Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Σηματων, Ελεγχου και Ρομποτικης

Σχεδιασμός ρομποτικής κίνησης βάσει ανάδρασης δύναμης σε διαδραστικές εργασίες χειρισμού με εφαρμογή μεθόδων ενισχυτικής μάθησης

Διπλωματικη Εργασια

του

ΠΑΠΑΔΗΜΗΤΡΙΟΥ Χ. ΕΥΘΥΜΙΟΥ

Επιβλέπων: Κωνσταντίνος Τζαφέστας
Αναπληρωτής Καθηγητής

Αθήνα, Απρίλιος 2024

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Σηματων, Ελεγχου και Ρομποτικης

Σχεδιασμός ρομποτικής κίνησης βάσει ανάδρασης δύναμης σε διαδραστικές εργασίες χειρισμού με εφαρμογή μεθόδων ενισχυτικής μάθησης

Διπλωματικη Εργασια

του

ΠΑΠΑΔΗΜΗΤΡΙΟΥ Χ. ΕΥΘΥΜΙΟΥ

Επιβλέπων: Κωνσταντίνος Τζαφέστας
Αναπληρωτής Καθηγητής

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 26η Απριλίου 2024.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Κωνσταντίνος Τζαφέστας	Αθανάσιος Ροντογιάννης	Ιωάννης Κορδώνης
Αναπληρωτής Καθηγητής	Αναπληρωτής Καθηγητής	Επίκουρος Καθηγητής

Αθήνα, Απρίλιος 2024

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Σηματων, Ελεγχου και Ρομποτικης

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Πτυχιακή μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Πτυχιακή Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

(Υπογραφή)

.....
Ευθύμιος Παπαδημητρίου

26 Απριλίου 2024

Περίληψη

Η αυτοματοποιημένη ρομποτική συγκομιδή καρπών αποτελεί σημαντικό πεδίο εφαρμογής για επιδέξιους ρομποτικούς μηχανισμούς χειρισμού εύθραυστων και ευαίσθητων αντικειμένων. Ειδικά το αντικείμενο της συγκομιδής μανιταριών έχει ειδικές απαιτήσεις και η εφαρμογή επιδέξιων ρομποτικών συστημάτων για την αυτοματοποίηση κάποιων σταδίων της διαδικασίας έχει αποκτήσει σημαντικό ενδιαφέρον κατά τα τελευταία χρόνια. Η συγκομιδή συνδυάζει πολλούς διαφορετικούς τομείς όπως ο Επιδέξιος Χειρισμός και η Ενισχυτική Μάθηση, για την επίτευξη μιας σύνθετης διαδικασίας εκρίζωσης του μανιταριού, χωρίς ταυτόχρονα να του προκληθεί ζημιά. Σε αυτή τη Διπλωματική Εργασία υλοποιείται μια μέθοδος Ενισχυτικής Μάθησης Αγνώστου Μοντέλου, η Episodic Linear Semi-gradient SARSA, για το συνδυασμό των ενεργειών της στρέψης και κάμψης (γύρω από συγκεκριμένο άξονα), που αποτελούν θεμελιώδεις ανεξάρτητες κινήσεις για την συμβατική δράση εκρίζωσης και συγκομιδής, αποφεύγοντας την επιβολή μεγάλων ροπών στο μανιτάρι μέσω κινήσεων που αντιτίθενται στη δυναμική του. Οι ροπές θα γίνονται αντιληπτές μέσω ανάδρασης δύναμης. Η συγκεκριμένη δράση επενέργησης σε δύο βαθμούς ελευθερίας ανάγεται αρχικά στην εκπαίδευση ενός κυκλικού πράκτορα να δραπετεύει από έναν διδιάστατο διάδρομο, ο οποίος ορίζεται από τοίχους. Η διείσδυση προκαλεί δυνάμεις επαναφοράς. Ο διάδρομος αντιπροσωπεύει τη δυναμική εκρίζωσης του μανιταριού, και οι δύο διαστάσεις τους δύο βαθμούς ελευθερίας στο πραγματικό πρόβλημα. Στη συνέχεια, ακολουθεί εφαρμογή της μεθόδου σε πραγματική διάταξη ενός αντικειμένου που προσομοιάζει ένα μανιτάρι, με εκπαίδευση του ρομποτικού βραχίονα Panda.

Λέξεις Κλειδιά

Επιδέξιος ρομποτικός χειρισμός, ανάδραση δύναμης, δυναμική εκρίζωσης και συγκομιδής μανιταριών, ενισχυτική μάθηση αγνώστου μοντέλου, προσαρμοστική εξερεύνηση και ρυθμός μάθησης, αλγόριθμος μάθησης SARSA, προσέγγιση συνάρτησης, Radial Basis Function, προσομοίωση, ρομποτικός βραχίονας Panda

Abstract

The automated robotic harvesting is a significant application field for dexterous robotic handling mechanisms of delicate and sensitive objects. Especially the harvesting of mushrooms entails specific requirements, and the application of dexterous robotic systems for automating certain stages of the process has gained significant interest in recent years. Harvesting combines plenty of different domains, such as Dexterous Manipulation and Reinforcement Learning, aiming to carry out the complex process of mushroom outrooting without inflicting damage on it. In this Thesis, a Model-free Reinforcement Learning method is implemented, called Episodic Linear Semi-gradient SARSA, for combining the twisting and bending (about specific axis) actions, which are fundamental independent moves for conventional harvesting, while avoiding excessive torques on the mushroom because of movements that are not compatible to its dynamics. The agent will be notified about the torques through force feedback. This task of 2 DoF action will be firstly implemented in a simplified simulation environment, where a round agent learns to exit from a 2-dimensional maze, which is defined by walls. Penetrating those walls provokes restoring forces. The maze represents the outrooting dynamics of the mushroom, and the 2 dimensions are an expression of the 2 DoF of the real problem. Following the simulation, the method is used on a real object that resembles a mushroom, by training the Panda robot arm.

Keywords

Dexterous robotic manipulation, force/tactile feedback, mushroom outrooting and harvesting dynamics, model-free reinforcement learning, adaptive exploration and learning rate, SARSA learning algorithm, function approximation, Radial Basis Function, simulation, Panda robot manipulator

στους γονείς μου

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω τον καθηγητή κ. Κωνσταντίνο Τζαφέστα για την επίβλεψη αυτής της διπλωματικής εργασίας και για την ευκαιρία που μου έδωσε να την εκπονήσω στο Intelligent Robotics and Automation Laboratory. Επίσης ευχαριστώ ιδιαίτερα τον υποψήφιο διδάκτορα Παρασκευά Οικονόμου για την καθοδήγησή του και για την εξαιρετική συνεργασία που είχαμε. Τέλος θα ήθελα να ευχαριστήσω τους γονείς μου για την αγάπη, την καθοδήγηση και την ηθική συμπαράσταση που μου προσέφεραν όλα αυτά τα χρόνια, σε κάθε βήμα της ζωής μου.

Αθήνα, Απρίλιος 2024

Ευθύμιος Παπαδημητρίου

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Κατάλογος Αλγορίθμων

Κεφάλαιο 1

Εισαγωγή

Η Ρομποτική είναι ένας ταχύτατα εξελισσόμενος κλάδος, ο οποίος παίζει όλο και μεγαλύτερο ρόλο στην ζωή των ανθρώπων. Στις μέρες μας γίνεται ορατή η χρήση ρομποτικών συστημάτων, συμβάλλοντας στον αυτοματισμό μέσα στην κοινωνία. Νέες τεχνολογίες συνεχώς μοντελοποιούνται και δοκιμάζονται, κάνοντας πραγματικότητα σύνθετες προοπτικές ρομπότ, όπως τα αυτόνομα οχήματα και τα ανθρωποειδή (humanoids). Με την εκτόξευση στην έρευνα και χρήση της τεχνητής νοημοσύνης (Artificial Intelligence), έχουν δημιουργηθεί νέοι ορίζοντες στη σύγχρονη ρομποτική, αποτελώντας πλέον δύο πεδία άρρηκτα συνδεδεμένα μεταξύ τους.

1.1 Επιδέξιος Χειρισμός και Τεχνητή Νοημοσύνη

Ο Επιδέξιος Χειρισμός (Dexterous Manipulation) είναι ένα πεδίο της Ρομποτικής όπου χειριστές ή δάχτυλα συνεργάζονται για να πιάνουν και να χειρίζονται αντικείμενα. Το επίκεντρο όλης της διαδικασίας αποτελεί το ίδιο το αντικείμενο, δίνοντας έμφαση στη συμπεριφορά και τις δυνάμεις πάνω σε αυτό. Εφόσον ο χειρισμός αντικειμένων αποτελεί κυρίως μια ανθρώπινη δραστηριότητα, ο σχεδιασμός ρομποτικών χεριών για σχετικά tasks συνήθως έχει ανθρωπομορφικά χαρακτηριστικά (π.χ. multi-fingered). Ο Επιδέξιος Χειρισμός έχει σύγχρονους χώρους έρευνας και εφαρμογών, με παραδείγματα όχι μόνο στα ανθρωποειδή ρομπότ, αλλά και τόσο στην προσθετική χειρουργική, όσο και στην επίτευξη στόχων σε περιβάλλοντα επικίνδυνα για τους ανθρώπους. Παρ' όλα αυτά, η μίμηση της φυσιολογικής ανθρώπινης δραστηριότητας και επιδεξιότητας είναι ακόμα μια πρόκληση, και παρουσιάζει δυσκολίες [5]. Ο κατ' εξοχήν αποτελεσματικός τρόπος εκμάθησης στοιχείων της ανθρώπινης αντίληψης και δραστηριότητας για μηχανές είναι η Τεχνητή Νοημοσύνη. Για την αναγνώριση των χαρακτηριστικών του κάθε μοναδικού αντικειμένου συμβάλλει η Όραση Υπολογιστών (Computer Vision), η οποία, μεταξύ άλλων, προσφέρει αλγόριθμους για αναγνώριση σχήματος, διαστάσεων και πόζας. Επιτακτική κρίνεται ακόμα η χρήση Ενισχυτικής Μάθησης (Reinforcement Learning - RL), ένα είδος Μηχανικής Μάθησης (Machine Learning) κατά το οποίο η μηχανή αλληλεπιδρά με το περιβάλλον και μαθαίνει λαμβάνοντας επιβράβευση ανάλογα με την καταλληλότητα της εκάστοτε δράσης του. Συγκεκριμένα, μπορεί να αποτελέσει τον τρόπο εκμάθησης της επίτευξης ενός task με επιτυχία, ελέγχοντας την υφιστάμενη δύναμη πάνω στο αντικείμενο [6]. Ένας επιπλέον τρόπος να επιταχυνθεί η μάθηση, αλλά και να αναβαθμιστεί η απόδοσή της, είναι η εισαγωγή δεδομένων από ανθρώπινη επίδειξη (human

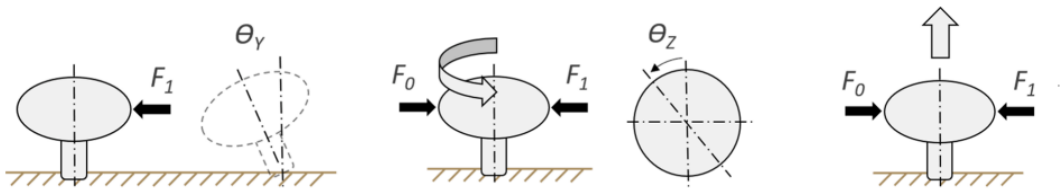
demonstration), που λειτουργούν ως οδηγίες για την επίτευξη του στόχου του ρομπότ [7]. Για τη συλλογή δεδομένων που θα καταστήσουν την αντίληψη του περιβάλλοντος χρησιμοποιούνται κάμερες και αισθητήρες. Επίσης, αναπτύσσονται νέες τεχνολογίες κατασκευής εύκαμπτων και ελαστικών ρομποτικών χεριών με τα κατάλληλα υλικά (Soft Robotics), που συμβάλλουν στην ευαισθησία κατά την αλληλεπίδραση με το αντικείμενο.

1.2 Επιδέξιος Χειρισμός για Συγκομιδή Μανιταριών

Ένα βασικό μέρος της βιομηχανίας των μανιταριών είναι η συγκομιδή, η οποία όχι μόνο επιδρά στην ποιότητα και την ποσότητα των συλλεγόμενων μανιταριών, αλλά και στο κόστος παραγωγής [8]. Η αυτοματοποίηση μπορεί να επιταχύνει σημαντικά αυτή τη διαδικασία [9], όμως τα μανιτάρια που συλλέγονται είναι δύσκολο να είναι σε καλή κατάσταση λόγω της μεταχείρισής τους από τη μηχανή/ρομπότ, και συχνά καταλήγουν να είναι κατεστραμμένα και χαλασμένα [8, 10]. Το ζήτημα της αυτοματοποίησης της συγκομιδής μανιταριών είναι επομένως πολυδιάστατο, και αφορά τόσο τον εντοπισμό του μανιταριού και την εκτίμηση του σχήματος και των διαστάσεών του, όσο και της επιδέξιας συλλογής του χωρίς αυτό να καταστραφεί (αξίζει να σημειωθεί πως αφού κάθε μανιτάρι είναι μοναδικό, και η διαδικασία της ασφαλούς συλλογής του θα διαφέρει ανάλογα).

Όπως αναφέρθηκε παραπάνω, η Τεχνητή Νοημοσύνη μπορεί να συνεισφέρει στα δύο προηγούμενα προβλήματα, στο πρώτο με αλγορίθμους Όρασης Υπολογιστών [8], και στο δεύτερο με μεθοδους Μηχανικής Μάθησης, και ιδιαίτερα Ενισχυτικής Μάθησης.

Για να ερευνηθεί πώς μπορεί να προσεγγιστεί η ασφαλής συγκομιδή ενός μανιταριού με χρήση Ενισχυτικής Μάθησης, χρειάζεται πρώτα να γίνουν κατανοητές οι δράσεις και οι συνδυασμοί τους που μπορούν να φέρουν το επιθυμητό αποτέλεσμα. Συγκεκριμένα, η συμβατική μέθοδος εκρίζωσης ενός μανιταριού από το έδαφος είναι ένας συνδυασμός τριών θεμελιωδών κινήσεων, της στρέψης (twisting), της κάμψης (bending) και της ανύψωσης (lifting), μια διαδικασία όμως που μπορεί να προσθέσει πολυπλοκότητα στο σχεδιασμό ενός end-effector για αυτό το σκοπό. Η στρέψη πραγματοποιείται με περιστροφή του μανιταριού γύρω από τον κεντρικό άξονά του, ενώ η κάμψη με περιστροφή γύρω από το σύνδεσμο στελέχους-υποστρώματος [3].



(α') *Κάμψη* (β') *Στρέψη* (γ') *Ανύψωση*

Εικόνα 1.1: *Οι 3 θεμελιώδεις κινήσεις για την συγκομιδή μανιταριού, όπου F_0, F_1 οι δυνάμεις του αντίχειρα και του δείκτη κατά τη διαδικασία, και Θ_Y, Θ_Z οι γωνίες περιστροφής περί τους 2 άξονες κατά τη διαδικασία (από [3])*

Έχει βρεθεί ότι ο συνδυασμός κάμψης και στρέψης μπορεί να έχει τα καλύτερα αποτελέσματα στην εκρίζωση του μανιταριού από ρομπότ [11]. Για τη μέτρηση και τον έλεγχο των

δυνάμεων που ασκούνται πάνω στο μανιτάρι κατά τη διαδικασία, γίνεται χρήση αισθητήρων και inertial measurement unit (IMU) [3].

Το task, λοιπόν, μπορεί να αντιμετωπιστεί μερικώς ως μια εφαρμογή Ενισχυτικής Μάθησης με συγκεκριμένο χώρο δράσεων (twist-bend) και παρατηρήσεων (δυνάμεις που μετρώνται από τους αισθητήρες), όπου πρέπει να επιτυγχάνεται γενίκευση για μανιτάρια με διαφορετικές διαστάσεις, αρχική πόζα, ρίζα κ.λπ.).

1.3 Αντικείμενο της Διπλωματικής

Η παρούσα διπλωματική εργασία, βασιζόμενη στο παραπάνω θέμα, ασχολείται με την εκμάθηση επιδέξιου χειρισμού ενός αντικείμενου που προσομοιάζει στη δυναμική διαδικασία εκρίζωσης ενός μανιταριού, σε ένα ρομποτικό βραχίονα. Ειδικότερα, πραγματεύεται την υλοποίηση και δοκιμή μεθόδων Ενισχυτικής Μάθησης για την εφαρμογή συνδυασμού στρέψης και κάμψης γύρω από συγκεκριμένο άξονα στο αντικείμενο με ανάδραση δύναμης, αποφεύγοντας την επιβολή μεγάλων ροπών σε αυτό.

Στην πειραματική διάταξη, η βάση του αντικειμένου συνδέεται με κινητήρες που ακολουθούν την κίνηση, μέσω νημάτων και ελατηρίων, στους άξονες της στρέψης και της κάμψης. Μη επιθυμητοί συνδυασμοί των τελευταίων δεν ακολουθούνται από τους κινητήρες, με αποτέλεσμα την επιμήκυνση των ελατηρίων που επιφέρουν συγκριτικά μεγάλες ασκούμενες ροπές πάνω στο αντικείμενο (προσομοιάζοντας έτσι τη δυναμική ενός τυχαίου μανιταριού), τις οποίες καλούμαστε να αποτρέψουμε. Ακόμα, δοκιμάζεται η χρήση αισθητήρων για τη μέτρηση των ασκούμενων ροπών κάθε χρονική στιγμή (μετασχηματισμένων στους άξονες της στρέψης και της κάμψης). Η διάταξη παρουσιάζεται αναλυτικά στην ενότητα 6.1.

Οι μέθοδοι μάθησης πρέπει:

- να είναι online, δηλαδή να δημιουργούν ένα μοντέλο το οποίο μαθαίνει και ανανεώνεται σε κάθε βήμα, γιατί είναι ανάγκη αυτό να προσαρμόζεται άμεσα στα ερεθίσματα που δέχεται αντιδρώντας κατάλληλα στην επιβολή μεγάλων ροπών,

- να είναι αγνώστου μοντέλου (model-free), δηλαδή να μην χρειάζονται πλήρη γνώση του περιβάλλοντος, αφού η δυναμική του αντικειμένου είναι άγνωστη.

Η πρώτη δοκιμή της υλοποίησης των μεθόδων γίνεται σε περιβάλλον προσομοίωσης. Παρουσιάζεται μια απλοποιημένη αλγοριθμική μορφή του ζητούμενου task, που πραγματεύεται την εκμάθηση ενός κυκλικού αντικειμένου να βρίσκει το δρόμο για την έξοδο μέσα σε έναν άγνωστο διδιάστατο διάδρομο, γνωρίζοντας μόνο τις δυνάμεις που ασκούνται πάνω του όταν παρεκκλίνει εκτός διαδρόμου, και οι οποίες τείνουν να το επαναφέρουν σε αυτόν. Έτσι, οι δυνάμεις προσομοιάζουν τις πραγματικές ροπές που δέχεται το αντικείμενο/μανιτάρι, και οι κινήσεις στους 2 άξονες προσομοιάζουν τη στρέψη και την κάμψη, ως μια αντιστοίχιση δύο ανεξάρτητων βαθμών ελευθερίας. Η αξιολόγηση του κάθε αλγορίθμου Ενισχυτικής Μάθησης γίνεται με βάση τρία χαρακτηριστικά:

- την επιτυχία στην επιτέλεση του στόχου, δηλαδή της εξόδου του κυκλικού αντικειμένου από το διάδρομο (που αντιστοιχίζεται στην εφαρμογή των θεμιτών συνδυασμών στρέψης

και κάμψης στην πραγματική διάταξη), καθώς και τη συνολική απόδοση και ευκολία σε αυτή την επίτευξη,

- την ταχύτητα της σύγκλισης, καθώς είναι επιθυμητή μια γρήγορη σύγκλιση, αλλά και προσαρμογή σε άλλα περιβάλλοντα,

- στην αμεσότητα της αντίδρασης, αφού είναι σημαντικό να γίνονται αμέσως οι κατάλληλες κινήσεις για να κρατείται χαμηλό το μέτρο της συνολικής δύναμης (ροπής) που ασκείται στο αντικείμενο.

Στη συνέχεια, η επικρατέστερη μέθοδος εφαρμόζεται σε πραγματική διάταξη για την εκπαίδευση του ρομποτικού χεριού Panda της Franka Emika.

1.4 Σχετική Έρευνα και Συνεισφορά της Εργασίας μας

Η αυτοματοποίηση της συγκομιδής μανιταριών αποτελεί ένα σύγχρονο πεδίο έρευνας με αρκετές προσεγγίσεις. Μια από τις πιο σχετικές με τη δική μας εργασία είναι η [12], η οποία πραγματεύεται ένα παρόμοιο πρόβλημα εκρίζωσης μανιταριού (in-hand) μέσω Ενισχυτικής Μάθησης Αγνώστου Μοντέλου συνδυάζοντας τις θεμελιώδεις κινήσεις της στρέψης και της ανύψωσης, και χρησιμοποιώντας έναν αλγόριθμο μεθόδου Δράστη-Κριτή, τον Continuous Actor Critic Learning Automaton (CACLA). Επίσης, η εργασία εφαρμόζει την ίδια αναγωγή του προβλήματος σε απόδραση ενός αντικειμένου από ένα διδιάστατο διάδρομο.

Η δική μας εργασία ασχολείται κυρίως με την υλοποίηση μιας μεθόδου Ενισχυτικής Μάθησης Αγνώστου Μοντέλου με βασική προϋπόθεση την ελαχιστοποίηση του χρόνου και της υπολογιστικής δύναμης της εκπαίδευσης, έτσι ώστε να διεξάγουμε πειράματα σε πραγματική διάταξη. Πέρα από την υλοποίηση του βασικού αλγορίθμου μας, ενσωματώνουμε στη μέθοδο μάθησης μια μορφή προσαρμοστικής εξερεύνησης και ρυθμού μάθησης, στοιχεία που αποδεικνύονται αποτελεσματικά στη μείωση των απαιτούμενων συνολικών βημάτων εκπαίδευσης, βελτιώνοντας παράλληλα την απόδοση του αλγορίθμου στο δικό μας πρόβλημα.

1.5 Οργάνωση του Τόμου

Η παρούσα διπλωματική εργασία αποτελείται από 7 κεφάλαια.

Στο κεφάλαιο 2 γίνεται μια περιεκτική παρουσίαση του θεωρητικού υπόβαθρου που θα χρειαστεί για την κατανόηση των υλοποιήσεών μας, ξεκινώντας από τον ορισμό των Μαρκοβιανών Διαδικασιών Λήψης Αποφάσεων και καταλήγοντας σε ανάλυση των στοιχείων των μεθοδων που χρησιμοποιήσαμε.

Στο κεφάλαιο 3 καταγράφεται το πρόβλημα το οποίο καλούμαστε να επιλύσουμε και εξάγονται οι απαιτήσεις για την υλοποίησή μας. Επιπλέον, αναλύονται οι επιλογές των αλγορίθμων μας, οι οποίοι περιγράφονται μαζί με περαιτέρω πληροφορίες για αυτούς.

Στο κεφάλαιο 4 περιγράφεται το περιβάλλον προσομοίωσης που χρησιμοποιήσαμε, και αναλύονται οι ακριβείς μέθοδοι και υλοποιήσεις που αναπτύχθηκαν σε αυτό, βασιζόμενοι στο γενικό κορμό που αναφέρεται στο κεφάλαιο 3.

Στο κεφάλαιο 5 επεξηγούνται οι παράμετροι που επιλέχθηκαν για την προσομοίωση, καθώς και παρατίθενται και σχολιάζονται τα αποτελέσματα της εφαρμογής της.

Στο κεφάλαιο 6 παρουσιάζεται η διάταξη του πραγματικού πειράματος (ρομπότ & χώρος εργασίας), οι αλλαγές στις παραμέτρους και οι συμβιβασμοί που χρειάζονται για τη νέα εφαρμογή των μεθόδων, οι περιορισμοί κατά την εκτέλεση, τα πειραματικά αποτελέσματα και ο σχολιασμός τους.

Τέλος, το κεφάλαιο 7 περιλαμβάνει τον επίλογο και τις μελλοντικές επεκτάσεις της διπλωματικής εργασίας.

Μέρος Ι

Θεωρητικό Μέρος

Κεφάλαιο 2

Θεωρητικό υπόβαθρο - Ενισχυτική Μάθηση

Στο κεφάλαιο αυτό παρουσιάζονται βασικές έννοιες και μέθοδοι στην Ενισχυτική Μάθηση, πληροφορίες χρήσιμες για την κατανόηση των αλγορίθμων που χρησιμοποιήθηκαν στην εργασία.

Ενισχυτική Μάθηση ονομάζεται η μάθηση, μέσω αλληλεπίδρασης με το περιβάλλον, των βέλτιστων δράσεων (actions), και συγκεκριμένα της σύνδεσης καταστάσεων (states) με δράσεις, με σκοπό τη μεγιστοποίηση ενός σήματος ανταμοιβής (reward signal), το οποίο μπορεί να παίρνει τη μορφή επιβράβευσης ή ποινής. Ο πράκτορας (agent), δηλαδή αυτός που μαθαίνει να δρα, εκπαιδεύεται μέσω της δοκιμής για να μεγιστοποιηθεί η ανταμοιβή, και πρέπει τόσο να έχει κάποια επίγνωση σχετικά με το περιβάλλον, όσο και να μπορεί να το επηρεάσει μέσω των δράσεών του. Συχνά, αναζητούνται οι δράσεις οι οποίες θα συμβάλλουν στο σύνολο των επόμενων ανταμοιβών αντί για την αμέσως επόμενη, αναπτύσσοντας μια λογική μακροπρόθεσμης μεγιστοποίησης. Ο πράκτορας, λοιπόν, χρειάζεται επιπλέον να θέτει έναν ή περισσότερους στόχους μέσω του σήματος ανταμοιβής, το οποίο του στέλνει μια αριθμητική τιμή ανταμοιβής σε κάθε διακριτό βήμα (iteration) εκπαίδευσης, διαφοροποιώντας έτσι τα θετικά από τα αρνητικά συμβάντα. Το σήμα ανταμοιβής μπορεί να είναι στοχαστική συνάρτηση της κατάστασης και της δράσης της εκάστοτε δεδομένης στιγμής [1].

Η ιδέα αυτού του είδους μάθησης εμπνέεται από την ίδια την ανθρώπινη εμπειρία, όπου η εξάσκηση για ένα σκοπό δημιουργεί δεδομένα σχετικά με τα επιμέρους αίτια, δράσεις και συνέπειες, στοιχεία τα οποία χρησιμοποιούνται για την επίτευξη του σκοπού. Αυτή η ικανότητα της Ενισχυτικής Μάθησης για εκπαίδευση από αλληλεπίδραση με το περιβάλλον και δημιουργία δεδομένων μέσω εμπειρίας, καθώς και η προσπάθεια μεγιστοποίησης του σήματος ανταμοιβής, τη διαφοροποιεί από την Επιβλεπόμενη και τη Μη-επιβλεπόμενη Μάθηση [1].

Στην Ενισχυτική Μάθηση ορίζεται ακόμα το Μοντέλο (Model) του περιβάλλοντος, το οποίο μιμείται τη συμπεριφορά του περιβάλλοντος, και υπάρχει μόνο στα συστήματα Ενισχυτικής Μάθησης όπου το περιβάλλον είναι γνωστό [1]. Άλλα στοιχεία θα παρατεθούν παρακάτω.

2.1 Μαρκοβιανές Διαδικασίες Λήψης Αποφάσεων

Οι Μαρκοβιανές Διαδικασίες Λήψης Αποφάσεων (Markov Decision Processes - MDPs) αποτελούν τον τρόπο περιγραφής ενός περιβάλλοντος στην Ενισχυτική Μάθηση, και σχεδόν όλα τα προβλήματά της μπορούν να μοντελοποιηθούν με MDPs [2].

2.1.1 Μαρκοβιανή Ιδιότητα

Οι MDPs βασίζονται στη Μαρκοβιανή Ιδιότητα (Markov Property). Έτσι, αν S_t η κατάσταση που βρίσκεται ένα σύστημα τη χρονική στιγμή t, και $\mathbb{P}[S_2|S_1]$ η πιθανότητα μετάβασης από την κατάσταση 1 στην 2 ισχύει ότι:

Μια κατάσταση S_t είναι Μαρκοβιανή αν και μόνο αν:

$$\mathbb{P}[S_{t+1}|S_1, ..., S_t] = \mathbb{P}[S_{t+1}|S_t] \quad [2]$$

Η ιδιότητα αυτή εκφράζει ότι η πιθανότητα μετάβασης του συστήματος σε μια κατάσταση, όταν βρίσκεται σε μια Μαρκοβιανή, είναι ανεξάρτητη των προηγούμενων καταστάσεων στις οποίες αυτό βρισκόταν στο παρελθόν, ότι δηλαδή η πληροφορία της τρέχουσας κατάστασης αρκεί για το μέλλον [2].

Με βάση τα παραπάνω, για μια Μαρκοβιανή κατάσταση s και την διαδοχική κατάσταση s' ορίζεται:

- η πιθανότητα μετάβασης κατάστασης $\mathcal{P}_{ss'} = \mathbb{P}[S_{t+1} = s'|S_t = s]$,
- ο πίνακας μετάβασης κατάστασης $\mathcal{P} = \begin{bmatrix} \mathcal{P}_{11} & \cdots & \mathcal{P}_{1n} \\ \vdots & \ddots & \vdots \\ \mathcal{P}_{n1} & \cdots & \mathcal{P}_{nn} \end{bmatrix}$ [2].

2.1.2 Μαρκοβιανές Διαδικασίες

Μια Μαρκοβιανή Διαδικασία (Markov Process) ή Μαρκοβιανή Αλυσίδα (Markov Chain) είναι μια σειρά τυχαίων καταστάσεων με τη Μαρκοβιανή Ιδιότητα.
Αν $\mathcal{S}$ το (πεπερασμένο) σύνολο αυτών των καταστάσεων και $\mathcal{P}$ ο αντίστοιχος πίνακας μετάβασης κατάστασης, τότε η Μαρκοβιανή Διαδικασία είναι η πλειάδα (tuple) $< \mathcal{S}, \mathcal{P} >$ [2].

Η Μαρκοβιανή Ιδιότητα και οι Μαρκοβιανές Διαδικασίες αποτελούν τη βάση αυτών που θα συζητηθούν παρακάτω στην παρούσα ενότητα.

2.1.3 Μαρκοβιανές Διαδικασίες με Ανταμοιβή (MRPs)

Με την προσθήκη της έννοιας της αξίας, μια Μαρκοβιανή Διαδικασία $< \mathcal{S}, \mathcal{P} >$ γίνεται Μαρκοβιανή Διαδικασία με Ανταμοιβή (Markov Reward Process - MRP) $< \mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma >$, όπου $\mathcal{R}$ μια συνάρτηση ανταμοιβής, με $\mathcal{R}_s = \mathbb{E}[R_{t+1}|S_t = s]$, και $\gamma \in [0, 1]$ ένας συντελεστής μείωσης. Γίνεται κατανοητό, λοιπόν, ότι σε κάθε κατάσταση υπάρχει μια ανταμοιβή [2].

Η χρήση του συντελεστή γ (discount factor) θα εξηγηθεί μέσα από τον ορισμό της απόδοσης (return). Η απόδοση G ορίζεται από τη σχέση:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{2.1}$$

και εκφράζει τη συνολική ανταμοιβή που θα λάβει το σύστημα από τη χρονική στιγμή t και μετά (όπου $R_{t+1}, R_{t+2}, R_{t+3}, ...$ η σειρά επερχόμενων ανταμοιβών για τον πράκτορα) [2].

Ο συντελεστής γ, λοιπόν, καθορίζει πόσο θα συμβάλλουν οι μελλοντικές ανταμοιβές στην απόδοση που "βλέπει" το σύστημα τη χρονική στιγμή t. Στις ακραίες περιπτώσεις όπου:

- $\gamma \to 0$ έχουμε μυωπική αξιολόγηση, δηλαδή λαμβάνεται υπ' όψιν μόνο η παρούσα ανταμοιβή, χωρίς να υπάρχει κάποια εικόνα για το μέλλον,
- $\gamma \to 1$ έχουμε διορατική αξιολόγηση, δηλαδή λαμβάνονται υπ' όψιν σε πολύ μεγάλο βαθμό και οι μελλοντικές ανταμοιβές.

Με τη χρήση του συντελεστή γ ελέγχεται η συμβολή των μελλοντικών ανταμοιβών, προσμετράται η αβεβαιότητα για το μέλλον και αποφεύγεται ο απειρισμός της απόδοσης λόγω πιθανών κύκλων στη Μαρκοβιανή Αλυσίδα [1, 2].

Μπορούμε να ορίσουμε την απόδοση (return) σε μια χρονική στιγμή t βάσει της απόδοσης την επόμενη χρονική στιγμή ως εξής [1, 2]:

$$G_t = R_{t+1}+\gamma R_{t+2}+\gamma^2 R_{t+3}+\gamma^3 R_{t+4}+\ldots = R_{t+1}+\gamma(R_{t+2}+\gamma R_{t+3}+\gamma^2 R_{t+4}+\ldots) = R_{t+1}+\gamma G_{t+1} \quad (2.2)$$

Τον ορισμό της απόδοσης μπορεί να ακολουθήσει αυτός της Συνάρτησης Αξίας. Η Συνάρτηση Αξίας (Value Function) $v(s)$ εκφράζει την αναμενόμενη απόδοση ξεκινώντας από την κατάσταση s, μέσω της μαθηματικής σχέσης:

$$v(s) = \mathbb{E}[G_t|S_t = s]$$

είναι δηλαδή μια συνάρτηση εκτίμησης της αξίας κάθε κατάστασης [1, 2]. Με τη βοήθεια της εξ. (2.2) έχουμε:

$$v(s) = \mathbb{E}[R_{t+1} + \gamma G_{t+1}|S_t = s] = \mathbb{E}[R_{t+1} + \gamma v(t+1)|S_t = s] = \mathcal{R}_s + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'} v(s') \quad (2.3)$$

Η παραπάνω σχέση ονομάζεται Εξίσωση Bellman για MRPs [2], και θα αναλυθεί παρακάτω.

2.1.4 Από τις MRPs στις Μαρκοβιανές Διαδικασίες Λήψης Αποφάσεων (MDPs)

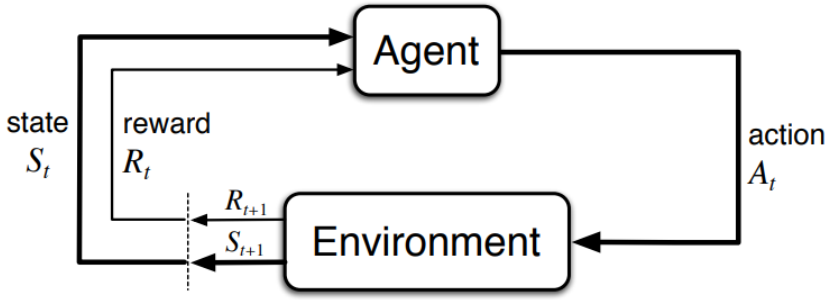
Αν σε μια MRP $< \mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma >$ προστεθεί η δυνατότητα αποφάσεων, τότε δημιουργείται μια Μαρκοβιανή Διαδικασία λήψης Αποφάσεων (Markov Decision Process - MDP) $< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma >$, όπου $\mathcal{A}$ ένα πεπερασμένο σύνολο δράσεων Α.
Η κάθε πιθανότητα μετάβασης κατάστασης πλέον ενσωματώνει τη δυνατότητα αποφάσεων ως εξής:

$$\mathcal{P}^a_{ss'} = \mathbb{P}[S_{t+1} = s'|S_t = s, A_t = a] = \sum_{r \in \mathcal{R}} \mathbb{P}[S_{t+1} = s', R_{t+1} = r|S_t = s, A_t = a] \quad (2.4)$$

όπως και η συνάρτηση ανταμοιβής:

$$\mathcal{R}^a_s = \mathbb{E}[R_{t+1}|S_t = s, A_t = a] = \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} \mathbb{P}[S_{t+1} = s', R_{t+1} = r|S_t = s, A_t = a] \quad (2.5)$$

Έτσι, έχουμε πλέον έναν πράκτορα ο οποίος μπορεί να δρα, και αλληλεπιδρά με το περιβάλλον του μέσω της ανταμοιβής [1, 2].



Σχήμα 2.1: *Η αλληλεπίδραση μεταξύ του πράκτορα και του περιβάλλοντος (από [1])*

Ο πράκτορας πρέπει να έχει ένα συγκεκριμένο τρόπο αποφάσεων, μια συμπεριφορά δράσης. Ορίζεται λοιπόν η πολιτική (policy) π ως εξής:

$$\pi(a|s) = \mathbb{P}[A_t = a|S_t = s]$$

εκφράζοντας μια αντιστοίχιση καταστάσεων του περιβάλλοντος με τις δράσεις που πρέπει να λάβει ο πράκτορας όταν βρίσκεται σε αυτές. Η πολιτική εξαρτάται μόνο από την τρέχουσα κατάσταση στην οποία βρίσκεται ο πράκτορας, αγνοώντας το παρελθόν [1, 2].

