Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Ανάπτυξη Κατανεμημένου Πολυπρακτορικού Συστήματος Από Κοινού Ενισχυτικής Μάθησης για την Αυτόνομη Ρύθμιση του Βηματισμού σε Τετράποδο Ρομπότ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΣΩΤΗΡΙΟΣ Α. ΑΠΟΣΤΟΛΟΠΟΥΛΟΣ

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

Αθήνα, Απρίλιος 2012

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Ανάπτυξη Κατανεμημένου Πολυπρακτορικού Συστήματος Από Κοινού Ενισχυτικής Μάθησης για την Αυτόνομη Ρύθμιση του Βηματισμού σε Τετράποδο Ρομπότ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΣΩΤΗΡΙΟΣ Α. ΑΠΟΣΤΟΛΟΠΟΥΛΟΣ

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

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 2η Απριλίου 2012.

.....	.....	.....
Κωνσταντίνος Τζαφέστας	Νικόλαος Μαράτος	Γεώργιος Στάμου
Αναπληρωτής Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Λέκτορας Ε.Μ.Π.

Αθήνα, Απρίλιος 2012

.....
Σωτήριος Α. Αποστολόπουλος

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

Ευχαριστίες

Με την εκπόνηση της διπλωματικής εργασίας μου, θα ήθελα να ευχαριστήσω τους ανθρώπους που με στήριξαν όλα αυτά τα χρόνια.

Αρχικά, θα ήθελα να ευχαριστήσω τον καθηγητή μου, κύριο Κωνσταντίνο Τζαφέστα, ο οποίος μου εμπιστεύτηκε αυτό το δύσκολο θέμα και υπήρξε δίπλα μου καθόλη τη διάρκεια της διπλωματική εργασίας. Οι συμβουλές του ήταν πολύτιμες, τόσο για την ολοκλήρωση της, όσο και για την ποιότητα των αποτελεσμάτων. Τον ευχαριστώ επίσης για την ώθηση που μου έδωσε να ασχοληθώ με το πεδίο της Μηχανικής Μάθησης και να καταλάβω τι επιθυμώ να ακολουθήσω μετά την ολοκλήρωση των σπουδών μου.

Επίσης, ευχαριστώ από καρδιάς τον κύριο Γιάννη Καρίγιαννη, διδακτορικό φοιτητή υπό την επίβλεψη του κύριου Τζαφέστα, για τον πολύτιμο χρόνο που μου αφιέρωσε, ο οποίος δεν του περίσσευε ιδιαίτερα, και τις πολύτιμες συμβουλές του. Τον ευχαριστώ κυρίως γιατί συμμερίστηκε την προσπάθεια και την αγωνία μου.

Ακόμη, θέλω να ευχαριστήσω όλους τους φίλους μου για τις όμορφες στιγμές που ζήσαμε αυτά τα χρόνια, για τη συμπαράσταση και την κατανόηση τους. Πολλές φορές πιστεύω ότι έκαναν πολλά περισσότερα από όσα περίμενα και πραγματικά νιώθω ευγνωμοσύνη.

Τέλος, θα ήθελα να ευχαριστήσω τους γονείς και την αδερφή μου για τις συμβουλές και την αγάπη τους. Έκαναν ό,τι μπορούσαν και στερήθηκαν πράγματα, για να μεγαλώσω όσο καλύτερα μπορούσα.

Σωτήρης Αποστολόπουλος

Περίληψη

Η κίνηση ρομπότ με πόδια έχει προσελκύσει το ενδιαφέρον αρκετών ερευνητών από διάφορα επιστημονικά πεδία παγκοσμίως, για περισσότερες από δύο δεκαετίες. Η σχεδίαση βέλτιστων προτύπων βάδισης για ένα βαδίζον ρομπότ συνιστά ένα ιδιαίτερα πολύπλοκο πρόβλημα, κυρίως λόγω του μεγάλου αριθμού παραμέτρων οι οποίες καθορίζουν τη συμπεριφορά κίνησης, καθιστώντας έτσι την εφαρμογή τυπικών αλγορίθμων αναζήτησης ακατάλληλη. Λύση στο πρόβλημα αυτό μπορούν να δώσουν μέθοδοι Μηχανικής Μάθησης, οι οποίες δε χρειάζονται κάποιο μοντέλο του συστήματος και μπορούν να μάθουν τις παραμέτρους με καλή προσέγγιση. Η εργασία η οποία παρουσιάζεται σε αυτή τη διπλωματική εργασία, εμπνεύστηκε από την ανάγκη ανάπτυξης μεθοδολογιών οι οποίες μπορούν να παρέχουν ιδιότητες αναπτυξιακής μάθησης και προσαρμογής συμπεριφοράς σε πολύπλοκους ρομποτικούς μηχανισμούς, όπως τα ρομπότ με πόδια. Ο στόχος είναι να προσδώσουν στους πολύπλοκους ρομποτικούς μηχανισμούς τη δυνατότητα να μαθαίνουν αυτόματα τον τρόπο με τον οποίο θα φέρνουν εις πέρας νέες εργασίες, ιδιαίτερα εργασίες δύσκολες σε μη – φιλικά και μη – δομημένα περιβάλλοντα για τα οποία τα κλασικά ρομπότ με τροχούς κρίνονται ακατάλληλα.

Στην παρούσα διπλωματική εργασία, προτείνουμε δύο νέες μεθόδους μάθησης παραμέτρων ενός τετράποδου ρομποτικού συστήματος, με σκοπό την κίνηση του με κάποια επιθυμητή ταχύτητα. Οι μέθοδοι αυτοί χρησιμοποιούν Ενισχυτική Μάθηση. Αρχικά προτείνουμε μία προσέγγιση ενός πράκτορα και στη συνέχεια, προτείνουμε μία προσέγγιση πολλών πρακτόρων. Συγκεκριμένα προσθέτουμε τέσσερις ακόμη πράκτορες, σε διαφορετικό επίπεδο, οι οποίοι μαθαίνουν κάποια τοπική συμπεριφορά και συγχρονίζονται από τον πράκτορα της πρώτης προσέγγισης. Το σύστημα αξιολογείται στην πλατφόρμα εξομοίωσης Webots™. Τα αποτελέσματα που προκύπτουν αποδεικνύουν ότι ο προτεινόμενος πολυπρακτορικός μηχανισμός μάθησης καθιστά ικανό το ρομπότ να κινηθεί με ένα ευσταθές πρότυπο βάδισης, επιτυγχάνοντας την επιθυμητή ταχύτητα με πολύ μικρό σφάλμα και υψηλό βαθμό γενίκευσης.

Λέξεις κλειδιά : Κίνηση Τετράποδου Ρομπότ, Ενισχυτική Μάθηση, Από κοινού Μάθηση, Πολυπρακτορικό Σύστημα, Κατανεμημένο Σύστημα, Πολυεπίπεδο Σύστημα

Abstract

Legged robot locomotion has attracted the interest of many researchers, from different scientific fields worldwide, for more than two decades now. Designing optimal gait patterns for a walking robot constitutes a particularly complex problem, notably due to the large number of parameters that govern locomotion behavior, making the application of typical search algorithms inappropriate. Solution to this problem can come from Machine Learning methods, that do not need a model of the system and are able to learn this set of parameters with a very good approximation. The work presented in this diploma thesis has been motivated by the need to develop methodologies that can provide complex robotic mechanisms, such as legged walking robots, with developmental learning and behavioral adaptability properties. The goal is to endow complex robot mechanisms with capacities to automatically learn how to fulfill new tasks, especially difficult tasks in hostile and unstructured environments for which classic wheeled mobile robots are unsuitable.

In this diploma thesis, we propose two new methods for parameter learning of a quadruped robot, in order to make it able to move with a desired velocity. These methods use Reinforcement Learning. At first, we suggest a single agent approach and later, we suggest a multiagent approach. Specifically, we introduce four additional agents, on a different level, that learn a local behavior and are coordinated by the agent of the first approach. The system is evaluated on the Webots simulation platform. The obtained results show that the proposed multi-agent learning mechanism enables the quadruped robot to achieve a stable gait pattern, reaching its goal velocity with very slight error and great degree of generalization.

Keywords : Quadrupedal Robot Locomotion, Reinforcement Learning, Joint Action Learning, Multiagent System, Distributed System, Multilevel System

Πίνακας περιεχομένων

1 *Εισαγωγή*

1.1 Ρομπότ με πόδια

Τα ρομπότ με πόδια έχουν αποδειχθεί πολλά υποσχόμενα συστήματα, ικανά να φέρνουν εις πέρας εργασίες τις οποίες ρομπότ με τροχούς αδυνατούν να καταφέρουν. Ακόμα πιο εντυπωσιακό είναι το γεγονός ότι το πεδίο αυτό αναπτύσσεται με πολύ ταχείς ρυθμούς, με συμβολή από πολλούς διαφορετικούς τομείς. Τις τελευταίες τρεις δεκαετίες έχουν αναπτυχθεί νέες μηχανές και μέθοδοι. Μεγαλύτερη συμβολή μπορούμε να παραδεχτούμε ότι έχει προέλθει από τη φύση. Εκεί, ο βηματισμός των ζώων έχει τελειοποιηθεί μέσω της εξέλιξης και αποτελεί το καλύτερο παράδειγμα μίμησης από τον άνθρωπο. Βιολόγοι, μηχανικοί και ιατροί προσπαθούν να κατανοήσουν τον τρόπο με τον οποίο προκύπτει ο βηματισμός σε αυτούς τους οργανισμούς και να το χρησιμοποιήσουν προς όφελος της ανθρωπότητας. Μεγάλη συμβολή πρέπει να αναγνωρίσουμε και στον τομέα της κατασκευής υλικών. Τα τελευταία χρόνια έχουν γίνει αρκετές προσπάθειες κατασκευής ευλύγιστων, αλλά και αρκετά ανθεκτικών υλικών, τα οποία θα ξεπεράσουν την ακαμψία των μεταλλικών συνδέσμων και θα προσδώσουν στα ρομπότ τη δυνατότητα ανάπτυξης υψηλών ταχυτήτων με ευστάθεια. Η μελέτη και η έρευνα πάνω στην κίνηση ρομπότ με πόδια έχει αποδείξει ότι η κίνηση τους είναι συνάρτηση αρκετών παραμέτρων. Όταν η πολυπλοκότητα του συστήματος αυξάνεται και οι ευριστικοί μηχανισμοί επίλυσης προβλημάτων φθάνουν στα όρια τους, μπορούμε να εφαρμόσουμε μεθόδους μηχανικής μάθησης, όπως η ενισχυτική μάθηση.

1.2 Αντικείμενο της διπλωματικής

Στην παρούσα διπλωματική εργασία επικεντρωνόμαστε σε πολιτικές συγχρονισμού των ποδιών ενός τετράποδου ρομπότ, με σκοπό να κινηθεί με μία συγκεκριμένη ταχύτητα. Ερευνούμε επίσης τον τρόπο με τον οποίο κάτι τέτοιο μπορεί να επιτευχθεί, θεωρώντας το συνολικό σύστημα πολυπρακτορικό. Συγκεκριμένα, υπάρχει ένας πράκτορας στο επίπεδο 1, ο οποίος συγχρονίζει την κίνηση των ποδιών και τέσσερις πράκτορες στο επίπεδο 2, ένας για κάθε πόδι, οι οποίοι μαθαίνουν τοπικές συμπεριφορές. Επομένως, έχουμε μία ετερογενή δομή.

Θεωρούμε ότι το πεδίο της Ενισχυτικής Μάθησης (Reinforcement Learning) παρέχει τους κατάλληλους αλγορίθμους βελτιστοποίησης για τη μάθηση μίας βέλτιστης πολιτικής συγχρονισμού τεσσάρων – ανεξάρτητων μεταξύ τους – τοπικών συμπεριφορών. Η ενισχυτική μάθηση έχει εφαρμοστεί με τεράστια επιτυχία σε αρκετά προβλήματα με μεγάλο χώρο κατάστασης και σε διάφορα πεδία που εκτείνονται από προβλήματα όπως η μάθηση βάδισης, μέχρι προβλήματα υπολογιστικής βιολογίας και βάσεων δεδομένων [1]. Αφού προτείνουμε τους δύο τρόπους επίτευξης του στόχου μας, τους αξιολογούμε σε ένα τετράποδο ρομπότ σε περιβάλλον εξομοίωσης.

Η σημαντικότερη πρόκληση στην έρευνα μας ήταν ο μεγάλος χώρος καταστάσεων – δράσεων και η έμφυτη πολυπλοκότητα ενός συστήματος κίνησης με τέσσερα πόδια. Επίσης, στην πολυπρακτορική προσέγγιση, η εισαγωγή τεσσάρων ακόμη πρακτόρων, αύξανε την πολυπλοκότητα του συστήματος σε αρκετό βαθμό, δεδομένου ότι κάθε πράκτορας αποφάσιζε την ατομική του δράση κάνοντας εκτιμήσεις για τις δράσεις των υπολοίπων.

Τέλος, να επισημάνουμε ότι η συμμετοχή των παραπάνω πρακτόρων στη μάθηση οδήγησε το σύστημα σε σχήματα βάδισης, οι οποίες μπόρεσαν να προσδώσουν στο ρομπότ την επιθυμητή ταχύτητα με πολύ μικρό σφάλμα.

1.2.1 Συνεισφορά

Στην παρούσα διπλωματική εργασία στόχος μας είναι η ανάπτυξη ενός συστήματος μάθησης το οποίο θα μάθει μία πολιτική συγχρονισμού και κάποιες τοπικές συμπεριφορές, οι οποίες θα καταστήσουν ικανό ένα τετράποδο ρομπότ να κινείται με διάφορες επιθυμητές ταχύτητες.

Παλαιότερες εργασίες μάθησης βάδισης σε ρομπότ με τέσσερα ή έξι πόδια έχουν γίνει στο [2], όπου όμως υπάρχει σύζευξη μεταξύ γειτονικών ποδιών και στο [3], όπου τα πόδια είναι αμοιβαία συζευγμένα. Στο [4] το κάθε module γνωρίζει και την κατάσταση των υπόλοιπων modules, στο [5] υπάρχει μερική γνώση της κατάστασης των άλλων ποδιών, ενώ στο [6] τα πόδια χωρίζονται σε 2 ομάδες των 3 ποδιών (εξάποδο), όπου κάθε ομάδα υλοποιεί έναν αλγόριθμο από κοινού μάθησης.

Στη δική μας προσέγγιση, κανένα πόδι δεν έχει κάποιο βαθμό σύζευξης με κάποιο άλλο. Επίσης, δε γνωρίζει κάτι για την κατάσταση κάποιου άλλου πράκτορα, δηλαδή το διάνυσμα κατάστασης του περιέχει μόνο τοπικές μεταβλητές και όχι μεταβλητές που αναφέρονται σε άλλους πράκτορες. Κάθε πράκτορας, με λίγα λόγια έχει το δικό του χώρο αναζήτησης. Το πολυπρακτορικό σύστημα έχει σχεδιαστεί με τέτοιο τρόπο ώστε όλοι οι πράκτορες να είναι ανεξάρτητοι μεταξύ τους και ο καθένας να υλοποιεί ένα δικό του αλγόριθμο από κοινού μάθησης. Έτσι, μειώνεται η πολυπλοκότητα του συστήματος και η διαδικασία μάθησης είναι πλήρως κατανεμημένη.

1.3 Οργάνωση κειμένου

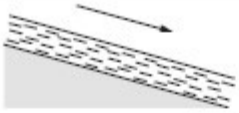
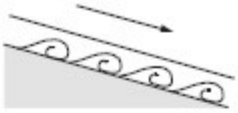
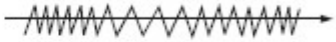
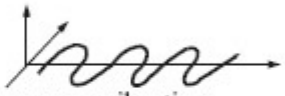
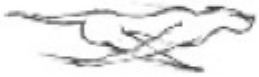
Η παρούσα εργασία χωρίζεται σε 7 κεφάλαια. Το κεφάλαιο 2 αναφέρεται στην κίνηση ρομπότ με τέσσερα πόδια. Αναφέρονται κριτήρια στατικής και δυναμικής ευστάθειας και περιγράφεται η κανονική βάδιση (standard gait). Στο κεφάλαιο 3 γίνεται μία εισαγωγή στο πεδίο της Ενισχυτικής Μάθησης και παρουσιάζονται οι σημαντικότερες κατηγορίες αλγορίθμων για συνεχείς και διακριτούς χώρους καταστάσεων. Στο κεφάλαιο 4 κάνουμε μία εισαγωγή στη μέθοδο της από κοινού μάθησης. Συγκεκριμένα, παρουσιάζουμε τον αλγόριθμο *Fictitious Play* και παραθέτουμε κάποιες πληροφορίες για τη σύγκλιση του. Στο κεφάλαιο 5 παραθέτουμε αναλυτικά τις δύο προσεγγίσεις που ακολουθήσαμε, παρουσιάζουμε και σχολιάζουμε τα αποτελέσματα μας. Στο κεφάλαιο 6 κάνουμε μία σύνοψη της εργασίας και παραθέτουμε κάποιες σκέψεις για μελλοντική έρευνα του συστήματος μας.

2 *Κίνηση τετράποδου ρομπότ*

Ένα κινούμενο ρομπότ χρειάζεται μηχανισμούς κίνησης ώστε να μπορεί να κινηθεί στο περιβάλλον στο οποίο βρίσκεται. Για το σκοπό αυτό υπάρχουν διάφοροι μηχανισμοί. Για παράδειγμα, υπάρχουν ρομπότ με ένα, δύο, τέσσερα και έξι πόδια ή ρομπότ με διάφορες διατάξεις τροχών. Στο κεφάλαιο αυτό θα επικεντρωθούμε στην κίνηση με πόδια. Θα περιγράψουμε την τετράποδη κίνηση και θα δώσουμε μερικά παραδείγματα βηματισμών.

2.1 Εισαγωγή

Οι μηχανισμοί κίνησης με πόδια είναι συνήθως εμπνευσμένοι από βιολογικά συστήματα, τα οποία έχουν εξελίξει με τα χρόνια τους μηχανισμούς αυτούς. Ο πίνακας παρακάτω περιγράφει μερικούς τέτοιους μηχανισμούς, οι οποίοι συναντώνται στη φύση. Όταν όμως καλούμαστε να ενσωματώσουμε αυτούς τους μηχανισμούς σε ρομπότ, ερχόμαστε αντιμέτωποι με αρκετά προβλήματα. Τα κυριότερα από αυτά είναι η μηχανική πολυπλοκότητα την οποία εμπεριέχουν τα πόδια, η ευστάθεια και η κατανάλωση ενέργειας.

Type of motion	Resistance to motion	Basic kinematics of motion
Flow in a Channel 	Hydrodynamic forces	Eddies 
Crawl 	Friction forces	Longitudinal vibration 
Sliding 	Friction forces	Transverse vibration 
Running 	Loss of kinetic energy	Oscillatory movement of a multi-link pendulum 
Jumping 	Loss of kinetic energy	Oscillatory movement of a multi-link pendulum 
Walking 	Gravitational forces	Rolling of a polygon (see figure 2.2) 

Πίνακας 2.1 : Μηχανισμοί κίνησης με πόδια, εμπνευσμένοι από τη φύση [7].

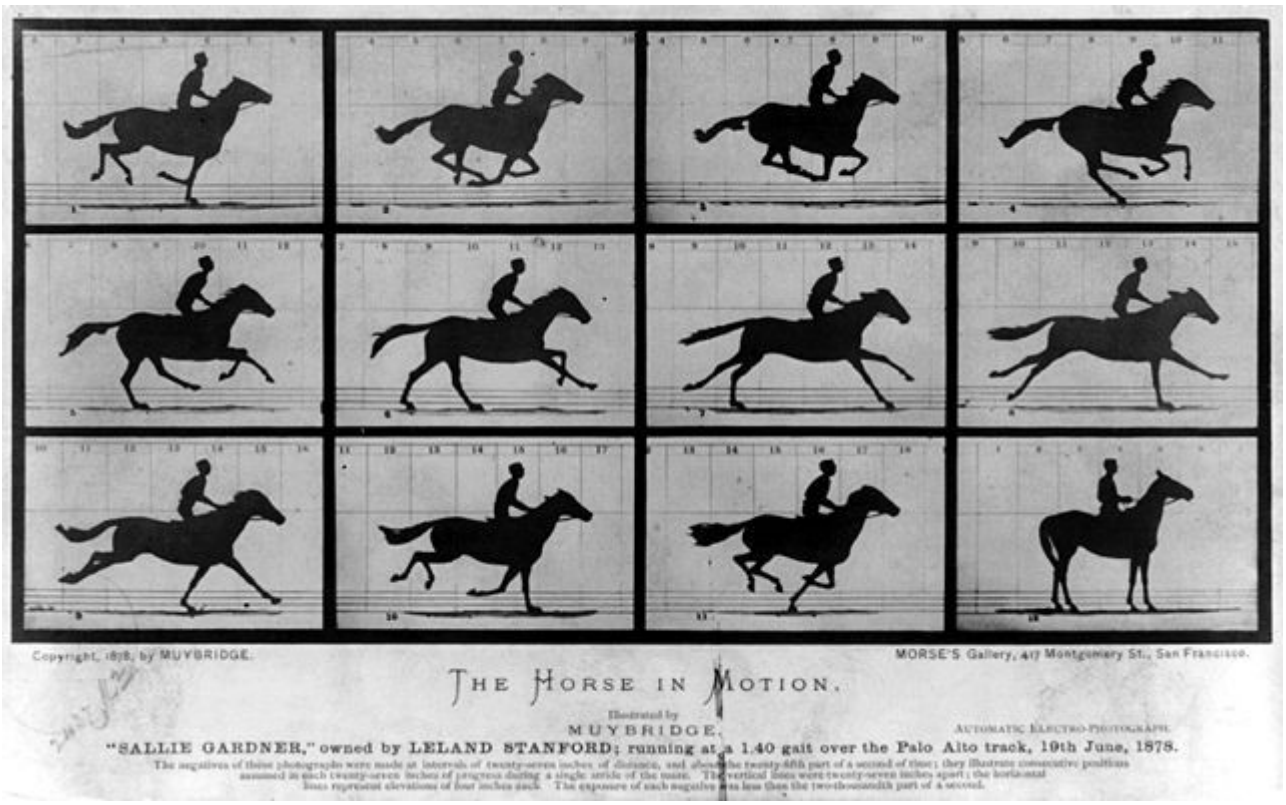
Παρά αυτά τα προβλήματα όμως, πολύ συχνά καλούμαστε να κατασκευάσουμε ρομπότ, τα οποία υιοθετούν μηχανισμούς κίνησης με πόδια. Ο λόγος είναι ότι ένα ρομπότ με πόδια μπορεί να κινηθεί σε πολύ περισσότερες επιφάνειες σε σύγκριση με ένα ρομπότ με τροχούς. Είναι ικανό να ανέβει σκάλες, να περάσει από κενά τόσο μεγάλα όσο ο διασκελισμός του και να κινηθεί σε επιφάνειες με πολλές ανωμαλίες.

Για να μπορέσει ένα ρομπότ με πόδια να κινηθεί, πρέπει να έχει τουλάχιστον δύο βαθμούς ελευθερίας σε κάθε πόδι (degrees of freedom – DOF). Κάθε βαθμός ελευθερίας αντιστοιχείται σε ένα σερβοκινητήρα, ο οποίος με τη σειρά του αποτελεί μία άρθρωση. Ένα τετράποδο λοιπόν χρειάζεται τουλάχιστον οκτώ σερβοκινητήρες, για να κινηθεί.

Πριν προχωρήσουμε στην ανάλυση της τετράποδης βάδισης, πρέπει να αναφέρουμε ότι ο πρωτοπόρος της επιστημονικής ανάλυσης και της κατάταξης των μηχανισμών κίνησης με τέσσερα

πόδια ήταν ο Milton Hildebrand [38]. Ο Hildebrand χώρισε την κίνηση του ποδιού σε δύο φάσεις, τη φάση υποστήριξης, όπου το πόδι βρίσκεται στο έδαφος και τη φάση μετάβασης, όπου βρίσκεται στον αέρα. Κάθε πόδι πρέπει να ολοκληρώσει και τις δύο φάσεις σε μία περίοδο, η οποία είναι ίδια και για τα τέσσερα πόδια, αλλιώς δεν μπορεί να προκύψει ένα επαναλαμβανόμενο μοτίβο. Επομένως, η βάδιση μπορεί να περιγραφεί μέσω της έναρξης και λήξης των δύο φάσεων τριών ποδιών, σε σχέση με τον κύκλο κίνησης ενός ποδιού αναφοράς.

Τέλος, παραθέτουμε την κλασική ακολουθία βάδισης ενός αλόγου του Eadweard Muybridge, η οποία θεωρείται ως μία πρώτη προσπάθεια μελέτης και κατανόησης της τετράποδης κίνησης.



Εικόνα 2.2 : Ακολουθία βάδισης αλόγου του Eadweard Muybridge.

2.2 Ευστάθεια

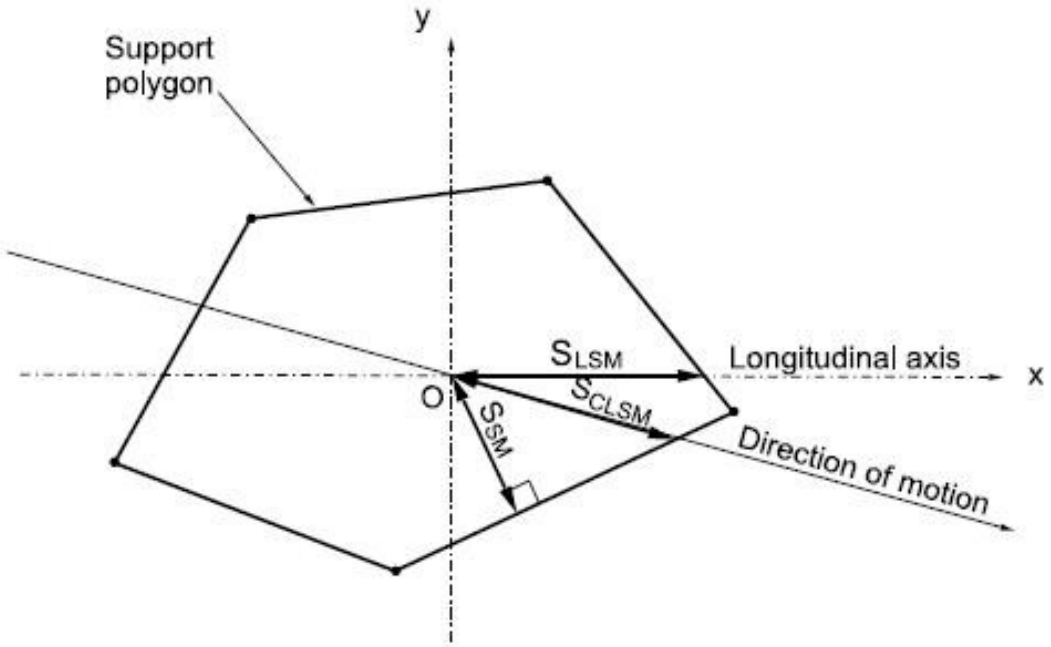
Η ευστάθεια των ρομπότ με πόδια χωρίζεται σε δύο κατηγορίες, τη στατική και τη δυναμική ευστάθεια. Παρακάτω παραθέτουμε μία περιγραφή τους.

2.2.1 Στατική ευστάθεια

Το πρώτο κριτήριο στατικής ευστάθειας διατυπώθηκε από τους McGhee και Frank [8], το 1968. Αφορούσε ένα ιδανικό ρομπότ με πόδια χωρίς μάζα, το οποίο κινούταν σε μία επίπεδη επιφάνεια με σταθερή ταχύτητα και σταθερή κατεύθυνση. Η μέθοδος της προβολής του κέντρου μάζας υποστηρίζει ότι ένα ρομπότ με πόδια είναι στατικά ευσταθές αν η οριζόντια προβολή του κέντρου μάζας του (center of mass – COM), βρίσκεται εντός του πολυγώνου υποστήριξης (support polygon). Το πολύγωνο υποστήριξης ορίζεται ως το κυρτό πολύγωνο, το οποίο σχηματίζεται αν ενώσουμε τα σημεία όπου πατούν τα πόδια του ρομπότ.

Υπάρχουν διάφορα μέτρα στατικής ευστάθειας, τα οποία θα περιγράψουμε επιγραμματικά.

- Το **περιθώριο στατικής ευστάθειας** (McGhee και Iswandhi [26], **static stability margin S_{SM}**), το οποίο ορίζεται ως η ελάχιστη από τις αποστάσεις των πλευρών του πολυγώνου υποστήριξης από την προβολή του COM (βλέπε σχήμα 2.3).
- Το **κατά μήκος περιθώριο ευστάθειας** (Zhang και Song[27], **longitudinal stability margin S_{LSM}**). Ορίζεται ως η ελάχιστη από τις αποστάσεις από την προβολή του COM στις μπροστινές και πισινές ακμές του πολυγώνου υποστήριξης, κατά μήκος του ρομπότ (βλέπε σχήμα 2.3). Το S_{LSM} είναι μία καλή προσέγγιση του S_{SM} και είναι πιο εύκολο να υπολογιστεί.
- Το **crab longitudinal stability margin S_{CLSM}** (Zhang και Song[28]). Ορίζεται ως η ελάχιστη από τις αποστάσεις από την προβολή του COM στις μπροστινές και πισινές ακμές του πολυγώνου υποστήριξης, κατά μήκος του άξονα κίνησης του ρομπότ (βλέπε σχήμα 2.3). Το S_{CLSM} είναι πιο κατάλληλο, όταν θεωρούμε ότι δεν έχουμε ένα ιδανικό ρομπότ και πλέον έχουμε να αντιμετωπίσουμε την επίδραση της αδράνειας κατά την επιτάχυνση.



