ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ, ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ

Συμβολή στην Ανάπτυξη Πολυπρακτορικής Αρχιτεκτονικής Αναπτυξιακού Ρομποτικού Ελέγχου στη Βάση Ασαφούς Ενισχυτικής Μάθησης: Εφαρμογή στον Επιδέξιο Ρομποτικό Χειρισμό

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Ιωάννης Ν. Καρύγιαννης

Αθήνα, Δεκέμβριος 2012

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ, ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ

Συμβολή στην Ανάπτυξη Πολυπρακτορικής Αρχιτεκτονικής Αναπτυξιακού Ρομποτικού Ελέγχου στη Βάση Ασαφούς Ενισχυτικής Μάθησης: Εφαρμογή στον Επιδέξιο Ρομποτικό Χειρισμό

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Ιωάννης Ν. Καρύγιαννης

Συμβουλευτική Επιτροπή:

Σπυρίδων Γ. Τζαφέστας Ομότιμος Καθηγητής Ε.Μ.Π.	Πέτρος Α. Μαραγκός Καθηγητής Ε.Μ.Π.	Κωνσταντίνος Σ. Τζαφέστας Επίκουρος Καθηγητής Ε.Μ.Π.

Προς έγκριση από την επταμελή εξεταστική επιτροπή στις 20 Δεκεμβρίου 2012.

..... Σπυρίδων Γ. Τζαφέστας Ομότιμος Καθηγητής Ε.Μ.Π.	..... Πέτρος Α. Μαραγκός Καθηγητής Ε.Μ.Π.	..... Κωνσταντίνος Σ. Τζαφέστας Επίκουρος Καθηγητής Ε.Μ.Π.

..... Γεώργιος Π. Παπαβασιλόπουλος Καθηγητής Ε.Μ.Π.	..... Ανδρέας Γ. Σταφυλοπάτης Καθηγητής Ε.Μ.Π.
..... Κωνσταντίνος Ι. Κυριακόπουλος Καθηγητής Ε.Μ.Π.	..... Αντώνιος Τζες Καθηγητής Πανεπιστημίου Πατρών

Αθήνα, Δεκέμβριος 2012.

.....
Ιωάννης Ν. Καρύγιαννης
Διπλωματούχος
Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών
του Πανεπιστημίου Concordia
Montreal - Quebec, Canada.

Περίληψη

Η παρούσα διατριβή, προτείνει μια ιεραρχική πολυπρακτορική αρχιτεκτονική εφαρμοσμένη στο πεδίο του επιδέξιου ρομποτικού χειρισμού. Η προτεινόμενη αρχιτεκτονική βασίζεται σε μία εμφωλευμένη ιεραρχική δομή, όπου κάθε πράκτορας σχηματίζει (τοπικά) εικόνα για τη συνολική (γενικευμένη) κατάσταση του συστήματος καθώς επίσης και για την εξέλιξη της εργασίας, μέσω μιας ανάδρομης (top-down / bottom-up) διαδικασίας. Με την οργάνωση των πρακτόρων σε ένα συγκεκριμένο πρότυπο εμφωλευμένης αρχιτεκτονικής, όπως αυτό που προτείνεται στο πλαίσιο της παρούσης διατριβής επιτρέπεται α) περαιτέρω επεκτασιμότητα σε σαφώς πιο σύνθετες κινηματικές τοπολογίες, και β) η μοντελοποίηση του συστήματος συνολικά με ένα τμηματικό (modular) και παράλληλα, δομημένο (structural) τρόπο. Η προτεινόμενη μεθοδολογία βασίζεται στην εφαρμογή μηχανισμών ασαφούς ενισχυτικής μάθησης, με σκοπό την εξέλιξη, σε τοπικό επίπεδο για κάθε πράκτορα, μιας αντιστοίχισης καταστάσεων - δράσεων σε ένα συνεχές πεδίο, δημιουργώντας με αυτό τον τρόπο ένα πολυπρακτορικό σύστημα το οποίο επιδεικνύει αναπτυξιακές ιδιότητες. Οι πράκτορες αντιστοιχούν σε ανεξάρτητους βαθμούς ελευθερίας του συστήματος, οι οποίοι επιτυγχάνουν να αποκτήσουν εμπειρία και να αναπτύξουν δεξιότητες σχετικές με την εκτέλεση συγκεκριμένων εργασιών συνεργατικού χειρισμού, μέσω μιας συνεχόμενης διαδικασίας εξερεύνησης (exploration) και αξιοποίησης (exploitation) του χώρου αντιστοίχισης καταστάσεων - δράσεων. Η παρούσα διατριβή μελετά την εφαρμογή της προτεινόμενης μεθοδολογίας πολυπρακτορικού αναπτυξιακού ελέγχου σε προβλήματα που προέρχονται από το χώρο του επιδέξιου ρομποτικού χειρισμού, ενώ παράλληλα εξετάζει την επεκτασιμότητα της συγκεκριμένης αρχιτεκτονικής σε συνεργατικά αυτοκινούμενα ρομποτικά συστήματα. Πιο συγκεκριμένα, εκτελέσθηκαν και παρουσιάζονται τρία σύνολα πειραματικών δοκιμών με στόχο την αξιολόγηση της προτεινόμενης μεθοδολογίας: 1) το πρώτο σύνολο αριθμητικών πειραμάτων θεωρεί την περίπτωση απλής ανοικτής κινηματικής αλυσίδας η οποία παρουσιάζει κινηματικούς πλεονασμούς (kinematic redundancies) ως προς τον επιθυμητό στόχο, 2) το δεύτερο πείραμα επεκτείνει περαιτέρω την προηγούμενη περίπτωση, θεωρώντας τρεις παράλληλες κινηματικές αλυσίδες οι οποίες συνεργατικά προσπαθούν να επιτύχουν σταθερή

ρομποτική λαβή, ενώ 3) το τελευταίο πείραμα εφαρμόζει την προτεινόμενη τοπολογία σε αυτοκινούμενα ρομπότ τα οποία πραγματοποιούν εργασία τύπου "box - pushing" (δηλαδή, από κοινού ώθηση χειριζόμενου αντικειμένου σε επιθυμητή θέση-στόχο). Οι πειραματικές αυτές δοκιμές αποσκοπούν στην αποτίμηση της ικανότητας που παρουσιάζει το προτεινόμενο πολυπρακτορικό σύστημα ως προς την αυτόνομη και προοδευτική απόκτηση συνεργατικών δεξιοτήτων μέσω μιας εσωτερικής διεργασίας μάθησης. Αυτή η εσωτερική διεργασία μάθησης δεν βασίζεται σε κάποιο εκ των προτέρων δεδομένο πλήρες μοντέλο της εκτελούμενης εργασίας, ούτε ακολουθεί κάποια στρατηγική καθολικής σχεδίασης δράσης βάσει ενός τέτοιου συνολικού μοντέλου. Τα πειραματικά αποτελέσματα που παρουσιάζονται στην παρούσα διατριβή δείχνουν την επεκτασιμότητα της προτεινόμενης εμφωλευμένης-ιεραρχικής αρχιτεκτονικής, όπου νέοι πράκτορες μπορούν αναδρομικά να προστεθούν στην τοπολογία καλύπτοντας διαφορετικούς βαθμούς ελευθερίας. Επιπλέον, αναλύονται χαρακτηριστικά γενίκευσης γνώσης καθώς και ευρωστίας της προτεινόμενης μεθοδολογίας κινηματικού ελέγχου σε απρόβλεπτες αστοχίες δομικών στοιχείων του ρομποτικού συστήματος. Τα πειραματικά αποτελέσματα τα οποία παρουσιάζονται υπογραμμίζουν τις δυνατότητες που παρέχει ένα τέτοιο κατανεμημένο σχήμα ρομποτικού ελέγχου, καταδεικνύοντας την επιτυχή εκτέλεση συνεργατικών κινήσεων οι οποίες οδηγούν το ρομποτικό σύστημα σε κινηματικές λύσεις συγκρίσιμες με τις θεωρητικά βέλτιστες (near-optimal). Αναλύοντας τα αποτελέσματα που προέκυψαν από την παρούσα διατριβή, διαφαίνεται ότι ένα τέτοιο κατανεμημένο πλαίσιο ρομποτικής μάθησης διαθέτει δυνητικά υψηλό βαθμό επεκτασιμότητας στον έλεγχο ρομποτικών συστημάτων τα οποία μπορεί να είναι κινηματικά πιο σύνθετα, αποτελούμενα από πολλαπλούς βαθμούς ελευθερίας τόσο σε ανοικτές όσο και σε κλειστές κινηματικές τοπολογίες.

Λέξεις–κλειδιά

Τεχνητή Νοημοσύνη, Αναπτυξιακή Ρομποτική, Ενισχυτική Μάθηση, Επιδέξιος Ρομποτικός Χειρισμός, Πολυπρακτορικά Ρομποτικά Συστήματα

Abstract

This thesis proposes a model-free learning mechanism based on a nested-hierarchical multi-agent architecture, which is applied in the context of dexterous robot manipulation control. In the proposed multi-agent system, each agent forms a local (partial) view of the global system state and task progress, through a recursive (top-down/bottom-up) learning process. By organizing the agents in a nested architecture, the goal is to facilitate modular scaling to more complex kinematic topologies, with loose control coupling among the agents. Reinforcement learning is applied within each agent, to evolve a local state-to-action mapping in a continuous domain, thus leading to a system that exhibits developmental properties. The agents correspond in fact to independent degrees-of-freedom (DOF) of the system, managing to gain experience over the task that they collaboratively perform by continuously exploring and exploiting their state-to-action mapping space. This thesis addresses problem settings in the domain of kinematic control of dexterous robot manipulation. Three sets of numerical experiments are performed: (i) the first one considers the case of a single-linkage open kinematic chain, presenting kinematic redundancies given the desired task-goal, (ii) the second experiment extends further on the previous case by considering three individual kinematic chains cooperatively acting to achieve a quasi-static multi-finger grasp, and (iii) the last experiment extends the proposed multi-agent framework to a control problem in the field of autonomous mobile robots, by considering two e-Puck robots performing a collaborative "box-pushing" task. The focal issue in all experiments is to assess the capacity of the proposed multi-agent system to progressively and autonomously acquire cooperative sensorimotor skills through a self-learning process, that is, without the use of any explicit model-based planning strategy. Generalization and robustness properties of the overall multi-agent system are also explored. Furthermore, these experiments aim to demonstrate the scaling properties of the proposed nested-hierarchical architecture, where new higher-level agents can be recursively added in the hierarchy to encapsulate individual active DOFs. The experimental results presented in this thesis demonstrate the feasibility of

such a distributed multi-agent control framework, showing that the solutions which emerge are plausible and near-optimal.

Keywords

Artificial Intelligence, Developmental Robotics, Reinforcement Learning, Dexterous Manipulation, Multiagent Robotic Systems

Αφιερώνεται
στην κόρη μου Ματίνα, στον γιο μου Νικόλα
& στην αγαπημένη μου σύζυγο Μαρία

Ευχαριστίες

Η διατριβή αυτή παρουσιάζει τα ερευνητικά αποτελέσματα μιας μακρόχρονης πορείας στο Εργαστήριο Ρομποτικής και Αυτοματισμού της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου. Σε αυτές τις γραμμές θα ήθελα καταρχήν να ευχαριστήσω τον επιβλέποντα καθηγητή μου, κ. Κωσταντίνο Τζαφέστα, Επίκουρο Καθηγητή ΕΜΠ για την πολύτιμη καθοδήγησή του, τόσο σε επιστημονικά θέματα όσο και στις δυσκολίες που προέκυψαν κατά τη διάρκεια αυτών των ετών. Η εμπιστοσύνη που έδειξε στο πρόσωπο μου ήταν μια διαρκής πηγή έμπνευσης αλλά και κίνητρο για την ολοκλήρωση της επιστημονικής μου οντότητας. Ειλικρινά τον ευχαριστώ γι' αυτή την ευκαιρία. Επίσης, τα υπόλοιπα μέλη της συμβουλευτικής επιτροπής, κ. Σπυρίδωνα Τζαφέστα, Ομότιμο Καθηγητή ΕΜΠ και κ. Πέτρο Μαραγκό, Καθηγητή ΕΜΠ, για τις υποδείξεις και τις παρατηρήσεις τους στη διάρκεια εκπόνησης της παρούσας διατριβής.

Φυσικά, σ' αυτό το ταξίδι δεν ήμουν μόνος. Είχα αρκετούς ανθρώπους δίπλα μου, τους οποίους θα ήθελα να ευχαριστήσω. Αρχικά λοιπόν, πρέπει να κάνω ιδιαίτερη αναφορά στο φίλο και συνάδελφο Θοδωρή Ρεκατσίνα. Η συμβολή του στη διατριβή μου και στα ερευνητικά αποτελέσματα ήταν θεμελιώδης, αφού ορισμένες πειραματικές υλοποιήσεις έγιναν με τη βοήθειά του. Επίσης, το φίλο και συνάδελφο Σωτήρη Αποστολόπουλο με τον οποίο πραγματοποιήσαμε πειραματικές υλοποιήσεις σε ρομποτικά τετράποδα. Θα ήθελα επίσης να ευχαριστήσω τους συναδέλφους στο εργαστήριο, τον Σπύρο Βελάνα και τον Νίκο Μήτσου, για τις πολύωρες συζητήσεις μας σε θέματα ερευνητικά καθώς και για το ευχάριστο και παραγωγικό κλίμα που είχαμε.

Επιπλέον, επιθυμώ να ευχαριστήσω θερμά τον κ. Παναγιώτη Κακλή, Καθηγητή ΕΜΠ, για το ενδιαφέρον, τη συμπαράσταση και τη βοήθεια που μου παρείχε τόσο σε επιστημονικό, όσο και σε προσωπικό επίπεδο.

Στην πορεία μου αυτή πρέπει φυσικά ν' αναφερθώ και στα άτομα τα οποία με στήριξαν με αμέριστη αγάπη και, πολλές φορές, υπομονή. Θα ήθελα να ευχα-

ριστήσω την αγαπημένη μου σύζυγο Μαρία και τα δύο ξεχωριστά παιδιά μου, οι οποίοι με στήριξαν ηθικά και ψυχολογικά σε όλες τις δύσκολες στιγμές που προέκυψαν σε αυτή τη δεκαετή πορεία. Τέλος, ποτέ δε θα είναι αρκετές οι ευχαριστίες μου προς τους γονείς μου Νίκο και Ματίνα, καθώς και την αγαπημένη μου αδερφή Κατερίνα. Η στήριξη, η αγάπη και η ουσιαστική συμπαράστασή τους, ήταν πάντα το εφαλτήριο για να επιδιώξω τους σημαντικότερους στόχους στη ζωή μου.

Ιωάννης Ν. Καρύγιαννης

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Κίνητρα και Στόχοι της Έρευνας

Τα σύγχρονα ρομποτικά συστήματα καλούνται να εκτελέσουν εργασίες σε περιβάλλοντα που είναι συνήθως, μερικώς παρατηρήσιμα, (*Partially Observable*), δυναμικά μεταβαλλόμενα (μη στατικά), στοχαστικά με σημαντικό βαθμό αβεβαιότητας, μη δομημένα (*Unstructured*)) και φυσικά συνεχή (*Continuous*). Εξ' αιτίας αυτής της πολυπλοκότητας που σχετίζεται με τις πραγματικές συνθήκες και τα χαρακτηριστικά του περιβάλλοντος ρομποτικής εργασίας, τα σύγχρονα ρομποτικά συστήματα είναι απαραίτητο να ενσωματώνουν σημαντικό βαθμό Νοημοσύνης, έτσι ώστε να δύνανται να μαθαίνουν γρήγορα και να λειτουργούν με ασφάλεια, ακρίβεια και αξιοπιστία. Μια πρώτη ερώτηση λοιπόν που προκύπτει από τα παραπάνω είναι τι ακριβώς εννοούμε με τον όρο Νοημοσύνη. Γενικά, με τον όρο Νοημοσύνη μπορεί να θεωρήσουμε ότι αναφερόμαστε συνήθως σε ένα σύνολο ιδιοτήτων οι οποίες περιλαμβάνουν: την ικανότητα λογικής σκέψης, την ικανότητα μάθησης μέσω της εμπειρίας και την προσαρμοστικότητα σε νέες καταστάσεις. Η διατύπωση όμως ενός ακριβούς ορισμού της Νοημοσύνης εμπίπτει στο πεδίο τόσο της φιλοσοφίας όσο και της ψυχολογίας: *'...Μια ιδιαίτερα γενική νοητική ικανότητα η οποία, μεταξύ άλλων, περιλαμβάνει την ικανότητα της λογικής σκέψης, σχεδιασμού, της επίλυσης προβλημάτων, της αφηρημένης σκέψης, της κατανόησης περίπλοκων ιδεών καθώς και της μάθησης μέσω εμπειριών... η νοημοσύνη αντικατοπτρίζει μια ευρύτερη και βαθύτερη ικανότητα για κατανόηση του περιβάλλοντός μας, να αντιλαμβανόμαστε τα πράγματα γύρω μας και στη συνέχεια να συμπεραίνουμε τον τρόπο με τον οποίο πρέπει να ενεργούμε..'*[45]. Μπορούμε να διαχωρίσουμε την Νοημοσύνη σε Βιολογική και Τεχνητή. Η Βιολογική Νοημοσύνη έχει στάδια εξέλιξης (*Gradation*)), αυτο-οργανώνεται με επαναλαμβανόμενο τρόπο και ασχολείται με την αντιμετώπιση προβλημάτων χρησιμοποιώντας λογικούς κανόνες, ενώ έχει τη δυνατότητα αφαιρετικής ικανότητας αντίληψης και συλλογιστικής. Πολύ βασικό στοιχείο αποτελεί το γεγονός ότι η Τεχνητή Νοημοσύνη βασίζεται στη χρήση υπολογιστικών μοντέλων και αρχών μέσα σε ένα πλαίσιο το οποίο είναι σαφώς καθορισμένο. Όταν το πλαίσιο παύει να είναι καθορισμένο, τότε αρχίζει να γίνεται αντιληπτή η εξαιρετική σημασία που αποκτούν ιδιότητες όπως η προσαρμοστικότητα και η ευελιξία της συμπεριφοράς σε ένα συνεχώς μεταβαλλόμενο και αβέβαιο κόσμο.

1.1.1 Αναπτυξιακή Ρομποτική

Ένας από τους βασικούς στόχους λοιπόν της ρομποτικής επιστήμης και τεχνολογίας είναι ο σχεδιασμός και η υλοποίηση ρομποτικών μηχανισμών και συστημάτων τα οποία θα δύνανται να εκτελούν πολύπλοκες εργασίες σε σύνθε-

τα, φυσικά περιβάλλοντα. Ο παραδοσιακός σχεδιασμός ανάπτυξης ρομποτικών συστημάτων δημιουργεί πολύ αποδοτικά υποσυστήματα, απαραίτητα για την αποτελεσματική λειτουργία ενός ρομπότ, όπως αισθητήριες διατάξεις ανίχνευσης συγκεκριμένων χαρακτηριστικών, μονάδες ελεγκτών, καθώς και μεθοδολογίες και στρατηγικές συμπεριφοράς και δράσης, οι οποίες όμως είναι συνήθως κατάλληλες μόνο για ένα σύνολο αισθητήρων και συστημάτων ελέγχου που συνθέτουν ένα συγκεκριμένο ρομποτικό περιβάλλον, το οποίο στη συνέχεια καλείται να καλύψει τις ανάγκες μιας πολύ συγκεκριμένης εργασίας. Πολύ συχνά, ακόμη και μικρές αλλαγές / αποκλίσεις στις υποθέσεις σύμφωνα με τις οποίες σχεδιάζονται αυτά τα τμήματα που συνθέτουν ένα ρομποτικό σύστημα, απαιτούν την επαναρύθμισή τους αν όχι τον πλήρη επανασχεδιασμό τους. Κάθε νέος συνδυασμός εργασίας, περιβάλλοντος και εξοπλισμού (*hardware*) του ρομπότ, παρουσιάζει ένα εγγενώς νέο μηχανικό πρόβλημα σχεδίασης και υλοποίησης. Επιπλέον, ρομποτικές συμπεριφορές οι οποίες έχουν σχεδιαστεί με προσέγγιση τύπου "hard - coded" συχνά αποτυγχάνουν όταν έχουν να αντιμετωπίσουν νέες, απρόβλεπτες καταστάσεις. Από τα παραπάνω λοιπόν προκύπτει ότι ο μηχανικός ρομποτικής πρέπει κατά τη φάση σχεδιασμού είτε να προβλέψει κάθε απρόοπτο, ή να τροποποιήσει το περιβάλλον του ρομπότ ώστε να περιορίσει το πεδίο δράσης του. Δεδομένου ότι αυτή είναι μια εξαιρετικά επίπονη και αβέβαιη διαδικασία, είναι δύσκολο να φανταστούμε ότι θα μπορούμε να έχουμε ρομπότ (τεχνητές οντότητες) που θα μπορούν να λειτουργούν αυτόνομα δίπλα σε ανθρώπους (βιολογικές οντότητες) σε μη περιορισμένα και μη δομημένα περιβάλλοντα, χωρίς να αυτοματοποιήσουμε την ίδια την διαδικασία της αντίληψης και ανάπτυξης σχετικής συμπεριφοράς, μέσω κατάλληλου συνδυασμού τεχνικών Μηχανικής Μάθησης (*Machine Learning*) [91]. Ουσιαστικά, στόχο εδώ αποτελεί η σχεδίαση και ανάπτυξη ρομποτικών συστημάτων που θα ενσωματώνουν κάποια μορφή δυνατοτήτων και λειτουργιών "αυτο-προσαρμοζόμενης" και *"αναπτυξιακής" μηχανικής μάθησης* (*Developmental Learning*) [86], επιτρέποντας στα ρομπότ να δημιουργούν τις δικές τους αισθητηριακές και κινητήριες αναπαραστάσεις και συμπεριφορές, προσαρμοζόμενα έτσι στις νέες καταστάσεις όπως και όταν εκείνες προκύπτουν.

Οι παραπάνω βασικές αρχές και ιδέες έχουν οδηγήσει στην ανάπτυξη μιας νέας προσέγγισης στο χώρο της ρομποτικής, η οποία επικεντρώνεται στην αυτόνομη αυτο-οργάνωση συστημάτων ρομποτικού ελέγχου τα οποία δεν εξαρτώνται από την εκάστοτε ανατιθέμενη ρομποτική εργασία. Η προσέγγιση αυτή καλείται συχνά υπό τον όρο *Αναπτυξιακή Ρομποτική* (*Developmental Robotics*)[48]. Η ιδέα της πηγάζει από την αναπτυξιακή ψυχολογία και την αναπτυξιακή νευρολογία. Η αναπτυξιακή ρομποτική προχωρά ένα βήμα παραπάνω από την κλασική προσέγγιση όπου το ρομποτικό σύστημα σχεδιάζεται για να δώσει λύση σε ένα συγκεκριμένο προκαθορισμένο πρόβλημα (όπως ο σχεδιασμός πορείας προς μια

επιθυμητή θέση-στόχο). Αντίθετα, η αναπτυξιακή ρομποτική διερευνά τις ικανότητες του ρομπότ σε αντίληψη, γνώση και συμπεριφορά που μπορεί το ίδιο να ανακαλύψει μέσω ενεργειών που εγκαινιάζονται από το ίδιο, και είναι βασισμένες στη δική του φυσική μορφολογία και τη δυναμική δομή του περιβάλλοντός του [86]. Σε αυτό το σημείο, ίσως θα ήταν σκόπιμο να δούμε ποια είναι η γενική απαίτηση που υπάρχει, έτσι ώστε να μπορέσουμε στη συνέχεια να αντιληφθούμε τις δυσκολίες που προκύπτουν. Πρέπει λοιπόν, να λάβουμε υπόψη ότι για τους ανθρώπους είναι επιθυμητό να ελέγχουν (ίσως καλύτερα, να επικοινωνούν με) τα ρομπότ μέσω υψηλού επιπέδου εντολών (*Higher Level Commands*), ενώ είναι επίσης εξαιρετικά κουραστικό για εκείνους να δίνουν λεπτομερείς εντολές για άμεσες ενέργειες κάθε δέκατο του δευτερολέπτου. Η δυσκολία λοιπόν που προκύπτει από αυτή την λειτουργική απαίτηση και η οποία αποτελεί ουσιαστική πρόκληση είναι η δυνατότητα να "προγραμματιστεί" ένα ρομπότ κατάλληλα ώστε να αντιλαμβάνεται και να εκτελεί τέτοιες υψηλού επιπέδου εντολές σε άγνωστα ανθρώπινα - φυσικά περιβάλλοντα.

Η αυτόνομη νοητική ανάπτυξη των ρομποτικών συστημάτων είναι ένα νέο πεδίο που ελκύει όλο και περισσότερο το ενδιαφέρον στη ρομποτική και την τεχνητή νοημοσύνη. Πρόσφατες εξελίξεις στη νευρολογία έχουν αμφισβητήσει την ιδέα ότι η δομή του εγκεφάλου και η αναπαράστασή της είναι σε μεγάλο βαθμό προκαθορισμένα από τα ανθρώπινα γονίδια. Το αναπτυξιακό πρόγραμμα στα ανθρώπινα γονίδια φαίνεται να είναι περισσότερο γενικής φύσης από ό,τι πίστευαν πολλοί, και επιτρέπει στους ανθρώπους να αναπτύσσουν το μυαλό τους από τη βρεφική ηλικία μέχρι την ενηλικίωση, μέσω εμπειριών σε πραγματικό χρόνο [29]. Βασικό κίνητρο για την έρευνά μας αποτελεί η σημαντική προσπάθεια η οποία έχει καταβληθεί τα τελευταία χρόνια για την κατασκευή υπολογιστικών μοντέλων που θα επέτρεπαν στα ρομπότ να αναπτύξουν τις ικανότητές τους αυτόνομα μέσω αλληλεπίδρασης με το περιβάλλον τους. Τέτοια συστήματα μπορούμε να τα χαρακτηρίσουμε ως "αναπτυξιακά" ρομποτικά συστήματα. Αυτό προϋποθέτει ότι η εσωτερική αρχιτεκτονική και αναπαράσταση του ρομπότ πρέπει να δημιουργείται αυτόματα και με αυξητικό τρόπο μέσω μιας τέτοιας "αναπτυξιακής διαδικασίας". Ο στόχος αυτού του νέου ερευνητικού πεδίου είναι να επιτρέψει τη δημιουργία και ανάπτυξη αυτόνομων ρομποτικών συστημάτων τα οποία θα δύνανται να "μεγαλώσουν νοητικά" και να αποκτήσουν νέες ικανότητες συλλογισμού, λήψης αποφάσεων και δράσης, μέσω της αλληλεπίδρασής τους με το ανθρώπινο περιβάλλον.

Οι μακροπρόθεσμοι στόχοι του ερευνητικού πλαισίου στο οποίο εντάσσεται και η παρούσα ερευνητική προσπάθεια, και το οποίο περιγράφεται στα υπόλοιπα τμήματα αυτής της διατριβής, είναι από τη μία, η συμβολή (με άμεσο τρόπο) στη δημιουργία νέων τεχνικών προς την ανάπτυξη αποδοτικών συστημάτων τεχνη-

τής νοημοσύνης, και από την άλλη (εμμέσως) η συνεισφορά στην καλύτερη κατανόηση διαδικασιών βιολογικής νοημοσύνης. Η παρούσα διατριβή εστιάζει στο συγκεκριμένο τομέα που αφορά την ανάπτυξη πολυπρακτορικών (*Multi-Agent*) συστημάτων ελέγχου για εφαρμογές επιδέξιου και γενικά σύνθετου ρομποτικού χειρισμού. Η κατανόηση της πολυπλοκότητας των διαδικασιών βιολογικής νοημοσύνης που συντελούν στην εκτέλεση επιδέξιων εργασιών μπορεί να βοηθηθεί σε σημαντικό βαθμό από την εμβάθυνση στη θεματολογία που μελετά η επιστήμη της τεχνητής νοημοσύνης και αντίστροφα, η τεχνητή νοημοσύνη μπορεί να ωφεληθεί ιδιαίτερα από τη γνώση των διαδικασιών που φαίνονται ότι ακολουθούνται στη φύση κατά τη δημιουργία και ανάπτυξη ευφυών συστημάτων ικανών να αντιμετωπίσουν το συγκεκριμένο επίπεδο πολυπλοκότητας. Με αυτό δεν εννοούμε ότι η τεχνητή νοημοσύνη πρέπει απαραίτητα να μιμηθεί ακριβώς τη βιολογική νοημοσύνη, ή ότι η βιολογική νοημοσύνη πρέπει να αποδειχθεί ότι λειτουργεί ακριβώς όπως τα καλύτερα συστήματα τεχνητής νοημοσύνης. Όμως, διαφαίνεται ότι πρέπει να υπάρχουν γενικές αρχές της νοημοσύνης τις οποίες μπορεί κανείς να αναζητήσει μέσω μιας τέτοιας συνέργειας στις μελέτες που αφορούν τόσο βιολογικά συστήματα όσο και συστήματα τεχνητής νοημοσύνης.

Έχοντας καθορίσει τις βασικές έννοιες με τις οποίες προσεγγίζουμε στην παρούσα διατριβή το πεδίο της τεχνητής νοημοσύνης, μπορούμε να προχωρήσουμε στη αναζήτηση μιας κύριας μεθοδολογίας ενσωμάτωσης νοημοσύνης στα ρομποτικά συστήματα τα οποία μελετούμε, η οποία βασίζεται στο ευρύτερο επιστημονικό πεδίο που είναι γνωστό ως "Συνδετισμός" (*Connectionism*) [53]. Σε αυτή την προσέγγιση, η νοημοσύνη πηγάζει από ένα μεγάλο αριθμό στοιχείων επεξεργασίας που διασυνδέονται όλα μαζί, με το καθένα να εκτελεί μια απλή λειτουργία (όπως για παράδειγμα τα Νευρωνικά Δίκτυα) [14]. Στην εργασία που παρουσιάζουμε εδώ εξετάζουμε την προσέγγιση του Συνδετισμού (*Connectionism*) από μια διαφορετική σκοπιά - αυτήν της Γνωσιακής Επιστήμης (*Cognitive Science)* (Γνωσιακή Επιστήμη: η μελέτη της λειτουργίας του εγκεφάλου με συνδυασμό διαφορετικών γνωσιακών αντικειμένων) [29]. Αυτή η προσέγγιση χρησιμοποιεί τον Συνδετισμό για να δώσει απαντήσεις σε ερωτήματα που σχετίζονται με την ανθρώπινη γνώση/αντίληψη, καλύπτοντας από τις αντιληπτικές διαδικασίες μέχρι και τις διαδικασίες συλλογιστικής *(Reasoning).* Ο Συνδετισμός - στο πλαίσιο της Γνωσιακής Επιστήμης - αποτελεί ουσιαστικά μια θεωρία επεξεργασίας πληροφοριών [53]. Αντίθετα με τα κλασικά συστήματα που χρησιμοποιούν σαφείς, συχνά λογικούς κανόνες διευθετημένους ιεραρχικά για το χειρισμό των συμβόλων με σειριακό τρόπο, τα συστήματα που βασίζονται στην προσέγγιση του Συνδετισμού στηρίζονται στην παράλληλη επεξεργασία υπο-συμβόλων, χρησιμοποιώντας στατιστικές ιδιότητες αντί για λογικούς κανόνες για τη μετατροπή των πληροφοριών. Επιπλέον, η προσέγγιση του Συν-

δετισμού βασίζει τα μοντέλα που δημιουργούνται στη γνώση και κατανόηση της νευροφυσιολογίας *(Neurophysiology)* του εγκεφάλου ενώ παράλληλα προσπαθεί να ενσωματώσει αυτές τις λειτουργικές ιδιότητες στη διεργασία για την απόκτηση της γνώσης [53]. Στην εργασία μας, υιοθετήσαμε αυτήν τη γενική προσέγγιση του Συνδετισμού στο συγκεκριμένο τομέα της Γνωσιακής Επιστήμης, της παράλληλης δηλαδή επεξεργασίας υπο-συμβόλων, προσπαθώντας να ενισχύσουμε τους ρομποτικούς μας πράκτορες με επαρκή βαθμό συμπεριφορών αυτο-επίβλεψης και αυτο-οργάνωσης, έτσι ώστε να μπορούν να ανταπεξέρχονται σε δυναμικά μεταβαλλόμενα και αβέβαια περιβάλλοντα.

Το πεδίο που καλύπτει η Τεχνητή Νοημοσύνη είναι ιδιαίτερα ευρύ. Ασχολείται με διάφορα είδη σχημάτων αναπαράστασης γνώσης, διαφορετικές μεθόδους για την επίλυση της αβεβαιότητας στα δεδομένα και τη γνώση, διαφορετικά σχήματα για την αυτοματοποιημένη μηχανική μάθηση και πολλά άλλα. Ανάμεσα στους τομείς εφαρμογής της Τεχνητής Νοημοσύνης είναι τα Έμπειρα Συστήματα, η Θεωρία Παιγνίων, η Επεξεργασία Φυσικής Γλώσσας, η Αναγνώριση Εικόνων και φυσικά η Ρομποτική. Μέσα στο καλά καθορισμένο πλαίσιο της Τεχνητής Νοημοσύνης και των πολυπρακτορικών ρομποτικών συστημάτων προσπαθούμε να καθορίσουμε ένα συγκεκριμένο πλαίσιο στο οποίο θα προσπαθήσουμε να εφαρμόσουμε μια νέα πολυπρακτορική αρχιτεκτονική η οποία θα συνδυάζει ικανότητες μάθησης, προσαρμοστικότητα σε δυναμικά περιβάλλοντα και τη συμπεριφορά που δε θα βασίζεται σε εκ των προτέρων γνωστά μοντέλα, και όλα αυτά εφαρμοσμένα στον συγκεκριμένο τομέα του επιδέξιου ρομποτικού χειρισμού. Η έννοια της δημιουργίας μιας πολυπρακτορικής δομής όπου, από τη μια μεριά, τα βασικά δομικά συστατικά της ακολουθούν μια κοινή, απλή και ιεραρχική αρχιτεκτονική, ενώ από την άλλη, αναπτύσσουν ικανότητες και συμπεριφορές με την πάροδο του χρόνου χωρίς να κάνουν χρήση κάποιας προκαθορισμένης στρατηγικής που να βασίζεται σε κάποιο εκ των προτέρων γνωστό μοντέλο, ανοίγει αναμφισβήτητα μεγάλες δυνατότητες. Επιπρόσθετα, η προσπάθειά μας είναι να απομονώσουμε συγκεκριμένες βασικές έννοιες από τους τομείς της πολυπρακτορικής θεωρίας, των αυτοκινούμενων ρομπότ *(Mobile Robotics)*, της ασαφούς λογικής, της μηχανικής μάθησης, του προσαρμοστικού ελέγχου και να τις συνθέσουμε με μεθοδικό τρόπο, δημιουργώντας έτσι ένα υβριδικό σύστημα ρομποτικού ελέγχου το οποίο θα αξιολογηθεί στον τομέα του επιδέξιου χειρισμού.

Έχοντας περιγράψει τις βασικές έννοιες που συνθέτουν το κίνητρο της ερευνητικής μας προσπάθειας, τα οφέλη που μπορεί να προκύψουν μέσω μιας τέτοιας προσέγγισης γίνονται περισσότερο φανερά. Με τη χρησιμοποίηση μιας προσαρμοστικής, πολυπρακτορικής αρχιτεκτονικής που δεν βασίζεται σε μοντέλα, σε συνδυασμό με την προσέγγιση του Συνδετισμού προσπαθούμε να επιτύχου-

με μια συμβολή στο πεδίο της αναπτυξιακής ρομποτικής που αποτελεί και τον απώτερο στόχο της ερευνητικής μας προσπάθειας. Ο στόχος αυτός οριοθετείται κατάρχάς ως ο σχεδιασμός μιας αρχιτεκτονικής ελέγχου που θα μπορεί να εφαρμοστεί σε ένα επιδέξιο ρομποτικό χειριστή, έτσι ώστε όταν το ρομπότ πραγματοποιεί εκκίνηση για πρώτη φορά να "εγκαινιάζει" μια συνεχή, αυτόνομη αναπτυξιακή διαδικασία. Αυτή η διαδικασία πρέπει να είναι μη επιβλεπόμενη, χωρίς προγραμματισμό, και ανεξάρτητη από την εργασία την οποία το ρομπότ δύναται να κληθεί να πραγματοποιήσει. Αναφερόμαστε, δηλαδή, σε μία αρχιτεκτονική η οποία πρέπει να δύναται να λειτουργήσει εξίσου καλά σε οποιαδήποτε ρομποτική πλατφόρμα, σε ένα σταθερό ρομποτικό βραχίονα, σε ένα επιδέξιο ρομποτικό χέρι, σε ένα ρομπότ με ρόδες, ή σε ένα ρομπότ με πόδια.

Η εγγενής αναπτυξιακή διαδικασία που εξετάζουμε περιέχει τρεις βασικούς άξονες: αφαιρετικότητα (*abstraction*), πρόβλεψη (*prediction*) και εσωτερικό κίνητρο (*Self-Motivation*). Σε ένα ρεαλιστικό δυναμικό περιβάλλον, ένα ρομπότ κατακλύζεται από μια συνεχής ροή αντιληπτικών πληροφοριών. Για να μπορέσει να χρησιμοποιήσει αυτές τις πληροφορίες αποτελεσματικά για τον καθορισμό ενεργειών, ένα ρομπότ πρέπει να έχει την ικανότητα να λειτουργεί αφαιρετικά έτσι ώστε να επικεντρώνεται στα πιο σημαντικά στοιχεία του περιβάλλοντος. Βασιζόμενο σε αυτές τις αφαιρέσεις, το ρομπότ πρέπει να μπορεί να προβλέψει τις πιθανές αλλαγές στο περιβάλλον του, στην πάροδο του χρόνου, έτσι ώστε να προχωρήσει από μια απλή, αντανακλαστική συμπεριφορά σε μια πιο σύνθετη, ηθελημένη συμπεριφορά. Πιο σημαντικά, όλη η διαδικασία θα μπορεί να ωθείται από κίνητρα που πηγάζουν εκ των έσω τα οποία ωθούν το σύστημα προς μεγαλύτερες αφαιρέσεις και πιο πολύπλοκες προβλέψεις. Πιστεύουμε επίσης ότι η αναπτυξιακή διαδικασία πρέπει να εφαρμόζεται με ιεραρχικό, αυτοδύναμο τρόπο, έτσι ώστε να έχει ως αποτέλεσμα την δημιουργία μιας γκάμας συνεχούς αυξανόμενης επιτηδευμένης συμπεριφοράς. Ξεκινώντας με μια βασική, δημιουργημένα έμφυτη συμπεριφορά, το ρομπότ μπορεί να ασκεί τους αισθητήρες και την κίνησή του, να χρησιμοποιεί τους μηχανισμούς αφαίρεσης και πρόβλεψης, ώστε να "ανακαλύψει" την απλή αντανακλαστική συμπεριφορά. Έπειτα, ένα Self-Motivated σχήμα ελέγχου θα εκμεταλλευόταν αυτές τις ανακαλύψεις για να αντικαταστήσει την εγγενή συμπεριφορά του ρομπότ. Αυτό αποτελεί το πρώτο στάδιο της διαδικασίας αυτοδυναμίας. Ο ίδιος εγγενής αναπτυξιακός αλγόριθμος μπορεί να εφαρμοστεί περιοδικά σε μετέπειτα στάδια, χρησιμοποιώντας την γνώση που έχει ανακαλυφθεί στα προηγούμενα στάδια. Αυτές οι ακολουθίες συμπεριφοράς που δημιουργούνται μπορεί να χρησιμοποιηθούν για να οδηγήσουν το ρομπότ σε μια σειρά καταστάσεων στο περιβάλλον, κάτι που θα το οδηγήσει να "επισκεφθεί" περισσότερο "ενδιαφέρουσες" καταστάσεις, όπως καθορίζεται από το εσωτερικό του μοντέλο κινήτρων.

Για να ανακεφαλαιώσουμε λοιπόν, προτείνουμε μια πολυ-επίπεδη, πολυπρακτορική αρχιτεκτονική διαδοχικής ανακάλυψης και ελέγχου για να διερευνήσουμε σχήματα αναπτυξιακής ρομποτικής στο συγκεκριμένο τομέα του επιδέξιου χειρισμού. Στην ακόλουθη ενότητα θα παρουσιάσουμε μια σύντομη επισκόπηση και ταξινόμηση της σχετικής έρευνας που έχει διεξαχθεί και που είναι ακόμη σε εξέλιξη στον επιστημονικό κλάδο τον οποίον προσεγγίζουμε στην παρούσα διατριβή.

1.2 Συνοπτική Επισκόπηση Ερευνητικού Αντικειμένου

Κινητήριο μοχλό στην παρούσα διατριβή αποτελεί η βασική επιδίωξη για την σχεδίαση και ανάπτυξη μιας αρχιτεκτονικής ρομποτικού ελέγχου η οποία θα είναι ιδιαίτερα κλιμακωτή και επιπλέον θα επιτρέπει τον έλεγχο κινηματικά πολύπλοκων ρομποτικών συστημάτων που περιέχουν πολλαπλούς, πιθανά πλεονάζοντες βαθμούς ελευθερίας, σχηματίζοντας ανοικτές ή κλειστές, σύνθετες κινηματικές αλυσίδες. Είναι σημαντικό σε αυτό το πρώιμο στάδιο της παρούσης διατριβής να πραγματοποιήσουμε μια συνοπτική επισκόπηση των διαφορετικών ερευνητικών προσπαθειών που έχουν συντελεστεί σε αυτόν τον τομέα, καθώς και μια αντίστοιχη ταξινόμηση του ερευνητικού αυτού πεδίου για να διευκολύνουμε την κατανόηση του πλαισίου στο οποίο έχει την φιλοδοξία να συνεισφέρει η ερευνητική μας προσπάθεια.

1.2.1 Ρομποτική Μηχανική Μάθηση

Σημείο αφετηρίας για την παρούσα διατριβή αποτελεί ο μηχανισμός μάθησης που θα πρέπει να υιοθετηθεί από το ρομποτικό μας σύστημα έτσι ώστε να μπορεί αυτό να λειτουργεί αρκετά γρήγορα και με ασφάλεια σύμφωνα με τους περιορισμούς που τίθενται από ένα περιβάλλον το οποίο αλλάζει συνεχώς και το οποίο δε δύναται να έχει συγκεκριμένα (εκ των προτέρων δεδομένα) συμπεριφορικά μοντέλα τα οποία θα διέπουν τη λειτουργία μέσα σε αυτό. Η μάθηση λαμβάνει χώρα καθώς ο πράκτορας παρατηρεί τη διαδικασία αλληλεπίδρασής του με τον κόσμο καθώς επίσης και την αντίστοιχη δική του διαδικασία λήψης αποφάσεων. Η κύρια ιδέα πίσω από τη μάθηση είναι ότι οι πληροφορίες που εμπίπτουν στην αντίληψη ενός ρομπότ πρέπει να χρησιμοποιούνται όχι μόνο για την πραγματοποίηση των διαφόρων περιστασιακών ενεργειών, αλλά και για τη βελτίωση της συνολικής ικανότητας του πράκτορα να λειτουργεί στο μέλλον. Ο τομέας της μηχανικής μάθησης συνήθως διαχωρίζει τρεις περιπτώσεις

μαθησιακών προβλημάτων τα οποία έχουν να αντιμετωπίσουν οι πράκτορες: επιβλεπόμενη (*Supervised*) [29] , μη-επιβλεπόμενη (*Unsupervised*) [29] και ενισχυτική μάθηση (*Reinforcement Learning*) [120], [62] . Το πρόβλημα στην επιβλεπόμενη μάθηση είναι ότι το σύστημα καλείται να μάθει μια λειτουργία μέσω παραδειγμάτων που αφορούν σε συγκεκριμένα δεδομένα εισόδου-εξόδου. Στην περίπτωση όπου το περιβάλλον μπορεί να παρατηρηθεί πλήρως, θα ισχύει πάντα ότι ο ρομποτικός πράκτορας μπορεί να παρατηρεί τα αποτελέσματα των ενεργειών του και επομένως θα μπορεί να χρησιμοποιεί επιβλεπόμενες μεθόδους μάθησης για να μαθαίνει να τις προβλέπει. Το πρόβλημα είναι πιο δύσκολο με αυτή την προσέγγιση όταν το περιβάλλον είναι μερικώς παρατηρήσιμο, γιατί τα άμεσα αποτελέσματα των ενεργειών του ρομπότ μπορεί να μην είναι άμεσα ορατά. Το πρόβλημα με τη μη-επιβλεπόμενη μάθηση έχει να κάνει με τη μάθηση προτύπων στην είσοδο δεδομένων όταν δεν έχουν δοθεί συγκεκριμένες τιμές για τα προσδοκώμενα εξερχόμενα (Reference Output) δεδομένα. Ένας ρομποτικός πράκτορας που μαθαίνει χωρίς καμιά επίβλεψη δεν μπορεί να μάθει τι πρέπει να κάνει γιατί δεν έχει πληροφορίες ως προς τι αποτελεί μια σωστή ενέργεια ή μια επιθυμητή κατάσταση. Αυτός ο τύπος μαθησιακού προβλήματος μελετάται πρωταρχικά στο πλαίσιο των πιθανολογικών συστημάτων λογικής. Η τρίτη κατηγορία αφορά το πρόβλημα της Ενισχυτικής Μάθησης [120], [29]. Η βασική ιδέα αυτής της προσέγγισης είναι ότι αντί ένας "δάσκαλος" να υποδεικνύει στο σύστημα ποια ενέργεια είναι η ενδεδειγμένη σε μια συγκεκριμένη κατάσταση, το σύστημα μαθαίνει με έμμεσο τρόπο μέσω μιας συγκεκριμένης ανάδρασης η οποία του απονέμεται σαν συνέπεια κάθε ενέργειας την οποία αυτό εκτελεί. Η ανάδραση αυτή ονομάζεται "ανταμοιβή" ή "ανταπόδοση"(*reward*)[8], και παίζει το σημαντικό ρόλο να καταδεικνύει στο ρομποτικό σύστημα αν μια συγκεκριμένη συμπεριφορά, σε δεδομένη κατάσταση, οδηγεί σε επιθυμητό αποτέλεσμα ή όχι, και συνεπώς εάν μπορεί αυτή να θεωρηθεί (εκ του αποτελέσματος) ως ενδεδειγμένη, ή όχι, για τη δεδομένη κατάσταση.

I) Ενισχυτική Μάθηση και Πολυπρακτορικά Συστήματα

Αναφέρεται στη διεθνή βιβλιογραφία ένας σημαντικός αριθμός ερευνητικών προσπαθειών που έχουν εφαρμόσει κάποιου είδους τεχνικές μηχανικής μάθησης σε πολυπρακτορικά ρομποτικά συστήματα. Οι εργασίες αυτές έχουν εφαρμοσθεί ως επί το πλείστον σε ένα συγκεκριμένο τομέα της ρομποτικής, τα αυτοκινούμενα ρομποτικά συστήματα (*Mobile Robotics*) [123], [76], [15], [9]. Η διαπίστωση της έλλειψης ουσιαστικά εφαρμογών πολυ-πρακτορικών συστημάτων στον τομέα των επιδέξιων ρομποτικών χειριστών είναι κάτι που αξίζει να σημειώσουμε και σίγουρα αποτέλεσε ένα πολύ ισχυρό κίνητρο για την ερευνητική μας προσπάθεια. Η Ενισχυτική Μάθηση είναι ένας ενεργός ερευνητικός τομέας στην

περιοχή της μηχανικής μάθησης, ο οποίος επίσης ελκύει την προσοχή από τους τομείς της θεωρίας αποφάσεων και του αυτομάτου ελέγχου. Αλγόριθμοι ενισχυτικής μάθησης προσπαθούν να λύσουν το συγκεκριμένο πρόβλημα: πώς ένας πράκτορας ή μια ομάδα πρακτόρων μπορούν να μάθουν να προσεγγίζουν την καλύτερη στρατηγική συμπεριφοράς, ενώ αλληλεπιδρούν άμεσα με το περιβάλλον. Η ενισχυτική μάθηση είναι μια τεχνική μάθησης που βασίζεται στην εμπειρία. Μια σωστά επιλεγμένη ενέργεια φέρνει επιβράβευση, αυξάνοντας την πιθανότητα να επιλεγεί η ίδια συμπεριφορά και στο μέλλον [120]. Από την άλλη, μια εσφαλμένη ενέργεια προκαλεί τιμωρία, μειώνοντας την πιθανότητα επανάληψής της. Ως αποτέλεσμα, ένας πράκτορας ή μια ομάδα πρακτόρων μαθαίνουν πώς από συγκεκριμένες καταστάσεις στις οποίες μεταβαίνουν επιλέγουν συγκεκριμένες ενέργειες οι οποίες μεγιστοποιούν την επιβράβευση που δέχονται.

Συγκεκριμένες τεχνικές ενισχυτικής μάθησης όπως Q-Learning [95], [75], [140] και Actor-Critic [123], [134], [52] έχουν εφαρμοστεί μέχρι τώρα με επιτυχία σε πολυπρακτορικές αρχιτεκτονικές οι οποίες στοχεύουν στον έλεγχο αυτοκινούμενων ρομπότ που λειτουργούν σε περιβάλλον που είναι πλήρως παρατηρήσιμο, όχι όμως και στο πλαίσιο εφαρμογών επιδέξιου ρομποτικού χειρισμού σε αβέβαιο και δυναμικά μεταβαλλόμενο περιβάλλον, όπως επιδιώκουμε στην ερευνητική μας εργασία. Επιπλέον, τεχνικές τύπου Actor-Critic στις περισσότερες περιπτώσεις εφαρμόζονται συνήθως σε συνδυασμό με Νευρωνικά Δίκτυα, όχι σε πολυπρακτορικό περιβάλλον, σε εφαρμογές πλοήγησης κινούμενων ρομπότ, χρησιμοποιώντας ένα "χάρτη αυτο-οργάνωσης" (*self-organizing map*) . Εργασίες που έχουν μέχρι σήμερα ανακοινωθεί [42], [104], [38], [87], [8], αναφέρονται σε εφαρμογές κατά τις οποίες το ρομπότ κινείται στο χώρο του για αρκετό χρόνο ώστε να τον καταγράψει, να αποκτήσει ικανοποιητικά δεδομένα και έπειτα να τα οργανώσει χρησιμοποιώντας τον αλγόριθμο του Kohonen [74]. Στη συνέχεια εφαρμόζεται ένα δίκτυο νευρώνων το οποίο υλοποιεί τον αλγόριθμο ενισχυτικής μάθησης με σκοπό να πλοηγηθεί το αυτοκινούμενο ρομπότ μέσα στον χάρτη. Ένας προφανής περιορισμός μιας τέτοιας μεθοδολογίας είναι το γεγονός ότι αν ο χώρος αλλάξει, η διαδικασία της μάθησης πρέπει να αρχίσει από την αρχή αφού ο χάρτης αυτο-οργάνωσης πρέπει να επαναπροσδιοριστεί.

Η κατανόηση πτυχών της ανθρώπινης συνεργατικής συμπεριφοράς αποτελεί σαφώς αντικείμενο έρευνας στο πεδίο των πολυπρακτορικών συστημάτων. Στη ρομποτική, σχετικά θέματα έχουν ερευνηθεί στα πλαίσια εργασιών που πραγματοποιήθηκαν σε αυτοκινούμενα ρομπότ [25], ρομποτικά χέρια, και πολλαπλούς συνεργαζόμενους ρομποτικούς χειριστές [26] [27] [29]. Συγκεκριμένα, στη εργασία [31], αναπτύχθηκαν πρωτόκολλα χειρισμού για μια ομάδα αυτοκινούμενων ρομπότ έτσι ώστε να συνεργατικά να επιτύχουν να σπρώξουν μεγάλα κιβώτια. Στο [30] μια αλγοριθμική δομή συντονίζει τη διαδικασία αναπροσανατολισμού αντικειμένων στο επίπεδο από ανεξάρτητους ρομποτικούς πράκτορες. Στο [29] παρουσιάζεται μια μελέτη όπου κατανεμημένες στρατηγικές συνεργασίας για τη

διαχείριση αντικειμένων απαιτούνται από μια ομάδα αυτοκινούμενων ρομπότ τα οποία έχουν ενσωματωμένες συμπεριφορές. Το κοινό σημείο όλων αυτών των εργασιών είναι ότι η κίνηση των αντικειμένων που υπόκεινται σε χειρισμό κάποιας μορφής, είναι κατ'ουσίαν στατικές ("quasi-static") και επιπλέον όλοι οι πράκτορες που εμπλέκονται στη σχετική εργασία έχουν προκαθορισμένα μοντέλα συμπεριφορών τα οποία τα συνδυάζουν κάνοντας χρήση συγκεκριμένων αρχιτεκτονικών (π.χ. υπαγωγή (subsumption)[32]). Επιπλέον της συνεργατικότητας, η ανθρώπινη συμπεριφορά επιδεικνύει χαρακτηριστικά εξέλιξης και ικανότητα αυτο-οργάνωσης. Αυτές οι μοναδικές ιδιότητες της ανθρώπινης συμπεριφοράς έχουν μελετηθεί επαρκώς στη διαδικασία σχεδιασμού ευφυών ρομποτικών συστημάτων τα οποία θα πρέπει να λειτουργούν συνεργατικά αυτόνομα και προσαρμοστικά στο περιβάλλον τους. Σε αυτό το πλαίσιο λοιπόν η χρήση τεχνικών προερχόμενων από την βιολογία όπως είναι η ενισχυτική μάθηση, η εξελικτική υπολογιστική (evolutionary computation), καθώς και τα ασαφή συστήματα συνθέτουν ένα νέο πολυσύνθετο και διαθεματικό ερευνητικό πεδίο στο οποίο στοχεύουμε να συνεισφέρουμε με την παρούσα διατριβή.

Η ενισχυτική μάθηση [1] [2] [3] αποτελεί ένα σημαντικό μηχανισμό απόκτησης δεξιοτήτων ρομποτικού χειρισμού. Μια άλλη προσέγγιση απόκτησης σχετικών δεξιοτήτων βασίζεται σε διαδικασίες τύπου Μάθησης μέσω Επίδειξης (Learning from Demonstration - LfD) [35] [36], αναφερόμενη επίσης στη βιβλιογραφία και ως Μάθηση μέσω Μίμησης (Learning by Imitation) [34]. Μέσω αυτής της διαδικασίας, η πολιτική την οποία χρειάζεται το ρομποτικό σύστημα να σχηματίσει για να λειτουργήσει τη μαθαίνει όχι μέσω εξερεύνησης και διάδρασης με το περιβάλλον αλλά μέσω της επίδειξης την οποία του παρέχει ο διδάσκων. Το σύνολο των παραδειγμάτων που παρουσιάζονται κατά τη διάρκεια της επίδειξης, ορίζονται ως ακολουθίες από ζεύγη κατάστασης-δράσης (state-action) τα οποία καταγράφονται κατά την εξέλιξη της επίδειξης συγκεκριμένης συμπεριφοράς στο ρομποτικό σύστημα. Να σημειωθεί σε αυτό το σημείο ότι στο πλαίσιο της μάθησης LfD η πολιτική η οποία σχηματίζεται περιλαμβάνει μόνο τις καταστάσεις της οποίες περιείχε η επίδειξη και προφανώς τις αντίστοιχες δράσεις.

Εμπνευσμένη από τις παραπάνω ερευνητικές κατευθύνσεις, η παρούσα διατριβή προσεγγίζει μηχανισμούς απόκτησης δεξιοτήτων για ρομποτικά συστήματα (skills) όχι μέσω επίδειξης αλλά μέσω εξερεύνησης και διάδρασης του συστήματος με το περιβάλλον του. Αυτό, στο πλαίσιο της πολυπρακτορικής τοπολογίας που προτείνουμε, επιτρέπει στον κάθε πράκτορα να αναπτύσσει τοπικά αισθητικο-κινητικές (sensori-motor) συμπεριφορές, στη βάση των ανταμοιβών που εισπράττει, και οι οποίες αποτελούν ποιοτικό κριτήριο της απόδοσης του συστήματος. Η προτεινόμενη λοιπόν αρχιτεκτονική βασίζεται σε μια εμφωλευμένη και ιεραρχική δομή, όπου ο κάθε πράκτορας συνθέτει και συντηρεί τοπικά μερική - τμηματική εικόνα για τη συνολική κατάσταση του συστήματος, κα-

θώς και της εξέλιξης της σχετικής εργασίας, μέσω μιας αναδρομικής (τύπου top-down / bottom-up) διαδικασίας.

Βασικό στόχο, λοιπόν, στην παρούσα διατριβή αποτελεί ο σχεδιασμός ενός αναπτυξιακού μηχανισμού ελέγχου, ο οποίος βασισμένος σε μια πολυπρακτορική ιεραρχική αρχιτεκτονική θα επιτρέπει σε ένα ρομποτικό σύστημα να αποκτά με αυτόνομο τρόπο νέες δεξιότητες. Τα πειράματα τα οποία παρουσιάζονται στη παρούσα διατριβή στοχεύουν στην αποτίμηση και αξιολόγηση της προτεινόμενης μεθοδολογίας σε τρείς φάσεις: (α) σε μια πρώτη φάση, σε προβλήματα κινηματικού ελέγχου ρομποτικών αλυσίδων με πλεονάζοντες βαθμούς ελευθερίας, (β) σε δεύτερη φάση σε προβλήματα στατικού ελέγχου παράλληλων κινηματικών αλυσίδων οι οποίες συνεργάζονται με στόχο την επίτευξη στατικής ρομποτικής λαβής, και τέλος (γ) σε προβλήματα συνεργατικού ελέγχου αυτοκινούμενων ρομποτικών οχημάτων.

II) Συνεχής Χώρος-Κατάστασης

Ένα βασικό εμπόδιο για τον τύπο του προβλήματος που προσπαθούμε να λύσουμε είναι η συνέχεια (Continuity) του Χώρου-Κατάστασης (State-Space) . Μια ενδιαφέρουσα προσέγγιση ως προς την αρχιτεκτονική ενός πράκτορα που χειρίζεται συνεχή Χώρο-Κατάστασης με μια αριθμητική μέθοδο ενισχυτικής μάθησης (χρησιμοποιώντας πρωτογενή, μη δομημένα, αισθητηριακά δεδομένα) παρουσιάζεται στο [52]. Για την εφαρμογή μιας τέτοιας προσέγγισης απαιτείται κατ΄ αρχάς εκ των προτέρων μια καταγραφή του χώρου εργασίας του ρομπότ. Με την εκ των προτέρων αυτή καταγραφή, επιλέγονται και καταχωρούνται συγκεκριμένα χαρακτηριστικά του χώρου κίνησης (πόρτες, διάδρομοι, δωμάτια κλπ.). Με βάση τη θέση του ρομπότ αναφορικά με αυτά τα χαρακτηριστικά δημιουργείται μια σειρά εμπειρικών κανόνων. Αυτοί οι κανόνες συνδέουν τα χαμηλά επίπεδα εισροής των αισθητηριακών σημάτων (sensory signals) με την υψηλού επιπέδου αντιληπτική γνώση (cognitive knowledge) που παρουσιάζει το ρομπότ. Μια ξεχωριστή μονάδα ελέγχου του κεντρικού ελεγκτή χειρίζεται ένα συγκεκριμένο χαρακτηριστικό του χώρου. Η αρχιτεκτονική που προκύπτει μοιάζει σαν ένα σύνολο "ειδικών" όπου ο καθένας είναι εξειδικευμένος στον έλεγχο μιας ξεχωριστής μονάδας του συστήματος. Στην περίπτωση αυτή, η μάθηση συντελείται στη βάση μιας μονοπρακτορικής δομής, υπό την έννοια ότι, μολονότι κάθε μονάδα ελέγχου εκπαιδεύεται να χειρίζεται συγκεκριμένη εργασία, ο κεντρικός ελεγκτής συμπεριφέρεται σαν ένας και μόνος πράκτορας.

Στην ερευνητική μας εργασία, η μεθοδολογία της ενισχυτικής μάθησης εφαρμόζεται σε ένα ασαφή χώρο καταστάσεων (*fuzzy state-space*). Επιδιώκουμε, κατ΄ αυτόν τον τρόπο, το σχεδιασμό μιας μεθοδολογίας ελέγχου που θα λειτουργεί σε ένα συνεχή χώρο, και η οποία θα επιτρέπει στους πράκτορες να

μαθαίνουν με την πάροδο του χρόνου πώς να εκτελούν, σε συνεργασία, αλληλουχίες συνεχών κινήσεων για να επιτύχουν τον στόχο τους, χωρίς πρότερη γνώση της εργασίας. Οι πράκτορες, οι οποίοι αντιστοιχούν σε αυτόνομους βαθμούς ελευθερίας στο ρομποτικό μας σύστημα, επιτυγχάνουν να αποκτήσουν εμπειρία στην εργασία που εκτελούν από κοινού, εξερευνώντας και αξιοποιώντας το δικό τους χώρο "μετασχηματισμού κατάστασης-δράσης" (state-to-action mapping space).

III) Ιεραρχικές Αρχιτεκτονικές Ρομποτικού Ελέγχου

Η έννοια της κατανομής δεξιοτήτων είναι στοιχείο το οποίο υιοθετήθηκε στα πλαίσια της εργασίας μας. Η κατανομή τώρα των δεξιοτήτων σε ξεχωριστούς πράκτορες, έτσι ώστε να προκύπτει μια πολυπρακτορική αρχιτεκτονική, αποτελεί βασική επιλογή στο πλαίσιο ελέγχου το οποίο προτείνουμε. Μια ερευνητική κατεύθυνση που έχει επηρεάσει ιδιαίτερα την εργασία μας, υπό το πρίσμα της συνέχειας του χώρου-κατάστασης, προέρχεται από το έργο του Doya [34]. Στο έργο αυτό προτείνεται ένας αλγόριθμος συνεχούς χώρου-κατάστασης με σκοπό τη διευκόλυνση της εφαρμογής τεχνικών ενισχυτικής μάθησης σε εφαρμογές ελέγχου πραγματικών συνθηκών. Για την αξιολόγηση του συγκεκριμένου πλαισίου, έχει εφαρμοστεί ένας μηχανισμός μάθησης Actor-Critic, ο οποίος υλοποιήθηκε με δίκτυα "Ακτινικών Συναρτήσεων Βάσης" (Radial Basis Functions), στο πρόβλημα ελέγχου ενός ανάστροφου εκκρεμούς με περιορισμένη ροπή.

Μια πολλά υποσχόμενη προσέγγιση που ονομάζεται "Ιεραρχική Πολυπρακτορική Ενισχυτική Μάθηση" έχει αναλυθεί εκτενώς στις [42], [36], [37], [8] με σκοπό πρώτον να βελτιώσει τις πολυπρακτορικές αρχιτεκτονικές για την εκτέλεση πολύπλοκων εργασιών, αναλύοντας αυτές σε υπο-στόχους, και δεύτερον για την επίτευξη πολλαπλών στόχων. Η μεθοδολογία την οποία προτείνουμε στην παρούσα διατριβή και στην οποία εστιάζει η ερευνητική μας προσπάθεια, ενσωματώνει συγκεκριμένες ιδέες που εισήχθησαν από την Ιεραρχική Πολυπρακτορική Ενισχυτική Μάθηση. Για παράδειγμα, χρησιμοποιούμε εξίσου την έννοια των στρατηγικών υψηλού επιπέδου που αυτόματα ανακαλύπτουν υπο-στόχους, ενώ οι στρατηγικές χαμηλού επιπέδου μαθαίνουν να ειδικεύονται σε διαφορετικούς υπο-στόχους. Προσπαθήσαμε να ενσωματώσουμε τέτοιες έννοιες σε μια εμφωλευμένη (nested-agent) αρχιτεκτονική, η οποία θα παρουσιαστεί στα επόμενα κεφάλαια, και να την εφαρμόσουμε στον τομέα του ελέγχου επιδέξιου ρομποτικού χειρισμού. Στο παράδειγμα που παρουσιάζεται στο Κεφ. 4 της διατριβής, οι σύνδεσμοι (links) του ρομποτικού χειριστή θεωρούνται ξεχωριστοί πράκτορες που αξιοποιούν μεθόδους ενισχυτικής μάθησης για να αναπτύξουν

συγκεκριμένες ικανότητες. Κατά συνέπεια, από αυτό το σημείο και έπειτα, ο ίδιος ο ρομποτικός χειριστής υφίσταται ως ένα πολυπρακτορικό περιβάλλον, όπου οι πράκτορες που τον απαρτίζουν αναλαμβάνουν ατομικούς στόχους, και μέσω της εμπειρίας καταφέρνουν να αποκτήσουν συγκεκριμένες ικανότητες / δεξιότητες. Αν και αυτόνομοι, οι πράκτορες είναι στενά συνδεδεμένοι ο ένας με τον άλλο εξαιτίας της φυσικής διασύνδεσης, κάνοντας ιδιαίτερα σημαντική την ακριβή συνεργασία και συντονισμό μεταξύ τους έτσι ώστε να επιτευχθεί η επιθυμητή συμπεριφορά του συστήματος. Η εργασία μας στο πλαίσιο του πολυπρακτορικού ρομποτικού περιβάλλοντος, προτείνει μια κατανεμημένη αρχιτεκτονική ρομποτικού ελέγχου που ενσωματώνει τεχνικές τεχνητής νοημοσύνης και κλασικές μεθοδολογίες ελέγχου, και βασίζεται στην υλοποίηση μιας ομάδας ρομποτικών πρακτόρων που μαθαίνει σε ένα συνεχή χώρο καταστάσεων.

1.2.2 Ευφυής Ρομποτικός Έλεγχος

Το είδος των προβλημάτων που προσπαθούμε να λύσουμε έχουν σημαντικό βαθμό πολυπλοκότητας και θέτουν στόχους που πρέπει να επιτευχθούν σε συνθήκες μεγάλης αβεβαιότητας. Είναι φανερό ότι οι μεθοδολογίες εύρωστου ελέγχου με προκαθορισμένη σταθερή ανατροφοδότηση (fixed feedback robust controllers) δεν μπορούν να αντιμετωπίσουν τα προβλήματα αυτά, στο συγκεκριμένο πλαίσιο που αναλύθηκε στις προηγούμενες παραγράφους. Αυτό που χρειάζεται είναι μια ενισχυμένη μεθοδολογία ελέγχου που θα χρησιμοποιεί αυτές τις συμβατικές προσεγγίσεις ελέγχου για να επιλύει προβλήματα ελέγχου σε ένα χαμηλότερο επίπεδο, πάνω στο οποίο πρέπει να δημιουργηθεί ένα ευφυές πλαίσιο το οποίο και θα το διαχειρίζεται. Αυτό έχει οδηγήσει στην ανάπτυξη υβριδικών συστημάτων ελέγχου, τα οποία ελέγχουν δυναμικές διαδικασίες συνεχούς χώρου κατάστασης (continuous-state) με ελεγκτές οι οποίοι λειτουργούν σε διακριτό χώρο (discreet-state). Στην περίπτωσή μας, το πρόβλημα αποκτά ακόμα μεγαλύτερες δυσκολίες αφού εισάγεται, όπως έχει ήδη αναφερθεί, και ο παράγοντας της επιδεξιότητας και πολυπλοκότητας στη διαδικασία χειρισμού.