Με την εισαγωγή της έννοιας της πολιτικής, ορίζονται τα ακόλουθα [1]:

- Η Συνάρτηση Αξίας μιας MDP ως η αναμενόμενη απόδοση ξεκινώντας από την κατάσταση s και ακολουθώντας την πολιτική π:

$$v_\pi(s) = \mathbb{E}_\pi[G_t|S_t = s] \overset{eq.(2.1)}{=} \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s] \tag{2.6}$$

- Η Συνάρτηση Δράσης-Αξίας (Action-Value Function) ως η αναμενόμενη απόδοση ξεκινώντας από την κατάσταση s, επιλέγοντας τη δράση a, και ακολουθώντας την πολιτική π:

$$q_\pi(s,a) = \mathbb{E}_\pi[G_t|S_t = s, A_t = a] \overset{eq.(2.1)}{=} \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s, A_t = a] \tag{2.7}$$

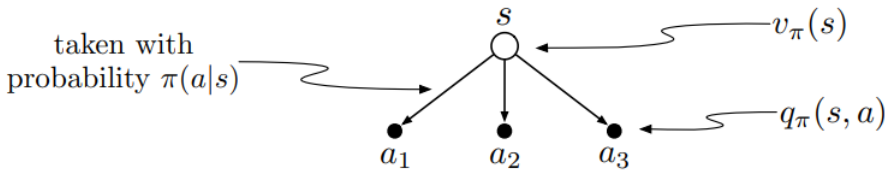
2.1.5 Οι εξισώσεις Bellman

Παρατηρούμε ότι η εξίσωση (2.3) μπορεί να είναι χρήσιμη, καθώς εκφράζει έναν πιο διαχειρίσιμο τρόπο υπολογισμού της Συνάρτησης Αξίας. Μπορεί τώρα να γενικευτεί για τη Συνάρτηση Αξίας και τη Συνάρτηση Δράσης-Αξίας υπό πολιτική π.
Εφόσον η πολιτική $\pi(a|s)$ είναι η πιθανότητα ο πράκτορας να επιλέξει τη δράση a όταν βρίσκεται στην κατάσταση s [1], από ιδιότητες πιθανοτήτων ισχύει:

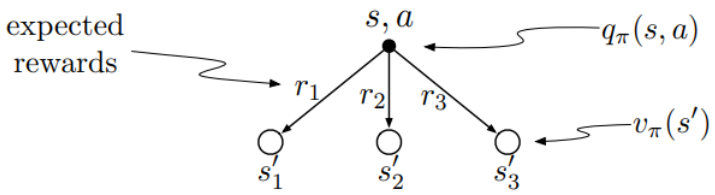
$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) q_\pi(s,a) \tag{2.8}$$

Από την άλλη, η q_π μπορεί να γραφεί ως συνάρτηση της v_π ως εξής [2]:

$$q_\pi(s,a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_\pi(s') \tag{2.9}$$



(α') v_π ως προς q_π



(β') q_π ως προς v_π

Σχήμα 2.2: *Διαγράμματα της σχέσης μεταξύ των v_π, q_π (από [1])*

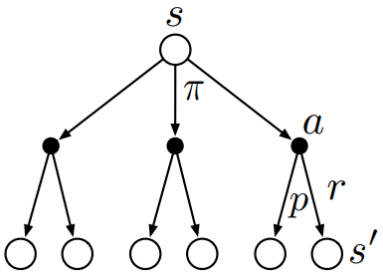
Συνδυάζοντας τις σχέσεις (2.8), (2.9) προκύπτουν οι εξισώσεις Bellman για τις συναρτήσεις v_π, q_π:

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s)(\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_\pi(s')) \tag{2.10}$$

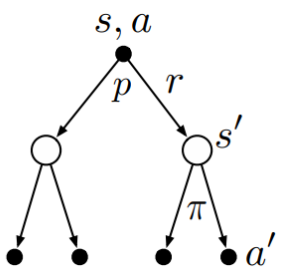
$$q_\pi(s,a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a \sum_{a' \in \mathcal{A}} \pi(a'|s') q_\pi(s',a') \tag{2.11}$$

όπου τα $\mathcal{P}_{ss'}^a, \mathcal{R}_s^a$ ορίζονται στις σχέσεις (2.4), (2.5) αντίστοιχα [2].

Αυτή η μαθηματική σύνδεση μεταξύ της Αξίας v_π (και αντίστοιχα της q_π) μιας κατάστασης με τις υπόλοιπες είναι πολύ σημαντική, καθώς αποτελεί τη βάση πολλών αλγορίθμων υπολογισμού και προσέγγισής της. Καθώς η απ' ευθείας επίλυση για το υπολογισμό αυτό είναι μη-αποδοτική σε πολυπλοκότητα [2], προτιμούνται επαναληπτικοί αλγόριθμοι.



(α') v_π ως προς q_π



(β') q_π ως προς v_π

Σχήμα 2.3: *Διαγράμματα των εξισώσεων Bellman (επέκταση των διαγραμμάτων του σχήματος 2.2) (από [1])*

2.1.6 Βέλτιστες συναρτήσεις και πολιτικές

Όπως προαναφέρθηκε, ένας πράκτορας ψάχνει μια πολιτική η οποία θα του αποφέρει τη μέγιστη συνολική ανταμοιβή. Αυτή η πολιτική λέγεται βέλτιστη. Πρώτα όμως πρέπει να γίνει κατανοητός ο τρόπος με τον οποίο συγκρίνονται δύο πολιτικές. Μια πολιτική π είναι καλύτερη ή ίση με μια άλλη π′ αν και μόνο αν η αναμενόμενη απόδοση (δηλαδή η αξία) της π είναι μεγαλύτερη ή ίση από της π′ για όλες τις καταστάσεις [1], ή αλλιώς:

$$\pi \geq \pi' \iff v_\pi(s) \geq v_{\pi'}(s) \forall s \in \mathcal{S}$$

Με βάση το παραπάνω, ισχύει:

Υπάρχει τουλάχιστον μια πολιτική η οποία είναι καλύτερη ή ίση με όλες τις άλλες. Αυτή ονομάζεται βέλτιστη πολιτική. Οι βέλτιστες πολιτικές σε ένα πρόβλημα Ενισχυτικής Μάθησης συμβολίζονται όλες με π_*, και όλες αντιστοιχούν στην ίδια Συνάρτηση Αξίας:

$$v_*(s) = max_\pi v_\pi(s), \forall s \in \mathcal{S} \text{ που ονομάζεται βέλτιστη Συνάρτηση Αξίας,}$$

και στην ίδια Συνάρτηση Δράσης-Αξίας:

$$q_*(s,a) = max_\pi q_\pi(s,a), \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$$

που ονομάζεται βέλτιστη Συνάρτηση Δράσης-Αξίας [1].

Έτσι, προκύπτουν οι Εξισώσεις Βελτιστοποίησης Bellman:

$$v_*(s) = max_a(\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_*(s')) \tag{2.12}$$

$$q_*(s,a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a max_{a'} q_*(s',a') \tag{2.13}$$

Οι Εξισώσεις Βελτιστοποίησης Bellman είναι μη-γραμμικές, άρα και εδώ προτιμούνται επαναληπτικοί αλγόριθμοι [2].

Με τη γνώση της q_* η σχεδίαση της βέλτιστης πολιτικής μπορεί να γίνει εύκολα με άπληστο (greedy) τρόπο ως εξής:

$$\pi_*(a|s) = \begin{cases} 1 & \text{αν } a = \underset{a \in \mathcal{A}}{argmax} q_*(s,a) \\ 0 & \text{αλλιώς} \end{cases}$$

Αξιοσημείωτο είναι ότι παρ' όλο που οι άπληστες μέθοδοι έχουν από φύση τους βραχυπρόθεσμη λογική, η χρήση της q_* κάνει αυτή τη λογική μακροπρόθεσμη (βάσει του ορισμού (2.7)) [1, 2].

2.2 Δυναμικός Προγραμματισμός

Σε ένα πρόβλημα Ενισχυτικής Μάθησης όπου είναι γνωστό το περιβάλλον με τη μορφή μιας MDP, τη λύση μπορεί να τη δώσουν αλγόριθμοι δυναμικού προγραμματισμού μέσω της χρήσης της Συνάρτησης Αξίας για την αναζήτηση καλών πολιτικών [1].

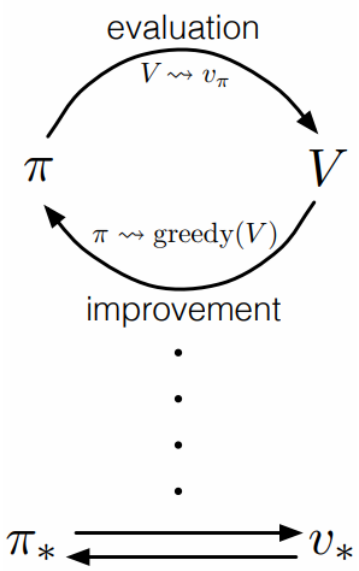
Ο δυναμικός προγραμματισμός (Dynamic Programming - DP) είναι μια αλγοριθμική τεχνική ανάλυσης ενός προβλήματος σε μικρότερα επανεμφανιζόμενα υποπροβλήματα, και αποθήκευσης των επιμέρους λύσεών τους για την επίλυση του αρχικού με αποδοτικό τρόπο. Εφαρμόζεται συνήθως σε προβλήματα βελτιστοποίησης [13].

Οι αλγόριθμοι δυναμικού προγραμματισμού στην Ενισχυτική Μάθηση χρησιμοποιούν τις εξισώσεις Bellman για τις εξής εργασίες:

- Αξιολόγηση Πολιτικής - Πρόβλεψη (Policy Evaluation - Prediction): Υπολογισμός της Συνάρτησης Αξίας v_π για μια τυχαία πολιτική π [1].

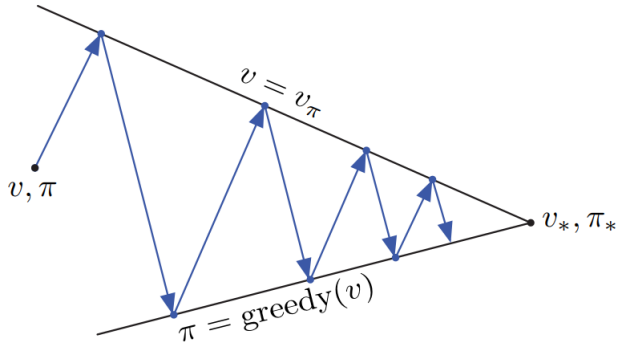
- Βελτίωση Πολιτικής (Policy Improvement): Έχοντας υπολογίσει τη Συνάρτηση Αξίας v_π για μια πολιτική, βελτίωση της πολιτικής αυτής, ώστε να συγκλίνει προς τη βέλτιστη π_*. Γίνεται με άπληστο τρόπο [1].

- Επανάληψη Πολιτικής (Policy Iteration): Μια αλληλουχία αξιολογήσεων και βελτιώσεων πολιτικής, ώστε μετά την βελτίωση της πολιτικής π σε μια καλύτερη π′, να υπολογιστεί η Συνάρτηση Αξίας v'_π, να ακολουθήσει μια εκ νέου βελτίωση κ.ο.κ. Η γενικότερη ιδέα της αλληλουχίας αξιολογήσεων και βελτιώσεων-ανανεώσεων πολιτικής (όχι απαραίτητα με τις συγκεκριμένες μεθόδους του δυναμικού προγραμματισμού που αναφέρθηκαν παραπάνω) ονομάζεται Γενικευμένη Επανάληψη Πολιτικής (Generalized Policy Iteration).



Σχήμα 2.4: *Γενικευμένη Επανάληψη Πολιτικής (από [1])*

Με αυτό τον τρόπο επιτυγχάνεται η σύγκλιση στις π_*, v_* [1].



Σχήμα 2.5: *Σύγκλιση των v, π (Γενικευμένη Επανάληψη Πολιτικής) (από [1])*

- Επανάληψη Αξίας (Value Iteration): Μια διαδικασία με τον ίδιο σκοπό με την επανάληψη πολιτικής. Διαφοροποιείται στο γεγονός ότι δεν εκτελείται κάθε φορά ολόκληρη αξιολόγηση πολιτικής που είναι ένας επαναληπτικός υπολογισμός, αλλά μόνο ένα βήμα αυτής (δηλαδή μόνο μια ενημέρωση κάθε κατάστασης). Αποδεικνύεται ότι και πάλι η σύγκλιση συμβαίνει επιτυχώς [1].

Στην παρούσα εργασία δεν είναι γνωστή η δυναμική του μοντέλου, και επομένως δε μπορούν να εφαρμοστούν αλγόριθμοι Δυναμικού Προγραμματισμού. Γι' αυτό το λόγο δεν θα αναλυθεί περαιτέρω.

2.3 Πρόβλεψη και Έλεγχος Αγνώστου Μοντέλου

Όταν δεν υπάρχει πλήρης γνώση του περιβάλλοντος, δηλαδή της MDPs που το περιγράφει, τότε, αντί για δυναμικό προγραμματισμό, χρησιμοποιούνται μέθοδοι μάθησης για την εκτίμηση των συναρτήσεων αξίας (πρόβλεψη - prediction) και τον εντοπισμό βέλτιστων πολιτικών (έλεγχος - control). Η διαδκασία της μάθησης γίνεται μέσω της εμπειρίας, δηλαδή μέσω μιας σειράς δειγμάτων από καταστάσεις, δράσεις και ανταμοιβές από το περιβάλλον, και μπορεί να χωρίζεται σε επεισόδια (σε κάθε επεισόδιο αυτή η δειγματοληψία ξεκινάει από την αρχή). Αυτού του είδους η Ενισχυτική Μάθηση ονομάζεται αγνώστου μοντέλου (model-free), και είναι ικανή να επιτύχει τη βέλτιστη συμπεριφορά [1].

2.3.1 Είδη Ενισχυτικής Μάθησης Αγνώστου Μοντέλου

Οι μέθοδοι Ενισχυτικής Μάθησης αγνώστου μοντέλου μπορούν να ταξινομηθούν με βάση διάφορα κριτήρια:

- Με βάση το σημείο που γίνεται η ανανέωση, η μάθηση μπορεί να γίνεται online ή offline. Στην online μάθηση η εκτίμηση της συνάρτησης αξίας και η βελτίωση της πολιτικής εκτελείται σε κάθε βήμα κάθε επεισοδίου. Από την άλλη, στην offline μάθηση οι διαδικασίες αυτές εκτελούνται μόνο στο τέλος κάθε επεισοδίου [1].

- Με βάση την πολιτική που ανανεώνεται, ο έλεγχος μπορεί να είναι πάνω στην πολιτική (on-policy) ή εκτός πολιτικής (off-policy). Στον έλεγχο πάνω στην πολιτική, βελτιώνεται η πολιτική η οποία ακολουθείται κατά την απόκτηση εμπειρίας, ενώ στον έλεγχο

εκτός πολιτικής οι δύο πολιτικές αυτές είναι διαφορετικές. Στην παρούσα εργασία εφαρμόστηκε έλεγχος πάνω στην πολιτική, οπότε παρακάτω θα αναλυθεί μόνο αυτή η μορφή του [1].

Ένας ακόμα διαχωρισμός είναι αυτός ανάμεσα στις μεθόδους Monte Carlo και Temporal-Difference.

2.3.2 Μέθοδος Monte Carlo

Η μέθοδος Monte Carlo βασίζεται στην εύρεση του μέσου όρου των αποδόσεων των δειγμάτων. Αφού για τον υπολογισμό των αποδόσεων χρειάζεται η ολοκλήρωση του εκάστοτε επεισοδίου, η εκτίμηση της συνάρτησης αξίας και η βελτίωση της πολιτικής συμβαίνουν στο τέλος αυτού. Η μέθοδος αυτή, λοιπόν, μπορεί να είναι μόνο offline.

Αναλυτικότερα, για την πρόβλεψη με Monte Carlo, η αξία κάθε κατάστασης υπολογίζεται ως ο μέσος όρος των αποδόσεων (παρελθοντικών και της πιο πρόσφατης) που παρατηρούνται (μέσω της εμπειρίας, σε κάθε επεισόδιο) μετά τις επισκέψεις σε αυτή την κατάσταση (και ακολουθώντας την πολιτική που έχουμε). Τελικά, αυτός ο μέσος όρος συγκλίνει στην αναμενόμενη τιμή. Μια πιο γενική μορφή αυτής της ιδέας εκφράζεται μαθηματικά στην παρακάτω διαδικασία που συμβαίνει στο τέλος κάθε επεισοδίου:

$$V(S_t) \leftarrow V(S_t) + a(G_t - V(S_t)) \tag{2.14}$$

όπου το a αποτελεί μια πιο ελεύθερη μετάφραση του αντιστρόφου αριθμού του πλήθους των αποδόσεων των οποίων το μέσο όρο θέλουμε να υπολογίσουμε. Έτσι, αντί να διαιρούμε με αυτό το πλήθος για να βρεθεί ο μέσος όρος, έχουμε τον αριθμό a ο οποίος ονομάζεται ρυθμός μάθησης (learning rate), μέσω του οποίου καθορίζεται πόση βαρύτητα έχει η παλαιότερη εμπειρία στην εκτίμηση της αξίας [1, 2].

Σχήμα 2.6: *Διάγραμμα Monte Carlo (από [1])*

Για τον έλεγχο (πάνω στην πολιτική), χρησιμοποιείται η λογική της Γενικευμένης Επανάληψης Πολιτικής (βλ. ενότητα 2.2). Όταν όμως δεν υπάρχει πλήρης γνώση του μοντέλου, η συνάρτηση αξίας δεν αρκεί για να καθορίσει μια πολιτική. Για αυτόν το λόγο, κατά το βήμα της πρόβλεψης γίνεται εκτίμηση της q_π αντί για την v_π, προκειμένου τελικά να συγκλίνει στην q_*. Υπολογίζεται ως ο (με την ευρύτερη έννοια) μέσος όρος των αποδόσεων που

παρατηρούνται μετά τις επισκέψεις στην κατάσταση S και παίρνοντας τη δράση A (και στη συνέχεια ακολουθώντας την πολιτική που έχουμε):

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + a(G_t - Q(S_t, A_t)) \tag{2.15}$$

είναι δηλαδή η ανάλογη διαδικασία της (2.14). Έτσι, η άπληστη επιλογή πολιτικής γίνεται με τον εξής τρόπο:

$$\pi'(s) = \underset{a \in \mathcal{A}}{argmax} Q(s, a)$$

Επειδή όμως γίνεται απλή δειγματοληψία καταστάσεων και δράσεων, υπάρχει ο κίνδυνος μέσω της άπληστης ντετερμινιστικής πολιτικής κάποια ζευγάρια καταστάσεων και δράσεων να μην προσπελαστούν ποτέ, και έτσι να γίνεται σύγκλιση σε κάποια υπο-βέλτιστη πολιτική, επειδή θα επιλέγονται πάντα οι συγκεκριμένες δράσεις που υπαγορεύονται από αυτή. Τη λύση σε αυτό το πρόβλημα δίνει η εξερεύνηση (exploration), δηλαδή η προσπέλαση και άλλων ζευγαριών καταστάσεων-δράσεων για να εξεταστεί αν πιθανόν είναι καλύτερα από αυτά της παρούσας πολιτικής που ακολουθείται.

Η πιο απλή υλοποίηση της ιδέας αυτής είναι η ε-greedy εξερεύνηση, κατά την οποία υπάρχει πιθανότητα ε να επιλεγεί μια δράση τυχαία. Έτσι, αν m το πλήθος των δράσεων που μπορούν να επιλεγούν, η πολιτική παίρνει την ακόλουθη μορφή:

$$\pi(a|s) = \begin{cases} \frac{\epsilon}{m} + 1 - \epsilon & \text{αν } a* = \underset{a \in \mathcal{A}}{argmax} Q(s, a) \\ \frac{\epsilon}{m} & \text{αλλιώς} \end{cases} \tag{2.16}$$

και ο έλεγχος πλέον ενσωματώνει την ανάλογη διαδικασία της ε-greedy βελτίωσης πολιτικής.

Είναι απαραίτητο να υπάρχει μια ισορροπία μεταξύ της εκμετάλλευσης (exploitation), δηλαδή της εύρεσης των καλύτερων αποφάσεων βάσει της εμπειρίας που ήδη έχουμε, και της εξερεύνησης (exploration), δηλαδή της αναζήτησης νέων πληροφοριών. Ένας συνήθης τρόπος να γίνει αυτό είναι η σταδιακή μείωση της πιθανότητας ε κατά την εκπαίδευση (π.χ. GLIE Monte Carlo Control).

Τέλος, όπως και στο Δυναμικό Προγραμματισμό, έτσι και εδώ κατά την πρόβλεψη δεν χρειάζεται να γίνεται επαναληπτικός υπολογισμός για την προσέγγιση της q_π, αλλά η διαδικασία μπορεί να σταματάει μετά από ένα επεισόδιο για να ακολουθήσει η βελτίωση της πολιτικής. Αποδεικνύεται ότι και πάλι η σύγκλιση στις βέλτιστες q_*, π_* συμβαίνει επιτυχώς [1, 2].

2.3.3 Μάθηση Temporal-Difference

Η μάθηση Temporal-Difference συνδυάζει τις ιδέες του Monte Carlo και του δυναμικού προγραμματισμού. Συγκεκριμένα, όπως η μέθοδος Monte Carlo, επιτρέπει τη μάθηση μέσω εμπειρίας, χωρίς πλήρη γνώση του περιβάλλοντος. Επιπλέον, όπως ο δυναμικός προγραμματισμός, βασίζει τις εκτιμήσεις της σε προηγούμενες εκτιμήσεις, χωρίς να χρειάζεται να περιμένει το τέλος των επεισοδίων (bootstrapping).

Παρακάτω θα παραθέσουμε την ιδέα μέσω της οποίας επιτυγχάνεται το bootstrapping.

Αρχικά, ισχύει ότι:

$$v_\pi(s) = \mathbb{E}_\pi[G_t|S_t = s] \overset{eq.(2.2)}{=} \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s] = \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s]$$

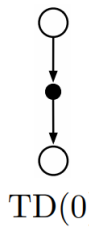
Με βάση την παραπάνω ιδιότητα, η πρόβλεψη, ανάλογα με την (2.14), μπορεί να γίνει ως εξής:

$$V(S_t) \leftarrow V(S_t) + a[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)] \tag{2.17}$$

Με αυτό τον τρόπο, η εκτίμηση της συνάρτησης αξίας χρειάζεται να περιμένει μόνο μέχρι το επόμενο βήμα (για την προσπέλαση της S_{t+1}) και όχι μέχρι το τέλος του επεισοδίου (αν υπάρχει), δηλαδή είναι εφικτή η online μάθηση. Το μειονέκτημα της μεθόδου αυτής, είναι ότι εισάγεται bias λόγω του bootstrapping.

Η ποσότητα $\delta_t = R_{t+1} + \gamma v_\pi(S_{t+1}) - V(S_t)$ είναι ένα σφάλμα που μετρά τη διαφορά μεταξύ της εκτιμώμενης $V(S_t)$ και της καλύτερης εκτίμησης $R_{t+1} + \gamma v_\pi(S_{t+1})$, και ονομάζεται TD σφάλμα (TD error).

Συγκεκριμένα, η παραπάνω μέθοδος είναι μια υποπερίπτωση μάθησης TD ονομάζεται TD(0) ή TD ενός βήματος (one-step TD), διότι βασίζεται στην εκτίμηση μόνο της επόμενης κατάστασης. Είναι υποκατηγορία των μεθόδων n-step TD και TD(λ) [1, 2].



Σχήμα 2.7: *Διάγραμμα TD(0) (από [1])*

Ο έλεγχος (πάνω στην πολιτική) με μάθηση TD(0) γίνεται όπως και στη μέθοδο Monte Carlo, δηλαδή γενικευμένη επανάληψη πολιτικής (χωρίς ακριβή σύγκλιση της Q στην q_π σε κάθε βήμα) με χρήση της συνάρτησης δράσης-αξίας, και με την εισαγωγή εξερεύνησης. Όμως, κατά την πρόβλεψη, με βάση την εξ. (2.17), έχουμε τη διαδικασία ανανέωσης:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + a[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)] \tag{2.18}$$

Σημειώνουμε εδώ ότι αν η κατάσταση S_{t+1} είναι τερματική, τότε η $Q(S_{t+1}, A_{t+1})$ ορίζεται ως 0 [1, 2].

Ο αλγόριθμος ελέγχου με μάθηση TD(0) ονομάζεται SARSA και θα αναφερθεί αναλυτικά στο επόμενο κεφάλαιο.

2.4 Προσέγγιση Συνάρτησης

Σε πολλά προβλήματα Ενισχυτικής Μάθησης ο χώρος των καταστάσεων ή/και των δράσεων έχει πολύ μεγάλες διαστάσεις, ή είναι συνεχής. Σε αυτές τις περιπτώσεις δε μπορούν να βοηθήσουν αυτούσια όσα παρουσιάζονται παραπάνω, γιατί αναφέρονται σε διακριτούς

χώρους οι οποίοι πρέπει να εξερευνηθούν, κάτι εξαιρετικά χρονοβόρο και κοστοβόρο, και μάλιστα αδύνατο στην περίπτωση συνεχούς χώρου.

Τη λύση σε αυτό το πρόβλημα δίνει η προσέγγιση συνάρτησης (function approximation), όπου η συνάρτηση αξίας αντί να αναπαρασταθεί με έναν πίνακα και να παίρνει διακριτές τιμές για κάθε κατάσταση, είναι μια συνάρτηση $\hat{v}(s, \mathbf{w}) \approx v_\pi(s)$, όπου $\mathbf{w}$ ένα διάνυσμα βαρών (παραμέτρων). Έτσι, μια ανανέωση της συνάρτησης επηρεάζει πολλαπλές καταστάσεις. Η $\hat{v}$ μπορεί να μοντελοποιηθεί με διάφορους τρόπους, από γραμμικός συνδυασμός χαρακτηριστικών μιας κατάστασης μέχρι βαθιά νευρωνικά δίκτυα και δέντρα αποφάσεων.

Ανάλογα μπορεί να προσεγγιστεί και η βέλτιστη συνάρτηση δράσης-αξίας ως $\hat{q}(s, a, \mathbf{w}) \approx q_*(s, a)$ [1].

2.4.1 Μέθοδοι Gradient-Descent

Οι μέθοδοι Gradient-Descent είναι δημοφιλείς μέθοδοι μάθησης για να επιτευθεί η προσέγγιση συνάρτησης. Σε αυτές, θεωρούμε ότι το διάνυσμα βαρών είναι ένα διάνυσμα στήλη $\mathbf{w} = [w_1, w_2, ..., w_d]^T$, με στοιχεία παραμέτρους που παίρνουν πραγματικές αριθμητικές τιμές, και ότι η $\hat{v}(s, \mathbf{w})$ είναι παραγωγίσιμη ως προς $\mathbf{w}$ για κάθε $s \in \mathcal{S}$.

Θέλουμε να ελαχιστοποιήσουμε το ελάχιστο τετραγωνικό σφάλμα μεταξύ της προσέγγισης $\hat{v}$ και της πραγματικής συνάρτησης αξίας $v_\pi(s)$. Αυτό επιτυγχάνεται αν σε κάθε διακριτό βήμα της εκπαίδευσης ανανεώνουμε τον διάνυσμα των βαρών ως εξής:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{1}{2} a \nabla [v_\pi(S_t) - \hat{v}(S_t, \mathbf{w}_t)]^2 = \mathbf{w}_t + a[v_\pi(S_t) - \hat{v}(S_t, \mathbf{w}_t)] \nabla \hat{v}(S_t, \mathbf{w}_t) \tag{2.19}$$

όπου a ο ρυθμός μάθησης (ρυθμός κλίσης) και, για μια συνάρτηση $f(\mathbf{w})$,
$\nabla f(\mathbf{w}) = [\frac{\partial f(\mathbf{w})}{\partial w_1}, \frac{\partial f(\mathbf{w})}{\partial w_2}, ..., \frac{\partial f(\mathbf{w})}{\partial w_3}]^T$.

Στα προβλήματα πρόβλεψης με άγνωστο περιβάλλον όμως η πραγματική συνάρτηση αξίας $v_\pi(S_t)$ είναι άγνωστη. Αντικαθίσταται οπότε από ένα στόχο (target), ο οποίος εξαρτάται από τον αλγόριθμο που χρησιμοποιείται. Για παράδειγμα:

- Σαν επέκταση της μεθόδου Monte Carlo (2.3.2), ορίζοντας ως στόχο την απόδοση G_t, έχουμε την gradient-descent μορφή της μεθόδου αυτής. Έτσι, η εξ. (2.19) γράφεται:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + a[G_t - \hat{v}(S_t, \mathbf{w}_t)] \nabla \hat{v}(S_t, \mathbf{w}_t) \tag{2.20}$$

- Σαν επέκταση της μάθησης TD(0), ορίζοντας ως στόχο την παράσταση $R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathbf{w}_t)$, έχουμε τη μέθοδο semi-gradient TD(0) που χρησιμοποιεί bootstrapping. Έτσι, η εξ. (2.19) γράφεται:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + a[R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathbf{w}_t) - \hat{v}(S_t, \mathbf{w}_t)] \nabla \hat{v}(S_t, \mathbf{w}_t) \tag{2.21}$$

Μέσω της ανανέωσης του διανύσματος βαρών μπορεί να επιτευχθεί τελικά προσέγγιση της v_π, η οποία προσέγγιση αποτελεί τη διαδικασία πρόβλεψης.

Για διαδικασία του ελέγχου (πάνω στην πολιτική) με τη μέθοδο TD(0) χρησιμοποιώντας προσέγγιση συνάρτησης (που αποτελεί την κύρια μέθοδο που χρησιμοποιήσαμε στην εργασία), αυτό που αλλάζει σε σχέση με όσα παρουσιάστηκαν στην υποενότητα 2.3.3 είναι η

αναπαράσταση της συνάρτησης δράσης-αξίας ως $\hat{q}(s, a, \mathbf{w}) \approx q_\pi(s, a)$, μαζί με ότι αυτό συνεπάγεται. Κατά την πρόβλεψη, λοιπόν, η εξίσωση ανανέωσης των βαρών (ανάλογη της (2.21)) είναι:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + a[R_{t+1} + \gamma\hat{q}(S_{t+1}, A_{t+1}, \mathbf{w}_t) - \hat{q}(S_t, A_t, \mathbf{w}_t)]\nabla\hat{q}(S_t, A_t, \mathbf{w}_t) \tag{2.22}$$

Η επιλογή δράσεων (πολιτική) γίνεται και εδώ με συνδυασμό του άπληστου τρόπου και εξερεύνησης (π.χ. ε-greedy πολιτική). Ως άπληστη επιλογή θεωρείται αυτή που μεγιστοποιεί την προσεγγιστική συνάρτηση $\hat{q}$ στην εκάστοτε κατάσταση ($A^*_{t+1} = argmax_a \hat{q}(S_{t+1}, a, \mathbf{w}_t)$) [1].

Σημειώνουμε εδώ ότι δεν επιτυγχάνεται σύγκλιση για κάθε μέθοδο ελέγχου με κάθε είδος μοντελοποίησης της προσέγγισης [1].

2.4.2 Γραμμικές Μέθοδοι

Οι γραμμικές μέθοδοι (linear methods) χρησιμοποιούνται ως προσεγγιστές για τη μοντελοποίηση της $\hat{v}$ (και ανάλογα της $\hat{q}$).

Αρχικά, θεωρούμε ένα διάνυσμα πραγματικών τιμών $\mathbf{x}(s) = [x_1(s), x_2(s), ..., x_d(s)]^T$ με στοιχεία τιμές συναρτήσεων $x_i : \mathcal{S} \to \mathbb{R}$, οι οποίες τιμές αποτελούν χαρακτηριστικά μιας κατάστασης $s \in \mathcal{S}$. Το πλήθος των στοιχείων είναι ίσο με αυτό των στοιχείων του διανύσματος βαρών $\mathbf{w}$. Το διάνυσμα $\mathbf{x}$ ονομάζεται διάνυσμα χαρακτηριστικών (feature vector) και στην ουσία αναπαριστά μια κατάσταση s. Αντίστοιχα για την προσέγγιση της $\hat{q}$ ορίζεται ανάλογα το διάνυσμα χαρακτηριστικών $\mathbf{x}(s, a) = [x_1(s, a), x_2(s, a), ..., x_d(s, a)]^T$ που αναπαριστά ζευγάρια καταστάσεων-δράσεων (προφανώς το d θα είναι το πλήθος των πιθανών ζευγαριών).

Σύμφωνα λοιπόν με την προσέγγιση των γραμμικών μεθόδων, η $\hat{v}$ ορίζεται ως γραμμικός συνδυασμός των βαρών, χρησιμοποιώντας το διάνυσμα $\mathbf{x}$, ως εξής:

$$\hat{v}(s, \mathbf{w}) = \mathbf{w}^T\mathbf{x}(s) = \sum_{i=1}^{d} w_i x_i(s)$$

και, ανάλογα, η $\hat{q}$ ως εξής:

$$\hat{q}(s, a, \mathbf{w}) = \mathbf{w}^T\mathbf{x}(s, a) = \sum_{i=1}^{d} w_i x_i(s, a) \tag{2.23}$$

Εφαρμόζοντας τα παραπάνω στις μεθόδους Gradient-Descent, παρατηρούμε ότι $\nabla\hat{v}(s, \mathbf{w}) = \mathbf{x}(s)$. Έτσι, στις εξισώσεις ανανέωσης βαρών (2.20), (2.21) μπορούμε να αντικαταστήσουμε την $\hat{v}(S_t, \mathbf{w}_t)$ με $\mathbf{w}_t^T\mathbf{x}_t$ (και την $\hat{v}(S_{t+1}, \mathbf{w}_t)$ με $\mathbf{w}_t^T\mathbf{x}_{t+1}$), και το $\nabla\hat{v}(S_t, \mathbf{w}_t)$ με $\mathbf{x}_t$.

Για την $\hat{q}$ ισχύουν οι ανάλογες αντικαταστάσεις, τις οποίες και μπορούμε να χρησιμοποιήσουμε στην εξίσωση ανανέωσης βαρών (2.22) [1].

Για την κατασκευή του διανύσματος χαρακτηριστικών $\mathbf{x}$, μπορούν να χρησιμοποιηθούν διάφορες συναρτήσεις, όπως πολυώνυμα, Fourier basis και Radial Basis Functions [1]. Στην παρούσα εργασία θα εφαρμοστούν οι τελευταίες.

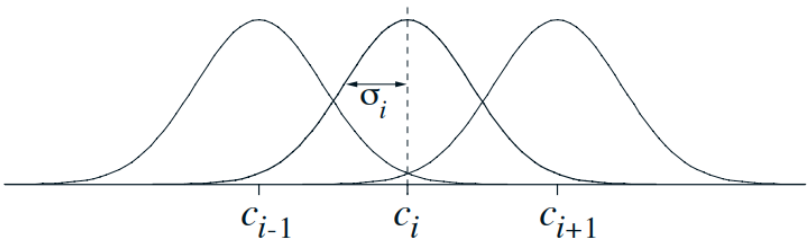
2.4.3 Κατασκευή Χαρακτηριστικών με Radial Basis Functions

Για τις συναρτήσεις $x_i(s)$ μπορούν να χρησιμοποιηθούν οι Radial Basis Functions (RBFs). Σύμφωνα με αυτή την υλοποίηση, κάθε $x_i(s)$ παίρνει τιμές στο διάστημα $[0,1]$ και υπολογίζεται μέσω μιας Γκαουσιανής συνάρτησης (Gaussian function) που εξαρτάται μόνο από την απόσταση μεταξύ της κατάστασης-μεταβλητής s και μιας κεντρικής κατάστασης c_i (μπορεί να υπολογιστεί με όποια μετρική ενδείκνυται κάθε φορά), και από ένα πλάτος σ_i, ως εξής:

$$x_i(s) = exp(-\frac{\|s - c_i\|^2}{2\sigma_i^2}) \tag{2.24}$$

Η κάθε κεντρική κατάσταση c_i λειτουργεί ως μέση τιμή, και το κάθε πλάτος σ_i ως τυπική απόκλιση.

Έτσι, αυτές οι γκαουσιανές συναρτήσεις δημιουργούν ένα συνεχή και παραγωγίσιμο χώρο.



Σχήμα 2.8: *Μονοδιάστατες RBFs (από [1])*

Η γραμμική προσέγγιση συνάρτησης με RBFs ως χαρακτηριστικά ονομάζεται RBF network [1].

Κεφάλαιο 3

Καταγραφή Απαιτήσεων και Επιλογή Αλγορίθμων

Στο προηγούμενο κεφάλαιο έγινε μια επισκόπηση των πιο βασικών εννοιών, θεωρημάτων και μεθόδων στην Ενισχυτική Μάθηση. Υπάρχουν ακόμα πολλές τεχνικές και αλγόριθμοι που εφαρμόζονται σε προβλήματα διαφορετικών φύσεων και απαιτήσεων. Για παράδειγμα, στην Ενισχυτική Μάθηση βάσει πολιτικής (policy-based RL), οι δράσεις επιλέγονται μέσω παραμετρικής πολιτικής χωρίς να χρειάζεται ο πράκτορας να συμβουλευτεί άμεσα τη συνάρτηση αξίας ή τη συνάρτηση δράσης-αξίας (δηλαδή η πολιτική δεν παράγεται άμεσα από αυτές τις συναρτήσεις) [1, 2], σε αντίθεση με την Ενισχυτική Μάθηση βάσει αξίας (value-based RL) στην οποία ανήκουν οι μέθοδοι που αναφέραμε στο προηγούμενο κεφάλαιο (οι μέθοδοι που βασίζουν την πολιτική στη συνάρτηση δράσης-αξίας ονομάζονται επίσης και μέθοδοι δράσης-αξίας [1]). Επιπλέον, στις μεθόδους Δράστη-Κριτή (Actor-Critic Methods) ο κριτής, δηλαδή η συνάρτηση αξίας, βοηθά τον δράστη, δηλαδή την παραμετρική πολιτική, να βελτιώνεται, χωρίς όμως αυτό να προκύπτει αναγκαστικά από τη μεγιστοποίηση της πρώτης [14]. Άλλα παραδείγματα αποτελούν οι μέθοδοι παρτίδας (batch methods) όπου η μάθηση γίνεται μέσω δειγμάτων τα οποία συλλέγονται εκ των προτέρων [15] προσεγγίζοντας τη λογική της Επιβλεπόμενης Μάθησης, καθώς και η βαθιά Ενισχυτική Μάθηση [16, 17].

Μέσα σε μια τόσο μεγάλη ποικιλία μεθόδων, για να βρεθεί ένας αλγόριθμος κατάλληλος για το αντικείμενο της εργασίας, πρέπει πρώτα αυτό να αναλυθεί και να εντοπιστούν οι απαιτήσεις που έχει, συνδέοντάς τες με αντίστοιχες μεθόδους.

3.1 Ανάλυση Προβλήματος και Καταγραφή Απαιτήσεων

Το task που θέλουμε να φέρουμε εις πέρας, όπως αναφέραμε στην εισαγωγή, αφορά τον επιδέξιο χειρισμό μέσω στρέψης και κάμψης ενός αντικειμένου που προσομοιάζει στη δυναμική διαδικασία εκρίζωσης ενός μανιταριού, με διαρκή σκοπό να μην ασκούνται πάνω σε αυτό μεγάλες ροπές (μέσω σύνδεσης με ελατήρια στη βάση του).

Η δυναμική του μανιταριού (δηλαδή το μοντέλο του περιβάλλοντος) είναι άγνωστη, άρα εφαρμόζουμε μάθηση Αγνώστου Μοντέλου.

Επίσης, θέτοντας ως χώρο καταστάσεων τις ροπές που μετρώνται πάνω στην επαφή του ρομποτικού βραχίονα με το αντικείμενο μετασχηματισμένες στους άξονες της στρέψης και της κάμψης, και ως χώρο δράσεων το συνδυασμό στρέψης και κάμψης, καταλαβαίνουμε ότι

αυτοί οι χώροι είναι συνεχείς και απαιτούν προσέγγιση συνάρτησης.