Σχήμα 2.3 : Πολύγωνο υποστήριξης και διάφορα περιθώρια στατικής ευστάθειας [9].

Τα παραπάνω κριτήρια ευστάθειας βασίζονται σε γεωμετρικά χαρακτηριστικά. Είναι ανεξάρτητα του ύψους του κέντρου μάζας του ρομπότ και δε λαμβάνουν υπόψιν τους κινηματικές και δυναμικές παραμέτρους. Όταν έχουμε να κάνουμε με ένα μη ιδανικό ρομπότ, οι παραπάνω παράμετροι πρέπει να ληφθούν υπόψιν. Ένα κριτήριο που συμπεριλαμβάνει τα παραπάνω είναι το **περιθώριο ενεργειακής ευστάθειας S_{ESM} (energy stability margin)**, το οποίο ορίζεται ως η ελάχιστη δυναμική ενέργεια που απαιτείται για να ανατρέψει το ρομπότ ως προς μία ακμή του πολυγώνου υποστήριξης. Ο τύπος που υπολογίζει αυτό το κριτήριο είναι ο

$$S_{ESM} = min(mgh_i), i = 1,\ldots,n_s \quad (2.1)$$

όπου h_i είναι η απόσταση του κέντρου μάζας του ρομπότ από το έδαφος και n_s το πλήθος το ποδιών που υποστηρίζουν το ρομπότ.

Το S_{ESM} είναι είναι πιο αποδοτικό για τη μέτρηση στατικής ευστάθειας. Είναι ένα ποιοτικό μέτρο της ενέργειας κρούσης που μπορεί να αντέξει το ρομπότ χωρίς να ανατραπεί και επίσης λαμβάνει υπόψιν του το ύψος του κέντρου μάζας. Όμως, δε λαμβάνει υπόψιν του δυναμικά φαινόμενα.

Μία παραλλαγή του S_{ESM} είναι το **κανονικοποιημένο περιθώριο ενεργειακής ευστάθειας (normalized energy stability margin S_{NESM})** το οποίο ορίζεται ως

$$S_{NESM} = \frac{S_{ESM}}{mg} = min(h_i), i = 1,..., n_s \quad (2.3)$$

Το S_{NESM} είναι κανονικοποιημένο ως προς το βάρος του ρομπότ και έχει αποδειχτεί ότι είναι το πιο αποδοτικό κριτήριο στατικής ευστάθειας.

Όταν όμως, προκύπτουν δυναμικά φαινόμενα κατά τη βάδιση, τα παραπάνω κριτήρια δεν μπορούν να κρίνουν την ευστάθεια του ρομπότ με ακρίβεια και επομένως χρειαζόμαστε κριτήρια δυναμικής ευστάθειας.

2.2.2 Δυναμική ευστάθεια

Το πρώτο κριτήριο δυναμικής ευστάθειας προτάθηκε από τον Orin [10] ως επέκταση της μεθόδου προβολής του κέντρου μάζας, για ρομπότ που εκτελεί crawl gait. Η μέθοδος αυτή ονομάζεται **μέθοδος του κέντρου πίεσης COP method (center of pressure method)** και υποστηρίζει ότι το ρομπότ είναι δυναμικά ευσταθές όταν η προέκταση του κέντρου μάζας του κατά μήκος της διεύθυνσης της συνισταμένης δύναμης σε αυτό, βρίσκεται εντός του πολυγώνου υποστήριξης. Το **περιθώριο δυναμικής ευστάθειας (dynamic stability margin S_{DSM})** ορίζεται λοιπόν, ως η ελάχιστη από τις αποστάσεις του COM από τις ακμές του πολυγώνου υποστήριξης.

Όταν ένα ρομπότ με πόδια ανατρέπεται, σημαίνει ότι έχουν χάσει την επαφή με το έδαφος όλα τα πόδια, εκτός από αυτά τα οποία σχηματίζουν τον άξονα περιστροφής. Μία συνισταμένη δύναμη $\mathbf{F_R}$ και ροπή $\mathbf{T_R}$ μεταξύ του ρομπότ και του εδάφους υπερνικάει το άθροισμα των δυνάμεων αντίδρασης με το έδαφος $\mathbf{F_{ri}}$ και τις ροπές που ασκούν γύρω από το κέντρο μάζας. Για να μπορέσει το ρομπότ να είναι ευσταθές, πρέπει να δημιουργηθεί μία συνισταμένη δύναμη και ροπή, οι οποίες θα ασκήσουν μία ροπή M_i γύρω από έναν άξονα i, η οποία θα αντισταθμίσει τις δυνάμεις και ροπές αποσταθεροποίησης. Όταν κάτι τέτοιο δεν μπορεί να συμβεί, το σύστημα είναι δυναμικά ασταθές.

Βασισμένοι στα παραπάνω, οι Lin και Song [11], αναδιατύπωσαν το περιθώριο δυναμικής ευστάθειας, ως η ελάχιστη από όλες τις ροπές *Mi* γύρω από κάθε άξονα περιστροφής του πολυγώνου υποστήριξης, κανονικοποιημένη ως προς το βάρος του συστήματος. Δηλαδή,

$$S_{DSM} = min(\frac{M_i}{mg}) = min(\frac{\boldsymbol{e_i} \cdot (\boldsymbol{F_R} \times \boldsymbol{P_i} + \boldsymbol{M_R})}{mg}) \quad (2.4)$$

όπου $\boldsymbol{P_i}$ είναι το διάνυσμα με αρχή το κέντρο μάζας και τέλος το i-οστό πόδι υποστήριξης και $\boldsymbol{e_i}$ είναι το μοναδιαίο διάνυσμα, το οποίο ταυτίζεται με την i-οστή ακμή του πολυγώνου υποστήριξης, κατά την ωρολογιακή φορά. Αν όλες οι ροπές είναι θετικές, τότε το σύστημα είναι δυναμικά ευσταθές.

Ένα άλλο κριτήριο δυναμικής ευστάθειας είναι το **κριτήριο ευστάθειας πτώσης $\boldsymbol{S_{TSJ}}$ (tumble stability judgement)** (Yoneda και Hiroshe [12]). Στο κριτήριο αυτό θεωρούμε ότι έχουμε πόδια χωρίς μάζα, επομένως οι δυνάμεις υποστήριξης και αντίδρασης από το έδαφος συμπίπτουν. Για να καταλήξουμε στη διατύπωση του κριτηρίου αυτού, χρειαζόμαστε τις εξισώσεις δυναμικής ισορροπίας του ρομπότ. Αυτές είναι

$$-\boldsymbol{F_R} = \boldsymbol{F_I} - \boldsymbol{F_G} \quad (2.5)$$
$$-\boldsymbol{M_R} = \boldsymbol{M_I} - \boldsymbol{M_G} \quad (2.6)$$

όπου $\boldsymbol{F_I}$ και $\boldsymbol{M_I}$ οι δυνάμεις και ροπές αδράνειας αντίστοιχα, $\boldsymbol{F_G}$ και $\boldsymbol{M_G}$, οι δυνάμεις και ροπές λόγω βαρύτητας και $-\boldsymbol{F_R}$ και $-\boldsymbol{M_R}$ οι δυνάμεις και ροπές αντίδρασης από το έδαφος. Η ροπή M_i γύρω από τον άξονα περιστροφής υπολογίζεται ως

$$M_i' = -\boldsymbol{M_R} \cdot \boldsymbol{e_i} - \boldsymbol{F_R} \times \boldsymbol{p_i} \cdot \boldsymbol{e_i} \quad (2.7)$$

Το S_{TSJ} υποστηρίζει ότι το σύστημα είναι δυναμικά ευσταθές αν υπάρχει κάποιο πόδι υποστήριξης j στη διεύθυνση της περιστροφής, το οποίο αποτρέπει το σύστημα από το να πέσει. Το S_{TSJ} δίνεται από τον τύπο

$$S_{TSJ} = min\left(\frac{|M_i'|}{mg}\right) \quad (2.8)$$

Μία παραλλαγή του S_{TSJ} είναι το κριτήριο **Leg-end Supporting Moment $\boldsymbol{S_{LESM}}$** (Zhou και άλλοι [13]), κατά το οποίο μετράμε τη συνισταμένη δύναμη και ροπή από αισθητήρες δύναμης στα πόδια. Έτσι αποφεύγουμε τα σφάλματα που δημιουργούνται στο S_{TSJ} εξαιτίας της αγνοήσης της δυναμικής των ποδιών.

Το 1996, οι Papadopoulos και Rey [13], πρότειναν το **περιθώριο ευστάθειας δύναμης – γωνίας** (**force – angle stability margin $\boldsymbol{S_{FASM}}$**). Το κριτήριο αυτό βρίσκει τη γωνία α_i μεταξύ της

συνισταμένης δύναμης, η οποία ασκείται από το κέντρο μάζας στο έδαφος ($\boldsymbol{F_R}$) – η αντίθετη από τη δύναμη αντίδρασης του εδάφους – και το διάνυσμα $\boldsymbol{R_i}$, το οποίο διέρχεται από το κέντρο μάζας και είναι κάθετο στον άξονα περιστροφής. Το σύστημα γίνεται ασταθές όταν αυτή η γωνία γίνει μηδενική. Το περιθώριο ευστάθειας δύναμης – γωνίας είναι το γινόμενο της γωνίας επί το μέτρο της συνισταμένης δύναμης $\boldsymbol{F_R}$, δηλαδή δίνεται από τη σχέση

$$S_{FASM} = min(\alpha_i)\left\|\boldsymbol{F_R}\right\| \quad (2.9)$$

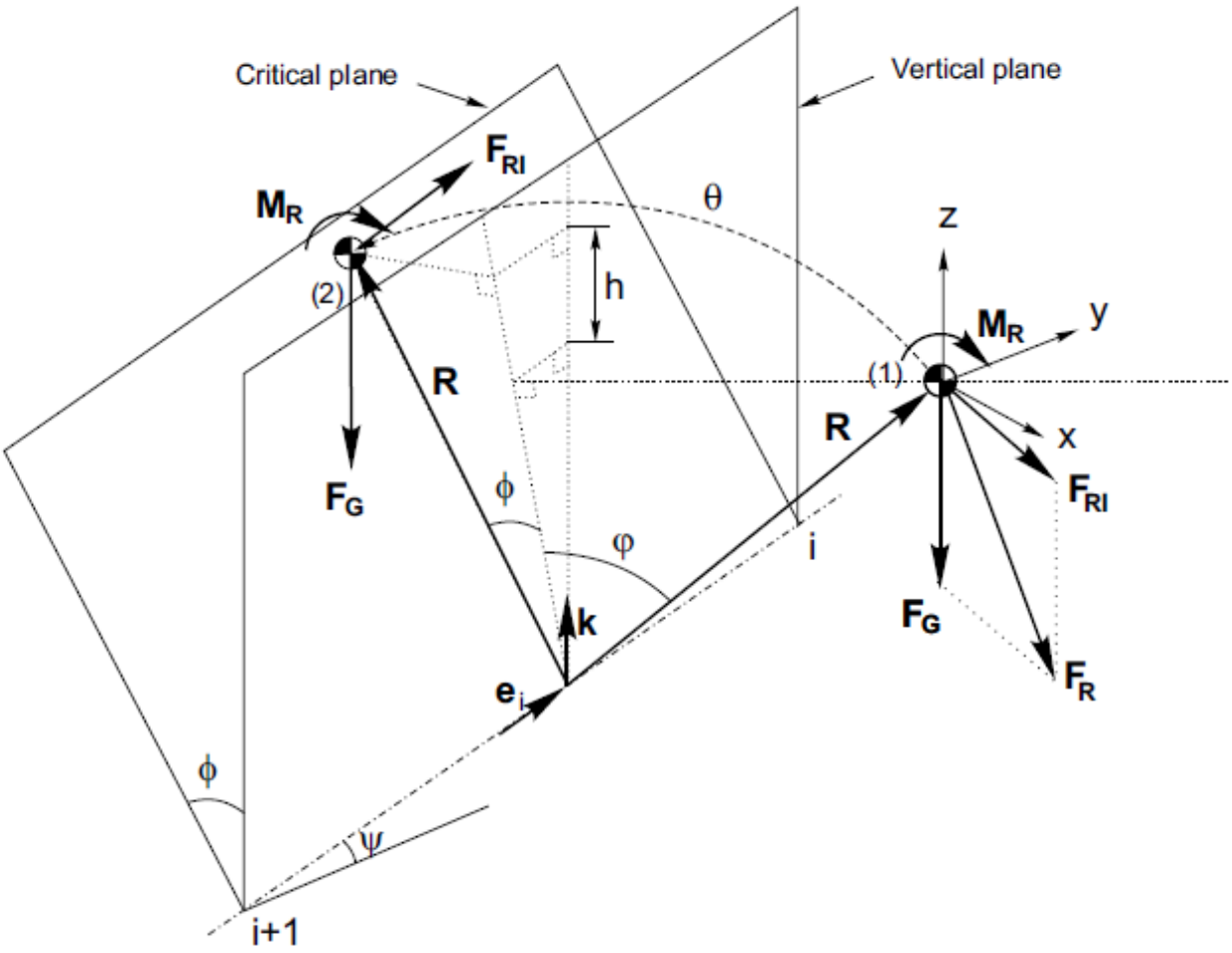
Το τελευταίο κριτήριο δυναμικής ευστάθειας είναι το **κανονικοποιημένο περιθώριο ευστάθειας δυναμικής ενέργειας** (**normalized dynamic energy stability margin S_{NDESM}**), των Garcia και Gonzalez de Santos[29]. Το κριτήριο αυτό ορίζεται ως το ελάχιστο ποσό ενέργειας το οποίο απαιτείται για να ανατρέψει το ρομπότ ως προς μία πλευρά του πολυγώνου υποστήριξης, κανονικοποιημένο ως προς το βάρος του ρομπότ. Η σχέση η οποία υπολογίζει το κριτήριο αυτό είναι η

$$S_{NDESM} = \frac{min(E_i)}{mg} \quad (2.10)$$

Η ποσότητα E_i είναι η μηχανική ενέργεια που απαιτείται για να ανατρέψει το ρομπότ ως προς την i-οστή πλευρά του πολυγώνου υποστήριξης. Η E_i δίνεται από τη σχέση

$$E_i = mg|\boldsymbol{R}|(\cos\phi - cos\varphi)cos\Psi + (\boldsymbol{F_{RI}}\cdot\boldsymbol{t})|\boldsymbol{R}|\theta + (\boldsymbol{M_R}\cdot\boldsymbol{e_i})\theta - \frac{1}{2}I_i\omega_i^2 \quad (2.11)$$

όπου $\boldsymbol{R}$ είναι το διάνυσμα το οποίο έχει τέλος το κέντρο μάζας και είναι κάθετο στην i-οστή πλευρά του πολυγώνου υποστήριξης. $\boldsymbol{F_{RI}}$ είναι η μη βαρυτική συνιστώσα της συνολικής δύναμης αντίδρασης από το έδαφος $\boldsymbol{F_R}$ που ασκείται από το κέντρο μάζας του ρομπότ στο έδαφος, I_i είναι η ροπή αδράνειας γύρω από τον άξονα περιστροφής i, ω_i είναι η γωνιακή ταχύτητα του κέντρου μάζας, Ψ είναι γωνία κλίσης της i-οστής πλευράς του πολυγώνου υποστήριξης και ϕ, φ και θ είναι οι γωνίες στροφής γύρω από τον άξονα περιστροφής i. φ είναι η γωνία που απαιτείται για να βρεθεί το κέντρο μάζας στο κάθετο επίπεδο, ϕ είναι η γωνία μεταξύ κάθετου και κρίσιμου επιπέδου (το επίπεδο όπου η συνολική ροπή που ασκείται στο κέντρο μάζας εξαφανίζεται). Τέλος, θ είναι το άθροισμα των γωνιών ϕ και φ. Το μοναδιαίο διάνυσμα $\boldsymbol{t}$ είναι εφαπτόμενο στην τροχιά του κέντρου μάζας και e_i είναι το μοναδιαίο διάνυσμα το οποίο διατρέχει το πολύγωνο υποστήριξης κατά την ωρολογιακή φορά (βλέπε σχήμα 2.4).



Σχήμα 2.4 : Γεωμετρική επεξήγηση για τον υπολογισμό του S_{NDESM} [9].

Οι τρεις πρώτοι όροι της εξίσωσης (2.11) είναι η δυναμική ενέργεια που απαιτείται για να ανατρέψει το ρομπότ. Η δυναμική αυτή ενέργεια προκαλείται από βαρυτικές και μη δυνάμεις και ροπές. Ο τέταρτος όρος είναι η κινητική ενέργεια που απαιτείται. Το S_{NDESM} είναι το μόνο κριτήριο δυναμικής ευστάθειας που λαμβάνει υπόψιν του εξωτερικές διαταραχές. Έχει αξιολογηθεί μέσω εξομοίωσης και αποδείχτηκε ότι ποσοτικοποιεί την ευστάθεια του ρομπότ με ακρίβεια σε ανώμαλες επιφάνειες, παρουσία εξωτερικών διαταραχών (και δυναμικών φαινομένων εξαιτίας ενός βραχίονα στο σώμα του ρομπότ).

2.3 Αργή στατική βάδιση

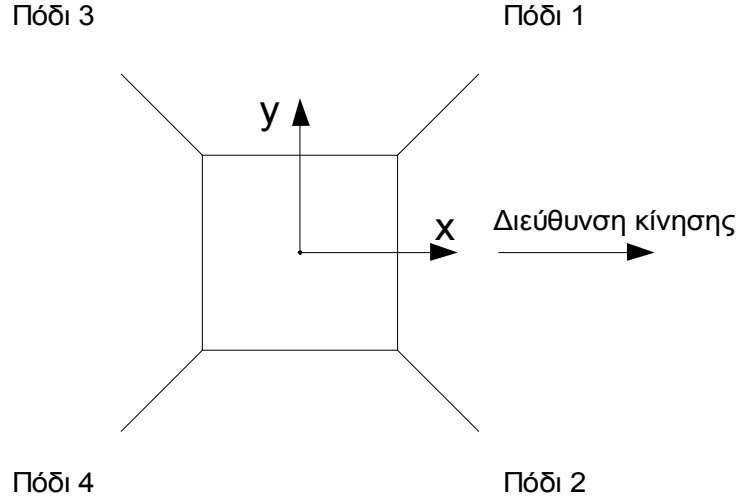
Στην ενότητα αυτή περιγράφουμε διάφορες χρήσιμες έννοιες και ορισμούς που θα μας οδηγήσουν στο φορμαλισμό της αργής στατικής βάδισης (standard gait, crawl gait). Στις επόμενες ενότητες, θα κάνουμε χρήση των ορισμών αυτών για να περιγράψουμε κι άλλους μηχανισμούς βάδισης.

2.3.1 Ακολουθίες γεγονότων

Στο πεδίο της κίνησης με πόδια, η βάδιση ορίζεται ως ένα επαναλαμβανόμενο πρότυπο τοποθέτησης ποδιών. Ένας πιο ακριβής ορισμός έχει δοθεί από τους Song και Waldron [15] και τον παραθέτουμε παρακάτω.

Ορισμός 2.1 Μία βάδιση ορίζεται από το χρόνο και τη θέση τοποθέτησης και ανύψωσης κάθε ποδιού, σε συγχρονισμό με την κίνηση του σώματος και στους έξι βαθμούς ελευθερίας του, με σκοπό τη μετατόπιση του σώματος από ένα σημείο σε ένα άλλο.

Για την κίνηση με πόδια, έχει εισαχθεί η έννοια της ακολουθίας γεγονότων. Ως γεγονός ορίζεται η ανύψωση ή τοποθέτηση ενός ποδιού. Για ένα ρομπότ με n πόδια, η τοποθέτηση του ποδιού i ορίζεται ως γεγονός i, ενώ η ανύψωση του ορίζεται ως γεγονός $i + n$. Επομένως, μπορούν να προκύψουν $2n$ διαφορετικά γεγονότα. Αν δύο γεγονότα προκύψουν ταυτόχρονα, η βάδιση καλείται ιδιόμορφη, σε αντίθεση με τη μη ιδιόμορφη βάδιση, όπου κάθε χρονική στιγμή συμβαίνει ένα γεγονός. Για ένα ρομπότ με n πόδια, το πλήθος των μη ιδιόμορφων βαδίσεων είναι $2n!$, όσες και οι μεταθέσεις $2n$ γεγονότων. Ορίζοντας το πρώτο γεγονός, υπάρχουν $(2n – 1)!$ μη ιδιόμορφες βαδίσεις, όπου για ένα ρομπότ με τέσσερα πόδια είναι 5040.



Σχήμα 2.5 : Κάτοψη και αρίθμηση των ποδιών ενός τετράποδου ρομπότ.

Παραδείγματα μη ιδιόμορφων βαδίσεων αποτελούν η αργή στατική βάδισης (standard gait, crawl gait) και η ασυνεχής βάδιση δύο και τεσσάρων φάσεων. Στην τελευταία, το σώμα του ρομπότ μετατοπίζεται μπροστά κάθε φορά που ολοκληρώνουν την ανύψωση και τοποθέτηση τους δύο πόδια (δύο φάσεις) ή κάθε φορά που η διαδικασία αυτή ολοκληρώνεται από ένα πόδι (τέσσερις φάσεις). Σημειώνεται, ότι η ανύψωση του ποδιού *i*, πρέπει να ακολουθηθεί από την τοποθέτηση του, δηλαδή μετά το γεγονός *i* συμβαίνει το γεγονός *i+n.*

Ένα τετράποδο, για να είναι στατικά ευσταθές, πρέπει να διατηρεί πάντα τρία πόδια σε φάση υποστήριξης, ή αλλιώς ένα μόνο πόδι σε φάση μετάβασης. Αυτό σημαίνει ότι, μετά την ανύψωση του ποδιού i (γεγονός i + n), πρέπει να συμβεί η τοποθέτηση του (γεγονός i). Αυτό το χαρακτηριστικό μειώνει δραματικά τον αριθμό των ευσταθών συνδυασμών σε (n–1)! , όπου για ένα ρομπότ με τέσσερα πόδια γίνεται έξι ακολουθίες γεγονότων. Αυτές είναι οι 5 – 1 – 6 – 2 – 7 – 3 – 8 – 4, 5 – 1 – 7 – 3 – 6 – 2 – 8 – 4, 5 – 1 – 7 – 3 – 8 – 4 – 6 – 2, 5 – 1 – 8 – 4 – 6 – 2 – 7 – 3, 5 – 1 – 6 – 2 – 8 – 4 – 7 – 3 και 5 – 1 – 8 – 4 – 7 – 3 – 6 – 2.

Οι McGhee και Frank [8], απέδειξαν ότι το βέλτιστο περιθώριο στατικής ευστάθειας επιτυγχάνεται από μία κανονική βάδιση, όπου μόλις τοποθετηθεί ένα πόδι, ανυψώνεται αμέσως το επόμενο στην ακολουθία γεγονότων, που παραθέτουμε παραπάνω. Ως κανονική ορίζεται η βάδιση, όπου ο λόγος του χρόνου υποστήριξης προς το συνολικό χρόνο μίας βάδισης (περίοδος), είναι ίδιος για όλα τα πόδια.

2.3.2 Ορισμοί χρήσιμοι για την περιγραφή βαδίσεων

Στην υποενότητα αυτή θα παραθέσουμε διάφορους ορισμούς, οι οποίοι θα μας βοηθήσουν να περιγράψουμε φορμαλιστικά διάφορους τύπους βαδίσεων. Οι ορισμοί αυτοί περιγράφονται και σχηματικά στο σχήμα 2.6.

<u>Ορισμός 2.2</u> Ο λόγος λειτουργίας β_i, του ποδιού i ορίζεται ως το κλάσμα του χρόνου όπου το πόδι βρίσκεται στο έδαφος. Αν το β_i είναι ίδιο για όλα τα πόδια, τότε η βάδιση είναι κανονική.

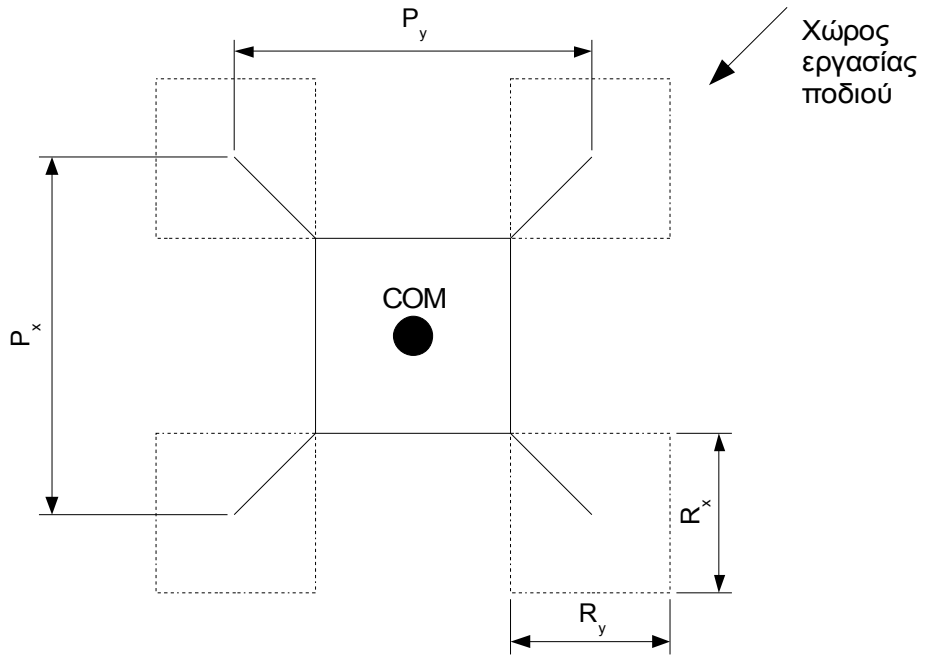
<u>Ορισμός 2.3</u> Η φάση του ποδιού i, φ_i, είναι ο χρόνος που καθυστερεί η τοποθέτηση του ποδιού i, ως προς την τοποθέτηση του ποδιού 1, κανονικοποιημένος ως προς το συνολικό χρόνο βάδισης $T_{locomotion}$. Στον ορισμό αυτόν, θεωρούμε ότι το πόδι 1 είναι το πόδι αναφοράς (αυτό που ανυψώνεται πρώτο).

<u>Ορισμός 2.4</u> Το μήκος κίνησης των ποδιών, R, είναι η απόσταση που διανύει ένα πόδι σε σχέση με το σώμα του ρομπότ, κατά τη φάση υποστήριξης. Το R πρέπει να βρίσκεται μέσα στο χώρο εργασίας του ρομπότ που ορίζεται από τις παραμέτρους R_x και R_y.

<u>Ορισμός 2.5</u> Η παράμετρος P ορίζεται ως η απόσταση μεταξύ των κέντρων των χώρων εργασίας γειτονικών ποδιών. P_x είναι η απόσταση για πόδια που βρίσκονται στην ίδια πλευρά του σώματος του ρομπότ και P_y είναι η απόσταση για πόδια που βρίσκονται σε αντίθετες πλευρές του σώματος του ρομπότ.

<u>Ορισμός 2.6</u> Το μήκος κίνησης του σώματος του ρομπότ λ, ορίζεται ως το μήκος του διανύει το κέντρο μάζας του ρομπότ κατά τη διάρκεια μίας βάδισης. Αν η βάδιση είναι περιοδική, τότε

$$\lambda = \frac{R}{\beta} \quad (2.12)$$



Σχήμα 2.6 : Ορισμοί αργής στατικής βάδισης

2.3.3 Φορμαλισμός αργής στατικής βάδισης

Με τους ορισμούς που μόλις δώσαμε και δεδομένου ότι οι χώροι εργασίας των ποδιών δεν επικαλύπτονται, δηλαδή $R \leq P$, μπορούμε να περιγράψουμε φορμαλιστικά την αργή στατική βάδιση.