I) Γενικές Αρχές Υβριδικού Ελέγχου

Η εφαρμογή κάποιου σχήματος υβριδικού ελέγχου αποτελεί μια βασική (και ουσιαστικά αναγκαία) επιλογή στην παρούσα διατριβή, ως προς την επίτευξη των ερευνητικών μας στόχων. Ένας παραδοσιακός ελεγκτής χαμηλού επιπέδου, συνδυαζόμενος με μία αυτόνομη μονάδα τεχνητής νοημοσύνης και λειτουργώντας από κοινού σε ένα ιεραρχικό σχηματισμό, συνθέτουν την εικόνα

του ελεγκτή που χρειαζόμαστε. Αυτή η προσέγγιση έχει υιοθετηθεί σχεδόν σε όλες τις προσπάθειες στο συγκεκριμένο τομέα ερευνών που εξετάζουμε [121], [18], [96], [57], [60]. Ωστόσο, η πλειονότητα των ερευνητικών αυτών προσπαθειών εστιάζουν αποκλειστικά σε εφαρμογές αυτοκινούμενων ρομπότ. Σε ένα γενικότερο ερευνητικό πλαίσιο, όπως αυτό που αναπτύσσεται στην παρούσα διατριβή, πρέπει να ορισθούν κάποια γενικά χαρακτηριστικά τα οποία καλείται να πληρεί ένας πρότυπος (μη-συνήθης) ρομποτικός ελεγκτής. Πιο συγκεριμένα, ένα τέτοιο ρομποτικό σύστημα ελέγχου πρέπει να ικανοποιεί ένα σύνολο βασικών προδιαγραφών και σχεδιαστικών απαιτήσεων, οι οποίες περιγράφονται πιο αναλυτικά ακολούθως:

- Να αντιδρά στο περιβάλλον - πρέπει ο ελεγκτής να είναι σε θέση να αντιδρά σε ξαφνικές αλλαγές στο περιβάλλον και να είναι ικανός να λαμβάνει υπόψη εξωτερικά περιστατικά με χρονικά όρια που είναι συμβατά με τη σωστή, αποτελεσματική και ασφαλή εκτέλεση των εργασιών του.

- Να επιδεικνύει νοήμονα συμπεριφορά - αυτό απαιτεί να γίνονται διάφοροι συμβιβασμοί με βάση τους κανόνες της κοινής λογικής ώστε ο ελεγκτής να επιδεικνύει ευφυή συμπεριφορά. Οι αντιδράσεις του ρομπότ σε εξωτερικά ερεθίσματα πρέπει να καθοδηγούνται από τους στόχους της κύριας εργασίας του.

- Να επιτυγχάνει λύση πολλαπλών στόχων - είναι αναπόφευκτο ότι θα προκύψουν καταστάσεις που θα απαιτούν αντικρουόμενες παράλληλες ενέργειες. Το σύστημα ελέγχου θα πρέπει να παρέχει τα μέσα για την εκπλήρωση αυτών των πολλαπλών στόχων.

- Σθεναρότητα ή Ευρωστία (Robustness) - το ρομπότ πρέπει να χειρίζεται ατελή εισροή δεδομένων, μη αναμενόμενες καταστάσεις και ξαφνικές βλάβες.

- Αξιοπιστία - η ικανότητα της λειτουργίας χωρίς αποτυχίες ή χειροτέρευση της απόδοσης στην πάροδο του χρόνου.

- Επαναπρογραμματισμός - πρέπει το σύστημα ελέγχου να είναι ικανό να επιτύχει πολλαπλές εργασίες που περιγράφονται σε κάποιο αφαιρετικό επίπεδο, αντί για μία μόνο εργασία.

- Τμηματικότητα (Modularity) - το σύστημα ελέγχου πρέπει να διαχωρίζεται σε μικρότερα υπο-συστήματα (ή υπο-μονάδες) που μπορούν να σχεδιαστούν, υλοποιηθούν, ελεγχθούν και συντηρηθούν ξεχωριστά και προσθετικά.

- Ευελιξία (Flexibility) - η πειραματική ρομποτική απαιτεί συνεχώς αλλαγές στο σχεδιασμό στη διάρκεια της φάσης υλοποίησης. Επομένως, απαιτούνται ευέλικτες δομές ελέγχου για να επιτραπεί στο σχεδιασμό να οδηγηθεί από την επιτυχία ή την αποτυχία των ατομικών στοιχείων.

- Προσαρμοστικότητα (Adaptability) - καθώς η κατάσταση του κόσμου αλλάζει δραματικά και απρόβλεπτα, το σύστημα ελέγχου πρέπει να είναι προσαρμόσιμο έτσι ώστε να μπορεί να αλλάζει ομαλά και ταχύτατα μεταξύ διαφορετικών στρατηγικών ελέγχου.

- Γενική χρήση λογικής σκέψης - απαιτείται ένας γενικός πράκτορας λήψης αποφάσεων υψηλού επιπέδου, υπεύθυνος για την κατανόηση της γενικής κατάστασης, ο οποίος θα εξετάζει τα λάθη που εισάγονται εξαιτίας της παρανόησης των αισθητήριων δεδομένων και θα συγχωνεύει τις μερικώς διαθέσιμες πληροφορίες.

II) Επιδέξιος Ρομποτικός Χειρισμός

Είναι ιδιαίτερα σημαντικό να υπογραμμίσουμε την έλλειψη μιας επαρκούς υβριδικής πολυπρακτορικής μεθοδολογίας ελέγχου στον τομέα του επιδέξιου ρομποτικού χειρισμού. Η ερευνητική μας προσπάθεια, όπως αυτή παρουσιάζεται στην παρούσα διατριβή, αποσκοπεί στο να συνεισφέρει στην ανάπτυξη ενός τέτοιου συγκεκριμένου πλαισίου ρομποτικού ελέγχου. Μια σχετική ερευνητική προσπάθεια παρουσιάζεται από τον Borst [16], όπου περιγράφεται μια μέθοδος για τη δημιουργία μιας διάταξης ρομποτικού χεριού για την ευσταθή λαβή και χειρισμό ενός αντικειμένου με δεδομένη ομάδα σημείων επαφής. Το ρομποτικό χέρι DLR-II που χρησιμοποιείται ως πλατφόρμα στην παραπάνω εργασία αποτελεί σίγουρα ένα από τα σημαντικότερα επιτεύγματα της σύγχρονης ρομποτικής και μηχατρονικής τεχνολογίας στον τομέα του επιδέξιου χειρισμού. Μια άλλη συναφής προσπάθεια παρουσιάζεται από τους Pollard, Hodgins [99], οι οποίοι περιγράφουν μια μέθοδο για την προσαρμογή ενός παραδείγματος εργασίας επιδέξιου χειρισμού, από ένα συγκεκριμένο αντικείμενο σε ένα νέο. Μια εξίσου εντυπωσιακή εργασία παρουσιάζεται από τους Martin, Ambrose, Diftler, Platt & Butzer [88], όπου παρουσιάζεται μια μέθοδος ελέγχου για την επαναληπτική βελτίωση της "ποιότητας" λαβής ενός αντικειμένου με άγνωστη γεωμετρία, χρησιμοποιώντας μόνο απτικές (haptic -tactile) αισθητηριακές πληροφορίες.

Αυτό που διαφαίνεται, σε όλες αυτές τις ιδιαίτερα σημαντικές ερευνητικές δραστηριότητες είναι η προσπάθεια για ανάπτυξη μιας υβριδικής προσέγγισης πολυπρακτορικής αρχιτεκτονικής, με στοιχεία τεχνητής νοημοσύνης, με εφαρμογή στο συγκεκριμένο, και πολύ απαιτητικό τομέα του ρομποτικού χειρισμού.

Εξετάζοντας τις σχετικές εργασίες επισκόπησης που αναφέρονται στην εξέλιξη που έχει πραγματοποιηθεί τα τελευταία χρόνια στον τομέα του επιδέξιου χειρισμού, αξίζει να αναφέρουμε κάποιες προτάσεις από την εργασία των Rodney Brooks, Leslie Kaelbling, Trevor Darrell & Push Singh [117] το Σεπτέμβριο του 2004: '*... Τι θα χρειαζόταν για να κατασκευάσουμε μια μηχανή ικανή για επιδέξιο χειρισμό αντικειμένων σε συνηθισμένο, μη δομημένο ανθρώπινο περιβάλλον· Πολλές από τις υποθέσεις που έγιναν από πρότερες εργασίες στον επιδέξιο χειρισμό πρέπει να επανεξεταστούν και πιθανότατα να εγκαταλειφθούν. Υποψιαζόμαστε ότι η λύση αυτού του προβλήματος θα απαιτήσει εντελώς νέες προσεγγίσεις...*', και συνεχίζει: '*... στη mobile ρομποτική υπήρξε ιδιαίτερη αλληλεπίδραση με το έργο που διεξάγεται στην τεχνητή νοημοσύνη, ειδικά το έργο στη μηχανική μάθηση και στις πιθανολογικές αναπαραστάσεις. Βλέπε για παράδειγμα το ενδιαφέρον σε θέματα ταυτόχρονου αυτοεντοπισμού θέσεως και χαρτογράφησης χώρου (simultaneous localization and mapping), μια κατηγορία τεχνικών που αφορά την πλοήγηση κινούμενων ρομπότ σε άγνωστα περιβάλλοντα. Όμως, είναι εντυπωσιακό το γεγονός ότι δεν υπάρχει καμιά σχέση μεταξύ του έργου που διεξάγεται στην τεχνητή νοημοσύνη και του έργου στον επιδέξιο χειρισμό. Είναι πιθανό ότι αυτό το κενό οφείλεται σε αυτό που φαίνεται να είναι μια μείωση στο έργο που αναφέρεται στον επιδέξιο χειρισμό κατά τη δεκαετία του 1990, όταν αναπτύχθηκαν πολλές από τις σύγχρονες πιθανολογικές θεωρίες μάθησης. Αυτές οι τεχνικές σχεδιάστηκαν για να μαθαίνουν μοντέλα του περιβάλλοντος καθώς και να δρουν με βάση αυτά, και επομένως είναι κατάλληλες για την περίπτωση του επιδέξιου χειρισμού σε μη δομημένα περιβάλλοντα. Μου φαίνεται ότι ο επιδέξιος χειρισμός είναι ένα είδος προβλήματος που μπορεί να ωφεληθεί ιδιαίτερα από τη μηχανική μάθηση...*'.

Πιστεύουμε λοιπόν, από τη μεριά μας, ότι ο τομέας του επιδέξιου ρομποτικού χειρισμού αποτελεί ένα από τα πιο ενδιαφέροντα και σημαντικά θέματα που αντιμετωπίζει η ρομποτική σήμερα, και πως μέσα από αυτά τα οποία προτείνουμε στο πλαίσιο της παρούσας διατριβής θα δώσουμε κάποιες κατευθύνσεις για περαιτέρω διερεύνηση, έτσι ώστε τελικά να οδηγηθούμε σε μεθοδολογίες οι οποίες θα αντιμετωπίζουν σε κάποιο βαθμό τις προκλήσεις που αναφέραμε παραπάνω.

1.3 Συνεισφορές της Διατριβής

Οι επιστημονικές συνεισφορές της παρούσας διατριβής εστιάζουν σε δύο επίπεδα. Το πρώτο επίπεδο περιλαμβάνει τον σχεδιασμό και την ανάπτυξη μιας πολυπρακτορικής αρχιτεκτονικής για τον έλεγχο ρομποτικών συστημάτων. Μέ-

σω μιας ιεραρχικής - εμφωλευμένης αρχιτεκτονικής, η οποία παρουσιάζει χαρακτηριστικά αυξημένης επεκτασιμότητας σε σύνθετα ρομποτικά προβλήματα, επιτυγχάνεται η δημιουργία ενός ελεγκτή ο οποίος επιδεικνύει σημαντική ευρωστία. Πιο συγκεκριμένα, το ρομποτικό σύστημα το οποίο ενσωματώνει την προτεινόμενη αρχιτεκτονική, επιδεικνύει σημαντική ανοχή σε απρόβλεπτες αστοχίες που ανακύπτουν σε δομικές του μονάδες.

Στο δεύτερο επίπεδο, η συνεισφορά της παρούσας διατριβής έγκειται στον σχεδιασμό και στην υλοποίηση μιας μεθοδολογίας ρομποτικής μάθησης βάσει τεχνικών ασαφούς ενισχυτικής μάθησης, προσαρμοσμένης σε συνεχή χώροκατάστασης. Με την προτεινόμενη μεθοδολογία αναπτυξιακής ρομποτικής μάθησης επιτρέπεται η αυτο-ανάπτυξη και η αυτο-οργάνωση του πολυπρακτορικού συστήματος χωρίς την ανάγκη εκ νέου μοντελοποίησης και επαναπρογραμματισμού.

Οι συνεισφορές της διατριβής και στα δύο επίπεδα, αξιολογήθηκαν σε δύο κατηγορίες προβλημάτων: α) στον κινηματικό έλεγχο ρομποτικών αλυσίδων με πλεονάζοντες βαθμούς ελευθερίας, και β) σε εφαρμογές συνεργαζόμενου ρομποτικού χειρισμού. Τα συγκριτικά αποτελέσματα, σε σχέση με συγκεντρωτικές (βάσει μοντέλου) προσεγγίσεις, δείχνουν την αποδοτικότητα της προτεινόμενης πολυπρακτορικής αρχιτεκτονικής κατανεμημένης μάθησης, τόσο σε σχέση με εγγενή χαρακτηριστικά επεκτασιμότητας και γενίκευσης της αποκτηθείσας γνώσης, όσο και σε σχέση με την ευρωστία σε πιθανές μη μοντελοποιημένες αστοχίες υπομονάδων του συστήματος.

Οι βασικοί στόχοι της παρούσας διατριβής, που συνιστούν την ουσιαστική επιστημονική συμβολή της, είναι οι ακόλουθοι:

1. *Ανάπτυξη ευέλικτης, κλιμακωτής, πολυπρακτορικής αρχιτεκτονικής ελέγχου, εφαρμόσιμης σε προβλήματα επιδέξιου ρομποτικού χειρισμού.*
   Στο πλαίσιο της παρούσας διατριβής, δημιουργήσαμε απλά, βασικά δομικά στοιχεία τα οποία αντιστοιχίζονται άμεσα σε απλές φυσικές οντότητες, όπως ρομποτικοί σύνδεσμοι ή τροχοί αυτοκινούμενων ρομπότ. Αυτές οι απλές ομοιογενείς οντότητες μπορούν να συνδυαστούν για να σχηματίσουν δομές που καταφέρνουν να επιλύουν πιο περίπλοκες εργασίες αλλά ταυτόχρονα να διατηρούν την απλότητά τους. Το αποτέλεσμα λοιπόν είναι η υλοποίηση μιας κατανεμημένης πολυπρακτορικής αρχιτεκτονικής η οποία είναι ιεραρχική, επιδεικνύει αρθρωτή δομή, και το πεδίο εφαρμογής της μπορεί να καλύπτει τον τομέα του επιδέξιου ρομποτικού χειρισμού.

2. *Ανάπτυξη μεθόδου ενισχυτικής ρομποτικής μάθησης εφαρμόσιμης σε*

συνεχή χώρο-κατάστασης.
Βασική συνεισφορά της παρούσας διατριβής είναι η ανάπτυξη μιας μεθόδου ενισχυτικής μάθησης εφαρμόσιμης σε συνεχή χώρο-καταστάσεων. Επιπλέον, μέσω της προτεινόμενης αναπτυξιακής μάθησης, το πολυπρακτορικό σύστημα επιτυγχάνει την προσαρμογή του σε δυναμικά μεταβαλλόμενα περιβάλλοντα, επιδεικνύοντας ικανότητες αυτο-οργάνωσης και αυτοανάπτυξης. Η συνεξελικτική μάθηση στην προτεινόμενη πολυπρακτορική τοπολογία, επιτυγχάνει την επίλυση του συνολικού (γενικευμένου) στόχου, μέσω των στρατηγικών που μαθαίνουν οι πράκτορες που συνθέτουν το σύστημα. Πιο συγκεκριμένα, ο μηχανισμός ενισχυτικής μάθησης που προτείνεται συνθέτει μια γενικευμένη αντιστοίχιση καταστάσεων-δράσεων για την επίλυση του συνολικού στόχου, μέσω μιας δυναμικής ανάλυσης του συνολικού στόχου σε υπο-στόχους, όπου για κάθε έναν από αυτούς ο κάθε πράκτορας καλείται να αναπτύξει τοπικές στρατηγικές (αντιστοιχίσεις καταστάσεων-δράσεων). Επιπλέον όφελος είναι ότι μέσω αυτού του μηχανισμού κατανέμεται το υπολογιστικό βάρος της μάθησης στο σύνολο των πρακτόρων που απαρτίζουν το σύστημα.

3. *Σχεδίαση ενός σχήματος υβριδικού ρομποτικού ελέγχου.*
Επόμενος στόχος που συνιστά ουσιαστική επιστημονική συμβολή της παρούσας διατριβής είναι ο επιτυχής συνδυασμός κλασσικής μεθοδολογίας ελέγχου, με υπολογιστικά μοντέλα τεχνητής νοημοσύνης, δημιουργώντας έτσι ένα υβριδικό σχήμα, το οποίο μπορεί να χρησιμοποιηθεί για τον έλεγχο συγκεκριμένων ρομποτικών διατάξεων (απλή κινηματική αλυσίδα τεσσάρων συνδέσμων, κινηματική αλυσίδα επτά συνδέσμων, πολυαρθρωτή ρομποτική λαβή, καθώς και για συνεργαζόμενα αυτοκινούμενα ρομπότ).

4. *Παρουσίαση και ανάλυση παραδειγμάτων εφαρμογής*
Τέλος, μέσω της παρουσίασης και της αντίστοιχης ανάλυσης παραδειγμάτων εφαρμογής, επιτυγχάνεται να αξιολογηθεί η προτεινόμενη μεθοδολογία αυτόνομης (αναπτυξιακής) ρομποτικής μάθησης δεξιοτήτων, σε προβλήματα ελέγχου επιδέξιου ρομποτικού χειριστή, καθώς και σε αυτοκινούμενα ρομπότ, ενσωματώνοντας την προτεινόμενη πολυπρακτορική τοπολογία. Τα παραδείγματα εφαρμογής που διερευνήθηκαν στην παρούσα διατριβή περιλαμβάνουν: (α) επίπεδο ρομποτικό χειριστή με πλεονάζοντες βαθμούς ελευθερίας (τοποθέτηση εργαλείου ρομποτικού χειριστή τεσσάρων συνδέσμων, και επέκταση σε ταυτόχρονη αποφυγή εμποδίων για ρομποτικό χειριστή επτά συνδέσμων), (β) πολυαρθρωτή ρομποτική λαβή και (γ) ρομποτικό χειρισμό αντικειμένου ("box-pushing") μέσω δύο συνεργαζόμενων αυτοκινούμενων οχημάτων.

1.4 Οργάνωση της Διατριβής

Η δομή της διατριβής ακολουθεί μια σειριακή λογική όπου το κάθε κεφάλαιο υποστηρίζεται από τα προηγούμενα. Πιο συγκεκριμένα, η οργάνωση της διατριβής είναι η εξής:

Το 2ο κεφάλαιο παρουσιάζει αναλυτικά την πολυπρακτορική αρχιτεκτονική που σχεδιάσαμε. Στο πρώτο μέρος του σχετικού κεφαλαίου, μελετάμε το θεωρητικό υπόβαθρο πάνω στο οποίο βασίζεται η μετέπειτα ανάλυση του πολυπρακτορικού μας συστήματος. Επιπλέον, αναλύουμε τα επιμέρους τμήματα που συνθέτουν έναν πράκτορα καθώς και την διασύνδεση μεταξύ τους, έτσι ώστε να επιτύχουμε το σχηματισμό μιας δομημένης πολυπρακτορικής τοπολογίας. Ακόμα, στο κεφάλαιο αυτό αναλύονται όλες οι πειραματικές πολυπρακτορικές τοπολογίες που αξιολογούνται (κινηματικές αλυσίδες και αυτοκινούμενα ρομπότ). Τέλος, παρουσιάζονται κάποια γενικά χαρακτηριστικά πολυπρακτορικών συστημάτων, τα οποία ενσωματώνει η αρχιτεκτονική που προτείνεται στην παρούσα διατριβή, όπως εξελικτική μάθηση, συνεργατική συμπεριφορά, επιμερισμένη επεξεργασία, ανοχή στα λάθη, και αυτονομία.

Το 3ο κεφάλαιο επικεντρώνεται στο μηχανισμό μάθησης που εφαρμόζουμε, σε συνάρτηση με την πολυπρακτορική αρχιτεκτονική που προτείνουμε στις διαφορετικές πειραματικές τοπολογίες που εξετάζουμε. Το πρώτο μέρος του 3ου κεφαλαίου αναλύει το θεωρητικό υπόβαθρο των προβλημάτων ενισχυτικής μάθησης ως προβλήματα Μαρκοβιανών διαδικασιών λήψης αποφάσεων. Στη συνέχεια του κεφαλαίου αναλύονται οι δύο κύριες οικογένειες αλγορίθμων ενισχυτικής μάθησης (Επανάληψη Πολιτικής / Επανάληψη Αξίας). Η ανάλυση στο κεφάλαιο αυτό επεκτείνεται στην περιγραφή του συγκεκριμένου μηχανισμού ενισχυτικής μάθησης για το σύνολο των πολυπρακτορικών τοπολογιών.

Το 4ο κεφάλαιο αναλύει το σχήμα ελέγχου για την προτεινόμενη πολυπρακτορική αρχιτεκτονική. Παρουσιάζει επίσης, αναλυτικά όλα τα παραδείγματα εφαρμογής (κινηματικές αλυσίδες, ρομποτική λαβή και αυτοκινούμενα ρομπότ), μέσω των οποίων προκύπτει μια συνολική αξιολόγηση της προτεινόμενης εμφωλευμένης-ιεραρχικής τοπολογίας καθώς και του αντίστοιχου αλγόριθμου μάθησης που αναπτύξαμε. Στο κεφάλαιο αυτό περιγράφονται και αναλύονται στο σύνολό τους όλα τα πειραματικά ευρήματα τα οποία προέκυψαν στα πλαίσια της παρούσας διατριβής.

Το 5ο κεφάλαιο περιλαμβάνει τα συμπεράσματα που προκύπτουν από την παρούσα διατριβή, ενώ υπογραμμίζει με τρόπο συγκριτικό τα ιδιαίτερα χαρακτηριστικά της προτεινόμενης πολυπρακτορικής τοπολογίας, σε σχέση με άλλες

ερευνητικές εργασίες. Τέλος, το κεφάλαιο ανακεφαλαιώνει τα γενικά ευρήματα που παρουσιάστηκαν στην παρούσα διατριβή ενώ παράλληλα προτείνει μελλοντικούς δρόμους για την επέκταση της συγκεκριμένης ερευνητικής προσπάθειας, καθώς και περαιτέρω θέματα τα οποία χρήζουν έρευνας.

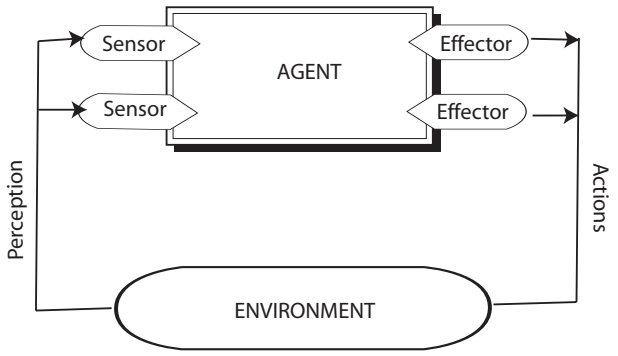
Κεφάλαιο 2

Πολυπρακτορική Ρομποτική Αρχιτεκτονική

2.1 Εισαγωγή

Παρά το γεγονός ότι ο όρος "Πράκτορας" (Agent) έχει χρησιμοποιηθεί ευρύτατα την τελευταία δεκαετία στα πλαίσια διαφορετικών προβλημάτων (και συχνά χωρίς να αιτιολογείται η χρήση του), δεν διαφαίνεται να έχει καθιερωθεί ένας ορισμός γενικά αποδεκτός από την επιστημονική κοινότητα. Θα προσπαθήσουμε ωστόσο να επισημάνουμε μερικά από τα βασικά χαρακτηριστικά του, τα οποία είναι σημαντικά σε συνάρτηση με την παρούσα διατριβή. Ως πράκτορα λοιπόν, μπορούμε να ορίσουμε οποιαδήποτε μονάδα μπορεί να θεωρηθεί ότι αντιλαμβάνεται το περιβάλλον μέσω αισθητήρων και επενεργεί σε αυτό το περιβάλλον μέσω επενεργητών (effectors) [36], κάτι που φαίνεται στο Σχήμα 2.1. Σε αυτό λοιπόν το πλαίσιο, η ευφυΐα πολύ συχνά χρησιμοποιείται ως επιθυμητή ιδιότητα ενός πράκτορα. Μπορούμε επίσης να πούμε ότι ένας λογικός ή νοήμων πράκτορας είναι εκείνος που πάντα επιλέγει τη σωστή δράση, σύμφωνα με κάποιο κριτήριο εξωτερικής απόδοσης και δεδομένης της τρέχουσας γνώσης του. Επομένως, το κίνητρο για μάθηση σε ένα λογικό πράκτορα μπορεί να αναλυθεί στην στρατηγική εφαρμογή της γνώσης που έχει αποκτηθεί για δικό του όφελος. Η αυτονομία είναι επίσης μια άλλη ιδιότητα που συχνά αποδίδεται στους πράκτορες. Ως αυτόνομος μπορεί να οριστεί ένας πράκτορας ο οποίος έχει έλεγχο και στις εσωτερικές του ιδιότητες (παραμέτρους) και στην συμπεριφορά του. Η απαίτηση για αυτονομία αυξάνεται εάν αναλογιστεί κανείς ότι κατά τη διάρκεια του σχεδιασμού ενός πράκτορα δεν θεωρείται εφικτή μια απολύτως επιτυχημένη πρόβλεψη των ιδιοτήτων και της ποικιλομορφίας του εξωτερικού περιβάλλοντος το οποίο θα χρειαστεί να αντιμετωπίσει ο πράκτορας. Στην εργασία [135] η ικανότητα προσαρμογής σε ποικίλες περιβαλλοντικές συνθήκες αναφέρεται ως ευέλικτη αυτονομία. Αν μια ομάδα A πρακτόρων επενεργεί στο ίδιο περιβάλλον ταυτόχρονα, αυτό το θεωρούμε ότι αποτελεί ένα πολυπρακτορικό σύστημα (multiagent system - MAS) [131] .

Τα τελευταία χρόνια τα πολυπρακτορικά συστήματα έχουν γίνει ολοένα και πιο σημαντικά σε διάφορους τομείς (τεχνητή νοημοσύνη, κατανεμημένα συστήματα και ρομποτική) με την εισαγωγή ζητημάτων συλλογικής νοημοσύνης και αλληλεπίδρασης. Με επίκεντρο την αυτονομία των πρακτόρων και τις αλληλεπιδράσεις που τους ενώνουν, έχουν ανακύψει πολλά ζητήματα που απαιτούν περαιτέρω έρευνα. Αυτά τα ζητήματα προκύπτουν κυρίως λόγω της διασύνδεση του πεδίου των πολυπρακτορικών συστημάτων με άλλους τομείς, όπως οι γνωσιακές επιστήμες, η εξελικτική επιστήμη, η νευρολογία και η φιλοσοφία. Πραγματοποιείται λοιπόν μία εκτεταμένη έρευνα η οποία έχει ως στόχο την ανάπτυξη πολυπρακτορικών συστημάτων τα οποία εξελίσσονται με την πάροδο του χρόνου, αποκτούν εμπειρία και προσαρμόζονται στις περιβαλλοντικές αλλαγές. Αυτό επιτυγχάνεται μέσω: α) της δυνατότητας αυτο-οργάνωσης των

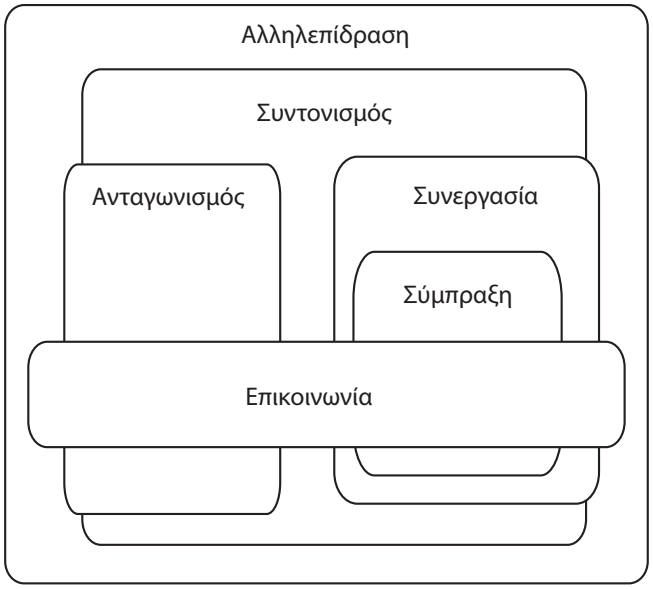


Σχήμα 2.1: *Πράκτορας που αλληλεπιδρά με το περιβάλλον*

πρακτόρων που συνθέτουν το σύστημα, β) μέσω της συνεργασίας των πρακτόρων μεταξύ τους και τέλος γ) μέσω της αλληλεπίδρασής τους με το περιβάλλον. Αποτέλεσμα αυτών των διεργασιών είναι τα πολυπρακτορικά συστήματα να αναπτύσσουν συμπεριφορές που τους επιτρέπουν να πραγματοποιούν εν τέλει περίπλοκες εργασίες.

Δεδομένου ότι δύναται να υπάρξουν καταστάσεις στις οποίες ένας πράκτορας μπορεί να λειτουργεί παραγωγικά μόνος του, η γενική υπόθεση είναι ότι θα πρέπει να αλληλεπιδρά με άλλους πράκτορες. Το συγκεκριμένο είδος αλληλεπίδρασης μπορεί να κυμαίνεται από έναν ανταγωνισμό πόρων έως μια μορφή κατανεμημένου κοινού σχεδιασμού. Το Σχήμα 2.2 δείχνει τα διαφορετικά επίπεδα αλληλεπίδρασης, καταδεικνύοντας ότι ανάμεσα στους πιθανούς τρόπους αλληλεπίδρασης που λαμβάνουν χώρα, το σημαντικότερο στοιχείο είναι η επικοινωνία.

Η σημασία της επικοινωνίας αυξάνει ακόμη περισσότερο σε ένα πολυπρακτορικό σύστημα που σχηματίζεται από ένα πλήθος πρακτόρων (με διαφορετικούς ατομικούς στόχους ο καθένας), όπου ο κάθε πράκτορας του συστήματος δύναται να αλλάζει δυναμικά συμπεριφορά, με τρόπο που δεν ήταν γνωστός κατά την αρχική φάση σχεδιασμού της πολυπρακτορικής τοπολογίας. Αντιλαμβανόμαστε ότι σε αυτό το σημείο φτάνουμε στα όρια των προκαθορισμένων πρωτοκόλλων επικοινωνίας και αλληλεπίδρασης καθώς επίσης και ότι η ικανότητα μάθησης συμπεριφορών με τρόπο δυναμικό αποκτά κρίσιμη σημασία.



Σχήμα 2.2: *Διάφορες μορφές (επίπεδα) αλληλεπίδρασης πρακτόρων*

2.2 Τι Είναι Πράκτορας

Πριν ασχοληθούμε με τον πολυπρακτορικό σχηματισμό, ας κατανοήσουμε καλύτερα τι σημαίνει "Πράκτορας" στο συγκεκριμένο πλαίσιο της παρούσης διατριβής. Ο πράκτορας είναι μια φυσική ή εικονική οντότητα η οποία: α) μπορεί να λειτουργεί σε ένα περιβάλλον, β) μπορεί να επικοινωνήσει άμεσα με άλλους πράκτορες, γ) ωθείται από μια σειρά τάσεων (με τη μορφή ατομικών σκοπών/στόχων που προσπαθεί να βελτιστοποιήσει), δ) έχει δικούς του πόρους, ε) μπορεί να αντιλαμβάνεται το περιβάλλον, στ) έχει μόνο μερική αναπαράσταση του περιβάλλοντος και ζ) εδραιώνει συμπεριφορές προς την ικανοποίηση των στόχων που θέτει, λαμβάνοντας υπόψη τους πόρους και τις δεξιότητες που είναι διαθέσιμες, ενώ εξαρτάται στην αντίληψη που αναπτύσσει και την επικοινωνία που πραγματοποιεί. Στις πειραματικές διατάξεις που μελετάμε στα πλαίσια της παρούσης διατριβής, ο πράκτορας ο οποίος δύναται να επενεργεί είναι: είτε ένας σύνδεσμος (εν γένει, ένα σύνολο εσωτερικών βαθμών ελευθερίας) ενός ρομποτικού χειριστή (μιας κινηματικής αλυσίδας), είτε ένας τροχός ενός αυτοκινούμενου οχήματος. Η έννοια της δράσης βασίζεται στο γεγονός ότι οι πράκτορες πραγματοποιούν ενέργειες ανεξάρτητα ο ένας από τον άλλο, οι οποίες μεταβάλλουν το περιβάλλον τους και επομένως τις μελλοντικές λήψεις αποφάσεων που θα χρειαστεί να πραγματοποιηθούν. Επίσης, μπορούν να επικοινωνούν μεταξύ τους με συγκεκριμένη μεθοδολογία η οποία ορίζεται από

την ιεραρχική αρχιτεκτονική η οποία προτείνεται. Η μεταξύ τους επικοινωνία αποτελεί έναν από τους κύριους μηχανισμούς με τον οποίο αλληλεπιδρούν οι πράκτορες. Στο συγκεκριμένο πλαίσιο που προτείνουμε, οι πράκτορες έχουν αυτονομία. Αυτό σημαίνει ότι δεν κατευθύνονται από εντολές που προέρχονται από εξωτερικό ως προς το σύστημα παρατηρητή αλλά από μια σειρά τάσεων ή στόχων που θέτουν οι ίδιοι, στο πλαίσιο πραγματοποίησης του συνολικού (γενικευμένου) στόχου του πολυπρακτορικού συστήματος.

Είναι ιδιαίτερα σημαντικό να υπογραμμίσουμε ότι η δομή του πράκτορα στην παρούσα διατριβή είναι ένα υβριδικό μοντέλο που συνδυάζει χαρακτηριστικά "γνωσιακών" (cognitive) καθώς και "αντανακλαστικών" (reactive) πρακτόρων. Οι γνωσιακοί πράκτορες έχουν τους στόχους τους, καθώς και τα αντίστοιχα σχέδια που τους επιτρέπουν να επιτυγχάνουν αυτούς τους στόχους. Κάθε πράκτορας έχει διαθέσιμη μια βάση γνώσης, που αποτελείται από όλα τα δεδομένα και το σχετικό μηχανισμό που απαιτείται για να ολοκληρώσει την εργασία του και για να χειρίζεται την αλληλεπίδραση με τους άλλους πράκτορες και το περιβάλλον. Ο αντανακλαστικός πράκτορας δεν χρειάζεται να είναι ατομικά νοήμων, για να επιδεικνύει το συνολικό σύστημα ευφυή συμπεριφορά. Οι γνωσιακοί πράκτορες είναι λογικοί, δηλαδή οι δράσεις που εγείρουν ακολουθούν τις αρχές της λογικής σε σχέση με τους στόχους που προσπαθούν να επιτύχουν. Οι αντανακλαστικοί πράκτορες από την άλλη, αντιδρούν μόνο σε ερεθίσματα από το περιβάλλον, ενώ η συμπεριφορά τους ελέγχεται ολοκληρωτικά από την τοπική κατάσταση του χώρου στον οποίο βρίσκονται. Οι γνωσιακοί πράκτορες, από τη μια, λόγω των επιτηδευμένων στοιχείων του χαρακτήρα τους και της ικανότητάς τους για λογική σκέψη όσον αφορά τον κόσμο στον οποίον εντάσσονται, μπορούν να λειτουργούν με σχετικά ανεξάρτητο τρόπο. Οι αντανακλαστικοί πράκτορες, από την άλλη, επιδεικνύουν πιο άκαμπτη συμπεριφορά, καθώς βασίζονται μόνο στους δικούς τους πόρους. Η δύναμή τους έγκειται στην ικανότητα τους να σχηματίζουν ομάδες. Έτσι οι αντανακλαστικοί πράκτορες αποκτούν ενδιαφέρον όχι τόσο πολύ σε ατομικό επίπεδο όσο σε επίπεδο πληθυσμών, και αναφορικά με τις ικανότητες για προσαρμογή και εξέλιξη που πηγάζουν από τις αλληλεπιδράσεις των μελών τους.

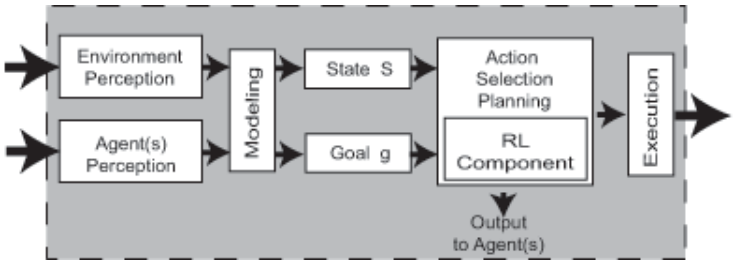
Ένας πράκτορας χαρακτηρίζεται τόσο από την αρχιτεκτονική του, όσο και από τη συμπεριφορά του. Η αρχιτεκτονική απεικονίζει το πώς τα διάφορα μέρη του πράκτορα μπορούν να συναρμολογηθούν ώστε να επιτύχει τις εργασίες που αναμένεται να πραγματοποιήσει. Επομένως, η αρχιτεκτονική ενός πράκτορα χαρακτηρίζει την εσωτερική δομή του, δηλαδή την οργάνωση και την διαρρύθμιση των διαφόρων στοιχείων (μονάδων) που τον συνθέτουν. Το άλλο σημαντικό χαρακτηριστικό του πράκτορα είναι η συμπεριφορά του. Η συμπεριφορά χαρακτηρίζει όλες τις ιδιότητες που ο πράκτορας εκδηλώνει στο περιβάλλον του,

με άλλα λόγια τις λειτουργίες του. Στο πρώτο τμήμα της επόμενης ενότητας θα εξετάσουμε τύπους αρχιτεκτονικής πρακτόρων που αναφέρονται στη διεθνή βιβλιογραφία, ενώ θα ολοκληρώσουμε αυτό το τμήμα με την ανάλυση της αρχιτεκτονικής δομής (τόσο σε επίπεδο πράκτορα όσο και σε επίπεδο συστήματος πρακτόρων) που προτείνουμε στο πλαίσιο της παρούσης διατριβής.

2.3 Μονοπρακτορική Εσωτερική Δομή

Βασικό στοιχείο σε έναν πράκτορα αποτελεί ο τρόπος με τον οποίο είναι οργανωμένος. Η οργάνωση μπορεί να οριστεί ως μια διευθέτηση σχέσεων μεταξύ συστατικών ή ατόμων που συνθέτουν μια μονάδα, ή ένα σύστημα, το οποίο επιδεικνύει ιδιότητες που δεν κατανοούνται στο επίπεδο των συστατικών ή των ατόμων. Η οργάνωση διασφαλίζει έναν σχετικά υψηλό βαθμό αλληλεξάρτησης και αξιοπιστίας, πιστοποιώντας έτσι το σύστημα με την πιθανότητα να διατηρηθεί για μια συγκεκριμένη χρονική περίοδο, παρόλη την τυχαία διαταραχή [35]. Ο όρος που γενικά χρησιμοποιείται για να περιγράψει την εσωτερική οργάνωση του πράκτορα είναι η αρχιτεκτονική. Η αρχιτεκτονική ενός πράκτορα χαρακτηρίζεται βασικά από τρεις παραμέτρους: τη σχεδιαστική προσέγγιση που επιλέχθηκε (λειτουργική ή αντικειμενοστραφής), τη δομή εξάρτησης (ιεραρχική ή ισότητας) και τον τρόπο διασύνδεσης, ο οποίος ενώνει τον ένα πράκτορα με τον άλλο (σταθερός, ποικίλος ή εξελικτικός). Βασιζόμενοι στο συνδυασμό των παραπάνω παραμέτρων μπορούμε να διακρίνουμε τις ακόλουθες αρχιτεκτονικές: Τμηματικές (Modular) αρχιτεκτονικές, αρχιτεκτονικές βασισμένες σε μοντέλα "Μαυροπίνακα" (Blackboard-based) , αρχιτεκτονικές Υπαγωγής (Subsumption), αρχιτεκτονικές Ανταγωνιστικές και τέλος "Συνδετικές" (connectionist) αρχιτεκτονικές.

Η Τμηματική αρχιτεκτονική είναι σίγουρα η πιο διαδεδομένη. Οι αρχιτεκτονικές που βασίζονται σε αυτή την προσέγγιση καθορίζουν μια διακριτή σειρά μονάδων που ενώνονται με προκαθορισμένες συνδέσεις και που έχουν συγκεκριμένη λειτουργία. Η προτεινόμενη στην παρούσα διατριβή αρχιτεκτονική όπως θα δούμε και στη συνέχεια παρουσιάζει τέτοια χαρακτηριστικά. Συνεχίζοντας, η αρχιτεκτονική "Μαυροπίνακα" βασίζεται στην ιδέα ότι οι υπο-μονάδες είναι ανεξάρτητες και δεν επικοινωνούν δεδομένα άμεσα, αλλά αλληλεπιδρούν εμμέσως. Η αρχιτεκτονική Υπαγωγής είναι η προσέγγιση κατά την οποία αποσυνθέτουμε έναν πράκτορα σε κάθετες υπο-μονάδες, η κάθε μία από τις οποίες είναι υπεύθυνη για έναν περιορισμένο τύπο συμπεριφοράς. Η αλληλεπίδραση μεταξύ των υπο-μονάδων είναι σταθερή, και επιτυγχάνεται μέσω μιας "σχέσης κυρίαρχου" (dominance relationship), όπως αυτή καθορίζεται στο στάδιο σχεδιασμού. Οι υπο-μονάδες εκτελούν τις εργασίες τους παράλληλα, και σε περίπτωση που δύο



Σχήμα 2.3: *Προτεινόμενη εσωτερική οργάνωση ενός πράκτορα*

υπο-μονάδες παράγουν αντικρουόμενα αποτελέσματα, τα δεδομένα που παρέχονται από την κυρίαρχη υπο-μονάδα είναι αυτά που υπερισχύουν και λαμβάνονται τελικώς υπόψη. Οι Ανταγωνιστικές αρχιτεκτονικές βασίζονται στο γεγονός ότι οι διασυνδέσεις μεταξύ των διαφόρων υπο-μονάδων δεν είναι σταθερές, κάτι που σημαίνει ότι επιτρέπουν κάποιο βαθμό μεταβλητότητας μεταξύ των συνδέσεων των υπο-μονάδων, παρέχοντας τη δυνατότητα της επιλογής ως προς το ποιές υπο-μονάδες να εισάγονται στην πραγματοποίηση μιας εργασίας. Οι Συνδετικές αρχιτεκτονικές βασίζονται σε ένα δίκτυο παρόμοιων μονάδων επεξεργασίας που διασυνδέονται όλες μαζί, με την κάθε μία να εκτελεί μία απλή λειτουργία. Κάθε μονάδα καθορίζει την λειτουργία της με βάση τις αξίες που λαμβάνει από τις άλλες μονάδες. Η προσέγγισή μας στη παρούσα διατριβή όπως θα δούμε στη συνέχεια προφανώς και καταδεικνύει ένα σύστημα συνδετικό (Connectionist System) όπου κάθε πράκτορας αντιστοιχεί σε ένα στοιχείο ενός μεγάλου δικτύου πρακτόρων.

Το σχήμα 2.3, αποτυπώνει την προτεινόμενη εσωτερική οργάνωση του πράκτορα, καθώς και τα σήματα διασύνδεσης και επικοινωνίας που ανταλλάσσονται μεταξύ των διαφορετικών υπο-μονάδων του πράκτορα καθώς επίσης ανάμεσα στον πράκτορα και το περιβάλλον, με σκοπό να επιτευχθεί ο επιθυμητός έλεγχος. Τα βασικά τμήματα αυτής της αρχιτεκτονικής, με τις αντίστοιχες εισόδους και εξόδους (inputs, outputs) καθώς και οι απαιτούμενες προϋποθέσεις περιγράφονται στη συνέχεια.

1. *Μονάδα Αντίληψης Περιβάλλοντος (Environment Perception)*
   Η συγκεκριμένη μονάδα είναι υπεύθυνη για την διασύνδεση του πράκτορα με το περιβάλλον του. Πιο συγκεκριμένα, μέσω αυτής της μονάδας καταγράφονται οι μετρήσεις από τους αισθητήρες και στη συνέχεια αντιστοιχίζονται σε παραμέτρους που ορίζουν την κατάσταση του πράκτορα.

   Input: Τα δεδομένα από τους αισθητήρες του πράκτορα

Output: Ένα σύνολο παραμέτρων που προκύπτουν από την επεξεργασία των αισθητηριακών δεδομένων, και υποστηρίζουν τον καθορισμό της κατάστασης του πράκτορα για τη συγκεκριμένη χρονική στιγμή.

Requirements: Ο συγκεκριμένος πράκτορας έχει ένα σύνολο αισθητήρων που δύναται να χρησιμοποιήσει (δυνητικά το ίδιο και οι άλλοι πράκτορες του συστήματος).

2. *Μονάδα Αντίληψης Πράκτορα/Πρακτόρων (Agent(s) Perception)*
Η συγκεκριμένη μονάδα είναι υπεύθυνη για την διασύνδεση του πράκτορα με το σύνολο των πρακτόρων που συνθέτουν το σύστημα (εν γένει, με το σύνολο των πρακτόρων που είναι "ορατοί" από το συγκεκριμένο πράκτορα του συστήματος). Μέσω αυτής της μονάδας συγκεντρώνονται και καταγράφονται οι σχετικές παράμετροι που προσδιορίζουν την κατάσταση στην οποία βρίσκονται, τη δεδομένη χρονική στιγμή, όλοι οι (ορατοί) πράκτορες του συστήματος. Στη συνέχεια η μονάδα αντίληψης επεξεργάζεται τα δεδομένα αυτά έτσι ώστε να καταφέρει να καθορίσει τις παραμέτρους οι οποίες θα προσδιορίζουν επίσης την κατάσταση του πράκτορα σε σχέση με τους υπόλοιπους.

   Input: Οι πολυπλεγμένες πληροφορίες από όλους τους άλλους ("ορατούς", "γειτονικούς" ή γενικά στατικά ή δυναμικά "διασυνδεδεμένους") πράκτορες που συνθέτουν την παρούσα αρχιτεκτονική. Οι πληροφορίες αυτές απαιτούνται για τον επαρκή προσδιορισμό της κατάστασης στην οποία βρίσκεται ο πράκτορας.

   Output : Η κατάσταση στην οποία βρίσκεται πλέον ο πράκτορας, όπως αυτή ορίζεται σε συνάρτηση με τις καταστάσεις στις οποίες βρίσκονται οι υπόλοιποι πράκτορες της αρχιτεκτονικής, δημιουργώντας έτσι τη συνολική δομή στην οποία θα αποτυπώνονται επακριβώς όλοι οι άλλοι πράκτορες του συστήματος.

3. *Μονάδα Μοντελοποίησης Εργασιών (Modeling)*
Η συγκεκριμένη μονάδα έχει ως αντικείμενο την επεξεργασία των δεδομένων που προέρχονται τόσο από τους αισθητήρες όσο και από τους άλλους πράκτορες του συστήματος τη δεδομένη χρονική στιγμή. Βάσει λοιπόν αυτών των στοιχείων, προσδιορίζεται η υφιστάμενη κατάσταση του συγκεκριμένου πράκτορα καθώς και ποιός είναι ο στόχος του για τη δεδομένη χρονική στιγμή. Έχοντας λοιπόν τα δύο στοιχεία α) υφιστάμενη κατάσταση του πράκτορα και β) ποιός είναι ο υφιστάμενος στόχος, δημιουργείται ένα τοπικό μοντέλο εργασίας για τον πράκτορα το οποίο εξελίσσεται με την πάροδο του χρόνου.

Input : Επεξεργασμένα δεδομένα τόσο από τους αισθητήρες όσο και από τους άλλους πράκτορες.

Output : Υφιστάμενη κατάσταση πράκτορα και στόχος για τη δεδομένη χρονική στιγμή.

Requirements : Στην προτεινόμενη πολυπρακτορική αρχιτεκτονική δεν υφίσταται κεντρικό μοντέλο εργασίας στο πράκτορα, πράγμα που θα σήμαινε δυσκολία στην προσπάθεια επέκτασης του πράκτορα να καλύψει και άλλες διαφορετικές εργασίες. Ο κάθε πράκτορας δημιουργεί αυξητικά ένα τοπικό μοντέλο εργασίας βασισμένο στην αλληλουχία των αλληλεπιδράσεών του με το περιβάλλον και τη σχετική ανάδραση που λαμβάνει από αυτό. Η αρχιτεκτονική της παρούσης διατριβής κάνει χρήση ενισχυτικής μάθησης (βλ. Κεφάλαιο 3) έτσι ώστε να βελτιστοποιεί το δυναμικά σχηματιζόμενο, τοπικό μοντέλο εργασίας.

4. *Μονάδα, Επιλογής Δράσης - Σχεδιασμού - Ενισχυτικής Μάθησης (Action Selection - Planning - RL)*

Η συγκεκριμένη μονάδα έχει ως αντικείμενο την επιλογή δράσης, λαμβάνοντας υπόψη την υφιστάμενη κατάσταση του πράκτορα καθώς επίσης και την προηγούμενη εμπειρία (εάν υφίσταται) και το στόχο που έχει προσδιοριστεί για τη δεδομένη χρονική στιγμή. Η μονάδα αυτή περιλαμβάνει και το κομμάτι της ενισχυτικής μάθησης (reinforcement learning - RL). Η διαδοχική επιλογή δράσεων (αποφάσεων) από πλευράς πράκτορα λειτουργεί ως εσωτερικός μηχανισμός μορφοποίησης μιας συγκεκριμένης δομής, η οποία περιλαμβάνει συσχετισμούς μεταξύ καταστάσεων από τις οποίες περνά ο πράκτορας, σε δράσεις που δύναται να επιλέξει ο πράκτορας και σχετικές ανταποδόσεις που έχουν προκύψει στο παρελθόν. Μέσω αυτού του μηχανισμού, προκύπτει με την πάροδο του χρόνου ένας σχεδιασμός για τις δράσεις που θα επιλέξει στο μέλλον ο πράκτορας, και συνεπώς έχουμε έναν πράκτορα ο οποίος σταδιακά μαθαίνει (χτίζοντας, αυξητικά, συσχετισμούς πιθανών καταστάσεων με δυνατές δράσεις).

Input: Υφιστάμενη κατάσταση του πράκτορα, στόχος για τη δεδομένη χρονική στιγμή.
Output : Επιλεγμένη δράση, νέα κατάσταση στην οποία μεταβαίνει ο πράκτορας, καθώς και ο στόχος προς επίτευξη.

Requirements : Η αρχιτεκτονική η οποία υιοθετούμε για τον έλεγχο αυτού του πολυπρακτορικού συστήματος λειτουργεί στη βάση συμπεριφορικών μοντέλων. Δεν υφίσταται γενική μονάδα σχεδιασμού δράσεων.

Εκείνο που υπάρχει είναι ένα τοπικό πλάνο δράσεων το οποίο είναι ενσωματωμένο σε μια γενική αντιστοίχιση κατάστασης-δράσης που υπάρχει. Η αρχιτεκτονική μας κάνει χρήση ενισχυτικής μάθησης έτσι ώστε να βελτιστοποιεί το δυναμικά σχηματιζόμενο, τοπικό μοντέλο δράσεων.

5. *Εκτελεστική Μονάδα (Execution)*
   Η συγκεκριμένη μονάδα έχει ως αντικείμενο να προκαλεί την κίνηση των επενεργητών του πράκτορα.
   Input: Τη δράση που τελικώς επιλέχθηκε.

   Output : Σήμα αναφοράς στον αντίστοιχο επενεργητή. Αναλυτική περιγραφή για τις λειτουργίες που πραγματοποιούνται σε αυτή τη μονάδα θα δούμε και σε επόμενο τμήμα αυτής της διατριβής (βλ. Κεφάλαιο 4).

2.4 Τι είναι το Πολυπρακτορικό Σύστημα

Έχοντας καθορίσει τι είναι πράκτορας, τώρα μπορούμε να εξετάσουμε τον όρο πολυπρακτορικό σύστημα (multi-agent system). Ένα σύστημα χαρακτηρίζεται ως πολυπρακτορικό όταν αποτελείται από τα ακόλουθα στοιχεία: α) ένα περιβάλλον E, δηλαδή, ένα χώρο που γενικά έχει κάποιο όγκο, β) μια σειρά αντικειμένων O, τα οποία είτε βρίσκονται, είτε δύναται να βρεθούν σε οποιαδήποτε χρονική στιγμή, σε μια θέση στο E, και μπορούν να γίνουν αντιληπτά, να δημιουργηθούν, να καταστραφούν και να τροποποιηθούν από τους πράκτορες, γ) μια συνάθροιση πρακτόρων A που αντιπροσωπεύουν τις ενεργές οντότητες του συστήματος, δ) ένα σύνολο σχέσεων R, οι οποίες ενώνουν αντικείμενα (και επομένως πράκτορες) μεταξύ τους, ε) μια σειρά λειτουργιών Op , που καθιστούν εύκολο για τους πράκτορες του A να αντιληφθούν, παράγουν, καταναλώσουν, μετατρέψουν και χειριστούν αντικείμενα από το O [35]. Το πολυπρακτορικό σύστημα το οποίο εξετάζουμε στην παρούσα διατριβή βρίσκεται σε ένα περιβάλλον το οποίο είναι μερικώς γνωστό και του οποίου ο όγκος δύναται να επεκταθεί ή να μειωθεί δυναμικά. Επιπλέον, το πολυπρακτορικό σύστημα περιλαμβάνει ένα σύνολο ενεργών πρακτόρων των οποίων ο αριθμός δύναται να μεταβληθεί λόγω απρόβλεπτων καταστάσεων που προκύπτουν, ενώ υπάρχει ένα αρχικό σύνολο σχέσεων, το οποίο μέσω μιας επαναλαμβανόμενης, εξελικτικής διαδικασίας διαφοροποιείται. Ουσιαστικά αυτό που πραγματοποιείται είναι ότι οι σχέσεις αυτές συνεχώς επαναπροσδιορίζονται και περιστασιακά απαλείφονται.

2.5 Προτεινόμενο Πολυπρακτορικό Πλαίσιο

Το προτεινόμενο πολυπρακτορικό πλαίσιο αποτελεί ουσιαστικά μια προσπάθεια δημιουργίας μιας υβριδικής αρχιτεκτονικής η οποία συνδυάζει τεχνικές κατανεμημένης τεχνητής νοημοσύνης με κλασικό ρομποτικό έλεγχο. Το ένα επίπεδο αποτελείται από ένα σύνολο ιεραρχικά οργανωμένων πρακτόρων το οποίο ενσωματώνει δυνατότητες Ενισχυτικής Μάθησης [7], ενώ το άλλο αποτελείται από ένα κλασικό τοπικό βρόχο ελέγχου με ανάδραση, υπεύθυνο να οδηγεί τους αντίστοιχους επενεργητές του συστήματος. Επιπλέον στόχος για το προτεινόμενο πολυπρακτορικό σύστημα είναι η δημιουργία ενός ρομποτικού συστήματος το οποίο θα επιδεικνύει τα ακόλουθα πολύ σημαντικά χαρακτηριστικά: α) ικανότητα αυτόνομης απόκτησης δεξιοτήτων, μέσω πειραματισμού, με τη χρήση μηχανισμών μη-επιβλεπόμενης μάθησης και χωρίς την εφαρμογή κάποιου προ-σχεδιασμένου πλάνου ενεργειών, β) τμηματικότητα (modularity) και άμεση επεκτασιμότητα σε τοπολογίες υψηλότερου βαθμού συνθετότητας, χωρίς να απαιτείται σημαντική τροποποίηση των δομών ελέγχου και της αλγοριθμικής υλοποίησης, γ) ευρωστία (robustness) σε μεταβολές της εσωτερικής δομής του συστήματος, καθώς επίσης δυνατότητα κατανεμημένης επεξεργασίας (decentralization) και παραλληλίας (parallelism). Πιστεύουμε λοιπόν ότι η χρήση πολυπρακτορικών αρχιτεκτονικών σε συνδυασμό με μεθοδολογίες ασαφούς ενισχυτικής μάθησης (που μελετάμε στο Κεφάλαιο 3) δημιουργεί ρομποτικά συστήματα τα οποία επιδεικνύουν τα παραπάνω χαρακτηριστικά. Τα στοιχεία και οι προϋποθέσεις για τη δημιουργία του προτεινόμενο πολυπρακτορικού πλαισίου συνοψίζονται στη συνέχεια:

1. Κάθε ρομποτική άρθρωση στην περίπτωση της κινηματικής αλυσίδας, όπως και κάθε τροχός στην περίπτωση των αυτοκινούμενων οχημάτων, ορίζεται ως ένας πράκτορας ο οποίος λειτουργεί σε τοπικό επίπεδο. Η πρόκληση στη συγκεκριμένη προσπάθεια έγκειται στη δημιουργία συνολικής επιδέξιας συμπεριφοράς σε επίπεδο συστήματος μέσω σταδιακής απόκτησης δεξιοτήτων στο τοπικό επίπεδο του πράκτορα. Να σημειώσουμε εδώ ότι κάθε πράκτορας στη συγκεκριμένη αρχιτεκτονική που προτείνουμε δύναται να ενσωματώνει (εμπεριέχει) περισσότερους του ενός βαθμούς ελευθερίας, όπως θα δούμε άλλωστε και στη συνέχεια.

2. Κάθε πράκτορας λειτουργεί τοπικά παρατηρώντας τις δράσεις των υπολοίπων πρακτόρων τη δεδομένη χρονική στιγμή και πραγματοποιώντας παράλληλα προβλέψεις για τις δράσεις που δυνητικά εκείνοι πραγματοποιούν. Επιπλέον, υπάρχει μέτρηση της συνολικής απόδοσης του πολυπρακτορικού συστήματος η οποία και κατευθύνει τη διαδικασία της ενισχυτικής μάθησης, μέσω κάποιας συνάρτησης ανταπόδοσης. Περισσότερα για

τη συνάρτηση ανταπόδοσης θα δούμε στο επόμενο κεφάλαιο της διατριβής όπου και δούμε και τη διαδικασία ενισχυτικής μάθησης στο σύνολό της. Επιγραμματικά σημειώνουμε επί του παρόντος ότι το πολυπρακτορικό σύστημα λαμβάνει ανταμοιβή (θετική ή αρνητική) για τις ενέργειες που πραγματοποιεί σε συνεχή κλίμακα (και όχι σε διακριτά χρονικά σημεία), γεγονός που του επιτρέπει μια συνεχή, δυναμικά προσαρμοζόμενη, συμπεριφορά.

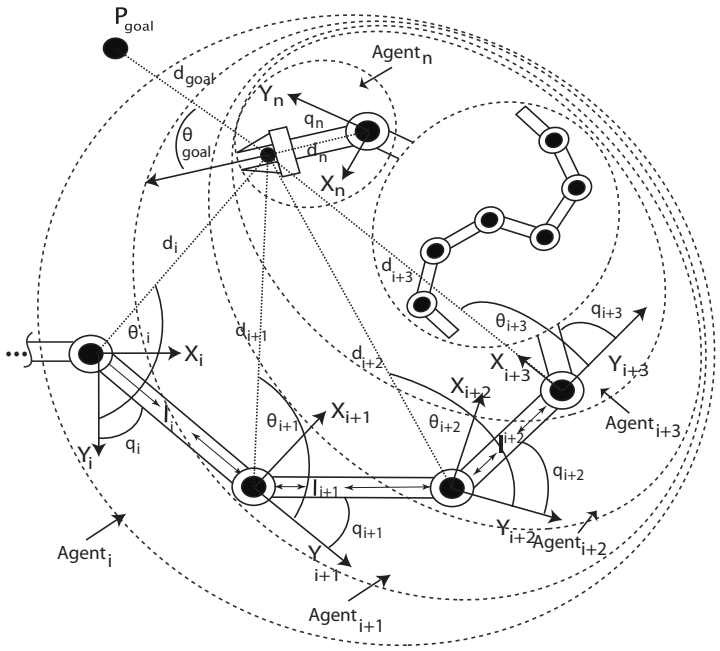
3. Η διαδικασία μάθησης και ανάπτυξης δεξιοτήτων για το πολυπρακτορικό σύστημα γίνεται επίσης σε συνεχή χώρο καταστάσεων. Η προτεινόμενη μεθοδολογία προβλέπει ένα αρχικό στάδιο ασαφοποίησης (fuzzification) των παραμέτρων που προσδιορίζουν την κατάσταση κάθε πράκτορα που υπάρχει στο σύστημα, ενώ στη συνέχεια, η επιλογή της δράσης πραγματοποιείται σε διακριτό χώρο δράσεων.

2.6 Πολυπρακτορικό Πλαίσιο Κινηματικής Αλυσίδας

Έχοντας περιγράψει το προτεινόμενο πολυπρακτορικό πλαίσιο, ας εξετάσουμε τη γενική περίπτωση ενός συστήματος (ρομποτικού χειριστή) το οποίο αποτελείται από n βαθμούς ελευθερίας (πράκτορες $i = 1, \ldots, n$), στη διάταξη η οποία παρουσιάζεται στο σχήμα 2.4. Για κάθε πράκτορα a_i ορίζουμε ως κατάσταση $S_i =< q_i, \theta_i, d_i, \vec{g_i} >$, όπου q_i είναι η υφιστάμενη γενικευμένη μετατόπιση της i άρθρωσης, θ_i είναι η τρέχουσα γωνία προσβολής του κέντρου του τελικού εργαλείου δράσης ως προς το τοπικό πλαίσιο αναφοράς του πράκτορα i, d_i η υφιστάμενη Ευκλίδεια απόσταση του i πράκτορα από το τελικό στοιχείο δράσης, και $\vec{g_i}$ το υφιστάμενο διάνυσμα το οποίο περιγράφει το στόχο στο χώρο εργασίας (task space) (δηλαδή τη θέση του στόχου σε σχέση με το τελικό στοιχείο δράσης).

Το διάγραμμα ροής το οποίο φαίνεται στο Σχήμα 2.5 αποτυπώνει σε επίπεδο πράκτορα, τη γενική δομή του αλγορίθμου πολυπρακτορικής ρομποτικής μάθησης, ο οποίος προτείνεται και αναλύεται στην παρούσα διατριβή.

Κάθε πράκτορας a_i είναι σε θέση να ορίσει συγκεκριμένες παραμέτρους που προσδιορίζουν την κατάστασή του και στη συνέχεια να προωθήσει αυτή την πληροφορία, με εμφωλευμένο (nested) μηχανισμό στους πράκτορες που βρίσκονται στο επόμενο επίπεδο, με στόχο να τους διευκολύνει στην διαδικασία προσδιορισμού των δικών τους καταστάσεων. Ο συγκεκριμένος μηχανισμός είναι top-down , ξεκινώντας από τον πράκτορα στη κορυφή της ιεραρχικής διάτα-



Σχήμα 2.4: *Επιδέξιος χειριστής n - βαθμών ελευθερίας*

ξης και προχωρώντας προς τους πράκτορες που βρίσκονται στα επόμενα επίπεδα. Κατά την αρχική φάση ο πράκτορας προσδιορίζει τις δικές του παραμέτρους άρθρωσης (Joint Parameters), καθώς επίσης και τα γεωμετρικά χαρακτηριστικά του (π.χ. $a_i < q_i, l_i >$), όπου q_i είναι η γενικευμένη μεταβλητή της άρθρωσης, και l_i είναι το μήκος του αντίστοιχου συνδέσμου. Όπως φαίνεται μέσω των παραμέτρων q_i, l_i, έχουμε μερικό προσδιορισμό της κατάστασης του πράκτορα a_i. Για να έχουμε συνολικό προσδιορισμό της κατάστασης του πράκτορα a_i απαιτούνται πρόσθετοι υπολογισμοί των παραμέτρων $< \theta_i, d_i, \vec{g_i} >$. Οι συγκεκριμένες παράμετροι δεν μπορούν να υπολογιστούν στη συγκεκριμένη φάση, δεδομένου ότι ο υπολογισμός τους απαιτεί πληροφορίες από τους υπόλοιπους πράκτορες που απαρτίζουν το πολυπρακτορικό περιβάλλον που εξετάζουμε, και οι οποίες πληροφορίες δεν είναι διαθέσιμες (ή προσβάσιμες) τη συγκεκριμένη χρονική στιγμή. Επομένως, δεδομένου ότι ο πράκτορας a_i δεν δύναται να δώσει συνολική λύση στο πρόβλημα προσδιορισμού της κατάστασής του (state definition problem), προωθεί την τμηματική λύση την οποία παρήγαγε στον πράκτορα a_{i+1}, ο οποίος βρίσκεται στο επόμενο επίπεδο. Ομοίως, ο πράκτορας a_{i+1} υπολογίζει εκείνες τις παραμέτρους οι οποίες μπορούν να υπολογιστούν και προωθεί την τμηματική λύση που σχηματίζεται στους πράκτορες που βρίσκονται στο πιο κάτω επίπεδο, στα πλαίσια πάντα της ιεραρχικής-εμφωλευμένης

```
Fuzzification of the State Space
Fuzzification of the Action Space

State Evaluation
      Agent = Agent(i), Tries to fully evaluate its state.
      If success
            Traverse back the Hierarchy of agents and provide them
            with all the available information in order for them to
            successfully define their state

            Go to Action Selection
      Else
            Pass Information gathered to Agent(i+1) and  i = i+1

            Go to State Evaluation

      Loop until all agents have evaluated their states.

Action Selection
      Agent = Agent(i),
      If NO experience exists in Agent
            Stochastically select Action = a(i)
            Stochastically estimate all other Agents actions
      Else if experience exist
            If Agent wants to explore
                  Select Action = not necessary the best action
                  Stochastically estimate all other Agents actions
            If Agent do not want to explore
                  Select Action = action that will generate the greater reward
                  Stochastically estimate all other Agents actions
            End
      End
      Agent = Agent(i+1)
      Loop until all agents have selected Action

Joint Action Execution
Reward Assigned to all Agents
Go Back to State Evaluation
```

Σχήμα 2.5: *Γενική δομή αλγορίθμου πολυπρακτορικής μάθησης*

(nested-hierarchical) δομής την οποία και προτείνουμε. Η διαδικασία αυτή επαναλαμβάνεται έως ότου ένας πράκτορας a_n καταφέρει επιτυχώς να υπολογίσει το σύνολο των παραμέτρων που προσδιορίζουν συνολικά την κατάστασή του. Στο επόμενο βήμα, η διαδικασία συνεχίζεται προς την αντίστροφη κατεύθυνση (bottom-up) από τον πράκτορα $a_{i+1} \rightarrow a_i$ τροφοδοτώντας τους πράκτορες στα πιο πάνω επίπεδα με τις πληροφορίες εκείνες που τους έλειπαν στην προηγούμενη φάση.