Ένα πολύ σημαντικό χαρακτηριστικό του αλγορίθμου μας πρέπει να είναι η ικανότητα να μαθαίνει γρήγορα να αντιδρά άμεσα στα ερεθίσματα του περιβάλλοντος, δηλαδή στις μεγάλες ροπές. Καταλαβαίνουμε λοιπόν πως η μάθηση είναι θεμιτό να γίνεται online, ώστε να έχουμε άμεση βελτίωση στη λήψη αποφάσεων σε κάθε βήμα.

Ο στόχος που έχουμε κατά τη μάθηση δεν παρουσιάζει πολύπλοκες σχέσεις, και για αυτόν το λόγο προτιμάται ένας σχετικά απλός αλγόριθμος, ο οποίος φυσικά να μπορεί να φέρει ικανοποιητικά αποτελέσματα. Η απλότητα του αλγορίθμου, μεταξύ άλλων, μπορεί να βοηθήσει στην επεξηγησιμότητά του (explainability), δηλαδή στην κατανόηση των αιτιών για τις οποίες ο αλγόριθμος παράγει ένα αποτέλεσμα, κάτι το οποίο είναι βοηθητικό για την επισκόπηση της συμπεριφοράς του και διόρθωση ή τη μορφοποίησή του, αν χρειαστεί. Τέλος, καθώς η εκπαίδευση του ρομπότ είναι μια διαδικασία που απαιτεί ενασχόληση με φυσική διάταξη και μπορεί να δεσμεύσει σημαντικό χρόνο, είναι κρίσιμο ο αλγόριθμος να μπορεί να συγκλίνει γρήγορα επιτυχώς, χωρίς την ανάγκη μεγάλου αριθμού επεισοδίων. Η χρήση περίπλοκων δομών και μεθόδων όπως Νευρωνικών Δικτύων και Βαθιάς Ενισχυτικής Μάθησης δεν ενδείκνυται, με βάση τις απαιτήσεις αυτές.

Σημειώνουμε εδώ ότι τις ίδιες απαιτήσεις έχουμε και για το περιβάλλον της προσομοίωσης (θα αναφερθούμε αναλυτικά σε αυτή στο επόμενο κεφάλαιο).

3.2 Επιλογή Αλγορίθμων

Εφόσον η συνάρτηση δράσης-αξίας αφορά κυρίως την αύξηση ή τη μείωση των εφαρμοζόμενων ροπών, η μορφή της εκτιμάται εξ αρχής ότι δε θα είναι πολύπλοκη και μπορεί να προσεγγιστεί σχετικά εύκολα. Έτσι, προτιμάται μια μέθοδος δράσης-αξίας. Ενδείκνυται μια διαδικασία βελτίωσης της ίδιας πολιτικής που εφαρμόζεται κατά τη διάρκεια της εκπαίδευσης, άρα η μέθοδος αυτή είναι πάνω στην πολιτική.

Ένας σχετικά απλός αλλά ταυτόχρονα δημοφιλής αλγόριθμος ο οποίος συνδυάζει όλα τα χαρακτηριστικά τα οποία αναφέρθηκαν σε αυτή και την προηγούμενη ενότητα είναι ο Linear Semi-gradient SARSA, ο οποίος αποτελεί μια βασική υλοποίηση TD(0) ελέγχου χρησιμοποιώντας προσέγγιση συνάρτησης με γραμμικές μεθόδους (η γραμμική προσέγγιση εξασφαλίζει τη σύγκλιση [1], όπως θα αναφέρουμε παρακάτω). Προτιμούμε μάθηση TD(0) από οποιαδήποτε άλλη TD μέθοδο, αφού μας ενδιαφέρει η άμεση online μάθηση για απόκριση στο εκάστοτε παρόν ερέθισμα.

Πρώτα δοκιμάζεται ο απλός αλγόριθμος SARSA στο περιβάλλον προσομοίωσης, διακριτοποιώντας το χώρο καταστάσεων και δράσεων προσεγγιστικά, για να εξετάσουμε τις δυνατότητες της μεθόδου στο πρόβλημα που έχουμε να επιλύσουμε. Στη συνέχεια, σειρά παίρνει η μετάβαση σε συνεχείς χώρους και η υλοποίηση του Linear Semi-gradient SARSA.

Ο αλγόριθμος SARSA είναι ευρέως χρησιμοποιούμενος σε προβλήματα Ενισχυτικής Μάθησης, όχι μόνο σε καθαρά αλγοριθμικές εργασίες [18], αλλά αποτελεί πολλές φορές και τη βάση νέων μεθόδων στη ρομποτική και όχι μόνο [19, 20].

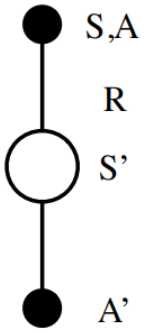
Ένας ακόμα παράγοντας για την επιλογή αλγορίθμου είναι η χρήση του SARSA ως βάση για αλγορίθμους σε προβλήματα όπου υπάρχουν μεγάλες (κατά απόλυτη τιμή) αρ-

νητικές ανταμοιβές, λόγω της μεγαλύτερης αποτελεσματικότητά του έναντι άλλων μεθόδων (Q-learning) [21]. Στο δικό μας task έχουμε τέτοιες ανταμοιβές λόγω της συχνής αύξησης των εφαρμοζόμενων στο αντικείμενο ροπών, κατά την εκπαίδευση, σε μεγαλύτερα από τα επιθυμητά επίπεδα (η συνάρτηση ανταμοιβής παρουσιάζεται αναλυτικά στην ενότητα 4.2).

3.3 Ο αλγόριθμος SARSA (on-policy TD control)

3.3.1 Επισκόπηση του SARSA

Ο αλγόριθμος SARSA είναι μια μέθοδος ελέγχου που χρησιμοποιεί μάθηση TD(0), όπως αναλύθηκε στην υποενότητα 2.3.3. Το όνομά του προκύπτει από το γεγονός ότι σε κάθε βήμα ανανέωσης της συνάρτησης δράσης αξίας χρησιμοποιούνται τα $S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}$ [1].



Σχήμα 3.1: *Διάγραμμα SARSA (από [2])*

Η εκπαίδευση οργανώνεται σε επεισόδια με διακριτά βήματα, και διαρκεί μέχρι η συνάρτηση δράσης-αξίας, και άρα και η πολιτική, να συγκλίνουν ικανοποιητικά.

Αλγοριθμος 3.1: *SARSA (από [1])*

Παράμετροι: ρυθμός μάθησης $a \in (0, 1]$, μικρό $\epsilon > 0$
1: **Αρχικοποίηση:** $Q(s, a), \forall s \in \mathcal{S}^+, a \in \mathcal{A}(s)$ αυθαίρετα, εκτός από $Q(\text{τερματική}, \cdot) = 0$
2: **Βρόχος** για κάθε επεισόδιο:
3: **Αρχικοποίηση** κατάστασης S
4: Διάλεξε δράση A από την S χρησιμοποιώντας πολιτική που παράγεται από την Q (*π.χ. ε-greedy*)
5: **Βρόχος** για κάθε βήμα του επεισοδίου:
6: Πάρε τη δράση A, παρατήρησε την ανταμοιβή R και την επόμενη κατάσταση S'
7: Διάλεξε δράση A' από την S' χρησιμοποιώντας πολιτική που παράγεται από την Q (*π.χ. ε-greedy*)
8: $Q(S, A) \leftarrow Q(S, A) + a[R + \gamma Q(S', A') - Q(S, A)]$
9: $S \leftarrow S', A \leftarrow A'$
10: **μέχρι** S τερματική

Στον ψευδοκώδικα 3.1 παρατηρούμε την οργάνωση του αλγορίθμου. Στη γραμμή 1 του κώδικα αρχικοποιείται η συνάρτηση δράσης αξίας η οποία αναπαρίσταται με έναν πίνακα

διακριτών τιμών για τα διάφορα ζευγάρια καταστάσεων-δράσεων. Στη συνέχεια, εκτελείται μια διεργασία στη λογική της γενικευμένης επανάληψης πολιτικής (βλ. ενοτ. 2.2). Έτσι, σε κάθε επεισόδιο, αφού πρώτα οριστεί η κατάσταση από την οποία ξεκινάει (γραμμή 3) και επιλεγεί η αρχική δράση του πράκτορα (γραμμή 4) με τρόπο που θα περιγράψουμε παρακάτω, εκτελείται μια σειρά ενεργειών σε κάθε βήμα, επαναληπτικά. Συγκεκριμένα, ο πράκτορας παίρνει τη δράση που έχει προεπιλέξει, και παρατηρεί την λαμβανόμενη ανταμοιβή και την επόμενη κατάσταση (γραμμή 6). Στη συνέχεια, επιλέγει την επόμενη δράση βάσει της πολιτικής που έχει, η οποία, όπως ξέρουμε, παράγεται από συνδυασμό εξερεύνησης και μεγιστοποίηση της συνάρτησης δράσης αξίας (γραμμή 7). Έπειτα ακολουθεί ένα βήμα πρόβλεψης στη λογική της μάθησης TD(0) όπως περιγράφεται στην εξίσωση (2.18) (γραμμή 8), δηλαδή η ανανέωση της συνάρτησης δράσης-αξίας Q ώστε να τείνει προς την q_π (όπου π η πολιτική που ακολουθεί ο πράκτορας). Τέλος, η νέα κατάσταση γίνεται πλέον η παρούσα, η νέα δράση γίνεται η προεπιλεγμένη (γραμμή 9), και επανεκκινείται η διαδικασία για το επόμενο βήμα του επεισοδίου.

3.3.2 Σύγκλιση του SARSA

Υπάρχουν δύο προϋποθέσεις ώστε ο SARSA να συγκλίνει με βεβαιότητα σε βέλτιστη πολιτική και συνάρτηση δράσης-αξίας:

- Να γίνει προσπέλαση όλων των ζευγαριών καταστάσεων-δράσεων άπειρες φορές και η πολιτική να συγκλίνει οριακά στην άπληστη. Ένας τρόπος με τον οποίο αυτά μπορούν να επιτευχθούν είναι με ε-greedy πολιτικές μειώνοντας το ε βάσει της σχέσης $\epsilon = \frac{1}{t}$ [1].

- Η ακολουθία του ρυθμού μάθησης ικανοποιεί τους ακόλουθους κανόνες:

$$\sum_{t=1}^{\infty} a_t = \infty,$$

$$\sum_{t=1}^{\infty} a_t^2 < \infty$$

  Ο πρώτος εγγυάται ότι ο συγκεκριμένος ρυθμός μάθησης μπορεί να υπερνικήσει τις όποιες αρχικές συνθήκες ή τυχαίες διακυμάνσεις. Ο δεύτερος εγγυάται ότι ο ρυθμός κάποια στιγμή θα μικρύνει κατάλληλα ώστε να εξασφαλιστεί η σύγκλιση [1].

3.4 Ο αλγόριθμος Episodic Linear Semi-gradient SARSA

3.4.1 Επισκόπηση του Episodic Linear Semi-gradient SARSA

Ο αλγόριθμος Episodic Linear Semi-gradient SARSA είναι μια μέθοδος ελέγχου που αποτελεί επέκταση του semi-gradient TD(0) [1] (βλ. υποενότητα 2.4.1). Είναι ουσιαστικά μια μετάβαση του αλγορίθμου SARSA στο συνεχή χώρο, με τις ανάλογες διαφοροποιήσεις όπου αυτές απαιτούνται. Η συνάρτηση δράσης αξίας προσδιορίζεται πλέον από τα βάρη της, τα οποία αρχικοποιούνται και ανανεώνονται ανάλογα. Η προσέγγιση της συνάρτησης δράσης-αξίας, όπως δηλώνει και το όνομα του αλγορίθμου, γίνεται με γραμμικές μεθόδους,

όπως περιγράφονται στην υποενότητα 2.4.2. Έτσι, το βήμα πρόβλεψης γίνεται με ανανέωση των βαρών στη γραμμή 10 του ψευδοκώδικα, όπως περιγράφεται και στην εξίσωση (2.22) εφαρμόζοντας τις αντικαταστάσεις λόγω των γραμμικών μεθόδων προσέγγισης (βλ. υποενοτ. 2.4.2). Ανανέωση βαρών γίνεται και στη γραμμή 7 του κώδικα, όταν ο πράκτορας παρατηρεί την τερματική κατάσταση, με την επισήμανση ότι η τιμή που παίρνει η συνάρτηση δράσης-αξίας σε αυτή την κατάσταση είναι 0.

Αλγοριθμος 3.2: *Episodic Linear Semi-gradient SARSA (από [1])*

Είσοδος: παραγωγίσιμη παραμετροποίηση συνάρτησης δράσης-αξίας $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Παράμετροι: ρυθμός μάθησης $a > 0$, μικρό $\epsilon > 0$
1: **Αρχικοποίηση** βαρών $\mathbf{w} \in \mathbb{R}^d$ αυθαίρετα (*π.χ.* $\mathbf{w} = \mathbf{0}$)

2: **Βρόχος** για κάθε επεισόδιο:
3: **Αρχικοποίηση** κατάστασης S και δράσης A (*π.χ. ε-greedy επιλογή*)
4: **Βρόχος** για κάθε βήμα του επεισοδίου:
5: Πάρε τη δράση A, παρατήρησε την ανταμοιβή R και την επόμενη κατάσταση S'
6: **Αν** S' τερματική:
7: $\mathbf{w} \leftarrow \mathbf{w} + a[R - \mathbf{w}^T\mathbf{x}(S, A)]\mathbf{x}(S, A)$
8: Πήγαινε στο επόμενο επεισόδιο
9: Διάλεξε δράση A' μέσω της $\mathbf{w}^T\mathbf{x}(S', \cdot)$ (*π.χ. ε-greedy επιλογή*)
10: $\mathbf{w} \leftarrow \mathbf{w} + a[R + \gamma\mathbf{w}^T\mathbf{x}(S', A') - \mathbf{w}^T\mathbf{x}(S, A)]\mathbf{x}(S, A)$
11: $S \leftarrow S', A \leftarrow A'$

3.4.2 Σύγκλιση του Episodic Linear Semi-gradient SARSA

Αποδεικνύεται ότι ο εν λόγω αλγόριθμος ελέγχου πάνω στην πολιτική, ο οποίος χρησιμοποιεί μέθοδο TD(0) και γραμμικές μεθόδους προσέγγισης συνάρτησης, συγκλίνει σε μια υπο-βέλτιστη λύση [1].

Μέρος II

Πειραματικό Μέρος

Κεφάλαιο 4

Περιγραφή Μεθόδου Μάθησης σε Περιβάλλον Προσομοίωσης

Η πρώτη υλοποίηση της εργασίας μας γίνεται σε περιβάλλον προσομοίωσης. Σε αυτό το κεφάλαιο θα το περιγράψουμε και θα αναλύσουμε όλα τα στοιχεία της υλοποίησης των μεθόδων Ενισχυτικής Μάθησης που ακολουθήσαμε.

4.1 Το Περιβάλλον Προσομοίωσης

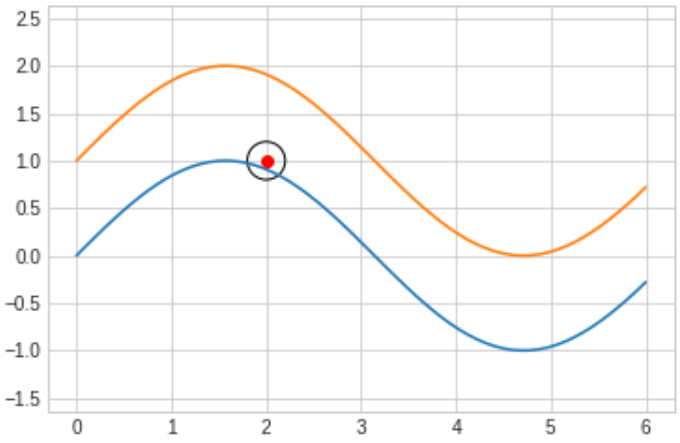
Όπως έχει προαναφερθεί, το αρχικό πρόβλημα του επιδέξιου χειρισμού, μέσω στρέψης και κάμψης, ενός αντικειμένου, με στόχο τον περιορισμό των εφαρμοζόμενων ροπών ανάγεται αρχικά σε ένα αλγοριθμικό πρόβλημα απόδρασης ενός κυκλικού αντικειμένου από έναν διδιάστατο διάδρομο.

4.1.1 Περιγραφή του Περιβάλλοντος

Το περιβάλλον αποτελείται από δύο συνεχείς πεπερασμένους τοίχους οι οποίοι σχηματιζουν ένα διάδρομο ανάμεσά τους. Για απλούστευση του προβλήματος, ορίζουμε για κάθε τοίχο μια τιμή του οριζόντιου άξονα να αντιστοιχεί το πολύ σε μια τιμή του κατακόρυφου άξονα, όπως σε μια συνάρτηση. Επίσης θέτουμε ως έξοδο του διαδρόμου το δεξί άκρο του. Με βάση αυτές τις πληροφορίες, ο γενικός προσανατολισμός που πρέπει να έχει ένα αντικείμενο για να καταφέρει να αποδράσει είναι προς τα δεξιά. Το αντικείμενο-πράκτορας είναι ένας κυκλικός δίσκος ακτίνας 0.2 μονάδων μήκους. Ένα παράδειγμα περιβάλλοντος προσομοίωσης παρουσιάζεται στο σχήμα 4.1. Αν το κυκλικό αντικείμενο διεισδύσει μέσα σε κάποιον από τους δύο τοίχους, αυτός του ασκεί μια δύναμη που τείνει να το επαναφέρει στο διάδρομο. Συγκεκριμένα, η δύναμη που ασκείται στο αντικείμενο προσομοιάζει την δύναμη επαναφοράς ενός ελατηρίου, και είναι της μορφής:

$$\vec{F} = -k \cdot \vec{p}_e$$

όπου $\vec{p}_e$ το διάνυσμα διείσδυσης του αντικειμένου στον εικονικό τοίχο. Η διείσδυση του αντικειμένου μετράται από το σημείο της περιφέρειάς του που έχει τη μεγαλύτερη διείσδυση, έως το πιο κοντινό στο κέντρο του αντικειμένου σημείο του τοίχου.



Σχήμα 4.1: *Παράδειγμα περιβάλλοντος προσομοίωσης*

Θα θέσουμε την παράμετρο του περιβάλλοντος $k = 1$, το οποίο σημαίνει ότι το μέτρο της δύναμης που ασκείται στον πράκτορα ισούται με το μέγεθος της διείσδυσής του σε τοίχο.

4.1.2 Μοντελοποίηση του Περιβάλλοντος με MDP

Το αντικείμενο-πράκτορας δεν έχει κάποια γνώση σχετικά με το διάδρομο, την έξοδό του, και τη μοντελοποίηση των δυνάμεων επαναφοράς, δηλαδή δεν έχει γνώση του περιβάλλοντος.

Η μόνη πληροφορία που έχει κάθε στιγμή είναι η δύναμη που του ασκείται, και συγκεκριμένα η συνιστώσα στον οριζόντιο άξονα F_x και στον κατακόρυφο άξονα F_y. Έτσι, ορίζουμε ως χώρο καταστάσεων την έκφραση της συνισταμένης δύναμης σε πολικές συντεταγμένες ως $[F, \theta_F]$, όπου $F = \sqrt{F_x^2 + F_y^2}$ το μέτρο της ολικής δύναμης, και $\theta_F = atan2(F_y, F_x) \in (-\pi, \pi]$ η γωνία που σχηματίζει το διάνυσμά της με τον θετικό οριζόντιο ημιάξονα. Η προτίμησή μας σε πολικές συντεταγμένες προκύπτει από την επιθυμία μας να διαχωρίσουμε το μέτρο από την κατεύθυνση της συνολικής δύναμης που δέχεται ο πράκτορας.

Ο πράκτορας προσπαθεί με την μετακίνησή του να φτάσει στην έξοδο του διαδρόμου. Η μετακίνηση αυτή αποτελεί σε κάθε βήμα τη δράση του, οπότε αν a_x η συνιστώσα της κίνησης στον οριζόντιο άξονα και a_y στον κατακόρυφο άξονα, θέτουμε ως χώρο δράσεων την ολική μετακίνηση σε πολικές συντεταγμένες ως $[a, \theta_a]$, όπου, ανάλογα με παραπάνω, $a = \sqrt{a_x^2 + a_y^2}$ και $\theta_a = atan2(a_y, a_x) \in (-\pi, \pi]$. Το μέτρο a της κίνησης του πράκτορα σε κάθε βήμα ορίζεται ίσο με 0.1 μονάδες μήκους, έτσι ώστε η διαδικασία της μάθησης να πραγματοποιηθεί ομαλά.

4.1.3 Αντιστοίχιση Προβλήματος Προσομοίωσης με το Πραγματικό Πρόβλημα

Θα κάνουμε ξεκάθαρο τώρα πώς το πρόβλημα στο περιβάλλον προσομοίωσης συνδέεται με αυτό της πραγματικής διάταξης. Όπως το πραγματικό task, έτσι και η προσομοίωση αποτελείται από κίνηση και ανάδραση ερεθισμάτων σε 2 βαθμούς ελευθερίας. Έτσι, οι δύο βαθμοί ελευθερίας στρέψη και κάμψη και οι αντίστοιχες ροπές, αντιπροσωπεύονται στην

προσομοίωση από τις κινήσεις και δυνάμεις στους δύο άξονες, x και y. Ο διάδρομος της προσομοίωσης αντιστοιχίζεται στους συνδυασμούς στρέψης και κάμψης κατά τους οποίους δεν ασκούνται στο αντικείμενο μεγάλες ροπές, οι οποίοι συνδυασμοί μπορούν να περιγραφούν και αυτοί από έναν νοητό διάδρομο. Επιπλέον, η δύναμη επαναφοράς που δέχεται ο πράκτορας στην προσομοίωση από λάθος συνδυασμό κίνησης στους δύο άξονες, ο οποίος τον φέρει εκτός διαδρόμου, μπορεί να αντιστοιχιστεί με την ανεπιθύμητη δύναμη επαναφοράς που ασκείται από τα ελατήρια στη βάση του πραγματικού αντικειμένου λόγω ενός συνδυασμού στρέψης και κάμψης που δε συνάδει με τη δυναμική του.

4.2 Συνάρτηση Επιβράβευσης

Η συνάρτηση επιβράβευσης αποτελεί το μέσο αξιολόγησης των αποφάσεων του πράκτορα. Είναι το άθροισμα δύο υπο-συναρτήσεων, οι οποίες επιτελούν διαφορετικούς σκοπούς.

- Η μια υπο-συνάρτηση αφορά το γενικό προσανατολισμό του πράκτορα κατά τη διάρκεια της προσπάθειας απόδρασης. Συγκεκριμένα, αφού έχουμε θεωρήσει ότι ο πράκτορας πρέπει να έχει μια τάση κίνησης προς τα δεξιά προκειμένου να επιτύχει το στόχο του, αυτός λαμβάνει θετική επιβράβευση όταν συμπεριφέρεται με τον ανάλογο τρόπο, δηλαδή όταν επιλέγει μια μετακίνηση με $\vartheta_a \in (-\frac{\pi}{2}, \frac{\pi}{2})$ (δηλαδή όταν $a_x > 0$). Επειδή όμως κατά την απόδραση ο πράκτορας πρέπει να επιλέγει και άλλες δράσεις όπου χρειάζεται, ώστε να μειώσει την δύναμη που μπορεί να του ασκείται, επιτρέπουμε ένα περιθώριο επιβράβευσης και εκτός του προαναφερθέντος διαστήματος. Επιλέγουμε αυτό το περιθώριο να είναι 0.2 rad, δηλαδή περίπου 11.5°σε κάθε άκρο του διαστήματος, άρα τελικά η επιβράβευση αφορά το διάστημα $(-\frac{\pi}{2} - 0.2, \frac{\pi}{2} + 0.2)$. Για γωνίες πιο κοντά στις 0°(δηλαδή στην κίνηση καθαρά προς τα δεξιά) έχουμε μεγαλύτερες επιβραβεύσεις, μοντελοποιώντας τη συμπεριφορά αυτή με μια συνημιτονοειδή συνάρτηση με μέγιστο μόνο στις 0°. Για τις γωνίες μετακίνησης εκτός του διαστήματος θετικής επιβράβευσης δεν υπάρχει κάποια ποινή (αρνητική επιβράβευση), καθώς μπορεί να αποτελούν αναγκαίες δράσεις για τη μείωση της εφαρμοζόμενης στον πράκτορα δύναμης. Η γενική (παραμετρική) μορφή της πρώτης υπο-συνάρτησης εκφράζεται μαθηματικά παρακάτω:

$$Reward_1 = \begin{cases} k_1 \cdot \cos(k_2 \cdot \vartheta_a) & \text{αν } \vartheta_a \in (-\frac{\pi}{2} - 0.2, \frac{\pi}{2} + 0.2) \\ 0 & \text{αλλιώς} \end{cases} \tag{4.1}$$

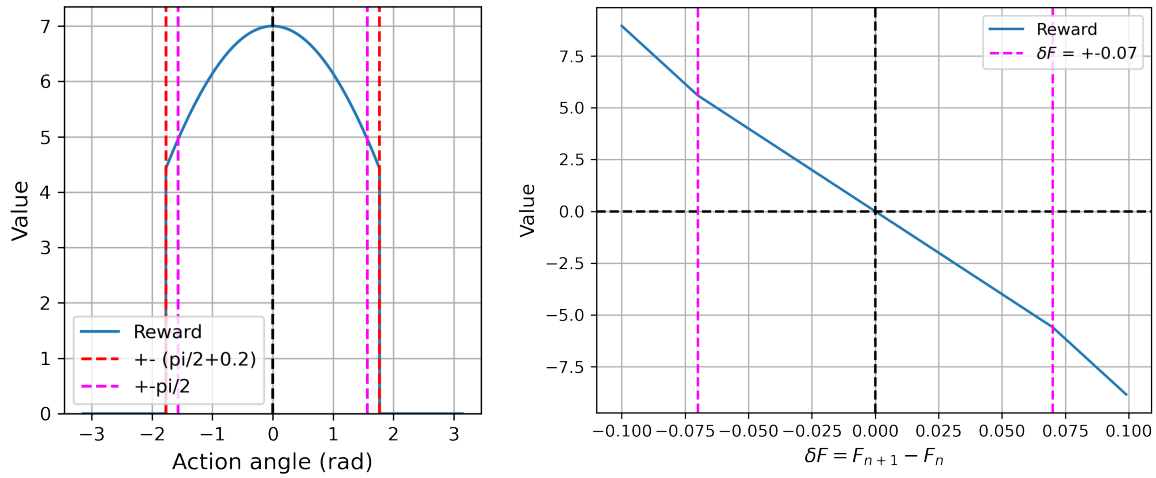
  όπου k_1, k_2 παράμετροι.

- Η δεύτερη υπο-συνάρτηση αφορά την επιθυμητή συμπεριφορά του πράκτορα να μειώνει την εφαρμοζόμενη πάνω του δύναμη μέσω της επαναφοράς στα όρια του διαδρόμου. Αυτό μπορεί να επιτευχθεί με χρήση της διαφοράς του μέτρου της δύναμης ανάμεσα στα δύο τελευταία βήματα του επεισοδίου $\delta F = F_{s+1} - F_s$, η οποία βάσει ορισμού μπορεί να προσφέρει θετική επιβράβευση λόγω μείωσης της δύναμης και αρνητική λόγω άυξησης. Μια ακόμα σημαντική λεπτομέρεια είναι η επιθυμία μας ο πράκτορας να μειώνει τη διείσδυσή του σε τοίχο (και άρα της δύναμης που δέχεται)

κρατώντας όμως την τάση που έχει προς τα δεξιά του διαδρόμου, όταν η διείσδυση είναι σχετικά μικρή. Για αυτόν το λόγο, ορίζουμε πιο απότομη κλίση (k_{big} αντί για k_{small}) όταν υπάρχει μεγάλη διαφορά μεταξύ των F_s, F_{s+1} (και άρα μεγάλη επιβράβευση για μεγάλη μείωση, μεγάλη ποινή για μεγάλη αύξηση), με σκοπό την αποτροπή των δράσεων που επιφέρουν μεγάλη αύξηση διείσδυσης και δύναμης και την ενθάρρυνση των δράσεων που επιφέρουν σημαντική μείωσή τους, όταν βέβαια ο πράκτορας βρίσκεται σε κατάσταση που το κάνει εφικτό. Με αυτό τον τρόπο, η τάση κίνησης προς τα δεξιά εφαρμόζεται κυρίως όταν η διείσδυση, και επομένως η δύναμη, είναι ελεγχόμενη. Ορίζουμε ως σημεία αλλαγής της κλίσης τις διαφορές δύναμης ±0.07 μονάδες δύναμης (που αντιστοιχούν σε ίση διαφορά διείσδυσης, δηλαδή το 35% της ακτίνας του κυκλικού πράκτορα). Τέλος, φροντίζουμε η υπο-συνάρτηση να είναι συνεχής για πιο αποτελεσματική εκπαίδευση. Η γενική μαθηματική έκφρασή της παρουσιάζεται παρακάτω:

$$Reward_2 = \begin{cases} k_{small} \cdot 0.07 + k_{big} \cdot (-\delta F - 0.07) & \text{αν } \delta F < -0.07 \\ k_{small} \cdot \delta F & \text{αν } -0.07 \leq \delta F \leq 0.07 \\ -k_{small} \cdot 0.07 - k_{big} \cdot (\delta F - 0.07) & \text{αν } \delta F > 0.07 \end{cases} \tag{4.2}$$

Η συνάρτηση επιβράβευσης, λοιπόν, ορίζεται ως $Reward = Reward_1 + Reward_2$. Στο σχήμα 4.2 γίνεται μια οπτικοποίηση των δύο υπο-συναρτήσεων που την αποτελούν, για τις πραγματικές παραμέτρους που επιλέξαμε για την υλοποίηση με Linear Semi-gradient SARSA (θα αναφερθούμε σε αυτό το θέμα στην ενότητα 5.1).



(α') *Υπο-συνάρτηση (4.1)* (β') *Υπο-συνάρτηση (4.2)*

Σχήμα 4.2: *Γραφικές παραστάσεις των υπο-συναρτήσεων που συνθέτουν την συνάρτηση επιβράβευσης*

4.3 Προσαρμοστική Εξερεύνηση & Ρυθμός Μάθησης

Όπως αναφέραμε παραπάνω, ο τρόπος διαχείρισης του ρυθμού μάθησης και της εξερεύνησης είναι πολύ σημαντικός για να επιτευχθεί σύγκλιση (βλ. υποενοτ. 3.3.2).

Σχετικά με το ρυθμό μάθησης, υπάρχει στη διεθνή βιβλιογραφία σημαντικός αριθμός ερευνητικών εργασιών που προσεγγίζουν το σχεδιασμό της στρατηγικής σχετικά με αυτόν, με σκοπό τη βελτίωση του ρυθμού σύγκλισης [22, 23]. Για παράδειγμα, μια προσέγγιση προσαρμοστικού (adaptive) ρυθμού μάθησης αφορά την χρήση συστήματος ασαφούς λογικής για την εύρεση των βέλτιστων τιμών για αυτόν [24]. Γενικά, η μεθοδολογία που θα ακολουθηθεί για τον ρυθμό μάθησης εξαρτάται κάθε φορά από το πρόβλημα που έχουμε να επιλύσουμε, τις απαιτήσεις, τις ιδιαιτερότητες και τα χαρακτηριστικά του. Στην εργασία [25] παρουσιάζεται μια μορφή μεταβλητού ρυθμού μάθησης που μπορεί να συνεισφέρει στην περίπτωση Ενισχυτικής Μάθησης πολλαπλών πρακτόρων.

Όσον αφορά την εξερεύνηση, επίσης υπάρχει πληθώρα μεθόδων που μπορούν να εφαρμοστούν. Τέτοιες τεχνικές εξερεύνησης που αφορούν τη βαθιά Ενισχυτική Μάθηση παρουσιάζονται για παράδειγμα στην εργασία [26]. Η επιλογή του κατάλληλου τύπου εξερεύνησης αφορά, όπως και ο ρυθμός μάθησης, το πρόβλημα στο οποίο θα την εφαρμόσουμε.

Στο δικό μας πρόβλημα καταλαβαίνουμε ότι η απλή προσέγγιση μιας μειούμενης με το πέρασμα των επεισοδίων εξερεύνησης δεν ενδείκνυται, καθώς μπορεί να υπάρξουν καταστάσεις οι οποίες πρεοσπελαύνονται πολύ λίγες φορές, και άλλες που ο πράκτορας τις επισκέπτεται πολύ περισσότερες (η μορφολογία του διαδρόμου μπορεί να κάνει κάποιες δυνάμεις να εφαρμόζονται πολύ πιο συχνά πάνω στον πράκτορα). Για αυτόν το λόγο, υλοποιούμε μια εξερεύνηση βασισμένη στο μέτρημα (count-based exploration). Σύμφωνα με αυτή την προσέγγιση, γίνεται καταγραφή της συχνότητας επίσκεψης του πράκτορα σε κάποιες καταστάσεις, έτσι ώστε να δοθεί προτεραιότητα εξερεύνησης σε αυτές που έχουν προσπελαστεί λιγότερες φορές ή καθόλου. Ένα θέμα που συχνά προκύπτει σε αυτή τη μέθοδο είναι η δυσκολία υλοποίησής του σε χώρο καταστάσεων (ή καταστάσεων-δράσεων) πολλών διαστάσεων, και για αυτό έχει διεξαχθεί έρευνα για την ανάπτυξη κατάλληλων για αυτό το πρόβλημα τεχνικών βασισμένων στη μέθοδο αυτή (π.χ. [27, 28]). Στη δική μας υλοποίηση ο χώρος καταστάσεων, παρ' όλο που είναι συνεχής, έχει σχετικά μικρό πλήθος διαστάσεων (διδιάστατος), και άρα δε χρειαζόμαστε κάποια πιο προηγμένη τεχνική. Χωρίζουμε όμως τον χώρο σε ¨γειτονιές¨ κοντινών σε μέτρο ή/και γωνία καταστάσεων (δυνάμεων) που αντιμετωπίζονται ενιαία στην καταγραφή των συχνοτήτων επίσκεψης.

Την ίδια λογική ακολουθούμε και για το ρυθμό μάθησης. Συγκεκριμένα, διατηρώντας την ίδια διαίρεση σε ¨γειτονιές¨, μειώνουμε το ρυθμό μάθησης όχι με βάση το χρόνο εκπαίδευσης, αλλά τοπικά σε κάθε μία ¨γειτονιά¨ με βάση το πλήθος επισκέψεων σε αυτή. Έτσι, κατά το βήμα πρόβλεψης χρησιμοποιείται η τιμή ρυθμού μάθησης που αντιστοιχεί στην εκάστοτε κατάσταση που βρίσκεται ο πράκτορας.

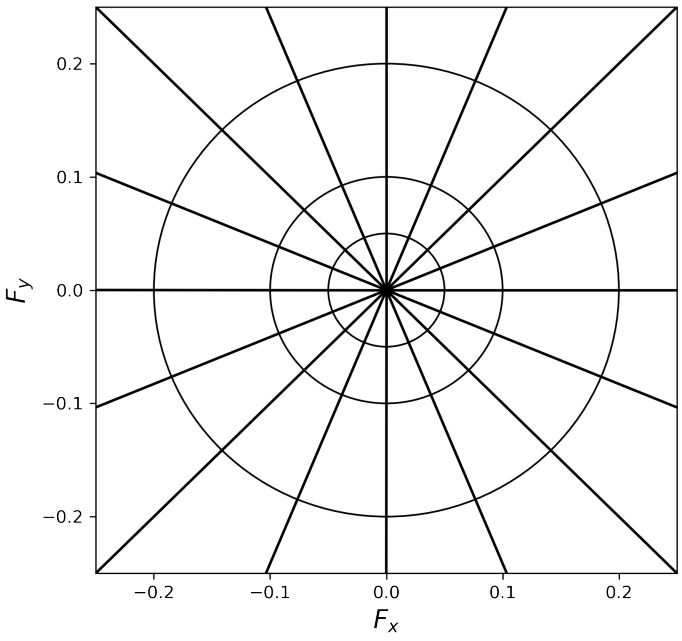
Ο ακριβής τρόπος που υλοποιούνται τα παραπάνω θα παρουσιαστεί αναλυτικά για κάθε αλγόριθμο παρακάτω (υποενοτ. 4.4.2, υποενοτ. 4.5.2).

4.4 Υλοποίηση με SARSA

Πρώτα επιχειρούμε υλοποίηση με SARSA σε μερικώς διακριτό χώρο καταστάσεων και δράσεων, ως δοκιμή που αξιολογεί την αποτελεσματικότητα της συγκεκριμένης μεθόδου μάθησης (για την παρουσίαση του αλγορίθμου βλ. ενοτ. 3.3).

4.4.1 Διακριτοποίηση των Χώρων Καταστάσεων και Δράσεων

Ο αλγόριθμος SARSA αναφέρεται σε διακριτό χώρο καταστάσεων και δράσεων. Για να τον εφαρμόσουμε, χωρίζουμε τον χώρο καταστάσεων σε 65 ¨γειτονιές¨ δυνάμεων, ανάλογα με το μέτρο και την γωνία τους. Η κατάσταση μηδενικής δύναμης αποτελεί μια ξεχωριστή γειτονιά, καθώς θεωρείται πολύ διαφορετική από τις άλλες (και επιθυμητή). Οι υπόλοιπες 64 είναι χωρισμένες, με βάση πολικές συντεταγμένες, με γωνιακά όρια κάθε $\frac{\pi}{8} rad$ και με ακτινικά όρια τις 0.05, 0.1 και 0.2 μονάδες δύναμης (δεν παίρνουμε πιο μεγάλα ακτινικά όρια γιατί οι δυνάμεις που είναι μεγαλύτερες από 0.2 μονάδες θεωρούνται πολύ πιο μεγάλες από το ανεκτό, και δεν έχει νόημα η περαιτέρω ακτινική διαίρεση). Ένα σχεδιάγραμμα της διαίρεσης σε ¨γειτονιές¨ απεικονίζεται στο σχήμα 4.3.

