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

Εκμάθηση δεξιοτήτων ρομποτικού χειρισμού συνδυάζοντας δεδομένα επίδειξης και τεχνικές ενισχυτικής μάθησης

Διπλωματικη Εργασια

του

ΣΤΥΛΙΑΝΟΥ ΚΟΤΣΟΒΟΛΗ

Επιβλέπων: Κωνσταντίνος Τζαφέστας
Αναπληρωτής Καθηγητής

Αθήνα, Ιούνιος 2021

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

Εκμάθηση δεξιοτήτων ρομποτικού χειρισμού συνδυάζοντας δεδομένα επίδειξης και τεχνικές ενισχυτικής μάθησης

Διπλωματικη Εργασια

του

ΣΤΥΛΙΑΝΟΥ ΚΟΤΣΟΒΟΛΗ

Επιβλέπων: Κωνσταντίνος Τζαφέστας
Αναπληρωτής Καθηγητής

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 24 Ιουνίου 2021.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Κωνσταντίνος Τζαφέστας	Πέτρος Μαραγκός	Χαράλαμπος Ψυλλάκης
Αναπληρωτής Καθηγητής	Καθηγητής	Λέκτορας

Αθήνα, Ιούνιος 2021

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)

.....
Κοτσόβολης Στυλιανός

24 Ιουνίου 2021

Περίληψη

Αντικείμενο της παρούσας διπλωματικής εργασίας είναι η ανάπτυξη μεθόδου επιδέξιου εσωτερικού (in-hand) ρομποτικού χειρισμού αντικειμένων, χρησιμοποιώντας ενισχυτική μάθηση και δεδομένα επίδειξης. Η έρευνα ξεκινά από το πρόβλημα της στοχαστικής βελτιστοποίησης αποφάσεων σε περιβάλλοντα με άγνωστη δυναμική, με την βοήθεια της ενισχυτικής μηχανικής μάθησης. Στη συνέχεια, μελετάμε μεθόδους χρήσης δεδομένων επίδειξης, με σκοπό τη βελτίωση της επίδοσης της ενισχυτικής μάθησης και τη μίμηση της ανθρώπινης συμπεριφοράς. Επικεντρωνόμαστε στο πρόβλημα της χωρίς μοντέλο μάθησης (model-free learning) συμπεριφορών in-hand χειρισμού αντικειμένων στα οποία είναι επιθυμητή η παρακολούθηση τροχιών των μεγεθών του και ο έλεγχος της ασκούμενης δύναμης στο αντικείμενο. Για τον σκοπό αυτό προτείνεται ένας νευρο-ελεγκτής, ο οποίος εκπαιδεύεται αρχικά με επιβλεπόμενη μάθηση από τα δεδομένα επίδειξης και στη συνέχεια με ενισχυτική μάθηση ώστε να βελτιστοποιηθεί περαιτέρω ως προς την ζητούμενη συμπεριφορά. Συγκεκριμένα, χρησιμοποιούμε ένα μοντέλο δράστη-κριτή (actor-critic) με αναπαραστάσεις νευρωνικών δικτύων για μία γκαουσιανή πολιτική και μία συνάρτηση αξίας, τα οποία εκπαιδεύονται κατά την ενισχυτική μάθηση με μία μέθοδο βελτιστοποίησης πολιτικής βάσει περιοχών εμπιστοσύνης. Χρησιμοποιούμε, εκτός των πληροφοριών των διατάξεων ρομπότ και αντικειμένου και τις δυνάμεις αλληλεπίδρασης μεταξύ τους, ως κύρια πληροφορία διατήρησης της ζητούμενης εσωτερικής λαβής επαφής αλλά και με στόχο τον έλεγχο των δυνάμεων που αναπτύσσονται. Εφαρμόζουμε πειραματικά την μέθοδο στο πρόβλημα λαβής και ανύψωσης αντικειμένου υπό συγκεκριμένη επιθυμητή τροχιά ύψους και προσανατολισμού από το ανθρωπομορφικό ρομποτικό χέρι ADROIT στο περιβάλλον προσομοίωσης Mujoco, με την βοήθεια δεδομένων επίδειξης που λαμβάνονται με τηλεχειρισμό του ρομπότ, χρησιμοποιώντας τον αισθητήρα Leap Motion. Τα αποτελέσματα των πειραμάτων επιβεβαιώνουν την μέθοδο που προτείνουμε, αναδεικνύοντας τις δυνατότητες γενίκευσης της τροχιάς που έχει μάθει το σύστημα σχετικά με τον χρόνο και το τελικό ύψος ανύψωσης, ενώ παράλληλα επικυρώνουν την συμβολή των δεδομένων επίδειξης στην απόδοση της μάθησης και των αισθητήρων δύναμης στην επιτυχία της ζητούμενης λαβής.

Λέξεις Κλειδιά

Ρομποτικός Χειρισμός, Επιδέξιος Χειρισμός Εσωτερικής Λαβής, Ενισχυτική Μάθηση, Μάθηση από Δεδομένα Επίδειξης, Παρακολούθηση Τροχιάς, Ανάδραση Δυνάμεων Επαφής, Τηλεχειρισμός

Abstract

The goal of this thesis is the development of a method for learning of dexterous in-hand manipulation robotic skills, using reinforcement learning and demonstration data. The research begins with the problem of stochastic decision optimization for environments with unknown dynamics, using reinforcement machine learning. Afterwards, we explore methods of exploiting demonstration data to improve the performance of reinforcement learning and mimic the human behavior. We then focus on the problem of model-free learning for in-hand object manipulation tasks, in which it is desirable to follow a trajectory of the object's pose and control the magnitude of the contact forces. For this purpose, a neuro-controller is proposed; firstly trained with supervised learning from the demonstration data and then with reinforcement learning, in order to further optimize the acquired behavior. Specifically, we use an actor-critic model with neural network representations for a gaussian policy and a value function, which are trained with a trust region policy optimization method. Besides the poses of the robot and the object, we also use the contact forces for maintaining the desired in-hand contact grasp and for controlling the magnitude of the forces that are being developed. We apply the method experimentally to the problem of grasping and lifting an object under a specific desired trajectory of height and orientation by the anthropomorphic robotic hand ADROIT in the simulation environment Mujoco, with the help of demonstration data obtained using a Leap Motion sensor device. The experimental results validate the proposed method, underlining the possibilities of generalization of the learned trajectory regarding time and final lifting height, while highlighting the contribution of the demonstration data in terms of efficiency and the contribution of tactile feedback in the success of the desired in-hand grasp.

Keywords

Robotic Manipulation, Dexterous In-hand Manipulation, Reinforcement Learning, Learning from Demonstration, Trajectory Following, Tactile Feedback, Teleoperation

στους γονείς μου

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω τον καθηγητή Κωνσταντίνο Τζαφέστα για την επίβλεψη αυτής της διπλωματικής εργασίας, την καθοδήγηση και την ευκαιρία που μου έδωσε να την εκπονήσω στο Εργαστήριο Ρομποτικής του Τομέα Σημάτων, Ελέγχου και Ρομποτικής. Επίσης ευχαριστώ ιδιαίτερα τον υποψήφιο διδάκτωρ Παρασκευά Οικονόμου για την βοήθεια του, τις ενδιαφέρουσες συζητήσεις και την εξαιρετική συνεργασία που είχαμε. Τέλος θα ήθελα να ευχαριστήσω τους γονείς μου για την υποστήριξη και την συμπαράσταση που μου προσέφεραν όλα αυτά τα χρόνια.

Αθήνα, Ιούνιος 2021

Κοτσόβολης Στυλιανός

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Ρομποτική και Χειρισμός Αντικειμένων

Η λέξη ¨ρομπότ¨ προέρχεται από την σλαβική λέξη "robota" που σημαίνει δουλεία ή εξαναγκασμένη εργασία. Ο όρος διαδόθηκε σε πολλές χώρες και γλώσσες ανά τον κόσμο και αποδόθηκε εν τέλη στα ρομπότ με την έννοια προγραμματισμένων μηχανών, ικανών να εκπονήσουν αυτόματα ένα σύνολο εργασιών. Η μελέτη των ρομπότ οδήγησε στην ανάπτυξη της επιστήμης της ρομποτικής, η οποία συνδυάζει γνώσεις από πολλές άλλες επιστήμες όπως για παράδειγμα η μηχανική, η ηλεκτρονική, η πληροφορική ακόμα και η ψυχολογία.

Τα ρομπότ σήμερα χρησιμοποιούνται σε ένα μεγάλο εύρος εργασιών από την βιομηχανία μέχρι την οικιακή χρήση, ενώ η μορφή τους ποικίλει και ανάλογα με την εφαρμογή και τον σκοπό για τον οποίο σχεδιάζονται, μπορεί να είναι ανθρωποειδή, βραχίονες, επιδέξια χέρια, ρομπότ εδάφους, ιπτάμενα και άλλα. Συγκεκριμένα, τα ρομποτικά χέρια αποτελούν σημαντικό τομέα της ρομποτικής πάνω στον οποίο, ιδιαίτερα τα τελευταία χρόνια, εργάζεται πληθώρα επιστημόνων.

Ο άνθρωπος χρησιμοποιεί τα χέρια του καθημερινά για τις περισσότερες από τις εργασίες του αφού είναι το μέσο εκείνο που του προσφέρει δυνατότητα απτικής αλληλεπίδρασης με το περιβάλλον του. Τα χέρια του ανθρώπου αποτελούν ένα πολύ σύνθετο και επιδέξιο για αυτόν εργαλείο, αποτέλεσμα της εξελικτικής διαδικασίας πολλών χιλιάδων ετών, με το οποίο πραγματοποιεί το μεγαλύτερο μέρος των εργασιών του καθημερινά. Η δυνατότητα εκτέλεσης χειρισμών από τα ρομπότ παρόμοιων με αυτούς του ανθρώπου μπορεί να δώσει τεράστιες δυνατότητες εφαρμογών ιδιαίτερα στον τομέα της αλληλεπίδρασης ρομπότ-ανθρώπου (Human-Robot Interaction).

Για τον λόγο αυτό ο επιδέξιος χειρισμός αντικειμένων στην ρομποτική, από απλές λαβές μέχρι σύνθετες εργασίες χειρισμού εργαλείων, αποτελεί ένα ευρύ πεδίο μελέτης και έρευνας. Το πεδίο αυτό είναι ένα από τα πιο σύνθετα και με πολλές προκλήσεις στην ρομποτική, τόσο λόγω της δυσκολίας κατασκευής χεριών που προσεγγίζουν λεπτομερώς τα χαρακτηριστικά και τις δυνατότητες του ανθρώπινου, όσο και λόγω της δυσκολίας ελέγχου ενός τόσο σύνθετου εργαλείου ώστε να πραγματοποιεί επιτυχημένα εργασίες.

Τα τελευταία χρόνια έχουν σχεδιαστεί σύνθετα ρομποτικά χέρια με 5 δάχτυλα και βαθμούς ελευθερίας ίσους ή ακόμη και περισσότερους από του ανθρώπινου χεριού. Σχεδιασμένα με ¨μαλακές¨ επαφές στα άκρα και εξοπλισμένα με συστήματα πνευματικής οδήγησης τα χέρια αυτά είναι ικανά να εκπονήσουν επιδέξιες εργασίες ανάλογες με αυτές που κάνει ο άν-

θρωπος. Ανάλογα διάφορες τεχνικές ελέγχου έχουν αναπτυχθεί και συνεχώς βελτιώνονται, ώστε να είναι δυνατός ο έλεγχος των χεριών αυτών με στόχο την πραγματοποίηση εργασιών.

1.2 Ερευνητικές Προσεγγίσεις

Ο επιδέξιος ρομποτικός χειρισμός αντικειμένων θεωρείται ένα από τα πιο σύνθετα προβλήματα ελέγχου. Τόσο η κατασκευή ρομποτικών χεριών που έχουν την δυνατότητα επιδέξιων κινήσεων παρόμοιων με το ανθρώπινο χέρι, όσο και ο έλεγχος τους είναι προβλήματα που απασχολούν σήμερα την επιστημονική κοινότητα. Κάποιες προσεγγίσεις [5], [6] στοχεύουν στην κατασκευή χεριών απλούστερων από το ανθρώπινο, ώστε να απλοποιηθεί το πρόβλημα ελέγχου.

Σε άλλες εργασίες γίνονται προσεγγίσεις στο πρόβλημα ελέγχου με βελτιστοποίηση τροχιάς βάσει μοντέλου [7], [8], [9]. Αυτές οι μέθοδοι έχουν επιτυχία σε περιβάλλοντα προσομοίωσης όπου είναι πιο εύκολο να προσεγγιστεί η δυναμική του περιβάλλοντος, όμως δυσκολεύονται να προσαρμοστούν σε πραγματικά ρομπότ και εργασίες που παρουσιάζουν περίπλοκη δυναμική επαφών μεταξύ ρομπότ και αντικειμένων, καθώς η προσέγγιση ενός ακριβούς μοντέλου είναι πολύ δύσκολη. Γενικά, οι κλασσικές προσεγγίσεις αυτόματου έλεγχου παρέχουν τις κατευθυντήριες γραμμές, όμως περιορίζουν κατά πολύ το είδος και την πολυπλοκότητα των εργασιών. Κύρια αίτια είναι η δυσκολία προγραμματισμού και ελέγχου ρομπότ με πολλούς βαθμούς ελευθερίας και η δυσκολία σχεδιασμού μοντέλων δυναμικής που προσεγγίζουν με ακρίβεια ρομπότ και περιβάλλον.

Τα τελευταία χρόνια η ανάπτυξη της μηχανικής μάθησης και των τριών κατηγοριών της: επιβλεπόμενη, μη επιβλεπόμενη και ενισχυτική μάθηση, έχουν δώσει την αφορμή σε πολλές ομάδες ερευνητών να εφαρμόσουν και να αναπτύξουν την γνώση αυτή πάνω στον τομέα του ελέγχου επιδέξιων ρομποτικών χεριών. Συγκεκριμένα, η ενισχυτική μάθηση (Reinforcement Learning) προσεγγίζοντας το πρόβλημα της βελτιστοποίησης αποφάσεων δίνει την δυνατότητα επίλυσης προβλημάτων, τα οποία είναι δύσκολο να μοντελοποιηθούν, ενώ επιτρέπει στο ρομπότ να μάθει να εκπονεί εργασίες, χωρίς γνώση της δυναμικής, μέσω αλληλεπίδρασης με το περιβάλλον με την λογική προσπάθειας-λάθους (trial-error).

Η ενισχυτική μάθηση παρουσιάζει εντυπωσιακά αποτελέσματα στο τομέα του ελέγχου ρομποτικών χειριστών. Η απαλοιφή της ανάγκης κατασκευής ενός μοντέλου της δυναμικής του περιβάλλοντος (model-free Reinforcement Learning)[10], [11] που επιτρέπει την επίλυση πιο δύσκολων προβλημάτων ελέγχου είναι ένας από τους κύριους λόγους που καθιστούν την ενισχυτική μάθηση ένα χρήσιμο και αποτελεσματικό εργαλείο. Επιπλέον, η χρήση βαθιών νευρωνικών δικτύων στην ενισχυτική μάθηση, που συνεπάγεται στην βαθιά ενισχυτική μάθηση, έχει αποτελεσματικά διεισδύσει στην μάθηση ρομποτικών χειριστών, καθώς μπορεί να αντεπεξέλθει στο πρόβλημα της μεγάλης διάστασης και των συνεχών χώρων καταστάσεων και ενεργειών του συστήματος μάθησης. Μέθοδοι βαθιάς ενισχυτικής μάθησης (DRL) [12], [13] που χρησιμοποιούν βαθιά νευρωνικά δίκτυα για τις προσεγγίσεις πολιτικής και συνάρτησης αξίας, έχουν χρησιμοποιηθεί με μεγάλη επιτυχία σε περίπλοκες εργασίες χειρισμού.

Πολλές ερευνητικές ομάδες προσεγγίζουν το πρόβλημα του ρομποτικού χειρισμού αντικειμένων χρησιμοποιώντας δεδομένα επίδειξης από κάποιον ειδήμων, σε συνδυασμό ή μη με ενισχυτική μάθηση. Αφενός, τα δεδομένα επίδειξης χρησιμοποιούνται για την βελτίωση

της απόδοσης, ιδιαίτερα σε περιβάλλοντα και χειριστές μεγάλης διάστασης. Αφετέρου, σκοπός των δεδομένων επίδειξης είναι η μάθηση κινήσεων οι οποίες είναι ανθρώπινες, καθώς οι αλγόριθμοι ενισχυτικής μάθησης χωρίς πρότερη γνώση μπορεί να οδηγήσουν στην εύρεση κινήσεων που επιτυχώς εκπονούν την επιθυμητή εργασία, αλλά είναι αντίθετες και περίεργες σε σχέση με την ανθρώπινη λογική. Συνεπώς, ένας ακόμα τομέας μάθησης, αυτός της ¨Μάθησης από δεδομένα επίδειξης¨, έχει αναπτυχθεί.

Σε αρκετές εργασίες έχουν χρησιμοποιηθεί Dynamic Motor Primitives (DMPs)[14], τα οποία χρησιμοποιούν δυναμικά συστήματα που περιγράφουν τις τροχίες επίδειξης. Ωστόσο, ένα βασικό μειονέκτημα της μεθόδου είναι ότι σε πραγματικά περιβάλλοντα είναι δύσκολο κανείς να λάβει τέλεια δεδομένα επίδειξης και συνεπώς χρειάζεται βελτίωση της συμπεριφοράς που βασίζεται σε αυτά. Για τον λόγο αυτό τα DMPs έχουν χρησιμοποιηθεί σε συνδυασμό με ενισχυτική μάθηση [15], [16], ώστε να γίνει κάποια περαιτέρω βελτιστοποίηση πάνω σε αυτά. Η αναπαράσταση ωστόσο της πολιτικής που ακολουθεί το σύστημα της ενισχυτικής μάθησης με DMPs είναι σχετικά απλή σε σχέση με τις αναπαραστάσεις αλγορίθμων βαθιάς ενισχυτικής μάθησης, και συνεπώς ενώ παρουσιάζουν καλά αποτελέσματα σε απλά προβλήματα υπάρχει δυσκολία εφαρμογής με την αύξηση των βαθμών ελευθερίας των ρομποτικών χειριστών αλλά και της πολυπλοκότητας των εργασιών που το ρομποτ επιχειρεί να λύσει.

Τα πιο εντυπωσιακά αποτελέσματα έχουν επιδείξει μέθοδοι βαθιάς ενισχυτικής μάθησης που χρησιμοποιούν συνδυαστικά κάποια μέθοδο μάθησης από δεδομένα επίδειξης. Οι μέθοδοι αυτές μπορεί είτε να είναι κάποια μορφή επιβλεπόμενης μάθησης είτε αντίστροφη ενισχυτική μάθηση κατά την οποία η βελτιστοποίηση γίνεται σύμφωνα με τα δεδομένα επίδειξης αντί για κάποια ορισμένη ποσότητα από τον προγραμματιστή. Στην εργασία [17] μία μέθοδος βαθιάς ενισχυτικής μάθησης με εκπαίδευση εκτός-πολιτικής (off-policy Reinforcement Learning) εμπλουτίζεται με δεδομένα εκπαίδευσης και εφαρμόζεται επιτυχημένα σε εργασίες αρκετά μεγάλης δυσκολίας από ένα ρομπότ 7 βαθμών ελευθερίας. Στην εργασία [2] χρησιμοποιείται βαθιά ενισχυτική μάθηση εντός πολιτικής (on policy) και εφαρμόζεται επιτυχημένα σε ακόμη πολυπλοκότερες εργασίες χειρισμού από ρομπότ 30 βαθμών ελευθερίας.

Σε αρκετές εργασίες [18], [19], [20], [21], [22] χρησιμοποιούνται στον χώρο κατάστασης χαρακτηριστικά από εικόνες/βίντεο (Visual Reinforcement Learning) κατά την διαδικασία της μάθησης (με ή χωρίς δεδομένα επίδειξης), τα οποία εισάγονται μέσω συνελικτικών νευρωνικών δικτύων CNN στην διαδικασία της ενισχυτικής μάθησης. Οι εργασίες αυτές αποτελούν μία διαφορετική προσέγγιση και δείχνουν πως ακόμη και οπτικά, μη επεξεργασμένα δεδομένα από μία κάμερα μπορούν να εκπαιδεύσουν ένα σύστημα βαθιάς ενισχυτικής μάθησης ακόμη και χωρίς άμεση γνώση των μεγεθών του περιβάλλοντος.

Ερευνητικές μέθοδοι έχουν επίσης αναπτυχθεί στην περιοχή του ¨εσωτερικού¨ χειρισμού αντικειμένων (in-hand manipulation), η οποία αποτελεί ένα ξεχωριστό επιμέρους πρόβλημα του ρομποτικού χειρισμού. Το ανθρώπινα χέρια (παλάμη και δάκτυλα) αποτελούνται από πολλούς περισσότερους βαθμούς ελευθερίας σε σχέση με τους βραχίονες και μας επιτρέπουν να χειριζόμαστε εύκολα και επιδέξια αντικείμενα μεγάλου εύρους σχήματος, βάρους μεγέθους κλπ. Ανθρωπομορφικά χέρια έχουν χρησιμοποιηθεί σε εργασίες εσωτερικών χειρισμών με χρήση βαθιάς ενισχυτικής μάθησης [23], [24] κατά την οποία χρησιμοποιείται ως

feedback του αντικειμένου χαρακτηριστικά του αντικειμένου (π.χ. θέση και προσανατολισμός) ή και οπτική πληροφορία (π.χ. από κάμερα).

Ωστόσο, εντοπίζεται στις μεθόδους αυτές μία αδυναμία καλής απόδοσης, η οποία εστιάζεται στην απουσία feedback που σχετίζεται με την δύναμη επαφής στο αντικείμενο. Η δυνατότητα του ανθρώπου να χειρίζεται αντικείμενα με τα δάκτυλα του βασίζεται σε πολύ μεγάλο βαθμό στην απτική πληροφορία που δέχεται από το αντικείμενο. Ομοίως και στα ρομποτικά χέρια η χρήση αισθητήρων μέτρησης της δύναμης επαφής με το αντικείμενο (tactile feedback) και η αξιοποίηση της πληροφορίας αυτής κατά την διαδικασία της ενισχυτικής μάθησης βοηθά σε μεγάλο βαθμού την απόδοση των αλγορίθμων [25], [26],[18]. Στην [27] η πληροφορία των δυνάμεων αξιοποιείται περαιτέρω, ώς μία εγγύηση διατήρησης της επαφής μεταξύ ρομποτ και αντικειμένου. Σε άλλες εργασίες [28], [29], η οι αισθητήρες δύναμης χρησιμοποιούνται ως αντιμετώπιση του προβλήματος της αβεβαιότητας. Όπως δηλαδή ένας άνθρωπος μπορεί να βασίσει τον χειρισμό ενός αντικειμένου στην αίσθηση της αφής χωρίς να το κοιτάει, έτσι και σε ρομποτικές πληροφορίες η απτική πληροφορία μπορεί να απαλείψει την ανάγκη οπτική πληροφορίας του αντικειμένου ή την γνώση μεγεθών όπως το μέγεθος. Συνεπώς, η απτική πληροφορία μπορεί να δώσει λύση σε προβλήματα χειρισμού αντικειμένων άγνωστων και διάφορων μεγεθών ή σχημάτων.

Ως προς τα προβλήματα εσωτερικών χειρισμών, αυτά ποικίλουν. Πολύ δημοφιλή είναι προβλήματα της Open AI [23] που χρησιμοποιούνται σαν προβλήματος εφαρμογής αλγορίθμων σε πολλές από τις εργασίες που αναφέραμε, στα οποία το χέρι, με την παλάμη σταθερή και τοποθετημένη προς τα πάνω, χειρίζεται ένα αντικείμενο (π.χ. κύβος, ράβδος,..) ¨χτυπώντας¨ το, ώστε να το φέρει σε κάποιο επιθυμητό προσανατολισμό. Άλλα προβλήματα [27] αφορούν λαβές και επαναπροσανατολισμό του αντικειμένου με κάποια στροφή. Οι εργασίες που ερευνώνται, ωστόσο, με εσωτερικές λαβές ποικίλουν και μπορούν να αφορούν οποιαδήποτε εργασίες χειρισμού από τα δάκτυλα του ρομπότ.

1.3 Αντικείμενο της Εργασίας

Αντικείμενο αυτής της διπλωματικής εργασίας είναι η μάθηση εσωτερικών (in-hand) επιδέξιων κινήσεων χειρισμού αντικειμένων από ένα ρομποτικό χέρι. Συγκεκριμένα, επιθυμούμε να πραγματοποιήσουμε λαβή και στην συνέχεια ανύψωση ενός αντικειμένου σε κάποιο ύψος χρησιμοποιώντας τα δάκτυλα του ρομποτικού χεριού. Η κίνηση αυτή, θέλουμε να γίνεται in-hand, δηλαδή με σταθερή θέση του βραχίονα πάνω από το αντικείμενο και χρησιμοποιώντας τις άκρες των δακτύλων ως επαφές ρομπότ-αντικειμένου. Μία τέτοια λαβή παρέχει πολλές δυνατότητες χειρισμού του αντικειμένου αναφορικά με την θέση του, τον προσανατολισμό του, αλλά και την ασκούμενη δύναμη, σε σχέση με ένα απλό grasp στο οποίο απλώς κλείνουν τα δάκτυλα και στην συνέχεια η ο βραχίωνας πραγματοποιεί την ανύψωση.

Κίνητρο του προβλήματος αποτελούν εργασίες χειρισμού που απαιτούν κάποια δεξιότητα από το ρομπότ. Για παράδειγμα, ευαίσθητα αντικείμενα όπως φυτά, τρόφιμα κλπ χρειάζονται ειδική μεταχείριση όταν ένα ρομπότ αλληλεπιδρά μαζί τους και μία απλή λαβή ίσως τραυματίσει το αντικείμενο. Εντοπίζουμε, επιπλέον το πρόβλημα, ότι ακόμη και μία in-hand λαβή όπως προτείνεται από τις υπάρχουσες προσεγγίσεις του προβλήματος ίσως να

μην είναι αρκετή. Ουσιαστικά, οι προσεγγίσεις αυτές στοχεύουν στον χειρισμό κάποιου αντικειμένου ώστε αυτό να καταλήγει σε κάποια τελική κατάσταση, για παράδειγμα έναν τελικό προσανατολισμό. Ωστόσο, η ενδιάμεσες καταστάσεις, δηλαδή ολόκληρη η κίνηση του αντικειμένου δεν λαμβάνονται υπ' όψιν. Έχοντας εντοπίσει το πρόβλημα αυτό, στοχεύουμε στην κατασκευή μίας μεθόδου που επιτρέπει την παρακολούθηση τροχιάς κάποιου ή κάποιων μεγεθών του αντικειμένου και που προσφέρει γενίκευση σε αυτές, ώστε να είναι δυνατή η εκτέλεση πλήθους επιθυμητών τροχιών. Επιπλέον, κατανοώντας την σημασία της απτικής πληροφορίας, χρησιμοποιούμε αισθητήρες δύναμης προκειμένου να ελέγχουμε την διατήρηση της επαφής ρομπότ-αντικειμένου ώστε να πραγματοποιηθεί η ζητούμενη λαβή επαφής των άκρων των δακτύλων και ελέγχουμε επιπλέον την ασκούμενη δύναμη, ώστε αυτή να διατηρείται εντός επιθυμητών ορίων.

Πιο συγκεκριμένα, στην παρούσα εργασία ασχολούμαστε με το πρόβλημα της ανύψωσης ενός αντικειμένου υπό συγκεκριμένη επιθυμητή προκαθορισμένη τροχιά ύψους, διατηρώντας παράλληλα το αντικείμενο οριζόντιο, ενώ παράλληλα επιθυμούμε να ελέγχουμε το εύρος των δυνάμεων που ασκούνται σε αυτό. Πειραματικά, χρησιμοποιούμε για τον σκοπό αυτό ένα ανθρωπομορφικό χέρι που προσομοιώνει τους βαθμούς ελευθερίας του ανθρώπινου χεριού, ώστε να είναι δυνατές επιδέξιες εργασίες. Όπως έχει ήδη διαπιστωθεί κλασσικές προσεγγίσεις ελέγχου, ιδιαίτερα πολυδιάστατων προβλημάτων δίνουν μη ικανοποιητικές λύσεις. Χρησιμοποιούμε, για τον λόγο αυτό, για την λύση του προβλήματος έναν νευρωνικό ελεγκτή ο οποίος εκπαιδεύεται με ενισχυτική μάθηση. Η ενισχυτική μάθηση, ωστόσο, αποτελεί ένα δύσκολο πρόβλημα, στο οποίο μεγάλες διάσταση του προβλήματος, όπως στην περίπτωση ενός ανθρωπομορφικού χεριού, επιδρούν σε μεγάλο βαθμό στην απόδοση της μάθησης. Για τον λόγο αυτό, χρησιμοποιούμε επιπλέον δεδομένα επίδειξης ώστε να μειώσουμε τον αριθμό των δειγμάτων που χρειάζονται κατά την διαδικασία της μάθησης και συνεπώς να βελτιώσουμε την απόδοση.

1.4 Οργάνωση της εργασίας

Η εργασία είναι οργανωμένη σε 7 βασικά κεφάλαια, το πρώτο από τα οποία είναι η εισαγωγή αυτή. Τα δύο επόμενα αποτελούν το πειραματικό υπόβαθρο, το 4ο την μεθοδολογία που προτείνουμε, το 5ο και το 6ο το πειραματικό μέρος και το τελευταίο είναι ο επίλογος.

Στο κεφάλαιο 2 γίνεται παρουσίαση τεχνικών ενισχυτικής μάθησης. Ξεκινώντας από την περιγραφή των μαρκοβιανών αλυσίδων αποφάσεων ως μαθηματικό εργαλείο μοντελοποίησης του προβλήματος της στοχαστικής βελτιστοποίησης αποφάσεων, συνεχίζουμε στους αλγορίθμους δυναμικού προγραμματισμού, οι οποίοι χρησιμοποιούν γνώση του μοντέλου του περιβάλλοντος. Στην συνέχεια, παρουσιάζεται η ενισχυτική μάθηση χωρίς γνώση του μοντέλου του περιβάλλοντος και περιγράφονται οι βασικοί αλγόριθμοι πρόβλεψης και ελέγχου εντός και εκτός πολιτικής. Έπειτα, αναλύονται προσεγγιστές, όπως τα νευρωνικά δίκτυα στην ενισχυτική μάθηση, και περιγράφονται αλγόριθμοι βαθιάς ενισχυτικής μάθησης.

Στο κεφάλαιο 3 παρουσιάζεται η βασική θεωρία της μάθησης από δεδομένα επίδειξης. Αρχικά, περιγράφουμε προσεγγίσεις όπου χρησιμοποιείται αναγωγή σε μάθηση με επίβλεψη και στην συνέχεια παρουσιάζεται η αντίστροφη ενισχυτική μάθηση.

Στο κεφάλαιο 4 παρουσιάζεται η μεθοδολογία που προτείνουμε για την επίλυση του

προβλήματος δηλαδή, την εσωτερική λαβή και ανύψωση ενός αντικειμένου από ένα ρομποτικό χέρι, ακολουθώντας κάποια επιθυμητή τροχιά ύψους και διατηρώντας τις ασκούμενες δυνάμεις εντός επιθυμητού εύρους.

Στο κεφάλαιο 5 περιγράφουμε την πειραματική διάταξη που χρησιμοποιήθηκε και τον τρόπο με τον οποίο αποκτήθηκαν τα δεδομένα επίδειξης που χρησιμοποιούνται στα πειράματα. Συγκεκριμένα περιγράφεται το περιβάλλον προσομοίωσης Mujoco, τα χαρακτηριστικά του ανθρωπομορφικού ρομποτικού χεριού ADROIT και η διαδικασία μέσω της οποίας λαμβάνουμε τα δεδομένα επίδειξης από τον αισθητήρα Leap Motion.

Στο κεφάλαιο 6 γίνεται παρουσίαση των πειραματικών αποτελεσμάτων της μεθόδου που προτείνουμε. Παρουσιάζουμε, την απόδοση του συστήματος που προτείνουμε, την επίδραση των δεδομένων επίδειξης στην διαδικασία της μάθησης, τα αποτελέσματα μάθησης μία τροχιάς κίνησης σε σχέση με την κλασσική προσέγγιση η οποία αφορά την επίτευξη ενός τελικού στόχου, τις δυνατότητες γενίκευσης του συστήματος σε σχέση με νέες τροχιές και την επίδραση των αισθητήρων δύναμης.

Τέλος, στο κεφάλαιο 7 και επίλογο της εργασίας γίνεται μία συζήτηση σχετικά με τα αποτελέσματα και τα συμπεράσματα που εξάγονται από αυτή την διπλωματική εργασία και προτείνονται μελλοντικές ερευνητικές επεκτάσεις και νέες κατευθύνσεις.

Κεφάλαιο 2

Ενισχυτική Μάθηση

Η ενισχυτική μάθηση αποτελεί το είδος μάθησης στο οποίο ένας πράκτορας μαθαίνει αλληλεπιδρώντας συνεχώς με το περιβάλλον του, μέσω ενός σήματος-ανταμοιβής το οποίο λαμβάνει από το περιβάλλον ως συνέπεια των πράξεων που επιλέγει. Η ιδέα της αλληλεπίδρασης με το περιβάλλον είναι βαθιά συνδεδεμένη με ίδια την φύση της μάθησης. Αποτέλεσε πρωτίστως αντικείμενο μελέτης της συμπεριφορικής ψυχολογίας και εισήχθη στο χώρο του αυτόματου ελέγχου και της τεχνητής νοημοσύνης αρχικά από τον Dr. Richard S. Sutton [1].

Η ιδέα της ενισχυτικής μάθησης σχετίζεται με την προτίμηση των έμβιων οργανισμών να επιλέγουν δράσεις, οι οποίες στο παρελθόν είχαν κάποια θετική συνέπεια για αυτούς και ομοίως να αποφεύγουν ενέργειες με αρνητική επίδραση. Όταν ένας βρέφος μαθαίνει να περπατά, για παράδειγμα, μέσα από πολλές προσπάθειες μαθαίνει να επιλέγει κινήσεις με τις οποίες στέκεται όρθιο περισσότερη ώρα και δεν πέφτει. Όταν κανείς ξεκινά να παίζει ένα παιχνίδι όπως για παράδειγμα το σκάκι, μαθαίνει μέσα από πολλές επαναλήψεις του παιχνιδιού κινήσεις και στρατηγικές που οδηγούν στην νίκη και ακολουθώντας τες γίνεται καλύτερος. Οργανωμένη πάνω σε αυτή την λογική, στην ενισχυτική μάθηση που χρησιμοποιείται στην τεχνητή νοημοσύνη, ένας πράκτορας (agent) μαθαίνει να λαμβάνει αποφάσεις προσπαθώντας να μεγιστοποιήσει ένα σήμα επιβράβευσης (reward signal). Αντίθετα με την επιβλεπόμενη μάθηση, δεν χρησιμοποιείται κάποιος δάσκαλος, που υποδεικνύει πως θα φτάσει ο πράκτορας τον στόχο του, αλλά εξερευνώντας το περιβάλλον του, αξιολογεί της πράξεις που επιλέγει και μαθαίνει ποιες από αυτές θα τον οδηγήσουν στο στόχο του.

Σημαντικά στοιχεία της ενισχυτικής μάθησης εκτός από τον πράκτορα και το περιβάλλον, είναι :

- Η πολιτική (policy), η οποία αφορά τον τρόπο που δρα ο πράκτορας. Βρισκόμενοι σε μια κατάσταση του περιβάλλοντος, η πολιτική ορίζει τις πράξεις που θα λάβει ο πράκτορας στην κατάσταση αυτήν.

- Το σήμα επιβράβευσης (reward signal), το οποίο καθορίζει τον στόχο του προβλήματος. Μετά από κάθε πράξη, ο πράκτορας λαμβάνει ένα reward (αριθμητική τιμή) από το περιβάλλον που σχετίζεται με την επίτευξη του στόχου, και δηλώνει ουσιαστικά πόσο καλή ή κακή ήταν η απόφαση του.

- Η συνάρτηση τιμής (value function). Η επιβράβευση αφορά μία βραχυπρόθεσμη αξιολόγηση της πράξης του πράκτορα. Ωστόσο, στα περισσότερα προβλήματα είναι αναγκαίο ο πράκτορας να αναπτύει πιο σύνθετες στρατηγικές και να πάρει αποφάσεις

που στοχεύουν στην μακροπρόθεσμη επίτευξη ενός στόχου, οι οποίες δεν οδηγούν αναγκαία σε μεγάλη άμεση επιβράβευση. Η συνάρτηση τιμής αναλαμβάνει τον σκοπό αυτό.

- Το μοντέλο του περιβάλλοντος (model), το οποίο αφορά το πως αντιδρά το περιβάλλον στις ενέργειες του πράκτορα, και μπορεί ή μπορεί να είναι ή να μην είναι γνωστό.

Σε αυτό το κεφάλαιο παρουσιάζουμε αλγορίθμους και μεθόδους που χρησιμοποιούνται στην ενισχυτική μάθηση. Για μία πιο αναλυτική περιγραφή των βάσεων της ενισχυτικής μάθησης παραπέμπουμε στο βιβλίο "An Introduction to Reinforcement Learning" [1].

2.1 Μαρκοβιανές διαδικασίες λήψης αποφάσεων (MDPs)

Για την περιγραφή του περιβάλλοντος ενός προβλήματος ενισχυτικής μάθησης χρησιμοποιούνται ως μαθηματικό μοντέλο οι Μαρκοβιανές διαδικασίες λήψης αποφάσεων (Markov Decision Processes - MDPs), οι οποίες αποτελούν αντικείμενο των στοχαστικών διαδικασιών. Αρχικά λοιπόν είναι σκόπιμο να περιγραφούν τα στοιχεία που συνθέτουν μία τέτοια αλυσίδα αποφάσεων.

Το κύριο δομικό στοιχείο μίας Μαρκοβιανής Διαδικασίας Λήψης Αποφάσεων είναι η Μαρκοβιανή Διαδικασία (Decision Process). Μία Μαρκοβιανή Διαδικασία είναι ένα ζεύγος διακριτών κατάστάσεων S και πιθανοτήτων μεταβάσεων P μεταξύ των καταστάσεων: $M =< S, P >$, για το οποίο ικανοποιείται η Μαρκοβιανή Ιδιότητα.

$$P[S_{t+1}|S_1, S_2, ..., S_t] = P[S_{t+1}|S_t]$$

Η Μαρκοβιανή ιδιότητα δηλώνει την ανεξαρτησία του παρόντος από το παρελθόν. Δεδομένου, δηλαδή, ότι το σύστημα βρίσκεται σε μία κατάσταση, η πιθανότητα να μεταβεί σε κάποια επόμενη κατάσταση δεν επηρεάζεται από τις καταστάσεις στις οποίες είχε βρεθεί στο παρελθόν.

Οι Μαρκοβιανές Διαδικασίες επεκτείνονται σε Μαρκοβιανές Διαδικασίες με επιβράβευση (Markov Reward Processes - MRPs), όταν κάθε κατάσταση παράγει ένα σήμα επιβράβευσης R_s, και συνεπώς μπορεί να αναπαρασταθεί ως μία τετράδα $M =< S, P, R, \gamma >$, όπου $R : S \rightarrow \mathbb{R}$, η συνάρτηση επιβράβευσης και $\gamma \in [0, 1]$, ένας συντελεστής μείωσης της επιβράβευσης, ο ρόλος του οποίου αναλύεται στην συνέχεια.

Σε ένα μαρκοβιανό πρόβλημα αποφάσεων, στόχος του πράκτορα είναι να παίρνει αποφάσεις ώστε να μεγιστοποιήσει την συσσώρευση μελλοντικών ανταμοιβών, η οποία μετράται με το σήμα αθροιστικών επιβραβεύσεων ή απόδοση G_t (return). Το σήμα αθροιστικών επιβραβεύσεων ή απόδοση είναι μία συνάρτηση της ακολουθίας ανταμοιβών και μπορεί να οριστεί με διάφορους τρόπους. Έστω ότι το σύστημα βρίσκεται στην χρονική στιγμή t. Τότε από εκεί και έπειτα θα λάβει μία ακολουθία σημάτων ανταμοιβής $R_{t+1}, R_{t+2}, R_{t+2}, ...$ Δύο από τους τρόπους που μπορεί να οριστεί η απόδοση G_t είναι:

- Μοντέλο πεπερασμένου ορίζοντα. Σε αυτή την περίπτωση προβλημάτων υπάρχει ένα τελικό χρονικό βήμα T, και η απόδοση δίνεται ως το άθροισμα όλων των επιβραβεύσε-

ων:

$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + ... + R_T$$

- Μοντέλο φθίνουσας ανταμοιβής. Σε προβλήματα στα οποία δεν υπάρχει φυσικό όριο τελικού βήματος ($T \to \infty$) ή θέλουμε να λάβουμε υπόψιν μελλοντικές επιβραβεύσεις, δίνοντας όμως προτεραιότητα στις άμεσα μελλοντικές, η απόδοση μπορεί να σχεδιαστεί με χρήση ενός συντελεστή μείωσης $\gamma \in [0, 1]$, ως:

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

  Ο συντελεστής γ ουσιαστικά καθορίζει το διάστημα που ο πράκτορας λαμβάνει υπόψιν τις ανταμοιβές. Στην ακραία περίπτωση $\gamma = 0$, ο πράκτορας συμπεριφέρεται μυωπικά, καθώς λαμβάνει υπόψιν μόνο την άμεση ανταμοιβή, ενώ όσο το γ αυξάνει και προσεγγίζει το 1, ο πράκτορας λαμβάνει υπόψιν και μεταγενέστερες ανταμοιβές.

Σχεδόν όλοι οι αλγόριθμοι ενισχυτικής μάθησης χρησιμοποιούν συναρτήσεις αξίας, προκειμένου να μετρούν πόσο καλή είναι μία κατάσταση. Η συνάρτηση αξίας έχει την έννοια των μελλοντικών επιβραβεύσεων που αναμένονται. Συνεπώς, η συνάρτηση αξίας καταστάσεων σε μία Μαρκοβιανή Διαδικασία με επιβραβεύσεις ορίζεται ως η εκτίμηση της απόδοση G_t.

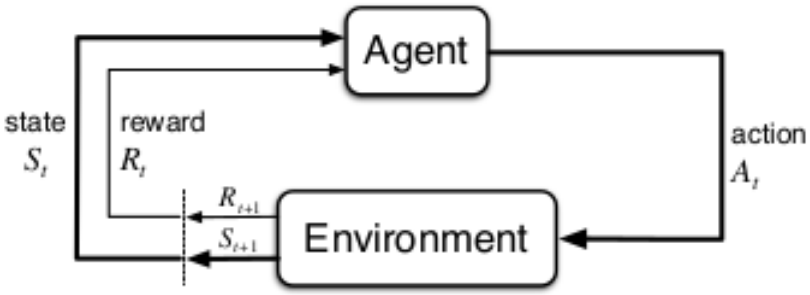
$$V(s) = \mathbb{E}[G_t|S_t = s] = \mathbb{E}[R_{t+1} + \gamma G_{t+1}|S_t = s] = ... = R_s + \gamma \sum_{s' \in S} P_{ss'} V(s')$$

Η αναδρομική αυτή έκφραση της συνάρτησης αξίας είναι μια πολύ σημαντική ιδιότητα των Μαρκοβιανών διαδικασιών που χρησιμοποιείται στους περισσότερους αλγορίθμους λύσης τους και ονομάζεται εξίσωση Bellman. Η εξίσωση Bellman μπορεί να λυθεί αναλυτικά ως $V = (I - \gamma P)^{-1} R$, με κόστος πολυπλοκότητας $O(n^3)$, το οποίο όμως δυσχεραίνει την λύση όταν οι καταστάσεις είναι πολλές. Συνεπώς η λύση προκύπτει από επαναληπτικές διαδικασίες, όπως ο Δυναμικός προγραμματισμός και η Ενισχυτική Μάθηση.

Εμπλουτίζοντας τα MRPs, με την δυνατότητα λήψης αποφάσεων από ένα σύνολο διακριτών πράξεων (actions) A, παράγονται οι Μαρκοβιανές Διαδικασίες Λήψης Αποφάσεων MDPs. Αυτές μπορούν να περιγραφούν από την πεντάδα $M =< S, A, P, R, \gamma >$, όπου P ο νέος πίνακας μετάβασης καταστάσεων, στην οποία ενσωματώνεται η λήψη δράσης $a \in A$ από τον πράκτορα $P^a_{ss'} = P[S_{t+1} = s'|S_t = s, A_t = a]$ και R η νέα συνάρτηση επιβραβεύσεων R^a_s με δράσεις.