Στην ολοκλήρωση της διαδικασίας αυτής η οποία εξελίσσεται με μια μορφή αναδρομής, κάθε πράκτορας έχει προσδιορίσει την κατάστασή του και συνεπώς το πολυπρακτορικό σύστημα έχει λύσει πλέον το πρόβλημα προσδιορισμού της συνολικής κατάστασής του ως σύστημα. Κάθε πράκτορας αρχικά λειτουργεί στοχαστικά (προβλεπτικά), δεδομένου ότι δεν υπάρχει προηγούμενη γνώση. Στο πλαίσιο αυτό, ο πράκτορας a_i αποφασίζει να πραγματοποιήσει μία τυχαία δράση α_i και την ίδια στιγμή πραγματοποιεί μία εκτίμηση για το τι δράσεις πιθανά οι άλλοι πράκτορες του συστήματος θα επιλέξουν. Έτσι λοιπόν κάθε πράκτορας ανεξάρτητα από τους άλλους επιλέγει δράσεις και την ίδια στιγμή πραγματοποιεί εκτίμηση για τις πιθανές επιλογές των υπολοίπων, μαθαίνοντας λοιπόν με αυτό τον τρόπο συλλογικές δράσεις (joint actions). Η διαδικασία αυτή έχει επίσης μορφή αναδρομής, διατρέχοντας την ιεραρχική δομή από επάνω προς τα κάτω, ξεκινώντας από τον πράκτορα $a_i \rightarrow a_{i+1}$ όπου a_{i+1} είναι ο(οι) πράκτορασ(ες) στο χαμηλότερο επίπεδο. Πιο αναλυτικά, ο a_i επιλέγει τη δράση α_i και εκτιμά ότι οι υπόλοιποι πράκτορες θα επιλέξουν $\alpha'_{i+1}, \alpha'_{i+2}, \ldots \alpha'_n$, αντίστοιχα. Στη συνέχεια, με τον ίδιο τρόπο ο a_{i+1} επιλέγει τη δράση α_{i+1} ενώ πραγματοποιεί μια άλλη εκτίμηση για τις επιλογές των άλλων πρακτόρων (π.χ. $\alpha''_{i+2} \ldots \alpha''_n$). Με την ολοκλήρωση αυτής της διαδικασίας το σύνολο των ανεξάρτητων δράσεων που επιλέχθηκαν αποτελούν την συλλογική δράση του πολυπρακτορικού συστήματος για τη δεδομένη χρονική στιγμή και είναι αυτή που τελικά εκτελείται. Το σύστημα λαμβάνει ανταπόδοση (θετική ή αρνητική) για τη συλλογική αυτή προσπάθεια.

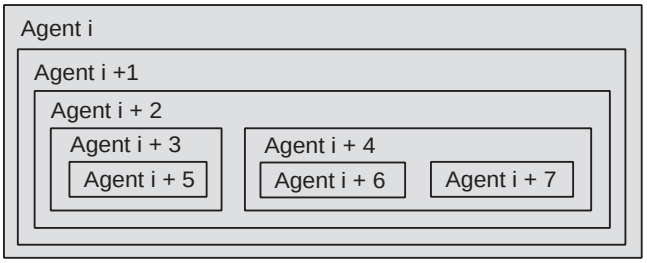
2.7 Πολυπρακτορική Ιεραρχία

Η σχετική εργασία [89] παρουσιάζει ιδιαίτερο ενδιαφέρον, κυρίως για το πλαίσιο των εννοιών που εισάγει παρά για την προσέγγιση που υιοθετεί. Πιο συγκεκριμένα, στην εργασία αυτή ένα πολυδακτυλικό ρομποτικό σύστημα ενσωματώνει μια πολυπρακτορική αρχιτεκτονική στον μηχανισμό συντονισμού των πολλαπλών δακτύλων. Στο πλαίσιο της παραπάνω εργασίας ο πράκτορας δεν είναι μια φυσική οντότητα. Όλο το σύστημα χωρίζεται σε διαφορετικές μονάδες

που δεν καταδεικνύουν οποιασδήποτε μορφής νοημοσύνη. Αυτές οι μονάδες ονομάζονται πράκτορες. Ο κάθε ένας από τους πράκτορες αποτελείται από δύο επίπεδα. Το ένα επίπεδο πραγματοποιεί υπολογισμούς ελέγχου, ενώ το άλλο επικοινωνεί με τους υπόλοιπους πράκτορες. Πρόκειται λοιπόν για πολυπρακτορική τοπολογία η οποία παρουσιάζει μια συμπεριφορά βάσει μοντέλου, χωρίς να επιδεικνύει δυνατότητες περαιτέρω εξέλιξης και προσαρμογής σε συνθήκες που δεν ήταν γνωστές κατά την αρχική φάση σχεδιασμού της συγκεκριμένης πολυπρακτορικής τοπολογίας. Στο πλαίσιο της παρούσης διατριβής το πολυπρακτορικό σύστημα επιδεικνύει μια διαφορετική σχέση ανάμεσα στους πράκτορες. Η αρχιτεκτονική που εμείς εφαρμόζουμε επιδεικνύει υβριδικά χαρακτηριστικά ιεραρχίας αλλά και ισότητας. Τα ιεραρχικά χαρακτηριστικά μας παρέχουν την ευρωστία ενός περισσότερου κλασικού κεντρικού συστήματος, ενώ τα χαρακτηριστικά της ισότητας την ευελιξία μιας κατανεμημένης αρχιτεκτονικής συνδετικής προσέγγισης (Connectionist Approach) .

Έχοντας περιγράψει ήδη την εσωτερική αρχιτεκτονική ενός πράκτορα, ας δούμε τώρα περισσότερο αναλυτικά την πολυπρακτορική δομή την οποία προτείνουμε. Το Σχήμα 2.6 παρουσιάζει σε διάγραμμα την προτεινόμενη αρχιτεκτονική ενώ το επόμενο Σχήμα 2.7 παρουσιάζει τη συγκεκριμένη αρχιτεκτονική εφαρμοσμένη σε ένα σύστημα ανοικτής κινηματικής αλυσίδας n βαθμών ελευθερίας. Η εμφωλευμένη πολυπρακτορική αρχιτεκτονική που εξετάζουμε έχει ως κύριο χαρακτηριστικό την ομοιογένεια αναφορικά με τη δομή όλων των πρακτόρων. Αυτά τα χαρακτηριστικά είναι εκείνα τα οποία δημιουργούν συνθήκες κατάλληλες για την δυνητική επέκταση ενός συστήματος που θα ενσωματώνει την προτεινόμενη αρχιτεκτονική σε πιο σύνθετες διατάξεις.

Πιο αναλυτικά, κάθε πράκτορας a_i "βλέπει" μόνο εκείνους τους πράκτορες που βρίσκονται ένα επίπεδο πιο κάτω στην ιεραρχική δομή. Αυτό σημαίνει ότι ο πράκτορας i κληροδοτεί / κληρονομεί τη γνώση την οποία έχει μόνο στους/από τους πράκτορες οι οποίοι είναι "ορατοί" ένα επίπεδο πιο κάτω (στην περίπτωση του σχήματος 2.6, αυτό σημαίνει μόνο τον πράκτορα a_{i+1}). Στην περίπτωση τώρα του πράκτορα a_{i+2}, οι ορατοί πράκτορες είναι μόνο οι a_{i+3} και a_{i+4}. Η προσέγγιση αυτή παρέχει στο πολυπρακτορικό μας σύστημα την ικανότητα αναζήτησης λύσης με αναδρομή από τον πράκτορα $a_i \rightarrow a_{i+1}$. Εάν ένας πράκτορας δεν μπορεί να λύσει το πρόβλημα (ή δεν μπορεί τουλάχιστον να συνδράμει στη λύση του), τότε περνά την πληροφορία η οποία έχει συλλεγεί στον (στους) πράκτορα(ες) που βρίσκονται στο αμέσως επόμενο επίπεδο. Όταν η διαδικασία αυτή καταλήξει σε έναν πράκτορα ο οποίος είναι σε θέση να δώσει λύση, τότε ο πράκτορας συνεισφέρει, και πηγαίνοντας προς τα πίσω η διαδικασία παρέχει γνώση σε όλους τους πράκτορες στην αλυσίδα.



Σχήμα 2.6: *Γενική δομή εμφωλευμένης-ιεραρχικής πολυπρακτορικής αρχιτεκτονικής*

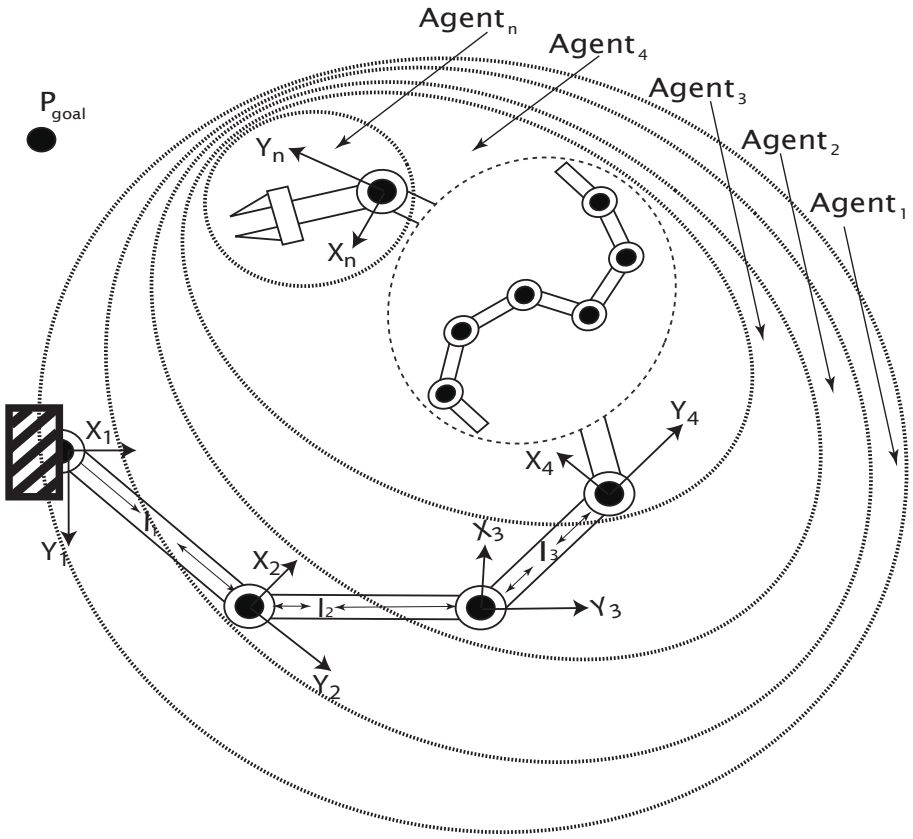
Εφαρμόζοντας αυτήν τη πολυπρακτορική αρχιτεκτονική στην περίπτωση της κινηματική αλυσίδας n βαθμών ελευθερίας, δημιουργούμε την ιεραρχική διάταξη με φωλιασμένους ομοιογενείς πράκτορες η οποία αποτυπώνεται στο Σχήμα 2.7. Κάθε πράκτορας ενσωματώνει συγκεκριμένες συμπεριφορές (αντιστοιχίσεις κατάστασης-δράσης) οι οποίες εξελίσσονται με τη χρήση αλγορίθμων ενισχυτικής μάθησης, ενώ ο συντονιστής πράκτορας είναι υπεύθυνος για την παρακολούθηση της απόδοσης του συστήματος καθώς και την ανατροφοδότηση των αποτελεσμάτων στους πράκτορες του συστήματος.

2.8 Πολυπρακτορικό Πλαίσιο Συνεργατικών Αυτοκινούμενων Ρομπότ

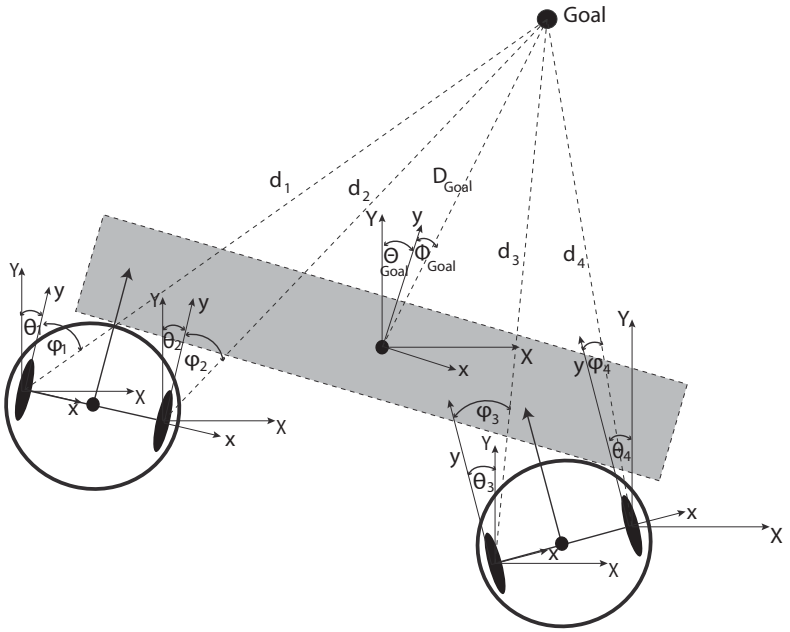
Ακολουθώντας το σχεδιασμό που είδαμε ήδη στην περίπτωση της κινηματικής αλυσίδας, η πολυπρακτορική αρχιτεκτονική επεκτείνεται στην περίπτωση των αυτοκινούμενων οχημάτων τα οποία επενεργούν σε ένα αντικείμενο έτσι ώστε να προκαλέσουν συνεργατικά, τη μετατόπισή του προς τη θέση-στόχο, όπως αποτυπώνεται στο Σχήμα 2.8. Η πρόκληση και σε αυτή την περίπτωση προβλήματος είναι η εύρεση μιας κατάλληλης περιγραφής της κατάστασης του πράκτορα. Ας εξετάσουμε το σχετικό σχήμα και ας δούμε τα θέματα τα οποία προκύπτουν.

Οι κύριες προϋποθέσεις-παραδοχές για την εφαρμογή της προτεινόμενης αρχιτεκτονικής στην περίπτωση των αυτοκινούμενων ρομπότ συνοψίζονται στη συνέχεια:

1. Αντιστοίχιση των πρακτόρων του συστήματος στους τροχούς των αυτοκινούμενων ρομπότ. Η πρόκληση σε αυτήν την περίπτωση είναι η δημιουργία



Σχήμα 2.7: *Πολυπρακτορική αρχιτεκτονική σε κινηματική αλυσίδα με n - βαθμούς ελευθερίας*



Σχήμα 2.8: *Δύο αυτοκινούμενα οχήματα ωθούν ένα αντικείμενο συνεργατικά στη θέση-στόχο πραγματοποιώντας εργασία τύπου "box-pushing"*

συνολικής (γενικευμένης), για το πολυπρακτορικό σύστημα, αντιστοίχισης κατάστασης-δράσης μέσω προοδευτικής απόκτησης τοπικής γνώσης σε επίπεδο κάθε πράκτορα.

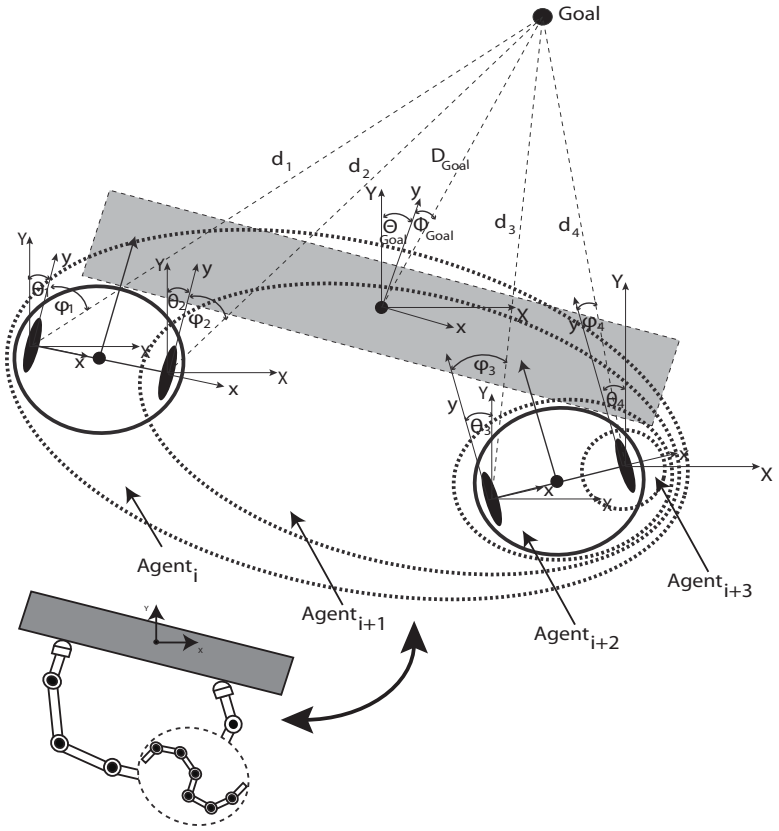
2. Η δεξιότητα την οποία οι πράκτορες-τροχοί πρέπει να αποκτήσουν αφορά τον τρόπο με τον οποίο θα συνεργαστούν έτσι ώστε και τα δύο ρομπότ να ωθήσουν το αντικείμενο, συνεργατικά στη θέση-στόχο. Αντιλαμβανόμαστε ότι η συγκεκριμένη δεξιότητα μπορεί να αναλυθεί σε δύο υπο-εργασίες α) διατήρηση της επαφής με το αντικείμενο, καθώς και β) συνεργατικός χειρισμός ώθηση/στρέψη του αντικειμένου έτσι ώστε να φτάσει στη θέση-στόχο. Ο προτεινόμενος μηχανισμός μάθησης, όπως και στην περίπτωση της κινηματικής αλυσίδας, δεν απαιτεί εκ των προτέρων (a priori) ανάλυση της εργασίας σε υπο-εργασίες, καθώς επίσης δεν ενσωματώνει συγκεκριμένο μηχανισμό ο οποίος να διαχειρίζεται την εναλλαγή από τη μια συμπεριφορά στην άλλη έτσι ώστε να καταφέρνει να διαχειρίζεται τις δύο υπο-εργασίες. Αντιθέτως, το πολυπρακτορικό σύστημα, δίχως την ύπαρξη κάποιου μοντέλου εργασίας, επιτυγχάνει να αποκτήσει την απαιτούμενη δεξιότητα για τη συνολική εργασία με έναν αφαιρετικό τρόπο.

3. Ο νευροδυναμικός μηχανισμός μάθησης ο οποίος ενσωματώθηκε και τον

οποίο θα αναλύσουμε στο επόμενο κεφαλαίο, δεν απαιτεί να προσδιορίσουμε στους πράκτορες κάποια συγκεκριμένα σημεία επαφής με το αντικείμενο. Ουσιαστικά τα σημεία μέσω των οποίων τα ρομπότ ωθούν το αντικείμενο δεν είναι προαποφασισμένα καθώς επίσης ούτε σταθερά. Επιπλέον το σχήμα του αντικειμένου δεν απαιτείται να είναι γνωστό ή γενικά να υπάρχει πληροφορία στους πράκτορες σχετικά με αυτό. Αυτά τα δύο σημεία είναι αρκετά σημαντικά, δεδομένου ότι η διαδικασία μάθησης είναι ανεξάρτητη των γεωμετρικών χαρακτηριστικών του αντικειμένου καθώς επίσης και των σημείων επαφής.

4. Κατανεμημένος ορισμός κατάστασης του συστήματος (distributed state definition) μεταξύ των πρακτόρων που το συνθέτουν. Δεδομένου ότι ο μηχανισμός μάθησης, όπως θα δούμε στο επόμενο κεφάλαιο, πρέπει να σχεδιαστεί έτσι ώστε να λειτουργεί σε ένα συνεχή χώρο καταστάσεων, το πρόβλημα το οποίο πρέπει να διαχειριστεί ο συγκεκριμένος μηχανισμός είναι η διαχείριση της μεγάλης διάστασης που παρουσιάζει ο χώρος καταστάσεων (dimensionality). Εκείνο το οποίο προτείνουμε λοιπόν είναι ο κατανεμημένος ορισμός της κατάστασης του συστήματος μεταξύ των πρακτόρων που το συνθέτουν, μέσω της εμφωλευμένης αρχιτεκτονικής. Αυτή η σχεδιαστική απόφαση οδηγεί σε μείωση ουσιαστικά του χώρου καταστάσεων προκαλώντας σημαντική επίπτωση στο υπολογιστικό κόστος του μηχανισμού μάθησης όπως θα δούμε στο Κεφάλαιο 4.

Οι πράκτορες είναι εμφωλευμένοι όπως και στην περίπτωση της κινηματικής αλυσίδας. Το πρώτο θέμα το οποίο συνιστά ουσιαστική διαφοροποίηση της παρούσας μοντελοποίησης σε σχέση με την τοπολογία της κινηματικής αλυσίδας είναι ότι δεν υπάρχει κάποιο σημείο του συστήματος το οποίο είναι σταθερό και το οποίο να λειτουργεί ως σημείο αναφοράς για τον προσδιορισμό του χώρου-κατάστασης. Με άλλα λόγια, ουσιαστικά το σύστημα το οποίο εξετάζουμε είναι ένα κινούμενο σύστημα. Διατηρώντας το ίδιο πλαίσιο πολυπρακτορικού σχεδιασμού με πριν (στην ιεραρχική δομή των πρακτόρων - τροχοί έναντι συνδέσμων - καθώς και το γεγονός ότι ο κάθε πράκτορας λειτουργεί και αποφασίζει ανεξάρτητα), μπορούμε να δούμε την πειραματική πλατφόρμα των αυτοκινούμενων ρομπότ ως μια κινηματική αλυσίδα με τα δύο άκρα της πλήρως ελεύθερα. Προφανώς, λοιπόν, έχουμε να αντιμετωπίσουμε μία μη-ολόνομη πολυρομποτική, πολυπρακτορική διάταξη, η οποία έχει στόχο να κατευθύνει το αντικείμενο στο οποίο επενεργούν οι πράκτορες στη θέση-στόχο. Το Σχήμα 2.9 αποτυπώνει αυτή την ιεραρχική εμφωλευμένη διάταξη των πρακτόρων που συνθέτουν το σύστημα. Κάθε πράκτορας βλέπει εκείνους τους πράκτορες οι οποίοι βρίσκονται σε χαμηλότερα επίπεδα στην ιεραρχική δομή. Το άλλο θέμα το οποίο παρουσιάζει ιδιαίτερο ενδιαφέρον είναι το γεγονός ότι προσαρμόζουμε έναν συγκεκριμένο



Σχήμα 2.9: *Οι τροχοί των αυτοκινούμενων ρομπότ αντιστοιχούν στους αυτόνομους πράκτορες του συστήματος*

ιεραρχικό συσχετισμό μεταξύ φυσικών οντοτήτων (τροχών-ρομπότ) οι οποίες, ανάλογα με τη συγκεκριμένη τοπολογία του προβλήματος, ενδέχεται να μην έχουν μεταξύ τους φυσική διασύνδεση (κάποιοι από τους τροχούς να ανήκουν σε διαφορετικά ρομπότ).

Η κατάσταση λοιπόν η οποία περιγράφει πλήρως το πολυπρακτορικό σύστημα ορίζεται στη συνέχεια. Για κάθε πράκτορα a_i, η κατάστασή του S_i ορίζεται ως ακολούθως:

$$S_i =< \theta_i, \phi_i, d_i, \omega_i, \Theta_{goal_i}, \Phi_{goal_i}, D_{goal_i} > \tag{2.1}$$

όπου i είναι ο δείκτης ο οποίος αναφέρεται στους ανεξάρτητους πράκτορες i, θ_i είναι ο προσανατολισμός του πράκτορα ως προς κάποιο οικουμενικό πλαίσιο αναφοράς $O-XY$ στο χώρο εργασίας, ϕ_i είναι ο προσανατολισμός του πράκτορα ως προς το στόχο, d_i είναι η απόσταση του πράκτορα από το στόχο, ω_i είναι η γωνιακή ταχύτητα του πράκτορα-τροχού, Θ_{goal_i} είναι ο προσανατολι-

σμός του αντικειμένου ως προς το οικουμενικό πλαίσιο αναφοράς $O-XY$ στο χώρο εργασίας, Φ_{goal_i} είναι ο προσανατολισμός του αντικειμένου ως προς το στόχο και D_{goal_i} είναι η απόσταση του αντικειμένου από το στόχο. Έχοντας ορίσει την κατάσταση λοιπόν για έναν πράκτορα του συστήματος είναι προφανής ο μηχανισμός γενίκευσης προσδιορισμού των καταστάσεων και των άλλων πρακτόρων.

2.9 Χαρακτηριστικά Πολυπρακτορικής Αρχιτεκτονικής

Στο τέλος αυτού του κεφαλαίου θα αναφέρουμε κάποια σημαντικά χαρακτηριστικά που επιδεικνύει η πολυπρακτορική αρχιτεκτονική την οποία προτείνουμε στην παρούσα διατριβή.

Συνεξελικτική (Co-evolutionary) Μάθηση

Στην πολυπρακτορική αρχιτεκτονική η έννοια της Συνεξελικτικής (co-evolutionary) μάθησης ορίζει ότι πολλές διαφορετικές στρατηγικές μπορούν να εξελιχθούν ταυτόχρονα καθώς και να μοιράζονται μεταξύ του πληθυσμού των πρακτόρων. Πιο συγκεκριμένα, τόσο στην εφαρμογή της κινηματικής αλυσίδας όσο και στην περίπτωση των αυτοκινούμενων οχημάτων, η συνεξελικτική μάθηση εφαρμόζεται στο πλαίσιο της αντιμετώπισης των υπο-στόχων που συνθέτουν τον συνολικό (γενικευμένο) στόχο. Αυτό που εννοούμε είναι ότι η αρχιτεκτονική μας υποστηρίζει την δυναμική ανάλυση του στόχου σε υπο-στόχους. Κατά συνέπεια, για αυτούς τους διαφορετικούς στόχους ο κάθε πράκτορας αναπτύσσει διαφορετικές τοπικές στρατηγικές για να τους επιτύχει.

Συνεργατική Συμπεριφορά

Οι πράκτορες μπορούν να μοιράζονται πληροφορίες έτσι ώστε να δύνανται να επιδεικνύουν συνεργατική συμπεριφορά. Για παράδειγμα, θα δούμε στο Κεφάλαιο 4 τις διαφορετικές πειραματικές διατάξεις που αξιολογούνται να εκτελούν εργασίες, όπου η συνεργατική συμπεριφορά μεταξύ των πρακτόρων που συνθέτουν το σύστημα είναι άκρως απαραίτητη. Οι εργασίες λοιπόν οι οποίες απαιτούν δυναμικό επιμερισμό καθηκόντων μεταξύ των διαφορετικών πρακτόρων, επιβάλλουν έμμεσα μία συνεργατική συμπεριφορά, την οποία η προτεινόμενη αρχιτεκτονική μπορεί να την διαχειριστεί.

Κατανεμημένη Επεξεργασία

Απόκριση σε πραγματικό χρόνο μπορεί να επιτευχθεί κατανέμοντας το υπολογιστικό βάρος του ελέγχου και της επεξεργασίας πληροφοριών στο σύνολο του πληθυσμού των πρακτόρων. Όπως ήδη αναφέραμε, δεν απαιτείται μόνο συνεργασία για την επίλυση των σύνθετων προβλημάτων, αλλά και η κατανομή του υπολογιστικού κόστους στο σύνολο των πρακτόρων. Καθώς ο αριθμός των πρακτόρων που απασχολείται αυξάνεται, συμβαίνει το ίδιο και στην πολυπλοκότητα του υπολογιστικού μοντέλου το οποίο υποστηρίζει τη λειτουργία τους. Το σημαντικό είναι ότι στις περισσότερες των περιπτώσεων αυτό δεν συμβαίνει γραμμικά. Προσπαθούμε λοιπόν να το εξισορροπήσουμε αυτό με τον καταμερισμό των απαραίτητων υπολογισμών που σχετίζονται τόσο με το μηχανισμό της μάθησης όσο και με το μηχανισμό έλεγχου σε κάθε ανεξάρτητη μονάδα που συνθέτει το σύστημα (και προφανώς εδώ με τον όρο μονάδα εννοούμε τον πράκτορα). Η κατανεμημένη επεξεργασία και το σχετικό υπολογιστικό κόστος αναλύεται στο Κεφάλαιο 4 της παρούσης διατριβής.

Ανεξαρτησία - Αλληλεξάρτησης

Ο βαθμός στον οποίο ένας πράκτορας βασίζεται στους υπόλοιπους, δεν είναι κάτι το οποίο προσδιορίζεται με ευκολία. Είναι ένα δύσκολο πρόβλημα το οποίο έχει να κάνει τόσο με το μηχανισμό αυτο-οργάνωσης που διαθέτει ο πράκτορας, όσο και με την εξέλιξη του πράκτορα μέσα στο σύστημα στο οποίο ανήκει. Δεν μπορούμε να έχουμε ανοχή σφάλματος στο σύστημά μας χωρίς να έχουμε σημαντικό βαθμό αλληλεξάρτησης και από την άλλη, χρειαζόμαστε την ανεξαρτησία για να ενισχύσουμε την αυτονομία του πράκτορα και να προωθήσουμε την έννοια της αυτο-εξέλιξης.

Ομοιογένεια

Τέλος, πρέπει να τονίσουμε ότι, ενώ η ειδίκευση μπορεί να επιτρέψει την πραγματοποίηση δύσκολων εργασιών και συχνά να οδηγήσει στη βελτίωση της συνολικής αποτελεσματικότητας, η έλλειψη όμως ομοιογένειας περιπλέκει την επεκτασιμότητα και τον έλεγχο. Στην προτεινόμενη αρχιτεκτονική, υιοθετούμε μια ομοιογενή δομή. Η ιδέα είναι να δημιουργήσουμε απλά, βασικά δομικά στοιχεία τα οποία να αντιστοιχίζονται άμεσα σε απλές φυσικές οντότητες, όπως οι ρομποτικοί σύνδεσμοι ή οι τροχοί των αυτοκινούμενων ρομπότ. Αυτές οι απλές ομοιογενείς οντότητες μπορούν να συνδυαστούν για να σχηματίσουν δομές που να καταφέρνουν να επιλύουν πιο περίπλοκες εργασίες αλλά ταυτόχρονα να διατηρούν την απλότητά τους και την συναρμολογησιμότητά τους.

Έχοντας περιγράψει την έννοια και τη δομή του πράκτορα στο πλαίσιο της προτεινόμενης πολυπρακτορικής αρχιτεκτονικής καθώς και τα ιδιαίτερα χαρακτηριστικά που επιδεικνύει η εμφωλευμένη-ιεραρχική τοπολογία που προτείνει η παρούσα διατριβή, θα προχωρήσουμε στο επόμενο κεφάλαιο στη μελέτη του προβλήματος της ρομποτικής μηχανικής μάθησης, ένα ιδιαίτερα δύσκολο και σίγουρα ενδιαφέρον πρόβλημα.

Κεφάλαιο 3

Μη-Επιβλεπόμενη Ρομποτική Μάθηση

3.1 Εισαγωγή

Το κεφάλαιο αυτό έχει ως στόχο την ανάλυση των λειτουργικών απαιτήσεων του μοντέλου ρομποτικής μάθησης όπως αυτές προκύπτουν μέσα από την παρούσα ερευνητική προσπάθεια. Αρχικά το κεφάλαιο ξεκινά με μία γενική περιγραφή του μηχανισμού Ενισχυτικής Μάθησης, βασικούς ορισμούς και ιδιότητές του, μεθόδους ενισχυτικής μάθησης που συναντάμε στη βιβλιογραφία, και τέλος εξειδικεύει την εφαρμογή του συγκεκριμένου μηχανισμού στο πλαίσιο του επιδέξιου ρομποτικού χειρισμού. Στη συνέχεια, επισημαίνονται σημεία όπου η ενισχυτική μάθηση κάνει χρήση στοιχείων που προέρχονται από τους τομείς του *Δυναμικού Προγραμματισμού* καθώς επίσης και των *Μαρκοβιανών Διαδικασιών Λήψης Αποφάσεων* (Markov Decision Process). Εν συνεχεία, καταγράφονται δύο μέθοδοι Ενισχυτικής Μάθησης, η Επανάληψη Πολιτικής (Policy Iteration) και η Επανάληψη Αξίας (Value Iteration) όπου για κάθε μία από αυτές αναλύεται ένας αντιπροσωπευτικός μηχανισμός μάθησης. Στη συνέχεια, έχοντας αναλύσει τους μηχανισμούς Ενισχυτικής Μάθησης, ακολουθεί ο ορισμός του *Χώρου Κατάστασης* (State - Space) μέσα στον οποίο κινούνται οι πράκτορες, πώς επιτυγχάνεται η συνέχειά του, η μείωση της διάστασής του, και τέλος ορίζεται ο *Χώρος Δράσης* (Action - Space) μαζί με τον αντίστοιχο μηχανισμό επιλογής δράσεων, για το συγκεκριμένο μοντέλο πολυπρακτορικής διάταξης το οποίο είδαμε ήδη στο προηγούμενο κεφάλαιο. Μέσω της διαδικασίας μηχανικής μάθησης που παρουσιάζουμε σε αυτό το κεφάλαιο επιτυγχάνεται η γενίκευση της γνώσης, γεγονός που επιτρέπει την επαναχρησιμοποίησή της στις περιπτώσεις εκείνες όπου οι στόχοι προς επίτευξη παρουσιάζουν κάποιου είδους συνάφεια. Το κεφάλαιο ολοκληρώνεται με την ανάλυση του μοντέλου ενισχυτικής μάθησης όπως αυτό διαμορφώθηκε στην εφαρμογή τόσο του επιδέξιου ρομποτικού χειριστή, όσο και στο πεδίο των αυτοκινούμενων ρομπότ.

3.2 Απαιτήσεις Μοντέλου Ρομποτικής Μάθησης

Έχοντας ήδη μελετήσει τη δομή της πολυπρακτορικής αρχιτεκτονικής την οποία προτείνουμε στην παρούσα διατριβή, και έχοντας επισημάνει τις βασικές προϋποθέσεις που επιβάλλει ο συγκεκριμένος τομέας εφαρμογής του επιδέξιου ρομποτικού χειρισμού καθώς και εκείνος των αυτοκινούμενων ρομπότ, το επόμενο βήμα είναι να καθορίσουμε τις αντίστοιχες απαιτήσεις που προκύπτουν μέσα από το πρόβλημα της ρομποτικής μάθησης, όπως αυτό διατυπώνεται στο συγκεκριμένο πλαίσιο, και στο οποίο σκοπεύουμε να προτείνουμε μια λύση. Μία αρχική διατύπωση της προσέγγισής μας είναι η ακόλουθη: *Η υλοποίηση*

μιας πολυπρακτορικής αρχιτεκτονικής όπου, κάθε πράκτορας ο οποίος θα αποτελεί τμήμα του συστήματος, πρέπει να επιδεικνύει σημαντικό βαθμό μαθησιακών ικανοτήτων, οι οποίες θα επιτρέπουν στο σύστημα με την πάροδο του χρόνου να αναπτύσσει δεξιότητες, να αποκτά εμπειρία, να αναγνωρίζει ομοιότητες μεταξύ διαφορετικών συμπεριφορών που έχουν ακολουθηθεί στο παρελθόν, να επεκτείνει εκείνες που είναι ήδη διαθέσιμες, και γενικά να επιδεικνύει συμπεριφορά Αναπτυξιακής Μάθησης (Developmental Learning) [48], [86].

Επομένως, έχοντας διαπιστώσει και εν συνεχεία διατυπώσει τις λειτουργικές απαιτήσεις που παρουσιάζονται στο πλαίσιο της προσπάθειας ανάπτυξης του συγκεκριμένου μοντέλου ρομποτικής μάθησης, προχωρούμε ένα βήμα ακόμα με σκοπό την αναγωγή των συγκεκριμένων αναγκών στο πλαίσιο του επιδέξιου ρομποτικού ελέγχου. Αυτός είναι και ο συγκεκριμένος τομέας όπου αποσκοπούμε να συνεισφέρουμε μέσω της παρούσης διατριβής.

Γενικά έχει καταγραφεί μια πολύ εκτενής και παράλληλα ουσιαστική ερευνητική προσπάθεια για τη δημιουργία ενός ικανοποιητικού μοντέλου ρομποτικής μάθησης για πολυπρακτορικά συστήματα με ειδικές εφαρμογές στον συγκεκριμένο τομέα των αυτοκινούμενων ρομπότ (Mobile Robots). Οι προσεγγίσεις σχεδιασμού που έχουν υιοθετηθεί έως τώρα, αναφορικά με ιδιότητες όπως η αυτο-οργάνωση και η μάθηση δεξιοτήτων, στον τομέα των κινητών ρομπότ είναι πολύ καλά εδραιωμένες και επιδεικνύουν σημαντικό βαθμό εύρωστης συμπεριφοράς και απόδοσης. Στην παρούσα διατριβή, διατυπώνουμε ορισμένες αρχές σχεδιασμού από το χώρο των κινητών ρομπότ, οι οποίες θα λειτουργήσουν ως υπόβαθρο για την καλύτερη κατανόηση της ανάλυσης που θα ακολουθήσει. Προφανώς, οι βασικές αυτές αρχές σχεδίασης, θα διευρυνθούν κατάλληλα έτσι ώστε να ικανοποιήσουν τις ιδιαίτερες απαιτήσεις που προκύπτουν από τον τομέα εφαρμογής που μας ενδιαφέρει και δεν είναι άλλος από αυτόν του επιδέξιου ρομποτικού χειρισμού. Ένας από τους κεντρικούς στόχους της παρούσας διατριβής, όπως έχει ήδη αναφερθεί, είναι η ανάπτυξη μίας μεθοδολογίας που θα επιτρέψει τη δημιουργία ενός πλαισίου μάθησης, αυτόνομων, προσαρμοζόμενων και αυτο-οργανούμενων πολυπρακτορικών ρομποτικών συστημάτων.

Στη συνέχεια λοιπόν, προσπαθούμε να ορίσουμε το πλαίσιο αυτό, το οποίο ενσωματώνοντας νευρο-δυναμικές προσεγγίσεις και πρακτικές από τον κλάδο του συνδετισμού στον τομέα του επιδέξιου χειρισμού και των αυτοκινούμενων ρομπότ, θα οδηγήσει σε ένα αρχικό παράδειγμα εφαρμογής αναπτυξιακών ρομποτικών συστημάτων.

3.3 Ενισχυτική Μάθηση (RL)

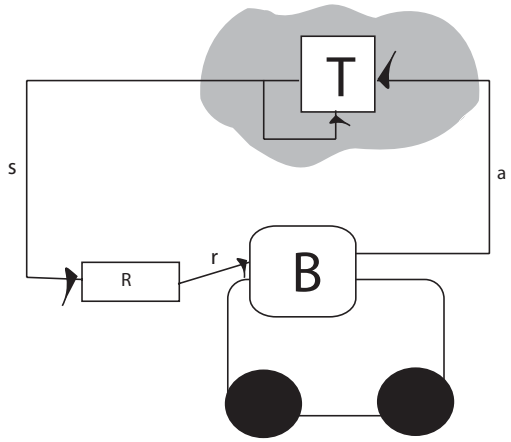
Τα φαινόμενα της ενισχυτικής μάθησης έχουν παρατηρηθεί σε ψυχολογικές μελέτες της συμπεριφοράς βιολογικών οργανισμών καθώς επίσης και στο πλαίσιο νευρολογικών ερευνών σχετικά, με το συντονισμό των νευρώνων (Neuromodulation) και εκείνων του εθισμού [29]. Ορίζουμε λοιπόν ως Ενισχυτική Μάθηση, τη διαδικασία εκείνη με την οποία βιολογικοί οργανισμοί, τεχνητοί πράκτορες και σύνθετα συστήματα μαθαίνουν να βελτιστοποιούν τη συμπεριφορά τους, εν όψει σχετικής ανταπόδοσης. Έχουν αναπτυχθεί διάφοροι αλγόριθμοι Ενισχυτικής Μάθησης οι οποίοι είναι στενά συνδεδεμένοι με τον ερευνητικό χώρο του *Δυναμικού Προγραμματισμού (Dynamic Programming)* [13], ο οποίος εστιάζει σε γενικότερες προσεγγίσεις βελτιστοποίησης διαδικασιών ελέγχου. Περισσότερα στοιχεία σχετικά με το Δυναμικό Προγραμματισμό και την Ενισχυτική Μάθηση θα δούμε και στη συνέχεια του Κεφαλαίου. Στο σημείο αυτό, πρέπει να τονισθεί αρχικά η βασική διαφοροποίηση του Δυναμικού Προγραμματισμού ως προς την Ενισχυτική Μάθηση, και η οποία δεν είναι άλλη από το γεγονός ότι ο Δυναμικός Προγραμματισμός προϋποθέτει δύο βασικά στοιχεία: τη γνώση όλων των πιθανοτήτων μετάβασης (Transition) από μία κατάσταση (State) σε άλλη για κάθε δυνατή δράση (Action), καθώς επίσης τη γνώση της ανταπόδοσης (Reward) η οποία προκύπτει από κάθε μετάβαση. Η Ενισχυτική Μάθηση έχει ένα συγκριτικό πλεονέκτημα λοιπόν, ότι δεν απαιτεί κανένα από τα δύο αυτά στοιχεία για την εφαρμογή της.

Το πλαίσιο μέσα στο οποίο μελετάται η ικανότητα μάθησης που επιδεικνύει ένας τεχνητός πράκτορας, σχετικά με την πραγματοποίηση κατάλληλων ενεργειών (δράσεων), σε απάντηση συγκεκριμένων ερεθισμάτων, στη βάση συσχετιζόμενων ως προς τα ερεθίσματα αυτά ανταποδόσεων ή τιμωριών, αποτελεί τον ερευνητικό χώρο όπου εστιάζει η *Συμπεριφορική Ψυχολογία* (*Behavioral Psychology*). Ο τομέας αυτός διαχωρίζεται σε *Classical Conditioning* (ή *Pavlovian* [53]) και *Instrumental Conditioning* [29], [53] . Στην περίπτωση του *Classical Conditioning* τα ενισχυτικά σήματα (*Reinforcers*) (είτε αυτά είναι ανταποδόσεις, είτε τιμωρίες) απονέμονται ανεξάρτητα των ενεργειών που εκτελεί ο πράκτορας. Στην περίπτωση του *Instrumental Conditioning* τα ενισχυτικά σήματα απονέμονται βάση των δράσεων του πράκτορα και είναι ουσιαστικά επακόλουθα των ενεργειών του. Η Ενισχυτική Μάθηση ως διαδικασία θα μπορούσαμε να θεωρήσουμε ότι αποτελείται από δύο σκέλη. Το πρώτο είναι εκείνο το οποίο εμπεριέχει την προσπάθεια του πράκτορα να δοκιμάζει διαφορετικές δράσεις, και με την πάροδο του χρόνου να μαθαίνει να απαντά σε ερεθίσματα αποκλειστικά και μόνο βασιζόμενος στις ανταποδόσεις (ή τις τιμωρίες) τις οποίες λαμβάνει και οι οποίες είναι άρρηκτα συνδεδεμένες με τις συγκεκριμένες δράσεις τις οποίες επέλεξε (*Trial & Error Process*). Ως

προς αυτό το κομμάτι, η Ενισχυτική Μάθηση θα μπορούσε να χαρακτηριστεί ως μια διαδικασία *Instrumental Conditioning*. Το δεύτερο σκέλος τώρα είναι εκείνο της διαδικασίας πρόβλεψης που χρησιμοποιεί η Ενισχυτική Μάθηση και το οποίο ονομάζεται *Μοντέλο Χρονικών Διαφορών* (*Temporal Difference Model*). Ως προς αυτό το δεύτερο σκέλος της, η Ενισχυτική Μάθηση χαρακτηρίζεται ως μια διαδικασία *Classical Conditioning*, δεδομένου ότι κάνει χρήση προγενέστερης εμπειρίας για την πραγματοποίηση ακριβέστερης πρόβλεψης για το μέλλον. Βλέπουμε ξεκάθαρα λοιπόν, ότι η Ενισχυτική Μάθηση παραπέμπει στη *Συμπεριφορική Ψυχολογία* πότε ως μια διαδικασία *Instrumental Conditioning* και πότε ως *Pavlovian* [14].

Στην αρχιτεκτονική την οποία και αναλύσαμε εκτενώς στο προηγούμενο κεφάλαιο αυτής της διατριβής, το σήμα ενίσχυσης που δέχεται ο πράκτορας, σύμφωνα πάντα με τις ενέργειες που πραγματοποιεί, είναι η μόνη ανάδραση που θα του επιτρέψει να δημιουργήσει με την πάροδο του χρόνου ένα συγκεκριμένο συμπεριφορικό μοντέλο. Σε αυτό το σημείο θα πρέπει να λάβουμε υπόψη το γεγονός ότι σύμφωνα με το μοντέλο ρομποτικής μάθησης που αναπτύσσουμε, ο τεχνητός πράκτορας δεν έχει στη διάθεσή του εξ΄ αρχής κάποιο μοντέλο συμπεριφοράς. Στο πλαίσιο αυτό λοιπόν έχουμε δύο περιπτώσεις. Την περίπτωση εκείνη όπου το σήμα ενίσχυσης δίνεται στον πράκτορα αμέσως μετά την πραγματοποίηση της οποιασδήποτε ενέργειας, γεγονός που κάνει τη μάθηση σχετικά πιο εύκολη. Η δεύτερη περίπτωση που συναντούμε είναι εκείνη όπου η ανταπόδοση ή η τιμωρία απονέμεται στον πράκτορα με βάση την ολοκλήρωση μιας συγκεκριμένης ακολουθίας ενεργειών και κατά συνέπεια εισάγεται μια χρονική καθυστέρηση. Η χρονική αυτή καθυστέρηση δύναται να είναι τμηματική ή συνολική. Στην περίπτωση της τμηματικής καθυστέρησης, η ανταπόδοση απονέμεται τμηματικά στο πράκτορα καθώς τα αντίστοιχα τμήματα της ακολουθίας ολοκληρώνονται, ενώ στη δεύτερη περίπτωση της συνολικής καθυστέρησης, η ανταπόδοση απονέμεται στον πράκτορα εφόσον η ακολουθία ενεργειών ολοκληρωθεί εξόλοκλήρου. Σε κάθε περίπτωση κατά την οποία η ανταπόδοση προς τον πράκτορα δεν είναι άμεση, η διαδικασία μάθησης για τον πράκτορα γίνεται σαφώς πιο δύσκολη. Στο τυπικό πλαίσιο ενισχυτικής μάθησης, ένας πράκτορας παρατηρεί επανειλημμένα την κατάσταση στην οποία βρίσκεται το περιβάλλον γύρω του, και εν συνεχεία επιλέγει και εκτελεί μια δράση. Η δραστηριότητα αυτή έχει δύο άμεσες επιπτώσεις. Πρώτη επίπτωση, να αλλάξει την κατάσταση του περιβάλλοντος και δεύτερη, ο πράκτορας να αποκτήσει ένα άμεσο αριθμητικό όφελος (θετικό ή αρνητικό) ως αποτέλεσμα της δράσης που πραγματοποίησε.

Στο Σχήμα 3.1 παρουσιάζεται ένα τυπικό μοντέλο ενισχυτικής μάθησης. Όπως απεικονίζεται και στο Σχήμα 3.1, σε κάθε βήμα αλληλεπίδρασης με το περιβάλλον ο πράκτορας λαμβάνει μία ένδειξη της υφιστάμενης κατάστασης s



Σχήμα 3.1: *Μοντέλο ενισχυτικής μάθησης*

του περιβάλλοντος. Στη συνέχεια, ο πράκτορας επιλέγει να πραγματοποιήσει μια δράση α. Η δράση αυτή αλλάζει την κατάσταση του περιβάλλοντος και η αξία αυτής της μετάβασης T του περιβάλλοντος, από τη μία κατάσταση στην άλλη, επικοινωνείται στον πράκτορα μέσω ενισχυτικών σημάτων r. Στόχος είναι η ανάπτυξη μιας μεθοδολογίας ή αλλιώς μιας συμπεριφοράς B, μέσω της οποίας ο πράκτορας θα καταφέρει να μάθει να επιλέγει ενέργειες τέτοιες ώστε να μεγιστοποιεί το μακροπρόθεσμο όφελός του. Η Συμπεριφορά B του πράκτορα, χαρακτηρίζεται από την πολιτική π την οποία θα αναπτύξει ο πράκτορας ως αποτέλεσμα του μηχανισμού Ενισχυτικής Μάθησης που θα υιοθετήσει. Τα οφέλη που λαμβάνει ο πράκτορας είναι υποκειμενικά, με την έννοια ότι, είναι η αποτίμηση του ίδιου του πράκτορα για την υφιστάμενη κατάστασή του. Η ικανότητα αυτή να αξιολογεί την εμπειρία του με αυτόν τον τρόπο είναι ουσιαστικό κομμάτι της λειτουργίας του πράκτορα. Τα οφέλη ουσιαστικά είναι εκείνα που καθορίζουν τι είναι εκείνο που προσπαθεί να επιτύχει ο πράκτορας. Για παράδειγμα, εάν η κίνησή του προς ένα συγκεκριμένο σημείο του αποφέρει μεγαλύτερα οφέλη, τότε προφανώς ο πράκτορας προσπαθεί να επιτύχει την προσέγγιση του σημείου και όχι την απομάκρυνση από αυτό. Συνεπώς, αυτό που πρέπει να μάθει ο πράκτορας είναι πώς να πραγματοποιεί δράσεις οι οποίες θα του αποφέρουν καλύτερα αποτελέσματα.

Τόσο η Ενισχυτική Μάθηση όσο και ο Δυναμικός Προγραμματισμός αποτελούν αλγοριθμικές μεθόδους για την επίλυση προβλημάτων στα οποία δράσεις εφαρμόζονται σε ένα σύστημα για μία εκτεταμένη χρονική περίοδο με σκοπό να

καταφέρει το συγκεκριμένο σύστημα να φτάσει σε κάποιο επιθυμητό στόχο. Η διαφορά τους έγκειται στο ότι οι μέθοδοι δυναμικού προγραμματισμού απαιτούν να υπάρχει μοντέλο το οποίο θα περιγράφει τη συμπεριφορά του συστήματος, ενώ οι μέθοδοι ενισχυτικής μάθησης δεν θέτουν τέτοια προϋπόθεση. Οι δράσεις πραγματοποιούνται σε διακριτό χρονικό βήμα και σε ένα σχήμα κλειστού βρόχου, πράγμα που σημαίνει ότι το αποτέλεσμα των προηγούμενων δράσεων παρατηρείται και λαμβάνεται υπόψη κατά την διαδικασία επιλογής νέων δράσεων. Στην πορεία διάδρασης του συστήματος με το περιβάλλον, ανταμοιβές απονέμονται και αξιολογούν σε κάθε βήμα την απόδοση της δράσης, με στόχο τη βελτιστοποίηση της μακροχρόνιας απόδοσης, η οποία αποτιμάται με βάση τη συνολική ανταμοιβή που συγκεντρώνεται καθ' όλη τη διάρκεια αυτής της διάδρασης.

Όταν υπάρχει μοντέλο του συστήματος, όπως έχει ήδη αναφερθεί, μπορούν να εφαρμοσθούν μέθοδοι Δυναμικού Προγραμματισμού. Πολλές όμως είναι εκείνες οι περιπτώσεις όπου δεν μπορεί να υπάρξει μοντέλο του συστήματος, είτε γιατί το σύστημα δεν είναι απόλυτα γνωστό εκ των προτέρων, είτε διότι δεν είναι επαρκώς κατανοητό ή ακόμα γιατί η δημιουργία ενός μοντέλου είναι ορισμένες φορές εξαιρετικά δαπανηρή. Σε αυτές λοιπόν τις περιπτώσεις οι μέθοδοι Ενισχυτικής Μάθησης είναι εξαιρετικά χρήσιμες δεδομένου ότι λειτουργούν χρησιμοποιώντας μόνο δεδομένα τα οποία αποκτούν από το ίδιο το σύστημα στο οποίο εφαρμόζονται, και χωρίς να απαιτούν την εκ των προτέρων ύπαρξη μοντέλου που να περιγράφει τη συμπεριφορά του συστήματος. Συνεπώς, ενσωματώνοντας ο πράκτορας ένα μηχανισμό Ενισχυτικής Μάθησης, είναι σε θέση να σχηματίσει πλέον μια συμπεριφορά. Η συμπεριφορά ενός πράκτορα ορίζεται από την πολιτική π την οποία ακολουθεί και η οποία είναι μια αντιστοίχιση καταστάσεων σε δράσεις. Στόχος του πράκτορα είναι μέσω αυτής τη πολιτικής που θα σχηματίσει να μεγιστοποιήσει το άθροισμα των ανταμοιβών που θα συγκεντρώσει στη πορεία της διάδρασής του με το περιβάλλον. Ξεκινώντας λοιπόν από τη χρονική στιγμή $t = 0$, δίνει ένα σχετικό βάρος στις ανταποδόσεις που λαμβάνει μέσω ενός συντελεστή ο οποίος μειώνεται καθώς το χρονικό βήμα αυξάνει:

$$\gamma^0 r_1 + \gamma^1 r_2 + \gamma^2 r_3 + \dots \tag{3.1}$$

Ο εκπτωτικός συντελεστής $\gamma \in [0, 1]$ αποτελεί ένα μέτρο για τη βαρύτητα που δίνεται στις ανταμοιβές που θα έρθουν στο μέλλον. Οι ανταμοιβές αυτές εξαρτώνται από την πορεία των καταστάσεων-δράσεων που θα ακολουθήσει στο σύστημα, πορεία η οποία αποτελεί και την πολιτική π που δημιουργεί:

$$s_0, \alpha_0 = \pi(s_0), s_1, \alpha_1 = \pi(s_1), s_2, \alpha_2 = \pi(s_2), \dots \tag{3.2}$$

Κάθε ανταμοιβή r_{t+1} είναι αποτέλεσμα της μετάβασης (s_t, α_t, s_{t+1}). Στόχος, παράλληλα και πρόκληση, για τους αλγορίθμους Ενισχυτικής Μάθησης αποτελεί η εύρεση της λύσης εκείνης η οποία βελτιστοποιεί τη *μακροχρόνια απόδοση* του συστήματος, κάνοντας χρήση όμως αποκλειστικά και μόνο ανταμοιβών που περιγράφουν την *άμεση απόδοση* του συστήματος. Συνεπώς πρέπει να βρεθεί η βέλτιστη πολιτική π^* η οποία μεγιστοποιεί τη σχέση (3.1) για κάθε αρχική κατάσταση. Για να βρεθεί η βέλτιστη αυτή πολιτική, πρέπει να υπολογιστούν οι μέγιστες ανταποδόσεις. Πιο συγκεκριμένα, με τη χρήση μεθόδων Ενισχυτικής Μάθησης πρέπει να βρεθεί η βέλτιστη συνάρτηση Q (Q-function) την οποία θα την αποτυπώνουμε ως Q^*, και η οποία θα αντιστοιχεί, για κάθε ζεύγος κατάστασης - δράσης (s, α) στην ανταπόδοση η οποία προκύπτει όταν έχουμε πραγματοποίηση της δράσης α, στην κατάσταση s και επιλέγοντας εν συνέχεια βέλτιστες δράσεις από το δεύτερο βήμα και μετά:

$$Q^*(s, \alpha) = \gamma^0 r_1 + \gamma^1 r_2 + \gamma^2 r_3 + \dots \tag{3.3}$$

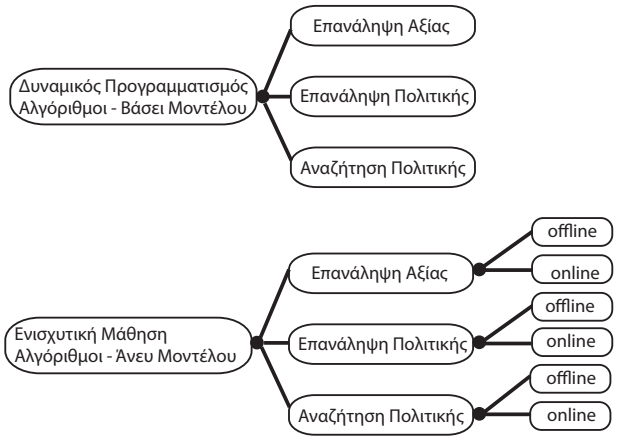
όπου $s_0 = s$, $\alpha_0 = \alpha$, και βέλτιστες δράσεις επιλέγονται για $s_1, s_2, \dots$.

Εν συνέχεια η βέλτιστη πολιτική προκύπτει με την επιλογή σε κάθε κατάσταση s μιας δράσης $\pi^*(s)$ η οποία μεγιστοποιεί τη βέλτιστη συνάρτηση Q για αυτήν την κατάσταση:

$$\pi^*(s) = \arg\max_{\alpha} Q^*(s, \alpha) \tag{3.4}$$

Για να δούμε λοιπόν πώς προκύπτει η βέλτιστη πολιτική, θυμίζουμε ότι η βέλτιστη συνάρτηση Q^* περιλαμβάνει μέγιστες ανταποδόσεις αρχίζοντας από το δεύτερο βήμα και μετά. Με την παραπάνω σχέση επιλέγεται εκείνη η δράση η οποία επιπλέον μεγιστοποιεί την απόδοση και για το πρώτο βήμα, κατά συνέπεια προκύπτει η συνολική μέγιστη ανταπόδοση.

Ο δυναμικός προγραμματισμός για να μπορέσει να βρει την βέλτιστη πολιτική απαιτεί ένα μοντέλο Μαρκοβιανής Διαδικασίας Λήψης Αποφάσεων (τα προβλήματα Μαρκοβιανών Διαδικασιών θα αναλυθούν στη συνέχεια), καθώς επίσης τις πιθανότητες μετάβασης και τη σχετική συνάρτηση ανταμοιβής [13] [101]. Από την άλλη πλευρά οι αλγόριθμοι Ενισχυτικής Μάθησης δεν χρειάζονται μοντέλο για να λειτουργήσουν [13] [7]. Πιο συγκεκριμένα χρησιμοποιούν δεδομένα τα οποία παράγονται κατά τη διαδικασία μάθησης με τη μορφή δειγμάτων. Από αυτά τα δείγματα ο μηχανισμός Ενισχυτική Μάθησης δημιουργεί μοντέλο το οποίο βέβαια παράγεται από σαφώς πιο περιορισμένο μέγεθος δεδομένων μετάβασης από μια κατάσταση σε άλλη (συγκριτικά πάντα με το Δυναμικό Προγραμματισμό ο οποίος, διαθέτοντας μοντέλο, μπορεί να δημιουργήσει



Σχήμα 3.2: *Αλγόριθμοι ενισχυτικής μάθησης και δυναμικού προγραμματισμού*

οποιονδήποτε αριθμό δειγμάτων μετάβασης). Προφανώς ένας τέτοιος περιορισμός καθιστά μεγαλύτερη την πρόκληση ως προς την επίλυση αυτού του τύπου προβλημάτων.

3.3.1 Αλγόριθμοι Ενισχυτικής Μάθησης

Οι αλγόριθμοι Ενισχυτικής Μάθησης και Δυναμικού Προγραμματισμού μπορούν να διαχωριστούν στις εξής κατηγορίες:

Όπως αποτυπώνεται στο Σχήμα 3.2, οι αλγόριθμοι Ενισχυτικής Μάθησης μπορούν να διαχωριστούν σε τρεις κατηγορίες από τις οποίες θα δούμε στη συνέχεια αναλυτικά μόνο τις δύο: την Επανάληψη Αξίας και την Επανάληψη Πολιτικής.

- **Αλγόριθμοι Επανάληψης Αξίας:**
  Η ομάδα αυτή των αλγορίθμων αναζητά τη βέλτιστη συνάρτηση αξίας, η οποία αποτελείται από τις μέγιστες ανταποδόσεις από κάθε κατάσταση, ή από κάθε ζεύγος κατάστασης-δράσης. Η βέλτιστη συνάρτηση αξίας χρησιμοποιείται στη συνέχεια για να προκύψει η βέλτιστη πολιτική.

- **Αλγόριθμοι Επανάληψης Πολιτικής:**
  Η ομάδα αυτή των αλγορίθμων αξιολογεί πολιτικές κτίζοντας πρώτα τις συναρτήσεις αξίας που τους αντιστοιχούν (αντί να προσπαθούν να σχηματίσουν απευθείας τη βέλτιστη συνάρτηση αξίας), και εν συνεχεία κάνουν χρήση αυτών των συναρτήσεων αξίας για να δημιουργήσουν νέες βελτιωμένες πολιτικές.

- **Αλγόριθμοι Αναζήτησης Πολιτικής:**
  Η ομάδα αυτή των αλγορίθμων κάνει χρήση τεχνικών βελτιστοποίησης για την απευθείας αναζήτηση της βέλτιστης πολιτικής.

Όπως βλέπουμε στο σχετικό σχήμα, υπάρχουν offline αλγόριθμοι Ενισχυτικής Μάθησης καθώς και online μηχανισμοί. Οι offline μηχανισμοί κάνουν χρήση δεδομένων τα οποία έχουν συλλεγεί εκ των προτέρων ενώ οι online μαθαίνουν να λύνουν το σχετικό πρόβλημα από τα δεδομένα που συλλέγουν μέσω της διάδρασης με το περιβάλλον στο οποίο λειτουργεί το σύστημα. Η πρόκληση την οποία έχουν να αντιμετωπίσουν τα συστήματα τα οποία ενσωματώνουν online Ενισχυτική Μάθηση είναι η αναζήτηση μιας ισορροπίας μεταξύ περαιτέρω εξερεύνησης/συλλογής δεδομένων και εκμετάλλευσης της υπάρχουσας γνώσης (exploration vs exploitation).

3.3.2 Προβλήματα Μαρκοβιανών Διαδικασιών Λήψης Αποφάσεων

Τα προβλήματα Ενισχυτικής Μάθησης αντιμετωπίζονται ως προβλήματα Μαρκοβιανής Διαδικασίας Λήψης Αποφάσεων (Markov Decision Process - MDP), με ντετερμινιστικές ή στοχαστικές μεταβάσεις κατάστασης. Ένα πρόβλημα MDP έχει τέσσερις συνιστώσες: (i) το χώρο-καταστάσεων (state-space) S, (ii) το χώρο δράσεων (actions) A , (iii)τη συνάρτηση μετάβασης η οποία μπορεί να είναι στοχαστική ή ντετερμινιστική ($T : S \times A \times S \rightarrow [0, 1]$), απεικονίζοντας την πιθανότητα μετάβασης από την κατάσταση s σε μία άλλη κατάσταση s' πραγματοποιώντας τη δράση α (ή πιο γενικά πώς αλλάζει μία κατάσταση ως αποτέλεσμα των δράσεων του πράκτορα), και τέλος (iv) τη συνάρτηση ανταπόδοσης $R : S \times A \rightarrow \Re$ (όπου $\Re$ το σύνολο τιμών που μπορεί να λάβει η συνάρτηση ανταπόδοσης) η οποία αποτυπώνει την άμεση απόδοση που προκύπτει από την εφαρμογή της δράσης α στην κατάσταση s. Ας υποθέσουμε λοιπόν ότι, η κατάσταση τη χρονική στιγμή t είναι s_t, και χαρακτηρίζει την υφιστάμενη κατάσταση του πράκτορα στο περιβάλλον. Πραγματοποιώντας τη

δράση α_t η κατάσταση αλλάζει σε s_{t+1}, μέσω της συνάρτηση $f : S \times A \rightarrow S$, όπου:

$$s_{t+1} = f(s_t, \alpha_t) \tag{3.5}$$

Την ίδια στιγμή ο πράκτορας λαμβάνει ένα ανταποδοτικό σήμα r_{t+1}, σύμφωνα με τη συνάρτηση ανταπόδοσης R, $R : S \times A \rightarrow \Re$ (όπου $\Re$ το σύνολο τιμών που μπορεί να λάβει η συνάρτηση ανταπόδοσης):

$$r_{t+1} = R(s_t, \alpha_t) \tag{3.6}$$

Αυτή η ανταμοιβή αξιολογεί τα άμεσα αποτελέσματα της δράσης α_t (δηλαδή τη μετάβαση από την κατάσταση s_t στην s_{t+1}), αλλά γενικά δεν δίνει καμία πληροφορία για τις μακροχρόνιες επιπτώσεις της συγκεκριμένης δράσης. Ο πράκτορας επιλέγει δράσεις σύμφωνα με την πολιτική π την οποία ακολουθεί ($\pi : S \rightarrow A$) χρησιμοποιώντας την ακόλουθη σχέση:

$$\alpha_t = \pi(s_t) \tag{3.7}$$

Με δεδομένα τα παρακάτω στοιχεία: τη συνάρτηση μετάβασης f και τη συνάρτηση ανταπόδοσης R, την υφιστάμενη κατάσταση s_t καθώς και την επιλεγείσα δράση α_t, μπορούν να προκύψουν τόσο η επόμενη κατάσταση s_{t+1} καθώς και η ανταπόδοση r_{t+1}. Η υπόθεση ότι οι ανταποδόσεις βασίζονται μόνο στην υφιστάμενη κατάσταση και την υφιστάμενη δράση, ονομάζεται Μαρκοβιανή Ιδιότητα (*Markov Property*), και διασφαλίζει ότι η περιγραφή της υφιστάμενης κατάστασης περιέχει όλες τις πληροφορίες εκείνες που σχετίζονται με την επιλογή μιας δράσης στην τρέχουσα αυτή κατάσταση.