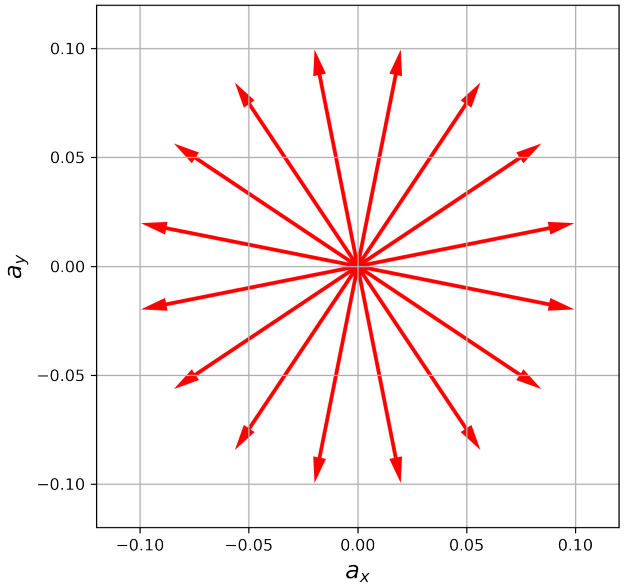


Σχήμα 4.3: *Διακριτοποίηση Χώρου Καταστάσεων*

Παρ' όλη την παραπάνω διαδικασία, η ανταμοιβή υπολογίζεται απο την πραγματική διαφορά των δυνάμεων των δύο τελευταίων βημάτων. Η διακριτοποίηση των καταστάσεων αφορά μόνο το ρόλο τους στη συνάρτηση δράσης-αξίας και όπου αυτή χρησιμοποιείται.

Με παρόμοιο τρόπο γίνεται και μια διακριτοποίηση στο χώρο δράσεων. Συγκεκριμένα, αφού γενικά στο πρόβλημά μας έχουμε θέσει το μέτρο της μετακίνησης σε κάθε βήμα ίσο με 0.1, ορίζουμε 32 διαφορετικές διακριτές κινήσεις που μπορούν να γίνουν σε κάθε βήμα,

οι οποίες διαφέρουν μαζί τους διαδοχικά κατά $\frac{\pi}{8} rad$. Η διακριτοποίηση αυτή απεικονίζεται στο σχήμα 4.4.



Σχήμα 4.4: *Διακριτοποίηση Χώρου Δράσεων*

Έτσι, η συνάρτηση δράσης-αξίας αναπαρίσταται με έναν πίνακα 65×32. Σημειώνουμε εδώ ότι σε περίπτωση που οι πραγματικές καταστάσεις δύο διαδοχικών βημάτων ανήκουν στην ίδια ¨γειτονιά¨ δεν υπάρχει πρόβλημα, επειδή για λόγους που θα εξηγήσουμε στην ενότητα 5.1, στο βήμα πρόβλεψης (γραμμή 8 του ψευδοκώδικα 3.1) θέτουμε $\gamma = 0$, έχουμε δηλαδή μυωπική αξιολόγηση όπου $\gamma \hat{q}(S_{t+1}, A_{t+1}, \mathbf{w}_t) = 0$.

4.4.2 Πολιτική & Εξερεύνηση

Η πολιτική που ακολουθούμε, δηλαδή με βάση την οποία γίνεται η επιλογή δράσεων (γραμμές 4, 7 του ψευδοκώδικα 3.1) είναι μια κλασσική ε-greedy (βλ. εξ. (2.16)), κατά την οποία υπάρχει μια πιθανότητα ε να επιλεγεί μια από τις 16 δράσεις τυχαία (εξερεύνηση), και πιθανότητα $1-\epsilon$ να επιλεγεί η δράση που μεγιστοποιεί τη συνάρτηση δράσης-αξίας (άπληστη επιλογή), δηλαδή αυτή που αντιστοιχεί στο μεγαλύτερο στοιχείο της στήλης που αντιστοιχεί στην κατάσταση στην οποία βρίσκεται ο πράκτορας, μέσα στον πίνακα Q. Ο παράγοντας ε, όπως και ο ρυθμός μάθησης α είναι διαφορετικοί για κάθε ¨γειτονιά¨ καταστάσεων, σύμφωνα με τη διακριτοποίηση που περιγράφουμε στην υποενότητα 4.4.1 (δηλαδή υπάρχουν 65 διακριτές τιμές για τον καθένα). Σε κάθε προσπέλαση μιας ¨γειτονιάς¨ αυξάνεται η τιμή ενός μετρητή για τη συγκεκριμένη, ο οποίος, όταν φτάνει σε ένα ορισμένο κατώφλι, σηματοδοτεί τη μείωση των ε, α για τη ¨γειτονιά¨, μέσω πολλαπλασιασμού με συντελεστές μείωσης (και ο συγκεκριμένος μετρητής επανεκκινείται). Με αυτό τον τρόπο επιτυγχάνεται η μέθοδος που περιγράψαμε στην ενότητα 4.3.

4.5 Υλοποίηση με Episodic Linear Semi-gradient SARSA

Η υλοποίηση αυτή (βλ. 3.4) αφορά την επέκταση της προηγούμενης μεθόδου σε συνεχή χώρο καταστάσεων και δράσεων, με χρήση Gradient-Descent. Ως εκ τούτου, κρατάμε τα χρήσιμα στοιχεία της υλοποίησης της ενότητας 4.4, κάνοντας τις απαραίτητες τροποποιήσεις και προσθήκες.

4.5.1 Υλοποίηση της Προσέγγισης Συνάρτησης

Όπως έχουμε αναφέρει, η προσέγγιση της συνάρτησης δράσης-αξίας γίνεται με γραμμικές μεθόδους, και χαρακτηρίζεται από την εξίσωση (2.23). Κατασκευάζουμε το διάνυσμα χαρακτηριστικών $\mathbf{x}$ με RBFs. Κάθε στοιχείο του διανύσματος εξαρτάται από την κατάσταση που βρίσκεται και τη δράση που επιλέγει ο πράκτορας ως το γινόμενο δύο RBFs όπου η μία αφορά την κατάσταση και η άλλη τη δράση. Για τον υπολογισμό τους δε χρησιμοποιούμε τη σχέση (2.24), αλλά τον τύπο της πολυδιάστατης Γκαουσιανής (παραλείποντας το συντελεστή $(2\pi)^{-k/2}\det(\mathbf{\Sigma})^{-1/2}$):

$$exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})) \tag{4.3}$$

όπου $\mathbf{x}$ η (πολυδιάστατη) μεταβλητή μας, $\boldsymbol{\mu}$ η μέση τιμή (ίδιων διαστάσεων με τη $\mathbf{x}$), και $\mathbf{\Sigma}$ ο πίνακας συνδιακύμανσης.

Οι RBFs που αφορούν την κατάσταση ορίζονται από καταστάσεις-κέντρα (μέσες τιμές), γύρω από τις οποίες "απλώνονται" στο χώρο. Η μεταβλητή μας είναι η κατάσταση στην οποία βρίσκεται ο πράκτορας $\mathbf{s} = [F_s, \partial_{Fs}]^T$ και η μέση τιμή είναι η κατάσταση-κέντρο $\mathbf{c}_s = [F_c, \partial_{Fc}]$ που αντιστοιχεί σε κάθε RBF. Έτσι, ο τύπος (4.3) παίρνει την ακόλουθη μορφή:

$$exp(-\frac{1}{2}(\mathbf{s}-\mathbf{c}_s)^T\mathbf{\Sigma}^{-1}(\mathbf{s}-\mathbf{c}_s))$$

Τονίζουμε ότι δεν υπολογίζουμε την αριθμητική διαφορά των γωνιών μεταξύ της κατάστασης s και του κέντρου c_s (θυμίζουμε ότι βρίσκονται στο διάστημα $(-\pi, \pi]$), καθώς το σύνολο αυτό είναι οργανωμένο κυκλικά. Αντί για αυτό, υπολογίζουμε την πραγματική διαφορά ως τη μικρότερη εκ των μετρήσεών της σύμφωνα και αντίθετα με τη φορά του ρολογιού (ούτως ή άλλως η τυπική απόκλιση που θέτουμε είναι αρκετά μικρότερη από αυτή που θα χρειαζόταν ώστε οι RBFs να παίρνουν μη-μηδενικές τιμές χρησιμοποιώντας την αριθμητική γωνιακή διαφορά).

Πρέπει να ορίσουμε τις κατάλληλες καταστάσεις-κέντρα ώστε να βρίσκονται ομοιόμορφα στο συνεχή χώρο. Έτσι, επιλέγουμε 33 κέντρα, ένα για την κατάσταση που αντιστοιχεί στη μηδενική δύναμη, και τα υπόλοιπα οργανωμένα σε ομόκεντρους κύκλους γύρω από το (0,0). Τα κέντρα αυτά αντιστοιχούν στις δυνάμεις με μέτρα 0.05, 0.09, 0.13, 0.17, και γωνίες που διαφέρουν διαδοχικά κατά $\frac{\pi}{4}$ rad (ξεκινώντας από 0 rad). Ο πίνακας συνδιακύμανσης για τις RBFs των καταστάσεων έχει διαστάσεις 2×2 και είναι διαγώνιος, αφού το μέτρο και η

γωνία των δυνάμεων είναι μεγέθη ασυσχέτιστα, δηλαδή:

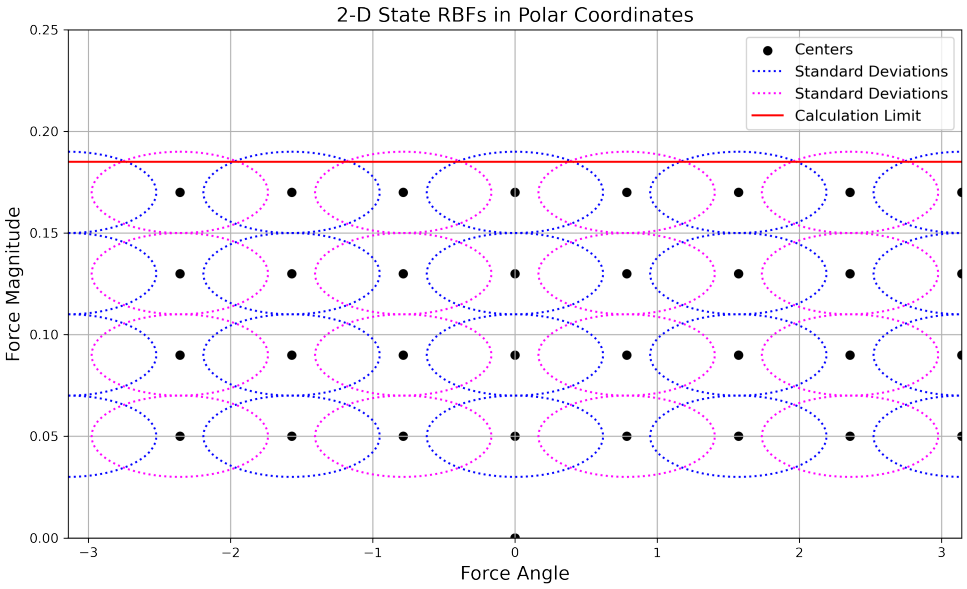
$$\mathbf{\Sigma}_s = \begin{bmatrix} \sigma_{00_s} & 0 \\ 0 & \sigma_{11_s} \end{bmatrix}$$

όπου το στοιχείο σ_{00_s} αφορά το μέτρο της δύναμης, ενώ το στοιχείο σ_{11_s} τη γωνία.

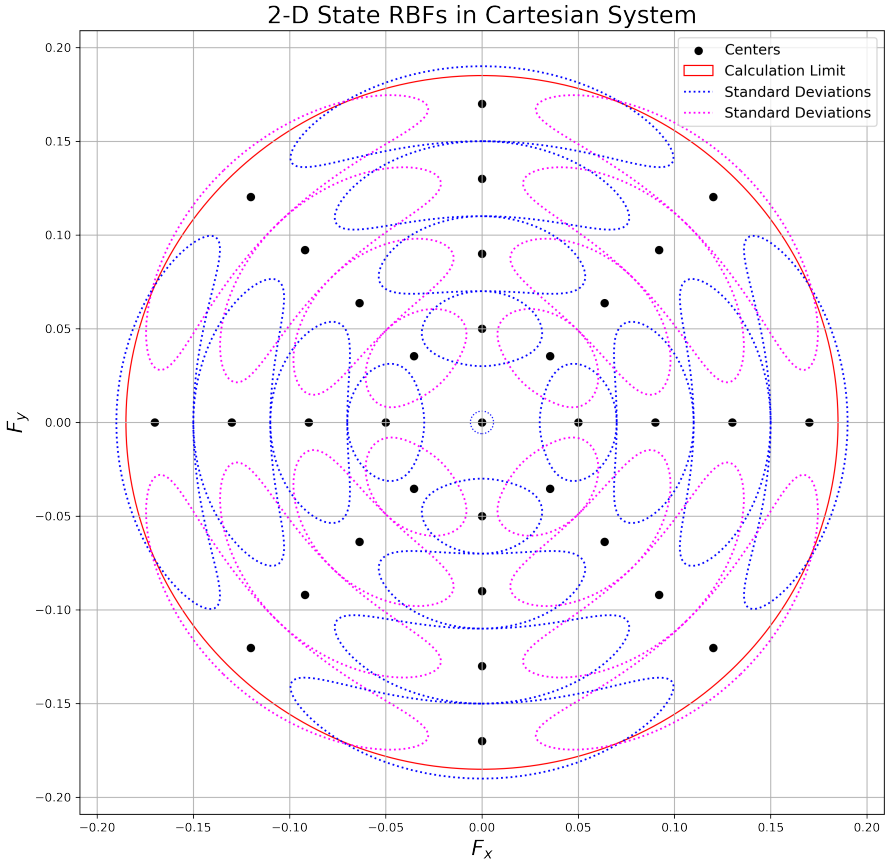
Για την κατάσταση μηδενικής δύναμης ισχύουν ξεχωριστοί κανόνες. Συγκεκριμένα, η γωνιακή διαφορά μεταξύ της μηδενικής δύναμης και αυτής που βρίσκεται κάποια στιγμή ο πράκτορας θεωρείται πάντα ίση με 0. Επομένως, για το κέντρο που αντιπροσωπεύει αυτή τη δύναμη έχουμε μονοδιάστατη (σε πολικές συντεταγμένες) *RBF* που εξαρτάται μόνο από τη διαφορά του μέτρου. Η *RBF* αυτή (με βάση τον τύπο (2.24)) χαρακτηρίζεται από τυπική απόκλιση $\sigma_{s_0} < \sigma_{00_s}$. Χρησιμοποιούμε διαφορετική τυπική απόκλιση για το συγκεκριμένο κέντρο, γιατί χαρακτηρίζει μια κατάσταση αρκετά διαφορετική από τις άλλες, αυτή κατά την οποία ο πράκτορας δεν δέχεται κάποια δύναμη ώστε να αντιδράσει κατάλληλα σε αυτή, και μεγαλύτερο σ θα δημιουργούσε σημαντικές εξαρτήσεις μεταξύ αυτής της κατάστασης με τις άλλες, κάτι το οποίο δεν είναι θεμιτό. Η επιλογή αυτή επαληθεύεται και από πειραματικά αποτελέσματα (βλ. υποενοτ. 5.3.3).

Επιπλέον, αν ο πράκτορας βρεθεί σε κατάσταση που αντιστοιχεί σε δύναμη μέτρου > 0.185, υπολογίζουμε τις *RBF* ως ίσες με τις αντίστοιχες της κατάστασης με την ίδια γωνία, και μέτρο ίσο 0.185. Με αυτό τον τρόπο δε χρειαζόμαστε παραπάνω κέντρα που να αντιστοιχούν σε μεγαλύτερα μέτρα, και οι αποστάσεις έχουν τάξη μεγέθους τέτοια ώστε η ανανέωση των βαρών να είναι υπολογίσιμη (δηλαδή δεν έχουμε πολύ μεγάλες αποστάσεις που θα μηδενίζουν όλες τις *RBFs*). Ταυτόχρονα, αυτό δεν αποτελεί πρόβλημα γιατί το μέτρο δύναμης 0.185 θεωρείται αρκετά μεγάλο για το task μας, κάτι το οποίο σημαίνει ότι σε καταστάσεις που αντιστοιχούν σε μέτρο ≥ 0.185 θέλουμε παρόμοια συμπεριφορά. Τέλος, όπως έχουμε αναφέρει για την προηγούμενη υλοποίηση, έτσι και εδώ, εφόσον έχουμε μυωπική αξιολόγηση ($\gamma = 0$), δεν μας πειράζει ούτε το γεγονός ότι διαδοχικές $\hat{q}$ μπορεί να είναι πολύ κοντά αριθμητικά (για διαδοχικές μεγάλες δυνάμεις).

Μια οπτικοποίηση των RBFs στο χώρο καταστάσεων παρουσιάζεται στο σχήμα 4.5, τόσο σε πολικές 4.5α′ όσο και σε καρτεσιανές 4.5β′ συντεταγμένες (για το σχεδιασμό τους χρησιμοποιούνται οι υπερπαράμετροι που επιλέχθηκαν, βλ. ενότ. 5.1). Οι τυπικές αποκλίσεις είναι σχεδιασμένες με δύο χρώματα για να είναι καλύτερα εμφανείς και να μην υπάρχει σύγχυση. Ακόμα, με κόκκινο χρώμα παρατηρούμε τις καταστάσεις που αντιστοιχούν σε μέτρα δυνάμεων ίσα με 0.185, πάνω από τα οποία διατηρούμε σταθερές τις τιμές των RBFs για την ίδια γωνία. Αξίζει να σημειωθεί ότι στο σχήμα 4.5α′ οι αριστερότερες τυπικές αποκλίσεις που εμφανίζονται μισές είναι αυτές που συνεχίζονται από τα κέντρα με γωνία π, καθώς όπως αναφέραμε οι γωνίες είναι οργανωμένες κυκλικά. Τέλος, το κενό που εμφανίζεται στο σχήμα 4.5β′ ανάμεσα στο κέντρο μηδενικής δύναμης και στα κοντινότερά του δεν μας πειράζει, καθώς οι καταστάσεις στις οποίες οι τιμές των RBFs είναι μικρές, είναι κυρίως αμελητέου μέτρου.



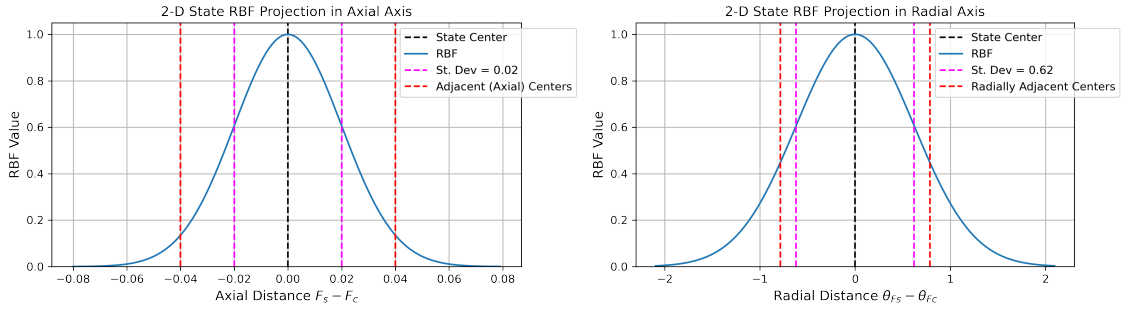
(α') *Σε πολικές συντεταγμένες*



(β') *Στο καρτεσιανό σύστημα*

Σχήμα 4.5: *RBFs στο Χώρο Κατασταάσεων*

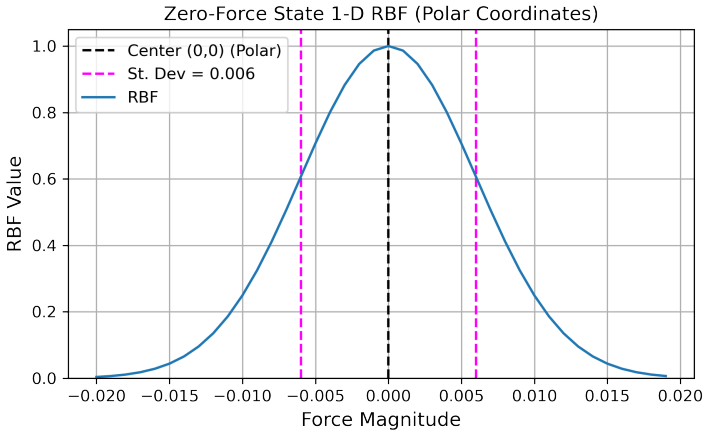
Στο σχήμα 4.6 απεικονίζουμε προβολές των διδιάστατων RBF τόσο στον άξονα του μέτρου (4.6α') όσο και στον άξονα της γωνίας (4.6β'). Έτσι, παρατηρούμε πώς μεταβάλλονται οι τιμές τους ανάλογα με τη διαφορά στο μέτρο ή στη γωνία αντίστοιχα, κρατώντας το άλλο μέγεθος σταθερό και ίσο (χρησιμοποιούνται οι υπερπαράμετροι που επιλέχθηκαν, βλ. ενότ. 5.1).



(α') *Προβολή στον άξονα του μέτρου* (β') *Προβολή στον άξονα της γωνίας*

Σχήμα 4.6: *Μονοδιάστατες προβολές των 2-D RBFs*

Παρακάτω (σχ. 4.7) βλέπουμε και τη μονοδιάστατη RBF (ως συνάρτηση της διαφοράς μέτρου) που αντιστοιχεί στο κέντρο μηδενικής δύναμης (για υπερπαραμ. βλ. ενότ. 5.1):



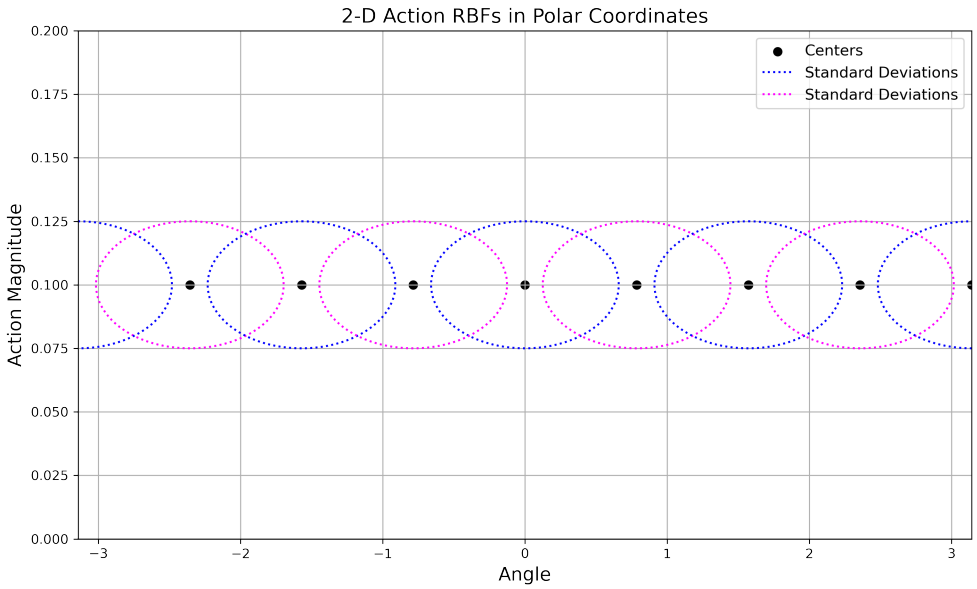
Σχήμα 4.7: *RBF γύρω από τη μηδενική δύναμη*

Για συναρτήσεις που αφορούν τη δράση ακολουθούμε ανάλογη στρατηγική. Επιστρέφοντας στον τύπο (4.3), η μεταβλητή μας είναι η δράση την οποία παίρνει ο πράκτορας $\mathbf{a} = [a_s, \vartheta_{a_s}]$, η μέση τιμή είναι η κάθε δράση-κέντρο $\mathbf{c}_a = [a_c, \vartheta_{a_c}]$ που έχουμε επιλέξει, και όπως πριν έχουμε έναν πίνακα συνδιακύμανσης

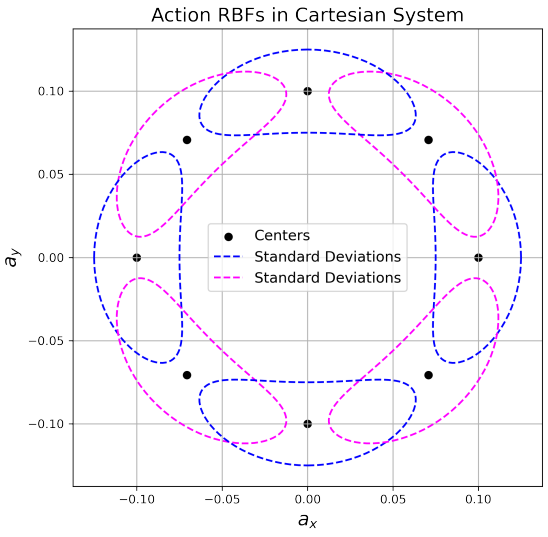
$$\mathbf{\Sigma}_a = \begin{bmatrix} \sigma_{00_a} & 0 \\ 0 & \sigma_{11_a} \end{bmatrix}$$

Επιλέγουμε 8 κέντρα c_{action} τοποθετώντας τα κυκλικά στις ίδιες γωνίες, αλλά σε ακτίνα αποκλειστικά 0.1, που είναι και το μέτρο μετακίνησης που έχουμε ορίσει. Καταλαβαίνει κανείς, λοιπόν, πως οι RBFs θα μπορούσαν να είναι μονοδιάστατες (σε πολικές συντεταγμένες, ως συνάρτηση της γωνιακής διαφοράς), όμως για γενίκευση προτιμούμε τυπικά να κρατήσουμε

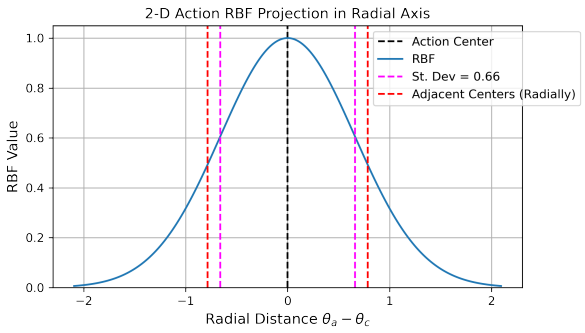
και τη συνεισφορά του μέτρου. Οι συναρτήσεις αυτές παρουσιάζονται στο σχήμα 4.8 (για υπερπαραμ. βλ. ενότ. 5.1).



(α′) *Σε πολικές συντεταγμένες*



(β′) *Στο καρτεσιανό σύστημα*



(γ′) *Προβολή στον άξονα των γωνιών*

Σχήμα 4.8: *RBFs στο Χώρο Δράσεων*

Έχουμε, λοιπόν 33 RBFs που αφορούν καταστάσεις, και 8 που αφορούν δράσεις. Έτσι, το διάνυσμα χαρακτηριστικών έχει διαστάσεις $(33 \cdot 8\,,\ 1) = (264\,,\ 1)$ (λαμβάνοντας δηλαδή υπ' όψιν όλους τους συνδυασμούς γινομένων).

4.5.2 Πολιτική & Γκαουσιανή Εξερεύνηση

Όπως και στον SARSA, έτσι και σε αυτόν τον αλγόριθμο ακολουθούμε ένα είδος ε-greedy πολιτικής (που εφαρμόζεται στις γραμμές 3, 9 του ψευδοκώδικα 3.2). Αντί όμως να επιλέγεται μια τυχαία κίνηση με μια πιθανότητα ε, εκτελούμε γκαουσιανή εξερεύνηση. Σύμφωνα με αυτή, κάθε φορά υπολογίζεται η δράση με μέτρο 0.1 μονάδες μήκους που μεγιστοποιεί τη συνάρτηση δράσης-αξίας (άπληστη επιλογή). Στη συνέχεια όμως η γωνία της δράσης επιλέγεται τυχαία από μια γκαουσιανή κατανομή με μέση τιμή την γωνία της βέλτιστης δράσης, και τυπική απόκλιση τον παράγοντα ε. Είναι λογικό ότι ξεκινάμε με μεγαλύτερο ε για να ενθαρρύνουμε την εξερεύνηση, το οποίο στη συνέχεια περιορίζεται με τρόπο ανάλογο με αυτόν της προηγούμενης μεθόδου (βλ. υποενοτ. 4.4.2, ενοτ. 4.3).

Συγκεκριμένα, έχουμε 33 τιμές ε και άλλες τόσες τιμές ρυθμού μάθησης α, όπου η κάθε μία αντιστοιχεί και εφαρμόζεται στη ¨γειτονιά¨ δυνάμεων που βρίσκεται πιο κοντά (σε πολικές συντεταγμένες) σε μια συγκεκριμένη κεντρική κατάσταση c (οι ίδιες που ορίστηκαν για το χώρο καταστάσεων, που είναι επίσης 33). Επιπλέον, ο μετρητής που σηματοδοτεί τη μείωση αυτών των τιμών μέσω κατωφλίου, δεν αυξάνεται διακριτά σε κάθε γειτονιά όπως στην προηγούμενη μέθοδο, αλλά με συνεχή τρόπο μέσω των τιμών RBFs των καταστάσεων (βλ. υποενοτ. 4.5.1). Για το tracking, δηλαδή, γίνεται μια πρόσθεση πινάκων ως εξής:

$$counter_{33\times1} += RBF\_states\_array_{33\times1}$$

Αυτή είναι μια λογική προσέγγιση των προσαρμοστικών ε, α, καθώς μια δύναμη αυξάνει το μετρητή που αντιστοιχεί σε κάθε κεντρική κατάσταση ανάλογα με το πόσο κοντά βρίσκεται σε αυτή.

Όσον αφορά την άπληστη επιλογή, αντί για τη μεγιστοποίηση της συνάρτησης δράσης-αξίας $\hat{q}$, ισοδύναμα επιχειρούμε την ελαχιστοποίηση της αντίθετης της συνάρτησης δράσης-αξίας, $-\hat{q}$, με χρήση της συνάρτησης `minimize` του module `scipy.optimize` της βιβλιοθήκης SciPy, και μέθοδο ελαχιστοποίησης Sequential Least SQuares Programming (SLSQP). Χρησιμοποιούμε μη-γραμμικό περιορισμό `NonlinearConstraint` για να διατηρήσουμε το μέτρο της δράσης ίσο με 0.1 [29].

4.5.3 Μάθηση με Διαταραχές

Η μετάβαση από ένα περιβάλλον προσομοίωσης στην πραγματική διάταξη συχνά δεν μπορεί να επιτευχθεί άμεσα. Αυτό συμβαίνει επειδή οι διαφορές που υπάρχουν μεταξύ της προσομοίωσης και του πραγματικού κόσμου μειώνουν την αποτελεσματικότητα των πολιτικών στον τελευταίο [30]. Για αυτόν το λόγο, έχουν αναπτυχθεί τεχνικές μεταφοράς από προσομοίωση στην πραγματικότητα (simulation-to-real transfer), οι οποίες χρησιμοποιούνται ώστε η μάθηση κατά την προσομοίωση να γίνει χρήσιμη για τον πραγματικό κόσμο. Στην ερευνητική εργασία [30] παρουσιάζονται αρκετές μέθοδοι, διαφορετικής λογικής μεταξύ τους, για να επιτευχθεί αυτό σε εφαρμογές Βαθιάς Ενισχυτικής Μάθησης (που μπορούν παρ' όλα αυτά να γενικευτούν για κάθε τύπο Ενισχυτικής Μάθησης). Μια από αυτές τις μεθόδους ονομάζεται Μάθηση με Διαταραχές (Learning with Disturbances), και αφορά την εισαγωγή διαταραχών σε δεδομένα του προβλήματος (π.χ. θόρυβος) για να αυξηθεί η σταθε-

ρότητα και η αποτελεσματικότητα των πρακτόρων [30]. Για παράδειγμα, στην εργασία [31] ερευνάται η αντιμετώπιση του χάσματος προσομοίωσης και του πραγματικού κόσμου λόγω διαφόρων σφαλμάτων στην αίσθηση, τη ρύθμιση και την αποτελεσματικότητα συνεργατικών ρομπότ, με εισαγωγή των αντίστοιχων διαταραχών στο περιβάλλον προσομοίωσης.

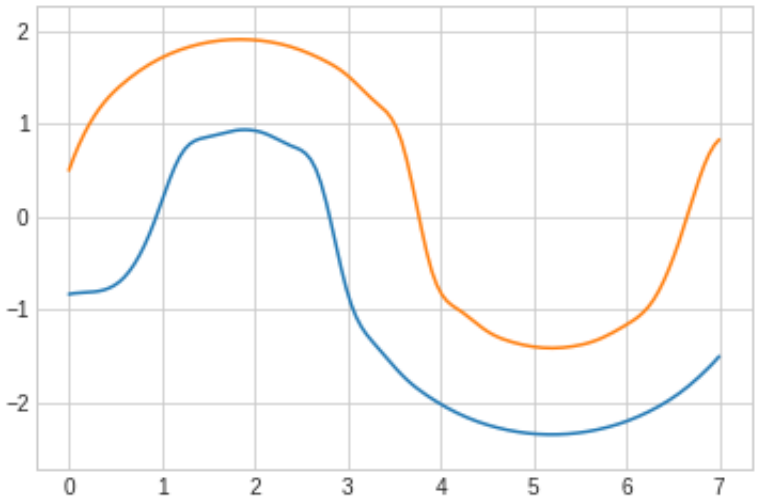
Στη δική μας εργασία, η εκπαίδευση του πραγματικού ρομπότ εκτελείται από την αρχή, και έτσι δε χρειάζεται η μεταφορά μάθησης από την προσομοίωση. Παρ' όλα αυτά, αξίζει να σημειωθεί ότι ενδέχεται στα πειράματα στο πραγματικό ρομπότ να εμφανίζονται σφάλματα στις μετρήσεις δυνάμεων από τους αισθητήρες. Έτσι, εμπνεόμενοι από τη μάθηση με διαταραχές, προσθέτουμε θόρυβο στις δυνάμεις που μετρώνται από τον πράκτορα στην προσομοίωση για να δημιουργήσουμε ένα ανάλογο φαινόμενο, με σκοπό να αξιολογήσουμε τη συμπεριφορά της μεθόδου μας στην εισαγωγή σφαλμάτων. Συγκεκριμένα, στις μετρούμενες δυνάμεις F_x, F_y προσθέτουμε λευκό θόρυβο μεγέθους $[-0.01, 0.01]$.

Κεφάλαιο 5

Αποτελέσματα Υλοποίησης σε Περιβάλλον Προσομοίωσης

Στο παρόν κεφάλαιο θα παραθέσουμε τα αποτελέσματα που προκύπτουν από τη διαδικασία εκπαίδευσης και επαλήθευσης στο περιβάλλον της προσομοίωσης, χρησιμοποιώντας τις υλοποιήσεις που αναλύθηκαν στο προηγούμενο κεφάλαιο.

Η εκπαίδευση διεξάγεται σε έναν ορισμένο διάδρομο. Θέλουμε ο πράκτορας να επιτύχει μια όσο το δυνατόν ομοιόμορφη και ολοκληρωμένη δειγματοληψία καταστάσεων του περιβάλλοντος (δηλαδή δυνάμεων πολλών διαφορετικών γωνιών και μέτρων). Για αυτόν το λόγο, επιλέγουμε ένα διάδρομο εκπαίδευσης ο οποίος να έχει πολλές διαφορετικές κλίσεις κατά μήκος του, που να καλύπτουν, όσο αυτό είναι δυνατόν, όλο το διάστημα γωνιών $(-\pi, \pi)$. Παρουσιάζεται στο σχήμα 5.1. Σημειώνουμε ότι οι πιο δύσκολες για τη μάθηση τροχιές διαδρόμων είναι οι απότομα κατακλινείς και επικλινείς, οι οποίες αντιτίθενται αρκετά από τη φύση τους στην τάση του πράκτορα να κινείται προς τη δεξιά κατεύθυνση.



Σχήμα 5.1: *Ο διάδρομος εκπαίδευσης (προσομοίωση)*

Πρώτα όμως, θα δικαιολογήσουμε την επιλογή των παραμέτρων της υλοποίησης.

5.1 Επιλογή και Σχολιασμός Υπερπαραμέτρων

Στον πίνακα 5.1 παρουσιάζονται οι τιμές των υπερπαραμέτρων που επιλέχθηκαν για κάθε μέθοδο. Στη συνέχεια, ακολουθεί σχολιασμός της στρατηγικής πίσω από τις επιλογές αυτές.

Πίνακας 5.1: *Υπερπαράμετροι της υλοποίησης σε περιβάλλον προσομοίωσης*

Παράμετρος	**Υλοποίηση SARSA**	**Υλοποίηση Episodic Linear Semi-gradient SARSA**
k_1 (εξ. (4.1))	7	7
k_2 (εξ. (4.1))	0.75	0.5
k_{small} (εξ. (4.2))	90	80
k_{big} (εξ. (4.2))	$1.4 \cdot 90$	$1.4 \cdot 80$
Συντελεστής γ (βλ. υποενοτ. 2.1.3, εξ. (2.18), (2.22))	0	0
Αρχικό ε	0.9	$\frac{2\pi}{3}$
Συντελεστής μείωσης ε	0.9	0.75
Αρχικό α	0.85	0.85
Συντελεστής μείωσης α	0.9	0.75
Κατώφλι μετρητή μείωσης ε,α	30	25 για την κατάσταση μηδενικής δύναμης, 4 για τις καταστάσεις με γωνίες 0 και $\pm\frac{\pi}{4} rad$, και 12 για τις υπόλοιπες
σ_{00_s}	-	0.02
σ_{11_s}	-	0.62
σ_{s_0}	-	0.006
σ_{00_a}	-	0.025
σ_{11_a}	-	0.66

Η επιλογή των υπερπαραμέτρων της συνάρτησης ανταμοιβής γίνεται έτσι ώστε να έχουμε συγκρίσιμα μεγέθη επιβράβευσης από δεξιά κίνηση και μείωση εφαρμοζόμενης δύναμης, και ποινής από αύξηση εφαρμοζόμενης δύναμης. Η επιθυμητή συμπεριφορά είναι ο πράκτορας να εκτελεί κίνηση γωνίας 0° όσο δεν δέχεται κάποια δύναμη, και να μειώνει γρήγορα την εφαρμοζόμενη δύναμη (δηλαδή τη διείσδυση) όταν αυτή είναι μεγάλη, χωρίς ταυτόχρονα να ταλαντώνεται μεταξύ καταστάσεων μηδενικής και μη δύναμης. Η επιθυμητή συμπεριφορά κατά την ύπαρξη διείσδυσης είναι δηλαδή η ακολούθηση του δεξιού τοίχου προς τη δεξιά κατεύθυνση διατηρώντας μικρά επίπεδα δύναμης.