Σε ένα μαρκοβιανό μοντέλο λήψης αποφάσεων υπάρχει ένας πράκτορας ο οποίος επιλέγει με κάποια πολιτική (policy) να επιλέγει ενέργειες-δράσεις. Εάν ο πράκτορας ακολουθεί την πολιτική π την χρονική στιγμή t τότε η $\pi(a|s)$ δηλώνει την πιθανότητα ο πράκτορας να επιλέξει την πράξη $A_t = a$ βρισκόμενος στην κατάσταση $S_t = s$, δηλαδή $\pi(a|s) = P[A_t = a|S_t = s]$. Η δράση αυτή οδηγεί το περιβάλλον σε νέα κατάσταση S_{t+1} και δίνει επιβράβευση R_{t+1}, όπου ο πράκτορας θα μεταβεί σε λήψη νέας δράσης βάσει της πολιτικής.

Επομένως, δοσμένου ενός MDP και μίας πολιτικής π, η ακολουθία καταστάσεων που προκύπτει $S_1, S_2, S_3, ...$ είναι μία μαρκοβιανή διαδικασία $M =< S, R^\pi >$, ενώ η ακολουθία καταστάσεων και επιβραβεύσεων $S_1, R_2, S_2, R_3, S_3, ...$ είναι μια μαρκοβιανή διαδικασία με επιβράβευση $M =< S, P^\pi, R^\pi, \gamma >$, όπου τα $P^\pi_{ss'}$ και R^π_s γράφονται $P^\pi_{ss'} = \sum_{a \in A} \pi(a|s) P^a_{ss'}$ και



Σχήμα 2.1: *Αλληλεπίδραση πράκτορα - περιβάλλοντος (από [1])*

$R_s^\pi = \sum_{a \in A} \pi(a|s) R_s^a$.

Μπορούμε τώρα να ορίσουμε την συνάρτηση αξίας για Μαρκοβιανές Διαδικασίες Λήψης Αποφασεων, υπό την πολιτική π ως το αναμενόμενο σήμα αθροιστικών επιβραβεύσεων ή αποδόσεων (expected return):

$$V_\pi(s) = \mathbb{E}[G_t|S_t = s] = \mathbb{E}[\sum_{k=1}^{\infty} \gamma^k R_{t+k+1}|S_t = s]$$

Ομοίως μπορούμε να ορίσουμε την αξία ο πράκτορας να επιλέξει την δράση a στην κατάσταση s, υπό την πολιτική π, η οποία ονομάζεται συνάρτηση πράξης-αξίας (action-value function).

$$Q_\pi(s, a) = \mathbb{E}[G_t|S_t = s, A_t = a] = \mathbb{E}[\sum_{k=1}^{\infty} \gamma^k R_{t+k+1}|S_t = s, A_t = a]$$

.

Οι δύο παραπάνω συναρτήσεις αξίας συνδέονται με την εξίσωση:

$$V_\pi(s) = \sum_{a \in A} \pi(a|s) Q_\pi(s, a)$$

Η εξισώσεις Bellman λοιπόν σε ένα MDP γίνονται:

$$V_\pi(s) = \sum_{a \in A} \pi(a|s) Q_\pi(s, a) = \sum_{a \in A} \pi(a|s)(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V_\pi(s'))$$

$$Q_\pi(s, a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V_\pi(s') = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \sum_{a' \in A} \pi(a'|s') Q_\pi(s', a')$$

Η λύση του προβλήματος συνίσταται στην εύρεση βέλτιστης πολιτικής π_* η οποία δίνει βέλτιστη συνάρτηση αξίας $V_*(s)$ και συνεπώς βέλτιστη συνάρτηση δράσης-αξίας $Q_*(s)$.Αυτές καθορίζουν και την βέλτιστη λύση που μπορεί να επιτευχθεί σε ένα μαρκοβιανό πρόβλημα λήψης αποφάσεων. Λαμβάνοντας το μέγιστο των συναρτήσεων αξίας, λοιπόν οδηγούμαστε στην βέλτιστη πολιτική:

$$V_*(s) = \max_\pi V_\pi(s) = \max_a (R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V_*(s'))$$

$$Q_*(s,a) = \max_\pi Q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \max_a Q_*(s',a')$$

,οι οποίες είναι οι εξισώσεις Βελτίστου Bellman.

2.2 Δυναμικός Προγραμματισμός

Ο όρος δυναμικός προγραμματισμός αναφέρεται στους αλγορίθμους που χρησιμοποιούνται για την εύρεση βέλτιστων πολιτικών σε προβλήματα, στα οποία θεωρούμε γνωστό το μοντέλο του περιβάλλοντος, δηλαδή τις μεταβάσεις καταστάσεων P, μιάς Μαρκοβιανής διαδικασίας λήψης αποφάσεων. Γενικά ο δυναμικός προγραμματισμός λύνει ένα σύνθετο πρόβλημα, χωρίζοντας το σε απλούστερα προβλήματα, και έτσι, συνδυάζοντας τις επιμέρους λύσεις των υποπροβλημάτων, προβαίνει στην λύση του συνολικού. Ωστόσο, ο δυναμικός προγραμματισμός δεν χρησιμοποιείται στην πράξη σε προβλήματα ενισχυτικής μάθησης τόσο λόγω του γεγονότος ότι απαιτεί μοντέλο δυναμικής του περιβάλλοντος και επιπλέον λόγω του ότι δεν εφαρμόζεται σε συνεχή περιβάλλοντα καταστάσεων ή δράσεων. Παρουσιάζεται, ωστόσο, στην συνέχεια, προς κατανόηση των αλγορίθμων ενισχυτικής μάθησης.

2.2.1 Αξιολόγηση Πολιτικής (Πρόβλεψη)

Εαν θέλουμε να υπολογίσουμε την συνάρτηση αξίας μίας πολιτικής, τότε μπορούμε να επαναληπτικά σε κάθε βήμα k να ανανεώνουμε το $V(s)$ για κάθε $s \in S$ με χρήση των την αξία των επόμενων καταστάσεων s' που έχουν υπολογιστεί σε προηγούμενο βήμα. Οι ανανεώσεις αυτές γίνονται σύγχρονα οπότε μπορούμε να γράψουμε:

$$V_{k+1}(s) = \sum_{a \in A} \pi(a|s)(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V_k(s'))$$

$$v^{k+1} = R^\pi + \gamma P^\pi v^k$$

Αποδεικνύεται ότι για $k \to \infty$ η συνάρτηση αξίας αυτή συγκλίνει στην V_π. Το πρόβλημα αξιολόγησης της πολιτικής (Policy Evaluation) με υπολογισμό ή εκτίμηση της συνάρτησης αξίας ή συνάρτησης αξίας καταστάσεων ονομάζεται πρόβλεψη.

2.2.2 Βελτίωση Πολιτικής (Έλεγχος)

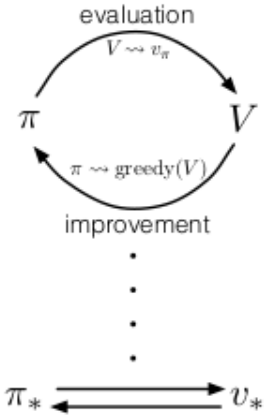
Το επόμενο βήμα, γνωρίζοντας την συνάρτηση αξίας με την μέθοδο αξιολόγησης είναι να βελτιώσουμε την πολιτική (Policy Improvement), ώστε να βρούμε τελικά μία βέλτιστη πολιτική π_*. Το πρόβλημα αυτό στον δυναμικό προγραμματισμό και στην ενισχυτική μάθηση αναφέρεται ως πρόβλημα ελέγχου. Στον δυναμικό προγραμματισμό μπορεί να επιτευχθεί με άπληστο (greedy) τρόπο. Συγκεκριμένα, έχοντας μία πολιτική π σε κάθε κατάσταση επιλέγουμε κάθε φορά ντετερμινιστικά την δράση που μας οδηγεί σε επόμενη κατάσταση με την μέγιστη συνάρτηση αξίας και ανανεώνουμε έτσι την πολιτική σε νέα π':

$$\pi'(s) = \arg\max_a Q_\pi(s,a) = \arg\max_a \sum_{s' \in S} P_{ss'}^a V_\pi(s')$$

.

Η επαναληπτικής διαδικασία βημάτων αξιολόγησης και βελτίωσης της πολιτικής με ντετερμινιστικό τρόπο ονομάζεται επανάληψη πολιτικής (policy iteration) και αποδεικνύεται ότι συγκλίνει στην βέλτιστη πολιτική π_*.

Η διαδικασία αυτή της αλληλουχίας δύο βημάτων είναι δομικό συστατικό πολλών αλγορίθμων ενισχυτικής μάθησης και αναφέρεται ως γενικευμένη επανάληψη πολιτικής (Generalized policy iteration). Ο πράκτορας, δηλαδή, δρα με βάση μία πολιτική, η οποία αξιολογείται και ανανεώνεται βάση κάποιας μορφής συνάρτησης αξίας, η οποία με την σειρά της καθορίζεται και υπολογίζεται βάσει της παρούσας πολιτικής που ακολουθεί ο πράκτορας.



Σχήμα 2.2: *Γενικευμένη Επανάληψη Πολιτικής (Generalized Policy Iteration) (από [1])*

2.2.3 Επανάληψη Αξίας

Ένα μειονέκτημα της επανάληψης πολιτικής στον δυναμικό προγραμματισμό είναι ότι σε κάθε βήμα γίνεται αξιολόγηση πολιτικής policy evaluation, η οποία είναι από μόνη της επαναληπτική διαδικασία, προκειμένου να συγκλίνει η πολιτική που υπολογίζεται στην V_π. Το ερώτημα που τίθεται είναι εάν χρειάζονται όλα τα βήματα στην αξιολόγηση πολιτικής και συνεπώς η σύγκλιση στην πραγματική τιμή της συνάρτησης αξίας ή εάν μπορούμε να επιλέξουμε μόνο με κάποια βήματα την επόμενη ανανέωσης της πολιτικής (με βάση την μέθοδο βελτίωσης πολιτικής). Αποδεικνύεται ότι η ιδιότητα σύγκλισης στην βέλτιστη πολιτική διατηρείται ακόμη και με ένα βήμα στην διαδικασία αξιολόγησης πολιτικής. Η μέθοδος αυτή ονομάζεται επανάληψη αξίας (Value Iteration). Εφόσον έχουμε ένα μόνο βήμα αξιολόγησης και στην συνέχεια επιλογής της νέας πολιτικής όπως ορίζει η βελτίωση πολιτικής, δηλαδή επιλέγοντας την πολιτικής με μέγιστη συνάρτηση αξίας, μπορούμε να γράψουμε την ανανέωση ως:

$$V_{k+1}(s) = \max_{a \in A}(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V_k(s'))$$

ή χρησιμοποιώντας πίνακες:

$$v^{k+1} = \max_{a \in A}(R^a + \gamma P^a v^k)$$

2.3 Πρόβλεψη σε MDPs άγνωστου μοντέλου

Όταν το μοντέλο P μιάς μαρκοβιανής διαδικασίας είναι άγνωστο ο δυναμικός προγραμματισμός αποτυγχάνει να δώσει λύση. Για τον λόγο αυτό χρησιμοποιείται, η ενισχυτική μάθηση, με την οποία γίνεται εκτίμηση των συναρτήσεων αξίας, μέσω αλληλεπιδράσεων με το περιβάλλον ώστε να βρεθούν οι βέλτιστες πολιτικές, χωρίς γνώση για το μοντέλο του περιβάλλοντος (model-free reinforcement learning).

2.3.1 Μέθοδος Monte Carlo

Θα θέλαμε να υπολογίσουμε την συνάρτηση αξίας $V_\pi(s) = \mathbb{E}[G_t|S_t = s]$, μίας πολιτικής π η οποία λαμβάνει αποφάσεις με κάποιον τρόπο, χωρίς όμως να γνωρίζουμε το μοντέλο της Μαρκοβιανής Διαδικασίας Λήψης Αποφάσεων. Το πρόβλημα λοιπόν είναι πως δεν γνωρίζουμε την απόδοση G_t εξαρχής, αλλά αντίθετα θα πρέπει μέσω αλληλεπίδρασης με το περιβάλλον να εκτιμηθεί. Με άλλα λόγια, χωρίς γνώση του μοντέλου, είναι αδύνατον ο πράκτορας να γνωρίζει εξαρχής που θα καταλήξει μία αλληλουχία πράξεων, αφού δεν γνωρίζει τις μελλοντικές καταστάσεις στις οποίες θα μεταβεί.

Με την μέθοδο Monte Carlo ο πράκτορας πραγματοποιεί ένα ολόκληρο επεισόδιο βάσει μίας πολιτικής και συλλέγει μία ακολουθία καταστάσεων, δράσεων, και επιβραβεύσεων (s, a, r). Με βάση την ακολουθία αυτή, εφόσον είναι γνωστές οι επιβραβεύσεις που αποκτήθηκαν σε κάθε κατάσταση, η συνάρτηση αξίας σε μία κατάσταση μπορεί να εκτιμηθεί ως μέσος όρος των καινούριων και των παλιών πληροφοριών που έχουμε για αυτήν. Με δεδομένη λοιπόν μία πρότερη εκτίμηση της και μία νέα ακολουθία (s, a, r), η συνάρτηση αξίας ανανεώνεται με τον κανόνα :

$$V(S_t) \leftarrow V(S_t) + \frac{1}{N(S_t)}(G_t - V(S_t))$$

,όπου $N(S_t)$ ο αριθμός επισκέψεων στην κατάσταση S_t.

Σε μη στατικά προβλήματα είναι πιο χρήσιμο να μην χρησιμοποιούμε τον κανόνα του μέσο όρου αλλά έναν ρυθμό μάθησης, ώστε να μην λαμβάνονται υπόψιν πολύ παλαιές επισκέψεις στις καταστάσεις, εάν το περιβάλλον αλλάξει.

$$V(S_t) \leftarrow V(S_t) + a(G_t - V(S_t))$$

Η μέθοδος Monte Carlo είναι μία πολύ απλή μέθοδος ανανέωσης, η οποία μαθαίνει από ολοκληρωμένα επεισόδια, εφόσον απαιτεί όλη την ακολουθία γεγονότων. Έχει το πλεονέκτημα ότι είναι unbiased εκτίμηση, αλλά και το μειονέκτημα ότι μπορεί να εφαρμοστεί μόνο σε επεισοδιακά MDP's, που έχουν δηλαδή κάποιο φυσικό τέλος.

2.3.2 Μάθηση Temporal Difference (TD(0))

Αντί να περιμένουμε το τέλος του επεισοδίου, όπως υποδεικνύει η μέθοδος Monte Carlo μπορούμε μετά από μία μόνο παρατήρηση να κάνουμε μια εκτίμηση για την συνάρτηση αξίας χρησιμοποιώντας την εκτίμηση για την συνάρτηση αξίας στην επόμενη κατάσταση. Με βάση, λοιπόν, την αναδρομική σχέση $V_\pi(s) = \mathbb{E}R_{t+1} + \gamma V_\pi(S_{t+1})$ μπορούμε να χρησιμοποιήσουμε

εκτίμηση και για την $V_\pi(S_{t+1})$. Έτσι η απόδοση G_t μπορεί να προσεγγιστεί ως $R_{t+1}+\gamma V(S_{t+1})$, ποσότητα η οποία ονομάζεται Temporal Difference Target. Ανανεώνουμε λοιπόν την γνώση που έχουμε για την συνάρτηση αξίας προς αυτή την κατεύθυνση:

$$V(S_t) \leftarrow V(S_t) + \alpha(R_{t+1} + \gamma V(S_{t+1}) - V(S_t))$$

Η ποσότητα ανανέωσης $\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$ ονομάζεται TD error και ο παραπάνω αλγόριθμος πρόβλεψης TD(0). Η λογική που χρησιμοποιεί της ανανέωσης μίας εκτίμησης από κάποια άλλη εκτίμηση (της επόμενης κατάστασης που παρατηρεί) ονομάζεται bootstraping και συναντάται σε πολλούς αλγορίθμους ενισχυτικής μάθησης. Έχει αποδειχθεί ότι με αρκετά μικρό ρυθμό μάθησης και δεδομένη πολιτική π η συνάρτηση αξίας που υπολογίζει συγκλίνει στην V_π. Επιπλέον, ο TD(0) έχει το πλεονέκτημα σε σχέση με τον Monte Carlo ότι δίνει την δυνατότητα online ανανέωσης ακόμα και σε προβλήματα που δεν γνωρίζουμε εάν τελειώνουν, ή ακόμα σε προβλήματα με πολύ μεγάλα επεισόδια δίνει την δυνατότητα ανανέωσης της γνώσης πριν το τέλμα τους, αλλά και το μειονέκτημα ότι εισάγει bias εφόσον χρησιμοποιεί εκτίμηση για την ανανέωση μίας εκτίμησης (bootstraping).

2.3.3 Μάθηση TD(λ)

Από την μία πλευρά η απόδοση G_t που χρησιμοποιεί η μέθοδος Monte Carlo μέσω sampling, όπως και το πραγματικό TD error : $\delta_t^{(real)} = R_{t+1} + \gamma V_\pi(S_{t+1}) - V_\pi(S_t)$ είναι unbiased αφού υπολογίζουν ακριβώς το G_t. Από την άλλη, το TD error δ_t που χρησιμοποιεί ο TD(0) εισάγει bias αφού η ανανέωση της υπάρχουσας προσέγγισης γίνεται βάση προσέγγισης (bootsraping). Ωστόσο, επειδή ο TD(0) βασίζεται σε μόνο ένα βήμα μετάβασης, δράσης και επιβράβευσης χαρακτηρίζεται από μικρότερο variance σε σχέση με τον Monte Carlo.

Οι δύο παραπάνω αλγόριθμοι αποτελούν ακραίες καταστάσεις, με την έννοια ότι ο πρώτος περιμένει μέχρι το τέλος και χρησιμοποιεί όλα τα βήματα για να κάνει μία ανανέωση ενώ ο δεύτερος λαμβάνει υπόψιν μόνο ένα βήμα. Ωστόσο, η πρόβλεψη μπορεί να γίνει με βάση οποιοδήποτε ν-βήμα χρησιμοποιώντας την ν-οστή απόδοση:

$$G_t^{(n)} = R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{n-1} R_{t+n} + \gamma^n V(S_{t+n})$$

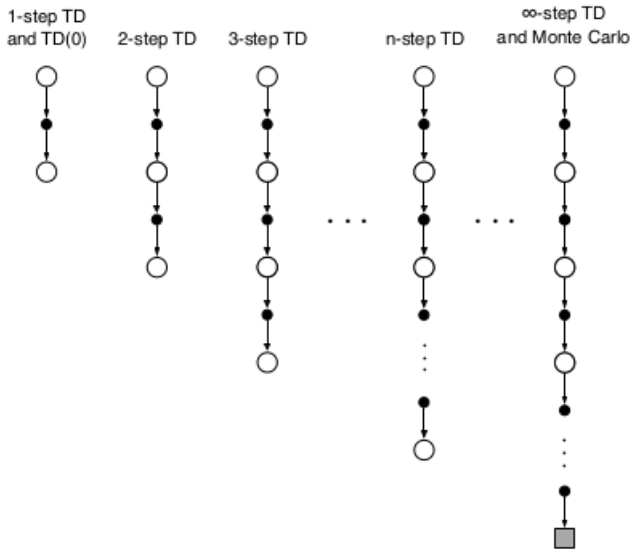
και συνεπώς η συνάρτηση αξίας μπορεί να ανανεωθεί ως:

$$V(S_t) \leftarrow V(S_t) + \alpha(G_t^{(n)} - V(S_t))$$

Παρατηρούμε ότι ο Monte Carlo και ο TD(0) χρησιμοποιούν τις αποδόσεις $G_t^{(\infty)}$ και $G_t^{(1)}$ αντίστοιχα, δηλαδή τις ακραίες περιπτώσεις. Μπορούμε να αναπτύξουμε περαιτέρω αυτή την ιδέα συνδυάζοντας όλες τις ν-οστές αποδόσεις, χρησιμοποιώντας έναν παράγοντα $\lambda \in [0, 1]$:

$$G_t^{\lambda} = (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} G_t^{(n)}$$

Ο αλγόριθμος αυτός ονομάζεται εμπρόσθιος (Forward) TD(λ), επειδή υπολογίζει την απόδοση βάσει των επόμενων καταστάσεων. Ωστόσο, υπάρχει, όπως και στον Monte Carlo,



Σχήμα 2.3: *Διάγραμμα Backup μεθόδων που χρησιμοποιούν ν-οστά βήματα (από [1])*

το πρόβλημα της αναμονής των επιβραβεύσεων επόμενων καταστάσεων, το οποίο θα θέλαμε να απαλείψουμε. Την λύση δίνει ο οπίσθιος (Backward) TD(λ),ο οποίος χρησιμοποιεί ίχνη δικαιοδοσίας (eligibility traces). Σκοπός τους είναι να συμπεριλάβουν στην ανανέωση την συχνότητα εμφάνισης και το πόσο πρόσφατα έχει εμφανιστεί η κάθε κατάσταση (frequency and recency heuristics). Κάθε κατάσταση λοιπόν, έχει ένα ίχνος $E_t(s)$ αρχικοποιημένο στο 0, το οποίο ανανεώνεται σύμφωνα με τον κανόνα:

$$E_t(s) = \gamma\lambda E_{t-1}(s) + \mathbb{1}(S_t = s)$$

Η ανανέωση της συνάρτησης αξίας κάθε κατάστασης γίνεται ανάλογα με το ίχνος δικαιοδοσίας της και ανάλογα με το TD error δ_t:

$$V(s) \leftarrow V(s) + \alpha\delta_t E_t(s)$$

Παρατηρούμε ότι για $\lambda = 0$ στην ανανέωση λαμβάνεται υπόψιν μόνο η τρέχουσα κατάσταση $V(s) \leftarrow V(S_t) + \alpha\delta_t$, η οποία είναι πανομοιότυπη με την ανανέωση του TD(0), ενώ για $\lambda = 1$, η συνολική ανανέωση που έχει γίνει μετά το τέλος ενός επεισοδίου είναι ίδια με την ανανέωση του Monte Carlo.

2.4 Έλεγχος σε MDPs άγνωστου μοντέλου

Έχοντας λύσει το πρόβλημα της πρόβλεψης της συνάρτηση αξίας για μία δεδομένη πολιτική π, προχωράμε στην διαδικασία ανανέωσης της με στόχο την εύρεση βέλτιστης πολιτικής από τον πράκτορα. Η λογική εναλλαγής των βημάτων πρόβλεψης και ελέγχου είναι ίδια με την γενικευμένη επανάληψη πολιτικής που περιγράφηκε για τον δυναμικό προγραμματισμό (Generalized Policy Iteration).

2.4.1 Έλεγχος πάνω στην πολιτική (on-policy control)

Στην λογική on-policy ανανεώνουμε την πολιτική π βάσει της γνώσης που αποκτήθηκε από την ίδια πολιτική π. Στον δυναμικό προγραμματισμό αυτό γινόταν με άπληστο greedy τρόπο, δηλαδή $\pi = greedy(V)$, επιλέγοντας την δράση που μας οδηγεί στην μέγιστη συνάρτηση αξίας.

Ας υποθέσουμε ότι χρησιμοποιούμε Monte Carlo πρόβλεψη. Η ανανέωση της πολιτικής δεν μπορεί να γίνει, όπως στον δυναμικό προγραμματισμό, με χρήσης της συνάρτησης αξίας V δηλαδή ως:

$$\pi'(s) = \arg\max_a \sum_{s' \in S} P^a_{ss'} V(s')$$

,διότι αυτό απαιτεί γνώση του μοντέλου $P^a_{ss'}$. Για την ανανέωση, επομένως, χρησιμοποιούμε την συνάρτηση αξίας-δράσης και με βάση τον δυναμικό προγραμματισμό θα ήταν:

$$\pi'(s) = \arg\max_a Q(s, a)$$

Ο παραπάνω άπληστος τρόπος ανανέωσης παρουσιάζει, ωστόσο, ένα πρόβλημα όταν εργαζόμαστε χωρίς γνώση του μοντέλου. Με άγνωστες τις μεταβάσεις, η πολιτική με αυτόν τον τρόπο ανανέωσης δεν θα συγκλίνει στην βέλτιστη. Ειδικότερα, υπάρχει η πιθανότητα να μην επιλεχθεί ποτέ η βέλτιστη δράση, διότι επιλέγεται συνεχώς μία υποβέλτιστη δράση, η οποία στην παρούσα φάση του αλγορίθμου να έχει υψηλότερη εκτίμηση για την συνάρτηση αξίας της. Συνεπώς, είναι αναγκαίο να εξερευνηθεί ο χώρος των καταστάσεων και δράσεων για να υπάρχει σύγκλιση στην π_*. Ένας τρόπος εξερεύνησης είναι, αντί να επιλέγουμε κάθε φορά την ενέργεια από την παρούσα βέλτιστη πολιτική (exploitation) να εισάγουμε έναν παράγοντα-πιθανότητα ϵ, κατα την οποία επιλέγεται τυχαία κάποια άλλη δράση (exploration). Ο τρόπος αυτός εξερεύνησης ονομάζεται ε-άπληστη εξερεύνηση (ϵ-greedy exploration).

Σύμφωνα, με τον τρόπο αυτό στον Monte Carlo έλεγχο πολιτικής, ο πράκτορας ακολουθώντας μία πολιτική π συλλέγει ένα επεισόδιο $S_1, A_1, R_2, ..., S_T$ με βάση ένα σύνολο m δυνατών δράσεων και προβαίνει σε διαδοχικά βήματα ανανέωσης της συνάρτησης αξίας-δράσης Monte Carlo Policy Evaluation και ανανέωσης της πολιτικής ϵ-greedy Policy Improvement:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{1}{N(S_t, A_t)}(G_t - Q(S_t, A_t))$$

$$\pi(a|s) = \begin{cases} \frac{\epsilon}{m} + 1 - \epsilon & \text{if } a^* = \arg\max_a Q(s, a), \\ \frac{\epsilon}{m} & \text{if } otherwise. \end{cases}$$

Όσον αφορά την σύγκλιση στην βέλτιστη πολιτική, αποδεικνύεται ότι ο Monte Carlo Policy Control συγκλίνει εφόσον είναι άπληστη στο όριο με άπειρη εξερεύνηση (Greedy in the Limit with Infinite Exploration - GLIE), το οποίο συνεπάγεται επίσκεψη στις διάδες κατάστασης-δράσης άπειρες φορές: $\lim_{k\to\infty} N_k(s, a) = \infty$ και σύγκλιση σε άπληστη πολιτική: $\lim_{k\to\infty} \pi_k(a|s) = \mathbb{1}(a = \arg\max_{a' \in A} Q(s, a'))$.

Ομοίως, μπορούμε να χρησιμοποιήσουμε Temporal Difference Control, με χρήση Temporal Difference Evaluation για την πρόβλεψη της συνάρτησης αξίας-δράσης $Q(s, a)$ και

στην συνέχεια ϵ-greedy policy improvement. Ο κανόνας ανανέωσης του βήματος πρόβλεψης, σύμφωνα με την μάθηση TD(0) θα είναι:

$$Q(S,A) \leftarrow Q(S,A) + \alpha(R + \gamma Q(S',A') - Q(S,A))$$

, όπου A η δράση που επιλέγεται στην κατάσταση S βάσει της πολιτικής π και δίνει επιβράβευση R, και A' η δράση που θα έπαιρνε η πολιτική π στην S'. Για την βελτίωση της πολιτικής (έλεγχος) ακολουθείται βήμα ϵ-greedy policy improvement. Ο αλγόριθμος αυτός ονομάζεται SARSA, λόγω των μεταβλητών που εμπεριέχει και συγκλίνει στην βέλτιστη $q_*(s,a)$ υπό τις προϋποθέσεις ότι έχουμε GLIE ακολουθία πολιτικών $\pi_t(a|s)$ και τηρούνται οι κανόνες Robbins-Monro της ακολουθίας του ρυθμού μάθησης α_t :

$$\sum_{t=1}^{\infty} \alpha_t = \infty$$

και

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

Συνεχίζοντας, με την ίδια λογική της TD(λ) πρόβλεψης η ανανέωση της συνάρτηση αξίας-δράσης μπορεί να γίνει σύμφωνα με το ν-όστό βήμα, οδηγώντας στον αλγόριθμο n-step SARSA:

$$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha(q_t^{(n)} - Q(S_t,A_t))$$

,όπου $q_t^{(n)} = R_{t+1} + \gamma R_{t+2} + .. + \gamma^{n-1} R_{t+n} + \gamma^n Q(S_{t+n})$

Όλες οι παραπάνω Q-αποδόσεις των ν-οστών βημάτων μπορούν να συνδυαστούν με χρήση του παράγοντα βάρους $(1-\lambda)\lambda^{n-1}$ στην απόδοση $q_t^{\lambda} = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} q_t^{(n)}$ και επομένως η ανανέωση να γίνει:

$$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha(q_t^{\lambda} - Q(S_t,A_t))$$

,ο οποίος είναι ο πρόσθιος SARSA(λ). Ανάλογα μπορούμε με χρήση ιχνών δικαιοδοσίας να λάβουμε τον οπίσθιο SARSA(λ). Αρχικοποιώντας ίχνη $E_0(s,a) = 0$ και κάθε φορά ανανεώνοντας:

$$E_t(s,a) = \gamma\lambda E_{t-1}(s,a) + \mathbb{1}(S_t = s, A_t = a)$$

Σύμφωνα με αυτά ανανεώνουμε την συνάρτηση αξίας-δράσης αναλογικά με τα ίχνη αυτά και το TD-σφάλμα δ_t:

$$Q(s,a) \leftarrow Q(s,a) + \alpha\delta_t E_t(s,a)$$

2.4.2 Έλεγχος εκτός πολιτικής (off-policy control)

Στην μάθηση εκτός πολιτικής αξιολογούμε μία πολιτική $\pi(a|s)$ μέσω υπολογισμούς της $v_\pi(s)$ ή της $q_\pi(s,a)$, ακολουθώντας όμως μία διαφορετική πολιτική $\mu(a|s)$. Αυτό μπορεί να είναι χρήσιμο, διότι με αυτόν τον τρόπο ο πράκτορας μπορεί να μάθει παρατηρώντας

ανθρώπους ή άλλους πράκτορες, μπορεί να χρησιμοποιήσει γνώση από παλαιές πολιτικές, μπορεί να μάθει για την βέλτιστη πολιτική ακολουθώντας κάποια εξερευνητική πολιτική και γενικά μπορεί να μάθει για πολλές πολιτικές ακολουθώντας μόνο μία.

Η λογική με την οποία επιτυγχάνεται αυτό, είναι η εκτίμηση μίας διαφορετικής κατανομής από την παρούσα, το οποίο ονομάζεται Importance Sampling:

$$\mathbb{E}_{X\sim P}[f(X)] = \sum P(X)f(X) = \sum Q(X)\frac{P(X)}{Q(X)}f(X) = \mathbb{E}_{X\sim Q}[\frac{P(X)}{Q(X)}f(X)]$$

Με τον τρόπο αυτό μπορούμε χρησιμοποιώντας επεισόδια που παράγει μία πολιτική συμπεριφοράς $\mu(a|s)$ και αποδόσεις G_t που υπολογίζονται από τις επιβραβεύσεις αυτών των επεισοδίων, να αξιολογήσουμε μία διαφορετική πολιτική στόχο $\pi(a|s)$ με συνάρτηση αξίας $V(s)$ χρησιμοποιώντας την ακόλουθη απόδοση με Importance Sampling:

$$G_t^{\pi/\mu} = \frac{\pi(A_t|S_t)\pi(A_{t+1}|S_{t+1})\dots\pi(A_T|S_T)}{\mu(A_t|S_t)\mu(A_{t+1}|S_{t+1})\dots\mu(A_T|S_T)}G_t$$

Με βάση αυτή την απόδοση η ανανέωση μπορεί να γίνει για την πολιτική στόχο $\pi(a|s)$, δηλαδή:

$$V(S_t) \leftarrow V(S_t) + \alpha(G_t^{\pi/\mu} - V(S_t))$$

Ωστόσο, αυτή η μέθοδος απαιτεί μη μηδενικές τιμές της μ, όπου η π είναι μη μηδενική και επιπλέον έχει πολύ μεγάλο variance, το οποίο οδηγεί στην μη χρησιμοποίηση στην πράξη.

Για να μειώσουμε το variance, μπορούμε να χρησιμοποιήσουμε TD Importance Sampling αντί για Monte Carlo Importance Sampling, δηλαδή ανανεώσεις που στηρίζονται μόνο στο επόμενο βήμα:

$$V(S_t) \leftarrow V(S_t) + \alpha(\frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}(R_{t+1} + \gamma V(S_{t+1})) - V(S_t))$$

Διαφορετικά, μπορούμε, χωρίς Importance Sampling, να επιλέγουμε δράσεις $A_t \sim \mu(a|s)$ σύμφωνα με την πολιτική μ οι οποίες οδηγούν σε επιβράβευση R_{t+1} και επόμενη κατάσταση S_{t+1}. Για να γίνει πρόβλεψη για την πολιτική π, μπορούμε να χρησιμοποιήσουμε την ανανέωση του SARSA, αλλά με στόχο μάθησης που αφορά την π. Για το λόγο αυτό, αντί να χρησιμοποιήσουμε ως στόχο την $Q(S_{t+1}, A_{t+1})$ που αφορά την πολιτική μ, χρησιμοποιούμε στον στόχο την δράση A' που θα λάμβανε υποθετικά ο πράκτορας στην κατάσταση S_{t+1} αν ακολουθούσε την πολιτική π. Έτσι, η συνάρτηση αξίας-δράσης της π μπορεί να ανανεωθεί με τον κανόνα:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma Q(S_{t+1}, A') - Q(S_t, A_t))$$

Η πρόβλεψη αυτή ονομάζεται πρόβλεψη Q-learning. Για το συνολικό πρόβλημα πρόβλεψης-ελέγχου, επιτρέπουμε στις δύο πολιτικές π (πολιτική-στόχος) και μ (πολιτική-συμπεριφοράς) να ανανεωθούν, η πρώτη δρώντας greedy $\pi(a|s) = \arg\max_{a'} Q(S_t, a')$ και η δεύτερη ϵ-greedy. Τότε ο Q-learning στόχος μπορεί να απλοποιηθεί ως

$$R_{t+1} + \gamma Q(S_{t+1}, A') = R_{t+1} + \gamma Q(S_{t+1}, \arg\max_{a'} Q(S_{t+1}, a')) = R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a')$$

Βάσει αυτού ο κανόνας ανανέωσης του Q-learning control ή SARSAMAX, ο οποίος αποδεικνύεται ότι συγκλίνει στην βέλτιστη πολιτική μπορεί να γραφεί ως:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

2.5 Προσέγγιση της Συνάρτησης Αξίας

Πολλά προβλήματα που αντιμετωπίζονται με χρήση ενισχυτικής μάθησης έχουν τεράστια διάσταση καταστάσεων (π.χ. εικόνες), ενώ σε άλλα ο χώρος καταστάσεων ή και δράσεων είναι συνεχής. Μέχρι τώρα, οι αλγόριθμοι που παρουσιάστηκαν, έχουν διακριτούς χώρους και συνεπώς η συνάρτηση αξίας αναπαριστάται ως ένας πίνακας v και η συνάρτηση δράσης-αξίας ως ένας πίνακας q (tabular methods). Η εξερεύνηση ολόκληρου του χώρου καταστάσεων για την εύρεση βέλτιστης πολιτικής όταν έχουμε μεγάλες διαστάσεις είναι πρακτικά αδύνατη, καθώς ο χρόνος που χρειάζεται για να βρούμε την συνάρτηση αξίας για κάθε κατάσταση είναι πάρα πολύ μεγάλος και ο χώρος που χρειαζόμαστε για την αποθήκευση τέτοιων πινάκων μπορεί να είναι πάρα πολύ μεγάλος Για συνεχείς χώρους προφανώς η χρήση πινάκων είναι αδύνατη. Προς αντιμετώπιση των προβλημάτων αυτών, οι συναρτήσεις αξίας προσεγγίζονται με χρήση παραμέτρων $w \in \mathbb{R}^d$, ώστε: $\hat{v}(s, w) \approx v_\pi(s)$ ή $\hat{q}(s, a, w) \approx q_\pi(s, a)$. Οι προσεγγιστές συνάρτησης (function approximators) που μπορούν να χρησιμοποιηθούν είναι διαφόρων ειδών, όπως γραμμικός συνδυασμών των χαρακτηριστικών, νευρωνικά δίκτυα, δέντρα αποφάσεων κλπ. Ιδιαίτερα τις τελευταίες δεκαετίες με την ανάπτυξη υλικού υπολογιστικής δύναμης για την υποστήριξη βαθιών νευρωνικών δικτύων, αυτά αποτελούν την πιο δημοφιλή και αποδοτική μορφή προσεγγιστών, που οδηγεί στην βαθιά ενισχυτική μάθηση.

Χρειαζόμαστε λοιπόν μεθόδους μάθησης για τις παραμέτρους w του προσεγγιστή. Η πιο διαδεδομένη μέθοδος, η οποία χρησιμοποιείται ευρέως στην επιβλεπόμενη μάθηση και στα νευρωνικά δίκτυα, είναι η Gradient Descent. Ορίζουμε μία παραγωγίσιμη συνάρτηση των βαρών $J(w)$, έστω το ελάχιστο τετραγωνικό σφάλμα μεταξύ εκτίμησης του προσεγγιστή και πραγματική τιμή της συνάρτησης αξίας $J(w) = \mathbb{E}_\pi[(v_\pi(S) - \hat{v}(S, w))^2]$ και με χρήση της κλίσης $\nabla_{\mathbf{w}} J(w)$ βρίσκουμε ένα τοπικό ελάχιστο της $J(w)$ με ανανεώσεις βαρών $\Delta w = \alpha(v_\pi(S) - \hat{v}(S, w))\nabla_w \hat{v}(S, w)$.

Η πρόβλεψη στις προσεγγίσεις γίνεται βάσει των μεθόδων που αναφέραμε στην model-free πρόβλεψη (Παράγραφος 2.3): Monte Carlo, TD(0), TD(λ), αντικαθιστώντας την $v_\pi(S)$ με τον αντίστοιχο στόχο (target) του κάθε αλγορίθμου.

Monte Carlo

$\Delta w = \alpha(G_t - \hat{v}(S_t, w))\nabla_w \hat{v}(S_t, w)$

TD(0)

$\Delta w = \alpha(R_{t+1} + \gamma \hat{v}(S_{t+1}, w) - \hat{v}(S_t, w))\nabla_w \hat{v}(S_t, w)$

TD(λ) Forward

$\Delta w = \alpha(G_t^\lambda - \hat{v}(S_t, w))\nabla_w \hat{v}(S_t, w)$

TD(λ) Backward

$\delta_t = R_{t+1} + \gamma\hat{v}(S_{t+1}, w) - \hat{v}(S_t, w)$

$E_t = \gamma\lambda E_{t-1} + \nabla_{\mathbf{w}}\hat{v}(S_t, w)$

$\Delta w = \alpha\delta_t E_t$

Με τον τρόπο αυτό μπορούμε να προσεγγίσουμε την συνάρτηση αξίας (πρόβλεψη). Για να έχουμε όμως ανανέωση της πολιτικής χρειαζόμαστε, όπως έχουμε περιγράψει, και βήμα ελέγχου. Σύμφωνα με τα προηγούμενα πρέπει πρώτα να κάνουμε πρόβλεψη για την συνάρτηση αξίας-δράσης, όπου εδώ χρησιμοποιούμε τον προσεγγιστή $\hat{q}(s, a, w)$. Οι ανανεώσεις των βαρών γίνεται όπως και στην συνάρτηση αξίας ανάλογα με τον αλγόριθμο. Έχουμε δηλαδή:

Monte Carlo

$\Delta w = \alpha(G_t - \hat{q}(S_t, A_t, w))\nabla_w\hat{q}(S_t, A_t, w)$

TD(0)

$\Delta w = \alpha(R_{t+1} + \gamma\hat{q}(S_{t+1}, A_{t+1}, w) - \hat{q}(S_t, A_t, \mathbf{w}))\nabla_w\hat{q}(S_t, A_t, w)$

TD(λ) Forward

$\Delta w = \alpha(G_t^{\lambda} - \hat{q}(S_t, A_t, w))\nabla_w\hat{q}(S_t, A_t, w)$

TD(λ) Backward

$\delta_t = R_{t+1} + \gamma\hat{q}(S_{t+1}, A_{t+1}, w) - \hat{q}(S_t, A_t, w)$

$E_t = \gamma\lambda E_{t-1} + \nabla_w\hat{q}(S_t, A_t, w)$

$\Delta w = \alpha\delta_t E_t$

Μετά το βήμα πρόβλεψης ακολουθεί σύμφωνα με την γενικευμένη επανάληψη πολιτικής (Generalized Policy Iteration) το βήμα ελέγχου. Ακολουθώντας την μέθοδο ϵ-greedy policy improvement επιλέγεται ως ακόλουθη η ενέργεια $a = \arg\max_{a'} \hat{q}(s, a', w)$ με πιθανότητα $1 - \epsilon$ και με πιθανότητα ϵ μία τυχαία ενέργεια.

Είναι σκόπιμο να αναφερθεί ότι η σύγκλιση σε κάποιο τοπικό ελάχιστο και συνεπώς η σύγκλιση σε υπο-βέλτιστη λύση, δεν επιτυγχάνεται πάντα στους παραπάνω αλγορίθμους. Συγκεκριμένα για τους παραπάνω on-policy αλγορίθμους που χρησιμοποιούνται γραμμικοί προσεγγιστές υπάρχει σύγκλιση σε όλες τις περιπτώσεις, ενώ για μη γραμμικούς συγκλίνει μόνο ο Monte Carlo, ενώ οι TD(0) και TD(λ) αποκλίνουν. Για τον λόγο αυτό η προσέγγιση συνάρτησης αξίας στην πράξη συνδυάζεται με τις μεθόδους κλίσης πολιτικής που παρουσιάζονται στην συνέχεια.

Παρόμοια, ο έλεγχος μπορεί να πραγματοποιηθεί και με την λογική (off-policy). Περιγράφουμε σύντομα τον αλγόριθμο Deep Q-Network (DQN), ο οποίος είναι μία προσαρμογή Q-Learning μεθόδων με χρήση προσεγγιστών και αποτελεί την βάση όλων των off policy αλγορίθμων. Ο DQN χρησιμοποιεί Replay Buffer, δηλαδή αποθηκεύει τις μεταβάσεις που συναντάει $s_t, a_t, s_{t+1}, r_{t+1}$ σε μία μνήμη και σε κάθε επανάληψη κάνει δειγματοληψία ενός τμήματος της μνήμης (minibatch) και αναβαθμίζει τις παραμέτρους βάσει αυτών. Με αυτόν τον τρόπο οι αναβαθμίσεις πραγματοποιούνται βάσει των δεδομένων που έχει συναντήσει ο πράκτορας καθ' όλη την αλληλεπίδραση του με το περιβάλλον. Οι αναβαθμίσεις γίνονται όπως στην Q-learning μάθηση, όμως χρησιμοποιείται για την προσέγγιση της Q ένα νευρωνικό δίκτυο με παραμέτρους w, ενώ για τον υπολογισμό του Q-learning στόχου χρησιμοποιούνται οι προηγούμενες τιμές του δικτύου w^-. Συνολικά στον DQN πραγματοποιούνται τα εξής βήματα:

- Λαμβάνεται η δράση a_t ως ϵ-greedy βάση της $Q(\cdot, a_t)$.
- Αποθηκεύεται η παρατήρηση $s_t, a_t, s_{t+1}, r_{t+1}$ στην μνήμη (Replay Buffer).
- Δειγματοληπτείται ένα minibatch s, a, s', a' από την μνήμη (Replay Buffer).
- Υπολογίζονται οι Q-learningστόχοι $Q_{target} = r + \gamma \max_{a'} Q(s', a'; w^-)$
- Εκτέλεση βήματος gradient descent για την ανανέωση των παραμέτρων w με συνάρτηση κόστους την διαφορά δικτύου $Q-$στόχου και προσέγγισης της Q:

$$LS(w) = (r + \gamma \max_{a'} Q(s', a'; w^-) - Q(s, a; w))^2$$

- Ανανέωση των παλαιών παραμέτρων $w^- = w$

Ο DQN είναι ένας από τους πιο γνωστούς αλγορίθμους ενισχυτικής μάθησης, κυριώς λόγω της ευστάθειας του με νευρωνικά δίκτυα άλλα και της χρήσης minibatch από μνήμη Replay Buffer για την αναβάθμιση. Η μέθοδος αυτή, στην οποία τα παλαιά δεδομένα που αποκτήθηκαν χρησιμοποιούνται ξανά στην εκπαίδευση ονομάζεται Experience Replay και χρησιμοποιείται σχεδόν σε όλους τους off policy αλγορίθμους βαθιάς ενισχυτικής μάθησης. Οι off policy μέθοδοι είναι μεν αποδοτικοί καθώς χρησιμοποιούν και δεδομένα από παλαιές πολιτικές στις ανανεώσεις του, αλλά παρουσιάζουν ωστόσο σημαντικές δυσκολίες συγκλίσεις σε προβλήματα με αυξημένες διαστάσεις των χώρων κατάστασεις και δράσεων.