3.3.3 Βελτιστοποίηση Πολιτικής

Όπως έχει ήδη αναφερθεί, στόχος της ενισχυτικής μάθησης είναι η εύρεση της βέλτιστης πολιτικής η οποία θα μεγιστοποιεί την ανταπόδοση από οποιαδήποτε αρχική κατάσταση s_0. Ως ανταπόδοση ορίζουμε τη συνολική συνάθροιση των ανταμοιβών κατά μήκος κάποιας πορείας ξεκινώντας από την αρχική κατάσταση s_0. Κατά ένα τρόπο συνοπτικό λοιπόν αναπαριστά την ανταμοιβή που λαμβάνει ο πράκτορας μακροπρόθεσμα. Υπάρχουν διαφορετικοί τύποι ανταποδόσεων, ανάλογα με τον τρόπο που γίνεται η συνάθροιση των ανταμοιβών [14]. Στην παρούσα διατριβή εφαρμόζουμε την Εκπτωτική Ανταπόδοση Απείρου Ορίζοντα (Infinite Horizon Discounted Return):

$$R^{\pi}(s_0) = \sum_{t=0}^{\infty} \gamma^t r_{t+1} = \sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(s_t)) \tag{3.8}$$

σε συνδυασμό με την εξίσωση (3.6), όπου $\gamma \in [0, 1]$ είναι ο εκπτωτικός συντελεστής και $s_{t+1} = f(s_t, \pi(s_t))$ για $t \geq 0$. Ο εκπτωτικός συντελεστής μπορεί να ερμηνευτεί ως μέσο για τον πράκτορα να σταθμίζει την αυξημένη αβεβαιότητα των μελλοντικών ανταμοιβών. Για παράδειγμα, εάν το γ είναι πολύ μικρό σημαίνει ότι ο πράκτορας δεν λαμβάνει σοβαρά υπόψη του ανταμοιβές τις οποίες εισπράττει μετά από πολλά χρονικά βήματα (δηλαδή μελλοντικές ανταμοιβές). Γενικά η επιλογή του συντελεστή γ αποτελεί μια απόφαση που πρέπει να εξισορροπεί την ταχύτητα σύγκλισης του αλγορίθμου Ενισχυτικής Μάθησης και την ποιότητα της λύσης που θα προκύψει. Στόχος λοιπόν είναι η μεγιστοποίηση της μακροχρόνιας ανταπόδοσης που θα λάβει ο πράκτορας, κάνοντας χρήση μόνο της άμεσης ανάδρασης την οποία εισπράττει, την ανταμοιβή δηλαδή που παίρνει σε κάθε ένα χρονικό βήμα. Αυτός ο στόχος, βεβαίως, δημιουργεί ένα πολύ ενδιαφέρον επιστημονικό πρόβλημα, το οποίο είναι γνωστό στην ερευνητική κοινότητα με τον όρο "Ανταμοιβές με Υστέρηση" (Delayed Rewards) [8]: δράσεις οι οποίες λαμβάνονται στο παρόν, δυνητικά επηρεάζουν την επίτευξη καλών ανταμοιβών στο μέλλον, αλλά, οι άμεσες ανταμοιβές δεν παρέχουν καμία απολύτως πληροφορία για αυτές τις μακροπρόθεσμες επιπτώσεις.

3.3.4 Συναρτήσεις Αξίας και Εξισώσεις Bellman

Μια μέθοδος για το διαχωρισμό ή την αξιολόγηση των πολιτικών που ακολουθεί ένας πράκτορας βασίζεται στις συναρτήσεις αξίας (Value Functions). Υπάρχουν δύο τύποι συναρτήσεων αξίας: Συναρτήσεις αξίας Κατάστασης-Δράσης (State-Action Value Functions) ή αλλιώς Συναρτήσεις Q (Q-functions) και οι Συναρτήσεις αξίας Κατάστασης (State Value Functions) ή αλλιώς Συναρτήσεις V (V-functions). Στην παρούσα διατριβή θα δούμε μόνο συναρτήσεις Q ως μέσο αξιολόγησης μια πολιτικής π. Ορίζουμε λοιπόν ότι η συνάρτηση Q, όπου $Q^{\pi} : S \times A \rightarrow \Re$, για μια πολιτική π, δίνει την ανταπόδοση που αποκτάται όταν ξεκινώντας από μια κατάσταση s, εφαρμόζεται συγκεκριμένη δράση α, και εν συνεχεία ακολουθείται η συγκεκριμένη πολιτική π:

$$Q^{\pi}(s, \alpha) = R(s, \alpha) + \gamma R^{\pi}(f(s, \alpha)) \tag{3.9}$$

Εδώ, $R^{\pi}(f(s, \alpha))$ είναι η ανταπόδοση από την επόμενη κατάσταση $f(s, \alpha)$. Αυτή η συνοπτική εξίσωση (3.9) προκύπτει γράφοντας αρχικά την $Q^{\pi}(s, \alpha)$ αναλυτικά ως εκπτωτικό άθροισμα ανταμοιβών το οποίο προκύπτει πραγματοποιώντας τη δράση α στην κατάσταση s και στη συνέχεια ακολουθώντας την πολιτική π:

$$Q^{\pi}(s,\alpha) = \sum_{t=0}^{\infty} \gamma^t R(s_t, \alpha_t) \tag{3.10}$$

όπου $(s_0, \alpha_0) = (s, \alpha)$, $s_{t+1} = f(s_t, \alpha_t)$, για $t \geq 0$, και $\alpha_t = \pi(s_t)$ για $t \geq 1$. Στη συνέχεια ο πρώτος όρος διαχωρίζεται από το άθροισμα:

$$Q^{\pi}(s,\alpha) = R(s,\alpha) + \sum_{t=1}^{\infty} \gamma^t R(s_t, \alpha_t) \tag{3.11}$$

$$Q^{\pi}(s,\alpha) = R(s,\alpha) + \gamma \sum_{t=1}^{\infty} \gamma^{t-1} R(s_t, \pi(s_t)) \tag{3.12}$$

$$Q^{\pi}(s,\alpha) = R(s,\alpha) + \gamma R^{\pi}(f(s,\alpha)) \tag{3.13}$$

όπου ο ορισμός από την εξίσωση (3.8) για την ανταπόδοση χρησιμοποιήθηκε στο τελευταίο στάδιο. Έτσι προκύπτει η εξίσωση (3.9).
Ως βέλτιστη συνάρτηση Q^* ορίζεται η καλύτερη συνάρτηση Q, η οποία προκύπτει από την αξιολόγηση όλων των πολιτικών:

$$Q^*(s,\alpha) = \max_{\pi} Q^{\pi}(s,\alpha) \tag{3.14}$$

Οποιαδήποτε πολιτική π^* είναι βέλτιστη (μεγιστοποιεί δηλαδή την απόδοση που εισπράττει ο πράκτορας που την εφαρμόζει), εάν επιλέγει σε κάθε κατάσταση τη δράση εκείνη με τη μεγαλύτερη βέλτιστη αξία Q:

$$\pi^*(s) \in \arg\max_{\alpha} Q^*(s,\alpha) \tag{3.15}$$

Γενικά για κάθε δεδομένη συνάρτηση Q, μια πολιτική π η οποία ικανοποιεί την:

$$\pi(s) \in \arg\max_{\alpha} Q(s,\alpha) \tag{3.16}$$

λέγεται "άπληστη" (greedy) πολιτική ως προς τη συνάρτηση Q. Αυτό σημαίνει ότι, για να βρούμε τη βέλτιστη πολιτική, πρέπει πρώτα να βρεθεί η βέλτιστη συνάρτηση αξίας Q^* και εν συνεχεία μέσω της (3.15) να υπολογιστεί μια greedy πολιτική ως προς την Q^*.

Ένα θεμελιώδη μηχανισμό εύρεσης βέλτιστης συνάρτησης αξίας στους αλγορίθμους ενισχυτικής μάθησης, αποτελεί η εξίσωση Bellman, η οποία αναλύεται

για το λόγο αυτό στη συνέχεια. Η εξίσωση του Bellman είναι ένας αναδρομικός μηχανισμός ο οποίος διατυπώνει ότι η αξία της συνάρτησης Q^{π} για την πραγματοποίηση της δράσης α στη κατάσταση s, ακολουθώντας την πολιτική π, ισούται με το άθροισμα της άμεσης ανταμοιβής που προκύπτει από την πραγματοποίηση της συγκεκριμένης δράσης και την εκπτωτική αξία της συνάρτησης Q^{π} η οποία προκύπτει ακολουθώντας την πολιτική π στην επόμενη κατάσταση:

$$Q^{\pi}(s,\alpha) = R(s,\alpha) + \gamma Q^{\pi}(f(s,\alpha), \pi(f(s,\alpha))) \tag{3.17}$$

Η εξίσωση Bellman προκύπτει ως εξής ξεκινώντας από την εξίσωση (3.12):

$$Q^{\pi}(s,\alpha) = R(s,\alpha) + \gamma \sum_{t=1}^{\infty} \gamma^{t-1} R(s_t, \pi(s_t)) \tag{3.18}$$

$$Q^{\pi}(s,\alpha) = R(s,\alpha) + \gamma [R(f(s,\alpha), \pi(f(s,\alpha))) + \gamma \sum_{t=2}^{\infty} \gamma^{t-2} R(s_t, \pi(s_t))] \tag{3.19}$$

$$Q^{\pi}(s,\alpha) = R(s,\alpha) + \gamma Q^{\pi}(f(s,\alpha), \pi(f(s,\alpha))) \tag{3.20}$$

όπου $(s_0, \alpha_0) = (s, \alpha)$, $s_{t+1} = f(s_t, \alpha_t)$ για $t \geq 0$ και $\alpha_t = \pi(s_t)$ για $t \geq 1$. Η εξίσωση βελτιστοποίησης Bellman ορίζει ότι για τη βέλτιστη συνάρτηση Q^*, η βέλτιστη αξία μιας δράσης α, η οποία πραγματοποιήθηκε στην κατάσταση s, ισούται με το άθροισμα της άμεσης ανταμοιβής και της εκπτωτικής βέλτιστης αξίας η οποία προέκυψε μέσω της επιλογής της καλύτερης δυνατής δράσης στην επόμενη κατάσταση:

$$Q^*(s,\alpha) = R(s,\alpha) + \gamma \max_{\alpha'} Q^*(f(s,\alpha), \alpha') \tag{3.21}$$

3.3.5 Αλγόριθμος Επανάληψης Αξίας

Οι τεχνικές επανάληψης αξίας κάνουν χρήση της εξίσωσης βελτιστοποίησης Bellman έτσι ώστε αναδρομικά να υπολογίσουν μια βέλτιστη συνάρτηση αξίας, απ' όπου στη συνέχεια θα προκύψει μια βέλτιστη πολιτική. Χαρακτηριστικός αλγόριθμος επανάληψης αξίας είναι ο μηχανισμός Q-Learning [128] [129]. Ο μηχανισμός αυτός ξεκινά από τυχαία συνάρτηση Q, έστω Q_0, και στη συνέχεια την ανανεώνει βάσει των δεδομένων τα οποία συγκεντρώνει παρατηρώντας τις μεταβάσεις από τη μία κατάσταση σε άλλη και τις σχετικές ανταμοιβές

που προκύπτουν. Τα δεδομένα τα οποία συγκεντρώνει είναι 4άδες της μορφής $(s_t, \alpha_t, s_{t+1}, r_{t+1})$. Μετά από κάθε μετάβαση, η συνάρτηση Q ανανεώνεται, κάνοντας χρήση των παραπάνω δεδομένων ως ακολούθως:

$$Q_{t+1}(s_t, \alpha_t) = Q_t(s_t, \alpha_t) + \lambda_t[r_{t+1} + \gamma \max_{\alpha'} Q_t(s_{t+1}, \alpha') - Q_t(s_t, \alpha_t)] \quad (3.22)$$

όπου λ_t είναι ο ρυθμός μάθησης. Ο όρος στις αγκύλες είναι η χρονική διαφορά (Temporal Difference) μεταξύ της νέας εκτίμησης $r_{t+1} + \gamma \max_{\alpha'} Q_t(s_{t+1}, \alpha')$ της βέλτιστης αξίας Q για (s_t, α_t), και της τρέχουσας εκτίμησης $Q_t(s_t, \alpha_t)$. Καθώς ο αριθμός των μεταβάσεων t προσεγγίζει το άπειρο, ο μηχανισμός Q-Learning ασυμπτωτικά συγκλίνει στη βέλτιστη Q^*, εφόσον ο χώρος καταστάσεων καθώς και ο χώρος δράσεων είναι διακριτός και πεπερασμένος (finite) και επιπλέον ισχύουν οι ακόλουθοι περιορισμοί [129] [125] [59]:

- Το άθροισμα $\sum_{t=0}^{\infty} \lambda_t^2$ παράγει πεπερασμένη (finite) τιμή, ενώ το άθροισμα $\sum_{t=0}^{\infty} \lambda_t$ τείνει στο άπειρο.
- Όλα τα ζεύγη κατάστασης-δράσης μπορούν να είναι επισκέψιμα (ή αλλιώς να επιλέξιμα) απεριόριστα συχνά.

Ο πρώτος περιορισμός δεν είναι δύσκολο να ικανοποιηθεί. Για παράδειγμα, μια τυπική ικανοποιητική επιλογή είναι η $\lambda_t = 1/t$. Στην πράξη βέβαια απαιτείται επιπλέον ρύθμιση της παραμέτρου, διότι επηρεάζει τον αριθμό των μεταβάσεων που απαιτείται έτσι ώστε να προκύψει ικανοποιητική λύση. Συνεπώς, το συμπέρασμα είναι ότι η τιμή της παραμέτρου λ εξαρτάται από το πρόβλημα το οποίο έχουμε να επιλύσουμε. Ο δεύτερος περιορισμός είναι ότι ο πράκτορας πρέπει να έχει μη-μηδενική πιθανότητα να επιλέξει οποιαδήποτε δράση σε κάθε κατάσταση στην οποία μεταβαίνει, πράγμα που ουσιαστικά επιτρέπει την εξερεύνηση του χώρου καταστάσεων από τον πράκτορα. Βέβαια, όπως έχει ήδη αναφερθεί, ο πράκτορας θα πρέπει να είναι σε θέση να ισορροπεί μεταξύ περαιτέρω εξερεύνησης ή εκμετάλλευσης της τρέχουσας γνώσης. Αυτό μπορεί να πραγματοποιηθεί μέσω ενός κλασικού μηχανισμού επιλογής δράσεων για την περίπτωση μάθησης Q-Learning, ο οποίος καλείται " ϵ-Greedy εξερεύνηση" [120], και ο οποίος προβλέπει για την επιλογή δράσης τα ακόλουθα:

$$\alpha_t = \begin{cases} \alpha = \arg\max_{\alpha'} Q_t(s_t, \alpha') & \text{με πιθανότητα } 1 - \epsilon_t \\ \alpha \text{ τυχαία δράση στο σύνολο A} & \text{με πιθανότητα } \epsilon_t \end{cases} \quad (3.23)$$

όπου $\epsilon_t \in (0, 1)$ είναι η πιθανότητα εξερεύνησης στο χρονικό βήμα t.

Στη συνέχεια και στο Σχήμα 3.3 παρουσιάζεται ο αλγόριθμος Q-Learning με ϵ-Greedy μηχανισμό εξερεύνησης.

Select Discount Factor γ
Select Exploration Coefficient $\{\epsilon_t\}_{t=0}^{\infty}$
Select Learning Rate $\{\lambda_t\}_{t=0}^{\infty}$
Initialize Q-function, e.g. $Q_0 \leftarrow 0$
Measure Initial State s_0
for every time step $t = 0, 1, 2, \ldots$ **do**
. . . .

$$\alpha_t = \begin{cases} \alpha = \arg\max_{\alpha'} Q_t(s_t, \alpha') & \text{με πιθανότητα } 1 - \epsilon_t \text{ (exploit)} \\ \alpha \text{ τυχαία δράση στο σύνολο A} & \text{με πιθανότητα } \epsilon_t \text{ (explore)} \end{cases}$$

. . . . **apply** α_t, **measure** s_{t+1} **and reward** r_{t+1}
. . . . $Q_{t+1}(s_t, \alpha_t) = Q_t(s_t, \alpha_t) + \lambda_t[r_{t+1} + \gamma \max_{\alpha'} Q_t(s_{t+1}, \alpha') - Q_t(s_t, \alpha_t)]$
end for

Σχήμα 3.3: *Αλγόριθμος Q-Learning*

3.3.6 Αλγόριθμος Επανάληψης Πολιτικής

Έχοντας εξετάσει το μηχανισμό επανάληψης αξίας ως μια κατηγορία αλγορίθμων Ενισχυτικής Μάθησης, στη συνέχεια μελετούμε το μηχανισμό επανάληψης πολιτικής. Οι αλγόριθμοι επανάληψης πολιτικής αξιολογούν πολιτικές μέσω της δημιουργίας των αντίστοιχων συναρτήσεων αξίας και εν συνεχεία, κάνοντας χρήση αυτών των συναρτήσεων αξίας, βρίσκουν νέες βελτιωμένες πολιτικές. Ας σκεφτούμε έναν αλγόριθμο offline, ο οποίος αξιολογεί πολιτικές στη βάση των συναρτήσεων αξίας τους (Q-functions). Ο αλγόριθμος ξεκινά κάνοντας χρήση μιας τυχαίας πολιτικής. Σε κάθε βήμα υπολογίζεται η συνάρτηση Q^{π_t} της τρέχουσας πολιτικής. Το βήμα αυτό λέγεται στάδιο αξιολόγησης πολιτικής. Η πραγματοποίηση της αξιολόγησης πολιτικής πραγματοποιείται με την επίλυση της εξίσωσης Bellman την οποία έχουμε ήδη περιγράψει. Μετά την αξιολόγηση της πολιτικής π_t, μια νέα πολιτική π_{t+1}, η οποία είναι Greedy ως προς τη συνάρτηση αξίας Q^{π}, βρίσκεται μέσω:

$$\pi_{t+1}(s) = \arg\max_{\alpha} Q^{\pi_t}(s, \alpha) \tag{3.24}$$

Αυτό το στάδιο λέγεται βελτίωση πολιτικής. Στη συνέχεια παρουσιάζουμε ένα χαρακτηριστικό αλγόριθμο Ενισχυτικής Μάθησης ο οποίος ανήκει στη οικογένεια αλγορίθμων επανάληψης πολιτικής. Ο αλγόριθμος SARSA είναι ένας online αλγόριθμος ο οποίος προτάθηκε από τους Rummery and Niranjan (1994) ως η εναλλακτική στον αλγόριθμο που είδαμε προηγουμένως (Επανάληψη Αξίας Q-Learning). Το όνομα SARSA προκύπτει από τα αρχικά των

στοιχείων που συνθέτουν τα δεδομένα που χρησιμοποιεί ο αλγόριθμος, δηλαδή: State - Action - Reward - (Next) State - (Next) Action. Η ορολογία η οποία θα χρησιμοποιηθεί είναι η ακόλουθη $(s_t, \alpha_t, r_t, s_{t+1}, \alpha_{t+1})$. Ο αλγόριθμος ξεκινά με μια τυχαία αρχική συνάρτηση αξίας Q, έστω Q_0, και σε κάθε βήμα την ανανεώνει κάνοντας χρήση αυτής της εξίσωσης:

$$Q_{t+1}(s_t, \alpha_t) = Q_t(s_t, \alpha_t) + \lambda_t[r_{t+1} + \gamma Q_t(s_{t+1}, \alpha_{t+1}) - Q_t(s_t, \alpha_t)] \quad (3.25)$$

όπου λ_t είναι ο ρυθμός μάθησης. Ο όρος στις αγκύλες είναι η χρονική διαφορά (Temporal Difference) μεταξύ της νέας εκτίμησης $r_{t+1} + \gamma Q_t(s_{t+1}, \alpha_{t+1})$ της αξίας Q για (s_t, α_t), και της τρέχουσας εκτίμησης $Q_t(s_t, \alpha_t)$. Να σημειώσουμε εδώ ότι αυτή η χρονική διαφορά δεν είναι η ίδια που είδαμε στην περίπτωση του αλγορίθμου Q-Learning. Πιο συγκεκριμένα, στην περίπτωση του μηχανισμού Q-Learning η χρονική διαφορά υπολογίζεται με βάσει τη μέγιστη αξία Q στην επομένη κατάσταση, ενώ ο μηχανισμός SARSA υπολογίζεται με βάσει την αξία Q της δράσης που πραγματικά έλαβε ο πράκτορας στην επόμενη κατάσταση. Αυτό σημαίνει ότι ο SARSA πραγματοποιεί online αξιολόγηση της υφιστάμενης πολιτικής που έχει ακολουθηθεί (δίχως συγκεκριμένο μοντέλο αξιολόγησης). Στη συνέχεια και στο Σχήμα 3.4 παρουσιάζεται ο αλγόριθμος SARSA με ϵ-Greedy μηχανισμό εξερεύνησης. Σε αυτόν τον αλγόριθμο, δεδομένου ότι η ανανέωση στο χρονικό βήμα t εμπλέκει και τη δράση α_{t+1}, αυτή η δράση πρέπει να έχει επιλεγεί πριν πραγματοποιηθεί η ανανέωση της συνάρτησης Q.

Select Discount Factor γ
Select Exploration Coefficient $\{\epsilon_t\}_{t=0}^{\infty}$
Select Learning Rate $\{\lambda_t\}_{t=0}^{\infty}$
Initialize Q-function, e.g. $Q_0 \leftarrow 0$
Measure Initial State s_0

$$\alpha_0 = \begin{cases} \alpha = \arg\max_{\alpha'} Q_t(s_t, \alpha') & \text{με πιθανότητα } 1-\epsilon_0 \textbf{ (exploit)} \\ \alpha \textbf{ τυχαία δράση στο σύνολο A} & \text{με πιθανότητα } \epsilon_0 \textbf{ (explore)} \end{cases}$$

for every time step $t = 0, 1, 2, \ldots$ **do**
. . . . **apply** α_t**, measure** s_{t+1} **and reward** r_{t+1}
. . . .

$$\alpha_{t+1} = \begin{cases} \alpha = \arg\max_{\alpha'} Q_t(s_{t+1}, \alpha') & \text{με πιθανότητα } 1-\epsilon_{t+1} \textbf{ (exploit)} \\ \alpha \textbf{ τυχαία δράση στο σύνολο A} & \text{με πιθανότητα } \epsilon_{t+1} \textbf{ (explore)} \end{cases}$$

. . . . $Q_{t+1}(s_t, \alpha_t) = Q_t(s_t, \alpha_t) + \lambda_t[r_{t+1} + \gamma Q_t(s_{t+1}, \alpha_{t+1}) - Q_t(s_t, \alpha_t)]$

Σχήμα 3.4: *Αλγόριθμος SARSA*

3.4 Συνεχής Χώρος-Καταστάσεων

Ο συνεχής Χώρος-Καταστάσεων είναι απαίτηση η οποία προκύπτει από την προσπάθεια την οποία κάνουμε να προσαρμόσουμε νευροβιολογικά μοντέλα σε προβλήματα ελέγχου. Παρόλο που το θεωρητικό κλασικό πλαίσιο της ενισχυτικής μάθησης είναι διακριτό, η μεθοδολογία αυτή στην παρούσα διατριβή ενσωματώνεται σε προβλήματα λήψης αποφάσεων σε συνεχές πεδίο χώρου-καταστάσεων.

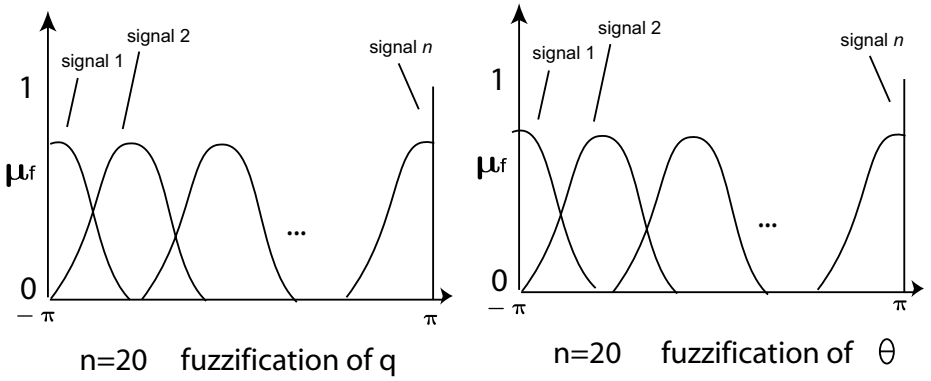
Προς την κατεύθυνση αυτή, αναφέρονται στη βιβλιογραφία αρκετές σχετικές ερευνητικές προσπάθειες. Στην εργασία [73] , περιγράφεται μια περίπτωση εφαρμογής του αλγόριθμου Actor-Critic , όπου η χρησιμοποίηση σε αυτή νευρωνικών δικτύων για την υλοποίηση της σχετικής μονάδας του Actor και της αντίστοιχης του Critic, επιτυγχάνει ευσταθή ελέγχου ενός ανάστροφου εκκρεμούς με δύο συνδέσμους. Ωστόσο, σε αντίθεση με την προσέγγιση η οποία προτείνεται στην παρούσα διατριβή, αυτό το σύστημα ελέγχου δεν βασίζεται σε κάποια κατανεμημένη ή πολυπρακτορική αρχιτεκτονική, αντιμετωπίζεται δηλαδή ως ένας και μόνο πράκτορας. Συναφής προσέγγιση έχει ακολουθηθεί στην ερευνητική εργασία που παρουσιάζεται [130] για την λύση του ίδιου προβλήματος. Σε αυτήν την περίπτωση, ο χώρος-καταστάσεων και ο χώρος-δράσεων παρουσιάζουν συνέχεια και το σύστημα εφαρμόζει μία ειδική κατηγορία του μηχανισμού μάθησης Action-Critic η οποία χρησιμοποιεί νευρωνικούς προσεγγιστές (Neural Approximators) και ο οποίος μηχανισμός καλείται Randomized Policy Optimizer [130]. Μία επιπλέον περίπτωση όπου το πρόβλημα αντιμετωπίζεται σε συνεχή χρόνο και χώρο σε συνδυασμό με χρονική διαφορά και Actor-Critic μέθοδο παρουσιάζεται στην εργασία [34] από τον Kenji Doya και αξιολογείται σε μια εργασία ελέγχου της ταλάντωσης ενός εκκρεμούς με περιορισμένη ροπή. Και οι δύο μονάδες Actor και Critic έχουν υλοποιηθεί μέσω νευρωνικών δικτύων τύπου Ακτινικών Συναρτήσεων Βάσης (Radial Basis Function-RBF). Άλλο πρόβλημα ενισχυτικής μάθησης το οποίο συνδυάζει πολυπρακτορικό περιβάλλον σε συνεχές χρονικό πεδίο εφαρμογής παρουσιάζεται στην εργασία [42] στο πλαίσιο εφαρμογής ελέγχου αυτοκινούμενων ρομπότ σε ευέλικτα βιομηχανικά συστήματα.

Συνοψίζοντας λοιπόν, τονίζουμε ότι η ερευνητική προσπάθεια της παρούσης διατριβής στοχεύει, μεταξύ άλλων, να καλύψει εν δυνάμει την απουσία εφαρμογής μεθόδων συνεχούς ενισχυτικής μάθησης στο πεδίο του επιδέξιου ρομποτικού ελέγχου μέσω πολυπρακτορικών συστημάτων. Στη συνέχεια, προχωρούμε στον ορισμό του χώρου-κατάστασης, όπως αυτός προκύπτει από το πρόβλημα το οποίο προσπαθούμε να επιλύσουμε.

3.4.1 Ασαφοποίηση Χώρου-Κατάστασης

Όπως έχουμε ήδη αναφέρει, η ενισχυτική μάθηση παρουσιάζει μία πολύ ουσιαστική αδυναμία η οποία σχετίζεται με την αύξηση του αριθμού της διάστασης του χώρου-καταστάσεων. Ο αριθμός των παραμέτρων ο οποίος θα πρέπει να αποτελέσει αντικείμενο μάθησης για τον πράκτορα ενός συστήματος, μεγαλώνει εκθετικά με το μέγεθος ακόμη και της πιο συνεπτυγμένης κωδικοποίησης κατάστασης. Η προσέγγιση η οποία θα ακολουθηθεί έτσι ώστε να δημιουργηθεί το πλαίσιο εκείνο το οποίο θα επιτρέψει να δοθεί μια λύση στο συγκεκριμένο πρόβλημα, είναι εκείνη της διακριτοποίησης του χώρου κατάστασης και η εισαγωγή της ιδέας της αφαιρετικής προσέγγισης του χρόνου (temporal abstraction) [29]. Μέσω αυτής της προσέγγισης δεν λαμβάνονται αποφάσεις σε κάθε χρονική στιγμή t. Αυτό που πραγματοποιείται είναι η λήψη αποφάσεων οι οποίες για να ολοκληρωθούν απαιτούν ένα παρατεταμένο χρονικό ορίζοντα. Αυτό αποτελεί ένα στοιχείο το οποίο έχουμε υιοθετήσει στο μηχανισμό μάθησης που εφαρμόζουμε και το οποίο έχει ήδη εφαρμοστεί στο τομέα των αυτοκινούμενων ρομποτικών συστημάτων με σημαντικό ποσοστό επιτυχίας [42].

Έχοντας λοιπόν αναφερθεί στην ανάγκη μείωσης της διάστασης του χώρου-καταστάσεων, προχωρούμε με την περιγραφή της μεθοδολογίας την οποία προτείνουμε. Σύνηθες εργαλείο για την δημιουργία ενός συνεχούς πεδίου από τη μία και διατήρησης των διαστάσεων του χώρου σε διαχειρίσιμα επίπεδα, αποτελεί η ασαφής λογική [103]. Με την εισαγωγή ασαφοποίησης, επιτυγχάνουμε τη συνέχεια που απαιτείται χωρίς να δημιουργούνται μεγάλοι πίνακες αποθήκευσης των παραμέτρων του πράκτορα. Κάθε παράμετρος η οποία ανήκει στην κατάσταση s ασαφοποιείται με χρήση συγκεκριμένων συναρτήσεων που είναι γνωστές ως συναρτήσεις συμμετοχής (*Membership Function*). Ένα ασαφές σύνολο f του υπερσυνόλου U είναι ένα σύνολο το οποίο ορίζεται με τη συνάρτηση συμμετοχής (membership function) $\mu_f : U \rightarrow [0, 1]$. Η τιμή $\mu_f(U)$ για το ασαφές σύνολο f καλείται τιμή συμμετοχής (membership value ή grade of membership) του $u \in U$. Η ουσιαστική έννοια της συνάρτησης συμμετοχής είναι ο βαθμός ή αλλιώς το ποσοστό κατά το οποίο το u ανήκει στο συγκεκριμένο ασαφές σύνολο f. Όταν το υπερσύνολο U είναι άπειρο τότε το ασαφές σύνολο f μπορεί να εκφραστεί ως $f = \int_u \mu_f(u_i)/u_i$. Αναφέρουμε εδώ ότι μια πολύ αναλυτική περιγραφή σχετικά με τα ασαφή σύνολα παρατίθεται στις αναφορές [124], [126], [103] . Πιο συγκεκριμένα, στην περίπτωση των προβλημάτων που εξετάζουμε στην παρούσα διατριβή, εάν θεωρήσουμε τις παραμέτρους της άρθρωσης $< q, \theta, d, \vec{g} >$, οι οποίες ανήκουν στο s, έχουμε συναρτήσεις συμμετοχής της μορφής εκείνης που απεικονίζονται στο Σχήμα 3.5. Με αντίστοιχο μηχανισμό πραγματοποιείται η ασαφοποίηση του χώρου δράσεων. Στις παραγράφους που ακολουθούν εξειδικεύουμε πλέον το μηχανισμό μάθησης και επιλογής δράσης, καθώς και τη συνάρτηση ανταπόδοσης στις συγκεκριμένες περιπτώσεις των



Σχήμα 3.5: *Παραδείγματα ασαφοποίησης παραμέτρων q και θ, μεταβλητές του $s \in S$*

κινηματικών αλυσίδων καθώς και των αυτοκινούμενων ρομπότ.

3.5 Ενισχυτική Μάθηση για Κινηματικές Αλυσίδες

Στις παραγράφους που ακολουθούν αναλύουμε το μηχανισμό μάθησης ο οποίος ενσωματώνεται στην περίπτωση μιας κινηματικής αλυσίδας. Αμέσως μετά παρουσιάζουμε τον αντίστοιχο μηχανισμό μάθησης τον οποίο και ενσωματώνουμε στην εφαρμογή της προτεινόμενης αρχιτεκτονικής στο πλαίσιο των αυτοκινούμενων ρομπότ.

Η μάθηση γενικά ως διαδικασία είναι άμεσα εξαρτημένη από την εμπειρία την οποία έχει ένας πράκτορας και πολύ συχνά χαρακτηρίζεται από την αλλαγή συμπεριφοράς η οποία προκαλείται σε αυτόν από την εξάσκηση κάποιας πρακτικής. Σε αυτήν την παράγραφο θα αναλύσουμε το μηχανισμό μάθησης τον οποίο υιοθετούμε για την κινηματική αλυσίδα, καθώς και τον τρόπο με τον οποίο το πολυπρακτορικό σύστημά μας δημιουργεί γνώση. Η ενισχυτική μάθηση έχει εφαρμοστεί σε σημαντικό αριθμό περιπτώσεων [123][76][60][113][115][114], κυρίως σε αυτοκινούμενα ρομπότ. Στην περίπτωση που εξετάζουμε, η ενισχυτική μάθηση εφαρμόζεται σε ένα αρκετά διαφορετικό πεδίο, εκείνο του πολυπρακτορικού επιδέξιου ρομποτικού χειρισμού. Στη σχετική εργασία [89], παρουσιάζεται μια πολυπρακτορική διάταξη για τον έλεγχο ενός πολυδακτυλικού ρομποτικού χεριού το οποίο δημιουργεί γνώση σε επίπεδο πράκτορα, συνδυάζοντας προϋπάρχοντα μοντέλα συμπεριφορών, χωρίς όμως την ενσωμάτωση κάποιου μηχανισμού ενισχυτικής μάθησης. Στις περιπτώσεις [110] και [111], ενισχυτική μάθηση εφαρμόζεται σε επιδέξιους ρομποτικούς χειριστές οι οποίοι όμως δεν

είναι πολυπρακτορικοί. Αντιθέτως, το ερευνητικό πλαίσιο της παρούσης διατριβής, όπως έχει ήδη αναφερθεί, αποτελεί μια προσπάθεια δημιουργίας μιας ιεραρχικής-εμφωλευμένης πολυπρακτορικής αρχιτεκτονικής η οποία, κάνοντας χρήση μεθόδων ενισχυτικής μάθησης προσαρμοσμένων σε συνεχή χώρο καταστάσεων και δράσεων αποσκοπεί να διευκολύνει την απόκτηση γνώσης και την ανάπτυξη δεξιοτήτων σε συστήματα επιδέξιου ρομποτικού χειρισμού.

Ας υποθέσουμε λοιπόν ότι σε ένα σύνολο n ομογενών (homogeneous) πρακτόρων, κάθε πράκτορας $i \in n$ έχει στη διάθεσή του ένα πεπερασμένο (finite) σύνολο διακριτών δράσεων A_i από τις οποίες καλείται να επιλέξει. Οι πράκτορες λειτουργούν επαναληπτικά μέσα στο πλαίσιο του συγκεκριμένου περιβάλλοντος, όπου κάθε πράκτορας καλείται ανεξάρτητα από τους υπόλοιπους να επιλέξει - εκτελέσει μία διακριτή δράση. Στη σχετική εργασία [62], η διαδικασία της ενισχυτικής μάθησης ορίζεται ως πρόβλημα το οποίο ο πράκτορας καλείται να επιλύσει μέσω της ανάπτυξης σχετικής συμπεριφοράς, η οποία θα προκύψει μέσω αλληλεπιδράσεων με το δυναμικά μεταβαλλόμενο περιβάλλον. Κατά τη διαδικασία ενισχυτικής μάθησης, ένας πράκτορας ουσιαστικά αντιμετωπίζει ένα πρόβλημα Μαρκοβιανής Διαδικασίας Λήψης Αποφάσεων (Markov Decision Process (MDP)). Ένα πρόβλημα MDP έχει τέσσερις συνιστώσες: καταστάσεις, δράσεις, μεταβάσεις και ανταποδόσεις. Πιο συγκεκριμένα, ένα MDP είναι ένα σύνολο τεσσάρων στοιχείων (S, A, T, r), όπου το S υποδεικνύει ένα πεπερασμένο σύνολο καταστάσεων, A ορίζει το χώρο δράσεων, T είναι η πιθανολογική (probabilistic) συνάρτηση μετάβασης $T : S \times A \times S \rightarrow [0, 1]$, η οποία ορίζει την πιθανότητα μετάβασης από την τρέχουσα κατάσταση s σε μία νέα κατάσταση s' όταν εφαρμόζεται συγκεκριμένη δράση α, και $r : S \times A \rightarrow \Re$ είναι η συνάρτηση ανταπόδοσης που ορίζει την τιμή της ανταπόδοσης η οποία προκύπτει από την εφαρμογή συγκεκριμένης δράσης α σε συγκεκριμένη αντίστοιχη κατάσταση s.

Ας προχωρήσουμε στον ορισμό λοιπόν της κατάστασης του συστήματός μας. Δεδομένης της πολυπρακτορικής αρχιτεκτονικής την οποία και είδαμε σε προηγούμενο κεφάλαιο, τόσο η κατάσταση του κάθε ανεξάρτητου πράκτορα όσο και η κατάσταση του συστήματος στο σύνολό του δύνανται να οριστούν ως: $< q_i, \theta_i, d_i, \vec{g_i} >$ (όπου i είναι ο δείκτης που αναφέρεται στον κάθε ανεξάρτητο πράκτορα i). Για την περίπτωση της κινηματικής αλυσίδας τεσσάρων βαθμών ελευθερίας, ο αντίστοιχος ορισμός της κατάστασης για τον κάθε ανεξάρτητο πράκτορα και γενικότερα για το σύνολο του συστήματος είναι ο ακόλουθος: $S_t = \{< q_1, \theta_1, d_1, \vec{g_1} >, < q_2, \theta_2, d_2, \vec{g_2} >, < q_3, \theta_3, d_3, \vec{g_3} >, < q_4, \theta_4, d_4, \vec{g_4} >\}$, τη δεδομένη χρονική στιγμή t. Γενική τάση για το σύνολο των πρακτόρων αποτελεί η επιλογή δράσεων κατάλληλων να μεγιστοποιούν την προσδοκώμενη ανταπόδοση. Κάθε πράκτορας συνεισφέρει με την δική του δράση στο σχημα-

τισμό αυτού που αποκαλούμε συλλογική δράση του συστήματος και η οποία θα είναι αυτή που θα εφαρμόζεται στο περιβάλλον έτσι ώστε να προσδιορίζεται στη συνέχεια η αντίστοιχη μετάβαση. Ο στόχος του πολυπρακτορικού συστήματος είναι να βρεθεί η πολιτική εκείνη η οποία θα μεγιστοποιεί σε βάθος χρόνου το άθροισμα των εκπτωτικών ανταποδόσεων (sum of discounted rewards) του πολυπρακτορικού συστήματος.

Προχωρώντας περαιτέρω την ανάλυση της μαθησιακής διαδικασίας που ενσωματώσαμε στην προτεινόμενη αρχιτεκτονική και στο συγκεκριμένο πλαίσιο της κινηματικής αλυσίδας, εξετάζουμε στη συνέχεια εν συντομία κάποια βασικά στοιχεία από το χώρο της θεωρίας παιγνίων [90]. Μια στοχαστική (randomized) πολιτική για ένα πράκτορα i, είναι μια κατανομή $\pi \in \Delta(A_i)$ (όπου $\Delta(A_i)$ είναι σύνολο κατανομών στο σύνολο των δράσεων A_i του πράκτορα). Συνεπώς, $\pi(a^i)$ υποδεικνύει την πιθανότητα ο πράκτορας i να επιλέξει την δράση a^i. Μια πολιτική π ορίζεται, λοιπόν, ως ντετερμινιστική εφόσον $\pi(a^i) = 1$ για κάποιες $a^i \in A_i$. Το σύνολο των πολιτικών (είτε αυτές είναι ντετερμινιστικές, είτε στοχαστικές) για κάθε πράκτορα i ορίζεται ως Προφίλ Πολιτικής, $\Pi = \{\pi_i : i \in n\}$, όπου n είναι το σύνολο των πρακτόρων. Εφόσον κάθε πολιτική $\pi \in \Pi$ είναι ντετερμινιστική, μπορούμε να θεωρήσουμε το Προφίλ Πολιτικής Π ως μια κοινή/συλλογική δράση (joint action). Καλούμε Μειωμένο Προφίλ για ένα πράκτορα i, ένα προφίλ πολιτικής το οποίο περιλαμβάνει όλους τους πράκτορες του συστήματος εκτός του i (και ορίζεται ως Π_{-i}). Με δεδομένο ένα Π_{-i}, μια πολιτική π_i είναι η καλύτερη δυνατή απόκριση ενός πράκτορα i εάν η προσδοκώμενη αξία του προφίλ πολιτικής $\Pi_{-i} \cup \{\pi_i\}$ είναι η μέγιστη για τον πράκτορα i, δηλαδή ο πράκτορας i δε θα μπορούσε να τα πάει καλύτερα κάνοντας χρήση οποιασδήποτε άλλης πολιτικής π_i'.

3.5.1 Αλγόριθμος Μάθησης Fuzzy Q-Learning

Στα πλαίσια της ενισχυτικής μάθησης για την τοπολογία της κινηματικής αλυσίδας, εφαρμόζεται μία τροποποιημένη έκδοση του αλγορίθμου Q-learning, η οποία ενσωματώνει αρχές Ασαφούς Λογικής. Ο βασικός αλγόριθμος μάθησης Q-learning [128], όπως ήδη επιγραμματικά αναφέρθηκε, βρίσκει μία βέλτιστη πολιτική μέσω της μεγιστοποίησης του συνόλου των ανταποδόσεων που συγκεντρώνει σε βάθος χρόνου ένα σύστημα. Έτσι λοιπόν, ορίζεται η αξία $Q(s_t, \alpha_t)$ για κάθε ζεύγος κατάστασης s_t και δράσης α_t τη χρονική στιγμή t με την σχέση (3.22). Το σύνολο των αξιών αποθηκεύονται σε δομές τύπου πινάκων ανεύρεσης (LookUp Tables). Λόγω της επιθυμητής συνέχειας του χώρου καταστάσεων, ένας τέτοιος πίνακας μπορεί να δημιουργηθεί μόνο προσεγγιστικά. Υιοθετούμε λοιπόν, το μηχανισμό fuzzy Q-learning [44], με στόχο

να προσεγγίσουμε τον πίνακα αξιών Q. Σημαντικός αριθμός εργασιών, στη βιβλιογραφία, αναφέρεται στην προσέγγιση του σχετικού πίνακα αξιών Q, κάποιες κάνοντας χρήση νευρωνικών δικτύων [106], CMAC [119], άλλες εφαρμόζοντας ανάλυση παλινδρόμησης (locally weighted regression) [118] ενώ άλλες μέσω μηχανισμών ασαφούς λογικής [47]. Στην παρούσα διατριβή, ο μηχανισμός fuzzy Q-learning έχει τροποποιηθεί στο σημείο επιλογής δράσης, λαμβάνοντας υπόψη το πολυπρακτορικό χαρακτήρα του συστήματος, έτσι ώστε να υποστηρίζει "Συλλογικές Δράσεις" (Joint Actions). Ο μηχανισμός εφαρμόζει μία συγκεκριμένη προσέγγιση χρησιμοποιώντας μια Βάση Κανόνων Κανόνων Ασαφούς Λογικής (Fuzzy Rule Base - FRB mechanism). Ουσιαστικά πρόκειται για μία συνάρτηση που αντιστοιχίζει το διάνυσμα των παραμέτρων που περιγράφουν (συνθέτουν) την κατάσταση του πράκτορα, σε μία συγκεκριμένη τιμή. Συνεπώς, είναι προφανές ότι απαιτείται να οριστεί ένα σύνολο τέτοιων κανόνων (Fuzzy Rules), κάθε ένας εκ των οποίων θα έχει την ακόλουθη μορφή:

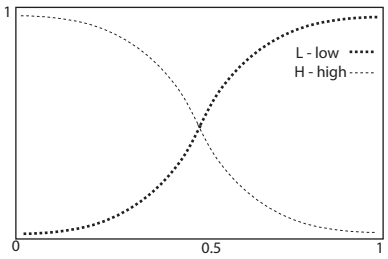
$$\text{Rule-i: IF } (s_1 \in F_1^i) \text{ AND } (s_2 \in F_2^i) \text{ AND } \ldots (s_n \in F_n^i)$$
$$\text{THEN } (\text{output } = \phi_t^i)$$

όπου $\{s_1, \ldots, s_n\}$ είναι οι n παράμετροι του σχετικού διανύσματος κατάστασης, $\{F_1^i, \ldots, F_n^i\}$ είναι ασαφείς συναρτήσεις συμμετοχής οι οποίες χρησιμοποιούνται από τον κανόνα i, και ϕ_t^i είναι η αξία που αντιστοιχεί τη χρονική στιγμή t, στο κανόνα i. Σύμφωνα με τη διαδικασία εξαγωγής συμπερασμάτων βάσει ασαφούς λογικής (fuzzy inference) κατά Tagaki-Sugeno [122], η αξία Q υπολογίζεται ως ακολούθως [50] [49]:

$$Q(s_t, \alpha_t) = \left\{ \frac{\sum_{i=1}^{K} O^i(s_t) \cdot \phi_t^i}{\sum_{i=1}^{K} O^i(s_t)} \right\} \tag{3.26}$$

όπου K είναι ο αριθμός των κανόνων, ενώ O^i είναι ο βαθμός ενεργοποίησης του κανόνα i, στη δεδομένη κατάσταση s_t.

Με στόχο την καλύτερη κατανόηση του μηχανισμού υπολογισμού των τιμών O^i , αναλύουμε στη συνέχεια ένα υποθετικό απλουστευμένο παράδειγμα, όπου η κατάσταση s_t ενός πράκτορα περιλαμβάνει τρεις μεταβλητές κατάστασης (s_1, s_2, s_3), κάθε μία εκ των οποίων περιγράφεται από δύο συναρτήσεις συμμετοχής, όπως ακριβώς αποτυπώνεται στο Σχήμα 3.6. Κάθε μεταβλητή κατάστασης περιγράφεται από δύο σιγμοειδείς καμπύλες των οποίων οι τιμές έστω ότι είναι H και L. Για τη συγκεκριμένη λοιπόν περίπτωση των τριών μεταβλητών με δύο συναρτήσεις συμμετοχής για κάθε μεταβλητή, έχουμε ένα σύνολο οκτώ



Σχήμα 3.6: *Δύο συναρτήσεις συμμετοχής για κάθε μεταβλητή κατάστασης στην περίπτωση μιας απλής κινηματικής αλυσίδας*

κανόνων που θα πρέπει να συμπεριληφθούν στο σχετικό μηχανισμό FRB. Το σύνολο των κανόνων $\{1 \ldots 8\}$, έχουν αντίστοιχα οκτώ βάρη $\{O^1 \ldots O^8\}$. Τα συγκεκριμένα βάρη υπολογίζονται με βάση το ακόλουθο σύνολο πιθανοτήτων: $Prob_H^{s1}, Prob_L^{s1}, Prob_H^{s2}, Prob_L^{s2}, Prob_H^{s3}, Prob_L^{s3}$. Πιο αναλυτικά, η πιθανότητα $Prob_L^{s1}$ ορίζεται ως ακολούθως:

$$Prob_L^{s_1} = \frac{1}{1 + \varrho \cdot \ e^{-\sigma \cdot s_1}} \tag{3.27}$$

όπου ϱ και σ είναι σταθερές τιμές που ορίζονται κατάλληλα, για κάθε μεταβλητή κατάστασης s_1 και: $Prob_H^{s1} = 1 - Prob_L^{s1}$.

Στη συνέχεια, η τιμή $O^i(s_t)$ μπορεί να υπολογιστεί κάνοντας χρήση της παρακάτω εξίσωσης:

$$O^i(s_t) = \prod_{j=1}^{n} Prob_{F_j^i}^{s_j} \tag{3.28}$$

όπου n είναι ο αριθμός των μεταβλητών κατάστασης που έχει η συγκεκριμένη κατάσταση s_t, ενώ i είναι ο συγκεκριμένος κανόνας. Συνεπώς, για το συγκεκριμένο παράδειγμα το οποίο εξετάζουμε, έχουμε αριθμό μεταβλητών κατάστασης $n = 3$ και αριθμό κανόνων $K = 8$:

$$\begin{pmatrix} O^1(s_t) = Prob_H^{s1} \cdot Prob_H^{s2} \cdot Prob_H^{s3} \\ O^2(s_t) = Prob_H^{s1} \cdot Prob_H^{s2} \cdot Prob_L^{s3} \\ \vdots \\ O^8(s_t) = Prob_L^{s1} \cdot Prob_L^{s2} \cdot Prob_L^{s3} \end{pmatrix}$$

Εν συνεχεία θα πρέπει να βρούμε το μηχανισμό με τον οποίο θα ανανεώνονται οι τιμές των ϕ. Αυτό επιτυγχάνεται υπολογίζοντας αρχικά τις μεταβολές δ των τιμών Q, όπως αυτές προκύπτουν από την εξίσωση (3.22), και στη συνέχεια υπολογίζοντας τις σχετικές μεταβολές μέσω της εξίσωσης (3.26). Πιο

συγκεκριμένα από την εξίσωση (3.22), έχουμε:

$$\delta = \Delta Q = \lambda(r_t + \gamma \max_{\alpha} Q(s_{t+1}, \alpha) - Q(s_t, \alpha_t)) \tag{3.29}$$

Στη συνέχεια, οι μεταβολές των τιμών ϕ προκύπτουν μέσω υπολογισμού της παραγώγου, ως προς ϕ (Gradient Calculation), της εξίσωσης (3.26):

$$\Delta\phi_t^i = \Delta Q \cdot \frac{\partial Q(s_t, \alpha_t)}{\partial \phi_t^i} = \delta \cdot \frac{O^i(s_t)}{\sum_{i=1}^{K} O^i(s_t)} \tag{3.30}$$

Ο ακόλουθος ψευδοκώδικας συνοψίζει τον αλγόριθμο fuzzy Q-learning, περιγράφοντας τα στάδια που τον απαρτίζουν. Σημειώνεται ότι ο μηχανισμός επιλογής "Συλλογικών Δράσεων" (Joint Action Selection Mechanism - JASM) παρουσιάζεται σε επόμενη παράγραφο, μαζί με τη σχετική συνάρτηση ανταπόδοσης και τον αντίστοιχο αλγόριθμο JASM.

Initialize ϕ^i **values for all** $i = 1 \dots K$ **rules**
WHILE *Epoch* $\neq$ *Final Epoch*
Current state $s_t = s_0$
Current joint action $\alpha_t = \alpha_0$
. . **WHILE** $t \neq$ *Final Trial step* t
. . . . **Execute joint action** α_t
. . . . **Get next state** s_{t+1}
. . . . **Get reward** r_{t+1}
. . . . **Select joint action** α_{t+1} **by using the JASM**
. . . . **Calculate** $Q(s_t, \alpha_t)$ **by using Eq. (3.26)**
. . . . **Calculate** ΔQ **by using Eq. (3.29)**
. . . . **Calculate** $\Delta\phi_t^i$ **by using Eq. (3.30)**
. . . . **Update** ϕ_t^i **by using** $\Delta\phi_t^i$
. . . . $s_t = s_{t+1}$
. . . . $\alpha_t = \alpha_{t+1}$
. . **End WHILE**
End WHILE

Σχήμα 3.7: *Αλγόριθμος Fuzzy Q-Learning*

Σε ένα συνεχή χώρο καταστάσεων, ο αριθμός των παραμέτρων που καλείται να μάθει ο πράκτορας αυξάνει εκθετικά καθώς ο αριθμός των καταστάσεων μεγαλώνει. Για να επιτύχουμε την επιθυμητή συνέχεια στον χώρο-καταστάσεων χωρίς τη δημιουργία πινάκων ανεύρεσης μεγάλων διαστάσεων, για τη σχετική

αποθήκευση όλων των μεταβλητών του πράκτορα, κάθε μεταβλητή η οποία συμμετέχει στο προσδιορισμό της κατάστασης του πράκτορα και του συστήματος (γωνιακές μετατοπίσεις αρθρώσεων, Ευκλίδειες αποστάσεις και όλα τα σχετικά σήματα που απαιτούνται) ασαφοποιείται με τη χρήση συγκεκριμένων συναρτήσεων οι οποίες όπως είπαμε καλούνται Συναρτήσεις Συμμετοχής (Membership Functions). Στην συγκεκριμένη περίπτωση της κινηματικής αλυσίδας, έχουμε έξι μεταβλητές κατάστασης που ασαφοποιούνται, όπως περιγράφεται στην παράγραφο 3.4.1 μέσω ενός συνόλου σιγμοειδών συναρτήσεων (Σχήμα 3.6). Για παράδειγμα, η γωνιακή θέση της άρθρωσης q_i και η μεταβλητή θ_i έχουν 20 σήματα η κάθε μεταβλητή ενώ κατά την ίδια λογική η μεταβλητή d_i έχει 13 σήματα, ενώ το διάνυσμα $\vec{g}_i$ έχει 7 σήματα για κάθε μία από τις μεταβλητές που το συνθέτουν. Στη συνέχεια περιγράφονται ο μηχανισμός επιλογής δράσης καθώς και η συνάρτηση ανταπόδοσης.

3.5.2 Μηχανισμός Επιλογής Δράσης και Συνάρτηση Ανταπόδοσης

Ο μηχανισμός επιλογής δράσης είναι εξαιρετικά περίπλοκος στην περίπτωση εκείνη όπου η βέλτιστη συλλογική δράση (Joint Action) δεν είναι μοναδική. Στην περίπτωση όπου η συλλογική δράση που επιλέγει να εκτελέσει το πολυπρακτορικό σύστημα είναι τυχαία ή δεν έχει επιλεγεί με αντικειμενικά κριτήρια, τότε δημιουργείται ο κίνδυνος η επιλογή όχι μόνο να μην είναι η βέλτιστη αλλά πιθανόν να είναι μία τελείως μη συντονισμένη κοινή / συλλογική δράση. Συνεπώς, το πρόβλημα που έχουμε να λύσουμε είναι η επιλογή μιας κοινής συλλογικής δράσης για το σύνολο των πρακτόρων του συστήματος που παρουσιάζουμε. Το πρόβλημα αυτό δύναται να προσεγγιστεί με διάφορους τρόπους. Ένας πιθανός τρόπος είναι μέσω επικοινωνίας μεταξύ των πρακτόρων του συστήματος [116]. Άλλος τρόπος είναι με την εδραίωση κανόνων ή συνθηκών που θα περιορίζουν / προσδιορίζουν συμπεριφορές έτσι ώστε να διασφαλίζεται ο επιθυμητός συντονισμός. Εκείνο το οποίο εμείς εφαρμόζουμε στο πλαίσιο της παρούσης διατριβής οδηγεί σε συντονισμένες δράσεις μεταξύ των πρακτόρων μέσω μιας διαδικασίας επαναλαμβανόμενης πραγματοποίησης μιας συγκεκριμένης εργασίας από το σύνολο των πρακτόρων. Αυτό λοιπόν το οποίο γίνεται κατά τη διαδικασία επιλογής δράσης είναι ότι ο κάθε πράκτορας a_i διατηρεί στη μνήμη του μια μεταβλητή η οποία υποδηλώνει τον αριθμό των περιπτώσεων που έχει επιλεγεί στο παρελθόν η συγκεκριμένη δράση από τον ίδιο (αλλά και από τους υπόλοιπους συνεργαζόμενους πράκτορες). Η ιδέα αυτή μολονότι είναι εξαιρετικά απλή, ορισμένες φορές είναι αρκετά αποτελεσματική, και είναι γνωστή ως Fictitious Play (FP) [41],[19]. Πιο συγκεκριμένα, κάθε πράκτορας a_i δημιουργεί ένα μετρητή $C^i(\alpha_k^j)$, για κάθε πράκτορα a_j ο οποίος είναι ορατός από τον

a_i, ο οποίος υποδεικνύει πόσες φορές ο πράκτορας a^j έχει επιλέξει τη δράση $\alpha_k^j \in A_j$ στο παρελθόν. Όταν λοιπόν μια εργασία ανατίθεται στο πολυπρακτορικό σύστημα, ο πράκτορας a_i αντιλαμβάνεται τη σχετική συχνότητα με την οποία ο πράκτορας a_j πραγματοποιεί τις διάφορες κινήσεις του, ως μια ένδειξη για την υφιστάμενη πολιτική του συγκεκριμένου πράκτορα. Πιο αναλυτικά, για κάθε πράκτορας j, ο a_i υποθέτει ότι ο πράκτορας a_j πραγματοποιεί τη δράση $\alpha_k^j \in A_j$ με πιθανότητα:

$$Pr^i(\alpha_k^j) = \frac{C^i(\alpha_k^j)}{\sum_{b^j \in A_j} C^i(b^j)} \tag{3.31}$$

Αξίζει να σημειωθεί ότι στην πλειοψηφία των περιπτώσεων (στη θεωρία παιγνίων) υποτίθεται ότι κάθε πράκτορας δύναται να παρατηρεί τις δράσεις τις οποίες πραγματοποιούν οι συνεργαζόμενοι με αυτόν πράκτορες, με απόλυτη βεβαιότητα. Ο αντίστοιχος μηχανισμός τον οποίο ενσωματώνουμε στο ερευνητικό πλαίσιο της παρούσης διατριβής, είναι σχετικά πιο γενικός, επιτρέποντας σε κάθε πράκτορα να έχει την δυνατότητα παρατήρησης των δράσεων που επιλέγουν οι άλλοι πράκτορες (και κατά συνέπεια της συλλογικής δράσης την οποία το σύστημα επιλέγει) με μία στοχαστική προσέγγιση. Δεδομένου λοιπόν ότι η επιλογή δράσης είναι σαφώς πιο δύσκολη στις περιπτώσεις εκείνες όπου οι πράκτορες δεν γνωρίζουν τις ανταποδόσεις οι οποίες σχετίζονται με τις διάφορες κοινές δράσεις, και με δεδομένο πλέον ότι το πρόβλημα το οποίο παρουσιάζουμε ανήκει σε αυτή την κατηγορία, αντιλαμβανόμαστε πλέον ξεκάθαρα ότι η χρήση μεθόδων ενισχυτικής μάθησης από τους πράκτορες είναι μια απόλυτα δικαιολογημένη επιλογή, όπου χρησιμοποιώντας εμπειρίες του παρελθόντος επιτυγχάνεται προσέγγιση των προσδοκώμενων ανταποδόσεων που θα προκύψουν από την πραγματοποίηση τόσο μεμονωμένων δράσεων από κάθε πράκτορα όσο και από την πραγματοποίηση κοινών δράσεων από το σύνολο των πρακτόρων.

Ο αλγόριθμος Q-learning τον οποίο ήδη αναφέραμε είναι εκείνος που στη παρούσα διατριβή ενσωματώθηκε στο γενικότερο μηχανισμό μάθησης της κινηματικής αλυσίδας. Έτσι λοιπόν ένας πράκτορας πραγματοποιεί μία εκτίμηση όλων των δράσεων που δύναται να πραγματοποιήσει, επιλέγει κάποια, βάσει της προσδοκώμενης αξίας αυτής της δράσης και η οποία υπολογίζεται μέσω μιας συνάρτησης επιλογής, εν συνεχεία παρατηρεί τη σχετική ανταπόδοση η οποία προκύπτει, και τέλος ανανεώνει την τιμή Q-value, ή αλλιώς την αρχική εκτίμηση που είχε για τη συγκεκριμένη δράση. Έχουμε λοιπόν έναν πράκτορα ο οποίος ανανεώνει τις εκτιμήσεις $Q(s,\alpha)$ ως ακολούθως: $Q(s,\alpha) \leftarrow Q(s,\alpha) + \lambda(r - Q(s,\alpha))$ όπου η εκτέλεση της δράσης α είχε ως συνέπεια να προκύψει η ανταπόδοση r. Τέλος, η παράμετρος $\lambda \in [0,1]$ είναι ο ρυθμός μάθησης του πράκτορα.

Όπως ήδη αναφέραμε, ο πράκτορας μέσω ενός σύνθετου μηχανισμού επιλέγει σε κάθε χρονικό βήμα t, να εκτελέσει μία δράση, από το σύνολο των δράσεων που δύναται να πραγματοποιήσει. Ο μηχανισμός επιλογής δράσεων τον οποίο ενσωματώνουμε στο ερευνητικό πλαίσιο της παρούσης διατριβής, είναι ο ϵ-decreasing, ο οποίος ουσιαστικά είναι μία παραλλαγή του μηχανισμού ϵ-greedy [136]. Πιο συγκεκριμένα, μέσω του μηχανισμού ϵ-decreasing ο πράκτορας δύναται, για ένα σαφώς ορισμένο χρονικό διάστημα της συνολικής περιόδου μάθησης, να επιλέγει δράσεις οι οποίες δεν είναι απαραίτητα βέλτιστες. Η δυνατότητα αυτή, του επιτρέπει να πραγματοποιήσει εξερεύνηση του χώρου καταστάσεων-δράσεων. Η δραστηριότητα αυτή είναι εξαιρετικά σημαντική για την συνολική πορεία της μάθησης του πράκτορα, διότι τον ωθεί στην αναζήτηση και πιθανόν στην εξεύρεση ακόμα καλύτερων λύσεων από εκείνες που έχει ήδη ανακαλύψει. Στη συνέχεια, η εξερεύνηση αυτή περιορίζεται και ο πράκτορας πλέον (βάσει της τρέχουσας κατάστασης στην οποία βρίσκεται) επιλέγει πάντα την καλύτερη δυνατή δράση. Στη συνέχεια, θα δούμε αναλυτικά το μηχανισμό επιλογής δράσεων ϵ-decreasing.

Ο μηχανισμός ξεκινά με τον υπολογισμό της πιθανότητας ο πράκτορας a_i, στο χρονικό βήμα t, να πραγματοποιήσει εξερεύνηση: $\epsilon \cdot (T(t)-1)/(T_{max}-1)$, όπου ϵ είναι ο συντελεστής που ορίζει το ποσοστό του χρόνου ως προς τη συνολική διάρκεια της περιόδου μάθησης, όπου θα επιτραπεί στον πράκτορα να πραγματοποιήσει εξερεύνηση. Για παράδειγμα, εάν ο πράκτορας εκτελεί εξερεύνηση κατά το 20% του συνολικού χρόνου μάθησης, τότε $\epsilon = 0.2$. Στη συνέχεια, σε κάθε χρονικό βήμα t, η επιλογή μιας δράσης πραγματοποιείται χρησιμοποιώντας την κατανομή Boltzmann, όπου με πιθανότητα $1-\epsilon \cdot (T(t)-1)/(T_{max}-1)$, κάθε πράκτορας a_i επιλέγει την δράση α^i με την μεγαλύτερη εκτιμώμενη πιθανότητα $\pi(\alpha^i)$:

$$\pi(\alpha^i) = \frac{e^{\frac{E(\alpha^i)}{T}}}{\sum_{\alpha_j^i \in A_i} e^{\frac{E(\alpha_j^i)}{T}}} \tag{3.32}$$

όπου $E(\alpha^i)$ είναι η προσδοκώμενη αξία της δράσης α^i ενώ η παράμετρος T καλείται "θερμοκρασία" και η τιμή της μειώνεται με την πάροδο του χρόνου έτσι ώστε η πιθανότητα αξιοποίησης υφιστάμενης γνώσης να αυξάνεται, ενώ παράλληλα η πιθανότητα αναζήτησης νέας γνώσης να μειώνεται. Πιο συγκεκριμένα, η παράμετρος "θερμοκρασία" προσδιορίζει την πιθανότητα του πράκτορα να διερευνήσει νέες δράσεις: όταν η παράμετρος T είναι αυξημένη, ακόμα και εάν η τιμή $E(\alpha^i)$ μίας δράσης είναι υψηλή, ο πράκτορας δύναται να επιλέξει μία δράση η οποία να φαίνεται λιγότερο επιθυμητή. Αυτή η στρατηγι-

κή διερεύνησης είναι ιδιαίτερα σημαντική σε στοχαστικά περιβάλλοντα όπως αυτό το οποίο εξετάζουμε. Για να έχουμε έναν αποτελεσματικό μηχανισμό διερεύνησης, η "θερμοκρασία" στα αρχικά στάδια της διαδικασίας είναι υψηλή. Στη συνέχεια, η "θερμοκρασία" μειώνεται για να ευνοηθεί η διαδικασία αξιοποίησης της γνώσης που έχει αποκτηθεί, δεδομένου ότι με την πάροδο του χρόνου αυξάνεται η πιθανότητα οι πράκτορες πλέον να έχουν ανακαλύψει τις πραγματικές αξίες των διαφορετικών δράσεων που έχουν στη διάθεσή τους. Η "θερμοκρασία" ορίζεται σε συνάρτηση με τον αριθμό των επαναλήψεων ως ακολούθως: $T(t) = 1 + T_{max} * e^{-st}$ όπου t είναι το χρονικό βήμα, $s \in (0, 1)$ είναι ο ρυθμός βαθμιαίας μείωσης (Decay Factor) και T_{max} είναι η αρχική "θερμοκρασία".

Στη συνέχεια, ας δούμε τον ορισμό της συνάρτησης $E(\alpha^i)$. Η παρουσία πολλαπλών πρακτόρων, όπου κάθε ένας μαθαίνει ταυτόχρονα με τους άλλους, είναι ένα δυνητικό εμπόδιο ως προς την επιτυχημένη εφαρμογή του σχετικού αλγορίθμου (αλλά και γενικά μεθόδων ενισχυτικής μάθησης) σε πολυπρακτορικές διατάξεις, όπως αυτή που εξετάζουμε. Όταν ένας πράκτορας i μαθαίνει την αξία των δράσεων που επιλέγει, παρουσία άλλων πρακτόρων, μαθαίνει σε ένα μη-στατικό (non-stationary) περιβάλλον. Συνεπώς, η σύγκλιση των αξιών Q δεν μπορεί να είναι εγγυημένη. Δύο είναι οι πιθανοί τρόποι μέσω των οποίων ο μηχανισμός Q-learning μπορεί να εφαρμοστεί σε ένα πολυπρακτορικό σύστημα. Ο πρώτος είναι μέσω της εφαρμογής αλγορίθμου μάθησης ανεξάρτητων δράσεων, ενώ ο δεύτερος μέσω της εφαρμογής αλγορίθμου μάθησης συλλογικών δράσεων (Independent Action Learner - Joint Action Learner Algorithm) [25]. Στην περίπτωση του αλγορίθμου μάθησης ανεξάρτητων δράσεων, κάθε πράκτορας μαθαίνει τις αξίες Q των δράσεων που εκείνος πραγματοποιεί ανεξάρτητα από το πώς δρουν οι άλλοι πράκτορες. Ο αλγόριθμος μάθησης συλλογικών δράσεων, από την άλλη, ορίζει ότι οι πράκτορες δεν μαθαίνουν τις αξίες Q των δικών τους μεμονωμένων δράσεων αλλά τις αξίες Q των συλλογικών δράσεων που επιλέγουν, ως ένα σύνολο πρακτόρων. Κάθε πράκτορας σε ένα τέτοιο σύστημα διατηρεί εκτιμήσεις σχετικά με τις πολιτικές των άλλων πρακτόρων. Έτσι, ο πράκτορας i αποτιμά ότι η προσδοκώμενη αξία $E(\alpha^i)$ της δράσης α^i θα είναι:

$$E(\alpha^i) = \sum_{\alpha^{-i} \in A_{-i}} \left\{ Q(s, \alpha^{-i} \cup \{\alpha^i\}) \prod_{\alpha_k^j \in \alpha^{-i}} [Pr^i(\alpha_k^j)] \right\} \tag{3.33}$$

Στη παραπάνω εξίσωση, A_{-i} υποδεικνύει το σύνολο όλων των πιθανών συλλογικών δράσεων οι οποίες δύνανται να πραγματοποιηθούν από το σύνολο των πρακτόρων οι οποίοι θεωρούνται "ορατοί" από τον πράκτορα a_i, στα πλαίσια

της εμφωλευμένης ιεραρχικής δομής που έχουμε, $\alpha^{-i} \in A_{-i}$ είναι μία τέτοια συλλογική δράση την οποία πραγματοποίησαν το σύνολο των πράκτορων, και $\alpha_k^j \in \alpha^{-i}$ είναι μια ανεξάρτητη δράση την οποία πραγματοποίησε ένας πράκτορας a_j ο οποίος ανήκει στο σύνολο των πρακτόρων. Έχοντας αναλύσει τον μηχανισμό επιλογής "Συλλογικών Δράσεων", ο ακόλουθος ψευδοκώδικας συνοψίζει τον αλγόριθμο Joint Action Selection Mechanism - JASM, περιγράφοντας τα στάδια που τον απαρτίζουν.