Συγκεκριμένα, η φάση κάθε ποδιού κατά την αργή στατική βάδιση είναι

$$\begin{aligned} \varphi_1 &= 0 \\ \varphi_2 &= \frac{1}{2} \\ \varphi_3 &= \beta \\ \varphi_4 &= F(\beta - \frac{1}{2}) \end{aligned} \qquad (2.13)$$

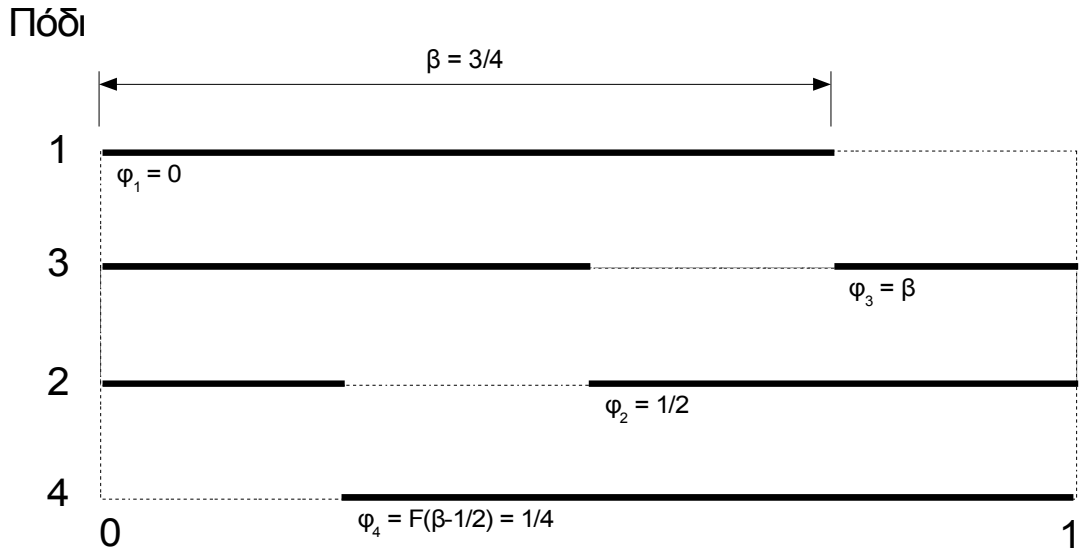
όπου F είναι η κλασματική συνάρτηση που ορίζεται ως

$$Y = F(X) = \begin{cases} \textit{το κλασματικό μέρος του } X & \textit{αν } X \geq 0 \\ 1 - \textit{το κλασματικό μέρος του} |X| & \textit{αν } X < 0 \end{cases} \qquad (2.14)$$

Έχει αποδειχθεί ότι το κριτήριο S_{LSM} είναι βέλτιστο για την αργή στατική βάδιση, όταν $¾ \leq \beta < 1$. Συγκεκριμένα, το S_{LSM} δίνεται από τη σχέση

$$S_{LSM} = (\beta - \frac{3}{4})\lambda, \quad \frac{3}{4} \leq \beta < 1 \qquad (2.15)$$

Η συνθήκη $¾ \leq \beta < 1$ είναι υποχρεωτική για να διατηρείται η στατική ευστάθεια. Κάθε πόδι πρέπει να βρίσκεται σε φάση υποστήριξης για τουλάχιστον ¾ του συνολικού χρόνου βάδισης $T_{locomotion}$. Το παρακάτω διάγραμμα βάδισης εξηγεί καλύτερα τη σχέση 2.13. Στο διάγραμμα αυτό η συνεχής γραμμή δείχνει ότι το πόδι βρίσκεται σε φάση υποστήριξης. Όταν διακόπτεται δείχνει το πόδι βρίσκεται σε φάση μετάβασης και όταν συνεχίζεται πάλι, το πόδι αρχίζει πάλι τη φάση υποστήριξης.



Σχήμα 2.7 : Διάγραμμα βάδισης αργής στατικής βάδισης. Ο χρόνος είναι κανονικοποιημένος ως προς τη περίοδο $T_{locomotion}$.

2.4 Ταχεία βάδιση – Trot gait

Η ταχεία βάδιση αποτελεί έναν ιδιόμορφο δυναμικό μηχανισμό κίνησης. Δυναμικοί μηχανισμοί βάδισης ορίζονται οι μηχανισμοί που δεν μπορούμε να μελετήσουμε την ευστάθεια τους με χρήση κριτηρίων στατικής ευστάθειας, αφού πλέον δεν υπάρχει πολύγωνο υποστήριξης. Κατά την ταχεία βάδιση, σε φάση υποστήριξης μπορούν να βρεθούν μόνο δύο πόδια, για ένα ποσοστό β του συνολικού χρόνου βάδισης τους $T_{locomotion}$. Υπάρχουν επίσης δύο διαστήματα όπου κανένα πόδι δε βρίσκεται στο έδαφος. Αυτά είναι από β έως φ_2 και από $\varphi_2+\beta$ έως 1. Πιο συγκεκριμένα, ισχύει

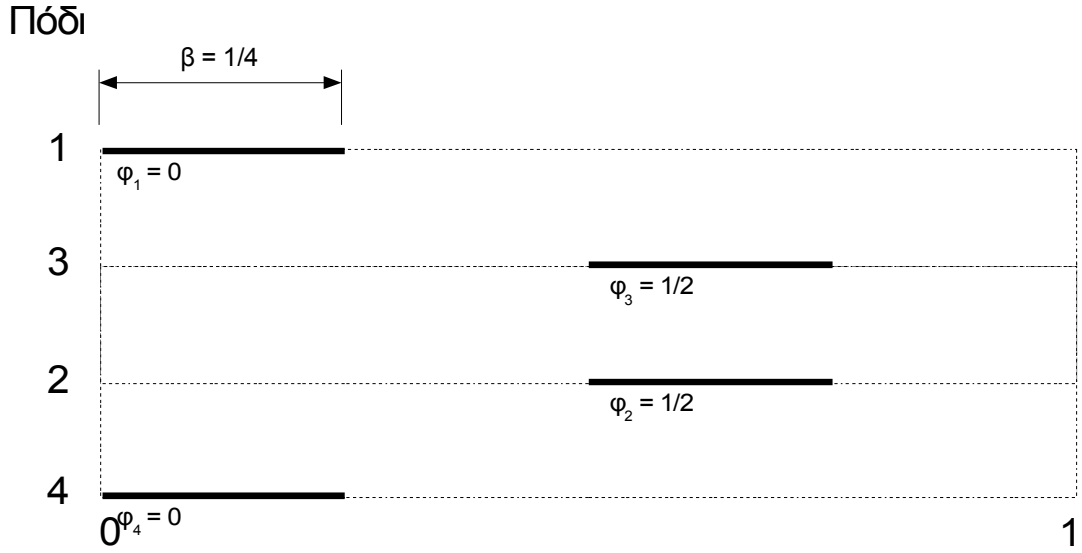
$$\begin{aligned} \varphi_1 &= 0 \\ \varphi_2 &= \frac{1}{2} \\ \varphi_3 &= \frac{1}{2} \\ \varphi_4 &= 0 \end{aligned} \qquad (2.16)$$

Συνήθως επιλέγεται η βάδιση να είναι κανονική, δηλαδή

$$\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta \quad (2.17)$$

όπου $\beta \leq ½$.

Ο συγκεκριμένος τύπος βάδισης επιτρέπει την επίτευξη μεγαλύτερων ταχυτήτων από την αργή στατική βάδιση. Πλέον όμως χρειαζόμαστε πιο πολύπλοκους μηχανισμούς ελέγχου και μηχανισμούς απορρόφησης κραδασμών (αποσβεστήρες). Παρακάτω παραθέτουμε το διάγραμμα βάδισης της ταχείας βάδισης. Χρησιμοποιούμε $\beta = 25\%$.



Σχήμα 2.8 : Διάγραμμα ταχείας βάδισης. Ο χρόνος είναι κανονικοποιημένος ως προς τη περίοδο $T_{locomotion}$.

Στο παραπάνω διάγραμμα παραθέσαμε τον τετραγωνικό τύπο της ταχείας βάδισης, όπου τα ζεύγη ποδιών 1 – 4 και 2 – 3 είναι έχουν ίδιες φάσεις φ_i. Στον απλό τύπο ταχείας βάδισης, οι φάσεις των των ποδιών στα ζεύγη 1 – 4 και 2 – 3 έχουν μία μικρή διαφορά.

2.5 Καλπασμός – Gallop gait

Ο καλπασμός αποτελεί κι αυτός ένα δυναμικό τρόπο βάδισης. Κατά τον καλπασμό μόνο ένα πόδι βρίσκεται σε φάση υποστήριξης για ένα ποσοστό β του συνολικού χρόνου βάδισης $T_{locomotion}$. Για τις φάσεις των ποδιών ισχύει

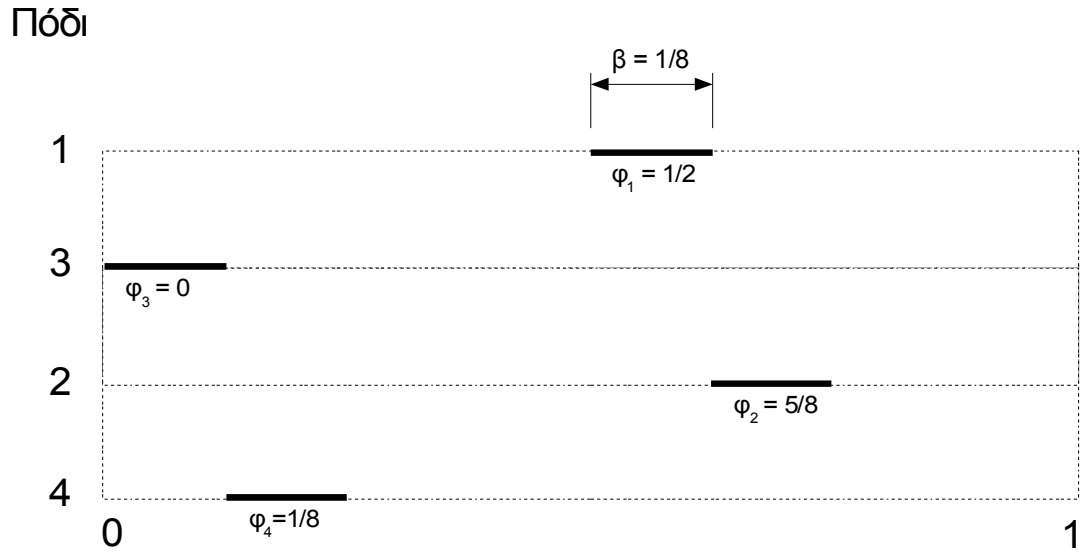
$$\begin{aligned} \varphi_1 &= \frac{1}{2} \\ \varphi_2 &= \varphi_1 + \beta = \frac{1}{2} + \beta \\ \varphi_3 &= 0 \\ \varphi_4 &= \beta \end{aligned} \quad (2.18)$$

Συνήθως κι αυτή η βάδιση επιλέγεται να είναι κανονική, δηλαδή

$$\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta \quad (2.19)$$

όπου $\beta \leq 1/4$.

Ο καλπασμός θεωρείται ο ταχύτερος τύπος βάδισης. Υπάρχουν κι εδώ δύο διαστήματα όπου κανένα πόδι δε βρίσκεται σε φάση υποστήριξης. Αυτά είναι από $\varphi_4+\beta$ έως φ_1 και από $\varphi_2+\beta$ έως 1. Όμως, ο έλεγχος των παραμέτρων βάδισης γίνεται ακόμα πιο δύσκολος και οι κραδασμοί είναι πιο έντονοι. Παρακάτω παραθέτουμε το διάγραμμα βάδισης του καλπασμού για β = 12.5%.



Σχήμα 2.9 : Διάγραμμα βάδισης καλπασμού. Ο χρόνος είναι κανονικοποιημένος ως προς τη περίοδο $T_{locomotion}$.

3 *Ενισχυτική μάθηση*

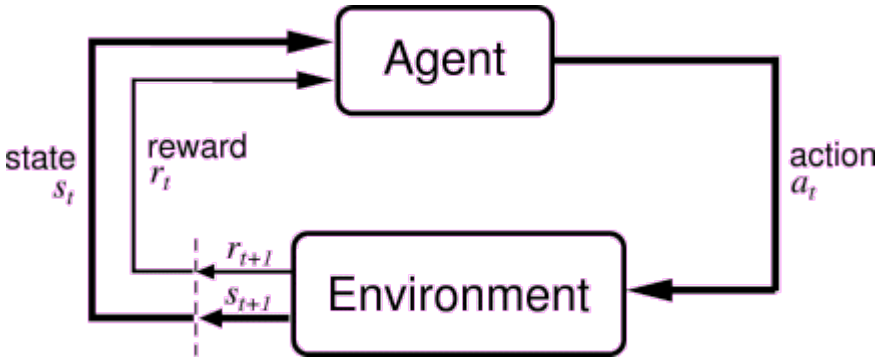
Η ιδέα της μάθησης μέσω της αλληλεπίδρασης μας με το περιβάλλον είναι ίσως η πρώτη που μας έρχεται στο μυαλό, όταν σκεφτόμαστε τη φύση της μάθησης. Ο άνθρωπος μπορεί να επέμβει στο περιβάλλον μέσω της αισθησιοκινητικής σύνδεσης του με αυτό. Αλληλεπιδρώντας λοιπόν με αυτό μπορεί να αποκτήσει γνώση για σχέσεις αιτίας – αποτελέσματος, για τις επιπτώσεις των δράσεων του (μέσω των οποίων αλληλεπιδρά με το περιβάλλον) και τι να πράξει ώστε να επιτύχει τους στόχους του. Οι R. Sutton και A. Barto έχτισαν τη θεωρία της ενισχυτικής μάθησης στο [16], η οποία βασίζεται στην παραπάνω ιδέα. Στο κεφάλαιο αυτό κάνουμε μία συνοπτική παρουσίαση της θεωρίας και των αλγορίθμων στις οποίες βασίζεται η ενισχυτική μάθηση.

3.1 Το πρόβλημα της ενισχυτικής μάθησης

Η ενισχυτική μάθηση είναι μία μέθοδος μη επιβλεπόμενης μηχανικής μάθησης. Ο πράκτορας δε γνωρίζει εκ των προτέρων τις βέλτιστες δράσεις του, αλλά προσπαθεί να τις μάθει αλληλεπιδρώντας συνεχώς με το περιβάλλον. Το περιβάλλον με τη σειρά του δίνει επιβραβεύσεις, δηλαδή αριθμητικές τιμές τις οποίες ο πράκτορας επιδιώκει να μεγιστοποιήσει με την πάροδο του χρόνου.

Πιο συγκεκριμένα, ο πράκτορας αλληλεπιδρά με το περιβάλλον σε διακριτές χρονικές στιγμές, $t = 0, 1, 2, 3...$ Κάθε χρονική στιγμή t βρίσκεται σε κάποια κατάσταση, η οποία αποτελεί μία αναπαράσταση του περιβάλλοντος. Η κατάσταση αυτή συμβολίζεται $s_t \in \boldsymbol{S}$, όπου $\boldsymbol{S}$ είναι το σύνολο των πιθανών καταστάσεων στις οποίες μπορεί να βρεθεί. Σε κάθε κατάσταση πρέπει να επιλέξει μία δράση $a_t \in A(s_t)$, όπου $\boldsymbol{A}(s_t)$ το σύνολο διαθέσιμων δράσεων στην κατάσταση s_t. Επιλέγοντας μία

δράση, ο πράκτορας μεταβαίνει σε μία νέα κατάσταση την επόμενη χρονική στιγμή. Σαν αποτέλεσμα της δράσης αυτής, λαμβάνει από το περιβάλλον μία πραγματική αριθμητική τιμή επιβράβευσης, r_{t+1}, και ο πράκτορας βρίσκεται σε μία νέα κατάσταση s_{t+1}. Στο σχήμα 3.1 βλέπουμε την αναπαράσταση αυτής της αλληλεπίδρασης.



Σχήμα 3.1 : Σχηματική αναπαράσταση της αλληλεπίδρασης ενός πράκτορα με το περιβάλλον του [16].

Ο πράκτορας κάθε χρονική στιγμή εφαρμόζει μία αντιστοίχιση από καταστάσεις σε πιθανότητες να επιλέξει κάποια από τις διαθέσιμες δράσεις. Αυτή η αντιστοίχιση ονομάζεται πολιτική και οι μέθοδοι της ενισχυτικής μάθησης εξηγούν τους τρόπους με τους οποίους ο πράκτορας αλλάζει την πολιτική του, ως αποτέλεσμα της εμπειρίας που αποκτάει αλληλεπιδρώντας με το περιβάλλον. Με λίγα λόγια, ο πράκτορας προσπαθεί να μεγιστοποιήσει μία συνάρτηση κέρδους με σκοπό να επιλέγει πάντα βέλτιστες δράσεις. Εκτελώντας όμως, πάντα μία «άπληστη» (greedy) πολιτική μπορεί να οδηγηθούμε σε υποβέλτιστες λύσεις. Για το λόγο αυτό, πρέπει να εξερευνήσουμε το χώρο των δράσεων – να αποφασίσουμε δηλαδή τυχαία την επόμενη δράση – με σκοπό να βρούμε δράσεις, οι οποίες μπορεί να αυξήσουν το συνολικό κέρδος σε βάθος χρόνου. Αυτό είναι και μία από τις μεγαλύτερες προκλήσεις της ενισχυτικής μάθησης, ο συμβιβασμός μεταξύ της αναζήτησης νέας γνώσης και της εκμετάλλευσης της ήδη υπάρχουσας.

Στο σημείο αυτό πρέπει να τονιστεί ότι το πρόβλημα της ενισχυτικής μάθησης ικανοποιεί την ιδιότητα του Markov, αφού η μετάβαση από μία κατάσταση στην επόμενη, εξαρτάται μόνο από την τελευταία και όχι όλες τις προηγούμενες. Δηλαδή, δεδομένης της κατάστασης s που βρισκόμαστε και της δράσης a που θα εκτελέσουμε, η πιθανότητα να βρεθούμε σε κάποια από τις επόμενες δυνατές καταστάσεις είναι

$$P_{ss'}^{\alpha} = Pr\{s_{t+1} = s' | s_t = s, \alpha_t = \alpha\} \quad (3.1)$$

Έτσι λοιπόν το πρόβλημα της ενισχυτικής μάθησης μπορεί να αναχθεί σε μία αλυσίδα Markov, όπου ο πράκτορας προσπαθεί να προσεγγίσει τον πίνακα μετάβασης $\boldsymbol{T : S \times A \times S \rightarrow [0, 1]}$, ο οποίος δηλώνει την πιθανότητα να βρεθούμε από μία κατάσταση *s* σε μία νέα *s′*, εκτελώντας τη δράση *α*.

Οι αλγόριθμοι ενισχυτικής μάθησης μπορούν να εφαρμοστούν αποδοτικά και χωρίς να υπάρχει κάποιο μοντέλο του περιβάλλοντος. Για το λόγο αυτό παρουσιάζουν ιδιαίτερο ενδιαφέρον και υπάρχει αρκετή έρευνα πάνω σε αυτό το πεδίο.

3.2 Στόχοι και επιβράβευση

Στην ενισχυτική μάθηση, ο στόχος του πράκτορα μοντελοποιείται ως το σήμα επιβράβευσης r_t, το οποίο λαμβάνει από το περιβάλλον κάθε χρονική στιγμή. Δηλαδή προσπαθεί να μεγιστοποιήσει το συνολικό άθροισμα επιβραβεύσεων το οποίο λαμβάνει σε βάθος χρόνου και όχι άμεσα. Για το λόγο αυτό, πρέπει να σχεδιάσουμε τη συνάρτηση επιβράβευσης με τέτοιο τρόπο, ώστε η μεγιστοποίηση του συνολικού αθροίσματος επιβράβευσης να είναι ταυτόσημη με την επίτευξη του στόχου.

Η ιδέα της χρήσης ενός σήματος επιβράβευσης για τη μοντελοποίηση του στόχου ενός πράκτορα είναι από τις πιο χαρακτηριστικές της ενισχυτικής μάθησης. Αν και αυτός ο τρόπος μοντελοποίησης στόχων φαίνεται αρχικά περιοριστικός, στην πραγματικότητα έχει αποδειχθεί ότι είναι πολύ ευέλικτος και ευρέως χρησιμοποιούμενος.

Τα προβλήματα ενισχυτικής μάθησης χωρίζονται σε δύο κατηγορίες, στα προβλήματα πεπερασμένου και άπειρου ορίζοντα. Στα προβλήματα πεπερασμένου ορίζοντα, η άθροιση των επιβραβεύσεων γίνεται για πεπερασμένο αριθμό βημάτων, ενώ στα προβλήματα άπειρου ορίζοντα η άθροιση γίνεται επ ’ αόριστον.

3.2.1 Προβλήματα πεπερασμένου ορίζοντα

Στα προβλήματα πεπερασμένου ορίζοντα, κάθε χρονική στιγμή *t* ο πράκτορας προσπαθεί να

μεγιστοποιήσει το συνολικό άθροισμα ανταμοιβών R_t για τα τελευταία N-t βήματα, όπου

$$R_t = r_{t+1} + r_{t+2} + r_{t+3} + ... + r_N \quad (3.2)$$

Το σήμα επιβράβευσης όμως είναι τυχαία μεταβλητή λόγω της ύπαρξης θορύβου. Για το λόγο αυτό ζητάμε τη μεγιστοποίηση της εκτιμώμενης τιμής

$$E\{\sum_{i=t+1}^{N} r_i\} \quad (3.3)$$

Αυτά τα προβλήματα μπορούν να λυθούν αναλυτικά με δυναμικό προγραμματισμό, όμως αρκετές φορές η πολυπλοκότητα του προβλήματος, η οποία εξαρτάται από τον αριθμό βημάτων N και τον πλήθος του συνόλου καταστάσεων S, καθιστά τη χρήση του απαγορευτική, καθιστώντας αναγκαία την εφαρμογή τεχνικών όπως αυτή της ενισχυτικής μάθησης.

3.2.2 Προβλήματα άπειρου ορίζοντα

Στα προβλήματα αυτά περιλαμβάνονται περιπτώσεις στις οποίες ο αριθμός των πιθανών καταστάσεων είναι αρκετά μεγάλος. Στα προβλήματα αυτά ο πράκτορας προσπαθεί να μεγιστοποιήσει την ποσότητα

$$E\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}\} \quad (3.4)$$

Η παράμετρος γ παίρνει τιμές στο διάστημα [0, 1] και καθορίζει την αξία που έχουν αυτήν τη χρονική στιγμή μελλοντικές επιβραβεύσεις. Μια επιβράβευση η οποία θα ληφθεί σε k χρονικές στιγμές από τώρα, αξίζει γ^{k-1} φορές την τιμή που θα είχε, αν τη λάμβανε ο πράκτορας τώρα.

Αν $\gamma = 0$, ο πράκτορας είναι « μυωπικός », δηλαδή τον ενδιαφέρει η μεγιστοποίηση μόνο των άμεσων επιβραβεύσεων που λαμβάνει, δηλαδή προσπαθεί να μάθει πώς να διαλέγει τη δράση α_t, η οποία θα μεγιστοποιήσει μόνο την άμεση επιβράβευση, χωρίς να λαμβάνει καθόλου υπόψιν τις επόμενες. Αν όντως κάθε δράση α_t επηρεάζει μόνο την άμεση επιβράβευση – και καθόλου τις μελλοντικές – τότε ο πράκτορας μπορεί να μεγιστοποιήσει την (3.4). Εν γένει όμως, μία τέτοια

συμπεριφορά, μας αποτρέπει να λάβουμε στο μέλλον επιβραβεύσεις, οι οποίες θα οδηγούσαν σε βέλτιστη τιμή της (3.4). Όσο η παράμετρος *γ προσεγγίζει* το 1, ο πράκτορας λαμβάνει περισσότερο υπόψιν του τις μελλοντικές επιβραβεύσεις.

3.3 Συναρτήσεις αξίας

Σχεδόν όλοι οι αλγόριθμοι ενισχυτικής μάθησης βασίζονται στην εκτίμηση συναρτήσεων αξιών, δηλαδή συναρτήσεων καταστάσεων (ή ζευγαριών κατάστασης – δράσης), οι οποίες είναι ενδεικτικές του πόσο καλό είναι για τον πράκτορα να βρίσκεται σε αυτή την κατάσταση. Το « πόσο καλό » καθορίζεται από το συνολικό άθροισμα επιβραβεύσεων, το οποίο θα λάβει στο μέλλον ο πράκτορας και το οποίο εξαρτάται από τις δράσεις τις οποίες θα λάβει. Από αυτές τις συναρτήσεις λοιπόν, ο πράκτορας μπορεί να αποφασίζει δράσεις οι οποίες θα τον οδηγήσουν σε βάθος χρόνου στο στόχο του και έτσι προκύπτουν οι επονομαζόμενες πολιτικές δράσης ή απλώς πολιτικές (policies) του πρακτορικού συστήματος.

Η πολιτική, π, είναι μία αντιστοίχιση από κάθε κατάσταση $s \in S$, και δράση $\alpha \in A(S)$ στην πιθανότητα $\pi(s, \alpha)$, δηλαδή στην πιθανότητα να αποφασίσει τη δράση α, όταν είναι στην κατάσταση s. Επομένως, η αξία $V^{\pi}(s)$, μίας κατάστασης s, όταν ακολουθεί την πολιτική π, είναι το εκτιμώμενο συνολικό άθροισμα επιβραβεύσεων που θα λάβει, αν ξεκινήσει από την κατάσταση s και ακολουθήσει την πολιτική π μετά. Για μία αλυσίδα Markov, η $V^{\pi}(s)$ ορίζεται ως

$$V^{\pi}(s) = E_{\pi}\{R_t | s_t = s\} = E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\} \quad (3.5)$$

δεδομένου ότι ακολουθεί την πολιτική π, όπου $E_{\pi}\{\}$ συμβολίζει την εκτιμώμενη αξία. Τονίζουμε ότι η αξία της τερματικής κατάστασης είναι πάντα 0. Η $V^{\pi}(s)$ καλείται συνάρτηση κατάστασης – αξίας για την πολιτική π.

Όμοια, η αξία μίας δράσης α σε μία κατάσταση s, όταν ακολουθούμε την πολιτική π, συμβολίζεται με $Q^{\pi}(s,\alpha)$ και ορίζεται ως

$$Q^{\pi}(s, \alpha) = E_{\pi}\{R_t | s_t = s, \alpha_t = \alpha\} = E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, \alpha_t = \alpha\} \quad (3.6)$$

και είναι το εκτιμώμενο συνολικό άθροισμα επιβραβεύσεων που θα λάβει, αν ξεκινήσει από την κατάσταση *s,* εκτελέσει τη δράση *α* και ακολουθήσει την πολιτική *π* μετά.

Οι τιμές αυτών των συναρτήσεων προσεγγίζονται τόσο καλύτερα, όσο περισσότερη είναι η γνώσει που συγκεντρώνει ο πράκτορας. Όταν το γινόμενο $|S|\cdot|A|$ είναι μικρό, όπου $|\cdot|$ ορίζεται ο πληθάριθμος του συνόλου, οι τιμές $Q^{\pi}(s,\alpha)$ αποθηκεύονται σε πίνακες. Όσες περισσότερες φορές εκτελέσει ο πράκτορας μία δράση *α* σε μία κατάσταση *s*, τόσο καλύτερα προσεγγίζει την αξία αυτών των ζευγαριών. Όταν το γινόμενο αυτό γίνει πολύ μεγάλο, τότε χρησιμοποιούμε άλλες τεχνικές, π.χ. προσέγγιση συναρτήσεων (function approximation).

Αποδεικνύεται επίσης, ότι οι συναρτήσεις $V^{\pi}(s)$ και $Q^{\pi}(s,\alpha)$ ικανοποιούν την αναδρομική εξίσωση Bellman, δηλαδή

$$V^{\pi}(s)=\sum_{\alpha}\pi(s,\alpha)\sum_{s'}P^{\alpha}_{ss'}[R^{\alpha}_{ss'}+\gamma V^{\pi}(s')] \quad (3.7)$$

$$Q^{\pi}(s,\alpha)=\sum_{s'}P^{\alpha}_{ss'}[R^{\alpha}_{ss'}+\gamma\sum_{\alpha'}\pi(s',\alpha')Q^{\pi}(s',\alpha')] \quad (3.8)$$

όπου $P^{\alpha}_{ss'}$ η πιθανότητα μετάβασης από την κατάσταση *s* στην *s′*, αν εκτελέσουμε τη δράση *α* και $R^{\alpha}_{ss'}$ η άμεση επιβράβευση που λαμβάνουμε μεταβαίνοντας από την κατάσταση *s* στην *s′*, εκτελώντας τη δράση *α*.

Η εξίσωση Bellman προσθέτει τις επιβραβεύσεις όλων των πιθανών μεταβάσεων από μία κατάσταση *s* σε μία άλλη *s′,* πολλαπλασιασμένα με την πιθανότητα αυτής της μετάβασης, την οποία μας δίνει ο πίνακας μετάβασης του μοντέλου Markov.