2.6 Κλίση πολιτικής - Policy Gradient

Μέχρι τώρα η πολιτική που επιλέγει ο πράκτορας καθοριζόταν από την συνάρτηση αξίας (π.χ. ϵ-greedy). Μπορούμε, ωστόσο, αντί να χρησιμοποιήσουμε συνάρτηση αξίας, να παραμετροποιήσουμε κατευθείαν την πολιτική ως μία συνάρτηση

$$\pi_\theta(s, a) = P[a|s, \theta]$$

. Οι αλγόριθμοι που χρησιμοποιούν αυτή την προσέγγιση ονομάζονται Policy Gradient αλγόριθμοι. Τα πλεονεκτήματα αυτών των βάσει πολιτικής αλγορίθμων (policy-based) σε σχέση με αυτών της προσέγγισης με συνάρτηση αξίας (value-based) είναι οι καλύτερες ιδιότητες σύγκλισης, η αποτελεσματικότητα τους σε συνεχής ή χώρους μεγάλης διάστασης και η δυνατότητα εκμάθησης στοχαστικών πολιτικών, ενώ τα μειονεκτήματα είναι η σύγκλιση συνήθως σε τοπικό και όχι ολικό ελάχιστο και το μεγάλο variance στην αξιολόγηση τους.

Για να αξιολογήσουμε την ποιότητα μίας πολιτικής $\pi_\theta(s, a)$ χρησιμοποιούμε μία αντικειμενική συνάρτηση $J(\theta)$. Σε επεισοδιακά περιβάλλοντα μπορούμε να χρησιμοποιήσουμε την συνάρτηση αξίας στην αρχική κατάσταση δηλαδή $J_1(\theta) = V_{\pi^\theta}(s_1) = \mathbb{E}_{\pi_\theta}[v_1]$. Σε συνεχή περιβάλλοντα μπορούμε να χρησιμοποιήσουμε είτε την μέση τιμή αξίας $J_{avV}(\theta) = \sum_s d^{\pi_\theta}(s) V^{\pi_\theta}(s)$ ή την μέση επιβράβευση ανα βήμα $J_{avR}(\theta) = \sum_s d^{\pi_\theta}(s) \sum_a \pi_\theta(s, a) R_s^a$, όπου $d^{\pi_\theta}(s)$ είναι μία στατική κατανομή των καταστάσεων της αλυσίδας Markov για το π_θ.

Το πρόβλημα τώρα έχει γίνει ένα πρόβλημα βελτιστοποίησης, όπου αναζητείται το θ που μεγιστοποιεί την $J(\theta)$. Αυτό μπορεί να επιλυθεί με διάφορες μεθόδους, η δημοφιλέστερη των

οποίων, η οποία χρησιμοποιείται στους policy-based αλγορίθμους είναι η Gradient ascent, κατά την οποία, οι παράμετροι ∂ ανανεώνονται κατά βήμα $\Delta\partial = a\nabla_\partial J(\partial)$, όπου a ένας ρυθμός μάθησης.

Υπολογίζοντας αναλυτικά την κλίση και θεωρώντας πολιτική π_∂ διαφορίσιμη στις μη μηδενικές τιμές και γνωστή κλίση $\nabla_\partial \pi_\partial(s,a)$ μπορούμε να γράψουμε

$$\nabla_\partial \pi_\partial(s,a) = \pi_\partial(s,a)\frac{\nabla_\partial \pi_\partial(s,a)}{\pi_\partial(s,a)} = \pi_\partial(s,a)\nabla_\partial \log \pi_\partial(s,a)$$

Η συνάρτηση $\nabla_\partial \log \pi_\partial(s,a)$ ονομάζεται συνάρτηση επιτυχίας (score function).

Θεώρημα κλίσης πολιτικής Policy gradient Theorem

Για κάθε διαφορίσιμη πολιτική $\pi_\partial(s,a)$ και κάθε μίας από τις αντικειμενικές συναρτήσεις $J = J_1, J_{avR}\frac{1}{1-\gamma}J_{avV}$ η policy gradient είναι:

$$\nabla_\partial J(\partial) = \mathbb{E}_{\pi_\partial}[\nabla_\partial \log_{\pi_\partial(s,a)} Q^{\pi_\partial}(s,a)]$$

Σύμφωνα με το θεώρημα κλίσης πολιτικής και χρησιμοποιώντας σύμφωνα την Monte Carlo unbiased εκτίμηση G_t (απόδοση) για την Q^{π_∂} καταλήγουμε στην ανανέωση

$$\Delta\partial_t = a\nabla_\partial \log \pi_\partial(s_t,a_t)G_t$$

Ο αλγόριθμος αυτός ονομάζεται REINFORCE ή Vanilla Policy Gradient [30].

Αλγοριθμος 2.1: *REINFORCE*

```
Initialize ∂ arbitarily
for each episode s_1, a_1, r_2, ..., s_{T-1}, a_{T-1}, r_T do
    for t=1 to T-1 do
        ∂ ← ∂ + a∇_∂ log π_∂(s_t, a_t) G_t
    end for
end for
```

Για διακριτούς χώρους δράσεων χρησιμοποιείται συχνά ως πολιτική η Softmax Policy. Αν $y_a(s)$ η έξοδος του προσεγγιστή, για παράδειγμα ενός νευρωνικού δικτύου, και τ παράμετρος που καθορίζει την εξερεύνηση:

$$\pi_\partial(a|s) = \frac{e^{y_a(s)/\tau}}{\sum_{a'} e^{y_{a'}(s)/\tau}}$$

Η πιο χρήσιμη πολιτική για συνεχή περιβάλλοντα που χρησιμοποιείται ευρέως είναι η γκαουσιανή πολιτική. Η γκαουσιανή πολιτική δίνει δράσεις $a \sim N(\mu(s), \Sigma)$, όπου Σ μπορεί να είναι σταθερός πίνακας παραμέτρων ή επίσης παράμετροι που μαθαίνονται, και χρησιμοποιείται για εξερεύνηση. Η μέση τιμή είναι η έξοδος του προσεγγιστή, για παράδειγμα η έξοδος ενός νευρωνικού δικτύου.

2.7 Μέθοδοι Δράστη-Κριτή (Actor-Critic)

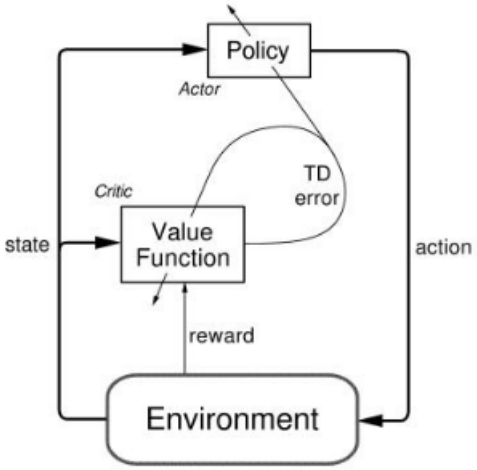
Το πρόβλημα με την μέθοδο Monte Carlo policy gradient είναι η μεγάλη διασπορά (variance) που εισάγεται μέσω της εκτίμησης που χρησιμοποιεί G_t. Για την λύση του προβλήματος αυτού, χρησιμοποιείται συνδυασμός των μεθόδων Policy Gradient και Value Function Approximation, χρησιμοποιώντας, εκτός από τον ¨δράστη¨ (παραμετροποιημένη πολιτική), άλλο ένα σετ παραμέτρων για τον ¨κριτή¨ (παραμετροποιημένη συνάρτηση αξίας-δράσης) $Q_w(s,a) \approx Q^{\pi_\theta}(s,a)$. Επομένως οι μέθοδοι δράστη-κριτή ακολουθούν μία προσεγγιστική κλίση πολιτικής:

$$\nabla_\theta J(\theta) \approx \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a) Q_w(s,a)]$$

και συνεπώς η ανανέωση των παραμέτρων θ του δράστη γίνεται:

$$\Delta\theta_t = \alpha \nabla_\theta \log \pi_\theta(s_t, a_t) G_t$$

Από την άλλη, η ανανέωση των παραμέτρων του κριτή w είναι ένα πρόβλημα πρόβλεψης με προσέγγιση συνάρτησης αξίας και επομένως χρησιμοποιούνται γνωστές μέθοδοι που περιγράφηκαν στην παράγραφο 2.5 : Monte Carlo, TD(0), TD(λ).



Σχήμα 2.4: *Γενική αρχιτεκτονική σχήματος δράστη-κριτή (actor-critic) (από [1])*

Για να μειωθεί το variance που εισάγεται μέσω του κριτή αφαιρείται από την συνάρτηση δράσης-αξίας μία ποσότητα, που ονομάζεται Baseline $B(s)$. Αυτή δεν έχει επίδραση στην εκτίμηση της κλίσης, δηλαδή $\mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a) B(s)] = 0$, όμως έχει σημαντική επίδραση στην μείωση του variance. Μία φυσική επιλογή του baseline είναι η $B(s) = V^{\pi_\theta}(s)$ και σύμφωνα με αυτή ορίζουμε την πλεονεκτική συνάρτηση (Advantage function) ως $A^{\pi_\theta}(s,a) = Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)$ και έτσι η κλίση πολιτικής γίνεται:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a) A^{\pi^\theta}(s,a)]$$

Επομένως, για την προσέγγιση των συναρτήσεων αξίας της πλεονεκτικής συνάρτησης

έχουμε:

$$A^{\pi_\theta}(s,a) \approx Q_w(s,a) - V_v(s,a)$$

Μία unbiased εκτίμηση της πλεονεκτικής συνάρτησης είναι το πραγματικό TD-σφάλμα δ^{π_θ}, αφού:

$$\mathbb{E}_{\pi_\theta}[\delta^{\pi_\theta}|s,a] = \mathbb{E}_{\pi_\theta}[r + \gamma V^{\pi_\theta}(s')|s,a] - V^{\pi_\theta}(s,a) = Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s) = A^{\pi_\theta}(s,a)$$

Με αυτή την προσέγγιση μπορούμε να γράψουμε την κλίση της αντικειμενικής συνάρτησης και αντίστοιχα την ανανέωση των παραμέτρων της πολιτικής ως:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a)\delta^{\pi_\theta}]$$

$$\Delta\theta = \alpha(r + \gamma V_v(s_{t+1}) - V_v(s_t))\nabla_\theta \log \pi_\theta(s_t,a_t)$$

Παρατηρούμε ότι πρακτικά δεν χρειάζονται 2 σετ παραμέτρων για τον κριτή, αλλά μόνο ένα για την συνάρτηση αξίας. Η ανανέωση αυτή είναι σύμφωνη με την λογική του αλγορίθμου TD(0), επειδή κατά την ανανέωση της πολιτικής λαμβάνεται υπόψιν το επόμενο μόνο βήμα. Κατ' όμοιο τρόπο, η ανανέωση μπορεί να γίνει σύμφωνα με τον Monte-Carlo στόχο δηλαδή την απόδοση G_t :

$$\Delta\theta = \alpha(G_t - V_v(s_t))\nabla_\theta \log \pi_\theta(s_t,a_t)$$

ή μπορεί να γίνει με bootstraping όλων των ν-οστών βημάτων, όπως στον αλγόριθμος TD(λ) για την πρόβλεψη της συνάρτησης αξίας. Ο αλγόριθμος αυτός για την προσέγγιση της πλεονεκτικής συνάρτησης με συνδυασμό όλων των ν-στών βημάτων ονομάζεται Generalized Advantage Estimation ή GAE [31] και είναι ο πλέον χρησιμοποιούμενος. Η εκτίμηση GAE μπορεί να εκφραστεί συναρτήσει του TD-σφάλματος $\delta_t^V = r_t + \gamma V(s_{t+1}) - V(s_t)$ ως:

$$\hat{A}_t^{GAE(\gamma,\lambda)} = (1-\lambda)(\hat{A}_t^{(1)} + \lambda\hat{A}_t^{(2)} + \lambda^2\hat{A}_t^{(3)} + \ldots) = \ldots = \sum_{t=0}^{\infty}(\gamma\lambda)^t\delta_{t+1}^V$$

Η εκτίμηση της συνάρτησης αξίας V με παραμέτρους v είναι ένα πρόβλημα πρόβλεψης και μπορεί να λυθεί με οποιονδήποτε τρόπο από αυτούς που έχουμε αναφέρει, δηλαδή Monte Carlo, TD(0), TD(λ) forward ή backward.

2.8 Natural Policy Gradient, TRPO

Οι αλγόριθμοι που έχουν αναφερθεί έως τώρα με τις μεθόδους κλίσης πολιτικής και δράστη-κριτή χρησιμοποιούν την κανονική κλίση (Vanilla Gradient) $g = \nabla_\theta J(\theta)$. Σε αυτές χρησιμοποιούμε έναν αυθαίρετο ρυθμό μάθησης και ανανεώνουμε την πολιτική κατά αυτό το βήμα προς την κατεύθυνση της κλίσης. Στην ενισχυτική μάθηση, ωστόσο η επιλογή του ρυθμού μάθησης είναι αρκετά δύσκολη και πολύ σημαντική για τον εξής λόγο. Ένα μεγάλο

βήμα ανανέωσης μπορεί να μεταφέρει την πολιτική μακριά από την προηγούμενη και να την κάνει χειρότερη, σε σημείο ίσως χρειάζονται πολλά ακόμα δείγματα για να επανέλθει. Αντίθετα ένα μικρό βήμα θα αλλάξει ελάχιστα την πολιτική και συνεπώς η σύγκλιση θα καθυστερεί. Αυτό οδηγεί την Vanilla Policy Gradient να χρειάζεται δεκάδες εκατομμύρια επαναλήψεις για αρκετά απλά προβλήματα.

Οι μέθοδοι βελτιστοποίησης μπορούν να χωριστούν σε γραμμικής αναζήτησης(line search) και περιοχής εμπιστοσύνης (Trust Region). Οι μέθοδοι Gradient Descent είναι μέθοδοι γραμμικής αναζήτησης γιατί πρώτα βρίσκουμε την κατεύθυνση και μετά ανανεώνουμε με ένα βήμα. Στις μεθόδους Trust Region πρώτα βρίσκουμε το όριο μέσα στο οποίο θέλουμε να γίνει η ανανέωση και στην συνέχεια βρίσκουμε την βέλτιστη ανανέωση. Ο αλγόριθμος Trust Region Policy Optimization [11] χρησιμοποιεί αυτή την λογική και λύνει το πρόβλημα του αυθαίρετου βήματος ανανέωσης της Policy Gradient.

Σύμφωνα με την Policy Gradient θέλουμε να διαφορίσουμε την αντικειμενική συνάρτηση

$$L^{PG}(\theta) = \mathbb{E}_t[\log \pi_\theta(a_t|s_t)A_t]$$

Ο TRPO μετρά την αλλαγή της πολιτικής μέσω της Kullback-Leibner (KL) divergence η οποία δίνει την διαφορά μεταξύ δύο κατανομών δεδομένων:

$$D_{KL}(P, Q) = \mathbb{E}_x[\frac{P(x)}{Q(x)}]$$

Αρχικά, ο TRPO εκφράζει την αντικειμενική συνάρτηση σε σχέση με την παλιά πολιτική $\pi_{\theta_{old}}$ μεImportance Sampling ως

$$L(\theta, \theta_{old}) = \mathbb{E}_t[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}A_t]$$

Ουσιαστικά η βελτιστοποίηση της πολιτικής είναι ίδιο πρόβλημα με την βελτιστοποίηση της πολιτικής ως προς την προηγούμενη. Οι δύο αντικειμενικές συναρτήσεις έχουν ίδιο gradient στο $\theta = \theta_{old}$ (Κανόνας Αλυσίδας).

Επιπλέον η αντικειμενική συνάρτηση επαυξάνεται με έναν όρο ποινής $-CD_{KL}^{max}(\pi_i, \pi)$ και η συνολική αντικειμενική συνάρτηση που βελτιστοποιείται είναι:

$$L^{TRPO}(\theta, \theta_{old}) = \mathbb{E}_t[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}A_t] - CD_{KL}^{max}(\pi_\theta, \pi_{\theta_{old}})$$

Η παράμετρος C μπορεί να υπολογιστεί αναλυτικά, ώστε να εγγυάται η μονοτονική βελτίωση της πολιτικής, δηλαδή η εγγύηση ότι πάντα μεταβαίνουμε σε καλύτερες πολιτικές. Ωστόσο, στην πράξη η C είναι πολύ μεγάλη που οδηγεί σε μεγάλη ποινή και επομένως πολύ μικρούς ρυθμούς μάθησης. Για τον λόγο αυτό μεταχειρίζεται σαν υπερπαράμετρος. Για τον ίδιο λόγο, στην πράξη χρησιμοποιούμε την μέση KL-Divergence αντί της μέγιστης. Επιπλέον, μία πιο χρήσιμη έκφραση της αντικειμενικής συνάρτησης είναι να γραφεί ως υπό συνθήκη πρόβλημα (Lagrangian Duality) ως

$$maximize_\theta L(\theta, \theta_{old}) = maximize_\theta \mathbb{E}_t[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}A_t]$$

subject to

$$\mathbb{E}_t[D_{KL}[\theta, \theta_{old}]] \leq \delta$$

Το παραπάνω πρόβλημα μπορεί να λυθεί αναλυτικά αλλά στην πράξη χρησιμοποιείται συχνά η προσεγγιστική λύση με Natural Policy Gradient [32]. Σε αυτήν χρησιμοποιείται προσέγγιση γραμμική για την αντικειμενική συνάρτηση και τετραγωνική για την KL-Divergence.

$$L(\theta, \theta_{old}) \approx g^T(\theta - \theta_{old})$$

και

$$\overline{D_{KL}}[\theta, \theta_{old}] \approx \frac{1}{2}(\theta - \theta_{old})^T F(\theta - \theta_{old})$$

,όπου g η κλίση της πολιτικής και F ο Hessian της KL divergence δηλαδή η κλίση δευτέρου βαθμού. Ο πίνακας αυτός ονομάζεται Fisher Information Matrix και αποδεικνύεται ότι μπορεί να υπολογιστεί ως:

$$F = \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s, a) \nabla_\theta \log \pi_\theta(s, a)^T]$$

Με την προσέγγιση αυτή το πρόβλημα μετατρέπεται σε:

$$\theta_{k+1} = \arg\max_\theta g^T(\theta - \theta_k)$$

, **subject to**

$$\frac{1}{2}(\theta - \theta_k)^T F(\theta - \theta_k)$$

Η υπό συνθήκη εξίσωση αυτή μπορεί να λυθεί αναλυτικά δίνοντας:

$$\theta_{k+1} = \theta_k + \sqrt{\frac{2\delta}{g^T F^{-1} g}} F^{-1} g$$

Το βήμα ανανέωσης σύμφωνα με αυτή είναι

$$a = \sqrt{\frac{2\delta}{g^T F^{-1} g}}$$

και η ανανέωση γίνεται προς την κατεύθυνση

$$g^{(nat)} = F^{-1} g$$

η οποία ονομάζεται φυσική κλίση πολιτικής (Natural Policy Gradient).

Τέλος, προκειμένου λόγω της προσέγγισης που χρησιμοποιήθηκε, προκειμένου να διασφαλιστεί η συνθήκη της KL-Divergence το βήμα a μειώνεται σύμφωνα με έναν παράγοντα μείωσης $b \in (0, 1)$

$$a = b^j \sqrt{\frac{2\delta}{g^T F^{-1} g}}$$

με $a \in (0, 1)$ και j τον μικρότερο θετικό ακέραιο ώστε να ικανοποιείται η KL-συνθήκη και η μονοτονική βελτίωση:

$$\mathbb{E}[D_{KL}[\theta, \theta_{old}]] \leq \delta \text{ and } L(\theta, \theta_{old}) \geq 0$$

Ο TRPO λοιπόν εγγυάται ανανέωση της πολιτικής προς καλύτερες πολιτικές χρησιμοποιώντας μία κλίση δευτέρου βαθμού. Επιπλέον, η υπερπαράμετρος d που αφορά το όριο της KL συνθήκης είναι επιπλέον πολύ πιο εύκολο να ρυθμιστεί απ' ότι το βήμα a στην Policy Gradient. Αυτό συμβαίνει γιατί, ουσιαστικά μέσω αυτού μεταφράζουμε την αλλαγή στην πολιτική σε αλλαγή των παραμέτρων του μοντέλου. Πρακτικά, η παράμετρος δ κανονικοποιεί το βήμα ως $a = \sqrt{\frac{2\delta}{g^T F^{-1} g}}$ και συνεπώς μπορούμε να χρησιμοποιήσουμε τον αλγόριθμο χωρίς ιδιαίτερο ή καθόλου tuning σε διαφορετικά μοντέλα και προβλήματα.

Κεφάλαιο 3

Μάθηση από Δεδομένα Επίδειξης (Imitation Learning)

Σε πολλά προβλήματα επιζητείται ο πράκτορας να αναπτύξει μία πολιτική μαθαίνοντας από δεδομένα που επιδεικνύει κάποιος δάσκαλος/ειδήμων (expert). Αυτό μπορεί να είναι θεμιτό είτε γιατί επιζητούμε την μίμηση του δασκάλου είτε για να επιταχύνουμε την μάθηση. Στο κεφάλαιο αυτό περιγράφουμε βασικούς αλγορίθμους που έχουν προταθεί για την μίμηση από κάποιον δάσκαλο.

3.1 Ο αλγόριθμος Behavioral Cloning

Στον αλγόριθμο Behavioral Cloning [33], [34] η ενισχυτική μάθηση ανάγεται σε πρόβλημα επιβλεπόμενης μάθησης. Ο expert σύμφωνα με κάποια πολιτική π^* λαμβάνει δράσεις a^* σε καταστάσεις s^*. Συμβολίζουμε ως $d^t_{\pi^*}$ την κατανομή των καταστάσεων που ακολουθεί ο expert στο βήμα t και συνεπώς η μέση κατανομή των καταστάσεων είναι $d^* = P(s, \pi^*) = \frac{1}{T}\sum_{t=1}^{T} d^*_t$. Τότε το επιβλεπόμενο πρόβλημα μάθησης είναι η ελαχιστοποίηση της συνάρτησης κόστους, όσον αφορά τα δεδομένα επίδειξης:

$$\arg\min_{\partial} \mathbb{E}_{(s,a^*)\sim d^*} L_{BC}(a, \pi_\partial(s))$$

Η συνάρτηση κόστους L_{BC} μπορεί να είναι οποιαδήποτε συνάρτηση κόστους ενός προβλήματος παλινδρόμησης. Ουσιαστικά, λοιπόν, ο αλγόριθμος Behavioral Cloning εφαρμόζει επιβλεπόμενη μάθηση με χαρακτηριστικά (features) τις καταστάσεις και ετικέτες (labels) τις δράσεις του δασκάλου. Η μέθοδος αυτή έχει τα πλεονεκτήματα της απλότητας και της αποτελεσματικότητας, ωστόσο έχει κάποια πολύ σημαντικά μειονεκτήματα. Εάν η πολιτική που έχει εκπαιδευτεί λάβει μία κακή απόφαση, τότε μπορεί να οδηγηθούμε σε μία κατάσταση που δεν υπάρχει στο σύνολο δεδομένων του δασκάλου και συνεπώς η πολιτική δεν γνωρίζει κάποιον σωστό τρόπο για να συνεχίσει από εκεί και να επανέλθει στον χώρο συμπεριφοράς του δασκάλου. Ας πάρουμε ως παράδειγμα την μάθηση από δεδομένα επίδειξης σε ένα αυτόνομο αυτοκίνητο, για το οποίο ο δάσκαλος έχει δείξει μία σωστή οδηγική συμπεριφορά σε έναν δρόμο. Μετά την εφαρμογή του Behavior Cloning, όταν η πολιτική που έχει μάθει ο πράκτορας εφαρμοστεί, υπάρχει η πιθανότητα λόγω του ότι δεν υπάρχει 100% επιτυχία στην διαδικασία της επιβλεπόμενης μάθησης, το αυτοκίνητο να ξεφύγει από το κέντρο του δρόμου

στο οποίο οδηγούσε ο δάσκαλος στην επίδειξη. Τότε, καθώς τα δεδομένα δεν καλύπτουν περιοχές στην άκρη του δρόμου ο πράκτορας δεν θα γνωρίζει την σωστή ενέργεια που θα πρέπει να πάρει και το αυτοκίνητο θα τρακάρει με μεγάλη πιθανότητα.

Το λάθος λοιπόν στον Behavior Cloning μπορεί είναι καταστροφικό. Έτσι χρησιμοποιείται αυτούσιος σε περιπτώσεις που γνωρίζουμε με κάποιο τρόπο, συνήθως σε περιβάλλοντα με μικρά επεισόδια, πως το ένα βήμα δεν θα έχει μεγάλη απόκλιση από τον χώρο των δεδομένων του δασκάλου, ή σε περιπτώσεις, στις οποίες ο δάσκαλος καλύπτει το μεγαλύτερο μέρος του χώρου καταστάσεων, συνήθως σε περιβάλλοντα με μικρή διάσταση καταστάσεων, ώστε να μην υπάρχουν άγνωστες περιοχές.

Αντίθετα, ο αλγόριθμος αυτός δεν συνίσταται αυτούσιος, για περιπτώσεις που χρειαζόμαστε κάποιον μακροπρόθεσμο προγραμματισμό (long-term planning) και που επιθυμείται η βελτιστοποίηση κάποιου μακροπρόθεσμου στόχου. Ωστόσο, χρησιμοποιείται συχνά σαν αρχικοποίηση της πολιτικής, η οποία στην συνέχεια εκπαιδεύεται περαιτέρω με ενισχυτική μάθηση. Σε πολλά προβλήματα ρομποτικών εφαρμογών, έχει δειχθεί ότι ενώ τόσο η ενισχυτική μάθηση από το μηδέν όσο και ο Behavior Cloning δίνουν ανικανοποιητικές λύσεις, ο συνδυασμός τους αποτελεί μία πολύ αποτελεσματική μέθοδο μάθησης.

3.2 Μάθηση πολιτικής μέσω διαδραστικού δασκάλου

Μία ιδέα για την επίλυση του προβλήματος του Behavioral Cloning που με κάποιο κακό βήμα μπορεί να οδηγήσει σε αποτυχία του πράκτορα να λύσει το πρόβλημα, είναι η διαδραστική εμπλοκή του δασκάλου στην διαδικασία. Πιο συγκεκριμένα, ο πράκτορας ζητάει από τον δάσκαλο συνεχώς να επέμβει και να δείξει ποια είναι η σωστή συμπεριφορά με νέα δεδομένα επίδειξης. Έχουμε λοιπόν την ακόλουθη επαναληπτική διαδικασία, η οποία σε κάθε επανάληψη συμπεριλαμβάνει τρία βασικά βήματα:

- Συλλογή Δεδομένων από Επίδειξη Δασκάλου
- Επιβλεπόμενη μάθηση
- Άσκηση πολιτικής στο περιβάλλον

Σημειώνουμε ότι ουσιαστικά ο Behavioral Cloning αποτελεί μία μόνο επανάληψη αυτού του βήματος, αφού μαθαίνει μόνο από το πρώτο σύνολο δεδομένων που δίνει ο δάσκαλος, χωρίς περαιτέρω εμπλοκή του.

Ένας τρόπος για να υλοποιηθεί αυτή η ιδέα είναι να αθροίζονται σε κάθε επανάληψη, τα νέα δεδομένα που αποκτούνται με το σύνολο των δεδομένων που έχουν αποκτηθεί στο παρελθόν (Data Aggregation). Σύμφωνα με αυτό, σε κάθε βήμα εκπαίδευσης m πραγματοποιείται η εισαγωγή των νέων δεδομένων επίδειξης d_m στο σύνολο των προηγούμενων: $d_1 \cup ... \cup d_m$. Ένας πολύ γνωστός αλγόριθμος αυτής της κατηγορίας είναι ο DAgger[35] από Data Aggregation.

Αλγοριθμος 3.1: *Dagger*

1: Train π_∂ from expert data $D = \{s_1, a_1, s_T, a_T\}$
2: Rollout π_∂ to get $D_\pi = s_1, .., s_K$
3: Ask expert to label D_π with actions a_t
4: Aggregate Datasets $D \leftarrow D \cup D_\pi$

Διαφορετικά, χρησιμοποιώντας ελαφρώς διαφορετική προσέγγιση, σε κάθε βήμα m η πολιτική π'_m εκπαιδεύεται πάνω στο d_m και στην συνέχεια γίνεται aggregation με τις προηγούμενες πολιτικές ώστε:

$$\pi_m = \beta\pi'_m + (1-\beta)\pi_{m-1}$$

Αντί δηλαδή, το άθροισμα να αφορά τα νέα δεμένα που εισάγονται κάθε φορά, το άθροισμα αφορά την νέα πολιτική που εκπαιδεύτηκε πάνω στα νέα δεδομένα. Αυτή η κατηγορία ονομάζεται μάθησης με διαδραστική επίδειξη ονομάζεται Policy Aggregation και βασικοί εκπρόσωποι είναι οι αλγόριθμοι SEARN, SMILe [36], [37] και άλλοι.

3.3 Αντίστροφη Ενισχυτική Μάθηση (Inverse RL)

Μία διαφορετική προσέγγιση στο πρόβλημα της μίμησης ενός δασκάλου είναι η εφαρμογή κάποιου αλγορίθμου ενισχυτικής μάθησης, όπου η επιβράβευση δεν είναι σταθερή συνάρτηση σχεδιασμένη από εμάς, αλλά μαθαίνεται βάσει των δεδομένων που έχει δώσει ο δάσκαλος. Η κατηγορία αυτή ονομάζεται αντίστροφη ενισχυτική μάθηση.

Στην διαδικασία Markov λήψης αποφάσεων $M =< S, A, P, R, \gamma >$ οι επιβραβεύσεις R είναι άγνωστες και μαθαίνονται από κάποιον δάσκαλο που δίνει ένα σύνολο δεδομένων

$$D = \{\tau_1, ..., \tau_m\} \sim \pi^*$$

Στόχος της αντίστροφη ενισχυτικής μάθησης είναι η εύρεση πολιτικής η οποία μεγιστοποιεί μία συνάρτησης επιβράβευσης r^*, η οποία μαθαίνεται από τα δεδομένα επίδειξης:

$$\pi^* = \arg\max_{\pi\in\Pi} \mathbb{E}_\pi[r^*(s,a)]$$

3.3.1 MaxEnt IRL

Για την επίλυση του παραπάνω προβλήματος έχουν αναπτυχθεί μοντέλα μέγιστης εντροπίας, σύμφωνα με τα οποία ο δάσκαλος θεωρείται ότι δρα στοχαστικά, δειγματοληπτώντας τα trajectories που δίνει από την κατανομή Boltzmann, χρησιμοποιώντας τις επιβραβεύσεις ως συνάρτηση ενέργειας. Συμβολίζοντας με $\tau = \{s_1, a_1, ..., s_t, a_t, .., s_T\}$ την ακολουθία (trajectory), $D : \{\tau_i\} \sim \pi^*$ τα δεδομένα επίδειξης του δασκάλου και $R_\psi(\tau) = \sum_t r_\psi(s_t, a_t)$ την υπο μάθηση επιβράβευση, παίρνουμε την παρακάτω πιθανότητα για τα trajectories υπό τον δάσκαλο.

$$p(\tau) = \frac{1}{Z} e^{R_\psi(\tau)}$$

Με την μοντελοποίηση αυτή, ουσιαστικά θεωρείται ότι τα trajectories που έχουν μεγαλύτερη επιβράβευση είναι και αυτά την μεγαλύτερη πιθανότητα να εμφανιστούν και μάλιστα η σχέση αυτή είναι εκθετική. Το πρόβλημα αυτό μπορεί να λυθεί όταν είναι γνωστό το μοντέλου του περιβάλλοντος με δυναμικό προγραμματισμό ή gradient descent. Σύμφωνα με την δεύτερη μέθοδο, το πρόβλημα ανάγεται στην μεγιστοποίηση της log-πιθανοφάνειας:

$$\max_\psi L(\psi) = \max_\psi \sum_{\tau \in D} \log p_{r_\psi}(\tau) = \sum_{\tau in D} \log \frac{1}{Z} e^{(R_\psi(\tau))}$$

, όπου $Z = \int e^{R_\psi(\tau)} d\tau$

Αποδεικνύεται ότι η κλίση της L προκύπτει:

$$\nabla_\psi L = -\frac{1}{|D|} \sum_{\tau_d \in D} \frac{dr_\psi}{d\psi}(\tau_d) - \sum_s p(s|\psi) \frac{dr_\psi}{d\psi}(s)$$

Συνεπώς ανανεώνοντας την παράμετρο ψ προς αυτή την κατεύθυνση με κάποιο ρυθμό μάθησης παίρνουμε την εκτίμηση για την επιβράβευση βάσει των δεδομένων επίδειξης. Ο αλγόριθμος αυτός ονομάζεται MaxEnt IRL[38].

Αλγοριθμος 3.2: *MaxEnt IRL (από [38])*

1: Initialize ψ, gather demonstrations D.
2: **for** ... **do**
3: Solve for optimal policy $\pi(a, s)$ w.r.t. reward r_ψ.
4: Solve for state visitation frequencies $p(s|\psi)$
5: Compute gradient $\nabla_\psi L$ and update ψ
6: **end for**

Ωστόσο, ο αλγόριθμος αυτός έχει τα μειονεκτήματα, ότι χρειάζεται γνώση του μοντέλου και επιπλέον, πως η απαιτούμενη επίλυση για βέλτιστη πολιτική μέσα σε κάθε επανάληψη μειώνει σημαντικά την απόδοση, επιτρέποντας την χρήση του αλγορίθμου μόνο για απλά προβλήματα μικρής διάστασης. Λύση σε αυτά τα προβλήματα δίνουν αλγόριθμοι όπως οι Guided Cost Learning, GAIL κ.α.

3.3.2 Generative Adversarial Imitation Learning (GAIL)

Ένας πολύ δημοφιλής αλγόριθμος στην κατηγορία της αντίστροφης ενισχυτικής μάθησης είναι ο αλγόριθμος Generative Adversarial Imitation Learning (GAIL) [39], ο οποίος έχει άμεση σύνδεση με τα δίκτυα GAN [40] της επιβλεπόμενης μάθησης. Ένα GAN αποτελείται από δύο επιμέρους δίκτυα: το δίκτυο γεννήτορα (Generator) και το δίκτυο του διευκρινιστή (Discriminator). Έχοντας ένα σύνολο δεδομένων επίβλεψης, ο γεννήτορας προσπαθεί, με είσοδο τυχαίο θόρυβο, να παράγει στην έξοδο του δεδομένα που μοιάζουν με τα δεδομένα επίβλεψης. Από την άλλη, ο διευκρινιστής προσπαθεί να διαχωρίσει τα δεδομένα στην έξοδο του γεννήτορα από αυτά του δασκάλου. Μέσω αυτής της ανταγωνιστικής δράσης των δύο

δικτύων, επιτυγχάνεται η παραγωγή-γέννηση νέων δεδομένων στην έξοδο του γεννήτορα που μοιάζουν με αυτά που έχει δώσει ο δάσκαλος.

Την λογική αυτή ακολουθεί ο αλγόριθμος GAIL, στο πλαίσιο της αντίστροφης ενισχυτικής μάθησης. Το αντίστοιχο δίκτυο του γεννήτορα αποτελείται από αυτό της ενισχυτικής μάθησης όπου μαθαίνεται μία πολιτική. Ο διευκρινιστής, από την άλλη, προσπαθεί να ξεχωρίσει τα δεδομένα του δασκάλου από αυτά που δίνει η πολιτική. Ως αποτέλεσμα, όταν επιτυγχάνεται η σύγκλιση, η πολιτική παίρνει αποφάσεις οι οποίες οδηγούν σε αποφάσεις που μοιάζουν με αυτές του δασκάλου. Με άλλα λόγια, η έξοδος του διευκρινιστή που έχει το ρόλο του διαχωρισμού των ¨αληθινών¨ από τα ¨ψεύτικα¨ δεδομένα, αποτελεί ουσιαστικά την επιβράβευση για την πολιτική.

Πιο συγκεκριμένα στον GAIL η πολιτική π μαθαίνει να μιμείται την π^* του expert, μέσω ελαχιστοποίησης της Jensen - Shannon divergence μεταξύ των κατανομών κατάστασης-δράσης που παράγει ο expert και των αντίστοιχων που παράγονται από την πολιτική π

$$\min_{\pi} \max_{D \in (0,1)^{S \times A}} \mathbb{E}_{\pi}[\log D(s,a)] + \mathbb{E}_{\pi^*}[\log(1 - D(s,a))] - \lambda H(\pi)$$

, όπου D ο διευκρινιστής που πραγματοποιεί μια δυαδική ταξινόμηση για να ξεχωρίσει τα δείγματα που προέρχονται από τον expert και αυτά που προέρχονται από την πολιτική, ενώ η $H(\pi) = -\mathbb{E}_{\pi_\theta}[\log \pi_\theta(a|s)]$ ονομάζεται όρος causual entropy regularization και λ υπερπαράμετρος.

Αλγοριθμος 3.3: *Generative Adversarial Imitation Learning (GAIL) (από [39])*

1: Input: Expert trajectories $\tau_E \sim \pi_E$, initial policy and discriminator parameters θ_0, w_0
2: **for** i=0,1,2,... **do**
3: Sample trajectories $\tau_i \sim \pi_{\theta_i}$
4: Update the discriminator parameters from w_i to w_{i+1} with the gradient

$$\mathbb{E}_{\tau_i}[\nabla_w \log(D_w(s,a))] + \mathbb{E}_{\tau_E}[\nabla_W \log(1 - D_w(s,a))]$$

5: Take a policy step from θ_i to θ_{i+1}, using the TRPO rule with cost function $\log(D_{w+1}(s,a))$. Specifically, take a KL-constrained natural gradient step with:

$$\mathbb{E}_{\tau_i}[\nabla_w \log(\pi_\theta(a|s) Q(s,a))] - \lambda \nabla_\theta H(\pi_\theta)$$

where $Q(\bar{s},\bar{a}) = \mathbb{E}_{\tau_i}[\log(D_{w+1}(s,a))]$
6: **end for**

Η κατηγορία των αλγορίθμων αντίστροφης ενισχυτικής μάθησης, έχει το μεγάλο πλεονέκτημα της μίμησης ολόκληρης της συμπεριφοράς του δασκάλου, λύνοντας το αρκετά σύνθετο πρόβλημα του σχεδιασμού της συνάρτησης επιβράβευσης από τον άνθρωπο. Ωστόσο, είναι σκόπιμο να σημειωθεί ότι ο ανταγωνιστικός τρόπος που δρουν τα δίκτυα γεννήτορα και διευκρινιστή σε συνδυασμό με την ενισχυτική μάθηση, η οποία είναι από μόνη της ένα δύσκολο πρόβλημα, περιορίζουν τα προβλήματα που μπορεί να εφαρμοστεί, όσον αφορά την διάσταση του χώρου καταστάσεων και δράσεων.

Μεγάλη ομοιότητα με τον GAIL, παρουσιάζει ο αλγόριθμος Guided Cost Learning [41], ενώ έχουν παρουσιαστεί αρκετές βελτιώσεις του αλγορίθμου που δείχνουν ότι βελτιώνουν την

απόδοση , όπως οι AIRL [42], VAIL [43], EAIRL [44].

Κεφάλαιο 4

Περιγραφή Μεθοδολογίας στο Πλαίσιο της Εφαρμογής

Στο κεφάλαιο αυτό γίνεται ο περιγραφή του συστήματος που χρησιμοποιήθηκε για τον σκοπό της εφαρμογής. Όπως αναφέρθηκε και στην εισαγωγή της εργασίας, σκοπός είναι η μάθηση επιδέξιων κινήσεων ενός αντικειμένου. Συγκεκριμένα, ερευνούμε την λήψη (grasp) ενός αντικειμένου από το ρομπότ σε ένα επιθυμητό ύψος με in hand τρόπο, χρησιμοποιώντας μόνο τα δάχτυλα του χεριού και όχι κάποιον βραχίονα. Επιθυμούμε κατά την ανύψωση να χρησιμοποιήσουμε σαν επαφή ρομπότ-αντικειμένου τις άκρες των δαχτύλων και όχι απλώς να εγκλωβίσουμε το αντικείμενο ανάμεσα στα δάκτυλα, διότι μία τέτοια λαβή προσφέρει πολύ περισσότερες δυνατότητες χειρισμού. Επιπλέον, θέλουμε το σύστημα μας να δίνει την δυνατότητα εκτέλεσης επιθυμητών τροχιών του ύψους του αντικειμένου διατηρώντας τον προσανατολισμό του οριζόντιο, ενώ παράλληλα οι δυνάμεις που ασκούνται στο αντικείμενο να είναι εντός επιθυμητού εύρους. Περιγράφουμε λοιπόν στην συνέχεια την δομή του συστήματος μάθησης που προτείνουμε για τον σκοπό αυτό και τις δυνατότητες τις οποίες προσφέρει.

4.1 Περιγραφή χώρων κατάστασης και δράσεων

Το πρώτο βήμα στην περιγραφή του συστήματος μάθησης είναι η επιλογή των χαρακτηριστικών που χρησιμοποιούνται ως περιγραφή του περιβάλλοντος και των δυνατών ενεργειών που μπορεί να πάρει το ρομπότ, δηλαδή οι χώροι καταστάσεων και δράσεων αντίστοιχα. Αυτά αντιπροσωπεύουν ουσιαστικά την περιγραφή της Μαρκοβιανής διαδικασίας που χρησιμοποιείται ως περιγραφή του προβλήματος. Αντιμετωπίζουμε το πρόβλημα χωρίς γνώση του μοντέλου του συστήματος, δηλαδή της δυναμικής του περιβάλλοντος αλλά και του μοντέλου του ρομπότ. Η γνώση μας δηλαδή για το περιβάλλον προέρχεται από ένα σύνολο αισθητήρων και δεν χρησιμοποιούμε κάποιο μοντέλο για την εξαγωγή περαιτέρω γνώσης για άλλα χαρακτηριστικά του ρομπότ ή του περιβάλλοντος.

Ένα ρομποτικό χέρι, είναι μία αλυσίδα συνδέσμων και αρθρώσεων. Υποθέτουμε ότι οι αρθρώσεις ελέγχονται με κάποιον ελεγκτή θέσης-γωνίας και διαθέτουν επιπλέον αισθητήρες μέτρησης αυτών των γωνιών. Χρησιμοποιούμε, ως χώρο δράσεων για το σύστημα μας τις εντολές στις εισόδους των ελεγκτών θέσης των συνδέσμων του ρομπότ.

Όσον αφορά τον χώρο κατάστασης, αυτός πρέπει να περιγράφει το περιβάλλον. Κύριο συστατικό αυτού, είναι το ίδιο το ρομπότ. Χρησιμοποιούμε λοιπόν στον χώρο κατάστασης

τις γωνίες των συνδέσμων (joints), που προέρχονται από τους αισθητήρες θέσης στα σημεία αυτά. Με τον τρόπο, υπάρχει στον χώρο κατάστασης πληροφορία για την διάταξη του ρομπότ.

Επιπλέον, εφόσον θέλουμε να χειριστούμε κάποιο αντικείμενο, είναι προφανές ότι χρειάζεται να συμπεριληφθεί και πληροφορία για την κατάσταση του αντικειμένου, η διάταξη του οποίου θεωρούμε ότι είναι γνωστή από κάποιον αισθητήρα (π.χ. κάμερα). Συνεπώς, εισάγουμε στην κατάσταση την θέση του κέντρου βάρους του αντικειμένου και τον προσανατολισμό του, χρησιμοποιώντας αναπαράσταση με γωνίες euler (roll,pitch,yaw).

Μέχρι στιγμής, δεν έχουμε πληροφορία για τον τρόπο επαφής και αλληλεπίδρασης ρομπότ - αντικειμένου. Για παράδειγμα, δεν μπορούμε να γνωρίζουμε, εάν τα δάχτυλα διατηρούν επαφή με το αντικείμενο κατά τον χειρισμό του ή εάν απλά το αντικείμενο βρίσκεται ¨κλεισμένο¨ ανάμεσα στα δάχτυλα και στην παλάμη (power grasp). Εφόσον επιθυμούμε να χειριστούμε ένα αντικείμενο χρησιμοποιώντας μόνο τις άκρες των δακτύλων, θεωρούμε απαραίτητα κάποια πληροφορία που να μπορεί να εγγυηθεί την ύπαρξη της επαφής αυτής. Χρησιμοποιούμε για τον σκοπό αυτό αισθητήρες δύναμης (tactiles) στα ακροδάχτυλα του ρομπότ, τα οποία αποτελούν και μεταβλητές του χώρου κατάστασης. Η πληροφορία από τους αισθητήρες δύναμης, χρησιμεύει τόσο ως πληροφορία για την επαφή με το αντικείμενο που θα μας δώσει την επιθυμητή λαβή του αντικειμένου όσο και για την ίδια την τιμή της δύναμης που ασκείται, ώστε να είναι δυνατός ο έλεγχος της.