Initialize *Current_Agent* to *Root* of the hierarchy
WHILE ***Current_Agent*** $\neq$ **NULL do**
. . . . **Let** a_i = ***Current_Agent***
. . . . $\forall$ **agent** $\mathbf{a}_j$ **(visible from** a_i**), Calculate** $P^i(\alpha_k^j)$ **by using Eq. (3.31)** $\forall \alpha_k^j \in A_j$
. . . . $\forall$ **possible action** α^i**, Calculate** $E(\alpha^i)$ **by using Eq. (3.33)**
. . . . $\forall$ **possible action** α^i**, Calculate** $\pi(\alpha^i)$ **by using Eq. (3.32)**
. . . . **Select the action** α^i **(**ε**-decreasing scheme)**
. . . . **Lock action** α^i **to the overall Joint Action:** α_t
. . . . ***Current_Agent*** $\leftarrow$ ***Select_next_visible_Agent***
End WHILE
Return the Joint Action: α_t **for execution**
Receive Reward $R(t)$ **by using Eq. (3.34)**

Σχήμα 3.8: *Αλγόριθμος "Joint Action Selection Mechanism - JASM"*

Τέλος, η ανταπόδοση που λαμβάνει ένας πράκτορας τη χρονική στιγμή t, έχοντας επιλέξει τη συγκεκριμένη δράση α_t και μεταβαίνοντας σε μία νέα κατάσταση, ορίζεται από τη συνάρτηση ανταπόδοσης $R(t)$ η οποία αποτυπώνεται ως εξής:

$$\begin{cases} \texttt{if } (D_{goal}(t) \leq D_{\min}) \wedge (\Delta D_{goal}) \leq 0) \texttt{ then} \\ \quad R(t) = e^{-c \cdot (D_{goal}(t))} \\ \\ \texttt{if } (D_{goal}(t) > D_{\min}) \texttt{ then} \\ \quad R(t) = -2 \\ \\ \texttt{if } (D_{goal}(t) < D_{\min}) \wedge (\Delta D_{goal}) > 0) \texttt{ then} \\ \quad R(t) = -1 \end{cases} \tag{3.34}$$

όπου $D_{goal}(t)$ είναι η απόσταση από τον στόχο τη χρονική επανάληψη t, $D_{\min}$ είναι το κατώφλι απόστασης μετά από το οποίο οι πράκτορες αρχίζουν να λαμβάνουν ανταπόδοση, η παράμετρος $c \in (0, 1)$ προσδιορίζει μέσω της εκθετικής συνάρτησης πόσο άμεσα θα αντιδρά η συνάρτηση ανταπόδοσης στις δράσεις

του πράκτορα, ενώ ΔD_{goal} είναι ο ρυθμός μεταβολής απόστασης από το στόχο.

3.6 Ενισχυτική Μάθηση για Αυτοκινούμενα Ρομπότ

Στις παραγράφους που ακολουθούν αναλύουμε τον μηχανισμό μάθησης ο οποίος ενσωματώνεται στην περίπτωση χειρισμού αντικειμένου μέσω συνεργαζόμενων αυτοκινούμενων ρομπότ. Στη συγκεκριμένη πειραματική διάταξη, τα ρομπότ αναπτύσσουν συνεργατική συμπεριφορά μέσω των δράσεων που επιλέγουν οι τροχοί τους (οι οποίοι αποτελούν τους πράκτορες που συνθέτουν το συγκεκριμένο πολυπρακτορικό σύστημα, όπως είδαμε άλλωστε ήδη στο Κεφάλαιο 2 της παρούσης διατριβής). Έχοντας ορίσει ήδη στην παράγραφο 2.8 το σύνολο των μεταβλητών κατάστασης της συγκεκριμένης πολυπρακτορικής διάταξης, στη συνέχεια μελετούμε την μέθοδο γραμμικής συνάρτησης προσέγγισης (Linear Function Approximation) σε συνδυασμό με τον μηχανισμό ενισχυτικής μάθησης Χρονικής Διαφοράς (Temporal Difference learning - TD(λ)).

3.6.1 Αλγόριθμος Μάθησης TD(λ)

Ο όρος Χρονική Διαφορά προκύπτει από το γεγονός ότι ο συγκεκριμένος μηχανισμός μάθησης καταγράφει, και στην συνέχεια χρησιμοποιεί τις διαφορές στις εκτιμήσεις που πραγματοποιήθηκαν για την αξία Q μίας κατάστασης s, σε διαδοχικά χρονικά διαστήματα $(t, t+1)$, κατά την εξέλιξη της διαδικασίας μάθησης. Πιο συγκεκριμένα, με τον αλγόριθμο αυτό ο πράκτορας (τροχός του αυτοκινούμενου ρομπότ) πραγματοποιεί εκτίμηση για την αξία της μετάβασης από την κατάσταση s_t στην κατάσταση s_{t+1}. Επομένως, εάν η αξία της κατάστασης στην οποία βρίσκεται το σύστημα τη χρονική στιγμή t είναι $Q(s_t, \alpha_t)$, η εκτίμηση για την αξία της επόμενης κατάστασης στην οποία θα μεταβεί το σύστημα τη χρονική στιγμή $t+1$, μέσω της δράσης α_{t+1} θα είναι $Q(s_{t+1}, \alpha_{t+1})$, ενώ η ανανέωση της αξίας $Q(s_t, \alpha_t)$ θα έχει τη μορφή:

$$Q(s_t, \alpha_t) = Q(s_t, \alpha_t) + \eta \left(r_{t+1} + \gamma Q(s_{t+1}, \alpha_{t+1}) - Q(s_t, \alpha_t)\right) \tag{3.35}$$

όπου $\gamma \in [0, 1]$ είναι ο εκπτωτικός συντελεστής και $\eta \in [0, 1]$ είναι ο ρυθμός μάθησης. Ομοίως, για την επόμενη χρονική στιγμή $t+1$ η ανανέωση θα είναι:

$$Q(s_{t+1}, \alpha_{t+1}) = Q(s_{t+1}, \alpha_{t+1}) + \eta \left(r_{t+2} + \gamma Q(s_{t+2}, \alpha_{t+2}) - Q(s_{t+1}, \alpha_{t+1})\right) \tag{3.36}$$

η οποία λαμβάνει υπόψη της και την εκτίμηση της αξίας $Q(s_{t+2}, \alpha_{t+2})$. Έχοντας τώρα μια καλύτερη εκτίμηση για την αξία $Q(s_{t+1}, \alpha_{t+1})$, ο TD(λ) αλγόριθμος

πηγαίνει πίσω και βασιζόμενος σε αυτή τη νέα πληροφορία βελτιώνει τις προηγούμενες εκτιμήσεις για όλες τις αξίες $Q(s, \alpha)$:

$$Q(s, \alpha) = Q(s, \alpha) + \eta \delta_t e(s, \alpha) \tag{3.37}$$

όπου

$$\delta_t = r_{t+1} + \gamma Q(s_{t+1}, \alpha_{t+1}) - Q(s_t, \alpha_t) \tag{3.38}$$

Η συνάρτηση $e(s, \alpha)$ εισάγει μια πρόσθετη παράμετρο κατά τους υπολογισμούς που εκτελεί ο αλγόριθμος για κάθε ζεύγος κατάστασης-δράσης. Η παράμετρος αυτή καλείται ίχνος επιλογής (Eligibility Trace), και είναι ένα ποιοτικό χαρακτηριστικό για κάθε ζεύγος κατάστασης-δράσης. Επιπλέον, τα ίχνη επιλογής αποτελούν ένα είδος προσωρινής μνήμης για τις καταστάσεις που βρέθηκε ο πράκτορας και τις δράσεις που επέλεξε σε αυτές, στο άμεσο παρελθόν. Πιο συγκεκριμένα, τα ίχνη επιλογής ανανεώνονται σε δύο φάσεις. Αρχικά, κατά το χρονικό βήμα t, ανανεώνεται το ίχνος του υφιστάμενου ζεύγους κατάστασης-δράσης $e(s_t, \alpha_t)$ ως ακολούθως:

$$e(s_t, \alpha_t) = e(s_t, \alpha_t) + 1 \tag{3.39}$$

Στη συνέχεια, ανανεώνονται όλες οι αξίες Q μέσω της σχέσης (3.37), και εν συνεχεία ακολουθεί η δεύτερη ανανέωση τιμών για τα ίχνη επιλογής $e(s, \alpha)$, για όλα τα ζεύγη κατάστασης-δράσης (συμπεριλαμβανομένου και του υφιστάμενου ζεύγους κατάστασης-δράσης). Πιο αναλυτικά:

$$e(s, \alpha) = \begin{cases} 0 & \text{if } \alpha_{t+1} \neq \alpha^* \\ \gamma \lambda e(s, \alpha) & \text{if } \alpha_{t+1} = \alpha^* \end{cases} \tag{3.40}$$

όπου $\lambda \in [0, 1]$. Η τιμή της παραμέτρου λ καθορίζει ουσιαστικά το βαθμό εξασθένησης του κάθε ίχνους μάθησης στο χρόνο: μεγαλύτερες τιμές του λ, δηλαδή, οδηγούν σε ίχνη μάθησης τα οποία διατηρούνται ενεργά για μεγαλύτερο χρονικό διάστημα, ενώ για $\lambda = 1$ οδηγούμαστε ουσιαστικά σε μηχανισμούς παράλληλης μάθησης τύπου Monte Carlo.

Όπως απεικονίζεται στη σχέση (3.40), η δεύτερη ανανέωση τιμών εξαρτάται από την επιλογή της δράσης για το επόμενο χρονικό βήμα $t + 1$. Πιο συγκεκριμένα, εάν η δράση α_{t+1} που επιλέγουμε να εκτελέσουμε κατά το επόμενο χρονικό βήμα $t + 1$ είναι βέλτιστη α^*, τότε οι τιμές σε όλα τα ίχνη ανανεώνονται σύμφωνα με το δεύτερο σκέλος της σχέσης (3.40). Στην περίπτωση όπου $\alpha_{t+1} \neq \alpha^*$, τότε η δράση που επιλέγουμε να εκτελέσουμε για το επόμενο χρονικό βήμα δεν είναι βέλτιστη, συνεπώς πραγματοποιούμε εξερεύνηση του χώρου-καταστάσεων και οι τιμές για τα ίχνη ανανεώνονται σύμφωνα με το πρώτο σκέλος της σχέσης (3.40). Συνοπτικά ο αλγόριθμος περιγράφεται

στο Σχήμα. 3.9, όπου θεωρούμε ότι η μάθηση για να ολοκληρωθεί απαιτεί ένα σύνολο εποχών (Epochs), με διάρκεια η κάθε μία έναν συγκεκριμένο αριθμό χρονικών βημάτων (Time Steps). Τέλος, S είναι ο χώρος-καταστάσεων ενώ A είναι ο χώρος-δράσεων.

Initialize $Q(s,\alpha)$ **randomly** $\forall\ s \in S$ **and** $\alpha \in A(s)$
Initialize $e(s,\alpha)$ **with 0** $\forall\ s \in S$ **and** $\alpha \in A(s)$
WHILE *Epoches* $\neq$ *Final Epoch* **do**
. . . . . $s_t \leftarrow s_0, \alpha_t \leftarrow \alpha_0$
. . **WHILE** *Time Step* $\neq$ *Final Time Step* **do**
. . . . . **Execute the action** α_t
. . . . . **Observe the reward** r_{t+1} **and the next state** s_{t+1}
. . . . . **Choose an action** α_{t+1} **according to the state** s_{t+1}
. . . . . $\alpha^* \leftarrow \underset{\alpha \in A(s_{t+1})}{\text{argmax}}\ Q(s_{t+1},\alpha)$
. . . . . $\delta_t \leftarrow r_{t+1} + \gamma Q(s_{t+1},\alpha_{t+1}) - Q(s_t,\alpha_t)$
. . . . . $e(s_t,\alpha_t) \leftarrow e(s_t,\alpha_t) + 1$
. . . . . **FOR ALL**$(s,\alpha) \in S \times A(s)$
. . . . . $Q(s,\alpha) \leftarrow Q(s,\alpha) + \eta\delta_t e(s,\alpha)$
. . . . . **IF** $\alpha_{t+1} = \alpha^*$
. . . . . $e(s,\alpha) \leftarrow \gamma\lambda e(s,\alpha)$
. . . . . **ELSE**
. . . . . $e(s,\alpha) \leftarrow 0$
. . . . . **END IF**
. . . . . **End FOR**
. . . . . $s_t \leftarrow s_{t+1}$, $\alpha_t \leftarrow \alpha_{t+1}$
. . . . . **Increment** *Time Step*
. . **End WHILE**
. . **Increment** *Epoch*
End WHILE

Σχήμα 3.9: *Αλγόριθμος μάθησης TD(λ)*

3.6.2 Μέθοδος Γραμμικής Συνάρτησης Προσέγγισης

Στη συνέχεια, ας ορίσουμε την συνάρτηση προσέγγισης: $f_\theta(s) \simeq Q(s)$, όπου f_θ είναι μία συνάρτηση παραμετροποιημένη ως προς θ, και s είναι το διάνυσμα μεταβλητών κατάστασης. Αντί να ανανεώνουμε τις τιμές της συνάρτησης $Q(s)$ απευθείας, ανανεώνουμε τις τιμές θ. Πιο συγκεκριμένα, προσπαθούμε να υπολογίσουμε ένα διάνυσμα παραμέτρων $\theta \in \mathbb{R}^n$ μιας συνάρτησης προσέγγισης αξίας $Q_\theta : S \rightarrow \mathbb{R}$, έτσι ώστε $Q_\theta(s) = \theta^T O_s$ (όπου $O_s \in \mathbb{R}^n$ είναι ένα διάνυσμα - Feature Vector, το οποίο χαρακτηρίζει την κατάσταση s). Η μέθοδος κλίσης (Gradient Method) είναι μία τυπική προσέγγιση για την ανανέωση των τιμών της παραμέτρου θ, όπου ουσιαστικά οι ανανεώσεις αυτές είναι ανάλογες ως προς την κλίση (Gradient) μιας κατάλληλης αντικειμενικής συνάρτησης,

ως προς θ. Η συνάρτηση του μέσου τετραγωνικού σφάλματος (Mean Squared Error - MSE) ανάμεσα στην συνάρτηση προσέγγισης αξίας Q_θ και στην συνάρτηση πραγματικής αξίας Q, αποτελεί μία τυπική επιλογή αντικειμενικής συνάρτησης. Συνεπώς, ορίζουμε την σχετική αντικειμενική συνάρτηση ως εξής:

$$E(s_t) = (1/2)(Q(s_t) - f_\theta(s_t))^2 \tag{3.41}$$

υπολογίζοντας την κλίση της συνάρτησης E, έχουμε:

$$\nabla_\theta(E) = (Q(s_t) - f_\theta(s_t))(0 - \nabla_\theta f_\theta(s_t)) \tag{3.42}$$

και στη συνέχεια έχουμε:

$$-\nabla_\theta(E) = (Q(s_t) - f_\theta(s_t))\nabla_\theta f_\theta(s_t) \tag{3.43}$$

Έχοντας υπολογίσει την κλίση, η ανανέωση του διανύσματος των παραμέτρων θ είναι:

$$\theta_{t+1} \leftarrow \theta_t + \eta(Q(s_t) - f_\theta(s_t))\nabla_\theta f_\theta(s_t) \tag{3.44}$$

όπου $\eta \in [0, 1]$. Στη συνέχεια, ενσωματώνουμε αυτό το μηχανισμό ανανέωσης των τιμών θ στον μηχανισμό μάθησης TD(λ) βάσει γραμμικής συνάρτησης προσέγγισης, ο οποίος παρουσιάζεται στην επόμενη παράγραφο.

3.6.3 Μέθοδος Γραμμικής Συνάρτησης Προσέγγισης βάσει Κανόνων Ασαφούς Λογικής

Έχοντας περιγράψει ήδη την μέθοδο μάθησης TD(λ) στην παράγραφο 3.6.1, ας εστιάσουμε στην τροποποίηση της μεθόδου αυτής έτσι ώστε να καλύψει τις συγκεκριμένες ανάγκες της πολυπρακτορικής αρχιτεκτονικής για την πειραματική διάταξη των αυτοκινούμενων ρομπότ. Η συγκεκριμένη μέθοδος χρησιμοποιεί μια Βάση Κανόνων Ασαφούς Λογικής (Fuzzy Rule Base - FRB mechanism), όπως και στην περίπτωση που εξετάσαμε ήδη στην παράγραφο 3.5.1. Πιο αναλυτικά, μέσω της μεθόδου γραμμικής συνάρτησης προσέγγισης, ορίζουμε μία συνάρτηση f η οποία αντιστοιχίζει το διάνυσμα των μεταβλητών που περιγράφουν (συνθέτουν) την κατάσταση του πράκτορα, σε μία συγκεκριμένη τιμή. Πιο συγκεκριμένα, $f_\theta(s)$ είναι η συνάρτηση την οποία προσπαθούμε να προσεγγίσουμε και s είναι το διάνυσμα μεταβλητών-κατάστασης. Συνεπώς, απαιτείται να οριστεί ένα σύνολο τέτοιων κανόνων (Fuzzy Rules), κάθε ένας εκ των οποίων

θα έχει την ακόλουθη μορφή:

$$\text{Rule-i: IF } (s_1 \in A_1^i) \text{ AND } (s_2 \in A_2^i) \text{ AND } \ldots (s_n \in A_n^i)$$
$$\text{THEN } (\text{output} = O^i)$$

όπου $\{s_1, \ldots, s_n\}$ είναι οι μεταβλητές του σχετικού διανύσματος κατάστασης, $\{A_1^i, \ldots, A_n^i\}$ είναι ασαφείς συναρτήσεις συμμετοχής οι οποίες χρησιμοποιούνται από τον κανόνα i, ενώ O^i είναι η αξία που αντιστοιχεί στον κανόνα i. Συνεπώς, το αποτέλεσμα χρήσης του μηχανισμού FRB είναι ο σταθμισμένος μέσος όρος του O^i (weighted average):

$$f_\theta(s) = \theta_1 O^1 + \theta_2 O^2 + \theta_3 O^3 + \cdots + \theta_n O^n$$

όπου και πάλι με s συμβολίζουμε το διάνυσμα των n μεταβλητών κατάστασης (δηλαδή: $s = (s_1, s_2, s_3, \ldots, s_n)$).

Στα πλαίσια καλύτερης κατανόησης του μηχανισμού απόκτησης των τιμών O^i, αναλύουμε στην συνέχεια ένα υποθετικό απλουστευμένο παράδειγμα. Ας υποθέσουμε ότι έχουμε ένα διάνυσμα με τρεις μεταβλητές-κατάστασης $(s_1,\ s_2,\ s_3)$, κάθε μία εκ των οποίων περιγράφεται από δύο συναρτήσεις συμμετοχής, με τον ίδιο τρόπο που έχει αποτυπωθεί στο Σχήμα 3.6 για το αντίστοιχο παράδειγμα της Παραγράφου 3.5.1. Σε αυτό το σχήμα βλέπουμε ότι κάθε μεταβλητή-κατάστασης η οποία χρησιμοποιείται ως είσοδος (Input State Variable) στο μηχανισμό FRB, περιγράφεται μέσω δύο σιγμοειδών καμπυλών (οι τιμές έστω ότι είναι πάλι: HIGH και LOW). Για τη συγκεκριμένη περίπτωση των τριών μεταβλητών κατάστασης, έχουμε οκτώ κανόνες που θα πρέπει να συμπεριληφθούν στον σχετικό μηχανισμό FRB.

Όλοι οι κανόνες $\{1, \ldots, 8\}$ δίνουν αντίστοιχα σήματα εξόδου $\{O^1, \ldots, O^8\}$. Αυτά τα σήματα εξόδου στη συνέχεια υπολογίζονται με βάση το ακόλουθο σύνολο πιθανοτήτων: $P_{HIGH}^{s_1}, P_{LOW}^{s_1}, P_{HIGH}^{s_2}, P_{LOW}^{s_2}, P_{HIGH}^{s_3}, P_{LOW}^{s_3}$. Η πιθανότητα $P_{LOW}^{s_1}$ ορίζεται ως εξής:

$$P_{LOW}^{s_1} = \frac{1}{1 + c_i\ e^{-c_j s_1}} \tag{3.45}$$

όπου c_i και c_j είναι σταθερές τιμές που ορίζονται κατάλληλα για την συγκεκριμένη μεταβλητή s_1, ενώ: $P_{HIGH}^{s_1} = 1 - P_{LOW}^{s_1}$. Εν συνεχεία, η τιμή O^i μπορεί να υπολογιστεί ως εξής: $O^i = \prod_{j=1}^{k} P_{A_j^i}^{s_j}$, όπου k είναι ο αριθμός των μεταβλητών κατάστασης και i είναι ο αριθμός των κανόνων. Συνεπώς, στο συγκεκριμένο παράδειγμα που εξετάζουμε έχουμε $k = 3$ και $i = 8$ άρα οι εξισώσεις έχουν ως εξής:

$$\begin{cases} O^1 = P^{s_1}_{HIGH} \cdot P^{s_2}_{HIGH} \cdot P^{s_3}_{HIGH} \\ O^2 = P^{s_1}_{HIGH} \cdot P^{s_2}_{HIGH} \cdot P^{s_3}_{LOW} \\ \vdots \\ O^8 = P^{s_1}_{LOW} \cdot P^{s_2}_{LOW} \cdot P^{s_3}_{LOW} \end{cases}$$

Έχοντας υπολογίσει το σύνολο των O^i μπορούμε να υπολογίσουμε την συνάρτηση $f_\theta(s_1, s_2, s_3)$ καθώς και την κλίση της, $\nabla_\theta f_\theta(s_1, s_2, s_3)$, για κάθε μεταβλητή κατάστασης. Προκύπτει λοιπόν ότι για κάθε θ_i, η κλίση της συνάρτησης $\nabla_{\theta_i} f_\theta(s_1, s_2, s_3)$ είναι O^i.

Στην συνέχεια παρουσιάζουμε τον μηχανισμό μέσω του οποίου ανανεώνονται τα ίχνη επιλογής καθώς και οι τιμές των παραμέτρων θ, στην συγκεκριμένη περίπτωση της μεθόδου γραμμικής προσέγγισης TD(λ). Ο υπολογισμός της παράμετρου δ στο χρονικό βήμα t (σχέση (3.38)), είναι ως εξής:

$$\delta_t \leftarrow r_{t+1} + \gamma f_\theta(s_{t+1}) - f_\theta(s_t) \tag{3.46}$$

Τα ίχνη επιλογής υπολογίζονται ως ακολούθως:

$$e(s) \leftarrow \gamma\lambda e(s) + \nabla_\theta f_\theta(s) \tag{3.47}$$

Έχοντας ανανεώσει τα ίχνη επιλογής, μπορούμε πλέον στη συνέχεια να ανανεώσουμε τις τιμές του συνόλου των παραμέτρων θ ως ακολούθως:

$$\theta \leftarrow \theta + \eta\ \delta_t\ e(s) \tag{3.48}$$

Επανερχόμαστε στο παράδειγμα των τριών μεταβλητών κατάστασης με δύο συναρτήσεις συμμετοχής και έχουμε τα ακόλουθα:

$$f_\theta(s_1, s_2, s_3) = \theta_1 O^1 + \theta_2 O^2 + \theta_3 O^3 + \theta_4 O^4 + \theta_5 O^5 + \theta_6 O^6 + \theta_7 O^7 + \theta_8 O^8 \tag{3.49}$$

Τα 8 ίχνη επιλογής του συγκεκριμένου παραδείγματος υπολογίζονται μέσω της σχέσης (3.47) ως εξής:

$$e_i(s) \leftarrow \lambda\gamma\ e_i(s) + \nabla_{\theta_i} f_\theta(s_1, s_2, s_3) \quad (\text{για κάθε κανόνα } i = 1, \ldots 8) \tag{3.50}$$

απ'οπου προκύπτει η ακόλουθη σχέση ανανέωσης για τα ίχνη επιλογής:

$$e_i(s) \leftarrow \lambda\gamma\ e_i(s) + O^i \quad (\text{για κάθε κανόνα } i = 1, \ldots 8) \tag{3.51}$$

Τέλος, η ανανέωση των τιμών θ είναι η εξής:

$$\theta_i \leftarrow \theta_i + \eta\ \delta\ e_i(s) \quad (\text{για κάθε κανόνα } i = 1, \ldots 8) \tag{3.52}$$

Σε αυτό το σημείο ολοκληρώνεται το θεωρητικό μέρος της παρούσης διατριβής. Το επόμενο κεφάλαιο εστιάζει στην υλοποίηση της προτεινόμενης αρχιτεκτονικής ρομποτικού ελέγχου καθώς επίσης και σε μια σειρά πειραματικών εφαρμογών της προτεινόμενης πολυπρακτορικής διάταξης μαζί με την αντίστοιχη αξιολόγηση των αποτελεσμάτων που προκύπτουν.

Κεφάλαιο 4

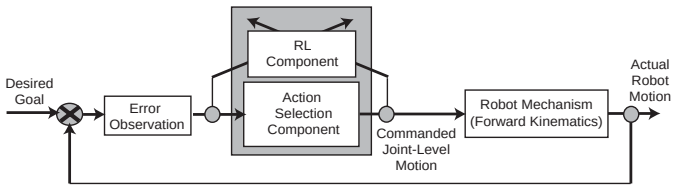
Παραδείγματα Εφαρμογής, Πειραματική Αξιολόγηση και Αποτελέσματα

4.1 Εισαγωγή

Έχοντας μελετήσει έως τώρα, τόσο το γενικότερο ερευνητικό πλαίσιο της εργασίας αυτής, όσο και τις ειδικές ερευνητικές κατευθύνσεις οι οποίες συνθέτουν τους στόχους της εργασίας μας, προχωρούμε στο επόμενο στάδιο το οποίο εστιάζει στην υλοποίηση της αρχιτεκτονικής ρομποτικού ελέγχου καθώς επίσης και σε μια σειρά πειραματικών εφαρμογών της προτεινόμενης πολυπρακτορικής διάταξης μαζί με την αντίστοιχη αξιολόγηση των αποτελεσμάτων που προκύπτουν. Για την αξιολόγηση της προτεινόμενης αρχιτεκτονικής, υιοθετούμε δύο βασικές κατηγορίες προβλημάτων (ρομποτικών εφαρμογών), σε κάθε μία από τις οποίες εκτελείται μια σειρά πειραματικών δοκιμών. Η πρώτη ομάδα πειραματικών δοκιμών αναφέρεται γενικά στον έλεγχο επιδέξιων ρομποτικών χειριστών, και περιλαμβάνει δύο στάδια: σε ένα πρώτο στάδιο προσομοιώνεται ο κινηματικός έλεγχος ενός επίπεδου ρομποτικού χειριστή τεσσάρων συνδέσμων, όπως αυτός που απεικονίζεται στο Σχήμα 2.4, ενώ σε ένα δεύτερο στάδιο, μοντελοποιούμε ρομποτικό χειριστή τριών πολυαρθρωτών δακτύλων, ο οποίος επιχειρεί στατική λαβή (quasi-static grasp) αντικειμένου (θεωρητικά άπειρης μάζας). Στη συνέχεια, η δεύτερη ομάδα πειραματικών δοκιμών αφορά τον έλεγχο συνεργατικών αυτοκινούμενων ρομποτικών οχημάτων. Η συγκεκριμένη πειραματική υλοποίηση που παρουσιάζεται στο κεφάλαιο αυτό περιλαμβάνει δύο αυτοκινούμενα ρομπότ τα οποία συνεργατικά προσπαθούν να πραγματοποιήσουν μια εργασία τύπου "box-pushing". Οι πλατφόρμες προσομοίωσης χρησιμοποιούνται συνθέτουν το πλαίσιο για μια ολοκληρωμένη αξιολόγηση όλων των βασικών μονάδων της αρχιτεκτονικής μας. Πιο συγκεκριμένα, το σύνολο των πειραματικών εφαρμογών περιλαμβάνουν αξιολόγηση της αντίστοιχης πολυπρακτορικής αρχιτεκτονικής η οποία είναι προσαρμοσμένη κάθε φορά στο εκάστοτε πειραματικό πλαίσιο (επιδέξιος χειρισμός / αυτοκινούμενα ρομπότ), της μεθοδολογίας ενισχυτικής μάθησης Q-Learning και του αντίστοιχου υβριδικού μοντέλου ρομποτικού ελέγχου. Το παρόν κεφάλαιο λοιπόν περιλαμβάνει τις διατάξεις των πειραμάτων μονής και πολλαπλής κινηματικής αλυσίδας, καθώς και τα αντίστοιχα αποτελέσματα τα οποία προέκυψαν [64] - [67]. Τέλος παρουσιάζεται ο σχεδιασμός εφαρμογής της προτεινόμενης αρχιτεκτονικής στο πλαίσιο εφαρμογής των αυτοκινούμενων ρομπότ [68].

4.2 Σχήμα RL Ρομποτικού Ελέγχου

Το προτεινόμενο πλαίσιο ρομποτικού ελέγχου το οποίο εξετάζουμε αποτελεί μια υβριδική δομή η οποία ενσωματώνει τρία επίπεδα. Το πρώτο είναι το επίπεδο *παρατήρησης σφάλματος* (error observation). Το δεύτερο είναι το επίπεδο *επιλογής δράσης* (action selection) και τρίτο το επίπεδο *μετάδοσης και*



Σχήμα 4.1: *Η προτεινόμενη RL αρχιτεκτονική ελέγχου*

ελέγχου κίνησης (robot mechanism) (βλ. Σχήμα 4.1). Το πρώτο επίπεδο λαμβάνει ως δεδομένα τον επιθυμητό στόχο για το πολυπρακτορικό σύστημα μαζί με την ανάδραση από το τελευταίο επίπεδο. Η επεξεργασία αυτών των στοιχείων παράγει το σύνολο της πληροφορίας που απαιτεί το επόμενο επίπεδο το οποίο είναι εκείνο της επιλογής δράσης. Όπως φαίνεται και στο σχετικό σχήμα, το επίπεδο επιλογής δράσης είναι συνδεμένο με τη μονάδα Ενισχυτικής Μάθησης. Αυτή η διασύνδεση επιτρέπει τη διεύρυνση των δυνατοτήτων του μηχανισμού επιλογής δράσεων έτσι ώστε να βελτιστοποιεί με την πάροδο του χρόνου τις ενέργειες που επιλέγει. Συγκεκριμένα, ο μηχανισμός ενισχυτικής δράσης απαιτεί δεδομένα τόσο από το επίπεδο παρατήρησης σφάλματος, όσο και από το επίπεδο επιλογής δράσης έτσι ώστε να υποστηρίξει αποτελεσματικά τη διαδικασία επιλογής καλύτερων δράσεων με την πάροδο του χρόνου. Συνεπώς το επίπεδο επιλογής δράσης παράγει, στο επίπεδο των πρακτόρων του συστήματος, δράσεις οι οποίες αντιστοιχούν σε εντολές κίνησης για το σύνολο των αρθρώσεων του ρομποτικού μηχανισμού, ενώ παράλληλα η μονάδα Ενισχυτικής Μάθησης ανατροφοδοτείται με την κατανομή πιθανοτήτων των διαφόρων δράσεων, η οποία πραγματοποιήθηκε στο συγκεκριμένο επίπεδο. Οι δράσεις οι οποίες προωθούνται στο τρίτο επίπεδο, υφίστανται επεξεργασία και προκαλούν την κίνηση των επενεργητών στις αρθρώσεις του ρομποτικού μηχανισμού.

Στη συνέχεια του κεφαλαίου, παρουσιάζουμε μία ολοκληρωμένη αξιολόγηση της προτεινόμενης αρχιτεκτονικής, την οποία πραγματοποιήσαμε μέσω μιας σειράς υπολογιστικών πειραμάτων. Η πρώτη ομάδα δοκιμών περιλαμβάνει πειράματα επιδέξιου ρομποτικού χειρισμού, ενώ η δεύτερη ομάδα δοκιμών εστιάζει σε εφαρμογή επί αυτοκινούμενων συνεργατικών ρομπότ. Στο πλαίσιο της πρώτης ομάδας πειραμάτων, εξετάζουμε αρχικά σειρά προσομοιώσεων απλής κινηματικής αλυσίδας τεσσάρων και επτά συνδέσμων, ενώ στη συνέχεια, προσομοιώνουμε τρία πολυαρθρωτά ρομποτικά δάκτυλα τα οποία μοντελοποιούμε αντίστοιχα με τρεις κινηματικές αλυσίδες τεσσάρων συνδέσμων η κάθε μία. Για την δεύτερη ομάδα πειραμάτων, εξετάζουμε δυο αυτοκινούμενα ρομπότ και αναλύουμε την εφαρμογή τόσο της προτεινόμενης πολυπρακτορικής αρχιτεκτονικής

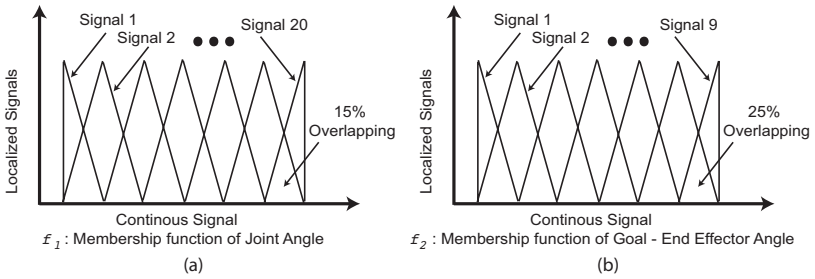
Πίνακας 4.1: Πειραματικές παράμετροι για την περίπτωση της απλής κινηματικής αλυσίδας

	Link 1	Link 2	Link 3	Link 4
Length	3 cm	2 cm	4 cm	3 cm
Initial Joint Angle	0°	20°	30°	40°
Step: Joint Angle	18°	18°	18°	18°
Step: Goal Angle	40°	40°	40°	40°
Overlap of function f_1	15%	15%	15%	15%
Overlap of function f_2	25%	25%	25%	25%

όσο και του σχετικού μοντέλου ενισχυτικής μάθησης. Η δεύτερη αυτή ομάδα πειραμάτων έχει στόχο να διερευνήσει την δυνατότητα αυτόνομης ανάπτυξης συνεργατικών δεξιοτήτων μεταξύ των πρακτόρων, με σκοπό την επίτευξη κοινού στόχου.

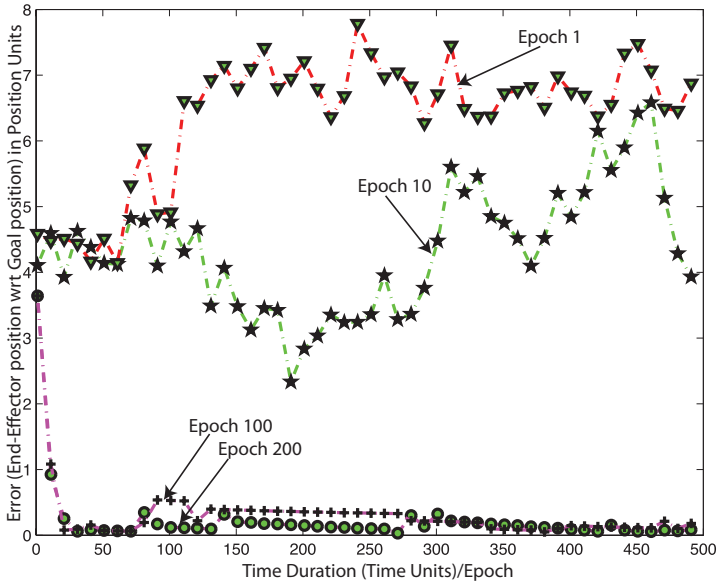
4.3 Απλή Κινηματική Αλυσίδα με Πλεονάζοντες Βαθμούς Ελευθερίας

Το σύνολο των παραμέτρων της διάταξης που χρησιμοποιήθηκαν στην προσομοίωση της απλής κινηματικής αλυσίδας, αποτυπώνονται στο Πίνακα 4.1. Για να επιτευχθεί η επιθυμητή συνέχεια του χώρου, τόσο σε επίπεδο γενικευμένων μεταβλητών των ρομποτικών αρθρώσεων (joint space) όσο και σε επίπεδο μεταβλητών του χώρου εργασίας (Task Space), κάνουμε χρήση συναρτήσεων τύπου f_1 και f_2 οι οποίες περιλαμβάνουν συναρτήσεις συμμετοχής, με σκοπό την ασαφοποίηση του συνόλου των παραμέτρων. Πιο συγκεκριμένα έχουμε, για κάθε γενικευμένη μετατόπιση άρθρωσης (Joint Angle Displacement) ένα σύνολο 20 σημάτων με εύρος 18° το κάθε ένα και με ποσοστό επικάλυψης 15%. Ομοίως, για τις μεταβλητές του χώρου εργασίας (Task Space) έχουμε ένα σύνολο 9 σημάτων με εύρος 40° το κάθε ένα και ποσοστό επικάλυψης 25% (βλ. Σχήμα 4.2). Ο στόχος του συγκεκριμένου πολυπρακτορικού συστήματος είναι διττός: α) να οδηγήσει σε λύσεις οι οποίες ικανοποιούν τους κινηματικούς περιορισμούς που δημιουργούνται λόγω της φυσικής διασύνδεσης και β) να οδηγήσει σε λύσεις οι οποίες αναπτύσσουν τη συνεργατικότητα καθώς επίσης και την επικοινωνία πληροφοριών μεταξύ των πρακτόρων, έτσι ώστε να επιτευχθεί η προσέγγιση της θέσης-στόχου, χωρίς προηγούμενη εμπειρία - γνώση, και χωρίς τη χρήση επιλύσιμου μοντέλου της προς επίτευξη εργασίας.



Σχήμα 4.2: *Ασαφοποίηση παραμέτρων: γωνία άρθρωσης, στόχος - τελικού στοιχείου δράσης*

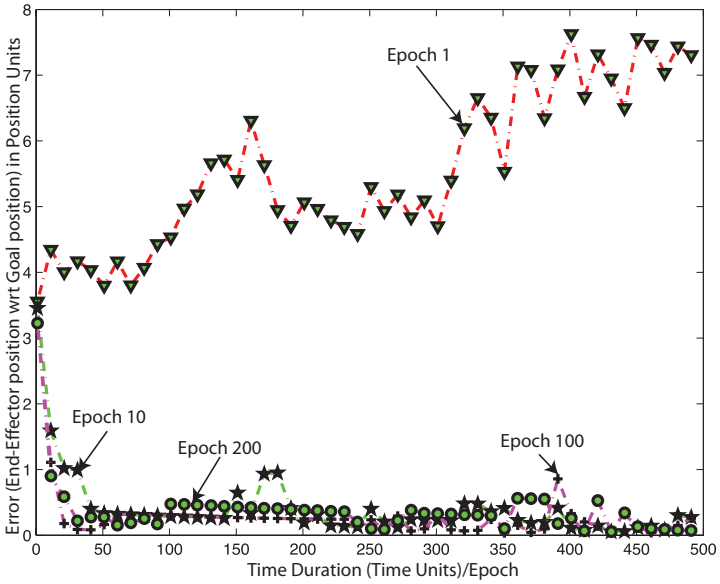
Αρχικά, μεσολαβεί το στάδιο εκπαίδευσης του πολυπρακτορικού συστήματος. Το σύστημα εκπαιδεύεται για ένα σύνολο 200 εποχών. Κάθε εποχή περιλαμβάνει χρονικό ορίζοντα 500 χρονικών μονάδων. Αυτό σημαίνει ότι η ομάδα των τεσσάρων πρακτόρων υπόκειται σε έναν περιορισμό: να επιλύσει το πρόβλημα ανεύρεσης του στόχου σε ένα χρονικό διάστημα 500 χρονικών μονάδων. Σε κάθε περίπτωση επιτυχούς ή αποτυχημένης προσπάθειας προσέγγισης του στόχου, η αποκτηθείσα γνώση αποθηκεύεται έτσι ώστε τελικά να αναμένεται από τους πράκτορες να βελτιώσουν τη συνολική συμπεριφορά τους στις εποχές που έπονται. Σημειώνουμε εδώ ότι αρχικά η συνεργατική συμπεριφορά των πρακτόρων δίνει περισσότερη έμφαση στην εξερεύνηση του χώρου αντιστοίχισης καταστάσεων - δράσεων. Κατά συνέπεια, για να διασφαλιστεί αυτό, οι παράμετροι στην εξίσωση $T(t) = 1 + T_{max} * e^{-st}$, ορίζονται έτσι ώστε να έχουμε υψηλά επίπεδα εξερεύνησης στο διάστημα των αρχικών 30 χρονικών μονάδων κάθε εποχής (το οποίο σημαίνει ότι ο πράκτορας σε αυτή την αρχική περίοδο ακολουθεί μια στρατηγική επιλογής δράσεων οι οποίες δύνανται να μην έχουν τις υψηλότερες τιμές Q-values), ενώ στο υπόλοιπο χρονικό ορίζοντα της συγκεκριμένης εποχής επιλέγει τις δράσεις εκείνες με τις υψηλότερες αξίες Q. Αυτό επιτυγχάνεται με το ρυθμό βαθμιαίας μείωσης (Decay Factor) να είναι $s = 0.35$, ενώ η μέγιστη 'θερμοκρασία' για $t \in [0 \dots 500]$ ορίζεται ως: $T_{max} = 100$. Επιπλέον, ο συντελεστής μάθησης είναι χαμηλός, δεδομένου ότι στο πείραμα επιλέγουμε μία αργή διαδικασία μάθησης (συγκεκριμένα $\lambda = 0.1$), και ο οποίος εν συνέχεια μειώνεται μετά από κάθε εποχή κατά ένα βήμα ίσο με 1/2050. Οι πράκτορες λοιπόν, μέσω αλληλεπίδρασης με το περιβάλλον, λαμβάνουν ανταποδόσεις (Rewards). Στη συγκεκριμένη προσομοίωση έχουμε $Dist_{min} = 3$, πράγμα που σημαίνει ότι οι πράκτορες δεν λαμβάνουν θετική ανταπόδοση εάν το τελικό στοιχείο δράσης δεν είναι σε απόσταση από το στόχο μικρότερη από εκείνη των 3 μονάδων μήκους.



Σχήμα 4.3: *Μέσο σφάλμα ανά χρονική μονάδα, και ανά εποχή, για συντελεστή βαθμιαίας μείωσης (decay factor)* $s = 0.35$

Το Σχήμα 4.3 συγκεντρώνει τα αποτελέσματα της συγκεκριμένης περιόδου εκπαίδευσης. Το αρχικό αυτό σύνολο των αποτελεσμάτων αποτυπώνει τη συμπεριφορά του πολυπρακτορικού συστήματος σε 4 διακριτές εποχές, καθώς το σφάλμα θέσης σταδιακά (κατά την εξέλιξη των διαφορετικών εποχών μάθησης) συγκλίνει στο μηδέν.

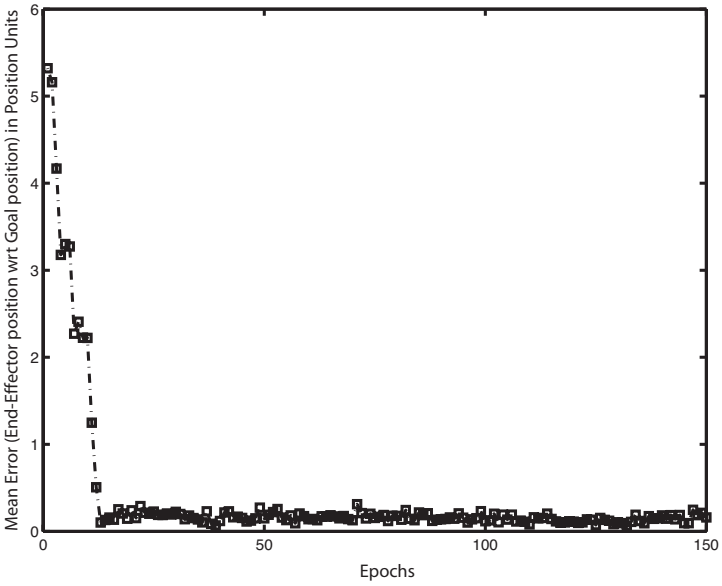
Πιο συγκεκριμένα, κατά τη διάρκεια της εποχής 1, παρατηρούμε ότι οι πράκτορες διερευνούν το χώρο τους. Αυτή η συμπεριφορά οδηγεί σε ένα μεγάλο σφάλμα θέσης (σχετικά πάντα με τη θέση-στόχο), το οποίο όπως φαίνεται και στο σχήμα παραμένει αμείωτο σε όλη τη διάρκεια της συγκεκριμένη εποχής. Αφήνουμε το πείραμα να εξελιχθεί για επιπλέον 10 εποχές. Στη συνέχεια, κατά τη διάρκεια της εποχής 10, βλέπουμε παρόμοια συμπεριφορά με ίσως κάποια σχετική βελτίωση στα μέσα της συγκεκριμένης εποχής, η οποία όμως καταλήγει και πάλι σε σημαντικό σφάλμα. Στις εποχές 100 και 200, όμως, η εξέλιξη της συνεργατικής συμπεριφοράς των πρακτόρων είναι προφανής, καθώς καταφέρνουν να μειώσουν το σφάλμα, επιλέγοντας σταδιακά δράσεις με υψηλές τιμές Q-Values. Στο Σχήμα 4.4, έχουμε τα αποτελέσματα τα οποία προκύπτουν κάνοντας το ίδιο πείραμα με μοναδική τροποποίηση το συντελεστή βαθμιαίας μείωσης (Decay Factor), ο οποίος έχει μεταβληθεί στη νέα τιμή $s = 0.75$. Τα αποτελέσματα δείχνουν ότι ήδη κατά την εποχή 10, το σφάλμα θέσης έχει μειω-



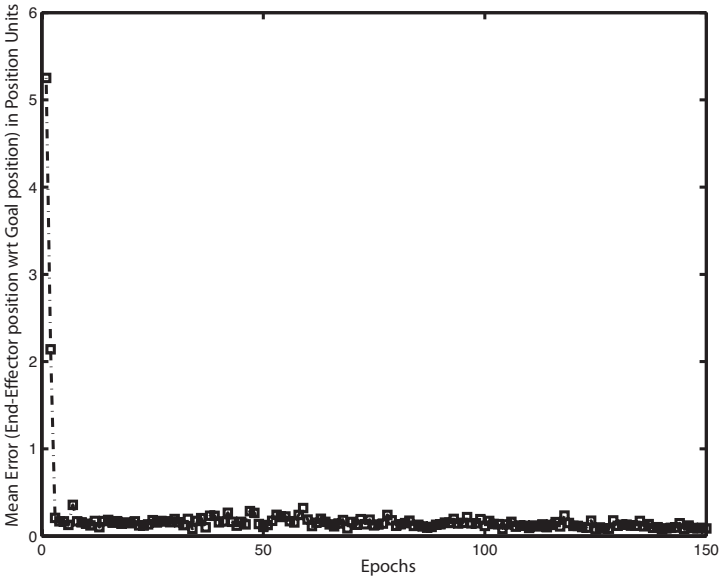
Σχήμα 4.4: *Μέσο σφάλμα ανά χρονική μονάδα, και ανά εποχή, για συντελεστή βαθμιαίας μείωσης (decay factor)* $s = 0.75$

θεί σημαντικά. Στις επόμενες εποχές προφανώς η συμπεριφορά των πρακτόρων και πάλι ισορροπεί σε επιλογή δράσεων με υψηλές τιμές Q-Values . Συγκεντρωτικά, τα αποτελέσματα της εξέλιξης του μέσου σφάλματος θέσης, για τις πρώτες 150 εποχές και για τους δύο προηγούμενους συντελεστές βαθμιαίας μείωσης (Decay Factors), $s = 0.35$ και $s = 0.75$, παρουσιάζονται στο Σχήμα 4.5 και στο Σχήμα 4.6 αντίστοιχα.

Έχοντας μελετήσει λοιπόν τη σύγκλιση σφάλματος της προτεινόμενης προσέγγισης, επόμενο σημαντικό θέμα αποτελεί η αξιολόγηση των λύσεων που προκύπτουν μέσω αυτού του μηχανισμού. Πιο συγκεκριμένα, χρησιμοποιώντας ως σήμα ανταμοιβής για το πολυπρακτορικό σύστημα, την απόσταση από τη θέση-στόχο, το ζήτημα που πρέπει να αξιολογηθεί είναι εάν η πορεία εξερεύνησης την οποία στοχαστικά υιοθέτησε το σύστημα θα συγκλίνει σε μια "βέλτιστη" λύση, επιλύοντας τους σχετικούς πλεονασμούς. Μολονότι η σύγκλιση αυτή ίσως δεν εξασφαλίζεται μέσω μαθηματικής απόδειξης, θεωρούμε ότι είναι εξαιρετικά σημαντικό να αξιολογήσουμε τις λύσεις που παράγει. Μία λύση λοιπόν θεωρείται βέλτιστη υπό την έννοια των ελαχίστων τετραγώνων (Least-Squares, LS), στην περίπτωση εκείνη που αντιστοιχεί σε μία γενικευμένη μετατόπιση κάθε βαθμού ελευθερίας (εν προκειμένω, άρθρωσης) της κινηματικής αλυσίδας, η οποία είναι η ελάχιστη δυνατή (πιο συγκεκριμένα, λύση η οποία απαιτεί την ελάχιστη δυνα-



Σχήμα 4.5: *Μέσο σφάλμα για τις πρώτες 150 εποχές, για συντελεστή βαθμιαίας μείωσης (decay factor)* $s = 0.35$



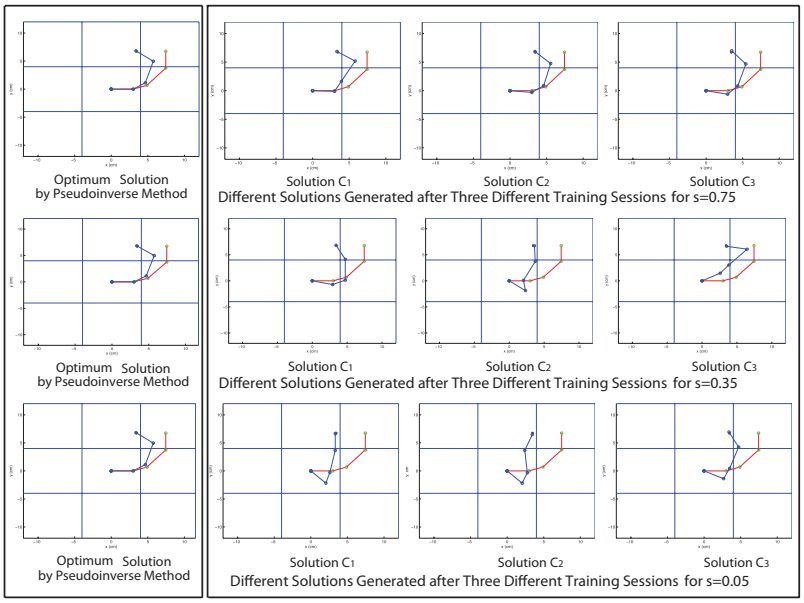
Σχήμα 4.6: *Μεσο σφάλμα για τις πρώτες 150 εποχές, για συντελεστή βαθμιαίας μείωσης (decay factor)* $s = 0.75$

Πίνακας 4.2: Διαφορετικές κινηματικές λύσεις παραγόμενες από την προτεινόμενη πολυπρακτορική αρχιτεκτονική, σε σύγκριση με τη λύση που παράγει η μέθοδος υπολογισμού της ψευδοανάστροφης J^+

	J^+	$s = 0.75$			$s = 0.35$			$s = 0.05$		
		c_1	c_2	c_3	c_1	c_2	c_3	c_1	c_2	c_3
Q_1	0	-2.0	-5.8	-11.8	-13.9	-39.0	29.4	-46.8	-46.9	-27.6
Q_2	24.1	63.9	41.9	57.8	39.1	137.3	22.6	120.8	119.0	93.2
Q_3	53.7	0.1	40.1	28.1	64.7	-33.3	-2.9	5.3	25.7	6.3
Q_4	59.8	84.1	59.8	55.9	26.1	27.2	120.3	10.4	-25.0	43.9

τή προσπάθεια στις αρθρώσεις), η οποία επιτυγχάνει κίνηση στο χώρο εργασίας κατά την διεύθυνση συντομότερης προσέγγισης του στόχου [12] [132]. Για την απόκτηση αυτής της θεωρητικά βέλτιστης διάταξης (με δεδομένη την αρχική διάταξη των αρθρώσεων του συστήματος) επιλύουμε το αντίστροφο κινηματικό πρόβλημα μιας επίπεδης (με πλεονάζοντες βαθμούς ελευθερίας) κινηματικής αλυσίδας, βασιζόμενοι στο υπολογισμό της ψευδοαντίστροφης J^+ της Ιακωβιανής μήτρας. Για την επίλυση του αντιστρόφου κινηματικού προβλήματος μιας κινηματικής αλυσίδας σαν αυτή που προσομοιώνουμε (τεσσάρων βαθμών ελευθερίας), κάνουμε χρήση επαναληπτικής μεθόδου επιλυμένης ταχύτητας, η οποία βασίζεται στον υπολογισμό της ψευδοαντιστρόφου J^+ της Ιακωβιανής μήτρας [55] [109]. Η Ιακωβιανή μήτρα J στη δική μας περίπτωση είναι ένας 3×4 πίνακας ο οποίος έχει για στήλες τα ακόλουθα διανύσματα $\vec{J}_1, \vec{J}_2, \vec{J}_3, \vec{J}_4$. Κάθε $\vec{J}_i$ διάνυσμα ($i = 1 \ldots 4$), δύναται να υπολογιστεί ως το διανυσματικό γινόμενο των εξής δύο στοιχείων: του διανύσματος περιστροφής του συνδέσμου i^{th}, και του διανύσματος που εκφράζει την απόσταση του τελικού στοιχείου δράσης από την αντίστοιχη i^{th} άρθρωση . Εν συνέχεια, μπορούμε να γράψουμε $\Delta\theta = J^+ \cdot \Delta p$, όπου το διάνυσμα $\Delta\theta$ αντιστοιχεί στις μεταβολές των γωνιακών μετατοπίσεων στις αρθρώσεις, οι οποίες οδηγούν το τελικό στοιχείο δράσης να κινηθεί αντίστοιχα κατά Δp, και όπου η ψευδοαντίστροφη Ιακωβιανή μήτρα J^+ ορίζεται ως: $J^T * (J * J^T)^{-1}$ (εφόσον ο πίνακας J είναι πλήρους τάξης - full rank). Κάνοντας λοιπόν χρήση αυτής της επαναληπτικής μεθόδου βρίσκουμε ένα βέλτιστο σύνολο γωνιακών μετατοπίσεων Q_i για $i = 1 \ldots 4$ όπως φαίνεται στο Πίνακα 4.2 (βέλτιστο σύνολο υπό την έννοια least-square (LS), με την ελάχιστη γωνιακή μετατόπιση ανά άρθρωση, κάθε συνδέσμου της κινηματικής αλυσίδας).

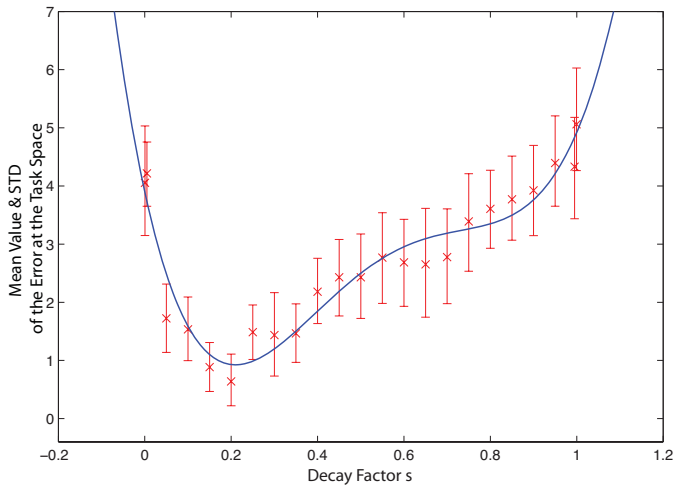
Ενδεικτικές αντίστοιχες λύσεις που προέκυψαν μέσω της προτεινόμενης αρ-



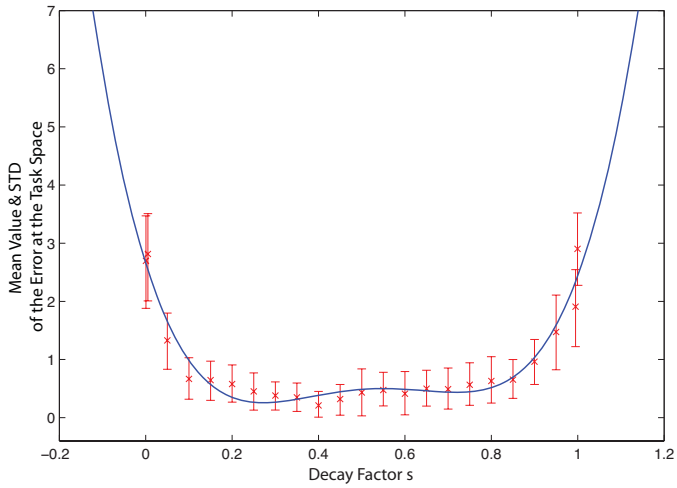
Σχήμα 4.7: *Κινηματικές λύσεις που προκύπτουν για* $s = 0.75$, $s = 0.35$ *και* $s = 0.05$

χιτεκτονικής αποτυπώνονται επίσης στο Πίνακα 4.2, και οι οποίες διαφοροποιούνται ανάλογα με την επιλογή του συντελεστή βαθμιαίας μείωσης (Decay Factor). Στο Σχήμα 4.7 παρουσιάζονται γραφικά τα αποτελέσματα του Πίνακα 4.2, όπου με τη μορφή απλών διαγραμμάτων, αποτυπώνεται το σύνολο των λύσεων που παράγει η προτεινόμενη αρχιτεκτονική. Οι λύσεις που προκύπτουν από την εφαρμογή της προτεινόμενης μεθόδου είναι εν γένει αρκετά ικανοποιητικές, εάν αξιολογηθούν ως προς την ομοιότητά τους με τη θεωρητικά LS-βέλτιστη διάταξη. Κάθε μία από αυτές τις λύσεις προκύπτει μετά από σχετική περίοδο εκπαίδευσης του πολυπρακτορικού συστήματος, η οποία έχει διάρκεια 200 εποχές, και πραγματοποιείται για τρεις διαφορετικούς συντελεστές βαθμιαίας μείωσης. Κατά συνέπεια παρουσιάζουμε ένα σύνολο εννέα λύσεων οι οποίες και περιλαμβάνονται στο σχετικό πίνακα.

Με σκοπό να αναλύσουμε περαιτέρω τις λύσεις που προκύπτουν πραγματοποιούμε μια στατιστική ανάλυση των σχετικών αποτελεσμάτων (σφάλμα σε σχέση με το ελάχιστο τετράγωνο της βέλτιστης διάταξης), για ένα σύνολο συντελεστών βαθμιαίας μείωσης (Decay Factors) $s \in [0.005 \ldots 0.9995]$. Οι μέσες τιμές (mean values) καθώς και οι τυπικές αποκλίσεις (standard deviations) του σφάλματος απεικονίζονται στα Σχήματα 4.8 και 4.9 σε δύο ενδιάμεσες φάσεις της διαδικασίας μάθησης, πιο συγκεκριμένα, μετά από 10 και 100 εποχές μάθησης αντίστοιχα. Τα αποτελέσματα επιδεικνύουν ότι για ένα ευρύ φάσμα τιμών

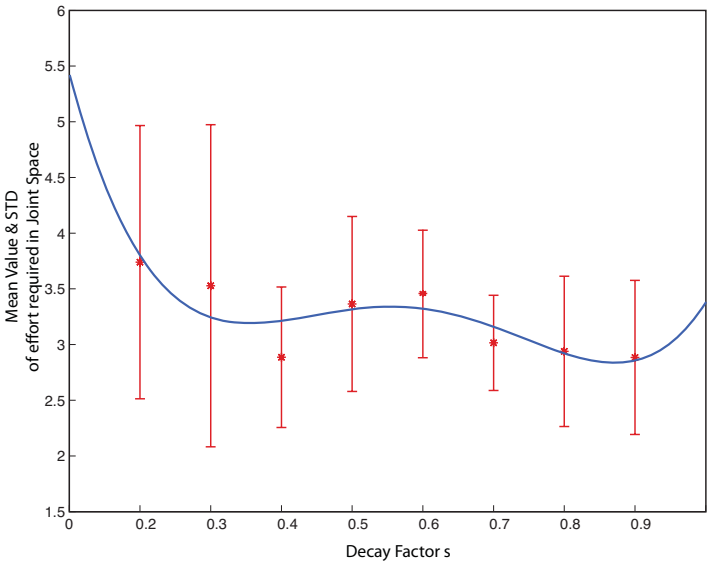


Σχήμα 4.8: *Μέση τιμή και τυπική απόκλιση του σφάλματος (σε σχέση με τη βέλτιστη διάταξη, υπό την έννοια των ελαχίστων τετραγώνων) για συγκεκριμένο εύρος συντελεστών βαθμιαίας μείωσης, μετά από* 10 *εποχές μάθησης*



Σχήμα 4.9: *Μέση τιμή και τυπική απόκλιση του σφάλματος (σε σχέση με τη βέλτιστη διάταξη, υπό την έννοια των ελαχίστων τετραγώνων) για συγκεκριμένο εύρος συντελεστών βαθμιαίας μείωσης, μετά από* 100 *εποχές μάθησης*

s, οι λύσεις οι οποίες στοχαστικά προέκυψαν από το πολυπρακτορικό σύστημα είναι πολύ κοντά στις βέλτιστες (near-optimal) υπό την έννοια των ελάχιστων τετραγώνων. Επιπλέον, στο Σχήμα 4.10 απεικονίζονται οι μέσες τιμές και η τυπική απόκλιση της προσπάθειας (το συνολικό άθροισμα των απολύτων τιμών των γενικευμένων μετατοπίσεων κάθε άρθρωσης, που πραγματοποιείται σε κά-

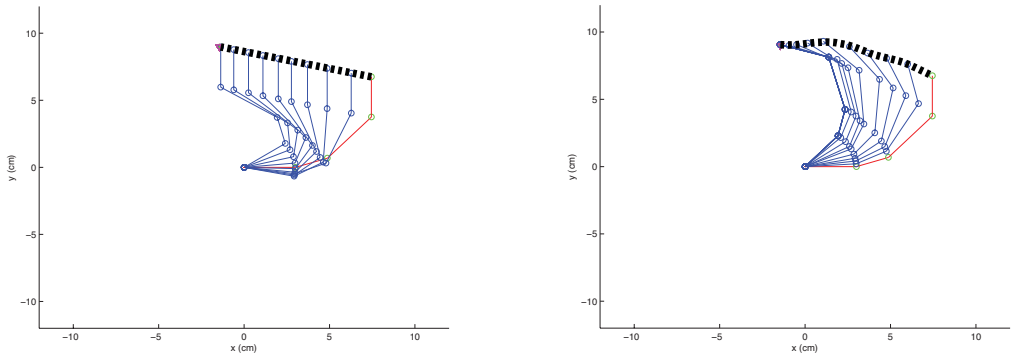


Σχήμα 4.10: *Μέση τιμή και τυπική απόκλιση της προσπάθειας (το συνολικό άθροισμα των απολύτων τιμών των γενικευμένων μετατοπίσεων κάθε άρθρωσης, που πραγματοποιείται σε κάθε χρονικό βήμα t, από όλους του πράκτορες με σκοπό την προσέγγιση της θέσης-στόχου) για συγκεκριμένο εύρος συντελεστών βαθμιαίας μείωσης, μετά από* 100 *εποχές μάθησης*

θε χρονικό βήμα t, από όλους του πράκτορες με σκοπό την προσέγγιση της θέσης-στόχου) μετά από 100 εποχές μάθησης, για ένα ευρύ φάσμα τιμών s.

Τα παραπάνω αποτελέσματα είναι εξαιρετικά ενθαρρυντικά, συγκρινόμενα με τα αντίστοιχα που προκύπτουν από τη θεωρητικά βέλτιστη (LS-optimal) λύση όπου προκύπτει μια μέση τιμή αθροιστικού σφάλματος ίση με 4.48 (rad) περίπου. Το παραπάνω εύρημα αξιολογείται επίσης και μέσω της παρατήρησης της πορείας των κινηματικών αλυσίδων στις δύο περιπτώσεις στο Σχήμα 4.11. Στο σχετικό Σχήμα, απεικονίζεται μία πορεία της κινηματικής αλυσίδας προς τη θέση-στόχο, μέσω της προτεινόμενης πολυπρακτορικής αρχιτεκτονικής (μετά από 100 εποχές μάθησης και συντελεστή $s = 0.5$). Η σχετική πορεία η οποία προκύπτει μέσω της πολυπρακτορικής αρχιτεκτονικής ελέγχου είναι απολύτως φυσική, συγκρινόμενη με την κίνηση που υπολογίζεται αναλυτικά μέσω της ψευδοαντιστρόφου της Ιακωβιανής μήτρας, όπως αποτυπώνεται στο ίδιο σχήμα. Αυτό είναι πολύ σημαντικό εύρημα το οποίο υποστηρίζει το συμπέρασμα περί της δυνατότητας αποτελεσματικής εφαρμογής της προτεινόμενης μεθοδολογίας σε προβλήματα επιδέξιου ρομποτικού χειρισμού.

