ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ

Εκμάθηση ρομποτικής κίνησης με χρήση αλγορίθμου ενισχυτικής μάθησης και ανάδραση δύναμης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Θεοφανίας Γ. Καράμπελα

Επιβλέπων : Κωνσταντίνος Σ. Τζαφέστας
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Σεπτέμβριος 2022

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ

Εκμάθηση ρομποτικής κίνησης με χρήση αλγορίθμου ενισχυτικής μάθησης και ανάδραση δύναμης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Θεοφανίας Γ. Καράμπελα

Επιβλέπων : Κωνσταντίνος Σ. Τζαφέστας
Αναπληρωτής Καθηγητής

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 6η Σεπτεμβρίου 2022.

.....	.....	.....
Κωνσταντίνος Τζαφέστας	Αλέξανδρος Ποταμιάνος	Χαράλαμπος Ψυλλάκης
Αναπληρωτής Καθηγητής Ε.Μ.Π.	Αναπληρωτής Καθηγητής Ε.Μ.Π.	Λέκτορας Ε.Μ.Π.

Αθήνα, Σεπτέμβριος 2022

.....

Θεοφανία Γ. Καράμπελα

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η ανάπτυξη επιδέξιων ρομποτικών λαβών στη γεωργία, και συγκεκριμένα στην καλλιέργεια μανιταριών, για τη συγκομιδή των προϊόντων και την προσεκτική λήψη αντικειμένων, αποτελεί σύγχρονη επιδίωξη στον τομέα της ρομποτικής. Τέτοιου είδους λαβές μπορούν να αναπτυχθούν με πλήθος μεθόδων, καθώς και με τη χρήση ποικίλων διαφορετικών ρομποτικών χειριστών, προκειμένου να προκύψουν τα επιθυμητά αποτελέσματα ως προς το είδος της λαβής, την ποιότητα του συλλεγόμενου προϊόντος και το κόστος της εργασίας, τόσο από άποψη πόρων, όσο και χρόνου. Σε αυτήν την διπλωματική εργασία παρουσιάζουμε τη χρήση μιας μεθόδου Actor-Critic, με εφαρμογή σε συνεχείς χώρους δράσεων και καταστάσεων, με στόχο την model-free εκπαίδευση ενός πράκτορα. Ο στόχος του πράκτορα είναι να καταφέρνει να εξέρχεται ενός διαδρόμου πεπερασμένου μήκους. Η γνώση του για τη δυναμική του περιβάλλοντος περιορίζεται αποκλειστικά στο πεδίο δυνάμεων-επαναφοράς που δέχεται από τα τοιχώματα του διαδρόμου. Η συγκεκριμένη εφαρμογή, αποτελεί απλούστευση του προβλήματος της επίτευξης ενός επιδέξιου εσωτερικού ρομποτικού χειρισμού (in-hand grasp), δεδομένου ότι η κίνηση του αντικειμένου μέσα στο διάδρομο, προσομοιάζει τις κινήσεις στροφής και μετατόπισης που καλείται να πραγματοποιήσει η ρομποτική λαβή κατά τη συγκομιδή των μανιταριών. Η εκπαίδευση των Actor-Critic, οι οποίοι αναπαρίστανται μέσω νευρωνικών δικτύων, γίνεται με τη χρήση Ενισχυτικής Μάθησης και πιο συγκεκριμένα, με χρήση Temporal Difference μάθησης, σε περιβάλλον αγνώστου μοντέλου.

Λέξεις κλειδιά

Ενισχυτική Μάθηση, Ρομποτική Λαβή, Ρομποτικός Χειρισμός Εσωτερικής Λαβής, Ρομποτική Κίνηση, Συνεχής Χώρος Δράσεων, Συνεχής Χώρος Καταστάσεων, Ανάδραση Δύναμης

Abstract

The development of dexterous robotic grasping in agriculture, and specifically in the mushroom cultivation, for harvesting products and carefully picking up objects, is a modern pursuit in the field of robotics. Such grippers can be developed by a number of methods, in addition to using a variety of different robotic operators to produce the desired results in terms of grip type, the quality of the collected product and the labor cost, both in terms of resources and time. In this thesis we present the use of an Actor-Critic method, with application to continuous state and action spaces, with the aim of model-free training of an agent. The goal of the agent is to exit a finite-length corridor. Its knowledge of the dynamics of the environment is limited exclusively to the force-feedback field it receives from the corridor walls. This specific application is a simplification of the problem of achieving an in-hand grasp, since the rotation and translation movement of the object in the corridor simulates the twisting and lifting movements of the robotic gripper when harvesting mushrooms. The training of the Actor-Critic, which are represented through neural networks, is accomplished by using Reinforcement Learning and more specifically, by Temporal Difference learning, with an unknown model of the environment.

Keywords

Reinforcement Learning, Robotic Grasping, Robotic In-hand Grasp Manipulation, Robotic Motion, Continuous Action Space, Continuous State Space, Force Feedback

στον παππού μου

Ευχαριστίες

Θα ήθελα πρωτίστως να ευχαριστήσω τον καθηγητή Κωνσταντίνο Τζαφέστα, τόσο για τη βοήθειά του και την καθοδήγηση που μου έδωσε κατά τη διάρκεια της εκπόνησης της εργασίας μου, όσο για την ευκαιρία που έδωσε να ασχοληθώ με θέματα του ερευνητικού του τομέα. Επιπλέον, ευχαριστώ τον υποψήφιο διδάκτωρ Παρασκευά Οικονόμου, για τις συμβουλές του επί του θέματος μελέτης της εργασίας. Τέλος, θα ήθελα να ευχαριστήσω πολύ τους γονείς μου για τη στήριξή τους και την εμπιστοσύνη που μου έχουν δείξει σε κάθε μου βήμα.

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Ρομποτική και εξέλιξη

Η ρομποτική ασχολείται με τη μελέτη των μηχανών που μπορούν να αντικαταστήσουν τον άνθρωπο σε διάφορες διεργασίες που εκτελεί, αυτοματοποιώντας τις κινήσεις του και κατ' επέκταση διευκολύνοντας την καθημερινότητά του. Με το πέρας των χρόνων, ο άνθρωπος έχει αναζητήσει μεθόδους που θα μπορέσουν να μιμηθούν τις κινήσεις και τη συμπεριφορά του. Η ρομποτική, επομένως, έχει βαθιές ιστορικές ρίζες.

Η βασική συνιστώσα ενός ρομπότ είναι το μηχανικό σύστημα (mechanical system), το οποίο εξοπλίζεται με μια συσκευή κίνησης (τροχοί, μηχανικά πόδια) και μια συσκευή χειρισμού (τελικά εργαλεία δράσης, μηχανικά χέρια). Η ικανότητα άσκησης μιας δράσης, τόσο μετακίνησης όσο και χειρισμού, παρέχεται από το σύστημα επενέργησης (actuation system), το οποίο κινεί τα μηχανικά μέρη του ρομπότ. Η ικανότητα για αντίληψη ανατίθεται σε ένα αισθητήριο σύστημα (sensory system) το οποίο μπορεί να αποκτήσει δεδομένα για την εσωτερική κατάσταση του μηχανικού συστήματος, όπως επίσης και στην εξωτερική κατάσταση του περιβάλλοντος. Τέλος, η ικανότητα σύνδεσης της δράσης στην αντίληψη με έναν ευφυή τρόπο παρέχεται από ένα σύστημα ελέγχου (control system) που μπορεί να δίνει εντολές για την εκτέλεση της δράσης σε σχέση με τους στόχους που ορίζονται από μια τεχνική σχεδιασμού (planning) εργασίας, όπως επίσης και από τους περιορισμούς που επιβάλλονται από το ρομπότ και το περιβάλλον στο οποίο βρίσκεται [6].

Η χρήση των ρομπότ στη σημερινή εποχή συναντάται σε πολλές εργασίες τόσο σε οικιακό, όσο σε βιομηχανικό επίπεδο. Τα τελευταία χρόνια, η έρευνα εστιάζει στην επίτευξη ρομποτικών εργασιών, όπως ρομποτικών λαβών (robotic grasping) και στη χειραγώγηση αντικειμένων (manipulation). Αυτό έχει ως αποτέλεσμα την εξέλιξη των άκαμπτων ρομποτικών χειριστών με εξειδικευμένες λαβές, σε επιδέξιους ανθρωποειδείς ρομποτικούς βραχίονες και χέρια που είναι ικανά να πιάνουν και να χειρίζονται ένα ευρύ φάσμα αντικειμένων της καθημερινότητας, εκτελώντας μια ποικιλία εργασιών σε περιβάλλοντα χωρίς συγκεκριμένη – προκαθορισμένη δομή.

Οι νέες απαιτήσεις συνοδεύονται από νέες προκλήσεις και ευκαιρίες. Ως απάντηση στην αβεβαιότητα στην αντίληψη του ρομπότ και στη δυναμική του περιβάλλοντος, αναφορικά στη φυσική αλληλεπίδραση με το ακαθόριστο περιβάλλον, έχουν ερευνηθεί και ακμάσει οι τομείς της σχεδίασης υλικού και ευφυούς λογισμικού. Ο ρομποτικός βραχίονας και το χέρι γίνονται πιο επιδέξια, απτικά, στιβαρά στην αβεβαιότητα, ανθεκτικά, ελαφριά και πιο φθηνά. Ταυτόχρονα, το τρέχον έξυπνο λογισμικό δίνει έμφαση στην ευελιξία, την προσαρμοστικότητα και την ικανότητα μάθησης. Παράλληλα, μεγάλη πρόοδος έχει σημειωθεί στην κατασκευή νέων ρομποτικών χεριών και αισθητήρων, σχετικά με το πώς τα απτικά σήματα επιτρέπουν την εξερεύνηση ενός νέου αντικειμένου και κατ' επέκταση ενός νέου περιβάλλοντος, χάρη στην ανθρώπινη ικανότητα να πιάνει αντικείμενα [7].

Σήμερα, πολλοί ερευνητές του τομέα της ρομποτικής σε όλο τον κόσμο είναι αφοσιωμένοι στα ερευνητικά θέματα που προτείνει το Technical Committee. Πολλά διεθνή έργα, όπως τα έργα της Ευρωπαϊκής Ένωσης GRASP [8] και Hand Embodied [41] έχουν χρηματοδοτηθεί για να απαντηθούν οι νέες ανακύπτουσες προκλήσεις. Πλατφόρμες ρομποτικών χεριών ανοιχτού πηγαίου κώδικα (Yale OpenHand [9], Open Hand Project [10]), προσομοιωτές λαβής (GraspIt! [11], OpenGRASP [12]) και βάσεις δεδομένων ανθρώπινων και ρομποτικών χεριών (HandCorpus [13], Columbia Grasp Database [14], Human Grasping Database [15]) εμφανίστηκαν τα τελευταία χρόνια και δημιούργησαν βάσεις για καινοτομία στο χώρο των ρομποτικών λαβών.

1.2 Τα είδη της μάθησης

Στα πλαίσια της εν λόγω εργασίας, κρίνεται χρήσιμη και απαραίτητη μια σύντομη, αλλά ταυτόχρονα περιεκτική εισαγωγή στην Ενισχυτική Μάθηση, με σχετική εμβάθυνση στα εργαλεία που θα χρησιμοποιηθούν στη συνέχεια της μελέτης.

Η ενισχυτική μάθηση διαφέρει από την επιβλεπόμενη μάθηση, η οποία μελετάται εκτενώς στον τομέα της μηχανικής μάθησης τη δεδομένη περίοδο. Η επιβλεπόμενη μάθηση αφορά στη μάθηση από ένα εκπαιδευμένο σετ παραδειγμάτων – δεδομένων, τα οποία διαθέτουν περιγραφή – ετικέτα (label) για το περιεχόμενό τους και παρέχονται από έναν εξωτερικό παρατηρητή. Κάθε δεδομένο αποτελεί περιγραφή μιας κατάστασης μαζί με την περιγραφή της σωστής δράσης που πρέπει να λάβει το σύστημα όταν βρίσκεται σε αυτήν την κατάσταση. Η δράση αυτή συχνά ταυτίζεται με τον εντοπισμό μιας κατηγορίας στην οποία ανήκει η κάθε κατάσταση (classification problem). Το αντικείμενο αυτού του είδους μάθησης είναι η γενίκευση ή κατηγοριοποίηση των αποκρίσεών του, με σκοπό να ενεργεί σωστά σε καταστάσεις για τις οποίες δεν διατίθεται σετ εκπαίδευσης. Η επιβλεπόμενη μάθηση αποτελεί ένα σημαντικό είδος μηχανικής μάθησης, αλλά δεν ενδείκνυται σε περιπτώσεις μάθησης μέσω αλληλεπίδρασης. Σε διαδραστικά προβλήματα, δεν είναι πάντα πρακτικό να λαμβάνονται παραδείγματα επιθυμητής συμπεριφοράς, τα οποία να είναι αντιπροσωπευτικά και σωστά για κάθε κατάσταση στην οποία ο πράκτορας θα κληθεί να λάβει μια απόφαση. Σε περιπτώσεις αχαρτογράφητων περιοχών, ο πράκτορας θα πρέπει να μάθει να λαμβάνει αποφάσεις έπειτα από την εμπειρία που θα έχει ο ίδιος αποκτήσει μετά την εκπαίδευσή του [42].

Μια άλλη κατηγορία μηχανικής μάθησης είναι η μη – επιβλεπόμενη μάθηση. Αυτό το είδος μάθησης σχετίζεται με την εύρεση κάποιας κρυμμένης δομής σε συλλογές δεδομένων χωρίς ετικέτα (unlabeled data). Ενώ η επιβλεπόμενη και η μη – επιβλεπόμενη μάθηση μοιάζουν συμπληρωματικές τεχνικές και φαίνεται να περιγράφουν πλήρως τη μηχανική μάθηση, η ενισχυτική μάθηση αποτελεί ξεχωριστό είδος και δεν υπάγεται σε καμία εκ των ανωτέρω κατηγοριών. Ενώ φαινομενικά μοιάζει στη λογική της μη – επιβλεπόμενης μάθησης, δεδομένου ότι δεν βασίζεται σε παραδείγματα σωστών συμπεριφορών, η ενισχυτική μάθηση διαθέτει εξ ολοκλήρου διαφορετική τεχνική [42].

Εν ολίγοις, η λογική της Ενισχυτικής Μάθησης συμπυκνώνεται στο ότι αποτελεί μια προσπάθεια μεγιστοποίησης μίας αριθμητικής επιβράβευσης (reward), μέσω κατάλληλης αντιστοίχισης των καταστάσεων (states) στις πιθανές δράσεις (actions) που μπορούμε να επιλέξουμε. Στην κάθε

δράση αντιστοιχεί μια επιβράβευση ή μια ποινή, η οποία φανερώνει την ορθότητα της συγκεκριμένης δράσης και τη συνεισφορά της στη μεγιστοποίηση της συνολικής επιβράβευσης. Αυτή η μέθοδος αντικαθιστά την έννοια του επιβλέποντα που συναντήσαμε στην επιβλεπόμενη μάθηση. Κατά τη διάρκεια της μάθησης, η επιλογή των βέλτιστων δράσεων δεν είναι γνωστή, εντούτοις καλούμαστε να την ανακαλύψουμε παρατηρώντας ποιες από αυτές μας δίνουν μεγαλύτερη επιβράβευση. Δεν γνωρίζουμε εκ των προτέρων το πόσο καλή ή κακή είναι η επιλογή μιας δράσης, ωστόσο το ανακαλύπτουμε έπειτα από ορισμένα βήματα, στο τέλος μιας προσπάθειας εύρεσης μιας ικανοποιητικής ακολουθίας βημάτων αποφάσεων. Στις πιο απαιτητικές και πολύπλοκες περιπτώσεις Ενισχυτικής Μάθησης, η δράση που επιλέγουμε μπορεί να μην επηρεάζει μόνο την άμεση επιβράβευση αλλά και την επόμενη κατάσταση στην οποία θα βρεθούμε, και κατ' επέκταση, και τις επερχόμενες επιβραβεύσεις.

Μία από τις πιο συναρπαστικές πτυχές της σύγχρονης ενισχυτικής μάθησης είναι οι ουσιαστικές και γόνιμες αλληλεπιδράσεις της με άλλους κλάδους της μηχανικής και της επιστήμης. Επιπλέον, η ενισχυτική μάθηση έχει αλληλεπιδράσει έντονα και με την ψυχολογία και τη νευροεπιστήμη, με αμφίδρομα ουσιαστικά οφέλη. Από όλες τις μορφές μηχανικής μάθησης, η ενισχυτική μάθηση είναι η πιο κοντινή στο είδος της μάθησης που παρατηρείται στους ανθρώπους και τα ζώα, ενώ πολλοί από τους βασικούς αλγόριθμους της ενισχυτικής μάθησης εμπνεύστηκαν αρχικά από βιολογικά συστήματα μάθησης [42].

Έχοντας ως στόχο να καταφέρουμε να οργανώσουμε το πρόβλημα της Ενισχυτικής Μάθησης, χρησιμοποιούμε – όπως θα δούμε και στη συνέχεια – στοιχεία και ιδέες από τη θεωρία δυναμικών συστημάτων. Πιο συγκεκριμένα, κλειδί στη διαχείριση του εν λόγω προβλήματος αποτελεί ο Βέλτιστος Έλεγχος μερικώς-γνωστών Μαρκοβιανών Διαδικασιών λήψης Αποφάσεων.

1.3 Ρομποτική και Grasping μανιταριών

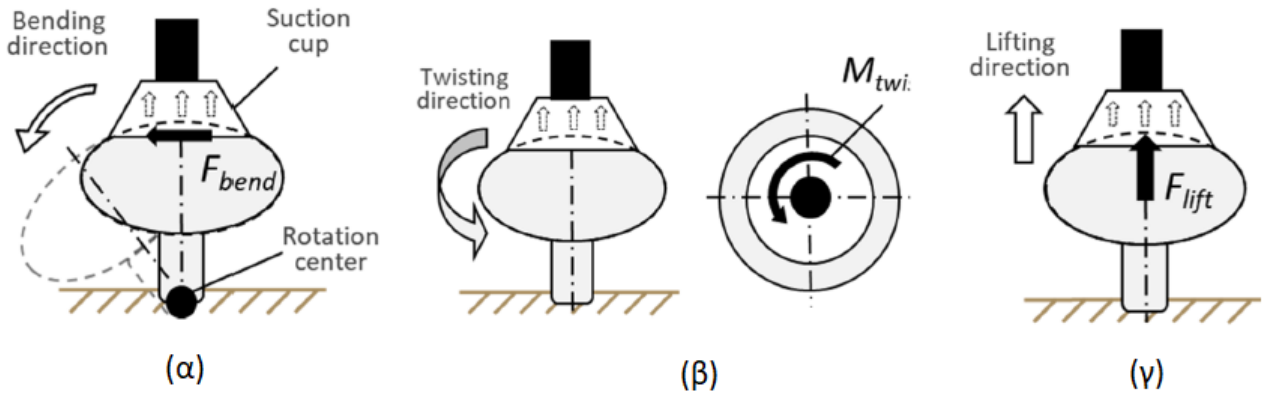
Στο χώρο της ρομποτικής, όλο και περισσότερες εφαρμογές αφορούν στην επίτευξη επιδέξιων ρομποτικών λαβών. Τόσο σε βιομηχανικά περιβάλλοντα, όσο και στη γεωργία, η χρήση αυτοματισμών μπορεί να συνεισφέρει στην εξοικονόμηση χρόνου και κόστους παραγωγής. Συγκεκριμένα, αναφορικά στη γεωργία, ένας μεγάλος αριθμός από ογκώδεις μηχανές συγκομιδής έχουν εισαχθεί και χρησιμοποιούνται στη βιομηχανία μανιταριών από τα τέλη της δεκαετίας του 1970. Η μαζική συγκομιδή έχει το πλεονέκτημα της υψηλής απόδοσης, η οποία επιτρέπει στους καλλιεργητές μανιταριών να συγκομίζουν την ίδια ποσότητα καλλιέργειας μόλις σε λίγα λεπτά, αντί για πολλές ημέρες [52]. Συνήθως, τα μανιτάρια αναπτύσσονται σε συμπλέγματα, επομένως, η διαδικασία της αποκόλλησης ενός μανιταριού από τη βάση του αποτελείται από έναν συνδυασμό κινήσεων στροφής, κάμψης και ανύψωσης [53].

Η μαζική μηχανική συγκομιδή είναι μη επιλεκτική και μπορεί να βλάψει την ποιότητα των μανιταριών. Ένα ρομποτικό σύστημα με επιλεκτική και ακριβή συγκομιδή είναι απαραίτητο για την επίτευξη καλύτερης ποιότητας καλλιέργειας [53]. Ένα ρομποτικό σύστημα συγκομιδής πρέπει πρώτα να ολοκληρώσει μια σειρά εργασιών για να καταφέρει να διαθέτει μανιτάρια υψηλής ποιότητας. Σε αυτές τις εργασίες συμπεριλαμβάνεται η αναγνώριση του μεγέθους και της τοποθεσίας της καλλιέργειας, η ήπια και προσεκτική λήψη των μανιταριών, με στόχο την μη – παραμόρφωσή και

τη μη – μόλυνσή τους, η κοπή των μπαστουνιών, τα οποία βρίσκονται στους στήμονες των μανιταριών, και η προσεκτική τοποθέτησή τους σε ένα δοχείο.

Σε σχετικές έρευνες, έχει βρεθεί πως η καλύτερη απόδοση συλλογής μανιταριών μπορεί να επιτευχθεί χρησιμοποιώντας την κάμψη ως μέθοδο αποκόλλησης σε σύγκριση με τη στροφή [54]. Επιπλέον μελέτες αποδεικνύουν ότι η συμβατική μέθοδος συγκομιδής απαιτεί περισσότερες αλλαγές προσανατολισμού για την αποκόλληση ενός μανιταριού, γεγονός που μπορεί να δημιουργήσει πολυπλοκότητα στη σχεδίαση ενός ρομποτικού end-effector με την αντίστοιχη επιδεξιότητα [53].

Λόγου χάριν, στην περίπτωση που η ρομποτική λαβή είναι ένα κύπελο κενού (vacuum cup), η λήψη του μανιταριού μπορεί να γίνει με τους τρείς ακόλουθους τρόπους. Για την κίνηση της κάμψης, τα μανιτάρια συλλέγονται με περιστροφή για μια συγκεκριμένη γωνία γύρω από τη σύνδεση στελέχους-υποστρώματος, με στόχο να σπάσουν τη σύνδεση (Σχ. 1.α). Για την περιστροφική κίνηση, τα μανιτάρια συλλέγονται με περιστροφή γύρω από τον κεντρικό άξονα του μανιταριού (Σχ. 1.β). Τέλος, για την ανυψωτική κίνηση, θεωρητικά, η μόνη απαραίτητη γνώση είναι αυτή του μεγέθους του μανιταριού (Σχ. 1.γ), [53].



Σχήμα 1: Λήψη μανιταριού με χρήση ενός κυπέλου κενού, με τρείς διαφορετικές μεθόδους: α) κάμψης, β) περιστροφής, και γ) ανύψωσης. Η F_{bend} αναπαριστά τη δύναμη κατά την κάμψη, η M_{twist} τη ροπή κατά την περιστροφή και η F_{lift} τη δύναμη κατά την ανύψωση, από [53].

Για να αποφευχθεί η ανάγκη κατασκευής πάρα πολλών ρομποτικών χειριστών για την εκμάθηση grasping πολιτικών και χειρισμού αντικειμένων, η χρήση προσομοιώσεων είναι μια πολύ ελκυστική εναλλακτική λύση. Οι προσομοιώσεις μπορούν να χρησιμοποιηθούν επειδή παρέχουν μια άφθονη πηγή δεδομένων με άψογους σχολιασμούς. Η μέθοδος domain randomization εφαρμόζει πολλές τυχαιότητες στις παρατηρήσεις ή στη δυναμική του συστήματος προσομοίωσης, έτσι ώστε η μετάβαση από την προσομοίωση στον πραγματικό κόσμο, να φανεί σαν μια ακόμα περίπτωση εφαρμογής τυχαιότητας στις παραπάνω παραμέτρους [55].

Μια πολλά υποσχόμενη και ισχυρή τεχνική για αυτόματη απόκτηση πολιτικών ελέγχου grasping μέσω trial-and-error, αποτελεί η Βαθιά Ενισχυτική Μάθηση (Deep RL). Μάλιστα, οι εκπαιδευμένες πολιτικές παρουσιάζουν ικανότητες γενίκευσης σε νέα αντικείμενα [55]. Γενικά, για να λειτουργούν αποτελεσματικά τα ρομπότ θα πρέπει να είναι ικανά να πραγματοποιούν γενικεύσεις και να

εκμεταλλεύονται τη γνώση που έχουν αποκτήσει από παλαιότερες εμπειρίες, ώστε να δομούν τη διαδικασία της μάθησης για νέες εργασίες. Πιο συγκεκριμένα, το ρομπότ μπορεί να δημιουργήσει μια αναπαράσταση του περιβάλλοντος βάσει αντικειμένων, δηλαδή, τμηματοποιώντας το σε αντικείμενα και, στη συνέχεια, εκτιμώντας τις τιμές των ιδιοτήτων τους. Αυτή η αναπαράσταση υποστηρίζει την επαναχρησιμοποίηση δεξιοτήτων επιτρέποντας στο ρομπότ να γενικεύει αποτελεσματικά παρόμοια αντικείμενα όταν τα συναντά σε διαφορετικές εργασίες. Η γενίκευση μεταξύ διαφορετικών εργασιών μπορεί να επιτευχθεί με τη δημιουργία αντιστοιχιών μεταξύ αντικειμένων και περιβαλλόντων σε επίπεδο σημείου. Αυτές οι αντιστοιχίες μπορούν, στη συνέχεια, να χρησιμοποιηθούν για την απευθείας χαρτογράφηση των δεξιοτήτων χειρισμού μεταξύ των διαφορετικών εργασιών ή για την ολοκλήρωση αναπαραστάσεων υψηλού επιπέδου [56].

Συχνά, είναι πιο αποτελεσματικό και ισχυρό να χρησιμοποιούμε χαρακτηριστικά που αντιπροσωπεύουν ομάδες αντικειμένων ως συνολική οντότητα. Ο εντοπισμός μεμονωμένων αντικειμένων μπορεί να μην είναι απαραίτητος, ενώ μπορεί ακόμη και να προσθέσει επιπλέον πολυπλοκότητα στη μάθηση, με αποτέλεσμα η μάθηση να γενικεύεται ελάχιστα ή να είναι λιγότερο ισχυρή. Η αντίληψη του ρομπότ για το περιβάλλον χωρίζεται ευρέως σε παθητική και διαδραστική αντίληψη (passive and interactive perception), με τη βασική διαφορά να είναι το εάν αλληλεπιδρά ή όχι φυσικά με το περιβάλλον. Η διαδραστική αντίληψη μπορεί επίσης να χρησιμοποιηθεί ως σήμα επίβλεψης για την εκμάθηση εκτίμησης ιδιοτήτων, χρησιμοποιώντας παθητική αντίληψη, διαδικασία γνωστή και ως αυτο-επιβλεπόμενη μάθηση [56].

1.4 Στόχος της εργασίας

Στην παρούσα εργασία θα πραγματοποιηθεί μια απλούστευση του προβλήματος της in-hand επιδέξιας λήψης ενός μανιταριού από ένα ανθρωπομορφικό ρομποτικό χέρι. Η in-hand κίνηση υπαγορεύει τη σταθερή θέση του βραχίονα του ρομπότ και τον χειρισμό του αντικειμένου αποκλειστικά με τα δάχτυλα του χεριού. Πιο συγκεκριμένα, θα μελετηθεί ο συνδυασμός των κινήσεων στροφής και ανύψωσης για την επίτευξη του παραπάνω στόχου. Ο τρόπος που θα γίνει αυτό απλοποιείται στην εκπαίδευση ενός αντικειμένου, ώστε να καταφέρνει να αποδράσει από έναν διάδρομο πεπερασμένου μήκους, με τη χρήση αλγορίθμου ενισχυτικής μάθησης και ανάδραση δύναμης. Η αντιστοιχία των δύο προβλημάτων θα εξηγηθεί αναλυτικά σε επόμενο κεφάλαιο.

1.5 Οργάνωση της εργασίας

Η εν λόγω εργασία έχει οργανωθεί σε συνολικά 8 κεφάλαια, εκ των οποίων ένα κεφάλαιο βρίσκεται στα παραρτήματα. Στο πρώτο κεφάλαιο συναντάμε την εισαγωγή, στα επόμενα τρία το θεωρητικό υπόβαθρο της εργασίας, στο 5^{ο} παρουσιάζεται η περιγραφή του πειράματος που πραγματοποιήσαμε, στο 6^{ο} τα πειραματικά αποτελέσματα και στο 7^{ο} ο επίλογος της εργασίας. Στο κεφάλαιο των παραρτημάτων συναντάμε μια προτεινόμενη αναλυτικότερη παρουσίαση της άμεσης επέκτασης του προβλήματος που μελετάμε στην εργασία.

Στο κεφάλαιο 2 παρουσιάζουμε αναλυτικά την ενισχυτική μάθηση, την οποία χρησιμοποιούμε στο πείραμά μας. Στο πλαίσιο αυτό, μελετάμε αρχικά τις Μαρκοβιανές Διαδικασίες λήψης αποφάσεων, οι οποίες συνδράμουν στην περιγραφή των προβλημάτων ενισχυτικής μάθησης. Στην πορεία, μελετάμε το δυναμικό προγραμματισμό, ως μέθοδο αντιμετώπισης προβλημάτων ακολουθιακής λογικής βάσει μοντέλου. Έπειτα, αναλύουμε τα προβλήματα ενισχυτικής μάθησης με άγνοια του μοντέλου του συστήματος. Στη συνέχεια, παρουσιάζουμε μεθόδους προσέγγισης συνάρτησης αξίας, κλίσης πολιτικής και τη λογική Δράστη-Κριτή που θα συναντήσουμε και στην πορεία.

Στο κεφάλαιο 3 αναλύεται η λογική και το κίνητρο πίσω από τον αλγόριθμο CACLA που χρησιμοποιούμε στην εφαρμογή μας. Στην αρχή, παρουσιάζουμε τον αντίστοιχο αλγόριθμο σε περιπτώσεις διακριτού χώρου δράσεων και καταστάσεων, ACLA και στη συνέχεια, την προσαρμογή του στους συνεχείς χώρους, CACLA.

Στο κεφάλαιο 4 γίνεται μια σύντομη ανάλυση των μεθόδων επίτευξης ευσταθών και επιδέξιων ρομποτικών λαβών. Πιο συγκεκριμένα, παρουσιάζονται διαφορετικοί τρόποι για να μπορέσει να εντοπιστεί και να ληφθεί ένα αντικείμενο ενδιαφέροντος από ένα ρομποτικό χέρι.

Στο κεφάλαιο 5 περιγράφουμε αναλυτικά το πρόβλημα που μελετήσαμε στην παρούσα εργασία, προσδιορίζοντας πλήρως κάθε βήμα που ακολουθήσαμε για την προσέγγισή του.

Στο κεφάλαιο 6 γίνεται παρουσίαση και σχολιασμός των πειραματικών αποτελεσμάτων της μεθόδου που ακολουθήσαμε. Πιο αναλυτικά, παρουσιάζουμε την απόδοση της εκπαίδευσης του αλγορίθμου που χρησιμοποιήσαμε στο πρόβλημά μας και οπτικοποιούμε τα αποτελέσματά της παρουσιάζοντας τις τροχιές που ακολουθεί το αντικείμενο προς μάθηση στην πειραματική μας διάταξη.

Στο κεφάλαιο 7, το οποίο αποτελεί επίλογο της εργασίας προτείνονται ορισμένες βελτιστοποιήσεις στις διαδικασίες μας και συζητούνται τα συμπεράσματα και οι επεκτάσεις της μελέτης μας.

Τέλος, στο κεφάλαιο των παραρτημάτων παρουσιάζεται σύντομα, αλλά αναλυτικότερα η άμεση προέκταση του προβλήματός μας στο πρόβλημα επίτευξης ευσταθούς και επιδέξιας ρομποτικής λήψης αντικειμένων, και συγκεκριμένα μανιταριών, καθώς επίσης παρουσιάζεται σύντομα η μέθοδος Denavit Hartenberg που χρησιμοποιούμε στο ίδιο κεφάλαιο.

Κεφάλαιο 2

Ενισχυτική Μάθηση

Η ενισχυτική μάθηση είναι μια υπολογιστική προσέγγιση για την κατανόηση και την αυτοματοποίηση της λήψης αποφάσεων. Διακρίνεται από άλλες υπολογιστικές προσεγγίσεις λόγω της έμφασης που δίνει στην εκπαίδευση ενός πράκτορα, ο οποίος μαθαίνει μέσα από την άμεση αλληλεπίδρασή του με το περιβάλλον του, χωρίς να απαιτούνται πλήρη μοντέλα του περιβάλλοντος ή η επίβλεψη του πράκτορα. Η ενισχυτική μάθηση χρησιμοποιεί τις Μαρκοβιανές διαδικασίες λήψης αποφάσεων για να περιγράψει την αλληλεπίδραση ενός μαθητευόμενου – πράκτορα και του περιβάλλοντος, μέσα από καταστάσεις, δράσεις και ανταμοιβές. Αυτό το πλαίσιο είναι ένας απλός τρόπος αναπαράστασης των βασικών χαρακτηριστικών του προβλήματος της τεχνητής νοημοσύνης. Αυτά τα χαρακτηριστικά σχετίζονται με την αίσθηση της αιτίας και του αποτελέσματος, την αίσθηση της αβεβαιότητας και του μη-ντετερμινισμού και την ύπαρξη σαφών στόχων [42].

2.1 Μαρκοβιανές Διαδικασίες λήψης αποφάσεων

Οι Μαρκοβιανές διαδικασίες λήψης αποφάσεων (Markov Decision Processes – MDPs) αποτελούν μαθηματικό κλειδί στην αναπαράσταση των προβλημάτων ενισχυτικής μάθησης. Ο φορμαλισμός τους, επιτρέπει την περιγραφή των προβλημάτων στα οποία η μάθηση επιτυγχάνεται μέσω αλληλεπίδρασης με το περιβάλλον, ειδικά όταν αυτό είναι πλήρως παρατηρήσιμο. Επιπλέον, είναι ιδανικές για την περιγραφή προβλημάτων συνεχούς λήψης αποφάσεων, στα οποία οι δράσεις δεν επηρεάζουν μόνο τις άμεσες επιβραβεύσεις, αλλά και επόμενες καταστάσεις, και μέσω αυτών, τις επόμενες – μελλοντικές επιβραβεύσεις. Σχεδόν κάθε πρόβλημα ενισχυτικής μάθησης μπορεί να περιγραφεί μέσω των Μαρκοβιανών διαδικασιών λήψης αποφάσεων. Πιο συγκεκριμένα, τα προβλήματα βέλτιστου ελέγχου σχετίζονται με συνεχείς MDPs, ενώ ακόμα και μερικώς παρατηρήσιμα προβλήματα μπορούν να αντιμετωπιστούν με χρήση MDPs.

Η ιδιότητα στην οποία βασίζεται η λογική των MDPs είναι η Μαρκοβιανή ιδιότητα, η οποία υπαγορεύει πως «το μέλλον είναι ανεξάρτητο από το παρελθόν, με δεδομένο το παρόν». Πιο αναλυτικά, μια κατάσταση S_t είναι Μαρκοβιανή αν και μόνο αν:

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

Μέσω της παραπάνω ιδιότητας φαίνεται πως στην περίπτωση που γνωρίζουμε μια κατάσταση, τότε παύει πλέον να μας ενδιαφέρει ό,τι έχει προηγηθεί, καθώς επίσης φαίνεται πως η τρέχουσα κατάσταση είναι επαρκώς ενδεικτική για το μέλλον.

Για μια Μαρκοβιανή κατάσταση s και μια ακόλουθή της s', η πιθανότητα μετάβασης κατάστασης ορίζεται ακολούθως:

$$\mathcal{P}_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s]$$

Ενώ ο πίνακας μετάβασης κατάστασης $\mathcal{P}$ έχει ως στοιχεία του τις πιθανότητες μετάβασης κατάστασης από κάθε κατάσταση s, σε κάθε ακόλουθη κατάσταση s' όπως φαίνεται παρακάτω, όπου η κάθε γραμμή του αθροίζεται στη μονάδα:

$$\mathcal{P} = \begin{bmatrix} \mathcal{P}_{11} & \cdots & \mathcal{P}_{1n} \\ \vdots & \ddots & \vdots \\ \mathcal{P}_{n1} & \cdots & \mathcal{P}_{nn} \end{bmatrix}$$

Αναφορικά στις Μαρκοβιανές διαδικασίες (Markov Processes) / Μαρκοβιανές αλυσίδες (Markov Chains), αποτελούν τυχαίες διαδικασίες χωρίς μνήμη και αναπαρίστανται ως μια τούπλα $\langle \mathcal{S}, \mathcal{P} \rangle$, όπου:

- $\mathcal{S}$: πεπερασμένο σύνολο καταστάσεων
- $\mathcal{P}$: πίνακας πιθανοτήτων μετάβασης κατάστασης, $\mathcal{P}_{ss'}$

Οι Μαρκοβιανές διαδικασίες επιβράβευσης (Markov Reward Processes - MRP) αποτελούν επέκταση των Μαρκοβιανών αλυσίδων, με προσθήκη της έννοιας της αξίας. Στην περίπτωση αυτή, αναπαρίστανται ως μια τούπλα $\langle \mathcal{S}, \mathcal{P}, R, \gamma \rangle$, όπου:

- $\mathcal{S}$: πεπερασμένο σύνολο καταστάσεων
- $\mathcal{P}$: πίνακας πιθανοτήτων μετάβασης κατάστασης, $\mathcal{P}_{ss'}$
- R: συνάρτηση επιβράβευσης, $R_s = \mathbb{E}[R_{t+1} | S_t = s]$
- γ: συντελεστής έκπτωσης (discount factor), $\gamma \in [0,1]$

Οι Μαρκοβιανές διαδικασίες λήψης αποφάσεων (Markov Decision Processes) αποτελούν επέκταση των Μαρκοβιανών διαδικασιών επιβράβευσης, με προσθήκη της έννοιας της απόφασης, δηλαδή της επιλογής μιας δράσης. Στην περίπτωση αυτή, αναπαρίστανται ως μια τούπλα $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma \rangle$, όπου:

- $\mathcal{S}$: πεπερασμένο σύνολο καταστάσεων
- $\mathcal{A}$: πεπερασμένο σύνολο δράσεων
- $\mathcal{P}$: πίνακας πιθανοτήτων μετάβασης κατάστασης, $\mathcal{P}^a_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a]$
- R: συνάρτηση επιβράβευσης, $R^a_s = \mathbb{E}[R_{t+1} | S_t = s, A_t = a]$
- γ: συντελεστής έκπτωσης (discount factor), $\gamma \in [0,1]$

Στις Μαρκοβιανές διαδικασίες λήψης αποφάσεων το αντικείμενο προς εκπαίδευση ονομάζεται πράκτορας. Κάθε τι με το οποίο αλληλεπιδρά ο πράκτορας και δεν περιλαμβάνει τον εαυτό του, αποτελεί το περιβάλλον της μάθησης. Αυτή η αλληλεπίδραση είναι συνεχής και στοχεύει στην απόκτηση εμπειρίας από τον πράκτορα για το περιβάλλον. Η λογική αυτή πηγάζει άμεσα από τον τρόπο που ένας άνθρωπος, λόγου χάρη, μαθαίνει να περπατάει. Έπειτα από επανειλημμένες προσπάθειες και μια σειρά επιτυχιών και αποτυχιών, πλέον επιλέγει τις δράσεις εκείνες που τον οδήγησαν στη διατήρηση της ισορροπίας του και στη βέλτιστη επίτευξη του στόχου του. Με τον ίδιο ακριβώς τρόπο, ο πράκτορας της ενισχυτικής μάθησης ανταλλάσσει συνεχώς πληροφορίες με το περιβάλλον, λαμβάνει επιβραβεύσεις ή ποινές για τις πράξεις του και μαθαίνει να παίρνει όλο και καλύτερες αποφάσεις στο μέλλον.

Σε μια πεπερασμένη MDP, κάθε σύνολο καταστάσεων, δράσεων και επιβραβεύσεων ($\mathcal{S}, \mathcal{A}$ και R) έχει πεπερασμένο αριθμό στοιχείων. Στην περίπτωση αυτή, οι τυχαίες μεταβλητές R_t και S_t έχουν καλώς ορισμένες διακριτές κατανομές πιθανότητας, οι οποίες εξαρτώνται μόνο από την προηγούμενη κατάσταση και δράση. Επομένως, για συγκεκριμένες τιμές αυτών των τυχαίων μεταβλητών, $s' \in \mathcal{S}$ και $r \in R$, ορίζεται η πιθανότητα εμφάνισης αυτών των τιμών τη χρονική στιγμή t, δεδομένων συγκεκριμένων τιμών για την προηγούμενη κατάσταση και δράση:

$$p(s', r|s, a) \doteq \Pr\{S_t = s', R_t = r|S_{t-1} = s, A_{t-1} = a\}, \forall s', s \in \mathcal{S}, r \in R, a \in \mathcal{A}(s) \quad (3)$$

Η συνάρτηση p αποτελεί τη δυναμική μιας MDP, και ισχύει ότι $p: \mathcal{S} \times R \times \mathcal{S} \times \mathcal{A} \to [0,1]$. Εναλλακτικά, η συνάρτηση p μπορεί να γραφεί και ως συνάρτηση με τρία ορίσματα ως ακολούθως:

$$p(s'|s, a) \doteq \Pr\{S_t = s', |S_{t-1} = s, A_{t-1} = a\} = \sum_{r \in R} p(s', r|s, a),$$

όπου $p: \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0,1]$. Στο πλαίσιο αυτό, μπορούμε να πραγματοποιήσουμε και τον υπολογισμό της εκτιμώμενης επιβράβευσης για το ζεύγος κατάσταση – δράση:

$$r(s, a) \doteq \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a] = \sum_{r \in R} r \sum_{s' \in \mathcal{S}} p(s', r|s, a),$$

όπου $r: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, αλλά και για την τριπλέτα κατάσταση – δράση – επόμενη-κατάσταση:

$$r(s, a, s') \doteq \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a, S_t = s'] = \sum_{r \in R} r \frac{p(s', r|s, a)}{p(s'|s, a)}$$

Στα πλαίσια της εν λόγω εργασίας θα γίνει χρήση της συνάρτησης p με τέσσερα ορίσματα.

Η απόδοση (return) G_t αποτελεί τη συνολική επιβράβευση, με την έκπτωση, από τη χρονική στιγμή t και έπειτα:

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

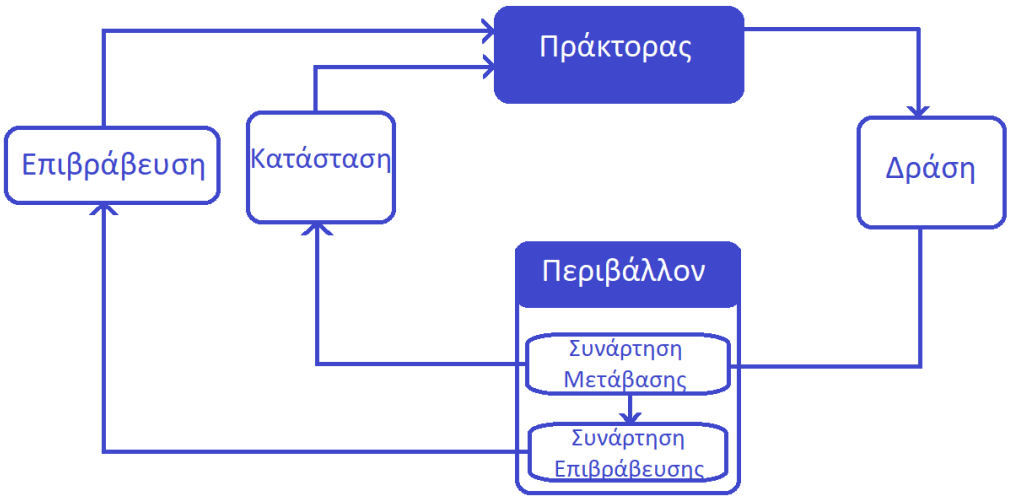
$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \cdots) = R_{t+1} + \gamma G_{t+1} \quad (1)$$

Στην περίπτωση που έχουμε γ κοντά στο 0, τότε έχουμε «μυωπική» αξιολόγηση, ενώ στην περίπτωση που το γ τείνει προς το 1, έχουμε «διορατική» αξιολόγηση.

Η αιτία της έκπτωσης – γ είναι πολυπαραγοντική καθώς χωρίς αυτή, δεν θα είχαμε πλήρη αντιπροσώπευση της αβεβαιότητας που έχουμε για το μέλλον, αλλά και σε προβλήματα οικονομικής φύσεως είτε ανθρώπινης συμπεριφοράς, υπάρχει μεγαλύτερη προτίμηση και ενδιαφέρον στην άμεση επιβράβευση – ανταμοιβή.

Ως συνάρτηση αξίας (value function) μιας MRP ορίζουμε την αναμενόμενη απόδοση, με εκκίνηση από την κατάσταση s:

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



Σχήμα 2.1: Αλληλεπίδραση πράκτορα – περιβάλλοντος σε μια MDP, προσαρμοσμένο από [42]

Όπως απεικονίζεται και στο παραπάνω σχήμα, σε κάθε βήμα του αλγορίθμου της ενισχυτικής μάθησης ακολουθούνται εναλλάξ και διαδοχικά, τρία βήματα από τον πράκτορα και τρία από το περιβάλλον. Στην αρχή ο πράκτορας εκτελεί μια δράση A_t και το περιβάλλον τη λαμβάνει. Στη συνέχεια, το περιβάλλον εκπέμπει μια παρατήρηση O_t, την οποία λαμβάνει ο πράκτορας. Τέλος, το περιβάλλον στέλνει στον πράκτορα την επιβράβευση R_t που αντιστοιχεί στην δράση A_t. Συνεπώς, μέσω αυτής της διαδικασίας αλληλεπίδρασης πράκτορα και περιβάλλοντος επιτυγχάνεται μια αντιστοίχιση δράσεων, A_t, σε καταστάσεις, S_t. Οι καταστάσεις του περιβάλλοντος είναι αυτές που καθορίζουν τι πρόκειται να συμβεί στα επόμενα βήματα. Μια

κατάσταση πληροφορίας (information state), δηλαδή μια Μαρκοβιανή κατάσταση, περιέχει χρήσιμες πληροφορίες για την ιστορία H_t (history), όπου $S_t = f(H_t)$.

