Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Πεμπτο Εργαστηριο
Νευροασαφης Έλεγχος

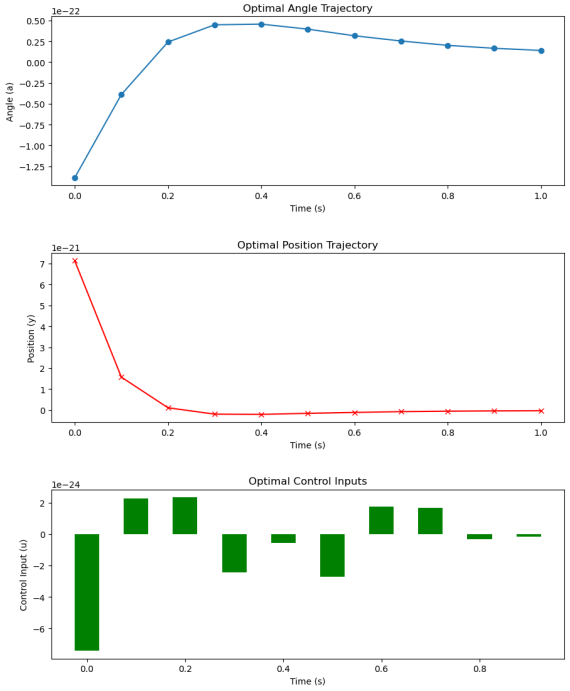
Αναστασία Χριστίνα Λίβα
03119029

Θέμα 1

Παίρνω τα εξής αποτελέσματα:

Listing 1: Optimal Control and State Trajectory

```
Optimal control inputs: [-7.44184868e-24  2.26398560e-24  2.34247015e
    ↪ -24 -2.43048299e-24
 -5.79003765e-25 -2.72710916e-24  1.76064012e-24  1.66596890e-24
 -3.39843485e-25 -1.68662358e-25]
Optimal state trajectory: [[-1.38807953e-22  7.13158267e-21]
 [-3.88724753e-23  1.57249670e-21]
 [ 2.41948147e-23  1.09330118e-22]
 [ 4.48661256e-23 -1.95394180e-22]
 [ 4.56888986e-23 -2.05500293e-22]
 [ 3.96662224e-23 -1.57951907e-22]
 [ 3.16961845e-23 -1.12127012e-22]
 [ 2.54292492e-23 -7.72238339e-23]
 [ 2.01842939e-23 -5.42165295e-23]
 [ 1.66636469e-23 -3.98003684e-23]
 [ 1.40728800e-23 -3.23834707e-23]]
```



Σχήμα 1: Plots

Θέμα 2

Η μέθοδος Monte Carlo για τον υπολογισμό του π αξιοποιεί την αρχή της τυχαίας δειγματοληψίας για την προσέγγιση της τιμής του π. Αρχικά, θεωρούμε ένα τετράγωνο με πλευρά 2 μονάδων (συνολική επιφάνεια 4 τετραγωνικών μονάδων) που περιέχει έναν εγγεγραμμένο κύκλο με ακτίνα 1 μονάδα. Το εμβαδόν του κύκλου είναι π. Διασκορπίζοντας τυχαία σημεία μέσα στο τετράγωνο, η αναλογία των σημείων που πέφτουν εντός του κύκλου προς το σύνολο των σημείων αντιστοιχεί στην αναλογία του εμβαδού του κύκλου προς το εμβαδόν του τετραγώνου, δηλαδή $\pi/4$. Για να υπολογίσουμε το π, πολλαπλασιάζουμε αυτή την αναλογία με το 4. Για παράδειγμα, αν 40 σημεία βρίσκονται μέσα στον κύκλο και 10 εκτός, τότε:

$$\pi = 4 \times \left(\frac{\text{σημεία εντός κύκλου}}{\text{συνολικά σημεία}} \right) = 4 \times \frac{40}{40 + 10} = 3,2.$$

Όσο αυξάνεται ο αριθμός των σημείων, τόσο πιο ακριβής γίνεται η προσέγγιση του π.