Κατά συνέπεια, μολονότι δεν ισχυριζόμαστε ότι το προτεινόμενο πολυπρακτορικό σύστημα μάθησης εγγυάται βέλτιστες λύσεις στο συγκεκριμένο πρόβλημα, δείξαμε ότι το σύστημα στοχαστικά συγκλίνει σε ένα εύρος λύσεων



(α) Θεωρητικά LS-βέλτιστη μέσω ψευδοαντιστρόφου (β) Λύση μέσω Πολυπρακτορικής Προσέγγισης

Σχήμα 4.11: *Η κίνηση της κινηματικής αλυσίδας κατά την εργασία προσέγγισης της θέσης-στόχου. (α) Θεωρητικά βέλτιστη λύση LS-optimal (pseudo-inverse). (β) Παράδειγμα λύσης πολυπρακτορικής προσέγγισης, χωρίς μοντέλο, η οποία προέκυψε μετά από* 100 *εποχές εκπαίδευσης (με συντελεστή* $s = 0.5$*).*

οι οποίες όχι μόνο επιλύουν αποτελεσματικά τους κινηματικούς πλεονασμούς, αλλά επιπλέον, όπως ανέδειξαν οι πειραματικές δοκιμές, προκύπτουν συγκρίσιμες με τις θεωρητικά βέλτιστες λύσεις για ένα ευρύ φάσμα συντελεστών βαθμιαίας μείωσης. Συνεπώς, η προτεινόμενη μέθοδος ενισχυτικής ρομποτικής μάθησης η οποία υιοθετήθηκε στην παρούσα διατριβή στατιστικά και προοδευτικά ανταμείβει δράσεις και συγκλίνει προς τις συμπεριφορές εκείνες οι οποίες οδηγούν ταχύτερα (πιο άμεσα) σε ελαχιστοποίηση του σφάλματος απόστασης, δημιουργώντας αποτελεσματικές αντιστοιχίσεις κατάστασης - δράσης, οι οποίες ανταμοίβοντας κινήσεις τύπου ελάχιστης προσπάθειας (Minimum-Effort) επιτυγχάνουν το στόχο και επιλύουν τους πλεονασμούς του συστήματος. Αυτό αποτελεί μια εξαιρετικά σημαντική πειραματική παρατήρηση η οποία υποστηρίζει την αρχική μας υπόθεση περί δυνατότητας εφαρμογής της προτεινόμενης μεθοδολογίας στο σχετικό πεδίο του επιδέξιου ρομποτικού χειρισμού.

Στο πιθανό ερώτημα, πώς καταφέρνει το πολυπρακτορικό σύστημα να επιλύει τους πλεονασμούς (redundancy) με μόνο τμηματική (τοπική) πληροφορία σχετική με το στόχο ενσωματωμένη σε κάθε πράκτορα, καθώς επίσης και το πώς αυτό πραγματοποιείται με τρόπο βέλτιστο, τα πειραματικά δεδομένα τα οποία προέκυψαν στο πλαίσιο της παρούσας εργασίας υποστηρίζουν τα εξής: 1) η αναπαράσταση της κατάστασης του συστήματος μεταξύ των επιμέρους πρακτόρων που το συνθέτουν, επαρκεί για την αναζήτηση και εύρεση λύσεων με έναν κατανεμημένο τρόπο, και 2) οι λύσεις που προκύπτουν ήταν αναμενόμενο να είναι κοντά σε βέλτιστες (υπό την έννοια των ελάχιστων τετραγώνων) λύσεις, δεδομένου ότι το σύστημα, κατά τη διάρκεια της εξερεύνησης, προσπα-

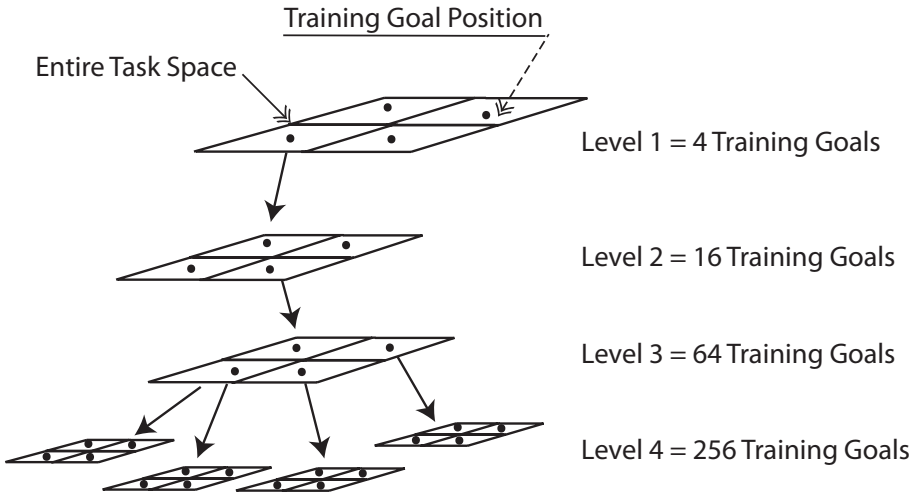
θεί να βρεί λύσεις οι οποίες συγκλίνουν στη θέση-στόχο γρήγορα, διότι αυτό αντιστοιχεί στο ελάχιστο αρνητικό κόστος που εισπράττει το πολυπρακτορικό σύστημα ως ανταμοιβή στο πλαίσιο της ενισχυτικής μάθησης που πραγματοποιεί.

Έχοντας ολοκληρώσει τη διαδικασία εκπαίδευσης, το επόμενο βήμα είναι να εξετάσουμε με ποιον τρόπο οι πράκτορες κάνουν χρήση των επιπέδων γνώσης που έχουν συλλέξει, και πώς χωρίς πρόσθετη εκπαίδευση είναι σε θέση να προσεγγίζουν νέους στόχους που παρουσιάζουν συνάφεια με εκείνους που έχουν εκπαιδευτεί.

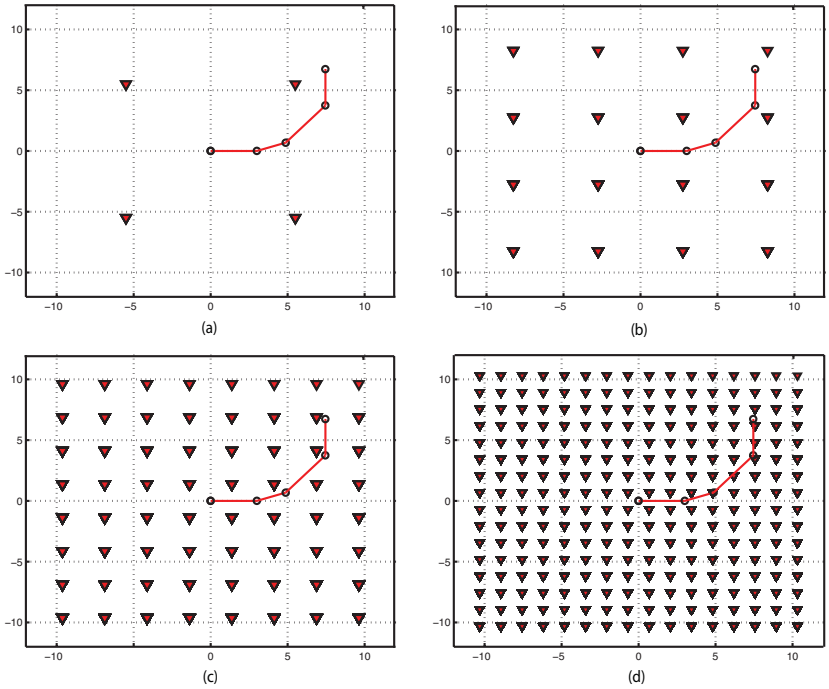
4.3.1 Εκπαίδευση Στόχου Πολλαπλής Ανάλυσης και Γενίκευση Γνώσης

Στη συνέχεια θα μελετήσουμε τις ιδιότητες γενίκευσης γνώσης που παρουσιάζει το προτεινόμενο πολυπρακτορικό σύστημα. Για το σκοπό αυτό δημιουργούνται πολλαπλά επίπεδα δεδομένων, τα οποία θα αποτελέσουν αντικείμενο μάθησης για το σύστημά μας (multiple layers of training data). Κάθε επίπεδο αντιπροσωπεύει μία διαφορετική ανάλυση (resolution)του χώρου εργασίας. Σκοπός των πειραματικών δοκιμών που αναλύονται στην παράγραφο αυτή είναι να διερευνηθεί η ικανότητα την οποία επιδεικνύει το προτεινόμενο σύστημα, ως προς την επέκταση/γενίκευση της γνώσης που αποκτήθηκε κατά το χρονικό διάστημα της εκπαίδευσής του. Ουσιαστικά, το κατά πόσο το πολυπρακτορικό σύστημα μπορεί να προσεγγίσει νέες θέσεις στόχους για τις οποίες δεν έχει προηγηθεί σχετική εκπαίδευση, στηριζόμενο αποκλειστικά στη χρήση, και κατά συνέπεια, γενίκευση, προγενέστερης γνώσης.

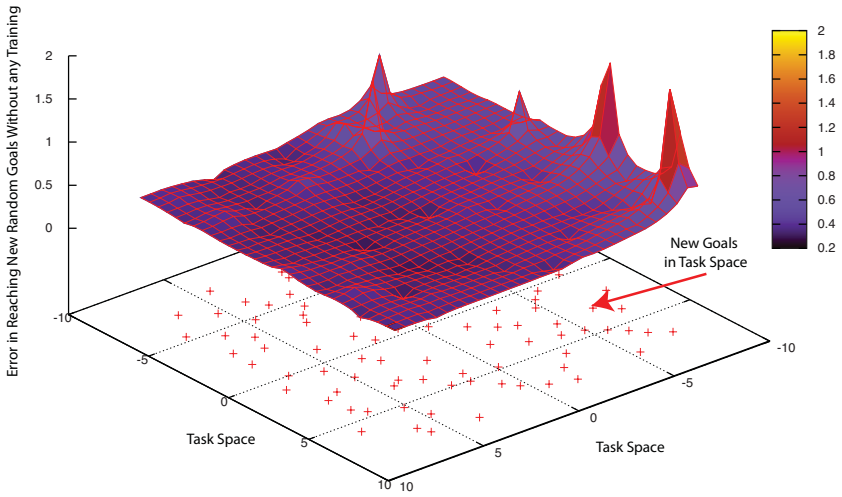
Αρχικά, ο χώρος εργασίας υποδιαιρείται με τρόπο αναδρομικό (recursive subdivision) σε επίπεδα αύξουσας ανάλυσης (levels of increasing resolution), δημιουργώντας μία δομή τύπου quad-tree, όπως αυτή που απεικονίζεται στο σχετικό Σχήμα 4.12. Ξεκινώντας αρχικά από το επίπεδο: *Level*-1, και το οποίο περιέχει τέσσερις κόμβους (nodes) ή αλλιώς τεταρτημόρια (quadrants), κινούμαστε προοδευτικά σε υψηλότερα επίπεδα ανάλυσης, φτάνοντας τελικά στο επίπεδο *Level*-4 το οποίο περιλαμβάνει 256 κόμβους. Κάθε κόμβος αντιστοιχεί σε μία θέση-στόχο, για την οποία το πολυπρακτορικό σύστημα θα εκπαιδευτεί. Συνεπώς, για κάθε ένα επίπεδο ανάλυσης που δημιουργούμε, το πολυπρακτορικό σύστημα κάθε φορά εκπαιδεύεται σε ένα διαφορετικό σύνολο θέσεων-στόχων. Το σύνολο αυτό προοδευτικά μεγαλώνει ως προς τις διαστάσεις του, ανάλογα με την ανάλυση που αντιστοιχεί στο συγκεκριμένο επίπεδο



Σχήμα 4.12: *Δημιουργία πολλαπλών επιπέδων ανάλυσης για εκπαίδευση του συστήματος*



Σχήμα 4.13: *Πολλαπλά επίπεδα ανάλυσης και αντίστοιχες θέσεις-στόχων*

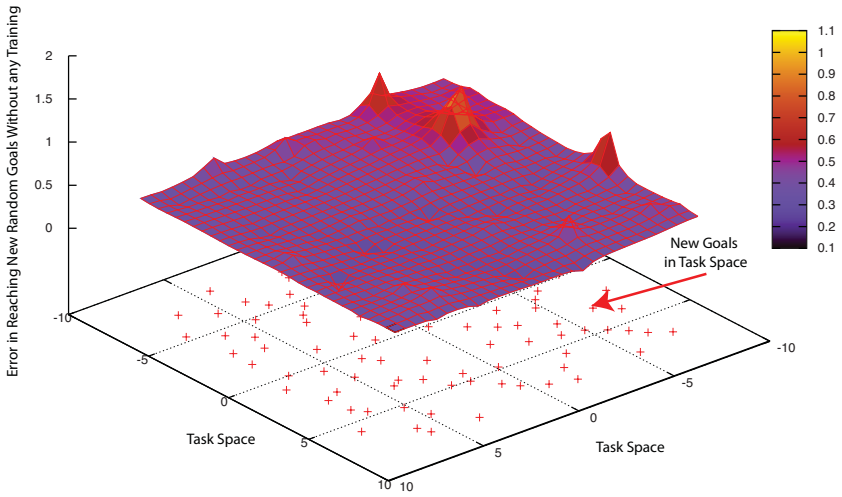


Σχήμα 4.14: *Σφάλμα προσέγγισης στόχων σε ολόκληρο το χώρο εργασίας, με ανάλυση 4 σημείων εκπαίδευσης*

(με άλλα λόγια, στο *Level*-1, οι τέσσερις κόμβοι αντιστοιχούν σε τέσσερις θέσεις στόχους που θα πρέπει να εκπαιδευτεί το σύστημα, για το επίπεδο *Level*-2, 16 θέσεις κτλ.)

Στα πλαίσια του συγκεκριμένου πειράματος που παρουσιάζουμε εδώ, διερευνούμε ένα εύρος αναλύσεων που ξεκινά από 4 έως 4^4 θέσεις στόχους, όπως αυτές απεικονίζονται στο αντίστοιχο Σχήμα 4.13. Για κάθε ένα από τα διαφορετικά επίπεδα ανάλυσης, μετά την ολοκλήρωση της διαδικασίας εκπαίδευσης για τις θέσεις που περιλαμβάνει το κάθε επίπεδο, παράγεται ένα νέο σύνολο το οποίο αποτελείται από 100 θέσεις-στόχους τυχαία κατανεμημένες κάνοντας χρήση των ακολουθιών Halton [94] έτσι ώστε να έχουμε ομοιογενή κατανομή των νέων στόχων. Στόχος μας είναι να διερευνήσουμε την ακρίβεια με την οποία, σε κάθε ένα διαφορετικό επίπεδο ανάλυσης, το πολυπρακτορικό σύστημα καταφέρνει να προσεγγίσει τις νέες θέσεις-στόχους (για τις οποίες δεν έχει εκπαιδευτεί) και οι οποίες έχουν προκύψει τυχαία, δίχως να πραγματοποιήσει κάποιο πρόσθετο στάδιο εκπαίδευσης.

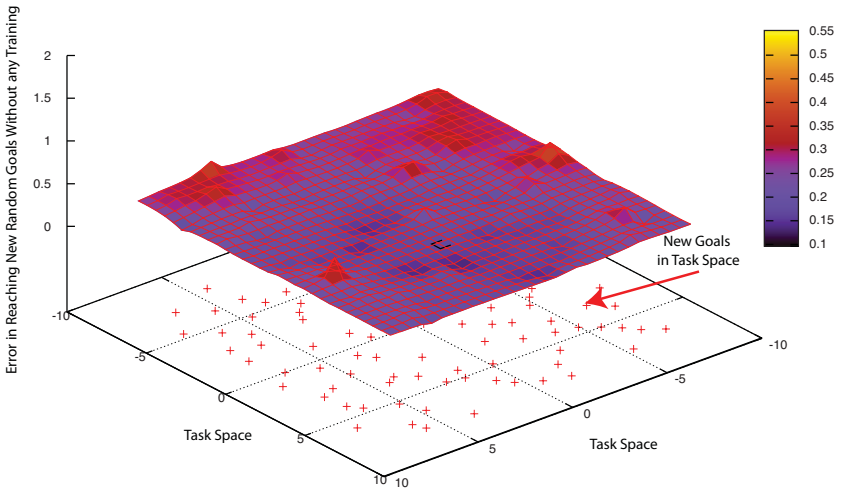
Η πειραματική διαδικασία περιλαμβάνει τα ακόλουθα στάδια: 1) Επιλέγουμε το επιθυμητό επίπεδο ανάλυσης, 2) Πραγματοποιούμε off-line εκπαίδευση στις θέσεις στόχους του συγκεκριμένου επιπέδου, 3) και στη συνέχεια, μετά την ολοκλήρωση της φάσης εκπαίδευσης, πραγματοποιούμε στο σύστημα διαδοχικά αιτήματα, on-line, να προσεγγίσει και τους 100 τυχαίους, νέους στόχους. Τα αποτελέσματα αυτής της διαδικασίας τα οποία συγκεντρώθηκαν και



Σχήμα 4.15: *Σφάλμα προσέγγισης στόχων σε ολόκληρο το χώρο εργασίας, με ανάλυση 16 σημείων εκπαίδευσης*

στα τρία διαφορετικά επίπεδα ανάλυσης, απεικονίζονται στα Σχήματα 4.14 έως 4.16. Παρατηρώντας τις σχετικές γραφικές παραστάσεις που προκύπτουν, είναι προφανές ότι ακόμα και στο επίπεδο με τη χαμηλότερη ανάλυση (το οποίο ουσιαστικά περιλαμβάνει μόνο 4 σημεία προς εκπαίδευση για ολόκληρο το χώρο εργασίας), το προτεινόμενο πολυπρακτορικό σύστημα καταφέρνει με μεγάλη επιτυχία να φτάσει σχεδόν όλους τους τυχαία παραγόμενους στόχους που του ζητήθηκαν. Είναι επίσης προφανές από τα αποτελέσματα τα οποία συγκεντρώθηκαν ότι η κατάσταση βελτιώνεται καθώς η ανάλυση των προς εκπαίδευση στόχων αυξάνεται από 4 σε 16 και τέλος σε 256.

Στη συνέχεια, το Σχήμα 4.17 απεικονίζει τη στατιστική ανάλυση των αποτελεσμάτων γενίκευσης γνώσης, για το σύνολο των διαφορετικών επιπέδων ανάλυσης. Παρατηρούμε ότι εκπαιδεύοντας το πολυπρακτορικό σύστημα μόνο με 4 θέσεις-στόχους, στο πρώτο επίπεδο ανάλυσης (*Level*-1), η μέση τιμή του σφάλματος είναι σημαντικά χαμηλή ($mean = 0.3814$), ενώ παρατηρούμε σημαντική τυπική απόκλιση ($std = 0.5596$). Σε κάθε περίπτωση, το γεγονός είναι ότι οι περισσότερες, τυχαίες, νέες θέσεις προσεγγίστηκαν με επιτυχία ακόμα και με το χαμηλότερο επίπεδο ανάλυσης ως υπόβαθρο αποθηκευμένης γνώσης για το σύστημα. Η μέση τιμή του σφάλματος φθίνει ομαλά και φτάνει, για το επίπεδο ανάλυσης των 64 σημείων (*Level*-3), στη τιμή $mean = 0.2366$, ενώ η αντίστοιχη τυπική απόκλιση μειώνεται στη τιμή $std = 0.2036$. Τέλος, παρατηρούμε ότι, στο συγκεκριμένο πείραμα, αυξάνοντας την ανάλυση από το επίπεδο ανάλυσης

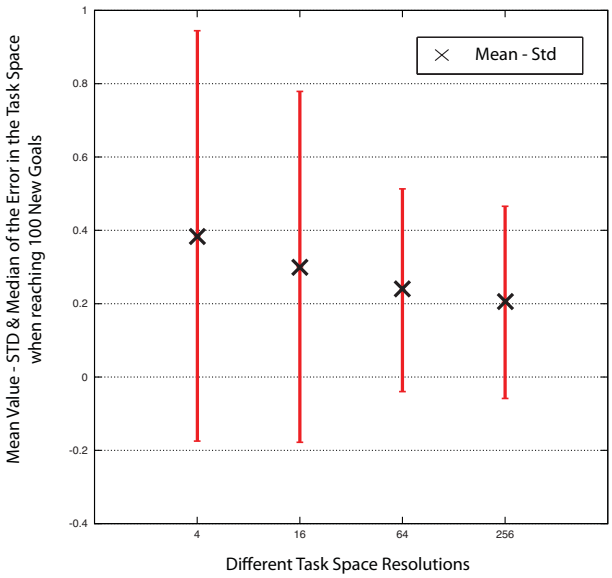


Σχήμα 4.16: *Σφάλμα προσέγγισης στόχων σε ολόκληρο το χώρο εργασίας, με ανάλυση 256 σημείων εκπαίδευσης*

Πίνακας 4.3: Χρόνος εκπαίδευσης σε συνάρτηση με την ανάλυση του χώρου εργασίας

Αρ. Σημείων Εκπαίδευσης	Χρόνος Εκπαίδευσης (λεπτά)
4 σημεία	54.9 (0.92 ώρες)
16 σημεία	334.2 (5.57 ώρες)
64 σημεία	667.9 (11.13 ώρες)
256 σημεία	5843.5 (97.39 ώρες)

Level-3 στο επίπεδο *Level*-4 (δηλαδή από τα 64 στα 256 σημεία) δεν βελτιώνει σημαντικά την ακρίβεια του συστήματος ως προς τη γενίκευση της γνώσης. Τα πειραματικά αυτά ευρήματα ξεκάθαρα υπογραμμίζουν την ουσιαστική δυνατότητα γενίκευσης την οποία επιδεικνύει η προτεινόμενη πολυπρακτορική αρχιτεκτονική. Ο πίνακας 4.3 παρουσιάζει τα υπολογιστικά κόστη, συγκεκριμένα το χρόνο εκτέλεσης (run time , σε ένα φορητό υπολογιστή, με έναν επεξεργαστή και συχνότητα ρολογιού στα 2.6 GHz) ο οποίος απαιτείται για την ολοκλήρωση της off-line εκπαίδευσης του πολυπρακτορικού συστήματος για τα διαφορετικά

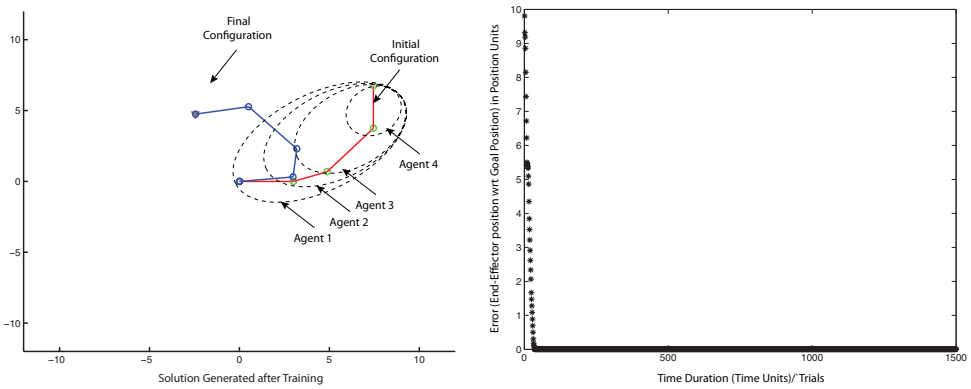


Σχήμα 4.17: *Μέση τιμή και τυπική απόκλιση σφάλματος προσέγγισης στόχου στο χώρο κατάστασης μετά από εκπαίδευση σε επίπεδα διαφορετικών αναλύσεων*

επίπεδα ανάλυσης. Μπορούμε να δούμε ότι το υπολογιστικό κόστος εκπαίδευσης αυξάνεται γραμμικά σε σχέση με τον αριθμό των σημείων που συνθέτουν το σύνολο εκπαίδευσης (δίνοντας έναν μέσο όρο στο χρόνο προσομοίωσης για κάθε σημείο εκπαίδευσης περίπου δεκαπέντε λεπτών). Σε κάθε περίπτωση, αυτό δεν αποτελεί ζήτημα δεδομένου ότι η διαδικασία εκπαίδευσης δε πραγματοποιείται σε πραγματικό χρόνο, αλλά είναι διεργασία που πραγματοποιείται off-line, και η οποία σαφώς μπορεί να επιταχυνθεί κάνοντας χρήση παράλληλων υπολογιστικών αρχιτεκτονικών. Πρέπει να σημειωθεί σε αυτό το σημείο ότι από τα αποτελέσματα τα οποία απεικονίζονται στο Σχήμα 4.17, προκύπτει ότι επιλέγοντας μία ανάλυση των επιπέδων 2 ή 3 (16 ή 64 σημεία εκπαίδευσης, σε αυτήν την περίπτωση) έχουμε ικανοποιητική κάλυψη ολόκληρου του χώρου εργασίας του ρομποτικού συστήματος.

4.3.2 Ευρωστία ως προς Αλλαγές στην Κινηματική Τοπολογία

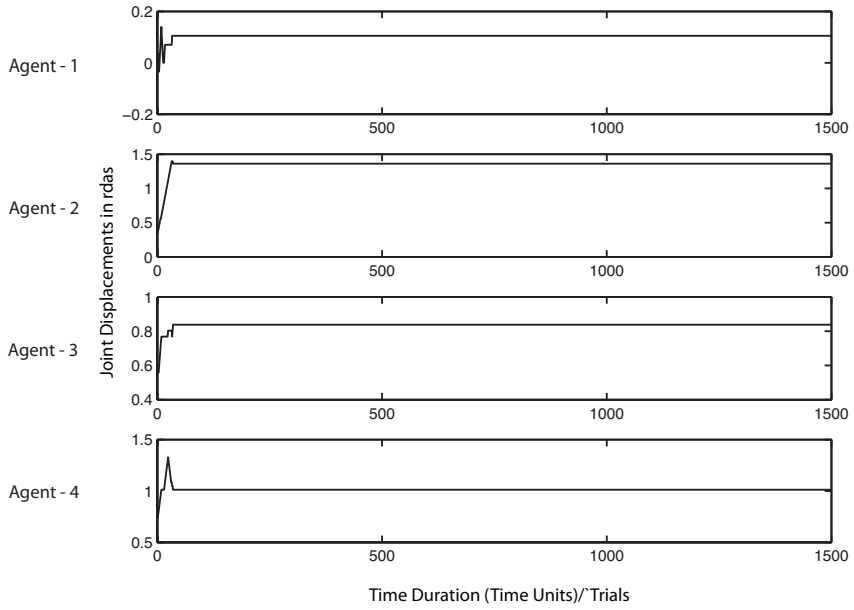
Το βασικό πλεονέκτημα της στρατηγικής κατανεμημένου πολυπρακτορικού ελέγχου (σε σύγκριση με μία τυπική μονοπρακτορική αρχιτεκτονική, η οποία λειτουργεί βάσει μοντέλου), είναι τα ιδιαίτερα χαρακτηριστικά ευρωστίας που επιδεικνύει. Σε αυτήν την παράγραφο, θα αξιολογήσουμε την ικανότητα της



Σχήμα 4.18: *Κινηματική λύση η οποία προέκυψε από το πολυπρακτορικό σύστημα και εξέλιξη του σφάλματος θέσης συναρτήσει του χρόνου (πλήρως λειτουργικό σύστημα - όλοι οι πράκτορες είναι ενεργοί).*

προτεινόμενης αρχιτεκτονικής ως προς τη διαχείριση μερικής / τμηματικής αστοχίας στο σύστημα, η οποία δύναται να προκληθεί σε πράκτορες (βαθμούς ελευθερίας) που συνθέτουν το σύστημα, καθώς και τη σχετική προσαρμοστικότητα που αναπτύσσει στις απρόβλεπτες μεταβολές της κινηματικής τοπολογίας. Με άλλα λόγια, στις περιπτώσεις εκείνες όπου κάποιοι πράκτορες αποτυγχάνουν κατά τη διάρκεια λειτουργίας του συστήματος, θέλουμε να διερευνήσουμε τον μηχανισμό με τον οποίο οι υπόλοιποι ενεργοί πράκτορες επιτυγχάνουν δίχως επανεκπαίδευση να συνεργαστούν έτσι ώστε να δώσουν μία ουσιαστική λύση στο πρόβλημα που προέκυψε και να ολοκληρώσουν τη σχετική εργασία.

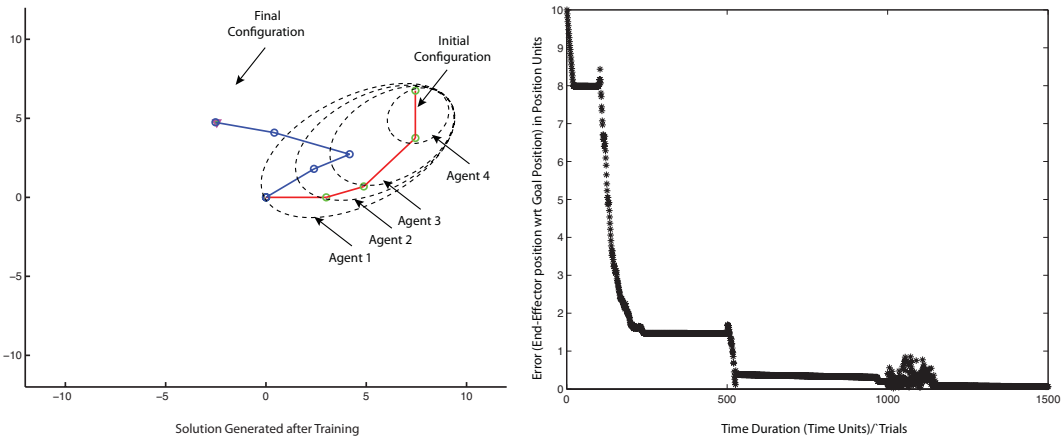
Για να αξιολογήσουμε τις ιδιότητες ευρωστίας της προτεινόμενης πολυπρακτορικής αρχιτεκτονικής, πραγματοποιούμε μία σειρά από αριθμητικές προσομοιώσεις. Αρχικά, το πολυπρακτορικό σύστημα εκπαιδεύεται στο να προσεγγίζει τη θέση-στόχο όπως ήδη περιγράψαμε σε προηγούμενο τμήμα αυτής της εργασίας. Μετά την ολοκλήρωση της εκπαίδευσης, το πλήρως λειτουργικό πολυπρακτορικό σύστημα είναι σε θέση να βρεί μία λύση και να μειώσει το σφάλμα θέσης του τελικού στοιχείου δράσης ως προς τη θέση-στόχο, όπως αυτό απεικονίζεται στο Σχήμα 4.18. Στο Σχήμα 4.19 απεικονίζονται οι δράσεις (joint displacements) όλων ανεξάρτητων πρακτόρων που συνθέτουν το σύστημα. Η επιτυχημένη συνεργασία των πρακτόρων οδήγησε σε κατάλληλες επιλογές δράσεων και επομένως στην ολοκληρωμένη πραγματοποίηση της εργασίας του πολυπρακτορικού ρομποτικού συστήματος. Να σημειωθεί εδώ ότι το σύστημα έχει τις ίδιες παραμέτρους μάθησης όπως και πριν ενώ ο αριθμός χρονικών μονάδων σε μία



Σχήμα 4.19: *Δράσεις (γωνιακές μετατοπίσεις) για όλους τους πράκτορες. Όλοι οι πράκτορες είναι ενεργοί και συνεργάζονται για την προσέγγιση της θέσης - στόχου.*

προσπάθεια είναι 1500 (time units).

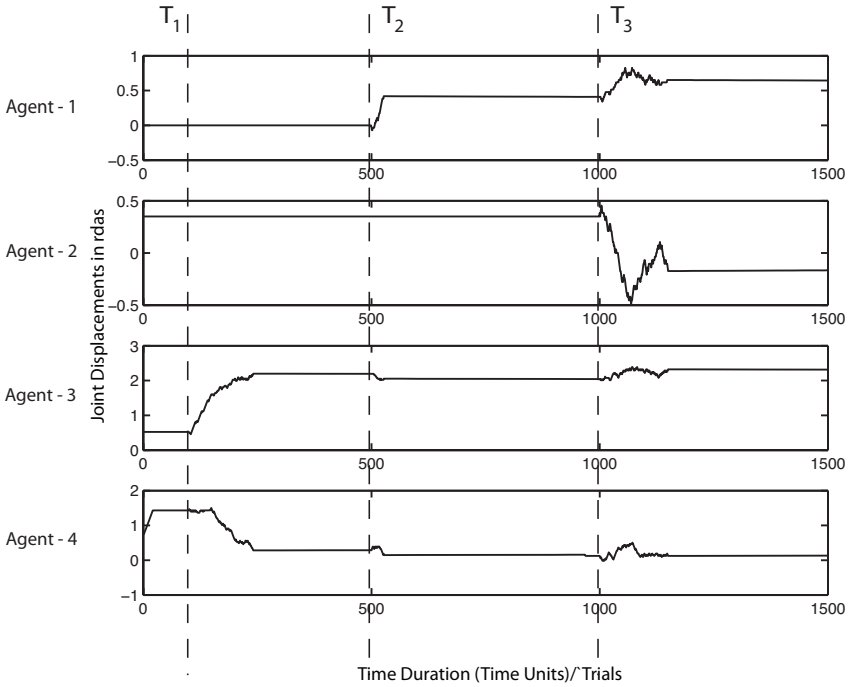
Στη συνέχεια, προσομοιώνεται ένας αριθμός περιπτώσεων όπου κάποιοι πράκτορες αποτυγχάνουν, με στόχο να αξιολογηθεί η ευρωστία που μπορεί να επιδείξει το σύστημα, ως προς την εμφάνιση τέτοιων απρόβλεπτων συμβάντων. Αρχικά, προσομοιώνουμε μία τμηματική αστοχία (failure) του συστήματος, από την οποία εν συνεχεία, προοδευτικά, υποθέτουμε ότι το σύστημα ανακάμπτει. Υποθέτουμε ότι, αρχικά, τη χρονική στιγμή $T = 0$ τρεις από τους πράκτορες που αποτελούν το σύστημα (συγκεκριμένα, οι πράκτορες 1, 2 και 3) είναι μη λειτουργικοί και μόνο ο πράκτορας 4 ανταποκρίνεται. Αυτή η περίπτωση αντιστοιχεί σε αποτυχία λήψης και εκτέλεσης οποιασδήποτε δράσης για τους συγκεκριμένους πράκτορες. Οι πράκτορες 1 εως 3 παραμένουν εγκλωβισμένοι στην αρχική τους γωνιακή θέση. Εν συνεχεία, τη χρονική στιγμή ($T_1 = 100$) ο πράκτορας 3 αρχίζει να ανταποκρίνεται, ενώ αργότερα, τη χρονική στιγμή ($T_2 = 500$) και ($T_3 = 1000$), οι πράκτορες 2 και 3, αντίστοιχα, ανακάμπτουν από την αποτυχία τους. Στόχος μας είναι να διερευνήσουμε τον τρόπο με τον οποίο, δίχως εξωτερική παρέμβαση (όπως επαναμοντελοποίηση,



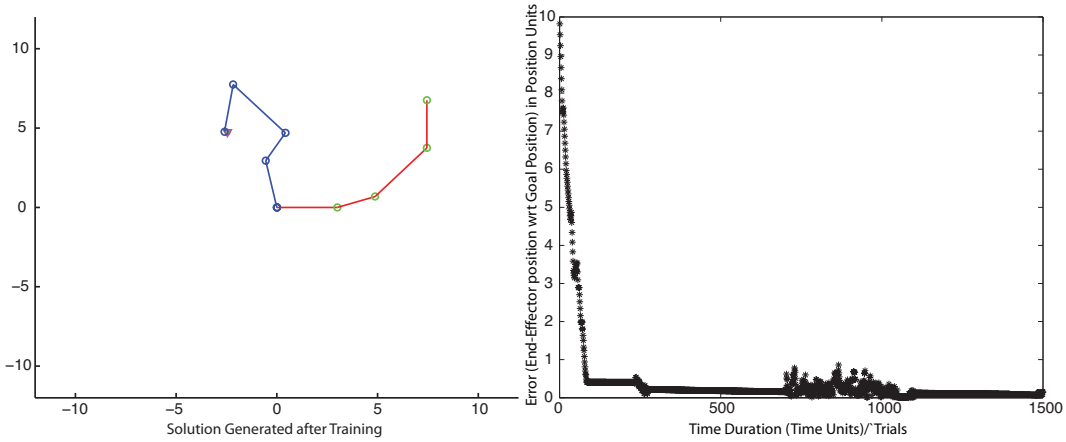
Σχήμα 4.20: *Κινηματική λύση η οποία προέκυψε από το πολυπρακτορικό σύστημα και εξέλιξη του σφάλματος θέσης συναρτήσει του χρόνου, μετά τις αποτυχίες και την προοδευτική ανάκαμψη των πρακτόρων 1 έως 3.*

μεταβολή λειτουργικού κώδικα, κλπ.), το πολυπρακτορικό σύστημα καταφέρνει να διαχειριστεί και ταυτόχρονα να ανταποκριθεί στην συγκεριμένη κατάσταση η οποία έχει προκύψει. Τα αποτελέσματα παρουσιάζουν εξαιρετικό ενδιαφέρον, ενώ επιδεικνύουν ξεκάθαρα ότι το πολυπρακτορικό σύστημα καταφέρνει, στο χρόνο εκτέλεσης (at run time), να βρει νέα λύση προσαρμοζόμενο στις απρόβλεπτες αλλαγές της κινηματικής τοπολογίας, οι οποίες προέκυψαν λόγω των τυχαίων αποτυχιών συγκεκριμένων πρακτόρων κατά τη διάρκεια λειτουργίας του συστήματος. Το Σχήμα 4.20 απεικονίζει τη νέα κινηματική λύση η οποία παράγεται από το πολυπρακτορικό σύστημα καθώς και την εξέλιξη του σφάλματος.Το Σχήμα 4.21 επιδεικνύει τις δράσεις που επιλέγει ο κάθε πράκτορας και πώς αυτές διαφοροποιούνται / προσαρμόζονται, για να ανταποκριθούν στις διαταραχές που προκαλούνται στο σύστημα από τα περιστατικά αστοχίας που καταγράφονται.

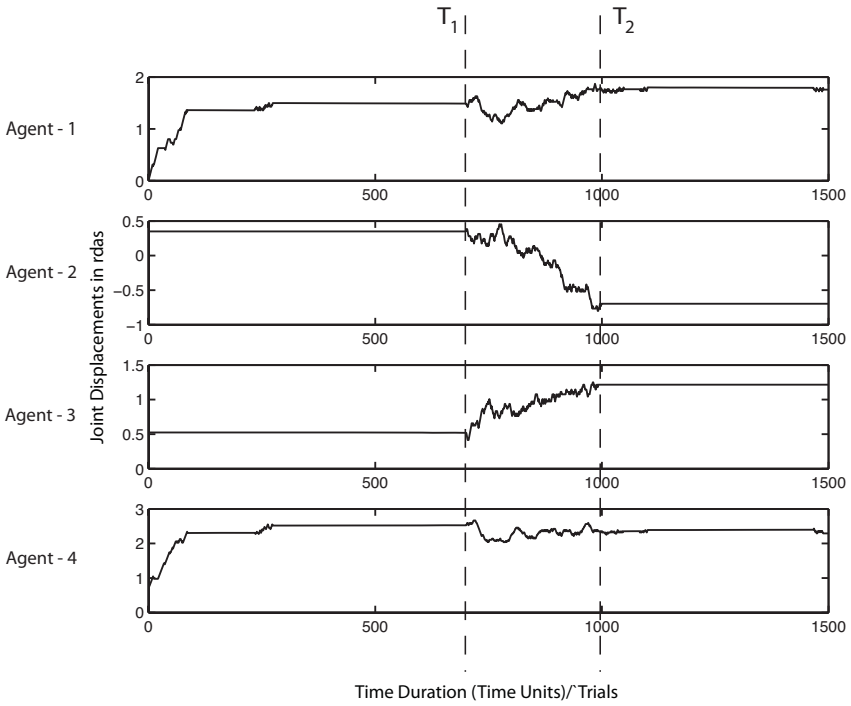
Στη συνέχεια, θα προσομοιώσουμε μία πιο σύνθετη κατάσταση αστοχίας. Στη συγκεκριμένη περίπτωση, δύο από τους πράκτορες (ο πράκτορας 2 και 3) παραμένουν ακινητοποιημένοι (στην αρχική τους θέση) από την αρχή της περιόδου εκτέλεσης της σχετικής εργασίας, αφήνοντας μόνο δύο πράκτορες (το πράκτορα 1 και 4) να λειτουργούν κανονικά από την αρχή της περιόδου εκτέλεσης της σχετικής εργασίας. Τη χρονική στιγμή T_1 $(T_1 = 700)$, οι πράκτορες 2 και 3 αρχίζουν να ανταποκρίνονται με τυχαίες γωνιακές μετατοπίσεις, για ένα περιορισμένο χρονικό διάστημα 300 χρονικών βημάτων. Στη συνέχεια, τη χρονική στιγμή T_2 $(T_2 = 1000)$, οι πράκτορες αυτοί περνούν για δεύτερη φορά σε κατάσταση εκτός λειτουργίας και παραμένουν σε αυτή μέχρι την ολοκλήρωση



Σχήμα 4.21: *Οι πράκτορες προσαρμόζουν τις δράσεις τους δυναμικά για να ανταποκριθούν στην απρόβλεπτη αποτυχία και εν συνεχεία ανάκαμψη των πρακτόρων 1 έως 3.*



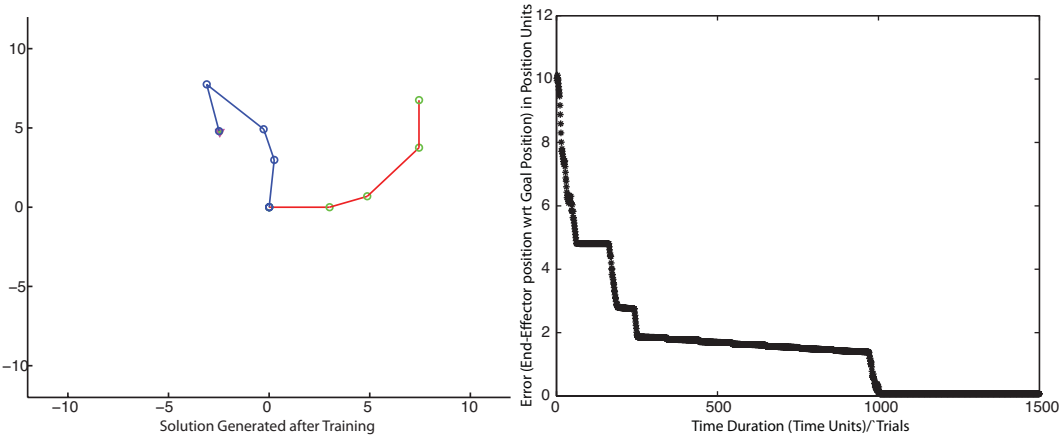
Σχήμα 4.22: *Κινηματική λύση η οποία προέκυψε από το πολυπρακτορικό σύστημα και εξέλιξη του σφάλματος θέσης συναρτήσει του χρόνου, κατά τη διαχείριση σύνθετης κατάστασης "fail-disturb-fail" για τους πράκτορες 2 και 3.*



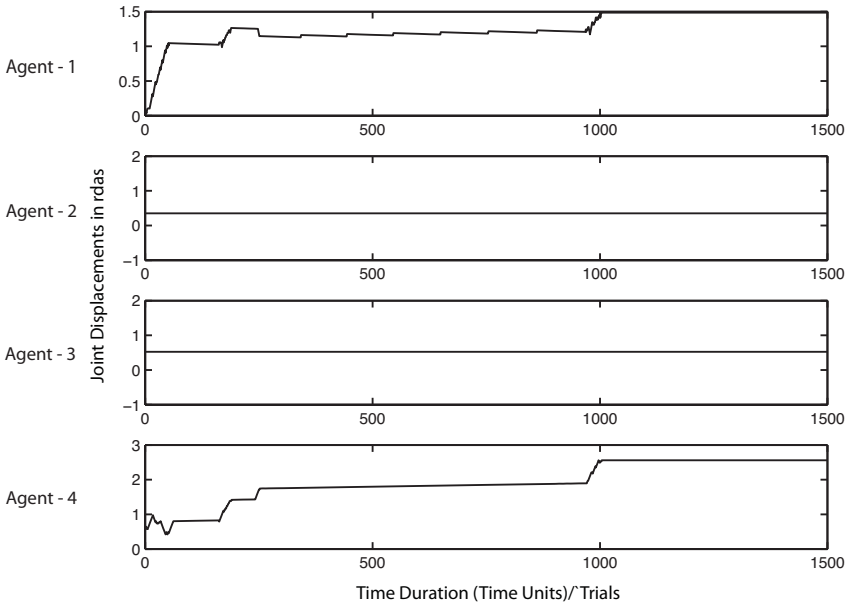
Σχήμα 4.23: *Οι πράκτορες 1 και 4 προσαρμόζουν τις δράσεις τους δυναμικά για να ανταποκριθούν στη διαχείριση της σύνθετης κατάστασης "fail-disturb-fail" για τους πράκτορες 2 και 3.*

της πειραματικής προσομοίωσης, στις νέες βέβαια γωνιακές θέσεις τις οποίες είχαν αποκτήσει κατά τη δεύτερη φάση ακινητοποίησής τους. Από τα αποτελέσματα τα οποία και απεικονίζονται στο Σχήμα 4.22, παρατηρούμε ότι το πολυπρακτορικό σύστημα καταφέρνει ξανά να βρει νέα κινηματική λύση και να προσεγγίσει τη θέση-στόχο με επιτυχία. Το αντίστοιχο Σχήμα 4.23 απεικονίζει ξεκάθαρα τις δράσεις όλων των πρακτόρων και ιδιαίτερα τον τρόπο με τον οποίο αυτοί που παραμένουν σε λειτουργία αυτοπροσαρμόζονται έτσι ώστε να ανταποκριθούν σε αυτή την σύνθετη κατάσταση "failure-disturbance-failure".

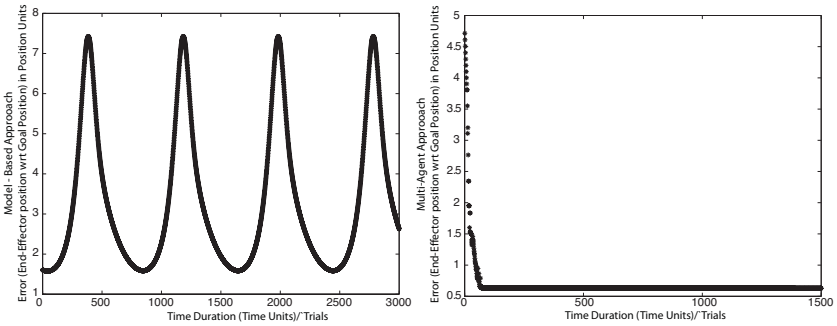
Για τη τελευταία πειραματική προσομοίωση, επαναλαμβάνουμε εν μέρει το προηγούμενο πείραμα, με μόνη διαφοροποίηση το γεγονός ότι οι πράκτορες 2 και 3 παραμένουν ακινοποιημένοι από την αρχή έως το τέλος της χρονικής περιόδου εξέλιξης της συγκεκριμένης εργασίας. Τα σχετικά Σχήματα 4.24 και 4.25 απεικονίζουν, για άλλη μία φορά ότι οι πράκτορες οι οποίοι παραμένουν σε λειτουργία, (συγκεκριμένα ο πράκτορας 1 και 4) καταφέρνουν να συνεργαστούν σε πραγματικό χρόνο, και να προσεγγίσουν τη θέση-στόχο, βρίσκοντας



Σχήμα 4.24: *Κινηματική λύση η οποία προέκυψε από το πολυπρακτορικό σύστημα και εξέλιξη του σφάλματος θέσης συναρτήσει του χρόνου, στην περίπτωση "complete fail" (fully blocked) των πρακτόρων 2 και 3.*



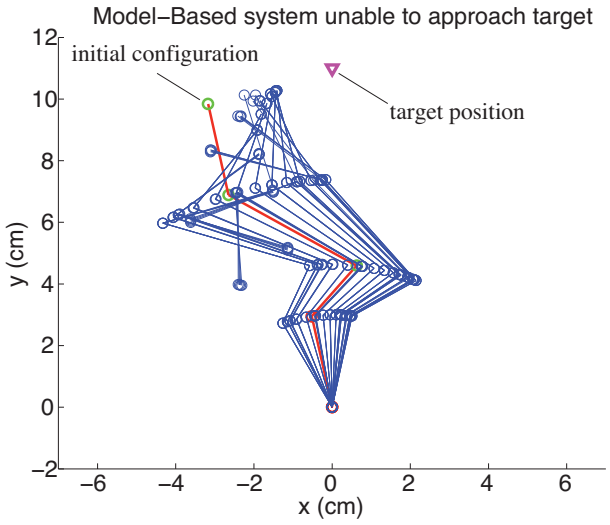
Σχήμα 4.25: *Οι πράκτορες 1 και 4 προσαρμόζουν τις δράσεις τους δυναμικά για να ανταποκριθούν στη διαχείριση της κατάστασης "complete fail" (fully blocked) των πρακτόρων 2 και 3.*



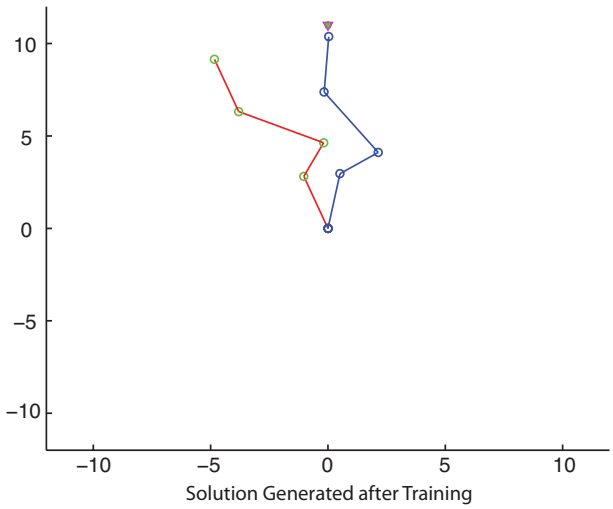
Σχήμα 4.26: *Συγκριτικά αποτελέσματα εξέλιξης σφάλματος θέσης (model-based vs. multi-agent approach) στην περίπτωση καθολικής αστοχίας των πρακτόρων 2 και 3 (fully blocked).*

μια καινούργια λύση η οποία ενσωματώνει και τις απρόβλεπτες αστοχίες που παρουσίασαν οι πράκτορες 2 και 3. Το σύνολο αυτών των αποτελεσμάτων επιδεικνύει ξεκάθαρα τα ιδιαίτερα χαρακτηριστικά ευρωστίας που παρουσιάζει το προτεινόμενο πολυπρακτορικό σύστημα, καθώς και τις δυνατότητες που παρέχει ως προς τη διαχείριση απρόβλεπτων και σύνθετων καταστάσεων αστοχίας, οι οποίες ουσιαστικά ισοδυναμούν με απότομες αλλαγές στη κινηματική τοπολογία και τις οποίες στρατηγικές ελέγχου οι οποίες λειτουργούν βάσει μοντέλου δε θα ήταν σε θέση να χειριστούν.

Με σκοπό την ανάδειξη των ιδιαίτερων χαρακτηριστικών ευρωστίας που επιδεικνύει το προτεινόμενο πολυπρακτορικό σύστημα το οποίο λειτουργεί δίχως να κάνει χρήση κάποιου μοντέλου (model-free approach), θα πραγματοποιήσουμε μια συγκριτική ανάλυση αυτής της στρατηγικής ελέγχου με την αντίστοιχη μονοπρακτορική στρατηγική η οποία λειτουργεί στη βάσει κάποιου συγκεκριμένου μοντέλου (model-based approach). Υποθέτουμε ξανά την ίδια κατάσταση αποτυχίας όπως και στη τελευταία πειραματική προσομοίωση. Πιο συγκεκριμένα, αντιμετωπίζουμε την κατάσταση κατά την οποία οι αρθρώσεις 2 και 3 παραμένουν πλήρως ακινητοποιημένες, με βασική διαφορά ότι εφαρμόζουμε βάσει μοντέλου (resolved motion-rate) κινηματικό έλεγχο κάνοντας χρήση της μεθόδου της ψευδοαντιστρόφου της Ιακωβιανής μήτρας (Jacobian matrix pseudo-inverse). Η εξέλιξη του σφάλματος θέσης απεικονίζεται στα αποτελέσματα του Σχήματος 4.26 και για τις δύο περιπτώσεις (model-based και multi-agent approach). Επιπλέον, η κίνηση της κινηματικής αλυσίδας και στις δύο περιπτώσεις απεικονίζεται στα Σχήματα 4.27 και 4.28. Από τα παραπάνω αποτελέσματα είναι προφανές πλέον ότι η κινηματική αλυσίδα, ενσωματώνοντας κινηματικό έλεγχο βάσει-μοντέλου, δεν είναι σε θέση να προσεγγίσει την επιθυμητή θέση-στόχο, ενώ η κίνησή της παρουσιάζει χαρακτηριστικά ταλάν-



Σχήμα 4.27: *Σύστημα Βάσει-Μοντέλου: Συγκριτικά αποτελέσματα (model-based vs. multi-agent approach) στην περίπτωση καθολικής αστοχίας των πρακτόρων 2 και 3 (fully blocked).*



Σχήμα 4.28: *Πολυπρακτορικό Σύστημα: Συγκριτικά αποτελέσματα (model-based vs. multi-agent approach) στην περίπτωση καθολικής αστοχίας των πρακτόρων 2 και 3 (fully blocked).*

τωσης μεταξύ συγκεκριμένων διατάξεων των αρθρώσεων της αλυσίδας. Αυτό προφανώς οφείλεται στο γεγονός ότι το υπάρχον κινηματικό μοντέλο είναι άκυρο στη συγκεριμένη κατάσταση και ένα σύστημα του οποίου η λειτουργία είναι βάσει-μοντέλου, δεν διαθέτει το μηχανισμό να διαχειριστεί (χωρίς επαναπρογραμματισμό και επανασχεδιασμό) τα απρόβλεπτα σφάλματα, δημιουργώντας λοιπόν κινήσεις οι οποίες δεν συνάδουν με τον στόχο στο χώρο εργασίας. Σε αντίθεση με αυτήν τη συμπεριφορά, το προτεινόμενο πολυπρακτορικό σύστημα καταφέρνει να βρει εφικτή λύση όπως αυτή απεικονίζεται στο σχετικό Σχήμα 4.28.

Μολονότι σε καμία περίπτωση δεν ισχυριζόμαστε ότι προσεγγίσεις και μεθοδολογίες οι οποίες δεν κάνουν χρήση μοντέλου, μπορούν να παρέχουν την ακρίβεια που διασφαλίζουν οι στρατηγικές που στηρίζονται σε συγκεκριμένο μοντέλο, τα παραπάνω αποτελέσματα επιδεικνύουν συγκεκριμένα ιδιαίτερα χαρακτηριστικά. Η προσαρμοστικότητα σε αλλαγές στην κινηματική τοπολογία καθώς και η ικανότητα του προτεινόμενου συστήματος να ανταπεξέλθει σε απρόβλεπτες και σύνθετες αστοχίες, χωρίς επαναπρογραμματισμό ή επανασχεδίαση, σαφώς αποτελούν λειτουργικές ιδιότητες τις οποίες η προτεινομένη αρχιτεκτονική καταφέρνει να διασφαλίσει.

4.3.3 Επεκτασιμότητα Πολυπρακτορικής Αρχιτεκτονικής σε Εργασίες με Περιορισμούς Κίνησης

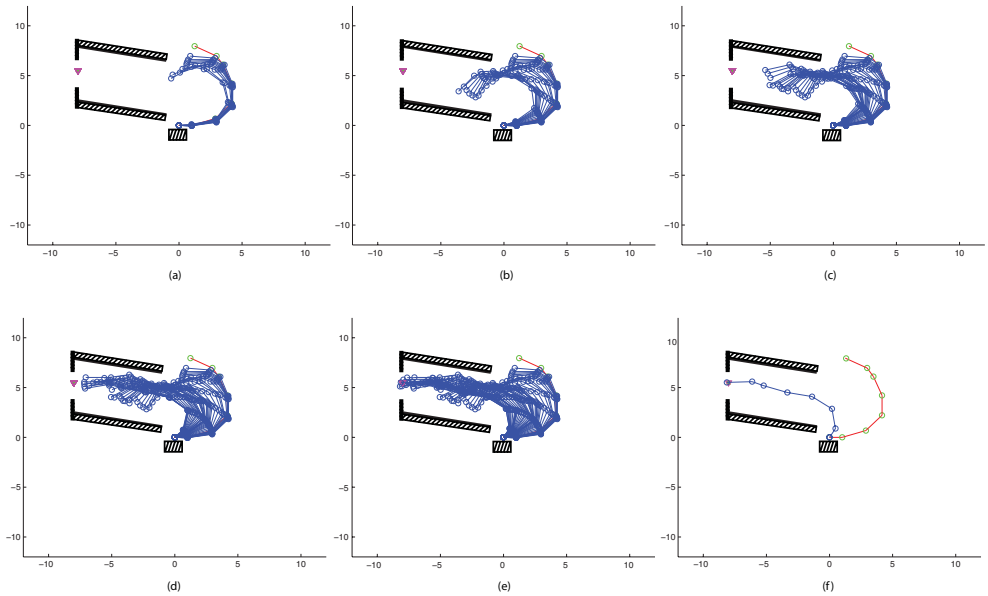
Στην προσπάθεια αποτίμησης της επεκτασιμότητας της προτεινόμενης αρχιτεκτονικής σε πιο σύνθετες τοπολογίες και διατάξεις, θα εξετάσουμε εφαρμογές σε κινήσεις που περιλαμβάνουν περιορισμούς. Πιο συγκεκριμένα, υποθέτουμε για το σύστημα την ακόλουθη υπό περιορισμούς κίνηση (Constrained Motion): μία κινηματική αλυσίδα 7 βαθμών ελευθερίας, η οποία ενσωματώνει την προτεινόμενη πολυπρακτορική αρχιτεκτονική, θα πρέπει να μάθει να προσεγγίζει τη θέση-στόχο η οποία αυτή τη φορά είναι στο τέλος ενός στενού διαδρόμου. Ο στόχος για το σύστημά μας είναι να βρει για την κινηματική αλυσίδα ένα σύνολο κινήσεων οι οποίες θα καταφέρουν να φέρουν το τελικό στοιχείο δράσης της αλυσίδας στη θέση-στόχο δίχως να υπάρχει κάποιο μοντέλο του περιβάλλοντος και προφανώς με ασφάλεια, δηλαδή, χωρίς συγκρούσεις της κινηματικής αλυσίδας με το περιβάλλον. Σε ένα τέτοιο πειραματικό σενάριο, μια τυπική κινηματική προσέγγιση βάσει-μοντέλου (μονοπρακτορική), θα πρέπει να εισάγει πρόσθετα κριτήρια βελτιστοποίησης απόστασης, γεγονός το οποίο θα έκανε την αναλυτική προσέγγιση επίλυσης του συγκεκριμένου προβλήματος, το οποίο περιλαμβάνει πλεονάζοντες βαθμούς ελευθερίας, πιο σύνθετο και σαφώς πιο ευαίσθητο σε πιθανά σφάλματα κατά τη διαδικασία δημιουργίας ενός σχετικού μοντέλου. Η προτεινόμενη κατανεμημένη πολυπρακτορική αρχιτεκτονική

επεκτείνεται με τρόπο πολύ φυσικό, έτσι ώστε να διαχειριστεί την προτεινόμενη τοπολογία με μόνη πρόσθετη απαίτηση την προσθήκη ενός επιπλέον όρου στη συνάρτηση ανταπόδοσης, του οποίου στόχος είναι η διαχείριση περιπτώσεων σύγκρουσης με τα εμπόδια του χώρου εργασίας που συνιστούν τους περιορισμούς κίνησης.

Η ανταπόδοση την οποία λαμβάνει ένας πράκτορας τη χρονική στιγμή t, έχοντας επιλέξει συγκεκριμένη δράση και έχοντας μεταβεί σε μία νέα κατάσταση, ορίζεται με παρόμοιο τρόπο όπως εκείνος της εξίσωσης (4.1) και με τη συνάρτηση ανταπόδοσης $R(t)$ στη συγκεκριμένη περίπτωση να επαναδιατυπώνεται ως εξής:

$$
\begin{cases}
\text{if} \quad (D_{goal}(t) \leq D_{\min}) \wedge (\Delta D_{goal}) \leq 0) \wedge (\forall\ a_i,\ C_{a_i} = 0) \qquad \text{then} \\
\qquad R(t) = e^{-c_1 \cdot \left(D_{goal}(t)\right)} \\
\text{if} \quad (D_{goal}(t) > D_{\min}) \qquad \text{then} \\
\qquad R(t) = -2 \\
\text{if} \quad (D_{goal}(t) < D_{\min}) \wedge (\Delta D_{goal} > 0) \qquad \text{then} \\
\qquad R(t) = -1 \\
\text{if} \quad (\exists\ a_i,\ C_{a_i} = 1) \qquad \text{then} \\
\qquad R(t) = -e^{-c_2 \cdot \left(\min D^{a_i}_{Corridor}(t)\right)}
\end{cases}
\tag{4.1}
$$

όπου $D_{goal}\,(t)$ είναι η απόσταση από το στόχο τη χρονική επανάληψη t, $D_{\min}$ είναι το κατώφλι απόστασης μετά από το οποίο οι πράκτορες αρχίζουν να λαμβάνουν ανταπόδοση, ΔD_{goal} είναι ο ρυθμός μεταβολής απόστασης από το στόχο, C_{a_i} είναι boolean flag που επισημαίνει ότι ο πράκτορας a_i είναι κοντά σε κατάσταση σύγκρουσης με το διάδρομο και τέλος $D^{a_i}_{Corridor}$ είναι η απόσταση στην οποία ο συγκεκριμένος πράκτορας a_i αντιλαμβάνεται τοπικά την ύπαρξη εμποδίου. Να επισημάνουμε εδώ ότι, στη συνάρτηση ανταπόδοσης, εάν περισσότεροι του ενός πράκτορες διαπιστώνουν να βρίσκονται κοντά σε σύγκρουση,

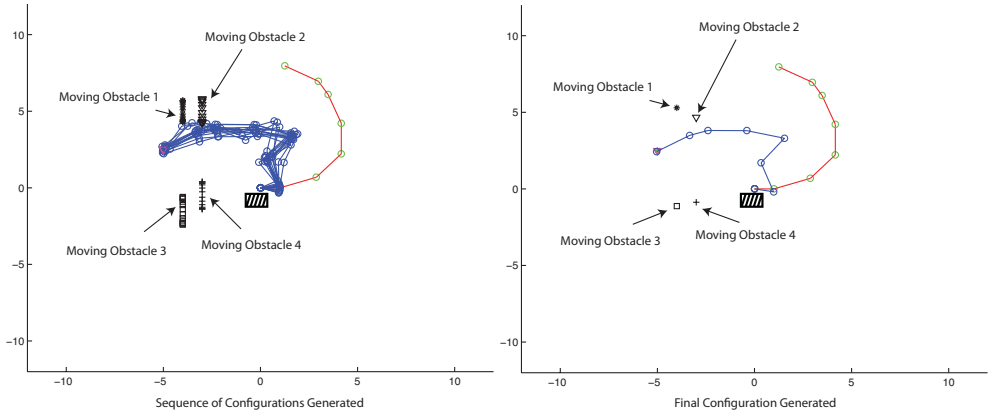


Σχήμα 4.29: *Επεκτασιμότητα αρχιτεκτονικής σε 7 dof κινηματική αλυσίδα η οποία εκτελεί κίνηση υπό περιορισμούς, διαμέσους στενής διόδου, χωρίς συγκρούσεις*

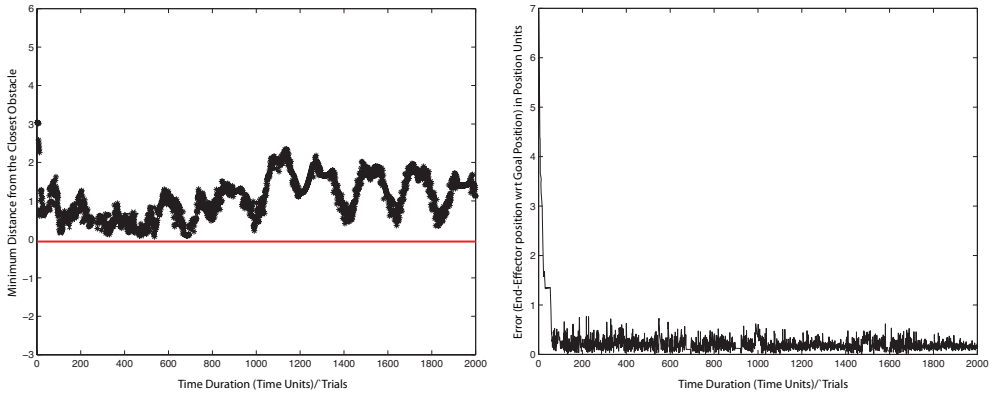
για να ενισχυθεί η συνολική αρνητική ανταπόδοση στο πολυπρακτορικό σύστημα, υπολογίζουμε τη τιμή $R(t)$ με βάση την τιμή $\min D^{a_i}_{Corridor}$ μεταξύ όλων των πρακτόρων που αντιλαμβάνονται τοπικά κάποια σύγκρουση.

Τα αποτελέσματα τα οποία προκύπτουν συγκεντρωτικά, από τη χρήση της πολυπρακτορικής αρχιτεκτονικής στη συγκεκριμένη τοπολογία, απεικονίζονται στο Σχήμα 4.29. Τα αποτελέσματα αυτά επιδεικνύουν ότι το πολυπρακτορικό σύστημα επιτυγχάνει να εξελίξει μία τέτοια συμπεριφορά η οποία αποτελεσματικά καθοδηγεί την κινηματική αλυσίδα έτσι ώστε να κινηθεί στο στενό διάδρομο και να φτάσει στη θέση-στόχο, με ασφάλεια και δίχως συγκρούσεις με το περιβάλλον.

