resoldre'l mitjançant aprenentatge per reforç, una branca de l'aprenentatge automàtic que busca crear algorismes que aprenguin autònomament a partir de l'experiència. Al lector amb coneixements tècnics de matemàtiques, programació i/o aprenentatge automàtic, així com aquell que tingui curiositat per aquests camps, li recomano visitar el següent enllaç, que conté una còpia de la memòria completa del treball juntament amb tots els programes: `github.com/RecursiveMagus/AstroIA_MasterThesis`

## 1. INTRODUCCIÓ

Quan un vehicle espacial orbita seguint una determinada trajectòria, haurà de realitzar constantment maniobres per tal de corregir la seva *actitud* (i.e. orientació i velocitat angular). Això pot ser degut a diversos motius, que depenen tant de l'òrbita en que es situï l'aparell com de la naturalesa de la tasca que vulgui desenvolupar.

Per exemple, alguns satèl·lits com ara els de telecomunicacions o cartografia disposen d'antenes i altres sensors que han d'estar sempre enfocats cap a determinats punts de la Terra, i el fregament amb L'atmosfera pot pertorbar la seva orientació. Altres, com el telescopi Gaia o el James Webb, reben contínuament l'impacte de petites partícules que produeixen un moviment de precessió que s'ha de contrarestar.

El **problema del control d'actitud** consisteix en determinar el moviment, o la seqüència de moviments (normalment en forma de *torques* generats per motors d'impuls feble) que permetin al vehicle espacial rectificar la seva orientació i velocitat angular (vegeu Figura S1).

Un *controlador* [1] és un algorisme que permet assolir aquest objectiu, normalment dictant a

[1] A la literatura sovint també s'anomena *llei de control* o simplement *control*.

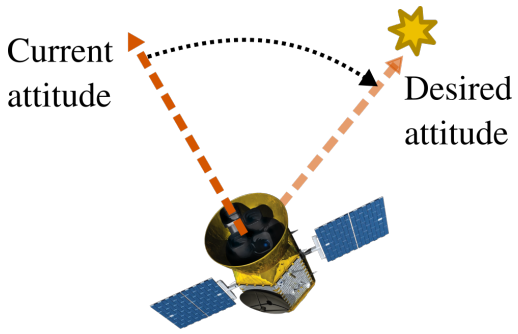


**Fig. S1.** El control d'actitud té com a objectiu determinar els moviments que permeten a un vehicle espacial rectificar la seva actitud (orientació i velocitat angular).

cada moment quin ha de ser el torque produït pels actuadors del vehicle. Existeixen moltes maneres i estratègies per dissenyar aquest tipus de controladors; en aquest treball, ens preguntem si l'aprenentatge per reforç és una eina adequada per construir-ne un.

## 2. REPRESENTACIÓ D'ACTITUD

Un sòlid rígid és un objecte tridimensional que no pot ser deformat per l'acció de forces externes. Malgrat que hom podria argumentar que un objecte d'aquest tipus no pot existir en la realitat (tot cos és, en certa mesura, deformable), nosaltres considerarem que el nostre vehicle espacial es un sòlid d'aquest tipus. Aquesta suposició facilita enormement els càlculs.

La representació matemàtica de l'orientació i velocitat angular d'un sòlid rígid s'anomena *actitud*. Existeixen diverses maneres de representar aquesta actitud, essent els quaternions una de les més habituals donat que estan ben definits i no tenen singularitats (al contrari del que ocórre amb altres representacions com els angles d'Euler).

### A. Sistemes de referència

Considerem l'espai real $\mathbb{R}^3$. Un sistema de referència $\mathcal{E} = \{O; \mathbf{e}_1, \mathbf{e}_2\mathbf{e}_3\}$ és un conjunt format per tres vectors ortonormals $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3 \in \mathbb{R}^3$ i un punt orígen $O \in \mathbb{R}^3$. Direm que un sistema de referència és *inercial* si el seu orígen no està accelerant i els seus eixos no estan rotant.