Η **πολιτική** στις MDPs αποτελεί μια αντιστοίχιση των καταστάσεων σε πιθανότητες επιλογής της κάθε ενδεχόμενης δράσης. Η πολιτική των MDP εξαρτάται πάντα από την τρέχουσα κατάσταση και συνεπώς είναι πλήρως ανεξάρτητη από την ιστορία. Εάν ο πράκτορας ακολουθεί μια πολιτική π σε μια χρονική στιγμή t, τότε η $\pi(a|s)$ είναι η πιθανότητα $A_t = a$, εάν $S_t = s$. Μάλιστα, οι πολιτικές είναι στάσιμες και δεν έχουν καμία εξάρτηση από το χρόνο. Επομένως, μπορούμε πλέον να αντιληφθούμε καλύτερα τη διαφορά των MDPs από τις MRPs, καθώς δεδομένης μιας MDP με σταθερή πολιτική, το πρόβλημά μας πλέον ορίζεται ως μια MRP. Υπό το πρίσμα της πολιτικής, ορίζουμε τη **συνάρτηση κατάστασης – αξίας** κάτω από μια πολιτική π (state-value function for a policy π) ως:

$$v_\pi \doteq \mathbb{E}_\pi[G_t|S_t = s] = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s] = \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s], \forall\, s \in S,$$

όπου το $\mathbb{E}_\pi[\cdot]$ αναπαριστά την αναμενόμενη τιμή μιας τυχαίας μεταβλητής, δεδομένου ότι ο πράκτορας ακολουθεί μια πολιτική π, και το t αποτελεί οποιαδήποτε χρονική στιγμή. Με παρόμοιο τρόπο ορίζουμε και την **συνάρτηση δράσης – αξίας** κάτω από μια πολιτική π (action-value function for a policy π) ως:

$$\begin{aligned} q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t|S_t = s, A_t = a] &= \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s, A_t = a] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1})|S_t = s, A_t = a] \end{aligned}$$

Από τα παραπάνω προκύπτει πως η συνάρτηση κατάστασης – αξίας θα μπορούσε να διασπαστεί σε δύο επιμέρους τμήματα με τον ακόλουθο τρόπο:

- R_{t+1}: άμεσης επιβράβευσης
- $\gamma v(R_{t+1})$: αξία με έκπτωση της διαδοχικής κατάστασης

$$v \doteq \mathbb{E}[G_t|S_t = s] \overset{(1)}{=} \mathbb{E}[R_{t+1} + \gamma G_{t+1}|S_t = s] = \mathbb{E}[R_{t+1} + \gamma v(S_{t+1})|S_t = s] = R_s + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'} v(s')$$

Εναλλακτικά:

$$\begin{aligned} v_\pi \doteq \mathbb{E}_\pi[G_t|S_t = s] \quad &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s] \\ &= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a) \left[r + \gamma \mathbb{E}_\pi[G_{t+1}|S_{t+1} = s']\right] \\ &= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a) \left[r + \gamma v_\pi(s')\right] \end{aligned} \tag{2}$$

Η έκφραση στην οποία καταλήξαμε για το v_π ονομάζεται **εξίσωση Bellman** για τις MRPs, η οποία στη γενική της μορφή αποτελεί μια γραμμική εξίσωση ($\boldsymbol{v} = \boldsymbol{R} + \gamma \boldsymbol{P}\boldsymbol{v}$, με τη χρήση πινάκων) και μπορεί να επιλυθεί άμεσα ($\boldsymbol{v} = \boldsymbol{R} + \gamma \boldsymbol{P}\boldsymbol{v} \Rightarrow (\boldsymbol{I} - \gamma \boldsymbol{P})\boldsymbol{v} = \boldsymbol{R} \Rightarrow \boldsymbol{v} = (\boldsymbol{I} - \gamma \boldsymbol{P})^{-1}\boldsymbol{R}$). Στην περίπτωση

μεγάλων MRPs (για μεγάλα n) η εν λόγω εξίσωση παρουσιάζει μεγάλη υπολογιστική πολυπλοκότητα $O(n^3)$.

Στην περίπτωση των πεπερασμένων MDPs μπορούμε να ορίσουμε ως βέλτιστη πολιτική π, εκείνη που είναι μεγαλύτερη ή ίση με κάθε άλλη πολιτική π', σε κάθε κατάσταση. Δηλαδή, $\pi \geq \pi'$ αν και μόνο αν $v_\pi(s) \geq v_{\pi'}(s), \forall\, s\epsilon\mathcal{S}$. Η βέλτιστη πολιτική είναι τουλάχιστον μια σε κάθε πρόβλημα, συμβολίζεται ως π_* και αντιστοιχεί στη βέλτιστη συνάρτηση κατάστασης – αξίας v_*, η οποία ορίζεται ως: $v_*(s) \doteq \max_\pi v_\pi(s)\,, \forall\, s\epsilon\mathcal{S}$. Οι βέλτιστες πολιτικές επίσης αντιστοιχούν και στην ίδια βέλτιστη συνάρτηση δράσης – αξίας q_*, η οποία ορίζεται ως: $q_*(s,a) \doteq \max_\pi q_\pi(s,a)\,, \forall\, s\epsilon\mathcal{S}, a\epsilon\mathcal{A}(s)$. Η χρησιμότητα της βέλτιστης συνάρτησης αξίας έγκειται στο ότι μέσω αυτής προσδιορίζεται η καλύτερη δυνατή απόδοσης μιας MDP. Για να προσδιοριστεί η βέλτιστη πολιτική, θα πρέπει να γίνει μεγιστοποίηση της συνάρτησης δράσης – αξίας:

$$\pi_*(s|a) = \begin{cases} 1, \alpha\nu\ a = \underset{a\epsilon\mathcal{A}}{\operatorname{argmax}}\, q_*(s,a) \\ 0, \sigma\varepsilon\ \kappa\acute{\alpha}\theta\varepsilon\ \acute{\alpha}\lambda\lambda\eta\ \pi\varepsilon\rho\acute{\imath}\pi\tau\omega\sigma\eta \end{cases}$$

Δεδομένου ότι το v_* είναι η συνάρτηση αξίας για μια δεδομένη πολιτική, θα πρέπει να ικανοποιεί τη συνθήκη που υπαγορεύεται από την εξίσωση Bellman (2). Η συνθήκη αυτή θα μπορεί πλέον να γραφεί ανεξάρτητα από την πολιτική π, αφού η v_* είναι η βέλτιστη, και ονομάζεται Εξίσωση Βελτιστοποίησης Bellman, ενώ ορίζεται ως εξής:

$$v_*(s) \doteq \max_{a\epsilon\mathcal{A}(s)} q_{\pi_*}(s) = \max_a \mathbb{E}_{\pi_*}[G_t|S_t = s, A_t = a] = \max_a \mathbb{E}_{\pi_*}[R_{t+1} + \gamma G_{t+1}|S_t = s, A_t = a]$$

$$\overset{(1)}{=} \max_a \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1})|S_t = s, A_t = a] = \max_a \sum_{s',r} p(s', r|s, a)\,[r + \gamma v_*(s')]$$

Με όμοιο τρόπο προκύπτει και η Εξίσωση Βελτιστοποίησης Bellman για το q_*:

$$q_*(s,a) = \mathbb{E}\left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \middle| S_t = s, A_t = a\right] = \sum_{s',r} p(s', r|s, a)\left[r + \gamma \max_{a'} q_*(s', a')\right]$$

Για πεπερασμένες MDP, η εξίσωση Βελτιστοποίησης του Bellman για το v_* έχει μοναδική λύση. Η εν λόγω εξίσωση είναι στην πραγματικότητα ένα σύστημα εξισώσεων, μια για κάθε κατάσταση. Άρα, αν υπάρχουν n καταστάσεις, τότε υπάρχουν n εξισώσεις με n αγνώστους. Αν η δυναμική p του περιβάλλοντος είναι γνωστή, τότε το σύστημα αυτό μπορεί να λυθεί για το v_* χρησιμοποιώντας οποιαδήποτε μέθοδο επίλυσης μη-γραμμικών συστημάτων. Εναλλακτικά, μπορεί κανείς να λύσει ένα σχετικό σύνολο εξισώσεων για το q_*. Όταν γίνει ο προσδιορισμός του v_*, είναι σχετικά εύκολος ο προσδιορισμός της βέλτιστης πολιτικής π_*. Για κάθε κατάσταση s, θα υπάρξουν μία ή περισσότερες ενέργειες στις οποίες επιτυγχάνεται το μέγιστο στην Εξίσωση Βελτιστοποίησης του Bellman. Οποιαδήποτε πολιτική που δεν εισάγει μηδενική πιθανότητα σε αυτές τις δράσεις, είναι βέλτιστη πολιτική. Ένας απλός τρόπος προσέγγισης της λογικής αυτής είναι να παρατηρήσει κανείς βήμα-βήμα τη λήψη αποφάσεων από τον πράκτορα. Εάν έχουμε τη βέλτιστη συνάρτηση αξίας, v_*, τότε οι δράσεις που εμφανίζονται μετά από μια αναζήτηση ενός βήματος θα είναι βέλτιστες. Μια άλλη οπτική είναι να σκεφτούμε πως κάθε πολιτική που είναι

άπληστη, σχετικά με τη βέλτιστη συνάρτηση αξιολόγησης v_*, είναι μια βέλτιστη πολιτική. Ο όρος άπληστος χρησιμοποιείται στην επιστήμη υπολογιστών για να περιγράψει οποιαδήποτε διαδικασία αναζήτησης ή απόφασης που επιλέγει εναλλακτικές μόνο βάσει τοπικών ή άμεσων εκτιμήσεων, χωρίς να εξετάζεται η πιθανότητα ότι μια τέτοια επιλογή μπορεί να εμποδίσει μελλοντική πρόσβαση σε ακόμη καλύτερες εναλλακτικές. Κατά συνέπεια, περιγράφει πολιτικές που επιλέγουν ενέργειες με βάση μόνο τις βραχυπρόθεσμες συνέπειές τους.

2.2 Δυναμικός Προγραμματισμός

Οι μέθοδοι βελτιστοποίησης που χρησιμοποιούνται στην Ενισχυτική Μάθηση χωρίζονται σε αυτές που βασίζονται σε κάποιο μοντέλο του περιβάλλοντος (model-based) και σε αυτές που δεν βασίζονται σε μοντέλο (model-free). Ο Δυναμικός Προγραμματισμός – ΔΠ (Dynamic Programming - DP) αποτελεί έναν τύπο αλγορίθμων οι οποίοι χρησιμοποιούνται στην αντιμετώπιση προβλημάτων ακολουθιακής λογικής (sequential problems), για τον υπολογισμό βέλτιστων πολιτικών, δεδομένου ενός τέλειου μοντέλου για το περιβάλλον (model-based, planning problems), ως μια Μαρκοβιανή Διαδικασία λήψης Αποφάσεων (MDP). Ο δυναμικός προγραμματισμός δίνει λύση σε περίπλοκα προβλήματα προσπαθώντας να τα διασπάσει σε απλούστερα υπο-προβλήματα, και στη συνέχεια συνδυάζοντας τις επιμέρους λύσεις που έχουν προκύψει από αυτά. Εντούτοις, ο ΔΠ δεν χρησιμοποιείται σε προβλήματα ενισχυτικής μάθησης, δεδομένου ότι δεν είναι εφικτή η τέλεια γνώση της δυναμικής και της επιβράβευσης για το περιβάλλον, καθώς επίσης δεν μπορεί να εφαρμοστεί σε περιβάλλοντα με συνεχή χώρο καταστάσεων ή/και δράσεων.

Για να μπορέσει να εφαρμοστεί η λογική του ΔΠ θα πρέπει το πρόβλημα που αντιμετωπίζουμε να ικανοποιεί ορισμένες ιδιότητες. Μια πρώτη ιδιότητα είναι αυτή της βέλτιστης υπο-δομής. Με άλλα λόγια, το πρόβλημα θα πρέπει να μπορεί να διασπαστεί σε δύο ή περισσότερα υπο-προβλήματα, με τη λύση των θυγατρικών προβλημάτων να υπαγορεύει και αυτήν του αρχικού. Για παράδειγμα, στην περίπτωση αναζήτησης του συντομότερου τρόπου προσέγγισης ενός σταθερού στόχου από ένα κινούμενο αντικείμενο, το πρόβλημα μπορεί να χωριστεί σε επιμέρους προβλήματα που αποτελούν την εύρεση συντομότερου μονοπατιού μεταξύ των αντικειμένων, με το σημείο εκκίνησης να είναι κάθε φορά το επόμενο σημείο στο οποίο βρίσκεται το κινούμενο αντικείμενο, έπειτα από την εκτέλεση ενός βήματος προς το στόχο. Μια δεύτερη ιδιότητα αποτελεί η επανεμφάνιση παρόμοιων υπο-προβλημάτων. Με τον τρόπο αυτό, η διάσπαση σε υπο-προβλήματα μπορεί να μειώσει την υπολογιστική δαπάνη, είτε γιατί συναντούμε τον ίδιο τρόπο λύσης σε πολλές υπο-περιπτώσεις, οπότε με προσωρινή αποθήκευση αυτής της λύσης, η επαναχρησιμοποίησή της είναι γρήγορη και εύκολη, είτε γιατί τα θυγατρικά προβλήματα είναι εν γένει απλούστερα από το αρχικό. Στην περίπτωση που το προς αντιμετώπιση πρόβλημα μπορεί να περιγραφεί από μια MDP, τότε και οι δύο παραπάνω ιδιότητες ικανοποιούνται. Συγκεκριμένα, η εξίσωση Bellman επιτρέπει την αναδρομική αποσύνθεση του κύριου προβλήματος, εφόσον πραγματοποιεί διάσπαση της βέλτιστης συνάρτησης αξίας, στη βέλτιστη συμπεριφορά για ένα βήμα και στη βέλτιστη συμπεριφορά μετά από αυτό το βήμα. Επιπλέον, αναφορικά στη δεύτερη ιδιότητα, η ύπαρξη της συνάρτησης αξίας είναι αυτή που επιτρέπει την αποθήκευση και

επαναχρησιμοποίηση των λύσεων. Αυτό συμβαίνει διότι η συνάρτηση αξίας συνιστά μια προσωρινή μνήμη στην οποία αποθηκεύεται κάθε χρήσιμη πληροφορία για την MDP. Με τον τρόπο αυτό, υπαγορεύει τον βέλτιστο τρόπο συμπεριφοράς ώστε να καταφέρουμε να πετύχουμε τη μέγιστη επιβράβευση, χωρίς να χρειάζεται να επαναλάβουμε τους υπολογισμούς που μας οδήγησαν σε αυτή.

2.2.1 Αξιολόγηση Πολιτικής (πρόβλεψη)

Στην περίπτωση που επιθυμούμε να υπολογίσουμε τη συνάρτηση κατάστασης – αξίας v_π για κάθε τυχαία πολιτική π, θεωρούμε μια ακολουθία κατά-προσέγγιση-συναρτήσεων-αξίας $v_0, v_1, v_2, \dots$, η κάθε μια εκ των οποίων αντιστοιχεί το $\mathcal{S}^+$ στο $\mathbb{R}$. Η αρχική προσέγγιση, v_0, επιλέγεται τυχαία (εκτός από το ότι η τελική κατάσταση, αν υπάρχει, θα πρέπει να έχει μηδενική τιμή), και κάθε διαδοχική προσέγγιση δίνεται με τη χρήση της εξίσωσης Bellman (2) ως κανόνα ανανέωσης:

$$v_{k+1}(s) \doteq \mathbb{E}_\pi[R_{t+1} + \gamma v_k(S_{t+1}) | S_t = s] = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)\,[r + \gamma v_k(s')], \forall s \in \mathcal{S} \Rightarrow$$

Εναλλακτικά: $v_{k+1}(s) = \sum_a \pi(a|s) \left(R_s^a + \gamma \sum_{s',r} \mathcal{P}_{ss}^a, v_k(s') \right)$

και $\boldsymbol{v}^{k+1} = \boldsymbol{R}^\pi + \gamma \boldsymbol{\mathcal{P}}^\pi \boldsymbol{v}^k$ με τη χρήση πινάκων.

Επομένως, στην περίπτωση της πρόβλεψης έχουμε ως είσοδο μια MDP ή μια MRP και η έξοδός μας είναι η v_π. Η ακολουθία των $\{v_k\}$ μπορεί να αποδειχθεί ότι συγκλίνει στη v_π καθώς $k \to \infty$. Ο συγκεκριμένος αλγόριθμος ονομάζεται επαναληπτική αξιολόγηση πολιτικής (Iterative Policy Evaluation).

2.2.2 Βελτίωση πολιτικής (έλεγχος)

Ο λόγος που επιθυμούμε να υπολογίσουμε τη συνάρτηση αξίας που αντιστοιχεί σε μια πολιτική είναι για να μπορέσουμε να βρούμε καλύτερες πολιτικές (Policy Improvement). Στην περίπτωσή μας, μέσω της προαναφερθείσας διαδικασίας έχουμε προσδιορίσει τη συνάρτηση αξίας v_π για μια τυχαία ντετερμινιστική πολιτική π. Σε κάποια κατάσταση s θέλουμε να μάθουμε αν θα ήταν καλύτερο να αλλάξουμε την πολιτική με το να επιλέξουμε ντετερμινιστικά μια δράση α η οποία δεν υπαγορεύεται από την $\pi(s)$. Αυτό που γνωρίζουμε είναι το πόσο καλή είναι η επιλογή της τρέχουσας πολιτικής από την κατάσταση s, η οποία είναι $v_\pi(s)$, αλλά δεν ξέρουμε πόσο καλύτερα ή χειρότερα θα ήταν αν αλλάζαμε πολιτική. Ένας τρόπος να απαντηθεί αυτός ο προβληματισμός είναι να θεωρήσουμε ότι επιλέγουμε τη δράση α στην κατάσταση s, η οποία μεγιστοποιεί τη συνάρτηση αξίας, $q_\pi(s, a)$, και στο εξής ακολουθούμε τη νέα πολιτική π'. Στην περίπτωση αυτή, η ανανέωση της πολιτικής γίνεται με άπλητο τρόπο, ως εξής:

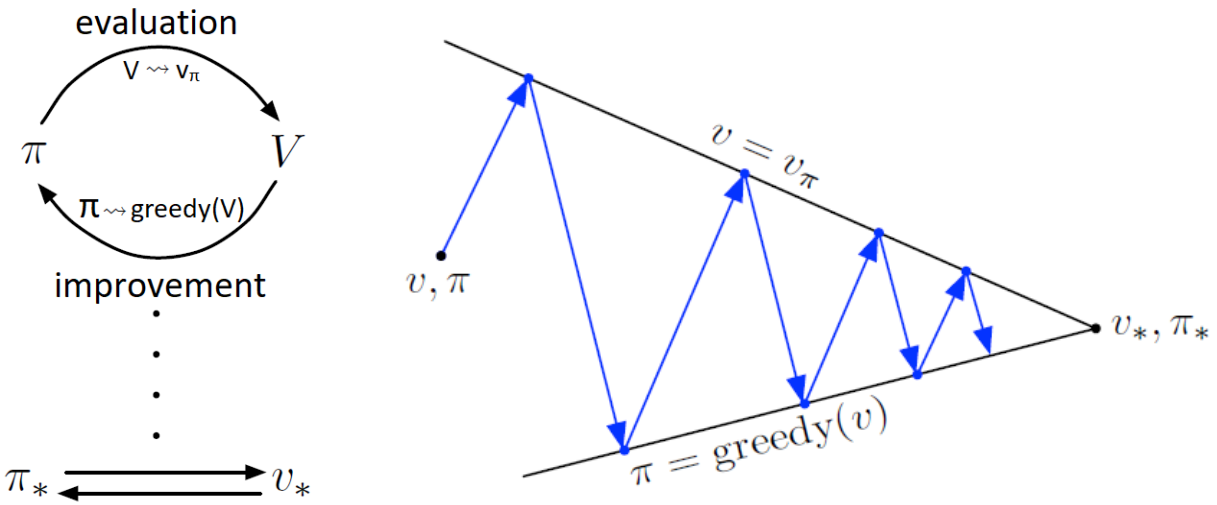
$$\begin{aligned}\pi'(s) &\doteq \underset{a\in\mathcal{A}}{arg\max}\, q_\pi(s,a)\\ &= \underset{a\in\mathcal{A}}{arg\max}\, \mathbb{E}[R_{t+1}+\gamma v_\pi(S_{t+1})|S_t=s,A_t=a]\\ &= \underset{a\in\mathcal{A}}{arg\max}\sum_{s',r} p(s',r|s,a)\,[r+\gamma v_\pi(s')]\end{aligned}$$

2.2.3 Επανάληψη πολιτικής

Ο τρόπος προσδιορισμού της βέλτιστης πολιτικής ονομάζεται Επανάληψη πολιτικής (Policy Iteration). Όταν μια πολιτική π έχει βελτιωθεί με χρήση της v_π σε π', μπορούμε στη συνέχεια να υπολογίσουμε την v_π, και να επαναλάβουμε τη διαδικασία της βελτίωσης, καταλήγοντας στην πολιτική π''. Με τον τρόπο αυτό μπορούμε να καταλήξουμε σε μια ακολουθία μονοτονικά βελτιούμενων πολιτικών και συναρτήσεων αξίας:

$$\pi_0 \xrightarrow{Αξιολόγηση} v_{\pi_0} \xrightarrow{Βελτίωση} \pi_1 \xrightarrow{Αξιολόγηση} v_{\pi_1} \xrightarrow{Βελτίωση} \pi_2 \xrightarrow{Αξιολόγηση} \dots \xrightarrow{Βελτίωση} \pi_* \xrightarrow{Αξιολόγηση} v_*$$

Γνωρίζουμε πως οι πεπερασμένες MDP έχουν και πεπερασμένο αριθμό πολιτικών, η παραπάνω διαδικασία θα πρέπει να συγκλίνει σε μια βέλτιστη πολιτική και σε μια βέλτιστη συνάρτηση αξίας, σε πεπερασμένο αριθμό επαναλήψεων. Επομένως, η επανάληψη πολιτικής απαρτίζεται από δύο ταυτόχρονες διαδικασίες, οι οποίες αλληλεπιδρούν μεταξύ τους. Μάλιστα, οι δύο αυτές διαδικασίες εναλλάσσονται, με την κάθε μια να ολοκληρώνεται πριν ξεκινήσει η άλλη, χωρίς όμως αυτό να είναι απαραίτητο. Αυτήν την αλληλεπίδραση μεταξύ Βελτίωσης-Πολιτικής και Αξιολόγησης-Πολιτικής την ονομάζουμε Γενικευμένη Επανάληψη Πολιτικής (Generalized Policy Iteration - GPI). Σχηματικά, τα βήματα που περιγράψαμε φαίνονται στις ακόλουθες εικόνες:



Σχήμα 2.2: Γενικευμένη Επανάληψη Πολιτικής, από [42]

Από τα σχήματα είναι προφανές ότι εάν σταθεροποιηθεί η διαδικασία αξιολόγησης και η διαδικασία βελτίωσης, τότε δεν παρατηρούνται άλλες αλλαγές στη συνάρτηση αξίας και στην πολιτική και επομένως έχουμε καταλήξει στις βέλτιστες τιμές τους.

2.2.4 Επανάληψη αξίας

Ένα μειονέκτημα της επανάληψης πολιτικής είναι ότι κάθε μια από τις επαναλήψεις της, περιέχει την αξιολόγηση πολιτικής, η οποία μπορεί να είναι ένας εκτεταμένος επαναληπτικός υπολογισμός που απαιτεί πολλαπλές σαρώσεις στον χώρο των καταστάσεων, προκειμένου να επιτευχθεί η σύγκλιση στην v_π. Στην πραγματικότητα, το βήμα αξιολόγησης πολιτικής μπορεί να περικοπεί με διάφορους τρόπους, χωρίς να χαθεί η εγγύηση σύγκλισης της επανάληψης πολιτικής. Μια σημαντική ειδική περίπτωση είναι όταν η αξιολόγηση πολιτικής διακόπτεται μετά από μία μόνο σάρωση, δηλαδή μια ενημέρωση κάθε κατάστασης. Αυτός ο αλγόριθμος ονομάζεται επανάληψη αξίας. Η αναπαράστασή του γίνεται με έναν συνδυασμό της βελτίωσης πολιτικής και της μη-πλήρους αξιολόγησης πολιτικής, επιλέγοντας τη δράση με τη μέγιστη συνάρτηση κατάστασης – αξίας, ως ακολούθως:

$$v_{k+1}(s) \doteq \max_{a\in\mathcal{A}} \mathbb{E}[R_{t+1} + \gamma v_k(S_{t+1}) | S_t = s, A_t = a] = \max_{a\in\mathcal{A}} \sum_{s',r} p(s', r|s, a)\,[r + \gamma v_k(s')]\,, \forall s \in \mathcal{S}$$

Εναλλακτικά: $v_{k+1}(s) = \max_{a\in\mathcal{A}} \left(R_s^a + \gamma \sum_{s',r} \mathcal{P}_{ss'}^a, v_k(s') \right)$

και $\boldsymbol{v}^{k+1} = \max_{a\in\mathcal{A}} \boldsymbol{R^a} + \gamma \boldsymbol{\mathcal{P}^a} \boldsymbol{v}^k$ με τη χρήση πινάκων.

Στην επανάληψη αξίας, για τυχαίο v_0 η ακολουθία των $\{v_k\}$ μπορεί να αποδειχθεί ότι συγκλίνει στη v_* καθώς $k \to \infty$. Παρατηρούμε επίσης, πως η ανανέωση της Επανάληψης Αξίας είναι παρόμοια με αυτή της Αξιολόγησης Πολιτικής, μόνο που χρειάζεται το μέγιστο πάνω σε κάθε δράση.

2.3 Bootstrapping και Εξερεύνηση – Εκμετάλλευση

Το bootstrapping – ως μέθοδος ΔΠ – κατά τη βελτιστοποίηση της συνάρτησης αξίας για μια αρχική κατάσταση, χρησιμοποιεί τις αναμενόμενες τιμές της επόμενης κατάστασης για να εμπλουτίσει την πρόβλεψη. Επομένως, η ενημέρωση των τιμών γίνεται με βάση ορισμένες εκτιμήσεις και όχι με βάση ακριβείς τιμές.

Ένας απαιτητικός και εύλογος προβληματισμός που ανακύπτει στην Ενισχυτική Μάθηση και όχι στα άλλα ήδη μάθησης που μελετήσαμε, αποτελεί η ισορροπία ανάμεσα στην εξερεύνηση και την εκμετάλλευση (exploration and exploitation trade-off). Πιο αναλυτικά, με στόχο τη μεγιστοποίηση των επιβραβεύσεων, ο «πράκτορας» της Ενισχυτικής Μάθησης θα πρέπει να προτιμά τις δράσεις που έχει ήδη δοκιμάσει στο παρελθόν και οι οποίες του εξασφαλίζουν μεγάλες τιμές

επιβραβεύσεων. Με τον τρόπο αυτό, ο πράκτορας – δράστης εκμεταλλεύεται την εμπειρία που έχει αποκτήσει στο «παρελθόν», για να καταφέρει να έχει καλύτερη επιβράβευση στο «μέλλον». Παράλληλα, θα πρέπει όμως και να εξερευνήσει νέες πιθανές δράσεις, οι οποίες ίσως να τον οδηγήσουν σε ακόμα καλύτερες αλληλουχίες επιβραβεύσεων. Επομένως, ο πράκτορας μπορεί να εκμεταλλευτεί τις γνώσεις που έχει αποκτήσει, με κίνδυνο να εγκλωβιστεί σε μια υπο-βέλτιστη αλληλουχία επιλογής δράσεων, ενώ από την άλλη, η ανεξέλεγκτη εξερεύνηση, μπορεί να μην οδηγήσει ποτέ στη μεγιστοποίηση της συνολικής επιβράβευσης. Συμπερασματικά, μια συνεργασία μεταξύ εκμετάλλευσης και εξερεύνησης θα μπορούσε να θεωρηθεί χρυσή τομή στο πρόβλημά μας. Στην περίπτωση ενός στοχαστικού προβλήματος Ενισχυτικής Μάθησης, η κάθε πιθανή δράση θα πρέπει να δοκιμάζεται πάνω από μια φορές, έτσι ώστε να μπορέσει ο πράκτορας να αποκτήσει μια αξιόπιστη εκτίμηση των επιβραβεύσεων που θα ακολουθήσουν.

Οι πιο κοινές στρατηγικές εξερεύνησης περιλαμβάνουν την προσθήκη κάποιας μορφής θορύβου στην επιλογή της δράσης, είτε με τη μορφή άμεσης διαταραχής των παραμέτρων της πολιτικής, είτε με δράση ϵ-greedy ή softmax σε διακριτούς χώρους δράσης και Gaussian θορύβου που προστίθεται κυρίως σε συνεχείς χώρους δράσης. Η τυχαία εξερεύνηση συχνά αποτυγχάνει να εξερευνήσει επαρκώς τον χώρο της πολιτικής, καθώς συνήθως περιορίζεται σε μια τοπική περιοχή κοντά στην τρέχουσα πολιτική [56].

2.4 Πρόβλεψη Αγνώστου Μοντέλου

Ένα πρόβλημα ενισχυτικής μάθησης μπορεί να τεθεί με ποικίλους τρόπους, ανάλογα με τις υποθέσεις σχετικά με το επίπεδο γνώσης που είναι αρχικά διαθέσιμο στον πράκτορα. Σε προβλήματα με πλήρη γνώση (complete knowledge), ο πράκτορας έχει ένα πλήρες και ακριβές μοντέλο της δυναμικής του περιβάλλοντος. Εάν το περιβάλλον μπορεί να μοντελοποιηθεί σε MDP, τότε ένα τέτοιο μοντέλο περιγράφεται από την πλήρη συνάρτηση δυναμικής p τεσσάρων ορισμάτων (3). Σε προβλήματα με ελλιπή γνώση, δεν είναι διαθέσιμο ένα πλήρες και τέλειο μοντέλο του περιβάλλοντος. Στην περίπτωση του Αγνώστου Μοντέλου (model-free) δεν γνωρίζουμε την MDP που περιγράφει το περιβάλλον του προβλήματός μας. Ο στόχος της πρόβλεψης, επομένως, είναι να εκτιμήσει τις συναρτήσεις αξίας μέσω αλληλεπίδρασης με το περιβάλλον, ώστε να προσδιοριστούν οι βέλτιστες πολιτικές.

2.4.1 Μέθοδος Monte Carlo

Στη μέθοδο Monte Carlo (MC Learning - MC) η μάθηση γίνεται απευθείας από την εμπειρία που αποκτά ο πράκτορας, δηλαδή από δείγματα συνόλων καταστάσεων, δράσεων, και επιβραβεύσεων από την πραγματική ή την μέσω-προσομοίωσης αλληλεπίδραση με το περιβάλλον. Συνεπώς, η μάθηση εξ ολοκλήρου μέσω της εμπειρίας μας δίνει τη χρήσιμη δυνατότητα να μπορούμε να

αντιμετωπίσουμε το πρόβλημα της άγνοιας για το μοντέλο του περιβάλλοντος. Εξίσου σημαντική είναι και η εμπειρία που αποκτούμε μέσω προσομοίωσης του περιβάλλοντος. Λόγω της μάθησης συναρτήσεων αξίας και βέλτιστων πολιτικών από την εμπειρία σε μορφή δειγμάτων επεισοδίων, οι μέθοδοι MC έχουν τρία πλεονεκτήματα σε σχέση με το Δυναμικό Προγραμματισμό. Πρώτον, μπορούν να χρησιμοποιηθούν για την εκμάθηση της βέλτιστης συμπεριφοράς απευθείας από την αλληλεπίδραση με το περιβάλλον, χωρίς μοντέλο της δυναμικής του περιβάλλοντος. Δεύτερον, μπορούν να χρησιμοποιηθούν με προσομοίωση ή δείγματα μοντέλων. Αυτό συμβαίνει γιατί σε πολλές εφαρμογές είναι εύκολο να προσομοιωθούν δείγματα επεισοδίων, ενώ είναι δύσκολο να κατασκευαστεί ένα σαφές μοντέλο πιθανοτήτων μετάβασης που απαιτείται από τις μεθόδους του ΔΠ. Τρίτον, είναι εύκολο οι μέθοδοι MC να επικεντρωθούν σε ένα μικρό υποσύνολο των καταστάσεων, καθώς μια περιοχή ειδικού ενδιαφέροντος μπορεί να αξιολογηθεί με ακρίβεια χωρίς να επιβαρύνει την ακριβή αξιολόγηση του υπόλοιπου συνόλου κατάστασης.

Ένα επιπλέον πλεονέκτημα των μεθόδων MC είναι ότι μπορεί να ζημιωθούν λιγότερο από παραβιάσεις της Μαρκοβιανής ιδιότητας. Αυτό συμβαίνει επειδή δεν ενημερώνουν τις εκτιμήσεις αξίας τους με βάση τις εκτιμήσεις αξίας των διαδοχικών καταστάσεων, δηλαδή δεν κάνουν bootstrap.

Η μέθοδος MC θα χρησιμοποιηθεί με στόχο τη μάθηση της συνάρτησης κατάστασης-αξίας, $v_\pi \doteq \mathbb{E}_\pi[G_t|S_t = s]$, δεδομένης μιας συγκεκριμένης πολιτικής, π, και έχοντας στη διάθεσή μας ένα σύνολο επεισοδίων, τα οποία έχουν αποκτηθεί ακολουθώντας την πολιτική π και διασχίζοντας την κατάσταση s. Ως επεισόδιο ορίζουμε την πρώτη ή την κάθε χρονική στιγμή t κατά την οποία επισκεπτόμαστε την κατάσταση s. Ως αξία-τιμή μιας κατάστασης θεωρούμε την αναμενόμενη επιβράβευση, δηλαδή την αναμενόμενη συσσωρευτική μελλοντική ανταμοιβή, με συνυπολογισμένη την έκπτωση γ, ξεκινώντας από αυτήν την κατάσταση. Ένας προφανής τρόπος εκτίμησής της μέσω της εμπειρίας είναι ο υπολογισμός του μέσου όρου των αποδόσεων (G_t) που παρατηρήθηκαν έπειτα από επισκέψεις σε αυτήν την κατάσταση. Όσο περισσότερες αποδόσεις παρατηρηθούν, τόσο περισσότερο θα συγκλίνει η μέση τιμή στην αναμενόμενη τιμή. Έστω η πρώτη επίσκεψη του πράκτορα στην κατάσταση s κατά τη διάρκεια ενός επεισοδίου, την οποία ονομάζουμε $first_{visit}$. Στη μέθοδο MC $first_{visit}$ εκτιμούμε την $v_\pi(s)$ ως το μέσο όρο των αποδόσεων που ακολουθούν της πρώτης επίσκεψης στην s, ενώ στην $every_{visit}$ μέθοδο εκτιμούμε την $v_\pi(s)$ υπολογίζοντας τον μέσο όρο των αποδόσεων μετά από όλες τις επισκέψεις στην s. Στην εν λόγω εργασία θα μελετήσουμε την $every_{visit}$ μέθοδο. Κάθε φορά που επισκεπτόμαστε την κατάσταση s, αυξάνουμε έναν μετρητή, $N(s)$, $N(S_t) \leftarrow N(S_t) + 1$ και υπολογίζουμε την τιμή που αναζητούμε ως:

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

Σε μη-στατικά προβλήματα, στα οποία ενδέχεται να υπάρξουν ξαφνικές αλλαγές στο περιβάλλον, δεν ενδείκνυται η παραπάνω ανανέωση της συνάρτησης αξίας με χρήση του μέσου όρου, αλλά ένας τύπος ανανέωσης που αγνοεί τα πολύ παλιά επεισόδια:

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

Η μέθοδος MC είναι μια διαδικασία που «κοιτάζει προς το μέλλον», δεδομένου ότι περιμένουμε μέχρι να ολοκληρωθεί ένα επεισόδιο για να δούμε την απόδοση που προέκυψε, και έπειτα επιστρέφουμε στην κατάσταση που επιθυμούμε να ενημερώσουμε, εφαρμόζοντας την αλλαγή που θα επιλέξουμε.

2.4.2 Μάθηση Temporal-Difference (TD)

Ο εν λόγω τρόπος μάθησης συνδυάζει ιδέες από τη μέθοδο Monte Carlo και το Δυναμικό Προγραμματισμό. Όπως και στην περίπτωση MC, η μάθηση πραγματοποιείται απευθείας από την εμπειρία, χωρίς ένα μοντέλο για τη δυναμική του συστήματος στη διάθεσή μας. Σε σχέση με τον ΔΠ, οι μέθοδοι TD ενημερώνουν τις εκτιμήσεις τους βασιζόμενοι εν μέρει σε άλλες εκτιμήσεις που έχουν ήδη μαθευτεί, χωρίς να περιμένουν για ένα τελικό αποτέλεσμα (bootstrap). Επομένως, έχουν τη δυνατότητα να μάθουν από ατελή επεισόδια, μέσω της τεχνικής bootstrapping.

2.4.3 Μέθοδος TD(0)

Η απλούστερη εκδοχή της TD μάθησης είναι ο αλγόριθμος TD(0) ή TD ενός-βήματος (one-step TD). Ο στόχος μας είναι να μάθουμε το v_π κάτω από μια πολιτική π. Αντί της απόδοσης G_t θεωρούμε την εκτιμώμενη απόδοση $R_{t+1} + \gamma V(S_{t+1})$, την οποία ονομάζουμε στόχο της μεθόδου TD (TD target), εξού και η μέθοδος bootstrapping. Επομένως, με την παραπάνω θεώρηση, ο κανόνας ανανέωσης θα έχει τη μορφή:

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

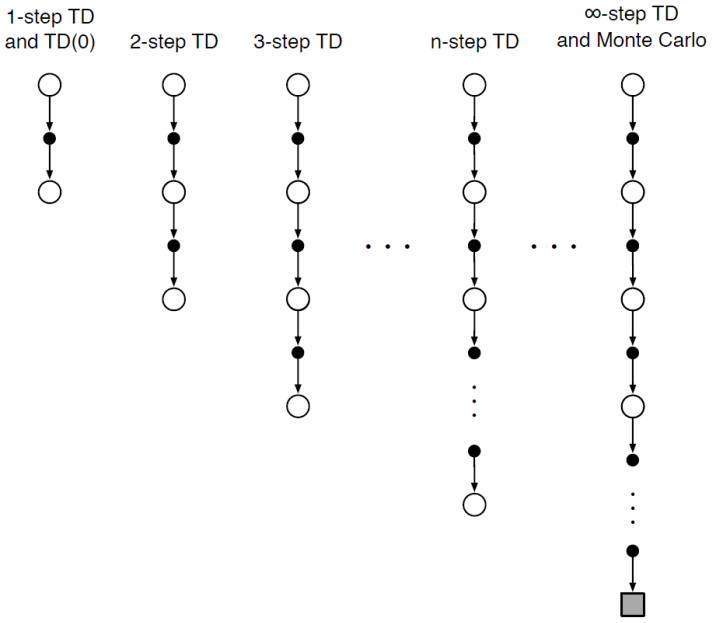
όπου το $R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$ αποτελεί το TD σφάλμα (TD error) και συμβολίζεται με δ_t.

Ένα σημαντικό πλεονέκτημα της TD μεθόδου είναι ότι μπορεί να μάθει online και δεν χρειάζεται, σε αντίθεση με τη μέθοδο MC, να περιμένει την ολοκλήρωση ενός επεισοδίου για να μάθει την απόδοση. Επομένως, ο TD μπορεί να μάθει από ημιτελείς ακολουθίες και να λειτουργήσει σε συνεχή (μη-τερματιζόμενα) περιβάλλοντα, σε αντίθεση με τον MC, ο οποίος μαθαίνει μόνο από πλήρης ακολουθίες και λειτουργεί σε επεισοδιακά περιβάλλοντα. Αυτό συμβαίνει γιατί σε εφαρμογές με συνεχείς εργασίες, χωρίς καν την ύπαρξη επεισοδίων, ορισμένες μέθοδοι MC πρέπει να αγνοούν τα επεισόδια στα οποία γίνονται πειραματικές ενέργειες, γεγονός που μπορεί να επιβραδύνει σημαντικά τη μάθηση. Από την άλλη, οι μέθοδοι TD είναι πολύ λιγότερο επιρρεπείς σε τέτοιου είδους προβλήματα επειδή μαθαίνουν από κάθε μετάβαση, ανεξάρτητα από τις επόμενες ενέργειες που λαμβάνονται. Εντούτοις, η χρήση της εκτίμησης της απόδοσης προσθέτει bias στον κανόνα ανανέωσης, μέσω του TD target. Αυτό μπορεί, όμως, να ερμηνευτεί και θετικά, δεδομένου ότι το TD target έχει πολύ μικρότερη διακύμανση από την απόδοση. Ο λόγος είναι ότι η απόδοση ($G_t = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T-1} R_T$) εξαρτάται από πολλές τυχαίες δράσεις, μεταβάσεις και επιβραβεύσεις, ενώ το TD target ($R_{t+1} + \gamma V(S_{t+1})$) εξαρτάται από μία

τυχαία δράση, μετάβαση και επιβράβευση. Τέλος, αποδεικνύεται ότι με μικρό ρυθμό μάθησης και δεδομένη πολιτική π, η πρόβλεψη του TD(0) συγκλίνει στο $v_\pi(s)$.

2.4.4 Μέθοδος TD(λ)

Οι μέθοδοι MC και TD(0) παρουσιάζουν ακραία αντίθετη λογική στον τρόπο που προσεγγίζουν το πρόβλημα, η πρώτη με απαίτηση ύπαρξης πλήρη επεισοδίων για την ανανέωση της τιμής του $V(S_t)$ και η δεύτερη με ανανέωση μόνο μετά από ένα βήμα του αλγορίθμου. Μια επέκταση του TD(0) θα μπορούσε να είναι το TD target να βλέπει n – βήματα στο μέλλον και όχι μόνο ένα. Στην περίπτωση που $n \to \infty$ τότε ο αλγόριθμος ισοδυναμεί με τον MC. Η σχηματική απεικόνιση του παραπάνω, καθώς και η περιγραφή του σε πίνακα φαίνονται στη συνέχεια:



Σχήμα 2.3: Από τον TD(0) στον Monte Carlo, από [42]

$n = 1$	$TD(0)$	$G_t^{(1)} = R_{t+1} + \gamma V(S_{t+1})$
$n = 2$		$G_t^{(2)} = R_{t+1} + \gamma R_{t+2} + \gamma^2 V(S_{t+2})$
⋮		⋮
$n = k$	$TD(k)$	$G_t^{(k)} = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^k V(S_{t+k})$
⋮		⋮
$n = \infty$	MC	$G_t^{(\infty)} = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T-1} R_T$

Πίνακας 2.1: Αποδόσεις n-βημάτων, για $n = 1,2, \dots, k, \dots, \infty$, από [44]

Οι μέθοδοι που χρησιμοποιούν ενημερώσεις n βημάτων εξακολουθούν να είναι μέθοδοι TD επειδή εξακολουθούν να αλλάζουν μια προηγούμενη εκτίμηση με βάση το πώς διαφέρει από μια μεταγενέστερη εκτίμηση. Στην περίπτωση του TD(n) η μεταγενέστερη εκτίμηση δεν είναι ένα βήμα αργότερα αλλά n βήματα αργότερα. Οι μέθοδοι αυτές, στις οποίες η διαφορά εκτείνεται σε n βήματα, ονομάζονται μέθοδοι TD n – βημάτων.

Στην TD(n) μέθοδο, η απόδοση θα εκτείνεται σε n – βήματα και θα έχει τη μορφή:

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

Ενώ, η ανανέωση θα είναι η εξής:

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

Μία επέκταση του παραπάνω θα μπορούσε να είναι ο υπολογισμός του μέσου όρου διαφόρων TD n – βημάτων, με διαφορετικά n. Έτσι μπορούμε να συνδυάσουμε πληροφορία από δύο ή περισσότερες διαφορετικές χρονικές στιγμές. Η λ – απόδοση, G_t^{λ}, (λ – return) συνδυάζει όλες τις αποδόσεις n – βημάτων, με τη χρήση βαρών $(1-\lambda)\lambda^{n-1}$ ως εξής:

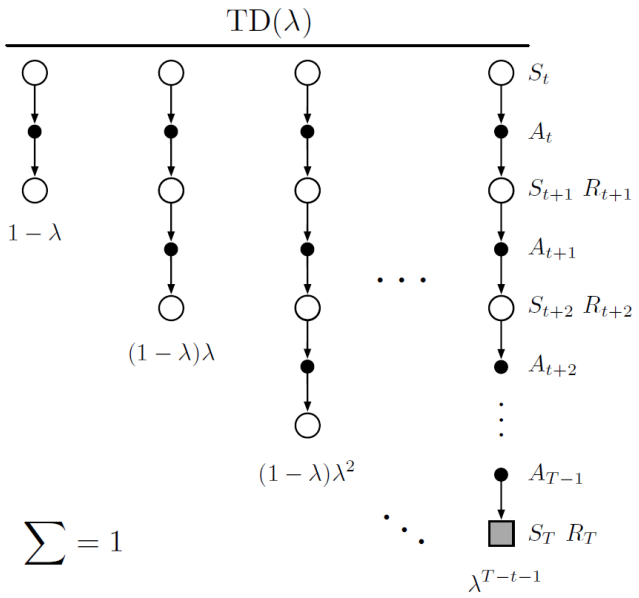
$$G_t^{\lambda} = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} G_t^{n}$$

Επομένως:

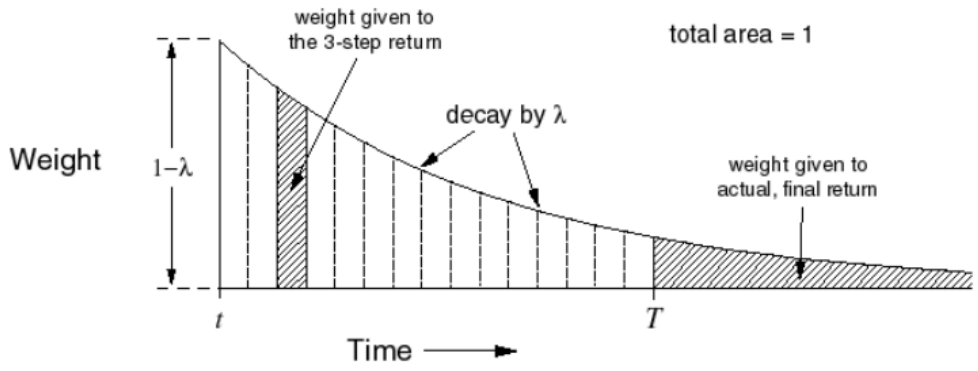
$$V(S_t) \leftarrow V(S_t) + a\big(G_t^{\lambda} - V(S_t)\big)$$

Ο εν λόγω αλγόριθμος ονομάζεται εμπρόσθιος TD(λ) αφού η απόδοση, G_t^{λ}, δίνεται με βάση τις επόμενες καταστάσεις.