Οι επιλογές των τυπικών αποκλίσεων εξυπηρετούν τον ίδιο σκοπό, δημιουργώντας επικαλύψεις μεταξύ των RBFs που επιτυγχάνουν αυτό το αποτέλεσμα καλύπτοντας όλο το συνεχή χώρο. Η μικρότερη τυπική απόκλιση για την κατάσταση μηδενικής δύναμης δικαιολογείται από το γεγονός ότι την θέλουμε σχετικά ανεξάρτητη από τις υπόλοιπες (βλ. υποενοτ. 5.3.3 για καλύτερη οπτικοποίηση).

Όπως έχουμε αναφέρει και σε προηγούμενο κεφάλαιο, θέτουμε $\gamma = 0$ διότι θέλουμε

μυωπική αξιολόγηση με άμεση αντίδραση στο παρόν ερέθισμα.

Η επιλογή του ε για τον αλγόριθμο των συνεχών χώρων καταστάσεων και δράσεων γίνεται έτσι ώστε στις πρώτες προσπελάσεις παρόμοιων καταστάσεων να περιλαμβάνονται όλες οι γωνίες εξερεύνησης στο $(-\pi, \pi]$ (υπενθυμίζουμε ότι χρησιμοποείται ως τυπική απόκλιση για γκαουσιανή εξερεύνηση, βλ. υποενοτ. 4.5.2).

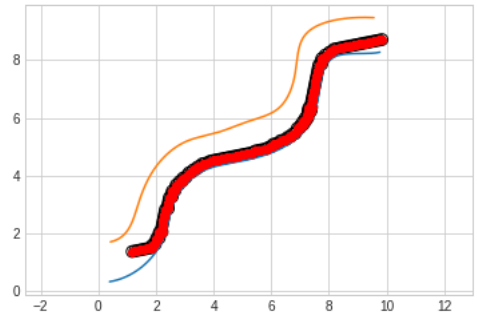
Τέλος, σχετικά με τα κατώφλια των μετρητών μείωσης του ε και του α (βλ. ενοτ. 4.3, υποενοτ. 4.4.2, 4.5.2), χρησιμοποιούμε ένα κοινό για όλες τις καταστάσεις στην υλοποίηση SARSA, κάτι που δε συμβαίνει στην άλλη υλοποίηση. Αυτό δικαιολογείται καθώς ο δυσανάλογος αριθμός δειγμάτων που έχουμε από διαφορετικές περιοχές καταστάσεων δεν επηρεάζει τον διακριτό χώρο, αλλά αντίθετα μπορεί να προσθέσει bias στον συνεχή, οπότε επιλέγουμε διαφορετικά κατώφλια για διαφορετικές περιοχές ώστε η καθεμία να συγκλίνει κατάλληλα μετά από ικανοποιητική αξιοποίηση των υπαρχόντων δειγμάτων.

Δυσανάλογα δείγματα έχουμε λόγω της φύσης του προβλήματος και της επίλυσης που έχουμε σχεδιάσει. Έτσι, αρχικά, οι καταστάσεις μηδενικής δύναμης στις οποίες βρίσκεται ο πράκτορας λόγω του μεγέθους του και της τάσης να τις επιδιώκει, είναι πολύ περισσότερες από τις υπόλοιπες, και άρα χρειάζεται μεγαλύτερο κατώφλι για την αξιοποίηση περισσότερων δειγμάτων (ώστε να αποφευχθεί σύγκλιση σε κάποια υποβέλτιστη κατεύθυνση). Επιπλέον, οι καταστάσεις που βρίσκονται πιο κοντά στις γωνίες 0 και $\pm\frac{\pi}{4}rad$ είναι αυτές τις οποίες ο πράκτορας επισκέπτεται λιγότερο, αφού, λόγω της συνάρτησης ανταμοιβής, μαθαίνει να τις αποφεύγει γρήγορα με κίνηση προς τη γενικότερη δεξιά κατεύθυνση με την οποία επιτυγχάνει επιβράβευση τόσο για την κατεύθυνση κίνησης όσο και για τη μείωση της εφαρμοζόμενης δύναμης. Γενικά, τα περισσότερα δείγματα μη-μηδενικής δύναμης προέρχονται από αλληλεπίδραση με τον τοίχο που εμφανίζεται κάθε φορά στα δεξιά του πράκτορα.

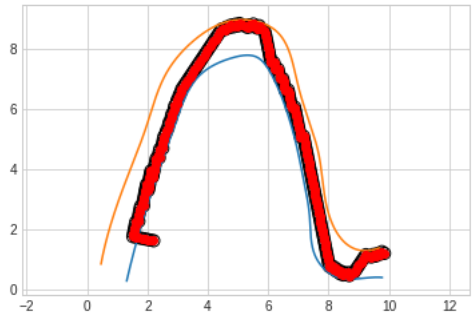
5.2 Αποτελέσματα υλοποίησης SARSA

Ξεκινάμε με την δοκιμή της υλοποίησης SARSA, της οποίας η επιτυχία τελικά ενθαρρύνει την ανάπτυξη της υλοποίησης συνεχών χώρων. Σημειώνουμε ότι δεν θα αναφέρουμε λεπτομέρειες εκπαίδευσης, αλλά μόνο τα αποτελέσματα του testing ενδεικτικά, καθώς αυτή η μέθοδος δεν είναι η κύρια εργασία μας.

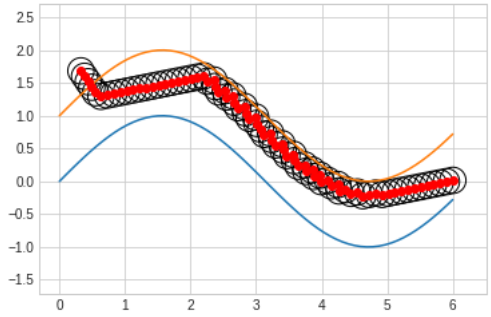
Στο σχήμα 5.2 παρουσιάζεται η δοκιμή του εκπαιδευμένου μοντέλου σε άγνωστους διαδρόμους διαφορετικών τροχιών. Ο πράκτορας φαίνεται να δραπετεύει επιτυχώς από τους διαδρόμους βαδίζοντας προς τη δεξιά κατεύθυνση μέχρι να διεισδύσει στον δεξί τοίχο και στη συνέχεια ακολουθώντας τον, όπως προείπαμε. Ακόμα, εμφανίζεται ικανός να εισέλθει στο διάδρομο όταν ξεκινά εκτός αυτού (βλ. σχ. 5.2β΄, 5.2γ΄, 5.2δ΄).



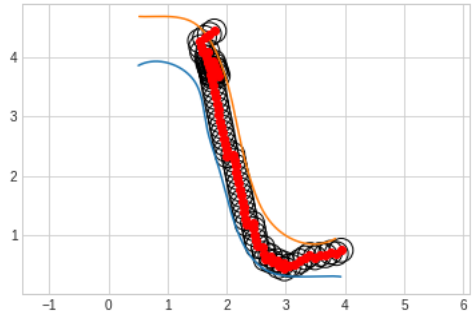
(α') *Επικλινής (κατά διαστήματα) διάδρομος*



(β') *Κοίλος διάδρομος*



(γ') *Ημιτονοειδής διάδρομος*



(δ') *Κατακλινής διάδρομος*

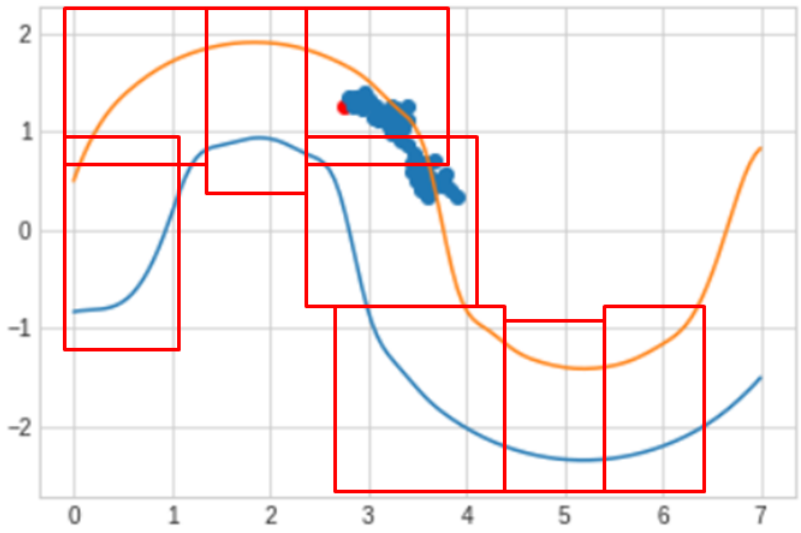
Σχήμα 5.2: *Διαδικασία testing της υλοποίησης με SARSA σε άγνωστους διαδρόμους (προσομοίωση)*

5.3 Αποτελέσματα υλοποίησης Episodic Linear Semi-gradient SARSA

Συνεχίζουμε με την κύρια μας μέθοδο η οποία αφορά συνεχείς χώρους καταστάσεων και δράσεων. Θα αναλύσουμε τόσο την εκπαίδευση όσο και τις δοκιμές σε άγνωστους διαδρόμους, και έπειτα θα δούμε πώς συμπεριφέρεται ο πράκτορας όταν εκπαιδεύεται με διαφορετικές υπερπαραμέτρους από αυτές που θέσαμε στην ενότητα 5.1.

5.3.1 Εκπαίδευση

Για καλύτερη δειγματοληψία καταστάσεων κατά τη διάρκεια της εκπαίδευσης, χωρίζουμε το χώρο μέσα στον οποίο εκκινούμε σε κάθε επεισόδιο τον πράκτορα μέσα σε 8 παράθυρα, κυκλικά. Η εκκίνηση του πράκτορα μπορεί να γίνει τόσο μέσα όσο και έξω από το διάδρομο, ώστε ο πράκτορας να λάβει ερεθίσματα πολλών διαφορετικών ειδών και να δοκιμαστεί η ικανότητά του να επανέρχεται σε αυτόν. Ο πράκτορας τοποθετείται σε αρχικό σημείο μέσα σε κάποιο παράθυρο τυχαία, υπό την προϋπόθεση ότι η αρχική δύναμη που δέχεται είναι μικρότερη από 0.3. Ο χωρισμός των παραθύρων εκκίνησης παρουσιάζεται στο σχήμα 5.3.

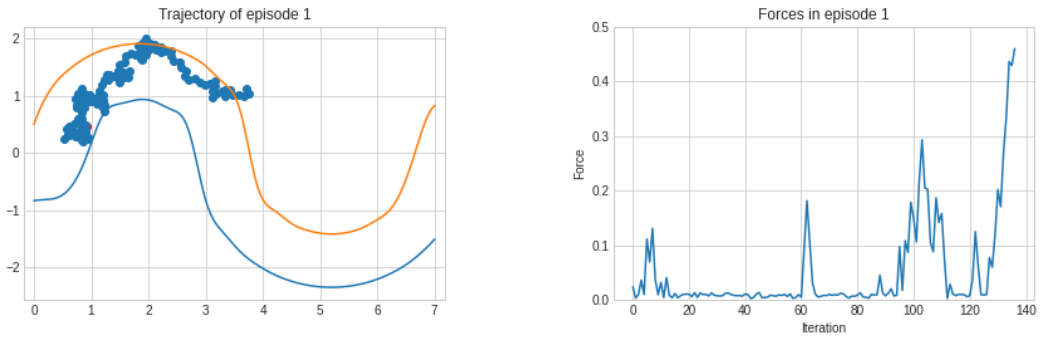


Σχήμα 5.3: *Τυχαιοποίηση και κατανομή της επιλογής σημείου εκκίνησης επεισοδίου*

Εκτελούμε συνολικά 50 επεισόδια, τα οποία ολοκληρώνονται είτε αν ο πράκτορας εξέλθει του διαδρόμου, είτε αν ξεπεράσει τα 200 βήματα (προς αποφυγή κάποιας προσωρινής παγίδευσης η οποία θα καθυστερήσει τη διαδικασία), είτε αν δεχτεί δύναμη μεγαλύτερη από 0.45 (οπότε και έχει βρεθεί σε σημείο αρκετά εκτός για να συνεχιστεί το επεισόδιο).

Παρακάτω φαίνονται κάποια ενδεικτικά επεισόδια της μάθησης τα οποία παρουσιάζουν ενδιαφέρον. Για κάθε επεισόδιο παραθέτουμε δύο γραφήματα, όπου το ένα δείχνει την τροχιά του κέντρου του κυκλικού αντικειμένου (πράκτορα) μέχρι την ολοκλήρωση (σημειώνουμε ότι η αρχική θέση απεικονίζεται με κόκκινο χρώμα, ενώ οι υπόλοιπες με μπλε), και το άλλο λειτουργεί επεξηγηματικά παρουσιάζοντας το μέτρο της δύναμης σε κάθε διακριτό βήμα εκπαίδευσης του επεισοδίου.

- 1ο Επεισόδιο:

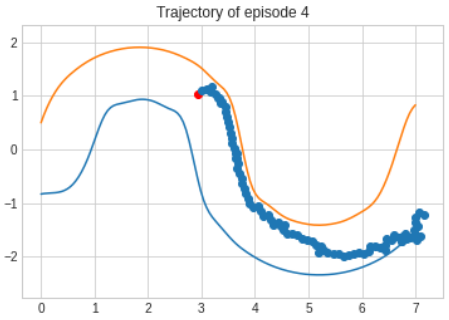


(α') *Τροχιά* (β') *Μετρούμενη από τον πράκτορα δύναμη*

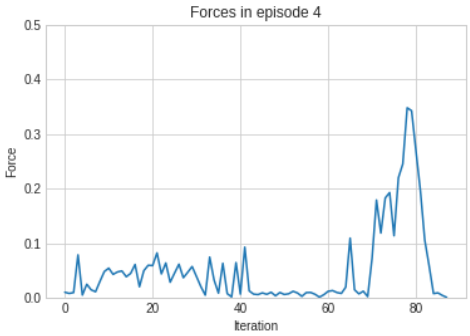
Σχήμα 5.4: *1ο επεισόδιο εκπαίδευσης (προσομοίωση)*

Στο πρώτο επεισόδιο (σχ. 5.4) δε μπορούμε να διακρίνουμε ακόμα την ικανότητα του πράκτορα να πλοηγείται μέσα στον διάδρομο, ο οποίος εκτελεί τις πρώτες ανανεώσεις βαρών με σημαντική εξερεύνηση. Τελικά, το επεισόδιο τερματίζεται λόγω δύναμης πάνω από το επιτρεπτό όριο.

- 4ο Επεισόδιο:



(α') *Τροχιά*

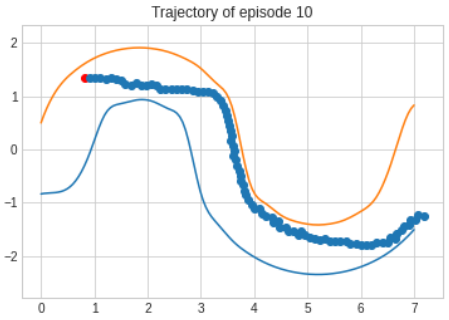


(β') *Μετρούμενη από τον πράκτορα δύναμη*

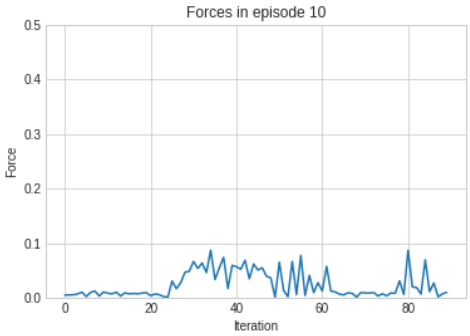
Σχήμα 5.5: *4ο επεισόδιο εκπαίδευσης (προσομοίωση)*

Παρατηρούμε όμως ότι ήδη από το 4ο επεισόδιο (σχ. 5.5) ο πράκτορας καταφέρνει να αντιδρά στις εφαρμοζόμενες δυνάμεις ώστε να τις μειώνει, και τελικά εξέρχεται του διαδρόμου επιτυχώς. Παρ' όλα αυτά, δεν έχει συγκλίνει ακόμα και ιδιαίτερα στα τελευταία βήματα παρουσιάζεται σημαντική αύξηση της δύναμης (που μπορεί να οφείλεται και σε εξερεύνηση).

- 10ο Επεισόδιο:



(α') *Τροχιά*



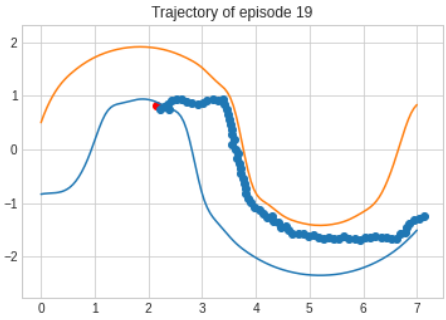
(β') *Μετρούμενη από τον πράκτορα δύναμη*

Σχήμα 5.6: *10ο επεισόδιο εκπαίδευσης (προσομοίωση)*

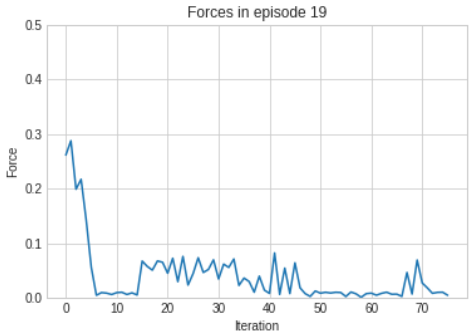
Στο 10ο επεισόδιο (σχ. 5.6) φαίνεται ότι ο πράκτορας έχει μάθει να ακολουθεί τον κατακλινή τοίχο πολύ ικανοποιητικά, και με αρκετά μικρές και λιγότερες απο προηγούμενα επεισόδια εφαρμοζόμενες δυνάμεις.

Μετά από αρκετές δοκιμές εκπαίδευσης του πράκτορα παρατηρούμε ότι περίπου σε αυτό το επεισόδιο η εκπαίδευση έχει επιτευχθεί σε πολύ μεγάλο ποσοστό, κάτι πολύ σημαντικό δεδομένου του γεγονότος ότι στο πραγματικό πείραμα δε θα έχουμε την πολυτέλεια για χρονοβόρες εκπαιδεύσεις. Παρ' όλα αυτά, εφόσον βρισκόμαστε ακόμα σε περιβάλλον προσομοίωσης, θα συνεχίσουμε την εκπαίδευση για τη σωστή δειγματοληψία όλου του χώρου καταστάσεων, ώστε να βελτιστοποιήσουμε λεπτομερώς το μοντέλο μας.

- 19ο Επεισόδιο:



(α') *Τροχιά*

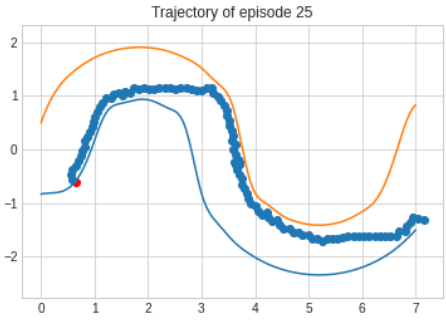


(β') *Μετρούμενη από τον πράκτορα δύναμη*

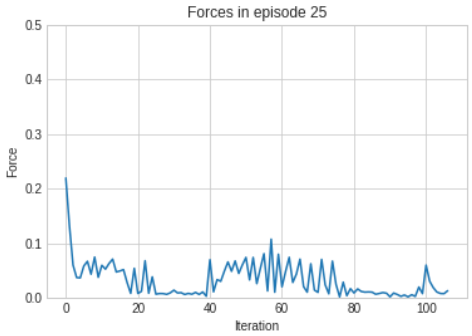
Σχήμα 5.7: *19ο επεισόδιο εκπαίδευσης (προσομοίωση)*

Στο 19ο επεισόδιο (σχ. 5.7) διακρίνεται, μεταξύ άλλων, η προσπάθεια του πράκτορα να μειώσει την αρχική δύναμη που του εφαρμόζεται, από το γεγονός ότι ξεκινά από θέση με διείσδυση. Η συμπεριφορά του σε αυτή τη θέση φαίνεται μέχρι στιγμής να είναι υπο-βέλτιστη, καθώς σχετικά γρήγορα (σε 5 βήματα) καταφέρνει να ελαχιστοποιήσει τη δύναμη (μπορεί να οφείλεται και σε εξερεύνηση).

- 25ο Επεισόδιο:



(α') *Τροχιά*



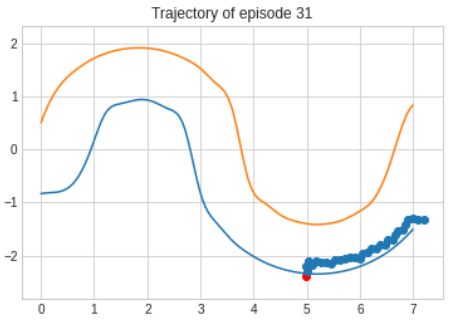
(β') *Μετρούμενη από τον πράκτορα δύναμη*

Σχήμα 5.8: *25ο επεισόδιο εκπαίδευσης (προσομοίωση)*

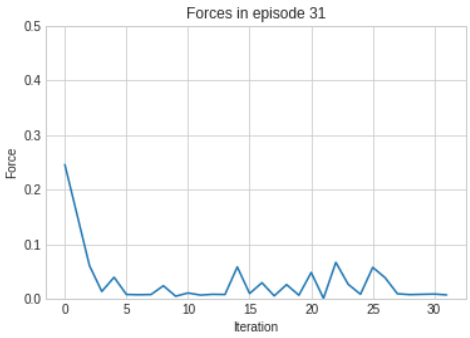
Στο 25ο επεισόδιο (σχ. 5.8) φαίνεται ότι έχουμε πολύ καλή συμπεριφορά τόσο για επικλινή όσο και για κατακλινή τοίχο. Τονίζουμε ότι όσο περισσότερο πλησιάζει η κλίση του τοίχου την κατακόρυφο, τόσο δυσκολότερο μπορεί να γίνει για τον πράκτορα να συμπεριφερθεί με κατάλληλο τρόπο, καθώς η κλίση αυτή αντιτίθεται στην τάση του πράκτορα να κινείται οριζόντια προς τα δεξιά.

Μπορούμε να παρατηρήσουμε, επιπλέον, ότι ενώ το επεισόδιο ξεκινά με τον πράκτορα να βρίσκεται σε θέση με διείσδυση, αυτός πολύ άμεσα δρα ώστε να επανέλθει στο διάδρομο (και άρα να ελαχιστοποιήσει την εφαρμοζόμενη δύναμη).

- 31ο Επεισόδιο:



(α') *Τροχιά*

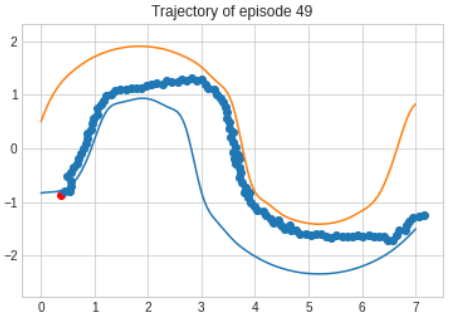


(β') *Μετρούμενη από τον πράκτορα δύναμη*

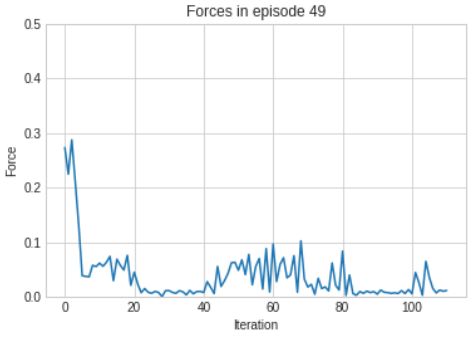
Σχήμα 5.9: *31ο επεισόδιο εκπαίδευσης (προσομοίωση)*

Και στο 31ο επεισόδιο (σχ. 5.9), όπως και στο 25ο (σχ. 5.8), παρατηρείται άμεση επαναφορά στο διάδρομο από την αρχική θέση διείσδυσης.

- 49ο Επεισόδιο:



(α') *Τροχιά*



(β') *Μετρούμενη από τον πράκτορα δύναμη*

Σχήμα 5.10: *49ο επεισόδιο εκπαίδευσης (προσομοίωση)*

Τέλος, στο 49ο επεισόδιο (σχ. 5.10) η επαναφορά στο διάδρομο από την αρχική θέση διείσδυσης φαίνεται να είναι υποβέλτιστη, αλλά η γενική συμπεριφορά είναι απόλυτα ικανοποιητική, και η εκπαίδευση έχει σχεδόν ολοκληρωθεί.

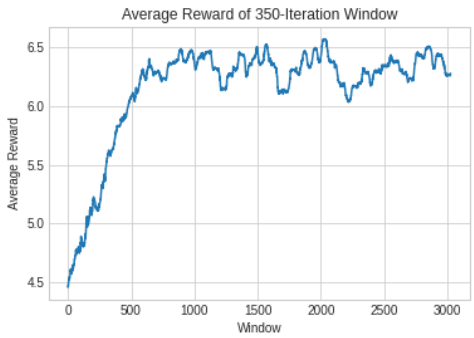
Υπενθυμίζουμε ότι ο αλγόριθμος συγκλίνει σε υποβέλτιστη λύση (βλ. υποενοτ. 3.4.2), κάτι το οποίο σημαίνει ότι δεν είναι κάθε εκπαίδευση το ίδιο επιτυχημένη, και μπορεί να παρουσιάζει διαφορετικά σημεία βελτιστότητας. Παρ' όλο που σχεδόν σε κάθε εκπαίδευση ο πράκτορας καταφέρνει να αποδράσει από το διάδρομο, η αποτελεσματικότητα στην ακολούθηση των τοίχων και η ακριβής κατεύθυνση της κίνησης όταν δε δέχεται κάποια δύναμη εξαρτώνται από τη σύγκλιση που συμβαίνει κάθε φορά. Φροντίσαμε η επιλογή των υπερπαραμέτρων να βοηθά την εκπαίδευση να πετυχαίνει όσο καλύτερη σύγκλιση γίνεται, και όλο και πιο συχνά.

Τονίζουμε επίσης ότι ο πράκτορας όταν δεν δέχεται δυνάμεις δεν ακολουθεί κάποια απόλυτα σταθερή κίνηση λόγω του θορύβου που έχουμε προσθέσει στη μέτρηση της δύναμης

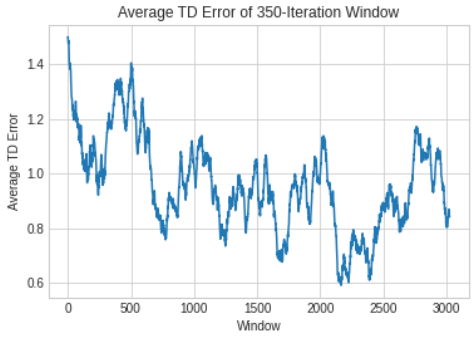
(βλ. υποενοτ. 4.5.3). Παρ' όλα αυτά, φαίνεται να εμφανίζει σημαντική ανοχή στο θόρυβο, σε βαθμό που να μην επηρεάζει αισθητά την αποτελεσματικότητα του μοντέλου.

Παρακάτω, στο σχήμα 5.11 μπορούμε να παρατηρήσουμε τις μέσες τιμές κάποιων μεγεθών της εκπαίδευσης από παράθυρα 350 βημάτων εκπαίδευσης που διαφέρουν μεταξύ τους διαδοχικά 1 βήμα (με την παραθύρωση αυτή θα εξομαλύνουμε τα γραφήματα των μεγεθών αυτών, τα οποία μπορεί να παρουσιάζουν σημαντικές διακυμάνσεις λόγω των διαφορετικών ερεθισμάτων).

(α') *Μέση μετρούμενη από τον πράκτορα δύναμη*



(β') *Μέση λαμβανόμενη ανταμοιβή*



(γ') *Μέσο TD σφάλμα*

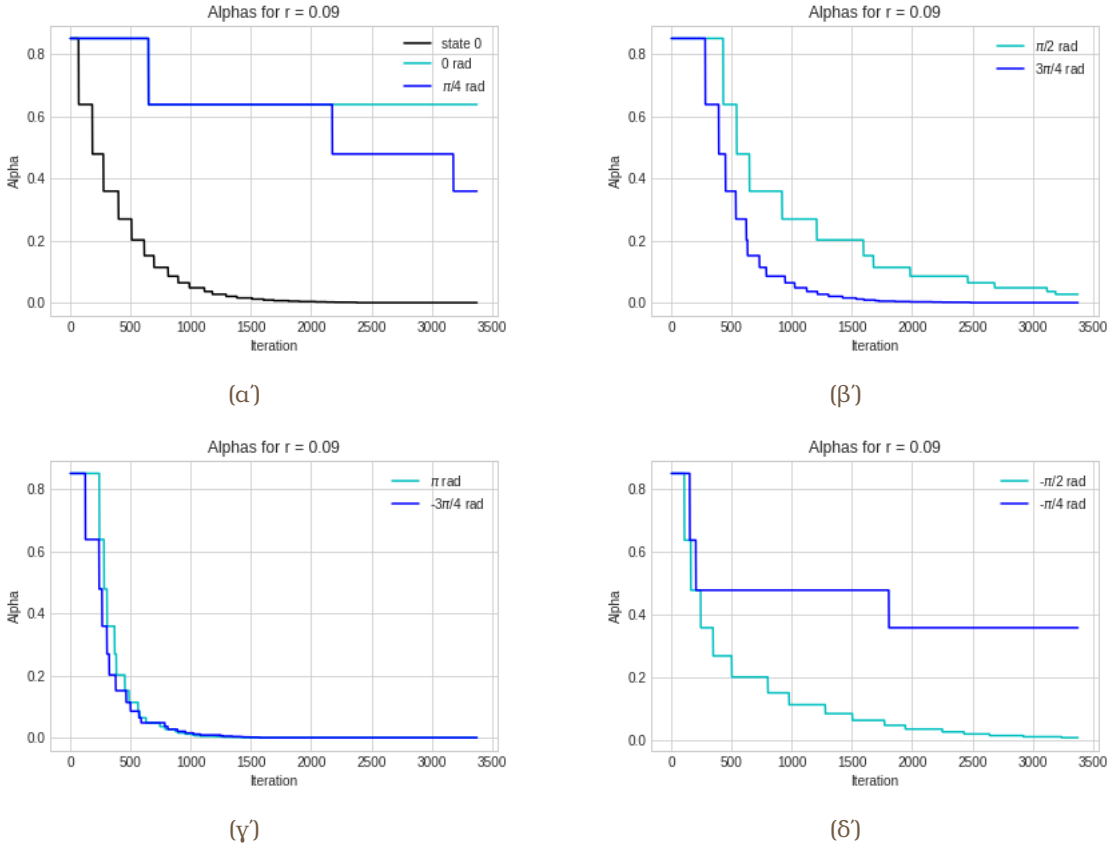
Σχήμα 5.11: *Μέσοι όροι μεγεθών σε παράθυρα 350 βημάτων (διαδοχικά κατά 1 βήμα)*

Βλέπουμε ότι πολύ γρήγορα η μέση μετρούμενη δύναμη (μαζί με το θόρυβο) και η μέση λαμβανόμενη ανταμοιβή μειώνεται και αυξάνεται αντίστοιχα. Αυτό είναι πολύ σημαντικό, καθώς φαίνεται ότι αν δεν χρειαζόμαστε μεγάλη λεπτομέρεια μπορούμε να σταματήσουμε την εκπαίδευση και σε λιγότερα επεισόδια.

Το μέσο TD σφάλμα, παρ' όλο που παρουσιάζει σημαντικές διακυμάνσεις, φαίνεται και αυτό μακροπρόθεσμα να μειώνεται.

Ακόμα, στο σχήμα 5.12 φαίνεται ενδεικτικά η μείωση των ρυθμών μάθησης α για τις διαφορετικές καταστάσεις κέντρα με μέτρο 0 και 0.09 κατά τη διάρκεια της εκπαίδευσης. Παρατηρούμε ότι παρ' όλο που θέσαμε διαφορετικό κατώφλι για κάθε μείωση, για κάποιες γωνίες ο ρυθμός μάθησης εξακολουθεί να συγκλίνει πολύ πιο γρήγορα απ' ότι άλλες. Αυτό όμως δεν μας πειράζει, αφού για τις γωνίες 0 και $\pm\frac{\pi}{4}$ η επιθυμητή συμπεριφορά είναι ξεκάθαρη από τη συνάρτηση ανταμοιβής (για δεξιά κατεύθυνση κίνησης έχουμε και μείωση της

εφαρμοζόμενης δύναμης), και ως εκ τούτου η ακριβής σύγκλιση δεν καθίσταται απαραίτητη. Και από αυτά τα σχήματα φαίνεται όμως ότι είναι δυνατόν να γίνει εκπαίδευση σε λιγότερα επεισόδια χωρίς σημαντική επιρροή στην αποτελεσματικότητα (τουλάχιστον στις καταστάσεις στις οποίες κατά κύριο λόγο βρίσκεται ο πράκτορας), εφόσον για τα κέντρα που αντιστοιχούν σε γωνίες πιο απαιτητικών κλίσεων επιτυγχάνεται γρήγορη σύγκλιση.



(α′) (β′)

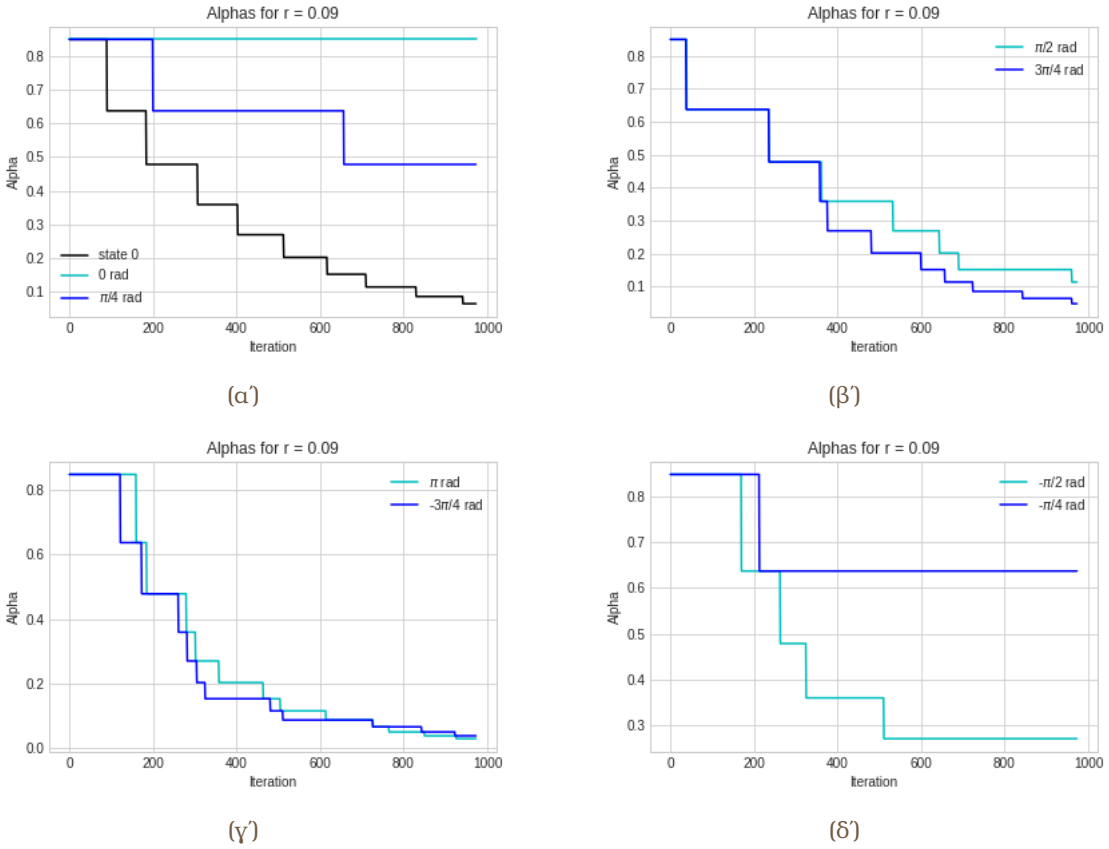
(γ′) (δ′)

Σχήμα 5.12: *Μείωση των τιμών του ρυθμού μάθησης κατά τη διάρκεια της εκπαίδευσης (μόνο για μέτρο δύναμης 0 και 0.09)*

Σημειώνουμε ότι η μείωση της εξερεύνησης ανά ¨περιοχή¨ είναι ανάλογη του παραπάνω σχήματος.

Εφόσον η υλοποίηση φαίνεται να μας το επιτρέπει, μειώνουμε τα επεισόδια μάθησης σε 12, χωρίς να αλλάξουμε κάποια παράμετρο. Στο σχήμα 5.13 παραθέτουμε ένα σχήμα ανάλογο του 5.12 για μια τέτοια εκπαίδευση. Παρατηρούμε για τις περιοχές καταστάσεων που συναντώνται πιο συχνά προλαβαίνουμε να έχουμε ικανοποιητική σύγκλιση. Παρ′ όλα αυτά, δε συμβαίνει το ίδιο για τις γωνίες $0, \pm\frac{\pi}{4}$ rad όπου έχουμε λιγότερες μειώσεις α (και άρα και ε) από πριν. Όμως, όπως προαναφέραμε, η συμπεριφορά σε κοντινές σε αυτά τα κέντρα καταστάσεις είναι αρκετά ξεκάθαρη ώστε να μη χρειάζεται μεγάλος αριθμός δειγμάτων.

Τα αποτελέσματα μιας τέτοιας εκπαίδευσης, όπως και της προηγούμενης (50 επεισόδια) θα παρουσιαστούν αναλυτικά στην επόμενη υποενότητα.