Un *sòlid rígid* és una distribució contínua de masses puntuals localitzades a una posició $\mathbf{r}$ respecte un cert sistema de referència inercial $\mathcal{I}$, la posició relativa dels quals és constant. És a dir, si $\mathbf{r}_a(t)$ i $\mathbf{r}_b(t)$ són les posicions de dues masses puntuals respecte $\mathcal{I}$ a un cert instant de temps $t$, llavors $||\mathbf{r}_a(t) - \mathbf{r}_b(t)|| = constant$.

Per tal de poder estudiar l'actitud del sòlid rígid, suposarem que existeix un segon sistema de referència, anomenat *sistema de referència del cos*, que està acoplat en tot moment al sòlid rígid. El denotarem per $\mathcal{B}$.

Considerem ara dos sistemes de referència qualsevols $\mathcal{A}$ i $\mathcal{B}$, i definim $R_{\mathcal{AB}}$ com la matriu de rotació que ens alinea aquests dos sistemes de referència. Pel teorema de rotació d'Euler, aquesta matriu és equivalent a una rotació d'angle $\theta$ al voltant d'algun eix $\mathbf{v}$ expressat en coordenades de $\mathcal{A}$. Així doncs, usem $R(\theta, \mathbf{v})$ per denotar la matriu de rotació. Suposarem que $\mathbf{v}$ és un vector unitari.

Definirem l'actitud d'un sòlid rígid com la matriu de rotació $R_{\mathcal{IB}}$ que ens permet alinear el sistema de referència $\mathcal{B}$ amb el sistema de referència inercial $\mathcal{I}$. Anomenarem aquestes matrius com a *matrius d'actitud*.

### B. Quaternions

Formalment, es defineix l'anell dels quaternions $\mathbb{H}$ com el generat per l'element real 1 i quatre elements imaginaris $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ que compleixen les següents propietats:

$$
\begin{aligned}
&1 \cdot \mathbf{i} = \mathbf{i}, && \mathbf{i} \cdot \mathbf{j} = -\mathbf{j} \cdot \mathbf{i} = \mathbf{k},\\
&1 \cdot \mathbf{j} = \mathbf{j}, && \mathbf{j} \cdot \mathbf{k} = -\mathbf{k} \cdot \mathbf{j} = \mathbf{i},\\
&1 \cdot \mathbf{k} = \mathbf{k}, && \mathbf{k} \cdot \mathbf{i} = -\mathbf{i} \cdot \mathbf{k} = \mathbf{j},\\
&\mathbf{i} \cdot \mathbf{i} = \mathbf{j} \cdot \mathbf{j} = \mathbf{k} \cdot \mathbf{k} = -1, &&
\end{aligned}
$$

Els elements $\mathfrak{q} \in \mathbb{H}$ es poden considerar vectors 4-dimensionals de la forma:

$$\mathfrak{q} = q_0 + q_1\mathbf{i} + q_2\mathbf{j} + q_3\mathbf{k},$$

amb $q_0, ..., q_3 \in \mathbb{R}$. Habitualment, però, representarem els quaternions com si fóssin vectors 4-dimensionals de $\mathbb{R}^4$, de la forma:

$$\mathfrak{q} = \begin{bmatrix} q_0 \\ q_1 \\ q_2 \\ q_3 \end{bmatrix} = \begin{bmatrix} q_0 \\ \mathbf{q} \end{bmatrix} = \begin{bmatrix} q_0 \\ \mathbf{q} \end{bmatrix}^T$$

a on $\mathbf{q}$ denota el vector de coeficients de les tres unitats imaginàries.
Donats dos quaternions $\mathfrak{a} = [a\,\mathbf{a}]^T$ i $\mathfrak{b} = [b\,\mathbf{b}]^T$, es defineix el seu producte com:

$$\mathfrak{a} \otimes \mathfrak{b} = [a_0 b_0 - \mathbf{a} \cdot \mathbf{b}, a_0\mathbf{b} + b_0\mathbf{a} + \mathbf{a} \times \mathbf{b}]^T$$

on $\cdot$ i $\times$ denoten el producte escalar i vectorial a $\mathbb{R}^3$, respectivament.

