Report - Tesi de' Salazar
Raggiungimento Target da 2m

Riccardo de' Salazar
proff. *Guido de Matteis, Alessandro Zavoli*
a.a. 2021-2022

12 Marzo 2022

Abstract

In questo report si confrontano due agenti allenati per il raggiungimento di un target posto a 2m di distanza. I due agenti differiscono per diversi coefficienti all'interno della funzione reward

1 Aspetti comuni

1.1 Azioni

Gli agenti hanno le stesse azioni, dai valori compresi tra $(-1, 1)$

$$(AddThrottle, Aileron, Elevator, Rudder)$$

Si tratta di pseudo comandi, diversamente dall'articolo che prevede quattro azioni, una per ogni motore. La caratteristica comune dei due sistemi di azioni è che in entrambi i casi avere quattro azioni nulle corrisponde alla condizione di hover

1.2 Funzione reset()

Nella seguente tabella si evidenziano le distribuzioni degli stati quando viene chiamata la funzione reset().
Insieme al set di stati che vengono definiti dalla funzione reset(), viene definito un raggio ε secondo l'espressione

$$\varepsilon = \varepsilon_0 \left(1 - \frac{n}{N_{tot}}\right)$$

dove ε_0 è un raggio iniziale, n è il numero di epsiodio corrente e N_{tot} è il numero totale di episodi.
Il numero di epsiodi totali nonè predicibile ad inizio allenamento : la funzione reset() viene richiamata ad ogni fine episodio, la fine di un episodio può avere diverse ragioni, ognuna determinata dalla funzione *isDone*. Quello che ho fatto è vedere, durante il primo allenamento svolto, il raggio risultante a fine allenamento e risettare il numero totale di episodi in modo da avere raggio nullo a fine allenamento. Gli angoli di Eulero vengono poi trasformati in quaternioni attraverso la conversione

Variabile	Media	Deviazione Std.
α	π	π
X	2cos α	0.02
Y	2sin α	0.02
Z	2cosα	0.02
u	0	0.025
v	0	0.025
w	0	0.025
p	0	0.0175
q	0	0.0175
r	0	0.0175
ϕ	0	0.44
θ	0	0.44
ψ	0	0.44

$$q0 = cos\left(\frac{\phi}{2}\right) cos\left(\frac{\theta}{2}\right) cos\left(\frac{\psi}{2}\right) + sin\left(\frac{\phi}{2}\right) sin\left(\frac{\theta}{2}\right) sin\left(\frac{\psi}{2}\right)$$

$$q1 = sin\left(\frac{\phi}{2}\right) cos\left(\frac{\theta}{2}\right) cos\left(\frac{\psi}{2}\right) - cos\left(\frac{\phi}{2}\right) sin\left(\frac{\theta}{2}\right) sin\left(\frac{\psi}{2}\right)$$

$$q2 = cos\left(\frac{\phi}{2}\right) sin\left(\frac{\theta}{2}\right) cos\left(\frac{\psi}{2}\right) + sin\left(\frac{\phi}{2}\right) cos\left(\frac{\theta}{2}\right) sin\left(\frac{\psi}{2}\right)$$
$$q3 = cos\left(\frac{\phi}{2}\right) cos\left(\frac{\theta}{2}\right) sin\left(\frac{\psi}{2}\right) - sin\left(\frac{\phi}{2}\right) sin\left(\frac{\theta}{2}\right) cos\left(\frac{\psi}{2}\right)$$

2 Aspetti Differenti

2.1 Osservazioni

Per gli agenti si sono impiegate 18 osservazioni

- Componenti (9) della matrice di rotazione da assi corpo ad assi terra, espressa utilizzando i quaternioni
- Velocità angolari : p, q, r
- Vettore Posizione : X, Y, Z
- Vettore delle velocità in assi corpo : u, v, w

Il vettore delle normalizzazioni è il seguente

$$(1, 1, 1, 1, 1, 1, 1, 1, 1, 30, 30, 30, 10, 10, 10, 20, 20, 20)$$

2.2 Architettura della Rete Neurale

Gli agenti impiega due reti neurali diverse per critico e attore:

- il critico ha una rete neurale avente due strati nascosti di **128 e 128** neuroni.
- l'attore ha una rete neurale avente tre strati nascosti di **32, 32 e 8** neuroni

La funzione di attivazione è la **rectified linear unit ReLU**.