Επιπλέον, εφόσον επιθυμούμε το ρομπότ να κινεί το αντικείμενο σε μία συγκεκριμένη τροχιά, είναι απαραίτητη η προσθήκη στον χώρο κατάστασης μίας μεταβλητής που να δίνει αυτή την δυνατότητα. Μία επιλογή θα ήταν, να χρησιμοποιήσουμε μία μεταβλητή χρόνου, ώστε κατά την διαδικασία μάθησης να είναι γνωστή η χρονική στιγμή και συνεπώς το αντίστοιχα σημείο της επιθυμητής τροχιάς. Το πρόβλημα, χρησιμοποιώντας τον χρόνο είναι ότι μαθαίνεται μία και μόνο τροχιά. Κατά την διαδικασία δηλαδή της μάθησης, η βελτιστοποίηση γίνεται ώστε την εκάστοτε χρονική στιγμή (η οποία βρίσκεται στον χώρο κατάστασης) να λαμβάνεται ενέργεια από το ρομπότ ώστε το ζητούμενο μέγεθος (εδώ το ύψος) να βρίσκεται στην επιθυμητή θέση της τροχιάς εκείνη την χρονική στιγμή. Αυτό σημαίνει ότι γίνεται μία αντιστοίχιση βέλτιστου σημείου ανά χρονική στιγμή. Εάν δηλαδή θέλουμε να εκτελεστεί μία διαφορετική τροχιά, το σύστημα θα πρέπει να εκπαιδευτεί ξανά στην νέα τροχιά. Αντί αυτού, προτείνουμε την εισαγωγή σημείου της τροχιά στον χώρο κατάστασης. Συγκεκριμένα, χρησιμοποιούμε την επιθυμητή τιμή που θέλουμε να έχει το μέγεθος/μεγέθη που ζητάμε την επόμενη χρονική στιγμή. Έτσι εφόσον σε αυτή την εφαρμογή θέλουμε να ελέγξουμε το ύψος του αντικειμένου, εισάγουμε στον χώρο κατάσταση το ύψος που θέλουμε να έχει το αντικείμενο την επόμενη χρονική στιγμή. Δοσμένης, δηλαδή, μιάς επιθυμητής τροχιάς $h_D(i), 0 <= i <= T-1$, η μεταβλητή του χώρου κατάσταση παίρνει διαδοχικά τις τιμές της τροχιάς: την χρονική στιγμή i θα έχει την τιμή $h_D(i+1)$. Αυτή ουσιαστικά αποτελεί την γνώση του συστήματος ώστε να πάρει τις σωστές αποφάσεις βάσει της τροχιάς που ζητάμε να ακολουθήσει το αντικείμενο.

Με τον τρόπο αυτό δεν γίνεται μία αντιστοιχία βέλτιστης απόφασης ανά χρονική στιγμή, αλλά βέλτιστης απόφασης ανά στόχο στην επόμενη χρονική στιγμή. Συνεπώς, αλλάζοντας τους στόχους αυτούς και επομένως την τροχιά, το σύστημα είναι δυνατόν να εκτελέσει πλήθος τροχιών στις οποίες πιθανώς να μην έχει εκπαιδευτεί, χωρίς να απαιτείται περαιτέρω εκπαίδευση. Εν γένη, η τροχιά θα μπορούσε να είναι πολυδιάστατη και να αφορά οποιοδήποτε

μέγεθος του αντικειμένου, αλλά σε αυτή την εφαρμογή μελετάμε την τροχιά του ύψους του.

Τέλος, αυξάνουμε τον χώρο κατάστασης με μία ακόμα μεταβλητή. Λόγω του ότι χειριζόμαστε ένα αντικείμενο, το ρομπότ μπορεί να μην είναι σε θέση να εκτελέσει την ζητούμενη τροχιά από την αρχή. Στο παράδειγμα ενός grasp πρέπει πρώτα το ρομπότ να πιάσει το αντικείμενο και μετά να εκτελέσει την ζητούμενη τροχιά. Προσθέτουμε συνεπώς μία μεταβλητή φάσης με συνεχείς τιμές στο [0,1] για το διάστημα πριν την εκτέλεση της τροχιάς. Αυτή ουσιαστικά δηλώνει πόσο κοντά βρισκόμαστε στην έναρξη της επιθυμητής τροχιάς. Η σημασία της μεταβλητής αυτής θα κατανοηθεί περισσότερο με τα πειραματικά αποτελέσματα.

Συνολικά λοιπόν δομούμε τον χώρο κατάστασης με τις εξής μεταβλητές:

- Τις αρθρώσεις - joints, που περιγράφουν την κατάσταση του ρομπότ
- Την θέση του κέντρου βάρους του αντικειμένου
- Τον προσανατολισμό του αντικειμένου
- Τις τιμές των δυνάμεων που ασκούνται στα ακροδάχτυλα και επομένως στο αντικείμενο
- Την επιθυμητή τιμή που θέλουμε να έχει το μέγεθος το οποίο μελετάμε την επόμενη χρονική στιγμή (εδώ το ύψος)
- Μία μεταβλητή φάσης για το διάστημα πριν από την εκτέλεση της τροχιάς

Οι τιμές αυτές αναπαριστώνται κάθε χρονική στιγμή t σε ένα διάνυσμα κατάστασης s_t, το οποίο παίρνει συνεχείς τιμές. Αντίστοιχα οι δράσεις του ρομπότ, παίρνουν επίσης συνεχείς τιμές και αναπαριστώνται σε ένα διάνυσμα δράσης a_t.

4.2 Συνάρτηση επιβράβευσης

Το επόμενο βήμα στην περιγραφή της μεθοδολογίας που χρησιμοποιούμε είναι η σχεδίαση του σήματος επιβράβευσης που χρησιμοποιούμε κατά την εκπαίδευση. Η επιβράβευση ουσιαστικά περιγράφει την ζητούμενη συμπεριφορά και είναι το αντικείμενο πάνω στην οποία γίνεται η βελτιστοποίηση των αποφάσεων του συστήματος μας.

Στην περίπτωσή μας, όπως αναφέραμε, ερευνάμε τη λαβή (grasp) ενός αντικειμένου σε κάποιο ύψος. Ωστόσο, δεν θέλουμε απλώς να σηκώσουμε το αντικείμενο σε ένα τελικό ύψος-στόχο, αλλά θέλουμε το αντικείμενό μας να ακολουθεί μία επιθυμητή τροχιά ύψους, έστω h_D. Επιπλέον, επιθυμούμε να διατηρούμε το αντικείμενο οριζόντιο, ώστε να αποτρέψουμε μία συμπεριφορά κατά την οποία το ρομπότ σηκώνει το αντικείμενο μόνο από την μία πλευρά, ώστε να αυξήσει το ύψος του κέντρου βάρος του. Για τον λόγο αυτό, χρησιμοποιούμε την αντίστοιχη γωνία ανύψωσης ή pitch των γωνιών προσανατολισμού του. Τέλος, επιθυμούμε να ελέγξουμε την δύναμη που ασκεί το ρομπότ στο αντικείμενο. Συνολικά, λοιπόν η επιβράβευση που χρησιμοποιούμε ορίζεται από τους τρεις επιμέρους όρους ύψους, pitch, δύναμης:

$$r(t) = -w_1 r_h(t) - w_2 r_{pitch}(t) - w_3 r_f(t)$$

Όπως περιγράψαμε στην παράγραφο 4.1, την χρονική στιγμή t ο χώρος κατάστασης περιέχει τον στόχο $h_D(t+1)$, που θέλουμε να έχει το αντικείμενο την επόμενη χρονική στιγμή, μετά την λήψη της απόφασης από το σύστημα. Συνεπώς, διαμορφώνουμε την επιβράβευση αξιολογώντας πόσο κοντά έφτασε το ύψος στην τιμή που ζητήσαμε με την απόφαση που πάρθηκε από την πολιτική. Επομένως, η επιβράβευση ορίζεται ως:

$$r_h(t) = |h(t) - h_D(t)|$$

Αντίστοιχα, για να κρατήσουμε το αντικείμενο οριζόντιο θα πρέπει η γωνία pitch του αντικειμένου να είναι μηδενική, χρησιμοποιώντας ως σημείο αναφορά το οριζόντιο επίπεδο. Συνεπώς, σχηματίζουμε τον αντίστοιχο όρο της επιβράβευσης ως:

$$r_{pitch}(t) = |pitch(t)|$$

Όσον αφορά την δύναμη άσκησης, όπως αναφέραμε επιθυμούμε να διασφαλίσουμε την επαφή του αντικειμένου με τα ακροδάχτυλα. Για τον λόγο αυτόν ζητάμε μέσω της επιβράβευσης η δύναμη άσκησης να είναι μεγαλύτερη από κάποια τιμή $f_{min,i}$, όπου το i αφορά την δύναμη στο εκάστοτε δάκτυλο, εισάγοντας δηλαδή ένα κάτω όριο. Επιπλέον, θέλουμε να ελέγξουμε το μέγεθος της δύναμης αυτής εισάγοντας ένα άνω όριο $f_{max,i}$. Σχηματίζουμε την επιβράβευση r_f χρησιμοποιώντας μία ποινή (βάρους w_3) εάν οι τιμές των δυνάμεων βρίσκονται εκτός των ορίων:

$$r_f(t) = \begin{cases} 0 & \text{if } f_{min,i} < f(t) < f_{max,i}, i = 1.2, \dots \\ 1 & \text{if } otherwise. \end{cases}$$

4.3 Μοντέλο

Για την αναπαράσταση της πολιτικής και της συνάρτησης αξίας χρησιμοποιούμε το μοντέλο δράστη-κριτή (παράγραφος 2.6), δηλαδή μία παραμετροποιημένη πολιτική $\pi_\theta(a|s)$ που λαμβάνει αποφάσεις και μία παραμετροποιημένη συνάρτηση αξίας $V_w(s)$, η οποία έχει τον ρόλο να κρίνει τις δράσεις που επιλέγονται από την πολιτική, συνδράμοντας στον υπολογισμό της κλίσης πολιτικής. Λόγω του ότι οι χώροι κατάστασης και δράσεις είναι συνεχείς χρειαζόμαστε έναν προσεγγιστή συνάρτησης (function approximator) τόσο για την πολιτική (δράστης) όσο και για την συνάρτηση αξίας (κριτής). Χρησιμοποιούμε δύο πλήρως συνδεδεμένα νευρωνικά δίκτυα που υλοποιούν αυτές τις λειτουργίες.

Όπως περιγράψαμε στο κεφάλαιο 2, η πολιτική θα πρέπει να είναι στοχαστική, ώστε να υπάρχει δυνατότητα εξερεύνησης. Χρησιμοποιούμε για τον λόγο αυτό μία γκαουσιανή πολιτική:

$$\pi(a|s,\theta) = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} \exp\left(-\frac{1}{2}(a - \mu(s,\theta_{NN}))^T\Sigma^{-1}(a - \mu(s,\theta_{NN}))\right)$$

Η έξοδος του πλήρως συνδεδεμένου νευρωνικού δικτύου του δράστη είναι η μέση τιμή της πολιτικής αυτής $\mu(s,\theta_{NN})$ ενώ ως μήτρα συνδιακύμανσης Σ χρησιμοποιούμε διαγώνιο διάνυσμα, του οποίου τα βάρη αποτελούν παράμετρους προς μάθηση. Η παραμετροποι-

ημένη πολιτική συνολικά, έχει ως παραμέτρους τα βάρη του νευρωνικού δικτύου και την μήτρα συνδιακύμανσης, δηλαδή $\theta = \theta_{NN} \cup \Sigma$. Το νευρωνικό δίκτυο του κριτή είναι και αυτό ένα πλήρως συνδεδεμένο νευρωνικό δίκτυο με είσοδο την κατάσταση s και έξοδο την τιμή της συνάρτηση αξίας $V_w(s)$.

4.4 Εκπαίδευση με Ενισχυτική Μάθηση

Για την εκπαίδευση του μοντέλου χρησιμοποιούμε on-policy ενισχυτική μάθηση και όχι off-policy λόγω των καλύτερων ιδιοτήτων σύγκλισης σε προβλήματα μεγάλων διαστάσεων χώρων κατάστασης και δράσης. Συγκεκριμένα χρησιμοποιούμε τον αλγορίθμου TRPO (παράγραφος 2.8), οποίος έχει ως αντικείμενο την βελτιστοποίηση σύμφωνα με την αντικειμενική συνάρτηση

$$maximize_{\theta} L(\theta, \theta_{old}) = maximize_{\theta} \mathbb{E}_t[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} A_t]$$

subject to

$$\mathbb{E}_t[KL[\pi_\theta(\cdot|s_t), \pi_{\theta_{old}}(\cdot|s_t)]] \leq \delta$$

όπου δ είναι το όριο που θέτουμε για την αλλαγή της πολιτικής (KL-Divergence) σε κάθε επανάληψη και μεταχειριζόμαστε σαν υπερπαράμετρο. Στην παραπάνω αντικειμενική συνάρτηση θα μπορούσαμε να χρησιμοποιήσουμε την Q συνάρτηση αξίας, αλλά όπως περιγράψαμε στην παράγραφο 2.7 η advantage function $A_t = Q_t - V_t$ στην οποία αφαιρείται από την Q η V ως baseline ώστε να μειωθεί το variance. Για την επίλυση της, χρησιμοποιούμε την Natural Policy Gradient, όπως περιγράψαμε στην παράγραφο 2.8, που δίνει λύση:

$$\theta_{k+1} = \theta_k + \sqrt{\frac{2\delta}{g^T F^{-1} g}} F^{-1} g$$

με g την Vanilla Policy Gradient και F τον Fisher Information Matrix. Για να φτάσουμε σε αυτή ακολουθούμε την ακόλουθη διαδικασία.

Σε κάθε επανάληψη k του αλγορίθμου, συλλέγουμε έναν αριθμό τροχιών με βάση την παρούσα στοχαστική πολιτική π_{θ_k}. Μετά την συλλογή των δειγμάτων υπολογίζουμε την παρακάτω ποσότητα $\nabla_\theta \log \pi(a_t|s_t, \theta_k)$ για κάθε ζευγάρι (s_t, a_t) από τα δείγματα των τροχιών που συλλέξαμε.

Στην συνέχεια υπολογίζουμε την πλεονεκτική συνάρτηση αξίας (Advantage Function), από τα δείγματα στην επανάληψη k και την εκτίμηση της συνάρτηση αξίας στην επανάληψη $k-1$ από το νευρωνικό δίκτυο του κριτή. Για τον υπολογισμό της πλεονεκτικής συνάρτησης αξίας χρησιμοποιούμε την Generalized Advantage Estimation (GAE) που περιγράψαμε στην παράγραφο 2.7. Σύμφωνα με αυτή:

$$A_k^\pi(s_t, a_t) = (1-\lambda)(A_t^{(1)} + \lambda A_t^{(2)} + \lambda^2 A_t^{(3)} + \ldots) = \ldots = \sum_{t=0}^{\infty} (\gamma\lambda)^t (r_t + \gamma V(s_{t+1}) - V(s_t))$$

όπου γ η υπερπαράμετρος του φθίνοντος αθροίσματος της απόδοσης (return) και λ υπερπαράμετρος συνδυασμού όλων των ν-οστών αποδόσεων, όπως περιγράψαμε στην παράγραφο

2.7.

Στην συνέχεια βάσει των τιμών της πλεονεκτικής συνάρτησης που υπολογίσαμε για τα N ζεύγη (s, a) μπορούμε να υπολογίσουμε την κανονική κλίση (Vanilla Policy Gradient), ως:

$$g = \frac{1}{N}\sum_{t=0}^{N} \nabla_\theta \log \pi_\theta(a_t|s_t) A^\pi(s_t, a_t)$$

Το επόμενο βήμα είναι ο υπολογισμός της Natural Policy Gradient $g^{nat} = F^{-1}g$. Συνεπώς, πρέπει να υπολογιστεί ο πίνακας Fisher information matrix:

$$F = \frac{1}{N}\sum_{t=0}^{N} \nabla_\theta \log \pi_\theta(a_t|s_t) \nabla_\theta \log \pi_\theta(a_t|s_t)^T$$

Ωστόσο, ο υπολογισμός του αντιστρόφου του είναι μια υπολογιστικά δύσκολη διαδικασία για έναν πίνακα με τόσες παραμέτρους όσες ενός μεγάλου νευρωνικού δικτύου, ενώ μπορεί να είναι και ασταθής. Για τον λόγο αυτό, υπολογίζουμε ολόκληρη την ποσότητα (Natural Policy Gradient) $g^{(nat)} = F^{-1}g$ χρησιμοποιώντας την Conjugate gradient.

Αλγοριθμος 4.1: *Conjugate Gradient*

Initialize:$x_0 \in R^n$ arbitary, $d_0 = g_0 = b - Qx_0$
for k=0,1,2,... **do**
  $a_k = \frac{g_k^T g_k}{d_k^T Q d_k}$
  $x_{k+1} = x_k + a_k d_k$
  $g_{k+1} = g_k - a_k d_k$
  If ($g_{k+1} < threshold$) then Break Loop
  $\beta_k = \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k}$
  $d_{k+1} = g_{k+1} + \beta_k d_k$
end for
Return x_{k+1}

Σύμφωνα με την μέθοδο Conjugate gradient (Αλγόριθμος 4.1), το πρόβλημα εύρεσης της λύσης x^* της εξίσωσης $Qx = b$ όπου $Q \in R^{nxn}$ θετικά ορισμένος πίνακας είναι ισοδύναμο με το ακόλουθο:

$$minimize_x[f(x) = (\frac{1}{2}x^T Qx - b^T x)]$$

εφόσον $f'(x) = Qx - b = 0$

Για διανύσματα $(d_0, d_1, ..., d_{n-1})$ ανεξάρτητα τα οποία είναι Q-conjugate, δηλαδή $d_i^T Q d_j = 0$ η λύση x^* είναι:

$$x = a_0 d_0 + a_1 d_1 + ... + a_{n-1} d_{n-1}$$

Επομένως:

$$d_i^T Q x^* = d_i^T Q(a_0 d_0 + a_1 d_1 + ... + a_{n-1} d_{n-1}) = a_i d_i^T Q d_i$$

ή

$$a_i = \frac{d_i^T Q x^*}{d_i^T Q_d i} = \frac{d_i^T b}{d_i^T Q d_i}$$

Έτσι με έναν αριθμό βημάτων n μπορούμε να βρούμε επαναληπτικά την λύση x^* προσθέτοντας κάθε φορά την ποσότητα $x_{i+1} = x_i + a_i d_i$ χωρίς τον υπολογισμό του αντιστρόφου του πίνακα Q. Ο αλγόριθμος Conjugate Gradient αποδεικνύεται ότι είναι μία Conjugate Direction μέθοδος, δηλαδή τα διανύσματα d_i είναι Q-conjugate και ότι συγκλίνει στην λύση x^* από τυχαία αρχικοποιημένη πρόβλεψη της λύσης: x_0.

Αλγοριθμος 4.2: *Trust Region Policy Optimization (Τροποποίηση από [45], [11])*

1: Initialize ∂ to ∂_0
2: **for** $k = 1 to K$ **do**
3: Collect trajectories $\tau^{(1)}, ..., \tau^{(N)}$ by rolling out the stochastic policy $\pi(\cdot, \partial_k)$
4: Compute $\nabla_\partial \log \pi(a_t|s_t; \partial_k)$ for each (s,a) pair along trajectories sample in iteration k.
5: Compute advantages A_k^π base on trajectories in iteration k and approximate value function V_{k-1}^π.
6: Compute policy gradient $g = \frac{1}{N} \sum_{t=0}^{N} \nabla_\partial \log \pi_\partial(a_t|s_t) A^\pi(s_t, a_t, t)$, where N is the number of (s,a) pairs.
7: Compute the Fisher matrix $F = \frac{1}{T} \sum_{t=0}^{T} \nabla_\partial \log \pi_\partial(a_t|s_t) \nabla_\partial \log \pi_\partial(a_t|s_t)^T$
8: Compute Natural Policy Gradient $g^{(nat)}$ using Conjugate Gradient
9: Compute Proposed Step $\Delta_k = \sqrt{\frac{2\delta}{g^T g^{(nat)}}} g^{(nat)}$
10: Perform Line Search and update $\partial_{k+1} = \partial_k + a^j \Delta_k$
11: Update parameters of value function in order to approximate $V_k^\pi(s_t) \approx G_t$
12: **end for**

Βρίσκουμε λοιπόν την Natural Policy Gradient προσεγγιστικά με χρήση της Conjugate Gradient. Αυτή η μέθοδος συχνά περιγράφεται ως Truncated Natural Policy Gradient (TNPG). Το τελευταίο βήμα για την ανανέωση του δικτύου πολιτικής σύμφωνα με όσα περιγράψαμε στην παράγραφο 2.8 για τον αλγόριθμο TRPO είναι η εύρεση του βήματος ανανέωσης. Αυτό μπορεί να βρεθεί επαναληπτικά (line search) με την χρήση μίας παραμέτρου a μείωσης του προτεινόμενου βήματος από την Natural Policy Gradient, ώστε να ικανοποιείται η KL συνθήκη και η μονοτονική ανανέωση. Όσον αφορά την KL συνθήκη, χρησιμοποιούμε την μέση τιμή βάσει των δειγμάτων που συλλέχθηκαν και όχι την max που προτείνεται θεωρητικά, διότι κάτι τέτοιο θα οδηγούσε σε υπερβολικά μικρά βήματα. Επομένως έχουμε για το τελικό βήμα ανανέωσης :

$$a_{TRPO} = a^j \sqrt{\frac{2\delta}{g^T F^{-1} g}}$$

με $a \in (0, 1)$ και j τον μικρότερο θετικό ακέραιο ώστε:

$$\overline{KL}[\pi_\partial, \pi_{\partial_{old}}] \leq \delta \text{ and } L(\partial, \partial_{old}) \geq 0$$

Έχοντας λοιπόν υπολογίσει την Natural Policy Gradient και το βήμα ανανέωσης ανανεώνουμε τα βάρη του δράστη ως:

$$\theta_{k+1} = \theta_k + \alpha_{TRPO} g^{nat}$$

Τέλος ανανεώνουμε το δίκτυο του κριτή χρησιμοποιώντας τον Monte Carlo στόχο (παράγραφος 2.5). Υπολογίζουμε, δηλαδή για κάθε κατάσταση s_t που έχουμε στα δείγματα που συλλέχθηκαν στις τροχιές με μήκος T της παρούσας επανάληψης, την απόδοση

$$G_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_t$$

και ανανεώνουμε τα βάρη w του δικτύου κριτή ως:

$w_{k+1} = w_k + \alpha(G_t - V(s_t, w))\nabla_w V(s_t, w)$

Παραθέτουμε συνολικά την υλοποίηση του αλγορίθμου TRPO που χρησιμοποιήσαμε στον Αλγόριθμο 4.2.

4.5 Προεκπαίδευση με Δεδομένα Επίδειξης

Η ενισχυτική μάθηση είναι ένα αρκετά δύσκολο πρόβλημα, στο οποίο η διάσταση των χώρων κατάστασης και δράσης καθώς και το ίδιο το πρόβλημα βελτιστοποίησης έχει πολύ μεγάλη επίδραση στην απόδοση. Ένα δύσκολο πρόβλημα με μεγάλη διάσταση ίσως καταλήξει σε υποβέλτιστη λύση ή σε κακή απόδοση χρησιμοποιώντας τεράστιο αριθμό δειγμάτων. Τα δεδομένα επίδειξης μπορούν να βοηθήσουν πολύ αποδοτικά προς επίλυση του προβλήματος αυτού.

Στην παρούσα μεθοδολογία χρησιμοποιούμε τα δεδομένα επίδειξης για προεκπαίδευση του συστήματος μας βάσει του αλγορίθμου Behavior Cloning. Επιλέγουμε τον αλγόριθμο αυτό, διότι θεωρούμε ότι τα δεδομένα επίδειξης που χρησιμοποιούμε δεν αποτελούν μία βέλτιστη λύση, αλλά απλώς μία αποδεκτή λύση, η οποία όμως μπορεί να βελτιωθεί.

Από τους αλγορίθμους μάθησης από δεδομένα επίδειξης που περιγράψαμε στο Κεφάλαιο 2, δεν μπορούμε να τα χρησιμοποιήσουμε με κάποιον αλγόριθμο διαδραστικού δασκάλου (παράγραφος 2.2) εφόσον τα δεδομένα επίδειξης που χρησιμοποιούμε δεν θεωρούμε ότι είναι ικανοποιητικά ώστε να διορθώσουν προς την βέλτιστη επιθυμητή συμπεριφορά.

Επιπλέον, ένας αλγόριθμος αντίστροφης ενισχυτικής μάθησης δεν θα ήταν κατάλληλος για μάθηση τροχιάς εφόσον η επιβράβευση μαθαίνεται βάσει των δεδομένων επίδειξης τα οποία ακολουθούν διαφορετικές-λανθασμένες τροχιές. Στην περίπτωση μας δηλαδή δεν θέλουμε να μιμηθούμε τα δεδομένα επίδειξης, αλλά θέλουμε απλώς να τα χρησιμοποιήσουμε ως μία αρχικοποίηση για το μοντέλο μας. Εκπαιδεύουμε λοιπόν αρχικά το μοντέλο μας με επιβλεπόμενη μάθηση σύμφωνα με τον αλγόριθμο Behavior Cloning και στην συνέχεια εκπαιδεύουμε με ενισχυτική μάθηση, ώστε να βρεθεί μία καλύτερη ή βέλτιστη συμπεριφορά. Είναι προφανές ότι η απόδοση της ενισχυτικής μάθησης που γίνεται στην συνέχεια σχετίζεται άμεσα με την ποιότητα των δεδομένων επίδειξης.

Συλλέγοντας λοιπόν δεδομένα επίδειξης, έχουμε τροχιές αποτελούμενες από ζεύγη κατάστασης-δράσης (s_t^D, a_t^D). Χρησιμοποιούμε τα ζεύγη αυτά ως ζεύγη χαρακτηριστικών-ετικετών για να πραγματοποιήσουμε παλινδρόμηση (regression) στις παραμέτρους της πολιτικής $\theta = \theta_{NN} \cup \Sigma$. Λόγω του ότι η πολιτική μας είναι μία κατανομή χρησιμοποιούμε ως συνάρτηση σφάλματος

την αρνητική log πιθανοφάνεια της γκαουσιανής:

$$L_{BC} = -ln\left(\frac{1}{\sqrt{(2\pi)^k|\Sigma|}}\exp\left(-\frac{1}{2}(a-\mu(s,\theta_{NN}))^T\Sigma^{-1}(a-\mu(s,\theta_{NN}))\right)\right)$$

,όπου μ η έξοδος του νευρωνικού δικτύου και a_t^D η δράση του expert στην κατάσταση s_t^D.

Κεφάλαιο 5

Συλλογή Δεδομένων Επίδειξης σε Περιβάλλον Εικονικής Πραγματικότητας

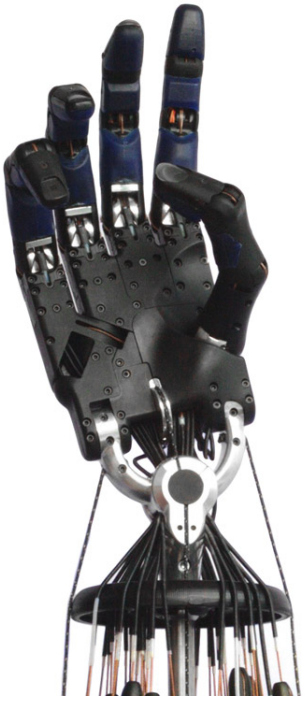
Σε αυτό το κεφάλαιο γίνεται περιγραφή του υλικού και του λογισμικού που χρησιμοποιήθηκε για τα πειράματα της εργασίας και του τρόπου συλλογής των δεδομένων επίδειξης. Συγκεκριμένα περιγράφουμε τον προσομοιωτή Mujoco και το ρομπότ ADROIT, το οποίο χρησιμοποιούμε στην προσομοίωση. Στην συνέχεια, αναλύουμε διαδικασία συλλογής των δεδομένων επίδειξης στο περιβάλλον προσομοίωσης, κατά την οποία χρησιμοποιούμε την συσκευή Leap Motion.

5.1 Το ρομπότ ADROIT

Το ρομπότ ADROIT [46] είναι ένα ανθρωπομορφικό ρομποτικό χέρι με βραχίονα, το οποίο αποτελείται από το ρομποτικό χέρι Shadow Hand (24 βαθμοί ελευθερίας) της Shadow Hand Company [4] και ένα βραχίονα (6 βαθμοί ελευθερίας). Το Shadow Hand Robot σχεδιάστηκε ώστε να αναπαριστά πιστά το ανθρώπινο χέρι και τους βαθμούς ελευθερίας του, με μέγεθος και σχήμα παρόμοιο με αυτό ενός ενήλικα άνδρα. Το χέρι του ADROIT έχει τον σκελετό του Shadow Hand, και είναι ενισχυμένο με κάποια αναβαθμισμένα χαρακτηριστικά ελέγχου με στόχο την καλύτερη απόδοση σε επιδέξιες εργασίες ρομποτικού χειρισμού.

Καθένα από τα τέσσερα δάχτυλα δείκτη, μέσου, παράμεσου, μικρού αποτελείται από 4 βαθμούς ελευθερίας στις αντίστοιχες αρθρώσεις joints ως εξής: 1 joint για την σύνδεση πρώτης-δεύτερης φάλαγγας, 1 joint για σύνδεση δεύτερης-τρίτης φάλαγγας και 2 joints (ή 1 universal) για την σύνδεση πρώτης φάλαγγας και μετακαρπίου. Ο αντίχειρας αποτελείται από 5 joints, ένα παραπάνω από τα υπόλοιπα τέσσερα που αποσκοπεί στην δυνατότητα περιστροφής του αντίχειρα στο ρομπότ. Επιπλέον υπάρχει ένα joint στο μετακάρπιο του μικρού δαχτύλου που αντιπροσωπεύει την δυνατότητα κάμψης της παλάμης, και δύο επιπλέον joints στον καρπό για τις δυνατότητες κάμψης-έκτασης και προσαγωγής-επαγωγής του. Τέλος, το ρομποτικό χέρι έχει προσαρμοστεί πάνω σε έναν ρομποτικό βραχίονα 6 βαθμών ελευθερίας, που αντιπροσωπεύουν τις δυνατότητες του ανθρώπινου αγκώνα και ώμου, με σκοπό την κίνηση του χεριού στον χώρο, δίνοντας ένα συνολικό σύστημα χεριού-βραχίονα 30 βαθμών ελευθερίας.

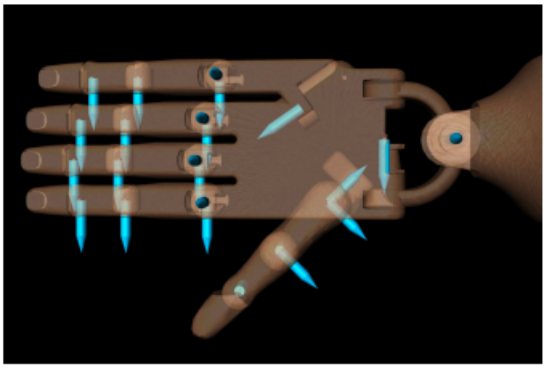
Το εύρος κίνησης των joints του χεριού έχει σχεδιαστεί σε αντιστοίχιση με τις ικανότητες των αρθρώσεων ενός ανθρώπινου. Το ADROIT διαθέτει πνευματικό σύστημα ενεργοποίησης,



Σχήμα 5.1: *Το ρομπότ Shadow Hand*

διαθέτοντας 40 τένοντες και κάθε joint ενεργοποιείται με αντίστοιχο ελεγκτή θέσης (position controller) της γωνίας του, ενώ για την μέτρηση της γωνίας καθενός joint χρησιμοποιούνται αισθητήρες Hall. Επιπλέον, το ρομπότ είναι εφοδιασμένο με αισθητήρες αφής (tactiles) για την μέτρηση των δυνάμεων επαφής. Ο αριθμός τους και η θέση τους ποικίλουν ανάλογα με την εφαρμογή. Κυριότεροι είναι οι αισθητήρες αφής στα ακροδάχτυλα (fingertips), ενώ μπορούν να προστεθούν και στους υπόλοιπους συνδέσμους του χεριού και στην παλάμη.

Στα πειράματα που γίνονται σε αυτή την εργασία χρησιμοποιούμε τα τρία δάχτυλα: αντίχειρα, δείκτη και μέσο (συνολικά 13 βαθμοί ελευθερίας) με τους ελεγκτές και αισθητήρες που βρίσκονται στις αρθρώσεις τους και επιπλέον τους αισθητήρες αφής (tactiles) που βρίσκονται στις άκρες των δακτύλων αυτών.

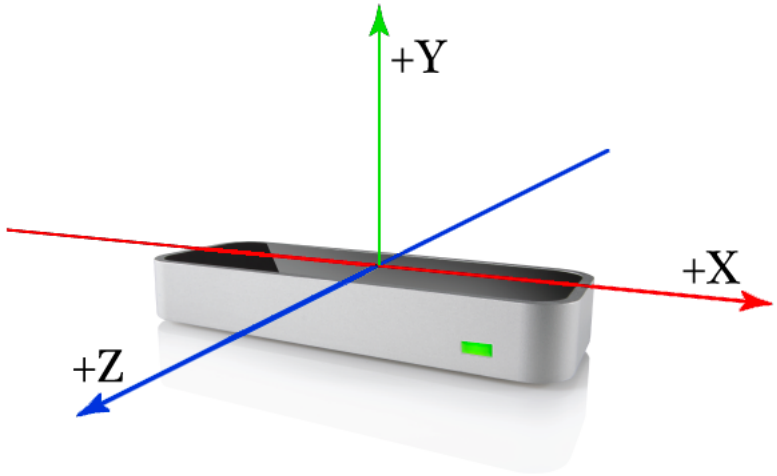


Σχήμα 5.2: *Βαθμοί ελευθερίας του χεριού ADROIT (από [2])*

5.2 Χειρισμός ρομπότ ADROIT σε Περιβάλλον Εικονικής Πραγματικότητας

Για την εφαρμογή αλγορίθμων μάθησης κινήσεων χειρισμού σε ρομποτικό χέρι έγινε χρήση του προσομοιωτή φυσικής Mujoco [47]. Η δυναμική επαφής, ο σχεδιασμός κινηματικής, δυναμικής και αισθητήρων που προσομοιώνουν σε μεγάλο βαθμό τον φυσικό κόσμο, τον καθιστά πολύ δημοφιλή για προσομοιώσεις σε προβλήματα που εμπεριέχουν πολύπλοκες επαφές μεταξύ αντικειμένων και συνεπώς σε προβλήματα χειρισμού αντικειμένων από ρομπότ.

Η διαδικασία μάθησης μίας συμπεριφοράς από κάποιον expert για την εκπόνηση μίας εργασίας από το ρομπότ, προϋποθέτει την καταγραφή των δεδομένων επίδειξης. Δημοφιλής τρόπος, που έχει χρησιμοποιηθεί σε άλλες εργασίες είναι ο τηλεχειρισμός του ρομποτικού χεριού μέσα στο περιβάλλον προσομοίωσης με χρήση του γαντιού Cyberglove III για την καταγραφή των δαχτύλων, του HTC vive tracker, για την καταγραφή της βάσης του χεριού και του HTC vive headset, για την στερεοσκοπική απεικόνιση του εικονικού περιβάλλοντος.



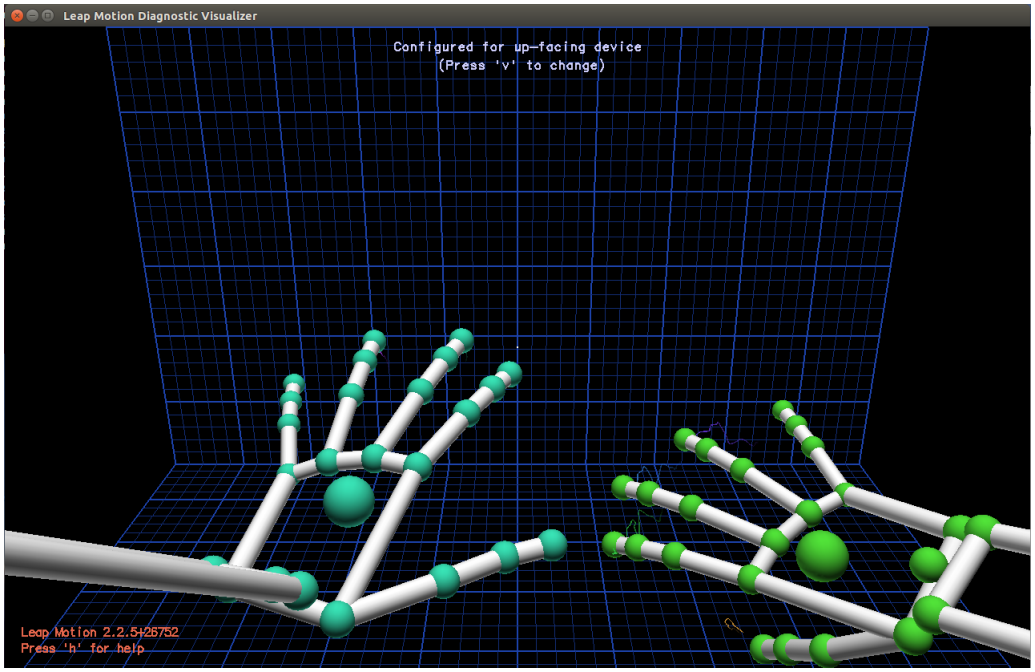
Σχήμα 5.3: *Η συσκευή Leap Motion και το σύστημα αξόνων του*

Διαφορετική επιλογή για την συλλογή δεδομένων που έχει χρησιμοποιηθεί σε εργασίες είναι συσκευές-¨ποντίκια¨ με πολλούς βαθμούς ελευθερίες για τον χειρισμό διατάξεων σε τρισδιάστατα περιβάλλοντα. Τέτοιες συσκευές είναι για παράδειγμα τα SpaceMouse της εταιρίας 3dconnection.

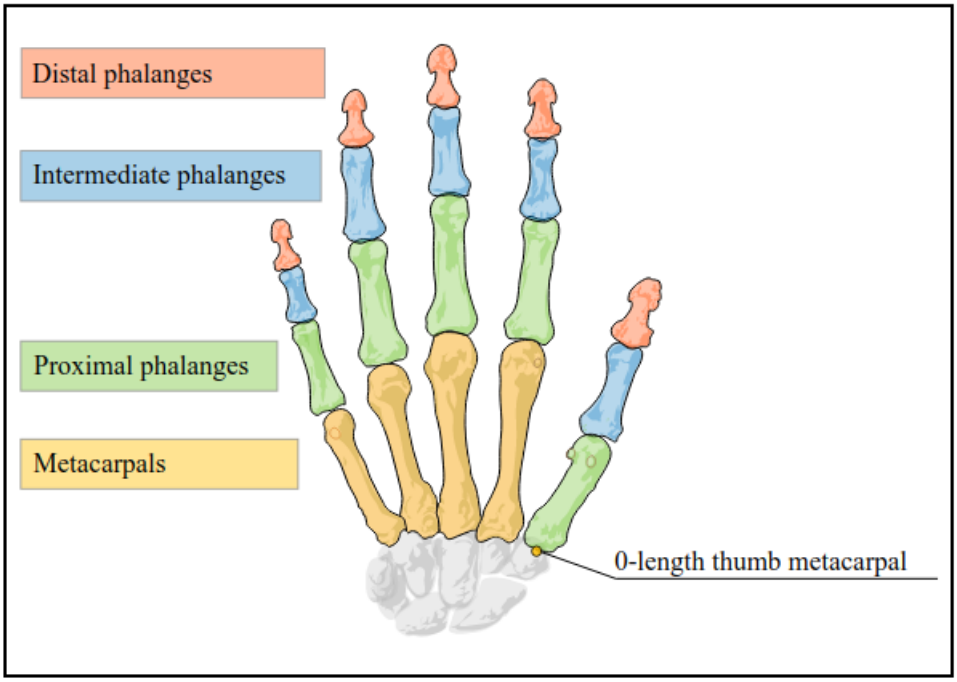
Σε αυτή την εργασία τα δεδομένα συλλέχθηκαν στο εικονικό περιβάλλον του προσομοιωτή Mujoco με χειρισμό του χεριού με τον αισθητήρα Leap Motion [3], μια συσκευή η οποία με χρήση υπερύθρων καταγράφει την θέση και την διάταξη του χεριού στον χώρο. Η συσκευή αυτή ειναι σημαντικά πιο φθηνή από αυτές που αναφέρθηκαν παραπάνω και δείχνουμε στην εργασία αυτή ότι μπορούμε να έχουμε ικανοποιητικά αποτελέσματα χωρίς σημαντικό κόστος εξοπλισμού.

Βάσει ενός συστήματος αξόνων πάνω στον αισθητήρα, παρέχονται ως δεδομένα οι θέσεις των αρθρώσεων των δακτύλων, τα διανύσματα των οστών των δακτύλων, καθώς και η θέση και ο προσανατολισμός της παλάμης. Με βάση τα δεδομένα που προσφέρει ο αισθητήρας Leap Motion μπορούμε να υπολογίσουμε τις γωνίες των αρθρώσεων του expert και στην συνέχεια

να τις προωθήσουμε ως είσοδο στους ελεγκτές θέσης του ρομπότ ADROIT, δημιουργώντας έτσι ένα περιβάλλον εικονικής πραγματικότητας, όπου ο χρήστης και στην περίπτωση μας ο expert, τηλε-χειρίζεται το ρομπότ.



Σχήμα 5.4: *Καταγραφή σκελετού από την συσκευή Leap Motion.*

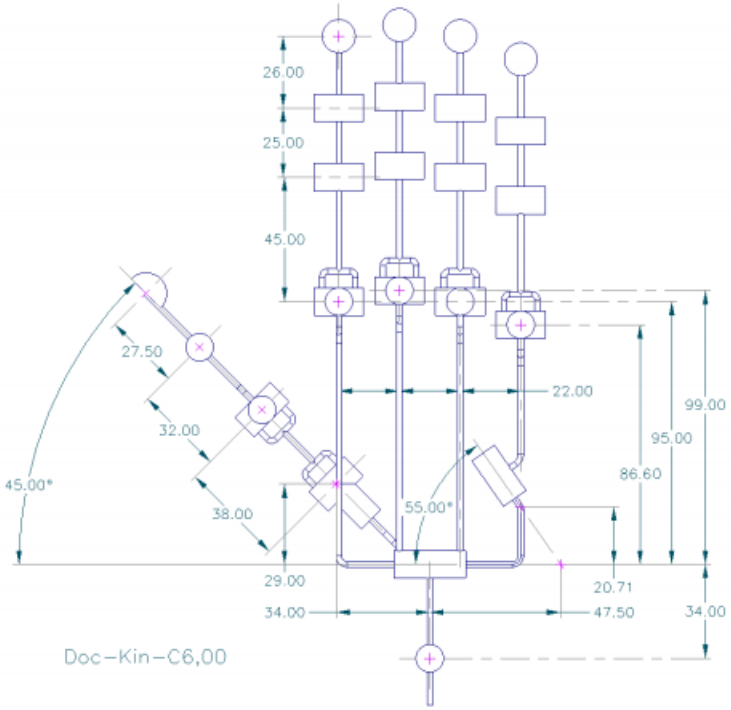


Σχήμα 5.5: *Τα οστά των δακτύλων που καταγράφονται από την συσκευή Leap Motion (από [3])*

5.3 Κινηματική Ανάλυση χεριού ADROIT

Για τον έλεγχο του ρομποτικού χεριού ADROIT σε περιβάλλον εικονικής προσομοίωσης πραγματοποιήσαμε μία μεταφορά των γωνιών των αρθρώσεων από το ανθρώπινο χέρι στο ρομποτικό χέρι. Στα πειράματά μας, όπως έχουμε αναφέρει, θέλουμε να πραγματοποιήσουμε μία εσωτερική λαβή κατά την οποία χρησιμοποιούμε μόνο τα δάκτυλα για το grasp ενός αντικειμένου. Συγκεκριμένα, χρησιμοποιούμε μόνο τρία δάχτυλα: τον αντίχειρα, τον μέσο και τον δείκτη και επομένως στην συνέχεια θα αναφερόμαστε μόνο σε αυτά και τους βαθμούς ελευθερίας τους.

Ο αισθητήρας Leap Motion, παρέχει δεδομένα για τις θέσεις των αρθρώσεων και για την θέση και προσανατολισμό της παλάμης. Χρησιμοποιήσαμε τα δεδομένα αυτά και το μοντέλο του ρομποτικού χεριού, ώστε να υπολογίσουμε τις αντίστοιχες γωνίες των αρθρώσεων του ανθρώπινου χεριού βάσει του μοντέλου του ρομποτικού χεριού Shadow Hand.