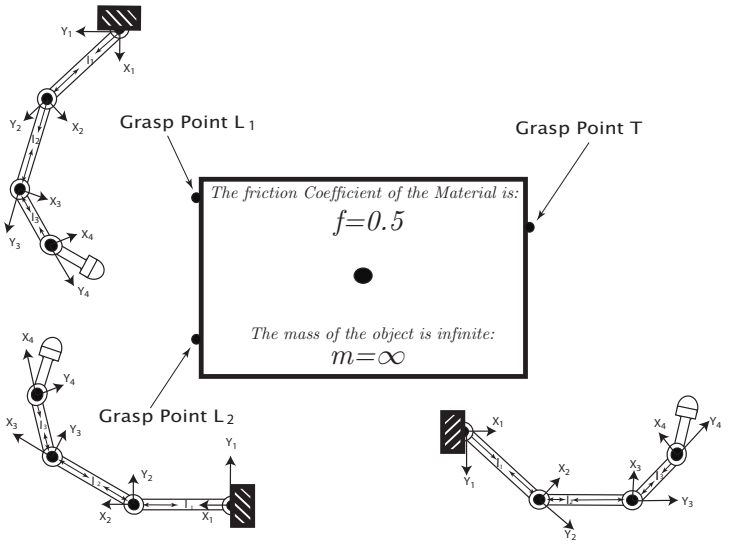
Στη συνέχεια, μελετούμε την περίπτωση μη-στατικού περιβάλλοντος, το οποίο περιλαμβάνει αντικείμενα τα οποία κινούνται. Στα Σχήματα 4.30 και 4.31, παρατηρούμε ότι το πολυπρακτορικό σύστημα καταφέρνει να μάθει να αποφεύγει τις συγκρούσεις και με τα τέσσερα εμπόδια τα οποία πραγματοποιούν ταλάντωση σε ολόκληρη τη διάρκεια της προσομοίωσης, διατηρώντας μία ελάχιστη απόσταση ασφαλείας από το σύνολο των εμποδίων (με μέση τιμή $mean = 1.0965$), ενώ την ίδια στιγμή διατηρεί συνεχή επαφή με το στόχο σε ολόκληρη τη διάρκεια εξέλιξης του πειράματος, όπως αυτό απεικονίζεται στο



Σχήμα 4.30: *Επεκτασιμότητα της αρχιτεκτονικής σε 7 dof κινηματική αλυσίδα η οποία εκτελεί κίνηση υπό περιορισμούς (χωρίς συγκρούσεις), σε μη στατικό περιβάλλον (συνδυασμός τεσσάρων κινούμενων εμποδίων)*



Σχήμα 4.31: *Το πολυπρακτορικό σύστημα διατηρεί επαφή με το στόχο (Σχήμα 4.31b), ενώ αποφεύγει συγκρούσεις με όλα τα κινούμενα εμπόδια διατηρώντας ελάχιστη απόσταση ασφαλείας κατά τη συνολική διάρκεια εκτέλεσης του πειράματος (Σχήμα 4.31a)*

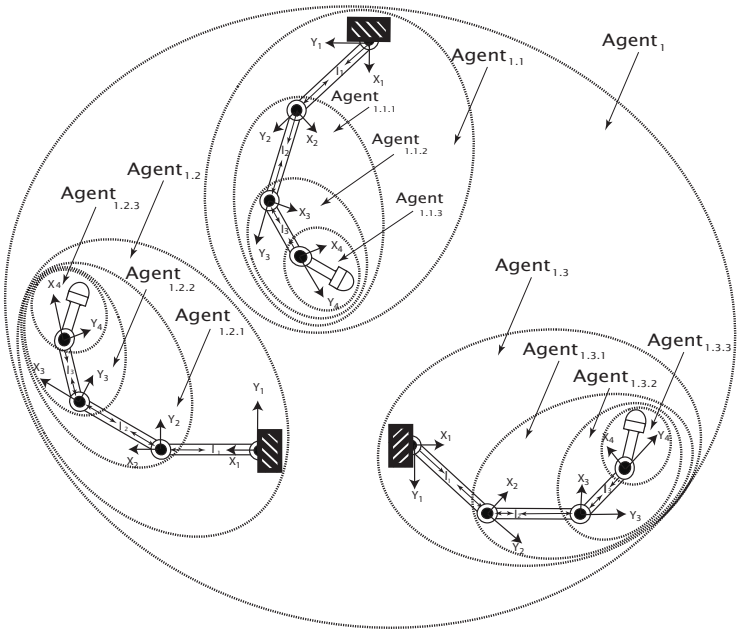


Σχήμα 4.32: *Λαβή τριών δακτύλων τύπου "quasi-static grasp"*

Σχήμα 4.31β (με μέση τιμή του σφάλματος θέσης: $mean = 0.2060$).

4.4 Πολυαρθρωτή Ρομποτική Λαβή

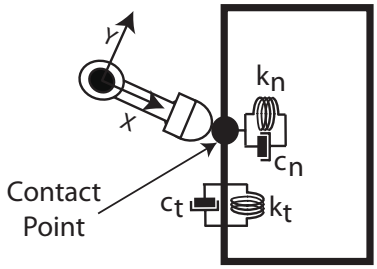
Έχοντας παρουσιάσει και αναλύσει, μέχρι τώρα, τα αποτελέσματα εφαρμογής της προτεινόμενης αρχιτεκτονικής σε μια απλή κινηματική αλυσίδα (την οποία αποτελούν τέσσερις σύνδεσμοι - πράκτορες, και οι οποίοι προσπαθούν να συνεργαστούν έτσι ώστε να προσεγγίσουν τη θέση-στόχο, χωρίς προηγούμενη εμπειρία ή προϋπάρχον συμπεριφορικό μοντέλο κίνησης), στη συνέχεια θα αξιολογήσουμε την ίδια αρχιτεκτονική σε πιο σύνθετη ως προς τη διάταξη των πρακτόρων πειραματική εφαρμογή. Συγκεκριμένα, μοντελοποιούμε ρομποτικό χειριστή τριών δακτύλων, ο οποίος επιχειρεί στατική λαβή (quasi-static grasp) αντικειμένου (θεωρητικά άπειρης μάζας) όπως φαίνεται στο Σχήμα 4.32. Να τονίσουμε εδώ ότι στα πλαίσια αυτού του πειράματος υποθέτουμε ότι πραγματοποιούμε κινηματικό έλεγχο του συστήματος, παραλείποντας από το πρόβλημα του ελέγχου τη δυναμική του αντικειμένου και των συνδέσμων, που προφανώς και εμπλέκονται. Τα αριθμητικά πειράματα τα οποία παρουσιάζονται στη συνέχεια είναι μια προσπάθεια αξιολόγησης της μαθησιακής ικανότητας ενός πολυπρακτορικού συστήματος το οποίο αποτελείται από 13 εμφωλευμένους πράκτορες, οι οποίοι πραγματοποιούν λαβή (grasp) χωρίς ολίσθηση σε κάποιο ακροδάκτυλο (fingertip) του χειριστή, και με επιθυμητή συνολική δύναμη και ροπή ίση με το μηδέν. Η αποτύπωση των κινηματικών αλυσίδων του Σχήματος 4.32, στο αντίστοιχο πολυπρακτορικό περιβάλλον απεικονίζεται στο Σχήμα



Σχήμα 4.33: *Πολυπρακτορική αναπαράσταση χειριστή τριών δακτύλων*

4.33. Αυτή η πολυπρακτορική διάταξη δείχνει τόσο την ιεραρχική όσο και την εμφωλευμένη φύση της προτεινόμενης αρχιτεκτονικής. Το σύνολο των παραμέτρων που χρησιμοποιήθηκαν για την προσομοίωση φαίνονται συγκεντρωτικά στο Πίνακα 4.4.

Βλέπουμε λοιπόν την αρχική διάταξη κάθε άρθρωσης της κινηματικής αλυσίδας καθώς επίσης τον κώνο τριβής σε κάθε σημειακή επαφή της λαβής (grasp point), μέσα στον οποίο θα πρέπει να μάθουν με λειτουργούν όλοι οι πράκτορες του συστήματος. Σημειώνουμε ότι όλα τα σημεία επαφής είναι μοντελοποιημένα κάνοντας χρήση γραμμικών ελαστικών μοντέλων με συγκεκριμένους συντελεστές απόσβεσης και ελατηρίου, όπως φαίνονται και στο σχετικό Σχήμα 4.34,



Σχήμα 4.34: *Προσομοίωση σημείου επαφής με κάθετους και εφαπτομενικούς συντελεστές απόσβεσης - ελατηρίου*

Πίνακας 4.4: Παράμετροι στατικής ρομποτικής λαβής

	Finger L_1	Finger L_2	Finger T
Num of Links (l_i)	4	4	4
Initial Joint Angle l_1	130°	180°	10°
Initial Joint Angle l_2	20°	340°	20°
Initial Joint Angle l_3	30°	330°	30°
Initial Joint Angle l_4	50°	310°	40°
Step: Joint Angle	18°	18°	18°
Step: Goal Angle	40°	40°	40°
Friction Coefficient	0.5	0.5	0.5

όπου k_t είναι η εφαπτομενική ενώ k_n είναι η κάθετη συνιστώσα του συντελεστή του ελατηρίου, ενώ c_n, c_t είναι οι αντίστοιχες συνιστώσες του συντελεστή απόσβεσης. Οι υπόλοιπες παράμετροι της σχετικής προσομοίωσης είναι ίδιες με εκείνες που αναφέρονται σε προηγούμενη παράγραφο. Οι τιμές των παραμέτρων ασαφοποίησης όσο και εκείνες του μηχανισμού μάθησης διατηρούν τις ίδιες τιμές. Το σύστημα εκπαιδεύεται για περίοδο 200 εποχών κάθε μία έχει διάρκεια 500 μονάδες χρόνου. Στο συγκεκριμένο χρονικό διάστημα οι πράκτορες θα πρέπει να μάθουν ταυτόχρονα να λύνουν δύο προβλήματα: α) Να προσεγγίζουν να σημεία επαφής (όπως ακριβώς και στο προηγούμενο πρόβλημα με τη μονή κινηματική αλυσίδα) και β) να συντονίσουν τις κινήσεις τους έτσι ώστε να εφαρμόζουν κατάλληλες δυνάμεις στα σημεία επαφής, πάντα μεταξύ των ορίων που θέτει ο κώνος τριβής, χωρίς φαινόμενα ολίσθησης, και για τα πλαίσια του συγκεκριμένου πειράματος να διασφαλίζουν συνολική δύναμη και ροπή ίση με το μηδέν.

Με δεδομένο αυτές τις νέες απαιτήσεις οι οποίες εισάγονται στο πείραμα, έχουμε προφανώς επαύξηση του προς επίτευξη στόχου για το πολυπρακτορικό σύστημα. Η επαύξηση αυτή συνεπάγεται προφανώς σχετική τροποποίηση της συνάρτησης ανταπόδοσης έτσι ώστε να είναι δυνατή η ορθή λειτουργία του μηχανισμού ενισχυτικής μάθησης. Η συνάρτηση ανταπόδοσης την οποία είδαμε σε προηγούμενη παράγραφο και η οποία καθοδηγούσε το μηχανισμό μάθησης έτσι ώστε η κινηματική αλυσίδα να μπορεί να προσεγγίζει τη θέση-στόχο, εξακολουθεί να ισχύει απλά ενσωματώνεται σε αυτή ένας κατάλληλος όρος R_i που αντιστοιχεί στους περιορισμούς των τριών υποεργασιών, πιο συγκεκριμένα: R_τ είναι η ανταπόδοση η οποία απονέμεται σε αντιστοιχία με το περιορισμό της συνολικής ροπής, R_f είναι η ανταπόδοση η οποία αντιστοιχεί στη ικανοποίηση του περιορισμού συνολικής δύναμης, και R_s είναι η ανταπόδοση η οποία απονέμεται

σε σχέση με τον περιορισμό του κώνου τριβής. Συνεπώς η νέα συνάρτηση ανταπόδοσης η οποία καθοδηγεί την τρέχουσα διαδικασία μάθησης έχει την εξής μορφή:

$$R(t) = \left\{ \begin{array}{l} \text{if } (Dist_{Goal}(t) \leq Dist_{min}) \wedge (\Delta Dist_{Goal}) \leq 0) \text{ then} \\ \quad R(t) = e^{-c*Dist_{Goal}(t)} + \sum_i R_i \\ \text{else if } (Dist_{Goal}(t) > Dist_{min}) \text{ then} \\ \quad R(t) = -2 + \sum_i R_i \\ \text{else if } (Dist_{Goal}(t) < Dist_{min}) \wedge (\Delta Dist_{Goal}) > 0) \text{ then} \\ \quad R(t) = -1 + \sum_i R_i \end{array} \right\} \quad (4.2)$$

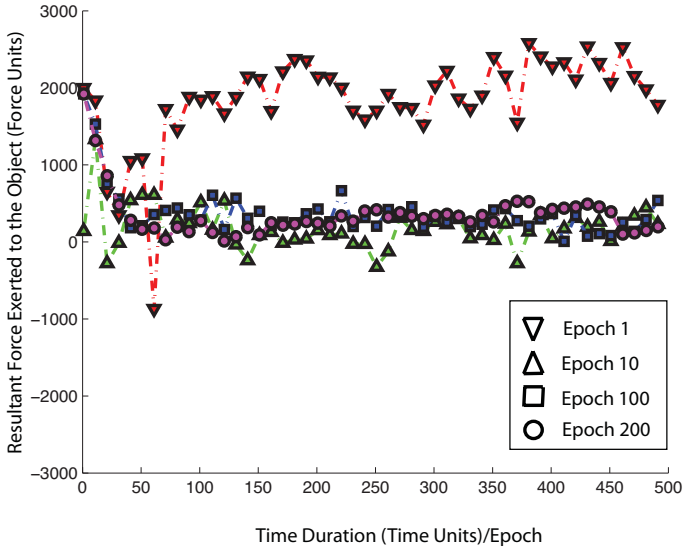
Οι νέοι όροι R_i της συνάρτησης ανταπόδοσης ($i = f$: περιορισμός δύναμης, $i = \tau$: περιορισμός ροπής, και $i = s$: περιορισμός ολίσθησης) ορίζονται συγκεκριμένα ως ακολούθως:

$$R_f = e^{-c*Net_{force}(t)}, R_\tau = e^{-c*Net_{torque}(t)}, R_s = e^{-c*Friction_{Cone}(t)}$$

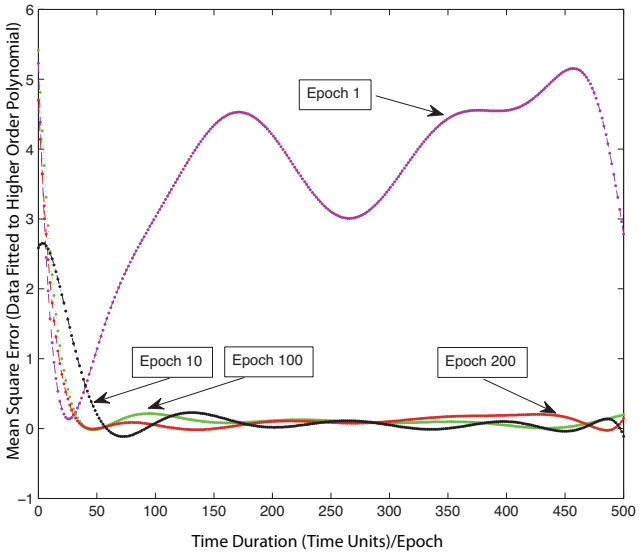
με τη μεταβλητή t να υποδεικνύει ως συνήθως την τρέχουσα χρονική στιγμή (διακριτό χρονικό βήμα).

Στο σύνολό της η πειραματική διαδικασία αποτελείται από τρεις υπο-ομάδες πειραμάτων, κάθε μία εκτελούμενη με διαφορετική τιμή συντελεστή βαθμιαίας μείωσης decay factor s ($s = 0.05$, $s = 0.35$, και $s = 0.75$). Τα αποτελέσματα τα οποία συγκεντρώθηκαν από αυτήν τη διαδικασία παρουσιάζονται στα Σχήματα 4.35 έως 4.40. Σε κάθε διάγραμμα παρουσιάζονται τα συγκεντρωτικά αποτελέσματα για τέσσερις διαφορετικές εποχές (1, 10, 100 και 200). Τα διαγράμματα αυτά αποτυπώνουν την εξέλιξη με την πάροδο του χρόνου της συνολικής εφαρμοζόμενης δύναμης (net force) καθώς και του μέσου σφάλματος δύναμης (με επιθυμητή συνολική δύναμη ίση με μηδέν, desired net force = 0) απεικονίζοντας έτσι τη μαθησιακή απόδοση του πολυπρακτορικού συστήματος καθώς και τη δυνατότητα της προτεινόμενης αρχιτεκτονικής για συνεργατική προσαρμοστικότητα στο πλαίσιο μιας εργασίας επιδέξιου ρομποτικού χειρισμού. Τέλος, στο Σχήμα 4.41 απεικονίζονται ενδεικτικές συγκεντρωτικές καμπύλες μάθησης του συστήματος, αποτυπώνοντας την εξέλιξη του μέσου τετραγωνικού σφάλματος (που επιτυγχάνεται στη διάρκεια μιας εποχής μάθησης) συναρτήσει της εποχής (για 1 έως 200 εποχές) και για τους τρεις διαφορετικούς συντελεστές βαθμιαίας μείωσης με τους οποίους εκτελέσθηκαν τα σχετικά πειράματα.

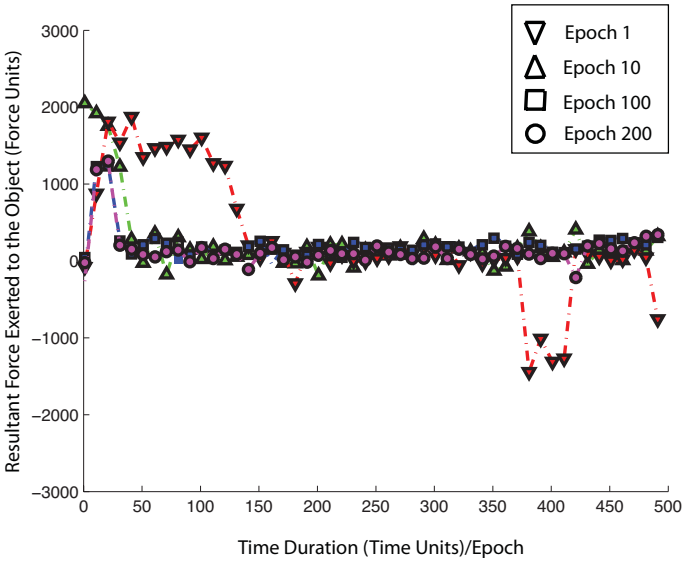
Σχετικά με το συντελεστή τριβής (ο οποίος στην περίπτωση την οποία και εξετάζουμε είναι $\mu = 0.5$), θα πρέπει να τονίσουμε ότι ορίζει τον κώνο τριβής και συνεπώς προσδιορίζει τα όρια μέσα στα οποία λειτουργούν τα ρομποτικά



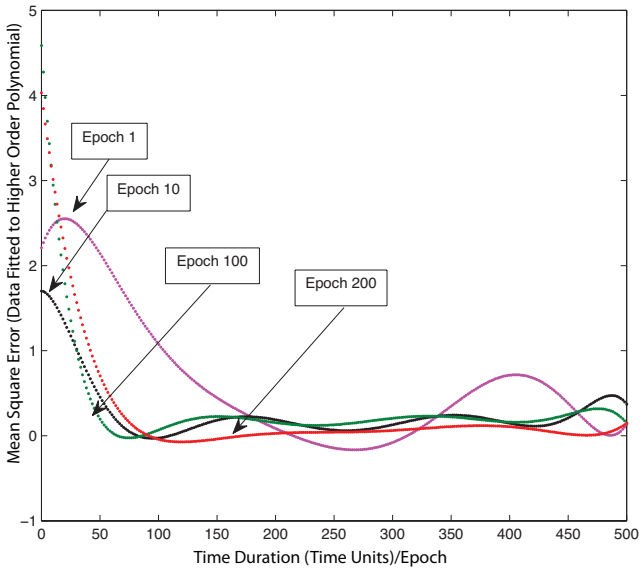
Σχήμα 4.35: *Συνολική δύναμη (Net Force), συναρτήσει του χρόνου, για διαφορετικές εποχές (με συντελεστή βαθμιαίας μείωσης $s = 0.05$)*



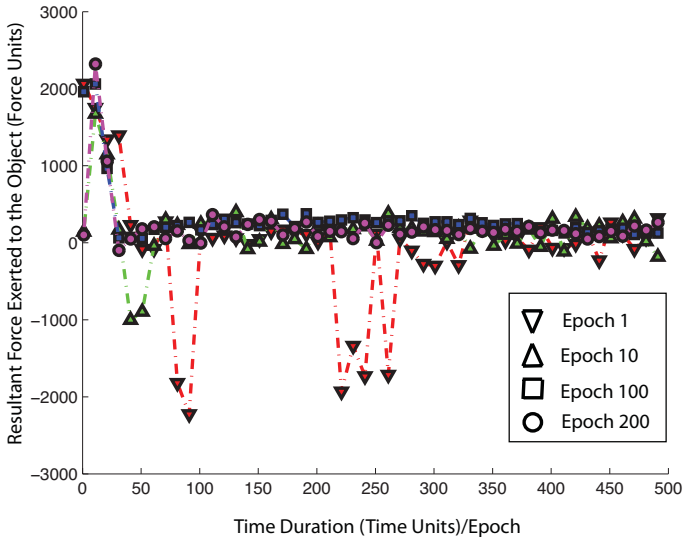
Σχήμα 4.36: *Μέσο τετραγωνικό σφάλμα για διαφορετικές εποχές (για συντελεστή βαθμιαίας μείωσης $s = 0.05$)*



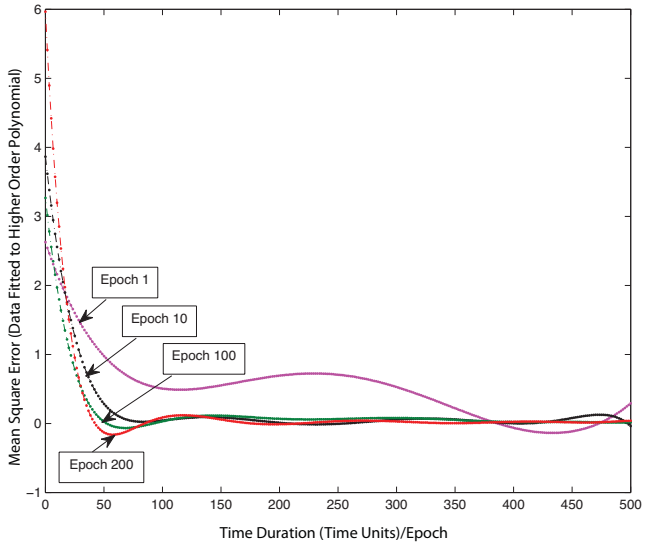
Σχήμα 4.37: *Συνολική δύναμη (Net Force), συναρτήσει του χρόνου, για διαφορετικές εποχές (με συντελεστή βαθμιαίας μείωσης $s = 0.35$)*



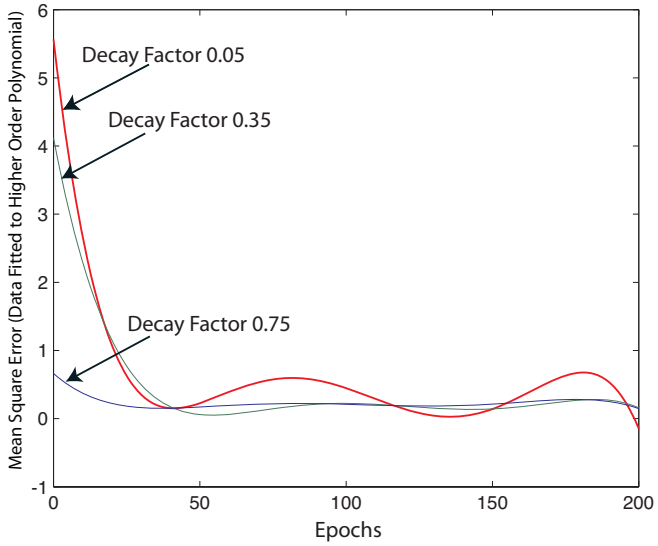
Σχήμα 4.38: *Μέσο τετραγωνικό σφάλμα για διαφορετικές εποχές (για συντελεστή βαθμιαίας μείωσης $s = 0.35$)*



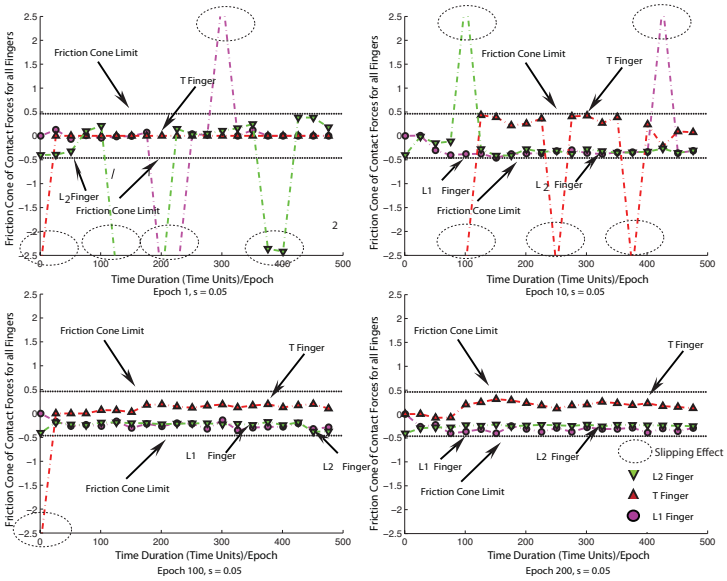
Σχήμα 4.39: *Συνολική δύναμη (Net Force), συναρτήσει του χρόνου, για διαφορετικές εποχές (με συντελεστή βαθμιαίας μείωσης $s = 0.75$)*



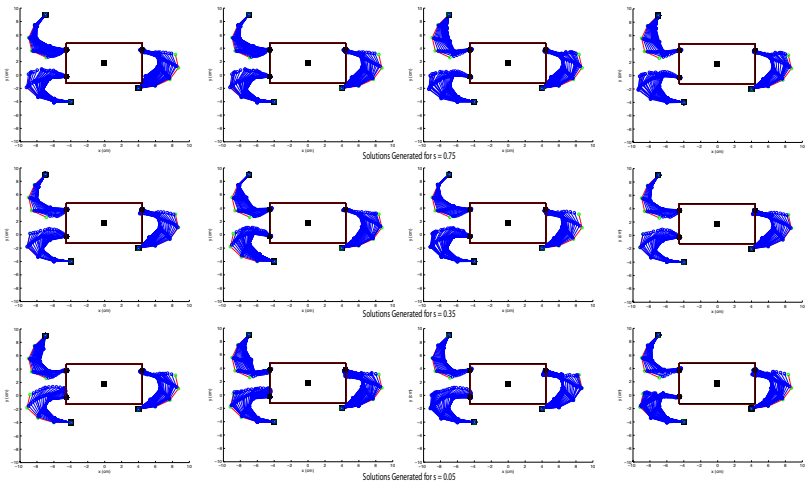
Σχήμα 4.40: *Μέσο τετραγωνικό σφάλμα για διαφορετικές εποχές (για συντελεστή βαθμιαίας μείωσης $s = 0.75$)*



Σχήμα 4.41: *Καμπύλες μάθησης του συστήματος: προσεγγιστική καμπύλη εξέλιξης μέσου τετραγωνικού σφάλματος συναρτήσει της εποχής μάθησης, για 3 διαφορετικούς συντελεστές βαθμιαίας μείωσης s*



Σχήμα 4.42: *Προσαρμογή του ακροδακτύλου (Fingertip) στα όρια του κώνου τριβής, για συντελεστή βαθμιαίας μείωσης* $s = 0.05$



Σχήμα 4.43: *Παραδείγματα παραγόμενων λαβών (για συντελεστές βαθμιαίας μείωσης* $s = 0.75, 0.35,$ *και* 0.05*)*

δάκτυλα (ουσιαστικά τα όρια μέσα στα οποία πρέπει να περιορίζονται οι δυνάμεις στα σημεία επαφής των δακτύλων). Το Σχήμα 4.42 αποτυπώνει τον τρόπο με τον οποίο οι δυνάμεις στα σημεία επαφής των δακτύλων των τριών κινηματικών αλυσίδων προοδευτικά προσαρμόζονται, κατά τη διάρκεια των εποχών που εξετάζουμε, μέσα στα όρια που προσδιορίζει ο συγκεκριμένος κώνος τριβής. Τέλος, στο Σχήμα 4.43, μπορούμε να δούμε πιθανές λύσεις που παράγει το πολυπρακτορικό σύστημα για διαφορετικές παραμέτρους προσομοίωσης. Η προτεινόμενη πολυπρακτορική διάταξη προχωρεί σε χρήση της γνώσης που έχει αποκτήσει, δίχως περαιτέρω εκπαίδευση, στην προσέγγιση και άλλων, νέων σημείων επαφής με τον ίδιο ακριβώς τρόπο, όπως και στην περίπτωση της απλής κινηματικής αλυσίδας.

4.5 Εφαρμογή σε Συνεργατικά Αυτοκινούμενα Ρομπότ

Στη τελευταία παράγραφο του παρόντος κεφαλαίου παρουσιάζουμε την πειραματική διάταξη των αυτοκινούμενων ρομπότ πάνω στη οποία βασίσθηκε το τελευταίο μέρος των πειραματικών δοκιμών αξιολόγησης της προτεινόμενης πολυπρακτορικής αρχιτεκτονικής ρομποτικού ελέγχου. Οι πειραματικές δοκιμές εκπονήθηκαν τόσο επί της πλατφόρμας προσομοίωσης (Webots 6.1.5) όσο και μέσω πραγματικών εργαστηριακών δοκιμών σε οχήματα τύπου e-Puck. Όπως έχει ήδη αναφερθεί στην αρχή του κεφαλαίου, στόχος των πειραματικών αυ-

τών δοκιμών είναι η αξιολόγηση του προτεινόμενου πολυπρακτορικού σχήματος και του αντίστοιχου πλαισίου μάθησης στο πεδίο των αυτοκινούμενων ρομπότ. Η εργασία box-pushing σχετίζεται με το γνωστό πρόβλημα "piano mover's problem" και η διατύπωσή του είναι η ακόλουθη: "*με δεδομένο ένα τυχαίο άκαμπτο πολυεδρικό περιβάλλον, αναζητούμε ένα συνεχή και ελεύθερο συγκρούσεων δρόμο (collision-free path) μέσω του οποίου θα είναι δυνατή η μεταφορά ενός αντικειμένου από μία αρχική θέση σε μία τελική επιθυμητή θέση-στόχο*". Αποδείχθηκε από τον Reif [102] ότι αυτό το πρόβλημα είναι πολυπλοκότητας τύπου PSPACE-hard (να σημειωθεί ότι ένα πρόβλημα είναι τάξης πολυπλοκότητας PSPACE όταν η επίλυσή του απαιτεί πολυωνυμικό χώρο αποθήκευσης κατά την εκτέλεση του αντίστοιχου αλγόριθμου [79]). Η εργασία λοιπόν, που καλούνται τα δύο αυτοκινούμενα ρομπότ να πραγματοποιήσουν είναι να αναπτύξουν τις συνεργατικές εκείνες δεξιότητες οι οποίες θα τους επιτρέψουν να ωθήσουν συντονισμένα το κουτί που χειρίζονται προς την επιθυμητή θέση-στόχο. Το υπό εξέταση σύστημα συνθέτουν τέσσερις διακριτοί πράκτορες: Ρομπότ 1 - αριστερός τροχός, Ρομπότ 1 - δεξιός τροχός Ρομπότ 2 - αριστερός τροχός και Ρομπότ 2 - δεξιός τροχός. Συνεπώς οι συγκεκριμένοι πράκτορες πρέπει συνεργατικά να επιτύχουν τον χειρισμό εκείνον ο οποίος θα επιτρέψει στα ρομποτικά οχήματα να επενεργήσουν με τέτοιο τρόπο πάνω στο αντικείμενο ώστε να φτάσουν το κουτί στον επιθυμητό στόχο δίχως προγενέστερη γνώση ή προϋπάρχον συμπεριφορικό μοντέλο της σχετικής εργασίας. Το περιβάλλον προσομοίωσης το οποίο χρησιμοποιήθηκε κατά την πρώτη φάση υπολογιστικών δοκιμών είναι εκείνο που παρέχεται μέσω του λογισμικού *Cyberbotics Webots*^{TM}, όπου το συνθετικό περιβάλλον που σχεδιάστηκε για το σκοπό αυτό είναι αυτό το οποίο απεικονίζεται στο Σχήμα 4.44. Αναλυτικά οι παράμετροι προσομοίωσης των αυτοκινούμενων ρομπότ αποτυπώνονται στο σχετικό πίνακα 4.5.

Η τοπολογία της προτεινόμενης πολυπρακτορικής αρχιτεκτονικής, προσαρμοσμένη στη περίπτωση των αυτοκινούμενων ρομπότ, έχει περιγραφεί αναλυτικά στη Παράγραφο 2.8. Επιπλέον, ο μηχανισμός μάθησης και ο αντίστοιχος αλγόριθμος, μαζί με το σύνολο των μεταβλητών κατάστασης της τοπολογίας των αυτοκινούμενων ρομπότ, έχει παρουσιαστεί ήδη αναλυτικά στην Παράγραφο 3.6. Έχοντας λοιπόν μελετήσει τόσο την προτεινόμενη αρχιτεκτονική όσο και τον αντίστοιχο μηχανισμό μάθησης για τα αυτοκινούμενα ρομπότ, ας εξετάσουμε τα πειραματικά ευρήματα που προέκυψαν.

Το σύστημα επιτρέπει τη ρύθμιση των παραμέτρων θ για μια σειρά εποχών κάθε μία διάρκειας 450 μονάδων χρόνου. Οι τιμές των παραμέτρων μάθησης είναι: ρυθμός μάθησης $\alpha = 0.00012$, ο εκπτωτικός συντελεστής $\gamma = 0.999$ ενώ $\lambda = 0.40$ και ο συντελεστής βαθμιαίας μείωσης $s = 0.005$ (Decay Factor). Τα αποτελέσματα των προσομοιώσεων απεικονίζονται στα Σχήματα 4.45 και

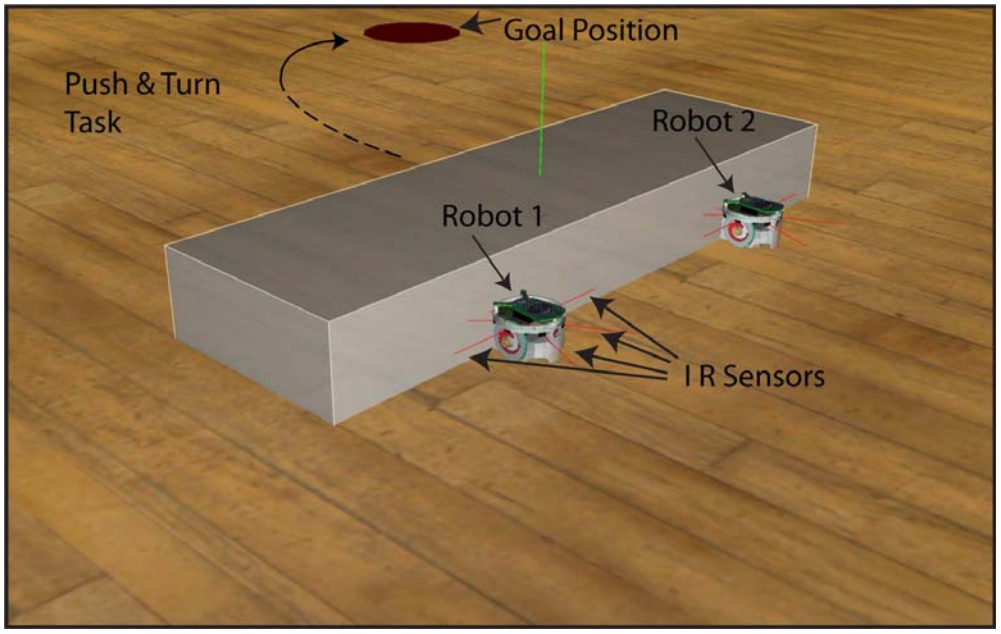
Πίνακας 4.5: Πειραματικές παράμετροι

	Robot 1	Robot 2	Object
Left/Right Wheel Radius	0.025 m	0.025 m	N/A
Left/Right Wheel Coulomb Friction	1	1	N/A
Distance Between the Wheels	0.09 m	0.09 m	N/A
Robot Radius	0.045 m	0.045 m	N/A
Robot Mass	0.5	0.5	N/A
Touch Sensor	Yes	Yes	N/A
Object Mass	N/A	N/A	0.5
Object Inertia	N/A	N/A	0.02417
Object Width	N/A	N/A	0.23 m
Object Length	N/A	N/A	0.82 m
Object Height	N/A	N/A	0.1 m
Object Coulomb Friction	N/A	N/A	1

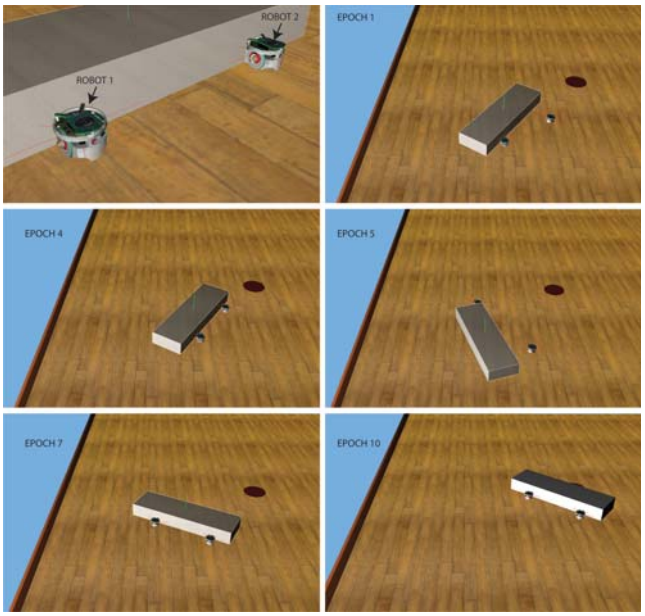
4.46. Όπως προκύπτει από το Σχήμα 4.46 (α) οι πράκτορες (τροχοί) επιτυγχάνουν με την πάροδο του χρόνου να αναπτύξουν τις δεξιότητες εκείνες που απαιτούνται έτσι ώστε τελικά, τα αυτοκινούμενα ρομποτικά οχήματα να ωθήσουν συνεργατικά το χειριζόμενο αντικείμενο προς τη θέση-στόχο.

Μέσα στις πέντε πρώτες εποχές οι πράκτορες καταφέρνουν να μειώσουν την απόσταση του αντικειμένου από τη θέση-στόχο δίχως κατάλληλο προσανατολισμό. Παρατηρούμε επίσης στο σχετικό Σχήμα 4.45 ότι κατά την αρχική περίοδο μάθησης τα αυτοκινούμενα ρομποτικά οχήματα αδυνατούν να διατηρήσουν συνεχή επαφή με το αντικείμενο. Η κατάσταση βελτιώνεται ουσιαστικά κατά τις επόμενες εποχές, καθώς φτάνοντας τις εποχές $7, 10, 20$ το αντικείμενο με επιτυχία οδηγείται στη θέση-στόχο με κατάλληλο προσανατολισμό. Το Σχήμα 4.46 (β) απεικονίζει τη βελτίωση του προσανατολισμού με τον οποίο το αντικείμενο προσεγγίζει το στόχο.

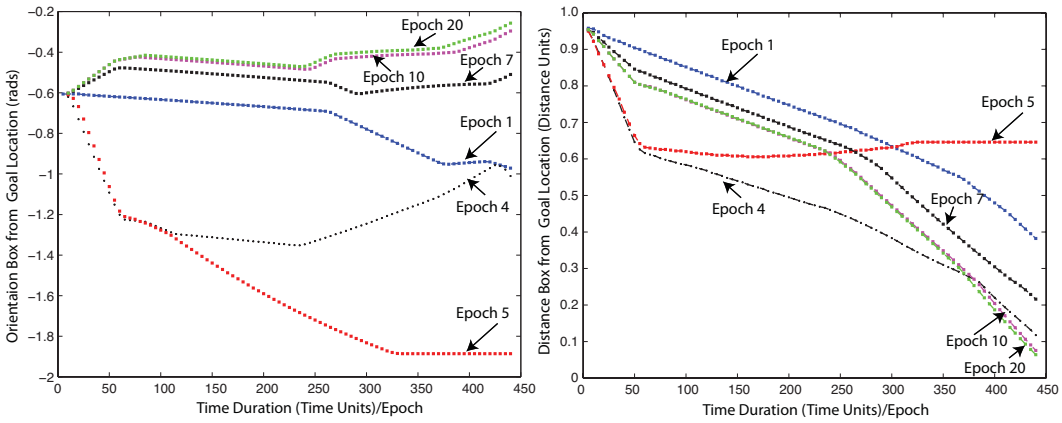
Με δεδομένο αυτές τις νέες απαιτήσεις οι οποίες εισάγονται από το συγκεκριμένο πείραμα, έχουμε προφανώς επαύξηση του προς επίτευξη στόχου για το πολυπρακτορικό σύστημα. Η επαύξηση αυτή συνεπάγεται προφανώς σχετική τροποποίηση της συνάρτησης ανταπόδοσης έτσι ώστε να είναι δυνατή η ορθή λειτουργία του μηχανισμού ενισχυτικής μάθησης. Η συνάρτηση ανταπόδοσης την οποία είδαμε και σε προηγούμενες παραγράφους και η οποία καθοδηγούσε το μηχανισμό μάθησης έτσι ώστε η κινηματική αλυσίδα να μπορεί να προσεγγίζει τη θέση-στόχο, καθώς και να επιτυγχάνεται στατική ισορροπία κατά την



Σχήμα 4.44: *Προσομοίωση δύο αυτοκινούμενων τύπου e-Puck που πραγματοποιούν εργασία "box-pushing" σε συγκεκριμένη θέση-στόχο*



Σχήμα 4.45: *Αποτελέσματα προσομοίωσης για διαφορετικές εποχές*



Σχήμα 4.46: *(α) Οι πράκτορες επιτυγχάνουν να στρέψουν το αντικείμενο προς τη θέση-στόχο. (β) Οι πράκτορες επιτυγχάνουν να μειώσουν την απόσταση του αντικειμένου από τη θέση-στόχο.*

πολυαρθρωτή ρομποτική λαβή, εξακολουθεί να ισχύει απλά επαυξάνεται κατάλληλα ώστε να ικανοποιεί τους περιορισμούς του συγκεκριμένου προβλήματος.

Πιο συγκεκριμένα, νέα τροποποιημένη συνάρτηση ανταπόδοσης συνδυάζει δεδομένα από τους αισθητήρες υπερύθρων (IR readings) των αυτοκινούμενων ρομπότ καθώς και δεδομένα σχετικά με την απόσταση και τον προσανατολισμό, προερχόμενα από το περιβάλλον προσομοίωσης. Συνεπώς, η νέα συνάρτηση ανταπόδοσης $R(t)$ για το χρονικό βήμα t, η οποία καθοδηγεί την τρέχουσα διαδικασία μάθησης, έχει την εξής μορφή:

$$R(t) = \begin{cases} R_1 - R_2 - R_3 - R_4 & \text{if } Condition_1 \\ 0.000001 & \text{else if } Condition_2 \\ R_1 - R_2 - R_3 & \text{else if } Condition_3 \\ -0.0015 & \text{else if } Condition_4 \\ -0.002 & \text{else } Condition_5 \end{cases} \tag{4.3}$$

όπου για τις επιμέρους συναρτήσεις $R_{1...4}$, έχουμε τα ακόλουθα:

$$\begin{cases} R_1 = c_1 \cdot \exp^{-r_1 \cdot Distance\ to\ Goal} \\ R_2 = c_2 \cdot \exp^{-r_2 \cdot IR\ Readings\ ePuck\ 1} \\ R_3 = c_3 \cdot \exp^{-r_3 \cdot IR\ Readings\ ePuck\ 2} \\ R_4 = c_4 \cdot Error\ in\ Orientation \end{cases} \tag{4.4}$$

με $c_1 \dots c_4$, και $r_1 \dots r_3$, σταθερές οι οποίες λαμβάνουν θετικές τιμές: $0 < c_1 \dots c_4 < 1$ και $0 < r_1 \dots r_3 < 1$. Παρατηρούμε ότι, η συνάρτηση R_4, σε αντίθεση με τις προηγούμενες συναρτήσεις R_1, R_2 και R_3, δεν περιλαμβάνει τον εκθετικό όρο. Ο λόγος είναι ότι το εύρος τιμών του σφάλματος προσανατολισμού, *Error in Orientation*, είναι μικρό και δεν επενεργεί σε σημαντικό

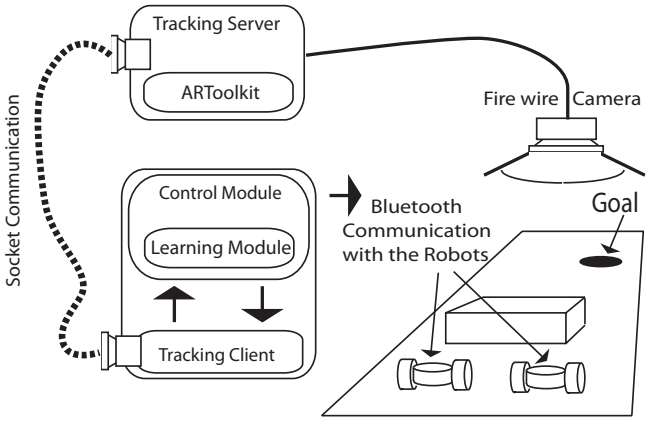
βαθμό. Όπως θα δούμε στη συνέχεια, δεν ισχύει το ίδιο για το εύρος τιμών της μεταβλητής *IR Reading*. Η επιλογή της εκθετικής συνάρτησης στο μηχανισμό ανταπόδοσης, όχι μόνο στην περίπτωση εδώ των αυτοκινούμενων οχημάτων, όπως περιγράφεται στη σχέση (4.4), αλλά και στις προηγούμενες περιπτώσεις των κινηματικών αλυσίδων που εξετάστηκαν ήδη, αποσκοπεί στο να πάρει το σύστημα άμεσα μηδενικές / αρνητικές ανταποδόσεις και με αυτό τον τρόπο να αντιδρά γρήγορα και πιο αποτελεσματικά στην τρέχουσα συμπεριφορά που ακολουθούν οι πράκτορες που το απαρτίζουν.

Στη συνέχεια, εξετάζουμε τις αντίστοιχες συνθήκες $Condition_{1...5}$ που ορίζονται ως ακολούθως:

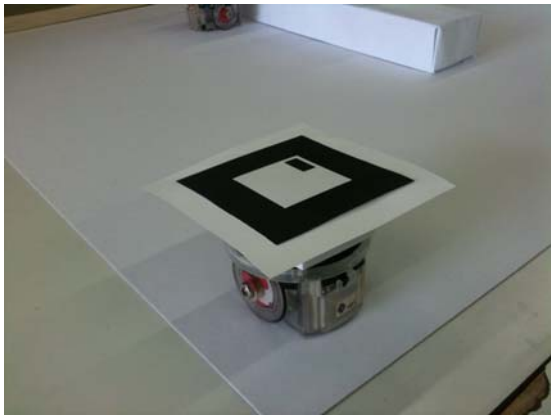
$$\begin{cases} Condition_1: & (\forall \text{ e-pucks}, contact == 1) \wedge (\Delta D < 0) \\ Condition_2: & (\forall \text{ e-pucks}, contact == 1) \wedge (\Delta D \geq 0) \\ Condition_3: & (\forall \text{ e-pucks } IR\ Readings == Within\ Range) \\ Condition_4: & (\exists \text{ e-puck } \ IR\ Readings == Within\ Range) \\ Condition_5: & (\forall \text{ e-pucks } IR\ Readings \ \ != Within\ Range) \end{cases} \tag{4.5}$$

όπου ΔD είναι ο ρυθμός μεταβολής της απόστασης του αντικειμένου από τη θέση-στόχο, *IR Reading*, είναι τα δεδομένα τα οποία καταγράφουν οι αισθήτηρες υπερύθρων των αυτοκινούμενων e-Puck, ενώ *Within Range* είναι το κατώφλι το οποίο διασφαλίζει μία ελάχιστη απόσταση των αυτοκινούμενων οχημάτων από το αντικείμενο έτσι ώστε τα δεδομένα να είναι αξιόπιστα. Το εύρος τιμών της μεταβλητής *IR Reading*, είναι $[0 \ldots 4000]$. Το κατώτατο όριο σηματοδοτεί ότι το ρομπότ είναι πολύ μακριά από το αντικείμενο ενώ στην αντίθετη περίπτωση, το ρομπότ είναι σε επαφή με το αντικείμενο. Στο πείραμα το οποίο πραγματοποιήσαμε υπάρχει ένα κατώφλι για τους αισθήτηρες υπερύθρων το οποίο είναι 800. Κάτω από αυτή την τιμή θεωρούμε ότι το ρομπότ δεν είναι κοντά στο αντικείμενο. Επιπλέον, οι τιμές των σταθερών που επιλέγησαν για τα πειράματα είναι οι εξής: $c_1 = 0.999$, $r_1 = 0.2$, $c_2 = 0.32$, $r_2 = 0.0015$, $c_3 = 0.32$, $r_3 = 0.0015$ και $c_4 = 0.1$. Τέλος, ορίζουμε το σφάλμα προσανατολισμού σε ένα εύρος τιμών: $0 \leq Error\ in\ Orientation \leq \pi/2$. Το σύστημα ρυθμίζει τις παραμέτρους θ για μια σειρά είκοσι εποχών κάθε μία διάρκειας 450 μονάδων χρόνου. Οι τιμές των παραμέτρων μάθησης είναι: ρυθμός μάθησης $\alpha = 0.00012$, ο εκπτωτικός συντελεστής $\gamma = 0.999$, ενώ $\lambda = 0.40$ και ο συντελεστής βαθμιαίας μείωσης $s = 0.005$ (Decay Factor).

Κατά την φάση μάθησης όλες οι μεταβλητές κατάστασης παρακολουθούνται μέσω του σχετικού περιβάλλοντος προσομοίωσης. Στην προσπάθεια εφαρμογής σε πραγματικά αυτοκινούμενα ρομπότ απαιτείται η δημιουργία μιας νέας αρχιτεκτονικής για τη διασύνδεση των μονάδων μάθησης και ελέγχου (Σχήμα 4.47). Βασικό στοιχείο σε αυτήν τη νέα αρχιτεκτονική ελέγχου είναι η υλοποίηση μιας



Σχήμα 4.47: *Σχηματική απεικόνιση της συνολικής αρχιτεκτονικής ελέγχου για την υλοποίηση σε πραγματικά αυτοκινούμενα οχήματα*



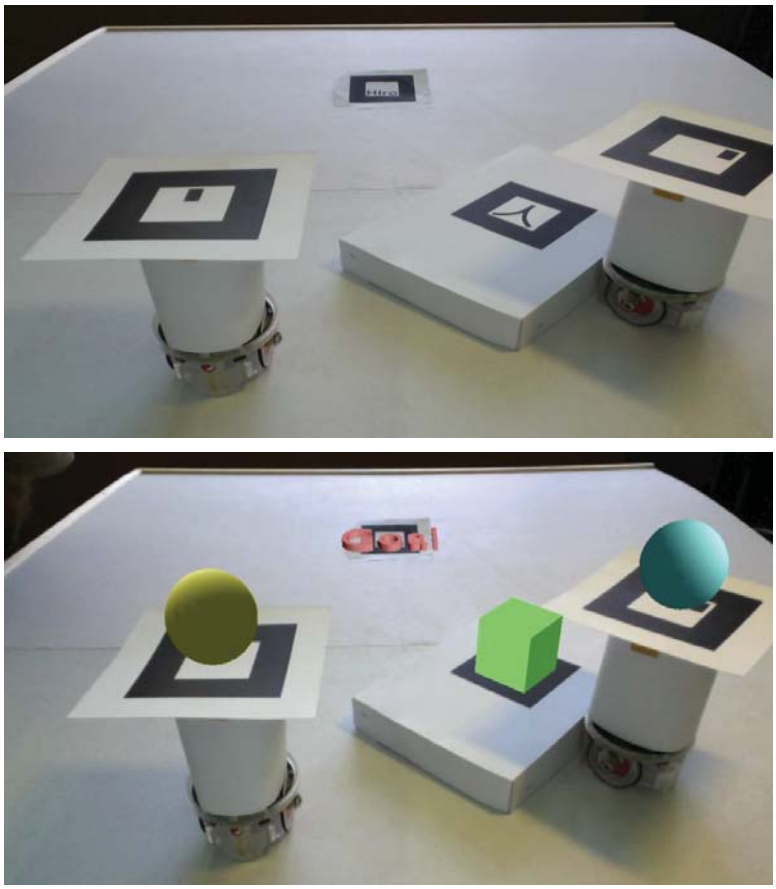
Σχήμα 4.48: *Η τοποθέτηση ενός marker σε ρομπότ τύπου e-Puck*

μονάδας οπτικής ιχνηλάτησης (Tracking Server) η οποία υλοποιείται με βάση το γνωστό λογισμικό υποστήριξης της ανάπτυξης εφαρμογών επαυξημένης πραγματικότητας (Augmented Reality - AR) ARToolkit [69]. Το τμήμα αυτό του λογισμικού θα παρέχει τις πληροφορίες εκείνες έτσι ώστε να είναι δυνατός ο υπολογισμός των παραμέτρων που απαιτούνται για τον προσδιορισμό της κατάστασης των πρακτόρων του συστήματος. Σε αυτό το σημείο σημειώνουμε ότι γίνεται μία κανονικοποίηση του πραγματικού χώρου εργασίας έτσι ώστε να είναι σε αντιστοιχία με το χώρο εργασίας της προσομοίωσης. Συνεχίζοντας, πρέπει να διευκρινίσουμε ότι η μονάδα οπτικής ιχνηλάτησης (Tracking Server) επιτυγχάνει να αναγνωρίζει τα ρομπότ, το αντικείμενο και τη θέση-στόχο μέσω συγκεκριμένων Markers (Σχήμα 4.48), που αντιστοιχούν σε κάθε ένα από τα παραπάνω αντικείμενα. Η παραπάνω αντιστοίχιση έχει γίνει εκ των προτέ-

ρων. Συνεπώς, μέσω κάμερας η οποία είναι συνδεμένη με τον Tracking Server, πραγματοποιείται επεξεργασία της εικόνας και εν συνεχεία υπολογίζεται η θέση και ο προσανατολισμός των αντικειμένων. Δεύτερο βασικό στοιχείο το οποίο χρειάστηκε να υλοποιηθεί στη νέα τοπολογία είναι ο Tracking Client. Το κομμάτι αυτό του λογισμικού ουσιαστικά αναλαμβάνει την επικοινωνία μεταξύ του Tracking Server και του λογισμικού του $Webots^{TM}$ όπου είναι υλοποιημένος και εκτελείται ο μηχανισμός μάθησης, ο οποίος χρειάζεται τις σχετικές παραμέτρους για τον προσδιορισμό της κατάστασης του κάθε πράκτορα. Σύμφωνα με τον μηχανισμό επιλογής δράσεων τον οποίο είδαμε σε προηγούμενη παράγραφο, το πολυπρακτορικό σύστημα επιλέγει δράσεις οι οποίες εν συνεχεία αποστέλλονται στους πραγματικούς πράκτορες - τροχούς μέσω ασύρματης επικοινωνίας τύπου bluetooth.

Εν συνεχεία, στο Σχήμα 4.49 απεικονίζεται η διάταξη του χώρου εργασίας με τους σχετικούς markers, τοποθετημένους στα δύο αυτοκινούμενα ρομπότ, στο αντικείμενο και στη θέση-στόχο, ενώ το Σχήμα 4.50 απεικονίζει την οπτική γωνία από την οποία θα παρακολουθείται η πειραματική διαδικασία. Βάζοντας σε λειτουργία όλα τα παραπάνω έχουμε την φυσική πειραματική διάταξη η οποία απεικονίζεται στο αμέσως επόμενο σχήμα 4.51. Στο επάνω αριστερό τμήμα της εικόνας είναι το παράθυρο της μονάδας οπτικής ιχνηλάτησης (Tracking Server) που καταγράφει τη θέση και τον προσανατολισμό των αντικειμένων. Στη συνέχεια, ο Tracking Client, ανοίγοντας ένα κανάλι επικοινωνίας τύπου Socket-Com στέλνει την πληροφορία στον ελεγκτή ο οποίος, συνθέτοντας την κατάσταση του συστήματος, στέλνει στους πράκτορες (τροχούς των αυτοκινούμενων οχημάτων) τις δράσεις που πρέπει να εκτελέσουν (γωνιακές ταχύτητες αναφοράς).

Στο Σχήμα 4.52 βλέπουμε μια σειρά στιγμιοτύπων τα οποία απεικονίζουν τους πράκτορες να επενεργούν στο αντικείμενο συνεργατικά και τελικά να καταφέρνουν να το ωθήσουν με επιτυχία στη θέση-στόχο, ενώ ο Tracking Server καταγράφει τη θέση και τον προσανατολισμό όλων των στοιχείων που βρίσκονται στο χώρο εργασίας. Όπως προκύπτει από τα πειραματικά ευρήματα του Σχήματος 4.52, το πολυπρακτορικό σύστημα κατάφερε να μάθει μία συνεργατική συμπεριφορά χειρισμού αντικειμένου (συνεργατική ώθηση του αντικειμένου προς τη θέση-στόχο) "off-line", ενώ στη συνέχεια επιτυγχάνει να εκτελέσει αυτή τη συμπεριφορά σε πραγματικό περιβάλλον μέσω των αυτοκινούμενων οχήματων τύπου e-Puck με επιτυχία. Κατά την εκτέλεση της εργασίας από τα πραγματικά ρομπότ, οι πράκτορες (τροχοί) επιλέγουν σε κάθε κατάσταση s στην οποία μεταβαίνουν, εκείνη τη δράση α με τη μεγαλύτερη αξία $max_{\alpha}Q(s,\alpha)$. Με τον τρόπο αυτό το πολυπρακτορικό σύστημα κάνει χρήση της γνώσης που απέκτησε κατά την περίοδο εκπαίδευσης που προηγήθηκε. Επιπλέον, στο Σχήμα 4.53 απεικονίζεται η κίνηση των ρομποτικών οχημάτων καθώς καταφέρνουν μέσω των δράσεων που επιλέγουν να ωθήσουν το αντικείμενο στη θέση-στόχο



Σχήμα 4.49: *Η συνολική πειραματική διάταξη, με τους αντίστοιχους markers τοποθετημένους στα ρομπότ και στο χειριζόμενο αντικείμενο, στις δύο διαφορετικές περιπτώσεις όπου α) ο μηχανισμός ιχνηλάτισης είναι ανενεργός και β) στην περίπτωση που αυτός ενεργοποιείται*

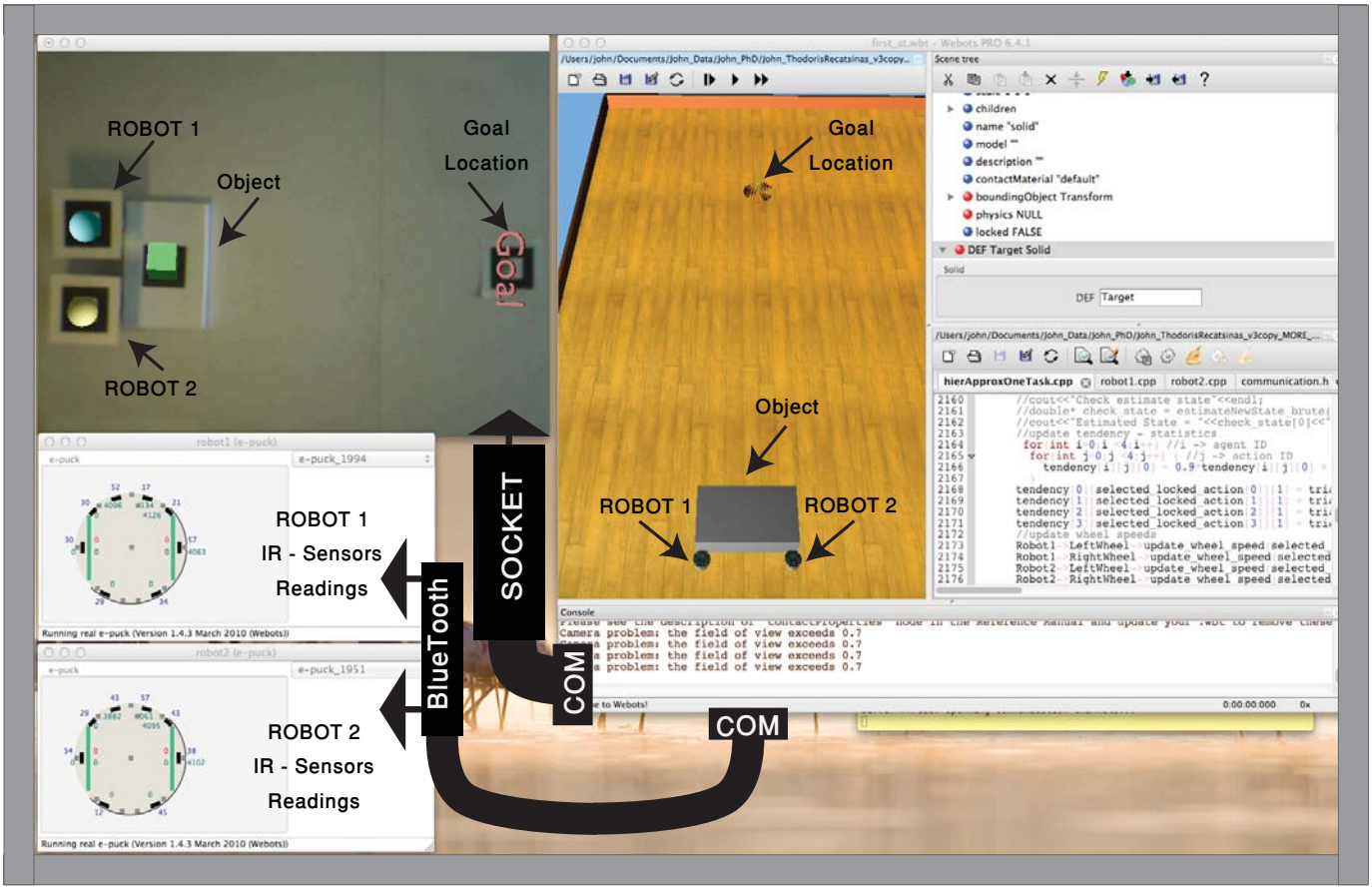
Σχήμα 4.50: *Η εικόνα της πειραματικής διάταξης όπως παρέχεται από την κάμερα η οποία βρίσκεται τοποθετημένη ακριβώς πάνω από τον χώρο εργασίας.*

έχοντας δώσει έναν διαφορετικό προσανατολισμό από αυτόν που είχε αρχικά.

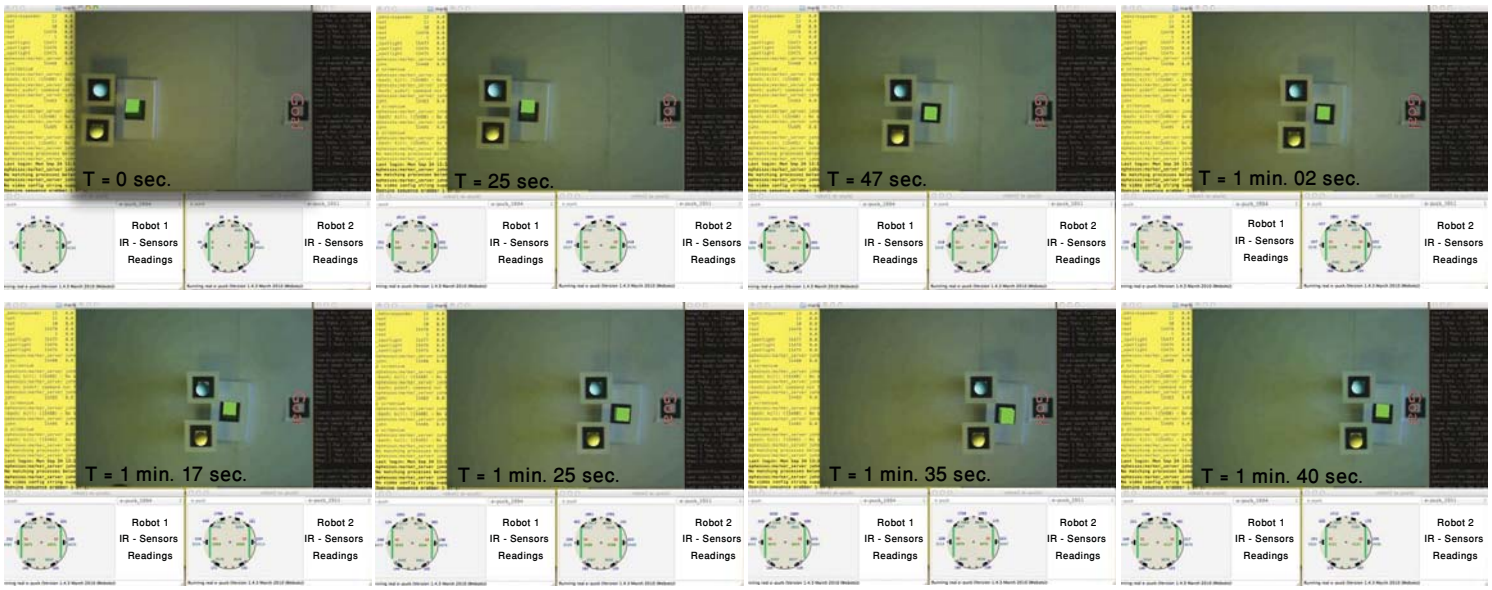
Συμπερασματικά, οι πειραματικές δοκιμές που εκτελέστηκαν κατέδειξαν την αποδοτικότητα της προτεινόμενης μεθόδου σε πραγματικό περιβάλλον εργασίας. Μέσω των πειραματικών αυτών αποτελεσμάτων, διαφαίνεται η επιτυχής προσαρμογή σε πραγματικό χώρο εργασίας των δεξιοτήτων ελέγχου οι οποίες έχουν αναπτυχθεί (off-line) από το πολυπρακτορικό σύστημα, μέσω της προτεινόμενης διαδικασίας ενισχυτικής ρομποτικής μάθησης.