2.3 Parametri dell'allenamento

- time step per la simulazione del modello di quadricottero : $0.01s$
- time step per ia valutazione e l'output della policy : $0.04s$
- time step massimi per episodio : 200 , corrispondente a $8s$
- Learning time steps : $10 * 10^6$
- Learning rate iniziale : $5 * 10^{-4}$
- Learning rate : Learning rate iniziale $* \left(1 - \frac{time}{Learningtimesteps}\right)$
- Cliprange iniziale : 0.35
- Cliprange : Cliprange iniziale $* \left(1 - \frac{time}{Learningtimesteps}\right)$
- Coefficiente di Entropia : $5 * 10^{-8}$
- GAE Factor : 0.99
- Numero di MiniBatch : 8
- Discount Factor : 0.9999
- Numero di Epoche di Apprendimento : 32
- Orizzonte Temporale (in steps) : 8192
- Numero di cpu : 6

3 Aspetti Differenti

3.1 Funzione Reward

La funzione reward in entrambi i casi è costituita da 3 elementi principali

- la norma del vettore degli errori di posizione X_{error}, Y_{error} e Z_{error}

$$||p_e|| = \sqrt{X_{error}^2 + Y_{error}^2 + Z_{error}^2}$$

- la norma del vettore degli errori sugli angoli ϕ_{error}, θ_{error} e ψ_{error}

$$||q_e|| = \sqrt{\phi_{error}^2 + \theta_{error}^2 + \psi_{error}^2}$$

- la norma del vettore delle azioni

$$||a|| = \sqrt{AddThrottle^2 + Aileron^2 + Elevator^2 + Rudder^2}$$

Il reward per l'**Agente 1** ha la seguente espressione

$$R = 1 - 1.3\left(\frac{||p_e||}{3.5}\right) - 0.9\left(\frac{||q_e||}{2.3232}\right) - 0.1\left(\frac{||a||}{2}\right)$$

Il reward per l'**Agente 2** ha invece la seguente espressione

$$R = 1 - 0.02\left(\frac{||p_e||}{3.5}\right) - 0.02\left(\frac{||q_e||}{2.3232}\right) - 0.02\left(\frac{||a||}{2}\right)$$

Nella funzione *isDone* viene aggiunto un termine *isarrived*, vero quando la posizione corrente del quadricottero è tale che la distanza dal target sia nimore del raggio ε. Quando questo termine è vero l'episodio viene terminato e il reward viene definito come segue.

$$R = R + (200 - numTimestepElasped) * \left(2 - \frac{\varepsilon}{2\varepsilon_0}\right)$$

il che significa che ogni volta che viene raggiunta la sfera di raggio ε viene agginto al reward corrente un termine proporzionale al numero di timesteps mancanti a fine episodio e proporzionale ad un termine crescente al diminuire del raggio della sfera. Questo secondo fattore viene inserito perchè si osserva che, in sua assenza, il miglior risultato in termini di allenamento viene raggiunto dall'agente quando la sfera è ancora molto grande, nel primo terzo dell'allenamento, e a seguire, il raggiungimento della sfera (avente nel frattempo raggio sempre più piccolo) avviene ad un numero di timestep più elevato e quindi il premio per il raggiungimento del target risulta minore. Il secondo fattore inoltre non ha massimo valore 1 ma 2 perchè negli istanti iniziali dell'allenamento l'agente deve capire che raggiungere il target corrisponde ad un guadagno maggiore: quello che succede con fattore avente massimo valore 1 è che l'agente trova indifferente o peggiorativo aver raggiunto il target negli istanti iniziali dell'episodio.
Il denominatore del secondo termine all'interno del secondo fattore del secondo termine a secondo memebro nell'ultima espressione è $2\varepsilon_0$ e non ε_0 perchè così l'agente osserva un vantaggio nel raggiungere il target anche a inizio allenamento quando la sfera è molto grande.
Il numero che normalizza il termine del reward sugli errori sugli angoli sono calcolati prendendo come massimi valori di offset quelli che secondo il reset randomico sono i valori delle variabili a due deviazioni standard dalla media. Il numero che normalizza i due termini del reward relativi all'errore di posizione e alle azioni sono calcolati rispetto al massimo valore che ogni termine può assumere ad inizio episodio.