3.4 Βέλτιστες συναρτήσεις αξίας

Σε μία πεπερασμένη διαδικασία απόφασης Markov ορίζεται μία σχέση μερικής διάταξης επί των πολιτικών. Μία πολιτική *π* ορίζεται ως καλύτερη από μία άλλη *π′* ή ίση, αν το αναμενόμενο άθροισμα μελλοντικών επιβραβεύσεων είναι μεγαλύτερο ή ίσο από το αντίστοιχο της *π′* για όλες τις καταστάσεις. Με λίγα λόγια,

$$\pi \geq \pi' \text{, εάν και μόνο εάν } V^{\pi}(s) \geq V^{\pi'}(s) \; \forall \, s \in S \quad (3.9)$$

Υπάρχει πάντα μία τουλάχιστον πολιτική η οποία είναι καλύτερη από όλες τις άλλες πολιτικές ή ίδια με αυτές. Αυτή η πολιτική ονομάζεται βέλτιστη πολιτική και θα τη συμβολίζουμε με π^*. Η συνάρτηση αξίας που αντιστοιχεί σε αυτή την πολιτική ονομάζεται βέλτιστη συνάρτηση αξίας και συμβολίζεται με V^*, δηλαδή

$$V^*(s) = \max_{\pi} V^{\pi}(s) \; \forall \, s \in S \quad (3.10)$$

Μπορεί να υπάρχουν διαφορετικές βέλτιστες πολιτικές, αλλά όλες έχουν την ίδια συνάρτηση V^*. Η συνάρτηση αξίας δράσης που αντιστοιχεί σε αυτή την πολιτική ονομάζεται βέλτιστη συνάρτηση αξίας δράσης και συμβολίζεται με Q^*, δηλαδή

$$Q^*(s,\alpha) = \max_{\pi} Q^{\pi}(s,\alpha) \; \forall \, s \in S \wedge \alpha \in A(s) \quad (3.11)$$

Η συνάρτηση αξίας V^* και αξίας δράσης Q^* ικανοποιούν την εξίσωση Bellman για τη βέλτιστη πολιτική π^* και ισχύουν ότι

$$\begin{aligned} V^*(s) &= \max_{\alpha} E\{r_{t+1} + \gamma V^*(s_{t+1}) | s_t = s, \alpha_t = \alpha\} \\ &= \max_{\alpha \in A(s)} \sum_{s'} P^{\alpha}_{ss'} [R^{\alpha}_{ss'} + \gamma V^*(s')] \end{aligned} \quad (3.12)$$

$$\begin{aligned} Q^*(s,\alpha) &= E\{r_{t+1} + \gamma \max_{\alpha'} Q^*(s_{t+1}, \alpha') | s_t = s, \alpha_t = \alpha\} \\ &= \sum_{s'} P^{\alpha}_{ss'} [R^{\alpha}_{ss'} + \gamma \max_{\alpha'} Q^*(s', \alpha')] \end{aligned} \quad (3.13)$$

Για μία πεπερασμένη διαδικασία απόφασης Markov, η (3.12) έχει μοναδική λύση ανεξάρτητη της πολιτικής. Η εξίσωση Bellman είναι στην πραγματικότητα ένα σύστημα εξισώσεων, μία για κάθε κατάσταση. Έτσι, αν υπάρχουν N καταστάσεις, θα υπάρχουν και N εξισώσεις με N αγνώστους. Αν η δυναμική του περιβάλλοντος είναι γνωστή (οι πιθανότητες μετάβασης και οι αντίστοιχες επιβραβεύσεις), τότε το σύστημα μπορεί να λυθεί αναλυτικά. Από τη λύση που θα πάρουμε, μπορούμε ακολουθώντας μία άπληστη πολιτική, να επιλέγουμε τη βέλτιστη δράση σε κάθε κατάσταση.

Δυστυχώς όμως, στη γενική περίπτωση, η δυναμική του μοντέλου Markov δεν είναι γνωστή και για την επίλυση της εξίσωσης Bellman χρησιμοποιείται η ενισχυτική μάθηση. Μάλιστα, αυτή η περίπτωση είναι η πιο συχνή σε προβλήματα βέλτιστου ελέγχου. Οι αλγόριθμοι ενισχυτικής μάθησης προσπαθούν να ανακαλύψουν τη δυναμική του μοντέλου βάσει των παρατηρήσεων που κάνουν για τις καταστάσεις που βρέθηκαν και τις δράσεις που τους οδήγησαν σε αυτές. Για να γίνει αυτό, ο πράκτορας πρέπει να εξερευνήσει το περιβάλλον του και να εκμεταλλευτεί τη γνώση που θα αποκτήσει για να ακολουθήσει μία βέλτιστη πολιτική.

3.5 Επιλογή δράσεων – Εξερεύνηση έναντι εκμετάλλευσης

Ο συμβιβασμός μεταξύ εξερεύνησης και εκμετάλλευσης αποτελεί μία μεγάλη πρόκληση. Ο πράκτορας θα πρέπει να να εξερευνήσει το περιβάλλον για να μάθει βέλτιστες (ή υποβέλτιστες) πολιτικές, αλλά θα πρέπει επίσης να αξιοποιήσει τη γνώση του, για να μεγιστοποιήσει το συνολικό άθροισμα επιβραβεύσεων που θα λάβει.

Υπάρχουν διάφορες μέθοδοι επιλογής δράσεων. Οι πιο ακραίες είναι αυτές της ομοιόμορφης τυχαίας επιλογής δράσεων σε κάθε βήμα και αυτή της άπληστης επιλογής, δηλαδή η επιλογή της δράσης α για την οποία ισχύει

$$\alpha = arg \max_{\alpha' \in A(s)} Q(s, \alpha') \quad (3.14)$$

Παρακάτω παραθέτουμε διάφορες μεθόδους επιλογής δράσεων που έχουν προταθεί.

3.5.1 Τυχαία (ομοιόμορφη) επιλογή

Όταν χρησιμοποιούμε τυχαία εξερεύνηση, συνήθως χρησιμοποιούμε τη μέθοδο *ε-greedy* ή την *ε-decreasing*. Οι μέθοδοι αυτοί αποσκοπούν στο να επιτευχθεί μία ισορροπία μεταξύ της εξερεύνησης και της εκμετάλλευσης και τις εξηγούμε παρακάτω.

3.5.1.1 ε-greedy

Η *ε-greedy* ελέγχει αυτήν την ισορροπία μέσω του παράγοντα *ε*, όπου *ε* ϵ [0,1]. Όταν ο πράκτορας καλείται να επιλέξει μία δράση, τότε θα επιλέξει σύμφωνα με την (3.14) με πιθανότητα *1-ε* και τυχαία με πιθανότητα *ε*.

3.5.1.2 ε-decreasing

Η *ε-decreasing* έρχεται να καλύψει ένα μειονέκτημα της *ε-greedy*. Μετά από κάποιο χρόνο μάθησης, όπου ο πράκτορας έχει προσεγγίσει αρκετά καλά το μοντέλο Markov του συστήματος, ενδεχομένως να μην είναι επιθυμητό να επιλέγουμε τυχαίες δράσεις. Κατά την *ε-decreasing* ο παράγοντας *ε* μειώνεται όσο αυξάνεται ο αριθμός επαναλήψεων της μάθησης. Με τον τρόπο αυτό προάγεται η εξερεύνηση στα αρχικά στάδια της μάθησης και η εκμετάλλευση στα μετέπειτα.

3.5.2 Επιλογή δράσεων βάσει των συναρτήσεων αξιών

Στη μέθοδο αυτή, όσο μεγαλύτερη είναι η αξία μίας δράσης, τόσο μεγαλύτερη είναι η πιθανότητα της να επιλεχθεί. Ισχύει δηλαδή ότι

$$Pr(\alpha_t=\alpha)=\frac{Q_t(s,\alpha)}{\sum_{\alpha'\in A(s)} Q_t(s,\alpha')} \quad (3.15)$$

Όπως πριν, με πιθανότητα *ε* επιλέγεται μία τυχαία δράση και με πιθανότητα *1-ε* μία δράση, βάσει της (3.15). Η κατανομή όμως δεν είναι τυχαία πλέον. Αυτή η μέθοδος μας διασφαλίζει ότι ο πράκτορας μπορεί και να εξερευνά και να εκμεταλλεύεται την ήδη υπάρχουσα γνώση του.

3.5.3 Κατανομή Boltzmann

Στη μέθοδο αυτή, η πιθανότητα επιλογής μίας δράσης είναι

$$Pr(\alpha_t=\alpha)=\frac{e^{Q_t(s,\alpha)/T}}{\sum_{\alpha'\in A(s)} e^{Q_t(s,\alpha')/T}} \quad (3.16)$$

Πάλι, με πιθανότητα *ε* επιλέγεται μία τυχαία δράση και με πιθανότητα *1-ε* μία δράση, βάσει της (3.16). Η παράμετρος *T* ονομάζεται θερμοκρασία και ισχύει *T > 0*. Η θερμοκρασία καθορίζει αν στην επιλογή απόφασης, εκμεταλλευόμαστε την ήδη υπάρχουσα γνώση ή επιλέγουμε τυχαία. Όταν Τ → 0, εκμεταλλευόμαστε την ήδη υπάρχουσα γνώση. Όταν Τ → ∞, έχουμε τυχαία επιλογή δράσεων (εξερεύνηση) από μία ομοιόμορφη κατανομή.

3.6 Επίλυση με δυναμικό προγραμματισμό

Ο όρος δυναμικός προγραμματισμός [30, 31], αναφέρεται σε μία οικογένεια αλγορίθμων, οι οποίοι μπορούν να χρησιμοποιηθούν για να υπολογίσουν βέλτιστες πολιτικές, δεδομένου ενός πλήρως γνωστού μοντέλου Markov. Επειδή στη γενική περίπτωση δεν έχουμε ένα πλήρες μοντέλο Markov, και λόγω του ότι οι αλγόριθμοι αυτοί είναι υπολογιστικά πολύπλοκοι, δε χρησιμοποιούνται για την επίλυση προβλημάτων ενισχυτικής μάθησης. Παρόλα αυτά, αποτελούν τη βάση των αλγορίθμων ενισχυτικής μάθησης και για το λόγο αυτό, τους παρουσιάζουμε.

Στο δυναμικό προγραμματισμό θεωρούμε ότι το περιβάλλον είναι ένα πεπερασμένο πρόβλημα απόφασης Markov. Θεωρούμε επίσης, ότι το σύνολο των καταστάσεων *S* και δράσεων *A(s)*, για *s* ϵ *S* είναι πεπερασμένο και ότι η δυναμική του περιβάλλοντος είναι δεδομένη. Δηλαδή θεωρούμε δεδομένες τις πιθανότητες μετάβασης

$$P^{\alpha}_{ss'} = Pr\{s_{t+1} = s' | s_t = s, \alpha_t = \alpha\} \quad (3.17)$$

και οι άμεσες εκτιμώμενες επιβραβεύσεις

$$R^{\alpha}_{ss'} = E\{r_{t+1} | \alpha_t = \alpha, s_t = s, s_{t+1} = s'\} \quad (3.18)$$

Όπως έχουμε ήδη αναφέρει, για κάθε *s* ϵ *S* ισχύει η εξίσωση (3.7)

$$V^{\pi}(s) = \sum_{\alpha} \pi(s, \alpha) \sum_{s'} P^{\alpha}_{ss'} [R^{\alpha}_{ss'} + \gamma V^{\pi}(s')] \quad (3.19)$$

Η ύπαρξη και μοναδικότητα της V^{π} εξασφαλίζονται για $\gamma < 1$ ή αν ακολουθώντας την πολιτική π, μπορούμε από όλες τις καταστάσεις να οδηγηθούμε σε κάποια τερματική.

Αν η δυναμική του περιβάλλοντος είναι πλήρως γνωστή, η (3.19) είναι ένα σύστημα $|S|$ γραμμικών εξισώσεων με $|S|$ αγνώστους (τις αξιές $V^{\pi}(s)$, $s \in S$). Για την εύρεση της βέλτιστης πολιτικής του συστήματος ακολουθούμε επαναληπτικές μεθόδους. Πιο συγκεκριμένα, θεωρούμε μία ακολουθία από προσεγγιστικές συναρτήσεις αξίας V^0, V^1, V^2,..., όπου κάθε μία κάνει μία αντιστοίχιση από το χώρο S^+ (S μαζί με την τερματική κατάσταση, αν το πρόβλημα είναι επεισοδιακό) στο χώρο των πραγματικών αριθμών. Η V^0 επιλέγεται αυθαίρετα (η τερματική κατάσταση όμως, αν υπάρχει, θα πρέπει να έχει αξία 0) και κάθε επόμενη προκύπτει από τον κανόνα ανανέωσης :

$$V_{k+1}(s) = E_{\pi}\{r_{t+1} + \gamma V_k(s_{t+1}) | s_t = s\}$$
$$= \sum_{\alpha} \pi(s,\alpha) \sum_{s'} P^{\alpha}_{ss'} [R^{\alpha}_{ss'} + \gamma V_k(s')], \forall s \in S \quad (3.20)$$

Αποδεικνύεται γενικά ότι η ακολουθία $\{V_k\}$ συγκλίνει στη V^{π} καθώς $k \to \infty$, υπό τις ίδιες προϋποθέσεις που εξασφαλίζουν την ύπαρξη της V^{π}.

Από τα παραπάνω προκύπτουν ο αλγόριθμος αξιολόγησης πολιτικής και ο αλγόριθμος επιλογής πολιτικής με δυναμικό προγραμματισμό. Οι αλγόριθμοι αυτοί παρουσιάζονται παρακάτω.

Αλγόριθμος 3.1 Αξιολόγηση πολιτικής με δυναμικό προγραμματισμό

Έστω πολιτική π
$V(s) \leftarrow 0$, για κάθε $s \in S$
Επανέλαβε
    $\Delta \leftarrow 0$
    Για κάθε $s \in S$:
        $u \leftarrow V(s)$
        $V(s) \leftarrow \sum_{\alpha} \pi(s,\alpha) \sum_{s'} P^{\alpha}_{ss'} [R^{\alpha}_{ss'} + \gamma V^{\pi}(s')]$
        $\Delta \leftarrow \max(\Delta, |u - V(s)|)$
μέχρι $\Delta < \theta$ (ένας μικρός θετικός αριθμός)
Επέστρεψε $V \approx V^*$

Αλγόριθμος 3.2 Επιλογή νέας πολιτικής με δυναμικό προγραμματισμό

Έστω πολιτική π

V(s) ← *τυχαίες τιμές,* για κάθε $s \in S$

Επανέλαβε

$\Delta \leftarrow 0$

Για κάθε $s \in S$:

$u \leftarrow V(s)$

$V(s) \leftarrow \sum_{\alpha} \pi(s,\alpha) \sum_{s'} P^{\alpha}_{ss'} [R^{\alpha}_{ss'} + \gamma V^{\pi}(s')]$

$\Delta \leftarrow \max(\Delta, |u - V(s)|)$

μέχρι $\Delta < \theta$ (ένας μικρός θετικός αριθμός)

policy-stable ← true

Για κάθε $s \in S$:

$b \leftarrow \pi(s)$

$\pi(s) \leftarrow \underset{\alpha}{argmax} \sum_{s'} P^{\alpha}_{ss'} [R^{\alpha}_{ss'} + \gamma V(s')]$

Αν $b \neq \pi(s)$, τότε *policy-stable* ← *false*

Αν policy-stable, τότε έξοδος, αλλιώς στο βήμα « Επανέλαβε »

3.7 Επίλυση με αναπαράσταση πινάκων

Όπως αναφέραμε ήδη, οι μέθοδοι δυναμικού προγραμματισμού εφαρμόζονται μόνο όταν είναι πλήρως γνωστό το μοντέλο Markov. Επειδή στη γενική περίπτωση δεν ισχύει κάτι τέτοιο, εφαρμόζουμε διαφορετικές μεθόδους.

Μία από αυτές είναι η μέθοδος επίλυσης με αναπαράσταση πινάκων. Στη μέθοδο αυτή, διατηρούμε στη μνήμη του υπολογιστή μία δομή πίνακα, η οποία διατηρεί τις τρέχουσες τιμές της συνάρτησης αξίας (ή δράσης – αξίας, ανάλογα το ζητούμενο του προβλήματος) για κάθε κατάσταση (ζεύγος δράσης – κατάστασης, αντίστοιχα). Η πολυπλοκότητα των αλγορίθμων εξαρτάται από τα μεγέθη $|S|$ και $|A(s)|$ ($|S||A(s)|$, για κάθε $s \in S$, αντίστοιχα) και για το λόγο αυτό ενδείκνυνται για επίλυση προβλημάτων ενισχυτικής μάθησης με μικρές τιμές για τα παραπάνω μεγέθη.

Πριν παρουσιάσουμε τις διάφορες μεθόδους επίλυσης με αναπαράσταση πινάκων, να αναφέρουμε

ότι οι αλγόριθμοι αυτοί, αποτελούν τη βάση των γενικευμένων αλγορίθμων, οι οποίοι μπορούν να εφαρμοστούν σε προβλήματα με μεγάλο χώρο αναζήτησης.

3.7.1 Επίλυση με προσομοίωση Monte Carlo

Οι μέθοδοι εφαρμόζονται κυρίως σε επεισοδιακά προβλήματα. Η κύρια ιδέα τους είναι ότι προσπαθούν να υπολογίσουν τη μέση τιμή των τιμών που λαμβάνει μία τυχαία μεταβλητή. Στην περίπτωση μας, δεδομένης μίας πολιτικής π, η μέθοδος Monte Carlo προσπαθεί να εκτιμήσει τις τιμές των συναρτήσεων $V(s)$ και $Q(s,\alpha)$. Για να το επιτύχει αυτό, διατηρεί τις μέσες τιμές των ενισχύσεων που λαμβάνει ο πράκτορας σε διαφορετικά επεισόδια.

Για την εφαρμογή μίας τέτοιας μεθόδου, δημιουργούμε ένα μεγάλο πλήθος δυνατών αλληλουχιών καταστάσεων, οι οποίες καταλήγουν σε μία τερματική κατάσταση. Έπειτα για κάθε κατάσταση που επισκέπτεται ο πράκτορας, διατηρεί τη μέση τιμή των ενισχύσεων που έλαβε. Όταν οι μέσοι όροι αρχίσουν να συγκλίνουν για όλες τις καταστάσεις, τότε ο πράκτορας έχει βρει τη βέλτιστη συνάρτηση αξίας V^*. Πλέον, η βέλτιστη πολιτική π^*, βρίσκεται λύνοντας το πρόβλημα βελτιστοποίησης

$$\alpha = arg\ \max_{\alpha \in A(s)} V^*(s) \quad (3.21)$$

3.7.2 Επίλυση με μεθόδους χρονικών διαφορών (temporal difference (TD) methods)

Η επίλυση με μεθόδους χρονικών διαφορών χρησιμοποιεί τη γνώση που αποκτά ο πράκτορας από την αλληλεπίδρασή του με το περιβάλλον, για να λύσει το πρόβλημα ενισχυτικής μάθησης. Δεδομένης μίας πολιτικής π, οι μέθοδοι χρονικών διαφορών ενημερώνουν την εκτίμηση V που έχουν για την V^π. Οι μέθοδοι επίλυσης με χρονικές διαφορές, όπως και οι μέθοδοι επίλυσης με προσομοίωση Monte Carlo, ενημερώνουν την εκτίμηση που έχουν για την $V(s_t)$, όταν επισκεφτούν τη μη τερματική κατάσταση s_t τη χρονική στιγμή t, βασιζόμενες στο τι συμβαίνει μετά από αυτήν την επίσκεψη.

Πιο συγκεκριμένα, οι TD μέθοδοι, χρειάζεται να περιμένουν μόνο μέχρι την επόμενη χρονική στιγμή, $t+1$, και όχι μέχρι το τέλος του επεισοδίου, όπως οι μέθοδοι Monte Carlo. Τη χρονική

στιγμή *t+1*, παρατηρούν την επιβράβευση r_{t+1} που λαμβάνουν και ενημερώνουν την εκτίμηση που έχουν για την $V(s_t)$. Η πιο απλή μέθοδος, γνωστή και ως *TD(0)*, είναι

$$V(s_t) \leftarrow V(s_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)] \quad (3.22)$$

όπου α ο ρυθμός μάθησης.

Επομένως, οι TD μέθοδοι, επιτρέπουν την ανανέωση των εκτιμήσεων συναρτήσεων αξιών μετά από κάθε επανάληψη του μοτίβου « κατάσταση – δράση – κατάσταση » της διαδικασίας μάθησης. Αυτό επιτρέπει στον πράκτορα να μαθαίνει καθώς ενεργεί (on-line learning).

Παρακάτω παρουσιάζουμε τον αλγόριθμο TD(0).

Αλγόριθμος 3.3 TD(0) για προσέγγιση της συνάρτησης V^{π}

Έστω πολιτική π
V(s) ← *τυχαίες τιμές,* για κάθε $s \in S$
Επανέλαβε (για κάθε επεισόδιο)
    s ← s_0
    Επανέλαβε (για κάθε χρονική στιγμή του επεισοδίου)
        α ← δράση που προκύπτει από την πολιτική *π* για την κατάσταση *s*
        Εκτέλεσε τη δράση, παρατήρησε την επιβράβευση r και την επόμενη κατάσταση *s′*
        V(s) ← *V(s)* + *α(r* + *γV(s′)* – *V(s))*
        s ← *s′*
    μέχρι να βρεθεί ο πράκτορας σε τερματική κατάσταση

Μόλις τελειώσει η ενημέρωση των εκτιμήσεων για την πολιτική *π*, την αξιολογούμε. Έπειτα, εφαρμόζουμε ξανά τον αλγόριθμο TD. Η παραπάνω διαδικασία ονομάζεται on-policy, καθώς σε κάθε επανάληψη της, γίνεται αξιολόγηση της πολιτικής π. Επίσης, αποδεικνύεται ότι συγκλίνει στη βέλτιστη πολιτική.

3.7.2.1 On-policy TD μέθοδοι

Όταν καλούμαστε να λύσουμε ένα πρόβλημα βελτιστοποίησης, πρέπει να προσεγγίσουμε τη

συνάρτηση αξίας δράσης $Q^{\pi}(s,a)$ για την πολιτική π, και όχι τη συνάρτηση αξίας $V^{\pi}(s)$. Επεκτείνοντας την σχέση (3.22), ο μηχανισμός ανανέωσης για τη συνάρτηση $Q^{\pi}(s,a)$ είναι

$$Q(s_t,\alpha_t) \leftarrow Q(s_t,\alpha_t) + \alpha(r_{t+1} + \gamma Q(s_{t+1},\alpha_{t+1}) - Q(s_t,\alpha_t)) \quad (3.23)$$

Από αυτό το μηχανισμό, προκύπτει ο αλγόριθμος SARSA (State – Action – Reward – State – Action) που παρουσιάζουμε στη συνέχεια.

Αλγόριθμος 3.4 Αλγόριθμος SARSA

Q(s,α) ← *τυχαίες τιμές*, για κάθε s ϵ *S, α ϵ A(s)*
Επανέλαβε (για κάθε επεισόδιο)
    s ← s_0
    Επέλεξε μία δράση *α* από την κατάσταση s, όπως αυτή προκύπτει από την *Q* (πχ, *ε-greedy*)
    Επανέλαβε (για κάθε χρονική στιγμή του επεισοδίου)
        Εκτέλεσε τη δράση, παρατήρησε την επιβράβευση r και την επόμενη κατάσταση *s′*
        α′ ← δράση από την κατάσταση s', όπως προκύπτει από την *Q* (πχ, *ε - greedy*)
        Q(s,α) ← *Q(s,α) + α(r + γQ(s′,α′) – Q(s,α))*
        s ← *s′*, *α* ← *α′*
    μέχρι να βρεθεί ο πράκτορας σε τερματική κατάσταση

Ο αλγόριθμος SARSA, αποδεικνύεται ότι συγκλίνει, αφού βασίζεται σε ανανέωση των εκτιμήσεων βάσει χρονικών διαφορών.

3.7.2.2 Off-policy TD μέθοδοι

Οι μέθοδοι off-policy, αποδεικνύεται ότι προσεγγίζουν τη βέλτιστη συνάρτηση αξίας δράσης Q^* και τη βέλτιστη πολιτική π^*, ανεξάρτητα από την πολιτική που ακολουθεί ο πράκτορας και για το λόγο αυτό αποτελούν ένα από τα πιο σημαντικά επιτεύγματα στην ενισχυτική μάθηση. Η πολιτική εξακολουθεί να επηρεάζει τη μάθηση, αφού καθορίζει ποια ζεύγη κατάστασης – δράσης επισκέπτονται από τον πράκτορα και ενημερώνονται. Παρ' όλα αυτά, το μόνο που χρειάζεται για τη σωστή σύγκλιση του αλγορίθμου είναι να συνεχίζουν να ενημερώνονται όλα τα ζεύγη

κατάστασης – δράσης. Ο μηχανισμός ανανέωσης για τη συνάρτηση *Q(s,a)* όταν ακολουθούμε off-policy μεθόδους είναι

$$Q(s_t, \alpha_t) \leftarrow Q(s_t, \alpha_t) + \alpha[r_{t+1} + \gamma \max_{\alpha} Q(s_{t+1}, \alpha_{t+1}) - Q(s_t, \alpha_t)] \quad (3.24)$$

Παρακάτω παρουσιάζουμε τον αλγόριθμο Q-Learning, ο οποίος χρησιμοποιεί την (3.24) ως μηχανισμό ανανέωσης.

Αλγόριθμος 3.5 Αλγόριθμος Q-Learning

Q(s,α) ← *τυχαίες τιμές*, για κάθε s ∈ *S*, *α* ∈ *A(s)*
Επανέλαβε (για κάθε επεισόδιο)
    s ← s_0
    Επανέλαβε (για κάθε χρονική στιγμή του επεισοδίου)
        α ← δράση που προκύπτει από την *Q* (πχ, *ε-greedy*)
        Εκτέλεσε τη δράση, παρατήρησε την επιβράβευση r και την επόμενη κατάσταση *s′*
        $Q(s,\alpha) \leftarrow Q(s,\alpha) + \alpha(r + \gamma \max_{\alpha'} Q(s',\alpha') - Q(s,\alpha))$
        s ← *s′*, *α* ← *α′*
    μέχρι να βρεθεί ο πράκτορας σε τερματική κατάσταση

Αποδεικνύεται, ότι ο αλγόριθμος Q-Learning συγκλίνει στη συνάρτηση αξιών Q^* και στη βέλτιστη πολιτική π^*.

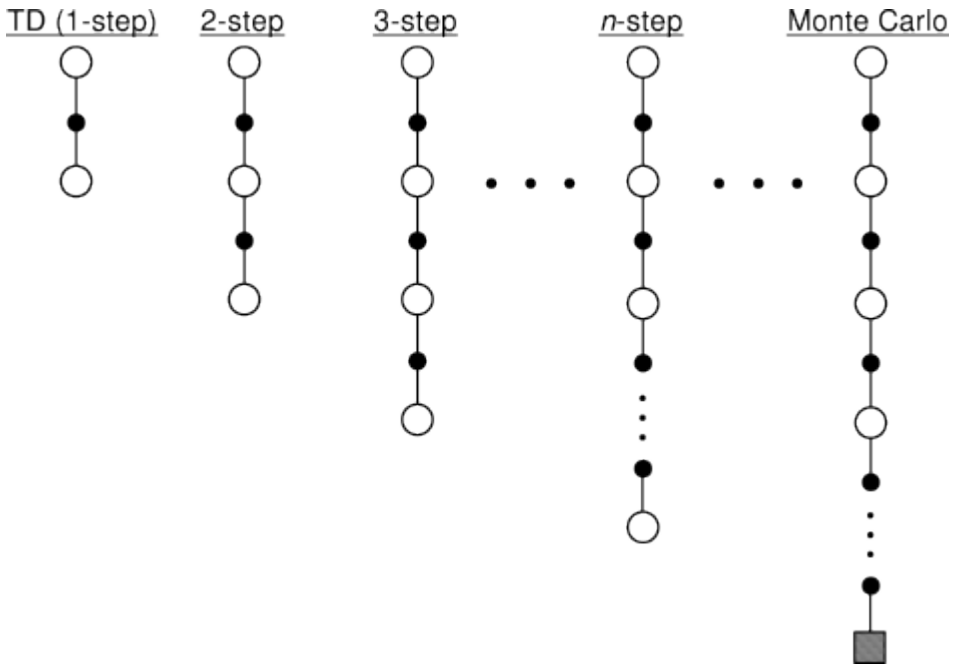
3.8 Eligibility traces και ο παράγοντας λ

Τα ίχνη επιλογής (eligibility traces) αποτελούν μία από τις ισχυρότερες δομές που έχουν προταθεί στο χώρο της ενισχυτικής μάθησης. Τα ίχνη επιλογής συνδέουν τις μεθόδους TD με τις μεθόδους Monte Carlo και αποτελούν ένα είδος προσωρινής μνήμης για τις καταστάσεις που βρέθηκε ο πράκτορας και τις δράσεις που επέλεξε σε αυτές, στο άμεσο παρελθόν. Οι αλγόριθμοι SARSA και Q-Learning, μπορούν να τροποποιηθούν ώστε να συμπεριλαμβάνουν τα ίχνη επιλογής.

3.8.1 Η TD πρόβλεψη n-βημάτων

Έστω ότι προσπαθούμε να προσεγγίσουμε μία συνάρτηση αξίας V^{π}. Όπως αναφέρθηκε σε

προηγούμενη ενότητα, οι μέθοδοι Monte Carlo διατηρούν μία προσεγγιστική τιμή βασισμένες σε όλη την ακολουθία επιβραβεύσεων που έλαβαν από την αρχική ως την τελική κατάσταση του επεισοδίου (full backup). Από την άλλη, οι μέθοδοι TD, διατηρούν μία προσεγγιστική τιμή βασισμένες στην αξία της προηγούμενης μόνο κατάστασης και του σήματος επιβράβευσης που μόλις έλαβε ο πράκτορας (1-step backup). Με τον ίδιο τρόπο, μπορούμε να διατηρούμε τις προσεγγιστικές τιμές βασισμένοι σε περισσότερες από μία παλιότερες καταστάσεις. Η ιδέα αυτή παρουσιάζεται στο σχήμα 3.2



Σχήμα 3.2 : Το φάσμα όλων των μεθόδων από την TD(0) μέχρι τη Monte Carlo [16].