4.6 Υπολογιστικό Κόστος

Στη συνέχεια θα πραγματοποιήσουμε μία ανάλυση ως προς το ζήτημα του υπολογιστικού κόστους της προτεινόμενης αρχιτεκτονικής (Computational Cost). Ο συνολικός χώρος καταστάσεων S του πολυπρακτορικού συστήματος συντίθεται από τους επιμέρους, τοπικούς (local), χώρους καταστάσεων S_1,S_2,... S_n των ανεξάρτητων πρακτόρων που απαρτίζουν την πολυπρακτορική τοπολογία που μελετάμε. Κάθε τοπικός χώρος κατάστασης ορίζεται από ένα αντίστοιχο ομοιογενές σύνολο μεταβλητών κατάστασης. Ας υποθέσουμε ότι το σύστημά μας αντιμετωπίζεται ως μία μονοπρακτορική τοπολογία, τότε θα ισχύουν τα ακόλουθα για το συνολικό χώρο κατάστασης S: $S = S_1 \times S_2 \times \cdots \times S_n$, γεγονός που σημαίνει ότι ο πληθάριθμος (cardinality) του χώρου κατάστασης κάνοντας χρήση μονοπρακτορικής τοπολογίας είναι: $|S| = |S_n|^n$. Όπως θα δούμε στη συνέχεια, υιοθετώντας την προτεινόμενη εμφωλευμένη, ιεραρχική πολυπρακτο-



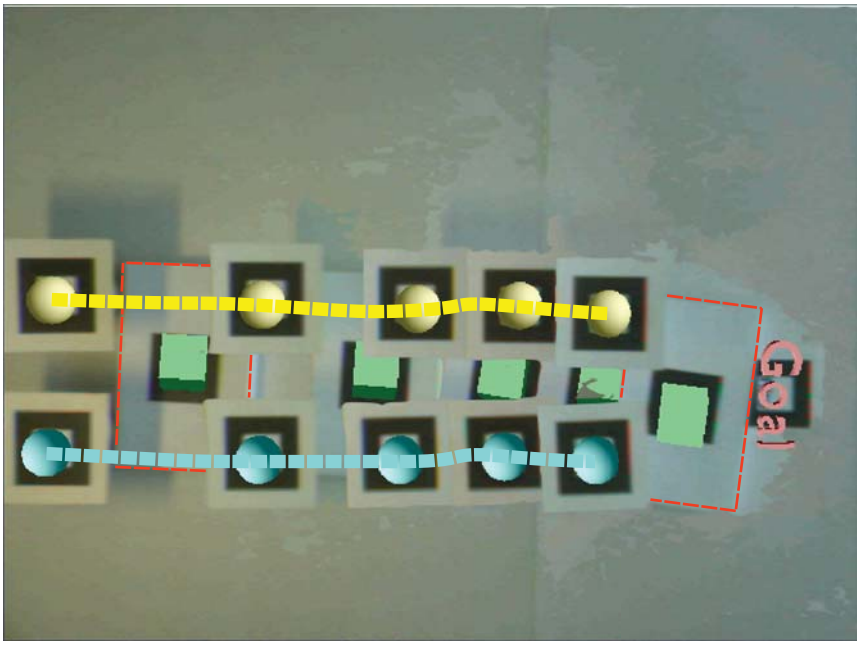
Σχήμα 4.51: *Συνολική πειραματική διάταξη για την υλοποίηση σε πραγματικά αυτοκινούμενα οχήματα*



Σχήμα 4.52: *Συνεργατικός χειρισμός αντικειμένου (εργασία τύπου "box-pushing") μέσω δύο πραγματικών αυτοκινούμενων ρομπότ τύπου e-Puck, με αποτέλεσμα την επιτυχή ώθηση του αντικειμένου προς τη θέση-στόχο. Επίβλεψη του πραγματικού χώρου εργασίας μέσω της μονάδας οπτικής ιχνηλάτησης*

ρική αρχιτεκτονική σε συνδυασμό με τον ομοιογενή ορισμό της κατάστασης κάθε πράκτορα, το υπολογιστικό κόστος του προβλήματος επανάληψης αξίας (value iteration problem) το οποίο επιλύουμε, μειώνεται σε σύγκριση με εκείνο της μονοπρακτορικής προσέγγισης.

Η συγκεκριμένη πολυπρακτορική αρχιτεκτονική ορίζεται από ένα σύνολο πρακτόρων οριζόμενων με την ίδια ομοιογενή εσωτερική δομή. Ουσιαστικά όλοι οι πράκτορες έχουν το ίδιο αριθμό μεταβλητών κατάστασης, με τον οποίο καταφέρνουν να προσδιορίσουν με μοναδικό τρόπο την κατάσταση στην οποία βρίσκονται για όλες τις πιθανές διατάξεις. Αυτό ουσιαστικά σημαίνει ότι ο πληθάριθμος (cardinality) $|\cdot|$ για κάθε τοπικό χώρο κατάστασης είναι ο ίδιος: $|S_1| = |S_2| = |S_3| = \cdots = |S_i| = |S|$ για κάθε πράκτορα i του συστήματος. Σύμφωνα με την προτεινόμενη εμφωλευμένη ιεραρχική αρχιτεκτονική, για να σχηματιστεί η επιθυμητή συλλογική δράση (Joint Action), κάθε πράκτορας είναι σε θέση να παρακολουθεί μόνο εκείνους τους πράκτορες οι οποίοι είναι στην αμέσως επόμενη βαθμίδα σε σχέση με αυτόν. Τέλος, ο πληθάριθμος του χώρου συλλογικών δράσεων (joint action space) είναι $|A|^i$, όπου i αντιστοιχεί στον αριθμό των πρακτόρων που συμμετέχουν στη συλλογική δράση στο συγκεκριμένο επίπεδο της ιεραρχίας, και $|A|$ αντιστοιχεί στον αριθμό των διακριτών



Σχήμα 4.53: *Ενδεικτικά ίχνη κίνησης ρομποτικών οχημάτων και αντικειμένου*

δράσεων που μπορεί να πραγματοποιήσει κάθε πράκτορας. Θεωρώντας ότι ο χώρος καταστάσεων είναι πεπερασμένος, ο αριθμός των ζευγών κατάστασης - δράσης που ανανεώνονται σε κάθε επανάληψη είναι: $|S| \cdot |A|^i$. Για να πραγματοποιηθεί η ανανέωση της αξίας ενος συγκεκριμένου ζεύγους κατάστασης - δράσης, απαιτείται απαρίθμηση όλων των στοιχείων του $|A|^i$. Συνεπώς, το κόστος σε κάθε επανάληψη είναι $|S| \cdot |A|^i \cdot |A|^i$ ή $|S| \cdot (|A|^i)^2$ ή $|S| \cdot |A|^{2i}$. Εάν υποθέσουμε ότι ο αλγόριθμος απαιτεί L επαναλήψεις και ότι το σύνολο των πρακτόρων είναι n, το συνολικό υπολογιστικό κόστος υπολογίζεται ως εξής:

$$L \cdot \sum_{i=1}^{n} \left\{ |S| \cdot (|A|^i)^2 \right\} =$$

$$= L \cdot |S| \sum_{i=1}^{n} (|A|^2)^i = L \cdot |S| \cdot \frac{(|A|^2)^{n+1} - |A|^2}{|A|^2 - 1} = L \cdot |S| \cdot \frac{|A|^{2(n+1)} - |A|^2}{|A|^2 - 1}$$

$$= L \cdot |S| \, |A|^{2n} \, \frac{|A|^2 - \frac{1}{|A|^{2(n-1)}}}{|A|^2 - 1} =$$

$$\simeq L \cdot |S| \cdot |A|^{2n} \cdot K \tag{4.6}$$

όπου για μεγάλο αριθμό πρακτόρων, όταν δηλαδή το n είναι μεγάλο, μπορούμε να υποθέσουμε ότι:

$$K \cong \frac{|A|^2}{|A|^2 - 1} \tag{4.7}$$

Η σύγκριση τώρα του παραπάνω κόστους με την περίπτωση εκείνη της μονοπρακτορικής τοπολογίας είναι εξαιρετικά απλή. Στη μονοπρακτορική αρχιτεκτονική, οι μεταβλητές κατάστασης, αντί να είναι κατανεμημένες μεταξύ των διαφόρων πρακτόρων, συγκεντρώνονται σε έναν και μόνο πράκτορα, γεγονός το οποίο οδηγεί τον πληθάριθμο του χώρου κατάστασης του συγκεκριμένου πράκτορα σε εκθετική αύξηση ως προς n (εν προκειμένω, ως προς τον αριθμό των βαθμών ελευθερίας του ρομποτικού συστήματος). Επομένως, το συνολικό υπολογιστικό κόστος σε αυτή την μονοπρακτορική προσέγγιση θα έχει ως εξής:

$$L \cdot |S|^n \cdot |A|^{2n} \tag{4.8}$$

Συγκρίνοντας τη σχέση (4.6) με τη σχέση (4.8) είναι ξεκάθαρο ότι, καθώς ο αριθμός n μεγαλώνει (στην μονοπρακτορική περίπτωση το n αναφέρεται ουσιαστικά στον αριθμό των βαθμών ελευθερίας του συστήματος), το υπολογιστικό κόστος στη μονοπρακτορική αρχιτεκτονική αυξάνεται εκθετικά, δεδομένου ότι ο πληθάριθμος του χώρου κατάστασης στη σχέση (4.8) υψώνεται στη δύναμη n. Στη συνέχεια, ας δούμε ένα παράδειγμα για την καλύτερη κατανόηση των όσων ήδη αναφέρθηκαν. Στην περίπτωση της ανοικτής κινηματικής αλυσίδας την οποία και είδαμε στη παράγραφο 4.3.3, το προτεινόμενο πολυπρακτορικό σύστημα αποτελείται από επτά εμφωλευμένους πράκτορες, συνεπώς $n = 7$. Ο χώρος κατάστασης κάθε πράκτορα όπως ήδη έχει αναφερθεί στο προηγούμενο κεφάλαιο αποτελείται από 6 μεταβλήτες κατάστασης. Ας υποθέσουμε ότι η κάθε μεταβλητή ασαφοποιείται με 8 συναρτήσεις συμμετοχής. Ο πληθάριθμος του χώρου κατάστασης για κάθε πράκτορα, σε αυτή την περίπτωση θα είναι: $|S| = 8^6$. Επιπλέον, όπως είδαμε επίσης στο προηγούμενο κεφάλαιο, κάθε πράκτορας έχει να επιλέξει μεταξύ τριών δράσεων, άρα: $|A| = 3$. Η κάθε εποχή εκπαίδευσης του συστήματος, στη διάρκεια της οποίας κάθε πράκτορας λειτουργεί, υποθέτουμε ότι έχει διάρκεια 1500 επαναλήψεις, άρα: $L = 1500$. Με βάση όλα τα παραπάνω στοιχεία, το υπολογιστικό κόστος της πολυπρακτορικής αρχιτεκτονικής είναι: $1500 \cdot 8^6 \cdot (3^2)^7 \cdot \frac{3^2}{3^2-1} = 2.1158 \times 10^{15}$. Στην αντίστοιχη περίπτωση της μονοπρακτορικής προσέγγισης το κόστος θα ήταν: 6.8663×10^{47}. Συνεπώς, καθώς ο αριθμός των πρακτόρων αυξάνεται, το υπολογιστικό όφελος το οποίο προκύπτει από την προτεινόμενη αρχιτεκτονική καθίσταται σημαντικό.

Κεφάλαιο 5

Συμπεράσματα - Μελλοντικές Κατευθύνσεις Έρευνας

5.1 Συμπεράσματα

Ολοκληρώνοντας την παρούσα διατριβή, είναι σκόπιμο επιγραμματικά να δοθούν τα συμπεράσματα τα οποία προέκυψαν.

1. Στο πλαίσιο της παρούσας διατριβής πραγματοποιήθηκε ο σχεδιασμός και η ανάπτυξη μιας ευέλικτης, κλιμακωτής, πολυπρακτορικής αρχιτεκτονικής ελέγχου, η οποία εφαρμόστηκε σε προβλήματα επιδέξιου ρομποτικού χειρισμού. Ουσιαστικά, υλοποιήθηκε μια κατανεμημένη πολυπρακτορική αρχιτεκτονική η οποία είναι ιεραρχική, επιδεικνύει αρθρωτή δομή, ενώ το πεδίο εφαρμογής της καλύπτει τον τομέα του επιδέξιου ρομποτικού χειρισμού και όχι μόνο. Η αρχιτεκτονική αυτή όπως διαφαίνεται από τα αποτελέσματα, λόγο των ιδιαίτερων χαρακτηριστικών πού ήδη αναφέραμε, δημιουργεί τις προϋποθέσεις εκείνες για να εφαρμοστεί και σε συνεργατικά αυτοκινούμενα ρομπότ, δίχως να απαιτεί προσαρμογές ή τροποποιήσεις, ενώ διασφαλίζει την επίτευξη των κατά περίπτωση στόχων.

2. Αναπτύχθηκε σχετική μέθοδος ενισχυτικής μάθησης εφαρμόσιμη σε συνεχή χώρο-κατάστασης. Η ανάγκη λειτουργίας μέσα σε δυναμικά μεταβαλλόμενο περιβάλλον, οδηγεί στην ανάγκη ύπαρξης μηχανισμού αυτοανάπτυξης και αυτο-οργάνωσης του ρομποτικού συστήματος. Την ανάγκη αυτή λοιπόν προσπαθήσαμε να καλύψουμε με το μηχανισμό αναπτυξιακής μάθησης τον οποίο και προτείνουμε στο πλαίσιο της παρούσας διατριβής.

3. Σχεδιάστηκε ένα σχήμα υβριδικού ρομποτικού ελέγχου. Ο συνδυασμός κλασσικής μεθοδολογίας ελέγχου, με υπολογιστικά μοντέλα τεχνητής νοημοσύνης, δημιούργησαν ένα υβριδικό σχήμα, το οποίο και χρησιμοποιήθηκε για τον έλεγχο των συγκεκριμένων πειραματικών διατάξεων (ανοικτή κινηματική αλυσίδα τεσσάρων και επτά συνδέσμων, πολυαρθρωτή ρομποτική λαβή καθώς και αυτοκινούμενα οχημάτα).

4. Πραγματοποιήθηκε ανάλυση σχετικών παραδειγμάτων εφαρμογής επιδέξιου ρομποτικού χειρισμού τα οποία ενσωμάτωσαν την παραπάνω τοπολογία. Πιο συγκεκριμένα, μέσω των διατάξεων ενός επίπεδου χειριστή τεσσάρων συνδέσμων με τέσσερις βαθμούς ελευθερίας, ενός επίπεδου χειριστή επτά συνδέσμων με επτά βαθμούς ελευθερίας καθώς και μιας πολυαρθρωτής ρομποτικής λαβής, αξιολογήθηκε τόσο η αρχιτεκτονική

όσο και ο σχετικός μηχανισμός μάθησης. Από τα αποτελέσματα των πειραμάτων φαίνεται ότι, το σύνολο των λύσεων οι οποίες προέκυψαν, είναι αρκετά κοντά στις βέλτιστες και αποτελούν μια θετική βάση για την περαιτέρω διερεύνηση του προτεινόμενου πλαισίου και σε άλλες τοπολογίες. Πραγματοποιήθηκαν σειρά πειραμάτων για την αξιολόγηση της δυνατότητας γενίκευσης γνώσης καθώς και της ευρωστίας που παρουσιάζει η προτεινόμενη αρχιτεκτονική σε απρόβλεπτα σφάλματα και αστοχίες των πρακτόρων που συνθέτουν το σύστημα.

5. Πραγματοποιήθηκαν επίσης πραγματικές πειραματικές δοκιμές σε συνεργαζόμενα αυτοκινούμενα ρομπότ τύπου e-Puck, με την εκτέλεση εργασίας συνεργατικού χειρισμού αντικειμένου (τύπου "box-pushing"). Τα πειραματικά αποτελέσματα κατέδειξαν την απόδοση της προτεινόμενης πολυπρακτορικής αρχιτεκτονικής στον έλεγχο του ρομποτικού συστήματος.

6. Η προτεινόμενη πολυπρακτορική τοπολογία σε συνδιασμό με το μηχανισμό αναπτυξιακής μάθησης επιτρέπει στο ρομποτικό σύστημα την επιλογή βέλτιστων συνεργατικών δράσεων χωρίς την ύπαρξη γενικευμένου μοντέλου εργασίας. Ο κάθε πράκτορας κάνοντας χρήση μόνο τοπικής πληροφορίας (αντιστοιχίσεις καταστάσεων-δράσεων) σχηματίζει (τοπικά) εικόνα για τη συνολική κατάσταση του συστήματος καθώς επίσης και για την εξέλιξη της εργασίας.

7. Υιοθετώντας την προτεινόμενη εμφωλευμένη, ιεραρχική πολυπρακτορική αρχιτεκτονική σε συνδυασμό με τον ομοιογενή ορισμό της κατάστασης κάθε πράκτορα, το υπολογιστικό κόστος του προβλήματος επανάληψης αξίας (value iteration problem) το οποίο επιλύουμε, μειώνεται καθώς κατανέμεται το υπολογιστικό βάρος της μάθησης στο σύνολο των πρακτόρων που απαρτίζουν το σύστημα.

8. Η προτεινόμενη κατανεμημένη πολυπρακτορική αρχιτεκτονική, δεδομένου ότι δεν λειτουργεί βάσει-μοντέλου, μπορεί δυνητικά να επεκτείνεται με τρόπο φυσικό σε πιο σύνθετες τοπολογίες και διατάξεις με πλεονάζοντες βαθμούς ελευθερίας, καθώς και σε κινήσεις που περιλαμβάνουν περιορισμούς. Η προοπτική αυτής της επεκτασιμότητας σε προβλήματα, όπου οι προσεγγίσεις βάσει-μοντέλου περιπλέκονται, αποτελεί σημαντικό χαρακτηριστικό της προτεινόμενης αρχιτεκτονικής. Επιπλέον, η μη ύπαρξη

μοντέλου επιτρέπει την προσαρμοστικότητα του πολυπρακτορικού συστήματος σε αλλαγές καθώς και να ανταπεξέλθει σε απρόβλεπτες και σύνθετες αστοχίες, χωρίς επαναπρογραμματισμό ή επανασχεδίαση.

5.2 Συζήτηση

Έχοντας ήδη πραγματοποιήσει μια εκτενή βιβλιογραφική επισκόπηση στο εισαγωγικό κεφάλαιο της παρούσας διατριβής, έχοντας ολοκληρώσει την ανάλυση της προτεινόμενης αρχιτεκτονικής, και τέλος έχοντας παρουσιάσει τα συμπεράσματα με βάση τα πειραματικά ευρήματα εφαρμογής, συνοψίζουμε στη συνέχεια τα κύρια σημεία της διατριβής σε σχέση με άλλες ερευνητικές προσπάθειες υπογραμμίζοντας τα σημεία που συνθέτουν την κύρια επιστημονική της συμβολή και πρωτοτυπία.

Η μάθηση μέσω επίδειξης (Learning from Demonstration - LfD) αποτελεί μια τυπική μεθοδολογία η οποία επιλύει παρόμοια προβλήματα με εκείνα που προσεγγίστηκαν στο πλαίσιο της παρούσας διατριβής. Μολονότι αυτές οι μεθοδολογίες σαφώς αποτελούν πολύ σημαντικές, state-of-the-art προσεγγίσεις στο συγκεκριμένο πεδίο (χρήσης δηλαδή της επίδειξης ως μέσω δημιουργίας ενος αρχικού υπόβαθρου γνώσης), διαφέρουν από την προσέγγιση την οποία εμείς προτείνουμε στο γεγονός ότι είναι καθαρά κεντρικά δομημένες (centralized), μονοπρακτορικές προσεγγίσεις. Η ιδιαίτερη συνεισφορά και καινοτομία της παρούσας διατριβής, σε σχέση με τις συγκεκριμένες μεθόδους που λειτουργούν στη βάση της μάθησης μέσω επίδειξης - LfD είναι ότι καταφέραμε να αποδείξουμε πώς ένα πολυπρακτορικό σύστημα, πώς ένα ρομποτικό πολυπρακτορικό σύστημα μπορεί, με ένα συντονισμένο τρόπο, να αναπτύξει σύνθετες συμπεριφορές (όπως για παράδειγμα, προσέγγιση στόχου, κίνηση με περιορισμούς, ρομποτική λαβή, κλπ.), ενώ ο κάθε πράκτορας που το απαρτίζει λειτουργεί στο δικό του τοπικό επίπεδο, συμμετέχοντας σε ένα συνεργατικό παίγνιο με τους υπόλοιπους πράκτορες (που βρίσκονται σε άλλα επίπεδα της ιεραρχίας). Παράλληλα, το ρομποτικό πολυπρακτορικό σύστημα μπορεί να πραγματοποιεί σε σημαντικό βαθμό γενίκευση γνώσης, επιδεικνύοντας ταυτόχρονα ιδιότητες προσαρμοστικότητας και ευρωστίας σε απρόβλεπτες (μη μοντελοποιημένες) αστοχίες, χωρίς να έχει προηγηθεί αρχικό στάδιο μάθησης "bootstrap" μέσω ανθρώπινης επίδειξης. Αναπαριστώντας τα διαφορετικά προβλήματα ρομποτικής μάθησης τα οποία παρουσιάζουμε στην παρούσα διατριβή ως συνεργατικά παίγνια, είναι μία προσέγγιση η οποία δεν έχει εφαρμοστεί σε κάποια προγενέστερη συναφή με το αντικείμενο εργασία.

Η παρούσα διατριβή συνδυάζει ερευνητική προσπάθεια προερχόμενη από τα πεδία του ελέγχου ρομποτικών κινηματικών αλυσίδων και της ενισχυτικής μάθησης με μεθοδολογίες που βασίζονται σε πολυπρακτορικά συστήματα. Στη συνέχεια θα δούμε συμπερασματικά τα σημεία στα οποία η προσέγγιση η οποία ακολουθήθηκε στην παρούσα διατριβή παρουσιάζει κοινά χαρακτηριστικά αλλά και ουσιαστικές διαφοροποιήσεις συγκριτικά με αντιπροσωπευτικές εργασίες προερχόμενες από τα δύο ανωτέρω ερευνητικά πεδία, εστιάζοντας σε σημεία καινοτομίας, και εν δυνάμει υπεροχής, της προτεινόμενης προσέγγισης.

5.2.1 Πολυπρακτορική Ενισχυτική Μάθηση

Στην σχετική εργασία [25] των Claus και Boutilier παρουσιάζεται ο τρόπος με τον οποίο ο μηχανισμός μάθησης Q-learning μπορεί να εφαρμοστεί σε πολυπρακτορικά συστήματα κάνοντας χρήση της προσέγγισης *Nash Equilibrium* για την περιγραφή της βέλτιστης συλλογικής δράσης. Η εμφωλευμένη αρχιτεκτονική που προτείνουμε βασίζεται στην ίδια διατύπωση, ενώ κάνοντας την υπόθεση ότι το πολυπρακτορικό σύστημα αποτελείται από ομοιογενείς πράκτορες που έχουν τον ίδιο χώρο κατάστασης, επιτυγχάνει να μειώσει την πολυπλοκότητα του σχετικού προβλήματος μάθησης. Στην εργασία [51] από τον Guestrin et al. παρουσιάζεται μια διαφορετική προσέγγιση πολυπρακτορικής ενισχυτικής μάθησης, όπου προτείνεται μία συγκεκριμένη μοντελοποίηση ως προς τον τρόπο με τον οποίο γίνεται η αναπαράσταση της πολυπρακτορικής τοπολογίας. Οι απαιτήσεις ως προς τον συντονισμό του συστήματος καταγράφονται μέσω ενός γράφου όπου οι κόμβοι του γράφου αντιστοιχούν στους πράκτορες του συστήματος ενώ οι συσχετισμοί για τον συντονισμό μεταξύ των πρακτόρων απεικονίζονται μέσω των ακμών που συνδέουν τους κόμβους του σχετικού γράφου. Η βέλτιστη συλλογική δράση του συστήματος προκύπτει μέσω αλγοριθμικής επεξεργασίας του γράφου. Η χρήση αυτής της προσέγγισης προϋποθέτει προφανώς την εκ των προτέρων γνώση της συνολικής τοπολογίας των πρακτόρων καθώς και των μεταξύ τους συσχετισμών. Σε αντίθεση, η ιεραρχική αρχιτεκτονική που προτείνεται στην παρούσα διατριβή δεν απαιτεί να είναι γνωστή εκ των προτέρων η συνολική τοπολογία των πρακτόρων. Παρέχει έναν εναλλακτικό μηχανισμό συντονισμού μεταξύ των πρακτόρων μέσω μιας ιεραρχικής αλυσίδας διασύνδεσης.

5.2.2 Πολυπρακτορική Κινηματική Αλυσίδα

Μεγάλο μέρος της ερευνητικής εργασίας στο πεδίο του σχεδιασμού ρομποτικού χειρισμού (Manipulation Planning) και κινηματικών αλυσίδων, καλύπτε-

ται από προσεγγίσεις που κάνουν χρήση αλγορίθμων τύπου *Randomized Path Planning* όπως *Probabilistic Roadmap Planner* [70] και *Rapidly Exploring Random Trees - RRTs* [80]. Πιο συγκεκριμένα, στην εργασία [137] μέσω των παραπάνω μηχανισμών, δημιουργείται ένα πλήθος πιθανών διατάξεων, ξεκινώντας από τυχαία παραγόμενες διατάξεις των κινηματικών αλυσίδων. Εν συνεχεία, πραγματοποιείται σε ένα δεύτερο επίπεδο ένα φιλτράρισμα κατά το οποίο, από τις διατάξεις που προκύπτουν απορρίπτονται εκείνες που οδηγούν σε σύγκρουση. Τέλος, αναγνωρίζονται γειτονικές διατάξεις και μέσω ενός επαναληπτικού μηχανισμού αναζήτησης επιτυγχάνεται να βρεθεί μια ακολουθία ενδιάμεσων διατάξεων που θα επιτρέπει την ασφαλή μετάβαση του ρομποτικού χειριστή στον επιθυμητό στόχο, χωρίς να γίνεται χρήση κάποιου συγκεκριμένου μοντέλου. Αδιαμφισβήτητα, αυτές οι προσεγγίσεις αποτελούν state-of-the-art προσεγγίσεις ως προς τη χρήση *Randomized Planners* για την υλοποίηση αλγορίθμων σχεδίασης εργασιών ρομποτικού χειρισμού (*Planning Tasks*). Όμως, σε όλες αυτές τις προσεγγίσεις υπάρχει κεντρικός έλεγχος και γνώση, για το σύνολο της τοπολογίας της κινηματικής αλυσίδας (ανοικτής ή κλειστής). Επιπλέον, οι μεθοδολογίες αυτές δεν είναι κατανεμημένες, ενώ η αρχιτεκτονική τους είναι καθαρά μονοπρακτορική. Επίσης, αξίζει να σημειωθεί ότι όλα τα εμπόδια είναι στατικά ενώ ο χώρος εργασίας είναι πλήρως παρατηρήσιμος (fully observable). Τέλος, θέματα απρόβλεπτης αστοχίας σε αρθρώσεις της κινηματικής αλυσίδας ή γενικότερα μεταβολές σε δομικά στοιχεία της τοπολογίας, δεν θίγονται.

Άλλη μια ενδιαφέρουσα εργασία παρουσιάζεται από τους Boularias και Peters [17], όπου επιτυγχάνουν την εκμάθηση εφαρμογής ρομποτικής λαβής σε αντικείμενα, κάνοντας χρήση αλγορίθμων βασισμένων σε γράφους. Οι περιορισμοί που επιβάλει η προτεινόμενη μεθοδολογία είναι οι ακόλουθοι: α) η ανάγκη να προηγηθεί στάδιο ανθρώπινης επίδειξης, β) η ύπαρξη δομημένης αναπαράστασης (γράφος), δηλαδή ενός μοντέλου το οποίο θα περιλαμβάνει όλες τις πιθανές καταστάσεις του συστήματος καθώς και τους συσχετισμούς αυτών των καταστάσεων με δράσεις και γ) ο γράφος θα πρέπει να περιλαμβάνει πληροφορίες σχετικά με ποιές καταστάσεις έχουν παρόμοιες βέλτιστες δράσεις. Η προσέγγιση που προτάθηκε και υλοποιήθηκε στην παρούσα διατριβή δεν απαιτεί το σύνολο αυτής της πληροφορίας, δεδομένου ότι αυτή η αντιστοίχιση καταστάσεων σε βέλτιστες δράσεις είναι αποτέλεσμα που προκύπτει δυναμικά και αυτόνομα για κάθε πράκτορα του συστήματος, μέσω μιας αυτόματης διαδικασίας δημιουργίας και εξέλιξης γνώσης. Στην παρούσα διατριβή προτείνουμε μια συνεργατική πολυπρακτορική διαδικασία εξερεύνησης του στόχου (όποιος και να είναι αυτός κάθε φορά). Αυτή η προσέγγιση επιτρέπει, όπως είναι προφανές, την διαχείριση απρόβλεπτων αποτυχιών ενώ παρέχει την ικανότητα αυτο-οργάνωσης του συστήματος για την ανάκαμψη από σφάλματα, μέσω των ανταποδόσεων που εισπράττει από το περιβάλλον. Τέλος, η προσέγγισής μας δεν απαιτεί πλήρως

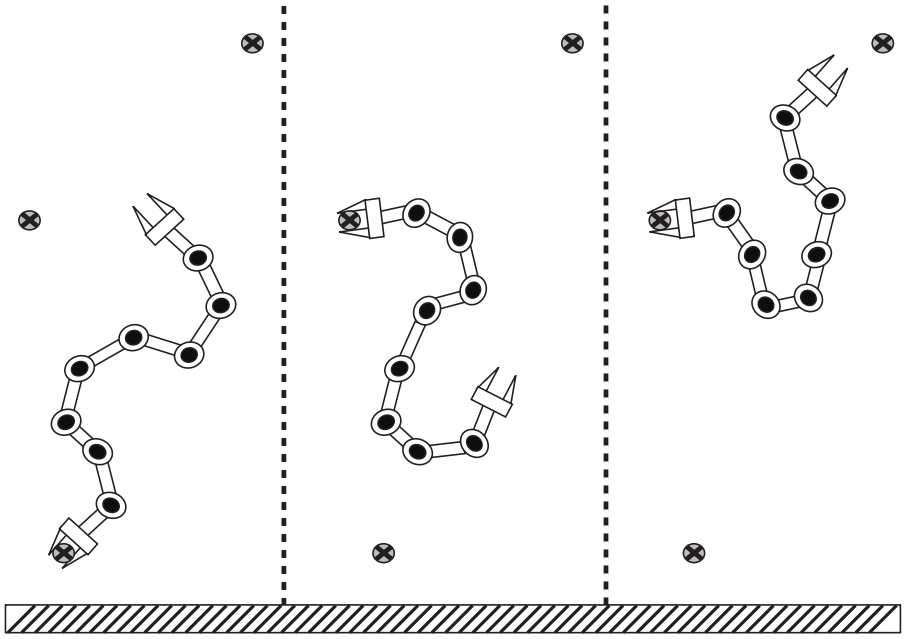
παρατηρήσιμο χώρο εργασία ενώ δεν περιορίζει την εφαρμοσιμότητά της σε στατικό περιβάλλον.

Στη συνέχεια, η τελευταία παράγραφος αυτής της διατριβής ανακεφαλαιώνει τα γενικά ευρήματα που παρουσιάστηκαν, συμπεριλαμβανομένων και των υφιστάμενων αδυναμιών του προτεινόμενου πλαισίου, ενώ παράλληλα προτείνει μελλοντικούς δρόμους για την επέκταση της συγκεκριμένης έρευνας, καθώς και περαιτέρω θέματα τα οποία χρήζουν έρευνας.

5.3 Περιορισμοί & Μελλοντικές Προεκτάσεις

Συνδυάζοντας μεθόδους Ενισχυτικής Μάθησης και κλασσικές προσεγγίσεις ελέγχου, στα πλαίσια ιεραρχικής πολυπρακτορικής αρχιτεκτονικής και ενός ασαποποιημένου χώρου καταστάσεων, προκύπτει ένα υβριδικό μοντέλο αναπτυξιακού ρομποτικού ελέγχου. Η μεθοδολογία η οποία προτάθηκε και υλοποιήθηκε στην παρούσα διατριβή δημιουργεί έναν συνεργατικό μηχανισμό αναζήτησης λύσεων ο οποίος αποδείχθηκε κατάλληλος για την επίλυση προβλημάτων επιδέξιου ρομποτικού ελέγχου. Το βασικό πλεονέκτημα της προτεινόμενης μεθοδολογίας είναι ότι δεν απαιτεί μοντέλο της προς επίτευξη εργασίας (για παράδειγμα στις πειραματικές διατάξεις που παρουσιάσαμε δεν υπάρχει αντίστροφο κινηματικό μοντέλο ή απλά μοντέλο κινηματικής λαβής). Επιπλέον, η προτεινόμενη πολυπρακτορική αρχιτεκτονική, λόγω της ιδιαίτερης ομοιογένειας που παρουσιάζει ως προς τα στοιχεία που τη συνθέτουν και της ιεραρχικής δομής της, διευκολύνει την ανάπτυξη μηχανισμών μάθησης για τον έλεγχο πιο σύνθετων ρομποτικών διατάξεων όπως αυτές που παρουσιάζονται στα Σχήματα 5.1 και 5.2.

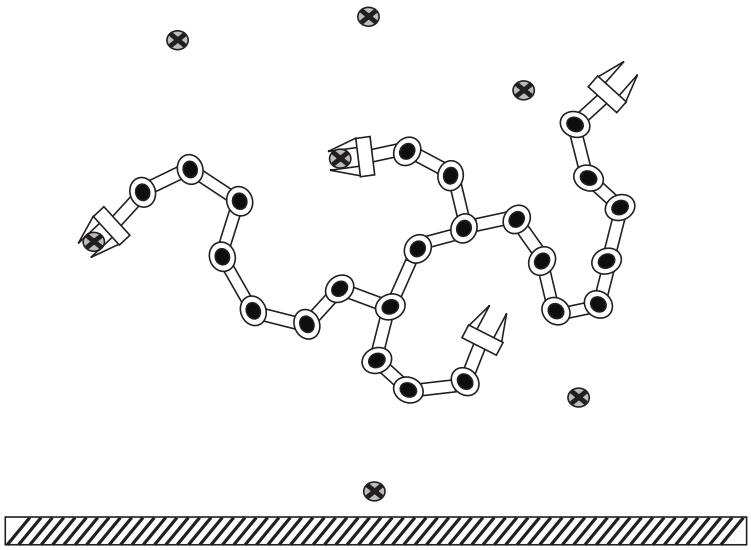
Η παρούσα διατριβή προτείνει μία εμφωλευμένη ιεραρχική πολυπρακτορική αρχιτεκτονική εφαρμοσμένη στο πεδίου του επιδέξιου ρομποτικού χειρισμού. Κύριος στόχος και παράλληλα κίνητρο για την παρούσα ερευνητική προσπάθεια αποτέλεσε η διερεύνηση της εφαρμοσιμότητας καινοτόμων μεθοδολογιών ρομποτικού ελέγχου, που δυνητικά θα επιδεικνύουν τα παρακάτω χαρακτηριστικά: (i) ικανότητα αναπτυξιακής μάθησης σύνθετων δεξιοτήτων, (ii) τμηματικότητα και άμεση επεκτασιμότητα σε προβλημάτα αυξημένης πολυπλοκότητας, (iii) ευρωστία σε μεταβολές της εσωτερικής δομής που χαρακτηρίζει ένα ρομποτικό σύστημα , (iv) κατανεμημένη λειτουργία καθώς και παραλληλία. Θεωρούμε λοιπόν ότι η χρήση πολυπρακτορικών αρχιτεκτονικών οι οποίες ενσωματώνουν ασαφή ενισχυτική μάθηση δύνανται να ικανοποιήσουν τις παραπάνω λειτουργικές απαιτήσεις.



Σχήμα 5.1: *Αναρριχώμενη ρομποτική αλυσίδα - (A)*

Ο περιορισμός της προσέγγισης που προτείνουμε, όπως άλλωστε και σε κάθε μεθοδολογία η οποία δεν κάνει χρήση μοντέλου, είναι κυρίως η περιορισμένη ακρίβεια των λύσεων που δίνει. Συνεπώς, συστήματα στα οποία η ακρίβεια είναι κρίσιμη παράμετρος για την ορθή λειτουργία τους (Critical Systems), δεν μπορούν να μοντελοποιηθούν αμιγώς από αρχιτεκτονικές οι οποίες δεν κάνουν χρήση κάποιου μοντέλου. Αυτό που μπορεί να προταθεί για την μοντελοποίηση συστημάτων που θέτουν αυστηρούς περιορισμούς ως προς την ακρίβεια είναι μία πολυεπίπεδη προσέγγιση. Σε αυτήν την προσέγγιση, εισάγουμε ένα υψηλότερο επίπεδο ελέγχου το οποίο ενσωματώνοντας μεθοδολογίες που δεν θα κάνουν χρήση μοντέλου (όπως ακριβώς η προσέγγιση που προτείνεται στην παρούσα διατριβή), θα μπορεί να διαχειρίζεται την αβεβαιότητα, παρουσιάζοντας χαρακτηριστικά αυτο-οργάνωσης και αυτο-ανάκαμψης από σφάλματα, επιδεικνύοντας ταυτόχρονα δυνατότητες γενίκευσης γνώσης. Στη συνέχεια, μέσω ενός δεύτερου επιπέδου ελέγχου (το οποίο θα καθοδηγείται από το προηγούμενο), κάνοντας χρήση μεθοδολογιών που θα λειτουργούν βάσει μοντέλου, θα διασφαλίζεται η ακρίβεια και η ευρωστία που απαιτεί η λειτουργία ενός κρίσιμου συστήματος.

Μολονότι στα περισσότερα προβλήματα τα οποία μελετήσαμε, υπάρχει μία



Σχήμα 5.2: *Αναρριχώμενη ρομποτική αλυσίδα - (B)*

φυσική συσχέτιση μεταξύ των βαθμών ελευθερίας, στους οποίους αντιστοιχίζονται οι διαφορετικοί πράκτορες, η προτεινόμενη μεθοδολογία μπορεί να επεκταθεί και σε προβλήματα που δεν παρουσιάζουν φυσική συσχέτιση. Ένα σχετικό παράδειγμα μελετήσαμε στην τελευταία πειραματική διάταξη, όπου ενσωματώσαμε την προσέγγισή μας σε δύο αυτοκινούμενα οχήματα τα οποία συνεργατικά πραγματοποίησαν μία εργασία τύπου "box-pushing". Θα πρέπει να σημειωθεί ότι είναι πιθανό, η τοπολογία των πρακτόρων να πρέπει να επαναπροσδιοριστεί (ως προς το βάθος της ιεραρχίας καθώς και τις σχετικές ομάδες των πρακτόρων που πρέπει να συγκροτηθούν), σε σχέση πάντα με την τοπολογία την οποία είδαμε στην περίπτωση των κινηματικών αλυσίδων. Ο περιορισμός της προτεινόμενης μεθοδολογίας και ο οποίος θα πρέπει να σημειωθεί είναι ότι ένας τέτοιος νέος επαναπροσδιορισμός της τοπολογίας δεν γίνεται με κάποιο αυτοματοποιημένο μηχανισμό.

Κάνοντας χρήση συναρτήσεων ανταπόδοσης μέσω των οποίων ενισχύονται οι δράσεις εκείνες που οδηγούν σε ταχύτερη μείωση του σφάλματος (είτε σφάλμα θέσης - είτε σφάλμα δύναμης) το πολυπρακτορικό σύστημα προοδευτικά συγκλίνει σε συμπεριφορές (αναπτύσσει τοπικά αντιστοιχίσεις κατάστασης-δράσης) οι οποίες αποτελεσματικά επιλύουν κινηματικούς πλεονασμούς ενώ επιτυγχάνουν την επιθυμητή συντονισμένη δράση. Θα πρέπει όμως να επισημάνουμε ότι σε όλες τις περιπτώσεις η ρύθμιση της συνάρτησης ανταπόδοσης πραγματοποιείται μέσω μη-αυτόματης διαδικασίας (Manual Tuning). Αυτό προφανώς

αποτελεί έναν περιορισμό της προτεινόμενης προσέγγισης, ενώ θα μπορούσε να αποτελέσει σίγουρα μια πιθανή μελλοντική προέκταση της παρούσας διατριβής ως προς τη διερεύνηση μεθοδολογιών τύπου *Apprenticeship Learning* [1], όπου η συνάρτηση ανταπόδοσης ανακτάται μέσω διαδικασίας ανθρώπινης επίδειξης κατά το πρώτο στάδιο της μάθησης.

Επιπλέον, μελλοντική προέκταση της παρούσας διατριβής θα μπορούσε να αποτελέσει η εφαρμογή του προτεινόμενου πλαισίου σε πιο γενικευμένους χειρισμούς τριών διαστάσεων, δεδομένου ότι η κίνηση του ρομποτικού χειριστή που εξετάσαμε μέχρι τώρα περιορίζεται στο επίπεδο. Η προέκταση αυτή θα επηρεάσει τον υφιστάμενο ορισμό του χώρου κατάστασης μόνο, χωρίς περαιτέρω αλλαγές στην προτεινόμενη αρχιτεκτονική και την τρέχουσα αλγοριθμική δομή.

Ο έλεγχος κινηματικής αλυσίδας με πλεονάζοντες βαθμούς ελευθερίας καθώς επίσης και η εργασία πολυδακτυλικής ρομποτικής λαβής είναι προβλήματα σχεδόν ισοδύναμης πολυπλοκότητας (εάν εξαιρέσουμε το δυναμικό μοντέλο) με το πρόβλημα βάδισης ενός δίποδου ρομπότ (Bipedal Walking Robot). Συνεπώς, η μοντελοποίηση των ποδιών ενός δίποδου ή τετράποδου ρομποτικού συστήματος, ως ανέξαρτητους πράκτορες είναι μία πολύ ενδιαφέρουσα μελλοντική προέκταση της παρούσας ερευνητικής προσπάθειας. Πιο συγκεκριμένα, αντικείμενο μελλοντικής έρευνας θα μπορούσε να αποτελέσει η προσαρμογή της παρούσας εμφωλευμένης αρχιτεκτονικής σε ένα τετράποδο ρομπότ, το οποίο θα πρέπει να μάθει εκείνες τις παραμέτρους βάδισης (Gait Parameters) που θα του επιτρέψουν να επιτύχει την επιθυμητή ταχύτητα και προσανατολισμό.

Θεωρούμε λοιπόν ότι το προτεινόμενο πολυπρακτορικό σύστημα, λόγω της ιδιαίτερης ομοιογένειας που παρουσιάζει ως προς τα στοιχεία που το συνθέτουν και της ιεραρχικής δομής του, διευκολύνει την ανάπτυξη μηχανισμών μάθησης για τον έλεγχο πιο σύνθετων ρομποτικών διατάξεων. Μέσω των εφαρμογών της προτεινόμενης αρχιτεκτονικής που αξιολογήθηκαν στο πλαίσιο της παρούσας διατριβής θεωρούμε ότι σημαντικά προβλήματα στο πεδίο του επιδέξιου ρομποτικού χειρισμού θα μπορέσουν να προσεγγιστούν με ένα νέο τρόπο που παρουσιάζει εξαιρετικό ενδιαφέρον (ως προς την τμηματικότητα, ευρωστία και επεκτασιμότητα που επιδεικνύει). Παρόμοια (και έως ένα βαθμό ισοδύναμα ως προς την πολυπλοκότητα) προβλήματα, όπως ο σχεδιασμός ρομποτικής λαβής (Grasp Planning), ο έλεγχος μετακίνησης (Locomotion Control), ή ο σχεδιασμός βέλτιστων προτύπων αναρρίχησης, βάδισης και μετακίνησης (Designing Optimal Climbing - Gaiting - Locomotion patterns), θα μπορούσαν να ενσωματώσουν την προτεινόμενη προσέγγιση μαζί με τα ιδιαίτερα χαρακτηριστικά που αυτή επιδεικνύει όπως η εξελικτική συνεργατική μάθηση καθώς και ο αναπτυξιακός ρομποτικός έλεγχος.

Βιβλιογραφία

[1] Abbeel P. and Ng A. Apprenticeship learning via inverse reinforcement learning. In: Proc. *21st International Conference on Machine Learning*, pp. 1-8, 2004.

[2] Ahmadabadi M. and Nakano E. A 'constrain and move' approach to distributed object manipulation. *IEEE Transactions on Robotics and Automation*, vol. 17, pp. 157-172, 2001.

[3] Argall B. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, vol. 57, pp. 469-483, 2009.

[4] Bakker B. and Schimidhuber J. Hierarchical reinforcement learning with subpolicies specializing for learned subgoals. In: *Neural Networks and Computational Intelligence*, pp. 125-130, 2004.

[5] Bakker B., Zhumatiy V., Gruener G., and Schmidhuber J. A robot that reinforcement-learns to identify and memorize important previous observations. In: Proc. *IEEE International Conference on Intelligent Robots and Systems*, (IROS'03), pp. 430-435, 2003.

[6] Bakker B., Zivkovic Z., and Krose B. Hierarchical dynamic programming for robot path planning. In: Proc. *IEEE International Conference on Intelligent Robots and Systems*, (IROS'05) pp. 3720-3725, 2005.

[7] Barto A., Sutton R., and Anderson C. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics - Part B*, vol. 13, no. 5, pp. 835-846, 1983.

[8] Barto A. and Mahadevan S. Recent advances in hierarchical reinforcement learning. *Discrete Events Dynamic Systems: Theory and Applications*, vol. 13, no. 4, pp. 341-379, 2003.

[9] Belker T., Beetz M., and Cremers A. Learning action models for the improved execution of navigation plans. *Robotics and Autonomous Systems*, vol. 38, no.3-4, pp. 137-148, 2002.

[10] Bellman R. *Dynamic Programming*. Princeton University Press, 1957.

[11] Bennett C. and Kessler J. *Hq learning: Discovering markovian subgoals for non-markovian reinforcement learning*. Technical report, University of Georgia, 2003.

[12] Ben-Israel A. and Greville T. *Generalised Inverses: Theory and Applications*. Springer, New York, 2003.

[13] Bertsekas D. *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont MA, 1995.

[14] Bertsekas D. and Tsitsiklis J. *Neuro-Dynamic Programming*. Athena Scientific, Belmont MA, 1996.

[15] Boada M., Barber R., and Salichs M. Visual approach skill for a mobile robot using learning and fusion of simple skills. *Robotics and Autonomous Systems*, vol. 38, pp. 157-170, 2002.

[16] Borst C., Fischer F., and Hirzinger M. Calculating hand configurations for precision and pinch grasps. In: Proc. *IEEE International Conference on Intelligent Robots and Systems*, (IROS'02) pp. 1553-1559, 2002.

[17] Boularias A., Kroemer O., and Peters J. Structured apprenticeship learning. In: Proc. *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2012.

[18] Brooks R. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, vol. 2, no. 1, pp. 14-23, 1986.

[19] Brown G. Iterative Solutions of Games by Fictitious Play. In: *Activity Analysis of Production and Allocation*, T.C. Koopmans (editor), John Wiley, 1951.

[20] Buffet O., Dutech A., and Charpillet F. Learning to weigh basic behaviors in scalable agents. newblock In: Proc. *1st International Joint Conference on Autonomous Agents & Multiagent Systems*, (AAMAS'02) pp. 1264-1265, 2002.

[21] Calafiore G., Indri M., and Bona B. Robot dynamic calibration: Optimal excitation trajectories and experimental parameter estimation. *Journal of Robotic Systems*, vol. 18, no. 2, pp. 55-68, 2001.

[22] Cao Y., Fukunaga A., Kahng A., and Meng F. Cooperative mobile robots: Antecedents and directions. In: Proc. *IEEE International Conference on Intelligent Robots and Systems*, (IROS'95) pp. 226-243, 1995.

[23] Cherif M. and Gupta K. Planning for in-hand dextrous manipulation. In: *Robotics, The Algorithmic Perspective*, L. Kavraki P Agarwal and M. Mason, (eds), Peters Publisher, 1998

[24] Chrisman L. Reinforcement learning with perceptual aliasing: The perceptual distinctions approach. In: Proc. *10th National Conference on Artificial Intelligence*, pp. 183-188, 1992.

[25] Claus C. and Boutilier C. The dynamics of reinforcement learning in cooperative multiagent systems. In: Proc. *15th National Conference on Artificial Intelligence*, pp. 746-752, 1998.

[26] Coelho J. and Grupen R. Learning in non-stationary conditions: A control theoretic approach. In: Proc. *17st International Conference on Machine Learning*, pp. 151-158, 2000.

[27] Coulom R. A model-based actor-critic algorithm in continuous time and space. In: Proc. *6th European Workshop on Reinforcement Learning*, 2003.

[28] Dayan P. and Geoffrey E. Feudal reinforcement learning. In *Advances in Neural Information Processing Systems 5*, S.J.Hanson J.D.Cowan and C.L.Giles, (eds), Morgan Kaufmann Publishers, 1993.

[29] Dayan P. and Abbott L. *Theoretical Neuroscience, Computational and Mathematical Modeling of Neural System*. MIT Press, Cambridge, MA, 2001.

[30] Dietterich T. Hierarchical reinforcement learning with the maxq value function decomposition, *Journal of Artificial Intelligence Research*, vol. 13, pp. 227-303, 1999.

[31] Digney B. Learning hierarchical control structure for multiple tasks and changing environments. In: *From Animals to Animats 5*, MIT Press, 1998.

[32] Donald D., Jennings J., and Rus D. Information invariant for distributed manipulation. *Journal of Robotics Research*, vol. 16, pp. 673-702, 1997.

[33] Dordevic G., Rasic M., Kostic D., and Potkonjak V. Representation of robot motion control skill. *IEEE Transactions on Systems, Man, and Cybernetics - Part C*, vol. 30, no. 2, pp. 219-238, 2000.

[34] Doya K. Temporal difference learning in continuous time and space. In *Advances in Neural Information Processing Systems 8*, D.S. Touretzky M.C. Mozer and Hasselmo (eds), MIT Press, 1996.

[35] Ferber J. *Multi-Agent Systems: an Introduction to Distributed Artificial Intelligence*. Addison Wesley, 1999.

[36] Fischer F. and Rovatsos M. An empirical semantics approach to reasoning about communication. *Engineering Applications of Artificial Intelligence*, vol. 15, no. 4, pp.809-823, 2005.

[37] Fischer F., Rovatsos M., and Wei G. Hierarchical reinforcement learning in communication-mediated multiagent coordination. In: Proc. *3rd International Joint Conference on Autonomous Agents & Multiagent Systems*, (AAMAS'04) pp. 1334-1335, 2004.

[38] Fischer F., Rovatsos M., and Weiss G. Acquiring and adapting probabilistic models of agent conversation. In: Proc. *3th International Joint Conference on Autonomous Agents & Multiagent Systems*, (AAMAS'05) pp. 106-113, 2005.

[39] Foster D., Morris R., and Dayan P. A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus*, vol.10, no.1, pp.1-16, 2000.

[40] Fuentes O. and Nelson R. Learning Dextrous Manipulation Strategies for Multifingered Robot Hands Using Evolution Strategy. *Machine Learning*, vol.31, pp.223-237, 1998.

[41] Fundenberg D. and Kreps D. *Lectures on learning and equilibrium in strategic-form games*. Technical Report, CORE Lecture Series, 1992.

[42] Ghavamzadeh M. and Mahadevan S. Continuous-time hierarchical reinforcement learning. In: Proc. *18th International Conference on Machine Learning*, (ICML'01), pp 186-193, 2001.

[43] Ghavamzadeh M. and Mahadevan S. A multiagent reinforcement learning algorithm by dynamically merging markov decision processes. In: Proc. *1st International Joint Conference on Autonomous Agents & Multiagent Systems*, (AAMAS'02), pp. 845-846, 2002.

[44] Glorennec P. and Jouffe L. Fuzzy q-learning. In: Proc. *6th IEEE International Conference on Fuzzy Systems*, pp.659-662, 1997.

[45] Gottfredson L. Mainstream science on intelligence: An editorial with 52 signatories history and bibliography. *Intelligence*, vol. 24, no.1, pp. 13-23, 1997.

[46] Graig J. *Introduction to Robotics, Mechanics and Control*. Addison Wesley, 2nd edition, 1997.

[47] Grefenstette J. and Schultz A. An evolutionary approach to learning in robots. In: Proc. *Machine Learning Workshop on Robot Learning*, pp.659-662, 1994.

[48] Grupen R. A developmental organization for robot behavior. In: Proc. *3rd Internationl Workshop on Epigenetic Robotics*, pp. 25-36, 2003.

[49] Grzegorz G. Enhancements of fuzzy q-learning algorithm. *Computer Science available at http://www.csci.agh.edu.pl/id/eprint/63*, vol. 7, 2005.

[50] Gu D. and Hu H. Accuracy based fuzzy q-learning for robot behaviours. In: Proc. *12th IEEE International Conference on Fuzzy Systems*, pp. 1126-1131, 2004.

[51] Guestrin C., Lagoudakis M., and Parr R. Coordinated reinforcement learning. In: Proc. *19th International Conference on Machine Learning*, (ICML'02), pp. 227-234, 2002.

[52] Hailu G. Symbolic structures in numeric reinforcement for learning optimum robot trajectory. *Robotics and Autonomous Systems*, vol. 37, pp. 53-68, 2001.

[53] Hebb D. *The Organization of Behavior: A Neuropsycological Theory*. John Wiley and Sons, 1949.

[54] Hesselroth T., Sarkar K., Patrick van der Smagt P., and Schulten K. Neural network control of a pneumatic robot arm. *IEEE Transactions on Systems, Man, and Cybernetics - Part A*, vol. 24, no. 1, pp. 28-37, 1994.

[55] Hollerbach J. and Ki C. Redundancy resolution of manipulators through torque optimization. *IEEE Journal of Robotics and Automation*, vol.3, pp.308-316, 1987.

[56] Hubern M. and Grupen R. A feedback control structure for on-line learning tasks. *IEEE Journal of Robotics and Automation*, vol. 22, pp. 303-315, 1997.

[57] Hwang K. and Tsai M. Tan S. Reinforcement learning to adaptive control of nonlinear systems. *IEEE Transactions on Systems, Man, and Cybernetics - Part B*, vol.33, no.3, pp.514-521, 2003.

[58] Hyde J., Tremblay M., and Cutkosky M. An object oriented framework of event-driven dextrous manipulation. In: Proc. *4th International Symposium on Experimental Robotics*, 1995.

[59] Jaakkola T., Jordan T., and Singh S. On the convergence of stochastic iterative dynamic programming algorithm. *Neural Computation*, vol. 6, no. 6, pp. 1185-1201, 1994.

[60] Iida M., Sugisaka M., and Shibata K. Application of direct-vision-based reinforcement learning to a real mobile robot. In: Proc. *9th International Conference on Neural Information Processing*, pp. 2556-2560, 2002.

[61] Irwig K. and Wobcke W. Multi agent reinforcement learning with vicarious rewards. *Electronic Transactions on Artificial Intelligence*, vol. 3, pp. 23-45, 1999.

[62] Kaelbling and L., Littman M., and Moore A. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, vol. 4, pp. 237-285, 1996.

[63] Kaiser M., Dillmann R., Friedrich H., Lin I., Wallner F., and Weckesser P. Learning coordination skills in multi-agent systems. In: Proc. *International Conference on Intelligent Robots and Systems*, pp. 1488-1495, 1996.

[64] Karigiannis J. and Tzafestas C. Multi-agent architecture with continuous reinforcement learning in fuzzy state-space for robot manipulation control. In: Proc. *Integrated Modeling and Analysis in Applied Control and Automation* (IMAACA 2005), 2005.

[65] Karigiannis J. and Tzafestas C. Multi-agent hierarchical architecture modeling kinematic chains employing continuous rl learning with fuzzified state space. In: Proc. *IEEE International Conference on Biomedical Robotics and Biomechatronics*, (BioRob'08), 2008.

[66] Karigiannis J., Rekatsinas T., and Tzafestas C. Fuzzy rule based neurodynamic programming for mobile robot skill acquisition on the basis of a nested multi-agent architecture. In: Proc. *IEEE International Conference on Robotics and Biomimetics*, (RoBio'10), 2010.

[67] Karigiannis J., Rekatsinas T., and Tzafestas C. Hierarchical multiagent architecture employing $TD(\lambda)$ learning with function approximators for robot skill acquisition. In: Proc. *IEEE International Conference on Applied Bionics and Biomechanics*, 2010.

[68] Karigiannis J., Rekatsinas T., and Tzafestas C. Developmental Learning of Cooperative Robot Skills: A Hierarchical Multi-Agent Architectures. In *Perception-action cycle: Models, architectures and hardware*, V. Cutsuridis, A. Hussain, J. Taylor (eds) Springer Series in Cognitive and Neural Systems, 2011.

[69] Kato H. and Billinghurst M. Marker tracking and hmd calibration for a video-based augmented reality conferencing system. In: Proc. *International Workshop on Augmented Reality* (IWAR'99), 1999.

[70] Kavraki L., Svestka P., Latombe J., and Overmars M. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Transactions on Robotics and Automation*, vol. 12, no. 4, pp. 566-580, 1996.

[71] Khatib O. Vehicle arm coordination and multiple mobile manipulator decentralized cooperation. In: Proc. *IEEE International Conference on Intelligent Robots and Systems*, (IROS'96), pp. 546-553, 1996.

[72] Kiguchi K., Nanayakkara T., Watanabe K., and Fukuda T. Multidimensional reinforcement learning using a vector q-net application to mobile robots. In: Proc. *FIRA Robot World Congress*, pp. 405-410, 2002.

[73] Kobori N., Suzuki K., Hartono P., and Hashimoto S. Learning to control a joint driven double inverted pendulum using nested actor critic algorithm. In: Proc. *9th International Conference on Neural Information Processing*, pp. 2610-2614, 2002.

[74] Kohonen T. *Self-Organizing Maps*. Springer, 1997.

[75] Kok J. and Vlassis N. Sparse tabular multiagent q-learning. In: Proc. *Annual Machine Learning Conference of Belgium & Netherlands*, pp. 65-71, 2004.

[76] Kondo T. and Ito K. A reinforcement learning using adaptive state space construction strategy for real autonomous mobile robots. In: Proc. *41st SICE Annual Conference*, pp.3139-3144. 2002.

[77] Kurazume R. and Hirose S. An experimental study of a cooperative positioning system. *Autonomous Robots*, vol.8, no.1, pp.43-52, 2000.

[78] Kuniyoshi Y., Rougeaux S., Ishii M., Kita N., Sakane S., and Kakikura M. Cooperation by observation: the framework and the basic task pattern. In: Proc. *International Conference on Robotics and Automation*, pp. 767-774, 1994.

[79] Lavalle S. *Planning Algorithms*. Cambridge University Press, 2006.

[80] Lavalle S. and Kuffner J. Rapidly-exploring random trees: Progress and prospects. In *In Algorithmic and Computational Robotics: New Direction*, B. Donald, K. Lynch and D. Rus (eds), Peters Publisher, 2001.

[81] Lee D. Behavioral context and coherent oscillations in the supplementary motor area. *Journal of Neuroscience*, vol.24, no. 18, pp.4453-4459, 2004.

[82] Lee H., Park J., and Joo Y. Comments on output tracking and regulation of nonlinear system based on takagi-sugeno fuzzy model. *IEEE Transactions on Systems, Man, and Cybernetics - Part B*, vol.33, no.3, pp.521-523, 2003.

[83] Lopes M. and Santos-Victor Jose. A developmental roadmap for learning by imitation in robots. *IEEE Transactions on Systems, Man, and Cybernetics - Part B*, vol.37, no.2, pp.308-321, 2007.

[84] Low K., Leow W., and Marcelo H. An ensemble of cooperative extended kohonen maps for complex robot motion tasks. *Neural Computation*, vol.17, no.6, pp.1411-1445, 2005.

[85] Luo Z., Ito T., Sugimoto N., Odashima T., and Hosoe S. Virtual impedance control for preshaping of a robot hand. In: Proc. *41st SICE Annual Conference*, pp.2317-2318, 2002.

[86] Lungarella M., Metta G., Pfeifer R., and Sandin G. Developmental robotics: A survey. *Connection Science*, vol.15, no.4, pp. 151-190, 2003.

[87] Makar R., Mahadevan S., and Ghavamzadeh M. Hierarchical multi-agent reinforcement learning. In: Proc. *5th International Conference on Autonomous Agents*, pp. 246-253, 2001.

[88] Martin T., Ambrose R., Diftler M., Platt R., and Butzer M. Tactile gloves for autonomous grasping with the nasa/darpa robonaut. In: Proc. *IEEE International Conference on Robotics and Automation*, (ICRA'04), pp.1713-1718, 2004.

[89] Matsui T., Omata T., and Kaniyoshi Y. Multi-agent architecture for controlling a multi-finger robot. In: Proc. *IEEE International Conference on Intelligent Robots and Systems*, pp.182-186, 1992.

[90] Myerson R. *Game Theory: Analysis of conflict*. Harvard University Press, 1991.

[91] Mitchell T. *Machine Learning*. McGraw Hill, 1997.

[92] Nakamura Y. *Advanced Robotics: Redundancy and Optimization*. Addison-Wesley, 1990.

[93] Namiki A., Imai Y., Ishikawa M., and Kaneko M. Development of a high-speed multifingered hand system and its application to catching. In: Proc. *IEEE International Conference on Intelligent Robots and Systems*, (IROS'03), pp. 2666-2671, 2003.

[94] Niederreiter H. *Random number generation and quasi-monte carlo methods*. Society for Industrial and Applied Mathematics, 1992.

[95] Park K., Kim Y., and Kim J. Modular q-learning based multi-agent cooperation for robot soccer. *Robotics ans Automation Systems*, vol.35, pp.109-122, 2001.

[96] Parr R. and Russell S. Reinforcement learning with hierarchies of machines. In: Proc. *10th Conference on Advances in Neural Information Processing Systems*, pp.1043-1049, 1997.

[97] Pirjanian P. Multiple objective behavior-based control. *Journal of Robotics and Autonomous Systems*, vol.31, no.1-2, pp. 53-60, 2000.

[98] Pollard N. Synthesizing grasps from generalized prototypes. In: Proc. *IEEE International Conference on Robotics and Automation*, pp.2124-2130, 1996.

[99] Pollard N. and Hodgins J. Generalizing demonstrated manipulation tasks. In: Proc. *Workshop on Algorithmic Foundation of Robotics*, pp.1-17, 2002.

[100] Pollard N. and Gilbert R. Tendon arrangement and muscle force requirements for humanlike force capabilities in a robotic finger. In: Proc. *IEEE International Conference on Robotics and Automation*, (ICRA '02), pp.3755-3762, 2002.

[101] Powell W. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. John Wiley, 2007.

[102] Reif J. Complexity of the mover"s problem and generalizations. In: Proc. *IEEE Symposium on Foundation of Computer Science*, 1979.

[103] Ross T. *Fuzzy Logic with Engineering Applications*. McGraw Hill, 1995.

[104] Rovatsos M., Fischer F., and Weiss G. Hierarchical reinforcement learning for communicating agents. In: Proc. *2nd European Workshop on Multiagent Systems*, pp. 593-604, 2004.

[105] Rovatsos M., Rahwan I., Fischer F., and Weiss G. Adaptive strategies for practical argument-based negotiation. In: Proc. *2nd International Workshop on Argumentation in Multi-Agent Systems*, 2005.

[106] Rummery G. and Niranjan M. *On-line q-learning using connectionist systems*. Technical Report, Cambridge University, 1994.

[107] Rus D. Coordinated manipulation of objects in a plane. *Algorithmica*, vol.19, pp. 129-147, 1997.

[108] Schaal S. Is imitation learning the route to humanoid robots. *Trends in Cognitive Sciences*, vol.3, no.6, pp. 233-242, 1999.

[109] Siciliano B., Sciavicco L., Villani L., and Oriolo G. *Robotics Modeling, Planning and Control*. Springer, 2009.

[110] Shibata K. and Ito K. Effect of force load in hand reaching movement acquired by reinforcement learning. In: Proc. *16th Conference on Advances in Neural Information Processing Systems*, pp.1444-1448, 2002.

[111] Shibata K. and Ito K. Hidden representation after reinforcement learning of hand reaching movement with variable link length. In: Proc. *International Conference on Neural Networks*, pp.2619-2624, 2003.

[112] Shibata K., Ito K., and Okabe Y. Direct-vision-based reinforcement learning in going to target task with an obstacle and with a variety of target sizes. In: Proc. *International Conference on Neural Networks and their Applications*, pp.95-102, 1998.

[113] Shibata K., Sugisaka M., and Ito K. Fast and stable learning in direct-vision-based reinforcement learning. In: Proc. *International Symposium on Artificial Life and Robotics*, pp.200-203, 2001.

[114] Shibata K. and Okabe Y. A robot that learns an evaluation function for acquiring of appropriate motions. In: Proc. World Congress in Neural Networks, pp.11-29, 1994.

[115] Shibata K. and Okabe Y. Smoothing-evaluation method in delayed reinforcement learning. 1995.

[116] Shoham Y. and Tennenholtz M. On the synthesis of useful social laws for artificial agent societies. In: Proc. *12th National Conference on Artificial Intelligence*, pp. 276-281, 1994.

[117] Singh P. *Reaching for dexterous manipulation*. September 2004. Technical Report, MIT University, 2003.

[118] Smart W. *Making Reinforcement Learning Work on Real Robots*. Phd. Thesis, Brown University, 2002.

[119] Sutton R. Generalization in reinforcement learning: Successful example using sparse coarse coding. *Advances in Neural Information Processing Systems*, vol.8, no.2, pp.1038-1044, 1996.

[120] Sutton R. and Barto A. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[121] Sutton R., Barto A., and Williams R. Reinforcement learning is direct adaptive optimal control. *IEEE Control Systems*, vol.12, no.2, pp.19-22, 1992.

[122] Takagi T. and Sugeno M. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics*, vol.15, no.1, pp.116-132, 1985.

[123] Takahashi T., Tanaka T., Nishida K., and Kurita T. Self-organization of place cells and reward-based navigation for a mobile robot. In: Proc. *International Conference on Neural Information Processing*, 2001.

[124] Tanaka K. *An Introduction to Fuzzy Logic for Practical Applications*. Springer, 1996.

[125] Tsitsiklis J. Asynchronous stochastic approximation and q-learning. *Machine Learning*, vol.16, pp.185-202, 1994.

[126] Tsoukalas L. and Uhrig R. *Fuzzy and Neural Approaches in Engineering*. John Wiley, 1997.

[127] Tzafestas S. *Introduction to Robotics, Mechanics and Control*. Addison Wesley, 1994.

[128] Watkins C. *Learning from Delayed Rewards*. PhD. Thesis, Kings College, 1989.

[129] Watkins C. and Dayan P. Q-learning. *Machine Learning*, vol.8, pp.279-292, 1992.

[130] Wawrzynski P. and Pacut A. A simple actor-critic algorithm for continuous environments. In: Proc. *10th IEEE International Conference on Methods and Models in Automation and Robotics*, pp.1143-1149, 2003.

[131] Weib G. *Multiagent Systems. A Modern Approach to Distributed Artificial Intelligence*. MIT Press, 1999.

[132] Whitney D. The mathematics of coordinated control of prosthetic arms and manipulators. *Journal of Dynamic Systems, Measurement and Control*, vol.94, no.4, pp.303-310, 1972.

[133] Williams R. and Baird L. *Analysis of some incremental variants of policy iteration: First steps toward understanding actor-critic learning systems.* Technical Report, Northeastern University, 1993.

[134] Worgotter F. Actor-critic models of animal control - a critique of reinforcement learning. In: Proc. *Internationl Symposium on Engineering on Intelligent Systems*, 2004.

[135] Wooldridge M. Intelligent agents. In *Multiagent Systems*, G. Weiss (editor), MIT Press, 1999.

[136] Wunder M., Littman M., and Babes M. Classes of multiagent q-learning dynamics with ϵ-greedy exploration. In: Proc. *19th International Conference on Machine Learning*, (ICML'10), pp.1167-1174, 2010.

[137] Yakey J., LaValle S., and Kavraki L. Randomized path planning for linkages with closed kinematic chains. *IEEE Transactions on Robotics and Automation*, vol. 17, no.6, pp.951-958, 2003.

[138] Yang J., Xu Y., and Chen C. Hidden markov model approach to skill learning and its application to telerobotics. *IEEE Transactions on Robotics and Automation*, vol. 10, no. 5, pp.621-631, 1994.

[139] Zamora J., Millan J., and Murciano A. Specialization in multi-agent systems through learning. *Biological Cybernetics*, vol.76, no.5, pp.375-382, 1997.

[140] Zamora J., Millan J., and Murciano A. Learning and stabilization of altruistic behaviors in multi-agent systems. In: Proc. *International Symposium on Computational Intelligence in Robotics and Automation*, pp.287-293, 1998.

[141] Zhang J., Collani Y., and Knoll A. Interactive assembly by a two-arm robot agent. *Robotics and Automation Systems*, vol.29, no.1, pp.91-100, 1999.