Rispetto all'articolo, ho moltiplicato per 10 i termini relativi al secondo agente. Questo perchè nell'articolo il reward non era complementare ad uno come avviene nel nostro caso. La struttura del reward $R = 1 - ...$ permette di avere una riscrittura del tipo impiegato in questo caso una volta raggiunto il target. Ho moltiplicato per 10 i coefficienti perchè si osserva che l'agente non trovi nella parte iniziale di allenamento un grande vantaggio nel raggiungere il target e che quindi tende a rimanere all'esterno delle target per proseguire l'episodio avendo una penalità per questo molto piccola rispetto ad uno.
L'aspetto che viene mantenuto nel secondo agente rispetto all'agente dell'articolo è che i tre termini, al netto delle normalizzazioni, hanno uguale peso all'interno del reward. Si osserva che questo aspetto ha un peso nei risultati finali

4 Confronto

Si sono svolte 25 simulazioni da posizione iniziale randomica e target in (0, 0, 0)
L'**Agente 1** ha una distanza media dal target a fine simulazione di $d1_{Target} = 0.01104m$
L'**Agente 2** ha una distanza media dal target a fine simulazione di $d2_{Target} = 0.01152m$

5 Analisi Risultati

Il Peso dei coefficienti all'interno del reward sui risultati finali è piuttosto marginale : in entrambi i casi il quadricottero arriva ad una distanza di circa un centimetro dal target e la differenza tra i due casi è mediamente al di sotto di un millimetro. Si riscontra un notevole miglioramento dei risultati rispeto ai casi svolti in assenza del target che si rimpicciolisce. Nel transiente si ha un comportamento mediamente migliore nel caso in cui i termini del reward sono pesati diversamente.