Η σχηματική απεικόνιση του εμπρόσθιου TD(λ) φαίνεται ακολούθως:



Σχήμα 2.4: TD(λ), λ-return, με χρήση eligibility traces. Εάν λ=0, η ανανέωση αποτελείται από ένα βήμα, την TD ανανέωση ενός-βήματος, ενώ αν λ=1, τότε η συνολική ανανέωση γίνεται σύμφωνα με τον κανόνα Monte Carlo, από [42]



Σχήμα 2.5: Τα βάρη που δίνονται στο λ-return, σε κάθε απόδοση των n-βημάτων, από [42]

Σχεδόν κάθε αλγόριθμος TD μπορεί να συνδυαστεί με ίχνη επιλεξιμότητας (eligibility traces) για να αποκτήσουμε μια πιο γενική μέθοδο, η οποία μπορεί να μάθει πιο αποδοτικά. Ο μηχανισμός είναι ένα διάνυσμα βραχυπρόθεσμης μνήμης, το ίχνος επιλεξιμότητας, $\boldsymbol{z}_t \in \mathbb{R}^d$, το οποίο είναι παράλληλο με το μακροπρόθεσμο διάνυσμα των βαρών, $\boldsymbol{w}_t \in \mathbb{R}^d$. Η ιδέα είναι ότι στην περίπτωση που ένα συστατικό του $\boldsymbol{w}_t$ συμμετέχει στην παραγωγή μιας εκτιμώμενης αξίας, τότε η αντίστοιχη θέση του $\boldsymbol{z}_t$ συγκεντρώνεται και στη συνέχεια αρχίζει να εξασθενεί. Έπειτα, η μάθηση θα προκύψει στο στοιχείο του $\boldsymbol{w}_t$, εάν προκύψει μη-μηδενικό TD-σφάλμα πριν το ίχνος πέσει στο μηδέν. Η παράμετρος μείωσης του ίχνους $\lambda \in [0,1]$ καθορίζει το ρυθμό με τον οποίο πέφτει η τιμή του ίχνους. Ένα σημαντικό πλεονέκτημα αυτής της μεθόδου είναι ότι απαιτεί τη χρήση μόνο ενός διανύσματος ιχνών και όχι την αποθήκευση των τελευταίων n διανυσμάτων χαρακτηριστικών.

Η χρήση των ιχνών επιλεξιμότητας συνδυάζει την πληροφορία του πόσο συχνά επισκεπτόμαστε μια κατάσταση και του ποια κατάσταση έχουμε επισκεφθεί πιο πρόσφατα, όπως φαίνεται στη συνέχεια:

$$\begin{cases} E_0(s) = 0 \\ E_t(s) = \gamma\lambda E_{t-1}(s) + \mathbb{1}(S_t = s) \end{cases}$$

Επομένως, έχοντας ένα ίχνος επιλεξιμότητας για κάθε κατάσταση, και $\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$, η ενημέρωση του $V(s)$ γίνεται ως εξής:

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

Στην ειδική περίπτωση όπου $\lambda = 0$, η ανανέωση γίνεται μόνο για την τρέχουσα κατάσταση και η μέθοδος ταυτίζεται με την TD(0):

$$\begin{cases} E_t(s) = \mathbb{1}(S_t = s) \\ V(s) \leftarrow V(s) + a\delta_t E_t(s) \end{cases}$$

Στην περίπτωση όπου $\lambda = 1$, η ανανέωση γίνεται στο TD(1) είναι ίδια με αυτή της μεθόδου MC.

2.5 Έλεγχος Αγνώστου Μοντέλου

Με στόχο τον προσδιορισμό βέλτιστης πολιτικής από τον πράκτορα, έπειτα από το βήμα της πρόβλεψης της συνάρτησης αξίας με δεδομένη πολιτική π, ακολουθεί το βήμα του ελέγχου-βελτίωσής της. Η λογική είναι παρόμοια με αυτήν που περιγράψαμε στο Δυναμικό Προγραμματισμό ως Γενικευμένη Επανάληψη Πολιτικής (GPI).

2.5.1 Έλεγχος πάνω στην Πολιτική

Ο έλεγχος πάνω στην πολιτική (on-policy control) σχετίζεται με τη μάθηση μιας πολιτικής π μέσω της εμπειρίας που αποκτήσαμε από την ίδια την πολιτική π, και ακολουθώντας τη συμπεριφορά για την οποία μαθαίνουμε. Η διαδικασία αξιολόγησης της πολιτικής που συναντήσαμε στο ΔΠ στη GPI ήταν άπληστη, με επιλογή της δράσης που οδηγεί στη μέγιστη συνάρτηση αξίας, $\pi = greedy(v)$. Όμως, στην περίπτωση αυτή δεν μας εγγυάται κάποιος ότι οι τροχιές που θα ακολουθούμε θα εξερευνήσουν όλο το χώρο καταστάσεων. Η άπληστη βελτίωση πολιτικής στο $V(s)$:

$$\pi'(s) \doteq \underset{a \in \mathcal{A}}{arg\max}(R_s^a + \mathcal{P}_{ss'}^a V(s'))$$

απαιτεί το μοντέλο της MDP, οπότε δεν μπορούμε να τη χρησιμοποιήσουμε στην περίπτωση του ελέγχου χωρίς μοντέλο.

Στη θέση της θα μπορούσαμε να χρησιμοποιήσουμε στην άπληστη βελτίωση πολιτικής μέσω της συνάρτησης δράσης – αξίας $Q(s, a)$:

$$\pi'(s) \doteq \underset{a \in \mathcal{A}}{arg\max}\, Q(s, a)$$

Στον εν λόγω αλγόριθμο, ξεκινάμε με μια συνάρτηση δράσης – αξίας και κάποια πολιτική π. Σε κάθε βήμα εκτελούμε μια αξιολόγηση πολιτικής σύμφωνα με τη μέθοδο MC, τρέχουμε κάποια επεισόδια, στο τέλος των οποίων διαθέτουμε την εκτίμηση για την πολιτική μας, $Q = q_\pi$. Από κάθε ζεύγος κατάστασης – δράσης παίρνουμε τη μέση απόδοση για να εκτιμήσουμε το $Q(s, a)$ και αυτό μας δείχνει πόσο καλό ήταν αυτό το συγκεκριμένο ζευγάρι. Στη συνέχεια, συμπεριφερόμαστε άπληστα σε σχέση με το Q, το οποίο μας δίνει μια νέα πολιτική, $\pi = greedy(Q)$. Με τη νέα πολιτική τρέχουμε τη διαδικασία για αρκετά ακόμα επεισόδια, έπειτα παίρνουμε το μέσο όρο κάθε ζευγαριού κατάστασης – δράσης, ξανά, και επαναλαμβάνουμε. Έπειτα από πεπερασμένο αριθμό βημάτων, η διαδικασία που περιγράψαμε θα περιμέναμε να μας οδηγήσει στα q^* και π^*. Εντούτοις, ο άπληστος τρόπος ανανέωσης της πολιτικής μας, ελλοχεύει τον κίνδυνο του να παγιδευτούμε σε μια υπο-βέλτιστη λύση, μη εξερευνώντας πλήρως και επαρκώς το χώρο των

καταστάσεων, με αποτέλεσμα να μην επιλέξουμε ποτέ τη βέλτιστη δράση. Ο λόγος που το πρόβλημα αυτό δεν προέκυψε στον ΔΠ είναι ότι στην περίπτωσή μας αλληλεπιδρούμε με το περιβάλλον και δεν το σαρώνουμε στην ολότητά του, όπως κάναμε στον ΔΠ, οπότε υπάρχει ο κίνδυνος να μην μπορέσουμε να το εξερευνήσουμε επαρκώς.

Η λύση στο παραπάνω πρόβλημα είναι η ϵ-άπληστη εξερεύνηση (ϵ-Greedy Exploration). Η μέθοδος αυτή αποτελεί την πιο απλή ιδέα έτσι ώστε να εξασφαλιστεί η συνεχής εξερεύνηση του χώρου καταστάσεων-δράσεων, με στόχο τη σύγκλιση στην π^*. Πιο συγκεκριμένα, για να το πετύχει αυτό, πριν την επιλογή της δράσης, με πιθανότητα ϵ θα επιλέξουμε μια τυχαία δράση, ενώ με πιθανότητα $1-\epsilon$ θα επιλέξουμε την άπληστη δράση, θεωρώντας m όλες τις δράσεις με μη-μηδενική πιθανότητα να συμβούν, δηλαδή:

$$\pi(a|s) = \begin{cases} {}^{\epsilon}/_{m} + 1 - \epsilon, & \alpha\nu\ a^* = \underset{a\in\mathcal{A}}{\operatorname{argmax}}\, Q(s,a) \\ {}^{\epsilon}/_{m}, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases}$$

Είμαστε σίγουροι πως για κάθε ϵ-άπληστη πολιτική π, η ϵ-άπληστη πολιτική π' σε σχέση με την q_π αποτελεί μια βελτίωση, $v_{\pi'}(s) \geq v_\pi(s)$. Η απόδειξη σκιαγραφείται στη συνέχεια:

$$q_\pi\big(s,\pi'(s)\big) = \sum_{a\in\mathcal{A}} \pi'(a|s) q_\pi(s,a) = {}^{\epsilon}/_{m} \sum_{a\in\mathcal{A}} q_\pi(s,a) + (1-\epsilon) \max_{a\in\mathcal{A}} q_\pi(s,a)$$

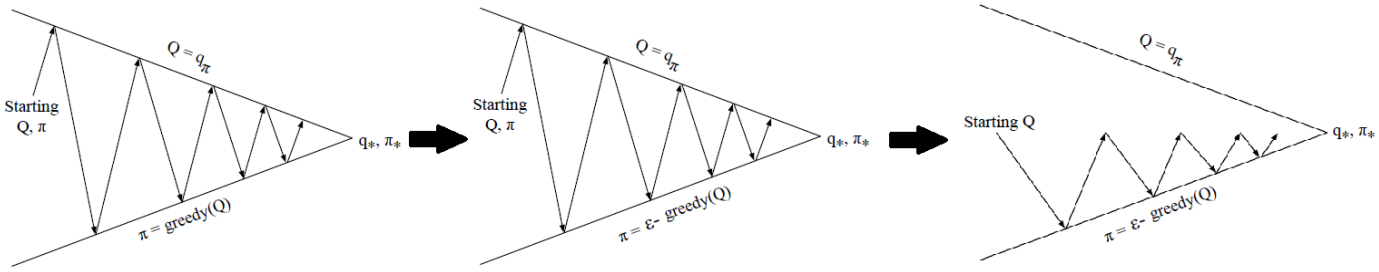
$$\xrightarrow{\tau o\ \mu\acute{\epsilon}\gamma\iota\sigma\tau o\ (\max_{a\in\mathcal{A}} q_\pi(s,a))\ \pi\rho\acute{\epsilon}\pi\epsilon\iota\ \nu\alpha\ \epsilon\acute{\iota}\nu\alpha\iota\ \mu\epsilon\gamma\alpha\lambda\acute{\upsilon}\tau\epsilon\rho o\ \alpha\pi\acute{o}\ \kappa\acute{\alpha}\theta\epsilon\ \sigma\tau\alpha\theta\mu\iota\sigma\mu\acute{\epsilon}\nu o\ \acute{\alpha}\theta\rho o\iota\sigma\mu\alpha\ \tau\omega\nu\ \tau\iota\mu\acute{\omega}\nu\ q_\pi}$$

$$\geq {}^{\epsilon}/_{m} \sum_{a\in\mathcal{A}} q_\pi(s,a) + (1-\epsilon) \sum_{a\in\mathcal{A}} \frac{\pi(a|s) - {}^{\epsilon}/_{m}}{1-\epsilon} q_\pi(s,a)$$

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

Επομένως, στην διαδικασία που περιγράψαμε παραπάνω, η βελτίωση πολιτικής θα γίνεται με ϵ-άπληστο τρόπο. Εντούτοις, δεν είναι απαραίτητο να αξιολογήσουμε πλήρως την πολιτική μας και να «φτάσουμε» μέχρι την $Q = q_\pi$ γραμμή. Μερικές φορές χρειάζεται να κάνουμε λίγα βήματα για να αξιολογήσουμε την πολιτική μας και να φτάσουμε στο σημείο που θα αποκτήσουμε τις πληροφορίες που χρειαζόμαστε έτσι ώστε να καταλήξουμε σε πολύ καλύτερη πολιτική. Επομένως, δεν χρειάζεται να συνεχίσουμε αυτή τη διαδικασία εκτελώντας επιπλέον επαναλήψεις μέχρι τέλους, όταν $Q = q_\pi$. Με άλλα λόγια, κατά το βήμα ελέγχου στον MC μπορούμε να βελτιώσουμε την πολιτική μας απευθείας μετά το πέρας ενός επεισοδίου.

Τα παραπάνω βήματα που παρουσιάσαμε και οι τροποποιήσεις από τη μια μέθοδο στην άλλη, οπτικοποιούνται στο ακόλουθο σχήμα:



Σχήμα 2.6: Η εξέλιξη από greedy Policy Improvement, σε ϵ-greedy (διάγραμμα 1ο → διάγραμμα 2ο). Η εξέλιξη από Monte Carlo Policy evaluation με $Q = q_\pi$, σε $Q \approx q_\pi$ (διάγραμμα 2ο → διάγραμμα 3ο). Απόδοση από [44].

Στην τελευταία προέκταση που κάναμε, θα πρέπει να ισορροπήσουμε δύο παράγοντες. Ο πρώτος αφορά την εξασφάλιση της εξερεύνησης και την μη απόκλιση καλύτερων περιπτώσεων, και ο δεύτερος την ασυμπτωτική σύγκλιση στην πολιτική στην οποία πλέον δεν εξερευνούμε περεταίρω, π^*. Ένας τρόπος για να πετύχουμε την ισορροπία είναι η μέθοδος Άπληστη στο Όριο με Απεριόριστη Εξερεύνηση (Greedy in the Limit with Infinite Exploration – GLIE), στην οποία κάθε ζεύγος κατάστασης – δράσης εξερευνιέται άπειρες φορές:

$$\lim_{k\to\infty} N_k(s,a) = \infty$$

και η πολιτική συγκλίνει σε μια άπληστη πολιτική:

$$\lim_{k\to\infty} \pi_k(a|s) = \mathbb{1}(a = \underset{a\in\mathcal{A}}{\operatorname{argmax}}\, Q_k(s,a'))$$

Στη μέθοδο αυτή, ο πράκτορας ακολουθεί μια πολιτική π και δειγματίζει ένα επεισόδιο $\{S_1, A_1, R_2, \dots, S_T\} \sim \pi$. Για κάθε κατάσταση S_t και δράση A_t σε ένα επεισόδιο:

$$N(S_t, A_t) \leftarrow N(S_t, A_t) + 1$$

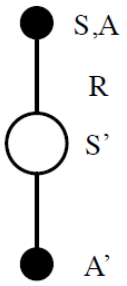
$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{1}{N(S_t, A_t)}(G_t - Q(S_t, A_t))$$

Ενώ, τέλος, βελτιώνουμε την πολιτική με βάση την νέα συνάρτηση δράσης-αξίας:

$$\epsilon \leftarrow {}^1\!/_k$$

$$\pi \leftarrow \epsilon - greedy(Q)$$

Μέσω αυτών των βημάτων ο GLIE MC συγκλίνει στην βέλτιστη συνάρτηση δράσης-αξίας, $Q(s,a) \to q_*(s,a)$.



Σχήμα 2.7:

Αλγόριθμος SARSA, από [44]

Πέραν του ελέγχου MC, μπορούμε να κάνουμε χρήση TD ελέγχου, με TD αξιολόγηση για να προβλέψουμε την $Q(s,a)$, ενώ το βήμα της βελτίωσης πολιτικής να είναι και πάλι ϵ-άπληστο, και η ενημέρωση της συνάρτησης δράσης-αξίας να γίνεται σε κάθε χρονικό βήμα ως ακολούθως:

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

Η λογική της ανανέωσης αυτής βασίζεται στο ότι η δράση A που επιλέγεται στην κατάσταση S σύμφωνα με την πολιτική π, δίνει επιβράβευση R, οδηγώντας σε μια νέα κατάσταση S', από την οποία επιλέγεται η δράση A'. Ο αλγόριθμος αυτός ονομάζεται SARSA, από τα γράμματα που χρησιμοποιούνται στην ανανέωση του Q.

Ο αλγόριθμος SARSA συγκλίνει στη βέλτιστη συνάρτηση δράσης-αξίας $Q(s,a) \rightarrow q_*(s,a)$ με την προϋπόθεση ότι διαθέτουμε GLIE ακολουθία πολιτικών $\pi_t(a|s)$ και ότι ικανοποιούνται οι νόμοι Robbins-Monro της ακολουθίας του ρυθμού μάθησης a_t:

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

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

Με την ίδια λογική που ακολουθήσαμε στην πορεία μας από τον TD(0) στον n-βημάτων TD έως τον TD(λ), έχουμε:

$n = 1$	$SARSA$	${q_t}^{(1)} = R_{t+1} + \gamma Q(S_{t+1})$
$n = 2$		${q_t}^{(2)} = R_{t+1} + \gamma Q_{t+2} + \gamma^2 Q(S_{t+2})$
⋮		⋮
$n = k$	$k - step$ $Q - return$	${q_t}^{(k)} = R_{t+1} + \gamma Q_{t+2} + \cdots + \gamma^k Q(S_{t+k})$
⋮		⋮
$n = \infty$	MC	${q_t}^{(\infty)} = R_{t+1} + \gamma Q_{t+2} + \cdots + \gamma^{T-1} R_T$

Πίνακας 2.2: Q-αποδόσεις n-βημάτων, για $n = 1,2, \dots, k, \dots, \infty$, από [44]

Με τη σειρά του, ο αλγόριθμος n-βημάτων SARSA πραγματοποιεί τις ανανεώσεις του $Q(s,a)$ προς την Q-απόδοση n-βημάτων:

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

Έτσι καταλήγουμε στον SARSA(λ) εμπρόσθιο αλγόριθμο, χρησιμοποιώντας κατά τα γνωστά τα βάρη $(1-\lambda)\lambda^{n-1}$ για την απόδοση ${q_t}^{\lambda} = (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}{q_t}^{(n)}$. Επομένως οδηγούμαστε στην ανανέωση:

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

Για την θεμελίωση του οπίσθιου αλγορίθμου, θα γίνει και πάλι χρήση των ιχνών επιλεξιμότητας με τον εξής τρόπο:

$$\begin{cases} E_0(s,a) = 0 \\ E_t(s,a) = \gamma\lambda E_{t-1}(s,a) + \mathbb{1}(S_t = s, A_t = a) \end{cases}$$

Επομένως, έχοντας ένα ίχνος επιλεξιμότητας για κάθε κατάσταση, και $\delta_t = R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)$, η ενημέρωση του $V(s)$ γίνεται ως εξής:

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

2.5.2 Έλεγχος εκτός της Πολιτικής

Στην περίπτωση ελέγχου εκτός πολιτικής (Off-policy Control), ο στόχος μας είναι να αξιολογήσουμε την πολιτική $\pi(a|s)$ για να υπολογίσουμε το $v_\pi(s)$ ή το $q_\pi(s,a)$, ακολουθώντας μια συμπεριφορική πολιτική μ(a|s), όπου $\{S_1, A_1, R_2, \dots, S_T\} \sim \mu$. Η σημαντικότητα αυτού του είδους ελέγχου έγκειται στο ότι έτσι θα μπορούμε να μαθαίνουμε παρατηρώντας ανθρώπους ή άλλους πράκτορες, να επαναχρησιμοποιούμε την εμπειρία που έχει παραχθεί από παλαιότερες πολιτικές $\pi_1, \pi_2, \dots, \pi_{t-1}$, να μαθαίνουμε για τη βέλτιστη πολιτική ενώ ακολουθούμε κάποια πολιτική εξερεύνησης και να μαθαίνουμε πολλαπλές πολιτικές, ενώ ακολουθούμε μόνο μία.

2.5.3 Importance Sampling

Τα παραπάνω μπορούν να επιτευχθούν με δύο μηχανισμούς, ο πρώτος εκ των οποίων είναι η λογική της Δειγματοληψίας Σπουδαιότητας - ΔΣ (Importance Sampling - IS), δηλαδή εκτιμώντας μια διαφορετική κατανομή, $Q(X)$, σε σχέση με αυτή που έχουμε, $P(X)$. Με άλλα λόγια λέμε πως η εκτίμηση που περιμένουμε είναι το άθροισμα κάποιων πιθανοτήτων $P(X)$ επί την επιβράβευση που μας επιστρέφεται $f(X)$:

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

Σχεδόν όλες οι μέθοδοι εκτός πολιτικής χρησιμοποιούν ΔΣ, μια γενική τεχνική για τον υπολογισμό αναμενόμενων τιμών κάτω υπό μια κατανομή δεδομένων δειγμάτων, από μια άλλη. Μπορούμε να εφαρμόσουμε τη ΔΣ στη μάθηση MC, εκτελώντας τη ΔΣ κατά μήκος μιας πλήρους τροχιάς. Ο τρόπος που εφαρμόζουμε τη ΔΣ στη μάθηση εκτός πολιτικής είναι σταθμίζοντας τις αποδόσεις

σύμφωνα με τη σχετική πιθανότητα οι τροχιές να προκύψουν υπό την πολιτική-στόχο και από τη συμπεριφορική πολιτική. Αυτός ο λόγος ονομάζεται πηλίκο ΔΣ (IS ratio). Δεδομένης μιας κατάστασης S_t, η πιθανότητα της επόμενης τροχιάς καταστάσεων-δράσεων $A_t, S_{t+1}, A_{t+1}, S_{t+2}, \dots, S_T$ υπό την πολιτική π είναι:

$$\begin{aligned}\Pr\{A_t, S_{t+1}, A_{t+1}, \dots, S_T | S_t, A_{t:T-1} \sim \pi\} \\ &= \pi(A_t|S_t)p(S_{t+1}|S_t, A_t)\pi(A_{t+1}|S_{t+1}) \dots p(S_{T+1}|S_{T-1}, A_{T-1}) \\ &= \prod_{k=t}^{T-1} \pi(A_k|S_k)p(S_{k+1}|S_k, A_k),\end{aligned}$$

όπου το p εδώ είναι η συνάρτηση πιθανότητας κατάστασης-μετάβασης της σχέσης (3). Έτσι, η σχετική πιθανότητα της τροχιάς κάτω από τις πολιτικές στόχου και συμπεριφοράς (το πηλίκο ΔΣ) είναι:

$$\rho_{t:T} = \frac{\prod_{k=t}^{T} \pi(A_k|S_k)p(S_{k+1}|S_k, A_k)}{\prod_{k=t}^{T} \mu(A_k|S_k)p(S_{k+1}|S_k, A_k)} = \prod_{k=t}^{T} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)}$$

Ενώ η απόδοση με ΔΣ θα είναι:

$$G_t^{\pi/\mu} = \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)} \frac{\pi(A_{t+1}|S_{t+1})}{\mu(A_{t+1}|S_{t+1})} \dots \frac{\pi(A_T|S_T)}{\mu(A_T|S_T)} G_t = \prod_{k=t}^{T} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)} G_t = \rho_{t:T} G_t$$

Επομένως, η ανανέωση θα γίνει προς το μέρος της διορθωμένης απόδοσης, $G_t^{\pi/\mu}$, ως εξής:

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

Η μέθοδος αυτή λειτουργεί υπό τον περιορισμό ότι η συμπεριφορική πολιτική, μ, θα πρέπει να έχει μη-μηδενικές τιμές όταν η πολιτική στόχος, π, είναι μη-μηδενική. Επιπλέον, όπως αντιλαμβανόμαστε, η εν λόγω μέθοδος έχει πολύ μεγάλη διακύμανση και στην πράξη δεν είναι χρήσιμη. Συνεπώς, η μάθηση MC δεν ενδείκνυται εκτός-πολιτικής.

Στη θέση της μεθόδου MC ΔΣ θα χρησιμοποιήσουμε την TD ΔΣ μάθηση, μέσω των TD targets. Στην περίπτωση της TD ΔΣ, θα σταθμίσουμε τα TD targets, $R_{t+1} + \gamma V(S_{t+1})$, με το πηλίκο ΔΣ. Υπενθυμίζουμε πως στο TD οι ανανεώσεις είναι ενός βήματος:

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

Εναλλακτικά, μπορούμε να εξετάσουμε το ενδεχόμενο μάθησης της συνάρτησης δράσης-αξίας, $Q(s, a)$, εκτός πολιτικής, χωρίς τη χρήση ΔΣ. Η μέθοδος αυτή ονομάζεται Q-μάθηση (Q-learning). Στην περίπτωση αυτή, η επόμενη δράση επιλέγεται χρησιμοποιώντας τη συμπεριφορική πολιτική $A_{t+1} \sim \mu(A_t|S_t)$. Επιπλέον, θεωρούμε μια εναλλακτική επόμενη δράση A'_{t+1}την οποία θα επιλέγαμε αν ακολουθούσαμε την πολιτική στόχο, π. Το βήμα της ανανέωσης της τιμής Q για την κατάσταση στην οποία ξεκινήσαμε και για τη δράση που επιλέξαμε, αποτελεί ανανέωση προς την τιμή της εναλλακτικής δράσης. Συνεπώς, όταν κάνουμε bootstrap, εκκινούμε από την αξία της

εναλλακτικής δράσης, επειδή αυτή είναι η τιμή που μας φανερώνει πόση αξία θα είχαμε στην πραγματικότητα από την πολιτική-στόχο μας:

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

Στο βήμα της βελτίωσης πολιτικής, επιτρέπουμε και στις δύο πολιτικές να ενημερωθούν. Η πολιτική στόχος, π, ενημερώνεται άπληστα σε σχέση με το $Q(s, a)$: $\pi(S_{t+1}) = \underset{a'}{\mathrm{arg\,max}}\, Q(S_{t+1}, a')$, ενώ η συμπεριφορική πολιτική, μ, ανανεώνεται με ϵ-άπληστο τρόπο σε σχέση με το $Q(s, a)$. Τότε, το Q-learning target γίνεται:

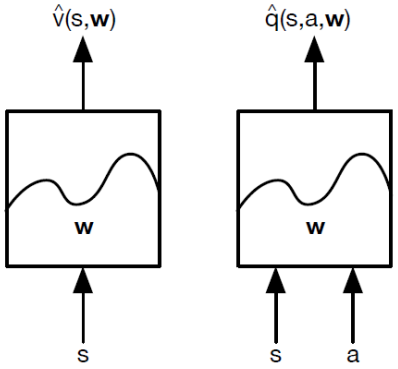
$$R_{t+1} + \gamma Q(S_{t+1}, A') = R_{t+1} + \gamma Q\left(S_{t+1}, \underset{a'}{\mathrm{arg\,max}}\, Q(S_{t+1}, a')\right) = R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a')$$

Ο έλεγχος στο Q-learning (εναλλακτικά SARSAMAX) συγκλίνει στη βέλτιστη συνάρτηση δράσης-αξίας, $Q(s, a) \rightarrow q_*(s, a)$ και η ανανέωση που υφίσταται το $Q(s, a)$ είναι:

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

2.6 Προσέγγιση Συνάρτησης Αξίας

Σε πολλά προβλήματα του πραγματικού κόσμου οι διαστάσεις του χώρου κατάστασης ή και του χώρου δράσεων είναι πάρα πολύ μεγάλες. Ειδικά στην περίπτωση συνεχούς χώρου κατάστασης ή και δράσεων, οι μεγάλες διαστάσεις δεν μπορούν να χειραγωγηθούν εύκολα από τους αλγορίθμους που έχουμε μελετήσει μέχρι πρότινος. Αυτό συμβαίνει καθώς ο τρόπος αναπαράστασης των συναρτήσεων αξίας, v, και δράσης-αξίας, q, στους προηγούμενους αλγορίθμους ήταν με πίνακες, των οποίων η κάθε θέση αντιστοιχεί σε μια διακριτή κατάσταση του χώρου κατάστασης ή και δράσεων. Ο νέος μας στόχος είναι να εντοπίσουμε λειτουργικές μεθόδους, ανεξάρτητα από τη διάσταση του χώρου κατάστασης ή και δράσεων. Η ιδέα της Προσέγγισης Συνάρτησης Αξίας – ΠΣΑ (Value Function Approximation – VFA) βασίζεται σε μια γενίκευση του χώρου κατάστασης ή και δράσεων και των σημείων αυτών που επισκεπτόμαστε συχνότερα. Με τον τρόπο αυτό δεν χρειάζεται να αποθηκεύουμε ξεχωριστά μια διαφορετική αξία για κάθε κατάσταση ή και δράση, αλλά να γενικεύουμε ορισμένες καταστάσεις ή και δράσεις και να τις αντιστοιχίζουμε στην ίδια αξία. Ορίζουμε, επομένως, μια ΠΣΑ, με τη χρήση παραμέτρων $\boldsymbol{w} \in \mathbb{R}^{\boldsymbol{d}}$, για τις συναρτήσεις αξίας που επιθυμούμε να προσεγγίσουμε:



$$\begin{cases} \hat{v}(s, \boldsymbol{w}) \approx v_\pi(s) \\ \text{ή } \hat{q}(s, a, \boldsymbol{w}) \approx q_\pi(s, a) \end{cases}$$

Σχήμα 2.8: Είδη VFA, από [44]

Υπάρχει μεγάλη ποικιλία στα είδη ΠΣΑ που μπορούμε να χρησιμοποιήσουμε και ενδεικτικά αναφέρουμε το γραμμικό συνδυασμό χαρακτηριστικών, τα νευρωνικά δίκτυα, τα δέντρα λήψης αποφάσεων, βάσεις Fourier κ.ά..

Πλέον η μάθηση εστιάζει στις παραμέτρους $\boldsymbol{w}$ του εκάστοτε προσεγγιστή. Από τις πιο διαδεδομένες μεθόδους, την οποία μάλιστα χρησιμοποιούμε κατ' εξοχήν στην επιβλεπόμενη μάθηση, αποτελεί η Κατάβαση Κλίσης (Gradient Descent – GD). Έστω μια παραγωγίσιμη συνάρτηση για το διάνυσμα παραμέτρων $\boldsymbol{w}$, την $J(\boldsymbol{w})$. Ο στόχος μας είναι να βρούμε το διάνυσμα παραμέτρων $\boldsymbol{w}$ για το οποίο ελαχιστοποιείται το μέσο τετραγωνικό σφάλμα ανάμεσα στην προσεγγιστική και στην πραγματική τιμή της συνάρτησης αξίας:

$$J(\boldsymbol{w}) = \mathbb{E}_\pi\left[\left(v_\pi(s) - \hat{v}(s, \boldsymbol{w})\right)^2\right]$$

Στη συνέχεια, προσδιορίζουμε την κλίση της συνάρτησης αυτής: $\nabla_w J(\boldsymbol{w}) = \left(\frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}_1} \ \frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}_2} \ \dots \ \frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}_n}\right)^T$. Για να μπορέσουμε να βρούμε ένα τοπικό ελάχιστο της $J(\boldsymbol{w})$, με χρήση της μεθόδου GD, υπολογίζουμε τις ανανεώσεις των βαρών – $\boldsymbol{w}$ ως εξής:

$$\Delta \boldsymbol{w} = -\frac{1}{2} a \nabla_w J(\boldsymbol{w}) = a \mathbb{E}_\pi\left[\left(v_\pi(s) - \hat{v}(s, \boldsymbol{w})\right) \nabla_w \hat{v}(s, \boldsymbol{w})\right],$$

όπου a είναι ο ρυθμός κλίσης. Στη Στοχαστική Κατάβαση Κλίσης (Stochastic Gradient Descent – SGD) δειγματοληπτούμε την κλίση και οι ανανεώσεις βαρών γίνονται: $\Delta \boldsymbol{w} = a(v_\pi(s) - \hat{v}(s, \boldsymbol{w})) \nabla_w \hat{v}(s, \boldsymbol{w})$, καθώς η αναμενόμενη ανανέωση ισούται με την πλήρη ανανέωση κλίσης.

Στη συνέχεια παρουσιάζουμε κάθε αλγόριθμο που έχουμε αναλύσει προηγουμένως, ανανεωμένο με την έννοια του ΠΣΑ:

		$Target$
$Monte\ Carlo$	$\Delta \boldsymbol{w} = a\left(G_t - \hat{v}(S_t, \boldsymbol{w})\right) \nabla_w \hat{v}(S_t, \boldsymbol{w})$	G_t
$TD(0)$	$\Delta \boldsymbol{w} = a\left(R_{t+1} + \gamma \hat{v}(S_t, \boldsymbol{w}) - \hat{v}(S_t, \boldsymbol{w})\right) \nabla_w \hat{v}(S_t, \boldsymbol{w})$	$R_{t+1} + \gamma \hat{v}(S_t, \boldsymbol{w})$
$Forward\ view\ linear\ TD(\lambda)$	$\Delta \boldsymbol{w} = a\left(G_t^\lambda - \hat{v}(S_t, \boldsymbol{w})\right) \nabla_w \hat{v}(S_t, \boldsymbol{w})$	G_t^λ

$Backward\ view\ linear\ TD(\lambda)$	$\delta_t = R_{t+1} + \gamma\hat{v}(S_t, \boldsymbol{w}) - \hat{v}(S_t, \boldsymbol{w})$ $E_t = \gamma\lambda E_{t-1} + \boldsymbol{x}(S_t)$ $\Delta \boldsymbol{w} = a\delta_t E_t$	$R_{t+1} + \gamma\hat{v}(S_t, \boldsymbol{w})$

Πίνακας 2.3: Value Function Approximation με Stochastic Gradient Descent για διάφορα είδη αλγορίθμων, από [44]

Με τα παραπάνω πραγματοποιείται η διαδικασία της πρόβλεψης, δηλαδή της προσέγγισης της συνάρτησης αξίας. Έπεται η διαδικασία ελέγχου και στη συνέχεια, η διαδικασία ανανέωσης της πολιτικής, με ϵ-άπληστο τρόπο.

Εναλλακτικά, η διαδικασία της πρόβλεψης θα μπορούσε να πραγματοποιηθεί και με στόχο την προσέγγιση της συνάρτησης δράσης-αξίας μέσω του προσεγγιστή $\hat{q}(s, a, \boldsymbol{w})$. Συγκεκριμένα, θα έχουμε:

$$J(\boldsymbol{w}) = \mathbb{E}_\pi[(q_\pi(s, a) - \hat{q}(s, a, \boldsymbol{w}))^2],$$

ενώ θα χρησιμοποιήσουμε SGD για τον εντοπισμό ενός τοπικού ελαχίστου:

$$\Delta \boldsymbol{w} = -\frac{1}{2} a\nabla_w J(\boldsymbol{w}) = a(q_\pi(s, a) - \hat{q}(s, a, \boldsymbol{w}))\nabla_w \hat{q}(s, a, \boldsymbol{w})$$

Στη συνέχεια παρουσιάζουμε κάθε αλγόριθμο που έχουμε αναλύσει προηγουμένως, ανανεωμένο με την έννοια του ΠΣΑ:

$Monte\ Carlo$	$\Delta \boldsymbol{w} = a(G_t - \hat{q}(S_t, A_t, \boldsymbol{w}))\nabla_w \hat{v}(S_t, A_t, \boldsymbol{w})$
$TD(0)$	$\Delta \boldsymbol{w} = a(R_{t+1} + \gamma\hat{q}(S_t, A_t, \boldsymbol{w}) - \hat{q}(s, a, \boldsymbol{w}))\nabla_w \hat{v}(S_t, A_t, \boldsymbol{w})$
$Forward\ view\ linear\ TD(\lambda)$	$\Delta \boldsymbol{w} = a\big(G_t^\lambda - \hat{q}(S_t, A_t, \boldsymbol{w})\big)\nabla_w \hat{v}(S_t, A_t, \boldsymbol{w})$
$Backward\ view\ linear\ TD(\lambda)$	$\delta_t = R_{t+1} + \gamma\hat{q}(S_t, A_t, \boldsymbol{w}) - \hat{v}(S_t A_t, , \boldsymbol{w})$ $E_t = \gamma\lambda E_{t-1} + \boldsymbol{x}(S_t)$ $\Delta \boldsymbol{w} = a\delta_t E_t$

Πίνακας 2.4: Action - Value Function Approximation με Stochastic Gradient Descent για διάφορα είδη αλγορίθμων, από [44]

Έπειτα από την παραπάνω διαδικασία της πρόβλεψης ακολουθεί το βήμα ελέγχου και στη συνέχεια, η διαδικασία ανανέωσης της πολιτικής, με ϵ-άπληστο τρόπο. Σύμφωνα με την ϵ-άπληστη βελτίωση πολιτικής:

$$a = \begin{cases} \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, \hat{q}(s, a', \boldsymbol{w}), & με\ πιθανότητα\ 1 - \epsilon \\ random, & με\ πιθανότητα\ \epsilon \end{cases}$$

Αναφορικά στη σύγκλιση των παραπάνω αλγορίθμων πρόβλεψης, ο MC συγκλίνει και εντός και εκτός πολιτικής, είτε με γραμμικό είτε με μη-γραμμικό προσεγγιστή. Ο TD(0) στην περίπτωση εντός πολιτικής συγκλίνει μόνο με γραμμικό προσεγγιστή, ενώ εκτός πολιτικής όχι. Επίσης δε συγκλίνει με μη-γραμμικό προσεγγιστή, ανεξαρτήτως πολιτικής. Τέλος, ο TD(λ) επίσης συγκλίνει στις ίδιες περιπτώσεις με τον TD(0).

Αναφορικά στη σύγκλιση των παραπάνω αλγορίθμων ελέγχου, καμία μέθοδος δε συγκλίνει με χρήση μη-γραμμικού προσεγγιστή. Μάλιστα, η Q-μάθηση δε συγκλίνει ούτε με τη χρήση γραμμικού προσεγγιστή. Αντιθέτως, στον έλεγχο MC και στον SARSA το αποτέλεσμα βρίσκεται γύρω από μια πλησίον-της-βέλτιστης συνάρτηση αξίας, χωρίς την εγγύηση ότι ο προσεγγιστής θα έχει βελτίωση πολιτικής, αν έχουμε γραμμικό προσεγγιστή.

2.6.1 Batch Methods

Μέχρι στιγμής έχουμε εξετάσει τον αλγόριθμο Κατάβασης Κλίσης (GD), ο οποίος είναι απλός στην εφαρμογή του, καθώς οι ανανεώσεις που εφαρμόζει δεν είναι υπολογιστικά απαιτητικές. Εντούτοις, η ανανέωση που πραγματοποιούμε γίνεται με βάση μια εμπειρία που βλέπουμε για πρώτη φορά, την οποία όμως δεν επαναχρησιμοποιούμε στη συνέχεια. Το γεγονός αυτό κάνει τον αλγόριθμο GD μη αποδοτικό ως προς τη χρήση και την εκμετάλλευση των δεδομένων. Λύση σε αυτό αποτελούν οι μέθοδοι παρτίδας (Batch Methods), οι οποίες αναζητούν την καλύτερη συνάρτηση αξίας που ταιριάζει στα δεδομένα μας, δηλαδή στην εμπειρία του πράκτορα. Αυτό το «ταίριασμα» της συνάρτησης αξίας θα μπορούσε να βασίζεται στη λογική ελαχίστων τετραγώνων, όπου θα προσπαθήσουμε να ελαχιστοποιήσουμε το άθροισμα των τετραγώνων των διαφορών μεταξύ του προσεγγιστή $\hat{v}(s,\boldsymbol{w})$ και της αξίας-στόχου v_t^π:

$$LS(\boldsymbol{w}) = \sum_{t=1}^{T} (v_t^\pi - \hat{v}(s,\boldsymbol{w}))^2 \xrightarrow[\mathbb{E}_{\mathcal{D}}:\, expectation]{\mathcal{D} = \{\langle s_1, v_1^\pi\rangle, \langle s_2, v_2^\pi\rangle, \dots, \langle s_T, v_T^\pi\rangle\}} \mathbb{E}_{\mathcal{D}}[(v_t^\pi - \hat{v}(s,\boldsymbol{w}))^2]$$

Εξέλιξη του παραπάνω αποτελεί η Στοχαστική Κατάβαση Κλίσης με Επανάληψη Εμπειρίας (Stochastic GR with Experience Replay). Στη μέθοδο αυτή διαθέτουμε την εμπειρία $\mathcal{D} = \{\langle s_1, v_1^\pi\rangle, \langle s_2, v_2^\pi\rangle, \dots, \langle s_T, v_T^\pi\rangle\}$ την οποία αποθηκεύουμε σε μια κρυφή μνήμη. Στη συνέχεια, σε κάθε χρονική στιγμή – βήμα, δειγματοληπτούμε μια κατάσταση και μια αξία από την εμπειρία μας $\langle s, v^\pi\rangle \sim \mathcal{D}$ και πραγματοποιούμε μια στοχαστική ανανέωση κλίσης (Stochastic Gradient Update):

$$\Delta\boldsymbol{w} = a\big(v^\pi - \hat{v}(S_t,\boldsymbol{w})\big)\nabla_w \hat{v}(S_t,\boldsymbol{w})$$

Τα βήματα που περιγράψαμε μας θυμίζουν λογική επιβλεπόμενης μάθησης. Σε κάθε περίπτωση, η διαδικασία αυτή συγκλίνει τη λύση ελαχίστων τετραγώνων: $\boldsymbol{w}^\pi = \underset{\boldsymbol{w}}{\operatorname{argmin}} LS(\boldsymbol{w})$.

2.6.2 Αλγόριθμος Deep Q-Networks (DQN)

Βασιζόμενοι στα προαναφερθέντα και συγκεκριμένα στην ιδέα της Επανάληψης Εμπειρίας και στην Q-μάθηση, θα περιγράψουμε τον αλγόριθμο Deep Q-Networks (DQN), ο οποίος είναι βάση της μάθησης εκτός πολιτικής. Ο DQN είναι ένας αλγόριθμος RL χωρίς μοντέλο, στον οποίο χρησιμοποιείται η τεχνική της βαθιάς μάθησης. Οι αλγόριθμοι DQN χρησιμοποιούν την Q-μάθηση για να μάθουν την καλύτερη δράση που πρέπει να κάνουν σε μια δεδομένη κατάσταση, ενώ μέσω ενός βαθιού Νευρωνικού Δικτύου (Neural Network – NN) ή ενός συνελικτικού NN πραγματοποιούν εκτίμηση της συνάρτησης αξίας Q. Η βασική ιδέα του αλγορίθμου σχετίζεται με την αποθήκευση όλων των μεταβάσεων που έχουμε παρατηρήσει έως το τρέχον βήμα, $(s_t, a_t, r_{t+1}, s_{t+1})$, σε μια Replay Memory $\mathcal{D}$. Σε κάθε βήμα επιλέγουμε μια δράση a_t σύμφωνα με μια ϵ-άπληστη πολιτική, σύμφωνα με τον προσεγγιστή μας. Στη συνέχεια, δειγματοληπτούμε τυχαία ένα minibatch από τις μεταβολές που έχουμε αποθηκεύσει στη μνήμη $\mathcal{D}$, (s, a, r, s'). Επιπλέον, διαθέτουμε ένα μεγάλο NN το οποίο πραγματοποιεί εκτίμηση για όλα τα Q-values, σύμφωνα με τις παλιές, σταθερές παραμέτρους w^-, $r + \gamma \max_{a'} Q(s', a'; w^-)$. Έπειτα, βελτιστοποιούμε το ελάχιστο τετραγωνικό σφάλμα ανάμεσα στα Q-targets και στην πρόβλεψη του Q-δικτύου:

$$\mathcal{L}_i(w_i) = \mathbb{E}_{s,a,r,s' \sim \mathcal{D}_i}\left[\left(r + \gamma \max_{a'} Q(s', a'; w_i^-) - Q(s, a; w_i)\right)^2\right]$$

Τέλος, ανανεώνουμε τις παλιές παραμέτρους $w^- = w$.

Η μέθοδος DQN είναι ευσταθής με τη χρήση NN, σε αντίθεση με τις μεθόδους SARSA και TD, διότι η χρήση Επανάληψης Εμπειρίας σταθεροποιεί τις μεθόδους των NN επειδή απο-συσχετίζει τις τροχιές (μέσω της τυχαίας δειγματοληψίας από τη μνήμη $\mathcal{D}$, η σειρά με την οποία προκύπτουν τα τμήματα μιας τροχιάς είναι τυχαία). Ένας επιπλέον λόγος είναι ότι κατά τον υπολογισμό των target της Q-μάθησης χρησιμοποιούμε τις παλιές, σταθερές παραμέτρους w^-, οι οποίες μας εξασφαλίζουν μια πιο ευσταθή αναβάθμιση και σταθεροποιούν το target μας. Για παράδειγμα, στην περίπτωση του TD, κάθε φορά που ανανεώνουμε τις παραμέτρους w, αλλάζουμε ελαφρώς και το target που έχουμε. Έτσι, στην περίπτωση μη-γραμμικών προσεγγιστών, μπορεί να καταλήξουμε σε αστάθεια.

2.7 Κλίση Πολιτικής

Οι μέθοδοι Κλίσης Πολιτικής (Policy Gradient) είναι πανταχού παρούσες σε αλγορίθμους Ενισχυτικής Μάθησης χωρίς μοντέλο (model-free). Η μέθοδος Κλίσης Πολιτικής είναι επίσης το τμήμα του «Δράστη» στις μεθόδους Δράστη-Κριτή που θα αναλύσουμε στη συνέχεια και χρησιμοποιούμε κατ' εξοχήν στην παρούσα εργασία.

Στην ουσία, οι μέθοδοι κλίσης πολιτικής ενημερώνουν την κατανομή πιθανοτήτων των δράσεων, έτσι ώστε οι δράσεις με υψηλότερη αναμενόμενη αξία να έχουν υψηλότερη τιμή πιθανότητας για μια παρατηρούμενη κατάσταση. Μέχρι πρότινος η πολιτική προέκυπτε άμεσα από τη συνάρτηση αξίας, λόγου χάρη με ϵ-άπληστο τρόπο. Στην παρούσα μέθοδο, όμως, θα παραμετροποιήσουμε άμεσα την πολιτική ως εξής:

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

όπου θ οι παράμετροι που χρησιμοποιήσαμε για να προσεγγίσουμε τη συνάρτηση αξίας ή τη συνάρτηση δράσης-αξίας: $V_\theta(s) \approx V^\pi(s), Q_\theta(s,a) \approx Q^\pi(s,a)$.

Οι μέθοδοι Ενισχυτικής Μάθησης βάσει Πολιτικής υπερτερούν αυτών βάσει Αξίας (value based RL) καθώς εμφανίζουν καλύτερη σύγκλιση, είναι αποδοτικές σε μεγάλες διαστάσεις ή σε συνεχείς χώρους δράσεων και μπορούν να μάθουν επίσης στοχαστικές πολιτικές. Αντιθέτως, συνήθως συγκλίνουν σε ένα τοπικό έναντι ολικού βελτίστου και η αξιολόγηση πολιτικής είναι συχνά ανεπαρκής και με υψηλή διακύμανση-απόκλιση.