Στο σημείο αυτό θα ορίσουμε ξανά την αντικειμενική συνάρτηση που προσπαθεί να μεγιστοποιήσει κάθε ένας από τους παραπάνω μηχανισμούς μάθησης. Ορίζουμε ως στόχο n βημάτων την έκφραση

$$R_t^{(n)} = r_{t+1} + \gamma r_{t+2} + \cdots + \gamma^{n-1} r_{t+n} + \gamma^n V(s_{t+n}), n \in \{1, \ldots, T-t\} \quad (3.25)$$

Η μέθοδος Monte Carlo έχει ως μηχανισμό ανανέωσης τον τύπο

$$V(s) \leftarrow V(s) + \alpha(R - V(s)) \quad (3.26)$$

όπου R η συνάρτηση επιβράβευσης. Η μέθοδος αυτή προσπαθεί να μεγιστοποιήσει την ποσότητα $R_t^{(T-t)}$ (full backup). Από την άλλη, η μέθοδος TD(0) χρησιμοποιεί ως μηχανισμό ανανέωσης των

τύπο

$$V(s) \leftarrow V(s) + \alpha(r + \gamma V(s') - V(s)) \quad (3.27)$$

όπου *r* η άμεση επιβράβευση. Αυτή η μέθοδος προσπαθεί να μεγιστοποιήσει την ποσότητα $R_t^{(1)}$ (1-step backup).

Αν θέλουμε να κινηθούμε στο φάσμα των μεθόδων του σχήματος 3.2, τότε χρησιμοποιούμε τον αλγόριθμο TD(λ), όπου $0 \leq \lambda \leq 1$, μία παράμετρος η οποία σχετίζεται με τον αριθμό των βημάτων για τα οποία διατηρεί backup ο αλγόριθμος. Η συνολική επιβράβευση δίνεται από τη σχέση

$$R_t^{\lambda} = (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} R_t^{(n)} = (1-\lambda) \sum_{n=1}^{T-t-1} \lambda^{n-1} R_t^{(n)} + \lambda^{T-t-1} R_t \quad (3.28)$$

Ο αλγόριθμος χρησιμοποιεί ένα σταθμισμένο μέσο όρο των backup τιμών για τα *n* βήματα της μάθησης, με την αξία κάθε βήματος *i* να πολλαπλασιάζεται με ένα παράγοντα λ^{i-1}. Παρατηρούμε ότι για $\lambda = 1$, ο αλγόριθμος TD(1) ταυτίζεται με τη μέθοδο Monte Carlo.

3.8.2 Μηχανιστική ερμηνεία των eligibility traces

Η μηχανιστική ερμηνεία είναι πιο απλή και ευκολότερα υλοποιήσιμη από τη θεωρητική που μόλις παρουσιάσαμε. Για το λόγο αυτό, μπορεί να ενσωματωθεί στους αλγορίθμους χρονικών διαφορών που έχουμε περιγράψει. Όπως αποδεικνύεται επίσης στο [16], η ερμηνεία αυτή, είναι ισοδύναμη με τη θεωρητική.

Για τη μηχανιστική θεώρηση των ιχνών επιλογής, διατηρούμε στη μνήμη του συστήματος μία επιπλέον μεταβλητή. Συμβολίζουμε το ίχνος επιλογής κάθε κατάστασης *s* τη χρονική στιγμή *t* ως $e_t(s) \in \mathbb{R}$, όπου $\mathbb{R}$ το σύνολο των πραγματικών αριθμών. Υπάρχουν δύο είδη ιχνών επιλογής, όσον αφορά τη μηχανιστική ερμηνεία, τα οποία μειώνονται σύμφωνα με τους κανόνες

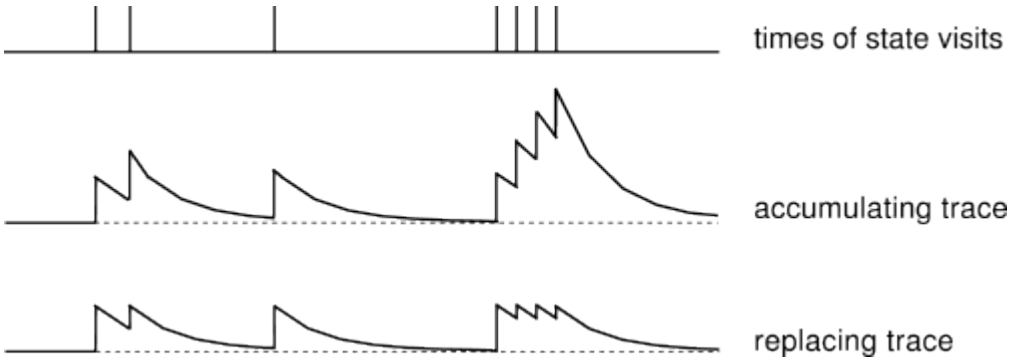
- *Συσσωρευτικά ίχνη επιλογής*

$$e_t(s)=\begin{cases}\gamma\lambda e_{t-1}(s), & s\neq s_t\\ \gamma\lambda e_{t-1}(s)+1, & s=s_t\end{cases},\ \lambda\in[0,1],\ s\in S \quad (3.29)$$

- *Ίχνη επιλογής με αντικατάσταση*

$$e_t(s)=\begin{cases}\gamma\lambda e_{t-1}(s), & s\neq s_t\\ 1, & s=s_t\end{cases},\ \lambda\in[0,1],\ s\in S \quad (3.30)$$

Τα ίχνη επιλογής θα μειώνονται κάθε χρονική στιγμή, εκτός από το ίχνος της τρέχουσας κατάστασης, το οποίο αυξάνεται κατά 1 στα συσσωρευτικά ίχνη επιλογής, ή γίνεται 1 στα ίχνη επιλογής με αντικατάσταση. Στο σχήμα 3.3 παρουσιάζονται οι κανόνες ανανέωσης (3.29) και (3.30).



Σχήμα 3.3 : Σχηματική αναπαράσταση των μηχανισμών ανανέωσης των συσσωρευτικών ιχνών επιλογής και των ιχνών επιλογής με αντικατάσταση [16]

Ενσωματώνοντας έναν από τους παραπάνω κανόνες ανανέωσης στη μέθοδο TD, προκύπτει ο μηχανισμός ανανέωσης

$$V_{t+1}(s_t)=V_t(s_t)+\alpha[r_{t+1}+\gamma V_t(s_{t+1})-V_t(s_t)]e_t(s_t) \quad (3.31)$$

3.8.3 Αλγόριθμοι χρονικών διαφορών με χρήση ιχνών επιλογής

Οι αλγόριθμοι TD, SARSA και Q-Learning, μπορούν να επεκταθούν με χρήση eligibility traces. Για την απλή περίπτωση του TD(*λ)*, ο κανόνας ανανέωσης των eligibility traces έχει παρουσιαστεί ήδη

και για αυτό το λόγο μπορούμε να τον παρουσιάσουμε αμέσως. Για τους άλλους δύο αλγορίθμους, πρέπει να αναφέρουμε πρώτα τους κανόνες ανανέωσης των ιχνών αντικατάστασης τους, οι οποίοι διαφέρουν λίγο από τους κανόνες που έχουμε ήδη περιγράψει.

3.8.3.1 Αλγόριθμος TD(λ)

Αλγόριθμος 3.6 Αλγόριθμος TD(λ)

V(s) ← *τυχαίες τιμές*, για κάθε s ϵ *S*
e(s) ← *0*, για κάθε s ϵ *S*
Επανέλαβε (για κάθε επεισόδιο)
    $s \leftarrow s_0$
    Επανέλαβε (για κάθε χρονική στιγμή του επεισοδίου)
        $\alpha \leftarrow$ δράση που προκύπτει από την πολιτική π για την κατάσταση s
        Εκτέλεσε τη δράση, παρατήρησε την επιβράβευση r και την επόμενη κατάσταση s'
        $\delta \leftarrow r + \gamma V(s') - V(s)$
        $e(s) \leftarrow e(s) + 1$
        Για κάθε s:
            $V(s) \leftarrow V(s) + \alpha\delta e(s)$
        $s \leftarrow s'$
    μέχρι να βρεθεί ο πράκτορας σε τερματική κατάσταση

3.8.3.2 Αλγόριθμοι SARSA(λ) και Watkin's Q(λ)

Στους κανόνες που περιγράψαμε προηγουμένως, τα ίχνη αντικατάστασης είναι συναρτήσεις της κατάστασης μόνο. Στους αλγορίθμους SARSA και Q-Learning, όμως, οι εκτιμήσεις που διατηρούμε είναι για ζεύγη κατάστασης – δράσης. Επομένως, πρέπει να τροποποιήσουμε τους κανόνες ανανέωσης (3.29) και (3.30) ώστε να λαμβάνουν υπόψη τα ζεύγη κατάστασης – δράσης. Η τροποποιήσεις αυτές παρουσιάζονται παρουσιάζονται παρακάτω, ανάλογα με το είδος τους.

- *Συσσωρευτικά ίχνη επιλογής*

$$e_t(s,\alpha)=\begin{cases}\gamma\lambda e_{t-1}(s,\alpha)+1, & s=s_t,\alpha=\alpha_t\\ \gamma\lambda e_{t-1}(s,\alpha), & s\neq s_t,\alpha\neq\alpha_t\end{cases},\ \lambda\in[0,1],\ s\in S,\ \alpha\in A(s) \quad (3.32)$$

- *Ίχνη επιλογής με αντικατάσταση*

$$e_t(s,\alpha)=\begin{cases}1+\gamma\lambda e_{t-1}(s,\alpha), & s=s_t,\alpha=\alpha_t\\ 0, & s=s_t,\alpha\neq\alpha_t\\ \gamma\lambda e_{t-1}(s,\alpha), & s\neq s_t\end{cases},\ \lambda\in[0,1],\ s\in S,\ \alpha\in A(s) \quad (3.33)$$

Αλγόριθμος 3.7 Αλγόριθμος SARSA(λ)

Q(s,α) ← *τυχαίες τιμές*, για κάθε s ϵ *S, α ϵ A(s)*

e(s,α) ← *0*, για κάθε s ϵ *S, α ϵ A(s)*

Επανέλαβε (για κάθε επεισόδιο)

s ← s_0, *α* ← α_0

Επανέλαβε (για κάθε χρονική στιγμή του επεισοδίου)

Εκτέλεσε τη δράση, παρατήρησε την επιβράβευση r και την επόμενη κατάσταση *s′*

Επέλεξε μία δράση *α′* από την *s*, όπως αυτή προκύπτει από την *Q* (πχ, *ε - greedy*)

δ ← *r* + *γQ(s′,α′)* – *Q(s,α)*

e(s,α) ← *e(s,α)* + *1*

Για κάθε *s, α:*

Q(s,α) ← *Q(s,α)* + *αδe(s,α)*

e(s,α) ← *γλe(s,α)*

s ← *s′*, *α* ← *α′*

μέχρι να βρεθεί ο πράκτορας σε τερματική κατάσταση

Αλγόριθμος 3.8 Watkin's Q(λ)

$Q(s,\alpha) \leftarrow$ *τυχαίες τιμές*, για κάθε s ϵ *S*, $\alpha \epsilon A(s)$

$e(s,\alpha) \leftarrow 0$, για κάθε s ϵ *S*, $\alpha \epsilon A(s)$

Επανέλαβε (για κάθε επεισόδιο)

$s \leftarrow s_0$, $\alpha \leftarrow \alpha_0$

Επανέλαβε (για κάθε χρονική στιγμή του επεισοδίου)

Εκτέλεσε τη δράση, παρατήρησε την επιβράβευση r και την επόμενη κατάσταση s'
Επέλεξε μία δράση α' από την s, όπως αυτή προκύπτει από την Q (πχ, ε - greedy)

$\alpha^* \leftarrow \arg\max_{b \epsilon A(s)} Q(s',b)$

$\delta \leftarrow r + \gamma Q(s',\alpha') - Q(s,\alpha)$

$e(s,\alpha) \leftarrow e(s,\alpha) + 1$

Για κάθε s, α:

$Q(s,\alpha) \leftarrow Q(s,\alpha) + \alpha\delta e(s,\alpha)$

Αν $\alpha' = \alpha^*$, τότε $e(s,\alpha) \leftarrow \gamma\lambda e(s,\alpha)$

αλλιώς $e(s,a) \leftarrow 0$

$s \leftarrow s'$, $\alpha \leftarrow \alpha'$

μέχρι να βρεθεί ο πράκτορας σε τερματική κατάσταση

3.9 Επίλυση με μηχανισμούς προσέγγισης συναρτήσεων (function approximation)

Οι μέθοδοι επίλυσης με μηχανισμούς προσέγγισης συναρτήσεων χρησιμοποιούνται σε περιπτώσεις, όπου έχουμε ένα περιβάλλον με μεγάλο πλήθος καταστάσεων και δράσεων. Σε τέτοια περιβάλλοντα προκύπτουν προβλήματα, όχι μόνο με το μέγεθος μνήμης που χρειαζόμαστε για τους πίνακες, αλλά και με τα δεδομένα και το χρόνο που χρειαζόμαστε για να προσεγγίσουμε τις τιμές των πινάκων. Με λίγα λόγια, το θέμα που προκύπτει είναι αυτό της γενίκευσης. Δηλαδή, με ποιον τρόπο μπορούμε να γενικεύσουμε τη γνώση που έχουμε λάβει από ένα μικρό υποσύνολο του χώρου καταστάσεων, σε ένα μεγαλύτερο υποσύνολο.

3.9.1 Προσέγγιση συνάρτησης αξίας V^π

Για να προσεγγίσουμε τη συνάρτηση αξίας V^π, διατηρούμε μία εκτίμηση της, V_t, σε κάθε χρονική

στιγμή t. Η V_t δεν αναπαρίσταται ως πίνακας, αλλά ως μία παραμετρική συναρτησιακή μορφή με παράμετρο ένα διάνυσμα $\boldsymbol{\theta}_t$.

Ως μέτρο απόδοσης υιοθετούμε την ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος

$$MSE(\vec{\theta}_t)=\sum_{s\in S} P(s)[V^{\pi}(s)-V_t(s)]^2 \quad (3.34)$$

όπου P είναι μία κατανομή η οποία σταθμίζει το σφάλμα σε διαφορετικές καταστάσεις. Η κατανομή αυτή είναι σημαντική, γιατί συνήθως δεν είναι εφικτό να μηδενίσουμε το σφάλμα σε όλες τις καταστάσεις. Μπορούμε όμως να πάρουμε καλύτερη προσέγγιση σε κάποιες καταστάσεις, εις βάρος άλλων καταστάσεων. Η κατανομή αυτή καθορίζει πώς θα γίνει αυτός ο συμβιβασμός.

3.9.1.1 Προσέγγιση συνάρτησης αξίας με μεθόδους απότομης κατάβασης

Οι μέθοδοι απότομης κατάβασης είναι από τις πιο ευρέως χρησιμοποιούμενες στην προσέγγιση συναρτήσεων και εφαρμόζονται ιδιαίτερα καλά στην ενισχυτική μάθηση.

Οι μέθοδοι αυτοί, μετακινούν το διάνυσμα παραμετροποίησης $\boldsymbol{\theta}$ κατά ένα μικρό ποσό, προς την κατεύθυνση αυτή που θα δώσει τη μεγαλύτερη μείωση του σφάλματος. Η μετακίνηση αυτή λαμβάνει χωρά σε κάθε χρονική στιγμή και έτσι προκύπτει ο παρακάτω κανόνας ανανέωσης

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \frac{1}{2}\alpha\nabla_{\theta_t}[V^{\pi}(s_t)-V_t(s_t)]^2 \\ &= \boldsymbol{\theta}_t + \alpha[V^{\pi}(s_t)-V_t(s_t)]\nabla_{\theta_t}V_t(s_t)\end{aligned} \quad (3.35)$$

όπου α είναι μία θετική σταθερά που καθορίζει το ποσό μετακίνησης, ή ο γνωστός ρυθμός μάθησης. Στην παραπάνω σχέση θεωρήσαμε ότι η κατανομή P είναι ομοιόμορφη.

Η μορφή της πλέον απότομης κατάβασης της μεθόδου TD(λ), χρησιμοποιεί την συνολική λ-επιβράβευση R_t^{λ}, ως προσέγγιση της $V_{\pi}(s_t)$. Επομένως, η ο κανόνας ανανέωσης (3.35), αλλάζει και γίνεται

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \frac{1}{2}\alpha\nabla_{\theta_t}[R_t^{\lambda}-V_t(s_t)]^2 \\ &= \boldsymbol{\theta}_t + \alpha[r+\gamma V(s_{t+1})-V_t(s_t)]\nabla_{\theta_t}V_t(s_t)\end{aligned} \quad (3.36)$$

και αν ενσωματώσουμε και το μηχανισμό των ιχνών επιλογής τότε ο κανόνας ανανέωσης γίνεται

$$\boldsymbol{\theta}_{t+1}=\boldsymbol{\theta}_t+\alpha\delta_t\boldsymbol{e}_t \quad (3.37)$$

όπου δ_t, είναι το συνηθισμένο TD σήμα σφάλματος,

$$\delta_t=r_{t+1}+\gamma V_t(s_{t+1})-V_t(s_t) \quad (3.38)$$

και $\boldsymbol{e}_t$, το διάνυσμα ιχνών επιλογής, όπου κάθε συνιστώσα του, αντιστοιχεί σε κάθε συνιστώσα του διανύσματος $\boldsymbol{\theta}_t$, το οποίο ανανεώνεται από τον κανόνα

$$\boldsymbol{e}_t=\gamma\lambda\boldsymbol{e}_{t-1}+\nabla_{\theta_t}V_t(s_t) \quad (3.39)$$

για τα συσσωρευτικά ίχνη επιλογής και από τον κανόνα

$$\boldsymbol{e}_t=max(\gamma\lambda\boldsymbol{e}_{t-1},\nabla_{\theta_t}V_t(s_t)) \quad (3.40)$$

για τα ίχνη επιλογής με αντικατάσταση. Σημειώνεται ότι και στους δύο κανόνες ισχύει $\boldsymbol{e}_0 = \boldsymbol{0}$. Παρακάτω παραθέτουμε το αλγόριθμο TD(λ), με προσέγγιση συνάρτησης αξίας.

Αλγόριθμος 3.9 Αλγόριθμος TD(λ) με προσέγγιση συνάρτησης αξίας

$\boldsymbol{\theta} \leftarrow$ *τυχαίες τιμές*
Επανέλαβε (για κάθε επεισόδιο)
    $\boldsymbol{e} \leftarrow 0$
    $s \leftarrow s_0$
    Επανέλαβε (για κάθε χρονική στιγμή του επεισοδίου)
        $\alpha \leftarrow$ δράση που προκύπτει από την πολιτική π για την κατάσταση s
        Εκτέλεσε τη δράση, παρατήρησε την επιβράβευση r και την επόμενη κατάσταση s'
        $\delta \leftarrow r + \gamma V(s') - V(s)$
        $\boldsymbol{e}=\gamma\lambda\boldsymbol{e}+\nabla_{\boldsymbol{\theta}}V(s')$
        $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\delta\boldsymbol{e}$
        $s \leftarrow s'$
    μέχρι να βρεθεί ο πράκτορας σε τερματική κατάσταση

3.9.1.2 Προσέγγιση συνάρτησης αξίας με γραμμικές μεθόδους

Όταν η V_t, είναι γραμμική συνάρτηση των παραμέτρων $\boldsymbol{\theta}_t$, τότε μπορούμε να την αναπαραστήσουμε ως
ένα εσωτερικό γινόμενο

$$V_t(s)=\boldsymbol{\theta}_t^T \boldsymbol{\varphi}_s=\sum_{i=1}^{n} \theta_t(i)\varphi_s(i) \quad (3.41)$$

Το διάνυσμα $\boldsymbol{\varphi}_s$, αντιπροσωπεύει την κατάσταση s, ονομάζεται διάνυσμα συναρτήσεων βάσης και έχει το ίδιο μέγεθος με το διάνυσμα παραμέτρων $\boldsymbol{\theta}_t$. Οι συναρτήσεις βάσεις αναφέρονται και ως χαρακτηριστικά του χώρου κατάστασης.

Από την (3.27) προκύπτει ότι

$$\nabla_{\theta_t} V_t(s)=\boldsymbol{\varphi}_s \quad (3.42)$$

Για την αναπαράσταση των συναρτήσεων βάσης μπορούν να χρησιμοποιηθούν διάφορες μέθοδοι. Οι πιο διαδεδομένη είναι η χρήση των κανονικοποιημένων Γκαουσιανών ακτινικών συναρτήσεων βάσεων (Gaussian Radial Basis Functions – RBFs). Ο τύπος της κανονικοποιημένης RBF δίνεται από τη σχέση

$$\bar{\varphi}_s(i)=\frac{\varphi_s(i)}{\sum_{j=1}^{n} \varphi_s(j)} \quad (3.43)$$

όπου η $\varphi_s(i)$ δίνεται από τη σχέση

$$\varphi_s(i)=\exp\left(\frac{-\|s-c_i\|^2}{2\sigma_i^2}\right) \quad (3.44)$$

Στην (3.43), σ είναι η τυπική απόκλιση της RBF και c_i το κέντρο της.

Υπάρχουν και άλλοι τύποι παραμετρικών μηχανισμών, όπως οι tile coding, coarse coding και Kanerva coding, οι οποίοι αναλύονται στο [16].

3.9.2 Προσέγγιση συνάρτησης αξίας δράσης Q^{π}

Η τεχνική προσέγγισης συνάρτησης αξίας V^{π}, που μόλις περιγράψαμε, μπορεί να επεκταθεί και για την προσέγγιση της συνάρτησης αξίας δράσης Q^{π}. Για να γίνει αυτό όμως, πρέπει να χρησιμοποιήσουμε ένα διάνυσμα $\boldsymbol{\varphi}_{s,a}$, το οποίο θα μας δίνει τα χαρακτηριστικά του χώρου κατάστασης – δράσης πλέον. Επομένως, η συνάρτηση αξίας δράσης Q^{π} δίνεται από το εσωτερικό γινόμενο

$$Q_t(s)=\vec{\theta}_t^{\,T}\vec{\varphi}_{s,a}=\sum_{i=1}^{n}\theta_t(i)\varphi_{s,a}(i) \quad (3.45)$$

Η επέκταση των αλγορίθμων SARSA(λ) και Watkins-Q(λ) με χρήση προσέγγισης συνάρτησης αξίας δράσης, δίνεται παρακάτω. Οι αλγόριθμοι παρουσιάζονται για δυαδικά χαρακτηριστικά του χώρου κατάστασης – δράσης.

Αλγόριθμος 3.10 Αλγόριθμος SARSA(λ) με προσέγγιση συνάρτησης αξίας δράσης

$\boldsymbol{\theta} \leftarrow$ *τυχαίες τιμές*

Επανέλαβε (για κάθε επεισόδιο)

$\boldsymbol{e} \leftarrow 0$

$s \leftarrow s_0$, $\alpha \leftarrow \alpha_0$

$F_\alpha \leftarrow$ σύνολα χαρακτηριστικών τα οποία είναι παρόντα στο ζεύγος *s, a*

Επανέλαβε (για κάθε χρονική στιγμή του επεισοδίου)

Για κάθε $i \in F_\alpha$

$e(i) \leftarrow e(i) + 1$ (συσσωρευτικά ίχνη επιλογής)

ή $e(i) \leftarrow e1$ (ίχνη επιλογής με αντικατάσταση)

Εκτέλεσε τη δράση, παρατήρησε την επιβράβευση r και την επόμενη κατάσταση *s*

$\delta \leftarrow r + \Sigma_{i \in F\alpha}\theta(i)$

$\alpha \leftarrow$ τυχαία δράση, η οποία ανήκει στο σύνολο $A(s)$ οπότε :

$F_\alpha \leftarrow$ σύνολα χαρακτηριστικών τα οποία είναι παρόντα στο ζεύγος *s, a*

$Q_\alpha \leftarrow \Sigma_{i \in F\alpha}\theta(i)$

ή α προκύπτει ως :

Για κάθε $\alpha \in A(s)$

$F_\alpha \leftarrow$ σύνολα χαρακτηριστικών τα οποία είναι παρόντα στο ζεύγος *s, a*

$Q_\alpha \leftarrow \Sigma_{i \in F\alpha}\theta(i)$

$\alpha \leftarrow \arg\max_\alpha Q_\alpha$

$\delta \leftarrow \delta + \gamma Q_\alpha$

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\delta\boldsymbol{e}$

$\boldsymbol{e} \leftarrow \gamma\lambda\boldsymbol{e}$

μέχρι να βρεθεί ο πράκτορας σε τερματική κατάσταση

Αλγόριθμος 3.11 Αλγόριθμος Watkins Q(λ) με προσέγγιση συνάρτησης αξίας δράσης και συσσωρευτικά ίχνη επιλογής

$\boldsymbol{\theta} \leftarrow$ *τυχαίες τιμές*

Επανέλαβε (για κάθε επεισόδιο)

$\boldsymbol{e} \leftarrow 0$

$s \leftarrow s_0$, $\alpha \leftarrow \alpha_0$

$F_\alpha \leftarrow$ σύνολα χαρακτηριστικών τα οποία είναι παρόντα στο ζεύγος *s, a*

Επανέλαβε (για κάθε χρονική στιγμή του επεισοδίου)

Για κάθε $i \in F_\alpha$

$e(i) \leftarrow e(i) + 1$ (συσσωρευτικά ίχνη επιλογής)

Εκτέλεσε τη δράση, παρατήρησε την επιβράβευση r και την επόμενη κατάσταση *s*

$\delta \leftarrow r + \Sigma_{i \in F\alpha}\theta(i)$

Για κάθε $\alpha \in A(s)$

$F_\alpha \leftarrow$ σύνολα χαρακτηριστικών τα οποία είναι παρόντα στο ζεύγος *s, a*

$Q_\alpha \leftarrow \Sigma_{i \in F\alpha}\theta(i)$

$\delta \leftarrow \delta + \gamma \max_\alpha Q_\alpha$

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\delta\boldsymbol{e}$

$\alpha \leftarrow$ τυχαία δράση, η οποία ανήκει στο σύνολο $A(s)$ οπότε $\boldsymbol{e} \leftarrow 0$

ή α προκύπτει ως :

Για κάθε $\alpha \in A(s)$

$Q_\alpha \leftarrow \Sigma_{i \in F\alpha}\theta(i)$

$\alpha \leftarrow \arg\max_\alpha Q_\alpha$

$\boldsymbol{e} \leftarrow \gamma\lambda\boldsymbol{e}$

μέχρι να βρεθεί ο πράκτορας σε τερματική κατάσταση

4 *Ενισχυτική μάθηση σε συνεργαζόμενα πολυπρακτορικά συστήματα*

Στα πολυπρακτορικά συστήματα, οι πράκτορες πρέπει να μάθουν να συντονίζουν τις δράσεις τους, ώστε να επιτύχουν το στόχο τους. Η ενισχυτική μάθηση είναι μία μέθοδος, η οποία μπορεί να τους βοηθήσει προς αυτή την κατεύθυνση. Έχοντας παρουσιάσει τη θεωρία της ενισχυτικής μάθησης για έναν πράκτορα, μπορούμε να την επεκτείνουμε και στην περίπτωση πολλών πρακτόρων. Το κεφάλαιο αυτό βασίζεται στη δουλεία των Claus και Boutilier [17], οι οποίοι μελέτησαν την επέκταση του αλγόριθμου Q-Learning, σε συνεργαζόμενα πολυπρακτορικά συστήματα. Επίσης, θα περιγράψουμε τον αλγόριθμο *Φανταστικού Παιγνίου (Fictitious Play – FP*) [18, 19], ο οποίος αποτελεί έναν απλό, αλλά αποδοτικό τρόπο να επιτύχουμε συνεργασία σε πολυπρακτορικά προβλήματα μάθησης.

4.1 Μάθηση σε συνεργατικά παίγνια

Στα συνεργατικά παίγνια, οι πράκτορες πρέπει να μάθουν να διαλέγουν δράσεις με σκοπό να επιτύχουν το στόχο τους. Αν οι πράκτορες επιλέγουν τυχαία δράσεις, τότε θα οδηγηθούν σε υποβέλτιστες λύσεις ή σε μη συνεργατική συμπεριφορά. Για να αποφευχθεί αυτή η κατάσταση, θα μπορούσαμε να επιτρέψουμε στους πράκτορες να επικοινωνούν μεταξύ τους [20] ή να θέσουμε περιορισμούς στη συμπεριφορά τους [21]. Μία άλλη λύση, την οποία θα περιγράψουμε παρακάτω στο κεφάλαιο αυτό, είναι η μάθηση μέσω του συνεχούς παιξίματος παιγνίου από τους ίδιους πράκτορες (αλγόριθμος *Φανταστικού Παιγνίου – Fictitious Play – FP*) [22, 23, 24, 25].