Σχήμα 5.13: *Μείωση των τιμών του ρυθμού μάθησης κατά τη διάρκεια μιας εκπαίδευσης 12 επεισοδίων (μόνο για μέτρο δύναμης 0 και 0.09)*

5.3.2 Testing

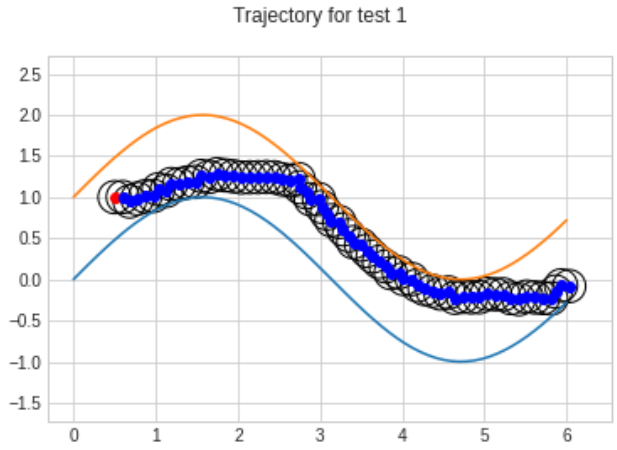
Δοκιμάζουμε το εκπαιδευμένο μοντέλο σε 4 άγνωστους για τον πράκτορα διαδρόμους, έναν ημιτονοειδή (σχ. 5.14), έναν επικλινή (σχ. 5.15), έναν κατακλινή (σχ. 5.16) και έναν οριζόντιο (σχ. 5.17). Σε κάθε μοντέλο που δοκιμάζουμε, χρησιμοποιούμε αυτούς τους 4 διαδρόμους, έτσι ώστε να έχουμε μια κοινή βάση σύγκρισης. Για αυτόν το λόγο τοποθετούμε τον πράκτορα στο ίδιο αρχικό σημείο σε κάθε δοκιμή στον ίδιο διάδρομο. Στους 3 πρώτους αυτό το σημείο βρίσκεται στην αρχή τους, χωρίς καθόλου αρχική διείσδυση. Στον τελευταίο (οριζόντιο) όμως βρίσκεται κοντά στο αριστερότερο σημείο του, και αρκετά εκτός από αυτόν, ώστε να δοκιμαστεί η ικανότητα του πράκτορα να επανέρχεται γρήγορα εντός ορίων, και στη συνέχεια να εκτελεί οριζόντια κίνηση προς τα δεξιά, παράλληλα με το διάδρομο.

Κατα το testing δεν εξερευνούμε καθόλου το χώρο, καθώς θέλουμε να αξιολογήσουμε αυτά που έχει μάθει. Επιπλέον, χρησιμοποιούμε έναν σταθερό μικρό ρυθμό μάθησης $a = 0.085$.

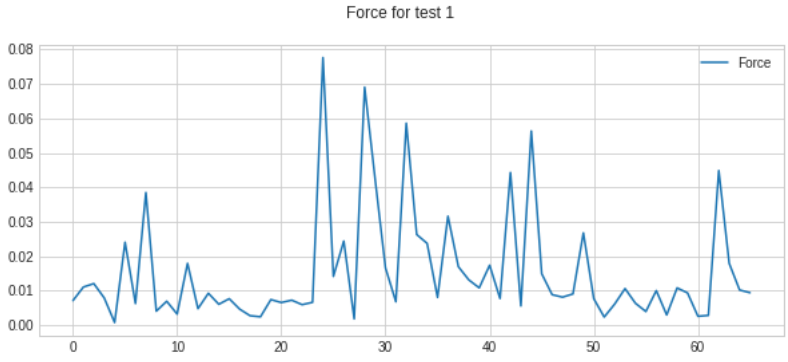
Αρχικά, παρουσιάζουμε το testing για την εκπαίδευση των 50 επεισοδίων που αναλύσαμε στην προηγούμενη υποενότητα. Τα αποτελέσματα είναι πολύ ικανοποιητικά σε όλες τις περιπτώσεις, με τον πράκτορα να ακολουθεί ομαλά την τροχιά του διαδρόμου και να οδηγείται γρήγορα στην έξοδο, να δέχεται αρκετά μικρές δυνάμεις, και να λαμβάνει μεγάλες ανταμοιβές.

Σημειώνουμε ότι στα γραφήματα των τροχιών του πράκτορα στο testing σχεδιάζουμε όλο

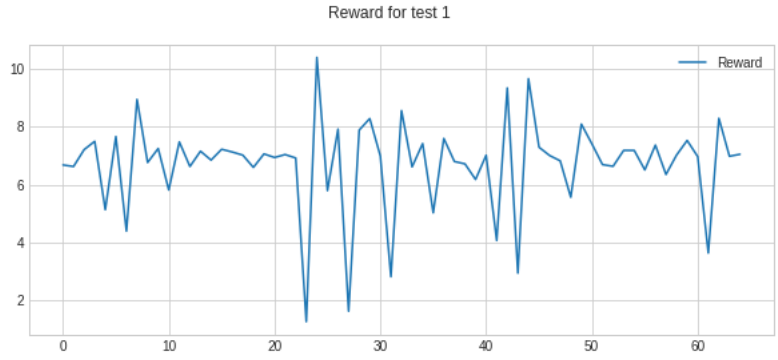
το κυκλικό αντικείμενο και όχι μόνο το κέντρο του, όπως στην εκπαίδευση, για καλύτερη οπτικοποίηση.



(α') *Τροχιά*



(β') *Δύναμη*

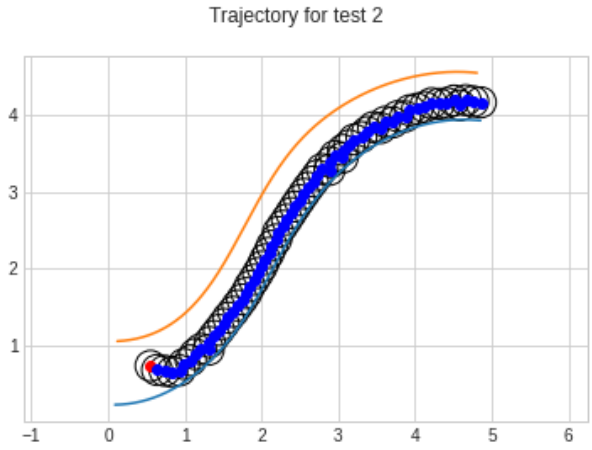


(γ') *Ανταμοιβή*

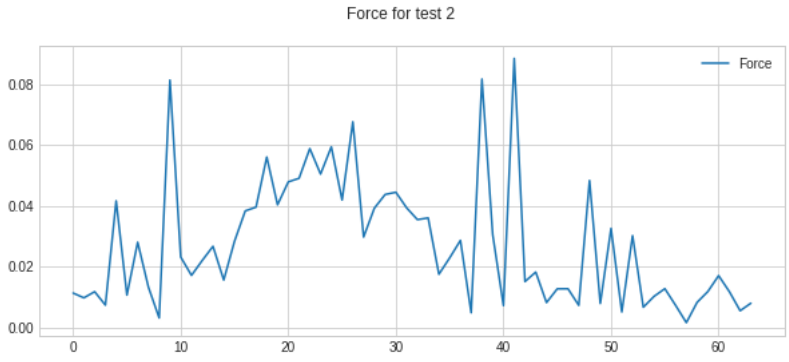
Σχήμα 5.14: *Testing σε ημιτονοειδή άγνωστο διάδρομο*

Στον ημιτονοειδή διάδρομο (σχ. 5.14) παρατηρούμε ότι ο πράκτορας μόλις σε περίπου 65 επεισόδια καταφέρνει να δραπετεύσει από το διάδρομο, εκτελώντας ακολούθηση του τοίχου που βρίσκεται κάθε φορά στα δεξιά του, και κινούμενος οριζόντια προς τα δεξιά όταν δεν είναι δίπλα σε τοίχο (έχει δηλαδή την αναμενόμενη συμπεριφορά). Βλέποντας το διάγραμμα της μετρούμενης δύναμης, γίνεται φανερό ότι η δύναμη διατηρείται σε χαμηλά επίπεδα, με

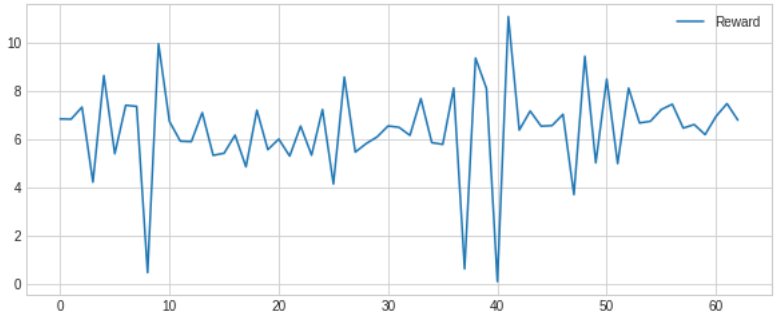
μόνο κάποια μεμονωμένα spikes, που αντιστοιχούν σε υποβέλτιστη τροχιά, η οποία παρ' όλα αυτά αποδεικνύεται αρκετά καλή για τη γρήγορη ολοκλήρωση της δοκιμής.



(α') *Τροχιά*



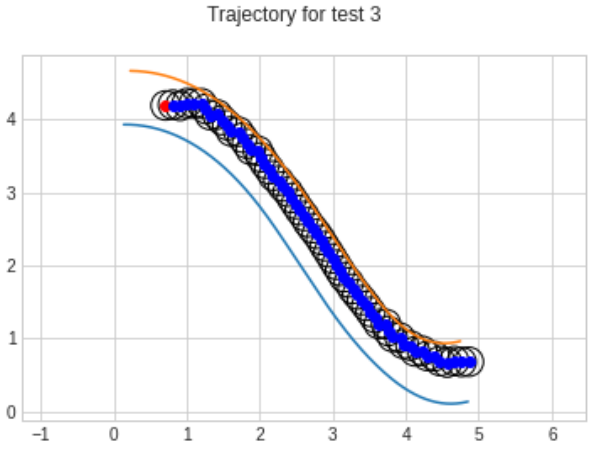
(β') *Δύναμη*



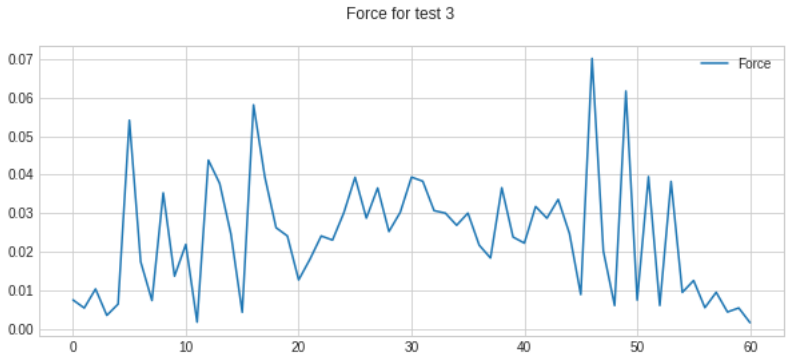
(γ') *Ανταμοιβή*

Σχήμα 5.15: *Testing σε επικλινή άγνωστο διάδρομο*

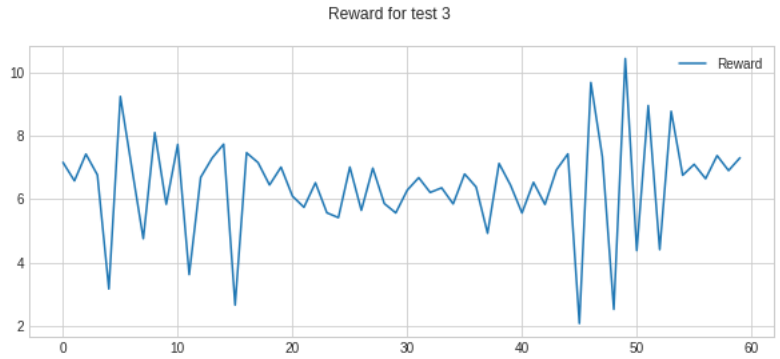
Στον επικλινή διάδρομο (σχ. 5.15) παρουσιάζεται η ικανότητα του πράκτορα να ακολουθεί μια δύσκολη τροχιά (βάσει της τάσης κίνησής του), και μάλιστα αρκετά λεπτομερώς σε μεγάλο μέρος του διαδρόμου. Αυτό γίνεται εμφανές, πέρα από το σχήμα της τροχιάς, και το σχήμα της δύναμης, όπου από το 11ο μέχρι περίπου το 38ο επεισόδιο δεν παρατηρούμε μεγάλα spikes.



(α΄) *Τροχιά*



(β΄) *Δύναμη*



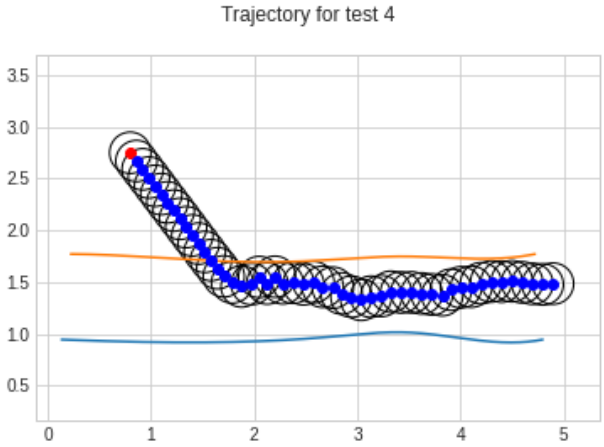
(γ΄) *Ανταμοιβή*

Σχήμα 5.16: *Testing σε κατακλινή άγνωστο διάδρομο*

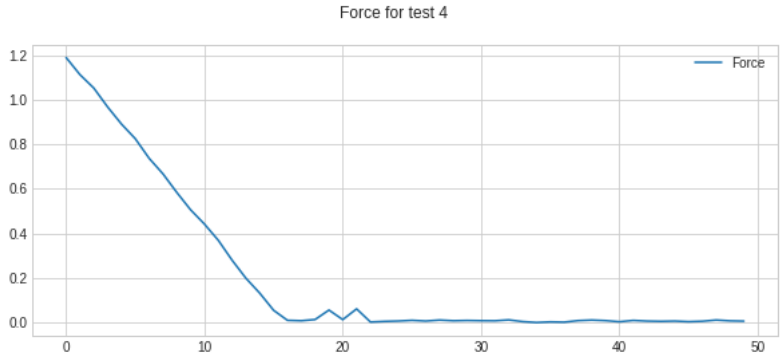
Στον κατακλινή διάδρομο (σχ. 5.16) έχουμε παρόμοια αποτελέσματα με τον επικλινή, και μάλιστα με ακόμα καλύτερη ακολούθηση του τοίχου.

Συμπεραίνουμε ότι η λεπτομερής ακολούθηση τοίχου, τόσο στον επικλινή όσο και στον κατακλινή διάδρομο γίνεται στο μέρος όπου η κλίση πλησιάζει περισσότερο την κατακόρυφο. Αυτό μπορεί να φαίνεται παράδοξο, αλλά δικαιολογείται από το γεγονός ότι ο διάδρομος εκπαίδευσης (βλ. σχ. 5.1) έχει αρκετά μεγάλο μέρος τέτοιων τροχιών, και άρα ο πράκτορας δέχεται περισσότερα δείγματα τέτοιων κλίσεων. Ένας ακόμα λόγος είναι ότι ο πράκτορας αρχικά αφιερώνει περισσότερα βήματα σε τέτοιες κλίσεις, ακριβώς λόγω της δυσκολίας α-

πόδρασης από αυτές, σε σύγκριση με τις κλίσεις που πλησιάζουν περισσότερο τον οριζόντιο άξονα). Ο τελευταίος είναι και ο λόγος που δεν μας πειράζει που ο αλγόριθμος έχει συγκλίνει καλύτερα σε αυτές τις τροχιές (εφόσον στις άλλες μπορεί να δραπετεύσει γρήγορα έτσι κι αλλιώς).



(α') *Τροχιά*



(β') *Δύναμη*

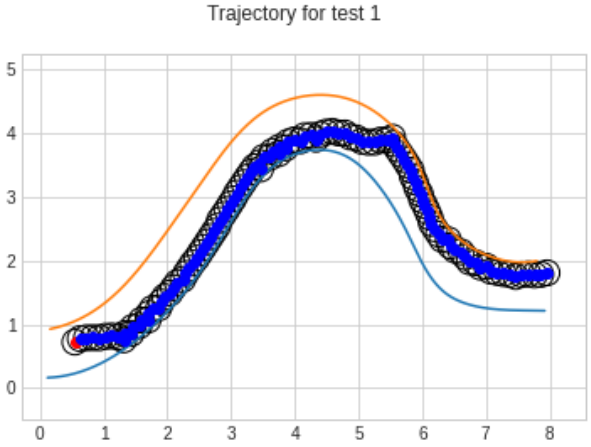
(γ') *Ανταμοιβή*

Σχήμα 5.17: *Testing σε οριζόντιο άγνωστο διάδρομο, με σημείο εκκίνησης εκτός αυτού*

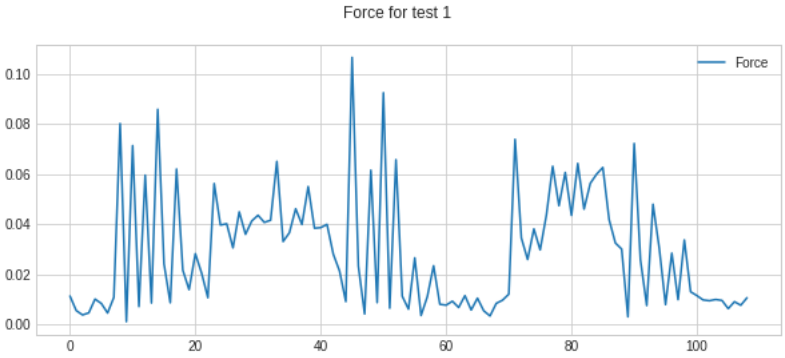
Στον οριζόντιο διάδρομο (σχ. 5.17) παρατηρούμε ότι ο πράκτορας άμεσα επιστρέφει εντός ορίων, κρατώντας μια μικρή τάση κίνησης προς τα δεξιά (όπως και θέλουμε), και στη συνέχεια εκτελεί μια αρκετά ικανοποιητική οριζόντια τροχιά, δεδομένου του θορύβου, στον

οποίο δείχνει να έχει αρκετή ανοχή, και της υποβελτιστότητας της σύγκλισης (από φύση του αλγορίθμου).

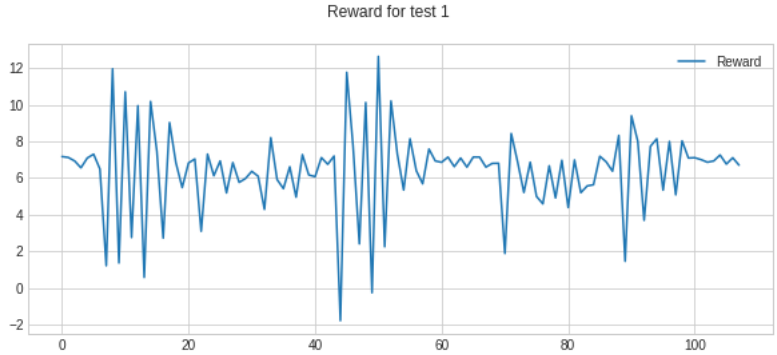
Για τη συγκεκριμένη εκπαίδευση παραθέτουμε και μια ακόμα δοκιμή (σχ. 5.18), σε διάδρομο αρχικά επικλινή και στη συνέχεια κατακλινή, για να αποδείξουμε ότι το μοντέλο δεν επηρεάζεται ιδιαίτερα από το περιβάλλον του κατά το testing, αλλά αντίθετα ο πράκτορας έχει την ικανότητα να αντιμετωπίζει μεγάλες αλλαγές στην κλίση του διαδρόμου.



(α') *Τροχιά*



(β') *Δύναμη*

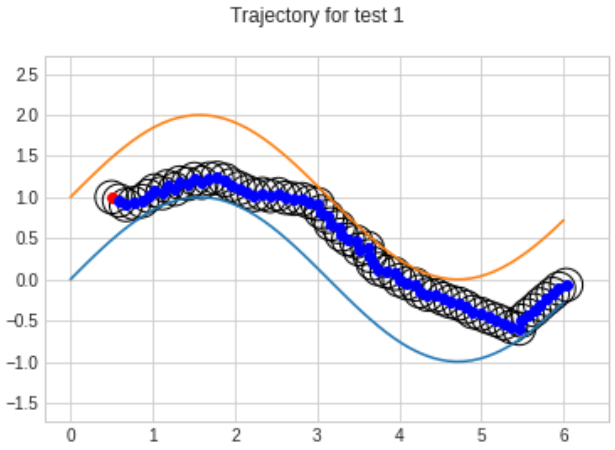


(γ') *Ανταμοιβή*

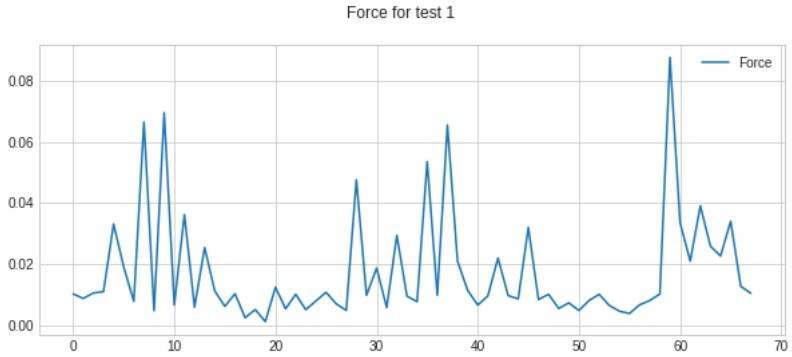
Σχήμα 5.18: *Testing σε άγνωστο διάδρομο που παρουσιάζει μεγάλη αλλαγή κλίσης*

Παρουσιάζουμε τώρα τα αποτελέσματα του testing στους 4 πρώτους διαδρόμους για την

εκπαίδευση των 12 επεισοδίων.



(α') *Τροχιά*

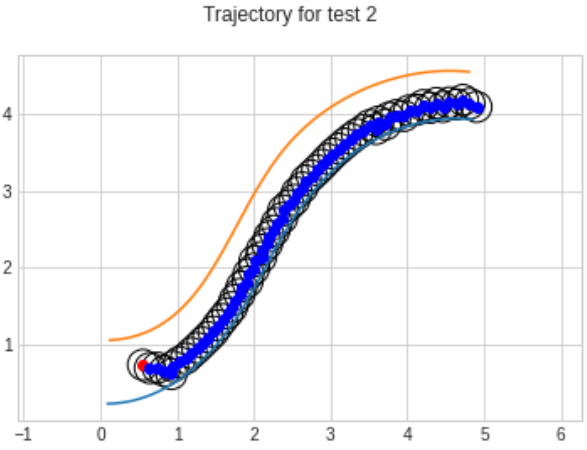


(β') *Δύναμη*

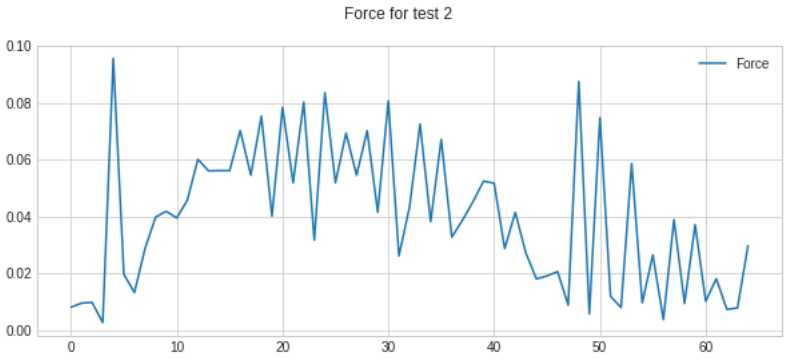
(γ') *Ανταμοιβή*

Σχήμα 5.19: *Testing σε ημιτονοειδή άγνωστο διάδρομο (εκπαίδευση 12 επεισοδίων)*

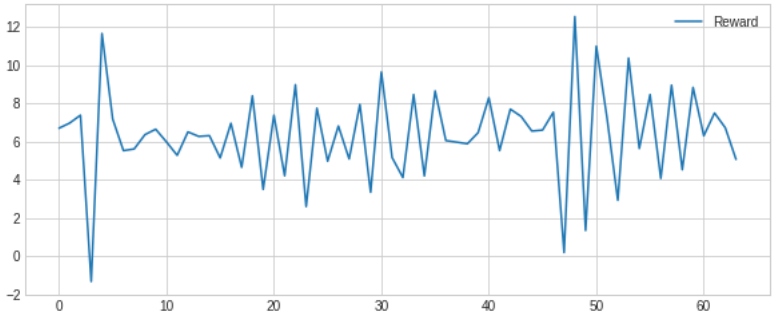
Στον ημιτονοειδή διάδρομο (σχ. 5.19) παρατηρούμε παρόμοια συμπεριφορά με την αντίστοιχη προηγούμενη (βλ. σχ. 5.14), με την ολοκλήρωση του task να διαρκεί και στις δύο περιπτώσεις περίπου 65 βήματα.



(α') *Τροχιά*



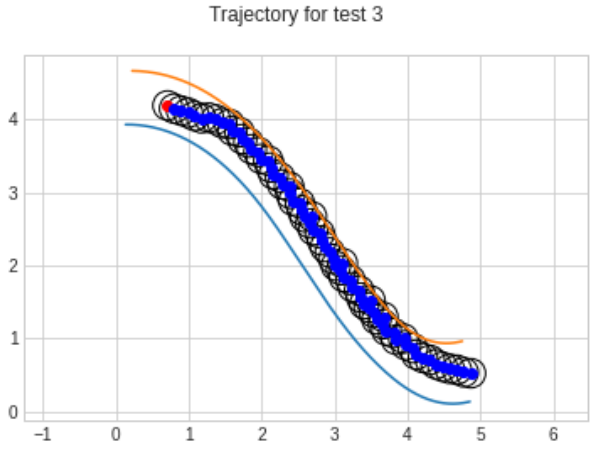
(β') *Δύναμη*



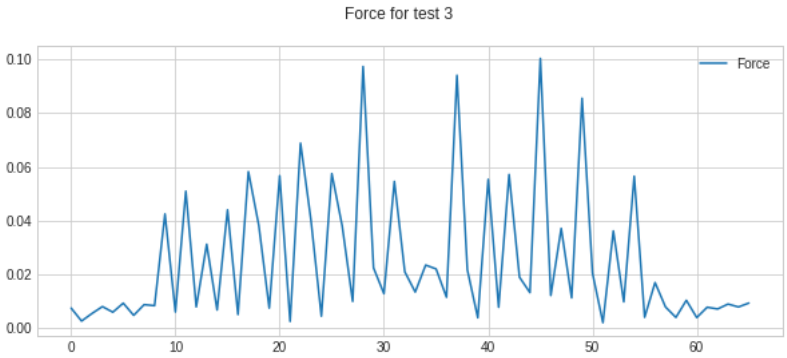
(γ') *Ανταμοιβή*

Σχήμα 5.20: *Testing σε επικλινή άγνωστο διάδρομο (εκπαίδευση 12 επεισοδίων)*

Στον επικλινή διάδρομο (σχ. 5.20) μπορεί να παρατηρούμε μια αρκετά μικρή αύξηση της δύναμης σε συγκεκριμένα σημεία σε σχέση με την προηγούμενη αντίστοιχη δοκιμή (βλ. σχ. 5.15), αλλά σα σύνολο η επίδοση φαίνεται να είναι καλή, διατηρώντας τα απαιτούμενα βήματα για την ολοκλήρωση σχετικά σταθερά ανάμεσα στις 2 δοκιμές (περίπου 65).



(α′) *Τροχιά*

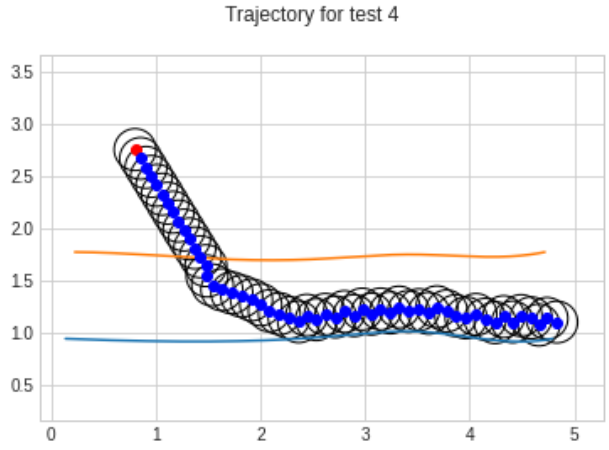


(β′) *Δύναμη*

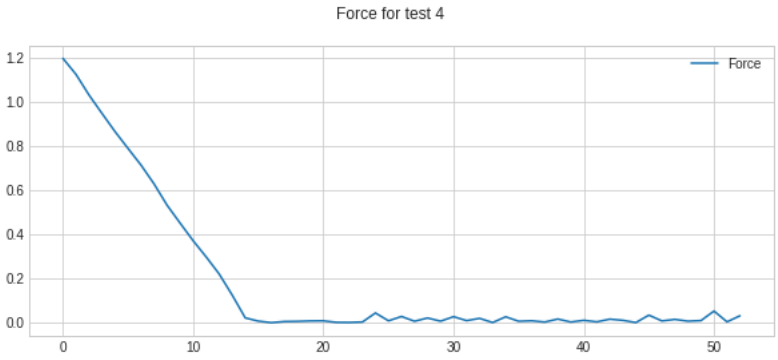
(γ′) *Ανταμοιβή*

Σχήμα 5.21: *Testing σε κατακλινή άγνωστο διάδρομο (εκπαίδευση 12 επεισοδίων)*

Στον κατακλινή διάδρομο (σχ. 5.21) βλέπουμε σχετική αύξηση της εφαρμοζόμενης δύναμης (και περισσότερα spikes) σε σχέση με την προηγούμενη αντίστοιχη δοκιμή (βλ. σχ. 5.16), καθώς και αύξηση των απαιτούμενων βημάτων (περίπου κατά 5). Παρ' όλα αυτά, ο πράκτορας εξακολουθεί να βρίσκει την έξοδο σχετικά γρήγορα. Η πτώση στην επίδοση αυτή είναι πιθανόν να οφείλεται σε μεμονωμένη υποβέλτιστη σύγκλιση, και δεν αποτελεί σημαντικό πρόβλημα.



(α΄) *Τροχιά*



(β΄) *Δύναμη*

(γ΄) *Ανταμοιβή*

Σχήμα 5.22: *Testing σε οριζόντιο άγνωστο διάδρομο, με σημείο εκκίνησης εκτός αυτού (εκπαίδευση 12 επεισοδίων)*

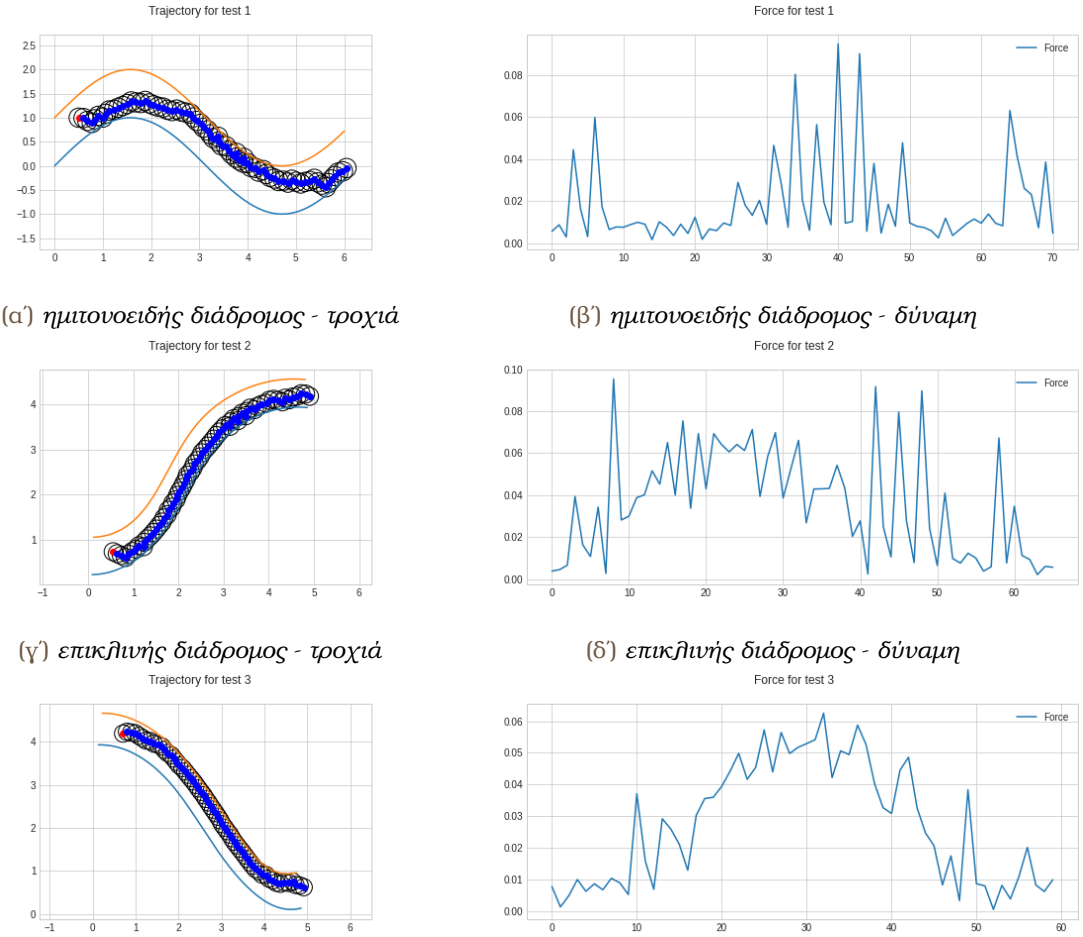
Τέλος, καλή συμπεριφορά βλέπουμε και στον οριζόντιο διάδρομο (σχ. 5.22), παρ' όλο που στην κατάσταση μηδενικής δύναμης δεν έγινε σύγκλιση στην ακριβή οριζόντια κίνηση προς τα δεξιά, αλλά κοντά σε αυτή (ως εκ τούτου και η μικρή αύξηση στα απαιτούμενα βήματα σε σχέση με την προηγούμενη αντίστοιχη δοκιμή, βλ. σχ. 5.17).

5.3.3 Αλλαγή Υπερπαραμέτρων

Σε αυτή την υποενότητα θα παρουσιάσουμε αποτελέσματα του testing μοντέλων που εκπαιδεύτηκαν σε 50 επεισόδια με μεταβολές σε κάποιες παραμέτρους, για να παρατηρήσουμε πως θα συμπεριφερθεί ο πράκτορας.

Θέλουμε η μέθοδός μας να είναι robust, δηλαδή να μην επηρεάζεται από συγκριτικά μικρές διαφορές στις παραμέτρους της μάθησης.

- Για αύξηση του σ_{00_s} σε 0.025 και του σ_{11_s} σε 0.67, όπως βλέπουμε στο σχήμα το μοντέλο εξακολουθεί να διατηρεί σε μεγάλο βαθμό την αποτελεσματικότητά του.



(α′) *ημιτονοειδής διάδρομος - τροχιά*

(β′) *ημιτονοειδής διάδρομος - δύναμη*

(γ′) *επικλινής διάδρομος - τροχιά*

(δ′) *επικλινής διάδρομος - δύναμη*

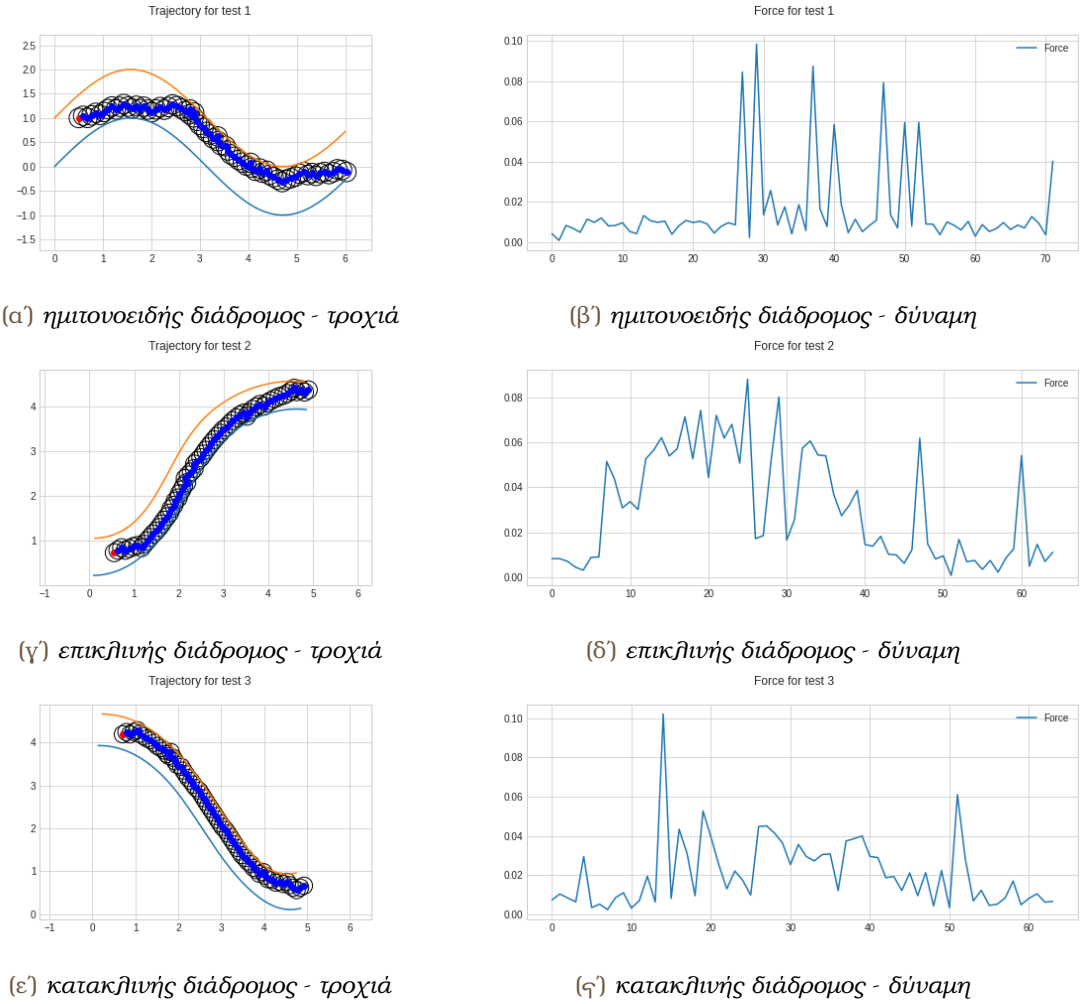
(ε′) *κατακλινής διάδρομος - τροχιά*

(ϛ′) *κατακλινής διάδρομος - δύναμη*

Σχήμα 5.23: *Testing για* $\sigma_{00_s} = 0.025$ *και* $\sigma_{11_s} = 0.67$

Μπορούμε να παρατηρήσουμε τόσο μια μικρή αύξηση της εφαρμοζόμενης δύναμης όσο και των συνολικών βημάτων κάθε δοκιμής. Παρ' όλα αυτά, παρατηρούνται σημεία πολύ καλής σύγκλισης (ιδιαίτερα σε κλίσεις πιο κοντά στην κατακόρυφο που αποτελούν, όπως προαναφέραμε, την πλειοψηφία των δειγμάτων) και συνολική ευκολία στην ολοκλήρωση του task.