Αρχικά θα ορίσουμε τις αντικειμενικές συναρτήσεις $J(\theta)$. Σε επεισοδιακά περιβάλλοντα μπορούμε να χρησιμοποιήσουμε την αρχική τιμή της συνάρτησης αξίας, $V^{\pi_\theta}(s_1)$: $J_1(\theta) = V^{\pi_\theta}(s_1) = \mathbb{E}_{\pi_\theta}[v_1]$. Σε συνεχή περιβάλλοντα μπορούμε να χρησιμοποιήσουμε τη μέση τιμή της αξίας: $J_{av^V}(\theta) = \sum_s d^{\pi_\theta}(s)\, V^{\pi_\theta}(s)$ ή τη μέση επιβράβευση ανά βήμα: $J_{av^R}(\theta) = \sum_s d^{\pi_\theta}(s) \sum_a \pi_\theta\,(s,a) R_s^a$, όπου $d^{\pi_\theta}(s)$ είναι μια σταθερή κατανομή των καταστάσεων της Μαρκοβιανής αλυσίδας για το π_θ.

Το πρόβλημά μας είναι ένα πρόβλημα βελτιστοποίησης, κατά το οποίο αναζητούμε το θ που μεγιστοποιεί την αντικειμενική συνάρτηση $J(\theta)$. Για να επιτευχθεί αυτός ο στόχος μπορούμε να χρησιμοποιήσουμε είτε αλγορίθμους μηδενικής τάξης (που δεν χρησιμοποιούν gradient), είτε πρώτης τάξης (με χρήση του gradient), είτε δεύτερης τάξης (με χρήση της hessian). Στην παρούσα εργασία θα εστιάσουμε στη μέθοδο κατάβασης κλίσης.

Στους αλγορίθμους κλίσης πολιτικής, αναζητούμε ένα τοπικό μέγιστο στην $J(\theta)$, ανεβαίνοντας την κλίση της πολιτικής, ως προς τις παραμέτρους θ: $\Delta\theta = \alpha \nabla_\theta J(\theta)$, όπου $\nabla_\theta J(\theta)$ είναι η κλίση πολιτικής: $\nabla_\theta J(\theta) = \left(\frac{\partial J(\theta)}{\partial \theta_1} \;\; \frac{\partial J(\theta)}{\partial \theta_2} \;\; \dots \;\; \frac{\partial J(\theta)}{\partial \theta_n}\right)^T$ και α ο ρυθμός μάθησης. Ένας τρόπος υπολογισμού των κλίσεων είναι με τη μέθοδο Πεπερασμένων Διαφορών (Finite Differences), σύμφωνα με την οποία για να αξιολογήσουμε την πολιτική κλίσης της $\pi_\theta(s,a)$, εκτιμούμε την k-οστή μερική παράγωγο της αντικειμενικής συνάρτησης ως προς το θ, για κάθε διάσταση $k \in [1,n]$, προσθέτοντας μια μικρή διαταραχή ϵ στο θ, στην k-οστή διάσταση: $\frac{\partial J(\theta)}{\partial \theta_k} \approx \frac{J(\theta + \epsilon u_k) - J(\theta)}{\epsilon}$, όπου u_k το μοναδιαίο διάνυσμα με μονάδα στη θέση του k-οστού στοιχείου και μηδέν σε κάθε άλλη θέση. Έπειτα, χρησιμοποιεί n αξιολογήσεις για να υπολογίσει την κλίση πολιτικής σε n διαστάσεις. Η εν λόγω μεθοδολογία είναι απλή, θορυβώδης και ανεπαρκής, αλλά μερικές φορές αποτελεσματική, καθώς λειτουργεί με τυχαίες πολιτικές, ακόμα κι αν δεν είναι διαφορίσιμες.

Εναλλακτικά, κάνουμε αναλυτικό υπολογισμό της κλίσης πολιτικής και θεωρούμε πως η πολιτική π_θ είναι διαφορίσιμη όπου είναι μη-μηδενική, καθώς επίσης πως γνωρίζουμε την κλίση $\nabla_\theta \pi_\theta(s,a)$. Ορίζουμε στη συνέχεια, τους λόγους πιθανότητας (Likelihood ratios):

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

όπου η συνάρτηση $\nabla_\theta log \pi_\theta(s,a)$ ονομάζεται Score Function.

Παρακάτω θα παρουσιάσουμε το Θεώρημα Κλίσης Πολιτικής, το οποίο γενικεύει την προσέγγιση των λόγων πιθανότητας σε πολυδιάστατες MDPs. Επιπλέον, αντικαθιστά τη στιγμιαία επιβράβευση r με τη μακροπρόθεσμη τιμή $Q(s,a)$.

Θεώρημα Κλίσης Πολιτικής -Policy Gradient Theorem

Για κάθε διαφορίσιμη πολιτική $\pi_\theta(s,a)$, για κάθε αντικειμενική συνάρτηση πολιτικής $J = J_1,\ J_{av^R}$ ή $\frac{1}{1-\gamma}J_{av^V}$, η πολιτική κλίσης είναι:

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

▪

Ο αλγόριθμος REINFORCE – Monte-Carlo κλίσης πολιτικής (MC Policy Gradient) αποτελεί χρήση του παραπάνω θεωρήματος, και ανανεώνει τις παραμέτρους με Stochastic Gradient Ascent, χρησιμοποιώντας την απόδοση $v_t \equiv G_t$ ως unbiased εκτίμηση της $Q^{\pi_\theta}(s,a)$:

$$\Delta\theta_t = a\nabla_\theta log\pi_\theta(s,a)v_t$$

```
function REINFORCE
    Initialise θ arbitrarily
    for each episode {s1, a1, r2, ..., sT-1, aT-1, rT} ~ πθ do
        for t=1 to T-1 do
            θ <- θ + α∇θlogπθ(st, at)vt
        end for
    end for
    return θ
end function
```

Πίνακας 2.5: Αλγόριθμος REINFORCE, από [43]

2.8 Μέθοδοι Δράστη-Κριτή

Η μέθοδος MC κλίσης πολιτικής εισάγει μεγάλη διακύμανση στις εκτιμήσεις, δεδομένης της χρήσης της απόδοσης $v_t \equiv G_t$ στην ανανέωση των παραμέτρων. Επομένως, αντί να χρησιμοποιήσουμε την απόδοση για την εκτίμηση της συνάρτησης δράσης-αξίας, πρόκειται να την εκτιμήσουμε χρησιμοποιώντας έναν κριτή, έναν προσεγγιστή συνάρτησης: $Q_w(s,a) \approx Q^{\pi_\theta}(s,a)$. Με τον τρόπο αυτό θα συνδυάσουμε την προσέγγιση της συνάρτησης αξίας (Value Function Approximation) με τις μεθόδους κλίσης πολιτικής (Policy Gradient).

Έχουμε πλέον δύο ομάδες παραμέτρων, αυτές του δράστη (actor) και αυτές του κριτή (critic). Η ιδέα είναι ότι θα προσαρμόσουμε τον δράστη, δηλαδή την πολιτική, στην κατεύθυνση που – σύμφωνα με τον κριτή – θα έχουμε μεγαλύτερη ανταμοιβή. Πιο αναλυτικά, ο κριτής πραγματοποιεί την ανανέωση των παραμέτρων w της συνάρτησης δράσης-αξίας και ο δράστης ανανεώνει τις παραμέτρους θ στη φορά που προτείνει ο κριτής.

Στις μεθόδους Actor-Critic (AC) κάνουμε χρήση μιας πολιτικής κλίσης κατά προσέγγιση, αντί της πραγματικής κλίσης πολιτικής:

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

ενώ η ανανέωση των παραμέτρων θ του δράστη γίνεται:

$$\Delta\theta = a\nabla_\theta log\pi_\theta(s,a)Q_w(s,a)$$

Το πρόβλημα της ανανέωσης των παραμέτρων w του κριτή είναι ένα πρόβλημα αξιολόγησης πολιτικής, με χρήση προσέγγισης συνάρτησης αξίας, που γίνεται με τις μεθόδους που έχουμε ήδη περιγράψει παραπάνω, Monte-Carlo policy evaluation, Temporal-Difference learning, TD(λ) κ.ά..

Με στόχο να βελτιώσουμε την παραπάνω διαδικασία, ο τρόπος μείωσης της διακύμανσης του εκτιμητή είναι η αφαίρεση μιας Baseline συνάρτησης $B(s)$ από την πολιτική κλίσης. Με τον τρόπο αυτό, μειώνουμε τη διακύμανση χωρίς να αλλάζουμε την εκτίμηση:

$$\mathbb{E}_{\pi_\theta}[\nabla_\theta log\pi_\theta(s,a)Q_w(s,a)] = \sum_s d^{\pi_\theta}(s)\sum_a \nabla_\theta\pi_\theta\,(s,a)B(s)$$

$$= \sum_s d^{\pi_\theta}(s)\,B(s)\nabla_\theta\sum_a \pi_\theta(s,a) = 0$$

Μια καλή προτεινόμενη baseline είναι η συνάρτηση κατάστασης-αξίας $B(s) = V^{\pi_\theta}(s)$. Ορίζουμε εκ νέου την πολιτική κλίσης χρησιμοποιώντας την advantage function $A^{\pi_\theta}(s,a) = Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)$:

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

ενώ η ανανέωση των παραμέτρων θ του δράστη γίνεται:

$$\Delta\theta = a\nabla_\theta log\pi_\theta(s,a)A^{\pi_\theta}(s,a)$$

Συνεπώς, με τη χρήση της advantage function, η οποία περιέχει την baseline, μειώνουμε σημαντικά τη διακύμανση στη μέθοδο πολιτικής κλίσης. Άρα, ο κριτής θα πρέπει πλέον να εκτιμά την advantage function, για παράδειγμα εκτιμώντας τις $V^{\pi_\theta}(s)$ και $Q^{\pi_\theta}(s,a)$, με τη χρήση δύο προσεγγιστών και δύο διανυσμάτων παραμέτρων και στη συνέχεια ανανεώνοντας και τις δύο συναρτήσεις αξίας (πχ. με TD learning):

$$\begin{cases} V_v(s) \approx V^{\pi_\theta}(s) \\ Q_w(s,a) \approx Q^{\pi_\theta}(s,a) \\ A(s,a) = Q_w(s,a) - V_v(s) \end{cases}$$

Το TD-σφάλμα δ^{π_θ} της πραγματικής συνάρτησης αξίας $V^{\pi_\theta}(s)$ είναι:

$$\delta^{\pi_\theta} = r + \gamma V^{\pi_\theta}(s') - V^{\pi_\theta}(s)$$

και αποτελεί unbiased εκτίμηση της advantage function:

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

Επομένως, μπορούμε να χρησιμοποιήσουμε το TD-σφάλμα δ^{π_θ} για να υπολογίσουμε την πολιτική κλίσης:

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

ενώ η ανανέωση των παραμέτρων θ του δράστη γίνεται:

$$\Delta\theta = a\nabla_\theta log\pi_\theta(s,a)\delta^{\pi_\theta}$$

Στην πράξη, μπορούμε να χρησιμοποιήσουμε ένα προσεγγιστικό TD-σφάλμα $\delta_v = r + \gamma V_v(s') - V_v(s)$. Παρατηρούμε, επομένως, πως αυτή η προσέγγιση απαιτεί μόνο ένα σύνολο παραμέτρων v. Μάλιστα, ο κριτής μπορεί να εκτιμήσει τη συνάρτηση αξίας $V_\theta(s)$ με διαφορετικά targets, ανάλογα τον αλγόριθμο που χρησιμοποιούμε:

		$Targets$
$Monte\ Carlo$	$\Delta\theta = a\big(v_t - V_\theta(s)\big)\varphi(s)$	$v_t \equiv G_t$
$TD(0)$	$\Delta\theta = a\big(r + \gamma V_\theta(s') - V_\theta(s)\big)\varphi(s)$	$r + \gamma V_\theta(s')$
$Forward\ view\ linear\ TD(\lambda)$	$\Delta\theta = a\left(v_t^\lambda - V_\theta(s)\right)\varphi(s)$	$\lambda - return\ v_t^\lambda$
$Backward\ view\ linear\ TD(\lambda)$	$\delta_t = r_{t+1} + \gamma V_\theta(s_{t+1}) - V_\theta(s_t)$ $e_t = \gamma\lambda e_{t-1} + \varphi(s_t)$ $\Delta\theta = a\delta_t e_t$	

όπου $\varphi(s)$ ένας γραμμικός προσεγγιστής.

Πίνακας 2.6: Ανανεώσεις παραμέτρων του κριτή για την εκτίμηση της συνάρτησης αξίας $V_\theta(s)$, από [44]

Ενώ, ο δράστης μπορεί να εκτιμήσει τη συνάρτηση αξίας $V_v(s)$ με διαφορετικά targets, ανάλογα τον αλγόριθμο που χρησιμοποιούμε:

		$Targets$
$Monte\ Carlo$	$\Delta\theta = a\big(v_t - V_v(s)\big)\nabla_\theta log\pi_\theta(s,a)$	$v_t \equiv G_t$
$TD(0)$	$\Delta\theta = a\big(r + \gamma V_v(s) - V_v(s)\big)\nabla_\theta log\pi_\theta(s,a)$	$r + \gamma V_v(s')$
$Forward\ view\ linear\ TD(\lambda)$	$\Delta\theta = a\left(v_t^\lambda - V_v(s)\right)\nabla_\theta log\pi_\theta(s,a)$	$\lambda - return\ v_t^\lambda$
$Backward\ view\ linear\ TD(\lambda)$	$\delta_t = r_{t+1} + \gamma V_v(s_{t+1}) - V_v(s_t)$ $e_t = \gamma\lambda e_{t-1} + \nabla_\theta log\pi_\theta(s,a)$ $\Delta\theta = a\delta_t e_t$	

Πίνακας 2.7: Ανανεώσεις παραμέτρων του δράστη για την εκτίμηση της συνάρτησης αξίας $V_v(s)$, από [44]

Μάλιστα, οι παραπάνω ανανεώσεις μπορούν να γίνουν online.

2.9 Φυσική Κλίση Πολιτικής

Μέχρι στιγμής, εκτιμούμε την κλίση πολιτικής δειγματοληπτώντας τον ίδιο το θόρυβο που έχουμε παράξει, δεδομένου ότι έχουμε θορυβώδεις πολιτικές, των οποίων θέλουμε τις εκτιμήσεις. Στην περίπτωση μιας Gaussian πολιτικής, όσο αυτή γίνεται πιο ευρεία, τόσο απειρίζονται οι αποκλίσεις της εκτίμησής της. Η εναλλακτική είναι να εργαστούμε άμεσα με την οριακή περίπτωση. Αντί να προσθέτουμε θόρυβο στην πολιτική μας και έπειτα να προσπαθήσουμε να τον μειώσουμε, καταλήγοντας σε κάτι περίπου ντετερμινιστικό, είναι προτιμότερο να ξεκινάμε θεωρώντας εξ αρχής ντετερμινιστικές πολιτικές. Επομένως, έχοντας μια ντετερμινιστική πολιτική, θα προσπαθήσουμε να προσαρμόσουμε τις παραμέτρους της, έτσι ώστε να μας μεγιστοποιήσει την αντικειμενική συνάρτηση που μελετήσαμε παραπάνω. Αν θεωρήσουμε την οριακή περίπτωση του θεωρήματος κλίσης πολιτικής, καταλήγουμε στην παρακάτω μορφή ανανέωσης:

$$\nabla_\theta^{nat}\pi_\theta(s,a) = G_\theta^{-1}\nabla_\theta\pi_\theta(s,a),$$

όπου G_θ είναι ο πίνακας πληροφορίας Fisher:

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

Η παραπάνω τεχνική ονομάζεται Ντετερμινιστική/Φυσική Κλίση Πολιτικής (Deterministic/Natural Policy Gradient) και λειτουργεί πολύ καλύτερα σε περιπτώσεις συνεχούς χώρου δράσεων σε σχέση με τη Στοχαστική Κλίση Πολιτικής.

Κεφάλαιο 3

Αλγόριθμος Continuous Actor Critic Learning Automaton

Η έρευνα στην Ενισχυτική Μάθηση έχει εστιάσει περισσότερο σε προβλήματα με συνεχή περιβάλλοντα παρά σε προβλήματα όπου και οι δράσεις μπορούν να επιλεγούν από συνεχή χώρο. Ο αλγόριθμος Continuous Actor Critic Learning Automaton - CACLA αποτελεί έναν τρόπο διαχείρισης περιπτώσεων όπου τόσο ο χώρος των καταστάσεων (state space) όσο των δράσεων (action space) είναι συνεχής. Συχνά, σε ορισμένα προβλήματα πραγματοποιείται διακριτοποίηση του χώρου δράσεων, γεγονός που ελλοχεύει κινδύνους, δεδομένου ότι μπορεί από το χώρο δράσεων να εξαιρεθούν τμήματα τα οποία είναι πιο σημαντικά από άλλα, τα οποία τελικά συμπεριλάβαμε. Επιπλέον, ενδέχεται να μην είναι εξαρχής προφανείς οι περιοχές ενδιαφέροντος στο χώρο δράσεων. Επίσης, η διακριτοποίηση του χώρου δράσεων δυσχεραίνει την ικανότητα γενίκευσης των προβλημάτων, ενώ η διαδικασία της μάθησης μπορεί να είναι πιο αργή όταν ο διακριτός χώρος δράσεων περιέχει πολλά στοιχεία. Συνεπώς, με στόχο την αντιμετώπιση των παραπάνω, προτείνεται ένας αλγόριθμος που δύναται να διαχειριστεί συνεχείς χώρους δράσεων [45].

3.1 Ο αλγόριθμος

Ο αλγόριθμος CACLA είναι model-free (εκτός μοντέλου), ο δράστης δεν γνωρίζει τη δυναμική του περιβάλλοντος και δεν χρειάζεται να τον περιμένουμε να διαμορφώσει μια εικόνα για αυτήν, προτού αρχίσει να εκπαιδεύεται. Επιπλέον, η προσπάθεια δημιουργίας ενός μοντέλου για την περιγραφή ενός προβλήματος πραγματικού κόσμου – πόσο μάλλον στην περίπτωση συνεχούς χώρου καταστάσεων και δράσεων – μπορεί εν δυνάμει να είναι πολύ σύνθετη, ενώ η εύρεση της βέλτιστης συμπεριφοράς μέσω της εκπαίδευσης να είναι μια πολύ πιο απλή διαδικασία. Για παράδειγμα, στην περίπτωση που ο πράκτορας έχει να επιλέξει ανάμεσα σε δύο δράσεις, εκ των οποίων η μια είναι εμφανώς καλύτερη από την άλλη, η προσπάθεια να μοντελοποιήσει πρώτα το περιβάλλον θα τον καθυστερήσει πολύ και θα του κοστίσει υπολογιστικά, σε σχέση με το να βρει γρήγορα και απλά τη βέλτιστη συμπεριφορά που πρέπει να ακολουθήσει [45].

Η εκδοχή του CACLA που θα χρησιμοποιήσουμε είναι ένας online αλγόριθμος, ο οποίος όμως, μπορεί εύκολα να εξελιχθεί σε batch αλγόριθμο. Τα πλεονεκτήματα του online αλγορίθμου είναι η καλύτερη απόδοση σε δυναμικά περιβάλλοντα και η ταχύτερη εκπαίδευση. Αντιθέτως, οι batch αλγόριθμοι αποδίδουν καλύτερα σε προβλήματα με μικρό αριθμό παρατηρήσεων (observations), όταν διαχειριζόμαστε πραγματικά αντικείμενα ή προσομοιώσεις του φυσικού κόσμου ή προβλήματα με μεγάλες υπολογιστικές απαιτήσεις.

Στο κεφάλαιο 2.1 ορίσαμε τη συνάρτηση p ως τη δυναμική μιας MDP, για την οποία ισχύει ότι $p: \mathcal{S} \times R \times \mathcal{S} \times \mathcal{A} \rightarrow [0,1]$ και μπορεί εναλλακτικά, να γραφεί και ως συνάρτηση με τρία ορίσματα $p(s'|s,a)$ όπου $p: \mathcal{S} \times \mathcal{S} \times \mathcal{A} \rightarrow [0,1]$. Στο τρέχον κεφάλαιο, θα συμβολίζουμε με T τη συνάρτηση

αυτή και θα την ονομάσουμε συνάρτηση μετάβασης, όπου το $T(s, a, s')$ μας δίνει την πιθανότητα $Pr(s'|s, a)$. Όπως έχουμε ήδη παρουσιάσει, ο στόχος της ενισχυτικής μάθησης είναι να εκπαιδευτεί ο πράκτορας μαθαίνοντας μια πολιτική επιλογής δράσεων $\pi: S \times A \rightarrow [0,1]$ που να μεγιστοποιεί τη συνολική επιβράβευση, ξεκινώντας από την κατάσταση s, επιλέγοντας τη δράση a, στη χρονική στιγμή t:

$$r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots = \sum_{i=0}^{\infty} \gamma^i r_{t+i}\,, \text{όπου } 0 \leq \gamma \leq 1 \; o\; discount\; factor$$

Θεωρούμε το discount factor ως χαρακτηριστικό του αλγορίθμου και όχι της MDP, καθώς ορισμένοι αλγόριθμοι αποδίδουν πολύ καλύτερα με συγκεκριμένους discount factors, ενώ άλλοι απαιτούν διαφορετικούς discount factors για την επίτευξη της βέλτιστης απόδοσης. Φυσικά, ένας διαφορετικός discount factor μπορεί να σημαίνει διαφορετική βέλτιστη πολιτική. Για το λόγο αυτό μετράμε την απόδοση με βάση την μέση ανταμοιβή-επιβράβευση και όχι τη μειωμένη κατά τον discount factor.

Όπως έχουμε αναλύσει στο προηγούμενο κεφάλαιο, η συνάρτηση αξίας, V^*, που αντιστοιχεί στη βέλτιστη πολιτική π^*, ικανοποιεί τη συνάρτηση βελτίστου Bellman:

$$V^*(s) = \max_a \sum_{s\prime} T(s, a, s')\,(R(s, a, s') + \gamma V^*(s'))$$

Επιπλέον, η συνάρτηση μετάβασης T είναι συνήθως άγνωστη. Θα κάνουμε χρήση TD μάθησης για να ανανεώσουμε τις τιμές της V: $V_{t+1}(s_t) = V_t(s_t) + a_t\delta_t$, με TD-σφάλμα: $\delta_t = r_t + \gamma V_t(s_{t+1}) - V_t(s_t)$ και ρυθμό μάθησης: $0 \leq a_t \leq 1$. Έχει αποδειχθεί ότι όταν αυτές οι τιμές αποθηκεύονται σε έναν πίνακα, χρησιμοποιώντας την παραπάνω TD-ανανέωση, θα καταλήξουν σε σύγκλιση των τιμών στις εκτιμώμενες αποδώσεις, για μια σταθερή πολιτική. Η σύγκλιση στη βέλτιστη πολιτική έχει αποδειχθεί, υπό ορισμένες συνθήκες, και για παραλλαγές αυτής της ανανέωσης που χρησιμοποιούν Q-values αντί για τιμές κατάστασης (state values), όπως το Q-Learning και το SARSA. Αυτές οι Q-values εξαρτώνται από καταστάσεις και δράσεις αντί απλά από καταστάσεις.

3.2 Συνεχείς χώροι

3.2.1 Συνεχείς χώροι κατάστασης

Στην περίπτωση του συνεχούς χώρου κατάστασης γίνεται χρήση Προσεγγιστών Συναρτήσεων (Function Approximators – FAs) τους οποίους παραμετροποιούμε κατάλληλα για να αποθηκεύσουμε τις τιμές των καταστάσεων που έχουμε επισκεφθεί και να γενικεύσουμε αυτές που δεν γνωρίζουμε. Στη συνέχεια, η ανανέωση εφαρμόζεται στις παραμέτρους των FAs.

3.2.2 Συνεχείς χώροι δράσης

Ένα πιο δύσκολο πρόβλημα είναι η επέκταση της RL σε συνεχείς χώρους δράσης. Ακόμα κι αν έχουμε μια καλή προσέγγιση της συνάρτησης αξίας, εξακολουθούμε να έχουμε το πρόβλημα ότι δεν μπορούμε να βρούμε επιπόλαια τη δράση που αντιστοιχεί στην υψηλότερη αξία, σε μια δεδομένη κατάσταση. Επομένως, θα θέλαμε ένας αλγόριθμος να εξάγει γρήγορα μια προσέγγιση της βέλτιστης δράσης, δεδομένης μιας συγκεκριμένης κατάστασης. Για το λόγο αυτό θα χρησιμοποιήσουμε ξανά FAs. Στη συνέχεια, το ερώτημα είναι πώς μπορεί να βελτιωθεί αυτή η προσέγγιση της βέλτιστης δράσης.

Μια λύση είναι να διαχειριστούμε σωστά τις δυνατότητες εξερεύνησης που διαθέτουμε, καθώς είναι ο μόνος τρόπος με τον οποίο μπορούμε να ανακαλύψουμε νέες, καλύτερες πολιτικές. Οι πιο διαδεδομένες μέθοδοι εξερεύνησης είναι η ϵ-greedy και η Gaussian. Στην πρώτη μέθοδο, επιλέγεται με πιθανότητα ϵ μια τυχαία δράση, ενώ με πιθανότητα $1-\epsilon$ επιλέγεται η greedy, τρέχουσα προσέγγιση της βέλτιστης δράσης. Με τον τρόπο αυτό, μειώνοντας τον παράγοντα ϵ, μειώνεται αντίστοιχα και η εξερεύνηση.

Εναλλακτικά, η Gaussian εξερεύνηση γύρω από την τρέχουσα προσέγγιση της βέλτιστης δράσης επιτυγχάνεται ορίζοντας μια Gaussian κατανομή γύρω από την εν λόγω προσέγγιση, με μέση τιμή την τιμή της προσέγγισης, και δειγματοληπτώντας τυχαία την κατανομή. Το αποτέλεσμα της δειγματοληψίας είναι η δράση που θα εκτελέσει ο πράκτορας.

3.3 Αλγόριθμος ACLA

Ο Actor Critic Learning Automaton (ACLA) είναι η βάση του CACLA και μπορεί εύκολα να επεκταθεί σε συνεχείς χώρους. Ξεκινάμε τη μελέτη μας με την tabular περίπτωση του αλγορίθμου, με διακριτές καταστάσεις και δράσεις, και όχι τη συνεχή του εκδοχή. Στην περίπτωση αυτή, διαθέτουμε δύο πίνακες για αποθήκευση τιμών. Ο πρώτος χρησιμεύει στην αποθήκευση των καταστάσεων, οι οποίες ανανεώνονται με TD-μάθηση και ο δεύτερος αποθηκεύει τις πιθανότητες να επιλεγεί η κάθε δράση, για κάθε κατάσταση. Οι αξίες των καταστάσεων συγκλίνουν στις μελλοντικές discounted επιβραβεύσεις, δεδομένης της τρέχουσας πολιτικής. Αν η εκτέλεση μιας δράσης επιφέρει θετική αλλαγή στην αξία μιας κατάστασης, τότε η δράση αυτή μπορεί δυνητικά να οδηγήσει σε μια μεγαλύτερη discounted επιβράβευση, και συνεπώς σε μια καλύτερη πολιτική. Επομένως, ενισχύουμε αυτή τη δράση:

$$\text{IF} \quad \delta_t > 0: \quad \text{increase}_t(\pi_t(s_t, a_t))$$

Η εντολή $\text{increase}_t(\pi_t(s_t, a_t))$ αυξάνει την πιθανότητα να επιλεχθεί η δράση a_t στην κατάσταση s_t, ενώ ταυτόχρονα μειώνει όλες τις πιθανότητες να επιλεγεί οποιαδήποτε άλλη δράση στην κατάσταση s_t. Σε αντίθετη περίπτωση, όταν $\delta_t < 0$, δεν πραγματοποιούμε μείωση της παραπάνω πιθανότητας, καθώς η λογική της αρνητικής ανάδρασης δεν μπορεί να εφαρμοστεί στην επέκταση του αλγορίθμου στο συνεχή χώρο. Η διαφορά μεταξύ του ACLA και των μεθόδων Actor-Critic που

είχαμε στην ενότητα 2.8 είναι ότι ο ACLA επιτρέπει την ανανέωση του Actor ανάλογα με το πρόσημο του TD-σφάλματος, ενώ οι υπόλοιποι αλγόριθμοι AC με βάση την ακριβή τιμή του TD-σφάλματος. Τα πλεονεκτήματα του ACLA έναντι των AC μεθόδων είναι ποικίλα, με βασική την δυνατότητα του ACLA να επεκταθεί σε συνεχείς χώρους. Επιπλέον, η παράμετρος μάθησης του Actor στον CACLA είναι αμετάβλητη σε διαφορετικές κλιμακώσεις της συνάρτησης επιβράβευσης.

Αναφορικά σε στοχαστικές πολιτικές, ο αλγόριθμος ACLA μαθαίνει την πολιτική που βελτιστοποιεί την πιθανότητα να λάβουμε θετική τιμή δ_t. Συνεπώς, συμπεραίνουμε ότι με τον τρόπο αυτό είναι πολύ πιθανή, για τον ACLA, η σύγκλιση σε υπο-βέλτιστη πολιτική. Εντούτοις, αποδεικνύεται ότι σε ντετερμινιστικά περιβάλλοντα το πρόβλημα αυτό δεν προκύπτει.

Για να επεκτείνουμε τον ACLA σε συνεχείς χώρους δράσεων, αντικαθιστούμε απλά τους πίνακες της tabular εκδοχής, με FAs που έχουν ως έξοδό τους την αξία και τη δράση που πρέπει να εκτελεστεί, δεδομένης μιας κατάστασης. Πιο συγκεκριμένα, ο Actor έχει ως έξοδο κάτι παρόμοιο με τη δράση a_t στην κατάσταση s_t, εάν η αξία της κατάστασης, $V(s_t)$, έχει υποστεί αύξηση. Μόνο στην περίπτωση αυτή ανανεώνουμε τις τιμές του Actor. Ο λόγος για τον οποίο συμβαίνει αυτό είναι ότι στην περίπτωση αρνητικών ανανεώσεων, ο αλγόριθμος θα ανανεωθεί μακριά από την τελευταία δράση, την οποία εξέλαβε ως κακή. Εντούτοις, η λογική αυτή ισοδυναμεί με την περίπτωση ανανέωσης προς κάποια άγνωστη δράση, η οποία δεν είναι απαραίτητα καλύτερη από την παρούσα προσέγγιση της βέλτιστης δράσης. Επομένως, ενδέχεται να έχουμε ανακαλύψει ότι η δράση a_t δεν είναι τέλεια, αλλά ακόμα δεν γνωρίζουμε αν υπάρχουν καλύτερες εναλλακτικές.

3.4 Κανόνες ανανέωσης CACLA

Οι ανανεώσεις των Actor και Critic γίνονται όπως προαναφέραμε με τη χρήση FAs. Έστω θ^V το διάνυσμα παραμέτρων ενός FA του Critic. Ο νόμος ανανέωσης που χρησιμοποιούμε για τον Critic, βασίζεται στην TD-μάθηση και έχει τη μορφή:

$$\theta^V_{i,t+1} = \theta^V_{i,t} + a\delta_t \frac{\partial V_t(s_t)}{\partial \theta^V_{i,t}} \tag{1}$$

όπου $\theta^V_{i,t}$ είναι το i-οστό στοιχείο του διανύσματος θ^V τη χρονική στιγμή t και $V_t(s)$ είναι η έξοδος του FA τη χρονική στιγμή t, έχοντας ως είσοδο την κατάσταση s.

Αναφορικά στον Actor, θεωρούμε θ^{A_C} το διάνυσμα παραμέτρων του FA του Actor και $A_{C_t}(s_t)$ την έξοδο του FA τη χρονική στιγμή t. Έτσι, προκύπτει ο κανόνας ανανέωσης των παραμέτρων του Actor ως εξής:

$$IF\ \delta_t > 0\ :\quad \theta^{A_C}_{i,t+1} = \theta^{A_C}_{i,t} + a\left(a_t - A_{C_t}(s_t)\right)\frac{\partial A_{C_t}(s_t)}{\partial \theta^{A_C}_{i,t}} \tag{2}$$

Για να τονίσουμε τα αποτελέσματα των ενεργειών που βελτιώνουν την αξία περισσότερο από το συνηθισμένο, μπορούμε να επεκτείνουμε την ενημέρωση. Αρχικά, αποθηκεύεται ένας τρέχων μέσος όρος της διακύμανσης του TD- σφάλματος:

$$var_{t+1} = (1 - \beta)var_t + \beta\delta_t^2$$

Χρησιμοποιώντας τη διακύμανση, μπορούμε να προσδιορίσουμε εάν μια δράση ήταν εξαιρετικά καλή. Έτσι, συνδέουμε τον αριθμό των ανανεώσεων προς μια δράση, με τον αριθμό των τυπικών αποκλίσεων που η αξία-στόχος βρίσκεται πάνω από την παλιά της τιμή. Πλέον, ο αριθμός των ανανεώσεων θα είναι γραμμικά εξαρτώμενος από την ποσότητα: $\delta_t / stddev_t = \delta_t / \sqrt{var_t}$, ενώ, όπως και προηγουμένως, για μη-θετικές τιμές του TD-σφάλματος, δεν θα πραγματοποιούνται ανανεώσεις στον Actor. Η εν λόγω επέκταση του CACLA ονομάζεται CACLA + Var.

Η συμπεριφορά του CACLA σε σύγκριση με τον ACLA διαφέρει ελαφρώς, επειδή το CACLA δεν αλλάζει τις ανεξάρτητες πιθανότητες των δράσεων, αλλά αντιθέτως ενημερώνει προς μια δράση που έχει βρεθεί ότι είναι καλύτερη από την παρούσα προσέγγιση για τη βέλτιστη δράση.

Algorithm: CACLA

```
Def __init__():
    Initialize critic
    Initialize actor
end
Def __call__():
    Observe current state O0
    Critic prediction for observation
    Actor prediction for observation
    Exploration around output of actor
    Perform explored action
    Compute reward
    Observe new current state O1
    Critic prediction for new observation
    Compute TD-error δt
    Update Critic [1]
    if δt > 0 then
        Update actor [2]
    end
end
```

Πίνακας 3.1: Αλγόριθμος CACLA, απόδοση από [45], [46].

Κεφάλαιο 4

Grasping

Η σύλληψη ενός αντικειμένου (grasping) περιλαμβάνει διάφορες φάσεις ξεκινώντας από την ανίχνευση της θέσης του αντικειμένου, έως την επιλογή της διαμόρφωσης λαβής με τελικό στόχο τη διατήρηση του αντικειμένου σταθερό στη λαβή του ρομπότ [2]. Η διαδικασία της σύλληψης του αντικειμένου είναι απαιτητική, ειδικά όταν υπάρχουν περιορισμένες ή καθόλου διαθέσιμες προηγούμενες πληροφορίες σχετικά με το αντικείμενο, και επομένως είναι απαραίτητο να βασιστούμε περισσότερο στην αντίληψη του ρομπότ σε πραγματικό χρόνο [5]. Αυτό μπορεί να συμβεί με τη χρήση αισθητήρων αφής στο ρομπότ. Μέσω της απτικής επαφής, μπορεί να συλλέξει χρήσιμες πληροφορίες. Στο πλαίσιο αυτό, το ρομπότ θα πρέπει να εκτελεί σχετικές ενέργειες, π.χ., μια ελεγχόμενη χειροκίνητη εξερεύνηση της επιφάνειας του αντικειμένου. Ωστόσο, μια μεγάλη πρόκληση για να βασιστεί κανείς στην ενεργή αντίληψη του ρομπότ, σε πραγματικό χρόνο, είναι η υποκείμενη αβεβαιότητα της ρομποτικής ανίχνευσης και επιλογής της δράσης της [2].

Για να καταλήξουμε, όμως, σε μια καλή ρομποτική λαβή και στη βελτιστοποίησή της, καθώς επίσης στην τοποθέτηση των δακτύλων ενός ρομποτικού χεριού n – δαχτύλων, χρειαζόμαστε:

- το τρισδιάστατο σχήμα του αντικειμένου στόχου,
- τους κινηματικούς περιορισμούς των δακτύλων του ρομπότ και
- ένα σχέδιο εργασιών (task planning) που αποτελείται από το επιθυμητό προφίλ κίνησης/δύναμης/ροπής που θα εφαρμοστεί στο αντικείμενο χειρισμού

Έτσι, εστιάζουμε στη βελτιστοποίηση της εκ-των-προτέρων-διαμόρφωσης της λαβής και του αρχικού μηχανισμού σύλληψης ενός αντικειμένου από έναν επιδέξιο μηχανισμό ρομποτικού χεριού, με n-δάχτυλα ή m-DoF (βαθμούς ελευθερίας) ανά δάχτυλο. Άρα, ο σκοπός είναι η συλλογή διαφορετικών συνόλων δεδομένων και η γενίκευση μιας τεχνικής grasping για την πραγματοποίηση ισχυρής λαβής σε άγνωστα ή μερικώς άγνωστα αντικείμενα [5].

Οι προσεγγίσεις για το grasping μπορούν να χωριστούν σε αναλυτικές (analytical) και βασισμένες-στη-μάθηση (learning-based) μεθόδους. Οι αναλυτικές μέθοδοι κατασκευάζουν λαβές που ικανοποιούν ορισμένες ιδιότητες grasp, όπως η κλειστότητα ως προς τη δύναμη [32], η συμβατότητα εργασιών και άλλα μέτρα σταθερότητας της λαβής [3]. Επιπλέον, οι αναλυτικές προσεγγίσεις δημιουργούν graspings που βασίζονται σε γεωμετρικά, κινηματικά ή/και δυναμικά μοντέλα αντικειμένων. Τέλος, οι περισσότερες αναλυτικές προσεγγίσεις δημιουργούν λαβές που ικανοποιούν ορισμένες ιδιότητες, χωρίς, όμως, να λαμβάνεται υπόψη ο τύπος λαβής.

Από την άλλη, οι learning-based μέθοδοι χρησιμοποιούν δεδομένα σύλληψης με ετικέτα (labelled data) για την εκπαίδευση ταξινομητών (classifiers), με στόχο να προβλέψουν την επιτυχία του grasping ή να προβλέψουν το επιθυμητό grasp, μέσω regression analysis. Σε σύγκριση με τις αναλυτικές προσεγγίσεις, οι learning-based μέθοδοι τείνουν να γενικεύονται καλά σε άγνωστα αντικείμενα, διαθέτοντας παρά μόνο λίγες πληροφορίες για αυτά. [3]

Με στόχο τη διαχείριση της αβεβαιότητας που προαναφέραμε σχετικά με την αίσθηση του ρομπότ και την επιλογή της δράσης του, εμπλουτίζουμε τις οπτικές πληροφορίες με την απτική διαδικασία εξερεύνησης, η οποία καθοδηγείται από ένα πιθανολογικό μοντέλο, όπως η Βελτιστοποίηση Bayes (Bayes Optimisation). Το ρομπότ εντοπίζει πρώτα τη θέση του αντικειμένου χρησιμοποιώντας δεδομένα από ένα νέφος σημείων (point cloud), τα οποία παράγονται από έναν αισθητήρα RGB-D. Στη συνέχεια, ξεκινά μια διαδικασία εξερεύνησης κατά την οποία το ρομποτικό χέρι αξιολογεί διαφορετικές διαμορφώσεις της λαβής (grasp configurations) που επιλέγονται από την Bayesian Optimization, με βάση μια μετρική grasping που υπολογίζεται από την απτική αίσθηση. Τέλος, αφού βρεθεί το καλύτερο grasp configuration, πραγματοποιείται η λήψη του αντικειμένου [2].

4.1 Είδη μάθησης

4.1.1 Ενισχυτική μάθηση (RL)

Σε αντίθεση με την επιβλεπόμενη μάθηση, οι μέθοδοι trial and error (δοκιμών και σφαλμάτων) του RL επιτρέπουν στα ρομπότ να μάθουν την ικανότητα αυτοεξερεύνησης, γεγονός που τα κάνει να έχουν μεγαλύτερη επιδεξιότητα. Ο λόγος που η ενισχυτική μάθηση κάνει τα ρομπότ πιο επιδέξια σε σχέση με την επιβλεπόμενη μάθηση είναι επειδή οι μέθοδοι εκπαίδευσης που ακολουθούν είναι εντελώς διαφορετικές. Η επιβλεπόμενη μάθηση, όπως έχουμε αναλύσει, είναι η ενημέρωση των παραμέτρων του μοντέλου μέσω ζευγών δειγμάτων και ετικετών μέχρις ότου να ελαχιστοποιηθεί η συνάρτηση απώλειας (loss function). Όμως, το μεγαλύτερο μειονέκτημα αυτής της μεθόδου εκπαίδευσης είναι ότι το σύνολο δεδομένων της έχει συνήθως μόνο μία ετικέτα ανά δείγμα. Δηλαδή, ένα σημείο λήψης (grasp point) μπορεί να αντιστοιχεί μόνο σε μία στάση λήψης (grasp pose), κάτι που στην πραγματικότητα προσθέτει πολλούς περιορισμούς στο μοντέλο. Αντίθετα, η ενισχυτική μάθηση σχετίζεται με την εύρεση της πολιτικής του grasping αντικειμένων, μέσω trial and error. Δεδομένου ότι το ρομπότ μπορεί να έχει εξερευνήσει πολλές δυνατότητες grasping κατά τη διάρκεια της εκπαίδευσής του και η τιμή της συνάρτησης ανταμοιβής του κάθε grasp pose δεν είναι κακή, τότε ο αλγόριθμος θα προσθέσει αυτές τις δυνατότητες στη συνάρτηση πολιτικής, έτσι ώστε οι πιθανές επιλογές λήψης να είναι περισσότερες. Εντούτοις, οι μέθοδοι που βασίζονται στην ενισχυτική μάθηση, ειδικά όταν χρησιμοποιούνται σε σύνθετες ρομποτικές εργασίες, έχουν σοβαρό πρόβλημα, καθώς ο χώρος εξερεύνησης του αλγορίθμου γίνεται εξαιρετικά μεγάλος ή οι grasp poses που απαιτούν trial and error είναι αμέτρητες [1].

4.1.2 Αυτο-επιβλεπόμενη μάθηση

Η αυτο-επιβλεπόμενη μάθηση (Self-supervised learning - SSL) είναι μια νέα μορφή μηχανικής μάθησης. Με τη χρήση της, αποφεύγουμε το μεγάλο κόστος συλλογής και επισήμανσης δεδομένων (labelling data), καθώς η εκπαίδευση γίνεται μέσω unlabelled δεδομένων. Μάλιστα, το

SSL – ως υποσύνολο της μάθησης χωρίς επίβλεψη – αρχικά μελετά ψευδο-ετικέτες που αντιπροσωπεύουν τα χαρακτηριστικά που εξάγονται από τα παρεχόμενα unlabelled δεδομένα. Στη μάθηση ρομποτικών εργασιών, η συλλογή συνόλων δεδομένων και ο χειροκίνητος σχολιασμός κάθε δείγματος είναι δαπανηρές διαδικασίες. Επιπλέον, το labelling κάθε δείγματος μπορεί να αποτελεί πρόκληση, ενώ το λάθος labelling μπορεί να δυσκολέψει την εκπαίδευση του μοντέλου. Για το λόγο αυτό, το SSL αναθέτει ετικέτες, οι οποίες βασίζονται στα χαρακτηριστικά του dataset, κατά τη διάρκεια της μελέτης, για να αποφευχθεί η ανακρίβεια της ανθρώπινης υποκειμενικότητας ή σφάλματος [1].

4.1.3 Βαθιά μάθηση

Από την άλλη, η βαθιά μάθηση (deep learning) απαιτεί μεγάλο όγκο δεδομένων εκπαίδευσης. Ωστόσο, είναι δύσκολο να συλλεχθεί όλος αυτός ο όγκος δεδομένων, ειδικά για την εκπαίδευση grasping-πολλών-δαχτύλων [4]. Εμπνευσμένοι από το Deep Q-learning network (DQN), στις μελέτες [39], [40], εκπαιδεύουν ένα CNN (Convolutional NN) ώστε να μάθει την Q-συνάρτηση και χρησιμοποιούν τη μέθοδο gradient Monte Carlo ως κανόνα ανανέωσης. Σε κάθε βήμα, δημιουργούνται αρκετά υποψηφία grasps, βάσει ιεραρχικής δειγματοληψίας και, στη συνέχεια, επιλέγεται ένα grasp pose σύμφωνα με την πολιτική που έχει μάθει το σύστημα. Επιπλέον, στη μελέτη [21], χρησιμοποιείται ένα δίκτυο που προτείνει grasp poses, βασισμένο στο PointNet++, το οποίο αντιστοιχεί σε κάθε σημείο του point cloud μια διαμόρφωση grasp και τη βαθμολογία της ποιότητάς του. Τέλος, οι μέθοδοι non-maximum suppression (NMS) και σταθμισμένης τυχαίας δειγματοληψίας (weighted random sampling) εφαρμόζονται στην έξοδο του παραπάνω δικτύου για την επιλογή του grasp που θα εκτελεστεί.

4.1.4 Ενεργητική μάθηση

Η ενεργητική μάθηση (active learning) αποτελεί υποσύνολο της μηχανικής μάθησης στο οποίο ο αλγόριθμος μάθησης μπορεί να ζητήσει άμεσα από το χρήστη να τοποθετήσει ετικέτες στα δεδομένα με τις επιθυμητές εξόδους. Η μέθοδος αυτή είναι πιο κατάλληλη σε περιπτώσεις που τα unlabelled δείγματα δεδομένων είναι πολλά, όταν απαιτούνται πολλά labelled δεδομένα για την εκπαίδευση ενός ακριβούς συστήματος επιβλεπόμενης μάθησης και όταν τα δείγματα δεδομένων μπορούν εύκολα να συλλεχθούν ή να συντεθούν [4].

Στη μελέτη [4], προτείνεται μια νέα προσέγγιση ενεργητικής μάθησης για να αντιμετωπιστεί το πρόβλημα της συλλογής δεδομένων για το grasp learning από χέρια με πολλά δάχτυλα. Η προσέγγισή της χρησιμοποιεί λιγότερα δείγματα εκπαίδευσης για να παράγει ποσοστά-επιτυχίας του grasping τα οποία να είναι συγκρίσιμα με τη μέθοδο της παθητικής επιβλεπόμενης μάθησης.