Donada una matriu de rotació de la forma $R(\theta, \mathbf{v})$, on $\theta$ és un angle i $\mathbf{v}$ un vector unitari, la podem expressar com a quaternió unitari fent servir la següent fòrmula:

$$\mathfrak{q} = \begin{bmatrix} cos(\theta/2) \\ sin(\theta/2)\mathbf{v} \end{bmatrix}$$

Aquest resultat és important donat que permet establir una relació directa entre els quaternions unitaris i les matrius de rotació. Per tant, si tenim una matriu d'actitud $R_{\mathcal{IB}}$, la podem expressar com a quaternió.

### C. Equació de rotació d'Euler

Els canvis en la velocitat angular ($\boldsymbol{\omega}$) d'un sòlid rígid, expressats en el sistema de referència del cos ($\mathcal{B}$) es poden descriure mitjançant l'equació d'Euler per a la dinàmica d'un sòlid rígid:

$$\dot{\boldsymbol{\omega}} = \begin{bmatrix} \dot{\omega}_x \\ \dot{\omega}_y \\ \dot{\omega}_z \end{bmatrix} = I^{-1}(T - \boldsymbol{\omega} \times (I \cdot \boldsymbol{\omega})), \tag{S1}$$

on $I \in \mathcal{M}_{3\times3}(\mathbb{R})$ representa el tensor d'inèrcia del sòlid rígid i $T = [T_x, T_y, T_z]^T \in \mathbb{R}^3$ representa el vector de torque que actua sobre el centroide del sòlid rígid. En la majoria de situacions, aquest vector $T$ representarà el torque de control generat pels actuadors de la nau, malgrat que en alguns casos també pot contenir les pertorbacions ambientals.

L'orientació d'un sòlid rígid es pot descriure mitjançant quaternions com la rotació que hi ha entre un sistema de referència inercial i el del cos. Els canvis en la orientació es poden representar mitjançant la següent fòrmula:

$$\dot{\mathfrak{q}} = \frac{1}{2}\mathfrak{q} \otimes \begin{bmatrix} 0 \\ \boldsymbol{\omega}. \end{bmatrix} \tag{S2}$$

Per tant, tenint en compte les equacions Eq. (S1) i Eq. (S2), podem descriure els canvis en la totalitat de l'actitud del sòlid rígid mitjançant l'equació d'espai d'estats:

$$\begin{cases} \dot{\mathfrak{q}} = \frac{1}{2}\mathfrak{q} \otimes \begin{bmatrix} 0 \\ \boldsymbol{\omega} \end{bmatrix}, \\ \\ \dot{\boldsymbol{\omega}} = I^{-1}(T - \boldsymbol{\omega} \times (I \cdot \boldsymbol{\omega})). \end{cases} \tag{S3}$$

El problema del control d'actitud té l'objectiu de, donada una actitud actual $(\mathfrak{q}, \boldsymbol{\omega})$, trobar la seqüència de maniobres en forma de vectors de torque $T$ que, al aplicar-se, condueixin el sistema definit a l'equació S3 a una certa actitud desitjada $(\mathfrak{q}_d, \boldsymbol{\omega}_d)$ en un temps finit (i, preferentment, curt). Un controlador és un algorisme que permet sol·lucionar aquest problema dictant el torque que cal produïr a cada instant. Nosaltres usarem un controlador basat en *aprenentatge per reforç* per tal d'assolir aquest objectiu.

## 3. APRENENTATGE PER REFORÇ

Durant la tesi de màster hem fet diversos experiments i provat diverses variacions per al problema d'aprenentatge per reforç. En aquesta secció discutirem el plantejament i els resultats de l'Experiment III de la memòria, el més significatiu, on s'intentava controlar un satèl·lit de forma cúbica en un entorn amb pertorbacions ambientals.

### A. Objectiu

L'objectiu d'aquest experiment és obtenir un controlador que, partint d'una orientació arbitrària i una velocitat angular aleatòria de, com a màxim, $\pm 3$ rad/s en tots els eixos del cos, ens arribi a l'actitud desitjada $\mathfrak{q}_d = [\pm 1, 0, 0, 0]$, $\boldsymbol{\omega}_d = [0, 0, 0]$.