Συγκεκριμένα, κάθε πράκτορας i διατηρεί ένα μετρητή $C_i^j(s,\alpha_j)$, για κάθε πράκτορα $j \in G$ και κάθε δράση $\alpha_j \in A_j$ του πράκτορα j, όπου G το σύνολο των πρακτόρων και A_j το σύνολο των δυνατών δράσεων του πράκτορα j. Ο μετρητής αυτός περιέχει τον αριθμό των φορών που ο πράκτορας j έλαβε τη δράση α_j, όταν ο πράκτορας i βρέθηκε στην κατάσταση s. Όταν ο πράκτορας i βρεθεί στην κατάσταση s και πρέπει να λάβει μία απόφαση, θεωρεί τις σχετικές συχνότητες των δράσεων κάθε πράκτορα j, ενδεικτικές τις τρέχουσας στρατηγικής που ακολουθεί. Επομένως, ο πράκτορας i θεωρεί ότι κάθε πράκτορας j θα λάβει την απόφαση $\alpha_j \in A_j$ με πιθανότητα

$$P_i^j(s,\alpha_j) = C_i^j(s,\alpha_j) / \sum_{b_j \in A_j} C_i^j(s,b_j) \quad (4.1)$$

Με τον τρόπο αυτό, ο πράκτορας i δημιουργεί ένα σύνολο στρατηγικών που θεωρεί ότι ακολουθούν οι υπόλοιποι πράκτορες και ονομάζεται μειωμένο προφίλ στρατηγικών Π_{-i}. Μετά από κάθε χρονική στιγμή t της διαδικασίας μάθησης, ο πράκτορας i ενημερώνει τους μετρητές που διατηρεί για κάθε άλλο πράκτορα κατάλληλα. Μπορούμε να θεωρήσουμε ότι οι μετρητές αυτοί αντικατοπτρίζουν τις πεποιθήσεις ενός πράκτορα για τις δράσεις που θα επιλέξουν οι υπόλοιποι πράκτορες.

Η ύπαρξη πολλών πρακτόρων, καθένας εκ των οποίων μαθαίνει ταυτόχρονα με τους υπόλοιπους, μπορεί να αποδειχθεί εμπόδιο στην επιτυχία του αλγορίθμου Q-Learning (ή της ενισχυτικής μάθησης γενικότερα). Όταν ένας πράκτορας μαθαίνει να εκτιμάει τις πράξεις του, παρουσία άλλων πρακτόρων, τότε βρίσκεται σε ένα μεταβαλλόμενο περιβάλλον. Επομένως, δεν υπάρχει εγγύηση όσον αφορά τη σύγκλιση των τιμών $Q(s,\alpha)$ κάθε πράκτορα. Μία απλή εφαρμογή του αλγορίθμου Q-Learning μπορεί να αποβεί επιτυχής, αν μπορούμε να διασφαλίσουμε ότι η στρατηγική κάθε πράκτορα, τελικά θα συγκλίνει.

4.2 Κατηγορίες αλγορίθμου Q-Learning σε πολυπρακτορικά περιβάλλοντα

Υπάρχουν δύο ξεχωριστοί τρόποι με τους οποίους μπορούμε να εφαρμόσουμε τον αλγόριθμο Q-Learning σε ένα πολυπρακτορικό σύστημα. Οι τρόποι αυτοί ονομάζονται *ανεξάρτητη μάθηση* και

από κοινού μάθηση.

4.2.1 Ανεξάρτητη μάθηση (independent learning – IL)

Θεωρούμε ότι ένας αλγόριθμος μάθησης σε πολυπρακτορικό περιβάλλον ανήκει στην κατηγορία της ανεξάρτητης μάθησης, όταν μαθαίνει τις τιμές $Q(s,\alpha)$ μόνο των δικών του δράσεων, χρησιμοποιώντας ως κανόνα ανανέωσης τη σχέση

$$Q(s,\alpha) \leftarrow Q(s,\alpha) + \alpha(r + \gamma Q(s',\alpha') - Q(s,\alpha)) \quad (4.2)$$

Η σχέση (4.2) έχει παρουσιαστεί και επεξηγηθεί ήδη στο κεφάλαιο 3 και παρουσιάζεται κι εδώ για λόγους καλύτερης παρουσίασης της θεωρίας.

Με λίγα λόγια, ο πράκτορας που χρησιμοποιεί έναν αλγόριθμο, ο οποίος ανήκει στην παραπάνω κατηγορία, εκτελεί τη δράση του α, λαμβάνει το σήμα επιβράβευσης του r και ενημερώνει τις τιμές του $Q(s,\alpha)$, χωρίς να λαμβάνει υπόψιν του τις δράσεις που εκτέλεσαν οι υπόλοιποι πράκτορες. Επομένως, η γνώση που αποκτά σε κάθε χρονική στιγμή t μπορεί να συνοψιστεί στη μορφή $\{\alpha,r\}$.

Η μέθοδος είναι κατάλληλη, όταν ο πράκτορας δε γνωρίζει την ύπαρξη και άλλων πρακτόρων, δεν μπορεί να αναγνωρίσει τις δράσεις τους ή δεν έχει λόγο να πιστεύει ότι οι άλλοι πράκτορες ακολουθούν κάποια στρατηγική. Είναι επίσης κατάλληλη, όταν ο πράκτορας επιλέξει να αγνοήσει τις πληροφορίες που θα μπορούσε να έχει για τους υπόλοιπους πράκτορες.

4.2.2 Από κοινού μάθηση (joint action learner – JAL)

Στην κατηγορία αυτή ανήκουν αλγόριθμοι που μαθαίνουν τιμές Q για κοινές δράσεις $\boldsymbol{\alpha}$. Η γνώση που αποκτά σε κάθε χρονική στιγμή t ένας πράκτορας που χρησιμοποιεί έναν αλγόριθμο, ο οποίος ανήκει στην παραπάνω κατηγορία συνοψίζεται στη μορφή $\{\boldsymbol{\alpha},r\}$. Αυτό σημαίνει ότι ο πράκτορας i, μπορεί να παρατηρήσει τις δράσεις των υπολοίπων πρακτόρων.

Όταν χρησιμοποιούμε JAL, οι στρατηγικές εξερεύνησης χρειάζονται προσοχή. Κάθε πράκτορας i, διατηρεί κάποια πεποίθηση για τη στρατηγική των άλλων πρακτόρων, τις οποίες χρησιμοποιεί για να καθορίσει τις σχετικές τιμές των ατομικών δράσεων τους. Με λίγα λόγια, ο πράκτορας i θεωρεί

ότι κάθε άλλος πράκτορας *j* θα επιλέξει τη δράση του σύμφωνα με τη πεποίθηση του *i* για τον *j*. Γενικότερα, ο πράκτορας *i* αξιολογεί την εκτιμώμενη τιμή της ατομικής του δράσης α_i όταν βρίσκεται στην κατάσταση *s*, σύμφωνα με τη σχέση

$$EV(s,\alpha_i)=\sum_{\alpha_{-i}\in A_{-i}} Q(s,\alpha_{-i}\cup\{\alpha_i\})\prod_{j\neq i} P_i^j \quad (4.3)$$

Η παραπάνω εκτίμηση μπορεί να χρησιμοποιηθεί σε μία στρατηγική εξερεύνησης, όπως θα κάναμε και με την τιμή $Q(s,\alpha)$ στη μέθοδο IL. Παραδείγματος χάριν, αν χρησιμοποιούσαμε μία κατανομή Boltzmann, η πιθανότητα να διαλέγαμε τη δράση α^i, θα ήταν

$$\frac{e^{EV(s,\alpha_i)/T}}{\sum_{\alpha_i\in A_i} e^{EV(s,\alpha_i)/T}} \quad (4.4)$$

Οι Claus και Boutilier στο [17], κάνουν μία σύγκριση του Q-Learning στις δύο παραπάνω κατηγορίες σε ένα παίγνιο πίνακα (matrix game) με δύο πράκτορες. Ως JAL χρησιμοποίησαν τον αλγόριθμο FP (ή FP-Q), ο οποίος αποδείχθηκε ότι σύγκλινε πιο γρήγορα. Παρακάτω παραθέτουμε τον αλγόριθμο FP-Q.

4.1 Αλγόριθμος Fictitious Play-Q (FP-Q)

$Q(s,\alpha^{-i}\cup\alpha^j)$ ← τυχαίες τιμές

$C_i^j(s,\alpha_j)$ ← *0, για κάθε j* ϵ *α, για κάθε δράση* $\alpha \epsilon A_j$*, για κάθε* $s \epsilon S$

Αρχικοποίησε το ρυθμό μάθησης *α*

Επανέλαβε (για κάθε επεισόδιο)

$s \leftarrow s_0$, $\alpha \leftarrow \alpha_0$

Επανέλαβε (για κάθε χρονική στιγμή του επεισοδίου)

Εκτέλεσε τη δράση, παρατήρησε την επιβράβευση r, την επόμενη κατάσταση *s′* και τις δράσεις που εκτέλεσαν οι άλλοι πράκτορες

Επέλεξε μία δράση α'_{-i} από την *s′*, όπως αυτή προκύπτει από την *Q* (πχ, *ε* - greedy ή Boltzmann)

$Q(s,\alpha_{-i}\cup\alpha_j) \leftarrow Q(s,\alpha_{-i}\cup\alpha_j) + \alpha(r + \gamma EV(s',\alpha'_{-i}\cup\alpha'_j) - Q(s,\alpha_{-i}\cup\alpha_j))$

$C_i^j(s,\alpha_j) \leftarrow C_i^j(s,\alpha_j) + 1$, *για κάθε j* ϵ *G*

$s \leftarrow s'$, $a \leftarrow a'$

μέχρι να βρεθεί ο πράκτορας σε τερματική κατάσταση

Μείωσε το ρυθμό μάθησης *α* (και τη θερμοκρασία *T* της κατανομής Boltzmann, αν αυτή χρησιμοποιείται)

4.3 Σύγκλιση του αλγορίθμου FP-Q

Οι Claus και Boutilier στο [17] χρησιμοποίησαν δύο παίγνια πίνακα για να μελετήσουν τη σύγκλιση του FP-Q. Κατέληξαν στο ότι ο αλγόριθμος θα συγκλίνει σχεδόν σίγουρα όταν ικανοποιούνται οι παρακάτω συνθήκες:

- Ο ρυθμός μάθησης α πρέπει να μειώνεται με την πάροδο του χρόνου. Συγκεκριμένα, αν α_t ο ρυθμός μάθησης τη χρονική στιγμή t, τότε πρέπει να ισχύει

$$\sum_{t=0}^{n} \alpha_t = \infty \quad (4.5) \qquad \text{και} \qquad \sum_{t=0}^{n} \alpha_t^2 < \infty \quad (4.6)$$

  όπου n το συνολικό πλήθος των επεισοδίων της διαδικασίας μάθησης.

- Κάθε πράκτορας εκτελεί απείρως συχνά κάθε δράση του σε κάθε κατάσταση, την οποία επισκέπτεται απείρως συχνά.

- Η πιθανότητα ένας πράκτορας i να αποφασίσει μία δράση α είναι πάντα μη μηδενική.

- Η στρατηγική κάθε πράκτορα είναι να αξιοποιεί τη γνώση του. Δηλαδή,

$$\lim_{t \to \infty} P_t^i(X_t) = 0 \quad (4.7)$$

  όπου X_t, μία τυχαία μεταβλητή που υποδηλώνει το γεγονός ότι ο πράκτορας i αποφάσισε μία μη βέλτιστη δράση, βάσει των εκτιμήσεών του τη χρονική στιγμή t.

Οι δύο πρώτες συνθήκες είναι απαραίτητες για τη σύγκλιση του απλού Q-Learning και η τρίτη, αν υλοποιηθεί κατάλληλα, θα διασφαλίσει τη δεύτερη (πχ, με χρήση κατανομής Boltzmann για την επιλογή δράσεων με κατάλληλα μειούμενη παράμετρο θερμοκρασίας T). Διασφαλίζει, επίσης, ότι οι πράκτορες δε θα υιοθετήσουν μία ντετερμενιστική στρατηγική εξερεύνησης. Τέλος, η τέταρτη συνθήκη διασφαλίζει ότι οι πράκτορες θα εκμεταλλεύονται τη γνώση τους (περισσότερες πληροφορίες για τη σύγκλιση αλγορίθμων από κοινού μάθησης μπορούν να βρεθούν στο [17]).

5 *Υλοποίηση και αξιολόγηση*

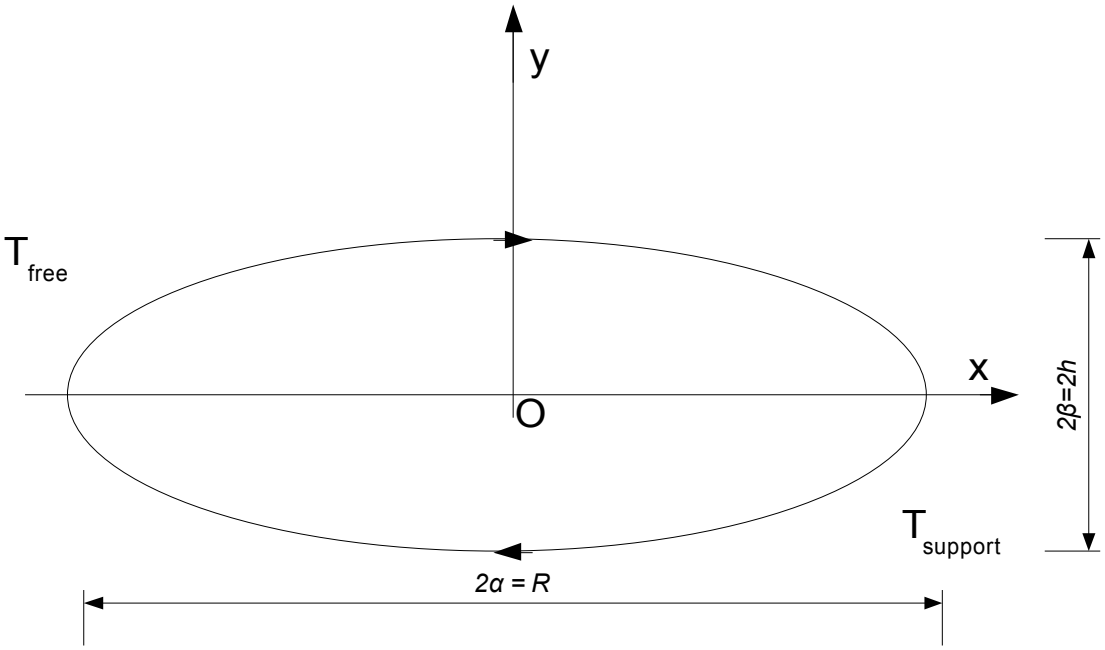
Στην εργασία μας προσπαθήσαμε να λύσουμε το πρόβλημα της μάθησης βάδισης σε ένα τετράποδο ρομπότ με δύο τρόπους. Ο πρώτος ήταν ο απλός : ένας πράκτορας καθόριζε την κίνηση των ποδιών και τη συνολική συμπεριφορά του συστήματος. Ο δεύτερος ήταν πιο πολύπλοκος : Μαζί με τον πράκτορα της πρώτης περίπτωσης, πλέον έχουμε και τέσσερις ακόμα, ένα για κάθε πόδι, οι οποίοι προσπαθούν να μάθουν κάποια τοπική συμπεριφορά του συστήματος.

Στο κεφάλαιο αυτό θα παρουσιάσουμε το σύστημα μάθησης που αναπτύξαμε, τις παραμέτρους του και τα αποτελέσματα του. Όλα τα πειράματα πραγματοποιήθηκαν στο περιβάλλον εξομοίωσης Webots™.

5.1 Εισαγωγή

5.1.1 Η καμπύλη κίνησης

Με τον όρο καμπύλη κίνησης αναφερόμαστε στην κίνηση του ποδιού του ρομπότ. Στο πείραμα μας χρησιμοποιήσαμε ως καμπύλη κίνησης μία έλλειψη, όπως μπορούμε να δούμε και στο σχήμα παρακάτω.



Σχήμα 5.1 : Καμπύλη κίνησης ποδιού.

Το μήκος κίνησης R ισούται με το μήκος του κύριου άξονα της έλλειψης, δηλαδή 2α και είναι παράμετρος προς μάθηση. Το ύψος h ισούται με το μισό του δευτερεύοντος άξονα, δηλαδή β και επελέγη ίσο με 0.02 m. Ο χρόνος της φάσης μετάβασης ισούται με T_{free} και ο χρόνος της φάσης υποστήριξης ισούται με $T_{support}$. Η συνολική διάρκεια βάδισης $T_{locomotion}$ επελέγη 3.072 s. Για τους χρόνους $T_{locomotion}$, T_{free} και $T_{support}$ ισχύουν

$$T_{locomotion} = T_{free} + T_{support} \quad (5.1)$$

$$D = \frac{T_{support}}{T_{locomotion}} \quad (5.2)$$

Η τιμή του συνολικού χρόνου βάδισης προκύπτει ως εξής: Επιλέγουμε το χρονικό βήμα (time step) της εξομοίωσης μας να ισούται με 32 ms. Επιλέγουμε επίσης να διακριτοποιήσουμε το χρόνο T_{free} σε 24 διαστήματα. Τέλος, επιλέξαμε το συνολικό χρόνο να ισούται με $4 \cdot T_{free}$ ώστε το σύστημα μας να έχει λόγο λειτουργίας D ίσο με 75%, όπως και η αργή στατική βάδιση που περιγράψαμε στο κεφάλαιο 2.

Οι παραμετρικές εξισώσεις της έλλειψης δίνονται από τις σχέσεις

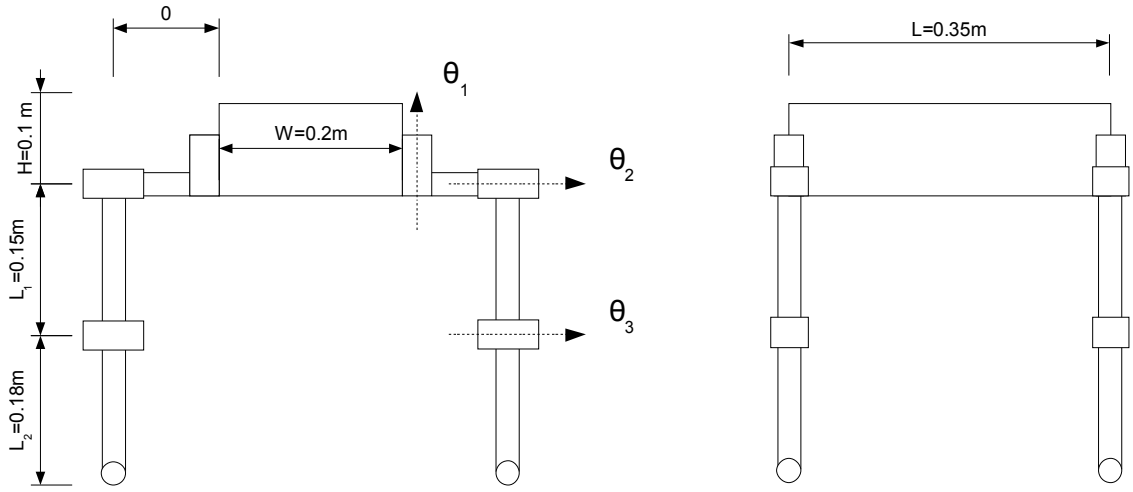
$$x(\theta) = \alpha cos\theta = \frac{R}{2}\cos(\theta)$$
$$y(\theta) = \alpha sin\theta = h\sin(\theta) = 0.02\sin(\theta) \qquad (5.3)$$

όπου $\theta \epsilon [0,2\pi]$.

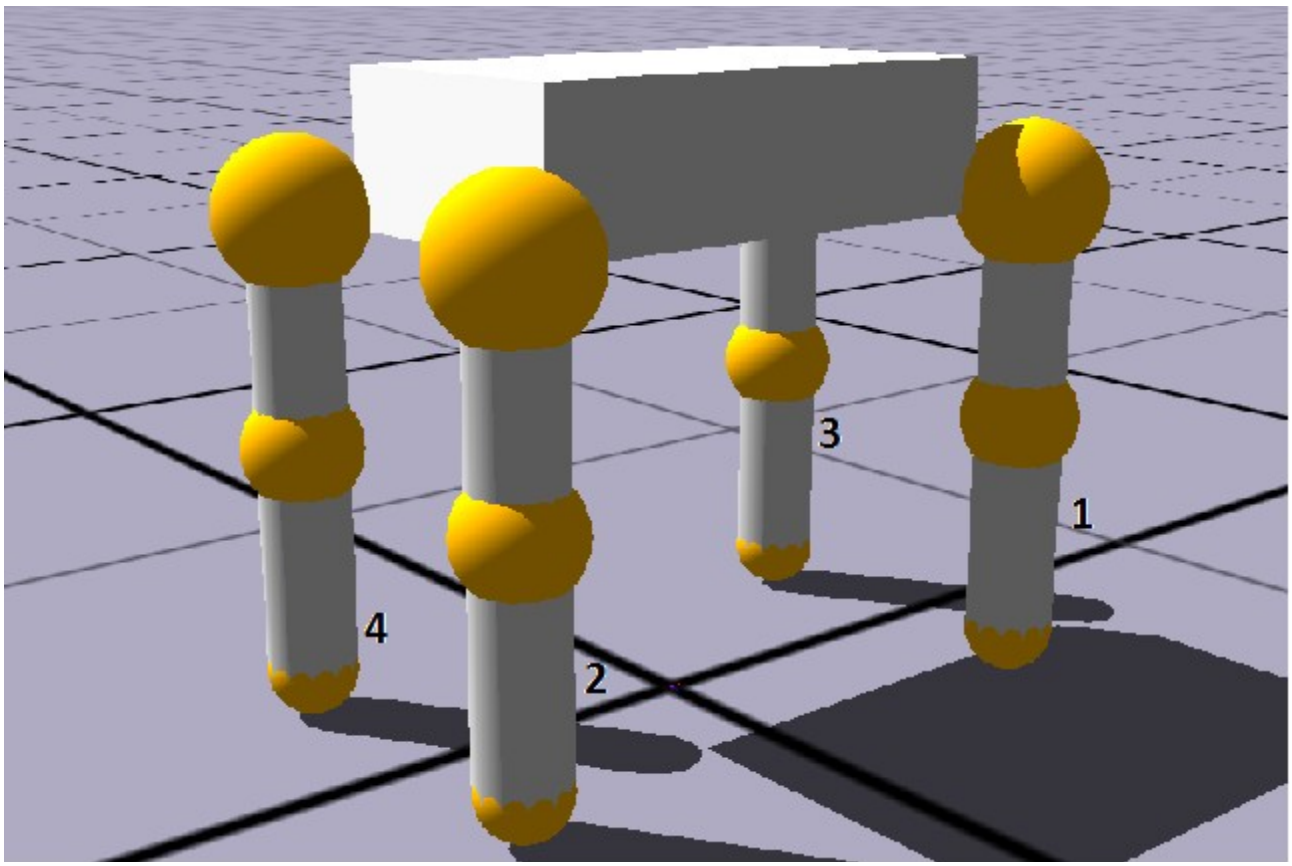
Προτιμήσαμε την έλλειψη ως προφίλ κίνησης, λόγω των λίγων παραμέτρων της. Επίσης, τα ελλειψοειδή προφίλ κίνησης έχουν προταθεί αρκετά στη βιβλιογραφία για πειράματα μάθησης βάδισης [32, 33, 34, 35].

5.1.2 Το ρομπότ

Το ρομπότ που χρησιμοποιήσαμε για την εργασία μας είναι ένα τετράποδο με τρεις βαθμούς ελευθερίας σε κάθε πόδι. Οι διαστάσεις και η δομή του φαίνονται στο σχήμα 5.2. Επίσης, στο σχήμα 5.3 φαίνεται μία εικόνα του από το περιβάλλον εξομοίωσης Webots™.



Σχήμα 5.2 : Το ρομπότ που χρησιμοποιήσαμε και οι διαστάσεις του. Αριστερά φαίνεται η πρόσοψη του ρομπότ και δεξιά η πλάγια όψη.



Σχήμα 5.3 : Εικόνα του ρομπότ από το περιβάλλον εξομοίωσης Webots™.Στο σχήμα φαίνεται και η αρίθμηση των ποδιών που θα χρησιμοποιηθεί αργότερα στα διαγράμματα βάδισης.

5.2 Μέθοδος ενός πράκτορα

Στην ενότητα αυτή θα περιγράψουμε τη μέθοδο του ενός πράκτορα. Συγκεκριμένα θα περιγράψουμε τον αλγόριθμο που χρησιμοποιήσαμε, τις παραμέτρους και το χώρο καταστάσεων – δράσεων. Τέλος, θα παραθέσουμε και θα σχολιάσουμε τα αποτελέσματα μας.

5.2.1 Χώρος καταστάσεων – δράσεων

5.2.1.1 Κατάσταση

Η κατάσταση στην οποία μπορεί να βρεθεί ο πράκτορας είναι η ταχύτητα v με την οποία κινήθηκε και το σχετικό σφάλμα Δv_{goal} σε σχέση με την επιθυμητή ταχύτητα v_{goal}. Επομένως,

$$s = \langle v, \Delta v_{goal} \rangle \quad (5.4)$$

Η διακριτοποίηση της παραμέτρου *v* είναι στο διάστημα [–0.1, 0.1] m/s με βήμα 0.01 m/s και της παραμέτρου Δv_{goal} στο διάστημα [– 100, 100] % με βήμα 5 %.

5.2.1.2 Δράσεις

Οι δράσεις του πράκτορα είναι τέσσερις : Η φάση του ποδιού φ_i για τα πόδια 2, 3 και 4 και το μήκος κίνησης ποδιού R. Θεωρούμε, ότι το πόδι 1 είναι το πόδι αναφοράς και η φάση του θεωρείται γνωστή και ίση με 0. Επομένως

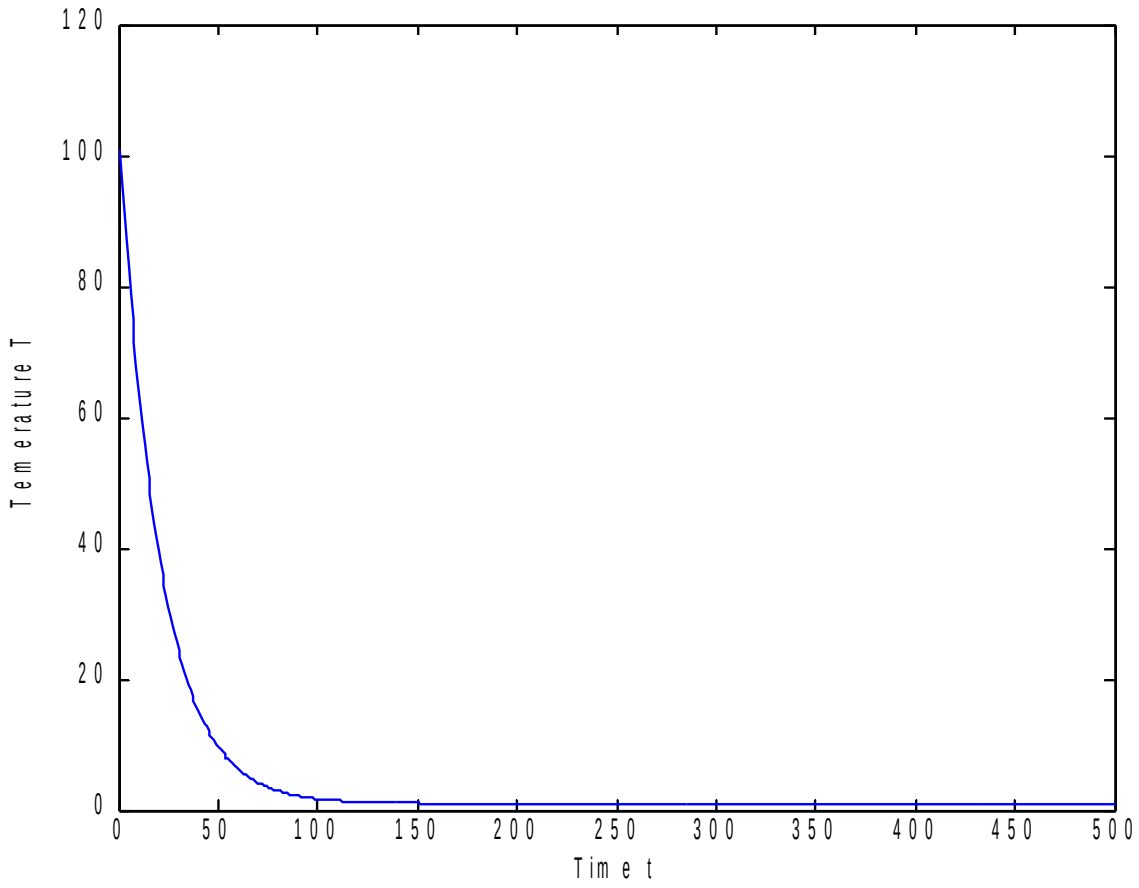
$$\alpha = \langle \varphi_{2,} \varphi_{3,} \varphi_{4,} R \rangle \quad (5.5)$$

Η διακριτοποίηση του φ_i είναι στο διάστημα [0, 3π/2] με βήμα π/4 και του *R* είναι στο διάστημα [0.15 m, 0.27 m] με βήμα 0.03 m.