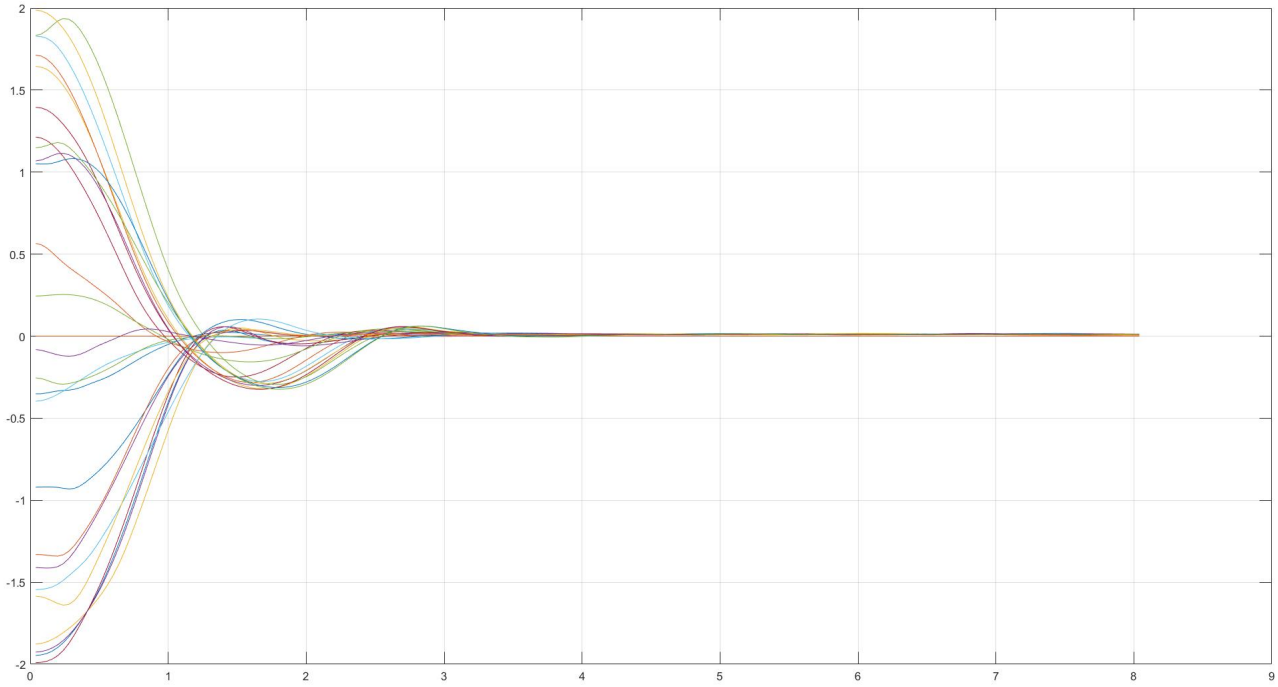

Figure 1: Coordinata X con Agente 1

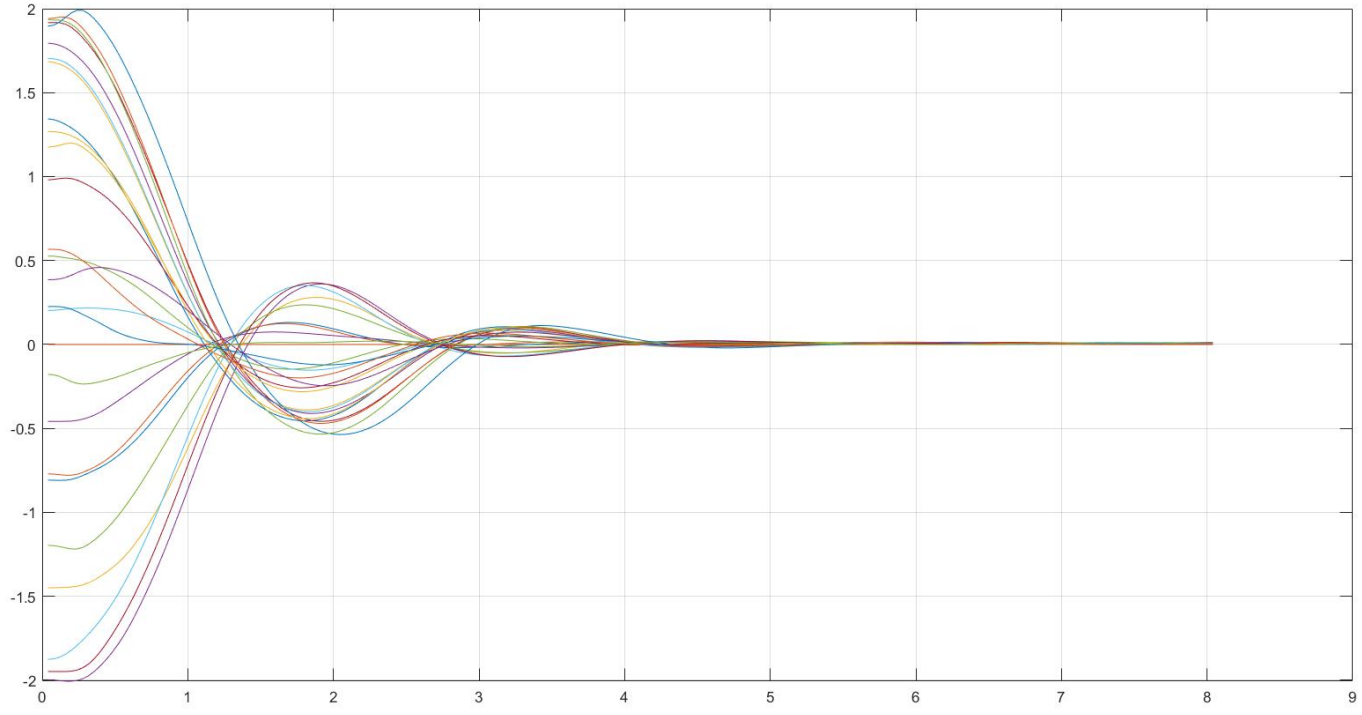