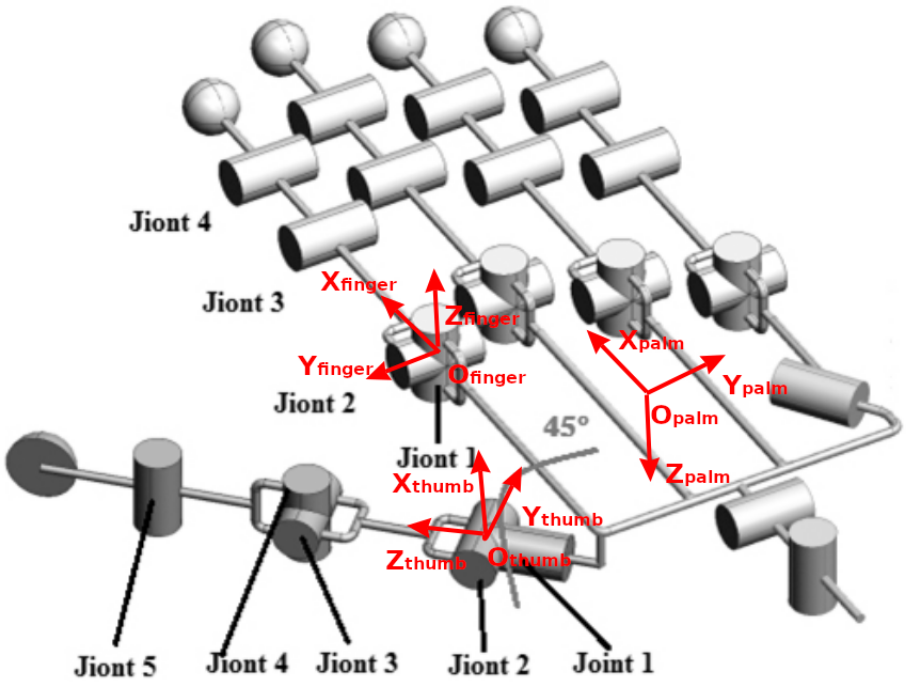


Σχήμα 5.6: *Διάγραμμα του χεριού Shadow Hand (από [4])*

Οι γωνίες των αρθρώσεων μπορούν να βρεθούν με αντίστροφη κινηματική ανάλυση. Ωστόσο, εφόσον εκτός από το άκρο των δαχτύλων γνωρίζουμε και όλες τις ενδιάμεσες θέσεις των αρθρώσεων αυτό δεν είναι απαραίτητο, αφού μπορούμε κάθε φορά να υπολογίζουμε τον μετασχηματισμό ως προς το σύστημα βάσης της κάθε άρθρωσης και να υπολογίζουμε στην συνέχεια τις γωνίες (βαθμούς ελευθερίας) της άρθρωσης αυτής βάσει του σημείου της θέσης της επόμενης άρθρωσης. Ακολουθεί η κινηματική ανάλυση για τα τρία δάχτυλα που χρησιμοποιήσαμε, τον αντίχειρα, τον δείκτη και τον μέσο. Όσον αφορά τον δείκτη και τον μέσο η ανάλυση είναι η ίδια.

Σαν σημειολογία για τους πίνακες μετασχηματισμών, χρησιμοποιούμε τον συμβολισμό

$R_{\Sigma_1}^{\Sigma_2}$ και $T_{\Sigma_1}^{\Sigma_2}$ για να δηλώσουμε τον μετασχηματισμό στροφής και τον ομογενή μετασχηματισμό αντίστοιχα, από το σύστημα συντεταγμένων Σ_1 στο σύστημα Σ_2. Για τα διανύσματα χρησιμοποιούμε την έκφραση p^{Σ} για να δηλώσουμε ότι το διάνυσμα p είναι εκφρασμένο ως προς το σύστημα Σ.



Σχήμα 5.7: *Συστήματα βάσης δαχτύλων και παλάμης*

Αρχικά υπολογίζουμε τους μετασχηματισμούς από το σύστημα του αισθητήρα στα συστήματα βάσης των δαχτύλων, όπως φαίνεται στο Σχήμα 5.7. Για να το επιτύχουμε αυτό χρησιμοποιούμε τον προσανατολισμό της παλάμης . Σύμφωνα με αυτόν ο η στροφή από το σύστημα του αισθητήρα στο σύστημα της παλάμης, η οποία δίνεται σαν πληροφορία, είναι:

$$R_{leap}^{palm} = \begin{bmatrix} \mathbf{n}^{\mathbf{leap}} & \mathbf{o}^{\mathbf{leap}} & \mathbf{a}^{\mathbf{leap}} \end{bmatrix}$$

Σύμφωνα με την στροφή αυτή υπολογίζουμε τον ομογενή μετασχηματισμό προς την βάση του αντίχειρα ως:

$$R_{leap}^{thumb} = R_{leap}^{palm} \cdot \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos(45^\circ) & -sin(45^\circ) \\ 0 & sin(45^\circ) & cos(45^\circ) \end{bmatrix}$$

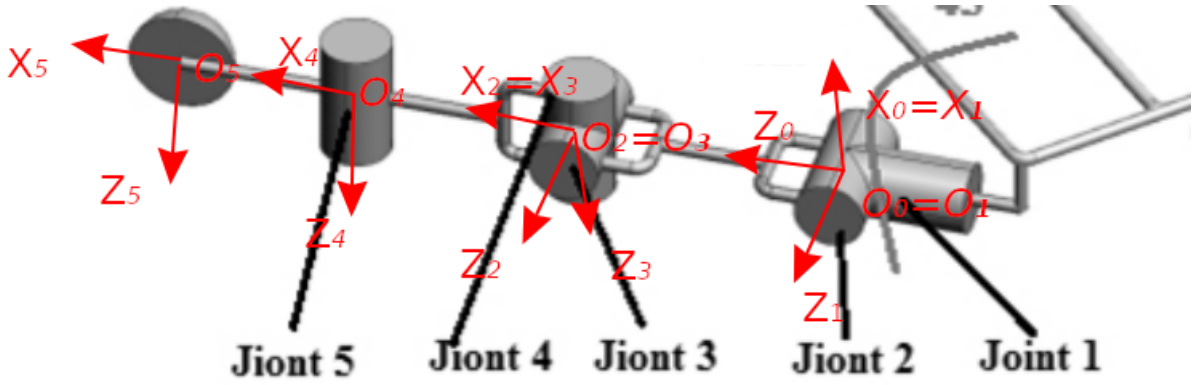
$$T_{leap}^{thumb_0} = \begin{bmatrix} R_{leap}^{thumb_0} & O_{thumb}^{leap} \\ 0 & 1 \end{bmatrix}$$

και προς την βάση των δακτύλων δείκτη και μέσου:

$$R_{leap}^{finger} = R_{leap}^{palm} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

$$T_{leap}^{thumb_0} = \begin{bmatrix} R_{leap}^{thumb_0} & O_{finger}^{leap} \\ 0 & 1 \end{bmatrix}$$

Έχοντας υπολογίσει τους μετασχηματισμού προς τις βάσεις του κάθε δαχτύλου κάνουμε κινηματική ανάλυση ως προς αυτές. Για να μην υπερφορτώσουμε τη σημειολογία το όνομα του δαχτύλου παραλείπεται και το σύστημα βάσης τώρα σημειώνεται σαν $\Sigma_0 = (O_0, X_0, Y_0, Z_0)$.



Σχήμα 5.8: *Συστήματα αρθρώσεων αντίχειρα με βάση την μέθοδο DH*

Υπολογίζουμε την μήτρα Denavit-Hartenberg του αντίχειρα σύμφωνα με το Σχήμα 5.8.

Πίνακας 5.1: *DH Matrix of Thumb*

joint i	a_i	α_i	d_i	ϑ_i
1	0	90°	0	q_1
2	l_1	0°	0	$q_2 + 90°$
3	0	-90°	0	q_3
4	l_2	0°	0	q_4
5	l_3	0	0°	q_5

Βάσει αυτής λαμβάνουμε τους ακόλουθους μετασχηματισμούς

$$T_0^1 = \begin{bmatrix} c1 & 0 & c1 & 0 \\ s1 & 0 & -c1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$T_1^2 = \begin{bmatrix} -s2 & -c2 & 0 & -s2l1 \\ c2 & -s2 & 0 & c2l1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$T_2^3 = \begin{bmatrix} c3 & 0 & -s3 & 0 \\ s3 & 0 & c3 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$T_3^4 = \begin{bmatrix} c4 & -s4 & 0 & c4l2 \\ s4 & -c4 & 0 & s4l2 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$T_4^5 = \begin{bmatrix} c5 & -s5 & 0 & c5l3 \\ s5 & -c5 & 0 & s5l3 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Αρχικά μπορούμε να υπολογίσουμε τις γωνίες q_1 και q_2 εφόσον γνωρίζουμε το κέντρο του συστήματος Σ_2 ως προς την βάση του αντίχειρα Σ_0, δηλαδή το διάνυσμα

$$O_2^{\Sigma_0} = (T_{leap}^{thumb})^{-1} O_2^{leap}$$

Υπολογίζοντας τον πίνακα:

$$T_{thumb_0}^{thumb_2} = \begin{bmatrix} -c1s2 & -c1c2 & s1 & -c1s2l1 \\ -s1s2 & -s1c2 & -c1 & -s1s2l1 \\ c2 & -s2 & 0 & c2l1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

απ' όπου παίρνουμε,

$$q2 = \arccos\left(\frac{O_{2,z}^{\Sigma_0}}{l1}\right)$$

$$q1 = \arctan\left(\frac{O_{2,y}^{\Sigma_0}}{O_{2,x}^{\Sigma_0}}\right)$$

Ομοίως για τις γωνίες $q3$ και $q4$ υπολογίζουμε τον πίνακα:

$$T_2^4 = \begin{bmatrix} c3c4 & -c3c4 & -s3 & c3c4l2 \\ s3c4 & -s3s4 & c3 & s3c4l2 \\ -s4 & -c4 & 0 & -s4l2 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

και υπολογίζοντας το διάνυσμα

$$O_4^{\Sigma_2} = (T_{leap}^{thumb} \cdot T_0^2)^{-1} O_4^{leap}$$

βρίσκουμε τις γωνίες ως:

$$q4 = \arcsin\left(-\frac{O_{4,z}^{\Sigma_2}}{l2}\right)$$

$$q3 = \arctan\left(\frac{O_{4,y}^{\Sigma_2}}{O_{4,x}^{\Sigma_2}}\right)$$

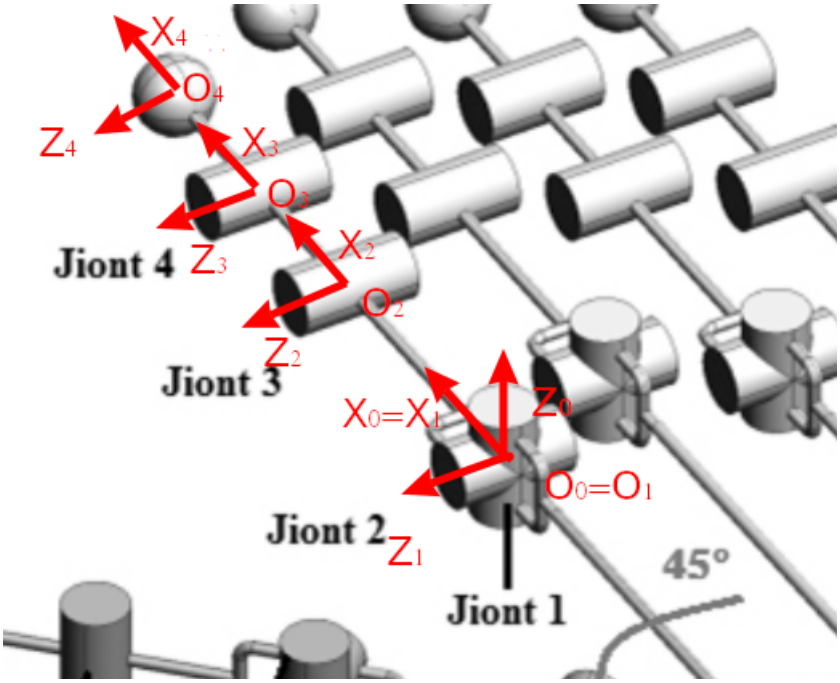
Για την $q5$ μπορούμε να συνεχίσουμε την ίδια διαδικασία αλλά εφόσον υπάρχει μόνο μία στροφή ενός βαθμού ελευθερία, η γωνία μπορεί να υπολογιστεί ως γωνία μεταξύ των διανυσμάτων (ως προς οποιοδήποτε σύστημα αξόνων) των δύο συνδέσμων.

$$q5 = angle(O_5^{leap} - O_4^{leap}, O_4^{leap} - O_2^{leap})$$

όπου,

$$angle(u, v) = \frac{u \cdot v}{|u| \cdot |v|}$$

Συνεχίζουμε με την ανάλυση του δείκτη και του μέσου. Παραθέτουμε μόνο την ανάλυση του ενός δακτύλου, εφόσον η ανάλυση είναι η ίδια για τα δύο δάκτυλα.



Σχήμα 5.9: *Συστήματα αρθρώσεων δείκτη με βάση την μέθοδο DH*

Ο Denavit-Hartenberg πίνακας βάση του Σχήματος 5.9 είναι ο ακόλουθος.

Πίνακας 5.2: *DH Matrix of Finger*

joint i	a_i	α_i	d_i	ϑ_i
1	0	-90°	0	q_1
2	l_1	0°	0	q_2
3	l_2	0°	0	q_3
4	l_3	0°	0	q_4

$$T_0^1 = \begin{bmatrix} c1 & 0 & -s1 & 0 \\ s1 & 0 & c1 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$T_1^2 = \begin{bmatrix} c2 & -s2 & 0 & c2l1 \\ s2 & -c2 & 0 & s2l1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$T_2^3 = \begin{bmatrix} c3 & -s3 & 0 & c3l2 \\ s3 & -c3 & 0 & s3l2 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$T_3^4 = \begin{bmatrix} c4 & -s4 & 0 & c4l3 \\ s4 & -c4 & 0 & s4l3 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Συνεπώς υπολογίζουμε τον:

$$T_0^2 = \begin{bmatrix} c1c2 & -c1c2 & -s1 & c1c2l1 \\ s1c2 & -s1s2 & c1 & s1c2l1 \\ -s2 & -c2 & 0 & -s2l1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

και από το διάνυσμα

$$O_2^{\Sigma_0} = (T_{leap}^{finger})^{-1} O_2^{leap}$$

υπολογίζουμε τις γωνίες $q2$ και $q3$:

$$q2 = \arcsin\left(-\frac{O_{2,z}^{\Sigma_0}}{l1}\right)$$

$$q1 = \arctan\left(\frac{O_{2,y}^{\Sigma_0}}{O_{2,x}^{\Sigma_0}}\right)$$

Οι γωνίες $q3$ και $q4$ μπορούν να υπολογιστούν ως γωνίες μεταξύ των διανυσμάτων (ως προς οποιοδήποτε σύστημα αξόνων) των αντίστοιχων συνδέσμων.

$$q3 = angle(O_3^{leap} - O_2^{leap}, O_2^{leap} - O_0^{leap})$$

$$q4 = angle(O_4^{leap} - O_3^{leap}, O_3^{leap} - O_2^{leap})$$

όπου,

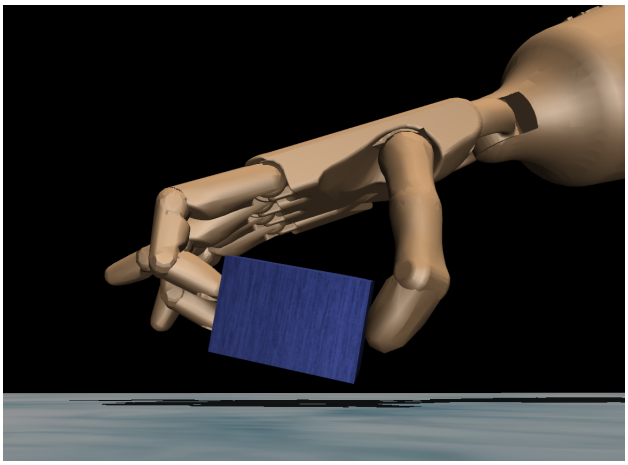
$$angle(u, v) = \frac{u \cdot v}{|u| \cdot |v|}$$

5.4 Συλλογή Δεδομένων Επίδειξης

Έχοντας υπολογίσει όλες τις γωνίες των δακτύλων που χρησιμοποιούμε, βάσει των δεδομένων που παρέχει η Leap Motion, μπορούμε δίνοντας τις τιμές αυτές ως είσοδο στους ελεγκτές θέσης γωνίας να δημιουργήσουμε ένα εικονικό περιβάλλον όπου είναι δυνατός ο online χειρισμός του ρομπότ από τον χρήστη.

Είναι σκόπιμο στο σημείο αυτό να επισημάνουμε ότι, λόγω της λειτουργίας του αισθητήρα με υπέρυθρες, το εύρος των δυνατών κινήσεων περιορίζεται στις διατάξεις του χεριού, για τις οποίες δεν υπάρχει ¨σύγκρουση¨ των διανυσμάτων που καταγράφονται πάνω στο επίπεδο του αισθητήρα, δηλαδή οι προβολές των οστών πάνω στο επίπεδο του αισθητήρα δεν θα πρέπει να τέμνονται. Πρακτικά, πολύ κλειστές λαβές και κινήσεις με διασταυρώσεις των δαχτύλων δεν μπορούν να καταγραφούν.

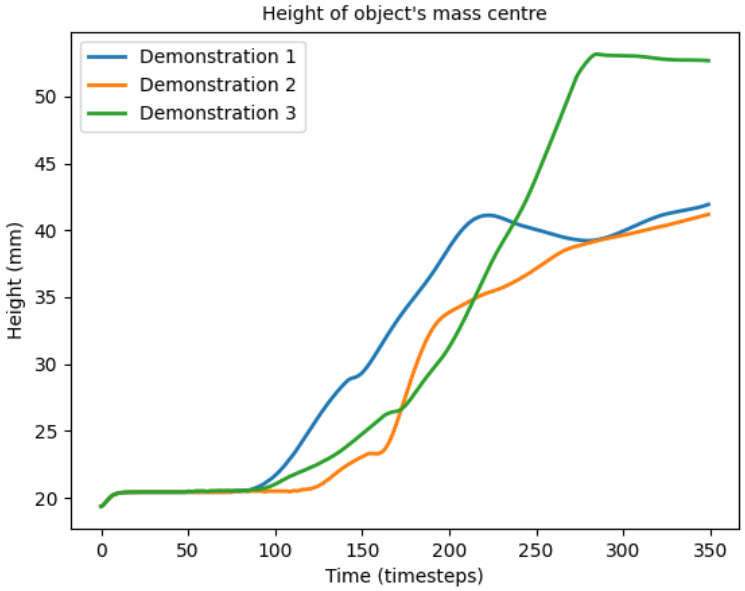
Σε αυτή την εργασία ασχολούμαστε με in hand χειρισμό ενός αντικειμένου, που σημαίνει ότι χρησιμοποιούμε μόνο τα δάχτυλα και όχι τον καρπό ή κάποιον βραχίωνα. Για τον λόγο αυτό ακινητοποιούμε την βάση του χεριού πάνω από το αντικείμενο το οποίο θέλουμε να χειριστούμε. Χρησιμοποιούμε στα πειράματα ως αντικείμενο έναν κύβο διαστάσεων [4cm,6cm,4cm], με την μεγάλη πλευρά των 6cm να είναι αυτή που βρίσκεται ανάμεσα στον αντίχειρα και τα άλλα δύο δάκτυλα. Όπως έχουμε αναφέρει, θέλουμε να καταγράψουμε δεδομένα επίδειξης στα οποία ανυψώνουμε οριζόντια ένα αντικείμενο με in hand λαβή. Χρησιμοποιώντας τον αντίχειρα, τον δείκτη και τον μέσο, σηκώνουμε το αντικείμενο όσο γίνεται σε οριζόντια θέση, τοποθετώντας τον αντίχειρα στην μία πλευρά του κύβου και τα άλλα δύο δάχτυλα στην αντίθετη πλευρά. Επιπλέον, ελέγχουμε την δύναμη που ασκούμε οπτικά, βλέποντας σε πραγματικό χρόνο τις τιμές των αισθητήρων δύναμης του ρομπότ.



Σχήμα 5.10: *In-hand λαβή του αντικειμένου - Δεδομένα επίδειξης*

Καταγράφουμε τα δεδομένα από 3 διαφορετικές τέτοιες λήψεις του κύβου. Συγκεκριμένα, καταγράφουμε κάθε χρονική στιγμή το joint space των τριών δακτύλων του ρομπότ που χρησιμοποιούμε, την θέση και τον προσανατολισμό του αντικειμένου, τις τιμές των αι-

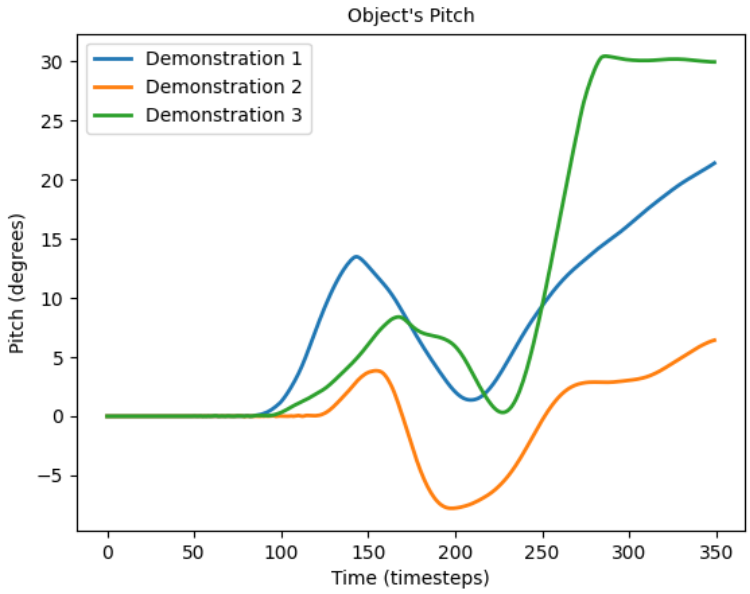
σθητήρων δύναμης και την δράσεις (είσοδοι στου ελεγκτές θέσης), δηλαδή το αντίστοιχο joint space του ανθρώπινου χεριού.



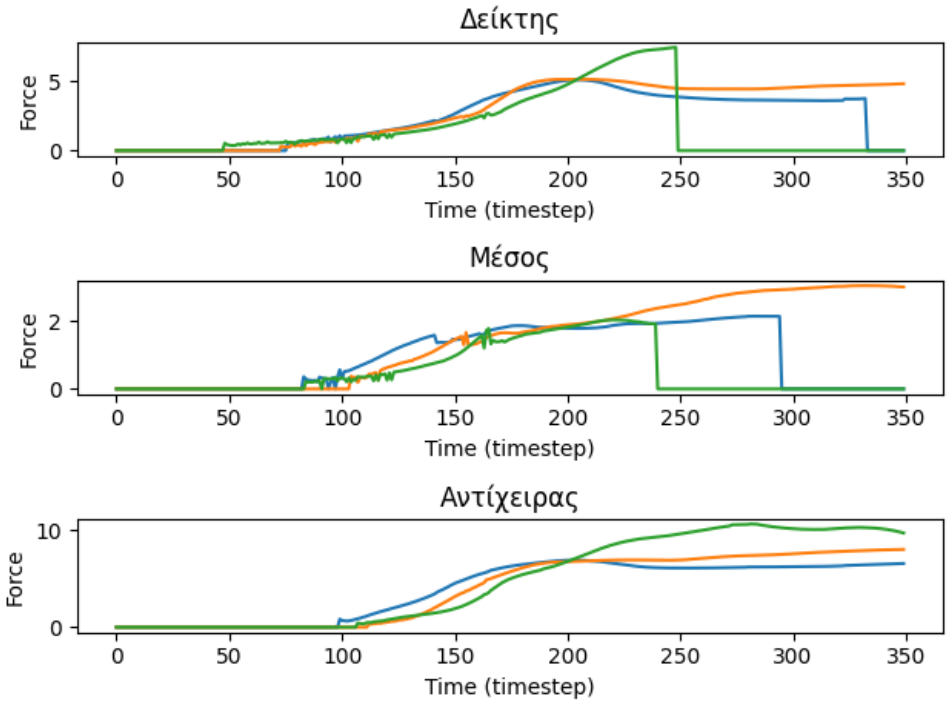
Σχήμα 5.11: *Ύψος αντικειμένου στα δεδομένα επίδειξης*

Με τα δεδομένα επίδειξης, καταφέρνουμε να σηκώσουμε το αντικείμενο με μία in hand λαβή (Σχήματα 5.10, 5.11). Ωστόσο, τα δεδομένα επίδειξης δεν είναι αυτό που ακριβώς επιθυμούμε, καθώς σε αυτά το ρομπότ σηκώνει το αντικείμενο στρέφοντας το ταυτόχρονα, χωρίς να διατηρεί τον οριζόντιο προσανατολισμό του. Από το Σχήμα 5.12 φαίνεται η αδυναμία να σηκώσουμε το αντικείμενο σταθερά οριζόντια. Επιπλέον, παρατηρώντας τις τροχιές της δύναμης (Σχήμα 5.13) που ασκείται στο αντικείμενο βλέπουμε ότι το πρώτο δάχτυλο του δείκτη ασκεί μεγαλύτερη δύναμη από το δεύτερο, ενώ υπάρχουν σημεία που η δύναμη φαίνεται μηδενική. Αυτό στην συγκεκριμένη περίπτωση δεν συμβαίνει επειδή χάνεται η επαφή, αλλά επειδή ο αντίστοιχος αισθητήρας δύναμης δεν καλύπτει το συγκεκριμένο σημείο του ακροδαχτύλου (μύτη) που γίνεται η επαφή.

Με άλλα λόγια, τα δεδομένα επίδειξης δεν είναι ακριβώς η επιθυμητή συμπεριφορά, αλλά μία συμπεριφορά κοντά στην επιθυμητή. Κάτι τέτοιο έχει ερευνητικό ενδιαφέρον, καθώς θέλουμε το σύστημα μας να μπορεί να μεταφερθεί σε ένα πραγματικό ρομπότ και με χρήση δεδομένων επίδειξης σε πραγματικό αντικείμενο. Σε μία τέτοια περίπτωση, το αντικείμενο μεν θα χειρίζεται σωστά από τον άνθρωπο, αλλά τα δεδομένα του ανθρώπινου χεριού αναπόφευκτα θα εμφανίζουν απόκλιση από το ρομπότ. Και στις δύο περιπτώσεις έχουμε δεδομένα που προσδίδουν μία αρχική γνώση για την εργασία την οποία θέλουμε να πραγματοποιήσουμε, και την οποία θέλουμε να βελτιστοποιήσουμε στην συνέχεια.



Σχήμα 5.12: *Pitch αντικειμένου στα δεδομένα επίδειξης*



Σχήμα 5.13: *Δυνάμεις επαφής των τριών δακτύλων στα δεδομένα επίδειξης*

Κεφάλαιο 6

Πειραματικά Αποτελέσματα

Στο κεφάλαιο αυτό παρουσιάζουμε τα πειραματικά αποτελέσματα της μεθοδολογίας που περιγράψαμε στο κεφάλαιο 4. Όπως περιγράψαμε στο Κεφάλαιο 5 της συλλογής των δεδομένων επίδειξης, τα πειράματα αφορούν την εσωτερική λαβή (in-hand grasp) και τον χειρισμό ενός ορθογώνιου παραλληλεπίπεδου με 3 δάκτυλα του ανθρωπομορφικού χεριού ADROIT στο περιβάλλον προσομοίωσης Mujoco.

Στο σημείο αυτό παρουσιάζουμε τις υπερπαραμέτρους που χρησιμοποιήθηκαν. Για την προσέγγιση της πολιτικής $\pi_\theta(a|s)$ και της συνάρτησης αξίας $V_w(s)$ χρησιμοποιήσαμε δύο νευρωνικά δίκτυα με 2 κρυφά layers των 32 νευρώνων στο καθένα. Για τις υπερπαραμέτρους του υπολογισμού της πλεονεκτικής συνάρτησης A_t και της απόδοσης G_t χρησιμοποιήθηκαν οι τιμές $\gamma = 0.995$ και $\lambda = 0.97$. Επίσης για τον υπολογισμό της Natural Policy Gradient χρησιμοποιήθηκαν 10 επαναλήψεις με την τεχνική Conjugate Policy Gradient και ως όριο της $KL - Divergence$ συνθήκης η τιμή $\delta = 0.05$. Ως ρυθμός μάθησης στην ανανέωση της συνάρτησης αξίας στην ενισχυτική μάθηση και της πολιτικής κατά την προεκπαίδευση με επιβλεπόμενη μάθηση (Behavior Cloning) χρησιμοποιήθηκε η τιμή 0.001. Τέλος, σε κάθε εποχή/επανάληψη της ενισχυτικής μάθησης έγινε συλλογή από 200 τροχιές (rollouts) και στην συνέχεια ανανέωση της πολιτικής και της συνάρτησης αξίας βάσει της μεθοδολογίας που παρουσιάστηκε στο κεφάλαιο 4.

Οι παραπάνω τιμές έχουν βρεθεί βιβλιογραφικά στις εργασίες που έχουμε αναφέρει ως ευρετικές τιμές (heuristics) που δουλεύουν ικανοποιητικά για τα περισσότερα προβλήματα. Πραγματοποιήθηκε, ωστόσο, μία μικρή προσαρμογή ώστε να καταλήξουμε σε αυτές τις τελικές τιμές που χρησιμοποιήθηκαν. Κρατάμε τις παραπάνω υπερπαραμέτρους σταθερές στα πειράματα που ακολουθούν, προκειμένου να είναι δυνατή η σύγκριση των μεγεθών.

Τα πειράματα που εκτελέστηκαν και παρουσιάζονται παρακάτω ακολουθούν την ακόλουθη λογική. Αρχικά, ξεκινάμε από την απλούστερη περίπτωση. Θέλουμε το ρομπότ να μάθει απλώς να σηκώνει οριζόντια το αντικείμενο χωρίς να ακολουθεί κάποια τροχιά και χωρίς να έχει πρότερη γνώση από δεδομένα επίδειξης. Στην συνέχεια προσθέτουμε την γνώση από δεδομένα επίδειξης και επιπλέον ερευνάμε πως επιδρούν οι αισθητήρες δύναμης στην περίπτωση αυτή και πραγματοποιούμε έλεγχο της δύναμης που ασκείται. Στην συνέχεια προχωράμε στην δικιά μας μεθοδολογία με την παρακολούθηση μίας τροχιάς ύψους του αντικειμένου και βλέπουμε πώς επιδρούν οι αισθητήρες δύναμης και η εισαγωγή μίας φάσης στον χώρο κατάστασης στο στάδιο πριν την άρση του αντικειμένου. Τέλος, βλέπουμε ξανά την περίπτωση περιορισμού της δύναμης που ασκείται στο αντικείμενο για την περίπτωση

παρακολούθησης τροχιάς.

6.1 Λαβή (Grasp) σε προκαθορισμένο ύψος

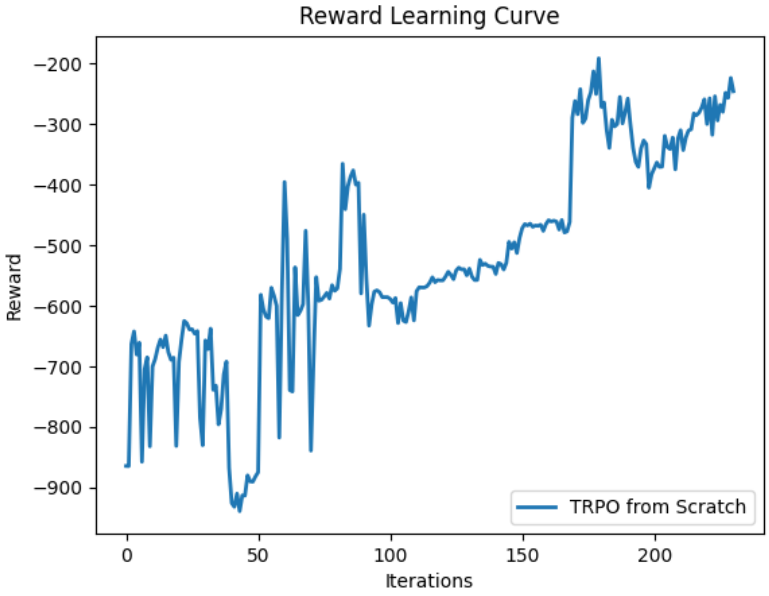
6.1.1 Ενισχυτική μάθηση χωρίς πρότερη γνώση

Σαν πρώτο πείραμα, θέλουμε να δούμε ποια συμπεριφορά βρίσκει το σύστημα μας χρησιμοποιώντας ενισχυτική μάθηση με τυχαία αρχικοποιημένη πολιτική, απλώς για την ανύψωση του αντικειμένου σε κάποιο ύψος-στόχο g, διατηρώντας οριζόντιο το αντικείμενο δηλαδή με ένα $pitch = 0$. Χρησιμοποιούμε σε αυτό το πείραμα ως χώρο κατάστασης μόνο τις αρθρώσεις του ρομπότ (αντίχειρα, δείκτη και μέσου), την θέση και τον προσανατολισμό του αντικειμένου και διαμορφώνουμε την επιβράβευση ως εξής:

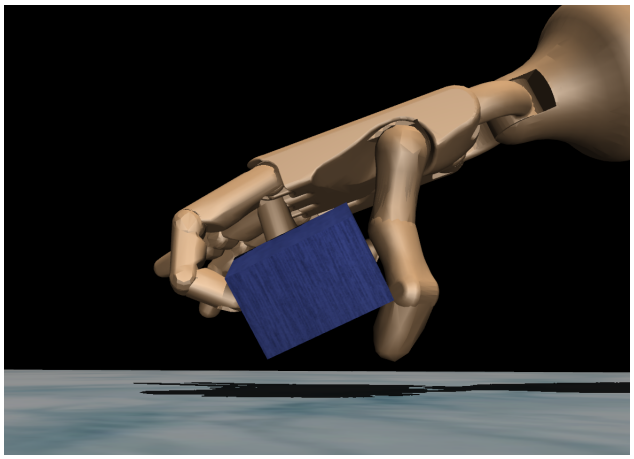
$$r = -w_1|h - g| - w_2|pitch|$$

Θέτουμε για το πείραμα αυτό ως επιθυμητό ύψος του κέντρου βάρους $g = 50mm$. Χρησιμοποιούμε ως βάρη της παραπάνω επιβράβευσης $w_1 = 0.1$ και $w2 = \frac{1}{70}$ μετρώντας το ύψος h σε χιλιοστά και την γωνία $pitch$ σε μοίρες. Παρουσιάζουμε στο Σχήμα 6.1 την καμπύλη μάθησης, η οποία αφορά το άθροισμα των επιβραβεύσεων σε ένα ενδεικτικό επισόδειο αξιολόγησης ανά επανάληψη. Η κάθε επανάληψη αφορά συλλογή 200 τροχιών και ανανέωση των νευρωνικών δικτύων πολιτικής και συνάρτησης αξίας όπως περιγράφηκε στο Κεφάλαιο 4. Μετά την ολοκλήρωση της φάσης εκπαίδευσης (200 επαναλήψεων) εκτελέσθηκε πείραμα για την αξιολόγηση της τελικής επίδοσης της μεθόδου για μία ενδεικτική εκτέλεση ενός επισοδείου ανύψωσης του αντικειμένου από το ρομποτικό χέρι. Σημειώνουμε ότι σε όλη την διαδικασία αξιολόγησης των πειραμάτων χρησιμοποιούμε ντετερμινιστικά την πολιτική, δηλαδή μόνο την έξοδο του αντίστοιχου νευρωνικού δικτύου, χωρίς την μήτρα συνδιακύμανσης. Στο σχήμα 6.2 παρουσιάζεται στιγμιότυπο της τελικής θέσης χεριού και αντικειμένου, ενώ στα Σχήματα 6.3 και 6.4 παρουσιάζονται οι τροχιές ύψους και γωνίας ανύψωσης (ως TRPO from Scratch αναφερόμαστε στην εκτέλεση του αλγορίθμου Trust Region Policy Optimization χωρίς πρότερη γνώση από δεδομένα επίδειξης).

Στις πρώτες 100 επαναλήψεις όπως φαίνεται από την καμπύλη μάθησης δεν υπάρχει ιδιαίτερη βελτίωση της πολιτικής. Στην συνέχεια η επιβράβευση αυξάνεται και το σύστημα μαθαίνει μία μέθοδο να σηκώσει το αντικείμενο με μία πρωτότυπη λαβή. Σταματάμε την εκπαίδευση μετά από 200 επαναλήψεις, καθώς παρατηρήσαμε ότι στο σημείο αυτό υπάρχει σταθεροποίηση της απόδοσης και ελέγχουμε τα αποτελέσματα. Το σύστημα έχει μάθει να πραγματοποιεί μία λαβή στο αντικείμενο και να το σηκώνει σε κάποιο ύψος. Η λαβή αυτή, ωστόσο, δεν είναι μία λαβή με επαφή των ακροδακτύλων στο αντικείμενο. Ο δείκτης του χεριού σπρώχνει το αντικείμενο σε ένα σημείο του αντίχειρα, εγκλωβίζοντας το σε αυτή την θέση. Αυτή η power grasp λαβή ουσιαστικά δεν επιτρέπει στο ρομποτ να ελέγξει το ύψους του αντικειμένου αλλά το σηκώνει απλώς σε κάποιο ύψος που σχετίζεται με το σημείο εγκλωβισμού του αντικειμένου ανάμεσα στα δάκτυλα (Σχήμα 6.2). Το τελικού ύψος του κέντρου βάρους του αντικειμένου, όπως βλέπουμε από την καμπύλη του ύψους (Σχήμα 6.3) δεν είναι το επιθυμητό (50mm), αλλά περίπου 60mm.

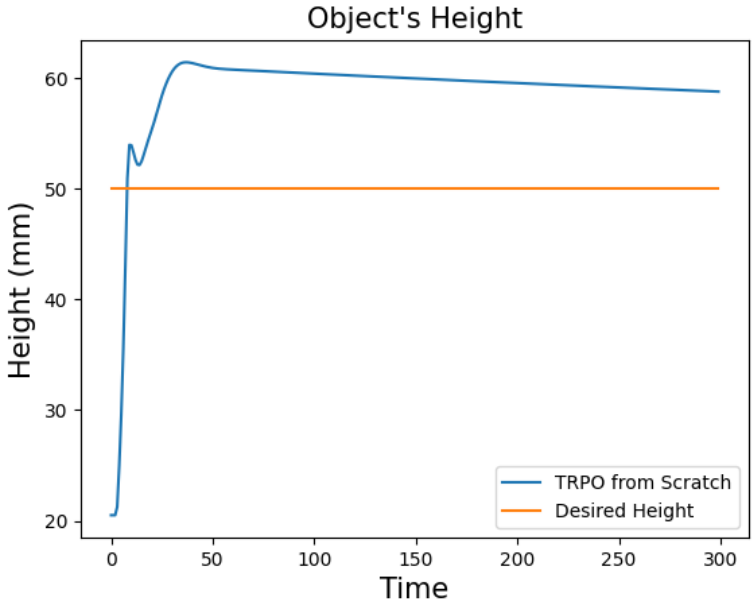


Σχήμα 6.1: *Καμπύλη μάθησης - Ενισχυτική Μάθηση χωρίς πρότερη γνώση*

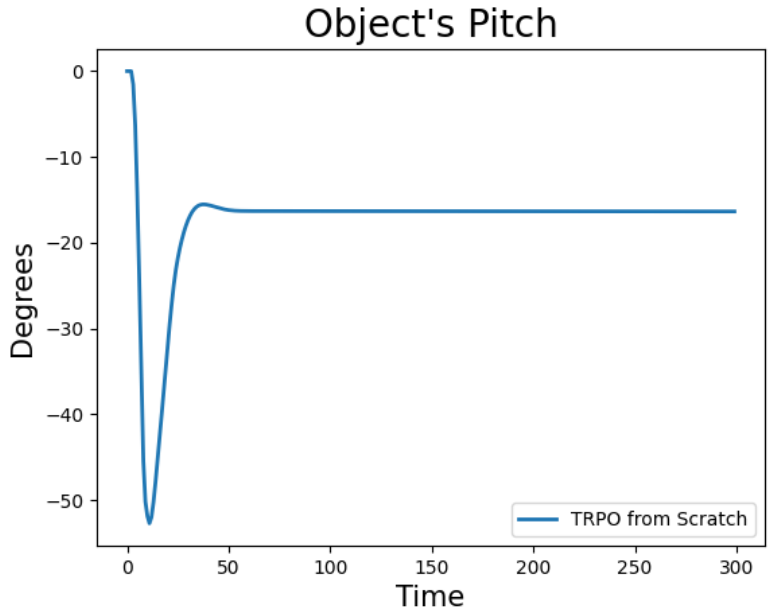


Σχήμα 6.2: *Τελική θέση grasp - Ενισχυτική μάθηση χωρίς πρότερη γνώση*

Επιπλέον, παρατηρούμε ότι ο προσανατολισμός του αντικειμένου δεν παραμένει οριζόντιος (Σχήμα 6.4). Αυτός ουσιαστικά καθορίζεται από την αντίστοιχη παράμετρο w_2 της επιβράβευσης. Αυξάνοντάς την παράμετρο αυτή θα μπορούσαμε να δώσουμε παραπάνω βάρος στο μέγεθος αυτό, ώστε να βοηθήσουμε το αντικείμενο να παραμένει όσο γίνεται οριζόντιο κατά την κίνηση. Ωστόσο, παρατηρήσαμε ότι στην περίπτωση αυτή χωρίς χρήση δεδομένων επίδειξης δεν ήταν δυνατόν να αυξήσουμε το βάρος w_2. Αυξάνοντας το είδαμε ότι το χέρι επιλέγει να μην σηκώνει καν το αντικείμενο. Αυτό συμβαίνει επειδή στην αρχή το αντικείμενο που βρίσκεται στο έδαφος είναι ήδη οριζόντιο. Συνεπώς, με ένα μεγάλο βάρος w_2 αυξάνουμε την ποινή για μη οριζόντιες θέσεις του αντικειμένου. Επομένως, το χέρι κατά την εξερεύνηση κουνώντας το αντικείμενο ελάχιστα χωρίς να το σηκώσει παίρνει μία μεγάλη ποινή και αποτρέπει τις ενέργειες που συνδέονται με αλληλεπίδραση χεριού-αντικειμένου. Αυτή ουσιαστικά είναι μία υποβέλτιστη λύση του συστήματος, καθώς δεν βρίσκει ποτέ λύσεις



Σχήμα 6.3: *Τροχία ύψους - Ενισχυτική Μάθηση χωρίς πρότερη γνώση*



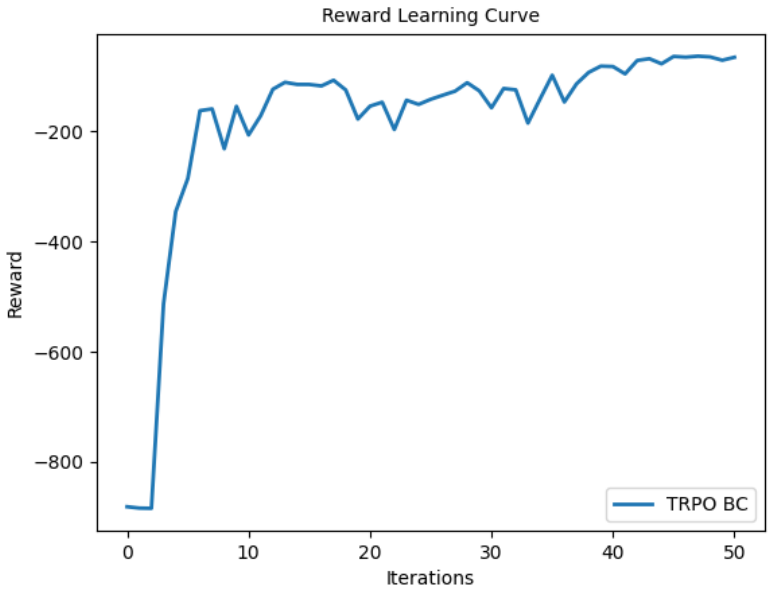
Σχήμα 6.4: *Τροχία pitch - Ενισχυτική Μάθηση χωρίς πρότερη γνώση*

που σχετίζονται με το grasp του αντικειμένου σε μεγάλο ύψος το οποίο θα έδινε καλύτερη επιβράβευση.

Βάσει λοιπόν των αποτελεσμάτων, βλέπουμε ότι η αποτυχία είναι πολλών επιπέδων. Η μάθηση είναι πολύ αργή, η λαβή δεν είναι η επιθυμητή, το αντικείμενο δεν ανυψώνεται στο επιθυμητό ύψος, αλλά σε κάποιο άλλο ύψος σχετιζόμενο με την λαβή που πραγματοποιεί, ενώ ουσιαστικά η βελτιστοποίηση ως προς την οριζόντια θέση του αντικειμένου δεν λαμβάνεται καν υπόψιν.