Θέμα 3

Στην παρούσα άσκηση, αντιμετωπίζουμε ξανά ένα πρόβλημα από την προηγούμενη σειρά ασκήσεων. Για το πρώτο ερώτημα, εφαρμόζοντας την εξής πολιτική:

```
[-1, -1, -1, -1, 1, 1, 1, 1, -1, -1]
```

που εμφανίστηκε στις περισσότερες από τις λύσεις του προβλήματος, προσομοιώνουμε την πορεία του συστήματος (για όλες τις αρχικές καταστάσεις)

```
Starting position: 1, Path: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Starting position: 2, Path: [2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Starting position: 3, Path: [3, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Starting position: 4, Path: [4, 3, 3, 2, 2, 2, 2, 1, 1, 1, 1]
Starting position: 5, Path: [5, 6, 6, 6, 6, 7, 8, 9, 8, 9, 9]
Starting position: 6, Path: [6, 7, 7, 7, 7, 8, 8, 9, 9, 8, 9]
Starting position: 7, Path: [7, 8, 8, 9, 8, 8, 8, 9, 8, 9, 8]
Starting position: 8, Path: [8, 8, 8, 9, 8, 8, 9, 9, 9, 8, 9]
Starting position: 9, Path: [9, 9, 8, 8, 9, 9, 9, 9, 9, 9, 8]
Starting position: 10, Path: [10, 10, 10, 10, 9, 8, 9, 8, 9, 8, 9]
```

Για την εφαρμογή του Q-learning σε αυτό το σύστημα, θα προσομοιώσουμε τις αλληλεπιδράσεις με το περιβάλλον σύμφωνα με τη δυναμική των αλυσίδων Markov. Αρχικά, θα θέσουμε τυχαίες τιμές Q, τις οποίες θα ενημερώνουμε επαναληπτικά βάσει των παρατηρούμενων μεταβάσεων και ανταμοιβών. Οι ανταμοιβές μπορούν να καθοριστούν ως αντίστροφες του κόστους, για παράδειγμα $r = -g(x)$, εφόσον δεν έχουμε ακριβείς ανταμοιβές. Θα θέσουμε αυθαίρετες αρχικές τιμές Q για κάθε συνδυασμό κατάστασης-ενέργειας. Στη συνέχεια, επαναλαμβάνουμε τη διαδικασία έως ότου επιτευχθεί σύγκλιση. Αρχικοποιούμε την κατάσταση και επιλέγουμε μια ενέργεια με βάση μια στρατηγική εξερεύνησης/εκμετάλλευσης (όπως η ϵ-greedy). Με την επιλεγμένη ενέργεια, παρατηρούμε την ανταμοιβή και την επόμενη κατάσταση. Ενημερώνουμε την τιμή Q χρησιμοποιώντας τον τύπο του Q-learning και λαμβάνοντας υπόψη την τρέχουσα κατάσταση.

Αυτή η προσέγγιση περιλαμβάνει έναν έλεγχο σύγκλισης συγκρίνοντας τις τιμές του πίνακα Q πριν και μετά από κάθε επεισόδιο. Εάν η μέγιστη διαφορά στις τιμές Q μεταξύ των επεισοδίων είναι μικρότερη από ένα καθορισμένο όριο, θεωρείται ότι ο αλγόριθμος έχει συγκλίνει και ο βρόχος σταματά. Αυτή η προσέγγιση επιτρέπει στον αλγόριθμο να εκτελείται μέχρι να μάθει αποτελεσματικά την βέλτιστη πολιτική, όπως υποδεικνύεται από τη σταθεροποίηση των τιμών Q. Η επιλογή του ορίου είναι κρίσιμη: ένα πολύ μικρό όριο μπορεί να οδηγήσει σε υπερβολικά μακρόχρονη λειτουργία του αλγορίθμου, ενώ ένα πολύ μεγάλο όριο μπορεί να τερματίσει τη διαδικασία εκμάθησης πριν αυτή συγκλίνει στην βέλτιστη πολιτική. Οι παράμετροι για αυτήν την διαδικασία είναι:

- **Learning rate:** Επηρεάζει τον τρόπο με τον οποίο οι νέες πληροφορίες τροποποιούν τις υπάρχουσες τιμές Q. Ένα υψηλότερο learning rate προσαρμόζει τις τιμές πιο γρήγορα, αλλά ένα υπερβολικά υψηλό μπορεί να προκαλέσει αστάθεια και να εμποδίσει τη σύγκλιση.
- **Discount factor:** Καθορίζει τη σημασία των μελλοντικών ανταμοιβών σε σχέση με τις άμεσες.
- **Τιμή ϵ:** Εξισορροπεί την εξερεύνηση νέων καταστάσεων και την εκμετάλλευση των ήδη γνωστών καταστάσεων. Η αυξημένη εξερεύνηση μπορεί να αποτρέψει την πρόωρη σύγκλιση σε υπο-βέλτιστες λύσεις.

```
   [[-10.          -11.        ]
 [-10.          -12.9       ]
 [-11.          -15.48441356]
 [-12.89999391 -13.15975991]
 [-12.26200088  -9.63684206]
 [-13.63093391  -6.26315789]
 [ -9.63684211  -4.73684211]
 [ -6.26315789  -5.26315789]
 [ -4.73684211  -6.73684211]
 [ -5.26315789  -6.73684211]]
```