4.2 Συμπλήρωμα σχήματος αντικειμένου

Η δημιουργία του point cloud εξαρτάται από τις συνθήκες φωτός στη σκηνή του πειράματος. Μερικές φορές το point cloud ορισμένων αντικειμένων είναι πολύ αραιό λόγω κακού φωτισμού ή επειδή η σκηνή είναι πολύ ακατάστατη. Σε αυτήν την περίπτωση, δεν συνιστάται η άμεση δειγματοληψία από αυτό. Το συμπλήρωμα σχήματος αντικειμένου γίνεται με σκοπό να αποκαταστήσουμε το αρχικό σχήμα του αντικειμένου, να βελτιώσουμε τις πληροφορίες που διαθέτουμε για το αντικείμενο και να βοηθήσουμε, εν τέλει, στη δημιουργία υποψήφιου grasp pose [1].

Η δειγματοληψία υποψήφιου grasp με βάση το object affordance είναι η μείωση του χώρου δειγματοληψίας στην περιοχή όπου το grasp είναι πιο πιθανό να πετύχει. Γενικά, η συμπλήρωση του σχήματος του αντικειμένου μπορεί να χρησιμοποιηθεί ως λύση στην περίπτωση που διαθέτουμε ένα point cloud κατώτερης ποιότητας. Για απλά, κανονικά αντικείμενα, αυτή η μέθοδος μπορεί να μην έχει πολύ νόημα. Ωστόσο, είναι ιδιαίτερα χρήσιμη για αντικείμενα που έχουν μεγάλη απόκλιση από το κέντρο βάρους τους και μπορούν να προκαλέσουν ζημιά στο τελικό στοιχείο δράσης του ρομπότ. Για τα συμμετρικά αντικείμενα, η επίδραση του shape completion μπορεί να μην είναι τόσο εμφανής, αλλά για τα ασύμμετρα αντικείμενα, αυτός ο τύπος μεθόδων είναι ιδιαίτερα ευνοϊκός [1].

Αν και οι υπάρχουσες μέθοδοι συμπληρώματος σχήματος συνήθως συνοδεύονται από υψηλή αβεβαιότητα, η επίγνωση του σχήματος του αντικειμένου είναι ικανή να διευκολύνει απίστευτα την ακρίβεια και την ευρωστία της δημιουργίας προτεινόμενων grasp poses [1].

4.3 Μορφές sampling

4.3.1 Geometry-based sampling

Οι geometry-based μέθοδοι δειγματοληψίας δημιουργούν υποψήφια grasp poses αλλάζοντας τυχαία, σε μια συγκεκριμένη εργασία (task), τη διαμόρφωση του grasp κάτω από φυσικούς και γεωμετρικούς περιορισμούς. Ωστόσο, η δειγματοληψία μερικών προτεινόμενων grasp poses είναι υπολογιστικά ακριβή. Η γενίκευση που μπορεί να επιτευχθεί με την geometry-based δειγματοληψία δεν είναι τόσο ισχυρή, ειδικά σε ακατάστατα περιβάλλοντα, καθώς συχνά αντιλαμβάνεται δύο αντικείμενα που βρίσκονται πολύ κοντά, σαν ένα, γεγονός που οδηγεί σε παράλογα αποτελέσματα της δειγματοληψίας με λαβή [1].

4.3.2 Learning-based sampling

Η learning-based μέθοδος δειγματοληψίας ολοκληρώνει τη μάθηση βασισμένη σε ένα σύνολο δεδομένων κατά τη διάρκεια της εκπαίδευσης. Οι learning-based μέθοδοι μπορούν να χωριστούν σε μάθηση με επίβλεψη ή μάθηση χωρίς επίβλεψη, σύμφωνα με το μοντέλο μάθησης. Για μεθόδους δειγματοληψίας που βασίζονται στην επιβλεπόμενη μάθηση, λαμβάνονται scene point clouds ή εικόνες RGB ως είσοδοι, για να εξαχθούν αποτελέσματα δειγματοληψίας με άμεσο τρόπο ή να εξαχθούν πρώτα τα κατάλληλα grasp points και στη συνέχεια να τοποθετηθεί η grasp posture στα σημεία αυτά. Η learning-based μέθοδος δειγματοληψίας μπορεί να βελτιώσει σημαντικά τη γενίκευση της δειγματοληψίας, ειδικά σε ακατάστατα περιβάλλοντα, μέσω των πλεονεκτημάτων των πιο συχνά εξαγόμενων χαρακτηριστικών, τα οποία χρησιμοποιούνται για εκπαίδευση στην επιβλεπόμενη μάθηση [1].

4.3.3 Learning-based object-aware sampling

Οι learning-based object-aware μέθοδοι δειγματοληψίας έχουν σημαντική βελτίωση στο ποσοστό επιτυχίας του τελικού grasp σε ακατάστατα περιβάλλοντα. Τα learning-based object-aware μοντέλα έχουν τις υψηλότερες δυνατότητες να πραγματοποιήσουν τα προτεινόμενα grasp. Ωστόσο, είναι συνήθως δύσκολο για το μοντέλο να προβλέψει το σημείο επαφής κάθε δακτύλου, στην περίπτωση τελικού στοιχείου δράσης πολλαπλών δακτύλων (με εξαίρεση τα suction cups και τους parallel-jaw grippers), γεγονός που συνήθως δυσκολεύει την απόδοση του μοντέλου σε ακατάστατα περιβάλλοντα [1].

Η learning-based μέθοδος, για παραγωγή υποψήφιων grasping, βασίζεται σε ένα εκπαιδευμένο μοντέλο, συνήθως σε ένα νευρωνικό δίκτυο, το οποίο λαμβάνει ως είσοδο το point cloud και ανάλογα με την πρόβλεψη της εξόδου του δικτύου, λαμβάνει το αποτέλεσμα της δειγματοληψίας του grasp pose. Αν και αυτή η μέθοδος είναι λιγότερο διαισθητική από την geometry-based, το νευρωνικό δίκτυο μπορεί να εξάγει πλουσιότερα χαρακτηριστικά στο κρυφό του στρώμα (hidden layer) για να βοηθήσει στη δειγματοληψία, μειώνοντας έτσι την υπολογιστική δυσκολία ανίχνευσης πιθανής σύγκρουσης με άλλα αντικείμενα [1].

Το learning-based grasp planning έχει γίνει δημοφιλές την τελευταία δεκαετία, λόγω της ικανότητάς του να γενικεύει καλά σε νέα αντικείμενα, όταν διατίθενται ελλιπείς πληροφορίες για αυτά. Εντούτοις, συχνά απαιτούνται πολλά δεδομένα για την εκπαίδευση ενός ισχυρού μοντέλου. Η συλλογή δεδομένων, όπως και η κατασκευή ενός συνόλου δεδομένων είναι δαπανηρή, γεγονός που καθιστά χρονοβόρα τη διαδικασία προετοιμασίας της μεθόδου [1].

4.3.4 Geometry-based object-aware sampling

Οι object-agnostic μέθοδοι λαμβάνουν το point cloud ως είσοδο και δημιουργούν τις προτάσεις των grasp από μοντέλα που έχουν μάθει, χωρίς να ανιχνεύουν το αντικείμενο στο point cloud. Η εν λόγω μέθοδος δειγματοληψίας δεν συνιστάται ιδιαίτερα. Η object-agnostic δειγματοληψία προτάθηκε στην εποχή που οι μέθοδοι οπτικής ανίχνευσης δεν ήταν αποτελεσματικές, ενώ η

έλλειψη πληροφοριών για τα αντικείμενα έχει σημαντική επίδραση στη δημιουργία αξιόπιστων grasp poses [1].

Από την άλλη, η geometry-based object-aware δειγματοληψία χρησιμοποιεί πρώτα τεχνικές όρασης υπολογιστών για τον εντοπισμό του αντικειμένου και, στη συνέχεια, δειγματοληπτεί τα υποψηφία grasp poses με βάση τον μειωμένο χώρο αναζήτησης (reduced searching space). Αυτή η μέθοδος έχει δείξει ότι η αξιοπιστία και η λογική των υποψηφίων grasp poses που δημιουργούνται έχουν βελτιωθεί αισθητά, ωστόσο, η υιοθέτηση περιορισμών που προσθέτει χειροκίνητα ο άνθρωπος στο βήμα δειγματοληψίας μπορεί να προκαλέσει τη δημιουργία ορισμένων ασταθών grasp poses και είναι υπολογιστικά δαπανηρή [1].

4.4 Function Optimization

Η Function Optimization (FO) ή Global Function Optimization (GFO) σχετίζεται με τον προσδιορισμό του ελαχίστου ή του μεγίστου μιας αντικειμενικής συνάρτησης (objective function). Δείγματα συλλέγονται από κάποιον τομέα και αξιολογούνται από την αντικειμενική συνάρτηση, η οποία επιστρέφει μια τιμή – κόστος για τα εν λόγω δείγματα. Ένα δείγμα ορίζεται ως ένα διάνυσμα μεταβλητών με προκαθορισμένο εύρος n-διαστάσεων. Στον χώρο αυτό πρέπει να γίνει δειγματοληψία και διερεύνηση προκειμένου να βρεθεί ο συνδυασμός μεταβλητών που θα έχουν ως αποτέλεσμα το καλύτερο κόστος. Η αντικειμενική συνάρτηση είναι συχνά εύκολο να προσδιοριστεί, αλλά μπορεί να είναι υπολογιστικά κοστοβόρος ο υπολογισμός της, ενώ μπορεί να οδηγήσει σε έναν θορυβώδη υπολογισμό του κόστους με την πάροδο του χρόνου. Η μορφή της αντικειμενικής συνάρτησης είναι άγνωστη και συχνά είναι σε μεγάλο βαθμό μη-γραμμική και πολυδιάστατη, καθώς ορίζεται από τον αριθμό των μεταβλητών εισόδου. Επιπλέον, η αντικειμενική συνάρτηση μερικές φορές καλείται «χρησμός» (oracle), δεδομένου ότι έχει την ικανότητα να δίνει μόνο απαντήσεις. Η FO είναι θεμελιώδης για τη μηχανική μάθηση. Οι περισσότεροι αλγόριθμοι μηχανικής μάθησης περιλαμβάνουν βελτιστοποίηση παραμέτρων (βαρών, συντελεστών κ.λπ.). Η βελτιστοποίηση αναφέρεται επίσης στη διαδικασία εύρεσης του καλύτερου συνόλου υπερ-παραμέτρων που διαμορφώνουν την εκπαίδευση ενός αλγορίθμου μηχανικής μάθησης [38].

4.4.1 Bayesian Optimization

Θεωρούμε τον αλγόριθμο Bayesian Optimization (BO) ως ένα από τα πιθανοτικά μοντέλα για την εξερεύνηση του ολικού βέλτιστου [30]. Ο κύριος στόχος είναι να βρεθεί μια καθολική μέθοδος βελτιστοποίησης, η οποία εστιάζει στην εύρεση της ελάχιστης-βέλτιστης τιμής για την αντικειμενική συνάρτηση $f: \mathbb{X} \to \mathbb{R}$, όπου $\mathbb{X}$ είναι ένας συμπαγής χώρος.

Η Bayesian Optimization είναι μια προσέγγιση που χρησιμοποιεί το θεώρημα πιθανοτήτων του Bayes για να κατευθύνει την αναζήτηση, προκειμένου να βρεθεί το ελάχιστο ή το μέγιστο μιας

αντικειμενικής συνάρτησης. Συγκεκριμένα, αποτελεί μια βελτιστοποίηση η οποία είναι χρησιμότερη στην περίπτωση αντικειμενικών συναρτήσεων που είναι σύνθετες ή/και θορυβώδεις. Όπως γνωρίζουμε, το θεώρημα του Bayes αποτελεί μια προσέγγιση για τον υπολογισμό της δεσμευμένης πιθανότητας:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Εναλλακτικά, μπορούμε να κανονικοποιήσουμε την τιμή του παρονομαστή και να απλοποιήσουμε την παραπάνω ιδιότητα: $P(A|B) \propto P(B|A)P(A)$. Αυτό είναι χρήσιμο δεδομένου ότι δεν μας ενδιαφέρει ο ακριβής υπολογισμός της δεσμευμένης πιθανότητας, αλλά η βελτιστοποίηση μιας ποσότητας. Η δεσμευμένη πιθανότητα που υπολογίζουμε, $P(A|B)$, καλείται μεταγενέστερη πιθανότητα (posterior probability), η αντίστροφη δεσμευμένη πιθανότητα, $P(B|A)$, αναφέρεται ως πιθανοφάνεια (likelihood) και η οριακή πιθανότητα, $P(A)$, αναφέρεται ως προηγούμενη πιθανότητα (prior probability). Επίσης, ορίζουμε ως Acquisition Function την τεχνική με την οποία χρησιμοποιείται το posterior για την επιλογή του επόμενου δείγματος από τον χώρο αναζήτησης [36].

Μπορούμε να επινοήσουμε συγκεκριμένα δείγματα $x = \{x_{1:n}\}$ και να τα αξιολογήσουμε χρησιμοποιώντας την αντικειμενική συνάρτηση $f(x_i)$ που επιστρέφει το κόστος ή το αποτέλεσμα για το δείγμα x_i, $z = \{z_{1:n}\}$. Τα δείγματα και το αποτέλεσμά τους συλλέγονται διαδοχικά και ορίζουν τα δεδομένα μας D, π.χ. $D_{1:n} = \{x_{1:n}, f(x_{1:n})\}$ και χρησιμοποιούνται για τον ορισμό του prior. Η συνάρτηση πιθανοφάνειας ορίζεται ως η πιθανότητα παρατήρησης των δεδομένων, με τη συνάρτηση $P(D_{1:n}|f)$. Αυτή η συνάρτηση πιθανοφάνειας θα αλλάξει καθώς συλλέγονται περισσότερες παρατηρήσεις. Για να ορίσουμε την posterior κατανομή: : $P(f|D_{1:n}) \propto P(D_{1:n}|f)P(f)$ [36]. Το posterior αντιπροσωπεύει όλα όσα γνωρίζουμε για την αντικειμενική συνάρτηση. Είναι μια προσέγγιση της αντικειμενικής συνάρτησης και μπορεί να χρησιμοποιηθεί για την εκτίμηση του κόστους διαφορετικών υποψηφίων δειγμάτων που μπορεί να θέλουμε να αξιολογήσουμε [37].

Μόλις συλλεχθούν πρόσθετα δείγματα και γίνει η αξιολόγησή τους μέσω της αντικειμενικής συνάρτησης $f(\cdot)$, προστίθενται στα δεδομένα D και, στη συνέχεια, ενημερώνεται το posterior, δηλαδή την προσέγγιση της αντικειμενικής συνάρτησης. Αυτή η διαδικασία επαναλαμβάνεται μέχρι να εντοπιστεί κάποιο άκρο της αντικειμενικής συνάρτησης (ελάχιστο ή μέγιστο), να εντοπιστεί ένα αρκετά καλό αποτέλεσμα ή να εξαντληθούν οι πόροι [36].

Αναφορικά στο grasping, ο αλγόριθμος BO λειτουργεί με στόχο την επιλογή των καλύτερων σημείων λήψης ενός αντικειμένου, καθώς σε κάθε επανάληψη προσανατολίζεται προς το ελάχιστο $|z^* - z_n|$. Εξετάζουμε αυτή τη διαδικασία σε δύο βασικά βήματα: Πρώτον, για κάθε είσοδο σημείου λήψης ενός αντικειμένου, δημιουργείται ένα πιθανό μοντέλο (στην περίπτωσή μας, μια Gaussian διαδικασία). Δεύτερον, χρησιμοποιώντας μια acquisition function a, επιλέγεται το μοντέλο με βάση το οποίο θα προσδιοριστεί το επόμενο σημείο για εξερεύνηση. Επομένως, η BO βοηθά στη βελτιστοποίηση του αριθμού των βημάτων που απαιτούνται για ένα ασφαλές grasping [2].

Από την άλλη, η Unscented Bayesian Optimization, που θα μελετήσουμε στη συνέχεια, μπορεί να βρει ασφαλέστερες λαβές αντικειμένων, λαμβάνοντας υπόψη την αβεβαιότητα στη ρομποτική ανίχνευση και στην εκτέλεση της λαβής [2].

4.4.2 Unscented Bayesian Optimization

Η Unscented Bayesian Optimization (UBO) είναι μια μέθοδος διάδοσης της μέσης τιμής και της συνδιακύμανσης μέσω ενός μη γραμμικού μετασχηματισμού [2]. Η βάση του αλγορίθμου είναι η καλύτερη διαχείριση μιας προσεγγιστικής κατανομής πιθανοτήτων αντί της προσέγγισης μιας αυθαίρετης μη-γραμμικής συνάρτησης (της αντικειμενικής συνάρτησης).

Ο unscented μετασχηματισμός χρησιμοποιεί ένα σύνολο ντετερμινιστικά επιλεγμένων δειγμάτων από την αρχική κατανομή (που ονομάζονται σημεία σίγμα) και τα μετασχηματίζει μέσω της μη γραμμικής συνάρτησης $f(\cdot)$. Στη συνέχεια, η μετασχηματισμένη κατανομή υπολογίζεται με βάση τον σταθμισμένο συνδυασμό των μετασχηματισμένων σημείων σίγμα. Το πλεονέκτημα του unscented μετασχηματισμού είναι ότι ο μέσος όρος και οι εκτιμήσεις συνδιακύμανσης της νέας κατανομής είναι σε συμφωνία με την τρίτη τάξη της σειράς Taylor της $f(\cdot)$, υπό την προϋπόθεση ότι η αρχική κατανομή είναι Gaussian prior ή μέχρι και επέκταση δεύτερης τάξης για οποιοδήποτε άλλο prior [31].

Το πλεονέκτημα της UBO έναντι της κλασικής BO είναι η ικανότητά της να εξετάζει την αβεβαιότητα στον χώρο εισόδου για να βρει τη βέλτιστη λαβή. Αυτό συμβαίνει επειδή στην UBO λαμβάνεται υπόψη ο θόρυβος εισόδου κατά τη διαδικασία λήψης αποφάσεων, για την εξερεύνηση και την επιλογή των περιοχών που είναι ασφαλείς. Επίσης, για τη διάσταση d, απαιτούνται $2d + 1$ σημεία σίγμα, που δείχνουν ότι το υπολογιστικό της κόστος είναι αμελητέο σε σύγκριση με άλλες μεθόδους, όπως η μέθοδος Monte Carlo ή η συνάρτηση Gauss, οι οποίες απαιτούν περισσότερα δείγματα [2], [31].

4.5 Αισθητήρες αφής

Η χρήση αισθητήρων αφής (Tactile sensors) στους ρομποτικούς χειριστές είναι σε θέση να αντισταθμίσει ορισμένα από τα προβλήματα της προσέγγισης των αντικειμένων μόνο με χρήση της όρασης υπολογιστών. Πράγματι, η δυνατότητα αντίληψης της αφής επιτρέπει στο ρομπότ να κατανοήσει πότε έχει επιτευχθεί η επαφή με το εκάστοτε αντικείμενο. Επιπλέον, του επιτρέπει να έχει καλύτερη αντίληψη των περιοχών του αντικειμένου με τις οποίες δεν μπορεί να έρθει σε επαφή, καθώς το αντικείμενο άπτεται σε άλλες επιφάνειες. Αυτό επιτυγχάνεται όταν το ίδιο το ρομπότ έρχεται σε επαφή με αυτές τις επιφάνειες [2]. Η σημαντική συνεισφορά των απτικών αισθητήρων στην πραγματοποίηση επιδέξιων ρομποτικών λαβών είναι εμφανής στο ότι στη βιβλιογραφία προτείνονται τεχνικές για τον έλεγχο της ολίσθησης και της σταθεροποίησης της λαβής των αντικειμένων χρησιμοποιώντας μόνο αισθητήρες αφής [24], [25].

Η απτική εξερεύνηση συνιστάται πρώτον από το κλείσιμο του ρομποτικού χεριού σε πολλά σημεία ενός αντικειμένου και δεύτερον από την αξιολόγηση της μετρικής του κάθε grasp. Στη συνέχεια, κατά τον υπολογισμό της μετρικής, και τη λήψη του αντικειμένου, δημιουργείται ένα διάνυσμα δύναμης για να επιτευχθεί η λαβή. Αυτή η δύναμη υπολογίζεται από τις συντεταγμένες της άκρης του δακτύλου έως το εικονικό πλαίσιο που βρίσκεται στη μέση των δακτύλων και του αντίχειρα [2].

4.6 Ανίχνευση αντικειμένου

Ο Random Sample Consensus (RANSAC), είναι ένας μη ντετερμινιστικός επαναληπτικός αλγόριθμος για την ανίχνευση αντικειμένων (Object Detection). Προσπαθεί να προσαρμόσει τα σημεία από το νέφος σημείων (point cloud) σε ένα μαθηματικό μοντέλο ενός κυρίαρχου επιπέδου. Στη συνέχεια, το RANSAC προσδιορίζει τα σημεία που δεν αποτελούν το μοντέλο κυρίαρχου επιπέδου. Αυτά τα σημεία που δεν ταιριάζουν στο επίπεδο μοντέλο (που ονομάζονται ακραία σημεία) συγκεντρώνονται για να σχηματίσουν ένα αντικείμενο. Επίσης, έχει οριστεί ένα ελάχιστο όριο για την αποφυγή ανίχνευσης μικροσκοπικών αντικειμένων και φιλτραρίσματος επιπλέον θορύβου [2], [25].

Από την άλλη ο αλγόριθμος GraspCNN αντιμετωπίζει την εκτίμηση του grasp ως πρόβλημα ανίχνευσης αντικειμένου. Πιο συγκεκριμένα, λαμβάνει εικόνες RGB ως είσοδο και προτείνει έναν κύκλο προσανατολισμένης διαμέτρου. Ο κύκλος και η προσανατολισμένη διάμετρος υποδεικνύουν την περιοχή λαβής και το άνοιγμα και τον προσανατολισμό κλεισίματός της, αντίστοιχα. Ο κύκλος αυτός που λαμβάνουμε από την RGB εικόνα, υπολογίζεται για προβολή στο point cloud [22].

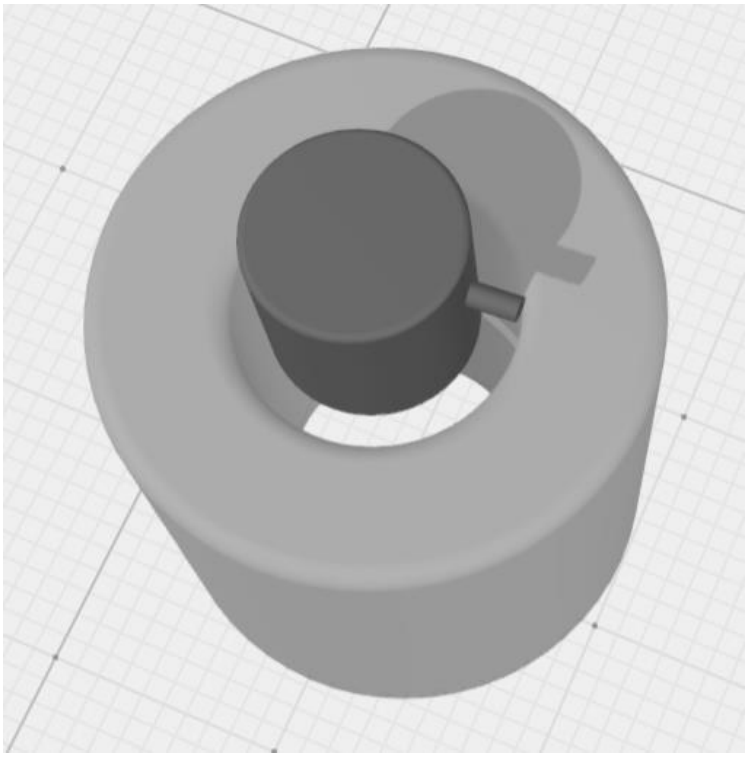
Κεφάλαιο 5

Περιγραφή προσομοίωσης και πειράματος

Στο τρέχον κεφάλαιο θα παρουσιάσουμε το σύστημα που χρησιμοποιήσαμε για να επιτευχθεί ο στόχος της εργασίας. Ο σκοπός μας είναι η εκμάθηση μιας ρομποτικής κίνησης με χρήση αλγορίθμου ενισχυτικής μάθησης και ανάδραση δύναμης. Η ανάδραση δύναμης αποτελεί ανάλογο των απτικών αισθητήρων στο πρόβλημα της επίτευξης επιδέξιας ρομποτικής λαβής για τη λήψη (grasp) ενός αντικειμένου (μανιταριού). Πιο συγκεκριμένα, στη μελέτη μας διαθέτουμε ένα ρομποτικό χέρι το οποίο πραγματοποιεί επιδέξιες κινήσεις με σκοπό να αποκολλήσει ένα μανιτάρι από τη βάση του. Συγκεκριμένα, οι κινήσεις που μπορεί να επιτελέσει είναι περιστροφής και ανύψωσης. Η εκμάθηση της κίνησης γίνεται χωρίς την επίγνωση της δυναμικής του περιβάλλοντος (model-free learning), με μοναδική γνώση των δυνάμεων που αντιτίθενται στην επίτευξη της κίνησης, μέσω των αισθητήρων αφής των άκρων των δαχτύλων του ρομπότ. Στη συνέχεια περιγράφουμε τη μοντελοποίηση του προβλήματος της λήψης μανιταριών από ένα ρομποτικό χέρι. Τέλος, περιγράφουμε τη δομή του συστήματος και του περιβάλλοντος μάθησης που χρησιμοποιήσαμε.

5.1 Περιγραφή πειραματικής διάταξης

Έστω ότι διαθέτουμε έναν παραμετροποιήσιμο διάδρομο, πεπερασμένου μήκους, με συμπαγείς τοίχους, με κλειστό αριστερό άκρο και ανοιχτό δεξί τέτοιο ώστε να εφάπτεται στο εσωτερικό της περιφέρειας ενός κυλινδρικού φλοιού, δημιουργώντας έναν εσωτερικό διάδρομο – εσοχή με ανιούσα πορεία. Έστω ένα συμπαγές κυλινδρικό αντικείμενο – το οποίο προσομοιάζει το μανιτάρι – το οποίο εφαρμόζει καλά στο εσωτερικό του κυλινδρικού φλοιού, και διαθέτει μια προεξοχή μικρού μήκους – συγκρίσιμου με το πλάτος του διαδρόμου – όπως φαίνεται στο Σχ. 5.1.1. Ο στόχος της εκπαίδευσης είναι να μπορέσει η ρομποτική λαβή, συγκρατώντας με in-hand τρόπο τον συμπαγή κύλινδρο, να μάθει να απεγκλωβίζει τον μικρό από τον μεγαλύτερο κύλινδρο, καθώς η προεξοχή του μικρού κυλίνδρου βγαίνει από τον εσωτερικά χαραγμένο διάδρομο του μεγάλου κυλίνδρου.

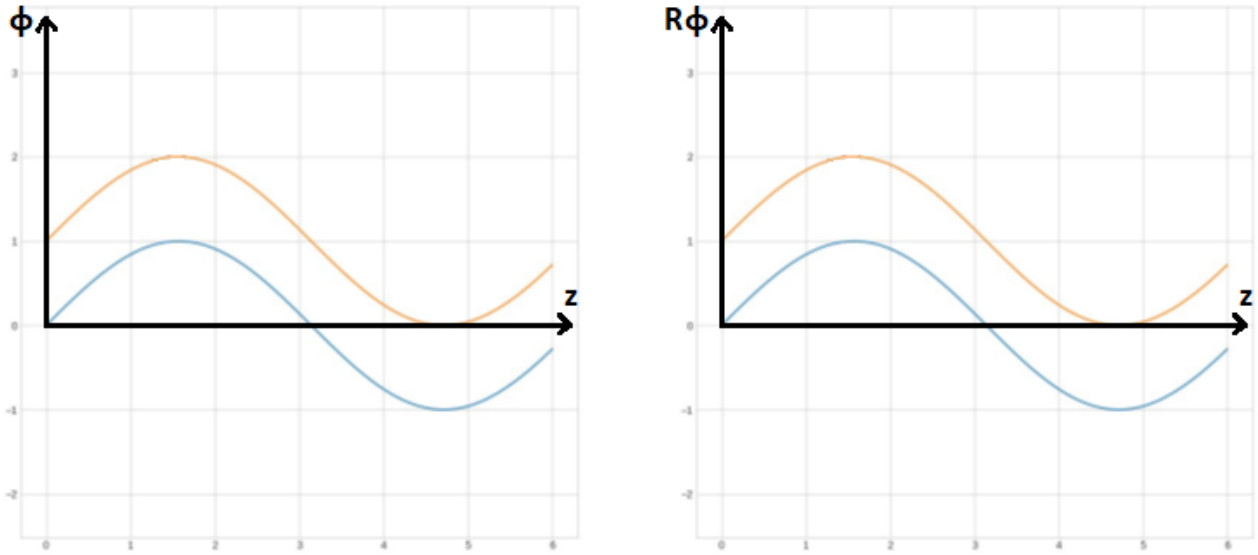


Σχήμα 5.1.1.: Απεικόνιση αντικειμένων του προβλήματος, (πραγματοποιήθηκε με χρήση προγράμματος vectary)

Το πέρας της προεξοχής του μικρού κυλίνδρου διαγράφει μια τροχιά πάνω στην εσωτερική επιφάνεια του κυλινδρικού φλοιού, ακτίνας R. Ο χώρος στον οποίο διαγράφεται η εν λόγω τροχιά είναι δύο διαστάσεων διότι η προεξοχή διατηρεί σταθερό μήκος, R. Οι ελεύθερες συντεταγμένες της κίνησης είναι η πολική γωνία, φ, και το ύψος, z. Η κίνηση που απαιτείται για τη συγκομιδή ενός μανιταριού συνδυάζει την ανύψωση (κατά τον άξονα z) και την περιστροφή του (σύμφωνα με τη γωνία φ), τις οποίες θα προσομοιάσουμε στο πρόβλημά μας, μέσω της μετατόπισης και της στροφής του πέρατος της προεξοχής του μικρού κυλίνδρου, αντίστοιχα. Όπως προαναφέραμε, η προεξοχή θα κινείται μέσα σε έναν διάδρομο δύο διαστάσεων, ο οποίος θα λαμβάνει χώρα πάνω στην εσωτερική κυλινδρική επιφάνεια του κυλινδρικού φλοιού. Ενδεικτικά η μορφή του μπορεί να είναι όπως στο Σχ. 5.1.2.(α).

5.1.1 Διάδρομος

Στο πρόβλημά μας θα μας φανεί χρήσιμο να εργαστούμε με συντεταγμένες που έχουν αποκλειστικά μονάδες μέτρησης μήκους. Για τον λόγο αυτό θα αντικαταστήσουμε την πολική γωνία, φ, με το μήκος τόξου, $R\varphi$. Το σχήμα του διαδρόμου θα έχει παρόμοια μορφή, διότι η πολική γωνία, φ, και το μήκος τόξου, $R\varphi$, συνδέονται με γραμμικό τρόπο, με συντελεστή αναλογίας R, όπου $R = \sigma\tau\alpha\theta\varepsilon\rho\acute{o}$. Για παράδειγμα, για $R = 1$ η μορφή του διαδρόμου θα δίνεται από το Σχ. 5.1.2.(β).



(α) (β)

Σχήμα 5.1.2. Ενδεικτική μορφή του διαδρόμου στο εσωτερικό του κυλίνδρου, (α) $\varphi = \varphi(z)$ και (β) $R\varphi = R\varphi(z)$, με ακτίνα $R = 1$

Η ανάδραση από τα τοιχώματα του διαδρόμου μπορεί να μοντελοποιηθεί θεωρώντας ότι το πέρας της προεξοχής του μικρού κυλίνδρου διαθέτει ελατήρια μήκους ρ, τα οποία όταν έρχονται σε επαφή με τα τοιχώματα του διαδρόμου, του ασκούν μια δύναμη επαναφοράς προς το διάδρομο. Τα ελατήρια αυτά χρησιμοποιούνται αποκλειστικά για τη μέτρηση της δύναμης από τα τοιχώματα του διαδρόμου και δεν καθορίζουν άμεσα την κίνηση του αντικειμένου (ταχύτητα, επιτάχυνση). Στη συνέχεια, θα κατασκευάσουμε το πεδίο δυνάμεων (force feedback).

5.1.2 Πεδίο δυνάμεων

Έστω ένα ορθοκανονικό σύστημα συντεταγμένων του οποίου ο άξονας z ταυτίζεται με τον άξονα συμμετρίας του κυλινδρικού φλοιού. Η θέση του κέντρου της προεξοχής του μικρού κυλίνδρου (προεξοχής) περιγράφεται από το διάνυσμα $\vec{r} = R\hat{\rho}(\varphi) + z\hat{z}$ σε κυλινδρικές συντεταγμένες. Η θέση του κοντινότερου σημείου του τοίχου από το κέντρο της προεξοχής περιγράφεται από το διάνυσμα $\overrightarrow{r_0} = R\hat{\rho}(\varphi_0) + z_0\hat{z}$ σε κυλινδρικές συντεταγμένες. Τότε, η δύναμη που θα δέχεται το πέρας της προεξοχής θα έχει διεύθυνση:

$$\hat{F} = \frac{\vec{F}}{|\vec{F}|} = \frac{\vec{r} - \overrightarrow{r_0}}{|\vec{r} - \overrightarrow{r_0}|}$$

και μέτρο:

$$|\vec{F}| = k(\rho - |\vec{r} - \overrightarrow{r_0}|)$$

Έπειτα, θα αναζητήσουμε το πεδίο δυνάμεων:

$$\vec{r} - \overrightarrow{r_0} = R\hat{\rho}(\varphi) + z\hat{z} - R\hat{\rho}(\varphi_0) - z_0\hat{z} \Rightarrow$$

$$\vec{r} - \overrightarrow{r_0} = R(\cos(\varphi)\,\hat{x} + \sin(\varphi)\,\hat{y} - \cos(\varphi_0)\,\hat{x} - \sin(\varphi_0)\,\hat{y}) + \delta z\hat{z}$$

Θεωρούμε ότι οι κινήσεις με τις οποίες το ρομποτικό χέρι κινεί τον μικρό κύλινδρο επιτρέπουν μικρές αλλαγές στην πολική γωνία, φ, και ότι ο διάδρομος έχει μικρό πλάτος. Επομένως, ισχύει ότι το $\boldsymbol{\delta\varphi = \varphi - \varphi_0}$ είναι πολύ μικρό ($\boldsymbol{\delta\varphi} < 20°$). Στη συνέχεια, αντικαθιστούμε $\varphi = \delta\varphi + \varphi_0$ στο $\vec{r} - \overrightarrow{r_0}$ και το αναπτύσσουμε κατά Taylor γύρω από το $\delta\varphi = 0$, κρατώντας τους δύο πρώτους όρους του αναπτύγματος:

$$\vec{r} - \overrightarrow{r_0} = R(\cos(\varphi_0)\,\hat{x} - \sin(\varphi_0)\,\delta\varphi\hat{x} + \sin(\varphi_0)\,\hat{y} + \cos(\varphi_0)\,\delta\varphi\hat{y} - \cos(\varphi_0)\,\hat{x} - \sin(\varphi_0)\,\hat{y}) + \delta z\hat{z} \Rightarrow$$

$$\vec{r} - \overrightarrow{r_0} = R\delta\varphi(-\sin(\varphi_0)\,\hat{x} + \cos(\varphi_0)\,\hat{y}) + \delta z\hat{z} \Rightarrow$$

$$\vec{r} - \overrightarrow{r_0} = R\delta\varphi\hat{\varphi} + \delta z\hat{z}$$

Επομένως, η κατεύθυνση και το μέτρο του πεδίου δυνάμεων είναι:

$$\hat{F} = \frac{\vec{F}}{|\vec{F}|} = \frac{1}{\sqrt{(R\delta\varphi)^2 + \delta z^2}}(R\delta\varphi, \delta z)$$

$$|\vec{F}| = k(\rho - \sqrt{(R\delta\varphi)^2 + \delta z^2})$$

Έχοντας προσδιορίσει το πεδίο δυνάμεων, μπορούμε να προσομοιώσουμε την προεξοχή με ένα αντικείμενο – πράκτορα που θα εκπαιδευτεί έτσι ώστε να εξέρχεται από έναν διάδρομο δύο διαστάσεων, στο επίπεδο (x, y), όπου $y = R\varphi$ και $x = z$, με πεδίο δυνάμεων:

$$\hat{F} = \frac{\vec{F}}{|\vec{F}|} = \frac{1}{\sqrt{\delta x^2 + \delta y^2}}(\delta x, \delta y)$$

$$|\vec{F}| = k(\rho - \sqrt{\delta x^2 + \delta y^2})$$

5.1.3 Σχήμα του πράκτορα

Έπειτα, θα δείξουμε κάτω από ποιες συνθήκες η προβολή της διατομής της προεξοχής κυλινδρικού σχήματος, στο επίπεδο $(R\varphi, z)$ είναι κύκλος. Έστω ότι η προεξοχή βρίσκεται στην κατεύθυνση του άξονα $x'x$ $(\varphi = 0)$ και ότι το σχήμα της στο επίπεδο $(R\varphi, z)$ είναι κύκλος με ακτίνα ρ.

$$(R\varphi)^2 + z^2 = \rho^2$$

Πραγματοποιώντας αλλαγή συντεταγμένων από κυλινδρικές σε καρτεσιανές, $\begin{cases} x = Rcos\varphi \\ y = Rsin\varphi \end{cases}$ και θεωρώντας μικρές γωνίες $\boldsymbol{\varphi}$ ($\boldsymbol{\varphi} < 20°$), βρίσκουμε $\begin{cases} x = R \\ y = R\varphi \end{cases}$. Επομένως, οι συντεταγμένες της προβολής της διατομής της προεξοχής ικανοποιούν τη σχέση:

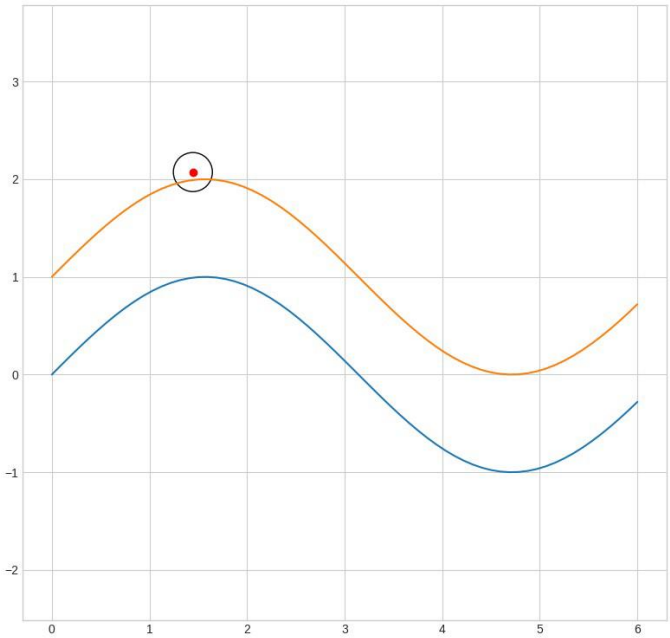
$$y^2 + z^2 = \rho^2 \text{ και } x = R$$

Παρατηρούμε ότι αν η κυλινδρική προεξοχή έχει μικρή ακτίνα ρ, και συνεπώς η γωνία φ παίρνει μικρές τιμές, τότε η προβολή της διατομής της προεξοχής στο επίπεδο $(R\varphi, z)$ είναι κύκλος. Επομένως, στο επίπεδο $(R\varphi, z)$ θα θεωρήσουμε κυκλικό πράκτορα, ο οποίος θα προσομοιάζει την προεξοχή του συμπαγούς κυλίνδρου.

5.2 Περιγραφή χώρου κατάστασης και δράσεων

Αρχικά, θα γίνει περιγραφή του χώρου καταστάσεων και δράσεων του συστήματός μας, δηλαδή του χώρου όλων των πιθανών διαμορφώσεων – καταστάσεων του συστήματος και του χώρου όλων των πιθανών δράσεων που μπορεί να επιλέξει ο πράκτορας σε ένα συγκεκριμένο περιβάλλον αντίστοιχα. Με τον τρόπο αυτό, γίνεται πλήρης περιγραφή της Μαρκοβιανής διαδικασίας που αναπαριστά το σύστημά μας. Στο πρόβλημά μας, δεν διαθέτουμε μοντέλο, δηλαδή δεν έχουμε πληροφορίες για τη δυναμική του συστήματος, ούτε για το αντικείμενό μας. Η σκιαγράφηση του περιβάλλοντος γίνεται μέσω της ανάδρασης δύναμης, δηλαδή των αισθητήρων που φέρει το αντικείμενό μας.

Θεωρούμε μια τυχαία μορφή διαδρόμου, με ημιτονοειδές σχήμα και μια τυχαία αρχική θέση του αντικειμένου – πράκτορα, η οποία επιλέγεται χειροκίνητα, όπως η ακόλουθη:



Σχήμα 5.2: Τυχαία μορφή διαδρόμου στο μοντέλο εκπαίδευσης, με τυχαία θέση του αντικειμένου

Η ανάδραση δύναμης έχει τη μορφή δύναμης επαναφοράς, όπως στο σύστημα ενός απλού αρμονικού ταλαντωτή:

$$\vec{F} = -k\vec{x}$$

όπου $\vec{x}$ είναι το διάνυσμα της διείσδυσης (penetration) του πράκτορα στον τοίχο του διαδρόμου. Επομένως, το $\vec{x}$ ορίζεται ως η απόσταση του σημείου του αντικειμένου με τη μεγαλύτερη διείσδυση, από το σημείο του τοίχου που είναι πλησιέστερο στο κέντρο του αντικειμένου.

Επομένως, στο πρόβλημά μας, ο χώρος κατάστασης, ο οποίος πρέπει να περιγράφει το περιβάλλον, είναι η δύναμη που δέχεται το αντικείμενο, κατά μέτρο και φορά:

$$s_t = \{\left|\vec{F}\right| = k \cdot penetration,\ \hat{F} = \frac{\vec{F}}{\left|\vec{F}\right|}\}$$

Αναφορικά στο χώρο των δράσεων, αυτός αποτελείται από την differential τη θέση του αντικειμένου στο επίπεδο: $a_t = \{p_x, p_y\}$.

5.3 Συνάρτηση επιβράβευσης

Στη συνέχεια, ορίζουμε τη συνάρτηση επιβράβευσης του συστήματος, η οποία συνδράμει σημαντικά στη διαδικασία της εκπαίδευσης του πράκτορα. Μέσω της επιβράβευσης, ο πράκτορας καλείται να αντιληφθεί αν και κατά πόσο είναι σωστές οι αποφάσεις που παίρνει. Επομένως, με τον τρόπο αυτό προσαρμόζει τις επόμενες αποφάσεις του και βελτιώνει τις επιλογές του.

Στο πρόβλημά μας, ο στόχος μας είναι η κατευθυνόμενη κίνηση του κυκλικού αντικειμένου μέσα στο διάδρομο, μέχρι την έξοδό του από αυτόν, και η μη διείσδυση στους τοίχους του. Συνεπώς, η συνάρτηση επιβράβευσης θα επικροτεί την κίνηση μέσα στο διάδρομο με κατεύθυνση προς την έξοδό του, την οποία έχουμε ορίσει προς τα δεξιά. Επιπλέον, ο πράκτορας θα λαμβάνει αρνητική επιβράβευση κάθε φορά που επιχειρεί να εισέλθει στους τοίχους του διαδρόμου. Η επιβράβευση, είτε θετική, είτε αρνητική, εξαρτάται από τους εξής τρείς παράγοντες:

- τη μετατόπιση του αντικειμένου,
- την κίνησή του προς τα δεξιά και
- το «χρόνο».

Πιο συγκεκριμένα, θεωρούμε τον υπολογισμό της επιβράβευσης μια πολύκλαδη συνάρτηση. Ο πρώτος παράγοντας που αναφέραμε, υπολογίζεται με βάση τη διαφορά της τρέχουσας και της προηγούμενης θέσης διείσδυσης του αντικειμένου:

$$differential_{position} = penetration_{previous} - penetration_{current}$$

Συνεπώς, αν $differential_{position} > 0$ τότε το αντικείμενο βρίσκεται έξω από το διάδρομο, μέσα σε κάποιον τοίχο, και με κατεύθυνση προς το διάδρομο. Άρα, παρόλο που δεν κινείται στο

εσωτερικό του διαδρόμου, η επιβράβευση που θα δοθεί στον πράκτορα θα είναι θετική, μόνο στην περίπτωση που μπαίνει στο διάδρομο με ταυτόχρονη προς τα δεξιά πορεία, καθώς το αντικείμενο τείνει να διορθώσει το λάθος στην τροχιά του. Όμως, η επιβράβευση θα είναι αρνητική σε περίπτωση που εισέλθει στο διάδρομο με κίνηση προς τα αριστερά, δηλαδή όταν η μεταβλητή $right_{movement} < 0$. Αντιθέτως, αν $differential_{position} < 0$ τότε η επιβράβευση θα είναι μονίμως αρνητική, δεδομένου ότι το αντικείμενο βρισκόταν και εξακολουθεί να βρίσκεται εκτός διαδρόμου και μάλιστα να απομακρύνεται από αυτόν. Τέλος, σε περίπτωση που $differential_{position} = 0$ τότε η επιβράβευση θα κριθεί από τους άλλους δύο παράγοντες που προαναφέραμε.

Αναφορικά στην κίνηση προς τα δεξιά, αυτή υπολογίζεται μέσω της μετατόπισης στον άξονα x. Ο παράγοντας αυτός συνεισφέρει θετικά στη συνάρτηση επιβράβευσης αν η μεταβλητή $right_{movement}$ έχει μη αρνητικό πρόσημο, και αρνητικά σε αντίθετη περίπτωση. Επιπλέον, στην περίπτωση που έχουμε $differential_{position} < 0$, δηλαδή κακή κίνηση του πράκτορα, ο παράγοντας $right_{movement}$ δεν συνεισφέρει στην επιβράβευση.

Ο παράγοντας του χρόνου, έχει μονίμως αρνητική συνεισφορά στην επιβράβευση, καθώς ο ρόλος του είναι να υπογραμμίσει την καθυστέρηση εύρεσης λύσης στο πρόβλημα. Η αρνητική του τιμή δεν είναι ίδια σε κάθε μια εκ των παραπάνω περιπτώσεων. Ο λόγος είναι ότι η καθυστέρηση θέλουμε να έχει μεγαλύτερη επίδραση όταν και η επιλογή της δράσης δεν είναι επιθυμητή. Επομένως, η τιμή του παράγοντα χρόνου είναι περισσότερο αρνητική όταν $differential_{position} < 0$.