Figure 2: Coordinata X con Agente 2

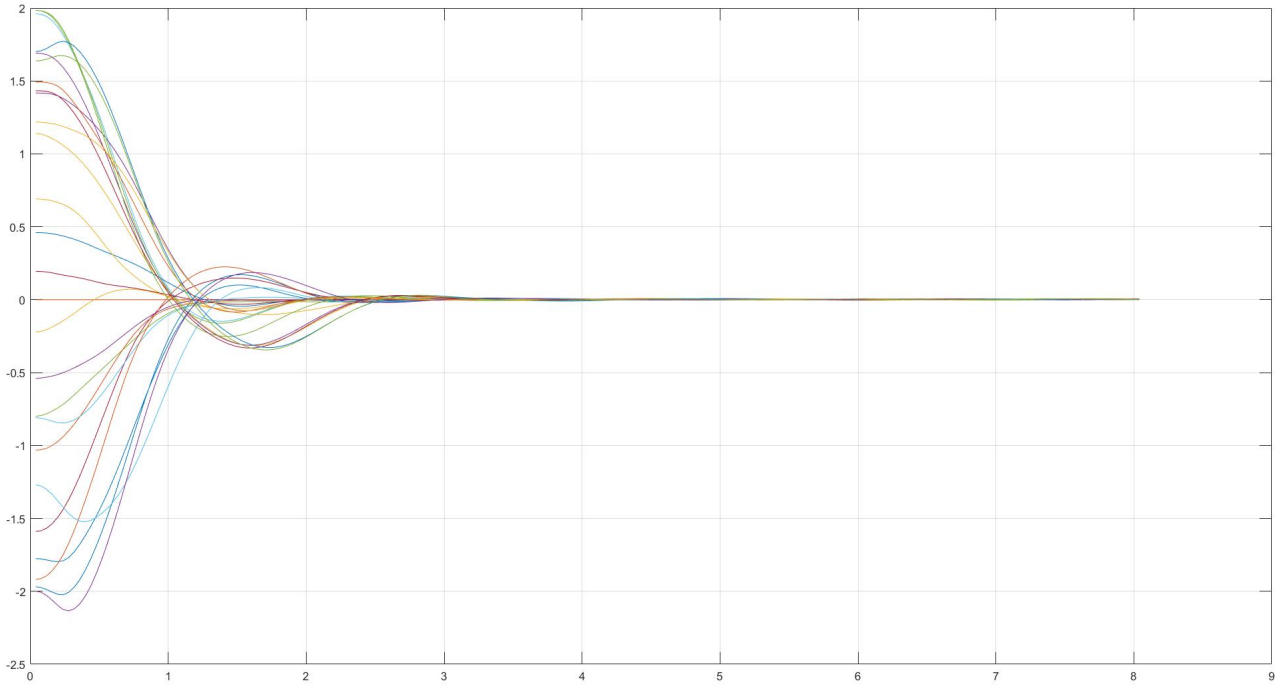

Figure 3: Coordinata Y con Agente 1