### B. Entorn de simulació, estats i accions

L'aprenentatge per reforç és un sub-camp del Machine Learning que busca maneres de crear agents intel·ligents capaços d'aprendre per si sols la manera òptima de desenvolupar una tasca en un entorn. En el nostre cas particular, l'entorn vindrà donat per la dinàmica definida a l'equació Eq. (S3), i l'agent serà el controlador del satèl·lit. La figura S2 mostra un diagrama del funcionament d'aquest entorn d'aprenentatge per reforç.

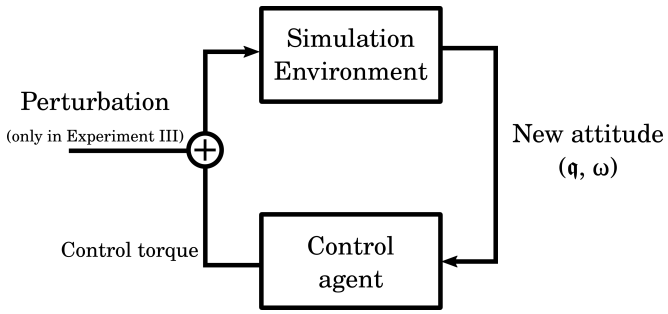


**Fig. S2.** Diagrama del funcionament del programa d'aprenentatge per reforç. L'entorn de simulació rep un torque, i integra l'equació d'espai d'estats per tal de trobar la nova actitud. Aquesta, es passada a l'agent de control el qual ha de determinar el torque que cal fer.

L'estat actual del satèl·lit a cada instant vindrà codificat per un vector de 7 elements. Els primers 4 representen la orientació (en quaternions), mentre que els altres 3 representen la velocitat angular (en rad/s) al voltant dels tres eixos del sistema de referència del cos:

$$[\overbrace{q_0, q_1, q_2, q_3}^{\text{orientation (quaternion)}}, \underbrace{\omega_x, \omega_y, \omega_z}_{\text{angular vel.}}] \tag{S4}$$

L'algorisme d'aprenentatge per reforç que hem usat per aquest experiment, i que discutirem més avall, és la versió discreta de l'algorisme Proximal-Policy Optimization. Això vol dir que hem de trobar una manera de discretitzar i codificar el torque de control en un conjunt d'accions discret.

Així doncs, definirem 31 accions, cada una de les quals correspondrà a un valor de Torque entre 1 i $10^{-4}$ en un dels tres eixos del cos:

| **Action number** | **Torque** $(N \cdot m)$ |
|---|---|
| 1 | $[0, 0, 0]$ |
| 2 and 3 | $[\pm 1, 0, 0]$ |
| 4 and 5 | $[0, \pm 1, 0]$ |
| 6 and 7 | $[0, 0, \pm 1]$ |
| 8 and 9 | $[\pm 10^{-1}, 0, 0]$ |
| 10 and 11 | $[0, \pm 10^{-1}, 0]$ |
| 12 and 13 | $[0, 0, \pm 10^{-1}]$ |
| $\vdots$ | $\vdots$ |
| 26 and 27 | $[\pm 10^{-4}, 0, 0]$ |
| 28 and 29 | $[0, \pm 10^{-4}, 0]$ |
| 30 and 31 | $[0, 0, \pm 10^{-4}]$ |

**Taula S1.** Accions disponibles juntament amb els seus torques respectius.

Considerarem que el nostre satèl·lit és un petit *cubesat* de 5 Kg de massa i 0.83 cm de costat. El seu tensor d'inèrcia serà, doncs,

$$I = 5 \cdot \frac{0.83^2}{6} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

### C. Pertorbacions

En aquest experiment tindrem dues fonts de pertorbacions:

- **Ambientals**: afegirem un torque aleatori al torque de control, que generarem a partir d'una distribució $Normal(0, 10^{-3})$. Aquests torques representaran pertorbacions ambientals tals com el fregament amb l'atmosfera o l'impacte amb petites partícules.
- **Forma i massa**: sumarem una matriu simètrica aleatòria $3 \times 3$ al tensor d'inèrcia, els valors de la qual es generaràn a partir d'una distribució Uniforme $U(0, 0.1)$.