Με τον τρόπο αυτό, παρουσιάζουμε τη μορφή της συνάρτησης επιβράβευσης, έχοντας ορίσει $differential_{position} = diff$, $right_{movement} = right$, $penetration_{new} = pen_{new}$, $penetration_{previous} = pen_{previous}$ και ως $time$ τη μεταβλητή του χρόνου:

$$r = \begin{cases} \begin{cases} a \cdot diff + b \cdot right + c \cdot time, & if\ right > 0 \\ -constant, & if\ right \leq 0 \end{cases}, & if\ pen_{new} = 0 \\ d \cdot pen_{new} + b \cdot right + c \cdot time, & if\ otherwise \end{cases}, \quad if\ diff > 0$$

$$\begin{cases} a \cdot diff + c \cdot time, & if\ pen_{previous} = 0 \\ -d \cdot pen_{new} + c \cdot time, & if\ otherwise \end{cases}, \quad if\ diff < 0$$

$$\begin{cases} -constant, & if\ right \leq 0 \\ constant, & if\ right > 0 \end{cases}, \quad if\ diff = 0$$

Όπου τα a, b, c, d, με $a, b, d > 0$ και $c < 0$ αποτελούν ακέραιες σταθερές και αντιπροσωπεύουν τη συνεισφορά των παραγόντων που αναφέραμε, στη συνάρτηση επιβράβευσης. Οι τιμές που δόθηκαν τόσο στα βάρη όσο και στους συντελεστές τους, έχουν επιλεγεί με λογική trail-and-error.

5.4 Είδος Μάθησης

Στο προς μελέτη πρόβλημα διαθέτουμε συνεχείς χώρους κατάστασης και δράσης. Συμπερασματικά, κάνουμε χρήση του αλγορίθμου CACLA που μελετήθηκε στο κεφάλαιο 3. Όπως έχουμε αναφέρει, στην περίπτωση συνεχούς χώρου κατάστασης, κάνουμε χρήση

παραμετροποιήσιμων Function Approximators (FAs), οι οποίοι αποθηκεύουν τις τιμές των καταστάσεων που παρατηρούμε και γενικεύουν για άγνωστες καταστάσεις. Στην εργασία μας, χρησιμοποιούμε Νευρωνικά Δίκτυα (NN) σαν FAs, τόσο για τον Actor όσο και για τον Critic, και συνεπώς η ανανέωση πραγματοποιείται στις παραμέτρους του NN, όπως αναφέραμε στην παράγραφο 3.5.

Για να μπορέσουμε να ορίσουμε τη νέα κατάσταση του αντικειμένου μας, δειγματοληπτούμε τις πιθανές μας δράσεις χρησιμοποιώντας Gaussian πολιτική εξερεύνησης: $V_{t+1}(s_t) = V_t(s_t) + a_t \cdot \delta_t$ όπου δ_t είναι το TD-error και a_t είναι η δράση που λαμβάνεται από μια Gaussian κατανομή με μέση τιμή A_{C_t}. Υπενθυμίζουμε ότι για τον actor θεωρούμε θ^{Ac} το διάνυσμα παραμέτρων του FA και $A_{C_t}(s_t)$ την έξοδο του FA τη χρονική στιγμή t. Η Gaussian πολιτική έχει τη μορφή:

$$\pi_t(s_t, a_t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(a - A_{C_t}(s_t)\right)^2}{2\sigma^2}}$$

όπου $\pi_t(s_t, a_t)$ είναι η πολιτική, ενώ π είναι η μαθηματική σταθερά. Επίσης, σ είναι η τυπική απόκλιση της Gaussian εξερεύνησης.

Για τον υπολογισμό του TD-error ($\delta_t = r_t + \gamma V_t(s_{t+1}) - V_t(s_t)$) χρησιμοποιήσαμε μια σειρά από διαφορετικά γ, των οποίων τη συμβολή θα παρατηρήσουμε στο επόμενο κεφάλαιο.

Πέραν της συνάρτησης επιβράβευσης, η οποία έχει differential μορφή, ο υπολογισμός της νέας θέσης του αντικειμένου είναι επίσης differential, δεδομένου ότι το αποτέλεσμα της Gaussian exploration προστίθεται στις συντεταγμένες του κέντρου του αντικειμένου όπως θα δούμε στην επόμενη παράγραφο.

5.5 Εκτέλεση αλγορίθμου

Στο σημείο αυτό θα παρουσιάσουμε αναλυτικά την εφαρμογή του CACLA στο πρόβλημά μας.

Η είσοδος των NN του actor και του critic θα είναι το observation, το οποίο αποτελείται από τρία στοιχεία, το μέτρο της δύναμης και τις δύο συνιστώσες του μοναδιαίου της διανύσματος: $obs = \{|\vec{F}|, \hat{F}_x, \hat{F}_y\}$. Με βάση το observation$_0$, ο critic κάνει την πρώτη πρόβλεψη της συνάρτησης αξίας, V_0 και ο actor προβλέπει την πρώτη δράση A_0, η οποία θα είναι δύο διαστάσεων (όπως έχουμε προαναφέρει, η έξοδος του actor NN είναι δύο διαστάσεων). Στη συνέχεια, πραγματοποιούμε Gaussian exploration γύρω από τη δράση A_0, με σκοπό να βρούμε την τελική δράση a_0. Πλέον, η δράση a_0 χρησιμοποιείται ως μετατόπιση ως προς την προηγούμενη θέση του αντικειμένου. Αφού υπολογιστεί η νέα θέση, με βάση την differential δράση, υπολογίζουμε τη δύναμη που δέχεται το αντικείμενο στη θέση αυτή, αναθέτοντας μια στο observation$_1$. Επιπλέον, για τη νέα θέση υπολογίζουμε την επιβράβευση που δεχόμαστε, σύμφωνα με τη διείσδυση που έχουμε σε κάποιον τοίχο. Στη συνέχεια, ο critic πραγματοποιεί μια νέα πρόβλεψη, σύμφωνα με το νέο observation που έχουμε, V_1. Έπειτα, υπολογίζουμε το TD-error, σύμφωνα με το οποίο θα αποφασίσουμε εάν θα ανανεώσουμε τον actor. Ανεξάρτητα από την τιμή του TD-error,

ανανεώνουμε τον critic με χρήση Gradient Descent, ως προς το TD-target. Τέλος, η ανανέωση του actor χρησιμοποιεί και αυτή Gradient Descent, προς την κατεύθυνση του a_0.

Η παραπάνω διαδικασία είναι επαναληπτική και συμβαίνει τόσες φορές όση είναι η τιμή της υπερπαραμέτρου των εποχών. Στο σημείο αυτό, αξίζει να σημειωθεί πως κατά την εκπαίδευση του πράκτορα, έχουμε θέσει μέγιστη επιτρεπόμενη διείσδυση έτσι ώστε να μην απειρίζεται η θέση του και ώστε να προλαβαίνει να εκπαιδευτεί σε λιγότερα βήματα, δίχως να εξερευνά ολόκληρο το χώρο των δύο διαστάσεων. Επιπλέον, πέραν της ανανέωσης των παραμέτρων του actor, όταν το TD-error είναι θετικό, πραγματοποιείται και η ανανέωση του observation_0, το οποίο θα πάρει την τιμή του observation_1 μόνο σε αυτήν την περίπτωση.

Στη συνέχεια παρατίθεται ο ψευδοκώδικας που περιγράφει η παραπάνω μέθοδος, ορίζοντας:

$$\begin{cases} input\_\dim = είσοδος\ NN \\ output\_\dim = έξοδος\ NN \\ alpha = lr_{actor} \\ beta = lr_{critic} \\ lr_{decay} \\ ef_{decay} = exploration_{decay} \\ e_{factor} = exploration_{factor} \end{cases}$$

```
Initialize Cacla(input_dim, output_dim, alpha, beta, gamma, lr_decay, ef_decay, e_factor)
input_dim = observation_space.shape
output_dim = action_space.shape
…
#get object position, radius & corridor walls
Initialize Simulation
#get first observation from force feedback function
observation0 = force(cx, cy, radius, coords1, coords2)
#done is a flag that indicates the termination of the task, aka we reached the foal position
WHILE(NOT done):
        #compute V0 value function from critic NN
        V0 = critic.predict(observation0)
        #compute action A0 from actor NN
        A0 = actor.predict(observation0)
        #perform Gaussian exploration around action A0 and find new action a0
        #action a0 will represent a new position (could be differential position as well)
        a0 = exploration(A0)
        #take action a0 at the simulated environment and plot the action
        simulation.step(a0)
        #update object position cx, cy
        #get second observation from force feedback function
```

```
        observation1 = force(cx, cy, radius, coords1, coords2)
        #reward computation according to penetration, which is given by force feedback
        function
        r = get_reward(penetration)
        #compute termination condition, aka variable done
        #compute V1 value function from critic NN
        V1 = critic.predict(observation1)
        #compute TD-error
        δ =reward + gamma * V1 –V0
        #fit the critic NN
        critic.fit(observation0, [reward + gamma * V1])
        IF δ>0 THEN:
            #fit the actor NN
            actor.fit(observation0, a0)
        ENDIF
ENDWHILE
```

Πίνακας 5.1: Ψευδοκώδικας CACLA που χρησιμοποιήθηκε στην εργασία

Κεφάλαιο 6

Πειραματικά αποτελέσματα

Στο κεφάλαιο αυτό θα παρουσιάσουμε και θα αναλύσουμε τα πειραματικά αποτελέσματα που προέκυψαν από την εφαρμογή του προβλήματος που περιγράψαμε στο κεφάλαιο 5. Τα πειράματά μας αφορούν στην εκπαίδευση ενός σφαιρικού αντικειμένου, προκειμένου να καταφέρει να βγει από έναν διάδρομο πεπερασμένου μήκους, ο οποίος εκτείνεται στο xy επίπεδο.

6.1 Προσδιορισμός υπερπαραμέτρων

Για τα νευρωνικά δίκτυα Actor και Critic έχει χρησιμοποιηθεί η ίδια δομή. Πιο συγκεκριμένα, έχουμε κάνει χρήση ενός ακολουθειακού μοντέλου με ένα στρώμα εισόδου και ένα εξόδου. Στο εσωτερικό στρώμα τοποθετήσαμε 100 νευρώνες και $ReLU$ συνάρτηση ενεργοποίησης, ενώ το στρώμα εξόδου έχει μια διάσταση για τον critic και δύο για τον actor, και γραμμική συνάρτηση ενεργοποίησης:

$$ReLU(x) = \begin{cases} x, x \geq 0 \\ 0, x < 0 \end{cases} \qquad\qquad Linear(x) = x$$

Όλα τα βάρη των NNs αρχικοποιούνται τυχαία με βάση μια κανονική κατανομή, με μηδενική μέση τιμή και διασπορά μικρότερη ή ίση της μονάδα. Στα νευρωνικά τόσο του Actor όσο και του Critic χρησιμοποιείται ο optimizer Adam, ο οποίος συνδυάζει τη λογική του RMSRoot και του Momentum optimizer. Επιπλέον, αρχικοποιούμε το ρυθμό μάθησης του κάθε NN στην τιμή 0.01 και στο πέρας κάθε επεισοδίου ανανεώνεται ως εξής: $lr_{new} = 0.997 \cdot lr_{old}$. Οι τιμές που δόθηκαν στο ρυθμό μάθησης και στο ρυθμό μείωσής του, έχουν επιλεγεί με λογική trail-and-error.

Για να μπορέσουμε να ορίσουμε τη νέα κατάσταση του αντικειμένου μας, πραγματοποιούμε Gaussian εξερεύνηση γύρω από την προτεινόμενη από τον Actor δράση. Όπως είδαμε στο κεφάλαιο 5, η Gaussian πολιτική εξερεύνησης έχει εκθετική μορφή ($\pi_t(s_t, a_t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(a - A_{C_t}(s_t)\right)^2}{2\sigma^2}\right)$), όπου $\pi_t(s_t, a_t)$ είναι η πολιτική, ενώ π είναι η μαθηματική σταθερά. Επίσης, σ είναι η τυπική απόκλιση της Gaussian εξερεύνησης, η οποία εκκινεί από μία τιμή 0.25, ενώ στο τέλος κάθε επεισοδίου ανανεώνεται ως εξής: $ef_{new} = 0.997 \cdot ef_{old}$. Οι τιμές που δόθηκαν στην τυπική απόκλιση σ και στο ρυθμό μείωσής της, έχουν επιλεγεί με λογική trail-and-error.

Για τον υπολογισμό του TD-error ($\delta_t = r_t + \gamma V_t(s_{t+1}) - V_t(s_t)$) χρησιμοποιήσαμε διάφορες τιμές του γ, τις οποίες επιλέξαμε σύμφωνα με τη βιβλιογραφία και θα μελετήσουμε αναλυτικά στα πειράματα που ακολουθούν.

Για τον υπολογισμό του reward έχουν δοθεί οι ακόλουθες τιμές:

$$r = \begin{cases} \begin{cases} 2 \cdot a \cdot diff + b \cdot right + c \cdot time, & if\ right > 0 \\ -2, & if\ right \leq 0 \end{cases}' & if\ pen_{new} = 0 \\ 4 \cdot d \cdot pen_{new} + b \cdot right + c \cdot time, & if\ otherwise \end{cases}, \quad if\ diff > 0$$

$$\begin{cases} a \cdot diff + 2 \cdot c \cdot time, & if\ pen_{previous} = 0 \\ -20 \cdot d \cdot pen_{new} + 2 \cdot c \cdot time, & if\ otherwise \end{cases}, \quad if\ diff < 0$$

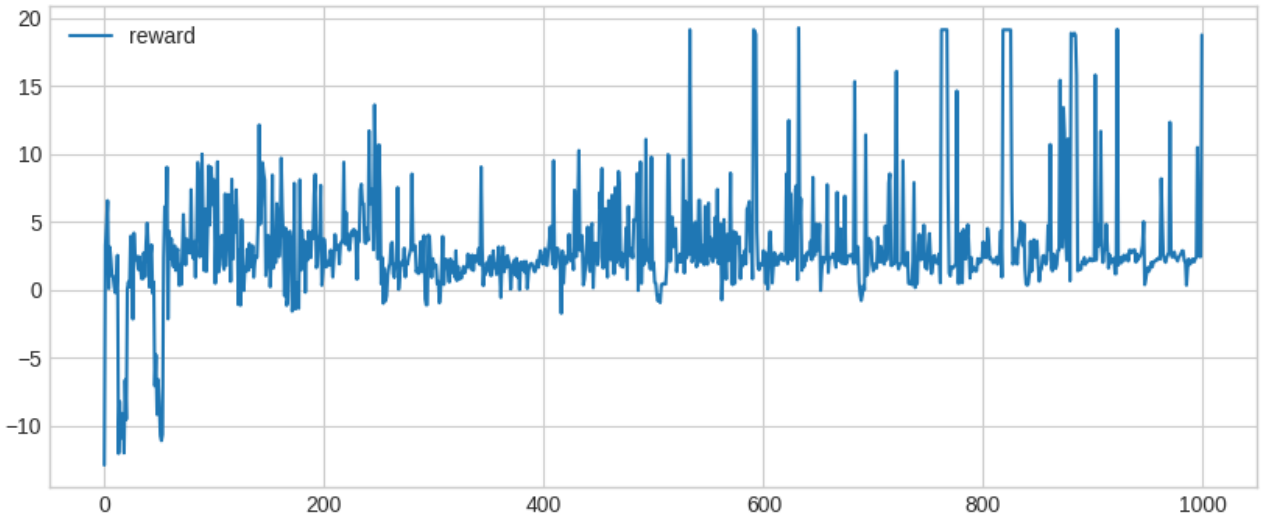
$$\begin{cases} -2, & if\ right \leq 0 \\ 20, & if\ right > 0 \end{cases}' \quad if\ diff = 0$$

Όπου τα $a, b, c, d,$ με $a, b, d > 0$ και $c < 0$ αποτελούν ακέραιες σταθερές με τιμές που έχουν επιλεγεί με λογική trail-and-error: $a = 100, b = 2, c = -1, d = 100.$

Τέλος, αναφέρουμε πως έπειτα από τη διαδικασία της Gaussian εξερεύνησης, αφού επιλεγεί η δράση που θα ακολουθήσει ο πράκτορας, πριν την εκτέλεση της δράσης, την κανονικοποιούμε έτσι ώστε το διάνυσμα της κίνησής του να έχει μέτρο 0.1. Η παραπάνω κανονικοποίηση επιλέχθηκε μέσω trial-and-error, έτσι ώστε ο πράκτορας να εκτελεί μικρά βήματα τόσο στην εκπαίδευση όσο και στο test του.

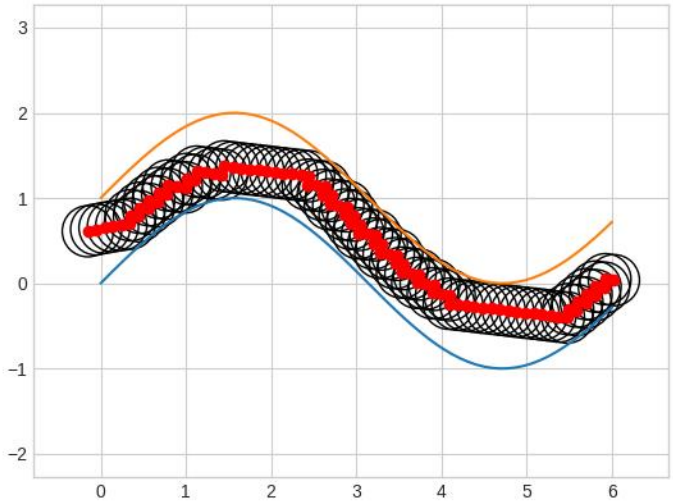
6.2 Παράθεση αποτελεσμάτων

Αρχικά, παρουσιάζουμε τα αποτελέσματα του πειράματός μας για τιμή γ=0.8, learning rate των NN ίσο με 0.01, learning rate decay ίσο με 0.997, exploration factor ίσο με 0.25 και exploration factor decay ίσο με 0.997. Τρέξαμε το πείραμα για 1000 επεισόδια, στο τέλος καθενός από τα οποία πραγματοποιήσαμε μείωση του learning rate και του exploration factor όπως προαναφέραμε. Παρακάτω φαίνεται η γραφική παράσταση της μέσης επιβράβευσης ανά επεισόδιο:

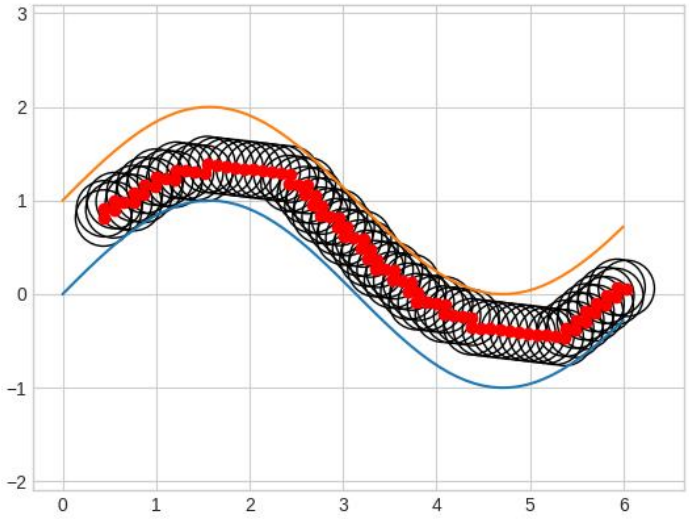


Σχήμα 6.1: Μέση επιβράβευση ανά επεισόδιο, για γ=0.8

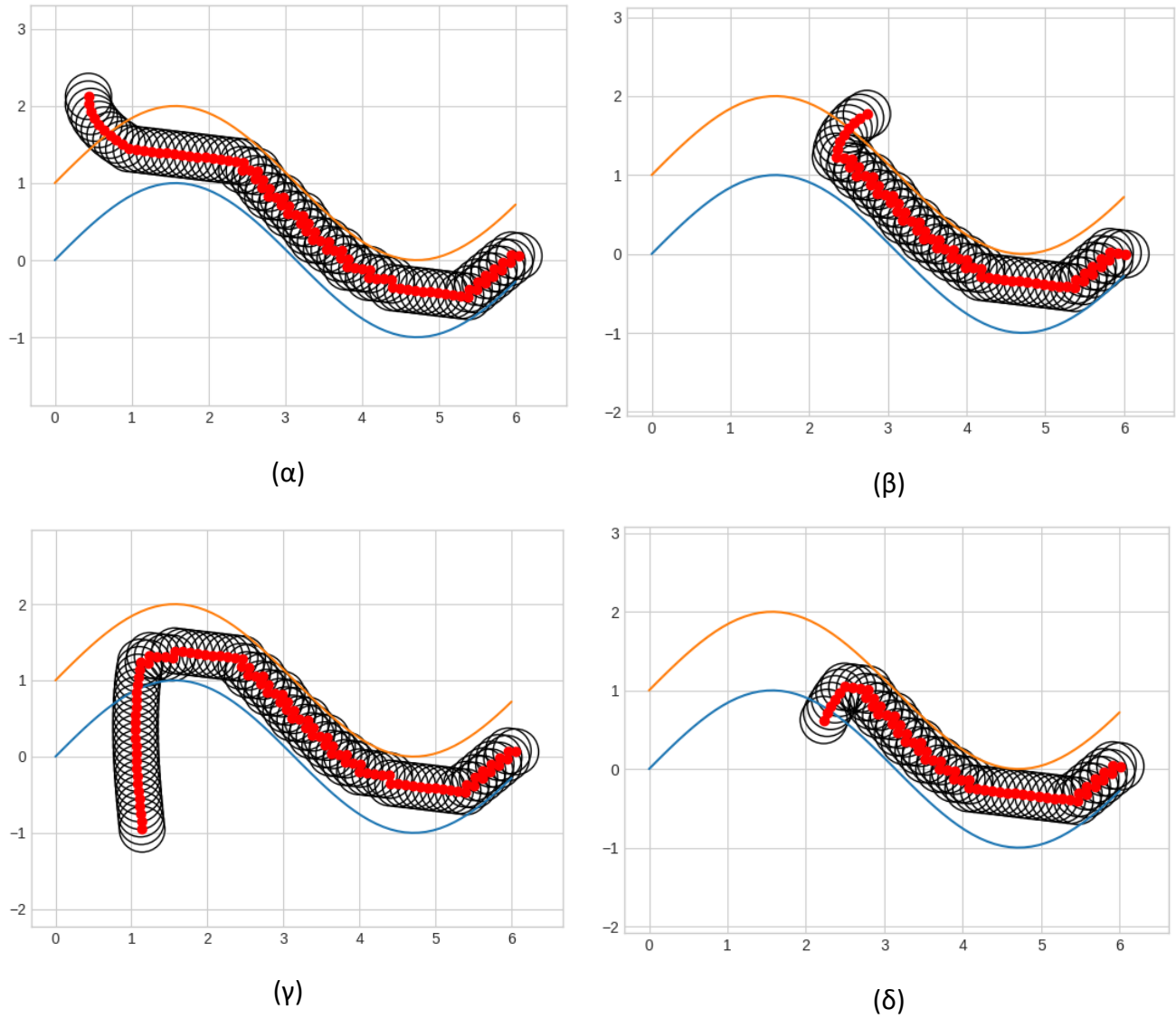
Πράγματι, από την παραπάνω γραφική απεικόνιση παρατηρούμε την αύξηση του μέσου reward έπειτα από τα πρώτα επεισόδια της εκπαίδευσης. Στη συνέχεια, μάλιστα, είναι εμφανής η σταθεροποίησή του γύρω από την τιμή $2.5 - 3$ της επιβράβευσης. Παρά τον θόρυβο που παρατηρούμε, φαίνεται η βελτίωση του μοντέλου και η βελτιούμενη και σταθερή πορεία της μάθησης. Επιπλέον, έπειτα από τα 500 βήματα της εκπαίδευσης παρατηρούμε ορισμένα αιχμηρά μέγιστα με μέση τιμή reward κοντά στο 20. Αυτό συμβαίνει καθώς στις εποχές αυτές ο πράκτορας καταφέρνει να βγει από το διάδρομο χωρίς να έρθει σχεδόν καθόλου σε επαφή με τους τοίχους του διαδρόμου. Ενώ, στις περιπτώσεις με χαμηλότερο αλλά θετικό μέσο reward, καταφέρνει να εξέλθει από το διάδρομο, αλλά πολλές φορές επιχειρεί να διεισδύσει ελαφρώς σε κάποιον τοίχο. Στα επόμενα διαγράμματα, παρουσιάζεται η οπτικοποίηση των κινήσεων και των τροχιών που ακολούθησε το αντικείμενο – πράκτορας κατά τη διάρκεια του test:



Σχήμα 6.2: Πορεία πράκτορα με σημείο εκκίνησης στην αρχή του διαδρόμου, αλλά έξω από αυτόν



Σχήμα 6.3: Πορεία πράκτορα με σημείο εκκίνησης στην αρχή του διαδρόμου, αλλά εντός αυτού

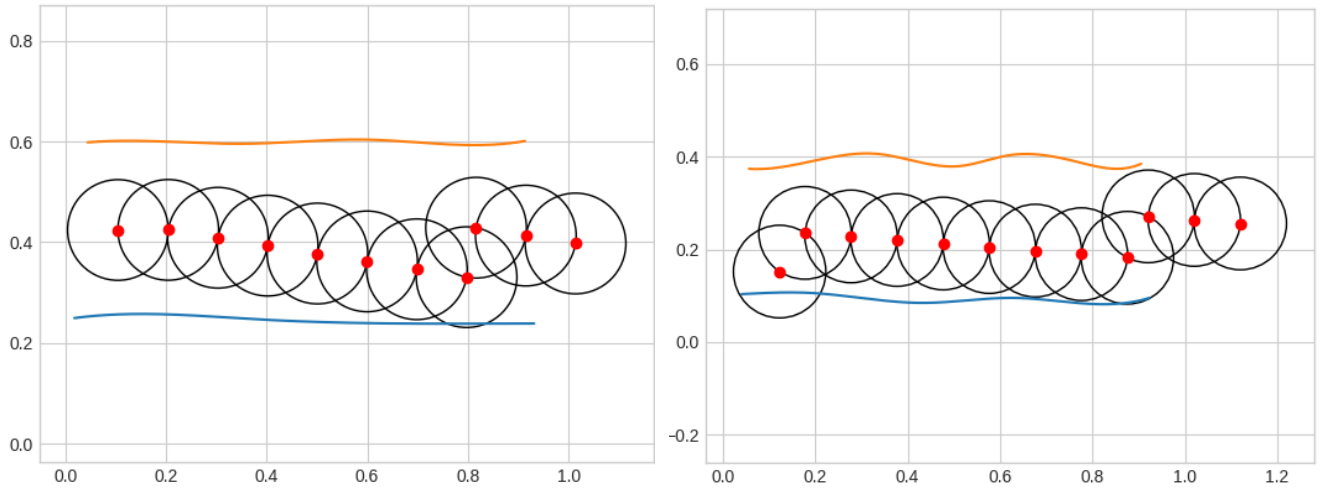


Σχήμα 6.4: Πορεία πράκτορα με σημείο εκκίνησης εκτός του διαδρόμου, για γ=0.8 (α) πάνω από την πρώτη ημιπερίοδο του πάνω τοίχου, (β) πάνω, πριν από το τέλος της πρώτης ημιπεριόδου του πάνω τοίχου, (γ) κάτω από την πρώτη ημιπερίοδο του κάτω τοίχου, και (δ) κάτω, πριν από το τέλος της πρώτης ημιπεριόδου του κάτω τοίχου

Από τις παραπάνω απεικονίσεις παρατηρούμε πως η διαδικασία της μάθησης λειτούργησε και σε περιπτώσεις όπου το σημείο εκκίνησης της πορείας του πράκτορα βρίσκεται εκτός του διαδρόμου, σε σημεία του χώρου τα οποία δεν έχουν συμπεριληφθεί στην εκπαίδευση. Κατά τη διάρκεια της εκπαίδευσης, ο πράκτορας έχει υποχρεωθεί να παραμένει ως επί το πλείστον στο εσωτερικό του διαδρόμου, ενώ του έχει επιτραπεί να πραγματοποιεί διείσδυση με μέγιστο βάθος 1.5. Στις παραπάνω προσομοιώσεις, ο πράκτορας έχει επί τούτου τοποθετηθεί εκτός του διαδρόμου σε βάθος διείσδυσης μεγαλύτερο του 1.5. Μπορούμε, επομένως, να συμπεράνουμε πως ο πράκτορας λαμβάνει σωστές αποφάσεις ακόμα και σε αυτές τις ακραίες περιπτώσεις, και άρα

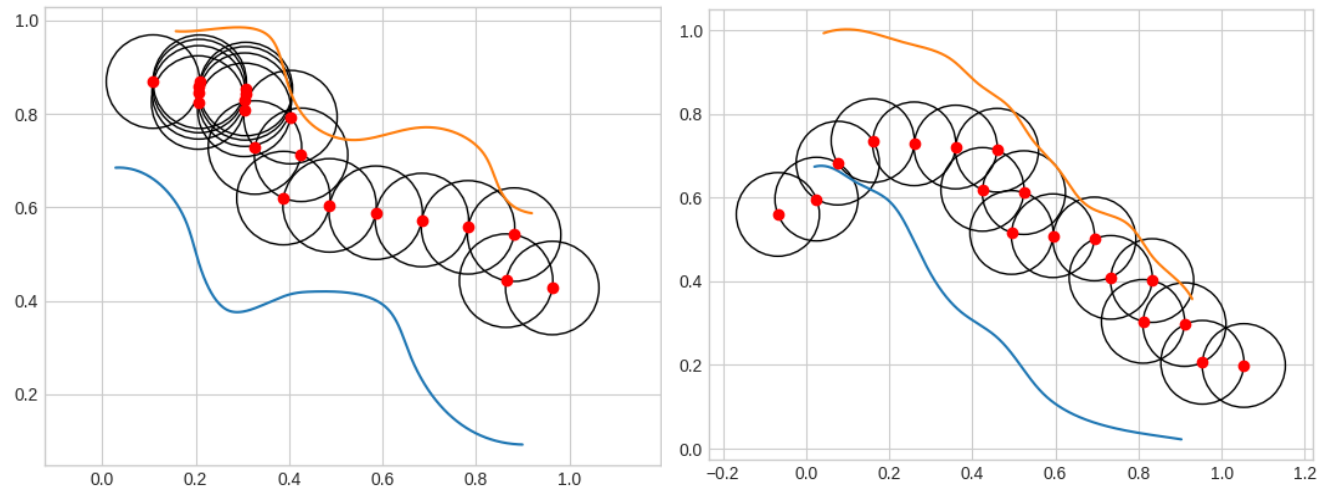
γενικεύει καλά τις αποφάσεις του σε όλο το χώρο. Συνεπώς, έχει κατανοήσει σε βάθος το μοντέλο του περιβάλλοντος.

Πέραν της γενίκευσης σε όλο το χώρο, η οποία δηλώνει την καλή αλληλεπίδραση του πράκτορα με το περιβάλλον, ο πράκτοράς μας έχει τη δυνατότητα να αποδώσει και σε νέα περιβάλλοντα, όπως βλέπουμε στη συνέχεια, όπου με χρήση μικρότερης ακτίνας και διαφορετικού σχήματος και μήκους διαδρόμου, ο πράκτορας κατάφερε να εκπληρώσει το στόχο του:



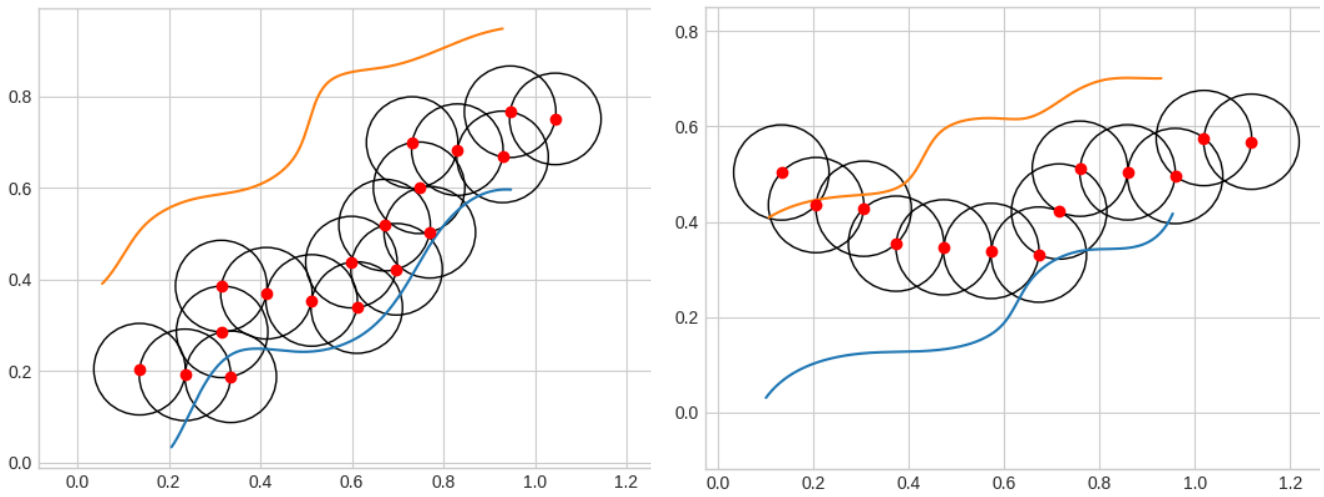
(α) (β)

Σχήμα 6.5: Πορεία πράκτορα σε νέο ευθύ διάδρομο, με διαφορετικά σημεία εκκίνησης: (α) μέσα στο διάδρομο, και (β) έξω από το διάδρομο



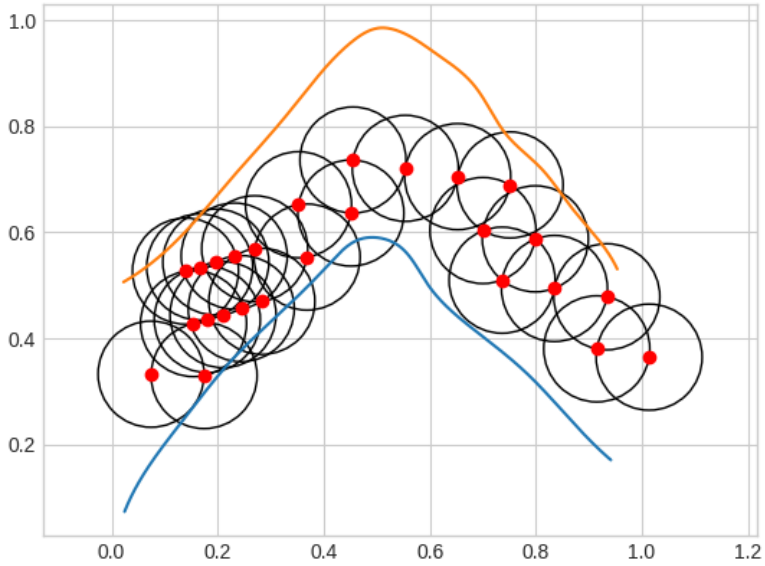
(α) (β)

Σχήμα 6.6: Πορεία πράκτορα σε νέο διάδρομο φθίνουσας κλίσης με καμπύλες, με διαφορετικά σημεία εκκίνησης: (α) μέσα στο διάδρομο, και (β) έξω από το διάδρομο



(α) (β)

Σχήμα 6.7: Πορεία πράκτορα σε νέο διάδρομο αύξουσας κλίσης με καμπύλες, με διαφορετικά σημεία εκκίνησης: (α) μέσα στο διάδρομο, και (β) έξω από το διάδρομο



Σχήμα 6.8: Πορεία πράκτορα σε νέο διάδρομο με απότομη αλλαγή κλίσης με σημείο εκκίνησης μέσα στο διάδρομο

Όπως παρατηρούμε από τις παραπάνω απεικονίσεις, η επιλογή των νέων διαδρόμων έγινε με στόχο να ελέγξουμε τη γενίκευση της εκπαίδευσης σε νέα περιβάλλοντα, με νέα είδη και σχήματα διαδρόμων και νέο μέγεθος του πράκτορα. Συγκεκριμένα, από το Σχ. 6.5. παρατηρούμε πως στην περίπτωση ενός περίπου ευθύ διαδρόμου, ο πράκτορας παίρνει σωστές αποφάσεις που του δίνουν τη μεγαλύτερη επιβράβευση, ενώ όταν πραγματοποιήσει διείσδυση σε κάποιον τοίχο, παίρνει την απόφαση αυτή που τον οδηγεί στην άμεση μείωση αυτής της διείσδυσης και στην ταυτόχρονη συνέχιση της πορείας του προς τα δεξιά.

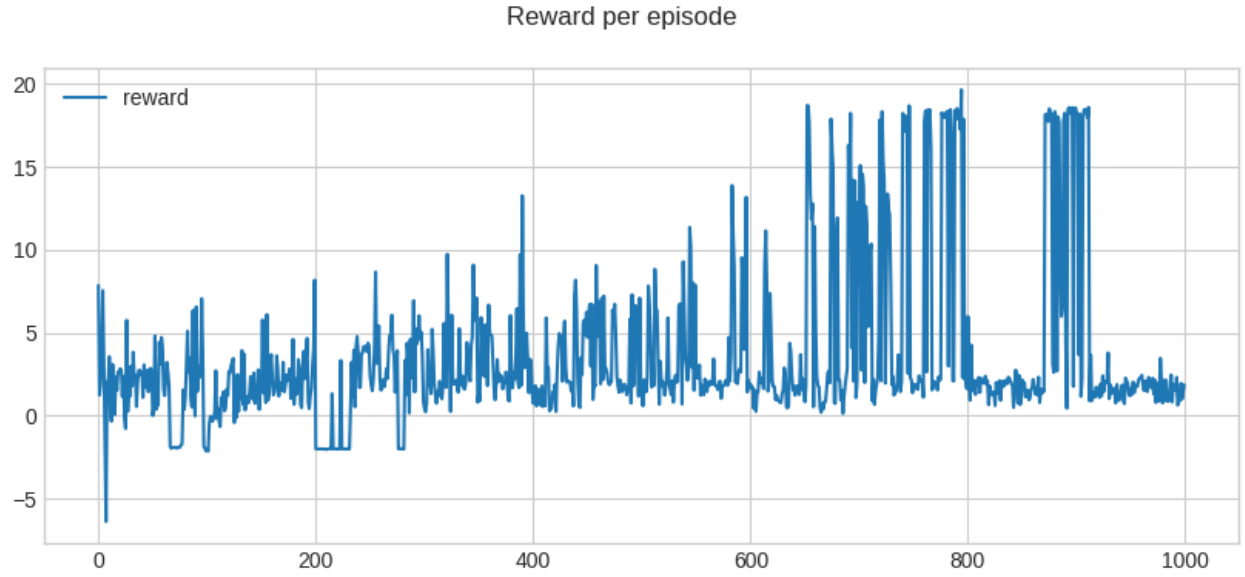
Στο Σχ. 6.6. παρουσιάζεται η πορεία του πράκτορα μέσα σε δύο νέους διαδρόμους με τα κοινά τους χαρακτηριστικά να είναι η φθίνουσα κλίση τους και η εμφάνιση καμπυλών στους τοίχους τους. Μάλιστα, ο διάδρομος του σχήματος 6.6.α διαθέτει έντονες καμπύλες με μεγάλη κλίση, οι οποίες εμφανίζονται σε μια μικρή έκταση του διαδρόμου. Στις περιπτώσεις αυτές παρατηρούμε πως ο πράκτορας προσπαθεί να ξεφύγει από τις εν λόγω περιοχές, πραγματοποιώντας κινήσεις δεξιά-αριστερά έως ότου σταματήσει να δέχεται δύναμη επαναφοράς από τον τοίχο, και τελικά να συνεχίσει την κίνησή του μέσα στο διάδρομο.

Στο Σχ. 6.7. φαίνεται η αντίθετη περίπτωση του διαδρόμου που μελετήσαμε στο Σχ. 6.6., δεδομένου ότι ο τρέχον διάδρομος διαθέτει αύξουσα κλίση. Παρατηρούμε πως είτε ο πράκτορας ξεκινήσει μέσα είτε έξω από το διάδρομο, καταφέρνει να βγει από αυτόν, ενώ πάντα διορθώνει την πορεία του όταν συγκρούεται με τον τοίχο ή διεισδύει σε αυτόν.

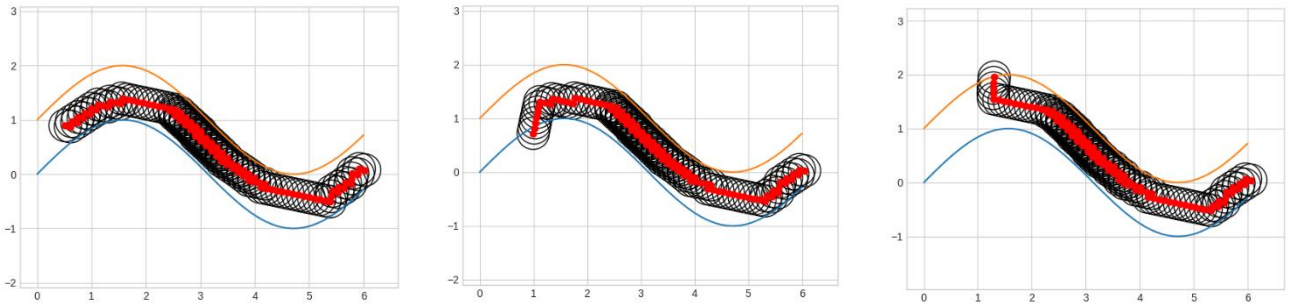
Τέλος, στο Σχ. 6.8. απεικονίζεται ένας καμπύλος διάδρομος με απότομη αλλαγή κλίσης, σε μικρή έκταση διαδρόμου. Παρόλη την απότομη αλλαγή, τη στενότητα του διαδρόμου και το σύντομο μήκος του, ο πράκτορας επιτυγχάνει και πάλι, εξερχόμενος από το διάδρομο.

Το γεγονός ότι σε οποιαδήποτε μορφή διαδρόμου, με οποιαδήποτε τροποποίηση σε σχέση με τον αρχικό διάδρομο, ο πράκτορας ακολουθεί σωστή πορεία και ερμηνεύει σωστά τις δυνάμεις που δέχεται από οποιοδήποτε σημείο των τοίχων, αποδεικνύει την αποτελεσματικότητα της εκπαίδευσής του και συνεπώς του αλγορίθμου, CACLA, που χρησιμοποιούμε.

Στην πορεία, στο Σχ. 6.9 παραθέτουμε το διάγραμμα της μέσης επιβράβευσης ανά επεισόδιο για τιμή γ=0.0, καθώς και ορισμένες τροχιές που διέγραψε ο πράκτορας στο διάδρομο κατά το test στο Σχ. 6.10.



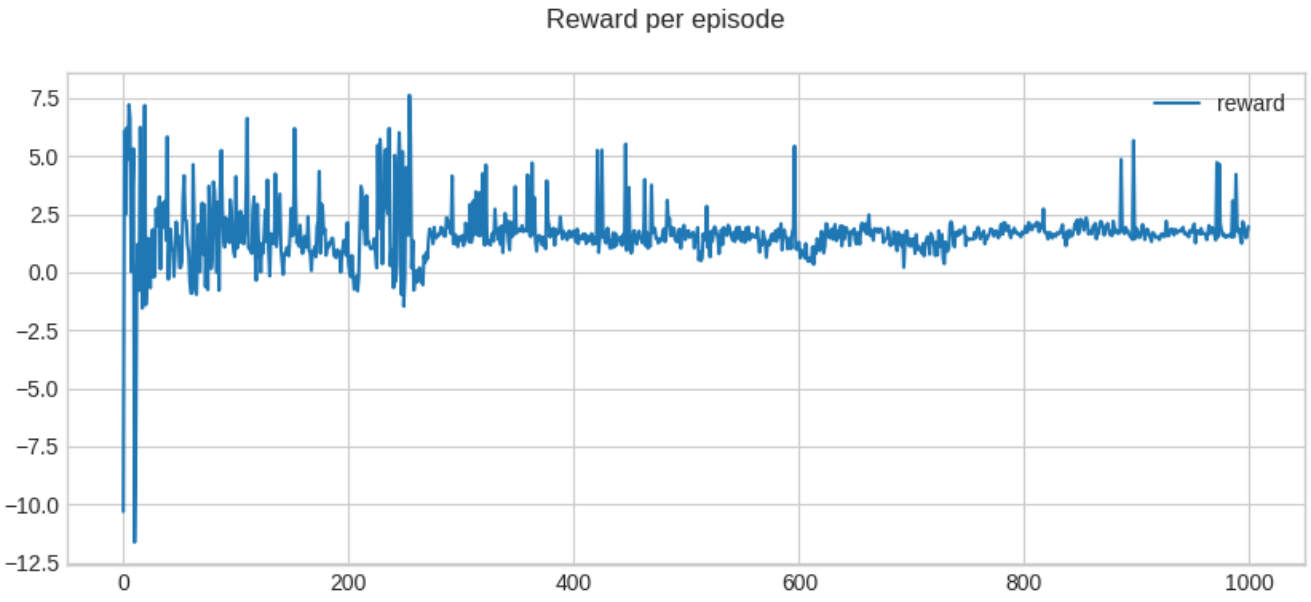
Σχήμα 6.9: Μέση επιβράβευση ανά επεισόδιο, για γ=0.0



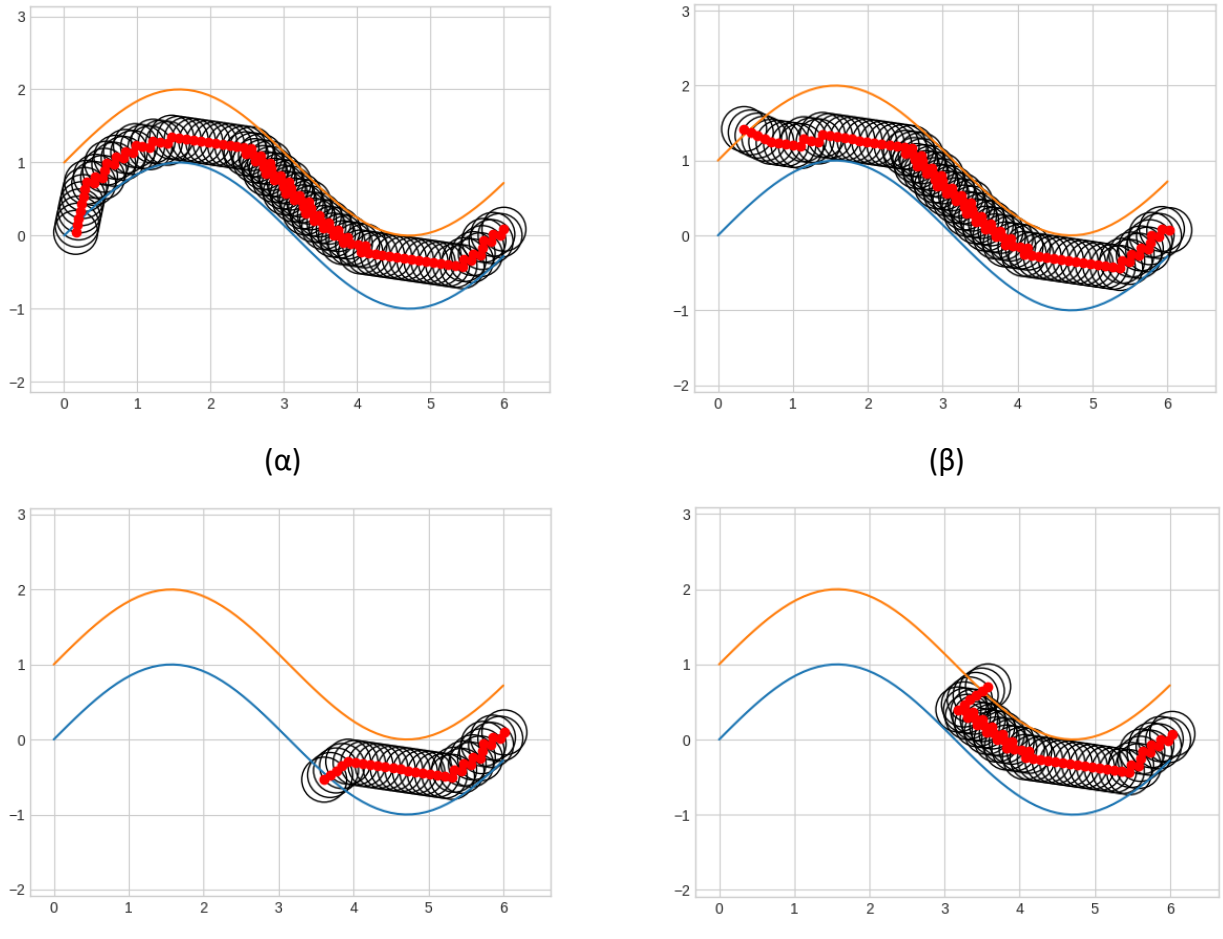
(α) (β) (γ)

Σχήμα 6.10: Πορεία πράκτορα με διαφορετικό σημείο εκκίνησης, για γ=0.0 (α) από την αρχή του διαδρόμου, (β) κάτω, κατά την πρώτη ημιπερίοδο του κάτω τοίχου, και (γ) πάνω, κατά την πρώτη ημιπερίοδο του πάνω τοίχου

Στη συνέχεια, στο Σχ. 6.11 παραθέτουμε το διάγραμμα της μέσης επιβράβευσης ανά επεισόδιο για τιμή γ=0.2, καθώς και ορισμένες τροχιές που διέγραψε ο πράκτορας στο διάδρομο κατά το test στο Σχ. 6.12.