6.1.2 Ενισχυτική μάθηση και προεκπαίδευση με δεδομένα επίδειξης

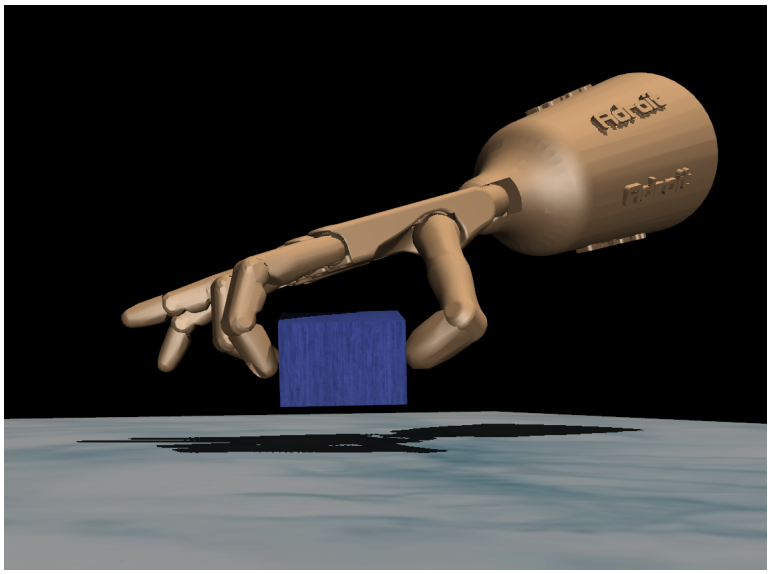
Στο επόμενο πείραμα κρατάμε τον χώρο καταστάσεων ως έχει. Χρησιμοποιούμε τα δεδομένα επίδειξης ως αρχικοποίηση της πολιτικής σύμφωνα με τον αλγόριθμο Behavior Cloning, όπως περιγράφηκε στην παράγραφο 4.5. Τα δεδομένα επίδειξης δίνουν μία καλή αρχικοποίηση στο σύστημα μας, η οποία ακολουθείται από ενισχυτική μάθηση. Επομένως, μπορούμε τώρα να αυξήσουμε το βάρος w_2 της επιβράβευσης, εφόσον λόγω των δεδομένων επίδειξης το σύστημα δεν μένει σε μία υποβέλτιστη κατάσταση που επιλέγει να μην σηκώσει καν το αντικείμενο όπως περιγράψαμε προηγουμένως. Χρησιμοποιούμε μία τιμή $\frac{1}{20}$, δίνοντας σημαντικό βάρος και στο να μένει το αντικείμενο οριζόντιο. Παρουσιάζουμε στην συνέχεια στα σχήματα 6.5 έως 6.9, τα αποτελέσματα των δύο καλύτερων περιπτώσεων που επιτύχαμε με και χωρίς δεδομένα επίδειξης, όπου ως TRPO from Scratch αναφερόμαστε στην μάθηση χωρίς πρότερη γνώση και ως TRPO BC στην μάθηση με προεκπαίδευση από δεδομένα επίδειξης με τον αλγόριθμο Behavior Cloning. Σημειώνουμε εδώ, ότι λόγω του ότι η επιβράβευση είναι διαφορετική συνάρτηση για τις δύο αυτές περιπτώσεις δεν έχει νόημα να συγκρίνουμε τις καμπύλες μάθησης ποσοτικά, επομένως στο σχήμα 6.5 παρουσιάζουμε μόνο την περίπτωση με χρήση δεδομένων επίδειξης, ενώ η αντίστοιχη χωρίς την πρότερη αυτή γνώση έχει παρουσιαστεί στο σχήμα 6.1.



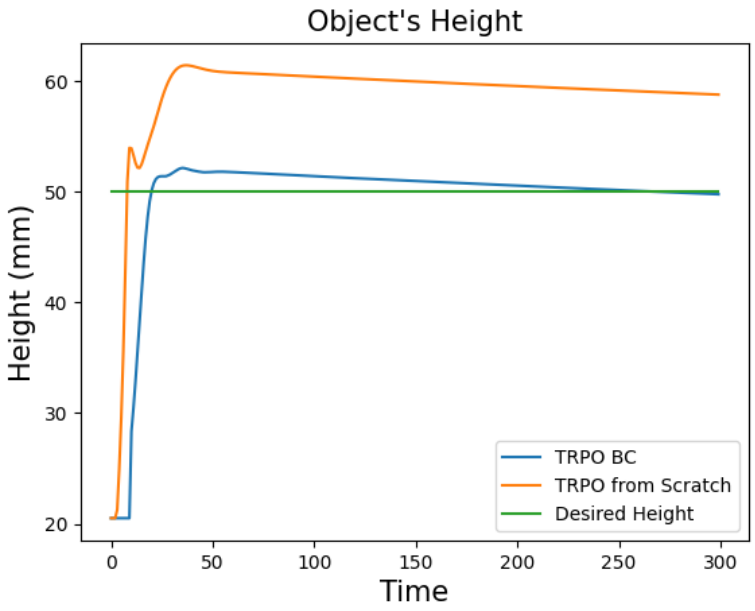
Σχήμα 6.5: *Καμπύλη μάθησης - Προεκπαίδευση με δεδομένα επίδειξης*

Όπως είδαμε στο προηγούμενο πείραμα (παράγραφος 6.1.1) χωρίς δεδομένα επίδειξης χρειάζεται 100-200 επαναλήψεις ώστε να πετύχει μία λαβή του αντικειμένου η οποία ούτε είναι η επιθυμητή λαβή επαφής των ακροδακτύλων, ούτε πετυχαίνει επιθυμητό ύψος και προσανατολισμό. Από την άλλη, με δεδομένα επίδειξης βλέπουμε ότι έχουμε πολύ γρηγορότερη σύγκλιση σε περίπου 10 με 20 επαναλήψεις (Σχήμα 6.5).

Με χρήση των δεδομένων επίδειξης παρατηρούμε ότι επιτυγχάνεται η επιθυμητή λαβή του αντικειμένου με επαφή των άκρων των δακτύλων ακόμα και χωρίς αισθητήρες δύναμης. Σχετικά με τις δυνάμεις, βλέπουμε ότι ακολουθείται η συμπεριφορά των δεδομένων επίδει-



Σχήμα 6.6: *Τελική θέση λαβής - Προεκπαίδευση με δεδομένα επίδειξης*

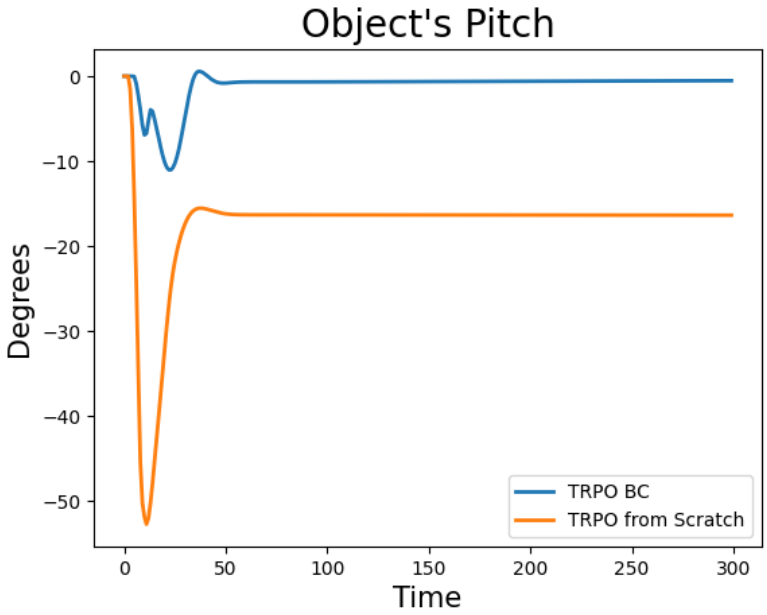


Σχήμα 6.7: *Σύγκριση τροχιάς ύψους αντικειμένου με και χωρίς δεδομένα επίδειξης*

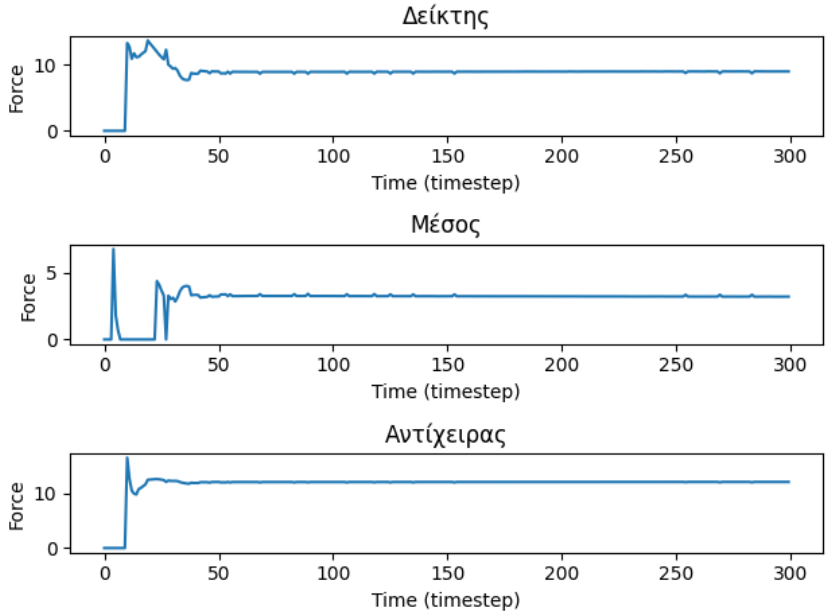
ξης να χρησιμοποιείται περισσότερο ο δείκτης από τον μέσο (Σχήμα 6.9). Σημειώνουμε ότι η μηδενική τιμή του μέσου δακτύλου στην περίπτωση αυτή δεν σημαίνει ότι χάνεται η επαφή, αλλά χρησιμοποιεί σε εκείνο σημείο την μύτη του δακτύλου στην οποία δεν φτάνει ο αισθητήρας δύναμης.

Σημειώνουμε, επιπλέον ότι από το πείραμα αυτό φαίνεται πως αυτή η λαβή μπορεί να ελέγξει την κίνηση του αντικειμένου και δεν το εγκλωβίζει απλώς σε μία κατάσταση μέσα στην παλάμη. Επιπλέον, βλέπουμε ότι το αντικείμενο σηκώνεται στο επιθυμητό ύψος το οποίο έχουμε ζητήσει μέσω της επιβράβευσης ενώ το pitch του αντικειμένου είναι μικρότερο και στο τέλος της κίνησης αφού το σηκώσει είναι σχεδόν μηδενικό.

Ωστόσο, βλέπουμε ότι το ρομπότ σηκώνει πολύ γρήγορα, σχεδόν ακαριαία το αντικείμενο.



Σχήμα 6.8: *Σύγκριση τροχιάς pitch αντικειμένου με και χωρίς δεδομένα επίδειξης*



Σχήμα 6.9: *Τροχιές δύναμης επαφής με χρήση δεδομένων επίδειξης*

Αυτό γίνεται διότι η επιβράβευση αφορά έναν τελικό στόχο. Συνεπώς όσο πιο γρήγορα φτάσει το σύστημα τον στόχο τόσο μεγαλύτερη επιβράβευση θα πάρει. Το γεγονός ότι η βελτιστοποίηση γίνεται και ως προς την ταχύτητα επίτευξης της ζητούμενης εργασίας δεν είναι πάντα θεμιτό, καθώς δεν μας επιτρέπει να ελέγξουμε την κίνηση του αντικειμένου μέχρι να φτάσει στον στόχο. Αυτός είναι και το κύριο πρόβλημα στο οποίο εστιάζουμε στην εργασία αυτή και επιχειρούμε να λύσουμε με την παρακολούθηση κάποιας τροχιάς.

Μπορούμε επίσης να παρατηρήσουμε ότι λόγω της μεγάλης ταχύτητας υπάρχει overshoot στα μεγέθη, τόσο στο ύψος το οποίο στην αρχή φτάνει παραπάνω από το επιθυμητό όσο και

στο pitch του αντικειμένου, το οποίο αν και έχει μειωθεί συνεχίζει να φτάνει σε μία τιμή περίπου 10 μοιρών κατά την διάρκεια του grasp. Ουσιαστικά επειδή έχουμε να κάνουμε με έναν τελικό στόχο ύψους, η ποινή (αρνητική επιβράβευση) που αφορά αυτόν τον τελικό στόχο αρχικά είναι πολύ μεγάλη και συνεπώς μέχρι αυτή να γίνει μικρή, δηλαδή το αντικείμενο να σηκωθεί κοντά στο επιθυμητό ύψος, η επιβράβευση που αφορά τα υπόλοιπα μεγέθη φαίνεται να μην λαμβάνεται σημαντικά υπόψιν και έτσι δεν μπορούμε να ρυθμίσουμε τα μεγέθη εύκολα μέχρι να φτάσουμε στον τελικό στόχο.

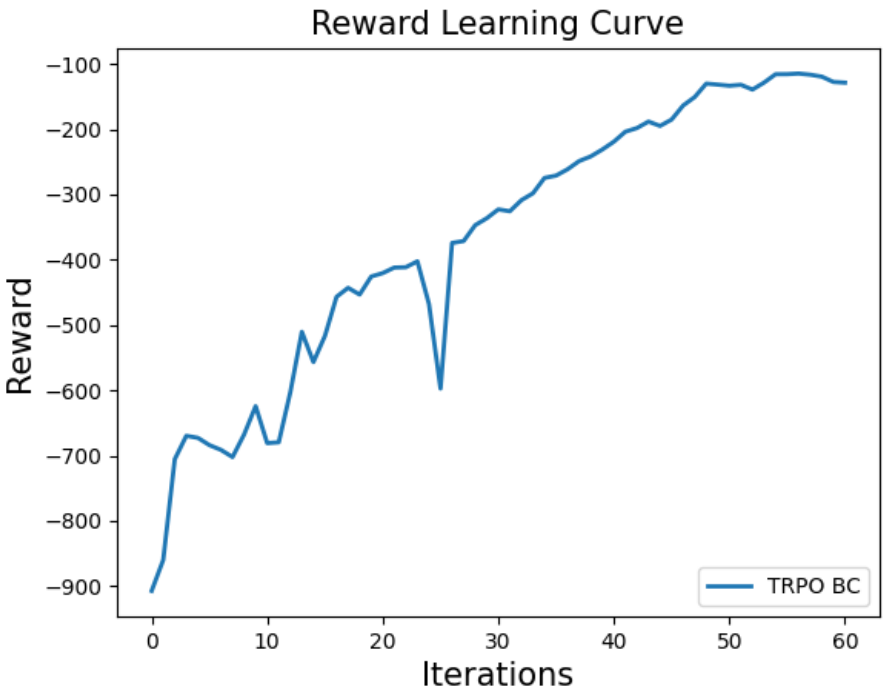
6.1.3 Χρήση αισθητήρων δύναμης

Στο πείραμα αυτό θέλουμε να διαμορφώσουμε με κάποιο επιθυμητό τρόπο την άσκηση δύναμης στο αντικείμενο. Προσθέτουμε λοιπόν στον χώρο καταστάσεις τις τιμές από τους αισθητήρες δύναμης που βρίσκονται στα άκρα των τριών δακτύλων και διαμορφώνουμε την επιβράβευση ως εξής:

$$r = -w_1|h - g| - w_2|pitch| - w_3 r_f$$

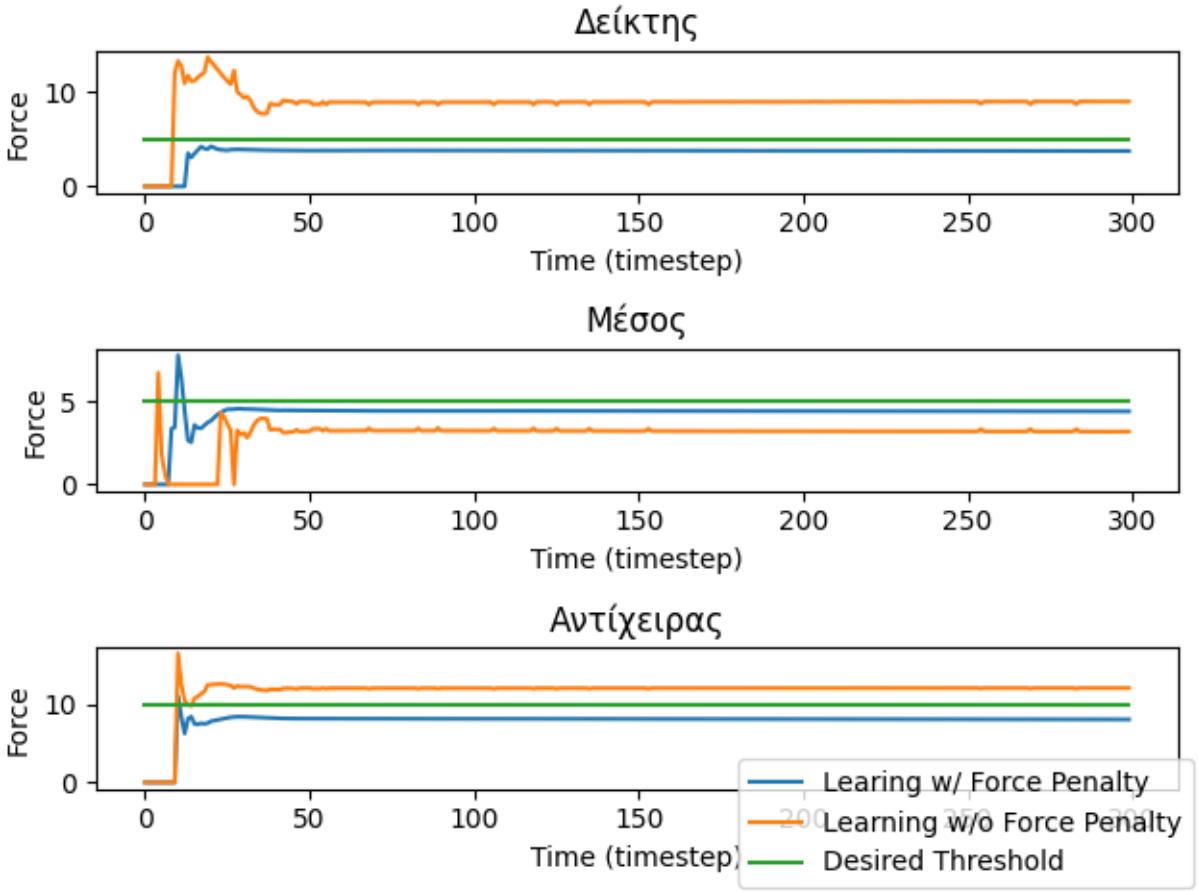
με

$$r_f = \begin{cases} 0 & \text{if } f_{min,i} < f_i < f_{max,i} \text{ for } i = 1,2,3 \\ 1 & \text{if } otherwise. \end{cases}$$



Σχήμα 6.10: *Καμπύλη Μάθησης - Grasp με περιορισμό των δυνάμεων*

Χρησιμοποιούμε τα βάρη w_1 και w_2 όπως στο προηγούμενο πείραμα με τιμές 0.1 και $\frac{1}{20}$ αντίστοιχα για στόχο $g = 50$ και $w_3 = 1.5$ αντιπροσωπεύει την ποινή έξω από τα όρια των επιθυμητών δυνάμεων (thresholds). Έχουμε ένα κάτω όριο ώστε να διατηρείται η επαφή



Σχήμα 6.11: *Σύγκριση δυνάμεων κατά το grasp με και χωρίς έλεγχο της δύναμης*

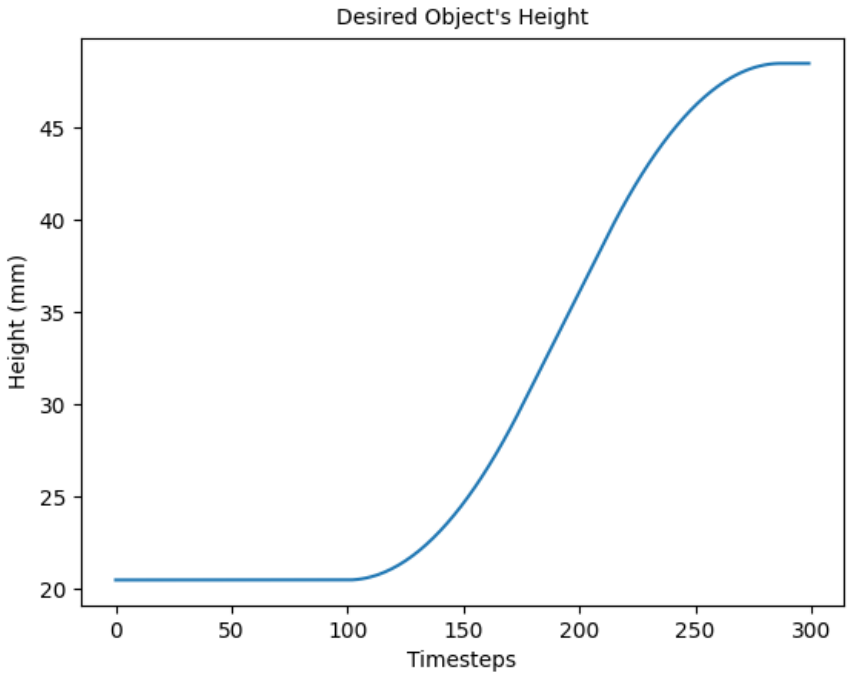
με το αντικείμενο και ένα άνω όριο ώστε να περιορίσουμε την δύναμη που ασκείται εντός επιθυμητών τιμών. Για το κάτω όριο χρησιμοποιούμε μία μικρή τιμή $f_{min,i} = 0.1$, $i = 1, 2, 3$ ώστε να διασφαλίζεται η επαφή. Σχετικά με το άνω όριο, θέλουμε να περιορίσουμε την συνολική δύναμη, η οποία προκύπτει ουσιαστικά από την δύναμη του αντίχειρα και να μοιράσουμε στην αντίθετη πλευρά του αντικειμένου την δύναμη στον δείκτη και στον μέσο. Χρησιμοποιούμε για τον αντίχειρα μία τιμή $f_{max,1} = 10$ και για τον δείκτη και μέσο μία τιμή $f_{max,2} = f_{max,3} = 5$. Στο Σχήμα 6.10 παρουσιάζεται η καμπύλη μάθησης για το πείραμα αυτό και στο Σχήμα 6.11 οι τροχιές δυνάμεων μετά την σύγκλιση του αλγορίθμου με και χωρίς ποινή δύναμης (w/ vs w/o tactiles).

Αρχικά, βλέπουμε ότι σε σχέση με το προηγούμενο πείραμα χρειάζονται περισσότερες επαναλήψεις για να μάθει να διατηρεί τις τιμές των δυνάμεων εντός των επιθυμητών ορίων. Κάτι τέτοιο είναι απολύτως λογικό, δεδομένου αρχικά ότι ζητάμε να βελτιστοποιηθεί μία ακόμη παράμετρος και κατά δεύτερον και σημαντικότερο ότι ζητάμε μία αρκετά διαφορετική συμπεριφορά σε σχέση με τα δεδομένα επίδειξης. Συνεπώς, χρειάζεται περισσότερη εξερεύνηση (exploration) από πλευράς ενισχυτικής μάθησης και επομένως περισσότερες επαναλήψεις. Τέλος παρατηρούμε και σε αυτήν την περίπτωση ότι ενώ τα επιθυμητά όρια τηρούνται εν τέλη, στην αρχή υπάρχει ξανά overshoot, λόγω του ότι υπάρχει ένας τελικός στόχος ύψους.

6.2 Επιδέξια Λαβή (Grasp) με Παρακολούθηση Προκαθορισμένης Τροχιάς Ύψους

Είδαμε στα προηγούμενα πειράματα, ότι καταφέρνουμε να σηκώσουμε το αντικείμενο με την ζητούμενη in hand λαβή σε ένα επιθυμητό ύψος. Ωστόσο, το κύριο μειονέκτημα είναι ότι με τον τρόπο αυτό δεν μπορούμε να ελέγξουμε ούτε τον τρόπο με τον οποίο φτάνει στο ύψος αυτό, αλλά ούτε τα υπόλοιπα μεγέθη τα οποία κάνουν overshoot από τις επιθυμητές τιμές στο διάστημα μέχρι το αντικείμενο να φτάσει στο επιθυμητό ύψος. Συνεχίζουμε, λοιπόν, με την μεθοδολογία που προτείνουμε, η οποία αφορά την παρακολούθηση τροχιάς και επιχειρεί να λύσει τα προβλήματα αυτά. Χρησιμοποιούμε στα επόμενα πειράματα ως επιθυμητή μία πολυωνυμική τροχιά τριών φάσεων με φάση επιτάχυνσης, σταθερής ταχύτητας και επιβράδυνσης. Επιπλέον στα επόμενα πειράματα χρησιμοποιούμε προεκπαίδευση με δεδομένα επίδειξης, καθώς όπως φάνηκε από την παράγραφο 6.1.1 χρειάζονται πάρα πολλές επαναλήψεις όταν χρησιμοποιούμε μάθηση χωρίς πρότερη γνώση και δεν έχει ιδιαίτερο νόημα να επαναλάβουμε κάποιο πείραμα με ενισχυτική μάθηση από το μηδέν ξανά.

Κατά την πολυωνυμική τροχιά που χρησιμοποιούμε, οι φάσεις επιτάχυνσης και επιβράδυνσης διαρκούν 40% του χρόνου τις τροχιάς και η φάση σταθερής ταχύτητας το 20%. Οι παράμετροι της φάσης επιτάχυνσης και επιβράδυνσης επιλέγονται ώστε να υπάρχει συνέχεια στην ταχύτητα. Για να είναι η ζητούμενη τροχιά κοντά στα δεδομένα επίδειξης (παράγραφος 5.4) επιλέγουμε τον συνολικό χρόνο ανύψωσης να διαρκεί 200 timesteps και αφήνουμε 100 timesteps στην αρχή χωρίς να ζητήσουμε να σηκωθεί το αντικείμενο ώστε το ρομπότ να πραγματοποιήσει το κλείσιμο των δακτύλων και να σχηματίσει την λαβή. Επιλέγουμε για την ζητούμενη τροχιά ένα ύψος $4.8cm$. Ονομάζουμε, στην συνέχεια την τροχιά αυτή h_D.



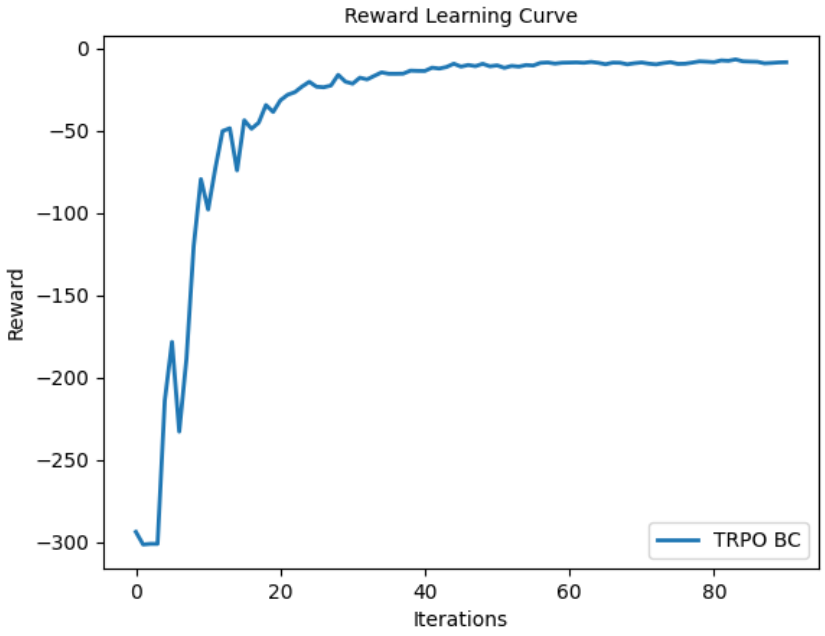
Σχήμα 6.12: *Επιθυμητή τροχιά ύψους*

6.2.1 Χωρίς χρήση Αισθητήρων Δύναμης

Σαν πρώτο πείραμα, θέλουμε να δούμε την συμπεριφορά του συστήματος χωρίς να χρησιμοποιήσουμε τους αισθητήρες δύναμης. Ο χώρος κατάστασης λοιπόν αποτελείται από τα joints του ρομπότ, την θέση και τον προσανατολισμό του αντικειμένου και το επιθυμητό ύψος του αντικειμένου για την επόμενη χρονική στιγμή, όπως περιγράψαμε στην παράγραφο 4.1. Ορίζουμε την επιβράβευση, σύμφωνα με την παράγραφο 4.2 ως:

$$r(t) = -w_1|h(t) - h_D(t)| - w_2|pitch(t)|$$

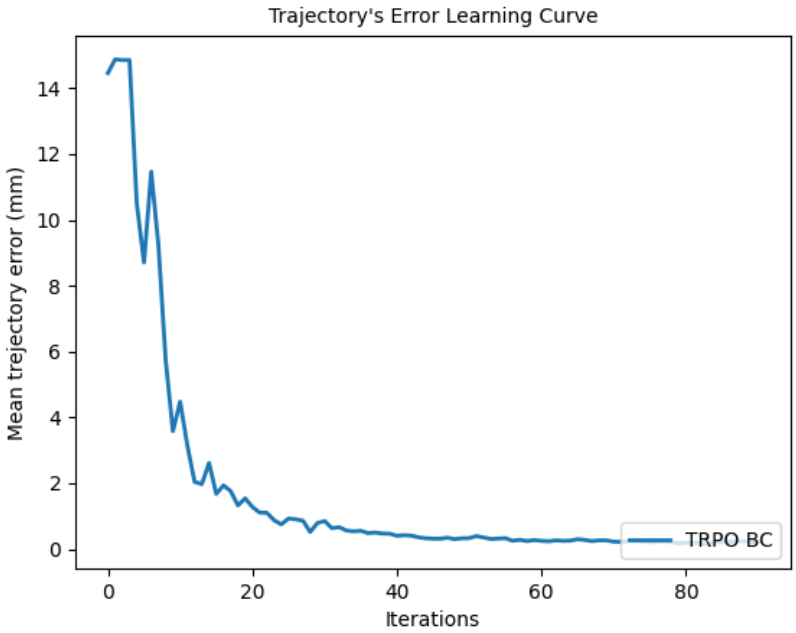
Χρησιμοποιούμε $w_1 = 0.1$ και $w_2 = \frac{1}{70}$. Σημειώνουμε ότι σε αυτήν την περίπτωση η τιμή $\frac{1}{70}$ του w_2 δεν είναι μικρή, ακόμα και αν είναι ίδια με αυτή που χρησιμοποιήθηκε στα πειράματα του grasp χωρίς παρακολούθηση τροχιάς. Ο λόγος είναι ότι με έναν τελικό στόχο δημιουργείται μεταξύ των όρων της επιβράβευσης ένα απροσδιόριστο (για τον σχεδιαστή) trade-off ύψους/pitch που χρειαζόταν αρκετή ρύθμιση των βαρών της επιβράβευσης για να έχουμε ένα καλό αποτέλεσμα. Αντίθετα, τώρα το trade-off αφορά μία χρονική στιγμή και συνεπώς οι παραπάνω τιμές των δύο βαρών της επιβράβευσης μπορούν να μεταφραστούν στο ότι η ποινή ενός χιλιοστού ύψους είναι ισοδύναμη με ποινή επτά μοιρών του pitch. Με αυτόν τον τρόπο η ρύθμιση των παραμέτρων γίνεται εξαιρετικά πιό εύκολη, αφού ορίζεται μία αναλογία από εμάς.



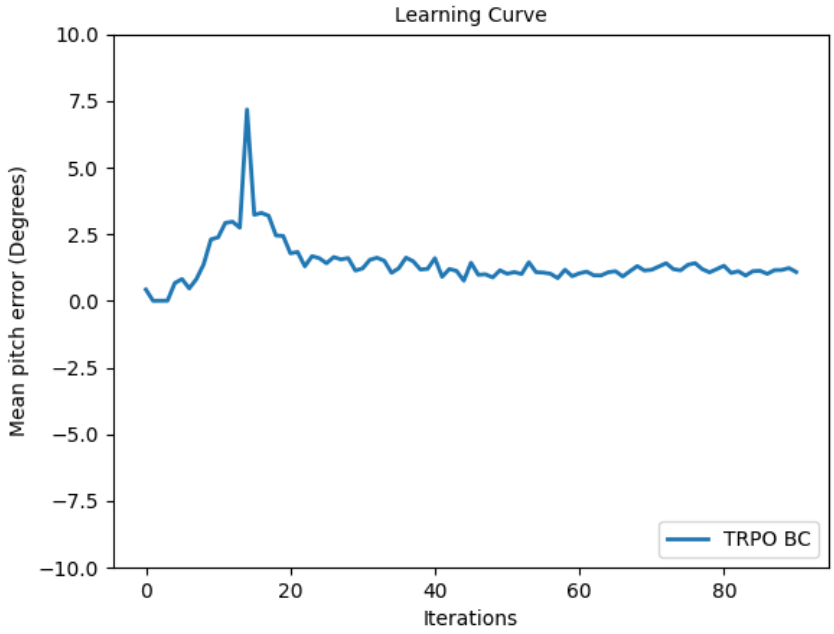
Σχήμα 6.13: *Καμπύλη επιβράβευσης - Παρακολούθηση τροχιάς ύψους χωρίς αισθητήρες δύναμης*

Για την αξιολόγηση της μεθόδου παρουσιάζουμε την καμπύλη μάθησης της επιβράβευσης (Σχήμα 6.13) και τις καμπύλες μάθησης ξεχωριστά για τα μεγέθη που υπάρχουν στην επιβράβευση, δηλαδή το σφάλμα της τροχιάς ύψος (Σχήμα 6.14) και της γωνίας ανύψωσης από το οριζόντιο επίπεδο (Σχήμα 6.15). Παρουσιάζουμε επιπλέον την συμπεριφορά μετά

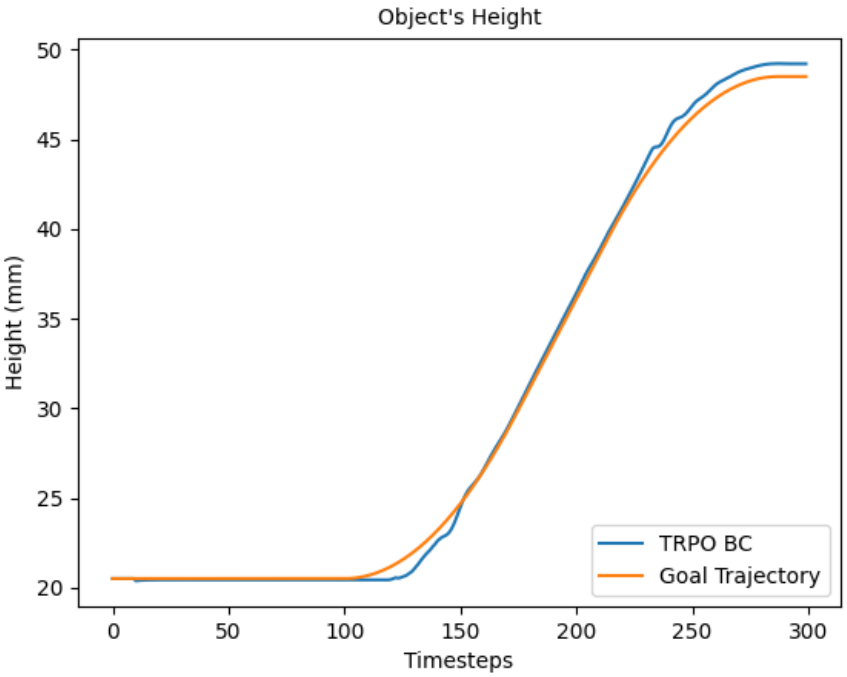
την σύγκλιση των καμπυλών μάθησης (50 επαναλήψεις), δηλαδή τις τροχιές ύψους, γωνίας ανύψωσης και δυνάμεων (Σχήματα 6.16 έως 6.18) που εκτελούνται σε ένα επεισόδιο αξιολόγησης.



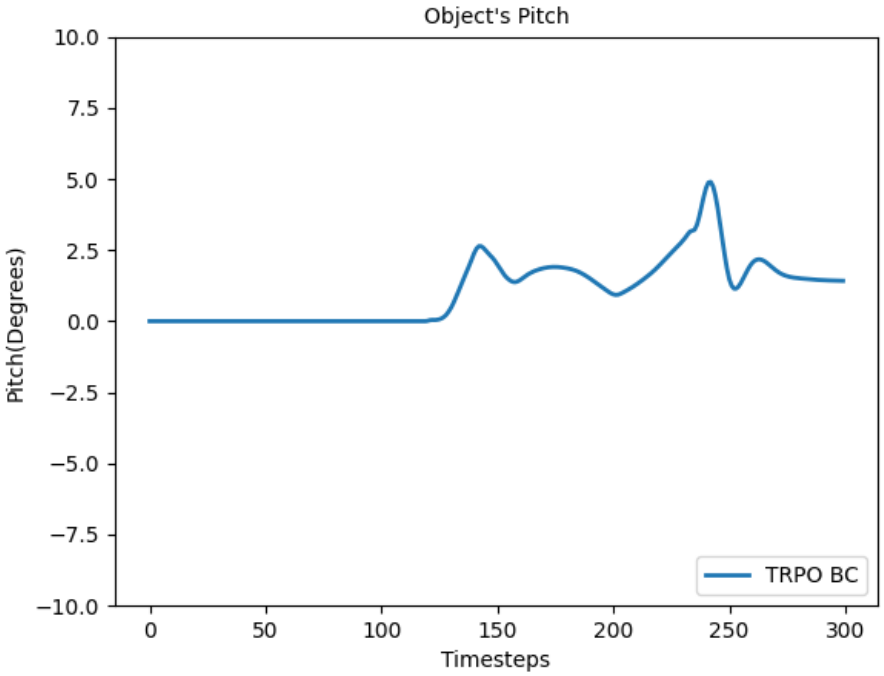
Σχήμα 6.14: *Καμπύλη μάθησης τροχιάς ύψους χωρίς αισθητήρες δύναμης*



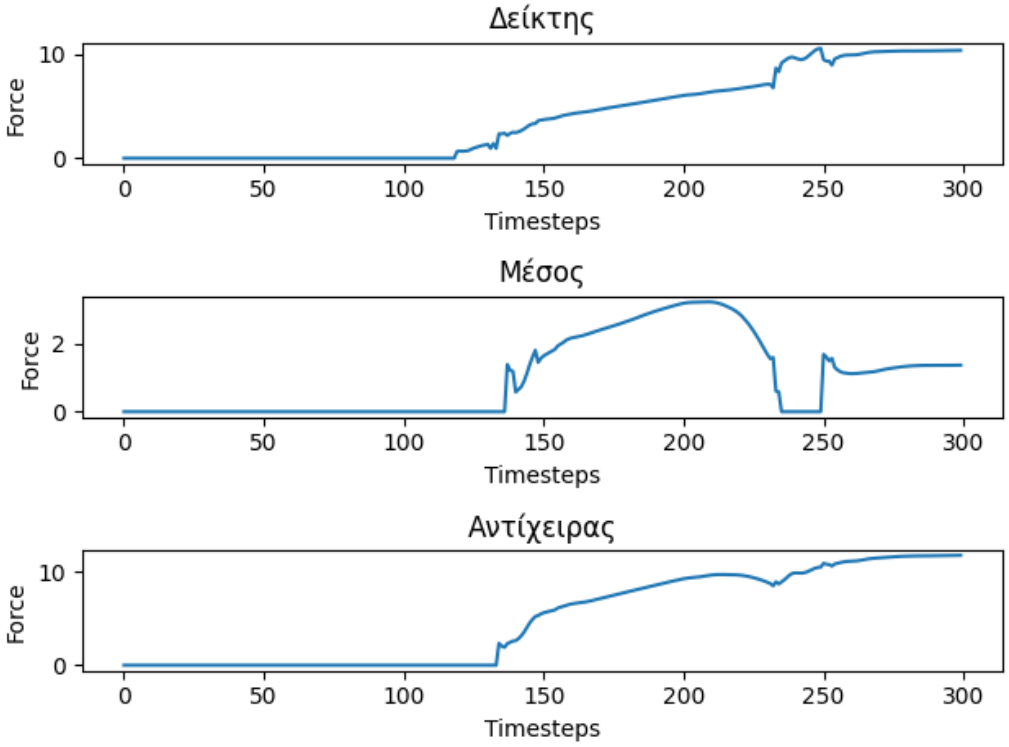
Σχήμα 6.15: *Καμπύλη μάθησης pitch χωρίς αισθητήρες δύναμης*



Σχήμα 6.16: *Παρακολούθηση τροχιάς ύψους - χωρίς αισθητήρες δύναμης*



Σχήμα 6.17: *Τροχιά Pitch - χωρίς αισθητήρες δύναμης*



Σχήμα 6.18: *Τροχιές δυνάμεων - χωρίς χρήση δυνάμεων στην μάθηση*

Από τα αποτελέσματα αυτά φαίνεται ότι μαθαίνεται η ζητούμενη παρακολούθηση τροχιάς ύψους. Ως πρώτο πείραμα της μεθόδου που προτείνουμε σχετικά με την παρακολούθηση τροχιάς παρατηρούμε τώρα, όπως είναι λογικό, ότι η βελτιστοποίηση δεν συμβαίνει και ως προς την ταχύτητα επίτευξης του τελικού ύψους. Με τον τρόπο, διορθώνονται και όλα τα overshoot που συνέβαιναν στα διάφορα μεγέθη. Ωστόσο, το κύριο πρόβλημα είναι ότι σε κάποιο σημείο χάνεται η επαφή από το μέσο δάκτυλο. Ως συνέπεια το αντικείμενο επιδέχεται ανεπιθύμητες διαταραχές στην τροχιά, το οποίο φαίνεται και ποιοτικά από την κίνηση και ποσοτικά από την ταλάντωση στην τροχιά του ύψους και την τροχιά του pitch στην αντίστοιχη χρονική στιγμή που χάνεται η επαφή. Επιπλέον, φαίνεται ότι η δύναμη διοχετεύεται ξαφνικά στον δείκτη, ασκώντας μεγαλύτερη δύναμη. Συμπερασματικά, μπορούμε να πούμε ότι με την μέθοδο αυτή έχουμε επιτυχία σε κάποιο βαθμό όσον αφορά την παρακολούθηση τροχιάς, αλλά η γενική συμπεριφορά ελέγχου του αντικειμένου δεν είναι η επιθυμητή, καθώς υπάρχει διακοπή της ζητούμενης λαβής επαφής.