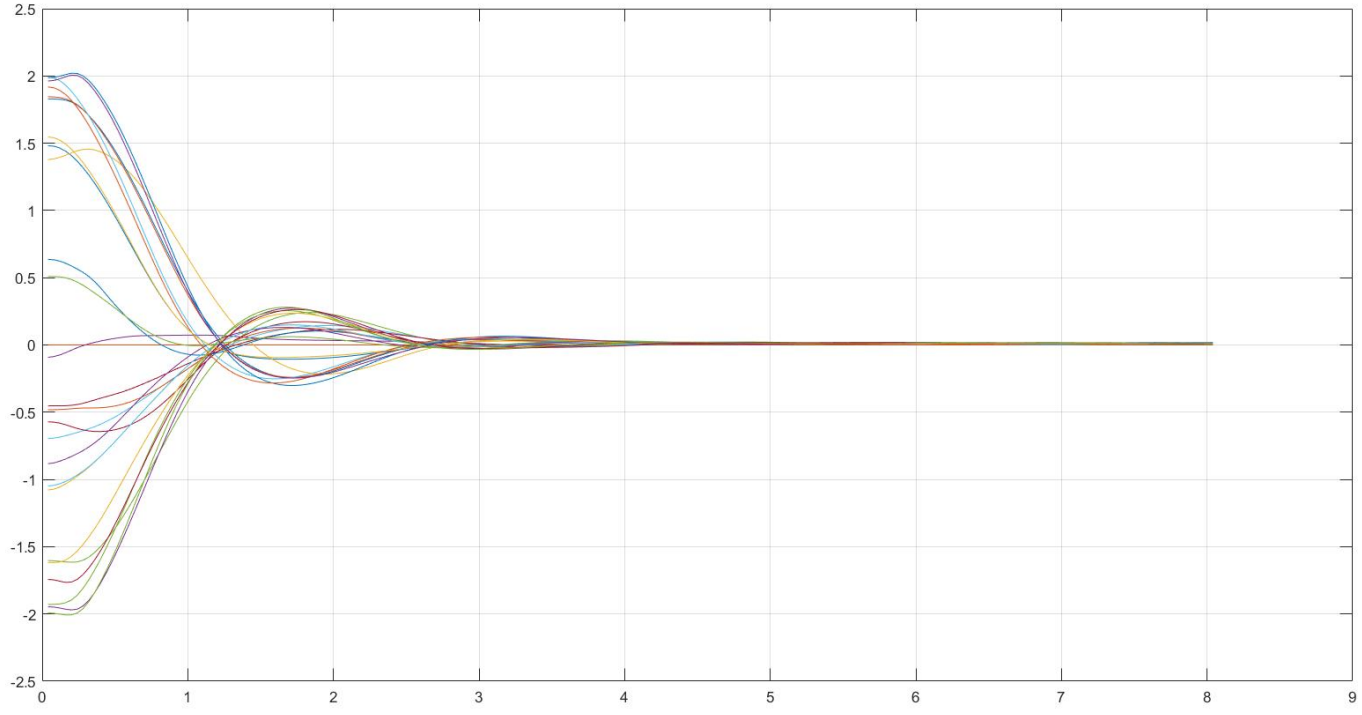

Figure 4: Coordinata Y con Agente 2

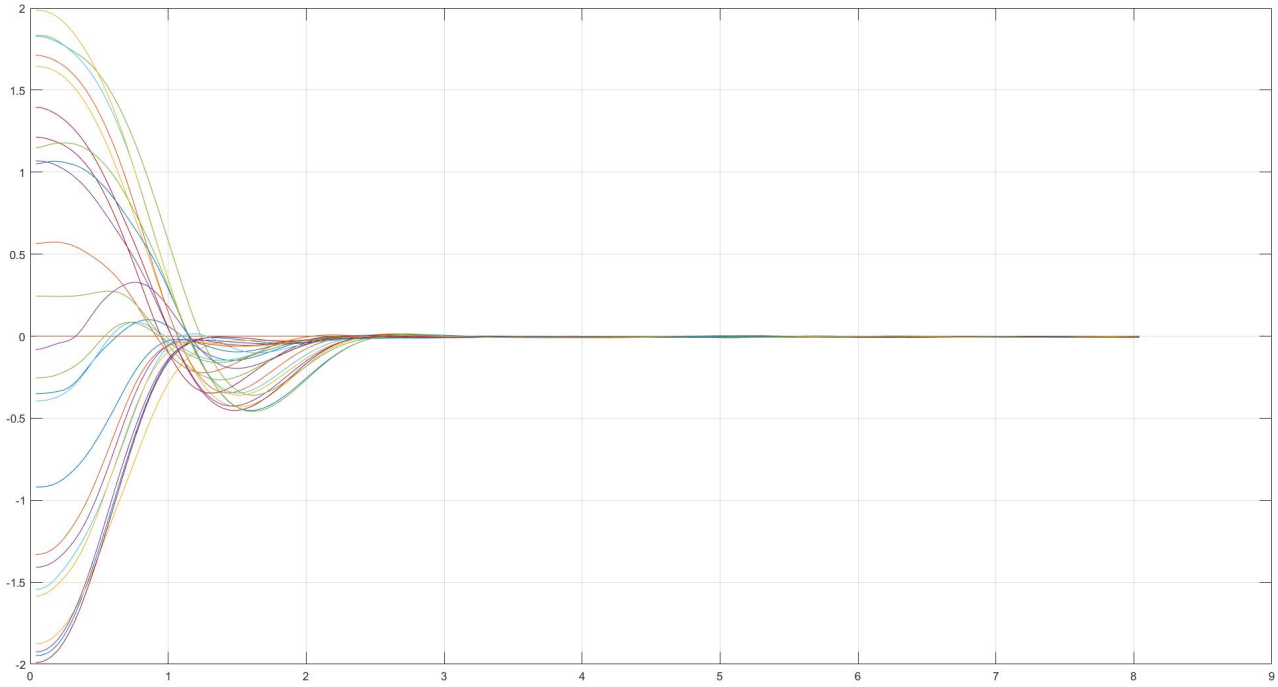