### D. Funció de recompensa

Hem dit que el nostre objectiu era aconseguir un controlador capaç de conduïr el satèl·lit a una actitud desitjada, entrenat mitjançant aprenentatge per reforç. No obstant, l'objectiu de qualsevol algorisme d'aprenentatge per reforç és sempre trobar una política [2] òptima que maximitzi la suma de recompenses a llarg plaç; per tant, cal definir una funció de recompensa adaptada al problema:

$$Reward = -\alpha \cdot (|q_1| + |q_2| + |q_3|) - \beta(|\omega_x| + |\omega_y| + |\omega_z|)$$

amb $\alpha, \beta \in \mathbb{R}$. Observem que aquesta funció està sempre entre $-\infty$ i 0, i creix a mesura que l'error es va fent més petit.

L'objectiu de l'agent és, doncs, aprendre una política òptima que determini, en funció de l'estat actual, quina acció (torque) s'ha d'aplicar per d'acabar maximitzant aquesta funció de recompensa a llarg plaç.

[2] Podem pensar una política com una "estratègia de joc"per realitzar una tasca de manera òptima

### E. Proximal-Policy Optimization

Existeixen molts algorismes d'aprenentatge per reforç. Per a aquest experiment hem utilitzat l'algorisme PPO (Proximal Policy Optimization) per tal de construir i entrenar l'agent. Aquest algorisme és complex i té uns fonaments teòrics bastant extensos; per tal de no allargar aquest document, únicament exposarem la idea intuïtiva que hi ha darrere aquest mètode. [3]

L'agent que controla el satèl·lit està composat per dues xarxes neuronals, que són objectes matemàtics que emulen la manera com els sistemes nerviosos naturals reben i processen informació. Aquestes dues xarxes són:

- El **Crític**, que intenta aproximar, a partir de l'experiència, el valor de cada estat. És a dir, en el nostre cas, la suma de recompenses que esperaríem obtenir si el satèl·lit assolís una actitud determinada i, a partir de llavors, seguís una política de control òptima. Per tant, el crític no determina quina acció cal fer, sinó que només aproxima un valor numèric corresponent a la recompensa esperada a llarg plaç.
- L'**Actor**. Amb l'ajuda del valor aproximat pel crític, l'actor intenta aprendre la política òptima que ajudi a controlar correctament el satèl·lit. A diferència del crític, l'actor sí que ordena quina és l'acció que s'ha de realitzar a cada moment.

Totes dues xarxes estàn compostes de 5 capes de 7, 128, 128 i 64 neurones. La última capa del crític té una única neurona, mentre que l'actor en té 31 (una per a cada acció).

### F. Episodis de simulació

L'agent s'entrena al llarg de 6.000 simulacions diferents. Cada un d'aquests episodis és una simulació que parteix d'unes condicions inicials aleatòries i dura 500 segons. L'agent ha d'intentar maximitzar la suma total de les recompenses al llarg d'aquest episodi.

La Figura S3 mostra el procediment complert d'entrenament del controlador. Les línies de color blau clar són la suma de les recompenses per a cada un dels episodis d'entrenament, mentre que la línia blau marí és la mitjana dels darrers 100 episodis. La imatge de la esquerra és el procés d'entrenament complert, mentre que la de la dreta mostra les recompenses després de l'episodi 1000.