- Εξίσου καλά αποτελέσματα παρατηρούμε για αύξηση του σ_{11_a} σε 0.72 (σχ. 5.24).



(α΄) *ημιτονοειδής διάδρομος - τροχιά*

(β΄) *ημιτονοειδής διάδρομος - δύναμη*

(γ΄) *επικλινής διάδρομος - τροχιά*

(δ΄) *επικλινής διάδρομος - δύναμη*

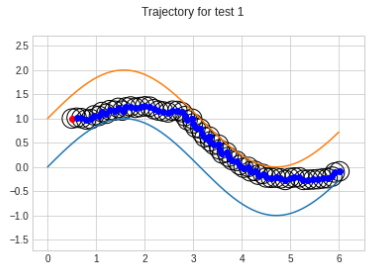
(ε΄) *κατακλινής διάδρομος - τροχιά*

(ϛ΄) *κατακλινής διάδρομος - δύναμη*

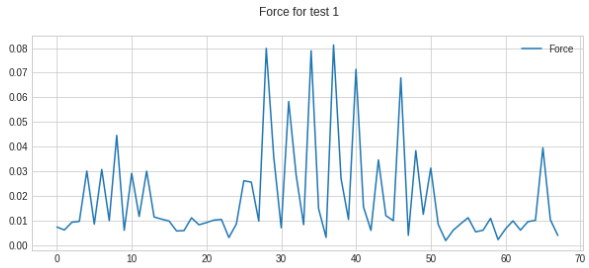
Σχήμα 5.24: *Testing για* $\sigma_{11_a} = 0.72$

Οι παρατηρήσεις μας για αυτές τις δοκιμές ταιριάζουν αρκετά με τις προηγούμενες.

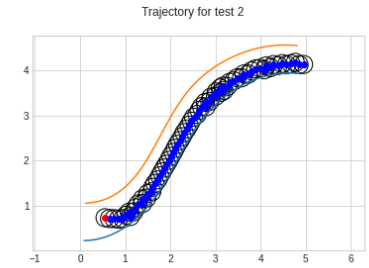
- Τέλος, παρατηρούμε στο σχήμα 5.25 ότι ούτε μια μικρή αύξηση του αρχικού ρυθμού μάθησης α σε 0.92 επηρεάζει αρνητικά σε μεγάλο βαθμό.



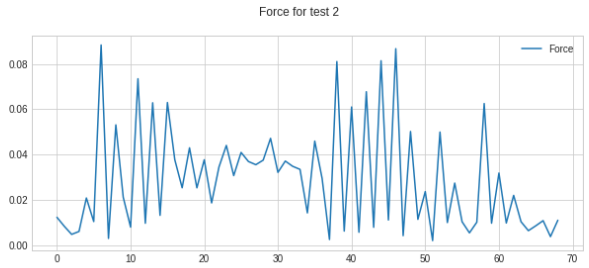
(α′) *ημιτονοειδής διάδρομος - τροχιά*



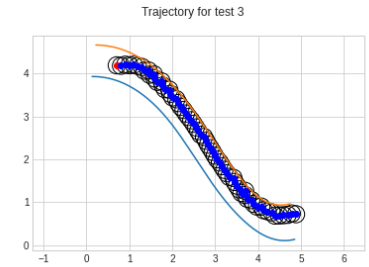
(β′) *ημιτονοειδής διάδρομος - δύναμη*



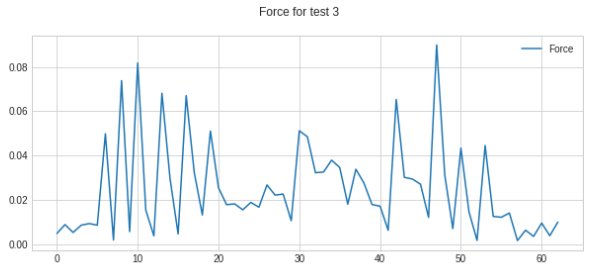
(γ′) *επικλινής διάδρομος - τροχιά*



(δ′) *επικλινής διάδρομος - δύναμη*



(ε′) *κατακλινής διάδρομος - τροχιά*



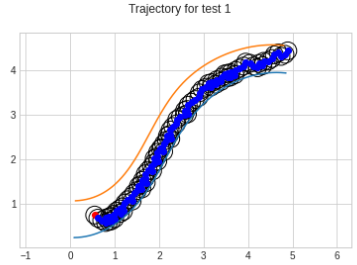
(ϛ′) *κατακλινής διάδρομος - δύναμη*

Σχήμα 5.25: *Testing για αρχικό* $a = 0.92$

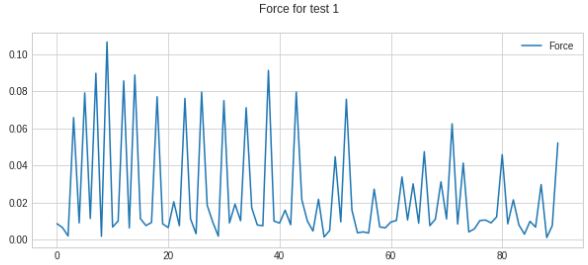
Και εδώ είναι εμφανείς αρκετές ατέλειες σε σχέση με τις δοκιμές με τις επιλεγείσες υπερπαραμέτρους. Ένας λόγος που πιθανόν να συμβαίνουν αυτές είναι οι μεγαλύτερες ανανεώσεις βαρών λόγω της αύξησης του ρυθμού μάθησης, κάτι που μπορεί να έχει επίδραση στη σύγκλιση. Η επίδοση, σε κάθε περίπτωση, παραμένει ικανοποιητική.

Άλλες μεταβολές, όμως, αρκετά μεγαλύτερης κλίμακας μπορούν να επηρεάσουν σημαντικά την απόδοση του αλγορίθμου. Ενδεικτικά:

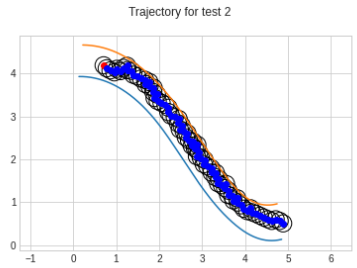
- Για μείωση της παραμέτρου k_1 της συνάρτησης επιβράβευσης (βλ. ενοτ. 4.2) από 7 σε 4, που υποδηλώνει σημαντική μείωση της συνεισφοράς της τάσης κίνησης προς τα δεξιά σε σχέση με τη συνεισφορά της αντίδρασης στις εφαρμοζόμενες δυνάμεις, η αποτελεσματικότητα μειώνεται αισθητά, καθώς παρ' όλο που ο πράκτορας καταφέρνει να δραπετεύσει από το διάδρομο, ταλαντώνεται συχνά ανάμεσα σε καταστάσεις μηδενικών και μη δυνάμεων (βλ. σχ. 5.26). Αυτό συμβαίνει διότι λαμβάνει μεγαλύτερη ανταμοιβή αν ελαττώνει με το γρηγορότερο τρόπο ακόμα και σχετικά μικρές δυνάμεις, από ότι αν εκτελέσει wall following διατηρώντας την τάση προς τα δεξιά και κρατώντας την δύναμη συνεχώς αρκετά μικρή.



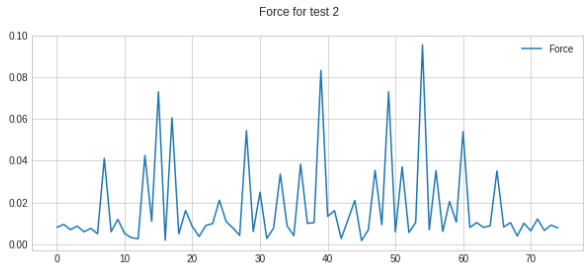
(α') *επικλινής διάδρομος - τροχιά*



(β') *επικλινής διάδρομος - δύναμη*



(γ') *κατακλινής διάδρομος - τροχιά*

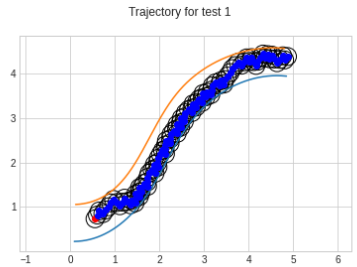


(δ') *κατακλινής διάδρομος - δύναμη*

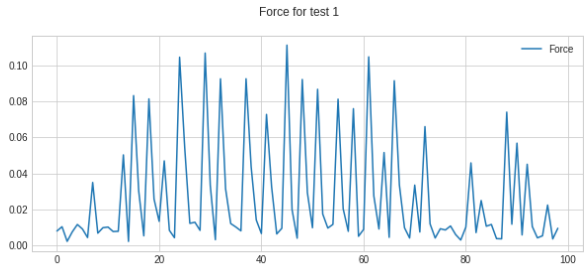
Σχήμα 5.26: *Testing για* $k_1 = 4$

Παρατηρείται μεγάλος αριθμός βημάτων συγκριτικά με τις προηγούμενες δοκιμές, καθώς και μεγάλα μέτρα εφαρμοζόμενων δυνάμεων με πολλά spikes.

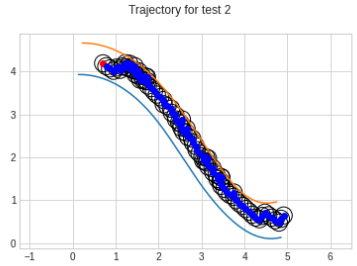
- Τέλος, παρόμοια αποτελέσματα εμφανίζονται και για πίνακα συνδιακύμανσης της κατάστασης μηδενικής δύναμης ίσο με των υπόλοιπων κέντρων (βλ. σχ. 5.27, όπου έχουμε πολύ μεγαλύτερη επικάλυψη μεταξύ της RBF της κατάστασης αυτής με τις άλλες, και άρα μεγαλύτερες εξαρτήσεις μεταξύ των καταστάσεων αυτών.



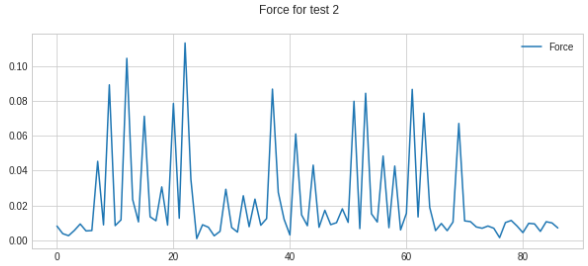
(α') *επικλινής διάδρομος - τροχιά*



(β') *επικλινής διάδρομος - δύναμη*



(γ') *κατακλινής διάδρομος - τροχιά*



(δ') *κατακλινής διάδρομος - δύναμη*

Σχήμα 5.27: *Testing για κοινό* $\mathbf{\Sigma}$ *για όλες τις καταστάσεις-κέντρα*

Μάλιστα, μαζί με την ίδια αναποτελεσματικότητα της προηγούμενης περίπτωσης, εμφανίζεται και μεγάλη αστάθεια στη συμπεριφορά του πράκτορα όταν δεν δέχεται κάποια δύναμη, καθώς και επιρροή από το θόρυβο.

5.3.4 Συνεισφορά της προσαρμοστικής εξερεύνησης & ρυθμού μάθησης

Σε αυτή την υποενότητα θα παρουσιάσουμε τη βελτίωση που παρατηρείται τόσο στην εκπαίδευση όσο και στο testing με χρήση της προσαρμοστικής εξερεύνησης και ρυθμού μάθησης (βλ. ενοτ. 4.3, υποενοτ. 4.5.2).

Συγκεκριμένα, εκτελούμε 10 εκπαιδεύεις 12 επεισοδίων με και χωρίς τη χρήση αυτής της μεθόδου, μαζί με testing στους 4 βασικούς διαδρόμους που έχουμε αναφέρει (βλ. υποενοτ. 5.3.2) και παραθέτουμε στατιστικά στοιχεία που προκύπτουν από αυτές. Στις εκπαιδεύσεις χωρίς χρήση της προσαρμοστικής μεθόδου, ο ρυθμός μάθησης και η εξερεύνηση είναι μειούμενοι με το πέρασμα των επεισοδίων (default μέθοδος).

Αρχικά, αναφέρουμε ότι στην υλοποίηση με προσαρμοστικά ε, α, στις 10 εκπαιδεύσεις το πλήθος επεισοδίων εκπαίδευσης όπου ο πράκτορας βρίσκει την έξοδο έχει μέσο όρο $\mu = 9.7$ και τυπική απόκλιση $\sigma = 1$, έναντι $\mu = 9.1$ και $\sigma = 1.375$ χωρίς αυτά. Φαίνεται λοιπόν ότι με τη μέθοδο που χρησιμοποιούμε, εμφανίζεται ένα μεγαλύτερο ποσοστό επιτυχίας κατά τη διάρκεια της εκπαίδευσης, μάλιστα με μικρότερες αποκλίσεις από εκπαίδευση σε εκπαίδευση.

Παρακάτω, στον πίνακα 5.2 βλέπουμε ανάλογα στατιστικά στοιχεία σχετικά με το πλήθος βημάτων στην εκπαίδευση και στις δοκιμές.

Πίνακας 5.2: *Στατιστικά στοιχεία πλήθους βημάτων για εκπαιδεύσεις με και χωρίς προσαρμοστικά ε, α*

Προσαρμοστικά ε, α	Training	Ημιτονοειδής	Επικλινής	Κατακλινής	Οριζόντιος
Όχι	μ=1311.6 σ=182.87	μ=80 σ=6.309	μ=71.1 σ=5.924	μ=73.3 σ=7.058	μ=65.2 σ=11.7
Ναι	μ=1111.4 σ=163.7	μ=71.2 σ=6.18	μ=70.3 σ=9.92	μ=66.7 σ=3.69	μ=53.3 σ=5.88

Από τους παραπάνω αριθμούς γίνεται κατανοητό ότι με χρήση προσαρμοστικών ε, α μειώνεται αισθητά ο συνολικός αριθμός βημάτων κατά την εκπαίδευση, κάνοντας έτσι τη μάθηση γρηγορότερη. Επίσης, φαίνεται να μειώνεται και η απόκλιση από εκπαίδευση σε εκπαίδευση, κάτι πολύ σημαντικό αν δεν επιθυμούμε να κάνουμε διαδοχικές εκπαιδεύσεις μέχρι να πετύχουμε μια καλή σύγκλιση.

Παρόμοιες παρατηρήσεις έχουμε και για τα στατιστικά στοιχεία των δοκιμών, κάτι που σηματοδοτεί την καλύτερη επίδοση του πράκτορα με χρήση προσαρμοστικών ε, α, με μικρότερη απόκλιση μεταξύ διαφορετικών δοκιμών (και εκπαιδεύσεων) σε συγκεκριμένο διάδρομο. Εξαίρεση στο τελευταίο αποτελούν οι δοκιμές σε επικλινή διάδρομο όπου έχουμε μεγαλύτερο σ, κάτι όμως που οφείλεται σε μια μεμονωμένη αρκετά κακή σύγκλιση του αλγορίθμου.

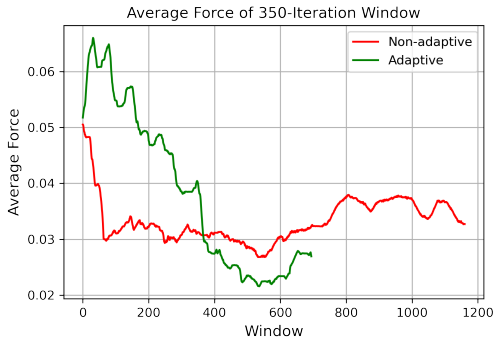
Επιπλέον, υπολογίζουμε αυτά τα στατιστικά και για κάποιους μέσους όρους στοιχείων του αλγορίθμου κατά το testing (πίνακας 5.3).

Πίνακας 5.3: *Στατιστικά στοιχεία μέσων δυνάμεων και ανταμοιβών για εκπαιδεύσεις με και χωρίς προσαρμοστικά ε, α (testing)*

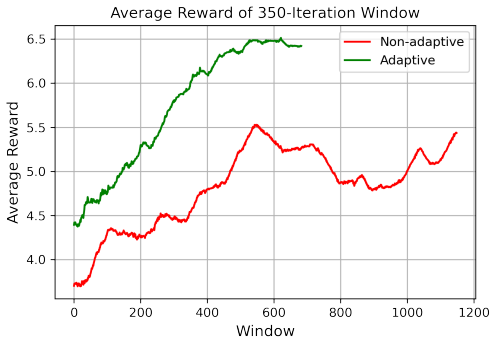
Προσαρμοστικά ε, α	Μέση δύναμη	Μέση ανταμοιβή
Όχι	μ=0.05 σ=0.00354	μ=6.529 σ=0.2
Ναι	μ=0.058 σ=0.0096	μ=6.73 σ=0.154

Παρατηρούμε αρκετά παρόμοια στοιχεία στις 2 υλοποιήσεις. Πιο συγκεκριμένα, η χρήση προσαρμοστικών ε, α καταφέρνει να αυξάνει ελαφρώς τη μέση ανταμοιβή, μειώνοντας επίσης την απόκλιση μεταξύ των εκπαιδεύσεων. Από την άλλη, η εφαρμοζόμενη δύναμη και η αντίστοιχη απόκλιση αυξάνονται ελαφρώς, αλλά αυτό δεν σηματοδοτεί απαραίτητα κάτι αρνητικό, αλλά πιθανόν να συμβαίνει λόγω του γεγονότος ότι όταν έχουμε μια λεπτομερή ακολούθηση τοίχου, ο πράκτορας δέχεται συνέχεια δυνάμεις (ελεγχόμενες), σε σχέση με άλλες χειρότερες συγκλίσεις.

Τέλος, για τη σύγκριση των δύο μεθόδων θα παραθέσουμε μαζί γραφήματα μέσων όρων στοιχείων μιας μέσης μάθησης για κάθε μέθοδο σε παράθυρα 350 βημάτων, με τη διαδικασία που περιγράψαμε για το σχήμα 5.11.



(α') *Μέση μετρούμενη από τον πράκτορα δύναμη*



(β') *Μέση λαμβανόμενη ανταμοιβή*

Σχήμα 5.28: *Σύγκριση μέσων όρων μεγεθών σε παράθυρα 350 βημάτων (διαδοχικά κατά 1 βήμα), σε μάθηση 12 επεισοδίων, με και χωρίς χρήση προσαρμοστικών ε, α*

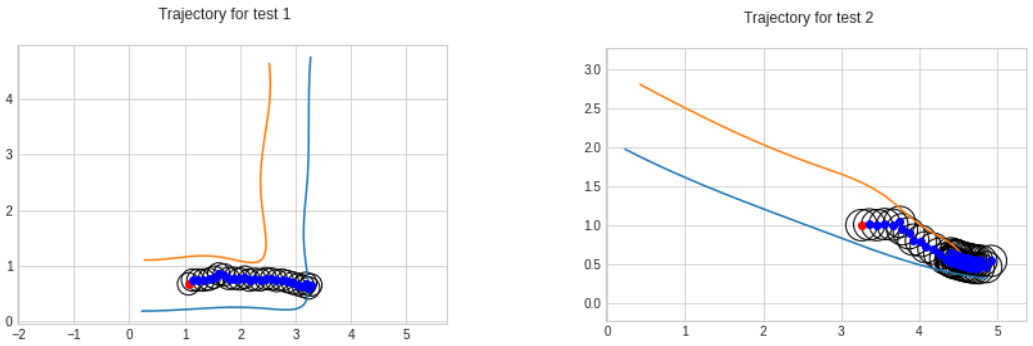
Παρατηρώντας τα δύο παραπάνω σχήματα μπορούμε να εξάγουμε το συμπέρασμα ότι η χρήση προσαρμοστικής εξερεύνησης και ρυθμού μάθησης προσφέρει γρηγορότερη και καλύτερη σύγκλιση, η οποία φαίνεται μέσω της αποτελεσματικότερης μείωσης της δύναμης και αύξησης της ανταμοιβής (μπορούμε να παρατηρήσουμε την ύπαρξη διαφορετικού πλήθους βημάτων, και άρα παραθύρων, σε κάθε μέθοδο).

5.3.5 Περιορισμοί της υλοποίησης

Παρ' όλα τα θετικά στοιχεία της υλοποίησής μας, υπάρχουν περιπτώσεις στις οποίες δεν μπορεί να ανταπεξέλθει, λόγω της αντίθεσής τους με τις παραδοχές που έχουμε πάρει από την αρχή της.

Έτσι, ο πράκτορας αδυνατεί να διασχίσει έναν κατακόρυφο διάδρομο, αφού πρώτον δεν έχει εκπαιδευτεί σε τέτοια κλίση, δεύτερον η κλίση αυτή αντιτίθεται πολύ στη φυσική τάση κίνησής του, και τρίτον (και σημαντικότερο) δεν μπορεί να διαχωρίσει αν η καλύτερη συμπεριφορά είναι η κίνηση προς τα πάνω ή προς τα κάτω, αφού γνωρίζει πως η έξοδος βρίσκεται προς τα δεξιά. Μια τέτοια δοκιμή φαίνεται στο σχήμα 5.29α'.

Ένας άλλος περιορισμός της υλοποίησης είναι ότι ο πράκτορας δεν μπορεί να συμπεριφερθεί σωστά σε οποιονδήποτε διάδρομο αποκτά τροχιά προς τα αριστερά, ακόμα και για μικρό διάστημα. Αυτό συμβαίνει γιατί, σε τέτοια περίπτωση, ο πράκτορας θα ακολουθήσει τον τοίχο ανάποδα, σύμφωνα με τη φυσική του τάση προς τα δεξιά. Μπορούμε να δούμε αυτή τη συμπεριφορά στο σχήμα 5.29β', όπου τελικά ο πράκτορας παγιδεύεται στην αρχή του διαδρόμου.



(α') *Κατακόρυφος διάδρομος*

(β') *Διάδρομος με τροχιά προς τα αριστερά*

Σχήμα 5.29: *Αποτυχίες της υλοποίησης*

Κεφάλαιο 6

Πειραματικά Αποτελέσματα

Μετά την αξιολόγηση της υλοποίησης σε περιβάλλον προσομοίωσης, σειρά παίρνει η διεξαγωγή πειράματος σε πραγματική διάταξη

6.1 Ρομποτικός Βραχίονας, Χώρος Εργασίας και Μηχανισμοί του Πειράματος

Για το πραγματικό πείραμα χρησιμοποιείται ο ρομποτικός βραχίονας πλεοναζόντων βαθμών ελευθερίας (redundant 7 DoF) Panda (εικόνα 6.1).

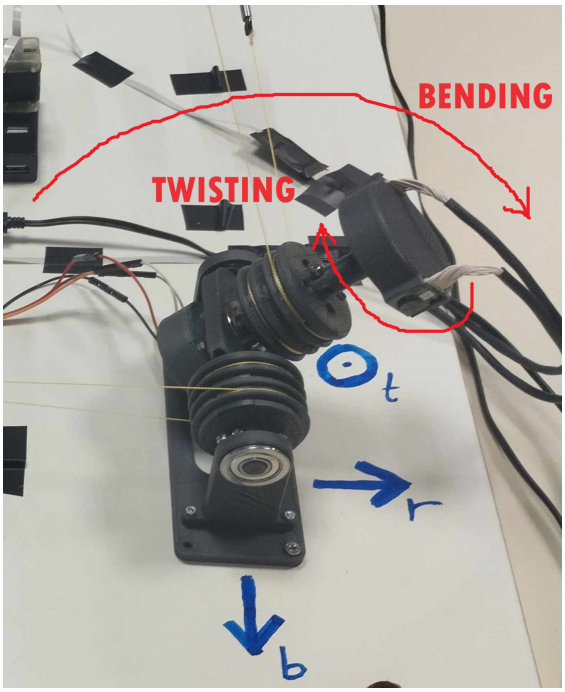
Εικόνα 6.1: *Ο ρομποτικός βραχίονας Panda του πειράματος*

Το εργαλείο του Panda είναι ένας παράλληλος gripper δύο δακτύλων, τα οποία ανοίγουν και κλείνουν εκτελώντας αντίθετες κινήσεις ταυτόχρονα. Ο gripper του Panda που χρησιμοποιήσαμε παρουσιάζεται στην εικόνα 6.2. Εσωτερικά του κάθε δακτύλου παρατηρούμε δύο βάσεις οι οποίες τυπώθηκαν για να εφαρμόζουν στους αισθητήρες της διάταξης.

Εικόνα 6.2: *Ο gripper του Panda του πειράματος*

Περισσότερες πληροφορίες για το ρομπότ παρατίθενται στο παράρτημα Α΄.

Ο βασικός μηχανισμός του χώρου εργασίας είναι το αντικείμενο που προσομοιάζει μανιτάρι, δηλαδή αυτό στο οποίο γίνεται ο επιδέξιος χειρισμός. Έχει δύο βαθμούς ελευθερίας, έναν στον άξονα της στρέψης και έναν στης κάμψης, επιτρέποντας συνδυασμό των δύο αυτών κινήσεων, που αποτελούν το χώρο δράσεών μας. Πάνω στην επιφάνειά του είναι τοποθετημένοι δύο αισθητήρες, στα σημεία που τον ακουμπούν τα δάκτυλα του gripper, ώστε να μετρώνται οι εφαρμοζόμενες ροπές, οι οποίες αποτελούν το χώρο καταστάσεών μας. Βλέπουμε το μηχανισμό στην εικόνα 6.3.

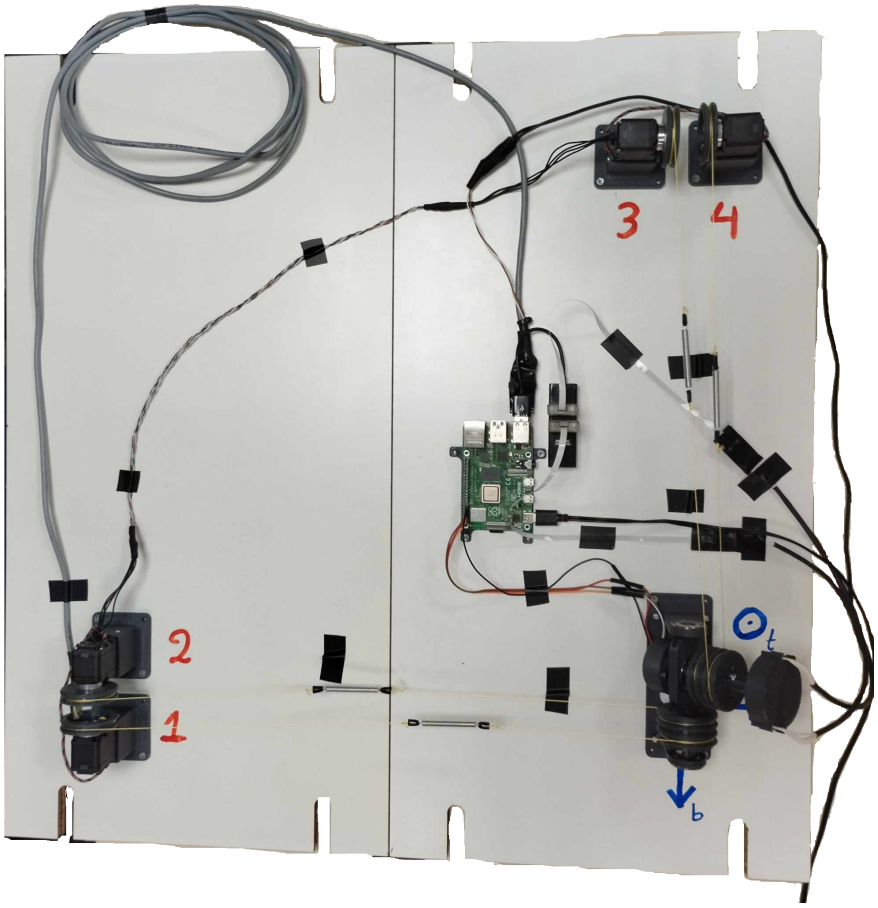


Εικόνα 6.3: *Το μηχανικό μανιτάρι του πειράματος*

Όπως παρατηρούμε στην παραπάνω εικόνα, στο χώρο εργασίας υπάρχουν ζωγραφισμένοι 3 άξονες. Στον αντίθετο του b εκτελείται η περιστροφική κίνηση της κάμψης και γύρω από έναν αντιπαράλληλο στον t εκτελείται η στρέψη. Ο άξονας b δεν αντιστοιχεί σε κάποιο βαθμό ελευθερίας, αλλά έχει σχεδιαστεί για πληρότητα, εκφράζοντας έναν άξονα κάμψης κάθετο

στον b. Το αντικείμενο είναι προσανατολισμένο στην αρχική θέση που έχει κατά τη διεξαγωγή του πειράματος, η οποία χαρακτηρίζεται από γωνία κάμψης $\frac{\pi}{4}$ από την κατακόρυφο. Επιπλέον, ξεκινώντας από την αρχική αυτή θέση έχει εύρος κάμψης $\frac{\pi}{4}$ προς τα αρνητικά, και εύρος στρέψης $\frac{\pi}{2}$ προς τα θετικά. Όταν ο προσανατολισμός του αντικειμένου γύρω από οποιονδήποτε από τους δύο άξονες κίνησης αντιστοιχεί με αυτόν της αρχικής θέσης, μια κάμψη ή στρέψη που δεν περιλαμβάνεται στο αντίστοιχο εύρος κινήσεων εμποδίζεται από τα μηχανικά όρια του αντικειμένου.

Στην εικόνα επίσης μπορεί κανείς να διακρίνει νήματα τα οποία τυλίγονται γύρω από τους δύο άξονες περιστροφής του μηχανισμού. Τα νήματα αυτά συνδέονται μέσω όμοιων ελατηρίων με κινητήρες, δημιουργώντας μια διάταξη που προσομοιάζει τη δυναμική του μηχανικού μανιταριού, όπως ακριβώς ο διάδρομος της προσομοίωσης. Το σύνολο της διάταξης του χώρου εργασίας φαίνεται στην εικόνα 6.4.



Εικόνα 6.4: *Η διάταξη του χώρου εργασίας του πειράματος*

Στην παραπάνω εικόνα, κάτω δεξιά διακρίνεται το αντικείμενο. Τα ζευγάρια αριθμών 1,2 και 3,4 αντιστοιχούν σε κινητήρες που λειτουργούν ανταγωνιστικά. Συγκεκριμένα, οι κινητήρες 1 και 2 θέτουν τα όρια της επιτρεπόμενης κάμψης. Έτσι, για συγκεκριμένη στρέψη ο κινητήρας 1 θέτει το άνω όριο κάμψης σε σχέση με την αρχική θέση, ενώ ο 2 το κάτω όριο. Κατά την ομαλή κάμψη μέσα στις επιτρεπόμενες τιμές, οι δύο κινητήρες ακολουθούν την κίνηση του αντικειμένου. Όταν ένα από τα δύο όρια παραβιαστεί, ο αντίστοιχος κινητήρας σταματάει να ακολουθεί την κίνηση, με αποτέλεσμα την επιμήκυνση του αντίστοιχου

ελατηρίου το οποίο προκαλεί ροπή στο αντικείμενο (η οποία επηρεάζει και αυτή στα σημεία επαφής με το ρομπότ). Την ίδια στιγμή, ο άλλος κινητήρας συνεχίζει να ακολουθεί την κίνηση για να μην προκληθούν αντίθετες ροπές. Ανάλογα συμπεριφέρονται και οι άλλοι δύο κινητήρες οι οποίοι θέτουν τα όρια της επιτρεπόμενης στρέψης, με τον 4 να θέτει το άνω όριο και τον 3 το κάτω. Οι δυνάμεις επαναφοράς μοντελοποιούνται από την επιμήκυνση των ελατηρίων (για τα οποία όπως γνωρίζουμε ισχύει ο νόμος του Hooke $F = -k \cdot x$).

Η διάταξη έχει και τη δυνατότητα να αυξάνει το stiffness του αντικειμένου, με μια ρύθμιση κατά την οποία ο κινητήρας του οποίου το όριο έχει παραβιαστεί δε σταματά μόνο να ακολουθεί την κίνηση, αλλά αρχίζει και εκτελεί την αντίθετη κίνηση, με αποτέλεσμα τη γρηγορότερη επιμήκυνση του ελατηρίου.

6.2 Προεργασία

Πριν τη διεξαγωγή του πειράματος χρειάστηκε calibration για το σχεδιασμό και την αποθήκευση της τροχιάς του εργαλείου προς την θέση και τον αρχικό προσανατολισμό του αντικειμένου, αλλά και για τον ορισμό των δύο περιστροφικών κινήσεων του προβλήματός μας (στρέψη & κάμψη).

Η στρέψη είναι μια κίνηση η οποία ορίστηκε εύκολα, καθώς αποτελεί τη roll περιστροφική κίνηση του εργαλείου. Η κάμψη ωστόσο δεν μπορεί να οριστεί τόσο εύκολα, καθώς αποτελεί μια κυκλική κίνηση γύρω από ένα εξωτερικό σημείο (που βρίσκεται στη βάση του αντικειμένου). Για τη χάραξη αυτής της τροχιάς κίνησης, έγινε δειγματοληψία σημείων πιάνοντας το αντικείμενο με το gripper και στη συνέχεια εκτελώντας την κίνηση με χειρονακτική βοήθεια, σε ελεύθερη λειτουργία των κινητήρων του ρομποτικού βραχίονα. Στη συνέχεια χρειάστηκε παρεμβολή σημείων, καθώς και προγραμματισμός του ρομπότ να εκτελεί τις κινήσεις του μεταξύ σημείων κυκλικά και όχι σε ευθεία γραμμή.

Ως φυσική τάση κίνησης θέτουμε τη θετική στρέψη (όπως τη θετική κίνηση στον οριζόντιο άξονα στο περιβάλλον προσομοίωσης). Επιπλέον, αποφασίζουμε, χάριν απλοποίησης, η ανάδραση δύναμης να μη γίνεται με τις ακριβείς δυνάμεις των αισθητήρων, αλλά με μετασχηματισμό τους στους άξονες στρέψης και κάμψης, ανάλογα με την προσομοίωση όπου μετρώνται οι κινήσεις στους δύο κάθετους άξονες.

6.3 Παραδοχές, Συμβιβασμοί και Περιορισμοί

Καθώς το πραγματικό task που έχουμε διαφέρει σε πολλά σημεία από την προσομοίωση, είναι αναγκαίο να κάνουμε τις κατάλληλες παραδοχές και συμβιβασμούς, αλλά και να κατανοήσουμε τους περιορισμούς που έχουμε.

Αρχικά, θεωρούμε ότι όπως στην προσομοίωση μπορούμε να μετασχηματίσουμε τις δύο δυνάμεις και δράσεις διαφορετικών αξόνων σε πολικές συντεταγμένες (μέτρο & γωνία), έτσι μπορούμε να συμπεριφερθούμε και με τις ροπές και περιστροφές στο πραγματικό πρόβλημα. Αυτό το λογικό άλμα δικαιολογείται από την αναγωγή του προβλήματος σε ένα νοητό διάδρομο με 2 βαθμούς ελευθερίας.

Ένα άλλο πρόβλημα που αντιμετωπίζουμε είναι ότι οι δράσεις που εξάγει ο αλγόριθμός μας για διεξαγωγή από το ρομπότ, δε μπορούν να πραγματοποιηθούν με ακρίβεια. Έτσι,

είναι πιθανό να έχουμε αποκλίσεις στις γωνίες στρέψης και κάμψης σε κάθε βήμα. Για την ομαλή λειτουργία του αλγορίθμου, μετά την πραγματοποίηση της δράσης σε κάθε βήμα λαμβάνουμε τις πραγματικές γωνίες στρέψης και κάμψης της κίνησης του ρομπότ, και κρατώντας ίδια τη νοητή τους γωνία σε πολικές συντεταγμένες, αλλάζουμε το μέτρο σε 5° (το επιθυμητό) για την χρήση τους σε ότι αφορά τον αλγόριθμο μάθησης.

Κάτι ακόμα που λαμβάνουμε υπ' όψιν είναι τα μηχανικά όρια του αντικειμένου (βλ. ενοτ. 6.1). Αν η δράση που εξάγει ο αλγόριθμος περιλαμβάνει αρνητική κάμψη ή στρέψη, όταν ο προσανατολισμός του αντικειμένου στον αντίστοιχο άξονα συμπίπτει με αυτόν της αρχικής θέσης, το ρομπότ δεν πρέπει να εκτελεί τη συγκεκριμένη κίνηση καθώς υπάρχει ο κίνδυνος καταστροφής του μηχανισμού. Σε τέτοια περίπτωση, λοιπόν, το ρομπότ λαμβάνει εντολή να μην εκτελέσει περιστροφή γύρω από τον αντίστοιχο άξονα. Προς αποφυγή αρνητικής επίδρασης στον αλγόριθμο μάθησης από αυτή την προσθήκη, σε τέτοια περίπτωση ορίζουμε την γωνία της επόμενης κίνησης μέσω γκαουσιανής εξερεύνησης (βλ. υποενοτ. 4.5.2 με μέση τιμή τη γωνία που προκύπτει από συνιστώσες την αντίθετη κίνηση αυτής που δεν διεξήχθη στο προηγούμενο βήμα και την έξοδο του αλγορίθμου στον άλλο βαθμό ελευθερίας (που δεν έχει το πρόβλημα). Αν δε μπορεί να γίνει καμία από τις δύο θεμελιώδεις κινήσεις, τότε ως μέση τιμή παίρνουμε τη γωνία που προκύπτει από συνιστώσες τις δύο αντίθετες κινήσεις. Με αυτό τον τρόπο υλοποιούμε έναν άμεσο τρόπο δραπέτευσης από τα μηχανικά όρια του αντικειμένου. Θέτουμε για αυτή την εξερεύνηση $\epsilon = \frac{\pi}{6} rad$.

Τέλος, κατά την εκτέλεση του πειράματος παρατηρούμε ότι οι αισθητήρες έχουν πολύ μεγάλο ποσοστό θορύβου στις μετρήσεις τους. Αυτό πιθανόν συμβαίνει λόγω της απαίτησης για πολύ αυστηρό calibration, κάτι το οποίο δεν είναι άμεσα εφικτό. Για το λόγο αυτό δε χρησιμοποιούνται, και ως ανάδραση δύναμης, αντί για τις μετασχηματισμένες (στους δύο άξονες) τιμές των μετρήσεών τους, δίνουμε τις αποκλίσεις στρέψης και κάμψης από τα επιθυμητά όρια (που είναι ανάλογες, και αντίθετης κατεύθυνσης, με τις ροπές που θεωρητικά ασκούνται στο αντικείμενο βάσει μοντελοποίησης).