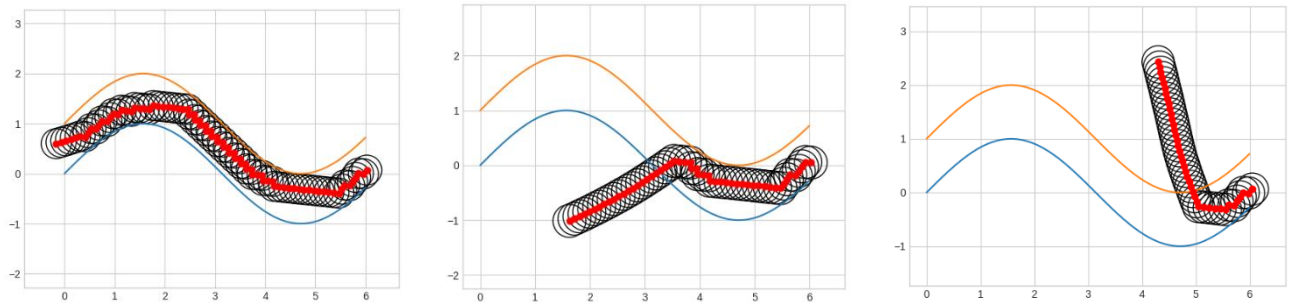
Σχήμα 6.11: Μέση επιβράβευση ανά επεισόδιο, για γ=0.2



Σχήμα 6.12: Πορεία πράκτορα με σημείο εκκίνησης εκτός του διαδρόμου, για γ=0.2 (α) κάτω από την πρώτη ημιπερίοδο του κάτω τοίχου, (β) πάνω, πριν από την πρώτη ημιπεριόδου του πάνω τοίχου, (γ) κάτω, μετά το τέλος της πρώτης ημιπεριόδου του κάτω τοίχου, και (δ) πάνω, μετά το τέλος της πρώτης ημιπεριόδου του πάνω τοίχου

Έπειτα, στο Σχ. 6.13 παραθέτουμε το διάγραμμα της μέσης επιβράβευσης ανά επεισόδιο για τιμή γ=0.5, καθώς και ορισμένες τροχιές που διέγραψε ο πράκτορας στο διάδρομο κατά το test στο Σχ. 6.14.

Σχήμα 6.13: Μέση επιβράβευση ανά επεισόδιο, για γ=0.5



(α) (β) (γ)

Σχήμα 6.14: Πορεία πράκτορα με διαφορετικό σημείο εκκίνησης, για γ=0.5 (α) από την αρχή του διαδρόμου, (β) κάτω, κατά την πρώτη ημιπερίοδο του κάτω τοίχου, και (γ) πάνω, κατά την δεύτερη ημιπερίοδο του πάνω τοίχου

Τέλος, παρουσιάζουμε το μέσο reward ανά επεισόδιο και την πορεία του πράκτορα στο διάδρομο, στην περίπτωση που πραγματοποιούμε ϵ-greedy exploration αντί για Gaussian, με $\gamma = 0.5$. Έχουμε υλοποιήσει το ϵ να ξεκινάει από την τιμή 1 και να μειώνεται εκθετικά με έναν παράγοντα 0.987 μέχρι να φτάσει την τιμή 0.001 στο επεισόδιο 500. Η παραπάνω μείωση υποδεικνύει ότι στο πρώτο επεισόδιο κάνουμε εξερεύνηση με πιθανότητα 100%, ενώ στο 500-οστό επιλέγεται μια τυχαία δράση με πιθανότητα 0.001. Τα εν λόγω αποτελέσματα παρουσιάζονται στα σχήματα 6.15. και 6.16..

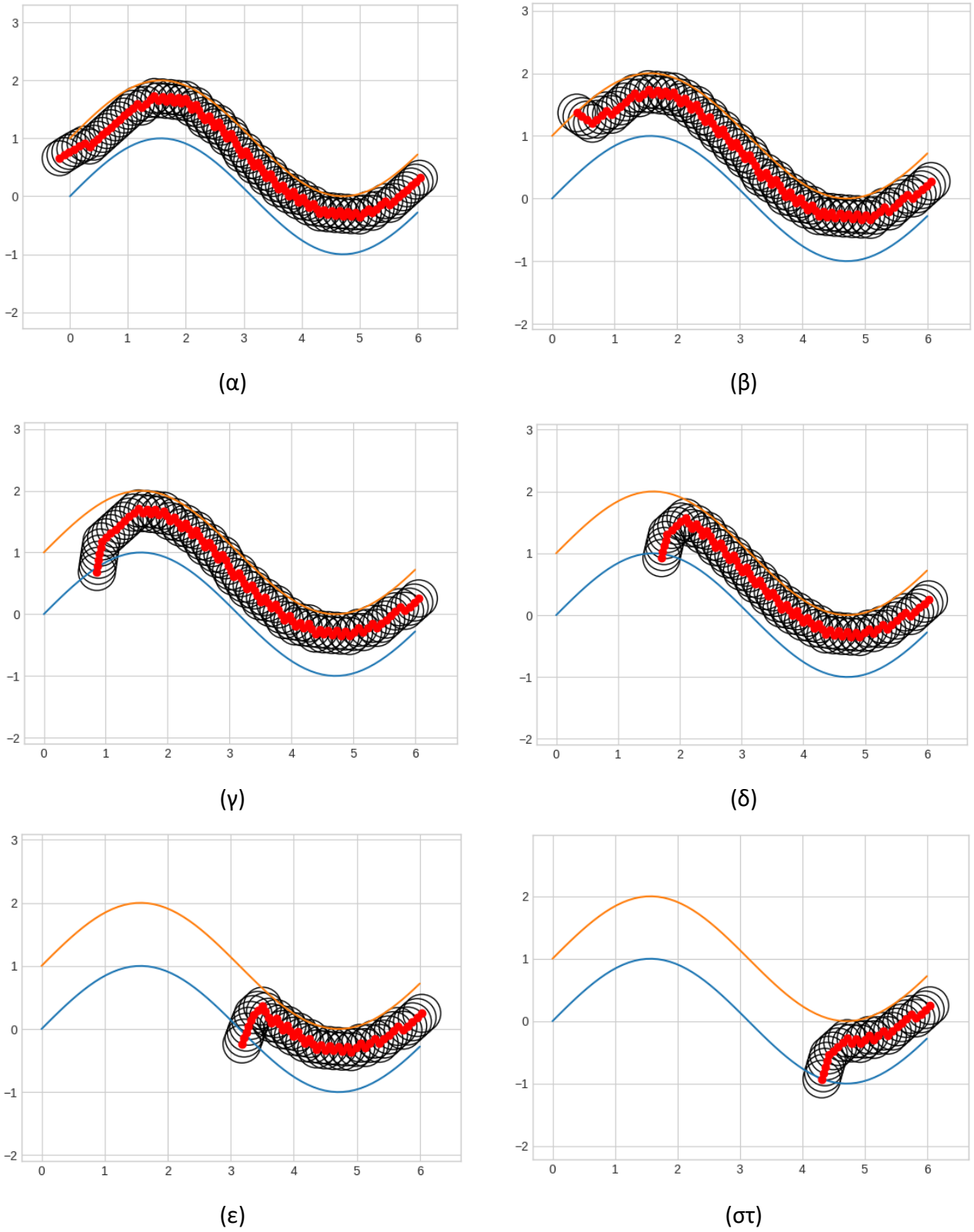
Σχήμα 6.15: Μέση επιβράβευση ανά επεισόδιο, για γ=0.5, με χρήση ϵ-greedy exploration

Από τις παραπάνω γραφικές παραστάσεις συμπεραίνουμε πως το μέσο reward αυξάνεται καθώς προχωράει η διαδικασία της εκπαίδευσης. Στη συνέχεια, μάλιστα, είναι εμφανής η σταθεροποίησή του σε μια θετική τιμή επιβράβευσης. Πιο συγκεκριμένα, στην περίπτωση της Gaussian exploration με γ=0.0 (Σχ. 6.9.) η σταθεροποίηση γίνεται γύρω από την τιμή 2.5, ενώ παρατηρούμε μια σχετική αστάθεια, με συχνές εμφανίσεις μεγαλύτερων επιβραβεύσεων. Στην περίπτωση όπου γ=0.2 (Σχ. 6.11.), παρατηρούμε και πάλι σταθεροποίηση γύρω από την ίδια τιμή, 2.5, εντούτοις ο θόρυβος γύρω από το σημείο σύγκλισης είναι αισθητά μειωμένος και η σύγκλιση πολύ πιο σταθερή. Τέλος, στην περίπτωση όπου γ=0.5 (Σχ. 6.13), παρατηρούμε και πάλι θόρυβο γύρω από την περιοχή σύγκλισης, 2.5. Σε κάθε μια εκ των παραπάνω περιπτώσεων, παρόλα τα αιχμηρά μέγιστα και το θόρυβο που συχνά παρατηρούμε στη σύγκλιση του reward, η συμπεριφορά του είναι σταθερή για κάθε τιμή του γ, με καλύτερα αποτελέσματα στην περίπτωση που γ=0.2. Τα αιχμηρά μέγιστα αιτιολογούνται στις περιπτώσεις που το αντικείμενο βρίσκεται αποκλειστικά μέσα στο διάδρομο και άρα δεν του ασκείται καμία δύναμη, και λόγω της μη ταχείας μείωσης του exploration factor, δεν επιλέγει πάντα την ίδια δράση. Αυτό έχει ως αποτέλεσμα, ορισμένες φορές να αποφεύγει τους τοίχους (υψηλότερο reward) και άλλες να έρχεται σε επαφή με αυτούς και στη συνέχεια να κινείται σχεδόν εφαπτομενικά με αυτούς, με εναλλάξ μέσα-έξω βήματα (χαμηλότερο reward).

Αναφορικά στην ϵ-greedy πολιτική εξερεύνησης, παρατηρούμε πολύ καλύτερη σύγκλιση του reward, χωρίς θόρυβο και αιχμηρά μέγιστα. Μάλιστα, η σύγκλιση της επιβράβευσης είναι γύρω από υψηλότερη τιμή, 12.5.

Αναφορικά στις τροχιές που διαγράφει ο πράκτορας, είναι όλες αποτελεσματικές και επιτυχείς, δεδομένου ότι καταφέρνει πάντα να εξέλθει του διαδρόμου, ανεξάρτητα με το σημείο εκκίνησής του. Μάλιστα, παρατηρούμε επιθυμητή συμπεριφορά ακόμα και στις περιπτώσεις που εκκινεί από

σημεία στα οποία δεν έχει εκπαιδευτεί, δεδομένου ότι κατά το train δεν του επιτρέπεται να πραγματοποιήσει διείσδυση στον τοίχο, μεγαλύτερη από 1.5.



(α) (β)

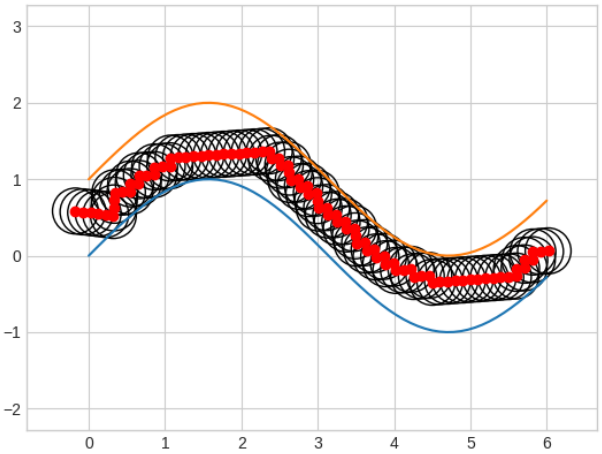
(γ) (δ)

(ε) (στ)

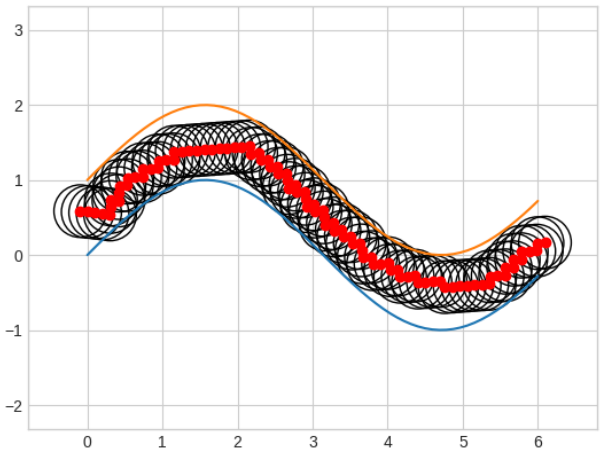
Σχήμα 6.16: Πορεία πράκτορα με διαφορετικό σημείο εκκίνησης, για γ=0.5 και ϵ-greedy exploration (α) από την αρχή του διαδρόμου, (β) πάνω από την είσοδο του διαδρόμου, (γ) κάτω,

από την αρχή της πρώτης ημιπεριόδου του κάτω τοίχου, (δ) κάτω από το μέσο της πρώτης ημιπεριόδου του κάτω τοίχου, (ε) κάτω από την αρχή της δεύτερης ημιπεριόδου του κάτω τοίχου και (στ) κάτω από το μέσο της δεύτερης ημιπεριόδου του κάτω τοίχου

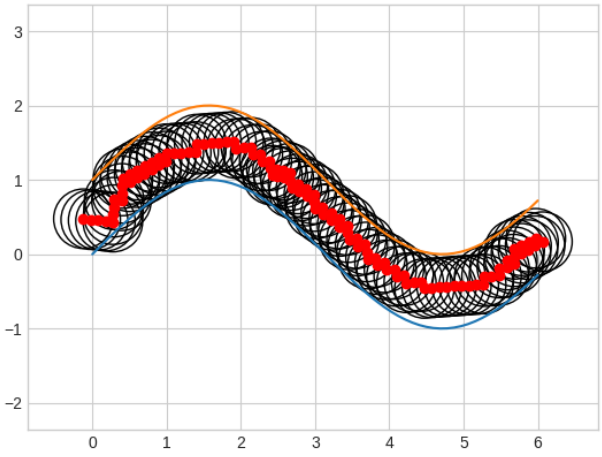
Τέλος, στο Σχ. 6.17. παραθέτουμε ορισμένες τροχιές του πράκτορα, από τη διαδικασία του test του, σε διαφορετικά είδη διαδρόμων και με διαφορετικές ακτίνες για το κυκλικό αντικείμενο, για να μπορέσουμε να αξιολογήσουμε τη δυνατότητα γενίκευσης του αλγορίθμου.



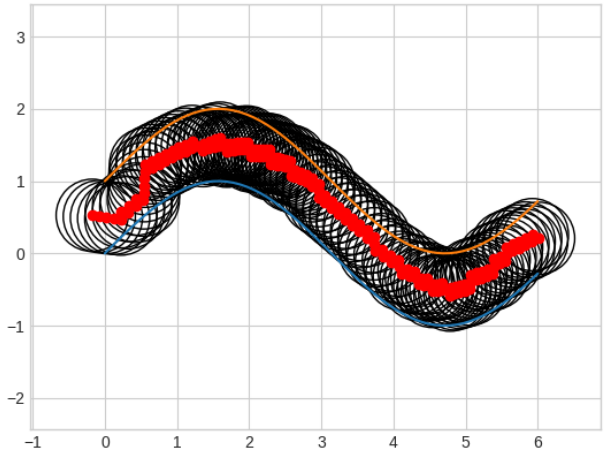
(α)



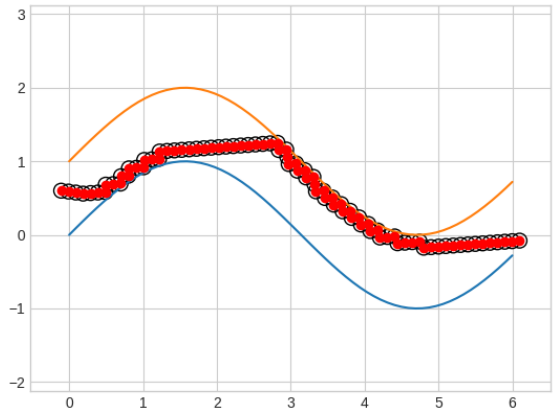
(β)



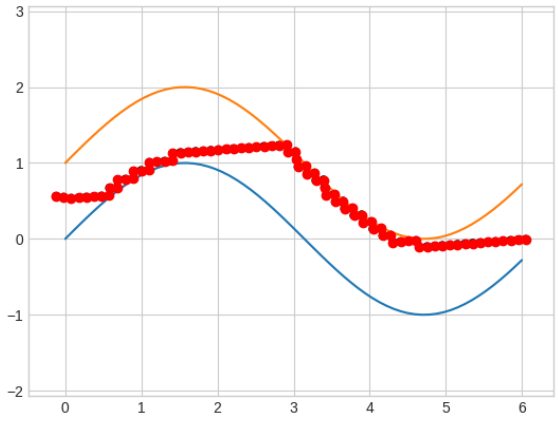
(γ)



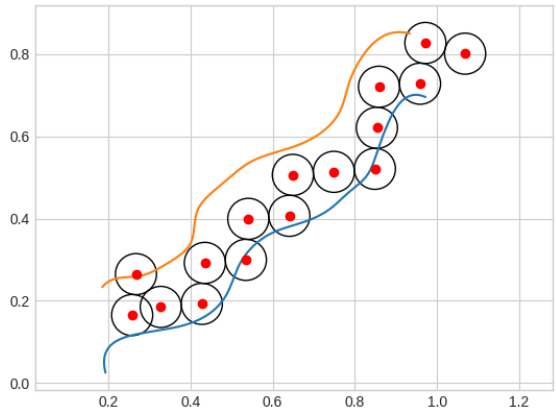
(δ)



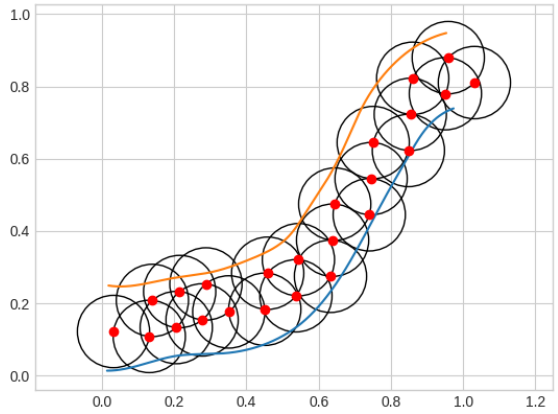
(ε)



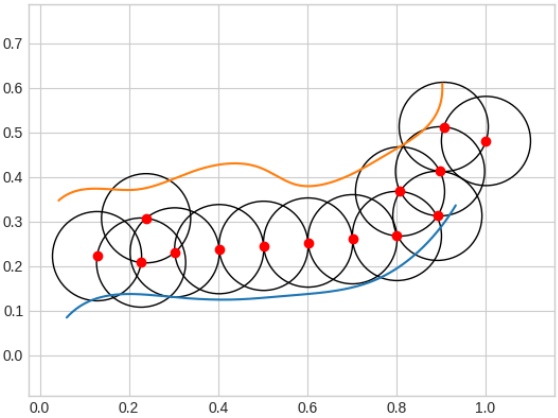
(στ)



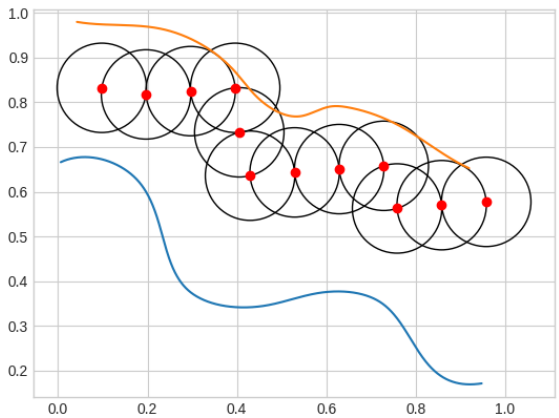
(ζ)



(η)



(θ)



(ι)

Σήμα 6.17: Τροχιές του πράκτορα για έλεγχο γενίκευσης της εκπαιδευτικής διαδικασίας: στο διάδρομο εκπαίδευσης με ακτίνες (α) 0.3, (β) 0.35, (γ) 0.4, (δ) 0.5, (ε) 0.1, (στ) 0.05 και σε νέο διάδρομο με ακτίνες (ζ) 0.05, (η) 0.1, (θ) 0.1, (ι) 0.1

Είναι εμφανές από τις παραπάνω γραφικές απεικονίσεις ότι η γενίκευση της εκπαιδευτικής διαδικασίας είναι επιτυχής τόσο σε νέα περιβάλλοντα, όσο και με τη χρήση νέου αντικειμένου.

Κεφάλαιο 7

Επίλογος

7.1 Συμπεράσματα

Στην παρούσα εργασία προσεγγίσαμε το πρόβλημα της in-hand επιδέξιας ρομποτικής λαβής με την επίλυση του προβλήματος της εξόδου ενός αντικειμένου από έναν διάδρομο πεπερασμένου μήκους, με μοναδική γνώση της δύναμης επαναφοράς που δεχόταν το αντικείμενο από τους τοίχους του διαδρόμου.

Από τις γραφικές απεικονίσεις που παρατέθηκαν στο κεφάλαιο 6, είναι εμφανής η αποτελεσματικότητα του αλγορίθμου CACLA. Παρατηρούμε πως μέσα σε έναν εύλογο αριθμό εποχών, ο πράκτορας εκπαιδεύεται ικανοποιητικά στον δοθέντα διάδρομο, ενώ παρουσιάζει εξίσου καλά αποτελέσματα γενίκευσης και σε άλλα είδη διαδρόμων. Επίσης, αποκτά ένα αντιπροσωπευτικό μοντέλο της δυναμικής του περιβάλλοντος, δεδομένου ότι ακόμα και όταν βρίσκεται σε σημεία του χώρου που δεν έχει επισκεφθεί ξανά, οι αποφάσεις που παίρνει είναι οι σωστές και στοχεύουν στη μεγιστοποίηση της συνάρτησης αξίας. Τέλος, η παραπάνω παρατήρηση υποδεικνύει και την καλή εξερεύνηση που έχει διεξαγάγει ο πράκτορας.

7.2 Προτεινόμενη μελλοντική επέκταση

Το πρόβλημα που μελετήσαμε στην εν λόγω εργασία αποτελεί απλούστευση του συνολικού, του οποίου η αντιμετώπιση είναι το επόμενο βήμα των πειραμάτων μας. Πιο συγκεκριμένα, πραγματοποιώντας την αντιστοίχιση από τη μια διάταξη στην άλλη, μπορούμε να εκπαιδεύσουμε ένα ανθρωπομορφικό ρομποτικό χέρι (όπως αυτό που θα μελετήσουμε στο Παράρτημα), έτσι ώστε να φέρει εις πέρας μια in-hand κίνηση που απαιτεί επιδέξιο ρομποτικό χειρισμό. Επομένως, μια προτεινόμενη επέκταση θα μπορούσε να είναι η προσαρμογή του αλγορίθμου που παρουσιάσαμε στον πραγματικό κόσμο, με στόχο την εκπαίδευση ενός συγκεκριμένου ρομποτικού χειριστή.

Επιπλέον, πέραν της χρήσης της απλής ενισχυτικής μάθησης για την επίτευξη της επιδέξιας ρομποτικής λαβής, μπορούμε να χρησιμοποιήσουμε ενισχυτική μάθηση, με δεδομένα επίδειξης. Τα εν λόγω δεδομένα μπορούν είτε να συλλεγούν από κάποια πιθανή πλατφόρμα δεδομένων, είτε μπορούν να παραχθούν σε περιβάλλον προσομοίωσης. Εναλλακτικά, μπορούν να παραχθούν με τη χρήση κάποιου ειδικά προσαρμοσμένου γαντιού, εφοδιασμένου με τους κατάλληλους αισθητήρες (tactile sensorized glove) [57], το οποίο να εκτελεί υποδειγματικά ορισμένες λαβές και να συγκρατεί ενδεικτικές τιμές για τις δυνάμεις που ασκούνται, τη ροπή, τη μετατόπιση και τον προσανατολισμό του αντικειμένου. Ενσωματώνοντας αυτά τα δεδομένα στον αλγόριθμο policy

gradient, όπως κάνει ο αλγόριθμος DAPG [58], μπορούμε να κάνουμε χρήση του transfer learning για να καταλήξουμε στην επίτευξη μιας επιδέξιας ρομποτικής λαβής.

Τέλος, όπως παρατηρήσαμε στα πειράματά μας, οι γραφικές παραστάσεις του μέσου reward ανά εποχή περιείχαν αρκετό θόρυβο, ενώ η σύγκλιση σε μια τιμή δεν ήταν επαρκώς εμφανής σε κάθε περίπτωση. Ως εκ τούτου, θα μπορούσαν να χρησιμοποιηθούν διαφορετικά είδη νευρωνικών δικτύων (μέθοδοι Deep RL με deep CNN) για την υλοποίηση του Actor και του Critic, με στόχο να γίνουν πιο ευσταθείς οι προβλέψεις τους.

Παράρτημα

1. Προέκταση

Μια προέκταση της εν λόγω μελέτης και της εφαρμογής του κεφαλαίου 5 θα μπορούσε να είναι η εφαρμογή του αλγορίθμου CACLA που περιγράψαμε στο κεφάλαιο 3, στο ρομποτικό χέρι Allegro, με σκοπό την επίτευξη μιας επιδέξιας ρομποτικής λαβής ενός αντικειμένου με in-hand τρόπο.

Πιο συγκεκριμένα, περιγράψαμε στο κεφάλαιο 5 την model-free εκπαίδευση ενός κυλινδρικού αντικειμένου – με μια προεξοχή – έτσι ώστε να βγαίνει από έναν χαραγμένο διάδρομο πεπερασμένου μήκους στο εσωτερικό ενός κυλινδρικού φλοιού.

1.1 Περιγραφή παραμέτρων

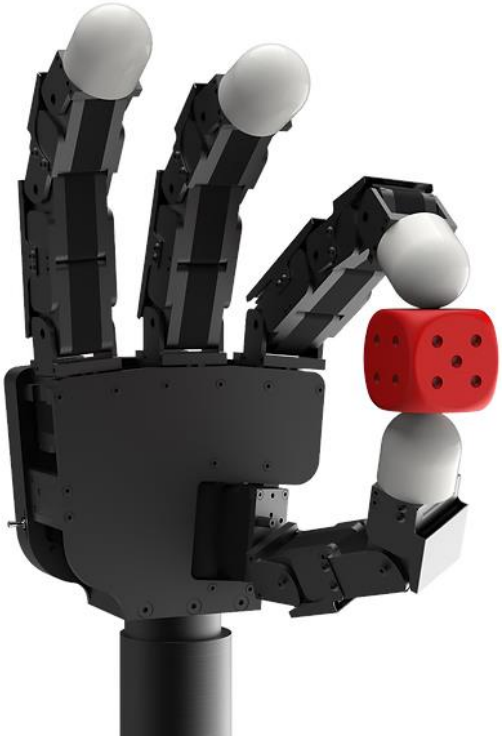
Αρχικά, θεωρούμε ότι γνωρίζουμε ποιο είναι το βέλτιστο grasp για το κυλινδρικό αντικείμενο και ασχολούμαστε μόνο με την εκπαίδευση της in-hand κίνησης. Όπως ακριβώς και στο πρόβλημα του απλού διαδρόμου, έτσι και στην επέκτασή του, χρησιμοποιούμε MDPs για την περιγραφή του. Δεν διαθέτουμε γνώση του μοντέλου του συστήματος, δηλαδή της δυναμικής του περιβάλλοντος και του ρομπότ. Η όποια γνώση αποκτούμε για το περιβάλλον, υπαγορεύεται από τους αισθητήρες που προαναφέραμε. Στο χώρο των δράσεων, θα χρησιμοποιήσουμε τους ελεγκτές θέσης που βρίσκονται στους συνδέσμους του ρομπότ. Αναφορικά στο χώρο κατάστασης, αυτός θα πρέπει να περιγράφει το περιβάλλον. Επομένως, απαραίτητη γνώση είναι οι γωνίες των αρθρώσεων των δαχτύλων που χρησιμοποιούμε, τις οποίες γνωρίζουμε μέσω των αισθητήρων στα σημεία αυτά. Επιπλέον, χρειαζόμαστε τη γνώση των αισθητήρων αφής που διαθέτει το σύστημά μας. Οι αισθητήρες των end-effectors χρησιμοποιούνται με στόχο να εξασφαλιστεί το grasp του αντικειμένου, και να ελεγχθεί η δύναμη που του ασκείται, με στόχο να μην παραμορφωθεί. Ταυτόχρονα, αποτελούν την ακριβή αντιστοίχιση του force-feedback που χρησιμοποιήσαμε στο κεφάλαιο 5. Στη συνέχεια, ορίζουμε τη συνάρτηση επιβράβευσης του συστήματος, η οποία θα έχει παρόμοια μορφή με αυτή που περιγράψαμε στο κεφάλαιο 5. Ως FAs χρησιμοποιούμε και πάλι NN με την ίδια δομή που περιγράψαμε στο κεφάλαιο 5. Τέλος, χρησιμοποιούμε Gaussian exploration γύρω από τη δράση που προβλέπει ο Actor.

2. Περιγραφή πλατφόρμας πειραμάτων

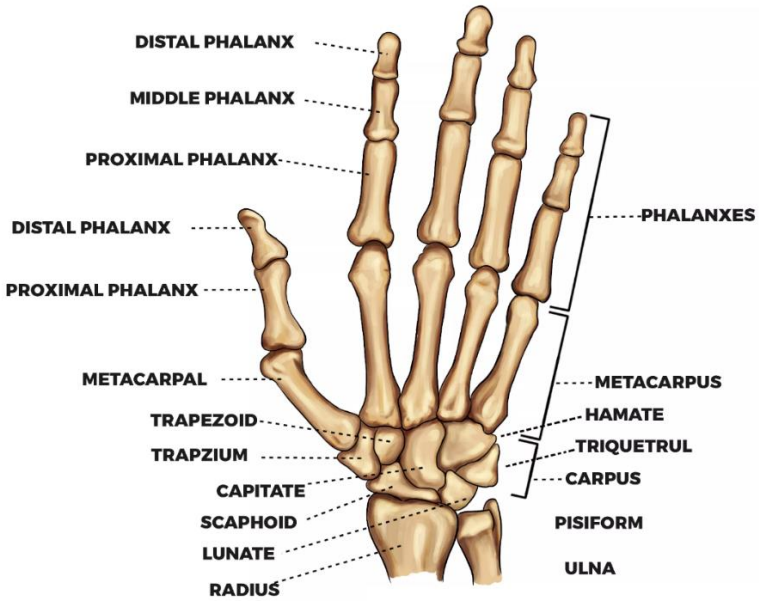
Στο εν λόγω κεφάλαιο γίνεται η περιγραφή του υλικού που χρησιμοποιήθηκε για τις προσομοιώσεις που πραγματοποιήθηκαν στα πλαίσια της εργασίας. Πιο συγκεκριμένα, θα γίνει αναλυτική περιγραφή του ρομποτικού χεριού Allegro, το οποίο χρησιμοποιήθηκε στα πειράματα της τρέχουσας μελέτης.

2.1 Το ρομποτικό χέρι Allegro

Το ρομπότ Allegro (Allegro Hand) [48] είναι ένα ανθρωπομορφικό ρομποτικό χέρι, 16 βαθμών ελευθερίας, κατασκευασμένο από την Wonik Robotics [49]. Το Allegro Hand αποτελείται από τέσσερα δάχτυλα και δεκαέξι ανεξάρτητες αρθρώσεις, ελεγχόμενες από ηλεκτρικό ρεύμα. Λόγω του χαμηλού του κόστους και της μεγάλης προσαρμοστικότητας που το χαρακτηρίζει, αποτελεί μια ιδανική πλατφόρμα τόσο για έρευνα στα πλαίσια εκτέλεσης ρομποτικών χειρισμών, όσο και για χρήση στη βιομηχανία. Η προσαρμοστικότητά του έγκειται στο ότι διαθέτει ποικιλία έτοιμων προς χρήση αλγορίθμων, χωρίς την προϋπόθεση ύπαρξης αισθητήρων, με την ικανότητα διαχείρισης αντικειμένων διαφορετικής γεωμετρίας. Επιπλέον, η πλατφόρμα του Allegro Hand υποστηρίζει τη δυνατότητα ελέγχου σε πραγματικό χρόνο, καθώς και διαδικτυακής προσομοίωσης.



Σχήμα Π.1: Το ρομποτικό χέρι Allegro Hand, από [48]



Σχήμα Π.2: Η ανατομία του ανθρώπινου χεριού [50]

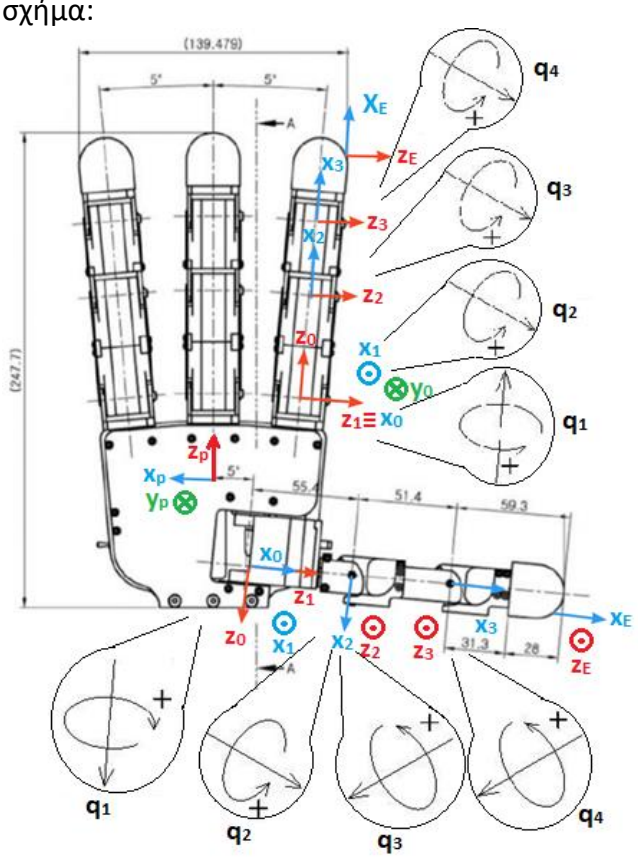
Το ρομποτικό χέρι, όπως φαίνεται και στην παραπάνω εικόνα, αποτελείται από τέσσερα δάχτυλα, εκ των οποίων, έναν αντίχειρα, έναν δείκτη, έναν μέσο και έναν μικρό. Ο αντίχειρας αποτελείται από τις εξής τέσσερις αρθρώσεις: 1 άρθρωση στη σύνδεση άπω και μεσαίας φάλαγγας, 2 αρθρώσεις στη σύνδεση μεσαίας φάλαγγας και μετακαρπίου και 1 άρθρωση στη σύνδεση μετακαρπίου και τραπεζοειδούς οστού. Εξαιρώντας των αντίχειρα, καθένα εκ των τριών λοιπών δαχτύλων του χεριού αποτελείται επίσης από τέσσερις βαθμούς ελευθερίας ως εξής: 1 άρθρωση μεταξύ εγγείας και μεσαίας φάλαγγας, 1 άρθρωση μεταξύ μεσαίας και άπω φάλαγγας και 2 αρθρώσεις μεταξύ εγγείας φάλαγγας και μετακαρπίου.

Στα πλαίσια της εν λόγω εργασίας γίνεται χρήση δύο εκ των δακτύλων του Allegro Hand, του αντίχειρα και του δείκτη.

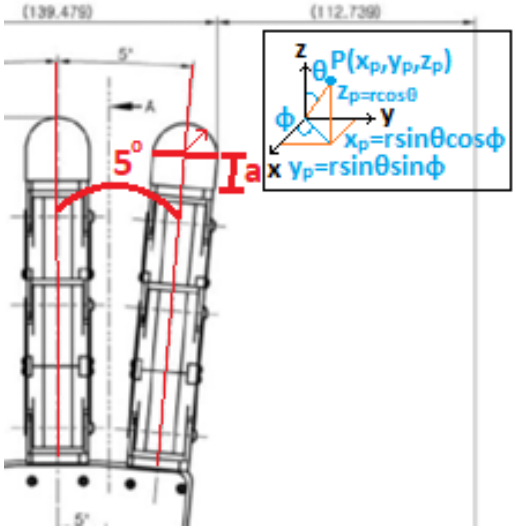
2.2 Κινηματική ανάλυση Allegro Hand

Για να επιτευχθεί ο ζητούμενος in-hand χειρισμός του αντικειμένου-στόχου από το ρομποτικό χέρι, θα πρέπει να έχει προηγηθεί η κινηματική ανάλυση του χεριού. Όπως έχει προαναφερθεί, θα γίνει χρήση δύο εκ των τεσσάρων δακτύλων του χεριού, του αντίχειρα και του δείκτη, για να πραγματοποιηθεί η λαβή και η κίνηση που μελετάται. Επομένως, η κινηματική ανάλυση θα εκπονηθεί για τα δύο αυτά δάχτυλα της διάταξης του Allegro Hand.

Επιπλέον, παρατίθεται η **<u>βέλτιστη</u>** τοποθέτηση αξόνων κατά D-H όπως φαίνεται στο παρακάτω σχήμα:



Σχήμα Π.3: Τοποθέτηση πλαισίων στο Allegro Hand βάσει της μεθόδου DH.



Σχήμα Π.4: Τοποθέτηση πλαισίων στον End-Effector του Allegro Hand βάσει της μεθόδου DH.

Παρακάτω φαίνεται ο **Πίνακας παραμέτρων της μεθόδου D-H** με βάση το Σχ.Π.3 και Π.4 για τον αντίχειρα:

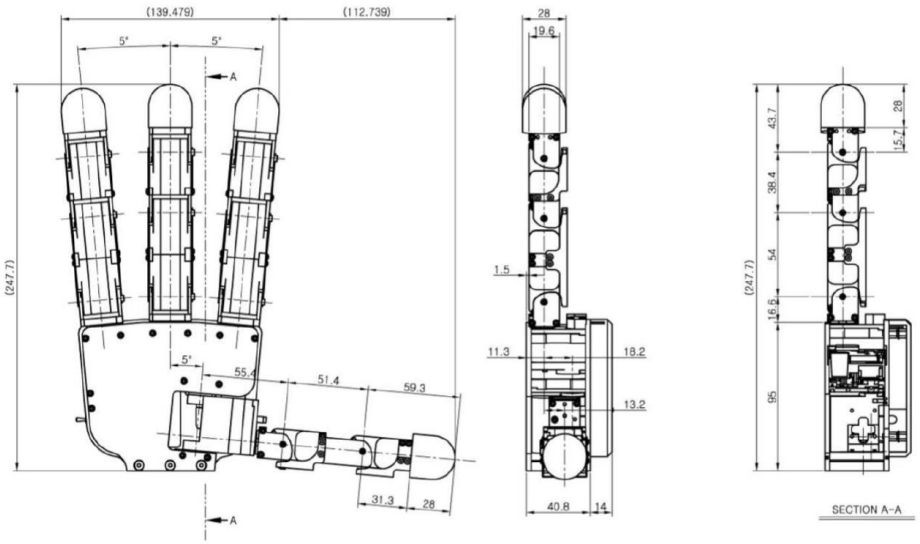
i	DH παράμετρος:	θ_i $Rotz_{i-1}$	d_i $Traz_{i-1}$	α_i $Rotx_i$	a_i $Trax_i$
1		π/2+q1	0	π/2	0
2		π/2+q2	55.4	π/2	0
3		π/2+q3	0	0	51.4
4 ≡ E		q4	r*sin(θ)*sin(ϕ)	0	31.3+a+r*cos(θ)

Πίνακας Π.1: Παράμετροι της μεθόδου DH για τον αντίχειρα του Allegro Hand.

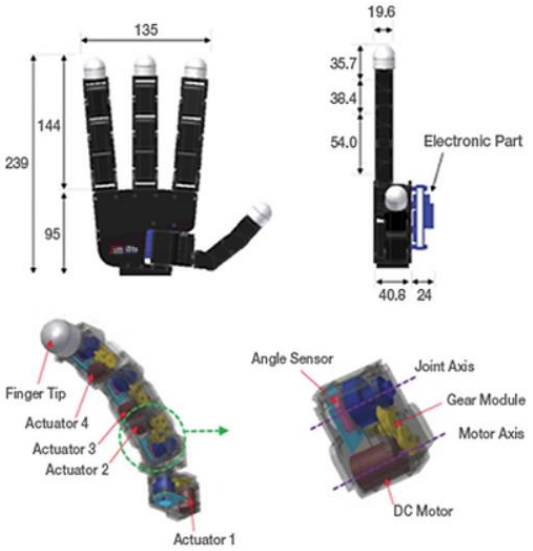
Παρακάτω φαίνεται ο **Πίνακας παραμέτρων της μεθόδου D-H** με βάση το Σχ.Π.3 και Π.4 για τον δείκτη:

i	DH παράμετρος:	θ_i $Rotz_{i-1}$	d_i $Traz_{i-1}$	α_i $Rotx_i$	a_i $Trax_i$
1		-π/2+q1	0	-π/2	0
2		-π/2+q2	0	0	54
3		q3	0	0	38.4
4 ≡ E		q4	r*sin(θ)*cos(ϕ)	0	15.7+a+r*cos(θ)

Πίνακας Π.2: Παράμετροι της μεθόδου DH για τον δείκτη του Allegro Hand.