6.2.2 Έλεγχος της επαφής με χρήση αισθητήρων δύναμης

Για να λύσουμε τα προβλήματα τα οποία δημιουργεί η απώλεια επαφής από το ένα δάκτυλο, εισάγουμε στον χώρο κατάστασης τους αισθητήρες δύναμης στις άκρες των τριών δακτύλων και ζητάμε τα δάχτυλα να ασκούν κάποια δύναμη στο αντικείμενο, αυξάνοντας την επιβράβευση ως εξής:

$$r(t) = -w_1|h(t) - h_D(t)| - w_2|pitch(t)| - w_3 r_f(t)$$

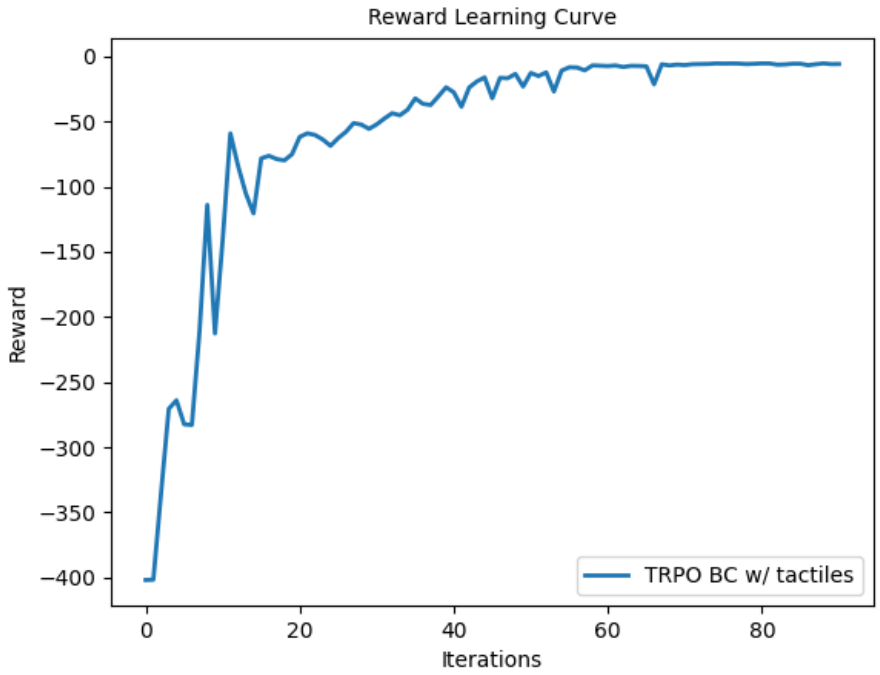
με

$$r_f(t) = \begin{cases} 0 & \text{if } f_{min,i} < f_i(t) \text{ for } i = 1,2,3 \\ 1 & \text{if } otherwise. \end{cases}$$

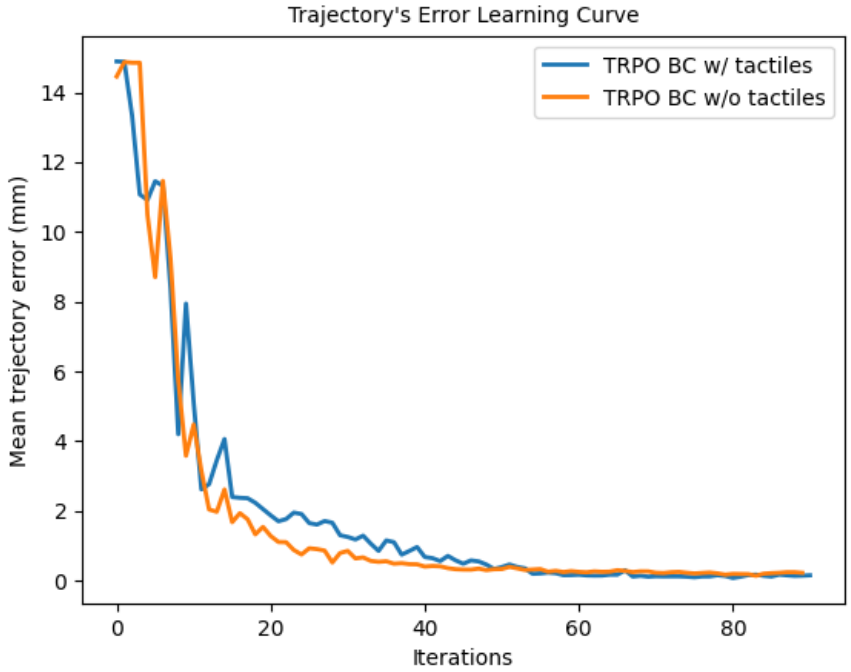
Επιθυμούμε με το πείραμα αυτό να διερευνήσουμε πως συνδράμει η διατήρηση της επαφής, δηλαδή το κάτω όριο και επομένως δεν ασχολούμαστε προς το παρόν με το άνω επιθυμητό όριο των δυνάμεων. Χρησιμοποιούμε όπως στο προηγούμενο πείραμα τις τιμές $w_1 = 0.1$, $w_2 = \frac{1}{70}$, $fmin, i = 0.1$, $i = 1,2,3$ και επιπλέον $w_3 = 0.5$, η οποία μπορεί να μεταφραστεί ως μία ποινή δύναμης που είναι ισοδύναμη με σφάλμα τροχιάς 5 χιλιοστών.

Για τις καμπύλες μάθησης παραθέτουμε όπως στα προηγούμενα πειράματα την καμπύλη του αθροίσματος των επιβραβεύσεων ανά επανάληψη (Σχήμα 6.19), η οποία ωστόσο δεν μπορεί να συγκριθεί με το προηγούμενο πείραμα στο οποίο δεν χρησιμοποιήσαμε αισθητήρες δύναμης καθώς έχει εισαχθεί μία ποινή δύναμης η οποία αλλάζει την επιβράβευση. Ωστόσο, συγκρίνουμε τις μεθόδους χρησιμοποιώντας τις καμπύλες μάθησης ανά μέγεθος, δηλαδή σφάλμα στις τροχιές ύψους, προσανατολισμού και ποσοστό διατήρησης της επαφής σε μία επανάληψη αξιολόγησης (Σχήματα 6.20 έως 6.22). Μετά την ολοκλήρωση της φάσης εκπαίδευσης (60 επαναλήψεις) και της σύγκλισης των καμπυλών εκμάθησης, εκτελέσθηκε σειρά πειραμάτων για την συγκριτική αξιολόγηση της τελικής επίδοσης της μεθόδου με και χωρίς τη χρήση δεδομένων από αισθητήρες δύναμης/αφής. Στα Σχήματα 6.23 έως 6.25 παρουσιάζονται τα αποτελέσματα από μια ενδεικτική εκτέλεση της εργασίας εσωτερικού χειρισμού ρομποτικής λαβής και ανύψωσης του αντικειμένου, συγκρίνοντας τις τροχιές ύψους, γωνίας ανύψωσης και δυνάμεων που προκύπτουν με έλεγχο ή μη των δυνάμεων επαφής (w/ vs w/o tactiles).

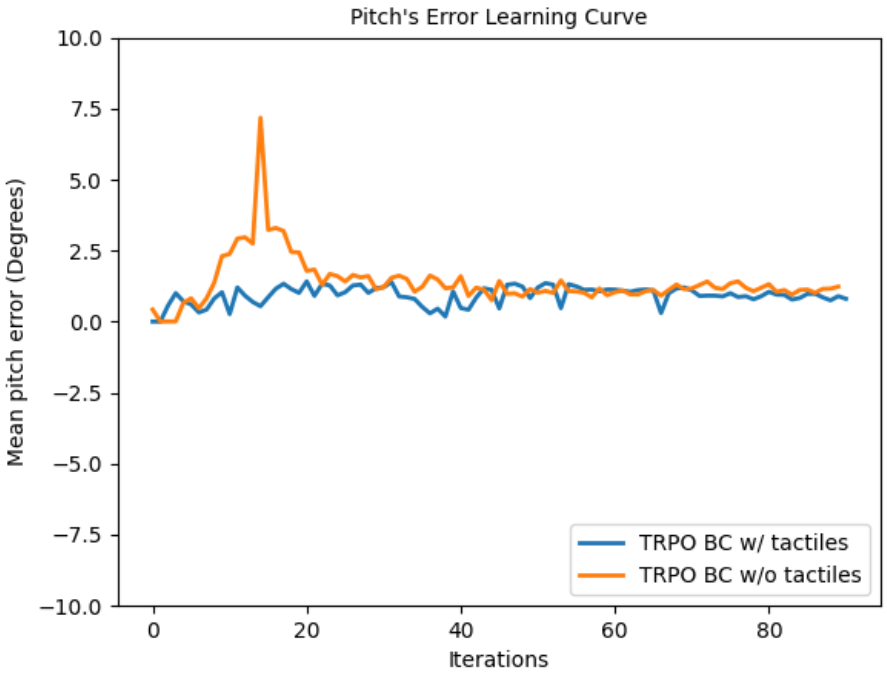
Από τις καμπύλες αυτές μπορούμε να εξάγουμε τα ακόλουθα συμπεράσματα. Με χρήση των αισθητήρων δύναμης και έλεγχο της επαφής κατά το grasp μέσω της επιβράβευσης, διορθώνεται η ανεπιθύμητη συμπεριφορά η οποία αφορά την απώλεια επαφής που δημιουργεί ποιοτικά διακοπή της ζητούμενης λαβής και ανεπιθύμητες διαταραχές στην τροχιά ανύψωσης του αντικειμένου και ποσοτικά διόρθωση της ταλάντωσης στα μεγέθη που μελετάμε. Επιπλέον, αναφορικά με τις δυνάμεις διορθώνεται η συμπεριφορά κατά την οποία η απώλεια επαφής από τον μέσο αναγκάζει τον δείκτη να ασκήσει απότομα μεγάλη δύναμη ώστε να κρατήσει το αντικείμενο στην επιθυμητή τροχιά. Ωστόσο, μπορούμε να δούμε από τις καμπύλες μάθησης ότι χρειάζονται περισσότερες επαναλήψεις για την σύγκλιση. Αυτό είναι απολύτως λογικό καθώς έχουμε εισάγει έναν ακόμη παράγοντα προς βελτιστοποίηση, οπότε χρειάζεται περισσότερη εξερεύνηση από μεριάς ενισχυτικής μάθησης ώστε να βρει επιθυμητές κινήσεις. Αυτό φαίνεται ουσιαστικά στην καμπύλη μάθησης των δυνάμεων, που αφορά την διατήρηση της επαφής. Στην περίπτωση που δεν χρησιμοποιούσαμε αισθητήρες δύναμης στην καλύτερη περίπτωση η επαφή διατηρείται περίπου στο 75% της τροχιάς, ενώ στην περίπτωση που χρησιμοποιούμε αισθητήρες δύναμης φτάνει το 100%.



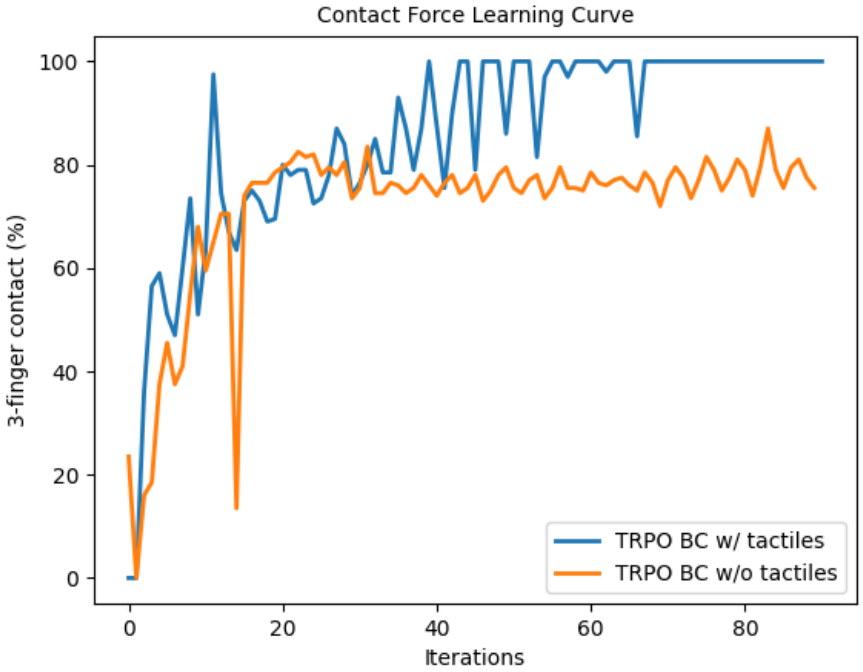
Σχήμα 6.19: *Καμπύλη επιβράβευσης - Χρήση αισθητήρων δύναμης για έλεγχο επαφής*



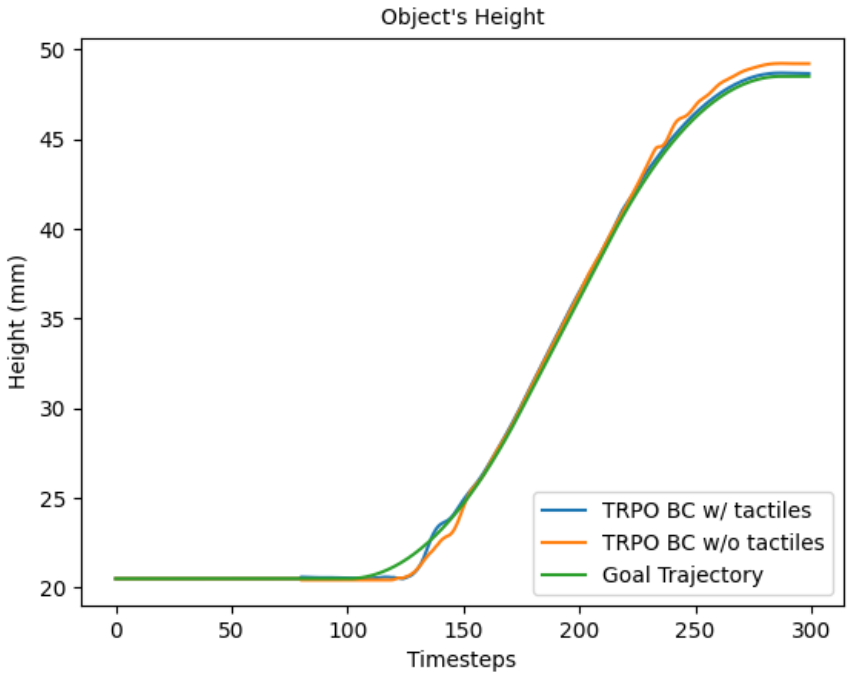
Σχήμα 6.20: *Καμπύλη μάθησης τροχιάς ύψους- Με και χωρίς έλεγχο επαφής*



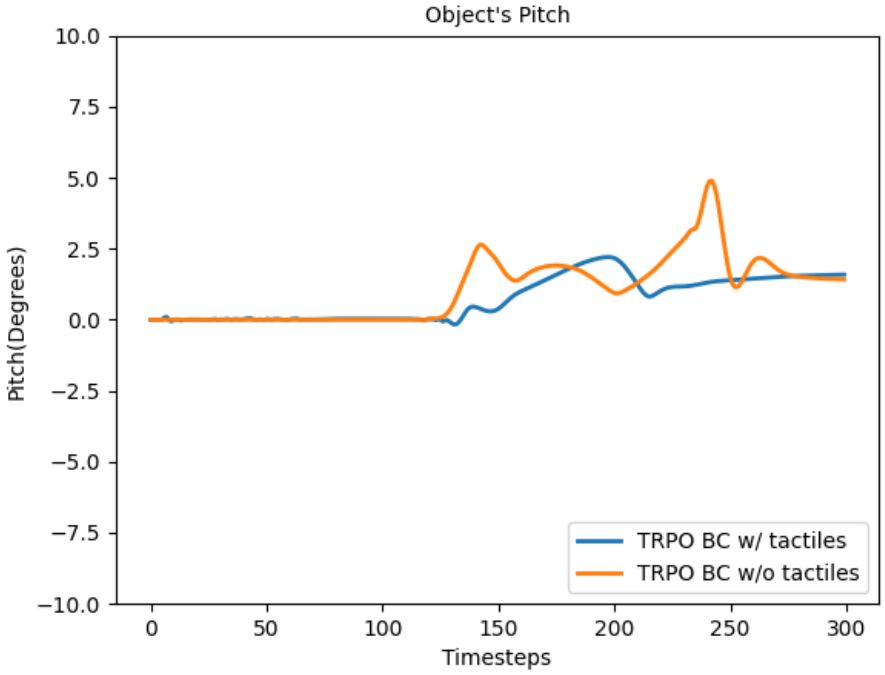
Σχήμα 6.21: *Καμπύλη μάθησης pitch- Με και χωρίς έλεγχο επαφής*



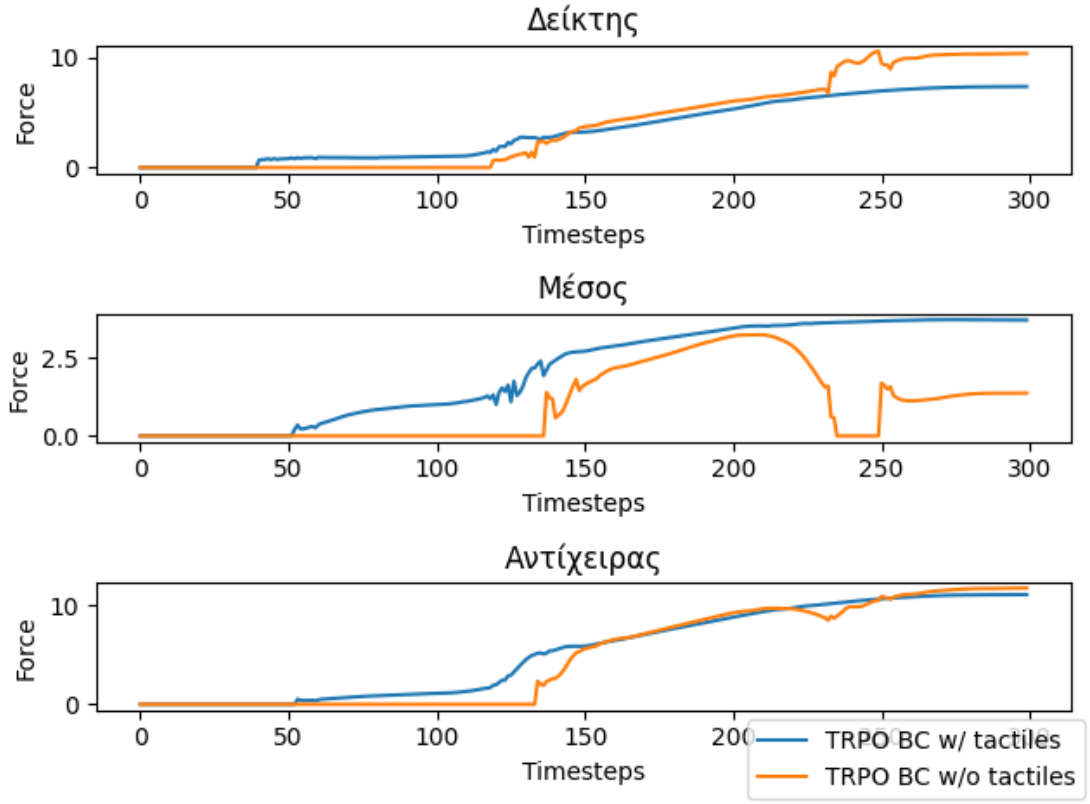
Σχήμα 6.22: *Καμπύλη μάθησης διατήρησης επαφής - Με και χωρίς έλεγχο επαφής*



Σχήμα 6.23: *Παρακολούθηση τροχιάς ύψους - Με και χωρίς έλεγχο επαφής*



Σχήμα 6.24: *Τροχιές pitch - Με και χωρίς έλεγχο επαφής*



Σχήμα 6.25: *Τροχιές δυνάμεων - Με και χωρίς έλεγχο επαφής*

6.2.3 Χρήση φάσης στο pre-grasp στάδιο

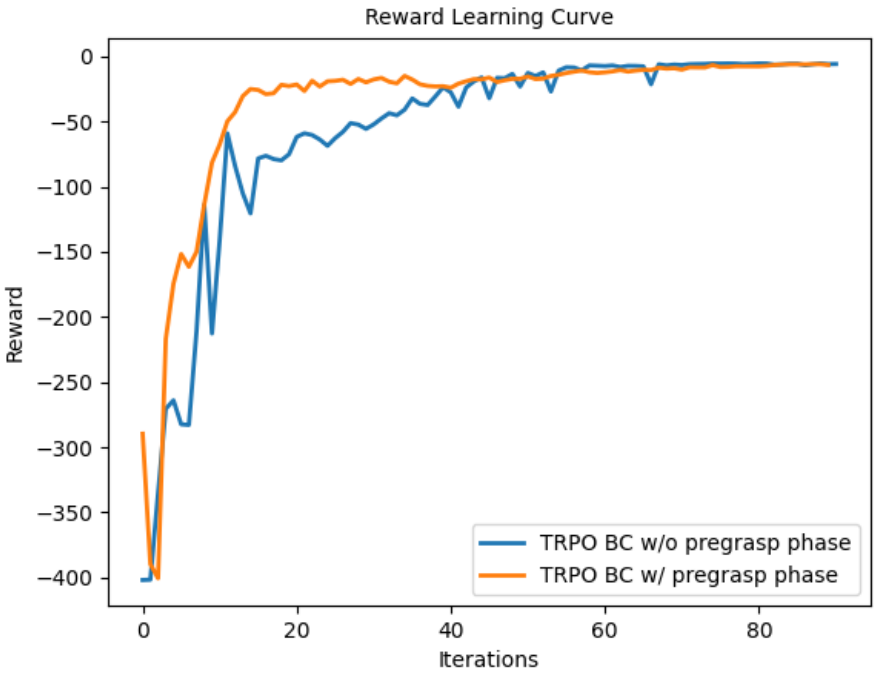
Τόσο με χρήση αισθητήρων δύναμης όσο και χωρίς αυτούς παρατηρήσαμε ότι υπάρχει μία καθυστέρηση στην ανύψωση του αντικειμένου. Η καθυστέρηση αυτή φυσικά είναι μεγάλη στις αρχικές επαναλήψεις και βελτιώνεται στην συνέχεια, κάτι το οποίο φαίνεται όμως να μαθαίνεται με δυσκολία. Αυτό φαίνεται σε μεγάλο βαθμό και από την καμπύλη μάθησης των δυνάμεων, η οποία αργεί αρκετά να συγκλίνει ώστε η επαφή να διατηρείται στο 100% της κίνησης. Η τροχιά αυτή αφορά την διακοπή της επαφής, αλλά και την επαφή στην αρχή της κίνησης. Για μεγάλο διάστημα δηλαδή το ρομπότ πιάνει το αντικείμενο μετά την έναρξη της επιθυμητής τροχιάς καθυστερώντας να το σηκώσει. Ακόμη και όμως, όταν το πιάνει πριν την έναρξη της επιθυμητής τροχιάς δεν το σηκώνει, αφού πιθανότατα δεν είναι σε θέση να το κάνει. Μέσω του χώρου κατάστασης η αρχή του grasp συμβαίνει όταν δοθεί επόμενος στόχος ύψους μεγαλύτερο αυτού που αφορά το αρχικό ύψος του αντικειμένου, δηλαδή στην περίπτωση μας τα 100 timesteps. Αυτό με άλλα λόγια γίνεται μέσα σε μία χρονική στιγμή και δεν έχει δοθεί στο σύστημα μία πρότερη γνώση ότι ακολουθεί εκκίνηση της διαδικασία ανύψωσης ώστε το ρομπότ να προετοιμαστεί (π.χ. να ασκήσει την απαιτούμενη δύναμη).

Για να εισάγουμε στο σύστημα μία γνώση της επικείμενης έναρξης του grasp εισάγουμε στον χώρο κατάστασης μία φάση κατά το στάδιο πριν το grasp, όπως περιγράψαμε στην ανάλυση της μεθοδολογίας μας στο Κεφάλαιο 4. Η φάση αυτή παίρνει συνεχείς τιμές στο

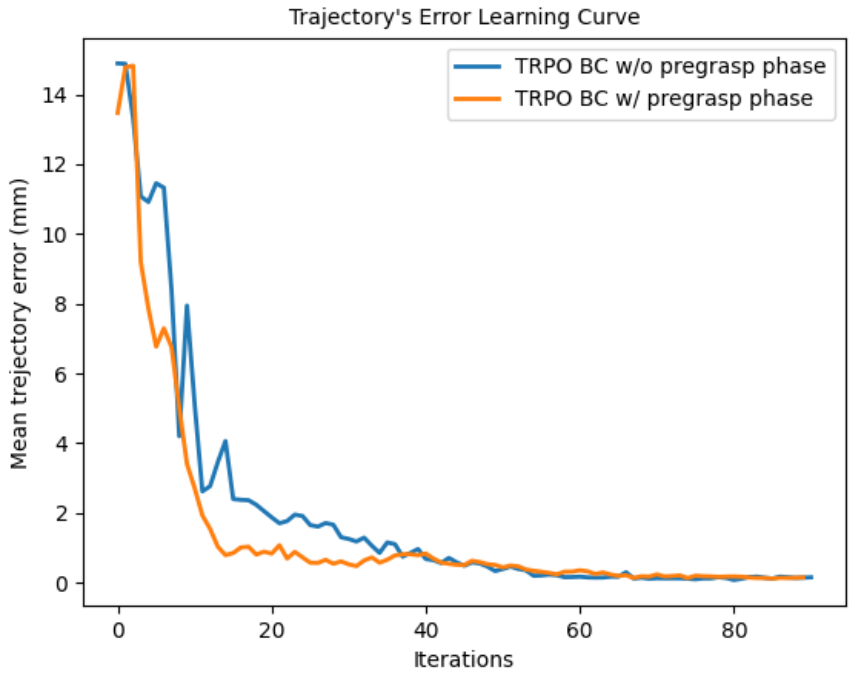
[0,1] και είναι ανάλογη του timestep ώστε να ξεκινά από 0 και να γίνεται 1 όταν ξεκινά η τροχιά ανύψωσης. Επιλέγουμε στο πείραμα αυτό να γίνεται 1 όταν το επιθυμητό ύψος του κέντρου βάρους έχει αυξηθεί κατά 0.1mm από το ύψος του κέντρου βάρους όταν το αντικείμενο βρίσκεται στο δάπεδο. Από εκείνη την στιγμή και έπειτα η φάση παραμένει 1 ώστε να μην παίζει κάποιο ρόλο κατά το grasp.

Στο ακόλουθο πείραμα διατηρούμε την επιβράβευση όπως ήταν στο προηγούμενο πείραμα και απλώς εισάγουμε στον χώρο κατάστασης την μεταβλητή φάσης, ο οποίος τώρα περιέχει τις αρθρώσεις, την θέση και τις γωνίες προσανατολισμού του αντικειμένου, τον επόμενο επιθυμητό στόχο της τροχιάς ύψους και την μεταβλητή φάσης. Μελετάμε την επίδρασή της, συγκρίνοντας τα αποτελέσματα σε σχέση με το προηγούμενο πείραμα (και στις δύο περιπτώσεις κάνουμε έλεγχο της επαφής μέσω αισθητήρων δύναμης). Στα Σχήματα 6.26 έως 6.29 παρουσιάζονται οι καμπύλες μάθησεις της επιβράβευσης συνολικά και ξεχωριστά για κάθε μέγεθος που μελετάμε (ύψος, γωνία ανύψωσης και ποσοστό διατήρησης της επαφής) συγκριτικά για τις περιπτώσεις που χρησιμοποιείται ή όχι η μεταβλητή φάσης (w/ vs w/o pregrasp phase). Επιπλέον παρουσιάζουμε στα Σχήματα 6.30, έως 6.32 τις τροχιές των μεγεθών αυτών σε ένα ενδεικτικό επισόδειο αξιολόγησης μετά την ολοκλήρωση της φάσης εκπαίδευσης (60 επαναλήψεις).

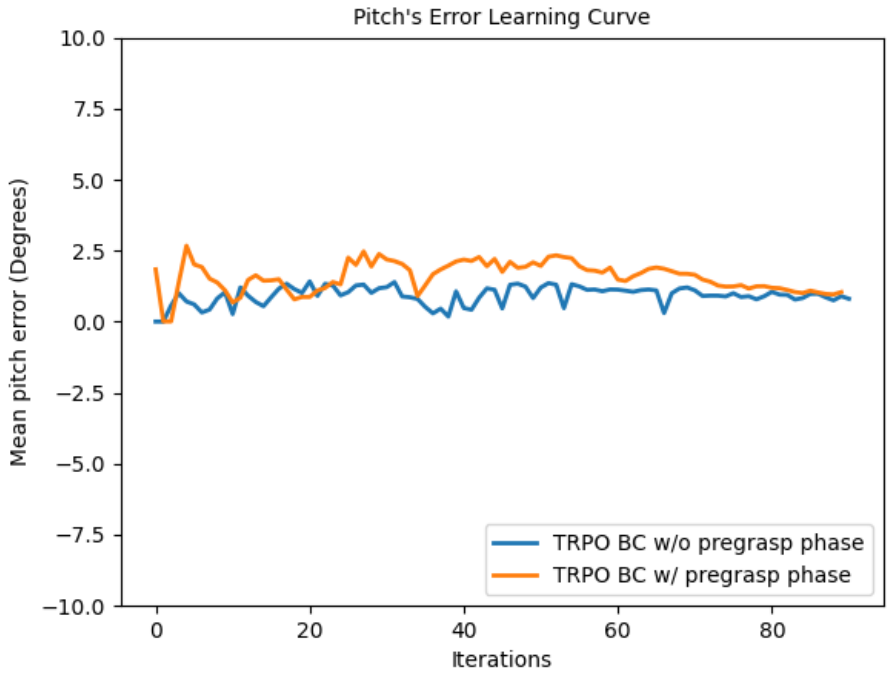
Από τις καμπύλες μάθησης παρατηρούμε ότι η εισαγωγή της φάσης πριν από την κίνηση βελτιώνει την απόδοση. Η μεγάλη αλλαγή παρατηρείται όπως περιμέναμε στην καμπύλη μάθησης των δυνάμεων. Η πληροφορία δηλαδή του πόσο κοντά βρίσκεται το σύστημα στην αρχή της επιθυμητής κίνησης που δίνει η φάση που προσθέσαμε βελτιώνει κατά πολύ την διατήρηση επαφής. Μπορούμε, να δούμε επιπλέον ότι αυτό το σύστημα ξεκινά από ένα πολύ υψηλό ποσοστό διατήρησης επαφής. Αυτό συμβαίνει γιατί η φάση αυτή εκτός της πληροφορίας που δίνει για την αρχή της κίνησης συγχρονίζει το στάδιο σχηματισμού της λαβής από τα δεδομένα επίδειξης. Επιπλέον, όσον αφορά τα υπόλοιπα μεγέθη, εφόσον η συμπεριφορά των δυνάμεων μαθαίνεται γρήγορα δεν χρειάζεται εξερεύνηση σε αυτόν τον τομέα και επομένως το σύστημα ¨ασχολείται¨ με την τροχιά ύψους εξαρχής. Επιπλέον, παρατηρώντας τις τροχιές των μεγεθών κατά την εκτέλεση επεισοδίου αξιολόγησης, βλέπουμε αρχικά ότι εξαφανίζεται η καθυστέρηση στην αρχή της κίνησης, ενώ η άσκηση δύναμης ξεκινά ακριβώς πριν από την αρχή της λαβής.



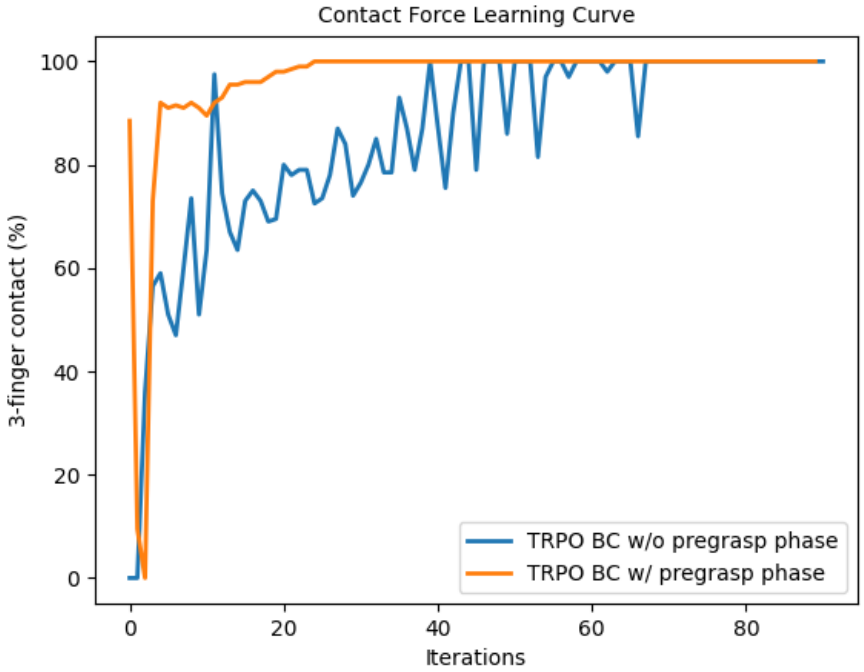
Σχήμα 6.26: *Καμπύλη επιβράβευσης - Με και χωρίς χρήση pregrasp φάσης*



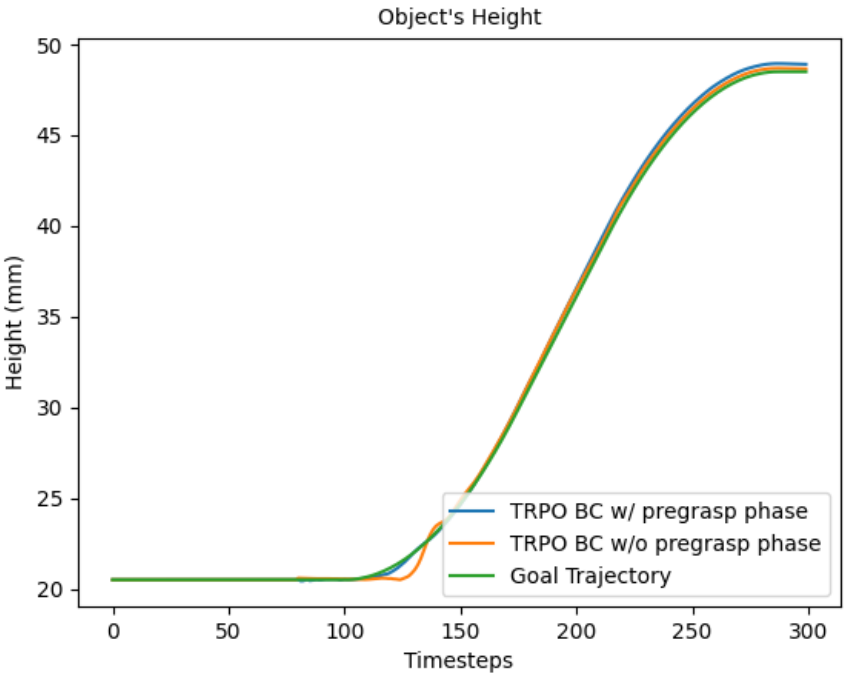
Σχήμα 6.27: *Καμπύλη μάθησης τροχιάς ύψους- Με και χωρίς χρήση pregrasp φάσης*



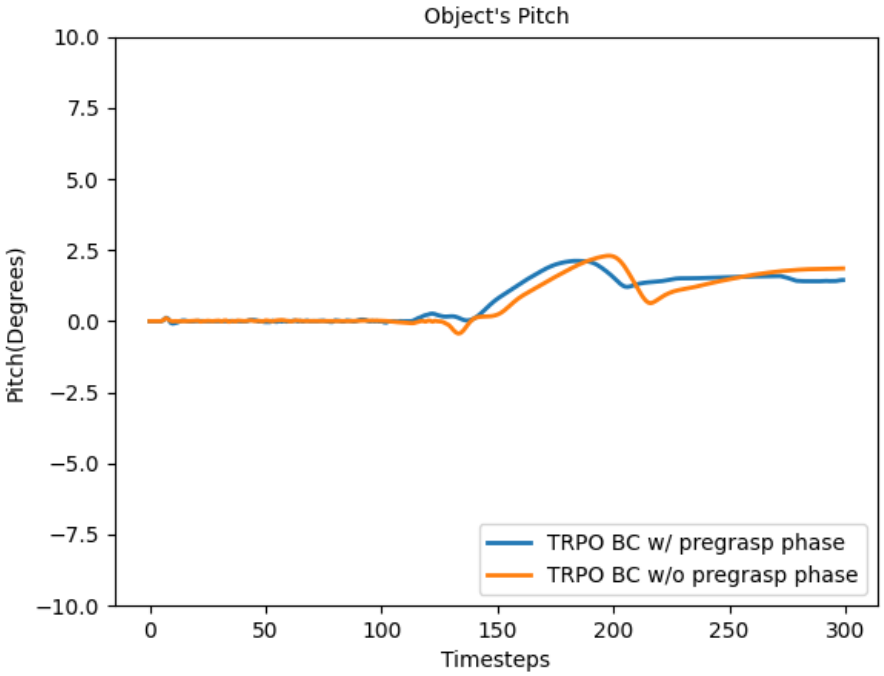
Σχήμα 6.28: *Καμπύλη μάθησης pitch- Με και χωρίς χρήση pregrasp φάσης*



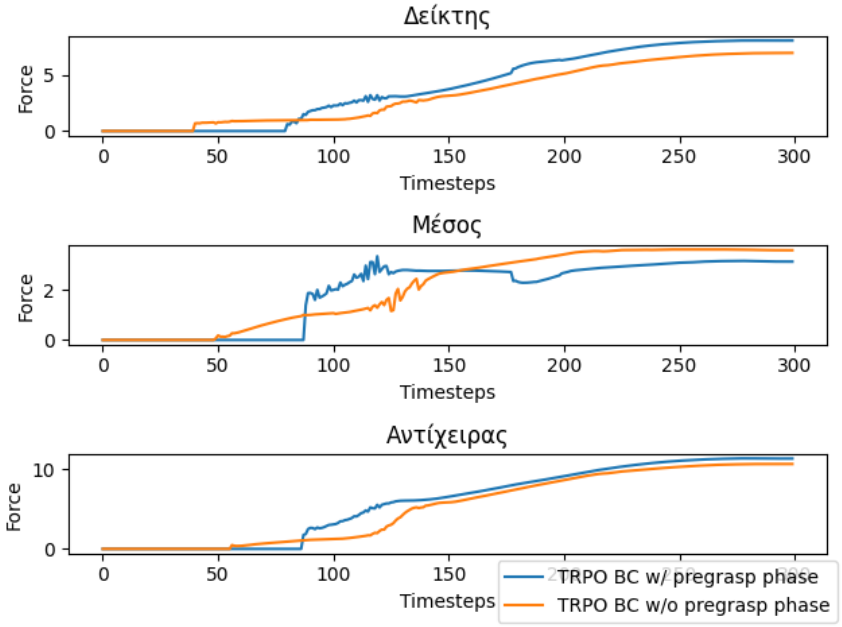
Σχήμα 6.29: *Καμπύλη μάθησης διατήρησης επαφής - Με και χωρίς χρήση pregrasp φάσης*



Σχήμα 6.30: *Παρακολούθηση τροχιάς ύψους - Με και χωρίς χρήση pregrasp φάσης*



Σχήμα 6.31: *Τροχιά pitch - Με και χωρίς χρήση pregrasp φάσης*



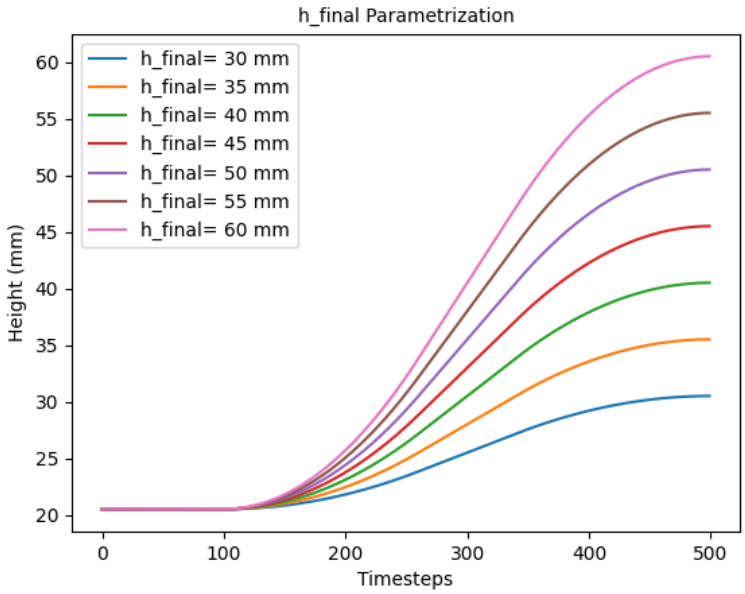
Σχήμα 6.32: *Τροχιές δυνάμεων - Με και χωρίς χρήση pregrasp φάσης*

6.2.4 Έλεγχος Γενίκευσης

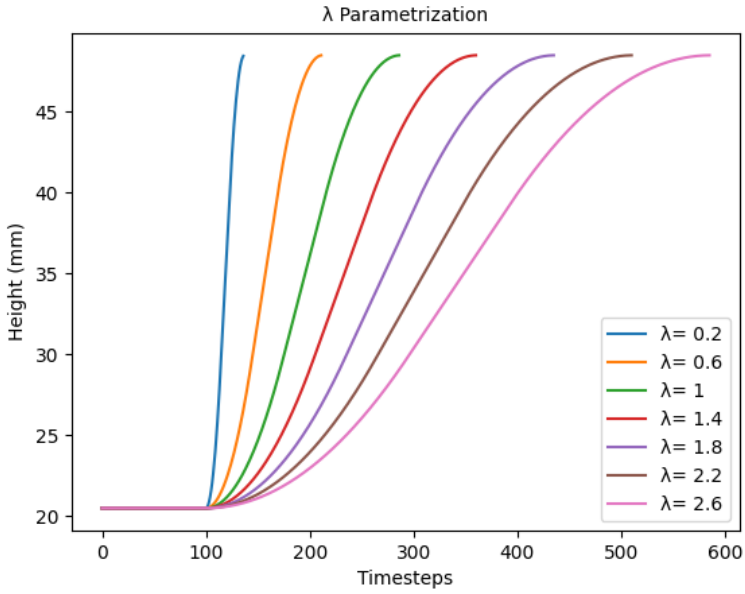
Το σύστημα που χρησιμοποιήσαμε στο τελευταίο πείραμα πετυχαίνει να εκτελέσει την επιθυμητή συμπεριφορά με την οποία ασχολούμαστε στην παρούσα εργασία, δηλαδή το grasp τριών δακτύλων ενός αντικείμενο ακολουθώντας μία τροχιά ύψους ορισμένη από τον χρήση διατηρώντας το αντικείμενο οριζόντιο και διατηρώντας την επαφή με αυτό. Όπως αναφέραμε, ωστόσο, στην περιγραφή της μεθοδολογίας που προτείναμε στο Κεφάλαιο 4, ο σκοπός είναι το σύστημα αυτό να προσφέρει κάποια γενίκευση όσον αφορά την τροχιά που δίνεται ως επιθυμητή. Ελέγχουμε λοιπόν στο σημείο αυτό τις δυνατότητες γενίκευσης του συστήματος μας παραμετροποιώντας την επιθυμητή τροχιά h_D αναφορικά με το ζητούμενο τελικό ύψος h_{final} και έναν πολλαπλασιαστικό παράγοντα λ σχετικό με τον χρόνο διάρκειας την κίνησης ανύψωσης (ανύψωση λ φορές πιο αργά) σύμφωνα με τον οποίο έγινε η εκπαίδευση. Η εκπαίδευσης δηλαδή του συστήματος, όπως αναφέραμε, έγινε με τις παραμέτρους $h_{final} = 48mm$ και $\lambda = 1$. Στα Σχήματα 6.33 και 6.34 φαίνεται πως επιδρούν διαφορετικές τιμές αυτής της παραμετροποίησης στην επιθυμητή τροχιά.

Χρησιμοποιώντας διάφορες τιμές για τις δύο αυτές παραμέτρους παραθέτουμε έναν πίνακα (Σχήμα 6.35) με το μέσο σφάλμα της τροχιάς που εκτελεί το σύστημα από την επιθυμητή στο αντίστοιχο επισόδιο αξιολόγησης και παραθέτουμε γραφικά τα αποτελέσματα σε ένα χρωματικό διάγραμα δύο διαστάσεων (Σχήμα 6.36), όπου το λ απεικονίζεται λογαριθμικά και το τελικό ύψος γραμμικά. Επιπλέον, παρουσιάζουμε στα Σχήματα 6.38 και 6.39 ενδεικτικά για διάφορες τιμές των παραμέτρων h_final και λ, την συμπεριφορά του συστήματος σε σχέση με την επιθυμητή τροχιά.

Όπως φαίνεται από τα αποτελέσματα υπάρχει καλή γενίκευση ως προς τον χρόνο. Το σύστημα είναι ικανό να ακολουθήσει τόσο γρηγορότερες τροχιές (μικρά λ) όσο και πιο αργές (μεγάλο λ). Βλέπουμε ωστόσο, ότι στις μεγάλες ταχύτητες υπάρχει μία καθυστέρηση στην



Σχήμα 6.33: *Παραμετροποίηση ως προς τελικό ύψος ανύψωσης*



Σχήμα 6.34: *Παραμετροποίηση ως χρόνο ανύψωσης*

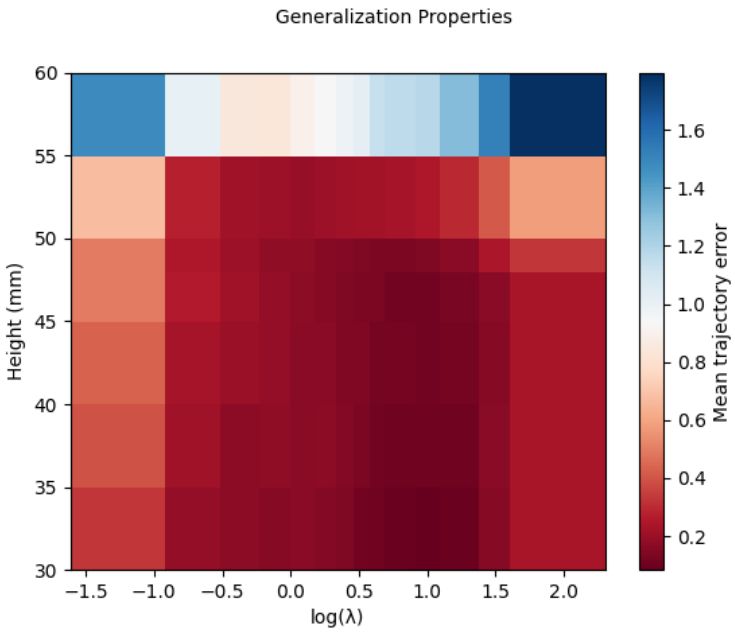
αρχή, πιθανότατα επειδή από την αδράνεια ζητάμε ξαφνικά πολύ μεγάλη ταχύτητα, ενώ στις πολύ αργές υπάρχει μία μικρή απόκλιση στο τελικό ύψος. Συνολικά, ωστόσο υπάρχει μία καλή γενίκευση του συστήματος ως προς τον χρόνο.