Ως μέθοδο επιλογής δράσεως χρησιμοποιήσαμε την κατανομή Boltzmann. Η παράμετρος θερμοκρασίας *T* μειωνόταν σύμφωνα με τη σχέση

$$T(t) = 1 + T_{max} e^{-ct} \quad (5.6)$$

όπου T_{max} = 100 και c = 0.05. Οι τιμές αυτές επιλέχθηκαν εμπειρικά, έτσι ώστε όλες οι δράσεις να έχουν ίδια πιθανότητα επιλογής στις πρώτες εποχές και στη συνέχεια να επιλέγονται αυτές με τη μεγαλύτερη τιμή συνάρτησης αξίας δράσης. Επίσης, η σχέση 5.6 ελέγχθηκε έτσι ώστε να μην προκαλεί αριθμητική υπερχείλιση κατά την ύψωση του αποτελέσματος της διαίρεσης *Q(s,a) / T* στον αριθμό του Euler, στη σχέση 3.16. Παρακάτω φαίνεται η καμπύλη μείωσης της θερμοκρασίας συναρτήσει του χρόνου *t* για μία εποχή. Στη δική μας περίπτωση, ο χρόνος κάθε εποχής είναι 500 δοκιμές.



Σχήμα 5.4 : Καμπύλη μείωσης της παραμέτρου θερμοκρασίας.

Σε κάθε δοκιμή, μία γεννήτρια τυχαίων αριθμών μας επέστρεφε έναν πραγματικό αριθμό στο διάστημα [0, 1]. Αν ισχύει η σχέση

$$\text{τυχαίος αριθμός} \le \varepsilon \frac{T(t)-1}{T_{max}} \quad (5.7)$$

τότε η δράση επιλέγεται ομοιόμορφα τυχαία. Αλλιώς, κάθε δράση έχει πιθανότητα να επιλεγεί

$$Pr(\alpha_t = \alpha) = \frac{e^{Q_t(s,\alpha)/T}}{\sum_{\alpha' \in A(s)} e^{Q_t(s,\alpha')/T}} \quad (5.8)$$

5.2.2 Αλγόριθμος μάθησης

Ο αλγόριθμος μάθησης που χρησιμοποιήσαμε είναι ο Watkin's Q(λ). Οι παράμετροι του ήταν :

- Συνολικός χρόνος μάθησης : 200 εποχές με 500 δοκιμές για κάθε εποχή
- Ρυθμός μάθησης α = 0.1, ο οποίος μειωνόταν σε κάθε εποχή γραμμικά, με ρυθμό

$$\Delta\alpha = \frac{\alpha}{10 \cdot (\pi\lambda\acute{\eta}\theta o\varsigma\ \varepsilon\pi o\chi\acute{\omega}\nu) + 1} \quad (5.9)$$

- Παράμετρος λ = 0.45.
- Παράμετρος γ = 0.999
- Παράμετρος ε = 0.2 (για την ε-greedy)

Η παράμετρος $\Delta\alpha$ επελέγη έτσι ώστε στις πρώτες εποχές της διαδικασίας μάθησης να προσπαθούμε να προσεγγίσουμε τις τιμές της συνάρτησης αξίας δράσης Q, ενώ στις τελευταίες να εφαρμόζουμε μία άπληστη στρατηγική, βάσει των τιμών που ήδη έχουμε. Επομένως, θέλουμε $\alpha \rightarrow 0$, στις τελευταίες εποχές. Ο ρυθμός μείωσης που επιλέξαμε εξυπηρετεί τον παραπάνω σκοπό.

Να τονίσουμε ότι σε κάθε δοκιμή το ρομπότ κινούταν για τρεις περιόδους κίνησης. Σε κάθε μία μετρούσαμε την ταχύτητα του και στο τέλος της δοκιμής υπολογίζαμε το μέσο όρο των τριών ταχυτήτων. Ο μέσος όρος ήταν και η ταχύτητα v με την οποία δεχόμασταν ότι κινήθηκε το ρομπότ. Χρησιμοποιήσαμε συσσωρευτικά ίχνη επιλογής. Η συνάρτηση επιβράβευσης είναι η παρακάτω.

Αν *υπήρξε πτώση ή κινήθηκε ανάποδα* r = –2

Αν *(* $|v - v_{goal}| <= \Delta$ *)*

Αν *(* $|v' - v_{goal}| >= |v - v_{goal}|$ *)* r = exp(-w| $v - v_{goal}$ |)

Αν *(* $std <= \Sigma$ *)* r = r + 1

Αν *(* $|v' - v_{goal}| < |v - v_{goal}|$ *)* r = –1

Αλλιώς r = -2

όπου v' η ταχύτητα με την οποία κινήθηκε στην προηγούμενη δοκιμή το ρομπότ και *std* η τυπική απόκλιση των $|v' - v_{goal}|$ και $|v - v_{goal}|$.

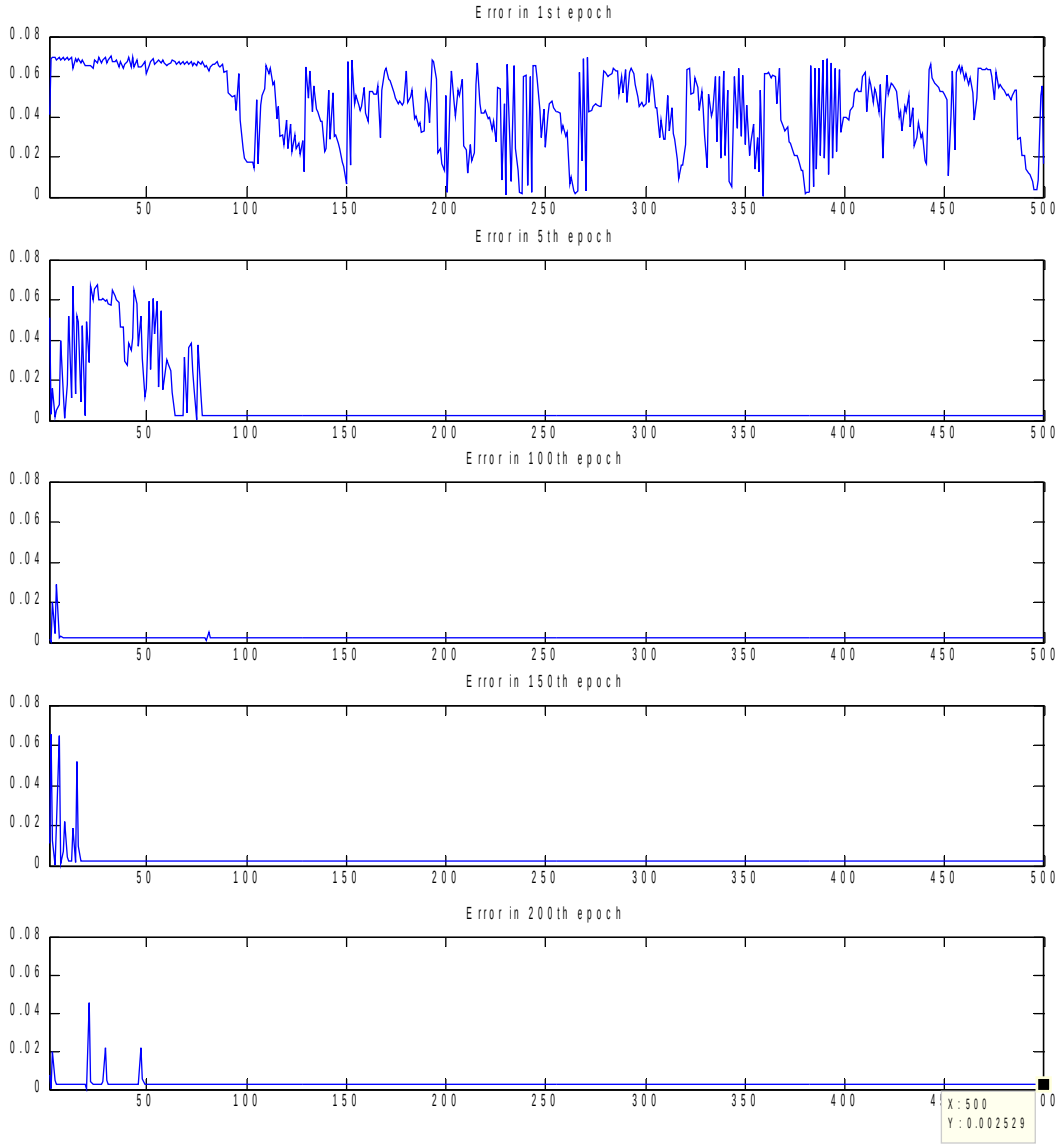
Δ είναι μία παράμετρος η οποία μειώνεται σε κάθε εποχή και Σ, w σταθερές. Η διαφορά $v - v_{goal}$ είναι σε m/s

Σχήμα 5.5 : Συνάρτηση επιβράβευσης για την περίπτωση ενός πράκτορα.

Συγκεκριμένα, όταν υπήρχε πτώση, το ρομπότ κινούταν προς τα πίσω ή η ταχύτητα του είχε διαφορά από την επιθυμητή μεγαλύτερη από Δ, τότε η επιβράβευση ήταν αρνητική και ίση με –2 . Αν η διαφορά από την επιθυμητή ήταν μικρότερη ή ίση με Δ, τότε έπρεπε να ελέγξουμε αν υπάρχει μείωση ή αύξηση της διαφοράς σε σχέση με την προηγούμενη. Αν υπήρχε αύξηση η επιβράβευση ήταν αρνητική και ίση με –1. Αν υπήρχε μείωση τότε η επιβράβευση προέκυπτε από την εκθετική συνάρτηση του σχήματος 5.5 και αν η τρέχουσα διαφορά με την προηγούμενη είχαν μικρή τυπική απόκλιση, τότε αυξάναμε την επιβράβευση που προέκυπτε από την εκθετική κατά 1.

5.2.3 Αποτελέσματα

Παρακάτω παρουσιάζουμε τα αποτελέσματα που πήραμε για την προσέγγιση του ενός πράκτορα. Σημειώνουμε ότι οι παράμετροι Δ και Σ της συνάρτησης επιβράβευσης ήταν 0.02 και 0.0001 αντίστοιχα. Ο ρυθμός μείωσης της παραμέτρου Δ είναι (0.02 – 0.01) / 200 = $5 \cdot 10^{-5}$. **Η επιθυμητή ταχύτητα είναι 0.07 m / s** και τέλος, ισχύει β = 75% για κάθε πόδι.

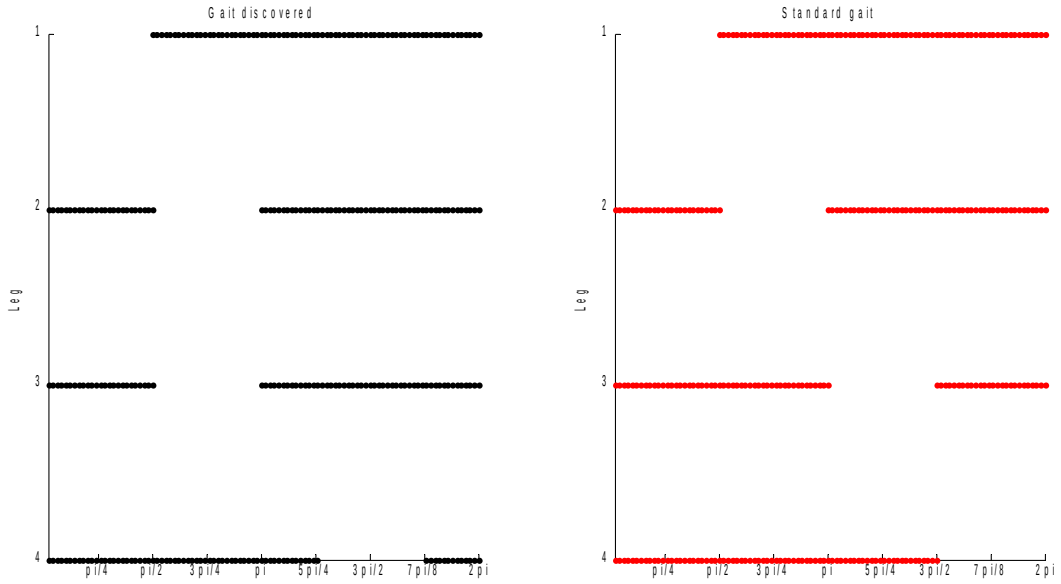


Στην αρχή παραθέτουμε το απόλυτο σφάλμα $| v - v_{goal} |$ για διάφορες εποχές της διαδικασίας μάθησης.*Σχήμα 5.6 : Απόλυτο σφάλμα για διάφορες εποχές της διαδικασίας μάθησης με επιθυμητή ταχύτητα* $v_{goal} = 0.07\ m/s$ *στην προσέγγιση ενός πράκτορα.*

Από το παραπάνω σχήμα παρατηρούμε ότι το σύστημα καταφέρνει και βρίσκει μία λύση στην 5η εποχή. Στην αρχή κάθε εποχής ενδέχεται να παρατηρήσουμε κάποια στοχαστική συμπεριφορά λόγω κάποιας τυχαίας επιλογής, όμως πλέον το σύστημα μπορεί να συγκλίνει αρκετά γρήγορα και να δώσει σχεδόν αμελητέο σφάλμα, $2.5 \cdot 10^{-3}$, χωρίς πτώση, όπως φαίνεται και στο σχήμα 5.6.

Τέλος, θα παραθέσουμε και το διάγραμμα βάδισης στο οποίο κατέληξε, σε σύγκριση με το

standard gait. Το standard gait παρατίθεται μαζί με τα αποτελέσματα μας, γιατί αποτελεί μία βέλτιστη λύση την οποία μπορεί να βρει το σύστημα μας, δεν αποτελεί όμως απόλυτο μέτρο σύγκρισης.



Σχήμα 5.7 : Το διάγραμμα βάδισης που ανακαλύφθηκε σε σύγκριση με το βέλτιστο standard gait για την προσέγγιση ενός πράκτορα.

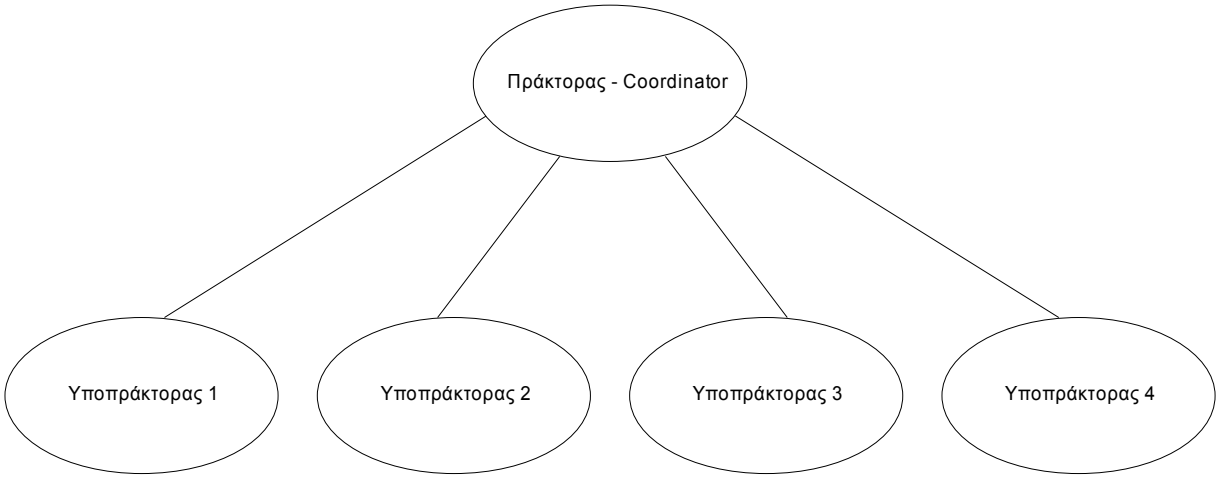
Το μήκος κίνησης ποδιού *R* στο οποίο κατέληξε το σύστημα είναι 0.15 m. Παρατηρούμε ότι το πόδι 2 βρίσκει τη βέλτιστη λύση του. Η βάδιση που ανακαλύφθηκε έχει ένα κρίσιμο διάστημα μεταξύ π/2 και π, όπου βρίσκονται μόνο δύο πόδια σε φάση υποστήριξης. Επίσης, είναι ιδιόμορφη, αφού στο διάστημα αυτό συμβαίνουν ταυτόχρονα τα γεγονότα ανύψωσης των ποδιών 2 και 3. Το σύστημα επομένως αποκτάει δυναμική συμπεριφορά στο παραπάνω διάστημα, όπου δεν υπάρχει πολύγωνο υποστήριξης. Βέβαια, υπάρχει ο περιορισμός ότι λόγω του σταθερού λόγου εργασίας β = 75%, το σύστημα δεν μπορεί να προσεγγίσει σε μεγάλο βαθμό δυναμικές βαδίσεις, όπως η ταχεία και ο καλπασμός. Όμως, καταφέρνει να προσδώσει στο ρομπότ την επιθυμητή ταχύτητα χωρίς πτώση.

5.3 Μέθοδος ενός πράκτορα και τεσσάρων υποπρακτόρων

Στην ενότητα αυτή θα περιγράψουμε τη δεύτερη μέθοδο η οποία προτείνεται στην παρούσα εργασία. Η μέθοδος αυτή βασίζεται σε μια ιεραρχική πολυπρακτορική δομή αποτελούμενη από

έναν πράκτορα-ρυθμιστή και τέσσερις υποπράκτορες (καθένας υποπράκτορας υπεύθυνος για τον έλεγχο κίνησης ενός ποδιού σε τοπικό επίπεδο). Η δομή αυτή βασίζεται στην ιεραρχική-εμφωλιασμένη (nested-hierarchical) πολυπρακτορική αρχιτεκτονική που έχει προταθεί από τους Γ. Καρίγιαννη και Κ. Τζαφέστα [36, 37]. Αυτή του ενός πράκτορα – ρυθμιστή, ο οποίος διευθύνει τέσσερις υποπράκτορες. Συγκεκριμένα θα περιγράψουμε τον αλγόριθμο που χρησιμοποιήσαμε, τις παραμέτρους και το χώρο καταστάσεων – δράσεων. Τέλος, θα παραθέσουμε και θα σχολιάσουμε τα αποτελέσματα μας.

Να τονίσουμε ότι, στον πράκτορα-ρυθμιστή δεν αλλάζει τίποτα σε σχέση με την πρώτη μέθοδο. Επίσης, όπως έχουμε ήδη περιγράψει, η πολυπρακτορική προσέγγιση συνιστά μία πολυεπίπεδη δομή. Ο πράκτορας (του πρώτου επιπέδου) βρίσκεται στο επίπεδο 1 και οι τέσσερις υποπράκτορες στο επίπεδο 2 (θεωρούμε ότι το επίπεδο 1 βρίσκεται υψηλότερα από το επίπεδο 2).



Σχήμα 5.8 : Η πολυπρακτορική προσέγγιση σχηματικά. Οι υποπράκτορες βρίσκονται στο επίπεδο 2 και ο πράκτορας – ρυθμιστής στο επίπεδο 1.

5.3.1 Χώρος καταστάσεων – δράσεων

5.3.1.1 Καταστάσεις

Η κατάσταση στην οποία μπορεί να βρεθεί ο υποπράκτορας είναι η ταχύτητα με την οποία κινείται το ρομπότ και το σχετικό σφάλμα ως προς την επιθυμητή ταχύτητα. Επομένως, η κατάσταση του υποπράκτορα είναι ίδια με την κατάσταση του πράκτορα – ρυθμιστή. Ο ρυθμιστής την προσδιορίζει και την προωθεί στους υποπράκτορες. Έτσι, έχουμε πάλι

$$s_{subagent} = s_{coordinator} = \langle v, \Delta v_{goal} \rangle \quad (5.7)$$

5.3.1.2 Δράσεις

Η δράση του υποπράκτορα είναι μία : ο λόγος λειτουργίας β_i του ποδιού στο οποίο αντιστοιχεί. Επομένως

$$\alpha_i = \langle \beta_i \rangle, i = 1,2,3,4 \quad (5.8)$$

Η διακριτοποίηση του β_i είναι στο διάστημα [1/2, 11/12] με βήμα 1/12.

Η μέθοδος επιλογής δράσης είναι ίδια ακριβώς με αυτήν του πράκτορα – ρυθμιστή. Ο αριθμός που επιστρέφεται από τη γεννήτρια τυχαίων αριθμών είναι κοινός και για τους πέντε πράκτορες.

5.3.2 Αλγόριθμος μάθησης

Ο αλγόριθμος μάθησης που χρησιμοποιήσαμε είναι ο FP-Q Learning. Οι παράμετροι του ήταν :

- Συνολικός χρόνος μάθησης : 200 εποχές με 500 δοκιμές για κάθε εποχή
- Ρυθμός μάθησης α = 0.1, ο οποίος μειωνόταν σε κάθε εποχή γραμμικά, με ρυθμό

$$\Delta\alpha = \frac{\alpha}{10 \cdot (\text{πλήθος εποχών}) + 1} \quad (5.6)$$

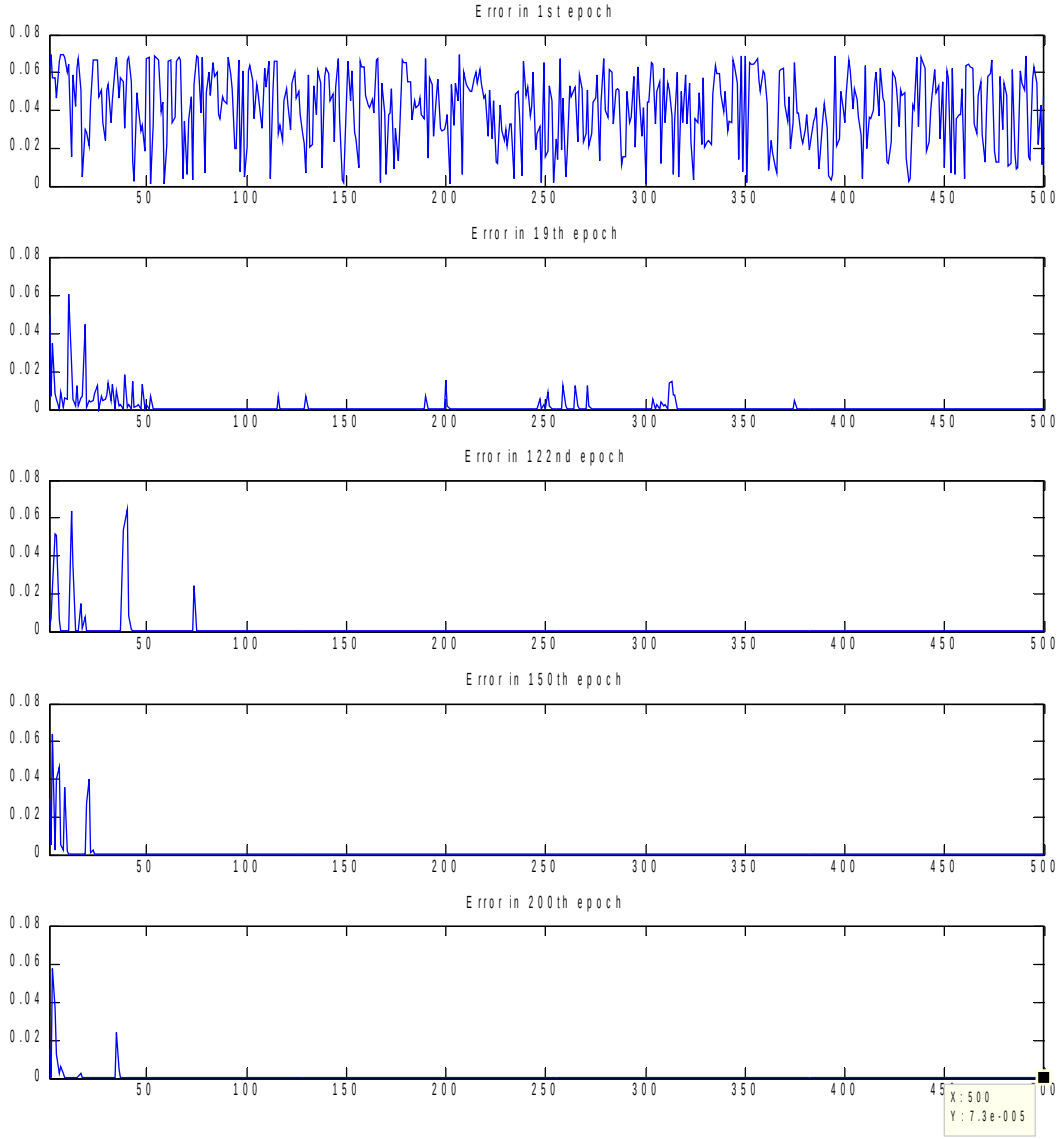
- Παράμετρος $\gamma = 0.999$
- Παράμετρος $\varepsilon = 0.2$ (για την ε-greedy)

Η συνάρτηση επιβράβευσης είναι ίδια με αυτή του πράκτορα – ρυθμιστή. Επομένως, και οι πέντε πράκτορες παίρνουν κοινή επιβράβευση.

5.3.3 Αποτελέσματα

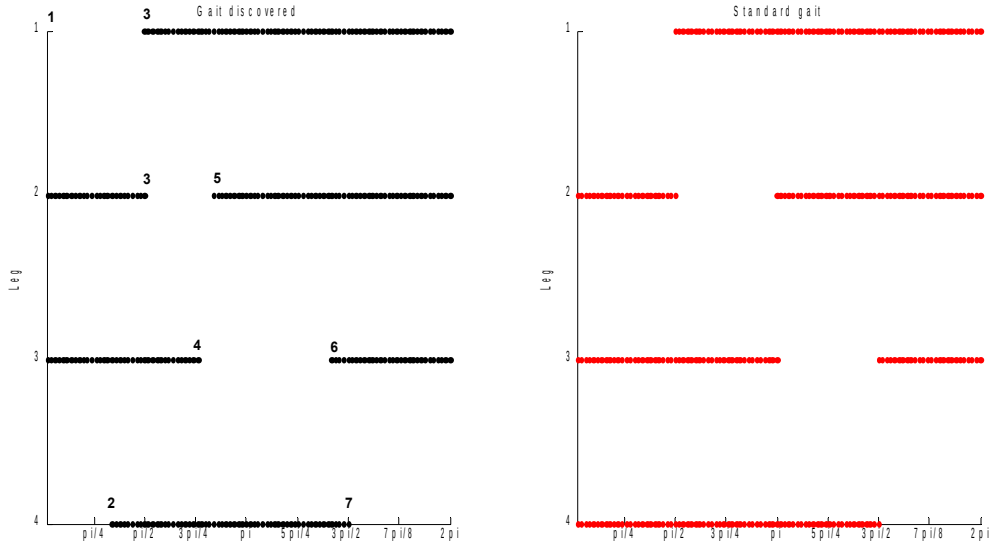
Παρακάτω παρουσιάζουμε τα αποτελέσματα που πήραμε. Σημειώνουμε ότι οι παράμετροι Δ και Σ της συνάρτησης επιβράβευσης ήταν 0.02 και 0.0001 αντίστοιχα. Ο ρυθμός μείωσης της παραμέτρου Δ είναι (0.02 – 0.01) / 200 = $5 \cdot 10^{-5}$. **Η επιθυμητή ταχύτητα είναι πάλι v_{goal} = 0.07 m/s**, ίδια με του πράκτορα – ρυθμιστή.

Σαν πρώτο αποτέλεσμα παραθέτουμε το σφάλμα σε διάφορες εποχές τις διαδικασίας μάθησης.



Σχήμα 5.9 : Απόλυτο σφάλμα για διάφορες εποχές της διαδικασίας μάθησης με επιθυμητή ταχύτητα $v_{goal} = 0.07\ m/s$ στην προσέγγιση ενός πράκτορα – ρυθμιστή και τεσσάρων υποπρακτόρων.

Στο πολυπρακτορικό σύστημα, η λύση βρίσκεται στην 19η εποχή, πιο αργά σε σύγκριση με την προσέγγιση ενός πράκτορα. Αυτό οφείλεται στη μεγαλύτερη πολυπλοκότητα του πολυπρακτορικού συστήματος και στο γεγονός ότι πρέπει να περάσει κάποιος χρόνος μάθησης, ώστε οι πράκτορες του δεύτερου επιπέδου – οι οποίοι μαθαίνουν από κοινού – να αρχίσουν να εκτιμούν σωστά τις δράσεις που θα λάβουν οι υπόλοιποι πράκτορες. Το σύστημα κατέληξε σε ένα αμελητέο σφάλμα, ίσο με $7.3 \cdot 10^{-5}$, μία πάρα πολύ καλή επίδοση, χωρίς να πέφτει. Επίσης, παραθέτουμε τη βάδιση στην οποία κατέληξε το σύστημα μας.