Σχήμα Π.5: Διαστάσεις και αποστάσεις στο Allegro Hand, από [48]



Σχήμα Π.6: Ηλεκτρονικά τμήματα του Allegro Hand, από [48]

Προσδιορισμός της **<u>Ευθείας Κινηματικής Εξίσωσης</u>** της δοθείσας διάταξης:

Αρχικά θα πρέπει να υπολογίσουμε τους μετασχηματισμούς από το σύστημα του κέντρου της παλάμης προς τα συστήματα βάσης των δαχτύλων που χρησιμοποιούμε, όπως φαίνεται στο Σχ.Π.3.:

$$R_{αντίχειρα}^{παλάμη} = R_{αντίχειρα}\begin{bmatrix}1 & 0 & 0\\ 0 & -1 & 0\\ 0 & 0 & -1\end{bmatrix}\begin{bmatrix}1 & 0 & 0\\ 0 & \cos(5°) & -sin(5°)\\ 0 & sin(5°) & cos(5°)\end{bmatrix},$$

$$T_{αντίχειρα}^{παλάμη} = \begin{bmatrix} R_{αντίχειρα}^{παλάμη} & \begin{matrix}x_{πα}\\ 0\\ z_{πα}\end{matrix}\\ \begin{matrix}0 & 0 & 0\end{matrix} & 1\end{bmatrix}$$

$$R_{δείκτη}^{παλάμη} = R_{δείκτη}\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}\cos(5°) & -sin(5° & 0\\ sin(5°) & \cos(5°) & 0\\ 0 & 0 & 1\end{bmatrix}, \qquad T_{δείκτη}^{παλάμη} = \begin{bmatrix} R_{δείκτη}^{παλάμη} & \begin{matrix}x_{πδ}\\ 0\\ z_{πδ}\end{matrix}\\ \begin{matrix}0 & 0 & 0\end{matrix} & 1\end{bmatrix}$$

2.3 Ορθή κινηματική ανάλυση (κινηματική εξίσωση και γεωμετρικό μοντέλο)

(Θεωρούμε l_3=0)

Υπολογίζουμε κατ' αρχήν τις ομογενείς μήτρες των διαδοχικών μετασχηματισμών μεταξύ των πλαισίων των συνδέσμων με βάση τον παρακάτω τύπο, κάνοντας χρήση της μεθόδου D-H:

$$A_i^{i-1} = \begin{bmatrix} cos\theta_i & -sin\theta_i \cdot cosa_i & sin\theta_i \cdot sina_i & \mathrm{a}_i \cdot cos\theta_i\\ sin\theta_i & cos\theta_i \cdot cosa_i & -cos\theta_i \cdot sina_i & \mathrm{a}_i \cdot sin\theta_i\\ 0 & sina_i & cosa_i & d_i\\ 0 & 0 & 0 & 1\end{bmatrix} \qquad (1)$$

Για τον **αντίχειρα**:

$$A_1^0 = \begin{bmatrix} cos\theta_1 & -sin\theta_1 \cdot cosa_1 & sin\theta_1 \cdot sina_1 & a_1 \cdot cos\theta_1 \\ sin\theta_1 & cos\theta_1 \cdot cosa_1 & -cos\theta_1 \cdot sina_1 & a_1 \cdot sin\theta_1 \\ 0 & sina_1 & cosa_1 & d_1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \Rightarrow$$

$$A_1^0 = \begin{bmatrix} -s_1 & 0 & c_1 & 0 \\ c_1 & 0 & s_1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$A_2^1 = \begin{bmatrix} cos\theta_2 & -sin\theta_2 \cdot cosa_2 & sin\theta_2 \cdot sina_2 & a_2 \cdot cos\theta_2 \\ sin\theta_2 & cos\theta_2 \cdot cosa_2 & -cos\theta_2 \cdot sina_2 & a_2 \cdot sin\theta_2 \\ 0 & sina_2 & cosa_2 & d_2 \\ 0 & 0 & 0 & 1 \end{bmatrix} \Rightarrow$$

$$A_2^1 = \begin{bmatrix} -s_2 & 0 & c_2 & 0 \\ c_2 & 0 & s_2 & 0 \\ 0 & 1 & 0 & 55.4 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$A_3^2 = \begin{bmatrix} cos\theta_3 & -sin\theta_3 \cdot cosa_3 & sin\theta_3 \cdot sina_3 & a_3 \cdot cos\theta_3 \\ sin\theta_3 & cos\theta_3 \cdot cosa_3 & -cos\theta_3 \cdot sina_3 & a_3 \cdot sin\theta_3 \\ 0 & sina_3 & cosa_3 & d_3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \Rightarrow$$

$$A_3^2 = \begin{bmatrix} -s_3 & -c_3 & 0 & -51.4 \cdot s_3 \\ c_3 & -s_3 & 0 & 51.4 \cdot c_3 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$A_E^3 = \begin{bmatrix} cos\theta_E & -sin\theta_E \cdot cosa_E & sin\theta_E \cdot sina_E & a_E \cdot cos\theta_E \\ sin\theta_E & cos\theta_E \cdot cosa_E & -cos\theta_E \cdot sina_E & a_E \cdot sin\theta_E \\ 0 & sina_E & cosa_E & d_E \\ 0 & 0 & 0 & 1 \end{bmatrix} \Rightarrow$$

$$A_E^3 = \begin{bmatrix} c_E & -s_3 & 0 & (31.3 + a + r \cdot cos\theta) \cdot c_E \\ s_E & c_3 & 0 & (31.3 + a + r \cdot cos\theta) \cdot s_E \\ 0 & 0 & 1 & r \cdot sin\theta \cdot sin\varphi \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$A_1^0 \cdot A_2^1 = \begin{bmatrix} -s_1 & 0 & c_1 & 0 \\ c_1 & 0 & s_1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} -s_2 & 0 & c_2 & 0 \\ c_2 & 0 & s_2 & 0 \\ 0 & 1 & 0 & 55.4 \\ 0 & 0 & 0 & 1 \end{bmatrix} \Rightarrow$$

$$A_2^0 = \begin{bmatrix} s_1 \cdot s_2 & c_1 & -s_1 \cdot c_2 & 55.4 \cdot c_1 \\ -c_1 \cdot s_2 & s_1 & c_1 \cdot c_2 & 55.4 \cdot s_1 \\ c_2 & 0 & s_2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$A_2^0 \cdot A_3^2 = \begin{bmatrix} s_1 \cdot s_2 & c_1 & -s_1 \cdot c_2 & 55.4 \cdot c_1 \\ -c_1 \cdot s_2 & s_1 & c_1 \cdot c_2 & 55.4 \cdot s_1 \\ c_2 & 0 & s_2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} . \begin{bmatrix} -s_3 & -c_3 & 0 & -51.4 \cdot s_3 \\ c_3 & -s_3 & 0 & 51.4 \cdot c_3 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \Rightarrow$$

$$A_3^0 = \begin{bmatrix} c_1 \cdot c_3 - s_2 \cdot s_3 \cdot s_1 & -c_1 \cdot s_3 - s_2 \cdot c_3 \cdot s_1 & -c_2 \cdot s_1 & 55.4 \cdot c_1 + 51.4 \cdot c_1 \cdot c_3 - 51.4 \cdot s_2 \cdot s_3 \cdot s_1 \\ s_2 \cdot c_1 \cdot s_3 + c_3 \cdot s_1 & s_2 \cdot c_1 \cdot c_3 - s_3 \cdot s_1 & c_2 \cdot c_1 & 51.4 \cdot s_2 \cdot c_1 \cdot s_3 + 51.4 \cdot c_3 \cdot s_1 + 55.4 \cdot s_1 \\ -c_2 \cdot s_3 & -c_2 \cdot c_3 & s_2 & -51.4 \cdot c_2 \cdot s_3 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$A_3^0 \cdot A_E^3 = \begin{bmatrix} c_1 \cdot c_3 - s_2 \cdot s_3 \cdot s_1 & -c_1 \cdot s_3 - s_2 \cdot c_3 \cdot s_1 & -c_2 \cdot s_1 & 55.4 \cdot c_1 + 51.4 \cdot c_1 \cdot c_3 - 51.4 \cdot s_2 \cdot s_3 \cdot s_1 \\ s_2 \cdot c_1 \cdot s_3 + c_3 \cdot s_1 & s_2 \cdot c_1 \cdot c_3 - s_3 \cdot s_1 & c_2 \cdot c_1 & 51.4 \cdot s_2 \cdot c_1 \cdot s_3 + 51.4 \cdot c_3 \cdot s_1 + 55.4 \cdot s_1 \\ -c_2 \cdot s_3 & -c_2 \cdot c_3 & s_2 & -51.4 \cdot c_2 \cdot s_3 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$
$$. \begin{bmatrix} c_E & -s_3 & 0 & (31.3 + a + r \cdot cos\theta) \cdot c_E \\ s_E & c_3 & 0 & (31.3 + a + r \cdot cos\theta) \cdot s_E \\ 0 & 0 & 1 & r \cdot sin\theta \cdot sin\varphi \\ 0 & 0 & 0 & 1 \end{bmatrix} \xrightarrow{31.3+a+r \cdot cos\theta = h} r \cdot sin\theta \cdot cos\varphi = d$$

$$A_E^0 = \begin{bmatrix} c_1 \cdot c_{3E} - s_2 \cdot s_1 \cdot s_{3E} & -c_1 \cdot s_{3E} - s_2 \cdot c_{3E} \cdot s_1 & -c_2 \cdot s_1 & 55.4 \cdot c_1 + 51.4 \cdot c_1 \cdot c_3 + c_1 \cdot h \cdot c_{3E} - c_2 \cdot d \cdot s_1 - 51.4 \cdot s_2 \cdot s_3 \cdot s_1 - s_2 \cdot h \cdot s_1 \cdot s_{3E} \\ s_2 \cdot c_1 \cdot s_{3E} + s_1 \cdot c_{3E} & -s_2 \cdot c_1 \cdot c_{3E} - s_{3E} \cdot s_1 & c_2 \cdot c_1 & c_2 \cdot c_1 \cdot d + 51.4 \cdot s_2 \cdot c_1 \cdot s_3 + s_2 \cdot c_1 \cdot h \cdot s_{3E} + 51.4 \cdot c_3 \cdot s_1 + h \cdot c_{3E} \cdot s_1 + 55.4 \cdot s_1 \\ -c_2 \cdot s_{3E} & -c_2 \cdot c_{3E} & s_2 & s_2 \cdot d - 51.4 \cdot c_2 \cdot s_3 - c_2 \cdot h \cdot s_E \cdot s_2 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Για τον **δείκτη**:

$$A_1^0 = \begin{bmatrix} cos\theta_1 & -sin\theta_1 \cdot cosa_1 & sin\theta_1 \cdot sina_1 & a_1 \cdot cos\theta_1 \\ sin\theta_1 & cos\theta_1 \cdot cosa_1 & -cos\theta_1 \cdot sina_1 & a_1 \cdot sin\theta_1 \\ 0 & sina_1 & cosa_1 & d_1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \Rightarrow$$

$$A_1^0 = \begin{bmatrix} s_1 & 0 & c_1 & 0 \\ -c_1 & 0 & s_1 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$A_2^1 = \begin{bmatrix} cos\theta_2 & -sin\theta_2 \cdot cosa_2 & sin\theta_2 \cdot sina_2 & a_2 \cdot cos\theta_2 \\ sin\theta_2 & cos\theta_2 \cdot cosa_2 & -cos\theta_2 \cdot sina_2 & a_2 \cdot sin\theta_2 \\ 0 & sina_2 & cosa_2 & d_2 \\ 0 & 0 & 0 & 1 \end{bmatrix} \Rightarrow$$

$$A_2^1 = \begin{bmatrix} s_2 & c_2 & 0 & 54 \cdot s_2 \\ -c_2 & s_2 & 0 & -54 \cdot c_2 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$A_3^2 = \begin{bmatrix} cos\theta_3 & -sin\theta_3 \cdot cosa_3 & sin\theta_3 \cdot sina_3 & a_3 \cdot cos\theta_3 \\ sin\theta_3 & cos\theta_3 \cdot cosa_3 & -cos\theta_3 \cdot sina_3 & a_3 \cdot sin\theta_3 \\ 0 & sina_3 & cosa_3 & d_3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \Rightarrow$$

$$A_3^2 = \begin{bmatrix} c_3 & -s_3 & 0 & 38.4 \cdot c_3 \\ s_3 & c_3 & 0 & 38.4 \cdot s_3 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$A_E^3 = \begin{bmatrix} cos\theta_E & -sin\theta_E \cdot cosa_E & sin\theta_E \cdot sina_E & a_E \cdot cos\theta_E \\ sin\theta_E & cos\theta_E \cdot cosa_E & -cos\theta_E \cdot sina_E & a_E \cdot sin\theta_E \\ 0 & sina_E & cosa_E & d_E \\ 0 & 0 & 0 & 1 \end{bmatrix} \Rightarrow$$

$$A_E^3 = \begin{bmatrix} c_E & -s_3 & 0 & (15.7 + a + r \cdot cos\varphi) \cdot c_3 \\ s_E & c_3 & 0 & (15.7 + a + r \cdot cos\varphi) \cdot s_3 \\ 0 & 0 & 1 & r \cdot sin\theta \cdot cos\varphi \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$A_2^0 = A_1^0 \cdot A_2^1 = \begin{bmatrix} s_1 & 0 & c_1 & 0 \\ -c_1 & 0 & s_1 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} s_2 & c_2 & 0 & 54 \cdot s_2 \\ -c_2 & s_2 & 0 & -54 \cdot c_2 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} s_2 \cdot s_1 & c_2 \cdot s_1 & c_1 & 54 \cdot s_2 \cdot s_1 \\ -s_2 \cdot c_1 & -c_2 \cdot c_1 & s_1 & -54 \cdot s_2 \cdot c_1 \\ c_2 & -s_2 & 0 & 54 \cdot c_2 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$A_3^0 = A_2^0 \cdot A_3^2 = \begin{bmatrix} s_2 \cdot s_1 & c_2 \cdot s_1 & c_1 & 54 \cdot s_2 \cdot s_1 \\ -s_2 \cdot c_1 & -c_2 \cdot c_1 & s_1 & -54 \cdot s_2 \cdot c_1 \\ c_2 & -s_2 & 0 & 54 \cdot c_2 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} c_3 & -s_3 & 0 & 38.4 \cdot c_3 \\ s_3 & c_3 & 0 & 38.4 \cdot s_3 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \Rightarrow$$

$$A_3^0 = \begin{bmatrix} s_1 \cdot s_{23} & c_{23} \cdot s_1 & c_1 & 54 \cdot s_2 \cdot s_1 + 38.4 \cdot s_{23} \cdot s_1 \\ -c_1 \cdot s_{23} & -c_1 \cdot c_{23} & s_1 & -54 \cdot s_2 \cdot c_1 - 38.4 \cdot s_{23} \cdot c_1 \\ c_{23} & -s_{23} & 0 & 54 \cdot c_2 + 38.4 \cdot c_{23} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$A_3^0 \cdot A_E^3 = \begin{bmatrix} s_1 \cdot s_{23} & c_{23} \cdot s_1 & c_1 & 54 \cdot s_2 \cdot s_1 + 38.4 \cdot s_{23} \cdot s_1 \\ -c_1 \cdot s_{23} & -c_1 \cdot c_{23} & s_1 & -54 \cdot s_2 \cdot c_1 - 38.4 \cdot s_{23} \cdot c_1 \\ c_{23} & -s_{23} & 0 & 54 \cdot c_2 + 38.4 \cdot c_{23} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\cdot \begin{bmatrix} c_E & -s_3 & 0 & (15.7 + a + r \cdot cos\varphi) \cdot c_3 \\ s_E & c_3 & 0 & (15.7 + a + r \cdot cos\varphi) \cdot s_3 \\ 0 & 0 & 1 & r \cdot cosw \\ 0 & 0 & 0 & 1 \end{bmatrix} \xrightarrow{15.7+a+r\cdot cos\theta=g} r \cdot sin\theta \cdot sin\varphi = f$$

$$A_E^0 = \begin{bmatrix} s_1 \cdot s_{23E} & s_1 \cdot c_{23E} & c_1 & c_1 \cdot f + 54 \cdot s_2 \cdot s_1 + g \cdot s_1 \cdot s_{23E} + 38.4 \cdot s_1 \cdot s_{23} \\ -c_1 \cdot s_{23E} & c_1 \cdot s_{23E} & s_1 & -54 \cdot s_2 \cdot c_1 + f \cdot s_1 - 38.4 \cdot c_1 \cdot s_{23} - c_1 \cdot g \cdot s_{23E} \\ c_{23E} & -s_{23E} & 0 & 54 \cdot c_2 + g \cdot c_{23E} + 38.4 \cdot c_{23} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Το <u>διάνυσμα θέσης</u> του Τ.Ε.Δ. εκφρασμένο σε καρτεσιανές συντεταγμένες $\boldsymbol{p_E}=[p_{Ex}\ p_{Ey}\ p_{Ez}]^T$ ως προς το πλαίσιο αναφοράς της βάσης είναι επομένως:

Για τον αντίχειρα:

$$\boldsymbol{p_E}(q_1,q_2,q_3,q_4)=\begin{bmatrix} p_{Ex} \\ p_{Ey} \\ p_{Ez} \end{bmatrix}=\begin{bmatrix} 55.4 \cdot c_1 + 51.4 \cdot c_1 \cdot c_3 + c_1 \cdot h \cdot c_{3E} - c_2 \cdot d \cdot s_1 - 51.4 \cdot s_2 \cdot s_3 \cdot s_1 - s_2 \cdot h \cdot s_1 \cdot s_{3E} \\ c_2 \cdot c_1 \cdot d + 51.4 \cdot s_2 \cdot c_1 \cdot s_3 + s_2 \cdot c_1 \cdot h \cdot s_{3E} + 51.4 \cdot c_3 \cdot s_1 + h \cdot c_{3E} \cdot s_1 + 55.4 \cdot s_1 \\ s_2 \cdot d - 51.4 \cdot c_2 \cdot s_3 - c_2 \cdot h \cdot s_E \cdot s_2 \end{bmatrix}$$

Για τον δείκτη:

$$\boldsymbol{p_E}(q_1,q_2,q_3,q_4) = \begin{bmatrix} p_{Ex} \\ p_{Ey} \\ p_{Ez} \end{bmatrix}=\begin{bmatrix} c_1 \cdot f + 54 \cdot s_2 \cdot s_1 + g \cdot s_1 \cdot s_{23E} + 38.4 \cdot s_1 \cdot s_{23} \\ -54 \cdot s_2 \cdot c_1 + f \cdot s_1 - 38.4 \cdot c_1 \cdot s_{23} - c_1 \cdot g \cdot s_{23E} \\ 54 \cdot c_2 + g \cdot c_{23E} + 38.4 \cdot c_{23} \end{bmatrix}.$$

Ο <u>προσανατολισμός</u> του Τ.Ε.Δ., ως προς το πλαίσιο αναφοράς της βάσης του αντίχειρα μπορεί να περιγραφεί από την 3x3 μήτρα στροφής:

$$R_E^0(q) = \begin{bmatrix} c_1 \cdot c_{3E} - s_2 \cdot s_1 \cdot s_{3E} & -c_1 \cdot s_{3E} - s_2 \cdot c_{3E} \cdot s_1 & -c_2 \cdot s_1 \\ s_2 \cdot c_1 \cdot s_{3E} + s_1 \cdot c_{3E} & -s_2 \cdot c_1 \cdot c_{3E} - s_{3E} \cdot s_1 & c_2 \cdot c_1 \\ -c_2 \cdot s_{3E} & -c_2 \cdot c_{3E} & s_2 \end{bmatrix}.$$

Ο <u>προσανατολισμός</u> του Τ.Ε.Δ., ως προς το πλαίσιο αναφοράς της βάσης του δείκτη μπορεί να περιγραφεί από την 3x3 μήτρα στροφής: $R_E^0(q) = \begin{bmatrix} s_1 \cdot s_{23E} & s_1 \cdot c_{23E} & c_1 \\ -c_1 \cdot s_{23E} & c_1 \cdot s_{23E} & s_1 \\ c_{23E} & -s_{23E} & 0 \end{bmatrix}$.

Όπου έχουν χρησιμοποιηθεί οι εξής συμβολισμοί:

- $cosq_{ij} = \cos(q_i + q_j)$
- $sinq_{ij} = \sin(q_i + q_j)$
- $\cos(a \pm b) = cosa \cdot cosb \mp sina \cdot sinb$
- $\sin(a \pm b) = sina \cdot cosb \pm cosa \cdot sinb$

Προσδιορισμός της **<u>Ιακωβιανής μήτρας</u>** της δοθείσας διάταξης:

2.4 Ορθή ευθεία διαφορική κινηματική ανάλυση (Ιακωβιανή Μήτρα)

Έστω ότι το ρομπότ βρίσκεται σε δοσμένη διάταξη: $\boldsymbol{q}$=[q_1 q_2 q_3]^T. Έστω επίσης ότι γνωρίζουμε το διάνυσμα των ταχυτήτων των αρθρώσεων $\dot{\boldsymbol{q}} = [\dot{q}_1\ \dot{q}_2\ \dot{q}_3\ ]^T$. Θα έχουμε επομένως ως γνωστόν για τη γραμμική ταχύτητα του Τ.Ε.Δ.: $\boldsymbol{v_E} = [\dot{p_{Ex}}\ \dot{p_{Ey}}\ \dot{p_{Ez}}\ ]^T = \boldsymbol{J}(\boldsymbol{q}) \cdot \dot{\boldsymbol{q}}$, όπου $\boldsymbol{J}(\boldsymbol{q})$ η 6x12 Ιακωβιανή μήτρα $\boldsymbol{J_L}$(για γραμμική ταχύτητα του Τ.Ε.Δ.) για την οποία έχουμε κατά τα γνωστά: $\boldsymbol{J_L}$ = [J_{L1} J_{L2} J_{L3} J_{LE}]^T, όπου J_{Li} (i=1,2,3,E) τα 3x1 διανύσματα στήλης της Ιακωβιανής, που δηλώνουν τη συνεισφορά κάθε βαθμού ελευθερίας *i* στη γραμμική ταχύτητα του Τ.Ε.Δ. του ρομπότ.

Για την κάθε άρθρωση του αντίχειρα έχουμε:

- *<u>Άρθρωση -1</u> (στροφική άρθρωση):* $\overrightarrow{\boldsymbol{J_{A_1}}} = \widehat{\boldsymbol{b_0}} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$,

$$\overrightarrow{\boldsymbol{J_{L_1}}} = \frac{\partial \overrightarrow{p_E(q_1,q_2,q_3,q_E)}}{\partial \boldsymbol{q_1}} = \begin{bmatrix} -55.4 \cdot s_1 - 51.4 \cdot s_1 \cdot c_3 - s_1 \cdot h \cdot c_{3E} - c_2 \cdot d \cdot c_1 - 51.4 \cdot s_2 \cdot s_3 \cdot c_1 - s_2 \cdot h \cdot c_1 \cdot s_{3E} \\ -c_2 \cdot s_1 \cdot d - 51.4 \cdot s_2 \cdot s_1 \cdot s_3 - s_2 \cdot s_1 \cdot h \cdot s_{3E} + 51.4 \cdot c_3 \cdot c_1 + h \cdot c_{3E} \cdot c_1 + 55.4 \cdot c_1 \\ 0 \end{bmatrix}.$$

- *<u>Άρθρωση -2</u> (στροφική άρθρωση):* $\overrightarrow{\boldsymbol{J_{A_2}}} = \widehat{\boldsymbol{b_1}} = A_2^0[1{:}3,3] = \begin{bmatrix} 55.4 \cdot c_1 \\ 55.4 \cdot s_1 \\ 0 \end{bmatrix}$,

$$\overrightarrow{\boldsymbol{J_{L_2}}} = \frac{\partial \overrightarrow{p_E(q_1,q_2,q_3,q_E)}}{\partial \boldsymbol{q_2}} = \begin{bmatrix} s_2 \cdot d \cdot s_1 - 51.4 \cdot c_2 \cdot s_3 \cdot s_1 - c_2 \cdot h \cdot s_1 \cdot s_{3E} \\ -s_2 \cdot c_1 \cdot d + 51.4 \cdot c_2 \cdot c_1 \cdot s_3 + c_2 \cdot c_1 \cdot h \cdot s_{3E} \\ c_2 \cdot d + 51.4 \cdot s_2 \cdot s_3 - c_2 \cdot h \cdot s_E \cdot c_2 \end{bmatrix}.$$

- *<u>Άρθρωση -3</u> (στροφική άρθρωση):* $\overrightarrow{\boldsymbol{J_{A_3}}} = \widehat{\boldsymbol{b_2}} = A_3^0[1{:}3,3] = \begin{bmatrix} 55.4 \cdot c_1 + 51.4 \cdot c_1 \cdot c_3 - 51.4 \cdot s_2 \cdot s_3 \cdot s_1 \\ 51.4 \cdot s_2 \cdot c_1 \cdot s_3 + 51.4 \cdot c_3 \cdot s_1 + 55.4 \cdot s_1 \\ -51.4 \cdot c_2 \cdot s_3 \end{bmatrix}$,

$$\overrightarrow{\boldsymbol{J_{L_3}}} = \frac{\partial \overrightarrow{p_E(q_1,q_2,q_3,q_E)}}{\partial \boldsymbol{q_3}} = \begin{bmatrix} -51.4 \cdot c_1 \cdot s_3 - c_1 \cdot h \cdot s_{3E} - 51.4 \cdot s_2 \cdot c_3 \cdot s_1 - s_2 \cdot h \cdot s_1 \cdot c_{3E} \\ 51.4 \cdot s_2 \cdot c_1 \cdot c_3 + s_2 \cdot c_1 \cdot h \cdot c_{3E} - 51.4 \cdot s_3 \cdot s_1 - h \cdot s_{3E} \cdot s_1 \\ -51.4 \cdot c_2 \cdot c_3 \end{bmatrix}.$$

- *<u>Άρθρωση -4</u> (στροφική άρθρωση):*

$$\overrightarrow{\boldsymbol{J_{A_E}}} = \widehat{\boldsymbol{b_3}} = A_E^0[1{:}3,3] = \begin{bmatrix} 55.4 \cdot c_1 + 51.4 \cdot c_1 \cdot c_3 + c_1 \cdot h \cdot c_{3E} - c_2 \cdot d \cdot s_1 - 51.4 \cdot s_2 \cdot s_3 \cdot s_1 - s_2 \cdot h \cdot s_1 \cdot s_{3E} \\ c_2 \cdot c_1 \cdot d + 51.4 \cdot s_2 \cdot c_1 \cdot s_3 + s_2 \cdot c_1 \cdot h \cdot s_{3E} + 51.4 \cdot c_3 \cdot s_1 + h \cdot c_{3E} \cdot s_1 + 55.4 \cdot s_1 \\ s_2 \cdot d - 51.4 \cdot c_2 \cdot s_3 - c_2 \cdot h \cdot s_E \cdot s_2 \end{bmatrix},$$

$$\overrightarrow{\boldsymbol{J_{L_E}}} = \frac{\partial \overrightarrow{p_E(q_1,q_2,q_3,q_E)}}{\partial \boldsymbol{q_E}} = \begin{bmatrix} -c_1 \cdot h \cdot s_{3E} - s_2 \cdot h \cdot s_1 \cdot c_{3E} \\ s_2 \cdot c_1 \cdot h \cdot c_{3E} - h \cdot s_{3E} \cdot s_1 \\ -c_2 \cdot h \cdot c_E \cdot s_2 \end{bmatrix}.$$

Επομένως η Ιακωβιανή μήτρα του αντίχειρα προκύπτει ως εξής:

$$\vec{\boldsymbol{J}} = \begin{bmatrix} \overrightarrow{\boldsymbol{J_{L_1}}} & \overrightarrow{\boldsymbol{J_{L_2}}} & \overrightarrow{\boldsymbol{J_{L_3}}} & \overrightarrow{\boldsymbol{J_{L_E}}} \\ \overrightarrow{\boldsymbol{J_{A_1}}} & \overrightarrow{\boldsymbol{J_{A_2}}} & \overrightarrow{\boldsymbol{J_{A_3}}} & \overrightarrow{\boldsymbol{J_{A_E}}} \end{bmatrix}$$

Για την κάθε άρθρωση του δείκτη έχουμε:

- *<u>Άρθρωση -1</u> (στροφική άρθρωση):* $\overrightarrow{\boldsymbol{J_{A_1}}} = \widehat{\boldsymbol{b_0}} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$,

  $\overrightarrow{\boldsymbol{J_{L_1}}} = \frac{\boldsymbol{\partial \overrightarrow{p_E(q_1,q_2,q_3,q_E)}}}{\boldsymbol{\partial q_1}} = \begin{bmatrix} c_1 \cdot s_{23E} & c_1 \cdot c_{23E} & -s_1 \\ s_1 \cdot s_{23E} & -s_1 \cdot s_{23E} & c_1 \\ 0 & 0 & 0 \end{bmatrix}$.

- *<u>Άρθρωση -2</u> (στροφική άρθρωση):* $\overrightarrow{\boldsymbol{J_{A_2}}} = \widehat{\boldsymbol{b_1}} = A_2^0[1{:}3,3] = \begin{bmatrix} 54 \cdot s_2 \cdot s_1 \\ -54 \cdot s_2 \cdot c_1 \\ 54 \cdot c_2 \end{bmatrix}$,

  $\overrightarrow{\boldsymbol{J_{L_2}}} = \frac{\boldsymbol{\partial \overrightarrow{p_E(q_1,q_2,q_3,q_E)}}}{\boldsymbol{\partial q_2}} = \begin{bmatrix} s_1 \cdot c_{23E} & -s_1 \cdot s_{23E} & 0 \\ -c_1 \cdot c_{23E} & c_1 \cdot c_{23E} & 0 \\ -s_{23E} & -c_{23E} & 0 \end{bmatrix}$.

- *<u>Άρθρωση -3</u> (στροφική άρθρωση):* $\overrightarrow{\boldsymbol{J_{A_3}}} = \widehat{\boldsymbol{b_2}} = A_3^0[1{:}3,3] = \begin{bmatrix} 54 \cdot s_2 \cdot s_1 + 38.4 \cdot s_{23} \cdot s_1 \\ -54 \cdot s_2 \cdot c_1 - 38.4 \cdot s_{23} \cdot c_1 \\ 54 \cdot c_2 + 38.4 \cdot c_{23} \end{bmatrix}$,

  $\overrightarrow{\boldsymbol{J_{L_3}}} = \frac{\boldsymbol{\partial \overrightarrow{p_E(q_1,q_2,q_3,q_E)}}}{\boldsymbol{\partial q_3}} = \begin{bmatrix} s_1 \cdot c_{23E} & -s_1 \cdot s_{23E} & 0 \\ -c_1 \cdot c_{23E} & c_1 \cdot c_{23E} & 0 \\ -s_{23E} & -c_{23E} & 0 \end{bmatrix}$.

- *<u>Άρθρωση -4</u> (στροφική άρθρωση):*

  $\overrightarrow{\boldsymbol{J_{A_E}}} = \widehat{\boldsymbol{b_3}} = A_E^0[1{:}3,3] = \begin{bmatrix} c_1 \cdot f + 54 \cdot s_2 \cdot s_1 + g \cdot s_1 \cdot s_{23E} + 38.4 \cdot s_1 \cdot s_{23} \\ -54 \cdot s_2 \cdot c_1 + f \cdot s_1 - 38.4 \cdot c_1 \cdot s_{23} - c_1 \cdot g \cdot s_{23E} \\ 54 \cdot c_2 + g \cdot c_{23E} + 38.4 \cdot c_{23} \end{bmatrix}$,

  $\overrightarrow{\boldsymbol{J_{L_E}}} = \frac{\boldsymbol{\partial \overrightarrow{p_E(q_1,q_2,q_3,q_E)}}}{\boldsymbol{\partial q_E}} = \begin{bmatrix} g \cdot s_1 \cdot c_{23E} \\ -c_1 \cdot g \cdot c_{23E} \\ -g \cdot s_{23E} \end{bmatrix}$.

Επομένως η Ιακωβιανή μήτρα του δείκτη προκύπτει ως εξής:

$$\vec{\boldsymbol{J}} = \begin{bmatrix} \overrightarrow{\boldsymbol{J_{L_1}}} & \overrightarrow{\boldsymbol{J_{L_2}}} & \overrightarrow{\boldsymbol{J_{L_3}}} & \overrightarrow{\boldsymbol{J_{L_E}}} \\ \overrightarrow{\boldsymbol{J_{A_1}}} & \overrightarrow{\boldsymbol{J_{A_2}}} & \overrightarrow{\boldsymbol{J_{A_3}}} & \overrightarrow{\boldsymbol{J_{A_E}}} \end{bmatrix}$$

Ακολουθώντας τη σύμβαση:
$\begin{cases} sinq_i = s_i \\ cosq_i = c_i \end{cases}$, i ∈ ℕ (Σ)

Το **<u>αντίστροφο διαφορικό κινηματικό μοντέλο</u>** των δαχτύλων του ρομποτικού βραχίονα ως προς τη γραμμική ταχύτητα του Τ.Ε.Δ. υπολογίζεται παρακάτω:

Έστω ότι το ρομπότ βρίσκεται σε μια δοσμένη διάταξη: $\boldsymbol{q}=[q_1\ q_2\ q_3]^T$ και γνωρίζουμε το διάνυσμα της γραμμικής ταχύτητας $\boldsymbol{v_E} = [\dot{p_{Ex}}\ \dot{p_{Ey}}\ \dot{p_{Ez}}\ ]^T$ του Τ.Ε.Δ. του, ως προς το πλαίσιο αναφοράς της βάσης.

Εάν η Ιακωβιανή μήτρα είναι αντιστρέψιμη, δηλαδή στην περίπτωση που $det(\vec{\boldsymbol{J}})\neq 0$, παίρνουμε τις ταχύτητες των αρθρώσεων:

$$\dot{\boldsymbol{q}} = \boldsymbol{J_L}^{-1} \cdot \boldsymbol{v_E} \Rightarrow \begin{bmatrix} \dot{q_1} \\ \dot{q_2} \\ \dot{q_3} \end{bmatrix} = \frac{1}{det(\overrightarrow{J_L})} \cdot adj(\overrightarrow{\boldsymbol{J_L}}) \cdot \begin{bmatrix} v_{Ex} \\ v_{Ey} \\ v_{Ez} \end{bmatrix} \quad (4.1)$$

Πρώτα γίνεται ο υπολογισμός της ορίζουσας του υποπίνακα της Ιακωβιανής μήτρας που αφορά στις γραμμικές ταχύτητες.

Έπειτα, για τον υπολογισμό του $adj(\overrightarrow{\boldsymbol{J_L}})$ θα πρέπει να προσδιορίσουμε τις υποορίζουσες του $\overrightarrow{\boldsymbol{J_L}}$.

Ο συγκεκριμένος μηχανισμός εμφανίζει **<u>ιδιόμορφες διατάξεις</u>** <u>ως προς τη γραμμική ταχύτητα του Τ.Ε.Δ.</u> στις τιμές που μηδενίζεται η $det(\overrightarrow{\boldsymbol{J_L}})$.

2.5 Περιβάλλον προσομοίωσης για το Allegro Hand

Για την εφαρμογή των αλγορίθμων μάθησης κινήσεων χειρισμού στο Allegro Hand μπορεί να χρησιμοποιηθεί το ROS (Robotic Operating System) [51]. Το ROS αποτελεί μια open-source πλατφόρμα που παρέχει βιβλιοθήκες για την δημιουργία και την προσομοίωση ρομποτικών εφαρμογών. Οι κύριες βιβλιοθήκες του ROS είναι προσανατολισμένες προς ένα σύστημα τύπου Unix, κυρίως λόγω της εξάρτησής τους από μεγάλες συλλογές λογισμικού ανοιχτού-κώδικα.

3. Μέθοδος Denavit – Hartenberg (D – H)

Σύμφωνα με τη μέθοδο **Denavit-Hartenberg (D-H)** γίνεται η τοποθέτηση των πλαισίων αναφοράς των συνδέσμων του βραχίονα. Οι κανόνες που ακολουθούνται είναι οι εξής:

- ✓ Ο άξονας z_i τοποθετείται στη διεύθυνση της i+1 άρθρωσης.
- ✓ Ο άξονας x_i τοποθετείται κάθετα στο επίπεδο των z_i και z_{i-1} αξόνων.
- ✓ Ο άξονας y_i τοποθετείται έτσι ώστε να έχουμε ορθοκανονικό σύστημα αξόνων και οι άξονες να διατάσσονται αντιωρολογιακά ($x_i \rightarrow y_i \rightarrow z_i$).
- ✓ Αν ο z_{i-1} τέμνεται με τον z_i πάνω στον z_{i-1}, τότε $O_i \equiv O_{i-1}$.
- ✓ Αν δεν υπάρχει συμβατική τοποθέτηση του πλαισίου βάσης στην q_1 άρθρωση, χρειαζόμαστε ενδιάμεσο πλαίσιο $O_{0'} – x_{0'}y_{0'}z_{0'}$ στη βάση, τέτοιο ώστε $\overrightarrow{O_{0'}z_{0'}} \sim q_1$.

Η τοποθέτηση του πλαισίου στο Τελικό Εργαλείο Δράσης (Τ.Ε.Δ.) γίνεται με τέτοιο τρόπο ώστε να μην χρειαστεί η προσθήκη επιπλέον βοηθητικού πλαισίου.

Ο προσδιορισμός του **<u>πίνακα παραμέτρων της μεθόδου D-H</u>** γίνεται σύμφωνα με τους παρακάτω κανόνες:

- ✓ **θ_i**: Πρώτα κοιτάμε ποια είναι η γωνία κατά τον z_{i-1}, που πρέπει να περιστραφεί ο x_{i-1} για να πέσει πάνω στον x_i.
  Στη συνέχεια ελέγχουμε αν υπάρχει περιστροφή γύρω από τον z_{i-1}.
- ✓ **d_i**: Κοιτάμε κατά τον άξονα z_{i-1} τί απόσταση διανύουμε για να εντοπίσουμε το επόμενο πλαίσιο.
- ✓ **α_i**: Κοιτάμε κατά τον άξονα x_i πόσο πρέπει να περιστραφεί ο z_{i-1} για να πέσει στον z_i.
- ✓ **a_i**: Κοιτάμε κατά τον άξονα x_i τί απόσταση διανύουμε για να μετακινηθούμε από το O_{i-1} στο O_i πλαίσιο.

Πηγές

[1] Haonan Duan, Peng Wang, Yayu Huang, Guangyun Xu1, Wei Wei and Xiaofei Shen, “Robotics Dexterous Grasping: The Methods Based on Point Cloud and Deep Learning”, Frontiers in Neurorobotics, Front. Neurorobot., 09 June 2021, https://doi.org/10.3389/fnbot.2021.658280, 2021

[2] Muhammad Sami Siddiqui, Claudio Coppola, Gokhan Solak and Lorenzo Jamone, “Grasp Stability Prediction for a Dexterous Robotic Hand Combining Depth Vision and Haptic Bayesian Exploration”, Frontiers in Robotics and AI, Front. Robot. AI, 12 August 2021, Sec. Robot and Machine Vision, https://doi.org/10.3389/frobt.2021.703869, 2021

[3] S. Arai, Z. Feng, F. Tokuda, A.S.Z. Purnomo, Y. Xu and K. Kosuge, “Deep Learning-based Fast Grasp Planning for Robotic Bin-picking by Small Data Set without GPU”, https://www.techrxiv.org/, 2020

[4] Qingkai Lu, Mark Van der Merwe, and Tucker Hermans, “Multi-Fingered Active Grasp Learning”, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA (Virtual), October 25-29, 2020

[5] Qingkai Lu and Tucker Hermans, “Modeling Grasp Type Improves Learning-Based Grasp Planning”, IEEE Robotics and Automation Letters, Volume: 4, Issue: 2, April 2019

[6] Bruno Siciliano, Lorenzo Sciavicco, Luigi Villani, Giuseppe Oriolo, “Robotics: Modelling, Planning and Control”, Springer, 2009

[7] IEEE Robotics & Automation Society, Technical Committee for Robotic Handa, Grasping and Manipulation, https://www.ieee-ras.org/robotic-hands-grasping-and-manipulation

[8] Emergence of Cognitive Grasping through Introspection, Emulation and Surprise https://www.csc.kth.se/grasp/

[9] Yale Openhand Project, https://www.eng.yale.edu/grablab/openhand/

[10] The Open Hand Project: A Low-Cost Robotic Hand https://www.indiegogo.com/projects/the-open-hand-project-a-low-cost-robotic-hand#/

[11] Graspit! Simulator, https://graspit-simulator.github.io/

[12] OpenGRASP Simulation Toolkit for grasping and dexterous manipulation http://opengrasp.sourceforge.net/

[13] Hand Corpus Open Repository, https://www.handcorpus.org/

[14] Corey Goldfeder, Matei Ciocarlie, Hao Dang and Peter K. Allen, “The Columbia Grasp Database”, https://www.cs.columbia.edu/~allen/PAPERS/icra_7page_pub.pdf

[15] Human Grasping Database, http://grasp.xief.net/

[16] Yu, Q., Shang, W., Zhao, Z., Cong, S., and Li, Z., “Robotic grasping of unknown objects using novel multilevel convolutional neural networks: from parallel gripper to dexterous hand”, IEEE Transactions on Automation Science and Engineering (New York, NY), 2020

[17] Wu, B., Akinola, I., and Allen, P. K. Pixel-attentive policy gradient for multi-fingered grasping in cluttered scenes. arXiv [Preprint]. arXiv:1903.03227., 2019

[18] Le, Q. V., Kamm, D., Kara, A. F., and Ng, A. Y., “Learning to grasp objects with multiple contact points”, 2010 IEEE International Conference on Robotics and Automation (Anchorage, AK: IEEE), 5062–5069. doi: 10.1109/ROBOT.2010.5509508, 2010

[19] Mahler, J., Liang, J., Niyaz, S., Laskey, M., Doan, R., Liu, X., et al. (2017). Dex-net 2.0: deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. arXiv [Preprint]. arXiv:1703.09312. doi: 10.15607/RSS.2017.XIII.058

[20] Mahler, J., Matl, M., Liu, X., Li, A., Gealy, D., and Goldberg, K.. “Dex- Net 3.0: computing robust vacuum suction grasp targets in point clouds using a new analytic model and deep learning,” 2018 IEEE International Conference on Robotics and Automation (ICRA) (Brisbane, QLD: IEEE), 1–8. doi: 10.1109/ICRA.2018.8460887, 2018

[21] Qin, Y., Chen, R., Zhu, H., Song,M., Xu, J., and Su, H. “S4g: amodal singleview single-shot se (3) grasp detection in cluttered scenes”, Conference on Robot Learning (Osaka: PMLR), 53–65, 2020

[22] Xu, Y., Wang, L., Yang, A., and Chen, L. GraspCNN: real-time grasp detection using a new oriented diameter circle representation. IEEE Access 7, 159322–159331. doi: 10.1109/ACCESS.2019.2950535, 2019

[23] Varley, J., DeChant, C., Richardson, A., Ruales, J., and Allen, P. “Shape completion enabled robotic grasping,” in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (Vancouver, BC: IEEE), 2442–2447. doi: 10.1109/IROS.2017.8206060, 2017

[24] James, J. W., and Lepora, N. F. Slip Detection for Grasp Stabilization with a Multifingered Tactile Robot Hand. IEEE Trans. Robot. 37, 506–519. doi:10.1109/TRO.2020.3031245, 2021

[25] Shaw-Cortez, W., Oetomo, D., Manzie, C., and Choong, P. Technical Note for "Tactile-Based Blind Grasping: A Discrete-Time Object Manipulation Controller for Robotic Hands". IEEE Robot. Autom. Lett. 5, 3475–3476. doi:10.1109/LRA.2020.2977585, 2020

[26] Zuliani, M. C. K., and Manjunath, B. The Multiransac Algorithm and its Application to Detect Planar Homographies. In IEEE International Conference on Image Processing doi:10.1109/icip.2005.1530351, 2005

[27] Coleman, D., Sucan, I., Chitta, S., and Correll, N. Reducing the Barrier to Entry of Complex Robotic Software: a Moveit! Case Study, arXiv:1404.3785, 2014

[28] Roa, M. A., and Suárez, R. Grasp Quality Measures: Review and Performance. Auton. Robot 38, 65–88. doi:10.1007/s10514-014-9402-3, 2015

[30] Brochu, E., Cora, V. M., and de Freitas, N. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modelling and Hierarchical Reinforcement Learning. CoRR abs/1012.2599, 2010

[31] Nogueira, J., Martinez-Cantin, R., Bernardino, A., and Jamone, L. Unscented Bayesian Optimization for Safe Robot Grasping. 1967–1972. doi:10.1109/IROS.2016.7759310, 2016

[32] V. D. Nguyen, "Constructing stable force-closure grasps," IEEE International Conference on Robotics and Automation, pp. 1368–1373, 1986.

[33] C. Ferrari and J. Canny, "Planning optimal grasps," IEEE International Conference on Robotics and Automation, pp. 2290–2295, 1992.

[34] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," The International Journal of Robotics Research, vol. 34, no. 4-5, pp. 705–724, 2015.

[35] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from RGBD images: Learning using a new rectangle representation," IEEE International Conference on Robotics and Automation, pp. 3304–3311, 2011.

[36] Eric Brochu, Vlad M. Cora and Nando de Freitas, "A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning", cite as: arXiv:1012.2599 [cs.LG], 2010

[37] Peter I. Frazier, "A Tutorial on Bayesian Optimization", cite as: arXiv:1807.02811 [stat.ML], 2018

[38] R. Horst, P.M. Pardalos and N.V. Thoai, Introduction to Global Optimization, Second Edition. Kluwer Academic Publishers, 2000.

[39] Mnih,V., Kavukcuoglu, K., Silver,D., Rusu, A. A.,Veness, J., Bellemare,M.G., et al. Human-level control through deep reinforcement learning. Nature 518, 529–533. doi: 10.1038/nature14236, 2015

[40] Gualtieri, M., and Platt, R. Learning 6-dof grasping and pick-place using attention focus. arXiv [Preprint]. arXiv:1806.06134, 2018

[41] The Hand Embodied, Seventh Framework Programme, https://cordis.europa.eu/project/id/248587

[42] Richard S. Sutton and Andrew G. Barto, "Introduction to Reinforcement Learning", second edition, The MIT press, Cambridge, Massachusetts, 1998

[43] Tingwu Wang, "Learning Reinforcement, Learning by Learning, REINFORCE", Machine Learning Group, University of Toronto

[44] David Silver, Lecture {1:10}, UCL Course on Reinforcement Learning, Advanced Topics, (COMPM050/COMPGI13), United Kingdom, 2015

[45] Hado van Hasselt and Marco A. Wiering, from “Intelligent Systems Group, Department of Information and Computing Sciences”, Utrecht University, Proceedings of the 2007 IEEE Symposium on Approximate Dynamic Programming and Reinforcement Learning – ADPRL, 2007

[46] Paris Oikonomou, Athanasios Dometios, Mehdi Khamassi, Costas Tzafestas, “Task Driven Skill Learning in a Soft-Robotic Arm”, 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, Prague, Czech Republic. hal-03275472v2, Sep 2021

[47] Paris Oikonomou, Mehdi Khamassi, Costas Tzafestas, “Periodic movement learning in a soft-robotic arm”, IEEE International Conference on Robotics and Automation (ICRA 2020), Paris (virtuel), France. 10.1109/ICRA40945.2020.9197035. hal-03435441, May 2020

[48] Allegro Hand Users Manual, Wonik Robotics, AllegroHandWiki, http://wiki.wonikrobotics.com/AllegroHandWiki/images/e/eb/V4_AllegroHandUsersManual_1.1.pdf

[49] Wonik Robotics, https://www.wonikrobotics.com/

[50] Paul Jarbet, “Hand and Wrist Anatomy”, https://murdochorthopaedic.com.au/our-surgeons/paul-jarrett/patient-information-guides/hand-wrist-anatomy/, 2016

[51] ROS 2 Documentation, http://wiki.ros.org/

[52] Azoyam, A., “Harvesting Feasibility Analysis of an Automated Mushroom System”, MS thesis, University of Georgia, GA, 2004

[53] Mingsen Huang, Long He, Daeun Choi, John Pecchia, Yaoming Li, “Picking dynamic analysis for robotic harvesting of Agaricus bisporus mushrooms”, 2021, Computers and Electronics in Agriculture 185 106145, ELSEVIER, journal homepage: www.elsevier.com/locate/compag, 2021

[54] Reed, J.N., Tillett, R.D., “Initial experiments in robotic mushroom harvesting”, Mechatronics 4 (3), 265–279, 1994

[55] Kilian Kleeberger, Richard Bormann, Werner Kraus, Marco F. Huber, “A Survey on Learning-Based Robotic Grasping”, SpringerLink, Robotics in Manufacturing (JN Pires, Section Editor), 2020

[56] Oliver Kroemer, Scott Niekum, George Konidaris, “A Review of Robot Learning for Manipulation: Challenges, Representations, and Algorithms”, Cornell University, arXiv:1907.03146v3 [cs.RO], https://doi.org/10.48550/arXiv.1907.03146, 2020

[57] Joo Chuan Yeo, Cassidy Lee, Zhiping Wang, Chwee Teck Lim, “Tactile sensorized glove for force and motion sensing”, IEEE SENSORS, Orlando, FL, USA, IEEE Catalog Number: CFP16SEN-POD, ISBN: 978-1-4799-8288-2, 2016

[58] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, Sergey Levine, “Learning Complex Dexterous Manipulation with Deep Reinforcement Learning and Demonstrations”, Cornell University, arXiv:1709.10087v2 [cs.LG], 2018