Αναφορικά με το ζητούμενο ύψος, παρατηρούμε ότι για μικρότερα ύψη από το ύψος πάνω στο οποίο έγινε η εκπαίδευση (48 χιλιοστά), επιτυγχάνεται η ζητούμενη τροχιά ενώ για μεγαλύτερη ύψη υπάρχει μεγάλο τελικό σφάλμα. Συγκεκριμένα για μεγαλύτερα ύψη το σύστημα μπορεί να φτάσει ύψη λίγο μεγαλύτερα από αυτό της εκπαίδευσης περίπου στα (55 χιλιοστά) και αδυνατεί να σηκώσει το αντικείμενο περαιτέρω.

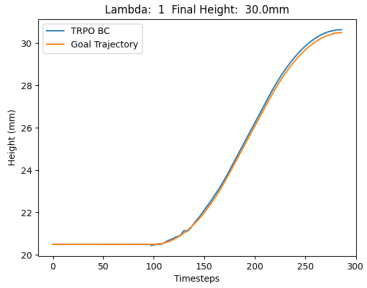
Mean trajectory error

λ\h(mm)	30	35	40	45	48	50	55	60
0.2	0.326	0.392	0.432	0.493	0.496	0.679	1.49	3.063
0.4	0.185	0.214	0.228	0.26	0.25	0.273	1.012	2.417
0.6	0.173	0.174	0.198	0.213	0.209	0.215	0.843	2.287
0.8	0.158	0.177	0.185	0.185	0.181	0.208	0.835	2.151
1	0.165	0.167	0.168	0.174	0.18	0.193	0.898	2.239
1.2	0.151	0.174	0.166	0.157	0.161	0.208	0.937	2.196
1.4	0.137	0.153	0.148	0.142	0.153	0.214	0.988	2.383
1.6	0.111	0.131	0.145	0.135	0.138	0.218	1.023	2.525
1.8	0.104	0.116	0.123	0.132	0.132	0.221	1.14	2.53
2.0	0.094	0.105	0.121	0.117	0.133	0.224	1.167	2.645
2.5	0.083	0.106	0.116	0.111	0.142	0.248	1.176	2.7
3	0.096	0.104	0.121	0.124	0.167	0.297	1.313	2.743
4	0.161	0.166	0.162	0.167	0.24	0.417	1.526	3.08
5	0.233	0.237	0.234	0.234	0.327	0.584	1.798	3.339
10	0.447	0.5	0.517	0.58	0.846	1.301	2.724	4.397

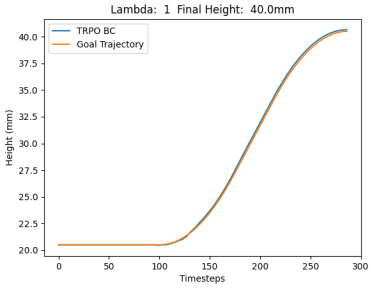
Σχήμα 6.35: *Μέση τιμή σφάλματος ανά χρόνο και τελικό ύψος ανύψωσης*



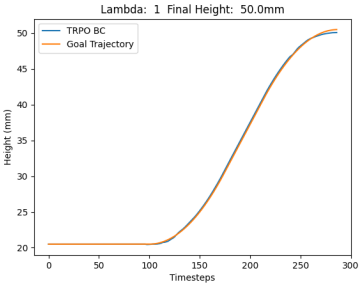
Σχήμα 6.36: *Χρωματικό διάγραμμα μέσου σφάλματος τροχιάς ύψους συναρτήσει τελικού ύψους και χρόνου ανύψωσης.*



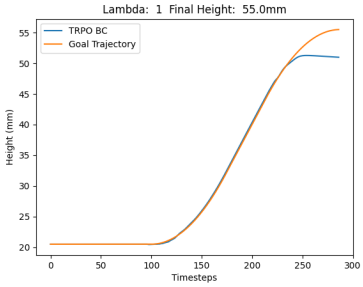
(α')



(β')

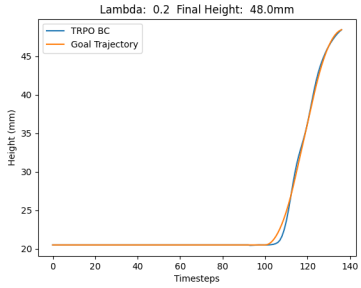


(γ')

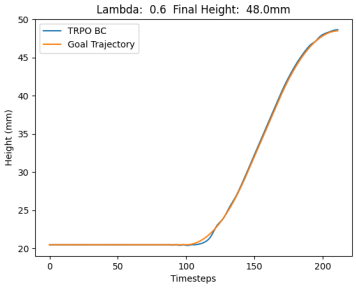


(δ')

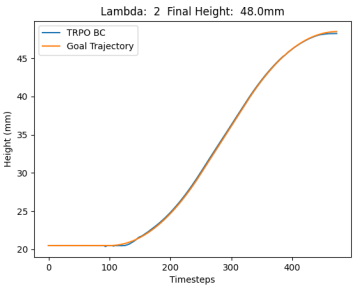
Σχήμα 6.37: *Αποτελέσματα γενίκευσης ως προς τελικό ύψος ανύψωσης*



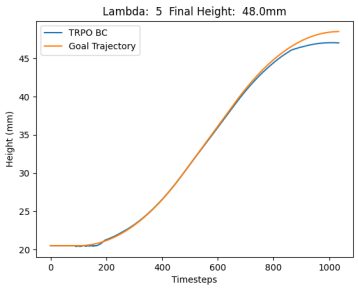
(α')



(β')



(γ')



(δ')

Σχήμα 6.38: *Αποτελέσματα γενίκευσης ως προς χρόνο ανύψωσης*

6.2.5 Περιορισμός της Δύναμης Επαφής

Στο επόμενο πείραμα ζητάμε οι δυνάμεις άσκησης στο αντικείμενο να μένουν εντός ενός επιθυμητού εύρους. Αντίστοιχο πείραμα έγινε και για το απλό grasp χωρίς παρακολούθηση τροχιάς (παράγραφος 6.1.3). Διαμορφώνουμε λοιπόν ανάλογα την συνάρτηση επιβράβευσης ως:

$$r(t) = -w_1|h(t) - h_D(t)| - w_2|pitch(t)| - w_3 r_f(t)$$

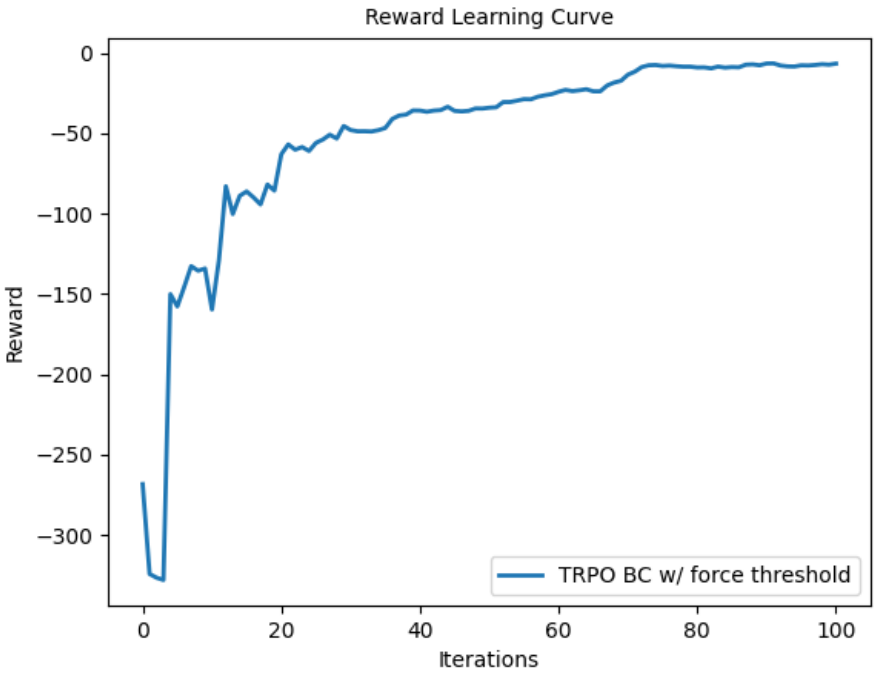
με

$$r_f = \begin{cases} 0 & \text{if } f_{min,i} < f_i(t) < f_{max,i} \text{ for } i = 1,2,3 \\ 1 & \text{if } otherwise. \end{cases}$$

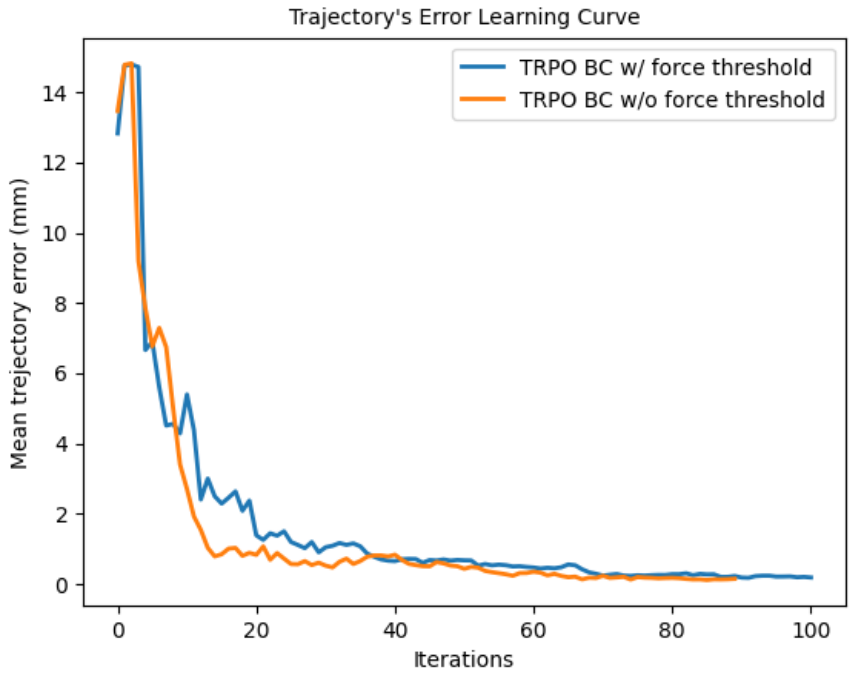
Χρησιμοποιούμε όπως πριν τις τιμές $w_1 = 0.1$, $w_2 = \frac{1}{70}$, $f_{min,i} = 0.1$ και επιπλέον $w_3 = 0.2$ για την ποινή του ορίου των δυνάμεων.

Για το άνω όρια θέτουμε όπως στο πείραμα της παραγράφου 6.1.3 όριο τιμής 10 στον αντίχειρα, που αντικατοπτρίζει την συνολική δύναμη και 5 στα δύο δάκτυλα, ώστε να μοιράσουμε την δύναμη. Όπως αναφέραμε, αυτή είναι μία διαφορετική συμπεριφορά από τα δεδομένα επίδειξης και συνεπώς χρειάζεται περισσότερη εξερεύνηση από πλευράς ενισχυτικής μάθησης. Παραθέτουμε τις καμπύλες μάθησης, όπως στα προηγούμενα πειράματα για την επιβράβευση συνολικά και ξεχωριστά για το ύψος, την γωνία ανύψωσης και το ποσοστό διατήρησης της τροχιάς εντός επιθυμητών ορίων στα σχήματα 6.39 έως 6.42. Καθώς το συγκεκριμένο πείραμα ελέγχου των τιμών των δυνάμεων αποτελεί ένα διαφορετικό πρόβλημα βελτιστοποίησης, δεν μπορούμε να συγκρίνουμε άμεσα τις καμπύλες μάθησης της επιβράβευσης με αυτές χωρίς τον έλεγχο των τιμών αυτών. Ωστόσο, παραθέτουμε συγκριτικά τις καμπύλες μάθησης που αφορούν ξεχωριστά τα μεγέθη που μελετάμε. Μετά την σύγκλιση καμπυλών εκμάθησης (80 επαναλήψεις), εκτελέσθηκε σειρά πειραμάτων για την συγκριτική αξιολόγηση την συμπεριφοράς των δυνάμεων με και χωρίς το άνω όριο (w/ vs w/o force threshold) (Σχήμα 6.43).

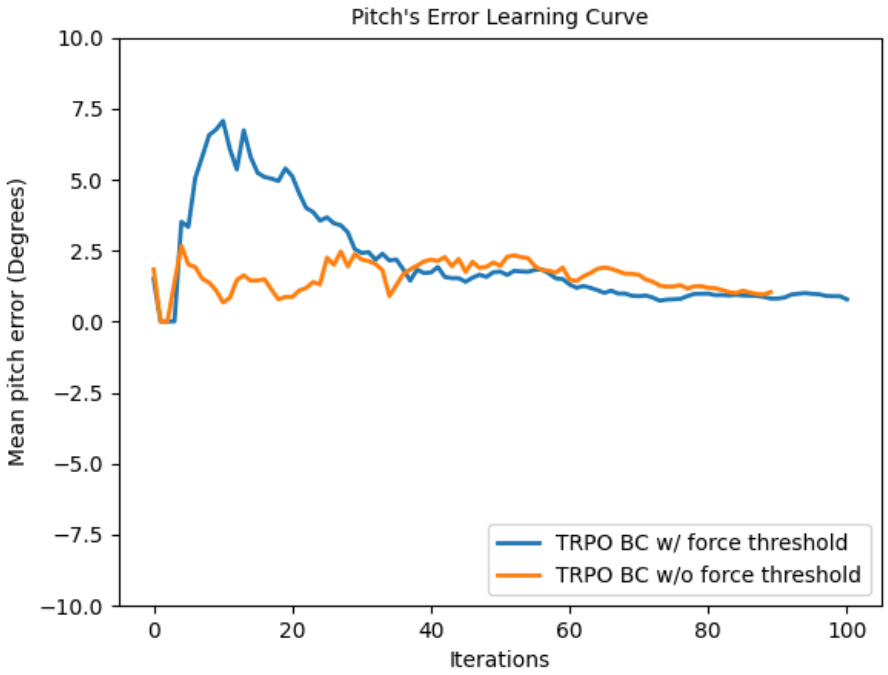
Συμπερασματικά, είναι προφανές ότι χρειάζονται περισσότερες επαναλήψεις ώστε να έχουμε το επιθυμητό αποτέλεσμα. Συγκεκριμένα, ενώ οι καμπύλη του σφάλματος τροχιάς συγκλίνει γρήγορα, η καμπύλη σύγκλισης για τις δυνάμεις αργεί να συγκλίνει και φτάνει το 100% περίπου στην εβδομηκοστή επανάληψη. Κάτι τέτοιο, όπως αναφέραμε είναι πολύ λογικό. Το σύστημα έχει μία καλή πρότερη γνώση από τα δεδομένα επίδειξης σχετική με την ανύψωση του αντικειμένου, και συνεπώς η βελτιστοποίηση σχετικά με το ύψος γίνεται σε μικρό αριθμό επαναλήψεων. Από την άλλη πλευρά, δεν υπάρχει καθοδήγηση από τα δεδομένα επίδειξης σχετικά με την συμπεριφορά των δυνάμεων που ζητάμε και επομένως χρειάζονται αρκετές επαναλήψεις ώστε να γίνει εξερεύνηση της επιθυμητής συμπεριφοράς των δυνάμεων επαφής.



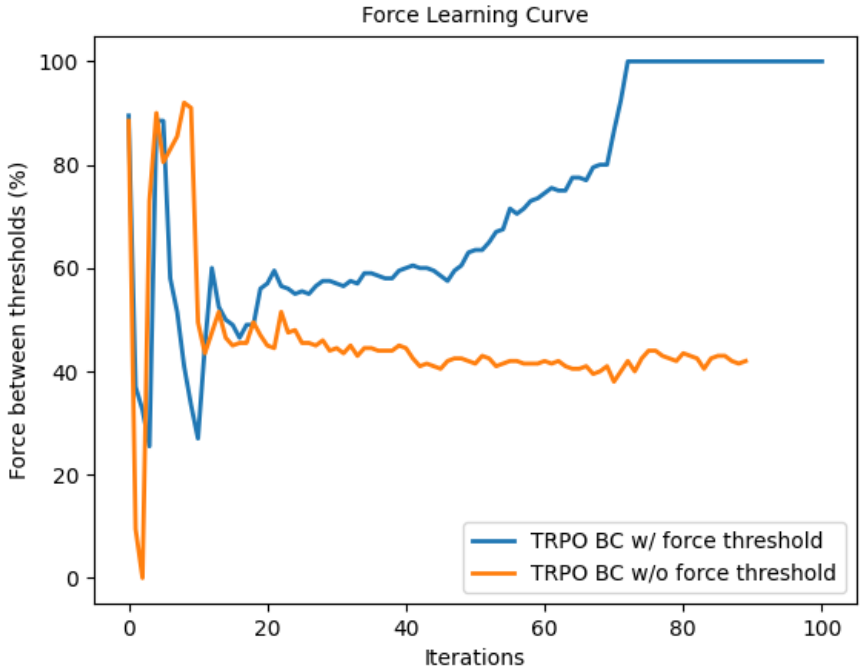
Σχήμα 6.39: *Καμπύλη επιβράβευσης - Περιορισμός δύναμης επαφής*



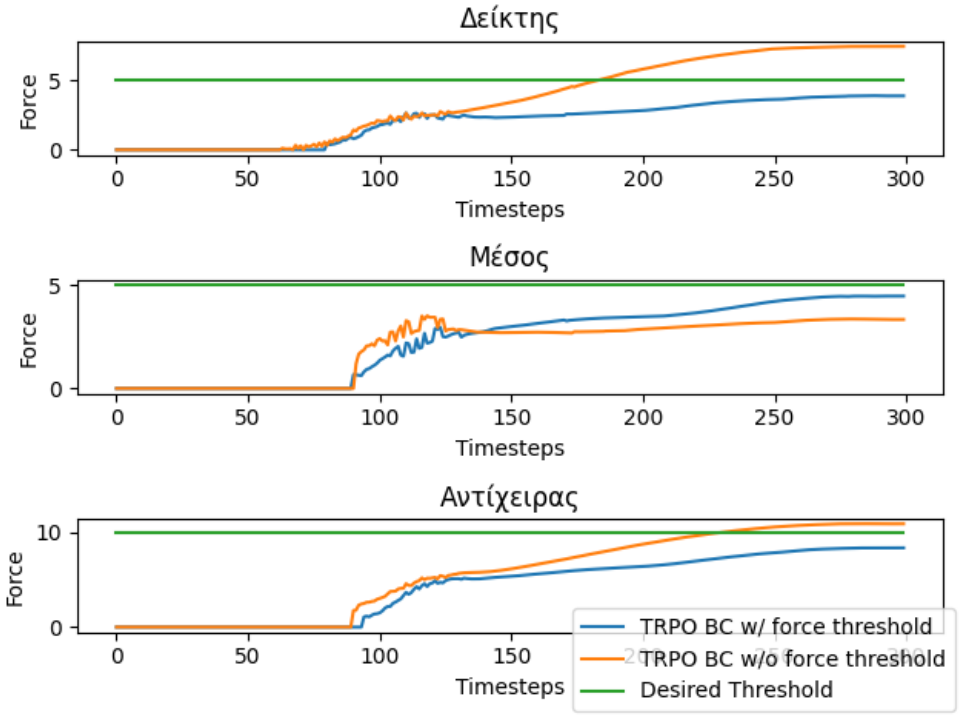
Σχήμα 6.40: *Καμπύλης μάθησης τροχιάς ύψους - Περιορισμός δύναμης επαφής*



Σχήμα 6.41: *Καμπύλης μάθησης pitch - Περιορισμός δύναμης επαφής*



Σχήμα 6.42: *Καμπύλης μάθησης δυνάμεων - Περιορισμός δύναμης επαφής*



Σχήμα 6.43: *Τροχιές δυνάμεων - Με και χωρίς περιορισμό στην δύναμη επαφής*

Κεφάλαιο 7

Επίλογος

7.1 Συμπεράσματα

Σε αυτή την εργασία έγινε προσέγγιση του προβλήματος μιας εσωτερικής λαβής (in hand grasp) και επιδέξιου χειρισμού με ένα ρομποτικό χέρι. Εστιάσαμε στο πρόβλημα επίτευξης λαβής με τα άκρα των δακτύλων ως επαφή ρομπότ-αντικειμένου και οριζόντια ανύψωση με παρακολούθηση προκαθορισμένης τροχιάς ύψους του αντικειμένου και έλεγχο των δυνάμεων που ασκούνται στο αντικείμενο ώστε να βρίσκονται σε επιθυμητό προκαθορισμένο όριο τιμών. Το πρόβλημα προσεγγίστηκε με ενισχυτική μάθηση και προεκπαίδευση του συστήματος με επιβλεπόμενη μάθηση. Στα πειράματα που έγιναν προς αξιολόγηση της μεθόδου χρησιμοποιήσαμε το ανθρωπομορφικό χέρι ADROIT και πραγματοποιήσαμε λαβή 3 δακτύλων με τον τρόπο που περιγράψαμε.

Από τα πειράματα αυτά μπορούμε να εξάγουμε πλήθος συμπερασμάτων. Αρχικά, γίνεται κατανοητό πως η ενισχυτική μάθηση χωρίς πρότερη γνώση, δηλαδή χωρίς χρήση δεδομένων επίδειξης, αποτυγχάνει να λύσει το πρόβλημα της ανύψωσης με την συγκεκριμένη ζητούμενη λαβή επαφής, ενώ η απόδοση δεν είναι ικανοποιητική. Συγκεκριμένα, χρειάστηκαν όπως είδαμε πάρα πολλά δείγματα ώστε το σύστημα να οδηγηθεί έστω σε μία υποβέλτιστη λύση, κατά την οποία τα δάκτυλα κλείνουν εγκλωβίζοντας το αντικείμενο σε κάποιο εσωτερικό σημείο της παλάμης. Η λαβή αυτή είναι μεν in hand αλλά όχι με επαφή των ακροδακτύλων. Κάτι τέτοιο, το οποίο εξαρχής έχουμε θεωρήσει ότι δεν προσφέρει πολλές δυνατότητες χειρισμού φαίνεται και από το αποτέλεσμα καθώς το αντικείμενο εγκλωβίζεται σε ένα τυχαίο ύψος, το οποίο δεν μπορεί να αλλάξει. Είδαμε λοιπόν στην συνέχεια ότι η επίδραση των δεδομένων επίδειξης είναι πολύ σημαντική, καθώς εκπαιδεύοντας μία αρχική πολιτική βάση αυτών, επιτυγχάνεται η ζητούμενη λαβή επαφής, ενώ στην συνέχεια χρειάζονται πολύ λιγότερα δείγματα για την εκπαίδευση με ενισχυτική μάθηση.

Στην συνέχεια προσθέτοντας τους αισθητήρες δύναμης μπορέσαμε να διατηρήσουμε τις δυνάμεις επαφής σε κάποιο επιθυμητό όριο, ισορροπώντας τις δυνάμεις από τον δείκτη και τον μέσο. Ωστόσο, παρατηρήσαμε ότι η βελτιστοποίηση σύμφωνα με έναν επιθυμητό στόχο δημιουργεί διάφορα προβλήματα. Αρχικά, δεν μπορούμε να ελέγξουμε την συμπεριφορά καθ' όλη την διάρκεια της κίνησης, παρά μόνο να ορίσουμε την επιθυμητή τελική κατάσταση. Ως αποτέλεσμα, είδαμε ότι η βελτιστοποίηση με αυτόν τον τρόπο οδηγεί και ως προς την ταχύτητα επίτευξης του στόχου όπως είναι λογικό, το οποίο με την σειρά του οδηγεί σε overshoot των τελικών επιθυμητών μεγεθών. Είδαμε, ότι όλα τα μεγέθη που μελετήσαμε,

το ύψος, η γωνία ανύψωσης από την οριζόντια θέση και οι δυνάμεις, ενώ καταλήγουν στις επιθυμητές τιμές ή όρια που έχουμε θέσει, τα υπερβαίνουν κατά την διάρκεια της κίνησης.

Κατανοώντας ότι αυτό είναι ένα κύριο πρόβλημα σε εργασίες χειρισμού ευαίσθητων αντικειμένων και μεγάλων απαιτήσεων σχετικά με την ακρίβεια, προτείναμε μία μέθοδο μάθησης βέλτιστης απόφασης ανά επόμενο χρονικό στόχο, και συγκεκριμένα για την εφαρμογή μας στόχο ύψους. Αξιολογώντας την μέθοδο βάσει των πειραμάτων που πραγματοποιήθηκαν, είδαμε ότι η μέθοδός μας καταφέρνει την επιθυμητή παρακολούθηση τροχιάς, ενώ παράλληλα έχει ιδιότητες γενίκευσης ως προς το τελικό ύψος και τον χρόνο ανύψωσης. Ουσιαστικά, το σύστημά μας μαθαίνοντας βέλτιστα ζεύγη κατάστασης-δράσης μπορεί θεωρητικά να έχει περισσότερες ιδιότητες γενίκευσης από τις δύο αυτές που αναφέραμε, αλλά στην συγκεκριμένη εφαρμογή της παρούσας εργασίας δεν εκρίθη σκόπιμη μια περαιτέρω διερεύνηση.

Είδαμε, επιπλέον στην διαδικασία αυτή την ουσιαστική συμβολή των αισθητήρων δύναμης ως μέσο και πληροφορία διατήρησης της επαφής των άκρων των δακτύλων. Στην απλή περίπτωση που δεν είχαμε παρακολούθηση τροχιάς δεν ήταν τόσο σημαντική καθώς η ανύψωση γινόταν πάρα πολύ γρήγορα, στην οποία ακόμη και μία διακοπή της επαφής δεν θα είχε πολύ μεγάλη επίδραση. Ωστόσο, κατά την παρακολούθηση της τροχιάς η διακοπή της επαφής είδαμε ότι οδηγούσε σε ένα ¨ταρακούνημα¨ του αντικειμένου το οποίο φαίνεται ως ταλάντωση στα μεγέθη του. Χρησιμοποιώντας την πληροφορία από τις δυνάμεις επαφής μπορέσαμε να διορθώσουμε την συμπεριφορά αυτήν, αλλά και να ελέγξουμε το εύρος των δυνάμεων που ασκούνται και να κατανείμουμε την δύναμη στα δάκτυλα του δείκτη και μέσου. Σε αυτό το τελευταίο πείραμα, έγινε φανερό ότι μία ζητούμενη συμπεριφορά διαφορετική από τα δεδομένα επίτευξης, χρειάζεται αρκετά περισσότερα δείγματα ώστε να επιτευχθεί, καθώς χρειάζεται όπως είναι αναμενόμενο περισσότερη εξερεύνηση από πλευράς ενισχυτικής μάθησης.

Τέλος, είδαμε ότι χωρίζοντας το πρόβλημα μέσω μίας μεταβλητής φάσης, σε ένα στάδιο σχηματισμού της λαβής και σε ένα στάδιο ανύψωσης καταφέραμε να βελτιώσουμε την απόδοση. Η βελτίωση αυτή οφείλεται στον σωστό σχηματισμό της διάταξης του χεριού την στιγμή που ξεκινά η ανύψωση, ο οποίος επιτυγχάνεται από την μεταβλητή φάσης ως πληροφορία έναρξης της κίνησης του αντικειμένου αλλά και πληροφορία που οδηγεί στον συγχρονισμό του πρώτου σταδίου ανάμεσα στα δεδομένα επίδειξης.

7.2 Μελλοντικές Επεκτάσεις

Η μεθοδολογία που αναπτύχθηκε σε αυτή την εργασία καθώς και η εφαρμογή μπορούν να επεκταθούν προς νέες κατευθύνσεις. Αρχικά, θα ήταν σημαντικό η παρούσα εφαρμογή να δοκιμαστεί σε πραγματικό περιβάλλον, που όπως είναι προφανές παρουσιάζει αρκετές προκλήσεις σε σχέση με ένα περιβάλλον προσομοίωσης, το οποίο ακόμα και εάν προσεγγίζει σε μεγάλο βαθμό τον φυσικό κόσμο δεν παύει να αποτελεί ένα μοντέλο του. Κάποιες ιδέες ως προς ένα πραγματικό περιβάλλον θα μπορούσαν να είναι μία πραγματική αλληλεπίδραση του ανθρώπου με πραγματικό αντικείμενο και μεταφορά της συμπεριφοράς αυτής στο ρομπότ. Με άλλα λόγια, η απαιτούμενη συμπεριφορά, όπως για παράδειγμα η προκαθορισμένη τροχιά και τα όρια των δυνάμεων που θέτουμε να προέρχονται από τα δεδομένα επίδειξης, ώστε το ρομπότ να μιμείται μία συμπεριφορά η οποία επιδεικνύεται από κάποιον

expert.

Επιπλέον, θα ήταν σκόπιμο η μεθοδολογία που προτείνουμε σχετικά με την μάθηση που αφορά παρακολούθηση τροχιών να δοκιμαστεί σε πολυδιάστατες τροχιές πολλών μεγεθών. Για παράδειγμα στην παρούσα εφαρμογή, θα μπορούσαμε αντί για τροχιές του ύψους να έχουμε τροχιές ολόκληρης της τρισδιάστατης θέσης, ή επίσης τροχιές του τρισδιάστατου προσανατολισμού, που θα αφορούν για παράδειγμα μία επιθυμητή τροχιά στροφής του αντικειμένου. Η μεθοδολογία που παρουσιάσαμε θεωρητικά επιτρέπει τις πολυδιάστατες τροχιές αλλά μία πειραματική αξιολόγηση είναι απαραίτητη.

Όσον αφορά τα δεδομένα επίδειξης, αυτά θα μπορούσαν να συνεισφέρουν εκτός από μία αρχικοποίηση της πολιτικής και στην διαδικασία της ενισχυτικής μάθησης. Κάτι τέτοιο θα βελτίωνε πιθανώς την απόδοση της μάθησης. Ένας τρόπος με τον οποίο θα μπορούσε να γίνει αυτό είναι με την εισαγωγή ενός όρου επιβλεπόμενης μάθησης στην κλίση πολιτικής (policy gradient), όπως για παράδειγμα γίνεται στον αλγόριθμο DAPG [2]. Ωστόσο, αυτό απαιτεί υψηλή ποιότητα δεδομένων επίδειξης κάτι το οποίο δεν επέτρεπε ο υλικός εξοπλισμός αυτής της εργασίας. Ένα γάντι δεδομένων, για παράδειγμα, θα κατέγραφε με αρκετά μεγαλύτερη ακρίβεια την κίνηση του ανθρώπου.

Βιβλιογραφία

[1] Richard S Sutton, Andrew G Barto και others. *Introduction to reinforcement learning*, τόμος 135. MIT press Cambridge, 1998.

[2] 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*. *arXiv preprint arXiv:1709.10087*, 2017.

[3] *Leap Motion Company*. https://developer.leapmotion.com/.

[4] *Shadow Hand Company*. https://www.shadowrobot.com.

[5] Raphael Deimel και Oliver Brock. *A novel type of compliant and underactuated robotic hand for dexterous grasping*. *The International Journal of Robotics Research*, 35(1-3):161–185, 2016.

[6] Abhishek Gupta, Clemens Eppner, Sergey Levine και Pieter Abbeel. *Learning dexterous manipulation for a soft robotic hand from human demonstrations*. *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, σελίδες 3786–3793. IEEE, 2016.

[7] Igor Mordatch, Zoran Popović και Emanuel Todorov. *Contact-invariant optimization for hand manipulation*. *Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation*, σελίδες 137–144, 2012.

[8] Michael Posa, Cecilia Cantu και Russ Tedrake. *A direct method for trajectory optimization of rigid bodies through contact*. *The International Journal of Robotics Research*, 33(1):69–81, 2014.

[9] Vikash Kumar, Yuval Tassa, Tom Erez και Emanuel Todorov. *Real-time behaviour synthesis for dynamic hand-manipulation*. *2014 IEEE International Conference on Robotics and Automation (ICRA)*, σελίδες 6808–6815. IEEE, 2014.

[10] Timothy Paul Lillicrap, Jonathan James Hunt, Alexander Pritzel, Nicolas Manfred Otto Heess, Tom Erez, Yuval Tassa, David Silver και Daniel Pieter Wierstra. *Continuous control with deep reinforcement learning*, 2017. ΥΣ Πατεντ Αππ. 15/217,758.

[11] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan και Philipp Moritz. *Trust region policy optimization*. *International conference on machine learning*, σελίδες 1889–1897, 2015.

[12] Shixiang Gu, Ethan Holly, Timothy Lillicrap και Sergey Levine. *Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates*. *2017 IEEE international conference on robotics and automation (ICRA)*, σελίδες 3389–3396. IEEE, 2017.

[13] A. Ghadirzadeh, A. Maki, D. Kragic και M. Björkman. *Deep predictive policy training using reinforcement learning*. *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, σελίδες 2351–2358. IEEE, 2017.

[14] Auke Jan Ijspeert, Jun Nakanishi και Stefan Schaal. *Movement imitation with nonlinear dynamical systems in humanoid robots*. *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292)*, τόμος 2, σελίδες 1398–1403. IEEE, 2002.

[15] Jens Kober και Jan R Peters. *Policy search for motor primitives in robotics*. *Advances in neural information processing systems*, σελίδες 849–856, 2009.

[16] Jan Peters και Stefan Schaal. *Reinforcement learning of motor skills with policy gradients*. *Neural networks*, 21(4):682–697, 2008.

[17] Mel Vecerik, Todd Hester, Jonathan Scholz, Fumin Wang, Olivier Pietquin, Bilal Piot, Nicolas Heess, Thomas Rothörl, Thomas Lampe και Martin Riedmiller. *Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards*. *arXiv preprint arXiv:1707.08817*, 2017.

[18] Divye Jain, Andrew Li, Shivam Singhal, Aravind Rajeswaran, Vikash Kumar και Emanuel Todorov. *Learning deep visuomotor policies for dexterous hand manipulation*. *2019 International Conference on Robotics and Automation (ICRA)*, σελίδες 3636–3643. IEEE, 2019.

[19] Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Xi Chen, Ken Goldberg και Pieter Abbeel. *Deep imitation learning for complex manipulation tasks from virtual reality teleoperation*. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, σελίδες 1–8. IEEE, 2018.

[20] Yuke Zhu, Ziyu Wang, Josh Merel, Andrei Rusu, Tom Erez, Serkan Cabi, Saran Tunyasuvunakool, János Kramár, Raia Hadsell, Nandode Freitas και others. *Reinforcement and imitation learning for diverse visuomotor skills*. *arXiv preprint arXiv:1802.09564*, 2018.

[21] Yevgen Chebotar, Mrinal Kalakrishnan, Ali Yahya, Adrian Li, Stefan Schaal και Sergey Levine. *Path integral guided policy search*. *2017 IEEE international conference on robotics and automation (ICRA)*, σελίδες 3381–3388. IEEE, 2017.

[22] Aravind Srinivas, Allan Jabri, Pieter Abbeel, Sergey Levine και Chelsea Finn. *Universal planning networks*. *arXiv preprint arXiv:1804.00645*, 2018.

[23] OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray και others. *Learning dexterous in-hand manipulation*. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

[24] Matthias Plappert, Marcin Andrychowicz, Alex Ray, Bob McGrew, Bowen Baker, Glenn Powell, Jonas Schneider, Josh Tobin, Maciek Chociej, Peter Welinder και others. *Multi-goal reinforcement learning: Challenging robotics environments and request for research*. *arXiv preprint arXiv:1802.09464*, 2018.

[25] Andrew Melnik, Luca Lach, Matthias Plappert, Timo Korthals, Robert Haschke και Helge Ritter. *Tactile sensing and deep reinforcement learning for in-hand manipulation tasks*. *IROS Workshop on Autonomous Object Manipulation*, 2019.

[26] Timo Korthals, Andrew Melnik, Marc Hesse και Jürgen Leitner. *Multisensory assisted in-hand manipulation of objects with a dexterous hand*. *2019 IEEE International Conference on Robotics and Automation Workshop on Integrating Vision and Touch for Multimodal and Cross-modal Perception,(ViTac) 2019, Montreal, CA, May 20-25, 2019*, 2019.

[27] Herke Van Hoof, Tucker Hermans, Gerhard Neumann και Jan Peters. *Learning robot in-hand manipulation with tactile features*. *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, σελίδες 121–127. IEEE, 2015.

[28] Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Edward H Adelson και Sergey Levine. *More than a feeling: Learning to grasp and regrasp using vision and touch*. *IEEE Robotics and Automation Letters*, 3(4):3300–3307, 2018.

[29] Hamza Merzić, Miroslav Bogdanović, Daniel Kappler, Ludovic Righetti και Jeannette Bohg. *Leveraging contact forces for learning to grasp*. *2019 International Conference on Robotics and Automation (ICRA)*, σελίδες 3615–3621. IEEE, 2019.

[30] Ronald J Williams. *Simple statistical gradient-following algorithms for connectionist reinforcement learning*. *Machine learning*, 8(3-4):229–256, 1992.

[31] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan και Pieter Abbeel. *High-dimensional continuous control using generalized advantage estimation*. *arXiv preprint arXiv:1506.02438*, 2015.

[32] Sham M Kakade. *A natural policy gradient*. *Advances in neural information processing systems*, σελίδες 1531–1538, 2002.

[33] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang και others. *End to end learning for self-driving cars*. *arXiv preprint arXiv:1604.07316*, 2016.

[34] Dean A Pomerleau. *Alvinn: An autonomous land vehicle in a neural network*. *Advances in neural information processing systems*, σελίδες 305–313, 1989.

[35] Stéphane Ross, Geoffrey Gordon και Drew Bagnell. *A reduction of imitation learning and structured prediction to no-regret online learning*. *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, σελίδες 627–635, 2011.

[36] Stéphane Ross και Drew Bagnell. *Efficient reductions for imitation learning*. *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, σελίδες 661–668, 2010.

[37] Alexandre Attia και Sharone Dayan. *Global overview of imitation learning*. *arXiv preprint arXiv:1801.06503*, 2018.

[38] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell και Anind K Dey. *Maximum entropy inverse reinforcement learning*. *Aaai*, τόμος 8, σελίδες 1433–1438. Chicago, IL, USA, 2008.

[39] Jonathan Ho και Stefano Ermon. *Generative adversarial imitation learning*. *Advances in neural information processing systems*, σελίδες 4565–4573, 2016.

[40] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville και Yoshua Bengio. *Generative adversarial nets*. *Advances in neural information processing systems*, σελίδες 2672–2680, 2014.

[41] Chelsea Finn, Sergey Levine και Pieter Abbeel. *Guided cost learning: Deep inverse optimal control via policy optimization*. *International conference on machine learning*, σελίδες 49–58, 2016.

[42] Justin Fu, Katie Luo και Sergey Levine. *Learning robust rewards with adversarial inverse reinforcement learning*. *arXiv preprint arXiv:1710.11248*, 2017.

[43] Xue Bin Peng, Angjoo Kanazawa, Sam Toyer, Pieter Abbeel και Sergey Levine. *Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow*. *arXiv preprint arXiv:1810.00821*, 2018.

[44] Ahmed H Qureshi, Byron Boots και Michael C Yip. *Adversarial imitation via variational inverse reinforcement learning*. *arXiv preprint arXiv:1809.06404*, 2018.

[45] Aravind Rajeswaran, Kendall Lowrey, Emanuel V Todorov και Sham M Kakade. *Towards generalization and simplicity in continuous control*. *Advances in Neural Information Processing Systems*, σελίδες 6550–6561, 2017.

[46] Vikash Kumar, Zhe Xu και Emanuel Todorov. *Fast, strong and compliant pneumatic actuation for dexterous tendon-driven hands*. *2013 IEEE international conference on robotics and automation*, σελίδες 1512–1519. IEEE, 2013.

[47] Emanuel Todorov, Tom Erez και Yuval Tassa. *Mujoco: A physics engine for model-based control*. *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, σελίδες 5026–5033. IEEE, 2012.

[48] Vikash Kumar, Abhishek Gupta, Emanuel Todorov και Sergey Levine. *Learning dexterous manipulation policies from experience and imitation*. *arXiv preprint arXiv:1611.05095*, 2016.

[49] Ilija Radosavovic, Xiaolong Wang, Lerrel Pinto και Jitendra Malik. *State-Only Imitation Learning for Dexterous Manipulation*. *arXiv preprint arXiv:2004.04650*, 2020.

[50] David Silver. *Reinforcement Learning*. University Lecture.

[51] Anusha Nagabandi, Kurt Konolige, Sergey Levine και Vikash Kumar. *Deep dynamics models for learning dexterous manipulation*. *Conference on Robot Learning*, σελίδες 1101–1112, 2020.

[52] Vikash Kumar, Emanuel Todorov και Sergey Levine. *Optimal control with learned local models: Application to dexterous manipulation*. *2016 IEEE International Conference on Robotics and Automation (ICRA)*, σελίδες 378–383. IEEE, 2016.

[53] *OpenAI*. https://openai.com/.

[54] Brian D Ziebart, J Andrew Bagnell και Anind K Dey. *Modeling interaction via the principle of maximum causal entropy*. 2010.

[55] Chelsea Finn, Paul Christiano, Pieter Abbeel και Sergey Levine. *A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models*. *arXiv preprint arXiv:1611.03852*, 2016.

[56] Jens Kober, J Andrew Bagnell και Jan Peters. *Reinforcement learning in robotics: A survey*. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

[57] Yevgen Chebotar, Oliver Kroemer και Jan Peters. *Learning robot tactile sensing for object manipulation*. *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, σελίδες 3368–3375. IEEE, 2014.

[58] Mayur Mudigonda, Pulkit Agrawal, Michael Deweese και Jitendra Malik. *Investigating deep reinforcement learning for grasping objects with an anthropomorphic hand*. 2018.

[59] Brandon Amos, Laurent Dinh, Serkan Cabi, Thomas Rothörl, Sergio Gómez Colmenarejo, Alistair Muldal, Tom Erez, Yuval Tassa, Nandode Freitas και Misha Denil. *Learning awareness models*. *arXiv preprint arXiv:1804.06318*, 2018.

[60] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford και Oleg Klimov. *Proximal policy optimization algorithms*. *arXiv preprint arXiv:1707.06347*, 2017.

[61] Ashvin Nair, Bob McGrew, Marcin Andrychowicz, Wojciech Zaremba και Pieter Abbeel. *Overcoming exploration in reinforcement learning with demonstrations*. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, σελίδες 6292–6299. IEEE, 2018.

[62] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel και Wojciech Zaremba. *Hindsight experience replay*. *Advances in neural information processing systems*, σελίδες 5048–5058, 2017.

[63] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine και Google Brain. *Time-contrastive networks: Self-supervised learning from video*. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, σελίδες 1134–1141. IEEE, 2018.

[64] Deepak Pathak, Parsa Mahmoudieh, Guanghao Luo, Pulkit Agrawal, Dian Chen, Yide Shentu, Evan Shelhamer, Jitendra Malik, Alexei A Efros και Trevor Darrell. *Zero-shot visual imitation*. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, σελίδες 2050–2053, 2018.

[65] Pratyusha Sharma, Lekha Mohan, Lerrel Pinto και Abhinav Gupta. *Multiple interactions made easy (mime): Large scale demonstrations data for imitation*. *arXiv preprint arXiv:1810.07121*, 2018.

[66] Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel και Sergey Levine. *Sfv: Reinforcement learning of physical skills from videos*. *ACM Transactions on Graphics (TOG)*, 37(6):1–14, 2018.

[67] Xue Bin Peng, Pieter Abbeel, Sergey Levine και Michielvan de Panne. *Deepmimic: Example-guided deep reinforcement learning of physics-based character skills*. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018.

[68] Γιώργος Βελέντζας. *Προσαρμοστική Ενισχυτική Μηχανική Μάθηση για την ανάπτυξη Ρομποτικών Δεξιοτήτων σε Δυναμικά Περιβάλλοντα*. Διπλωματική εργασία, Εθνικό Μετσόβιο Πολυτεχνείο, 2018.

[69] Αθανάσιος Μητράκης. *Εκπαίδευση Ρομποτικών Χειρονομιών Βάσει Πρωτογενών Δυναμικών Κινήσεων Σε Περιβάλλον Αλληλεπίδρασης Ανθρώπου-Ρομπότ*. Διπλωματική εργασία, Εθνικό Μετσόβιο Πολυτεχνείο, 2019.

[70] Meinard Müller. *Dynamic Time Warping, chapter 4*. *Information Retrieval for Music and Motion*, σελίδες 69–84.

[71] Meinard Müller. *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer, 2015.