Figure 5: Coordinata Z con Agente 1

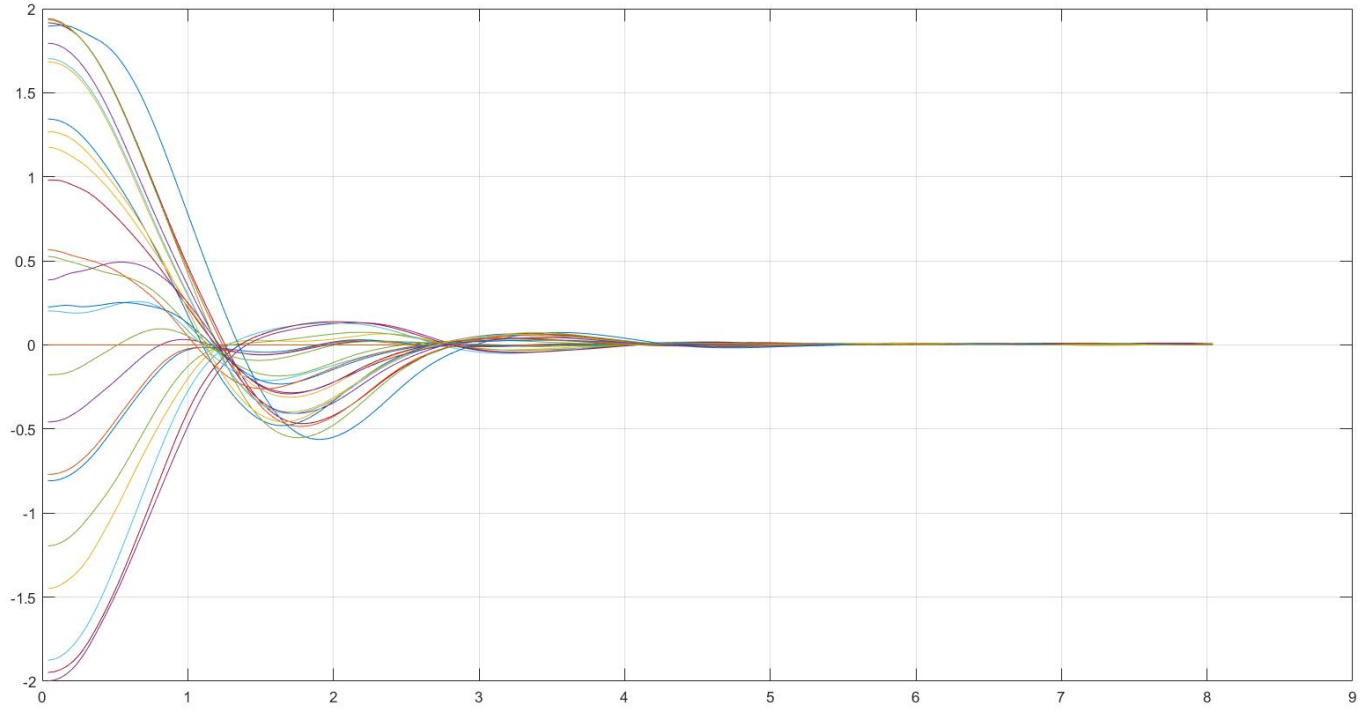

Figure 6: Coordinata Z con Agente 2