6.4 Tuning υπερπαραμέτρων

Όσον αφορά τις υπερπαραμέτρους της μεθόδου μάθησης, αυτές χρειάζονται αλλαγές έτσι ώστε να ταιριάξουν στα νέα δεδομένα και τάξεις μεγέθους του πραγματικού προβλήματος. Υπενθυμίζουμε, όπως αναφέρθηκε στην ενότητα 6.3, ότι εφαρμόζουμε ανάδραση διείσδυσης (από το νοητό διάδρομο που δημιουργείται), δηλαδή της συνισταμένης των αποκλίσεων κάμψης και στρέψης από τους επιθυμητούς συνδυασμούς, μέσω των τύπων των πολικών συντεταγμένων για το μέτρο και τη γωνία (βλ. υποενοτ. 4.1.2). Από εδώ και στο εξής, θα αναφερόμαστε σε αυτή την συνισταμένη ως διείσδυση (στους τοίχους του νοητού διαδρόμου).

Αρχικά, ορίζουμε τις 8° ως όριο ανώτατης διείσδυσης πάνω από το οποίο θα σταματάμε το επεισόδιο εκπαίδευσης για προστασία της διάταξης.

Αναλογικά, θέτουμε 33 κέντρα RBF των καταστάσεων: αυτό της μηδενικής διείσδυσης, και άλλα επιπλέον 32 που προκύπτουν από 4 ομόκεντρους κύκλους με διαδοχικές διαφορές ακτίνων 0.02264° διείσδυσης, αρχική ακτίνα 0.0283 και τελική ακτίνα 0.09622° διείσδυσης. Τα κέντρα αυτά τοποθετούνται στις ίδιες γωνίες με αυτά της προσομοίωσης, δηλαδή $0, \pm\frac{\pi}{4}, \pm\frac{\pi}{2}, \pm\frac{3\pi}{4}, \pi$ (σύνολο $8 \cdot 4 = 32$). Ακόμα, ως όριο πάνω από το οποίο κρατάμε τις τιμές

των RBFs ίδιες για σταθερή γωνία θέτουμε τις 6° (ανάλογα του 0.185 στην προσομοίωση, βλ. υποενοτ. 4.5.1).

Τα κέντρα των RBF των δράσεων τοποθετούνται στις ίδιες γωνίες, για σταθερό μέτρο περιστροφής (στρέψης & κάμψης) 5°.

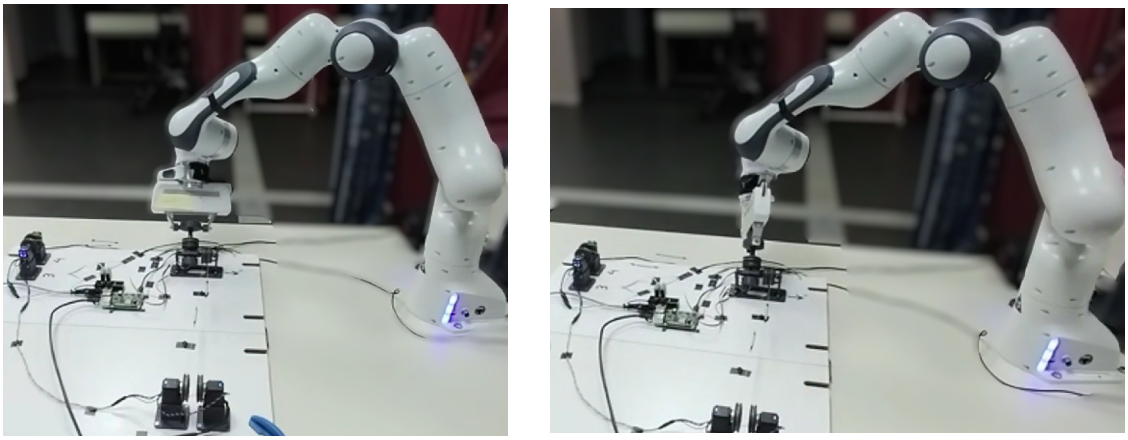
Σημειώνουμε ότι οι διαφορές στις μέχρι τώρα παραμέτρους αφορούν κυρίως ακτινικές αποστάσεις, οι οποίες, όπως είναι φυσιολογικό, είναι διαφορετικές στο πραγματικό πρόβλημα. Αντίθετα, ο γωνιακός χώρος παραμένει ίδιος, στο διάστημα $(-\pi, \pi]$. Με τον ίδιο τρόπο θα συνεχίσουμε το tuning στις υπόλοιπες υπερπαραμέτρους. Οι μεταβολές παρουσιάζονται στον πίνακα 6.1.

Πίνακας 6.1: *Διαφοροποιήσεις στις υπερπαραμέτρους μεταξύ προσομοίωσης και πραγματικού πειράματος*

Παράμετρος	**Προσομοίωση**	**Πραγματικό πείραμα**
k_{small} (εξ. (4.2))	80	147
k_{big} (εξ. (4.2))	$1.4 \cdot 80$	$1.55 \cdot 147$
Κατώφλι μετρητή μείωσης ε,α	25 για την κατάσταση μηδενικής δύναμης, 4 για τις καταστάσεις με γωνίες 0 και $\pm\frac{\pi}{4} rad$, και 12 για τις υπόλοιπες	17 για την κατάσταση μηδενικής δύναμης, 3 για τις καταστάσεις με γωνίες 0 και $\pm\frac{\pi}{4} rad$, και 7 για τις υπόλοιπες
σ_{00_s}	0.02	0.015
σ_{s_0}	0.006	0.005
σ_{00_a}	0.025	0.018

6.5 Παράθεση Αποτελεσμάτων

Πραγματοποιούμε μια εκπαίδευση συνολικά 9 επεισοδίων, ανάμεσα στα οποία αλλάζουμε τη δυναμική του αντικειμένου (δηλαδή το νοητό διάδρομο), προκειμένου να εξετάσουμε τη συμπεριφορά του πράκτορα. Τα επεισόδια ολοκληρώνονται σε σημεία που θέτουμε εμείς.



(α΄) (β΄)

Εικόνα 6.5: *Στιγμιότυπα εκπαίδευσης (πραγματικό πείραμα)*

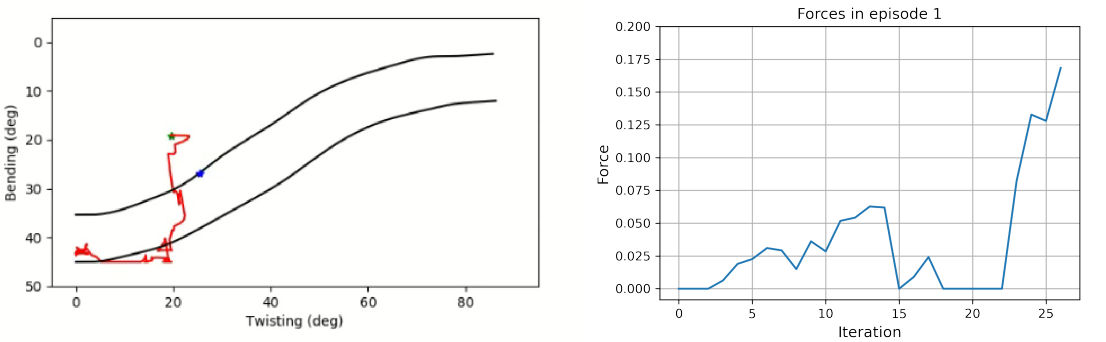
Απεικονίζουμε την τροχιά που εφαρμόζει το ρομπότ πάνω στο αντικείμενο με ένα διδι-

άστατο διάδρομο, ακριβώς όπως στην προσομοίωση, μόνο που αντί να έχουμε ένα κυκλικό αντικείμενο, έχουμε ένα σημείο το οποίο δείχνει τις εκάστοτε γωνίες που έχει το αντικείμενο-μανιτάρι στον άξονα στρέψης και κάμψης. Οι τοίχοι του διαδρόμου ορίζουν τα επιθυμητά όρια συνδυασμών γωνιών στρέψης και κάμψης. Ο οριζόντιος άξονας αντιστοιχεί στη στρέψη, ενώ ο κατακόρυφος στην κάμψη. Το τέλος του διαδρόμου σηματοδοτεί το τέλος του επεισοδίου.

Σημειώνουμε εδώ ότι κατά σύμβαση σχεδιάζουμε το σύστημα αξόνων έτσι ώστε ψηλότερα στον κατακόρυφο άξονα να αντιστοιχούν μικρότερες γωνίες κάμψης. Επίσης, τα γραφήματα δύναμης που θα παρουσιάσουμε παρακάτω ουσιαστικά αναπαριστούν την διείσδυση, όμως βασιζόμενοι στο ότι είναι ποσά ανάλογα, παρουσιάζονται ως δύναμης στα γραφήματα για καλύτερη κατανόηση του προβλήματος.

Παρουσιάζουμε παρακάτω την εκπαίδευση:

- Επεισόδιο 1:

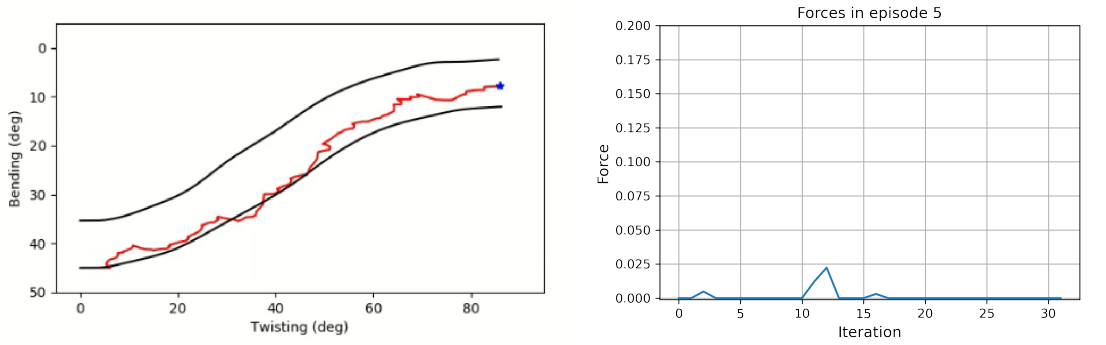


(α') *Τροχιά* (β') *Μετρούμενη από τον πράκτορα δύναμη*

Σχήμα 6.1: *1ο επεισόδιο εκπαίδευσης (πραγματικό πείραμα)*

Στο πρώτο επεισόδιο (σχ. 6.1) ο πράκτορας δεν παρουσιάζει επιτυχία, καθώς έχει μόλις ξεκινήσει να εκπαιδεύεται, και επίσης εξερευνά το χώρο. Το επεισόδιο ολοκληρώνεται περίπου στα 22 βήματα, λόγω μεγάλης υπέρβασης των ορίων.

- Επεισόδιο 5:

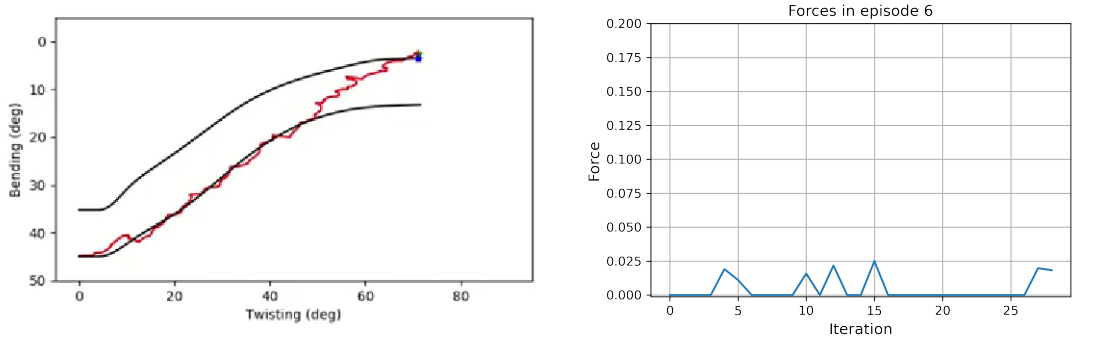


(α') *Τροχιά* (β') *Μετρούμενη από τον πράκτορα δύναμη*

Σχήμα 6.2: *5ο επεισόδιο εκπαίδευσης (πραγματικό πείραμα)*

Παρατηρούμε ότι μέχρι το 5ο επεισόδιο (σχ. 6.2) ο πράκτορας έχει εκπαιδευτεί πολύ ικανοποιητικά ώστε να ακολουθεί ομαλά τα όρια της δυναμικής του αντικειμένου. Η εφαρμοζόμενη δύναμη (δηλαδή η διείσδυση) έχει μειωθεί σημαντικά, και το επεισόδιο ολοκληρώνεται επιτυχώς περίπου στα 32 βήματα.

- Επεισόδιο 6:

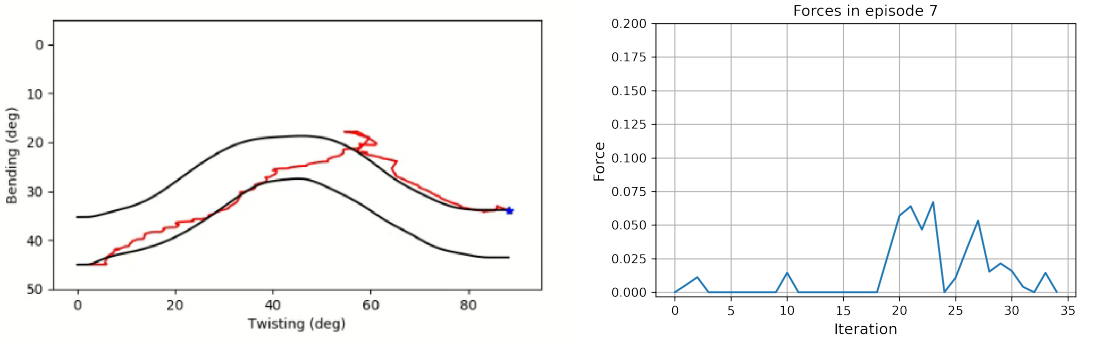


(α') *Τροχιά* (β') *Μετρούμενη από τον πράκτορα δύναμη*

Σχήμα 6.3: *6ο επεισόδιο εκπαίδευσης (πραγματικό πείραμα)*

Στο 6ο επεισόδιο (σχ. 6.3) αλλάζουμε τη δυναμική του αντικειμένου, όπως φαίνεται και από το νέο διάδρομο, συνεχίζοντας όμως την εκπαίδευση από το σημείο που την αφήσαμε. Ο πράκτορας φαίνεται να ανταποκρίνεται καλά στο διαφορετικό περιβάλλον, και κάνοντας ομαλή ακολούθηση των ορίων της νέας δυναμικής ολοκληρώνει επιτυχώς το task στα 29 βήματα.

- Επεισόδιο 7:



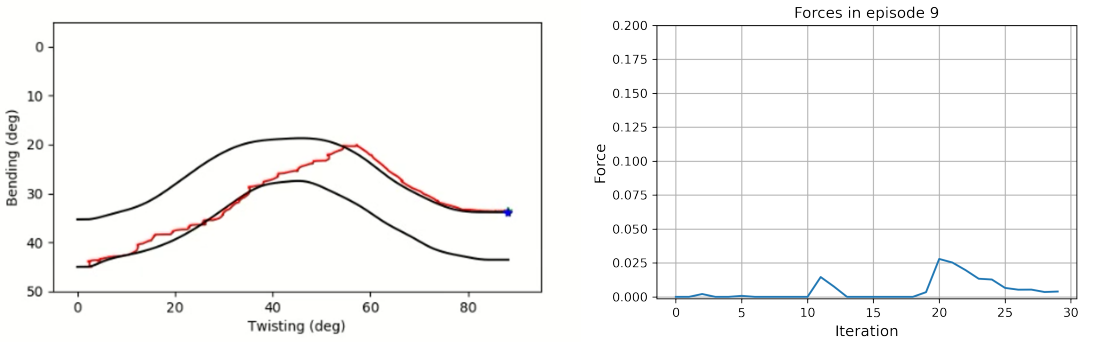
(α') *Τροχιά* (β') *Μετρούμενη από τον πράκτορα δύναμη*

Σχήμα 6.4: *7ο επεισόδιο εκπαίδευσης (πραγματικό πείραμα)*

Μέχρι στιγμής έχουμε εκπαιδεύσει το μοντέλο μόνο σε αύξηση της στρέψης αποκλειστικά με ταυτόχρονη μείωση της κάμψης. Αλλάζουμε λοιπόν το περιβάλλον ξανά, εισάγοντας μια δυναμική όπου κάποια στιγμή χρειάζεται ταυτόχρονη αύξηση της στρέψης και της κάμψης. Στο 7ο επεισόδιο λοιπόν (σχ. 6.4), ο πράκτορας ξεκινά διαγράφοντας ικανοποιητική τροχιά, στη συνέχεια όμως (περίπου στο 19ο βήμα) όπου χρειάζεται

αύξηση στρέψης και κάμψης ταυτόχρονα, κάτι στο οποίο δεν έχει εκπαιδευτεί μέχρι στιγμής, φαίνεται να καταλήγει εκτός ορίων (πιθανόν να εξερευνά το περιβάλλον), όμως εντέλει στη συνέχεια καταφέρνει να τερματίσει περίπου στα 35 βήματα. Στο διάγραμμα της δύναμης είναι εμφανές το μέρος όπου ο πράκτορας βρίσκεται σε σημεία για τα οποία δεν έχει εμπειρία, από την αυξημένη τιμή της.

- Επεισόδιο 9:



(α') *Τροχιά*

(β') *Μετρούμενη από τον πράκτορα δύναμη*

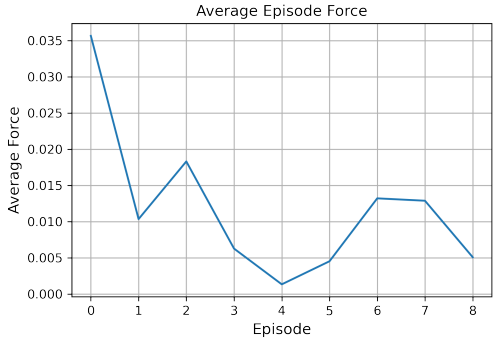
Σχήμα 6.5: *9ο επεισόδιο εκπαίδευσης (πραγματικό πείραμα)*

Στο τελευταίο επεισόδιο της εκπαίδευσης (σχ. 6.5) παρατηρούμε ότι ο πράκτορας έχει εκπαιδευτεί και για το μέρος της δυναμικής του αντικειμένου όπου πριν είχε μειωμένη απόδοση λόγω της έλλειψης εμπειρίας, ακολουθώντας την αυξημένης δυσκολίας δυναμική ομαλά, κρατώντας τη δύναμη σε χαμηλά επίπεδα, και τερματίζοντας εύκολα περίπου στα 30 βήματα.

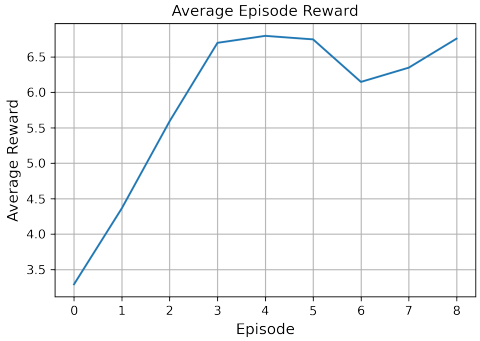
Αξιοσημείωτο είναι το γεγονός ότι παρ' όλο που ο πράκτορας, εν μέσω εκπαίδευσης, βρέθηκε σε περιβάλλον με άγνωστα ερεθίσματα ενώ η συμπεριφορά του είχε συγκλίνει σε άλλα, κατάφερε να τα ενσωματώσει στη συμπεριφορά του χωρίς να επηρεάσει την προηγούμενη μάθηση. Σε αυτό φαίνεται να παίζει σημαντικό ρόλο και η προσαρμοστική εξερεύνηση και ρυθμός μάθησης, καθώς επιτρέπει τη σύγκλιση σε σημεία όπου δεν έχει επιτευχθεί χωρίς να επηρεάζει ιδιαίτερα τα άλλα, προσφέροντας μια ανεξαρτησία των περιοχών μάθησης (χωρίς φυσικά αυτό να αντιτίθεται με τη λογική των συνεχών χώρων καταστάσεων και δράσεων).

Το παραπάνω στοιχείο είναι ιδιαίτερα σημαντικό, αφού στα ρομπότ η συνεχής μάθηση μπορεί να φανεί αποτελεσματική, εφόσον είναι δύσκολο να εφαρμόσουμε εξ αρχής μια εκπαίδευση που θα περιλαμβάνει όλες τις δυνατές υποπεριπτώσεις (πιθανόν να αποτελέσει μια ιδιαίτερα χρονοβόρα διαδικασία).

Τέλος, παραθέτουμε τα διαγράμματα της μέσης δύναμης, ανταμοιβής, και TD σφάλματος ανά επεισόδιο εκπαίδευσης (σχ. 6.6). Παρατηρούμε το αναμενόμενο, δηλαδή σταδιακή μείωση της δύναμης (διείσδυσης) και του TD σφάλματος και αύξηση της ανταμοιβής έως το 7ο επεισόδιο (σημειώνουμε ότι στα διαγράμματα τα επεισόδια έχουν το εύρος 0-8 αντί για 1-9), όπου και παρουσιάζονται διαταραχές λόγω του άγνωστου περιβάλλοντος, οι οποίες εντέλει στη δύναμη και στην ανταμοιβή εξομαλύνονται.



(α') *Μέση δύναμη*



(β') *Μέση λαμβανόμενη ανταμοιβή*

(γ') *Μέσο TD σφάλμα*

Σχήμα 6.6: *Μέσοι όροι μεγεθών μάθησης ανά επεισόδιο (στο πραγματικό πείραμα)*

Μέρος III

Επίλογος

Κεφάλαιο 7

Επίλογος

7.1 Συμπεράσματα

Στην εργασία αυτή ασχοληθήκαμε με την υλοποίηση μιας μεθόδου Ενισχυτικής Μάθησης αγνώστου μοντέλου, η οποία, μεταξύ άλλων, δεν απαιτεί μακροσκελή εκπαίδευση και εμφανίζει γρήγορα αποτελέσματα, έτσι ώστε να μπορεί να χρησιμοποιηθεί σε πραγματική διάταξη. Φάνηκε ότι ο αλγόριθμος Episodic Linear Semi-gradient SARSA είναι κατάλληλος για τις απαιτήσεις μας,

Ξεκινήσαμε ανάγοντας το πρόβλημα του επιδέξιου χειρισμού με ανάδραση δύναμης, σε ένα task απόδρασης από διδιάστατο διάδρομο (περιβάλλον προσομοίωσης). Στην προσομοίωση έγιναν όλες οι απαραίτητες δοκιμές προκειμένου η μέθοδός μας να λειτουργήσει όσο καλύτερα γίνεται.

Μια πολύ σημαντική προσθήκη μας στην υλοποίηση είναι αυτή της προσαρμοστικής εξερεύνησης και ρυθμού μάθησης, η οποία ταυτόχρονα μείωσε τα απαιτούμενα βήματα εκπαίδευσης και αύξησε την απόδοση του αλγορίθμου. Επιπλέον, βοηθά στις περιπτώσεις όπου θέλουμε να έχουμε μια διαρκή μάθηση, όπως σε καταστάσεις όπου συλλέγουμε δεδομένα συνεχώς σε πραγματικό χρόνο κατά την εργασία που πραγματοποιεί το ρομπότ, χωρίς να είναι δυνατή μια εκ των προτέρων ολοκληρωμένη εκπαίδευση.

Τέλος, με τα πειράματα στην πραγματική διάταξη δοκιμάστηκε η πραγματική συνεισφορά της μεθόδου μας στην πραγματικότητα, όπου πέτυχε πολύ ικανοποιητικά αποτελέσματα.

7.2 Μελλοντικές Επεκτάσεις

Η εργασία μας είναι μόνο ένα μέρος ενός μεγάλου πεδίου έρευνας, και υπάρχουν πολλές και διαφορετικές μελλοντικές επεκτάσεις για αυτή.

Μια από τις πιο σημαντικές επεκτάσεις οι οποία θα μπορούσε να εφαρμοστεί πιο άμεσα είναι η γενίκευση του task προς μια πιο ρεαλιστική tactile κατεύθυνση, όπου οι δυνάμεις που μετρώνται από τους αισθητήρες δε θα αναλύονται στους άξονες της στρέψης και της κάμψης, αλλά θα αποτελούν την κατάσταση του πράκτορα κάθε στιγμή, χωρίς κάποιον σημαντικό μετασχηματισμό. Αυτό το πρόβλημα φυσικά προσθέτει περισσότερες διαστάσεις στο πρόβλημα, και φέρει νέα ζητήματα προς επίλυση, όπως η επιλογή νέων παραμέτρων και η μείωση της απαιτούμενης υπολογιστικής δύναμης.

Επιπλέον, ωφέλιμη στο πρόβλημα που έχουμε να επιλύσουμε, σε μελλοντική έρευνα,

μπορεί να είναι η μάθηση από ανθρώπινη επίδειξη, η οποία μπορεί να επιταχύνει ακόμα περισσότερο την εκπαίδευση και τη σύγκλιση, μέσω δεδομένων χειροκίνητης επίτευξης του task που θα αποτελέσουν τη βάση της μετέπειτα εκπαίδευσης του ρομπότ. Η προσέγγιση αυτή μπορεί να εφαρμοστεί και εφαρμογή της εργασίας μας σε multi-fingered ρομποτικές διατάξεις, στις οποίες φαίνεται να βοηθά τόσο στη μείωση των δειγμάτων μάθησης και στο robustness, όσο και στη φυσικότητα των κινήσεων [32].

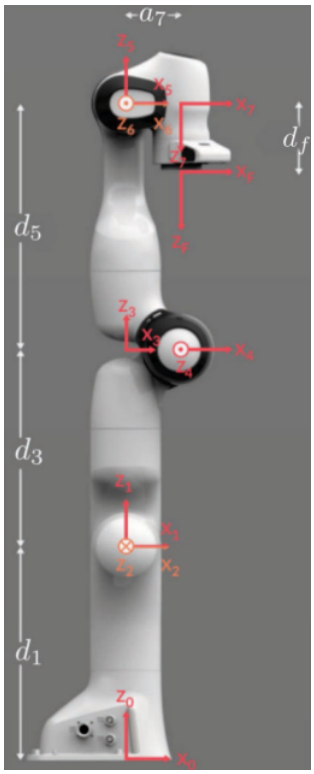
Μια ακόμα προτεινόμενη επέκταση είναι η γενίκευση του προβλήματος της προσομοίωσης για διαδρόμους όπου η γενική κατεύθυνση τους δε βρίσκεται κατ' ανάγκη προς τα δεξιά. Μια υλοποίηση για αυτό το πρόβλημα πιθανόν να απαιτεί καταγραφή των τοίχων τους οποίους συνάντησε πρόσφατα ο πράκτορας, ώστε να υπερνικηθεί η πιθανότητα παγίδευσής του. Με αυτή τη γενίκευση είναι λογικό ότι και η πραγματική ρομποτική κίνηση σε 2 βαθμούς ελευθερίας θα έχει πολύ μεγαλύτερες ελευθερίες.

Παραρτήματα

Παράρτημα Α΄

Το ρομπότ Panda

Το Panda της Franka Emika είναι ένας ρομποτικός βραχίονας πλεοναζόντων, και συγκεκριμένα 7 βαθμών ελευθερίας, μέσω 7 στροφικών αρθρώσεων, που επιτρέπει τη διεξαγωγή επιδέξιων κινήσεων.



Εικόνα Α΄.1: *Σχεδιασμός αξόνων στο Panda (από [4])*

Στην εικόνα Α΄.1 απεικονίζεται το Panda, μαζί με τους άξονες στις αρθρώσεις του. Είναι $d_1 = 0.333m$, $d_3 = 0.316m$, $d_5 = 0.384m$, $d_f = 0.107m$, $a_4 = 0.0825m$, $a_5 = -0.0825m$, $a_7 = 0.088m$ [4].

Με βάση την εικόνα αυτή, παραθέτουμε στον πίνακα Α΄.1 τις παραμέτρους Denavit-Hartenberg του Panda σύμφωνα με τη σύμβαση του Craig [33] (από [4]).

Πίνακας Α΄.1: *Παράμετροι Denavit-Hartenberg του Panda (από [4])*

Άρθρωση	a_i (m)	α_i (rad)	d_i (m)	ϑ_i
1	0	0	0.333	q_1
2	0	$-\frac{\pi}{2}$	0	q_2
3	0	$\frac{\pi}{2}$	0.316	q_3
4	0.0825	$\frac{\pi}{2}$	0	q_4
5	-0.0825	$-\frac{\pi}{2}$	0.384	q_5
6	0	$\frac{\pi}{2}$	0	q_6
7	0.088	$\frac{\pi}{2}$	0	q_7
Εργαλείο	0	0	0.107	0

Βιβλιογραφία

[1] Richard S. Sutton και Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, Massachusetts, 2η έκδοση, 2018.

[2] *UCL Course on RL - David Silver*. https://www.davidsilver.uk/teaching/. Ημερομηνία πρόσβασης: 14-02-2024.

[3] Mingsen Huang, Long He, Daeun Choi, John Pecchia και Yaoming Li. *Picking dynamic analysis for robotic harvesting of Agaricus bisporus mushrooms*. *Computers and Electronics in Agriculture*, 185:106145, 2021.

[4] Claudio Gaz, Marco Cognetti, Alexander Oliva, Paolo Robuffo Giordano και Alessandro De Luca. *Dynamic Identification of the Franka Emika Panda Robot With Retrieval of Feasible Parameters Using Penalty-Based Optimization*. *IEEE Robotics and Automation Letters*, 4(4):4147–4154, 2019.

[5] A.M. Okamura, N. Smaby και M.R. Cutkosky. *An overview of dexterous manipulation*. *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065)*, τόμος 1, σελίδες 255–262 ολ.1, 2000.

[6] Zhen Xie, Xinquan Liang και Canale Roberto. *Learning-based robotic grasping: A review*. *Frontiers in Robotics and AI*, 10, 2023.

[7] Daichi Saito, Kazuhiro Sasabuchi, Naoki Wake, Jun Takamatsu, Hideki Koike και Katsushi Ikeuchi. *Task-grasping from human demonstration*, 2022.

[8] Juan Valverde. *Harvesting and Processing of Mushrooms*, κεφάλαιο 13, σελίδες 261–270. John Wiley Sons, Ltd, 2017.

[9] Areg Azoyan. *Feasibility analysis of an automated mushroom harvesting system*. Μεταπτυχιακή διπλωματική εργασία, University of Georgia, 2004.

[10] Van Griensven LJLD. *The Cultivation of Mushrooms*. Mushroom Experimental Station: Horst, The Netherlands, 1988.

[11] J.N. Reed και R.D. Tillett. *Initial experiments in robotic mushroom harvesting*. *Mechatronics*, 4(3):265–279, 1994.

[12] Θ. Καράμπελα. *Εκμάθηση ρομποτικής κίνησης με χρήση αλγορίθμου ενισχυτικής μάθησης και ανάδραση δύναμης*. Διπλωματική εργασία, Εθνικό Μετσόβιο Πολυτεχνείο, 2022.

[13] T.H. Cormen, C.E. Leiserson, R.L. Rivest και C. Stein. *Introduction to Algorithms*. MIT Press, Cambridge, Massachusetts, 4η έκδοση, 2022.

[14] Ivo Grondman, Lucian Busoniu, Gabriel A. D. Lopes και Robert Babuska. *A Survey of Actor-Critic Reinforcement Learning: Standard and Natural Policy Gradients*. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):1291–1307, 2012.

[15] Sascha Lange, Thomas Gabel και Martin Riedmiller. *Batch Reinforcement Learning*, σελίδες 45–73. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[16] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage και Anil Anthony Bharath. *Deep Reinforcement Learning: A Brief Survey*. *IEEE Signal Processing Magazine*, 34(6):26–38, 2017.

[17] Hao nan Wang, Ning Liu, Yi yun Zhang, Da wei Feng, Feng Huang, Dong sheng Li και Yi ming Zhang. *Deep reinforcement learning: a survey*. *Frontiers of Information Technology & Electronic Engineering*, 21(12):1726–1744, 2020.

[18] Clark Kendrick Go, Bryan Lao, Junichiro Yoshimoto και Kazushi Ikeda. *A reinforcement learning approach to the shepherding task using SARSA*. *2016 International Joint Conference on Neural Networks (IJCNN)*, σελίδες 3833–3836, 2016.

[19] Deepak Ramachandran και Rakesh Gupta. *Smoothed Sarsa: Reinforcement learning for robot delivery tasks*. *2009 IEEE International Conference on Robotics and Automation*, σελίδες 2125–2132, 2009.

[20] Anderson Mesquita, Yuri Nogueira, Creto Vidal, Joaquim Cavalcante-Neto και Paulo Serafim. *Autonomous Foraging with SARSA-based Deep Reinforcement Learning*. *2020 22nd Symposium on Virtual and Augmented Reality (SVR)*, σελίδες 425–433, 2020.

[21] Hitoshi Iima και Yasuaki Kuroe. *Swarm reinforcement learning algorithms based on Sarsa method*. *2008 SICE Annual Conference*, σελίδες 2045–2049, 2008.

[22] Harsh Gupta, R. Srikant και Lei Ying. *Finite-Time Performance Bounds and Adaptive Learning Rate Selection for Two Time-Scale Reinforcement Learning*. *Advances in Neural Information Processing Systems*H. Wallach, H. Larochelle, A. Beygelzimer, F.d' Alché-Buc, E. Fox και R. Garnett, επιμελητές, τόμος 32. Curran Associates, Inc., 2019.

[23] Othmane Mounjid και Charles Albert Lehalle. *Improving reinforcement learning algorithms: Towards optimal learning rate policies*. *Mathematical Finance*, 34(2):588–621, 2024.

[24] A. Aziz Khater, Ahmad M. El-Nagar, Mohammad El-Bardini και Nabila M. El-Rabaie. *Online learning based on adaptive learning rate for a class of recurrent fuzzy neural network*. *Neural Computing and Applications*, 32(12):8691–8710, 2020.

[25] Michael Bowling και Manuela Veloso. *Multiagent learning using a variable learning rate*. *Artificial Intelligence*, 136(2):215–250, 2002.

[26] Pawel Ladosz, Lilian Weng, Minwoo Kim και Hyondong Oh. *Exploration in deep reinforcement learning: A survey*. *Information Fusion*, 85:1–22, 2022.

[27] Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck και Pieter Abbeel. *#Exploration: A Study of Count-Based Exploration for Deep Reinforcement Learning*. *Advances in Neural Information Processing Systems*I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan και R. Garnett, επιμελητές, τόμος 30. Curran Associates, Inc., 2017.

[28] Jarryd Martin, Suraj Narayanan Sasikumar, Tom Everitt και Marcus Hutter. *Count-Based Exploration in Feature Space for Reinforcement Learning*, 2017.

[29] *SciPy v1.12.0 Manual - scipy.optimize.minimize*. https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.minimize.html. Ημερομηνία πρόσβασης: 30-03-2024.

[30] Wenshuai Zhao, Jorge Peña Queralta και Tomi Westerlund. *Sim-to-real transfer in deep reinforcement learning for robotics: a survey*. *2020 IEEE symposium series on computational intelligence (SSCI)*, σελίδες 737–744. IEEE, 2020.

[31] Wenshuai Zhao, Jorge Peña Queralta, Li Qingqing και Tomi Westerlund. *Towards Closing the Sim-to-Real Gap in Collaborative Multi-Robot Deep Reinforcement Learning*, 2020.

[32] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov και Sergey Levine. *Learning Complex Dexterous Manipulation with Deep Reinforcement Learning and Demonstrations*, 2018.

[33] J.J. Craig. *Introduction to Robotics: Mechanics and Control*. Pearson, 2018.

Συντομογραφίες - Αρκτικόλεξα - Ακρωνύμια

βλ.	βλέπε
κ.λπ.	και λοιπά
κ.ο.κ.	και ούτω καθεξής
RL	Reinforcement Learning
IMU	Inertial Measurement Unit
MDP	Markov Decision Process
MRP	Markov Reward Process
DP	Dynamic Programming
RBF	Radial Basis Function

Απόδοση ξενόγλωσσων όρων

Απόδοση	Ξενόγλωσσος όρος
ανθρωποειδή	humanoids
Τεχνητή Νοημοσύνη	Artificial Intelligence
Επιδέξιος Χειρισμός	Dexterous Manipulation
Όραση Υπολογιστών	Computer Vision
Ενισχυτική Μάθηση	Reinforcement Learning
Μηχανική Μάθηση	Machine Learning
ανθρώπινη επίδειξη	human demonstration
στρέψη	twisting
κάμψη	bending
ανύψωση	lifting
αγνώστου μοντέλου	model-free
δράση	action
κατάσταση	state
ανταμοιβή	reward
πράκτορας	agent
βήμα	iteration
πλειάδα	tuple
μοντέλο	model
Μαρκοβιανή Διαδικασία	Markov Process
Μαρκοβιανή Ιδιότητα	Markov Property
Μαρκοβιανή Αλυσίδα	Markov Chain
απόδοση	return
πολιτική	policy
συνάρτηση αξίας	value function
συνάρτηση δράσης-αξίας	action-value function
Δυναμικός Προγραμματισμός	Dynamic Programming
αξιολόγηση πολιτικής	policy evaluation
πρόβλεψη	prediction
βελτίωση πολιτικής	policy improvement
επανάληψη πολιτικής	policy iteration
επανάληψη αξίας	value iteration
έλεγχος	control
πάνω στην πολιτικής	on-policy
εκτός πολιτικής	off-policy

ρυθμός μάθησης	learning rate
εξερεύνηση	exploration
εκμετάλλευση	exploitation
προσέγγιση συνάρτησης	function approximation
γραμμικές μέθοδοι	linear methods
διάνυσμα χαρακτηριστικών	feature vector
γκαουσιανή συνάρτηση	gaussian function
Ενισχυτική Μάθηση βάσει πολιτικής	policy-based Reinforcement Learning
Ενισχυτική Μάθηση βάσει αξίας	value-based Reinforcement Learning
Μέθοδοι Δράστη-Κριτή	Actor-Critic Methods
μέθοδοι παρτίδας	batch methods
επεξηγησιμότητα	explainability
προσαρμοστικός	adaptive
εξερεύνηση βασισμένη στο μέτρημα	count-based exploration
μάθηση με διαταραχές	learning with disturbances