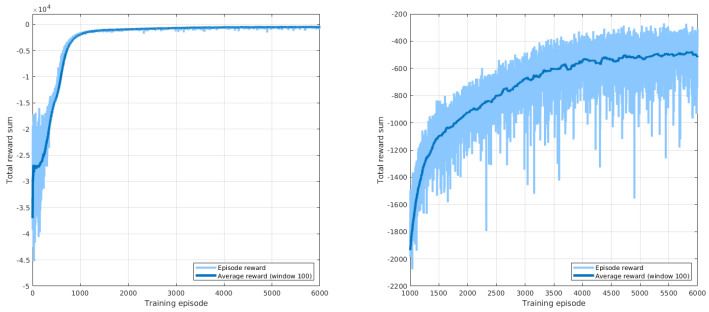


**Fig. S3.** Procés d'entrenament del controlador.

Podem observar que al principi del procés d'entrenament, el controlador té dificultats per controlar correctament el satèl·lit, i la suma de les recompenses és baixa. No obstant, al voltant de l'episodi 500 es comença a millorar els resultats, i s'aprèn una política òptima al voltant de l'episodi 1000 i la suma de recompenses s'estabilitza.

La Figura S4 mostra els resultats d'una simulació amb un controlador totalment entrenat al llarg de 3 segons de simulació. La gràfica superior mostra el canvi en el quaternió de la orientació, la gràfica del mig mostra l'error de l'orientació (en graus), i la gràfica inferior mostra la velocitat angular en els tres eixos. Observem que en totes tres gràfiques es tendeix a l'error nul.

[3] Per a una explicació formal i detallada de l'algorisme, recomano consultar l'article original a on es va formular: Schulman J., Wolski F., Dhariwal P., Radford A., and Klimov O. "Proximal Policy Opti- mization Algorithms", arxiv (Agost 2017).

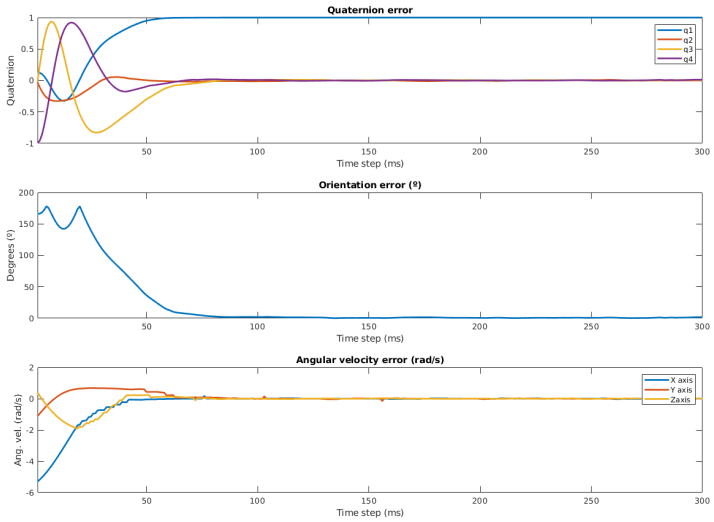


**Fig. S4.** Simulació amb un controlador totalment entrenat.

### G. Simulacions animades

Al següent enllaç poden trobar-se dos vídeos amb els resultats de dues simulacions diferents, la primera amb un controlador sense entrenar i la segona amb un controlador totalment entrenat:
`github.com/RecursiveMagus/AstroIA_MasterThesis/tree/main/Extra_videos`

## 4. CONCLUSIONS

La feina feta en aquest treball mostra que és possible entrenar un controlador mitjançant aprenentatge per reforç que sigui capaç de controlar l'actitud d'un satèl·lit, fins i tot davant la presència de pertorbacions ambientals.

En un futur es continuarà treballant i ampliant aquest projecte en dues direccions diferents:

- Primerament, s'intentarà aconseguir un controlador general que pugui funcionar amb diversos tensors d'inèrcia diferents que representin diversos vehicles espacials d'una mateixa categoria.
- S'intentarà modelitzar amb més precisió les característiques de la nau i limitacions a l'hora de produir torques.
- S'estudiarà afegir penalitzacions per al cost energètic i de combustible per realitzar determinades maniobres.
- Es compararà detalladament el comportament dels controladors basats en aprenentatge per reforç envers altres controladors dissenyats seguint metodologies més tradicionals.

***Isaac de Palau**, Matemàtic i Científic de Dades. Estudiant al Programa de Doctorat en Tecnologia de la Universitat de Girona.*