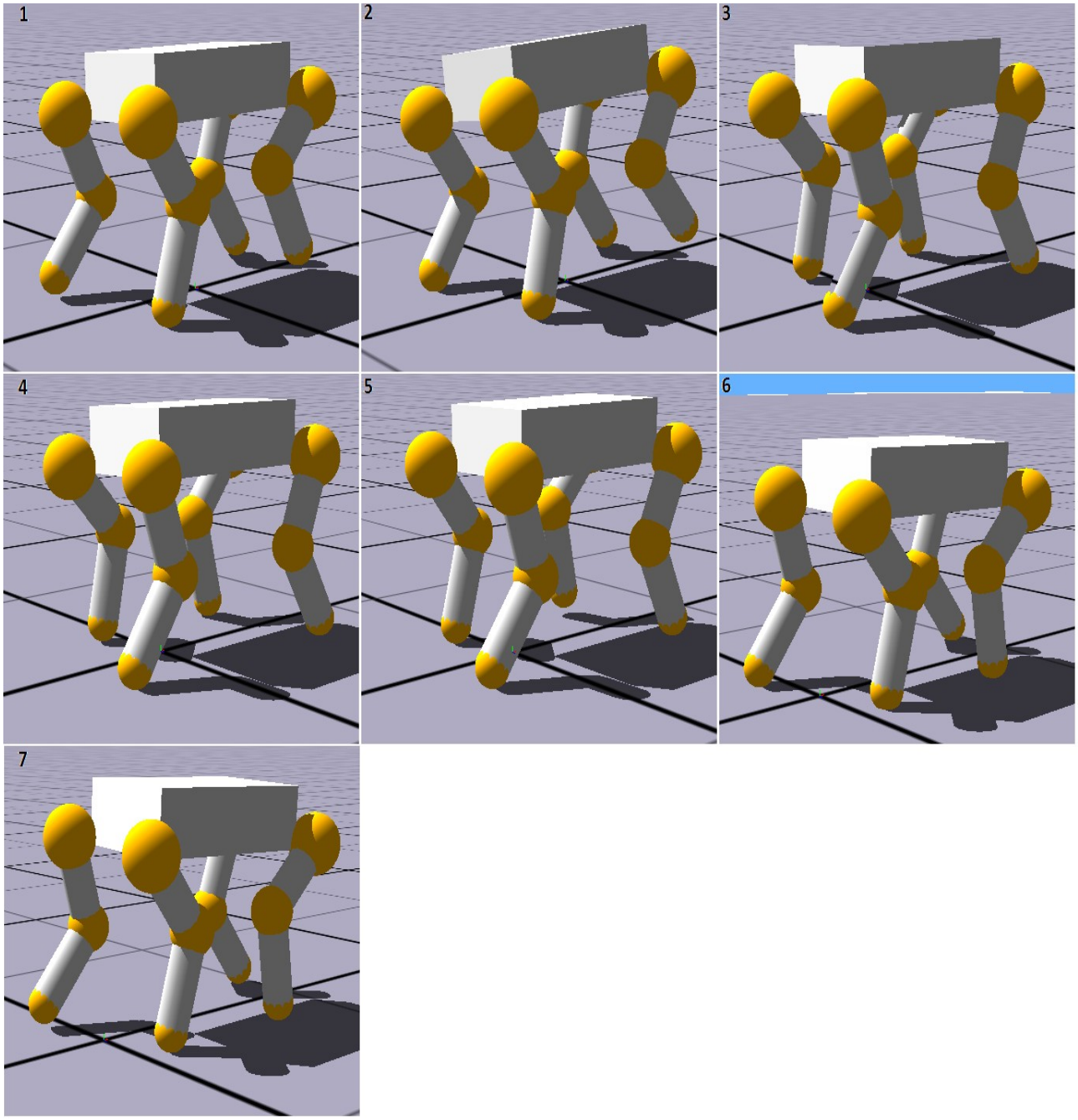
Σχήμα 5.10 : Το διάγραμμα βάδισης που ανακαλύφθηκε σε σύγκριση με το βέλτιστο standard gait για το πολυπρακτορικό σύστημα. Τα νούμερα στις παχιές γραμμές αντιστοιχούν στην αρίθμηση των στιγμιοτύπων του σχήματος 5.11 (παρακάτω). Ο αριθμός 3 εμφανίζεται δύο φορές για να υποδηλώσει ότι η τοποθέτηση του ποδιού 1 συμβαίνει ταυτόχρονα με την ανύψωση του ποδιού 2.

Όπως και στην προσέγγιση ενός πράκτορα, το σύστημα βρήκε μία λύση η οποία καταφέρνει να κινεί το ρομπότ με την επιθυμητή ταχύτητα με αμελητέο σφάλμα. Το κρίσιμο διάστημα, επίσης, είναι πολύ μικρό. Συγκεκριμένα, είναι από 0 έως 17π/50 , δηλαδή 17% της περιόδου. Σε αυτήν την προσέγγιση το ρομπότ κατάφερε να κινηθεί με πιο στατική συμπεριφορά κι επομένως με μεγαλύτερη ευστάθεια.
Τέλος, το μήκος κίνησης ποδιού *R*, στο οποίο συνέκλινε το σύστημα είναι 0.15 m.

Όπως παρατηρούμε στα σχήματα 5.6 και 5.9, το σφάλμα του πολυπρακτορικού συστήματος είναι μικρότερο από αυτό του συστήματος ενός πράκτορα. Αυτό εξαρχής ίσως να φαίνεται περίεργο, λόγω του ότι, ο χώρος αναζήτησης στην απλή προσέγγιση είναι μικρότερος σε σχέση με τη σύνθετη, αφού δε μαθαίνουμε τοπικές συμπεριφορές. Αυτή η διαφορά όμως, όπως φαίνεται, μας επιτρέπει να πάρουμε καλύτερα αποτελέσματα. Οι τοπικές συμπεριφορές, τις οποίες μαθαίνουν οι πράκτορες του δεύτερου επιπέδου, διορθώνουν κάποια σφάλματα κίνησης, τα οποία υπήρχαν στην

απλή προσέγγιση κι έτσι επιτρέπουν στο σύστημα να καταλήξει σε λύση με μεγαλύτερη ακρίβεια.

Τέλος, παραθέτουμε διάφορα σημαντικά στιγμιότυπα του συνολικού χρόνου κίνησης $T_{locomotion}$. Η αρίθμηση έχει γίνει κατά αντιστοιχία με τους αριθμούς στο διάγραμμα βάδισης του σχήματος 5.10.



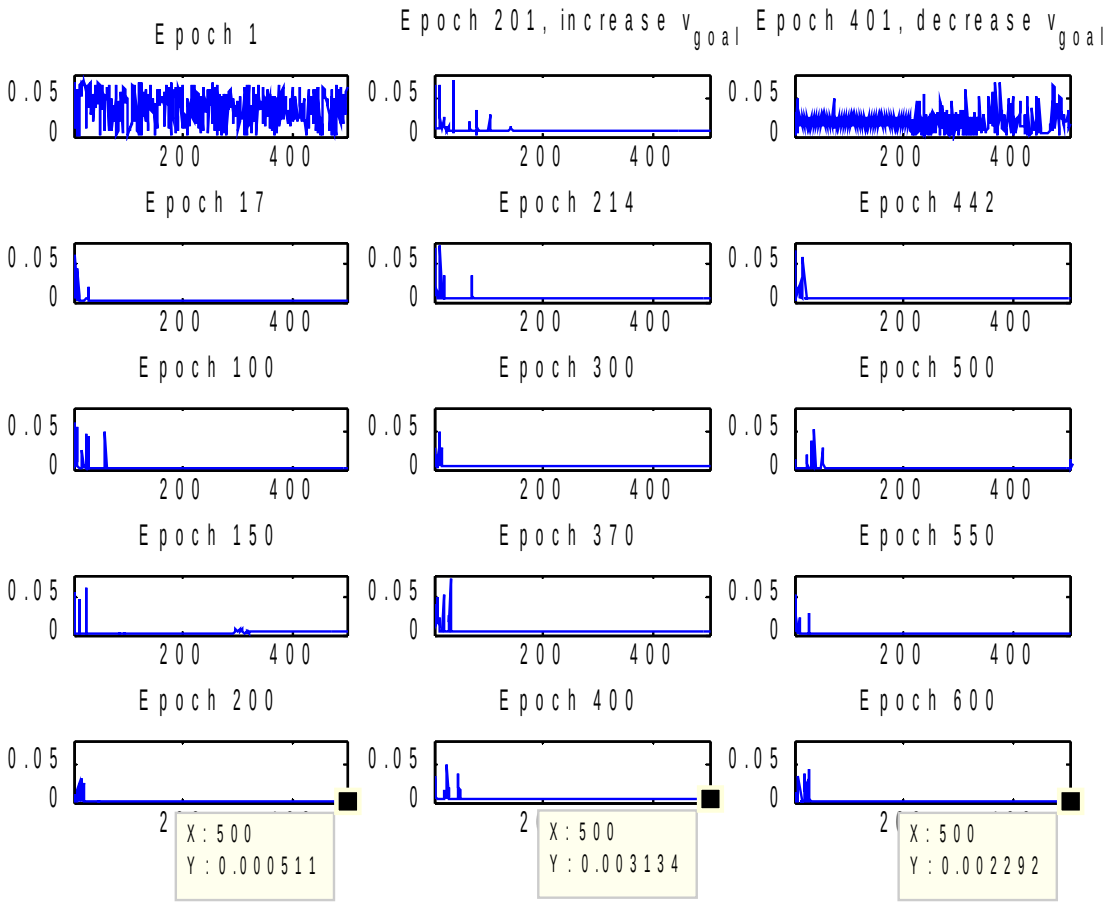
Σχήμα 5.11 : Στιγμιότυπα του συνολικού χρόνου κίνησης $T_{locomotion}$ από το περιβάλλον εξομοίωσης Webots™.

5.4 Μελέτη γενίκευσης

Στο σημείο αυτό θα περιγράψουμε τη μελέτη γενίκευσης που κάναμε για το σύστημα μας. Η γενίκευση συνίσταται στο να μπορούμε να μαθαίνουμε δράσεις για όλες τις καταστάσεις του χώρου καταστάσεων. Στην περίπτωση μας, συνίσταται στο να μπορεί το ρομπότ να κινηθεί και με άλλες επιθυμητές ταχύτητες.

Για να μελετήσουμε το βαθμό γενίκευσης του συστήματος μας, εκτελέσαμε το ακόλουθο πείραμα: Εκτελέσαμε την πολυπρακτορική προσέγγιση για 600 εποχές. Οι παράμετροι μάθησης ήταν ίδιοι με αυτούς που αναφέρονται στην ενότητα 5.4. Κάθε 200 εποχές, η επιθυμητή ταχύτητα άλλαζε και επαναρχικοποιούσαμε το ρυθμό μάθησης.

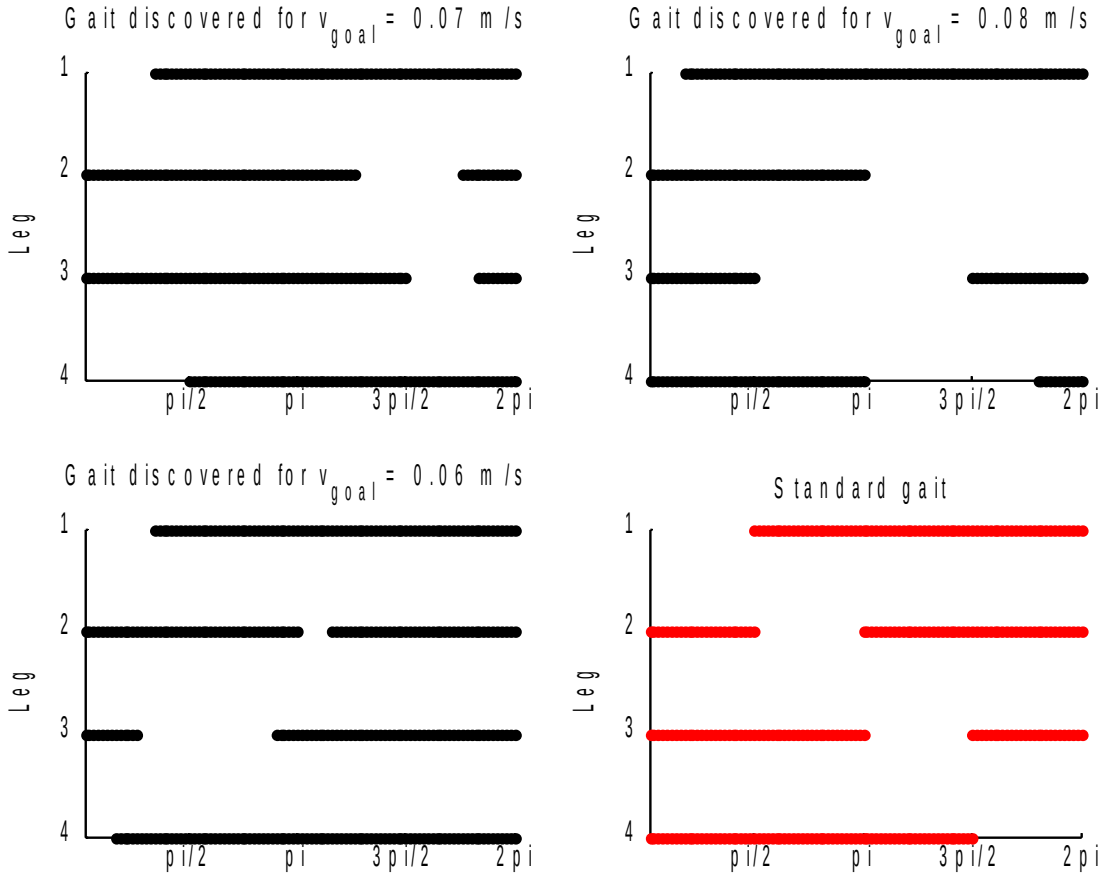
Πρώτα, παραθέτουμε το σφάλμα για διάφορες εποχές της διαδικασίας μάθησης.



Σχήμα 5.12 : Σφάλμα σε διάφορες εποχές της διαδικασίας μάθησης για τη μελέτη γενίκευσης.

Από το σχήμα 5.12 παρατηρούμε ότι το σύστημα αναπτύσσει μία προσαρμοστικότητα σε αλλαγές του στόχου του, δηλαδή καταφέρνει να μάθει τις δράσεις που θα κινήσουν το ρομπότ με την επιθυμητή ταχύτητα με μικρό σφάλμα, χωρίς πτώση.

Στη συνέχεια παραθέτουμε τη βάδιση στην οποία συνέκλινε το σύστημα για τις διάφορες επιθυμητές ταχύτητες.



Σχήμα 5.13 : Τα *διαγράμματα βάδισης που ανακαλύφθηκαν για κάθε επιθυμητή ταχυτητα σε σύγκριση με το βέλτιστο standard gait στη μελέτη γενίκευσης.*

Το μήκος κίνησης *R* στο οποίο συνέκλινε το σύστημα είναι 0.15 m για v_{goal} = 0.07 m/s και v_{goal} = 0.06 m/s και 0.21 m για v_{goal} = 0.08 m/s.

Παρατηρούμε ότι η πιο ασταθής βάδιση είναι αυτή που ανακαλύφθηκε για την ταχύτητα v_{goal} = 0.08 m/s. Συγκεκριμένα, έχει ένα κρίσιμο διάστημα από π έως 3π/2 όπου ένα πόδι βρίσκεται σε φάση υποστήριξης και από 3π/2 έως 9π/5, όπου βρίσκονται μόνο δύο πόδια σε φάσης υποστήριξης.

Συνολικά λοιπόν, το κρίσιμο διάστημα διαρκεί 40% της περιόδου. Για τις άλλες δύο παρατηρούμε καλύτερη συμπεριφορά. Για την ταχύτητα v_{goal} = 0.07 m/s υπάρχει ένα κρίσιμο διάστημα από 0 έως 17π/50 και από 3π/2 έως 44π/25, όπου μόνο δύο πόδια βρίσκονται σε φάση υποστήριξης. Συνολικά, το κρίσιμο διάστημα αντιστοιχεί σε 30% της περιόδου κίνησης. Όσον αφορά τη βάδιση για την ταχύτητα v_{goal} = 0.06 m/s, το κρίσιμο διάστημα είναι 0 έως 4π/25 και από 13π/50 έως 17π/50. Στα δύο αυτά διαστήματα έχουμε μόνο δύο πόδια σε φάση υποστήριξης. Εδώ, το ποσοστό του κρίσιμου διαστήματος ως προς την περίοδο είναι 12%.

Όπως φαίνεται από το σχήμα 5.13 και τα σχόλια της προηγούμενης παραγράφου, όσο μεγαλύτερη είναι η ταχύτητα που επιθυμούμε να μάθει το ρομπότ, τόσο πιο δυναμική γίνεται η βάδιση, αφού τα κρίσιμα διαστήματα αποκτούν μεγαλύτερη διάρκεια, κάτι το οποίο μπορεί να χαρακτηριστεί λογικό.

6 *Επίλογος*

6.1 Σύνοψη και συμπεράσματα

Η μάθηση βάδισης σε τετράποδα ρομπότ είναι ένα πολύ δύσκολο πρόβλημα – πιο δύσκολο από τη μάθηση σε εξάποδα – και η υπάρχουσα έρευνα έχει να επιδείξει διαφορετικές προσεγγίσεις και αλγορίθμους για τη λύση του. Κατά την έρευνα μας παρατηρήσαμε ότι οι πολυπρακτορικές προσεγγίσεις που προτείνονται, εμφανίζουν κάποιο βαθμό σύζευξης μεταξύ των πρακτόρων ή κάποιο βαθμό διαμοιρασμού πληροφορίας.

Το κύριο αποτέλεσμα της εργασίας μας είναι ότι η μάθηση βάδισης σε ένα τετράποδο ρομπότ μπορεί να επιτευχθεί με μία πολυπρακτορική, πλήρως κατανεμημένη on-line διαδικασία μάθησης. Επίσης, αποδείξαμε ότι το σύστημα μας επιδεικνύει μεγάλο βαθμό γενίκευσης, ζητώντας του να αλλάξει το στόχο του κατά τη μάθηση. Παρά την πολυπλοκότητα του συστήματος μας και της on-line μάθησης, το σφάλμα του συστήματος μας είναι αρκετά μικρό.

Αρχικά παρουσιάσαμε μία βιβλιογραφική μελέτη σχετική με την κίνηση τετράποδων ρομπότ, το πεδίο της Ενισχυτικής Μάθησης και την Από Κοινού Μάθησης. Πριν την εφαρμογή κάποιας προσαρμοστικής πολιτικής συγχρονισμού δράσεων για την κίνηση ενός τετράποδου ρομπότ, πρέπει να τροποποιήσουμε τις γενικές μεθόδους που παρουσιάστηκαν και να τις εξειδικεύσουμε για το σύστημα που θέλουμε να αναπτύξουμε. Κρίνεται λοιπόν αναγκαία η τροποποίηση κάποιων παραμέτρων όπως ο ρυθμός μάθησης των αλγορίθμων, αλλά και κάποιων συναρτήσεων όπως η

συνάρτηση επιβράβευσης.

Στη συνέχεια, παρουσιάσαμε τα αποτελέσματα της on-line διαδικασίας μάθησης που υολοποιήσαμε. Παρουσιάζονται τα αποτελέσματα της μεθόδου τόσο με έναν πράκτορα, όσο και με την προσθήκη τεσσάρων ακόμη, οι οποίοι δρουν σε τοπικό επίπεδο. Τέλος, παραθέσαμε τα αποτελέσματα της μελέτης γενίκευσης για το πολυπρακτορικό σύστημα.

Από αυτά προκύπτει ότι το σύστημα μας καταφέρνει να κινήσει το τετράποδο με μία επιθυμητή ταχύτητα με πολύ μικρό σφάλμα. Επίσης, έχει τη δυνατότητα να μεταβαίνει από μία ταχύτητα σε μία άλλη.

6.2 Μελλοντικές επεκτάσεις

Μία προφανής μελλοντική επέκταση της διπλωματικής μας εργασίας είναι η εφαρμογή του συστήματος μας σε ένα πραγματικό τετράποδο. Το παραπάνω παρουσιάζει μεγαλύτερες προκλήσεις καθώς τα δυναμικά φαινόμενα γίνονται εντονότερα και η εφαρμογή του συστήματος ίσως θα απαιτούσε τη χρήση κάποιου δυναμικού μοντέλου, το οποία θα τα λαμβάνει υπόψιν του. Το δυναμικό μοντέλο, επίσης, θα πρόσθετε κι άλλες παραμέτρους προς μάθηση.

Ιδιαίτερο ενδιαφέρον θα είχε και η χρήση του συστήματος που αναπτύξαμε για τη μάθηση περισσότερων – ή και διαφορετικών – παραμέτρων από αυτές που μάθαμε. Για παράδειγμα θα μπορούσαμε να ζητάμε από το σύστημα να μαθαίνει την μορφή του προφίλ κίνησης κάθε ποδιού (μέσω παραμετροποίησης), ώστε να κινείται με κάποια επιθυμητή ταχύτητα ή στάση (στην εργασία μας η μορφή της καμπύλης κίνησης θεωρήθηκε δεδομένη – έλλειψη).

Βιβλιογραφία

[1] Θεόδωρος Ρεκατσίνας, *Ανάπτυξη μιας Προσαρμοστικής Πολιτικής Αντικατάστασης Αρχείων με χρήση Ενισχυτικής Μάθησης*, Διπλωματική Εργασία, Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών ΕΜΠ, Ιούλιος 2010, artemis.ntua.ece/DT2010-0139.

[2] Svinin M., Yamada K., Ueda, K., Reinforcement learning approach to acquisition of stable gaits for locomotion robots, In Proc.: *1999 IEEE International Conference on Systems, Man, and Cybernetics. (*SMC '99), vol. 6, pp 936-941, 1999, doi:10.1109/ICSMC.1999.816678

[3] Bull, L., Fogarty, T.C., Mikami, S., & Thomas, J.G., Adaptive Gait Acquisition using Multi-agent Learning for Wall Climbing Robots, In *Automation and Robotics in Construction XII*, pp.80-86, 1995

[4] Pattie Maes, Rodney A. Brooks: Learning to Coordinate Behaviors. AAAI 1990: 796-802

[5] Murao H., Tamaki H., Kitamura S., Walking pattern acquisition for quadruped robot by using modular reinforcement learning, *2001 IEEE International Conference on Systems, Man, and Cybernetics*, vol. 3, pp.1402-1405, 2001, doi: 10.1109/ICSMC.2001.973478

[6] Zennir Y., Couturier P., Temps M.B., Distributed Reinforcement Learning of a Six-Legged Robot to Walk, In Proc.: *4th International Conference on Control and Automation,* pp.896-900, 12-12 June 2003, doi: 10.1109/ICCA.2003.1595152

[7] Siegwart Roland and Illah R. Nourbakhsh , *Introduction to autonomous mobile robots*, MIT Press Cambridge, 2004

[8] McGhee R., B. and Frank, A. A., On the stability properties of quadruped creeping gaits. *Mathematical Bioscience*, 3: 331-351, 1968

[9] González de Santos Pablo, Garcia Elena, Estremera Joaquin, *An Introduction to the Control of Four-legged Robots*, Springer, 2006

[10] D. E. Orin, R. B. McGhee, V. C. Jaswa, Interactive compute-control of a six-legged robot vehicle with optimization of stability, terrain adaptibility and energy, *1976 IEEE Conference on Decision and Control including the 15th Symposium on Adaptive Processes*, vol. 15, pp.382-391, Dec. 1976, doi: 10.1109/CDC.1976.267763

[11] Lin, B.-S., Song, S.-M., Dynamic modeling, stability and energy efficiency of a quadrupedal walking machine, In Proc.: *1993 IEEE International Conference on Robotics and Automation*, vol.3, pp.367-373, 2-6 May 1993, doi: 10.1109/ROBOT.1993.292201

[12] Yoneda K., Hirose S., Three dimensional stability criterion of intergratedmotion and manipulation, *Journal of Robotics ans Mechatronics*, 9(4), 267-274

[13] Zhou D., Low K.H., Zielinska, T., A stability analysis of walking robots based on leg-end supporting moments, In Proc.: *International Conference on Robotics and Automation, 2000, ICRA '00*, vol.3, pp.2834-2839, 2000, doi: 10.1109/ROBOT.2000.846457

[14] Papadopoulos, E.G., Rey D.A., A new measure of tipover stability margin for mobile manipulators, In Proc.: *1996 IEEE International Conference on Robotics and Automation*, vol.4, pp.3111-3116, 22-28 Apr 1996, doi: 10.1109/ROBOT.1996.509185

[15] Shin-Min Song, Kennth J. Waldron, Machines that walk: the adaptive suspension vehicle, MIT Press, 1989

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

[17] Caroline Claus and Craig Boutilier. 1998. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence* (AAAI '98/IAAI '98). American Association for Artificial Intelligence, Menlo Park, CA, USA, 746-752.

[18] D. Fundenberg and D. M. Kreps, Lectures on Learning and Equilibrium in Strategic Form Games, *CORE Foundation*, Louvain-La-Neuve, Belgium, 1992.

[19] G. W. Brown, Iterative Solution of Games by Fictitious Play, In T. C. Koopmans editor, *Activity Analysis of Production and Allocation*, Wiley, New York, 1951.

[20] G. Weiß. Learning to coordinate actions in multi-agent systems, In Proc.: *IJCAI-93*, pp.311–316, Chambery, FR, 1993.

[21] Y. Shoham and M. Tennenholtz. On the synthesis of useful social laws for artificial agent societies, In Proc.: *AAAI-92*, pp.276–281, San Jose, 1992.

[22] D. Fudenberg and D. K. Levine, Steady state learning and Nash equilibrium, *Econometrica*, 61(3):547–573, 1993.

[23] D. Fudenberg and D. M. Kreps. Lectures on Learning and Equilibrium in Strategic Form Games, *CORE Foundation*, Louvain-La-Neuve, Belgium, 1992.

[24] E. Kalai and E. Lehrer. Rational learning leads to Nash equilibrium. *Econometrica*, 61(5):1019–1045, 1993.

[25] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In Proc.: *11th Intl. Conf. on Machine Learning*, pp.157–163, New Brunswick, NJ, 1994.

[26] R. B. McGhee and G. I. Ishwandhi. Adaptive locomotion of a multilegged robot over rough terrain. *IEEE Trans. Systems, Man, and Cybernetics*, 9(4): 176–182, 1979, doi: 10.1109/TSMC.1979.4310180

[27] Chang-de Zhang and Shin-Min Song, Gaits and geometry of a walking chair for the disabled. *Journal of Terramechanics*, 26(3/4): 211-233, 1989, ISSN 0022-4898, doi: 10.1016/0022-4898(89)90037-2.

[28] Chang-de Zhang and Shin-Min Song, Stability analysis of wave-crab gaits of a quadruped. *Journal of Robotic Systems*, 7(2): 243-276, 1990.

[29] E. Garcia and P. Gonzalez de Santos, An improved energy stability margin for walking machines subject to dynamic effect, *Robotica*, 23(11): 13-20, 2005, doi: 10.1017/S0263574704000487

[30] D. Bertsekas, *Dynamic Programming and Optimal Control,* Athena Scientific, 2005

[31] Bellman, Richard, The theory of dynamic programming, *Bulletin of the American Mathematical Society*, 60: 503–516, 1954

[32] Jih-Gau Juang, Fuzzy neural network approaches for robotic gait synthesis, *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 30(4): 594-601, 2000

[33] Uchitane T., Hatanaka T., Uosaki K., Evolution strategies for biped locomotion learning using nonlinear oscillators, In Proc.: *SICE Annual Conference 2010*, pp.1458-1461, 2010, doi: 10.1109/3477.865178

[34] Jiaqi Zhang, Qijun Chen, Learning based gaits evolution for an AIBO dog, 2007 *IEEE Congress on Evolutionary Computation*, pp.1523-1526, 25-28 Sept. 2007, doi: 10.1109/CEC.2007.4424653

[35] Kohl N., Stone P., Policy gradient reinforcement learning for fast quadrupedal locomotion, In Proc.: *2004 IEEE International Conference on Robotics and Automation, ICRA '04*, vol.3, pp.2619- 2624, 26 April-1 May 2004, doi: 10.1109/ROBOT.2004.1307456

[36] Karigiannis J.N., Rekatsinas T.I., Tzafestas C.S, Fuzzy rule based neuro-dynamic programming for mobile robot skill acquisition on the basis of a nested multi-agent architecture, *2010 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp.312-319, 14-18 Dec. 2010, doi: 10.1109/ROBIO.2010.5723346

[37] John Karigiannis, Theodoros Rekatsinas, Costas S. Tzafestas, *Developmental Learning of Cooperative Robot Skills: A Hierarchical Multi-Agent Architecture*, Perception-reason-action cycle: Models, algorithms and systems, Springer USA, 2011, doi: 10.1007/978-1-4419-1452-1_16

[38] Hildebrand M., Vertebrate locomotion an introduction how does an animal's body move itself along?, *Bioscience* 39(39): 764–765, 1989

Αναφορές σε ιστοσελίδα

Sven Böttcher, Principles of robot locomotion, Seminar 'Human robot interaction', http://www2.cs.siu.edu/~hexmoor/classes/CS404-S09/RobotLocomotion.pdf

James Andrew Smith, Galloping, Bounding and Wheeled-LegModes of Locomotion on Underactuated Quadrupedal Robots, PhD Thesis, Department of Mechanical Engineering McGill University, November 2006
http://www.martinbuehler.net/theses/James_Smith_thesis.pdf

http://en.wikipedia.org/wiki/Gait

http://en.wikipedia.org/wiki/Horse_gait